This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3164339, IEEE Internet of Things Journal

BEV-SGD: Best Effort Voting SGD against Byzantine Attacks for Analog Aggregation based Federated Learning Over the Air

Xin Fan[1], *Student Member, IEEE*, Yue Wang[2], *Senior Member, IEEE*, Yan Huo[1], *Senior Member, IEEE*, and Zhi Tian[2], *Fellow, IEEE*

***Abstract*—As a promising distributed learning technology, analog aggregation based federated learning over the air (FLOA) provides high communication efficiency and privacy provisioning under the edge computing paradigm. When all edge devices (workers) simultaneously upload their local updates to the parameter server (PS) through commonly shared time-frequency resources, the PS obtains the averaged update only rather than the individual local ones. While such a concurrent transmission and aggregation scheme reduces the latency and communication costs, it unfortunately renders FLOA vulnerable to Byzantine attacks. Aiming at Byzantine-resilient FLOA, this paper starts from analyzing the channel inversion (CI) mechanism that is widely used for power control in FLOA. Our theoretical analysis indicates that although CI can achieve good learning performance in the benign scenarios, it fails to work well with limited defensive capability against Byzantine attacks. Then, we propose a novel scheme called the best effort voting (BEV) power control policy that is integrated with stochastic gradient descent (SGD). Our BEV-SGD enhances the robustness of FLOA to Byzantine attacks, by allowing all the workers to send their local updates at their maximum transmit power. Under worst-case attacks, we derive the expected convergence rates of FLOA with CI and BEV power control policies, respectively. The rate comparison reveals that our BEV-SGD 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 analysis.**

I. Introduction

Edge intelligence has been recognized as a key enabler of various Internet-of-Things (IoT) services and applications in next-generation wireless systems [2], [3]. Federated learning (FL) provides a promising paradigm for edge intelligence, by taking advantages of parallel computing at edge devices and privacy-aware access to rich distributed data [4]–[7]. To achieve communication-efficient FL, sparsification [8], [9], quantization [10]–[12] and infrequent uploading of local updates [13]–[17] are developed to reduce the amount of data to be digitally transmitted over wireless systems. However, the communication overhead and latency are still proportional to the number of local workers participated in FL over digital communication channels. To handle this issue, FL over the air (FLOA) is recently proposed as a new framework for distributed learning [18]–[29], which exploits the over-the-air computation (AirComp) principle [30], [31] for "one-shot" aggregation via local workers' simultaneous update transmission over 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 [30]–[32], which exactly fits the need of FL for utilizing only an average of all distributed local gradients but not the individual values.

By virtue of its communication-efficient gradient aggregation, FLOA has attracted growing interest from multiple research communities to advance its development from the perspectives of communications, optimization and machine learning, such as power control [18], [19], [21], [33], devices scheduling [19], [20], [29], gradient compression [23]–[27], beamforming design [22], [28], [34] and learning rate optimization [35]. For instance, a broadband analog aggregation scheme for power control and device scheduling in FLOA is proposed in [20], where a set of tradeoffs between communications and learning are discussed. In [18], [19], convergence analysis is provided to quantify the impact of AirComp on FL and then joint optimization of communication and learning is proposed for optimal power scaling and device scheduling. Considering energy-constrained local devices, an energy-aware device scheduling strategy is proposed in [29] to maximize the average number of workers scheduled for gradient update. For update compression, sparsification [26], [27], quantization [23] and compressive-sensing based methods [24], [25] are proposed to further improve communication efficiency. In multiple antennas scenarios, a joint design of device scheduling and beamforming is presented in [22] to maximize the number of selected workers under a given mean square error (MSE) requirement. Since hyper-parameters can also affect learning performance, a learning rate optimization scheme is

Manuscript received October 21, 2021; revised January 5, 2022; accepted March 20, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61871023 and Grant 61931001; in part by the Beijing Natural Science Foundation under Grant 4202054; in part by the National Science Foundation of the U.S. under Grant 1741338, Grant 1939553, Grant 2003211, Grant 2128596 and Grant 2136202; and in part by the Virginia Research Investment Fund (Commonwealth Cyber Initiative Grant) under Grant 223996. This article was presented in part at IEEE ICC Workshops, May 2022 [1]. The associate editor coordinating the review of this article and approving it for publication was Jian Ren. (*Corresponding author: Yan Huo.*)

Xin Fan and Yan Huo are with the School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: fanxin@bjtu.edu.cn; yhuo@bjtu.edu.cn).

