Federated Variance-Reduced Stochastic Gradient Descent With Robustness to Byzantine Attacks

Zhaoxian Wu, Qing Ling, Tianyi Chen, and Georgios B. Giannakis, *Fellow, IEEE*

***Abstract*—This paper deals with distributed finite-sum optimization for learning over multiple workers in the presence of malicious Byzantine attacks. Most resilient approaches so far combine stochastic gradient descent (SGD) with different robust aggregation rules. However, the sizeable SGD-induced stochastic gradient noise challenges discerning malicious messages sent by the Byzantine attackers from noisy stochastic gradients sent by the 'honest' workers. This motivates reducing the variance of stochastic gradients as a means of robustifying SGD. To this end, a novel Byzantine attack resilient distributed (Byrd-) SAGA approach is introduced for federated learning tasks involving multiple workers. Rather than the mean employed by distributed SAGA, the novel Byrd-SAGA relies on the geometric median to aggregate the corrected stochastic gradients sent by the workers. When less than half of the workers are Byzantine attackers, Byrd-SAGA attains provably linear convergence to a neighborhood of the optimal solution, with the asymptotic learning error determined by the number of Byzantine workers. Numerical tests corroborate the robustness to various Byzantine attacks, as well as the merits of Byrd-SAGA over Byzantine attack resilient distributed SGD.**

***Index Terms*—Distributed finite-sum optimization, Byzantine attacks, gradient noise, variance reduction.**

I. Introduction

With the rapid development of information technologies, the volume of distributed data increases explosively. Every day, numerous distributed devices including sensors, cellphones, computers, and vehicles, generate huge amounts of data, which are often forwarded to datacenters for further processing and learning tasks. However, collecting data from distributed devices and storing them in datacenters raise major privacy concerns [1]–[3]. Accounting for these concerns, federated learning has been advocated to provide a privacy-preserving, decentralized data processing and machine learning framework [4], [5]. Data in federated learning are kept private, and local computations are carried at the distributed devices. Updates of local variables (such as stochastic gradients, corrected stochastic gradients, and model parameters) are found using per-device private data, while the datacenter aggregates local variables and disseminates the aggregated result to the distributed devices.

Even though privacy is preserved, the distributed nature of federated learning makes it vulnerable to errors and adversarial attacks. Devices can then become unreliable in either computing or communicating, or, they can even be hacked by adversaries. As a result, compromised devices may send malicious messages to the datacenter, thus misleading the learning process [6], [7]. We will henceforth focus on the class of malicious attacks known as Byzantine attacks [8]. Different from fixed or random attacks, adversarial devices inflicting what are henceforth termed Byantine attacks are adaptive in the sense that they can arbitrarily bias their outputs and strategically inject false information in the distributed system, by colluding with themselves [9]. Robustifying federated learning against Byzantine attacks is of paramount importance for secure processing and learning.

To cope with Byzantine attacks in federated learning, several robust aggregation rules have been developed in recent years, mainly towards improving the distributed stochastic gradient descent (SGD) solver of the underlying optimization task. Through aggregating stochastic gradients with the geometric median [10], [11], median [12], trimmed mean [13], or iterative filtering [14], stochastic algorithms have been able to tolerate a small number of devices attacked by Byzantine adversaries. Other aggregation rules include Krum [15] that selects a stochastic gradient having the minimal cumulative squared distance from a given number of nearest stochastic gradients, and robust stochastic aggregation (RSA) [16] which aggregates models other than stochastic gradients through penalizing the differences between the local and global model parameters. Related works also include adversarial learning in distributed principal component analysis [17], escaping from saddle points in non-convex distributed learning under Byzantine attacks [18], and leveraging redundant gradients to improve robustness [19], [20].

Although robust SGD iterates can ensure convergence to a neighborhood of the attack-free optimal solution, this neighborhood size can be large when Byzantine attacks are carefully crafted [21]. Essentially, SGD suffers from the sizeable approximation error (noise) associated with stochastic gradients. This

Manuscript received December 29, 2019; revised May 19, 2020; accepted July 20, 2020. Date of publication July 31, 2020; date of current version September 1, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vincent Gripon. The work of Qing Ling was supported in part by NSF China under Grants 61573331 and 61973324 and in part by Fundamental Research Funds for the Central Universities. The work of Georgios B. Giannakis was supported by NSF under Grants 1509040, 1508993, 1711471, and 1901134. This article was presented in part at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Barcelona, Spain, May 4–8, 2020. *(Corresponding author: Qing Ling.)*

Zhaoxian Wu and Qing Ling are with the School of Data and Computer Science and Guangdong Province Key Laboratory of Computational Science, Sun Yat-Sen University, Guangzhou 510006, China (e-mail: wuzhx23@mail2.sysu.edu.cn; lingqing556@mail.sysu.edu.cn).

Tianyi Chen is with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: chent18@rpi.edu).

Georgios B. Giannakis is with the Department of Electrical and Computer Engineering and Digital Technology Center, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: georgios@umn.edu).

Digital Object Identifier 10.1109/TSP.2020.3012952

leads to the challenge of distinguishing malicious messages sent by Byzantine attackers from the noisy stochastic gradients sent by 'honest' devices.

In the face of this challenge, we posed the following question: *Is it possible to better distinguish the malicious messages from the stochastic gradients through reducing the stochastic gradient-induced noise?* Our answer will turn out to be in the affirmative. Intuitively, if the stochastic gradient noise is small, the malicious messages should be easy to identify; see also the illustrative example in Section II-D. This intuition suggests combining variance reduction techniques with robust aggregation rules to handle Byzantine attacks in federated learning.

Existing variance reduction techniques in stochastic optimization include mini-batch [22], stochastic variance reduced gradient [23], stochastic dual coordinate descent [24], stochastic recursive gradient algorithm [25], stochastic average gradient algorithm (SAGA) [26], and many others [27]–[30]. Among them, we are particularly interested in SAGA, which has been proven effective in finite-sum optimization. SAGA can also be implemented in a distributed manner [31]–[33], and hence it fits well the federated learning applications, where each device deals with a finite number of data samples.

Our proposed novel Byzantine attack resilient distributed (Byrd-) SAGA combines SAGA's variance reduction with robust aggregation to deal with the malicious attacks in federated finite-sum optimization setups. Instead of the mean employed by distributed SAGA, the datacenter in Byrd-SAGA relies on the geometric median to aggregate the corrected stochastic gradients sent by distributed devices. Through reducing the stochastic gradient-induced noise, Byrd-SAGA turns out to outperform the Byzantine attack resilient distributed SGD. When less than half of the workers are Byzantine attackers, the robustness of geometric median to outliers enables Byrd-SAGA to achieve provably linear convergence to a neighborhood of the optimal solution, and the asymptotic learning error is solely determined by the number of Byzantine workers. Numerical tests demonstrate the robustness of Byrd-SAGA to various Byzantine attacks.

II. Problem Statement

We start this section by specifying the federated finite-sum optimization problem in the presence of Byzantine attacks. We then elaborate on the limitation of Byzantine attack resilient distributed SGD, which motivates our subsequent development of Byrd-SAGA.

A. Federated Finite-Sum Optimization in the Presence of Byzantine Attacks

Consider a network with one master node (datacenter) and W workers (devices), among which B workers are Byzantine attackers with their identities unknown to the master node. Let $\mathcal{W}$ be the set of all workers, and $\mathcal{B}$ that of Byzantine attackers with respective cardinalities $|\mathcal{W}| = W$ and $|\mathcal{B}| = B$. The data samples are evenly distributed across the honest workers $w \notin \mathcal{B}$. Each honest worker has J data samples, and $f_{w,j}(x)$ denotes the loss of the j-th data sample at the honest worker w with respect to the model parameter $x \in \mathbb{R}^p$. We are interested in the finite-sum optimization problem

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

where

$$f_w(x) := \frac{1}{J}\sum_{j=1}^{J} f_{w,j}(x). \tag{2}$$

The main challenge of solving (1) is that the Byzantine attackers can collude and send arbitrary malicious messages to the master node so as to bias the optimization process. We aspire to develop a robust distributed stochastic algorithm to address this issue. Intuitively, when a majority of workers are Byzantine attackers, it is difficult to obtain a reasonable approximate solution to (1). For this reason, we will assume $B < \frac{W}{2}$ throughout, and prove that the proposed Byzantine attack resilient algorithm is able to tolerate attacks from up to half of the workers.

B. Sensitivity of Distributed SGD to Byzantine Attacks

When all workers are honest, a popular solver of (1) is SGD [34]. At time slot (iteration) k, the master node broadcasts x^k to workers. Upon receiving x^k, worker w uniformly at random chooses a local data sample with index i_w^k to obtain the stochastic gradient $f'_{w,i_w^k}(x^k)$ that then communicates back to the master node. Upon collecting stochastic gradients from all workers, the master node updates the model as

$$x^{k+1} = x^k - \gamma^k \cdot \frac{1}{W}\sum_{w=1}^{W} f'_{w,i_w^k}(x^k) \tag{3}$$

where γ^k is the non-negative step size. Note that the distributed SGD can be extended to its mini-batch version; whereby, each worker uniformly at random chooses a mini-batch of data samples per iteration, and communicates the averaged stochastic gradient back to the master node.

While the honest workers send true stochastic gradients to the master node, the Byzantine ones can send arbitrary malicious messages to the master node in order to perturb the optimization process. Let $\tilde{m}_w^k$ denote the message worker w sends to the master node at slot k, given by

$$\tilde{m}_w^k = \begin{cases} f'_{w,i_w^k}(x^k), & w \notin \mathcal{B}, \\ *, & w \in \mathcal{B} \end{cases} \tag{4}$$

where $*$ denotes an arbitrary $p \times 1$ vector. Then, the distributed SGD update (3) becomes

$$x^{k+1} = x^k - \gamma^k \cdot \frac{1}{W}\sum_{w=1}^{W} \tilde{m}_w^k. \tag{5}$$

Even when only one Byzantine attacker is present, the distributed SGD may fail. Consider that a Byzantine attacker w_b sends to the master node $\tilde{m}_{w_b}^k = -\sum_{w\neq w_b} \tilde{m}_w^k$, which yields $x^{k+1} = x^k$. In practice, Byzantine attackers can send more sophisticated messages to fool the master node, and thus bias the optimization process.

C. Byzantine Attack Resilient Distributed SGD

Recent works often robustify the distributed SGD by incorporating robust aggregation rules when the master node receives messages from the workers. Here, we will adopt and analyze the geometric median, even though alternative robust aggregation rules are also viable [10], [11].

With $\mathcal{Z}$ denoting a subset in a normed space, the geometric median of $\mathcal{Z}$ is

$$\underset{z\in\mathcal{Z}}{\text{geomed}}\{z\} := \arg\min_{y} \sum_{z\in\mathcal{Z}} \|y - z\|. \tag{6}$$

Using (6), the distributed SGD in (5) can be modified to its Byzantine attack resilient form as

$$x^{k+1} = x^k - \gamma^k \cdot \underset{w\in\mathcal{W}}{\text{geomed}}\{\tilde{m}_w^k\}. \tag{7}$$

In essence, the geometric median chooses a reliable vector to represent the received messages $\{\tilde{m}_w^k, w \in \mathcal{W}\}$ through majority voting. When the number of Byzantine workers $B < \frac{W}{2}$, the geometric median approximates reasonably well the mean of $\{\tilde{m}_w^k, w \notin \mathcal{B}\}$. This property enables the Byzantine attack resilient distributed SGD to converge to a neighborhood of the optimal solution [10], [11].

