2022 IEEE International Conference on Communications Workshops (ICC Workshops) | 978-1-6654-2671-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCWORKSHOPS53468.2022.9814495

Best Effort Voting Power Control for Byzantine-resilient Federated Learning Over the Air

Xin Fan[1], Yue Wang[2], Yan Huo[1], and Zhi Tian[2]
[1]School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing, China
[2]Department of Electrical & Computer Engineering, George Mason University, Fairfax, VA, USA
E-mails: {yhuo,fanxin}@bjtu.edu.cn, {ywang56,ztian1}@gmu.edu

***Abstract*—Analog aggregation based federated learning over the air (FLOA) provides high communication efficiency and privacy provisioning in edge computing paradigm. When all edge devices (workers) simultaneously upload their local updates to the parameter server (PS) through the commonly shared time-frequency resources, the PS can only obtains the averaged update rather than the individual local ones. As a result, such a concurrent transmission and aggregation scheme reduces the latency and costs of communication but makes FLOA vulnerable to Byzantine attacks. For the design of Byzantine-resilient FLOA, this paper starts from analyzing the channel inversion (CI) power control mechanism that is widely used in existing FLOA literature. Our theoretical analysis indicates that although CI can achieve good learning performance in the non-attacking scenarios, it fails to work well with limited defensive capability to Byzantine attacks. Then, we propose a novel scheme called the best effort voting (BEV) power control policy, integrated with stochastic gradient descent (SGD). Our proposed BEV-SGD improves the robustness of FLOA to Byzantine attacks, by allowing all the workers to send their local updates at their maximum transmit power. Under the strongest-attacking circumstance, we derive the expected convergence rates of FLOA with CI and BEV, respectively. The comparison reveals that our BEV outperforms its counterpart with CI in terms of better convergence behavior, which is verified by experimental simulations.**

***Index Terms*—Federated learning, analog aggregation, Byzantine attack, best effort voting, channel-inversion, convergence.**

I. Introduction

Federated learning (FL) is a promising paradigm of distributed learning for low-latency and privacy-aware access to rich distributed data [1]–[5]. To achieve communication-efficient FL, sparsification [6], quantization [7] and infrequent uploading of local updates [8], [9] are proposed to reduce the amount of data needed to transfer in digital wireless communications. However, the communication overhead and latency are still proportional to the number of involved local workers in FL over digital communication channels. To handle this issue, FL over the air (FLOA) is proposed as a distributed learning solution [10]–[18], which exploits the over-the-air computation (AirComp) principle [19] for "one-shot" aggregation via local workers' concurrent update transmission using the same time-frequency resources. Based on the inherent waveform superposition property of wireless multiple access channels (MAC), AirComp allows to directly collect the gradient aggregation among local workers via concurrent transmission and computation [19], which exactly fits the need of FL for only an average of all distributed local gradients but not the individual values.

Benefitting from communication-efficient gradient aggregation, FLOA as a cross-disciplinary topic has attracted growing research interests, such as power control [11], [14], [20], devices scheduling [11], [13], [18], gradient compression [10], [12], [16], [17], and beamforming design [15]. For instance, a broadband analog aggregation scheme for broadband power control and device scheduling in FLOA is proposed in [13], where a set of tradeoffs between communication and learning are derived. In [11], convergence analysis quantifies the impact of AirComp on FL and then a joint optimization of communication and learning is proposed for the optimal power scaling and device scheduling. Considering energy-constrained local devices, an energy-aware device scheduling strategy is proposed in [18] to maximize the average number of workers scheduled for gradient update. For update compression, sparsification [17], quantization [16] and compressive-sensing based methods are proposed to further improve communication efficiency [10], [12]. In multiple antennas scenarios, a joint design of device scheduling and beamforming is presented in [15] to maximize the number of selected workers under the given mean square error (MSE) requirements.

Besides, FLOA not only improves communication efficiency, but also enhances the data privacy thanks to its unaccessibility of individual local gradients, which thus avoids the risk of potential model inversion attack, e.g., deep leakage from gradients [21]. While FLOA closes the doors to deep leakage from gradients, it leaves the windows open for adversaries to perform Byzantine attacks as well. Byzantine-robust aggregation has been well studied for vanilla FL [22], most of which uses a screening method, such as geometric median [23], coordinate-wise median [24], coordinate-wise trimmed mean [24], Krum/Multi-Krum [25], Bulyan [26] and so on [22]. The basic idea of these existing screening methods is to exclude outliers while aggregating the local gradients. All of them hinge on knowing the individual values of local gradients, which is however not accessible in FLOA due to the analog superposition of all local gradients over the air. Thus, the existing Byzantine-robust methods designed for vanilla FL fail to work for FLOA, which motivates our work.

To the best of our knowledge, this is the first paper to study the Byzantine attacks that occurs in the over-the-air transmissions for FL. We aim to deeply understand how

Byzantine attacks affect FLOA and then provide the corresponding defense strategy. Our main technical contributions are three-fold.