Yue Wang and Zhi Tian are with the Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA 22030, USA (e-mail: ywang56@gmu.edu; ztian1@gmu.edu).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3164339, IEEE Internet of Things Journal

proposed for multi-antenna systems to further improve the MSE performance and the testing accuracy [35].

Beside its superior communication efficiency over conventional FL, FLOA also enhances the data privacy thanks to its inherent unaccessibility to individual local gradients, which thus prevent potential model inversion attacks, e.g., deep leakage from gradients [36]. While FLOA closes the doors to deep leakage from gradients, it leaves the windows open for adversaries to perform Byzantine attacks as well. In fact, even a single Byzantine fault may destroy FL. Byzantine-robust aggregation has been well studied for vanilla FL [37]–[40], most of which uses a screening method, such as geometric median [41]–[44], coordinate-wise median [38], coordinate-wise trimmed mean [38], Krum/Multi-Krum [45], Bulyan [46], [47], Zeno/Zeno++ [48], [49] and so on [37]. The basic idea of these screening methods is to exclude outliers while aggregating the rest of local gradients. All of them hinge on the knowledge on 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, existing Byzantine-robust methods designed for vanilla FL cannot be applied to FLOA, which motivates us to design a new Byzantine-resilient approach customized for FLOA.

To the best of our knowledge, there is no literature so far on the study of Byzantine attacks to the over-the-air transmissions, nor is there any design of counter-attack measures for FLOA. In this work, we aim to deeply understand how Byzantine attacks affect FLOA and then provide the corresponding defense strategy. Our main contributions are three-fold.

- Given the fact that most prior works on FLOA adopt channel inversion (CI) power control (or its variants) [18]–[20], [23]–[26], [29], [35], [50], [51], we first theoretically prove that the CI methods under fading channels can achieve performance approximating that of the ideal error-free case, which explains why it is widely used to overcome the transmission errors in FL. Meanwhile, our analysis reveals that the defensive capacity of CI is very limited against Byzantine attacks. Thus, we propose a new robust transmission policy to counter Byzantine attacks, named the best effort voting (BEV) power control policy, where local workers transmit their local gradients with their maximum power.
- To study the impact of Byzantine attacks to FLOA, we derive the transmission policy of intelligent Byzantine attackers, including the falsified gradients and transmit power, that can maximally deter the convergence of FLOA. As this is the strongest attack, it is meaningful to assess its impact on FLOA under various transmission policies, which in turn serves to illuminate the respective robustness level of these policies.
- To demonstrate the effectiveness of our proposed BEV method compared with the popular CI scheme under the strongest attacks, we provide the convergence analysis for both our BEV and the existing CI. Our theoretical results prove that BEV outperforms CI in terms of better convergence behavior under the strongest Byzantine attacks.

We also test the proposed method on image classification problems using the MNIST dataset. Simulation results show that the learning performance of BEV is slightly worse than that of CI when there are no Byzantine attacks, while BEV significantly outperforms CI in terms of the robustness to against Byzantine attacks. Thus, our theoretical analysis and simulation results suggest that BEV is preferred over CI in practical applications that are subject to Byzantine attacks.

The rest of this paper is organized as follows. The system model for FLOA is presented in Section II, where we provide two power control policies i.e., CI and BEV. The closed-form expressions of their expected convergence rate are derived to compare the performance of different power control policies in Section III, where we also delineate the strongest attack case for a Byzantine attacker. Simulation results are provided in Section IV, followed by conclusions in Section V.

II. System Model

A. Federated Learning Model

Consider a distributed computation model with one parameter server (PS) and U local workers. Each local worker stores K data points, which are independent and identically distributed (i.i.d.) samples drawn from a large dataset $\mathcal{D}$ [38]–[40]. The Byzantine-resilient issue for the non-i.i.d. case is more involved, which is left for future work. 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 the loss function associated with each data point $(\mathbf{x}_{i,k}, \mathbf{y}_{i,k})$, where $\mathbf{w} = [w^1, \ldots, w^D]$ of size D consists of the model parameters. 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 the model parameter vector $\mathbf{w}$ by minimizing

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

The minimization of $F(\mathbf{w})$ is typically carried out through stochastic gradient descent (SGD) algorithm. At the PS, 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 data sample, say the k-th sample. Some communication and aggregation scheme needs to be in place in order for the PS to acquire the sum of local gradients in (2) from local workers.

