Byzantine-Resilient Distributed Finite-Sum Optimization over Networks

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

***Abstract*—In this paper, we investigate the problem of distributed finite-sum optimization in presence of malicious attacks from Byzantine workers. Existing Byzantine-resilient algorithms often combine stochastic gradient descent (SGD) with various robust aggregation rules to handle malicious attacks. However, the large gradient noise of SGD brings difficulty to distinguish malicious messages sent by the Byzantine workers from noisy stochastic gradients sent by the honest workers. This fact motivates us to reduce the gradient noise so as to achieve better performance than Byzantine-resilient SGD. Therefore, we propose Byrd-SAGA, a Byzantine-resilient version of distributed SAGA to deal with the malicious attacks in the distributed finite-sum optimization setting. Byrd-SAGA uses geometric median to aggregate the corrected stochastic gradients sent by the distributed workers, other than uses mean in distributed SAGA. When less than half of the workers are Byzantine, the robustness of geometric median to outliers enables Byrd-SAGA to achieve provable linear convergence to a neighborhood of the optimal solution, where the size of neighborhood is determined by the number of Byzantine workers. Numerical experiments demonstrate the robustness of Byrd-SAGA to various Byzantine attacks, as well as the advantage of Byrd-SAGA over Byzantine-resilient SGD.**

***Index Terms*—Byzantine-resilience, distributed finite-sum optimization, variance reduction**

I. Introduction

With the fast development of information technologies, the volume of distributed big data increases explosively. Every day, distributed devices (e.g., sensors, cellphones, computers, vehicles, etc) generate a huge amount of data, which are often transmitted to datacenters for processing and learning. However, collecting the data from the distributed devices and storing them in the datacenters raise significant privacy concerns [1]–[3]. To address this issue, federated learning has been proposed as a new privacy-preserving distributed data processing and machine learning framework [4]. In federated learning, the data are kept privately by and the computation is assigned to the distributed devices. Iteratively, the distributed devices calculate their local variables (e.g., stochastic gradients, corrected stochastic gradients, models, etc) using the private data samples, while the datacenter aggregates the local variables and disseminates the aggregated result to the distributed devices.

Nevertheless, the distributed nature of federated learning makes it vulnerable to errors and attacks. Some of the distributed devices can be unreliable in either computation or communication, while some can be hacked by malicious attackers. These distributed devices may send arbitrary malicious messages to the datacenter, aiming at misleading the learning process [5]–[7]. We call these arbitrary malicious attacks as Byzantine attacks [8]. It is crucial to develop robust federated learning algorithms to handle these Byzantine attacks for secure processing and learning.

In view of the challenge in Byzantine-resilient federated learning, various robust aggregation rules have been developed in these years, mainly focused on improving the distributed stochastic gradient descent (SGD) method. Through aggregating stochastic gradients with geometric median [9], [10], median [11], trimmed mean [12] or iterative filtering [13], Byzantine-resilient distributed SGD is able to tolerate the attacks from a small number of Byzantine devices. Other aggregation rules include Krum [14] which selects a stochastic gradient having the minimal summation of squared distances from a given number of nearest stochastic gradients, and RSA [15] which aggregates models other than stochastic gradients through penalizing the differences between the local models. Related works also include [16] which considers adversarial learning in distributed principal component analysis, [17] which investigates how to escape from saddle points in non-convex distributed learning under Byzantine attacks, and [18], [19] which use redundant gradients at the datacenter to improve the robustness.

Although these Byzantine-resilient SGD methods are often guaranteed to reach a neighborhood of the Byzantine-free optimal solution, the size of the neighborhood can be large under well-designed Byzantine attacks [20]. Essentially, SGD suffers from large gradient noise. This disadvantage leads to the key difficulty in distinguishing the malicious messages sent by the Byzantine attackers from the noisy stochastic gradients sent by the honest devices.

Considering the deficiency of Byzantine-resilient distributed SGD, we ask: *Can we better distinguish the malicious messages from the stochastic gradients through reducing the gradient noise?* Our answer is *Yes*. When the

Zhaoxian Wu and Qing Ling are with School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, Guangdong 510006, China. Tianyi Chen is with Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, New York 12180, USA. Georgios B. Giannakis is with Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, Minnesota 55455, USA. Part of this paper has been submitted to IEEE International Conference on Acoustics, Speech, and Signal Processing, Barcelona, Spain, May 4-8, 2020. Corresponding Email: lingqing556@mail.sysu.edu.cn.

gradient noise is small, the malicious messages are easy to be identified (see the illustrative example in Section **??**). This observation suggests the combination of variance reduction techniques with robust aggregation rules to handle Byzantine attacks in federated learning.

Existing variance reduction techniques in stochastic optimization include mini-batch [21], SAG [22], SVRG [23], SAGA [24], SDCA [25], SARAH [26], Katyusha [27], to name a few. Among these algorithms, we are particularly interest in SAGA, which has been proven to be effective in finite-sum optimization. SAGA can also be implemented in a distributed manner [28]–[30], and is hence fit for the federated learning applications where every distributed device has a finite number of data samples.

In this paper, we propose Byrd-SAGA, which combines the variance reduction technique of SAGA with robust aggregation to deal with the malicious attacks in the distributed finite-sum optimization setting. In Byrd-SAGA, the data-center uses geometric median to aggregate the corrected stochastic gradients sent by the distributed devices, other than mean in distributed SAGA. Through reducing the gradient noise, Byrd-SAGA is able to achieve better performance than Byzantine-resilient distributed SGD. When less than half of the devices are Byzantine, robustness of the geometric median to outliers enables Byrd-SAGA to achieve provable linear convergence to a neighborhood of the optimal solution, where the size of neighborhood is determined by the number of Byzantine devices. Numerical experiments demonstrate the robustness of Byrd-SAGA to various Byzantine attacks.

II. Problem Statement

In this section, we describe the distributed finite-sum optimization problem in presence of Byzantine workers. We show the deficiency of Byzantine-robust distributed SGD algorithms, which motivates the development of Byrd-SAGA.

A. Distributed finite-sum optimization in presence of Byzantine workers

Consider a distributed network with one master node (data-acenter) and W workers (devices), among which B workers are Byzantine but their identities are unknown to the master node. Denote the set of all workers as $\mathcal{W}$ and the set of Byzantine workers as $\mathcal{B}$ (hence $|\mathcal{W}| = W$ and $|\mathcal{B}| = B$). The data samples are evenly distributed across the honest workers $w \notin \mathcal{B}$. Every honest worker has n data samples, and we use $f_{w,j}(x)$ to denote the loss function of the j-th data sample on the honest worker w with respect to the model $x \in \mathbb{R}^p$. We are interested in the finite-sum optimization problem in the form of

$$x^* = \arg\min_x \ f(x), \tag{1}$$

where

$$f(x) := \frac{1}{W-B}\sum_{w\notin B} f_w(x), \ f_w(x) := \frac{1}{n}\sum_{j=1}^{n} f_{w,j}(x). \tag{2}$$

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

B. Sensitivity of distributed SGD to Byzantine attacks

When all the workers are honest, one of the most popular algorithms to solve the distributed finite-sum optimization problem is SGD [31]. At time k, the master node sends the current model x^k to the workers. Upon receiving x^k, every worker w uniformly at random chooses a local data sample with index i_w^k to calculate a stochastic gradient $f'_{w,i_w^k}(x^k)$ and sends back to the master node. After collecting all the stochastic gradients from the workers, the master node updates the model by

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

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

While the honest workers send true stochastic gradients to the master node, the Byzantine workers may not do so. In fact, Byzantine workers can send arbitrary malicious messages to the master node, aiming at biasing the optimization process. We use m_w^k to denote the message sent from worker w to the master node at time k, given by

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

where $*$ represents an arbitrary p-dimensional vector. Then, the distributed SGD update (3) becomes

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

Even when only one Byzantine worker is present, the distributed SGD may fail. Let w_b be the Byzantine worker. It can send to the master node $m_{w_b}^k = -\sum_{w\neq w_b} m_w^k$ which makes $x^{k+1} = x^k$, or send $m_{w_b}^k = \infty$ which blows up the update. In these two simple examples, the master node

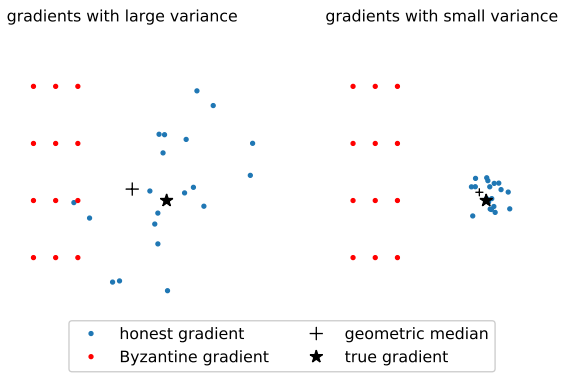

Fig. 1. Impact of gradient noise on geometric median-based robust aggregation. Blue dots denote stochastic gradients sent by the honest workers. Red dots denote malicious messages sent by the Byzantine workers. Plus signs denote the outputs of geometric median-based robust aggregation. Pentagrams denote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

can easily detect the Byzantine attacks. But in practice, the Byzantine workers can send more tricky messages to fool the master node, and bias the optimization process.

C. Byzantine-resilient distributed SGD

Recent works often robustify the distributed SGD by incorporating the robust aggregation rules when the master node receives messages from the workers. In particular, we will focus on the application and analysis of geometric median, while other robust aggregation rules are also viable [9], [10].