- Given the fact that most prior works on FLOA adopt channel inversion (CI) power control (or its variants) [10]–[13], [16]–[18], we theoretically prove that it can achieve performance approximating that of the ideal error-free case, which explains why it is widely used. However, its defensive capacity to Byzantine attacks is limited. Thus, we propose a new transmission policy, named the best effort voting (BEV) power control policy, where local workers transmit their local gradients with their maximum power.
- We theoretically prove that there exists a strongest attack for a Byzantine attacker to prevent FLOA from converging to the correct updating direction. As this is the strongest attack, which a Byzantine attacker may adopt to design its transmission policy.
- To demonstrate the effectiveness of our BEV compared with the existing most popular CI scheme under the strongest attacks, we provide the convergence analysis for both of them. We thus theoretically prove that BEV is better than CI in practice.

We also test the proposed method on the image classification problems using the MNIST dataset. The simulation results show that BEV is slightly weaker than CI when there are no Byzantine attacks, while BEV is much better than CI against Byzantine attacks. Thus, it is proved by theory and simulation that BEV is preferred over CI in practical applications where it is unpredictable whether there exists a Byzantine attack.

II. System Model

A. Federated Learning Model

We consider a distributed computation model with one parameter server (PS) and U local workers. Each local worker stores K data points, each of which is sampled independently from $\mathcal{D}$. That is, all workers have independent and identically distributed (i.i.d.) datasets [24]. Denote $(\mathbf{x}_{i,k}, \mathbf{y}_{i,k})$ as the k-th data of the i-th local worker. Let $f(\mathbf{w}; \mathbf{x}_{i,k}, \mathbf{y}_{i,k})$ denote a loss function of a parameter vector $\mathbf{w} = [w^1, \ldots, w^D] \in \mathcal{R}^D$ of dimension D associated with the data point $(\mathbf{x}_{i,k}, \mathbf{y}_{i,k})$. The corresponding population loss function is denoted as $F(\mathbf{w}) := \mathbb{E}_{\mathcal{D}}[f(\mathbf{w}; \mathbf{x}_{i,k}, \mathbf{y}_{i,k})]$. The PS and local workers collaboratively learn a model defined by the parameter $\mathbf{w}$ that minimizes the population loss:

$$\textbf{P1:} \quad \mathbf{w}^* = \arg\min_{\mathbf{w}} \quad F(\mathbf{w}). \tag{1}$$

The minimization of $F(\mathbf{w})$ is typically carried out through stochastic gradient descent (SGD) algorithm. The model parameter $\mathbf{w}_t$ at the t iteration is updated as

$$\text{(Model updating)} \quad \mathbf{w}_t = \mathbf{w}_{t-1} - \alpha \frac{\sum_{i=1}^{U} \mathbf{g}_{i,t}}{U}, \tag{2}$$

where α is the learning rate and $\mathbf{g}_{i,t} = \nabla f(\mathbf{w}_{t-1}; \mathbf{x}_{i,k}, \mathbf{y}_{i,k})$ is the local gradient computed at the i-th local worker using its randomly selected the k-th data sample.

We assume that N local workers are Byzantine attackers, and the remaining $M = U - N$ local workers are normal. To achieve (2), the PS needs to communicate with the local workers using some predefined protocol. The Byzantine attackers do not need to abide by this protocol and can send arbitrary messages to the PS. In particular, they may have complete knowledge of the learning system and algorithms, and can collude with each other.

B. Analog Aggregation Transmission Model

Assume symbol-level synchronization is achieved among the local workers through a synchronization channel [13]. To facilitate the power-control design, the transmitted symbols, denoted by $\tilde{\mathbf{g}}_{i,t} = [\tilde{g}_{i,t}^1, ..., \tilde{g}_{i,t}^d, ..., \tilde{g}_{i,t}^D]$, are standardized such that they have zero mean and unit variance, i.e., $\mathbb{E}[\tilde{g}_{i,t}^d (\tilde{g}_{i,t}^d)^H] = 1$. In this way, the power-control policy can be designed at the PS without knowledge of the specific transmitted symbols.

Since the statistics of the gradients may change over iterations, the standardization is needed for all communication rounds. Specifically, at the beginning of each communication round, each local worker estimates the mean and variance of the locally learnt gradient, denoted by $\bar{g}_{i,t} = \frac{1}{D}\sum_{d=1}^{D} g_{i,t}^d$ and $\epsilon_{i,t}^2 = \frac{1}{D}\sum_{d=1}^{D}(g_{i,t}^d - \bar{g}_{i,t})^2$, respectively. Then the locally estimated mean and variance are transmitted to the PS for global gradient statistics estimation by averaging.