D. Impact of Stochastic Gradient Noise on Robust Aggregation

In distributed SGD, the stochastic gradients evaluated by honest workers are noisy because of the randomness in choosing data samples. Due to the stochastic gradient noise however, it is not always easy to distinguish the malicious messages from the stochastic gradients using just the robust aggregation rules, e.g. the geometric median. Several existing works have recognized this issue. With carefully crafted Byzantine attacks, outputs of several Byzantine attack resilient SGD algorithms can be far away from the optimal solution [21]. In [11] and [19], the workers are divided into several groups, with averages taken within groups and the geometric median obtained across groups. This approach leads to reduced variance and thus enhanced ability to distinguish malicious messages. In [15], it is explicitly assumed that the ratio of the variance of stochastic gradients to the distance between iterate and optimal solution is upper-bounded.

Fig. 1 shows the impact of stochastic gradient noise on geometric median-based robust aggregation. When the stochastic gradients sent by honest workers have small variance, the gap between the true mean and the aggregated value is also small; that is, the same Byzantine attacks are less effective. We will quantify this statement in our analysis of Section IV-A.

Prompted by this observation, our key idea is to reduce the variance of stochastic gradients in order to enhance robustness to Byzantine attacks. In the Byzantine attack-free case, an effective approach to alleviating stochastic gradient noise in SGD is through variance reduction. By compensating for stochastic gradient noise, variance reduction techniques lead to faster convergence than SGD. For specificity, we will focus on SAGA, which reduces stochastic gradient noise for finite-sum optimization [26], and we will show how SAGA can also aid robust aggregation against Byzantine attacks.

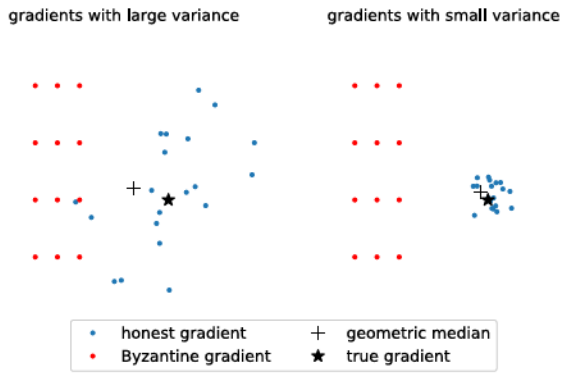


Fig. 1. Impact of stochastic gradient noise on geometric median-based robust aggregation. Blue dots denote stochastic gradients sent by the honest workers. Red dots denote malicious messages sent by the Byzantine workers. Plus signs denote the outputs of geometric median-based robust aggregation. Pentagrams denote the means of the stochastic gradients sent by the honest workers. Variance of the stochastic gradients from the honest workers is large in left and small in right.

III. Algorithm Development

In this section, we first introduce distributed SAGA with mean aggregation. Then, we propose Byrd-SAGA, which replaces mean aggregation by geometric median-based robust aggregation.

A. Distributed SAGA With Mean Aggregation

In distributed SAGA, each worker maintains a table of stochastic gradients for all of its local data samples [31], [32]. As in distributed SGD, the master node at slot k sends x^k to the workers, and every worker w uniformly at random chooses a local data sample with index i_w^k to find the stochastic gradient $f'_{w,i_w^k}(x^k)$. However, worker w does not send back $f'_{w,i_w^k}(x^k)$ to the master node. Instead, it corrects $f'_{w,i_w^k}(x^k)$ by first subtracting the previously stored stochastic gradient of the i_w^k-th data sample, and then adding the average of the stored stochastic gradients across local data samples. Then, worker w sends such a corrected stochastic gradient to the master node, and stores $f'_{w,i_w^k}(x^k)$ as the stochastic gradient of the i_w^k-th data sample in the table. After collecting the corrected stochastic gradients from all workers, the master node updates the model x^{k+1}.

To better describe distributed SAGA, let

$$\phi_{w,j}^{k+1} = \begin{cases} \phi_{w,j}^k, & j \neq i_w^k \\ x^k, & j = i_w^k \end{cases} \tag{8}$$

where $\phi_{w,j}^{k+1}$ is the iterate at which the most recent $f'_{w,j}$ is evaluated when slot k ends. Then, $f'_{w,j}(\phi_{w,j}^k)$ refers to the previously stored stochastic gradient of the j-th data sample prior to slot k on worker w, and

$$g_w^k := f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(\phi_{w,i_w^k}^k) + \frac{1}{J}\sum_{j=1}^{J} f'_{w,j}(\phi_{w,j}^k)$$

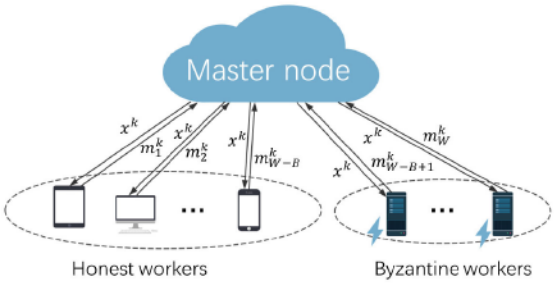


Fig. 2. Illustration of Byzantine attack resilient distributed SAGA. For the ease of illustration, the honest workers are from 1 to $W-B$ while the Byzantine attackers are from $W-B+1$ to W. But in practice, the identities of Byzantine attackers are unknown to the master node.

is the corrected stochastic gradient of worker w at slot k. The model update of SAGA is hence

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{W} \sum_{w=1}^{W} g_w^k \tag{9}$$

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

B. Distributed SAGA With Geometric Median Aggregation

Here, it is useful to recall that Byzantine workers may send to the master node malicious messages, other than the corrected stochastic gradients. To account for this, the message sent from worker w to the master node at slot k is expressed as

$$m_w^k = \begin{cases} g_w^k, & w \notin \mathcal{B}, \\ *, & w \in \mathcal{B} \end{cases} \tag{10}$$

where $*$ denotes an arbitrary $p \times 1$ vector. Similar to distributed SGD, distributed SAGA is also sensitive to Byzantine attacks. Our robust aggregation rule here is the geometric median. This leads to the proposed Byzantine attack resilient distributed (Byrd) form of SAGA in (9), that is given by

$$x^{k+1} = x^k - \gamma \cdot \underset{w \in \mathcal{W}}{\text{geomed}}\{m_w^k\}. \tag{11}$$

The proposed Byzantine attack resilient distributed SAGA, abbreviated as Byrd-SAGA, is listed step-by-step under Algorithm 1, and illustrated in Fig. 2. There are various implementations of the distributed SAGA. For example, [32] proposed to store the tables of stochastic gradients in the master node. The workers only need to upload the stochastic gradients and their indexes, while the master node performs the aggregation. This setup is also vulnerable to Byzantine attacks, since the Byzantine attackers may upload incorrect stochastic gradients. The proposed robust aggregation rule can also be applied therein.

Robust aggregations other than the geometric median are available, including the median [12], Krum [15], marginal trimmed mean [13], and iterative filtering [14]. In the median for instance, the aggregation outputs the element-wise median of $\{m_w^k, w \in \mathcal{W}\}$; while in the Krum, the aggregation outputs

$$\underset{w \in \mathcal{W}}{\text{Krum}}\{m_w^k\} = m_{w^*}, \; w^* = \arg\min_{w \in \mathcal{W}} \sum_{w \to w'} \|m_w^k - m_{w'}^k\|^2$$

Algorithm 1: Byzantine Attack Resilient Distributed SAGA.

Require: step size γ; number of workers W; number of data samples J on every honest worker w

Master node and honest workers initialize x^0

for all honest worker w **do**

  for $j \in \{1, \ldots, J\}$ **do**

    Initializes gradient storage $f'_{w,j}(\phi_{w,j}) = f'_{w,j}(x^0)$

  end for

  Initializes average gradient $\bar{g}_w^1 = \frac{1}{J}\sum_{j=1}^{J} f'_{w,j}(x^0)$

  Sends $\bar{g}_w^1$ to master node

end for

Master node updates $x^1 = x^0 - \gamma \cdot \text{geomed}_{w \in \mathcal{W}}\{\bar{g}_w^1\}$

for all $k = 1, 2, \cdots$ **do**

  Master node broadcasts x^k to all workers

  for all honest worker node w **do**

    Samples i_w^k from $\{1, \ldots, J\}$ uniformly at random

    Updates $m_w^k = f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(\phi_{w,i_w^k}) + \bar{g}_w^k$

    Sends m_w^k to master node

    Updates $\bar{g}_w^{k+1} = \bar{g}_w^k + \frac{1}{J}(f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(\phi_{w,i_w^k}))$

    Stores gradient $f'_{w,i_w^k}(\phi_{w,i_w^k}) = f'_{w,i_w^k}(x^k)$

  end for

  Master node updates $x^{k+1} = x^k - \gamma \cdot \text{geomed}_{w \in \mathcal{W}}\{m_w^k\}$

end for

where $w \to w'$ $(w \neq w')$ selects the indexes w' of the $W - B - 2$ nearest neighbors of m_w^k in $\{m_{w'}^k, w' \in \mathcal{W}\}$. Note that Krum needs to know B, the number of Byzantine attackers, in advance. In addition, other variance reduction techniques, such as mini-batch [22], are also available to alleviate the gradient noise. Here we opted for the combination of geometric median and SAGA. Extending the current work to other robust aggregation rules and variance reduction techniques, is in our future research agenda.

Remark 1: Computing the geometric median involves solving an optimization problem in the form of (6). Since it is costly to obtain the exact geometric median, one is typically satisfied with an ϵ-approximate value [35]. We say that z_ϵ^* is an ϵ-approximate geometric median of $\mathcal{Z}$ if

$$\sum_{z \in \mathcal{Z}} \|z_\epsilon^* - z\| \leq \inf_y \sum_{z \in \mathcal{Z}} \|y - z\| + \epsilon. \tag{12}$$

We shall show that the ϵ-approximation only slightly affects the convergence of Byrd-SAGA.

Remark 2: Every worker in Byrd-SAGA stores J stochastic gradients, where J is the number of local data samples. For this reason, Byrd-SAGA fits setups where workers have enough memory resources, such as federated learning among financial institutions, hospitals, or the Internet of Vehicles [4], [5]. In addition, our future work will pursue means of reducing Byrd-SAGA's storage requirements, as well as the communication overhead along the lines of [36] and [37].

IV. Theoretical Analysis

In this section, we theoretically justify the intuitive idea that reducing stochastic gradient noise helps identify malicious

messages in robust aggregation, specifically to the geometric median in this paper. We prove that our Byrd-SAGA converges to a neighborhood of the optimal solution at a linear rate under Byzantine attacks, and the asymptotic learning error is determined by the number of Byzantine attackers.