Let $\{z : z \in \mathcal{Z}\}$ be a subset in a normed space. The geometric median of $\{z\}$ is defined as

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

With this definition, we can modify the distributed SGD in (5) to a Byzantine-resilient form of

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

In essence, the geometric median chooses a reliable vector to represent the received messages $\{m_w^k\}$ through majority voting. When the number of Byzantine workers $B < \frac{W}{2}$, the geometric median provides a reasonable approximate to the mean of $\{m_w^k, w \notin \mathcal{B}\}$. This property enables the Byzantine-resilient distributed SGD to converge to a neighborhood of the optimal solution [9], [10].

D. Impact of gradient noise on robust aggregation

In distributed SGD, the stochastic gradients calculated by the honest workers are noisy because of the randomness in choosing data samples. Due to the existence of gradient noise, it is not always easy to distinguish the malicious messages from the stochastic gradients using the robust aggregation rules, such as geometric median. Several existing works have noticed this issue. Under well-designed Byzantine attacks, outputs of several Byzantine-resilient SGD algorithms can be far away from the optimal solution [20]. In [10] and [18], the workers are divided into several groups, such that averages are taken within the groups and geometric median is taken among the groups. This approach leads to reduced variance and thus stronger ability to distinguish malicious messages. In [14], it is explicitly assumed that that the ratio of the variance of stochastic gradients to the distance between iterate and optimal solution is upper-bounded.

Fig. 1 depicts the impact of gradient noise on geometric median-based robust aggregation. When the variance of the stochastic gradients sent by the honest workers is smaller, the gap between the true mean and the aggregated value is also smaller. That is to say, the same Byzantine attacks are less effective. We will theoretically justify this statement in the theoretical analysis.

Motivated by this fact, we propose to reduce the gradient noise in Byzantine-resilient SGD so as to achieve better robustness to Byzantine attacks. In the Byzantine-free case, an effective approach to alleviate the gradient noise of SGD is variance reduction. Through correcting the noise in the stochastic gradients, variance reduction techniques enable the algorithms to converge to the exact optimal solution with constant step sizes and achieve linear rates when the loss function $f(x)$ is strongly convex. In this paper, we focus on SAGA, which reduces gradient noise for finite-sum optimization [24], and will show that this technique effectively helps robust aggregation against Byzantine attacks.

III. Algorithm Development

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

A. Distributed SAGA with mean aggregation

In distributed SAGA, every worker maintains a table of stochastic gradients for all of its local data samples [28], [29]. Like the distributed SGD, at time k, the master node sends the current model x^k to the workers and every worker w uniformly at random chooses a local data sample with index i_w^k to calculate a 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 of all the local data samples. Then, worker w sends the 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 all of the corrected stochastic gradients from the workers, the master node updates the model x^{k+1}.

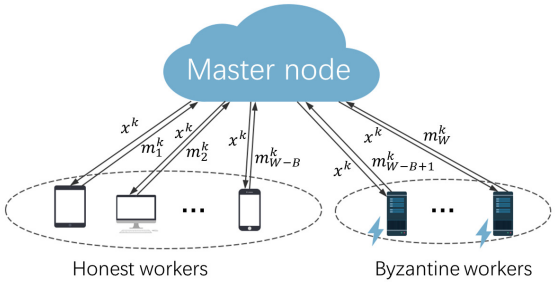

Fig. 2. Illustration of Byzantine-resilient distributed SAGA. For the ease of illustration, the honest workers are from 1 to $W-B$ while the Byzantine workers are from $W-B+1$ to W. But in practice, the identities of Byzantine workers are unknown to the master node.

To better describe distributed SAGA, define

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

Note that $\phi_{w,j}^{k+1}$ is the iterate at which the most recent $f'_{w,j}$ is evaluated after time k ends. Then, $f'_{w,j}(\phi_{w,j}^k)$ refers to the the previously stored stochastic gradient of the j-th data sample prior to time k on worker w, and $f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(\phi_{w,i_w^k}^k) + \frac{1}{n}\sum_{j=1}^n f'_{w,j}(\phi_{w,j}^k)$ is the corrected stochastic gradient of worker i at time k. The model update of SAGA is hence

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{W}\sum_{w=1}^{W}\Big(f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(\phi_{w,i_w^k}^k) + \frac{1}{n}\sum_{j=1}^{n} f'_{w,j}(\phi_{w,j}^k)\Big), \tag{9}$$

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

B. Distributed SAGA with geometric median aggregation

However, a Byzantine worker may send malicious messages, other than the corrected stochastic gradient, to the master node. We use m_w^k to denote the message sent from worker w to the master node at time k, given by

$$m_w^k = \begin{cases} f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(\phi_{w,i_w^k}^k) \\ \quad + \frac{1}{n}\sum_{j=1}^n f'_{w,j}(\phi_{w,j}^k), & w \notin \mathcal{B}, \\ *, & w \in \mathcal{B}, \end{cases} \tag{10}$$

where $*$ is an arbitrary p-dimensional vector. Similar to distributed SGD, distributed SAGA is also sensitive to Byzantine attacks. Here we propose to use geometric median as the robust aggregation rule. Thus, distributed SAGA in (9) can be modified to a Byzantine-resilient form of

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

The proposed Byzantine-resilient distributed SAGA, abbreviated as Byrd-SAGA, is outlined in Algorithm 1 and illustrated in Fig. 2. Note that there are various implementations of the distributed SAGA. For example, [29] proposes to storing the tables of the stochastic gradients in the master node. The workers only need to upload the stochastic gradients and their indexes, while the master node takes care of the aggregation. This setup is also vulnerable to Byzantine attacks, since the Byzantine workers may upload incorrect stochastic gradients. The proposed robust aggregation rule can also be applied therein.

Algorithm 1 Byzantine-Resilient Distributed SAGA

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

Master node and honest workers initialize x^0

for all honest worker w **do**

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

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

  end for

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

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

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

  end for

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

end for

There exist other robust aggregation rules besides geometric median, such as median [11], Krum [14], marginal trimmed mean [12], iterative filtering [13], etc. Take median and Krum as examples. In median, the aggregation outputs the element-wise median of the vectors $\{m_w^k\}$. In Krum, the aggregation outputs

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

where $w \to w'$ $(w \neq w')$ selects the indexes w' of the $W-B-2$ nearest neighbors of m_w^k in $\{m_{w'}^k\}$. Note that Krum needs to know B, the number of Byzantine workers, in advance. In addition, other variance reduction techniques, such as mini-batch [21], SAG [22], SVRG [23], SAGA [24], SDCA [25], SARAH [26] and Katyusha [27], are also available to alleviate the gradient noise.

Due to the page limit, in this paper we focus on the combination of geometric median and SAGA. Extending the current work to other robust aggregation rules, as well as various variance reduction techniques, is our future work.

Remark 1. *Note that calculating the geometric median needs to solve an optimization problem in the form of* (6)*. It is time-consuming to calculate the exact geometric median, and we are often satisfied with an ϵ-approximate value [32]. 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 inexact computation slightly affects the convergence of Byrd-SAGA.

IV. Theoretical Analysis

In this section, we theoretically justify the intuitive idea that reducing gradient noise helps identify malicious messages in robust aggregation (to be specific, geometric median in this paper). We also prove that the proposed Byrd-SAGA converges to a neighborhood of the optimal solution at a linear rate under Byzantine attacks, and the size of the neighborhood is determined by the number of Byzantine workers. Proofs of the theoretical results are given in appendices.

A. Importance of reducing gradient noise

The influence of gradient noise on the geometric median aggregation can be demonstrated by the following lemma.

Lemma 1. *(Concentration property) Let $\{z : z\in\mathcal{Z}\}$ 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*

$$\begin{aligned}&E\|\operatorname*{geomed}_{z\in\mathcal{Z}}\{z\}-\bar{z}\|^2 \\ \leq&2C_\alpha^2\frac{\sum_{z\notin\mathcal{Z}'}E\|z-Ez\|^2}{|\mathcal{Z}|-|\mathcal{Z}'|}+2C_\alpha^2\frac{\sum_{z\notin\mathcal{Z}'}\|Ez-\bar{z}\|^2}{|\mathcal{Z}|-|\mathcal{Z}'|},\end{aligned} \tag{13}$$

where

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

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

Assume that $\mathcal{Z}$ is the set of messages sent by all the workers in $\mathcal{W}$ and $\mathcal{Z}'$ is the set of malicious messages sent by the Byzantine workers in $\mathcal{B}$. Then, $\bar{z}$ denotes the true gradient (averaged expectation of the stochastic gradients) and the left-hand side of (13) is the variation of the geometric median with respect to the true gradient. The upper bound in the right-hand side of (13) consists of two terms. The first term 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 Byzantine-resilient SGD, the upper bound can be large due to the large gradient noise of SGD. Through reducing the gradient noise in terms of either inner variation or outer variation, we are able to achieve better accuracy under malicious messages.

B. Convergence of Byrd-SAGA

Now we show the convergence property of Byrd-SAGA and demonstrate its robustness to Byzantine attacks. We begin with several assumptions on the functions $\{f_{w,i}\}$.

Assumption 1. *(Strong convexity and Lipschitz continuous gradients) Each $f_{w,i}$ is μ-strongly convex and has L-Lipschitz continuous gradients. That is, for any $x,y\in\mathbb{R}^p$, we have*

$$f_{w,i}(x)\geq f_{w,i}(y)+\langle f'_{w,i}(y),x-y\rangle+\frac{\mu}{2}\|x-y\|^2, \tag{14}$$

and

$$\|f'_{w,i}(x)-f'_{w,i}(y)\|\leq L\|x-y\|. \tag{15}$$

Assumption 2. *(Bounded gradients) Each $f_{w,i}$ has bounded gradients. That is, for any $x\in\mathbb{R}^p$, we have*

$$\|f'_{w,i}(x)\|\leq r. \tag{16}$$

Assumption 3. *(Bounded inner variation) For any 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 as*

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

Assumption 4. *(Bounded outer variation) For any $x\in\mathbb{R}^p$, the variation of the aggregated gradients at the honest workers with respect to the overall gradient is upper-bounded as*

$$E_{w\notin\mathcal{B}}\|f'_w(x)-f'(x)\|^2\leq\delta^2. \tag{18}$$

Assumptions 1 and 2 are standard in convex analysis. Assumptions 3 and 4 bound the variations of stochastic gradients within and across the honest workers, respectively [33]. These two assumptions are mild. For instance, most of the existing works of Byzantine-resilient SGD assume that the stochastic gradients at the honest workers are independently and identically distributed (i.i.d.) with finite variance, such that σ^2 in Assumption 3 is finite and δ^2 in Assumption 4 is zero. With Assumption 4, we allow the the stochastic gradients to be non-i.i.d. at the honest workers.