Assume that N out of U local workers are Byzantine attackers, and the remaining $M = U - N$ local workers are normal. However, the Byzantine attackers do not need to follow this protocol and can send arbitrary messages to the PS. Even worse, these attackers may have complete knowledge of the learning system and algorithms, and can collude with each other. Further, the communications between the PS and local workers inevitably introduce channel noise, while Byzantine attackers could also exploit this opportunity to disrupt FLOA. Next, we will show that different predefined analog aggregation transmission protocols result in different performance of FLOA in the presence of Byzantine attacks.

B. Analog Aggregation Transmission Model

In FLOA, to exploit over-the-air computation for low-latency gradient aggregation, local gradients are amplitude-modulated for analog transmission and simultaneously transmitted from local workers to the PS through the same multi-access channel. Assume that symbol-level synchronization is achieved among the local workers through a synchronization channel [20]. 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., $\mathbb{E}[(\tilde{g}_{i,t}^d)^2] = 1, \forall i, t$. In this way, the power control policy can be designed at the PS without knowledge of the specific transmitted symbols. Note that the standardization factors are uniform for all local gradients and therefore can be inverted at the PS.

Since the statistics of the gradients may change over iterations, the standardization is executed in all communication rounds. Specifically, at the beginning of each communication round, each local worker estimates its 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. Given the received $\bar{g}_{i,t}$ 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-one vector with dimension equal to that of $\mathbf{g}_{i,t}$.

Considering only two symbols ($\bar{g}_t$ and ϵ_t^2) transmitted in each communication round, the individual locally estimated mean and variance are collected at the PS one by one. We assume that 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 certain transmit power $p_{i,t}$ (the design of power control on $p_{i,t}$ will be discussed later in this section). The transmission of each local worker is subject to the transmit power constraint:

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

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 of $\hat{\mathbf{g}}_{n,t}$ 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

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

Consider 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 signal 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 the first, second, and third terms correspond to normal workers, attackers and noise, respectively. In particular, $|h_{i,t}|$ is the channel gain from the i-th worker to the PS at the t-th iteration and $\mathbf{z}_t \sim \mathcal{N}(0, z^2\mathbf{I})$ is additive white Gaussian noise (AWGN) that is independent of the gradient updates. The channels follow independent Rayleigh fading, i.e., $h_{i,t} \sim \mathcal{CN}(0, \sigma_i^2)$. In this work, we assume that the 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 $\mathbf{y}_t$ in (6) from the local workers, the PS performs de-standardization to get the estimated aggregated gradient by inverting the standardization of (3) as follows:

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

where the first term corresponds to the aggregated gradients from normal local workers, the second plus the third terms denote the malignant contributions of Byzantine attackers to the gradient update, and the final term is from the noise.

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 discuss two transmit power allocation schemes for the design of $p_{i,t}$ that are adopted by normal local workers: the existing channel-inversion (CI) transmission [20], [23] and our proposed best effort voting (BEV) scheme.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3164339, IEEE Internet of Things Journal

1) Channel-Inversion Transmission Scheme: Given perfect known CSI, in the CI scheme [20], [23], channels are inverted by power control so that gradient parameters transmitted by different local workers are received with identical amplitudes, which leads to 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 set to

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

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

2) The Proposed Best Effort Voting Scheme: To counter intelligent Byzantine attackers, our idea is to let normal local workers try their best to combat the impact of potential Byzantine attacks so that FLOA converges to 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 which is independent to their CSI knowledge. 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 resilience against Byzantine attackers, which we will discuss next.

III. The Convergence Analysis

In this section, we compare the convergence performance of the aforementioned two power allocation schemes, CI and BEV. We first prove that there exists the strongest attack where a Byzantine attacker tries its best to prevent the convergence of FLOA. And then under such a circumstance, we derive the convergence rate of FLOA when applying the two transmission schemes, respectively.

A. Assumptions

To facilitate the convergence analysis, we make several standard assumptions on the loss function and the local gradient estimates. Note that our theoretical derivations do not assume convexity on the loss function. Therefore, our methodology is also applicable to the popular learning models of deep neural networks (DNNs).

Assumption 1: The loss function F is Lipschitz continuous and smooth, that is,

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

where L is a positive constant, referred to as the Lipschitz constant for the function $F(\cdot)$ [52].

Assumption 2: The stochastic local gradient estimates are independent and unbiased estimates of the global gradient with the variance [23], [53], 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. If 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 [20]

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

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

The above assumptions allow tractable convergence analysis.