A. Importance of Reducing Stochastic Gradient Noise

Here, we quantify the role of stochastic gradient noise on the geometric median aggregation. Towards this objective, consider the set of messages $\mathcal{Z}$ sent by all workers in $\mathcal{W}$, and the set $\mathcal{Z}'$ of malicious messages sent by the Byzantine attackers in $\mathcal{B}$. Further, let $\bar{z}$ denote the true gradient given by the ensemble average of stochastic gradients. Using these definitions, the ensuing lemma bounds the mean-square error of the geometric median relative to the true gradient.

Lemma 1. (Concentration property): Let $\mathcal{Z}$ be a subset of random vectors distributed in a normed vector space. If $\mathcal{Z}' \subseteq \mathcal{Z}$ and $|\mathcal{Z}'| < \frac{|\mathcal{Z}|}{2}$, then it holds that

$$E\|\underset{z\in\mathcal{Z}}{\text{geomed}}\{z\} - \bar{z}\|^2 \leq 2C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'} E\|z - Ez\|^2}{|\mathcal{Z}| - |\mathcal{Z}'|} + 2C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'} \|Ez - \bar{z}\|^2}{|\mathcal{Z}| - |\mathcal{Z}'|} \quad (13)$$

where

$$\bar{z} := \frac{\sum_{z\notin\mathcal{Z}'} Ez}{|\mathcal{Z}| - |\mathcal{Z}'|}$$

while $C_\alpha := \frac{2-2\alpha}{1-2\alpha}$, and $\alpha := \frac{|\mathcal{Z}'|}{|\mathcal{Z}|}$.

The left-hand side of (13) is the mean-square error of the geometric median relative to the true gradient, while the right-hand side is the sum of two terms. The first is determined by the variances of the local stochastic gradients sent by the honest workers (inner variation), while the second term is determined by the variations of the local gradients at the honest workers with respect to the true gradient (outer variation). In the Byzantine attack resilient SGD, the upper bound can be large due to the large stochastic gradient noise of SGD. Through reducing the stochastic gradient noise in terms of either inner variation or outer variation, we are able to attain improved accuracy under malicious attacks.

B. Convergence of Byrd-SAGA and Comparison With Byzantine Attack Resilient SGD

Here, we establish convergence of Byrd-SAGA, and theoretically justify that, through reducing the impact of inner variation, Byrd-SAGA enjoys superior robustness to Byzantine attacks. We begin with several needed assumptions on the functions $\{f_{w,j}, w \notin \mathcal{B}\}$.

Assumption 1. (Strong convexity and Lipschitz continuity of gradients): The function f is μ-strongly convex and has L-Lipschitz continuous gradients, which amounts to requiring that for any $x, y \in \mathbb{R}^p$, it holds that

$$f(x) \geq f(y) + \langle f'(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2 \quad (14)$$

and

$$\|f'(x) - f'(y)\| \leq L\|x - y\|. \quad (15)$$

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

$$\frac{1}{W - B}\sum_{w\notin\mathcal{B}} \|f'_w(x) - f'(x)\|^2 \leq \delta^2. \quad (16)$$

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

$$E_{i_w^k}\|f'_{w,i_w^k}(x) - f'_w(x)\|^2 \leq \sigma^2, \quad \forall w \notin B. \quad (17)$$

Assumption 1 is standard in convex analysis. Assumptions 2 and 3 bound the variation of gradients and the variation of stochastic gradients within the honest workers, respectively [38]. For instance, most of the existing Byzantine attack resilient SGD algorihtms assume that the stochastic gradients at the honest workers are independently and identically distributed (i.i.d.) with finite variance, such that the outer variation δ^2 in Assumption 2 is proportional to $1/J$ and the inner variation σ^2 in Assumption 3 is finite. In the analysis of Byzantine attack resilient SGD, *both* outer and inner variations must be bounded. Interestingly, inner variation will turn out not to impact Byrd-SAGA, and Assumption 3 will no longer be necessary in its analysis.

The presence of geometric median makes Byrd-SAGA analysis challenging. Specifically, for every honest worker $w \notin \mathcal{B}$, m_w^k is an unbiased estimate of $f'_w(x^k)$, meaning

$$E_{i_w^k}[m_w^k] = f'_w(x^k). \quad (18)$$

Averaging (18) over all honest workers $w \notin \mathcal{B}$, we have

$$\frac{1}{W - B}\sum_{w\notin\mathcal{B}} E_{i_w^k}[m_w^k] = \frac{1}{W - B}\sum_{w\notin\mathcal{B}} f'_w(x^k) = f'(x^k). \quad (19)$$

From (19), we observe that the mean of $\{m_w^k, w \notin \mathcal{B}\}$ is an unbiased estimate of $f'(x^k)$. Nevertheless, the geometric median of $\{m_w^k, w \notin \mathcal{B}\}$, even only over all the honest workers and calculated accurately, is a biased estimate of $f'(x^k)$. This is the main challenge in adapting the proof of SAGA to that of Byrd-SAGA. Note that [39] also encounters the gradient estimate bias, due to random shuffling in SAGA. However, the technique used in [39] is unable to handle the gradient estimate bias caused by geometric median here.

To simplify notation, we will henceforth use E to represent the expectation with respect to all random variables i_w^k.

The following theorem asserts that Byrd-SAGA converges to a neighborhood of the optimal solution x^* at a linear rate, with the asymptotic learning error determined by the number of Byzantine attackers.

Theorem 1: Under Assumptions 1 and 2, if the number of Byzantine attackers satisfies $B < \frac{W}{2}$ and the step size satisfies

$$\gamma \leq \frac{\mu}{8J^2C_\alpha L^2}$$

then for Byrd-SAGA with ϵ-approximate geometric median aggregation, it holds that

$$E\|x^k - x^*\|^2 \leq \left(1 - \frac{\gamma\mu}{2}\right)^k \Delta_1 + \Delta_2 \tag{20}$$

where

$$\Delta_1 := \|x^0 - x^*\|^2 - \Delta_2 \tag{21}$$

$$\Delta_2 := \frac{4}{\mu^2}\left(1 + 32J^3C_\alpha^2\gamma^2L^2\right)\left(4C_\alpha^2\delta^2 + \frac{2\epsilon^2}{(W-2B)^2}\right). \tag{22}$$

In (20), the constant of convergence rate is given by

$$1 - \frac{\gamma\mu}{2} \geq 1 - \frac{1}{16J^2C_\alpha\frac{L^2}{\mu^2}}$$

which is close to 1 when J (the number of data samples at each worker) and $\frac{L}{\mu}$ (the condition number of functions) are large. Observe that C_α is monotonically increasing when the portion of Byzantine attackers α increases. Therefore, (20) shows that Byrd-SAGA converges slower as the number of Byzantine attackers grows. Correspondingly, the theoretical upper bound of step size γ is small when J and C_α are large. The asymptotic learning error Δ_2 in (22) is also monotonically increasing when C_α (and hence the number of Byzantine attackers) increases.

To demonstrate the superior robustness of Byrd-SAGA, we also establish the convergence of Byzantine attack resilient SGD with constant step size as a benchmark. As in Theorem 1, the convergence of Byzantine attack resilient SGD is in the mean-square error sense. This is different from [11], where convergence is asserted in the high probability sense.

Theorem 2: Under Assumptions 1, 2 and 3, if the number of Byzantine attackers is $B < \frac{W}{2}$ and the step size satisfies

$$\gamma \leq \frac{\mu}{2L^2}$$

then for Byzantine attack resilient SGD with ϵ-approximate geometric median aggregation, it holds that

$$E\|x^k - x^*\|^2 \leq (1-\gamma\mu)^k\Delta_1' + \Delta_2' \tag{23}$$

where

$$\Delta_1' := \|x^0 - x^*\|^2 - \Delta_2' \tag{24}$$

$$\Delta_2' := \frac{2}{\mu^2}\left(4C_\alpha^2\sigma^2 + 4C_\alpha^2\delta^2 + \frac{2\epsilon^2}{(W-2B)^2}\right). \tag{25}$$

Let us ignore the approximation error in computing geometric median by setting $\epsilon = 0$, and compare the two asymptotic learning errors Δ_2 and Δ_2'. With the step size $\gamma \leq \mu/(8J^2C_\alpha L^2)$, the constant $1 + 32J^3C_\alpha^2\gamma^2L^2$ in Δ_2 is in the order of $O(1)$. Therefore, we deduce that

$$\Delta_2 = O\left(\frac{C_\alpha^2}{\mu^2}\delta^2\right) \quad \text{and} \quad \Delta_2' = O\left(\frac{C_\alpha^2}{\mu^2}(\sigma^2+\delta^2)\right).$$

Observe that Δ_2', the asymptotic learning error of Byzantine attack resilient SGD, is proportional to the sum of inner and outer variations. With all honest workers having a same data sample, we have $\sigma^2 = \delta^2 = 0$. In this case, the asymptotic learning error Δ_2' vanishes because the geometric median aggregation takes effect and attains the true gradient. However, when all honest workers share the same set of different data samples, the inner variation σ^2 is no longer zero and the asymptotic learning error Δ_2' can be large. In contrast, Byrd-SAGA effectively reduces the impact of inner variation, and is able to achieve smaller learning error.

V. Numerical Experiments

Here we present numerical experiments on convex and non-convex learning problems.[1] For each problem, we evenly distribute the dataset into $W - B = 50$ honest workers unless indicated otherwise. To account for malicious attacks, we additionally launch $B = 20$ Byzantine workers. We test the performance of the proposed Byrd-SAGA under three typical Byzantine attacks: Gaussian, sign-flipping and zero-gradient attacks [16], [40]. For a Gaussian attack, a Byzantine attacker $w \in \mathcal{B}$ draws its m_w^k from a Gaussian distribution with mean $\frac{1}{W-B}\sum_{w'\notin\mathcal{B}} m_{w'}^k$ and variance 30. For a sign-flipping attack, a Byzantine attacker $w \in \mathcal{B}$ sets its message as $m_w^k = u \cdot \frac{1}{W-B}\sum_{w'\notin\mathcal{B}} m_{w'}^k$, where the magnitude $u = -3$ is used in the numerical experiments. And for a zero-gradient attack, a Byzantine attacker $w \in \mathcal{B}$ sends $m_w^k = -\frac{1}{B}\sum_{w'\notin\mathcal{B}} m_{w'}^k$ so that the messages at the master sum up to zero. We use the algorithm in [35] to obtain the ϵ-approximate geometric median with $\epsilon = 1 \times 10^{-5}$.

A. ℓ_2-Regularized Logistic Regression

Consider the ℓ_2-regularized logistic regression cost, where each summand $f_{w,j}(x)$ is given by

$$f_{w,i}(x) = \ln\left(1 + \exp\left(-b_{w,i}\langle a_{w,i}, x\rangle\right)\right) + \frac{\rho}{2}||x||^2$$

with $a_{w,j} \in \mathbb{R}^p$ being the feature vector, $b_{w,j} \in \{-1, 1\}$ the label, and $\rho = 0.01$ a constant. We use the IJCNN1 and COVTYPE datasets.[2] IJCNN1 contains 49,990 training data samples of internal combustion engine outputs, each with $p = 22$ dimensions. COVTYPE contains 581,012 training data samples of forest cover types, each with $p = 54$ dimensions.