Upon receiving $\bar{g}_{i,t}$ and $\epsilon_{i,t}^2$, the PS averages all the local estimates to get the global estimates of the mean and variance of the gradient as $\bar{g}_t = \frac{1}{U}\sum_{i=1}^{U} \bar{g}_{i,t}$ and $\epsilon_t^2 = \frac{1}{U}\sum_{i=1}^{U} \epsilon_{i,t}^2$. Then the estimated $\bar{g}_t$ and ϵ_t^2 are broadcast back to the local workers and used for the standardization.

After receiving the standardization factors $\bar{g}_t$ and ϵ_t^2, each local worker performs the transmit signal standardization as follows:

$$\tilde{\mathbf{g}}_{i,t} = \frac{\mathbf{g}_{i,t} - \bar{g}_t \mathbf{1}}{\epsilon_t}, \tag{3}$$

where $\mathbf{1}$ is an all-1 vector, the dimension of which is the same as $\mathbf{g}_{i,t}$ and all the entries of which are 1.

Considering only two symbols transmitted in each communication round, the individual locally estimated mean and variance are collected at the PS one by one. We assume such communications for standardization are noise-free without introducing errors. Note that the Byzantine attackers know the designed standardization method, and they would send the true mean and variance of their local gradients to avoid exposing themselves during the standardization stage. Otherwise, the attackers may be easily detected and then filtered out by the PS, as the normal workers and Byzantine workers have i.i.d. datasets.

After standardization, all local workers transmit their standardized local gradients $\tilde{\mathbf{g}}_{i,t}$ to the PS with the transmit power $p_{i,t}$ to be designed in the sequel. The transmission of each local worker is subject to the transmit power constraint:

$$\mathbb{E}[\|p_{i,t}\tilde{\mathbf{g}}_{i,t}\|^2] = \mathbb{E}[p_{i,t}^2 \sum_{d=1}^{D} (\tilde{g}_{i,t}^d)^2] = D p_{i,t}^2 \leq p_i^{\max}, \quad \forall i. \tag{4}$$

Thus the power constraint boils down to $p_{i,t}^2 \leq \frac{p_i^{\max}}{D}$.

On the other hand, the Byzantine attackers can report any values $\hat{\mathbf{g}}_{n,t}$ as their gradient updates to the PS so as to skew

FL. The transmit power $\hat{p}_{n,t}$ of the n-th Byzantine attackers satisfies

$$\mathbb{E}[\|\hat{p}_{n,t}\hat{\mathbf{g}}_{n,t}\|^2] \leq p_n^{\max}, \quad \forall n. \tag{5}$$

Considering block fading channels, where the wireless channels remain unchanged within each iteration in FL but may change independently from one iteration to another. We define the duration of one iteration as one time block, indexed by t. At the t-th iteration, the received signals at the PS is given by

$$\mathbf{y}_t = \sum_{m=1}^{M} p_{m,t}|h_{m,t}|\tilde{\mathbf{g}}_{m,t} + \sum_{n=1}^{N} \hat{p}_{n,t}|h_{n,t}|\hat{\mathbf{g}}_{n,t} + \mathbf{z}_t, \tag{6}$$

where $|h_{i,t}|$ is the channel gain from the i-th worker to the PS at the t-th iteration and $\mathbf{z}_t \sim \mathcal{CN}(0, z^2\mathbf{I})$ is additive white Gaussian noise (AWGN). The channels follow independent Rayleigh fading, i.e., $h_{i,t} \sim \mathcal{CN}(0, \sigma_i^2)$ and we assume that channels are perfectly known at local workers and the PS. With perfect channel state information (CSI), the channel phase offset is compensated at the local workers before they transmit their gradient updates.

After receiving the signals from the local workers, the PS performs de-standardization to get the estimated aggregated gradient by inverting the standardization as follows

$$\begin{aligned}
\tilde{\mathbf{g}}_t &= \epsilon_t \mathbf{y}_t + \left(\sum_{i=1}^{U} p_{i,t}|h_{i,t}|\right)\bar{g}_t\mathbf{1} \\
&= \sum_{m=1}^{M} p_{m,t}|h_{m,t}|\mathbf{g}_{m,t} + \epsilon_t \sum_{n=1}^{N} \hat{p}_{n,t}|h_{n,t}|\hat{\mathbf{g}}_{n,t} \\
&+ \left(\sum_{n=1}^{N} p_{n,t}|h_{n,t}|\right)\bar{g}_t\mathbf{1} + \epsilon_t\mathbf{z}_t.
\end{aligned} \tag{7}$$

By using the estimated aggregated gradient, the global model parameters are updated at the t-th iteration by

$$\text{(Updating with estimated gradients)} \quad \mathbf{w}_t = \mathbf{w}_{t-1} - \alpha\tilde{\mathbf{g}}_t. \tag{8}$$

Next, we provide two transmit power allocation schemes for normal local workers: the existing channel-inversion (CI) transmission [13], [16] and our proposed best effort voting (BEV) scheme.