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

The use of geometric median in Byrd-SAGA brings difficulties to the analysis. To be specific, for every honest worker $w\notin\mathcal{B}$, m_i^k is an unbiased estimate of $f'_w(x^k)$, because

$$Em_i^k=f'_w(x^k). \tag{19}$$

Averaging (19) over all honest workers $w\notin\mathcal{B}$, we have

$$\frac{1}{W-B}\sum_{w\notin\mathcal{B}}Em_i^k=\frac{1}{W-B}\sum_{w\notin\mathcal{B}}f'_w(x^k)=f'(x^k). \tag{20}$$

From (20), we observe that the mean of m_i^k over all the honest workers $w\notin\mathcal{B}$ provides an unbiased estimate of

$f'(x^k)$. Nevertheless, the geometric median of m_i^k, even only over all the honest workers $w \notin \mathcal{B}$ and calculated accurately, is a biased estimate of $f'(x^k)$.

To handle the bias, the following lemma characterizes the error between an ϵ-approximate geometric median of $\{m_w^k\}$ and the overall gradient $f'(x^k)$ at time k.

Lemma 2. *Under Assumptions 2, 3 and 4, if the number of Byzantine workers $B < \frac{W}{2}$, then an ϵ-approximate geometric median of $\{m_w^k\}$, denoted by z_ϵ^*, satisfies*

$$E\|z_\epsilon^* - f'(x^k)\|^2 \leq \xi, \tag{21}$$

where

$$\xi := C_\alpha^2(4\sigma^2 + 16r^2 + 2\delta^2) + \frac{2\epsilon^2}{(W-2B)^2}, \tag{22}$$

$$C_\alpha := \frac{2-2\alpha}{1-2\alpha} \quad \textit{and} \quad \alpha := \frac{B}{W}. \tag{23}$$

The following theorem shows that Byrd-SAGA converges to a neighborhood of the optimal solution x^* at a linear rate, and the size of the neighborhood is determined by the number of Byzantine workers.

Theorem 1. *Under Assumptions 1, 2, 3 and 4, if the number of Byzantine workers $B < \frac{W}{2}$ and the step size $\gamma < \min\{\frac{2}{n\mu+32C_\alpha^2 L}, \frac{1}{8LC_\alpha^2}\}$, then for Byrd-SAGA with ϵ-approximate geometric median, it holds*

$$E\|x^k - x^*\|^2 \leq (1 - \frac{\gamma\mu}{2})^k \Delta_1 + \Delta_2, \tag{24}$$

where

$$\Delta_1 := \|x^0 - x^*\|^2 + 2\gamma n\left[f(x^0) - f(x^*)\right] - \Delta_2, \tag{25}$$

$$\Delta_2 := \left(\frac{8C_\alpha^2\delta^2}{\mu} + \frac{2\epsilon^2}{(W-2B)^2}\right)\gamma + \frac{4}{\mu^2}\left(\xi + \frac{2\epsilon^2}{(W-2B)^2}\right), \tag{26}$$

and ξ is a constant defined in (22)*.*

In (24), the constant of convergence rate is given by

$$1 - \frac{\mu}{2n\mu + 32C_\alpha^2 L} = 1 - \frac{1}{2n + 32C_\alpha^2 \frac{L}{\mu}},$$

which is close to 1 when n (the number of data samples at each worker) and $\frac{L}{\mu}$ (the condition number) are large. Observe that C_α is monotonically increasing when the portion of Byzantine workers α increases. Therefore, (24) demonstrates that Byrd-SAGA converges slower when more Byzantine workers are present. Correspondingly, the theoretical upper bound of step size γ is small when n and C_α are large. The size of the neighborhood Δ_2 in (26) is also monotonically increasing when C_α (and hence the number of Byzantine workers) increases.

V. Numerical Experiments

We conduct numerical experiments on convex and non-convex learning problems. Within every problem, we evenly distribute the dataset into $W - B = 50$ honest workers. For the case with Byzantine attacks, we additionally launch $B = 20$ Byzantine workers. We test the performance of the proposed Byrd-SAGA under three typical Byzantine attacks: Gaussian, max-value and zero-gradient attacks [15], [34]. With Gaussian attack, every Byzantine worker $w \in \mathcal{B}$ generates its m_w^k following a Gaussian distribution with mean $\frac{1}{W-B}\sum_{w'\notin\mathcal{B}} m_{w'}^k$ and variance 30. With max-value attack, every Byzantine worker $w \in \mathcal{B}$ sets its message as $m_w^k = u \cdot \frac{1}{W-B}\sum_{w'\notin\mathcal{B}} m_{w'}^k$, where the magnitude $u = 4$ is used in the numerical experiments. With zero-gradient attack, every Byzantine worker $w \in \mathcal{B}$ sends $m_w^k = -\frac{1}{B}\sum_{w'\notin\mathcal{B}} m_{w'}^k$ such that the messages received by the master node are summed to zero. We use the algorithm proposed in [32] to calculate ϵ-approximate geometric median with $\epsilon = 1 \times 10^{-5}$.

A. ℓ_2-regularized logistic regression

We consider the ℓ_2-regularized logistic regression problem, in which every $f_{w,i}(x)$ is in the form of

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

where $a_{w,i} \in \mathbb{R}^p$ is the feature vector, $b_{w,i} \in \{-1, 1\}$ is the label, and $\rho > 0$ is a regularizing parameter. In the numerical experiments, we fix $\rho = 0.01$. We use the IJCNN1 and COVTYPE datasets[1]. ICJNN1 contains 49,990 training data samples of $p = 22$ dimensions. COVTYPE contains 581,012 training data samples of $p = 54$ dimensions.

We first compare SGD, mini-batch SGD (BSGD) with batch size 50 and SAGA, using mean and geometric median aggregation rules. Comparing to SGD, BSGD enjoys smaller gradient noise but suffers from higher computational cost. In comparison, SAGA also reduces gradient noise, but its computational cost is in the same order as that of SGD. For every algorithm, we use the constant step size, which is tuned to achieve the best optimality gap $f(x^k) - f(x^*)$ for the Byzantine-free case. The performance of these algorithms on the IJCNN1 and COVTYPE datasets is depicted in Fig. 3 and Fig. 4, respectively. With Byzantine attacks, the three algorithms using mean aggregation all fail. Among the three algorithms using geometric median aggregation, Byrd-SAGA remarkably outperforms the other two, while BSGD is better than SGD. This fact suggests that the importance of variance reduction to handling Byzantine attacks. To be specific, for the IJCNN1 dataset, regarding the variance of honest messages, Byrd-SAGA, Byzantine-resilient BSGD and Byzantine-resilient SGD are in the order of 10^{-3}, 10^{-2} and 10^{-1}, respectively. For the COVTYPE dataset, Byrd-SAGA and Byzantine-resilient BSGD have the same order

[1]https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

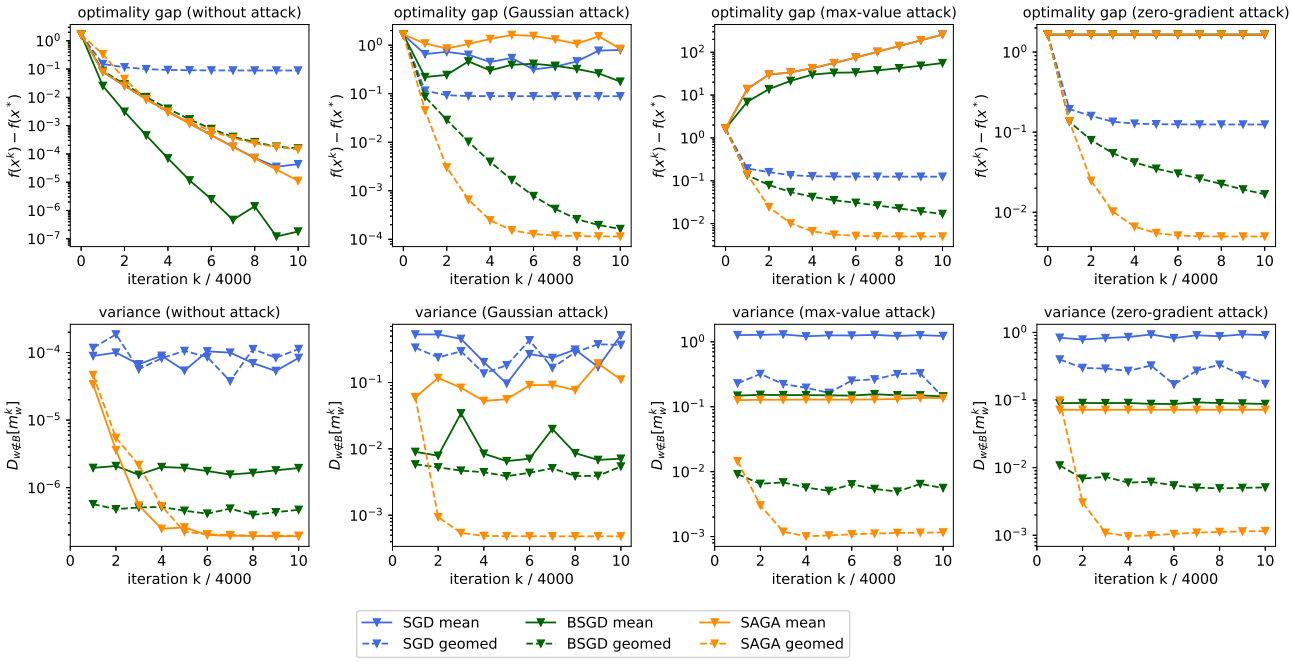

Fig. 3. Performance of the distributed SGD, mini-batch SGD (BSGD) and SAGA, with mean and geometric median (geomed) aggregation rules on IJCNN1 dataset. The step sizes are 0.02, 0.01 and 0.02, respectively. From top to bottom: optimality gap and variance of honest messages. From left to right: without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