We first compare SGD, mini-batch (B)SGD with batch size 50 and SAGA, using mean and geometric median aggregation rules. Compared to SGD, BSGD enjoys smaller stochastic gradient noise, but incurs higher computational cost. In comparison, SAGA also reduces stochastic gradient noise, but its computational cost is in the same order as that of SGD. For each algorithm, we adopt a constant step size, which is tuned to achieve the best optimality gap $f(x^k) - f(x^*)$ in the Byzantine-free scenario. The performance of these algorithms on the IJCNN1 and COVTYPE datasets is depicted in Fig. 3 and Fig. 4, respectively. With Byzantine attacks, all three algorithms using mean aggregation fail. Among the three using geometric median aggregation, Byrd-SAGA markedly outperforms the other two, while BSGD is better than SGD. This demonstrates

[1]The codes are available at https://github.com/MrFive5555/Byrd-SAGA
[2]https://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets

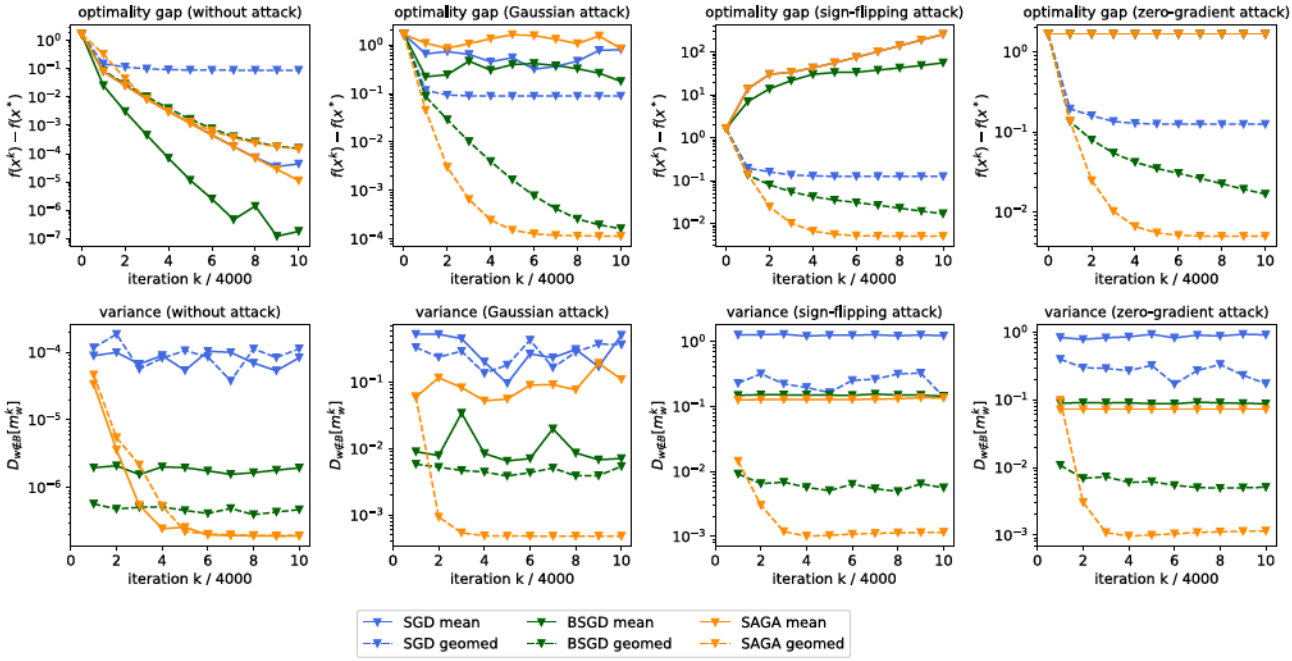


Fig. 3. Performance of the distributed SGD, mini-batch (B)SGD and SAGA, with mean and geometric median (geomed) aggregation rules on IJCNN1 dataset. The step sizes are 0.02, 0.01, and 0.02, respectively. SAGA geomed stands for the proposed Byrd-SAGA. From top to bottom: optimality gap and variance of honest messages. From left to right: without attack, Gaussian attack, sign-flipping attack, and zero-gradient attack.

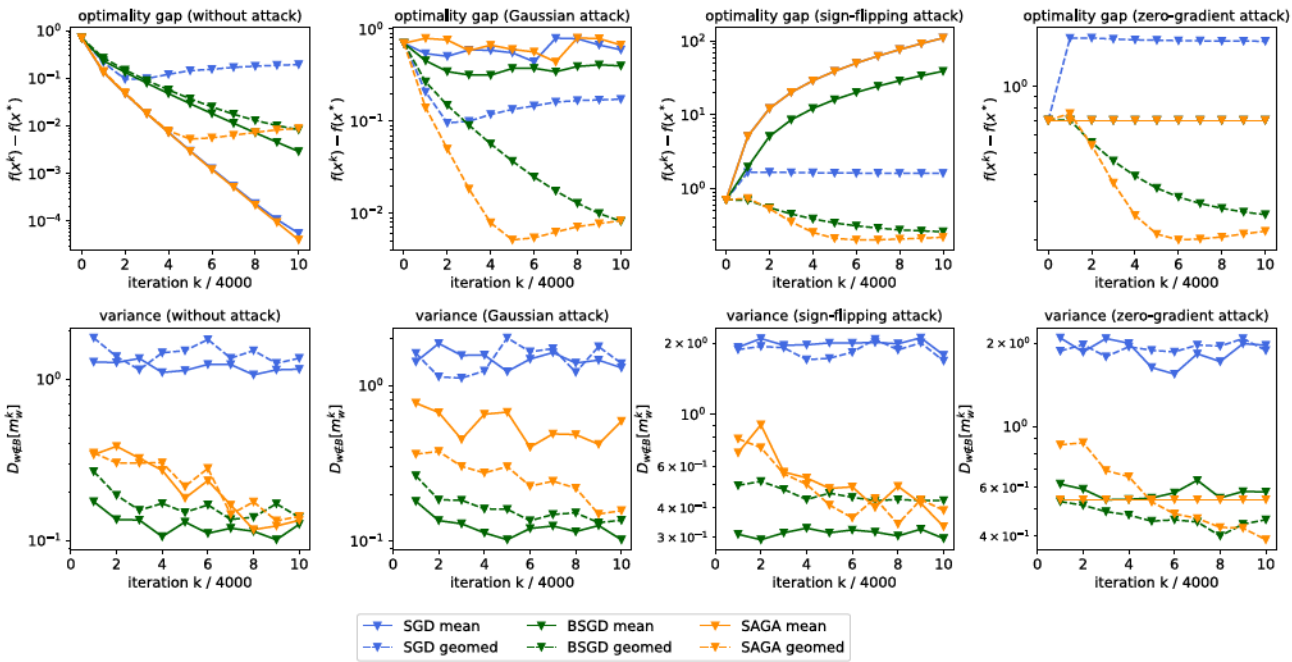


Fig. 4. Performance of the distributed SGD, mini-batch (B)SGD and SAGA, with mean and geometric median (geomed) aggregation rules on COVTYPE dataset. The step sizes are 0.01, 0.005, and 0.01, respectively. SAGA geomed stands for the proposed Byrd-SAGA. From top to bottom: optimality gap and variance of honest messages. From left to right: without attack, Gaussian attack, sign-flipping attack, and zero-gradient attack.

the importance of variance reduction to handling Byzantine attacks. Regarding the variance of honest messages in particular, Byrd-SAGA, Byzantine attack resilient BSGD and Byzantine attack resilient SGD are in the order of 10^{-3}, 10^{-2} and 10^{-1}, respectively, for the IJCNN1 dataset. For the COVTYPE dataset, Byrd-SAGA and Byzantine attack resilient BSGD have the same order of variance with respect to honest messages. In this case, Byrd-SAGA achieves similar optimality gap as Byzantine attack resilient BSGD, but converges faster because it is able to use a larger step size.

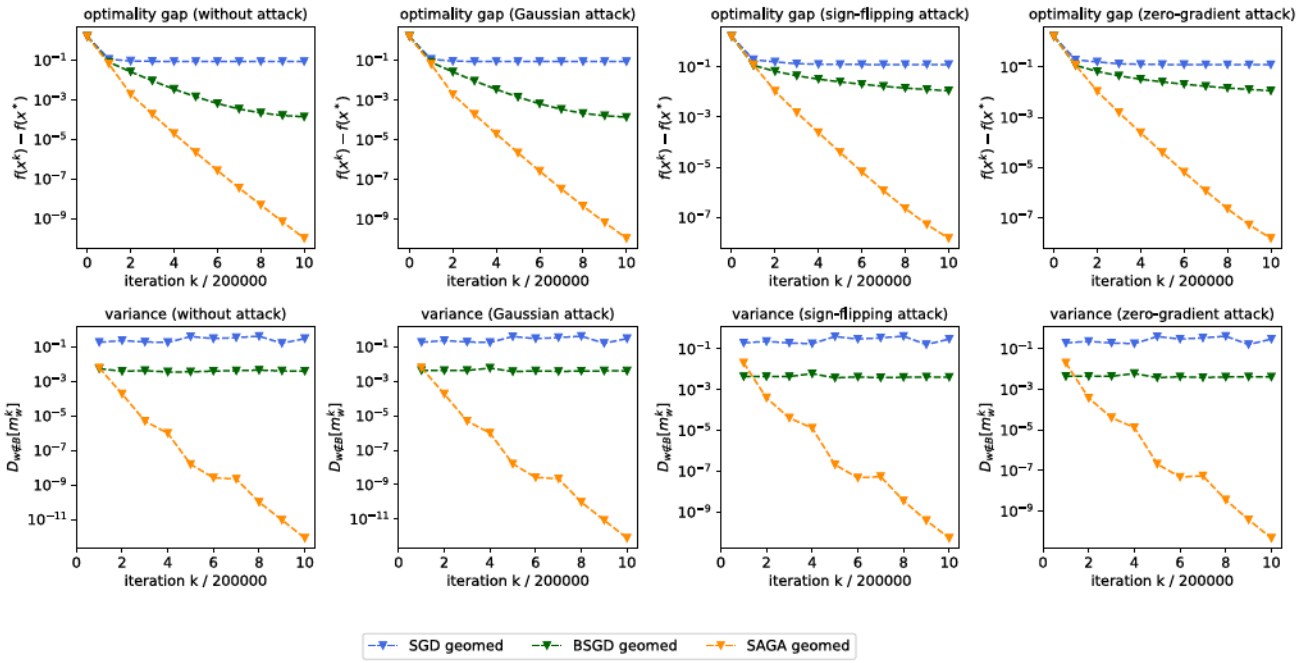


Fig. 5. Performance of the distributed SGD, mini-batch (B)SGD and SAGA, with geometric median (geomed) aggregation rule. Every honest worker has the whole IJCNN1 dataset and the outer variation $\delta^2 = 0$. The step sizes are 0.0004, 0.0002, and 0.0004, respectively. SAGA geomed stands for the proposed Byrd-SAGA. From top to bottom: optimality gap and variance of honest messages. From left to right: without attack, Gaussian attack, sign-flipping attack, and zero-gradient attack.