1) Channel-inversion Transmission Scheme: given perfect CSI, in the channel-inversion scheme [13], [16], channels are inverted by power control so that gradient parameters transmitted by different local workers are received with identical amplitudes, achieving amplitude alignment at the PS. The transmit power of the i-th local worker is given by $p_{i,t}^2 = \frac{b_t^2}{|h_{i,t}|^2}, \forall i$, where $b_t^2 = \min\{\frac{P_i^{\max}}{D}|h_{i,t}|^2, i = 1, 2, ..., U\}$ is a scaling factor used to satisfy the power constraint in (4).

It is evident that

$$\mathbb{E}[b_t^2] \geq P_0^{\max}\mathbb{E}[\min\{|h_{i,t}|^2, i = 1, 2, ..., U\}], \tag{9}$$

where $P_0^{\max} = \min\{\frac{P_i^{\max}}{D}, i = 1, 2, ..., U\}$. Hence we can set $b_t^2 = P_0^{\max}\mathbb{E}[\min\{|h_{i,t}|^2, i = 1, 2, ..., U\}]$ for the power allocation. Since the channel coefficient is Rayleigh distributed $h_{i,t} \sim \mathcal{CN}(0, \sigma_i^2)$, $|h_{i,t}|^2$ follows the exponential distribution with mean $\frac{1}{\lambda_i} = 2\sigma_i^2$. Thus, we have $\mathbb{E}[\min\{|h_{i,t}|^2, i = 1, 2, ..., U\}] = \frac{1}{\sum_{i=1}^{U}\lambda_i} \doteq \lambda$. As a result, for fulfilling the channel-inversion scheme in practice, the transmit power of the i-th local worker is

$$p_{i,t} = \frac{b_0}{|h_{i,t}|}, \quad \forall i, \tag{10}$$

where we set $b_t^2 = P_0^{\max}\lambda \doteq b_0^2$.

2) The Proposed BEV scheme: To counter intelligent Byzantine attackers, our idea is to let normal local workers try their best to combat the impact of potential Byzantine attackers and to guide the FL to converge in the right direction, which is therefore named as the best effort voting (BEV) scheme. In the BEV scheme, normal local workers transmit their local gradients by using their maximum transmit power. The transmit power of the i-th local worker in BEV scheme is given by

$$p_{i,t} = \sqrt{\frac{p_i^{\max}}{D}}, \quad \forall i. \tag{11}$$

Different power allocation schemes have different resistance against Byzantine attackers, we will discuss in the next section.

III. The Convergence Analysis

In this section, we compare the convergence performance of two different power allocation schemes. We first prove that there is a strongest attack that a Byzantine attacker can achieve to prevent the convergence of FLOA. And then under such a circumstance, we provide the convergence rate of FLOA for the two transmission schemes, respectively.

A. Assumptions

To facilitate the convergence analysis, we make several standard assumptions on the loss function and computed gradient estimates. Note that our developed theory is applicable to the popular deep neural networks (DNNs), since we do not assume a convex setting on the loss function.

Assumption 1: We assume the Lipschitz continuity and smoothness of the loss function F, and thus we get [27]

$$F(\mathbf{w}_t) \leq F(\mathbf{w}_{t-1}) + \mathbf{g}_t^T(\mathbf{w}_t - \mathbf{w}_{t-1}) + \frac{L}{2}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2. \tag{12}$$

where L is a positive Lipschitz constant.

Assumption 2: It is assumed that the stochastic local gradient estimates are independent and unbiased estimates of the global gradient with the bounded variance [16], i.e.,

$$\mathbb{E}(\mathbf{g}_{i,t}) = \mathbf{g}_t, \quad \forall i, t, \tag{13}$$

$$\mathbb{E}(\|\mathbf{g}_{i,t} - \mathbf{g}_t\|^2) \leq \delta^2, \quad \forall i, t, \tag{14}$$

where we consider the standard SGD in this work. When the mini-batched SGD with a size K_b is applied, then the variance is bounded by $\frac{\delta^2}{K_b}$.

Assumption 3: The standardization factors $\bar{g}_t$ and ϵ_t^2 are unbiased estimates of the global gradient with the bounded variance as follows [13]

$$\mathbb{E}[\bar{g}_t] = \frac{\sum_{d=1}^{D} g_t^d}{D}, \quad \forall t, \tag{15}$$

$$\epsilon_t \leq \epsilon, \quad \forall t. \tag{16}$$

The above assumptions allow tractable convergence analysis as follows.

B. The Strongest Attack Case of Byzantine Attacks

While the Byzantine attackers may send arbitrary signals to destroy FL, there exists the strongest attack that a Byzantine attacker can achieve to prevent the convergence of FLOA. Intuitively, since the Byzantine attackers want the global gradient to be updated at the PS in the opposite direction of what normal local workers expect, the Byzantine attackers would like to transmit $\hat{\mathbf{g}}_{n,t} = -\mathbf{g}_{n,t}$ to the PS with its maximum transmit power $\hat{p}_{n,t}$. Given the global model parameter $\mathbf{w}_{t-1}$, the Byzantine attackers compute the true gradient $\mathbf{g}_{n,t}$ by using their own data. In addition, the transmit power $\hat{p}_{n,t}$ satisfies the maximum power constraint, i.e., $\mathbb{E}[\|\hat{p}_{n,t}\hat{\mathbf{g}}_{n,t}\|^2] = p_n^{\max}$. This is the worst case considered in this paper and we theoretically demonstrate in the following **Theorem 1** that it is the strongest attack that a Byzantine attacker can achieve.