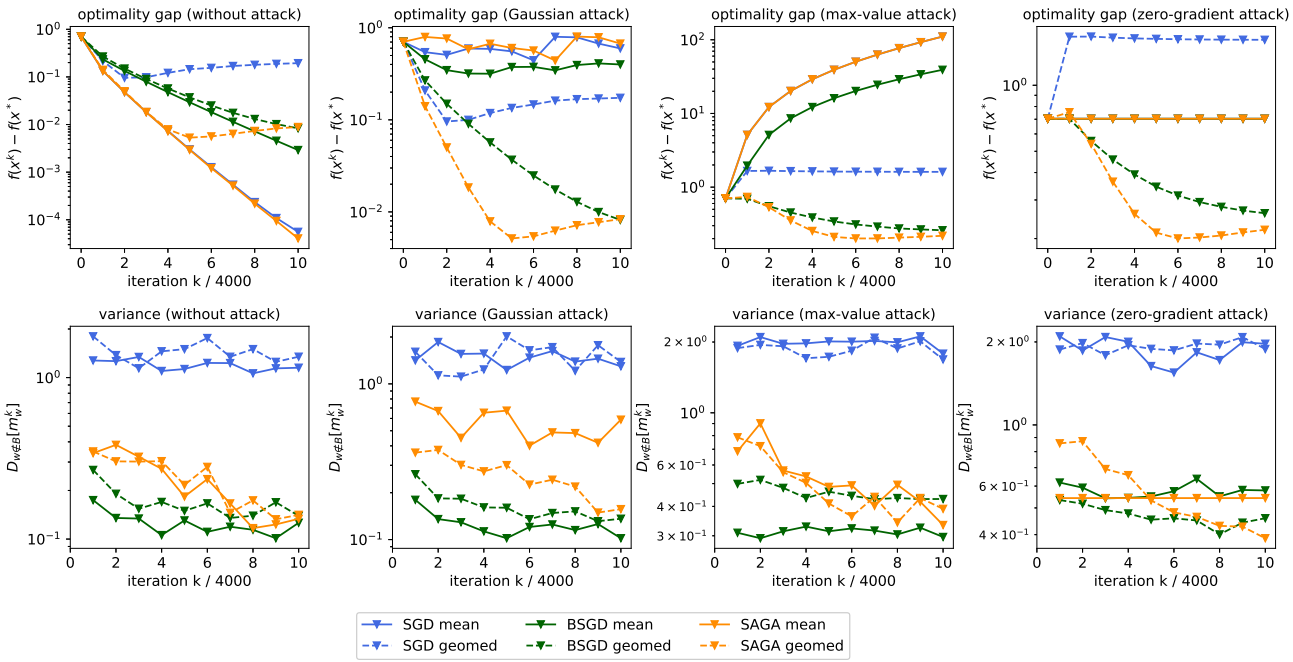

Fig. 4. Performance of the distributed SGD, mini-batch SGD (BSGD) and SAGA, with mean and geometric median (geomed) aggregation rules on COVTYPE dataset. The step sizes are 0.01, 0.005 and 0.01, respectively. From top to bottom: optimality gap and variance of honest messages. From left to right: without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

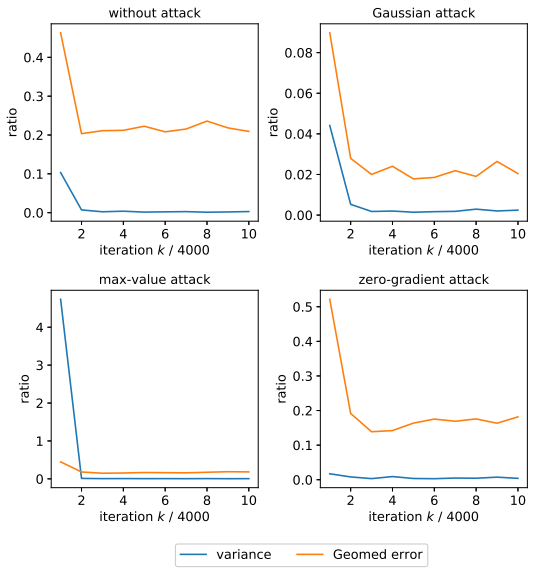

Fig. 5. Variance of the honest messages, as well as error between the geometric median of messages and the ground truth. The ratios are computed as those with corrected stochastic gradients to those with uncorrected stochastic gradients. The experiments are conducted on the IJCNN1 dataset.

of variance of honest messages. In this case, Byrd-SAGA achieves similar optimality gap as Byzantine-resilient BSGD, but converges faster because it is able to use a larger step size.

To further illustrate that Byrd-SAGA has stronger abilities of reducing gradient noise and distinguishing malicious messages than Byzantine-resilient SGD, we design the following second set of numerical experiments on the IJCNN1 dataset. At each iteration of Byrd-SAGA with step size 0.02, compute the variance of the honest corrected stochastic gradients, as well as the error between the geometric median of messages (with the honest corrected stochastic gradients and malicious messages) and the ground truth. Meanwhile, also compute the the variance of the honest uncorrected stochastic gradients, as well as the error between the geometric median of messages (with the honest uncorrected stochastic gradients and malicious messages) and the ground truth. We compute the ratios of the variances and the errors, shown in Fig. 5. Observe that the ratios are always less than 1 asymptotically, demonstrating that Byrd-SAGA does reduces gradient noise, which consequently helps reduce the error of geometric median outputs.

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. For the case with Gaussian attacks, Byrd-SAGA using geometric median achieves the best performance. For the case with max-value attacks and zero-gradient attacks,

TABLE I
ACCURACY OF SGD, MINI-BATCH SGD (BSGD) AND SAGA, WITH MEAN AND GEOMETRIC MEDIAN (GEOMED) AGGREGATION RULES.

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
max-value	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

Byrd-SAGA using Krum is the best, while that using geometric median also performs well. Note that Krum requires to know the exact number of Byzantine workers in advance, while geometric median and median do not need this prior knowledge.

B. Neural network training

We carry out a set of numerical experiments on a neural network with one hidden layer of 50 neurons and Tahn activation functions. We use this neural network for multi-class classification on the MNIST dataset, which has $60,000$ data with dimension $p = 784$. We compare SGD with step size 0.1, mini-batch SGD (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 record the final accuracy in Table 1. With mean aggregation, all the algorithms yields low accuracy under Byzantine attacks. BSGD and SAGA are both robust with the help of geometric median aggregation. Note that Byrd-SAGA has much lower iteration-wise computational cost comparing to Byzantine-resilient BSGD.

VI. CONCLUSIONS

In this paper, we propose Byrd-SAGA, a Byzantine-resilient distributed SAGA to solve the distributed finite-sum optimization problem with Byzantine attacks. Similar to SAGA, Byrd-SAGA corrects the stochastic gradient through variance reduction. At every iteration, distributed workers calculate their corrected stochastic gradients and send to the master node. But unlike SAGA, in Byrd-SAGA the master node aggregates the received messages using geometric median, other than mean. This robust aggregation rule guarantees the robustness of Byrd-SAGA in presence of Byzantine attacks. Indeed, we prove that, when less than half of the workers are Byzantine, Byrd-SAGA converges linearly to a neighborhood of the optimal solution, where the size of neighborhood is determined by the number of Byzantine workers. Through reducing the gradient noise, Byrd-SAGA has stronger ability to identify malicious messages comparing to the existing Byzantine-resilient SGD algorithms.

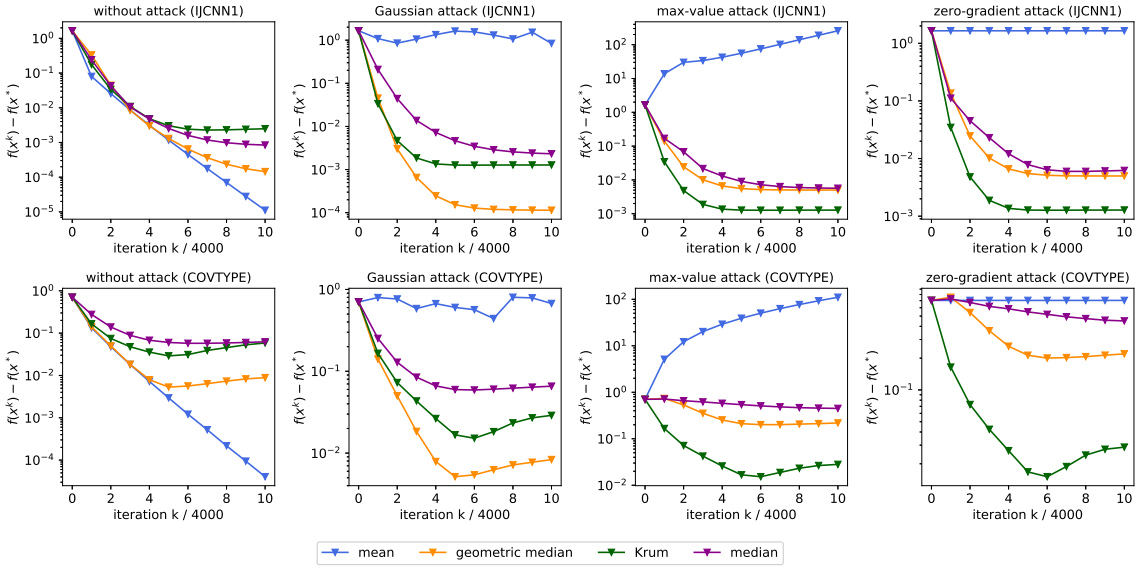

Fig. 6. Optimality gaps of the distributed SAGA with different aggregation rules: mean, geometric median, median and Krum. The step sizes are 0.02 and 0.01 for the IJCNN1 and COVTYPE datasets, respectively. From top to bottom: on IJCNN1 dataset and on COVTYPE dataset. From left to right: without attacks,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

According to the numerical experiments, combinations of other robust aggregation rules and other variance reduction techniques also demonstrate satisfactory robustness. We will leave the analysis of these algorithms as our future work. We are also interested in developing and analyzing Byzantine-resilient algorithms over decentralized networks [35], [36].