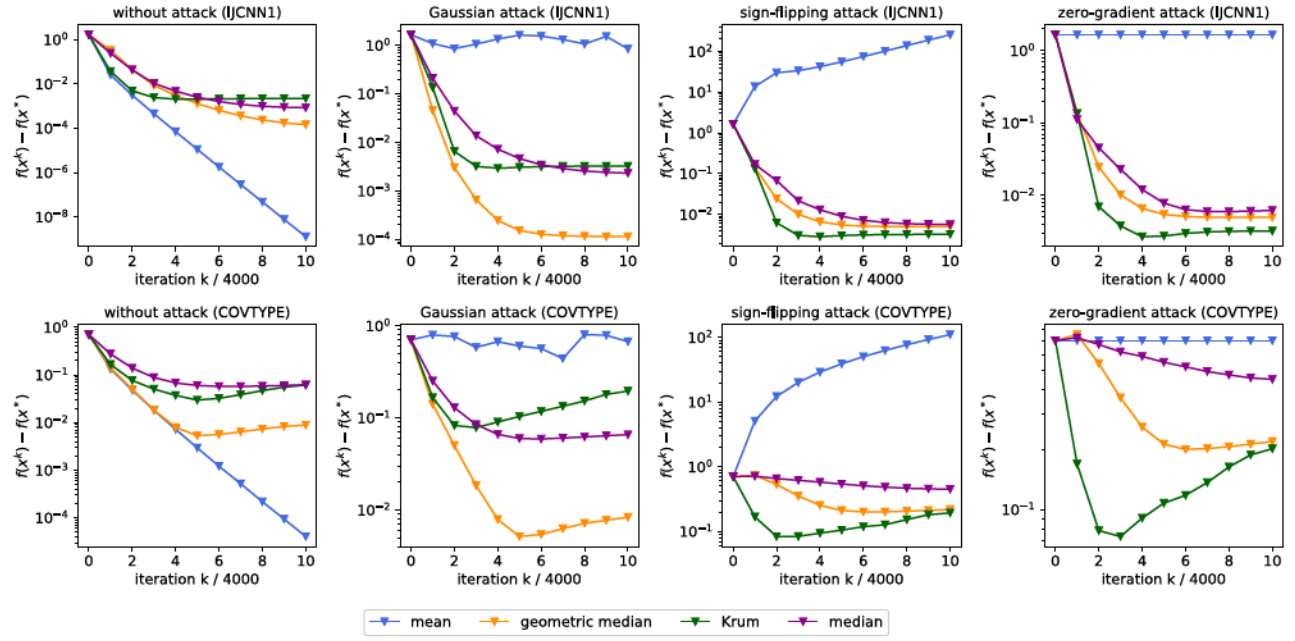


Fig. 6. Optimality gaps of distributed SAGA with different aggregation rules: mean, geometric median, median and Krum. The step sizes are 0.02 and 0.01 for the IJCNN1 and COVTYPE datasets, respectively. Curves of geometric median correspond to the proposed Byrd-SAGA. From top to bottom: on IJCNN1 dataset and on COVTYPE dataset. From left to right: without attack, with Gaussian attack, with sign-flipping attack, and with zero-gradient attack.

Theorem 1 establishes that when the outer variation $\delta^2 = 0$, the asymptotic learning error of Byrd-SAGA is zero, no matter how large the inner variation σ^2 is. In contrast, according to Theorem 2, the asymptotic learning error of Byzantine attack resilient SGD is still proportional to the inner variation σ^2. To validate these theoretical results, we conducted a second set of numerical experiments, where every honest worker has the whole IJCNN1 dataset. Therefore, $\delta^2 = 0$ and σ^2 remains the

TABLE I
ACCURACY OF SGD, MINI-BATCH (B)SGD AND SAGA, WITH MEAN AND GEOMETRIC MEDIAN (GEOMED) AGGREGATION RULES. SAGA GEOMED STANDS FOR THE PROPOSED BYRD-SAGA

attack	algorithm	mean acc (%)	geomed acc (%)
without	SGD	97.0	92.3
	BSGD	98.6	98.0
	SAGA	96.5	96.3
Gaussian	SGD	36.3	92.5
	BSGD	36.3	98.0
	SAGA	14.5	96.4
sign-flipping	SGD	0.11	0.03
	BSGD	0.16	90.3
	SAGA	0.12	86.4
zero-gradient	SGD	9.94	26.2
	BSGD	9.89	81.5
	SAGA	9.88	92.4

same as that in the first set of experiments. We compare SGD, BSGD with batch size 50 and SAGA, all using the geometric median aggregation rule. The results depicted in Fig. 5 corroborate the theoretical findings – the asymptotic learning error of Byrd-SAGA vanishes, while those of Byzantine attack resilient SGD and BSGD are the same as those shown in Fig. 3.

In the third set of numerical experiments, we compare the use of different aggregation rules in distributed SAGA: mean, geometric median, median, and Krum. As shown in Fig. 6, distributed SAGA using mean aggregation is the best in terms of the optimality gap $f(x^k) - f(x^*)$ when there are no Byzantine attacks. However, it fails under all kinds of attacks. With Gaussian attacks, Byrd-SAGA using geometric median achieves the best performance. With sign-flipping and zero-gradient attacks, Byrd-SAGA using Krum is the best, while that using geometric median also performs well. Note that Krum has to know the exact number of Byzantine attackers in advance, while geometric median and median do not need this prior knowledge.

B. Neural Network Training

Here we test training a neural network with one hidden layer of 50 neurons and "tanh" activation function, for multi-class classification on the MNIST dataset.[3] MNIST contains 60,000 training and 10,000 testing data samples of handwritten digits, each with $p = 784$ dimensions. We compare SGD with step size 0.1, BSGD with step size 0.5 and batch size 50, and SAGA with step size 0.1. We run the algorithms for 15,000 iterations, and report the final accuracy in Table I. With mean aggregation, all algorithms yield low accuracy in the presence of Byzantine attacks. With the help of geometric median aggregation, BSGD and SAGA are both robust and outperform SGD. Note that Byrd-SAGA exhibits a much lower per-iteration computational cost relative to Byzantine attack resilient BSGD.

VI. CONCLUSION

The present paper developed a novel Byzantine attack resilient distributed (Byrd-) SAGA approach to federated finite-sum optimization in the presence of Byzantine attacks. On a par with SAGA, Byrd-SAGA corrects stochastic gradients through variance reduction. Per iteration, distributed workers obtain their corrected stochastic gradients before uploading to the master node. Different from SAGA though, the master node in Byrd-SAGA aggregates the received messages using the geometric median rather than the mean. This robust aggregation markedly enhances robustness of Byrd-SAGA in the presence of Byzantine attacks. It was established that Byrd-SAGA converges linearly to a neighborhood of the optimal solution, with the asymptotic learning error determined solely by the number of Byzantine workers.

As confirmed by numerical tests, combinations with other robust aggregation rules also exhibit satisfactory robustness. In addition to investigating their analyses, our future research agenda also includes communication-efficient extensions of Byrd-SAGA [36], [37], as well as the development and analysis of Byzantine attack resilient algorithms over fully decentralized networks [41], [42].

APPENDIX A
PROOF OF LEMMA 1

The proof of Lemma 1 relies on the following lemma.

Lemma 2: Let $\mathcal{Z}$ be a subset of random vectors distributed in a normed vector space. If $\mathcal{Z}' \subseteq \mathcal{Z}$ and $|\mathcal{Z}'| < \frac{|\mathcal{Z}|}{2}$, then it holds that

$$E\|\underset{z\in\mathcal{Z}}{\text{geomed}}\{z\}\|^2 \leq C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'} E\|z\|^2}{|\mathcal{Z}| - |\mathcal{Z}'|} \tag{26}$$

where $C_\alpha := \frac{2-2\alpha}{1-2\alpha}$ and $\alpha := \frac{|\mathcal{Z}'|}{|\mathcal{Z}|}$.

Proof: With $z^* = \text{geomed}_{z\in\mathcal{Z}}\{z\}$ and $z \in \mathcal{Z}'$, it holds that $\|z^* - z\| \geq \|z\| - \|z^*\|$; and for all $z \notin \mathcal{Z}'$, we have $\|z^* - z\| \geq \|z^*\| - \|z\|$. Then, summing up $\|z^* - z\|$ over all $z \in \mathcal{Z}$ yields

$$\sum_{z\in\mathcal{Z}} \|z^* - z\| \geq \sum_{z\in\mathcal{Z}} \|z\| + (|\mathcal{Z}| - 2|\mathcal{Z}'|)\|z^*\| - 2\sum_{z\notin\mathcal{Z}'} \|z\|. \tag{27}$$

According to the definition of geometric median, it holds that

$$\sum_{z\in\mathcal{Z}} \|z^* - z\| = \inf_y \sum_{z\in\mathcal{Z}} \|y - z\| \leq \sum_{z\in\mathcal{Z}} \|z\|. \tag{28}$$

Combining the two inequalities, we arrive at

$$\|z^*\| \leq \frac{2\sum_{z\notin\mathcal{Z}'} \|z\|}{|\mathcal{Z}| - 2|\mathcal{Z}'|} = C_\alpha \frac{\sum_{z\notin\mathcal{Z}'} \|z\|}{|\mathcal{Z}| - |\mathcal{Z}'|} \tag{29}$$

and upon squaring both sides of the latter, we find

$$\|z^*\|^2 \leq C_\alpha^2 \frac{(\sum_{z\notin\mathcal{Z}'} \|z\|)^2}{(|\mathcal{Z}| - |\mathcal{Z}'|)^2} \leq C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'} \|z\|^2}{|\mathcal{Z}| - |\mathcal{Z}'|}. \tag{30}$$

Then taking expectations on both sides, yields (26), and completes the proof. ■

With Lemma 2, the proof of Lemma 1 is straightforward.

Proof: It follows readily from Lemma 2 that

$$E\|\underset{z\in\mathcal{Z}}{\text{geomed}}\{z\} - \bar{z}\|^2 = E\|\underset{z\in\mathcal{Z}}{\text{geomed}}\{z - \bar{z}\}\|^2 \leq C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'} E\|z - \bar{z}\|^2}{|\mathcal{Z}| - |\mathcal{Z}'|}. \tag{31}$$

[3] http://yann.lecun.com/exdb/mnist

Applying the inequality of $\|z-\bar{z}\|^2 \leq 2\|z-Ez\|^2 + 2\|Ez-\bar{z}\|^2$ to (31), yields

$$E\|\underset{z\in\mathcal{Z}}{\text{geomed}}\{z\} - \bar{z}\|^2 \leq 2C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'} E\|z-Ez\|^2}{|\mathcal{Z}|-|\mathcal{Z}'|} + 2C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'} E\|Ez-\bar{z}\|^2}{|\mathcal{Z}|-|\mathcal{Z}'|} \quad (32)$$

which completes the proof. ■

Appendix B
Lemma 3 and Its Proof

Since computing the accurate geometric median is difficult, we consider the ϵ-approximate geometric median in this paper. The following lemma is the ϵ-approximate counterpart of Lemma 2.

Lemma 3: Let $\mathcal{Z}$ be a subset of random vectors distributed in a normed vector space. If $\mathcal{Z}' \subseteq \mathcal{Z}$ and $|\mathcal{Z}'| < \frac{|\mathcal{Z}|}{2}$, it holds that

$$E\|z_\epsilon^*\|^2 \leq 2C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'} E\|z\|^2}{|\mathcal{Z}|-|\mathcal{Z}'|} + \frac{2\epsilon^2}{(|\mathcal{Z}|-2|\mathcal{Z}'|)^2} \quad (33)$$

where $C_\alpha := \frac{2-2\alpha}{1-2\alpha}$, $\alpha := \frac{|\mathcal{Z}'|}{|\mathcal{Z}|}$, and z_ϵ^* is an ϵ-approximate geometric median of $\mathcal{Z}$.