Theorem 1. *Considering SGD for the FL system deploying analog aggregation transmission with Byzantine attackers, the strongest attacks can be performed as*

$$\hat{\mathbf{g}}_{n,t} = -\mathbf{g}_{n,t}, \tag{17}$$

$$\hat{p}_{n,t} = \sqrt{\frac{p_n^{\max}}{(\bar{g}_t^2 + \epsilon_t^2)D}}. \tag{18}$$

Proof. All the proofs, which are omitted in this paper due to the page limit, can be found in our journal version at [28]: https://arxiv.org/abs/2110.09660. □

We consider the above strongest attacks in the following convergence analysis so as to evaluate the defensive efficiency of different transmission schemes.

C. The Convergence of SGD with Channel-inversion Transmission

With perfect CSI at each local worker, the channel inversion power control can be accurately performed. The resultant convergence rate of the CI transmission scheme under the strongest attacks is derived as follows.

Theorem 2. *Considering SGD for the FL system deploying analog aggregation transmission with the CI power control and N Byzantine attackers taking the strongest attacks as in* (17)-(18)*, the convergence rate is given by*

$$\mathbb{E}[\sum_{t=1}^{T}\frac{1}{T}\|\mathbf{g}_t\|^2)] \leq \frac{1}{\sqrt{T}}\left(\frac{2L\Omega_{CI}}{\omega_{CI}^2\bar{\alpha}}(F(\mathbf{w}_0) - F(\mathbf{w}^*)) + \bar{\alpha}\left(\delta^2 + \frac{1}{\Omega_{CI}}\epsilon^2 z^2\right)\right), \tag{19}$$

where $\omega_{CI} = Mb_0 - \sum_{n=1}^{N}\sqrt{\frac{\pi\sigma_n^2 p_n^{\max}}{2D}}$, $\Omega_{CI} = (U + N)(Ub_0^2 + \sum_{n=1}^{N}\frac{2\sigma_n^2 p_n^{\max}}{D})$. We set the learning rate $\alpha = \frac{\omega_{CI}}{L\Omega_{CI}\sqrt{T}}\bar{\alpha}$, where $\bar{\alpha}$ is a positive constant satisfying $\bar{\alpha} < 2\sqrt{T}$, and T is the cumulative number of the communication rounds. The convergence is guaranteed if $\omega_{CI} > 0$.

Proof. Please refer to our journal version [28]. □

Remark 1. Considering a small learning rate, the asymptotical convergence rate is dominated by $O(\frac{\Omega_{CI}}{\omega_{CI}^2\sqrt{T}})$. In addition, the convergence condition is given by $\omega_{CI} > 0$, which implies that the FL converges as long as we set a small enough learning rate. From this convergence condition, we can see that even one Byzantine attacker (once it has a very large transmit power or its channel gain is very large) can destroy the FL.

Remark 2. For a special case where all the local workers have the same maximum power (i.e., $p_i^{\max} = p^{\max}$, $\forall i$) and the independent and identically distributed channels (i.e., $\sigma_i = \sigma$, $\forall i$), we have the convergence condition $\omega_{CI} = (\frac{M}{\sqrt{U}} - \sqrt{\frac{N^2\pi}{4}})\sqrt{\frac{2p^{\max}\sigma^2}{D}} > 0$. Therefore, we conclude that the number of attackers in this special case should be no more than $\frac{U}{1+\sqrt{\pi U}}$ to make the CI scheme defend against the Byzantine attack.

Remark 3. As we can see, in the case of CI power control without Byzantine attackers, we get the fastest asymptotical convergence rate as $O(\frac{1}{\sqrt{T}})$, which is the same as the error-free (EF) case where we do not consider the influence of wireless channels and noises.

D. The Convergence of SGD with BEV Transmission

For our BEV transmission scheme under the strongest attacks, the resultant convergence rate is derived as following **Theorem 3**.