B. The Strongest Byzantine Attacks

While the Byzantine attackers may send arbitrary signals, there exists the strongest attack that a Byzantine attacker can achieve to prevent the convergence of FLOA. Intuitively, the Byzantine attackers would like to influence the global gradients at the PS along the opposite direction of that of normal local workers. To this end, the Byzantine attackers will transmit $\hat{\mathbf{g}}_{n,t} = -\mathbf{g}_{n,t}$ to the PS with its maximum transmit power $\hat{p}_{n,t}$. In particular, given the global model parameter $\mathbf{w}_{t-1}$, the Byzantine attackers compute its own gradient $\mathbf{g}_{n,t}$ by using their own local 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 that FLOA experiences in this work and we theoretically demonstrate in the following **Theorem 1** that it is the strongest attack that a Byzantine attacker can impose to deter the convergence of FLOA.

Theorem 1. *Employing SGD for the FL system deploying analog aggregation transmission in the presence of 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. The proof of **Theorem 1** is provide in Appendix A. □

Since the aforementioned strongest attack has been proved as the worst case that FLOA can experience, next we will evaluate the defense efficiency of different transmission schemes via convergence analysis. We adopt the well known strategy of relating the norm of the gradient to the expected improvement to show the convergence for non-convex optimization [23], [53], [54], i.e,

$$\min_{0,1,\ldots,T} \mathbb{E}[\|\mathbf{g}_t\|^2] \leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{T}\|\mathbf{g}_t\|^2\right] \leq \mathcal{O}(\frac{1}{T^q}), \tag{19}$$

where $q > 0$ is the order of the total number of the iterations T. As we can see, if (19) holds, the norm of the gradient is expected to converge to 0 as T increases to infinity, which means that FL converges asymptotically. The convergence rate depends on the order value q, which is a key parameter to be assessed next.

C. The Convergence of SGD with CI Transmission

With CSI at each local worker, the CI power control can be performed as (10). The resultant convergence rate of the CI transmission scheme under the strongest attacks is derived as follows.

Theorem 2. *For a FLOA system with SGD-based model updating, CI-based power control for normal workers, and N Byzantine attackers taking the strongest attacks as in* (17)-(18)*, the convergence rate is given by*

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

where

$$\omega_{CI} = Mb_0 - \sum_{n=1}^{N}\sqrt{\frac{\pi\sigma_n^2 p_n^{\max}}{2D}}, \tag{21}$$

$$\Omega_{CI} = (U+N)\left(Ub_0^2 + \sum_{n=1}^{N}\frac{2\sigma_n^2 p_n^{\max}}{D}\right), \tag{22}$$

and $\bar{\alpha} = \frac{L\Omega_{CI}\sqrt{T}}{\omega_{CI}}\alpha$ is a positive constant satisfying $\bar{\alpha} < 2\sqrt{T}$, and b_0 is initialized as in (10)*. The convergence is guaranteed if $\frac{\alpha^2 L}{2}\Omega_{CI} - \alpha\omega_{CI} < 0$, which imposes constraints on α, L, b_0, σ_n, $p_n^{\max}$, M, N, D.*

Proof. The proof of **Theorem 2** is provide in Appendix B. □

Remark 1. For a small learning rate, the asymptotic convergence rate is dominated by $O(\frac{\Omega_{CI}}{\omega_{CI}^2\sqrt{T}})$. In addition, the convergence condition is given by $\frac{\alpha^2 L}{2}\Omega_{CI} - \alpha\omega_{CI} < 0$, the proof of which is also provided in Appendix B. This condition imposes an upper bound on the learning rate in the form $\alpha < \frac{2\omega_{CI}}{L\Omega_{CI}}$. Further, when the learning rate is set to be small enough, α^2 approaches 0, and the FL converges under a simplified condition of $\omega_{CI} > 0$. From this convergence condition, we can see that even one Byzantine attacker can destroy the FLOA, if this attacker has a very large transmit power or its channel gain is very large, e.g., if $p_n^{\max}$ or σ_n^2 for any n is very large, it is hard to ensure $\omega_{CI} > 0$.

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.

When there are no Byzantine attackers, i.e., $N = 0$, we have the following **Lemma 1**.