REFERENCES

[1] R. Agrawal and R. Srikant, "Privacy-preserving data mining," In: Proceedings of SIGMOD, 2000

[2] J. Duchi, M. J. Wainwright, and M. I. Jordan, "Local privacy and minimax bounds: Sharp rates for probability estimation," In: Proceedings of NeurIPS, 2013

[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 Processing Magazine, vol. 35, no. 5, pp. 76–87, 2018

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

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

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

[7] Z. Yang, A. Gang, and W. U. Bajwa, "Adversary-resilient inference and machine learning: From distributed to decentralized," arXiv Preprint arXiv:1908.08649

[8] L. Lamport, R. Shostak, and M. Pease, "The Byzantine generals problem," ACM Transactions on Programming Languages and Systems, vol. 4, no. 3, pp. 382–401, 1982

[9] S. Minsker, "Geometric median and robust estimation in Banach spaces," Bernoulli, vol. 21, no. 4, pp. 2308-C2335, 2015

[10] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," In: Proceedings of SIGMETRICS, 2019

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

[12] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," In Proceedings of ICML, 2018

[13] L. Su and J. Xu, "Securing distributed machine learning in high dimensions," arxiv Preprint arXiv:1804.10140, 2018

[14] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," In Proceedings of NeurIPS, 2017

[15] 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 Proceedings of AAAI, 2019

[16] J. Feng, H. Xu, and S. Mannor, "Distributed robust learning," arXiv Preprint arXiv:1409.5937, 2014

[17] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Defending against saddle point attack in Byzantine-robust distributed learning," arXiv Preprint arXiv:1806.05358, 2018

[18] L. Chen, H. Wang, Z. Charles, and D. Papailiopoulos, "DRACO: Byzantine-resilient distributed training via redundant gradients," arXiv Preprint arXiv:1803.09877, 2018

[19] S. Rajput, H. Wang, Z. Charles, and D. Papailiopoulos, "DETOX: A redundancy-based framework for faster and more robust gradient aggregation," arXiv Preprint arXiv:1907.12205, 2019

[20] C. Xie, O. Koyejo, and I. Gupta, "Fall of empires: Breaking Byzantine-tolerant SGD by inner product manipulation," arXiv Preprint arXiv:1903.03936, 2019

[21] P. Goyal, P. Dollar, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," arXiv Preprint arXiv:1706.02677, 2017

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

[23] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," In Proceedings of NeurIPS, 2013

[24] A. Defazio, F. R. Bach, and S. Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," In Proceedings of NeurIPS, 2014

[25] S. Shalev-Shwartz and T. Zhang, "Stochastic dual coordinate ascent methods for regularized loss minimization," Journal of Machine Learning Research, vol. 14, no. 2, pp. 567–599, 2013

[26] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takac, "SARAH: A novel method for machine learning problems using stochastic recursive gradient," In Proceedings of ICML, 2017
[27] Z. Allen-Zhu, "Katyusha: The first direct acceleration of stochastic gradient methods," Journal of Machine Learning Research, vol. 18, no. 1, pp. 8194–8244, 2017
[28] C. Calauzenes and N. Le Roux, "Distributed SAGA: Maintaining linear convergence rate with limited communication," arXiv Preprint arXiv:1705.10405, 2017
[29] S. De and T. Goldstein, "Efficient distributed SGD with variance reduction," In Proceedings of ICDM, 2016
[30] 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 Proceedings of NeurIPS, 2015
[31] L. Bottou, "Large-scale machine learning with stochastic gradient descent," In Proceedings of COMPSTAT, 2010
[32] E. Weiszfeld and F. Plastria, "On the point for which the sum of the distances to n given points is minimum," Annals of Operations Research, vol. 167, no. 1, pp. 7–41, 2009
[33] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "D2: Decentralized training over decentralized data," In Proceedings of ICML, 2018
[34] F. Lin, Q. Ling, and Z. Xiong, "Byzantine-resilient distributed large-scale matrix completion," In Proceedings of ICASSP, 2019
[35] W. Ben-Ameur, P. Bianchi, and J. Jakubowicz, "Robust distributed consensus using total variation," IEEE Transactions on Automatic Control, vol. 61, no. 6, pp. 1550–1564, 2016
[36] Z. Yang and W. U. Bajwa, "BRIDGE: Byzantine-resilient decentralized gradient descent," arXiv Preprint arXiv:1908.08098, 2019
[37] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Springer, 2013

APPENDIX A
PROOF OF LEMMA 1

The proof of Lemma 1 relies on the following supporting lemma.

Lemma 3. *Let $\{z : z \in \mathcal{Z}\}$ be a subset in a normed vector space. If $\mathcal{Z}' \subseteq \mathcal{Z}$ and $|\mathcal{Z}'| < \frac{|\mathcal{Z}|}{2}$, then it holds*

$$E\|\underset{z\in\mathcal{Z}}{\text{geomed}}\{z\}\|^2 \leq C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'} E\|z\|^2}{|\mathcal{Z}| - |\mathcal{Z}'|}, \tag{27}$$

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

Proof: Let $z^* = \text{geomed}_{z\in\mathcal{Z}}\{z\}$. For all $z \in \mathcal{Z}'$, we have $\|z^* - z\| \geq \|z\| - \|z^*\|$. 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

$$\sum_{z\in\mathcal{Z}} \|z^* - z\| \geq \sum_{z\in\mathcal{Z}} \|z\| + (|\mathcal{Z}| - 2|\mathcal{Z}'|)\|z^*\| - 2\sum_{z\notin\mathcal{Z}'} \|z\|. \tag{28}$$

According to the definition of geometric median, it holds

$$\sum_{z\in\mathcal{Z}} \|z^* - z\| = \inf_y \sum_{z\in\mathcal{Z}} \|y - z\| \leq \sum_{z\in\mathcal{Z}} \|z\|. \tag{29}$$

Combining the two inequalities yields

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

Taking squares for both sides of the inequality above yields

$$\|z^*\|^2 \leq C_\alpha^2 \frac{\left(\sum_{z\notin\mathcal{Z}'} \|z\|\right)^2}{(|\mathcal{Z}| - |\mathcal{Z}'|)^2} \leq C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'} \|z\|^2}{|\mathcal{Z}| - |\mathcal{Z}'|}. \tag{31}$$

Then taking expectations for both sides yields (27) and completes the proof. ∎

With Lemma 3, the proof of Lemma 1 is straightforward.

Proof: According to Lemma 3, it holds

$$E\|\underset{z\in\mathcal{Z}}{\text{geomed}}\{z\} - \bar{z}\|^2 = E\|\underset{z\in\mathcal{Z}}{\text{geomed}}\{z - \bar{z}\}\|^2 \leq C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'} E\|z - \bar{z}\|^2}{|\mathcal{Z}| - |\mathcal{Z}'|}. \tag{32}$$

Applying the inequality of $\|z - \bar{z}\|^2 \leq 2\|z - Ez\|^2 + 2\|Ez - \bar{z}\|^2$ to (32) yields

$$E\|\underset{z\in\mathcal{Z}}{\text{geomed}}\{z\} - \bar{z}\|^2 \leq 2C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'} E\|z - Ez\|^2}{|\mathcal{Z}| - |\mathcal{Z}'|} + 2C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'} E\|Ez - \bar{z}\|^2}{|\mathcal{Z}| - |\mathcal{Z}'|}, \tag{33}$$

which completes the proof. ∎

APPENDIX B
LEMMA 4 AND ITS PROOF

Since computing an accurate geometric median is difficult, we consider ϵ-approximate geometric median in this paper. The following lemma is an ϵ-approximate version of Lemma 3.