Theorem 3. *Considering SGD for the FL system deploying analog aggregation transmission with BEV power control and N Byzantine attackers taking the strongest attacks as in* (17)-(18)*, the convergence rate is given by*

$$\mathbb{E}[\sum_{t=1}^{T}\frac{1}{T}\|\mathbf{g}_t\|^2)] \leq \frac{1}{\sqrt{T}}\left(\frac{2L\Omega_{BEV}}{\bar{\alpha}\omega_{BEV}^2}(F(\mathbf{w}_0) - F(\mathbf{w}^*)) + \bar{\alpha}\left(\delta^2 + \frac{1}{\Omega_{BEV}}\epsilon^2 z^2\right)\right), \tag{20}$$

where $\omega_{BEV} = \sum_{i=1}^{M}\sqrt{\frac{p_i^{\max}\pi}{2D}}\sigma_i - \sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}\pi}{2D}}\sigma_n$ and $\Omega_{BEV} = (U + N)\sum_{i=1}^{U}\frac{2\sigma_i^2 p_i^{\max}}{D}$. We set the learning rate $\alpha = \frac{\omega_{BEV}}{L\Omega_{BEV}\sqrt{T}}\bar{\alpha}$, where $\bar{\alpha}$ is a positive constant satisfying $\bar{\alpha} < 2\sqrt{T}$. The convergence is guaranteed if $\omega_{BEV} > 0$.

Proof. Please refer to our journal version [28]. □

Remark 4. If all the attackers and normal workers are isomorphic (the same case in *Remark 2*), our BEV can defend Byzantine attacks when $N \leq \frac{U}{2}$. Since $\frac{U}{2} \geq \frac{U}{1+\sqrt{\pi U}}$, our BEV scheme can defend against a larger number of Byzantine attackers than that of CI.

Remark 5. Considering a small learning rate, if both the CI scheme and our BEV scheme can converge, the asymptotical convergence rate is dominated by $O(\frac{\Omega}{\omega^2\sqrt{T}})$. The comparison between $O(\frac{\Omega_{CI}}{\omega_{CI}^2\sqrt{T}})$ and $O(\frac{\Omega_{BEV}}{\omega_{BEV}^2\sqrt{T}})$ depends on the specific parameters. Considering a large learning rate, if both CI and our BEV can converge, the asymptotical convergence rate is dominated by $O(\frac{1}{\Omega\sqrt{T}})$. Since $\Omega_{BEV} > \Omega_{CI}$, the convergence rate of BEV is faster than CI .

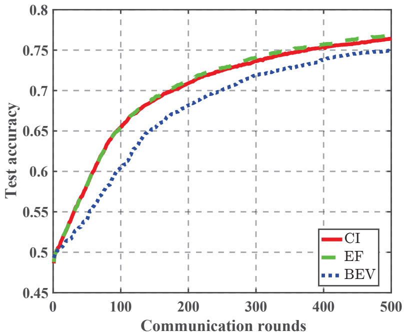

Fig. 1: The performance of BEV, CI and EF without Byzantine attacks.

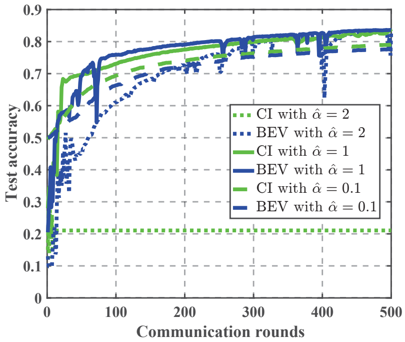

Fig. 2: The performance comparisons under an attacker with the lowest channel gain.

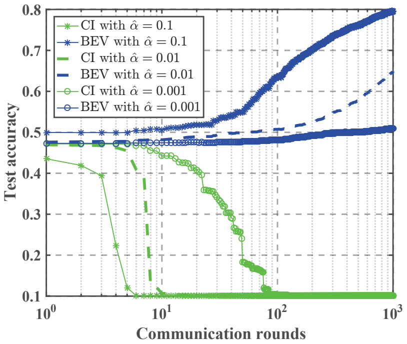

Fig. 3: The performance comparisons under an attacker with the highest channel gain.

Remark 6. When there are no Byzantine attackers, i.e., $N = 0$, we have $\omega_{BEV}^2 \leq \Omega_{BEV}$. For a small learning rate, the asymptotic convergence rate of BEV is dominated by $O(\frac{\Omega_{BEV}}{\omega_{BEV}^2 \sqrt{T}})$, which is slower than both the CI scheme and the EF case.

IV. Simulation Results

To evaluate the resilience of our proposed BEV scheme against Byzantine attacks, we provide the simulation results for an image classification task. Unless specified otherwise, the simulation settings are as follows. We consider that the FL system has $U = 10$ workers. The channels $h_{i,t}$'s are generated from $\mathcal{CN}(0, 1)$ for different i and t. We set the average receive SNR at local workers, defined as $\frac{P_i^{\max}}{Dz^2} = 10$ dB [16].

We consider the learning task of handwritten-digit identification using the well-known MNIST dataset[1] that consists of 10 classes ranging from digit "0" to "9". In our experiments, we train a multilayer perceptron (MLP) with a 784-neuron input layer, a 64-neuron hidden layer, and a 10-neuron softmax output layer. We adopt rectified linear unit (ReLU) as the activation function, and cross entropy as the loss function. The total number of parameters in the MLP is $D = 50890$. We randomly select 3000 distinct training samples and distribute them to all local workers as their different local datasets, i.e., $K_i = \bar{K} = 3000$, for any $i \in [1, U]$.