Proof: Because z_ϵ^* is an ϵ-approximate geometric median, it follows that

$$\sum_{z\in\mathcal{Z}} \|z_\epsilon^* - z\| \leq \inf_y \sum_{z\in\mathcal{Z}} \|y-z\| + \epsilon \leq \sum_{z\in\mathcal{Z}} \|z\| + \epsilon. \quad (34)$$

Notice that (27) remains valid here. Hence, we have

$$\|z_\epsilon^*\| \leq C_\alpha \frac{\sum_{z\notin\mathcal{Z}'} \|z\|}{|\mathcal{Z}|-|\mathcal{Z}'|} + \frac{\epsilon}{|\mathcal{Z}|-2|\mathcal{Z}'|}. \quad (35)$$

Squaring both sides of (35), leads to

$$\|z_\epsilon^*\|^2 \leq \left(C_\alpha \frac{\sum_{z\notin\mathcal{Z}'} \|z\|}{|\mathcal{Z}|-|\mathcal{Z}'|} + \frac{\epsilon}{|\mathcal{Z}|-2|\mathcal{Z}'|} \right)^2 \quad (36)$$

$$\leq 2C_\alpha^2 \left(\frac{\sum_{z\notin\mathcal{Z}'} \|z\|}{|\mathcal{Z}|-|\mathcal{Z}'|} \right)^2 + \frac{2\epsilon^2}{(|\mathcal{Z}|-2|\mathcal{Z}'|)^2} \quad (37)$$

$$\leq 2C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'} \|z\|^2}{|\mathcal{Z}|-|\mathcal{Z}'|} + \frac{2\epsilon^2}{(|\mathcal{Z}|-2|\mathcal{Z}'|)^2}. \quad (38)$$

Then taking expectations on both sides, yields (33), and completes the proof. ■

Appendix C
Lemma 4 and Its Proof

As we have indicated in Section IV-B, the main challenge in the proof of Byrd-SAGA is that the geometric median of $\{m_w^k, w\in\mathcal{W}\}$ is a biased estimate of the gradient $f'(x^k)$. To handle the bias, the following lemma characterizes the error between an ϵ-approximate geometric median of $\{m_w^k, w\in\mathcal{W}\}$ and $f'(x^k)$ per slot k.

Lemma 4: Consider Byrd-SAGA with ϵ-approximate geometric median aggregation. Under Assumptions 1 and 2, if the number of Byzantine attackers satisfies $B < \frac{W}{2}$, then an ϵ-approximate geometric median of $\{m_w^k, w\in\mathcal{W}\}$, denoted by z_ϵ^*, satisfies

$$E\|z_\epsilon^* - f'(x^k)\|^2 \leq 4C_\alpha^2 L^2 S^k + 4C_\alpha^2\delta^2 + \frac{2\epsilon^2}{(W-2B)^2} \quad (39)$$

where

$$C_\alpha := \frac{2-2\alpha}{1-2\alpha} \quad \text{and} \quad \alpha := \frac{B}{W} \quad (40)$$

while S^k is defined as

$$S^k := \frac{1}{W-B} \sum_{w\notin\mathcal{B}} \frac{1}{J} \sum_{j=1}^J \|x^k - \phi_{w,j}^k\|^2. \quad (41)$$

Proof: We begin with upper bounding the mean-square error $E\|m_w^k - f_w'(x^k)\|^2$, where $w\notin\mathcal{B}$. Using the definition of m_w^k in (10), we have for any $w\notin\mathcal{B}$ that

$$\begin{aligned} &E\|m_w^k - f_w'(x^k)\|^2 \\ &= E\|f_{w,i_w^k}'(x^k) - f_{w,i_w^k}'(\phi_{w,i_w^k}^k) \\ &\quad + \frac{1}{J}\sum_{j=1}^J f_{w,j}'(\phi_{w,j}^k) - f_w'(x^k)\|^2 \\ &= E\|f_{w,i_w^k}'(x^k) - f_{w,i_w^k}'(\phi_{w,i_w^k}^k)\|^2 \\ &\quad - \|f_w'(x^k) - \frac{1}{J}\sum_{j=1}^J f_{w,j}'(\phi_{w,j}^k)\|^2 \\ &\leq E\|f_{w,i_w^k}'(x^k) - f_{w,i_w^k}'(\phi_{w,i_w^k}^k)\|^2 \\ &\leq L^2 E\|x^k - \phi_{w,i_w^k}^k\|^2 \end{aligned} \quad (42)$$

where the second equality is due to variance decomposition $E\|a-Ea\|^2 = E\|a\|^2 - \|Ea\|^2$ with $a = f_{w,i_w^k}'(x^k) - f_{w,i_w^k}'(\phi_{w,i_w^k}^k)$, and $Ea = f_w'(x^k) - \frac{1}{J}\sum_{j=1}^J f_{w,j}'(\phi_{w,j}^k)$; while the last inequality comes from Assumption 1.

Next, we will derive an upper bound on $E\|z_\epsilon^* - f'(x^k)\|^2$. According to (33) in Lemma 3, (42), and Assumption 2, it holds that

$$\begin{aligned} &E\|z_\epsilon^* - f'(x^k)\|^2 \\ &\leq 2C_\alpha^2 \frac{1}{W-B}\sum_{w\notin\mathcal{B}} E\|m_w^k - f'(x^k)\|^2 + \frac{2\epsilon^2}{(W-2B)^2} \\ &\leq 4C_\alpha^2 \frac{1}{W-B}\sum_{w\notin\mathcal{B}} E\|m_w^k - f_w'(x^k)\|^2 \\ &\quad + 4C_\alpha^2 \frac{1}{W-B}\sum_{w\notin\mathcal{B}} \|f_w'(x^k) - f'(x^k)\|^2 + \frac{2\epsilon^2}{(W-2B)^2} \\ &\leq 4C_\alpha^2 \frac{1}{W-B}\sum_{w\notin\mathcal{B}} L^2 E\|x^k - \phi_{w,i_w^k}^k\|^2 \\ &\quad + 4C_\alpha^2\delta^2 + \frac{2\epsilon^2}{(W-2B)^2} \\ &= 4C_\alpha^2 L^2 S^k + 4C_\alpha^2\delta^2 + \frac{2\epsilon^2}{(W-2B)^2} \end{aligned} \quad (43)$$

which completes the proof. ■

APPENDIX D
LEMMA 5 AND ITS PROOF

In Lemma 4, the upper bound of $E\|z_\epsilon^* - f'(x^k)\|^2$ contains a time-varying term S^k. The following lemma characterizes the evolution of S^k.

Lemma 5: Consider Byrd-SAGA with ϵ-approximate geometric median aggregation. Under Assumptions 1, it holds that

$$ES^{k+1} \le 4J \cdot E\|x^{k+1} - x^k + \gamma f'(x^k)\|^2 + 4J\gamma^2 L^2 \|x^k - x^*\|^2 + \left(1 - \frac{1}{J^2}\right) S^k \quad (44)$$

where S^k is defined in (41).

Proof: For the expectation of ES^{k+1}, we have that

$$\begin{aligned} ES^{k+1} &= E\left(\frac{1}{W-B}\sum_{w\notin\mathcal{B}}\frac{1}{J}\sum_{j=1}^{J}\|x^{k+1} - \phi_{w,j}^{k+1}\|^2\right) \\ &\le E\left(\frac{1}{W-B}\sum_{w\notin\mathcal{B}}\frac{1}{J}\sum_{j=1}^{J}(1+\beta^{-1})\|x^{k+1} - x^k\|^2\right) \\ &\quad + E\left(\frac{1}{W-B}\sum_{w\notin\mathcal{B}}\frac{1}{J}\sum_{j=1}^{J}(1+\beta)\|x^k - \phi_{w,j}^{k+1}\|^2\right) \\ &= (1+\beta^{-1})\cdot E\|x^{k+1} - x^k\|^2 + (1+\beta)\left(1 - \frac{1}{J}\right) S^k \end{aligned} \quad (45)$$

where the inequality comes from $\|a+b\|^2 \le (1+\beta^{-1})\|a\|^2 + (1+\beta)\|b\|^2$ for any $\beta > 0$, and the third equality holds because at slot k, honest worker w uniformly at random chooses one out of J data samples. For the chosen data sample j, $\phi_{w,j}^{k+1} = x^k$; otherwise, $\phi_{w,j}^{k+1} = \phi_{w,j}^k$. Then it holds

$$E\left(\frac{1}{W-B}\sum_{w\notin\mathcal{B}}\frac{1}{J}\sum_{j=1}^{J}\|x^k - \phi_{w,j}^{k+1}\|^2\right) = \left(1 - \frac{1}{J}\right) S^k. \quad (46)$$

Using the fact that $f'(x^*) = 0$, the first term in the right-hand side of (45) can be bounded as

$$\begin{aligned} &\|x^{k+1} - x^k\|^2 \\ &= \|x^{k+1} - x^k + \gamma f'(x^k) - \gamma f'(x^k) + \gamma f'(x^*)\|^2 \\ &\le 2\|x^{k+1} - x^k + \gamma f'(x^k)\|^2 + 2\gamma^2\|f'(x^k) - f'(x^*)\|^2 \\ &\le 2\|x^{k+1} - x^k + \gamma f'(x^k)\|^2 + 2\gamma^2 L^2\|x^k - x^*\|^2 \end{aligned} \quad (47)$$

where the first inequality comes from $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, and the last inequality comes from Assumption 1.

Substituting (47) into (45), and choosing $\beta = 1/J$, we have

$$\begin{aligned} ES^{k+1} &\le (1+J)\cdot E\|x^{k+1} - x^k\|^2 + \left(1 - \frac{1}{J^2}\right) S^k \\ &\le 2J \cdot E\|x^{k+1} - x^k\|^2 + \left(1 - \frac{1}{J^2}\right) S^k \\ &\le 4J \cdot E\|x^{k+1} - x^k + \gamma f'(x^k)\|^2 \\ &\quad + 4J\gamma^2 L^2\|x^k - x^*\|^2 + \left(1 - \frac{1}{J^2}\right) S^k \end{aligned} \quad (48)$$

which completes the proof. ■

APPENDIX E
PROOF OF THEOREM 1

Proof: Let z_ϵ^* be the ϵ-approximate geometric median of $\{m_w^k, w \in \mathcal{W}\}$. We begin by manipulating $E\|x^{k+1} - x^*\|^2$ as

$$\begin{aligned} &E\|x^{k+1} - x^*\|^2 \\ &= E\|x^k - \gamma f'(x^k) - x^* + x^{k+1} - x^k + \gamma f'(x^k)\|^2 \\ &\le \frac{1}{1-\eta}\|x^k - \gamma f'(x^k) - x^*\|^2 \\ &\quad + \frac{1}{\eta}E\|x^{k+1} - x^k + \gamma f'(x^k)\|^2, \end{aligned} \quad (49)$$

where $0 < \eta < 1$, and the inequality comes from $\|a+b\|^2 \le \frac{1}{\eta}\|a\|^2 + \frac{1}{1-\eta}\|b\|^2$.