Lemma 4. *Let $\{z : z \in \mathcal{Z}\}$ 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*

$$E\|z_\epsilon^*\|^2 \leq 2C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'} E\|z\|^2}{|\mathcal{Z}| - |\mathcal{Z}'|} + \frac{2\epsilon^2}{(|\mathcal{Z}| - 2|\mathcal{Z}'|)^2}. \tag{34}$$

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 the z_ϵ^* is an ϵ-approximate geometric median, it holds

$$\sum_{z\in\mathcal{Z}} \|z_\epsilon^* - z\| \leq \inf_y \sum_{z\in\mathcal{Z}} \|y - z\| + \epsilon \leq \sum_{z\in\mathcal{Z}} \|z\| + \epsilon. \tag{35}$$

Notice that (28) remains valid here. Hence, we have

$$\|z_\epsilon^*\| \leq C_\alpha \frac{\sum_{z\notin\mathcal{Z}'}\|z\|}{|\mathcal{Z}|-|\mathcal{Z}'|} + \frac{\epsilon}{|\mathcal{Z}|-2|\mathcal{Z}'|}. \tag{36}$$

Taking squares for both sides of (36) leads to

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

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

$$\leq 2C_\alpha^2 \frac{\sum_{z\notin\mathcal{Z}'}\|z\|^2}{|\mathcal{Z}|-|\mathcal{Z}'|} + \frac{2\epsilon^2}{(|\mathcal{Z}|-2|\mathcal{Z}'|)^2}. \tag{39}$$

Then taking expectations for both sides yields (34) and completes the proof. ■

APPENDIX C
PROOF OF LEMMA 2

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

$$\begin{aligned}
&E\|m_w^k - f'(x^k)\|^2 \qquad (40)\\
=&E\|f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(\phi^k_{w,i_w^k}) + \frac{1}{n}\sum_{j=1}^n f'_{w,j}(\phi^k_{w,j}) - f'(x^k)\|^2\\
=&E\|f'_{w,i_w^k}(x^k) - f'_w(x^k) - f'_{w,i_w^k}(\phi^k_{w,i_w^k}) + \frac{1}{n}\sum_{j=1}^n f'_{w,j}(\phi^k_{w,j})\|^2 + \|f'_w(x^k) - f'(x^k)\|^2\\
\leq&2E\|f'_{w,i_w^k}(x^k) - f'_w(x^k)\|^2 + 2E\|f'_{w,i_w^k}(\phi^k_{w,i_w^k}) - \frac{1}{n}\sum_{j=1}^n f'_{w,j}(\phi^k_{w,j})\|^2 + \|f'_w(x^k) - f'(x^k)\|^2.
\end{aligned}$$

To derive the second equality, we use the variance decomposition $E\|a\|^2 = E\|a - Ea\|^2 + \|Ea\|^2$ with

$$a = f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(\phi^k_{w,i_w^k}) + \frac{1}{n}\sum_{j=1}^n f'_{w,j}(\phi^k_{w,j}) - f'(x^k),$$

where $Ea = f'_w(x^k) - f'(x^k)$ due to the fact that

$$E[f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(\phi^k_{w,i_w^k}) + \frac{1}{n}\sum_{j=1}^n f'_{w,j}(\phi^k_{w,j})] = f'_w(x^k),$$

as we have shown in (19). The inequality comes from $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. The first term at the right-hand side of (40) can be bounded with Assumption 3. For the second term, it holds

$$\begin{aligned}
&E\|f'_{w,i_w^k}(\phi^k_{w,i_w^k}) - \frac{1}{n}\sum_{j=1}^n f'_{w,j}(\phi^k_{w,j})\|^2 \qquad (41)\\
=&E\|\frac{1}{n}\sum_{j=1}^n \left(f'_{w,i_w^k}(\phi^k_{w,i_w^k}) - f'_{w,j}(\phi^k_{w,j})\right)\|^2\\
\leq&\frac{1}{n}\sum_{j=1}^n E\|f'_{w,i_w^k}(\phi^k_{w,i_w^k}) - f'_{w,j}(\phi^k_{w,j})\|^2\\
\leq&\frac{1}{n}\sum_{j=1}^n E\left(\|f'_{w,i_w^k}(\phi^k_{w,i_w^k})\| + \|f'_{w,j}(\phi^k_{w,j})\|\right)^2 \leq 4r^2.
\end{aligned}$$

Here the first inequality comes from $\|\frac{1}{n}\sum_{j=1}^n a_j\|^2 \leq \frac{1}{n}\sum_{j=1}^n \|a_j\|^2$ with

$$a_j = f'_{w,i_w^k}(\phi_{w,i_w^k}^k) - f'_{w,j}(\phi_{w,j}^k).$$

The second inequality comes from the triangle inequality. The third inequality uses the fact of bounded gradients in Assumption 4. Substituting (17) in Assumption 3 and (41) into (40), we obtain for any $w \in \mathcal{W}$ that

$$E\|m_w^k - f'(x^k)\|^2 \leq 2\sigma^2 + 8r^2 + \|f'_w(x^k) - f'(x^k)\|^2. \tag{42}$$

Denote z_ϵ^* as an ϵ-approximate geometric median of $\{m_w^k\}$ and $\tilde{z}_\epsilon^*$ as an ϵ-approximate geometric median of $\{m_w^k - f'(x^k)\}$. Now we derive an upper bound for $E\|z_\epsilon^* - f'(x^k)\|^2$. Since $\tilde{z}_\epsilon^* = z_\epsilon^* - f'(x^k)$, according to (34) in Lemma 4 and (42), it holds

$$\begin{aligned}
&E\|z_\epsilon^* - f'(x^k)\|^2 = E\|\tilde{z}_\epsilon^*\|^2 \\
\leq& 2C_\alpha^2 \frac{\sum_{w\notin\mathcal{B}} E\|m_w^k - f'(x^k)\|^2}{W-B} + \frac{2\epsilon^2}{(|\mathcal{Z}| - 2|\mathcal{Z}'|)^2} \\
\leq& 2C_\alpha^2 \left(2\sigma^2 + 8r^2 + \frac{1}{W-B}\sum_{w\notin\mathcal{B}} \|f'_w(x^k) - f'(x^k)\|^2\right) + \frac{2\epsilon^2}{(|\mathcal{Z}| - 2|\mathcal{Z}'|)^2}.
\end{aligned} \tag{43}$$

Note that $\frac{1}{W-B}\sum_{w\notin\mathcal{B}} \|f'_w(x^k) - f'(x^k)\|^2 = E_{w\notin\mathcal{B}}\|f'_w(x^k) - f'(x^k)\|^2$, which is no larger than δ^2 according to (18) in Assumption 4. Applying this fact to (43) immediately yields (24) and completes the proof. ■

APPENDIX D
LEMMA 5 AND ITS PROOF

On top of Lemma 2, the following lemma further gives an upper bound for the ϵ-approximate geometric median of $\{m_w^k\}$ at time k.

Lemma 5. *Under Assumptions 1 and 4, if the number of Byzantine workers $B < \frac{W}{2}$ and z_ϵ^* is an ϵ-approximate geometric median of $\{m_w^k\}$, then z_ϵ^* satisfies*

$$E\|z_\epsilon^*\|^2 \leq 4C_\alpha^2(1+\beta)S_1^k - 8C_\alpha^2\mu\beta S_2^k + 8C_\alpha^2 L(1+\beta^{-1})S_3^k + 4C_\alpha^2\delta^2 + \frac{2\epsilon^2}{(W-2B)^2}, \tag{44}$$

where β is a positive constant, while S_1^k, S_2^k and S_3^k are non-negative variables defined as

$$S_1^k := \frac{1}{W-B}\sum_{w\notin\mathcal{B}}\frac{1}{n}\sum_{j=1}^n \|f'_{w,j}(x^k) - f'_{w,j}(x^*)\|^2, \tag{45a}$$

$$S_2^k := f(x^k) - f(x^*), \tag{45b}$$

$$S_3^k := \frac{1}{W-B}\sum_{w\notin\mathcal{B}}[\frac{1}{n}\sum_{j=1}^n f_{w,j}(\phi_{w,j}^k) - f_w(x^*) - \frac{1}{n}\sum_{j=1}^n \langle f'_{w,j}(x^*), \phi_{w,j}^k - x^*\rangle]. \tag{45c}$$

Proof: Similar to the proof of Lemma 2, we begin with deriving an upper bound for $E\|m_w^k - f'_w(x^*)\|^2$ where $w \notin \mathcal{B}$.

Using the definition of m_w^k in (10), we have for any $w \notin \mathcal{B}$ that

$$
\begin{aligned}
& E\|m_w^k - f'_w(x^*)\|^2 \\
=& E\|f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(\phi_{w,i_w^k}^k) + \frac{1}{n}\sum_{j=1}^n f'_{w,j}(\phi_{w,j}^k) - f'_w(x^*)\|^2 \\
=& E\|f'_{w,i_w^k}(x^k) - f'_w(x^k) - f'_{w,i_w^k}(\phi_{w,i_w^k}^k) + \frac{1}{n}\sum_{j=1}^n f'_{w,j}(\phi_{w,j}^k)\|^2 + \|f'_w(x^k) - f'_w(x^*)\|^2 \\
=& E\|(f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(x^*)) - (f'_w(x^k) - f'_w(x^*)) - (f'_{w,i_w^k}(\phi_{w,i_w^k}^k) - f'_{w,i_w^k}(x^*)) \\
& + (\frac{1}{n}\sum_{j=1}^n f'_{w,j}(\phi_{w,j}^k) - f'_w(x^*))\|^2 + \|f'_w(x^k) - f'_w(x^*)\|^2 \\
\le& (1+\beta)E\|(f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(x^*)) - (f'_w(x^k) - f'_w(x^*))\|^2 + (1+\beta^{-1}) \\
& E\|(f'_{w,i_w^k}(\phi_{w,i_w^k}^k) - f'_{w,i_w^k}(x^*)) - (\frac{1}{n}\sum_{j=1}^n f'_{w,j}(\phi_{w,j}^k) - f'_w(x^*))\|^2 + \|f'_w(x^k) - f'_w(x^*)\|^2,
\end{aligned} \tag{46}
$$

where $\beta > 0$ is any constant. The second equality comes from the variance decomposition $E\|a\|^2 = E\|a - Ea\|^2 + \|Ea\|^2$ with