We evaluate our BEV scheme under different attacks, including 1) without any attacks, 2) only one attacker who is far to the PS, 3) only one attacker who is close to the PS, and 4) randomly selected several attackers. We compare two benchmarks with our BEV scheme, including 1) the CI scheme and 2) the FL under the ideal EF case where we do not consider the influence of wireless channels and noises.

A. Performance without Attacks

The EF case is set as the benchmark where the local gradients are perfectly aggregated at the PS, i.e., we set the channel $h_{i,t} = 1$ and the AWGN $\mathbf{z}_t = 0$. In Fig 1, we compare the performance of BEV with CI and EF without Byzantine attacks. We set adjusting factor of the learning rate, define $\hat{\alpha} = \frac{\bar{\alpha}}{L\sqrt{T}} = 0.1$. As we can see, the performance of CI is almost the same as EF. However, the performance of BEV is 2% loss compared to CI and EF. This results are in agreement with the theoretical analysis as indicated by Remark 6.

B. Performance under a Single Attacker with Weak Channel Gain

In Fig 2, we compare the performance of BEV with CI under Byzantine attacks. We consider the local worker whose channel gain is the lowest as the Byzantine attacker. The Byzantine attacker adopts the strongest attack to destroy FL. Under different adjusting factors of the learning rate $\hat{\alpha} = \frac{\bar{\alpha}}{L\sqrt{T}}$, we compare the performance of BEV with CI. Since the Byzantine attacker's channel gain is lowest, the overall impact of its attack to FLOA is relatively weak. In this case, both BEV and CI can converge, if a proper learning rate is selected. On the other hand, when the learning rate is not properly chosen, e.g., when $\hat{\alpha} = 2$ in Fig. 2, BEV can converge but CI fails. When $\hat{\alpha} = 1$, both BEV and CI can converge, but the convergence rate of BEV is faster than that of CI. This is because considering a large learning rate, the asymptotic convergence rate is dominated by $O(\frac{1}{\Omega\sqrt{T}})$ and $\Omega_{BEV} > \Omega_{CI}$. When $\hat{\alpha} = 0.1$, the performance of BEV is a little bit weaker in performance than CI. In practice, we prefer a large learning rate and hence BEV is superior than CI.

C. Performance under a Single Attacker with Large Channel Gain

In Fig 3, we compare the performance of BEV with CI under a Byzantine attacker whose channel gain is the highest. Due to the highest channel gain of the Byzantine attacker, we consider its attack as a strong attack. In this case, we compare the performance of BEV with CI under different $\hat{\alpha} = \frac{\bar{\alpha}}{L\sqrt{T}}$. Since the convergence condition $\omega_{CI} > 0$ is hard to guarantee, it can be seen from Fig 3 that CI cannot converge or coverage to a failure situation. As the decrease of $\hat{\alpha}$, it is useful for CI to converge to the right direction, but it still cannot defense the attack after a few iterations. On the other hand, BEV can still converge even in this strong attack case. Thus, if there is a strong attack, BEV is a better choice than CI. In addition, the convergence rate decreases as $\hat{\alpha}$ decreases. This implies that a larger learning rate is recommended under the condition of guaranteed convergence.

D. Performance with Multiple Randomly Selected Attackers

In Fig 4, we compare the performance of BEV with CI under the different number of Byzantine attackers. When the number of Byzantine attackers is less than 4, both BEV and

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

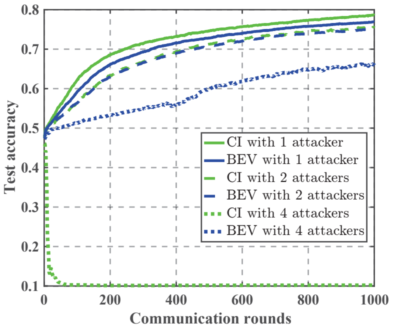

Fig. 4: The performance of BEV and CI with the different number of Byzantine attackers.

CI can converge, but the convergence rate decreases as the number of Byzantine attackers increases. When the number of Byzantine attackers is 4, i.e., $N > \frac{U}{1+\sqrt{\pi U}}$, CI can not converge to the correct direction, while BEV still converges in the correct direction but it converges at a slower rate. These results is consistent with the Remark 2 and Remark 4.