To bound the first term in the right-hand side of (49), we use that f_{w,i_w^k} is μ-strongly convex and has L-Lipschitz continuous gradients. Using also the fact that $f'(x^*) = 0$, we obtain

$$\begin{aligned} &\|x^k - \gamma f'(x^k) - x^*\|^2 \\ &= \|x^k - \gamma(f'(x^k) - f'(x^*)) - x^*\|^2 \\ &= \|x^k - x^*\|^2 - 2\gamma\langle f'(x^k) - f'(x^*), x^k - x^*\rangle \\ &\quad + \gamma^2\|f'(x^k) - f'(x^*)\|^2 \\ &\le \|x^k - x^*\|^2 - 2\gamma\mu\|x^k - x^*\|^2 + \gamma^2 L^2\|x^k - x^*\|^2 \\ &= (1 - 2\gamma\mu + \gamma^2 L^2)\|x^k - x^*\|^2. \end{aligned} \quad (50)$$

Here $\langle f'(x^k) - f'(x^*), x^k - x^*\rangle \ge \mu\|x^k - x^*\|^2$ because f is μ-strongly convex [43, Theorem 2.1.9]. Further, because f has L-Lipschitz continuous gradients, it holds that $\|f'(x^k) - f'(x^*)\|^2 \le L^2\|x^k - x^*\|^2$.

Substituting (50) into (49) yields

$$E\|x^{k+1} - x^*\|^2 \le \frac{1 - 2\gamma\mu + \gamma^2 L^2}{1-\eta}\|x^k - x^*\|^2 + \frac{1}{\eta}E\|x^{k+1} - x^k + \gamma f'(x^k)\|^2. \quad (51)$$

With $\eta = \gamma\mu/2$, as long as

$$\gamma^2 L^2 \le \frac{\gamma\mu}{2} \quad (52)$$

it follows that

$$\frac{1 - 2\gamma\mu + \gamma^2 L^2}{1-\eta} \le 1 - \gamma\mu.$$

Therefore, (51) can be rewritten as

$$E\|x^{k+1} - x^*\|^2 \le (1 - \gamma\mu)\|x^k - x^*\|^2 + \frac{2}{\gamma\mu}E\|x^{k+1} - x^k + \gamma f'(x^k)\|^2. \quad (53)$$

Then, we construct a *Lyapunov function* T^k as

$$T^k := \|x^k - x^*\|^2 + cS^k \tag{54}$$

where c is any positive constant. According to the definition in (41), we know S^k is non-negative. Therefore, T^k is also non-negative.

Substituting (44) and (53) into (54), it follows that

$$\begin{aligned} ET^{k+1} \leq\ & (1-\gamma\mu+4cJ\gamma^2L^2)\|x^k-x^*\|^2 \\ & + \left(\frac{2}{\gamma\mu}+4cJ\right)E\|x^{k+1}-x^k+\gamma f'(x^k)\|^2 \\ & + (1-\frac{1}{J^2})cS^k. \end{aligned} \tag{55}$$

According to Lemma 4, the second term on the right-hand side (55) can be bounded as

$$\begin{aligned} & E\|x^{k+1}-x^k+\gamma f'(x^k)\|^2 = \gamma^2 E\|z_\epsilon^* - f'(x^k)\|^2 \\ & \leq \gamma^2\left(4C_\alpha^2L^2S^k + 4C_\alpha^2\delta^2 + \frac{2\epsilon^2}{(W-2B)^2}\right). \end{aligned} \tag{56}$$

Hence, we have

$$\begin{aligned} ET^{k+1} \leq\ & (1-\gamma\mu+4cJ\gamma^2L^2)\|x^k-x^*\|^2 \\ & + \left(\left(1-\frac{1}{J^2}\right)c + \left(\frac{2}{\gamma\mu}+4cJ\right)4C_\alpha^2\gamma^2L^2\right)S^k \\ & + \gamma^2\left(\frac{2}{\gamma\mu}+4cJ\right)\left(4C_\alpha^2\delta^2+\frac{2\epsilon^2}{(W-2B)^2}\right). \end{aligned} \tag{57}$$

If we constrain the step size γ as

$$4cJ\gamma^2L^2 \leq \frac{\gamma\mu}{2} \tag{58}$$

the coefficient in front of $\|x^k - x^*\|^2$ satisfies

$$1-\gamma\mu+4cJ\gamma^2L^2 \leq 1-\frac{\gamma\mu}{2}.$$

Similarly, if γ and c are chosen such that

$$\frac{\gamma\mu}{2} < \frac{1}{4J^2} \quad \text{and} \quad 16JC_\alpha^2\gamma^2L^2 < \frac{1}{4J^2} \tag{59}$$

and

$$c = \frac{16J^2C_\alpha^2\gamma L^2}{\mu} \geq \frac{8C_\alpha^2\gamma L^2}{\mu(1/J^2-\gamma\mu/2-16JC_\alpha^2\gamma^2L^2)}$$

the coefficient in front of S^k satisfies

$$\begin{aligned} & \left(1-\frac{1}{J^2}\right)c + \left(\frac{2}{\gamma\mu}+4cJ\right)4C_\alpha^2\gamma^2L^2 \\ & = \left(1-\frac{1}{J^2}+16JC_\alpha^2\gamma^2L^2\right)c + \frac{8C_\alpha^2\gamma L^2}{\mu} \leq \left(1-\frac{\gamma\mu}{2}\right)c. \end{aligned}$$

Therefore, (57) becomes

$$\begin{aligned} & ET^{k+1} \\ & \leq \left(1-\frac{\gamma\mu}{2}\right)\|x^k-x^*\|^2 + \left(1-\frac{\gamma\mu}{2}\right)cS^k \\ & \quad + \gamma^2\left(\frac{2}{\gamma\mu}+4cJ\right)\left(4C_\alpha^2\delta^2+\frac{2\epsilon^2}{(W-2B)^2}\right) \\ & = \left(1-\frac{\gamma\mu}{2}\right)T^k \\ & \quad + \left(\frac{2\gamma}{\mu}+\frac{64J^3C_\alpha^2\gamma^3L^2}{\mu}\right)\left(4C_\alpha^2\delta^2+\frac{2\epsilon^2}{(W-2B)^2}\right). \end{aligned} \tag{60}$$

For simplicity, let also

$$\tilde{\Delta}_2 := \left(\frac{2\gamma}{\mu}+\frac{64J^3C_\alpha^2\gamma^3L^2}{\mu}\right)\left(4C_\alpha^2\delta^2+\frac{2\epsilon^2}{(W-2B)^2}\right). \tag{61}$$

Using telescopic cancellation on (60) from slot 1 to slot k, we arrive at

$$ET^k \leq (1-\frac{\gamma\mu}{2})^k\left[T^0 - \frac{2}{\gamma\mu}\tilde{\Delta}_2\right] + \frac{2}{\gamma\mu}\tilde{\Delta}_2. \tag{62}$$

Here and thereafter, the expectation is taken over i_w^k for all workers $w \notin \mathcal{B}$ and slots $t \leq k-1$.

The definition of the *Lyapunov function* in (54), implies that

$$E\|x^k - x^*\|^2 \leq ET^k \leq \left(1-\frac{\gamma\mu}{2}\right)^k\Delta_1 + \Delta_2 \tag{63}$$

where the constants Δ_1 and Δ_2 are defined as

$$\Delta_1 := \|x^0 - x^*\|^2 - \Delta_2 \tag{64}$$

$$\begin{aligned} \Delta_2 &:= \frac{2}{\gamma\mu}\tilde{\Delta}_2 \\ &= \frac{4}{\mu^2}\left(1+32J^3C_\alpha^2\gamma^2L^2\right)\left(4C_\alpha^2\delta^2+\frac{2\epsilon^2}{(W-2B)^2}\right). \end{aligned} \tag{65}$$

In our derivation so far, the step size γ must satisfy (52), (58) and (59), meaning that

$$\gamma \leq \min\left\{\frac{\mu}{2L^2}, \frac{\mu}{8\sqrt{2}J^{3/2}C_\alpha L^2}, \frac{1}{2J^2\mu}, \frac{1}{8J^{3/2}C_\alpha L}, \right\}.$$

Therefore, we simply choose

$$\gamma \leq \frac{\mu}{8J^2C_\alpha L^2}$$

and the proof is complete. ■

Appendix F
Proof of Theorem 2

Let z_ϵ^* denote the ϵ-approximate geometric median of $\{\tilde{m}_w^k, w \in \mathcal{W}\}$, where $\tilde{m}_w^k = f'_{w,i_w^k}(x^k)$ for $w \notin \mathcal{B}$, and arbitrary otherwise. Similar to the proof of Theorem 1, we first derive an upper bound on $E\|x^{k+1} - x^*\|$. Inequality (53) is still true for Byzantine attack resilient SGD with $\gamma < \mu/(2L^2)$, and the only difference is that $E\|x^{k+1} - x^k + \gamma f'(x^k)\|^2$ becomes

$$\begin{aligned} & E\|x^{k+1}-x^k+\gamma f'(x^k)\|^2 \\ & = \gamma^2 E\|z_\epsilon^* - f'(x^k)\|^2 \end{aligned}$$

$$\leq \gamma^2 \left(2C_\alpha^2 \frac{\sum_{w\notin\mathcal{B}} E\|f'_{w,i_w^k}(x^k) - f'(x^k)\|^2}{W-B} + \frac{2\epsilon^2}{(W-2B)^2} \right)$$

$$\leq \gamma^2 \left(2C_\alpha^2 \frac{1}{W-B} \sum_{w\notin\mathcal{B}} 2E\|f'_{w,i_w^k}(x^k) - f'_w(x^k)\|^2 \right)$$

$$+ \gamma^2 \left(2C_\alpha^2 \frac{1}{W-B} \sum_{w\notin\mathcal{B}} 2\|f'_w(x^k) - f'(x^k)\|^2 \right)$$

$$+ \frac{2\epsilon^2}{(W-2B)^2}$$

$$\leq \gamma^2 \left(4C_\alpha^2\sigma^2 + 4C_\alpha^2\delta^2 + \frac{2\epsilon^2}{(W-2B)^2} \right) \tag{66}$$

where first two inequalities are analogous to those in (43), while the last inequality comes from Assumptions 2 and 3. Therefore, for Byzantine attack resilient SGD, we have

$$E\|x^{k+1} - x^*\|^2$$

$$\leq (1-\gamma\mu)\|x^k - x^*\|^2 + \frac{2}{\gamma\mu} E\|x^{k+1} - x^k + \gamma f'(x^k)\|^2$$

$$\leq (1-\gamma\mu)\|x^k - x^*\|^2$$

$$+ \frac{2\gamma}{\mu} \left(4C_\alpha^2\sigma^2 + 4C_\alpha^2\delta^2 + \frac{2\epsilon^2}{(W-2B)^2} \right). \tag{67}$$

Here and thereafter, the expectation is taken over i_w^k for all workers $w \notin \mathcal{B}$, and slots $t \leq k-1$.