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

where $Ea = f'_w(x^k)$ as we have shown in (19). The inequality comes from $\|a+b\|^2 \le (1+\beta)\|a\|^2 + (1+\beta^{-1})\|b\|^2$ which holds for any $\beta > 0$. Observing that

$$
\begin{aligned}
E[f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(x^*)] &= f'_w(x^k) - f'_w(x^*), \\
E[f'_{w,i_w^k}(\phi_{w,i_w^k}^k) - f'_{w,i_w^k}(x^*)] &= \frac{1}{n}\sum_{j=1}^n f'_{w,j}(\phi_{w,j}^k) - f'_w(x^*),
\end{aligned}
$$

we use the variance decomposition $E\|a - Ea\|^2 = E\|a\|^2 - \|Ea\|^2$ twice and obtain an upper bound of (46) as

$$
\begin{aligned}
& E\|m_w^k - f'_w(x^*)\|^2 \\
\le& (1+\beta)E\|f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(x^*)\|^2 - (1+\beta)\|f'_w(x^k) - f'_w(x^*)\|^2 + \|f'_w(x^k) - f'_w(x^*)\|^2 \\
& + (1+\beta^{-1})E\|f'_{w,i_w^k}(\phi_{w,i_w^k}^k) - f'_{w,i_w^k}(x^*)\|^2 - (1+\beta^{-1})\|\frac{1}{n}\sum_{j=1}^n f'_{w,j}(\phi_{w,j}^k) - f'_w(x^*)\|^2 \\
\le& (1+\beta)E\|f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(x^*)\|^2 - \beta\|f'_w(x^k) - f'_w(x^*)\|^2 + (1+\beta^{-1})E\|f'_{w,i_w^k}(\phi_{w,i_w^k}^k) - f'_{w,i_w^k}(x^*)\|^2.
\end{aligned} \tag{47}
$$

According to Theorem 2.1.5 of [37], we have

$$
\begin{aligned}
\|f'_w(x^k) - f'_w(x^*)\|^2 &\ge 2\mu[f_w(x^k) - f_w(x^*) - \langle f'_w(x^*), x^k - x^*\rangle], \\
\|f'_{w,i_w^k}(\phi_{w,i_w^k}^k) - f'_{w,i_w^k}(x^*)\|^2 &\le 2L[f_{w,i_w^k}(\phi_{w,i_w^k}^k) - f_{w,i_w^k}(x^*) - \langle f'_{w,i_w^k}(x^*), \phi_{w,i_w^k}^k - x^*\rangle].
\end{aligned}
$$

Applying these two inequalities into (47) yields

$$
\begin{aligned}
& E\|m_w^k - f'_w(x^*)\|^2 \\
\le& (1+\beta)E\|f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(x^*)\|^2 - 2\mu\beta[f_w(x^k) - f_w(x^*) - \langle f'_w(x^*), x^k - x^*\rangle] \\
& + 2L(1+\beta^{-1})E[f_{w,i_w^k}(\phi_{w,i_w^k}^k) - f_{w,i_w^k}(x^*) - \langle f'_{w,i_w^k}(x^*), \phi_{w,i_w^k}^k - x^*\rangle].
\end{aligned} \tag{48}
$$

Now we derive an upper bound for $E\|m_w^k\|^2$. Using $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, we have

$$
E\|m_w^k\|^2 \le 2\|f'_w(x^*)\|^2 + 2E\|m_w^k - f'_w(x^*)\|^2. \tag{49}
$$

Noticing that $f'(x^*) = 0$ and applying Assumption 4, we know that

$$
\frac{1}{W-B}\sum_{w\notin\mathcal{B}}\|f'_w(x^*)\|^2 = E_{w\notin\mathcal{B}}\|f'_w(x^*) - f'(x^*)\|^2 \le \delta^2. \tag{50}
$$

Therefore, according to (34) in Lemma 4, it holds

$$\begin{aligned}&E\|z_\epsilon^*\|^2\\ \leq&2C_\alpha^2\frac{\sum_{w\notin\mathcal{B}}E\|m_w^k\|^2}{W-B}+\frac{2\epsilon^2}{(W-2B)^2}\\ \leq&4C_\alpha^2\frac{\sum_{w\notin\mathcal{B}}E\|m_w^k-f_w'(x^*)\|^2}{W-B}+4C_\alpha^2\delta^2+\frac{2\epsilon^2}{(W-2B)^2}.\end{aligned}\tag{51}$$

Substituting (48) into (51) yields

$$\begin{aligned}&E\|z_\epsilon^*\|^2\\ \leq&\frac{4C_\alpha^2(1+\beta)}{W-B}\sum_{w\notin\mathcal{B}}E\|f_{w,i_w^k}'(x^k)-f_{w,i_w^k}'(x^*)\|^2\\ &-\frac{8C_\alpha^2\mu\beta}{W-B}\sum_{w\notin\mathcal{B}}[f_w(x^k)-f_w(x^*)-\langle f_w'(x^*),x^k-x^*\rangle]\\ &+\frac{8C_\alpha^2L(1+\beta^{-1})}{W-B}\sum_{w\notin\mathcal{B}}E[f_{w,i_w^k}(\phi_{w,i_w^k}^k)-f_{w,i_w^k}(x^*)-\langle f_{w,i_w^k}'(x^*),\phi_{w,i_w^k}^k-x^*\rangle]\\ &+4C_\alpha^2\delta^2+\frac{2\epsilon^2}{(W-2B)^2}.\end{aligned}\tag{52}$$

For the first term at the right-hand side of (52), it holds

$$E\|f_{w,i_w^k}'(x^k)-f_{w,i_w^k}'(x^*)\|^2=\frac{1}{n}\sum_{j=1}^n\|f_{w,j}'(x^k)-f_{w,j}'(x^*)\|^2.\tag{53}$$

For the second term at the right-hand side of (52), it holds

$$\begin{aligned}&\frac{1}{W-B}\sum_{w\notin\mathcal{B}}[f_w(x^k)-f_w(x^*)-\langle f_w'(x^*),x^k-x^*\rangle]\\ =&f(x^k)-f(x^*)-\langle f'(x^*),x^k-x^*\rangle\\ =&f(x^k)-f(x^*).\end{aligned}\tag{54}$$

For the third term at the right-hand side of (52), it holds

$$\begin{aligned}&E[f_{w,i_w^k}(\phi_{w,i_w^k}^k)-f_{w,i_w^k}(x^*)-\langle f_{w,i_w^k}'(x^*),\phi_{w,i_w^k}^k-x^*\rangle]\\ =&\frac{1}{n}\sum_{j=1}^n f_{w,j}(\phi_{w,j}^k)-f_w(x^*)-\frac{1}{n}\sum_{j=1}^n\langle f_{w,j}'(x^*),\phi_{w,j}^k-x^*\rangle.\end{aligned}\tag{55}$$

Using these equalities to rewrite (52), we obtain (44). The non-negativeness of S_1^k and S_2^k is obvious. Observing that the left-hand side of (55) is non-negative due to the convexity of f_{w,i_w^k}, we know that S_3^k is non-negative and complete the proof. ■

APPENDIX E
PROOF OF THEOREM 1

Proof: Denote z_ϵ^* as an ϵ-approximate geometric median of $\{m_w^k\}$. We begin from manipulating $E\|x^{k+1}-x^*\|^2$ as

$$\begin{aligned}&E\|x^{k+1}-x^*\|^2\\ =&E\|x^k-\gamma z_\epsilon^*-x^*\|^2\\ =&\|x^k-x^*\|^2-2\gamma E\langle z_\epsilon^*,x^k-x^*\rangle+\gamma^2E\|z_\epsilon^*\|^2.\end{aligned}\tag{56}$$

The second term at the right-hand side of (56) can be bounded as

$$\begin{aligned}&E\langle z_\epsilon^*,x^k-x^*\rangle\\ =&\langle f'(x^k),x^k-x^*\rangle+E\langle z_\epsilon^*-f'(x^k),x^k-x^*\rangle.\\ \geq&\langle f'(x^k),x^k-x^*\rangle-\frac{\gamma}{2\eta}E\|z_\epsilon^*-f'(x^k)\|^2-\frac{\eta}{2\gamma}\|x^k-x^*\|^2,\end{aligned}\tag{57}$$

for any $\eta > 0$.

To bound the first term at the right-hand side of (57), we use the fact that f_{w,i_w^k} is μ-strongly convex and has L-Lipschitz continuous gradients. Define the function $g(x) = f_{w,i_w^k}(x) - \frac{\mu}{2}\|x\|^2$, which has $(L-\mu)$-Lipschitz continuous gradients. According to Theorem 2.1.5 of [37], for any $a, b \in \mathbb{R}^p$, we have

$$g(a) - g(b) - \langle g'(b), a - b\rangle \geq \frac{1}{2(L-\mu)}\|g'(b) - g'(a)\|^2.$$

Substituting the definition of $g(x)$, we know that

$$\begin{aligned}&\langle f'_{w,i_w^k}(a) - f'_{w,i_w^k}(b), a - b\rangle\\ \geq&\frac{\mu}{2}\|a-b\|^2 + \frac{1}{2L}\|f'_{w,i_w^k}(a) - f'_{w,i_w^k}(b)\|^2 + \frac{L-\mu}{L}[f_{w,i_w^k}(b) - f_{w,i_w^k}(a) - \langle f'_{w,i_w^k}(a), b - a\rangle].\end{aligned}$$

Letting $a = x^*$ and $b = x^k$, taking expectation over i_w^k for worker w, and averaging over all workers $w \notin \mathcal{B}$, we obtain

$$\begin{aligned}&\frac{1}{W-B}\sum_{w\notin\mathcal{B}} E\langle f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(x^*), x^k - x^*\rangle \\ \geq&\frac{\mu}{2}\|x^k - x^*\|^2 + \frac{1}{2L}\frac{1}{W-B}\sum_{w\notin\mathcal{B}} E\|f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(x^*)\|^2\\ &+ \frac{L-\mu}{L}\frac{1}{W-B}\sum_{w\notin\mathcal{B}} E[f_{w,i_w^k}(x^k) - f_{w,i_w^k}(x^*) - \langle f'_{w,i_w^k}(x^*), x^k - x^*\rangle].\end{aligned} \tag{58}$$