Lemma 1. *Employing SGD-based model updating for a FLOA system with the CI power control for normal local workers and no Byzantine attackers, the convergence rate is given by*

$$\mathbb{E}[\sum_{t=1}^{T} \frac{1}{T}\|\mathbf{g}_t\|^2)] \leq \frac{1}{\sqrt{T}}\left(\frac{2L}{\bar{\alpha}}(F(\mathbf{w}_0) - F(\mathbf{w}^*)) + \bar{\alpha}\left(\delta^2 + \frac{1}{U^2 b_0^2}\epsilon^2 z^2\right)\right), \tag{23}$$

where $\alpha = \frac{1}{LUb_0\sqrt{T}}\bar{\alpha}$.

Proof. When $N = 0$, we have $\omega_{CI}^2 = \Omega_{CI}$. Then setting $\alpha = \frac{\omega_{CI}}{L\Omega_{CI}\sqrt{T}}\bar{\alpha} = \frac{1}{LUb_0\sqrt{T}}\bar{\alpha}$, substituting α, ω_{CI} and Ω_{CI} into (20), we complete the proof. □

Remark 3. As we can see from (23), in the case of CI power control without Byzantine attackers, we get the fastest asymptotic 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. *Employing SGD-based model updating for a FLOA system with the BEV power control for normal workers and N Byzantine attackers taking the strongest attacks as in* (17)-(18)*, the convergence rate is given by*

$$\mathbb{E}[\sum_{t=1}^{T} \frac{1}{T}\|\mathbf{g}_t\|^2)] \leq \frac{1}{\sqrt{T}}\left(\frac{2L\Omega_{BEV}}{\bar{\alpha}\omega_{BEV}^2}(F(\mathbf{w}_0) - F(\mathbf{w}^*)) + \bar{\alpha}\left(\delta^2 + \frac{1}{\Omega_{BEV}}\epsilon^2 z^2\right)\right), \tag{24}$$

where

$$\omega_{BEV} = \sum_{i=1}^{M}\sqrt{\frac{p_i^{\max}\pi}{2D}}\sigma_i - \sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}\pi}{2D}}\sigma_n, \tag{25}$$

$$\Omega_{BEV} = (U+N)\sum_{i=1}^{U}\frac{2\sigma_i^2 p_i^{\max}}{D}, \tag{26}$$

and $\bar{\alpha} = \frac{L\Omega_{BEV}\sqrt{T}}{\omega_{BEV}}\alpha$ is a positive constant satisfying $\bar{\alpha} < 2\sqrt{T}$. The convergence is guaranteed if $\frac{\alpha^2 L}{2}\Omega_{BEV} - \alpha\omega_{BEV} < 0$, which imposes constraints on α, L, σ_i, $p_i^{\max}$, M, N, D.

Proof. The proof of **Theorem 3** is provide in Appendix C. □

Remark 4. The proof of the convergence condition $\frac{\alpha^2 L}{2}\Omega_{BEV} - \alpha\omega_{BEV} < 0$ is provided in Appendix C. This condition imposes an upper bound on the learning rate in the form $\alpha < \frac{2\omega_{BEV}}{L\Omega_{BEV}}$. Further, when the learning rate is set to be small enough, α^2 approaches 0, and the FL converges under a simplified condition of $\omega_{BEV} > 0$. 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. For a small learning rate, if both the CI scheme and our BEV scheme can converge, the asymptotic 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. For a large learning rate, if both the CI scheme and our BEV scheme 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 scheme is faster than that of the CI scheme.

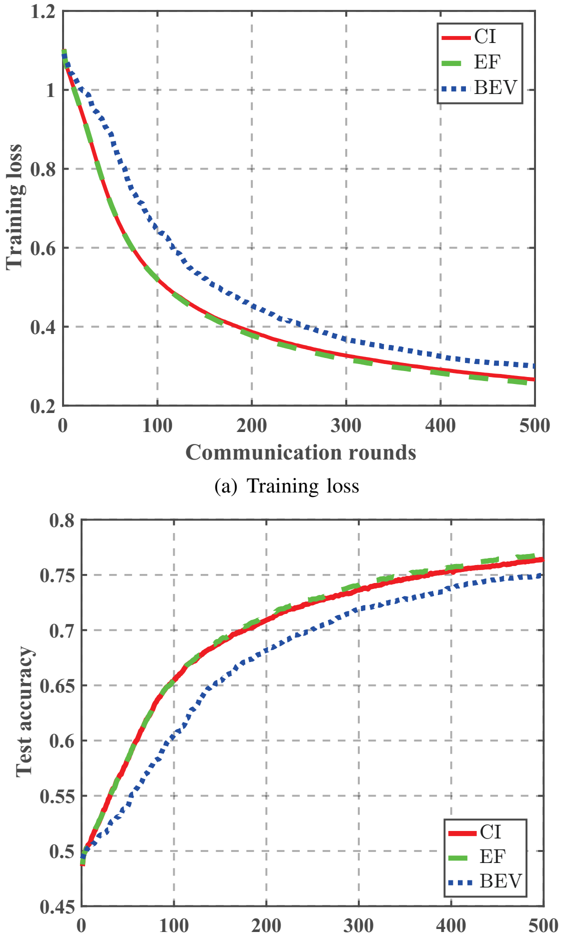
Remark 6. When there are no Byzantine attackers, i.e., $N = 0$, we have $\omega_{BEV}^2 \leq \Omega_{BEV}$. Considering a small learning rate, the asymptotic convergence rate of BEV is dominated by $O(\frac{\Omega_{BEV}}{\omega_{BEV}^2\sqrt{T}})$, which is slower than both the CI scheme and the EF case.