Using telescopic cancellation on (67) from slot 1 to slot k, we deduce that

$$E\|x^{k+1} - x^*\|^2 \leq (1-\gamma\mu)^k \Delta'_1 + \Delta'_2 \tag{68}$$

where Δ'_1 and Δ'_2 are defined as

$$\Delta'_1 := \|x^0 - x^*\|^2 - \Delta'_2 \tag{69}$$

$$\Delta'_2 := \frac{2}{\mu^2} \left(4C_\alpha^2\sigma^2 + 4C_\alpha^2\delta^2 + \frac{2\epsilon^2}{(W-2B)^2} \right) \tag{70}$$

and the proof is complete.

REFERENCES

[1] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proc. SIGMOD*, Dallas, TX, USA, May 2000, pp. 439–450.

[2] J. Duchi, M. J. Wainwright, and M. I. Jordan, "Local privacy and minimax bounds: Sharp rates for probability estimation," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, Stateline, NV, USA, Dec. 2013, pp. 1529–1537.

[3] L. Zhou, K. Yeh, G. Hancke, Z. Liu, and C. Su, "Security and privacy for the industrial Internet of Things: An overview of approaches to safeguard endpoints," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 76–87, Sep. 2018.

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

[5] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, *Federated Learn.* San Rafael, CA, USA: Morgan & Claypool, 2019.

[6] A. Vempaty, L. Tong, and P. K. Varshney, "Distributed inference with Byzantine data: State-of-the-art review on data falsification attacks," *IEEE Signal Process. Mag.*, vol. 30, no. 5, pp. 65–75, Aug. 2013.

[7] Y. Chen, S. Kar, and J. M. F. Moura, "The Internet of Things: Secure distributed inference," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 64–75, Sep. 2018.

[8] L. Lamport, R. Shostak, and M. Pease, "The Byzantine generals problem," *ACM Trans. Program. Lang. Syst.*, vol. 4, no. 3, pp. 382–401, Jul. 1982.

[9] Z. Yang, A. Gang, and W. U. Bajwa, "Adversary-resilient distributed and decentralized statistical inference and machine learning: An overview of recent advances under the Byzantine threat model," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 146–159, May 2020.

[10] S. Minsker, "Geometric median and robust estimation in Banach spaces," *Bernoulli*, vol. 21, no. 4, pp. 2308–2335, Nov. 2015.

[11] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," in *ACM. Measurement. Anal. Comput. Syst.*, Phonenix, AZ, USA, vol. 1, no. 2, pp. 44:1–44:25, Jul. 2017.

[12] C. Xie, O. Koyejo, and I. Gupta, "Generalized Byzantine-tolerant SGD," Feb. 2018, *arXiv:1802.10116*.

[13] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *Proc. Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018, pp. 5636–5645.

[14] L. Su and J. Xu, "Securing distributed machine learning in high dimensions," Apr. 2018, *ACM. Measurement. Anal. Comput. Syst.*, vol. 3, no. 1, pp. 12:1–12:41, Mar. 2019.

[15] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 118–128.

[16] L. Li, W. Xu, T. Chen, G. B. Giannakis, and Q. Ling, "RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets," in *Proc. AAAI*, Honolulu, HI, USA, Jan. 2019, pp. 1544–1551.

[17] J. Feng, H. Xu, and S. Mannor, "Distributed robust learning," Sep. 2014, *arXiv:1409.5937*.

[18] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Defending against saddle point attack in Byzantine-robust distributed learning," in *Proc. ICML*, Long Beach, CA, USA, Jun. 2019, pp. 7074–7084.

[19] L. Chen, H. Wang, Z. Charles, and D. Papailiopoulos, "DRACO: Byzantine-resilient distributed training via redundant gradients," in *Proc. ICML*, Stockholm, Sweden, Jul. 2018, pp. 902–911.

[20] S. Rajput, H. Wang, Z. Charles, and D. Papailiopoulos, "DETOX: A redundancy-based framework for faster and more robust gradient aggregation," in *Proc. NIPS*, Vancouver, BC, Canada, Dec. 2019, pp. 10320–10330.

[21] C. Xie, O. Koyejo, and I. Gupta, "Fall of empires: Breaking Byzantine-tolerant SGD by inner product manipulation," in *Proc. UAI*, Tel Aviv, Israel, Jul. 2019, pp. 83:1–83.10.

[22] P. Goyal *et al.*, "Accurate, large minibatch SGD: Training imagenet in 1 hour," Jun. 2017, *arXiv:1706.02677*.

[23] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Stateline, NV, USA, Dec. 2013, pp. 315–323.

[24] S. Shalev-Shwartz and T. Zhang, "Stochastic dual coordinate ascent methods for regularized loss minimization," *J. Mach. Learn. Res.*, vol. 14, no. 2, pp. 567–599, Feb. 2013.

[25] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takac, "SARAH: A novel method for machine learning problems using stochastic recursive gradient," in *Proc. Int. Conf. Mach. Learn.*, Sydney, Australia, Aug. 2017, pp. 2613–2621.

[26] A. Defazio, F. R. Bach, and S. Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Montreal, PQ, Canada, Dec. 2014, pp. 1646–1654.

[27] M. W. Schmidt, N. Le Roux, and F. R. Bach, "Minimizing finite sums with the stochastic average gradient," *Math. Program.*, vol. 162, no. 1–2, pp. 83–112, Mar. 2017.

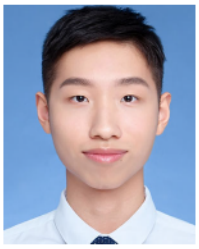
[28] Z. Allen-Zhu, "Katyusha: The first direct acceleration of stochastic gradient methods," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 8194–8244, Jun. 2017.

[29] A. Bietti and J. Mairal, "Stochastic optimization with variance reduction for infinite datasets with finite sum structure," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Long Beach, USA, Dec. 2017, pp. 1622–1632.

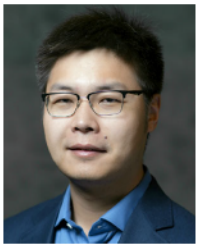
[30] K. Yuan, B. Ying, and A. H. Sayed, "COVER: A cluster-based variance reduced method for online learning," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Brighton, U.K., May 2019, pp. 3102–3106.

[31] C. Calauzenes and N. Le Roux, "Distributed SAGA: Maintaining linear convergence rate with limited communication," May 2017, *arXiv:1705.10405*.

[32] S. De and T. Goldstein, "Efficient distributed SGD with variance reduction," in *Proc. 16th Int. Conf. Data Mining*, Barcelona, Spain, Dec. 2016, pp. 111–120.
[33] S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. J. Smola, "On variance reduction in stochastic gradient descent and its asynchronous variants," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2015, pp. 2647–2655.
[34] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, Paris, France, Aug. 2010, pp. 177–186.
[35] E. Weiszfeld and F. Plastria, "On the point for which the sum of the distances to n given points is minimum," *Ann. Operations Res.*, vol. 167, no. 1, pp. 7–41, Mar. 2009.
[36] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, Fort Lauderdale, FL, USA, Apr. 2017, pp. 1273–1282.
[37] T. Chen, G. B. Giannakis, T. Sun, and W. Yin, "LAG: Lazily aggregated gradient for communication-efficient distributed learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Montreal, PQ, Canada, Dec. 2018, pp. 5055–5065.
[38] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "D2: Decentralized training over decentralized data," in *Proc. Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018, pp. 4855–4863.
[39] B. Ying, K. Yuan, and A. H. Sayed, "Variance-reduced stochastic learning under random reshuffling," *IEEE Trans. Signal Process.*, vol. 68, pp. 1390–1408, Jun. 2020.
[40] F. Lin, Q. Ling, and Z. Xiong, "Byzantine-resilient distributed large-scale matrix completion," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Brighton, U.K., May 2019, pp. 8167–8171.
[41] W. Ben-Ameur, P. Bianchi, and J. Jakubowicz, "Robust distributed consensus using total variation," *IEEE Trans. Autom. Control*, vol. 61, no. 6, pp. 1550–1564, Jun. 2016.
[42] Z. Yang and W. U. Bajwa, "BRIDGE: Byzantine-resilient decentralized gradient descent," Aug. 2019, *arXiv:1908.08098*.
[43] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Berlin, Germany: Springer, 2013.



Zhaoxian Wu received the B.E. degree in software engineering from the School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China, in 2020. He received the China National Scholarship in 2016. His research interest is distributed optimization.



Tianyi Chen received the B.Eng. degree in communication science and engineering from Fudan University in 2014, and the M.Sc. and Ph.D. degrees in electrical and computer engineering (ECE) from the University of Minnesota (UMN) in 2016 and 2019, respectively.

Since August 2019, he is with Department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute as an Assistant Professor. During 2017-2018, he has been a Visiting Scholar at Harvard University, the University of California, Los Angeles, and the University of Illinois Urbana-Champaign. He was a Best Student Paper Award finalist in the 2017 Asilomar Conference on Signals, Systems, and Computers. He received the National Scholarship from China in 2013, the UMN ECE Department Fellowship in 2014, and the UMN Doctoral Dissertation Fellowship in 2017.

His research interests lie in optimization and statistical signal processing with applications to machine learning and wireless networks.

Qing Ling received the B.E. degree in automation and Ph.D. degree in control theory and control engineering from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively. He was a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, Michigan Technological University, Houghton, MI, USA, from 2006 to 2009 and an Associate Professor with the Department of Automation, University of Science and Technology of China, from 2009 to 2017. He is currently a Professor with the School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China. His current research interest includes decentralized network optimization and its applications. He received the 2017 IEEE Signal Processing Society Young Author Best Paper Award as a Supervisor. He is an Associate Editor of the IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT AND IEEE SIGNAL PROCESSING LETTERS.

Georgios B. Giannakis (Fellow, IEEE) received the diploma in electrical engineering from the National Technical University of Athens, Greece, 1981. From 1982 to 1986, he was with the University of Southern California (USC), where he received the M.Sc. in electrical engineering, in 1983, M.Sc. in mathematics in 1986, and Ph.D. in electrical engineering in 1986. He was a Faculty Member with the University of Virginia from 1987 to 1998, and since 1999 he has been a Professor with the University of Minnesota, where he holds an ADC Endowed Chair, a University of Minnesota McKnight Presidential Chair in ECE, and serves as Director of the Digital Technology Center. His general interests span the areas of statistical learning, communications, and networking–subjects on which he has published more than 465 journal papers, 765 conference papers, 25 book chapters, two edited books and two research monographs. Current research focuses on Data Science, and Network Science with applications to the Internet of Things, and power networks with renewables. He is the co-inventor of 33 issued patents, and the co-recipient of 9 best journal paper awards from the IEEE Signal Processing (SP) and Communications Societies, including the G. Marconi Prize Paper Award in wireless communications. He also received the IEEESPS Nobert Wiener Society Award (2019); EURASIP's A. Papoulis Society Award (2020); Technical Achievement Awards from the IEEE-SPS (2000) and from EURASIP (2005); the IEEE ComSoc Education Award (2019); the G. W. Taylor Award for Distinguished Research from the University of Minnesota, and the IEEE Fourier Technical Field Award (2015). He is a Fellow of the National Academy of Inventors, the European Academy of Sciences, IEEE and EURASIP. He has served the IEEE in a number of posts, including that of a Distinguished Lecturer for the IEEE-SPS.