Notice that

$$\begin{aligned}&\frac{1}{W-B}\sum_{w\notin\mathcal{B}} E\langle f'_{w,i_w^k}(x^k) - f'_{w,i_w^k}(x^*), x^k - x^*\rangle \\ =&\langle f'(x^k) - f'(x^*), x^k - x^*\rangle = \langle f'(x^k), x^k - x^*\rangle,\\ &\frac{1}{W-B}\sum_{w\notin\mathcal{B}} E[f_{w,i_w^k}(x^k) - f_{w,i_w^k}(x^*) - \langle f'_{w,i_w^k}(x^*), x^k - x^*\rangle].\\ =&f(x^k) - f(x^*) - \langle f'(x^*), x^k - x^*\rangle = f(x^k) - f(x^*).\end{aligned} \tag{59}$$

Therefore, the first term at the right-hand side of (57) can be bounded as

$$\langle f'(x^k), x^k - x^*\rangle \geq \frac{\mu}{2}\|x^k - x^*\|^2 + \frac{1}{2L}S_1^k + \frac{L-\mu}{L}S_2^k, \tag{60}$$

where S_1^k and S_2^k are defined in Lemma 5. Substituting (57) and (60) into (56) yields

$$\begin{aligned}E\|x^{k+1} - x^*\|^2 \leq&(1 + \eta - \gamma\mu)\|x^k - x^*\|^2 - \frac{\gamma}{L}S_1^k - \frac{2\gamma(L-\mu)}{L}S_2^k\\ &+ \frac{\gamma^2}{\eta}E\|z_\epsilon^* - f'(x^k)\|^2 + \gamma^2 E\|z_\epsilon^*\|^2.\end{aligned} \tag{61}$$

Further substituting (22) in Lemma 2 and (44) in Lemma 5 yields

$$\begin{aligned}E\|x^{k+1} - x^*\|^2 \leq&(1 + \varepsilon - \gamma\mu)\|x^k - x^*\|^2 + (4\gamma^2 C_\alpha^2(1+\beta) - \frac{\gamma}{L})S_1^k\\ &- (2\gamma\frac{L-\mu}{L} + 8C_\alpha^2\mu\gamma^2\beta)S_2^k + 8C_\alpha^2\gamma^2 L(1+\beta^{-1})S_3^k + 4C_\alpha^2\gamma^2\delta^2 + \frac{\gamma^2}{\varepsilon}\xi + (\frac{\gamma^2}{\varepsilon} + \gamma^2)\frac{2\epsilon^2}{(W-2B)^2}.\end{aligned} \tag{62}$$

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

$$T^k := c\|x^k - x^*\|^2 + S_3^k, \tag{63}$$

where c is a positive constant. According to the definition in (45c), we know

$$S_3^k = \frac{1}{W-B}\sum_{w\notin\mathcal{B}}[\frac{1}{n}\sum_{j=1}^n f_{w,j}(\phi_{w,j}^k) - f_w(x^*) - \frac{1}{n}\sum_{j=1}^n\langle f'_{w,j}(x^*), \phi_{w,j}^k - x^*\rangle],$$

which is non-negative. Therefore, T^k is also non-negative.

Next, we compute the expectation of S_3^{k+1}. According to the updating rule of Byrd-SAGA, it holds

$$E[\frac{1}{n}\sum_{j=1}^{n} f_{w,j}(\phi_{w,j}^{k+1})] = \frac{1}{n} f_w(x^k) + (1-\frac{1}{n})\frac{1}{n}\sum_{j=1}^{n} f_{w,j}(\phi_{w,j}^k),$$

$$E[\frac{1}{n}\sum_{j=1}^{n}\langle f_w'(x^*), \phi_{w,j}^{k+1} - x^*\rangle] = \frac{1}{n}\langle f_w'(x^*), x^k - x^*\rangle + (1-\frac{1}{n})\frac{1}{n}\sum_{j=1}^{n}\langle f_{w,j}'(x^*), \phi_{w,j}^k - x^*\rangle.$$

Combining these two equalities yields

$$\begin{aligned} &E[\frac{1}{n}\sum_{j=1}^{n} f_{w,j}(\phi_{w,j}^{k+1}) - f_w(x^*) - \frac{1}{n}\sum_{j=1}^{n}\langle f_w'(x^*), \phi_{w,j}^{k+1} - x^*\rangle] \\ =&\frac{1}{n}\left[f_w(x^k) - f_w(x^*) - \langle f_w'(x^*), x^k - x^*\rangle\right] \\ &+ (1-\frac{1}{n})\left[\frac{1}{n}\sum_{j=1}^{n} f_{w,j}(\phi_{w,j}^k) - f_w(x^*) - \frac{1}{n}\sum_{j=1}^{n}\langle f_w'(x^*), \phi_{w,j}^k - x^*\rangle\right]. \end{aligned} \tag{64}$$

Averaging (64) over all workers $w \notin \mathcal{B}$ yields

$$ES_3^{k+1} = \frac{1}{n}S_2^k + (1-\frac{1}{n})S_3^k. \tag{65}$$

Here we use the definition of S_2^k in (45b) and the property $\frac{1}{W-B}\sum_{w\notin\mathcal{B}} f_w'(x^*) = f'(x^*) = 0$ such that

$$\frac{1}{W-B}\sum_{w\notin\mathcal{B}}\frac{1}{n}\left[f_w(x^k) - f_w(x^*) - \langle f_w'(x^*), x^k - x^*\rangle\right] = f(x^k) - f(x^*) = S_2^k.$$

Substituting (62) and (65) into (63), it follows that

$$\begin{aligned} &ET^{k+1} - (1+\eta-\gamma\mu)T^k \\ \le& c\gamma[4\gamma C_\alpha^2(1+\beta) - \frac{1}{L}]S_1^k + [\frac{1}{n} - 2c\gamma\frac{L-\mu}{L} - 8C_\alpha^2 c\mu\gamma^2\beta]S_2^k \\ &+ [\gamma\mu - \eta + 8C_\alpha^2 c\gamma^2 L(1+\beta^{-1}) - \frac{1}{n}]S_3^k + c\left(4C_\alpha^2\gamma^2\delta^2 + \frac{\gamma^2}{\eta}\xi + (\frac{\gamma^2}{\eta}+\gamma^2)\frac{2\epsilon^2}{(W-2B)^2}\right). \end{aligned} \tag{66}$$

Setting the constants as

$$\eta = \frac{\gamma\mu}{2}, \quad \beta = \frac{1-4C_\alpha^2\gamma L}{4C_\alpha^2\gamma L}, \quad c = \frac{1}{2\gamma(1-4C_\alpha^2\gamma\mu)n}$$

and constraining that

$$\begin{cases} 1 - 4C_\alpha^2\gamma L > \frac{1}{2}, \\ 1 - 4C_\alpha^2\gamma\mu > \frac{1}{2}, \\ \gamma < \frac{2}{n\mu + 32C_\alpha^2 L}, \end{cases} \quad \text{or equivalently} \quad \gamma < \min\{\frac{2}{n\mu + 32C_\alpha^2 L}, \frac{1}{8LC_\alpha^2}\}, \tag{67}$$

we know that the coefficients in front of S_1^k, S_2^k and S_3^k are all non-positive. Since S_1^k, S_2^k and S_3^k are all non-negative, dropping these terms yields

$$ET^{k+1} \le (1-\frac{\gamma\mu}{2})T^k + c\left(4C_\alpha^2\gamma^2\delta^2 + \frac{2\gamma}{\mu}\xi + (\frac{2\gamma}{\mu}+\gamma^2)\frac{2\epsilon^2}{(W-2B)^2}\right). \tag{68}$$

For the sake of simplicity, let

$$\tilde{\Delta}_2 = 4C_\alpha^2\gamma^2\delta^2 + \frac{2\gamma}{\mu}\xi + (\frac{2\gamma}{\mu}+\gamma^2)\frac{2\epsilon^2}{(W-2B)^2}. \tag{69}$$

Using telescopic cancellation on (68) from time 1 to time k yields

$$ET^k \le (1-\frac{\gamma\mu}{2})^k\left[T^0 - \frac{2c}{\gamma\mu}\tilde{\Delta}_2\right] + \frac{2c}{\gamma\mu}\tilde{\Delta}_2. \tag{70}$$

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

According the definition of the *Lyapunov function* in (63), it holds that

$$\begin{aligned} E\|x^k - x^*\|^2 &\leq \frac{1}{c} E T^k \leq (1 - \frac{\gamma\mu}{2})^k \left[\frac{T^0}{c} - \frac{2}{\gamma\mu}\tilde{\Delta}_2\right] + \frac{2}{\gamma\mu}\tilde{\Delta}_2 \\ &\leq (1 - \frac{\gamma\mu}{2})^k \Delta_1 + \Delta_2, \end{aligned} \tag{71}$$

where the last inequality comes from $\frac{1}{c} \leq 2\gamma n$ and the constants Δ_1 and Δ_2 are defined as

$$\Delta_1 := \|x^0 - x^*\|^2 + 2\gamma n \left[f(x^0) - f(x^*)\right] - \Delta_2, \tag{72}$$

$$\Delta_2 := \frac{2}{\gamma\mu}\tilde{\Delta}_2 = \left(\frac{8C_\alpha^2\delta^2}{\mu} + \frac{2\epsilon^2}{(W-2B)^2}\right)\gamma + \frac{4}{\mu^2}\left(\xi + \frac{2\epsilon^2}{(W-2B)^2}\right). \tag{73}$$

This completes the proof. ■