V. Conclusion

This paper studies the robustness of FLOA against Byzantine attacks. We provide analysis of convergence of different transmission schemes. Our analysis reveals the strongest attack that Byzantine attackers can achieve to prevent FLOA from converging in the correct direction. Through our convergence analysis, we find that, without Byzantine attackers, CI has the performance comparable to the ideal EF, while BEV has 2% performance loss. In the weakest Byzantine attack, considering a large learning rate, both CI and BEV can converge while BEV converges faster. If there is a strong Byzantine attack, the convergence of CI cannot guaranteed, but BEV can still converge. In practice, since it is impossible to determine the intensity of potential attacks, BEV is the better option because it performs well under various attack situations.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (Grants #61871023 and #61931001), Beijing Natural Science Foundation (Grant #4202054), the National Science Foundation of the US (Grants #1741338, #1939553, #2003211,#2128596 and #2136202), and the Virginia Research Investment Fund (Commonwealth Cyber Initiative Grant #223996).

References

[1] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, “A joint learning and communications framework for federated learning over wireless networks,” *IEEE Transactions on Wireless Communications*, 2020.

[2] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, “Energy efficient federated learning over wireless communication networks,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1935–1949, 2021.

[3] M. Chen, H. V. Poor, W. Saad, and S. Cui, “Convergence time optimization for federated learning over wireless networks,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2457–2471, 2020.

[4] ——, “Wireless communications for collaborative federated learning,” *IEEE Communications Magazine*, vol. 58, no. 12, pp. 48–54, 2020.

[5] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, “Distributed learning in wireless networks: Recent progress and future challenges,” *IEEE Journal on Selected Areas in Communications*, 2021.

[6] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, “Deep gradient compression: Reducing the communication bandwidth for distributed training,” *arXiv preprint arXiv:1712.01887*, 2017.

[7] Y. Liu, K. Yuan, G. Wu, Z. Tian, and Q. Ling, “Decentralized dynamic admm with quantized and censored communications,” in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 1496–1500.

[8] P. Xu, Z. Tian, Z. Zhang, and Y. Wang, “Coke: Communication-censored kernel learning via random features,” in *2019 IEEE Data Science Workshop (DSW)*. IEEE, 2019, pp. 32–36.

[9] P. Xu, Y. Wang, X. Chen, and Z. Tian, “Coke: Communication-censored decentralized kernel learning,” *Journal of Machine Learning Research*, vol. 22, no. 196, pp. 1–35, 2021.

[10] X. Fan, Y. Wang, Y. Huo, and Z. Tian, “Communication-efficient federated learning through 1-bit compressive sensing and analog aggregation,” in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2021, pp. 1–6.

[11] ——, “Joint optimization of communications and federated learning over the air,” *IEEE Transactions on Wireless Communications*, pp. 1–1, 2021.

[12] ——, “1-bit compressive sensing for efficient federated learning over the air,” *arXiv preprint arXiv:2103.16055*, 2021.

[13] G. Zhu, Y. Wang, and K. Huang, “Broadband analog aggregation for low-latency federated edge learning,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 491–506, 2019.

[14] X. Cao, G. Zhu, J. Xu, and K. Huang, “Optimal power control for over-the-air computation,” in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[15] K. Yang, T. Jiang, Y. Shi, and Z. Ding, “Federated learning via over-the-air computation,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2022–2035, 2020.

[16] G. Zhu, Y. Du, D. Gündüz, and K. Huang, “One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis,” *IEEE Transactions on Wireless Communications*, 2020.

[17] M. M. Amiri and D. Gündüz, “Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air,” *IEEE Transactions on Signal Processing*, vol. 68, pp. 2155–2169, 2020.

[18] Y. Sun, S. Zhou, and D. Gündüz, “Energy-aware analog aggregation for federated learning with redundant data,” *arXiv preprint arXiv:1911.00188*, 2019.

[19] B. Nazer and M. Gastpar, “Computation over multiple-access channels,” *IEEE Transactions on information theory*, vol. 53, no. 10, pp. 3498–3516, 2007.

[20] N. Zhang and M. Tao, “Gradient statistics aware power control for over-the-air federated learning,” *IEEE Transactions on Wireless Communications*, 2021.

[21] L. Zhu and S. Han, “Deep leakage from gradients,” in *Federated learning*. Springer, 2020, pp. 17–31.

[22] 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 Processing Magazine*, vol. 37, no. 3, pp. 146–159, 2020.

[23] S. Minsker, “Geometric median and robust estimation in banach spaces,” *Bernoulli*, vol. 21, no. 4, pp. 2308–2335, 2015.

[24] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, “Byzantine-robust distributed learning: Towards optimal statistical rates,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 5650–5659.

[25] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, “Machine learning with adversaries: Byzantine tolerant gradient descent,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 118–128.

[26] E.-M. El-Mhamdi and R. Guerraoui, “Fast and secure distributed learning in high dimension,” *arXiv e-prints*, pp. arXiv–1905, 2019.

[27] Y. Nesterov, “Introductory lectures on convex programming volume i: Basic course,” *Lecture notes*, vol. 3, no. 4, p. 5, 1998.

[28] X. Fan, Y. Wang, Y. Huo, and Z. Tian, “Bev-sgd: Best effort voting sgd for analog aggregation based federated learning against byzantine attackers,” *arXiv preprint arXiv:2110.09660*, 2021.