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 given as follows. The FLOA system has $U = 10$ workers. The wireless channels between the workers and the PS are modeled as i.i.d. Rayleigh fading, by generating $h_{i,t}$'s from the complex Gaussian distribution $\mathcal{CN}(0,1)$ for different i and t. The average receive SNR at local workers is set to be $\frac{P_i^{\max}}{Dz^2} = 10$ dB [23].

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 the MNIST dataset, a total of 60000 labeled training data samples and 10000 test samples. 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 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 from the PS, hence a weak attacker, 3) only one attacker who is close to the PS, hence a strong attacker, and 4) randomly selected several attackers. We compare with two benchmarks: 1) the CI scheme and 2) the FLOA under the ideal error-free case (EF) where we do not consider the influence of wireless channels and noise.



(a) Training loss

(b) Test accuracy

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

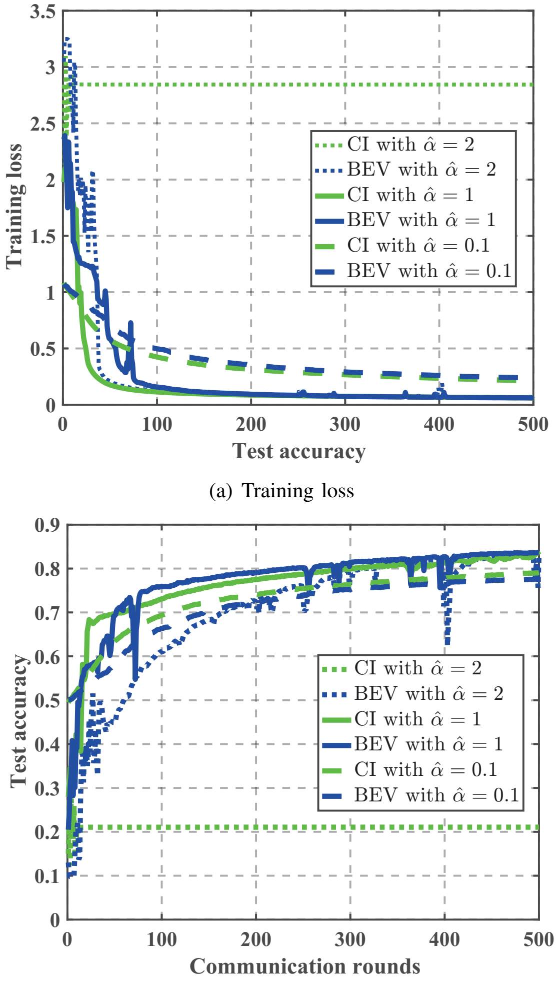
A. Performance without Attacks

The error-free 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. Considering $\alpha < \frac{\omega}{L\Omega}$ in *Remark 1* and *Remark 4*, we set the learning rate α such as its scaled version is $\hat{\alpha} = \frac{\bar{\alpha}}{L\sqrt{T}} = \frac{\Omega}{\omega}\alpha = 0.1$, where $\hat{\alpha}$ denotes the adjusting fact of α. As we can see from Fig. 1, the performance of CI is almost the same as EF. However, BEV experiences a 2% performance loss compared to CI and EF. This results are in agreement with our theoretical analysis in Theorem 3, which has been discussed in Remark 6. That is, CI converges a little faster than our BEV scheme, if and only if there exist no Byzantine attackers. However, practical learning applications of interest often operate in possible adversarial environments.

B. Performance under a Single Attacker with Weak Channel Gain

In Fig 2, we compare the performance of BEV with CI under a single Byzantine attack. Suppose that the attacker

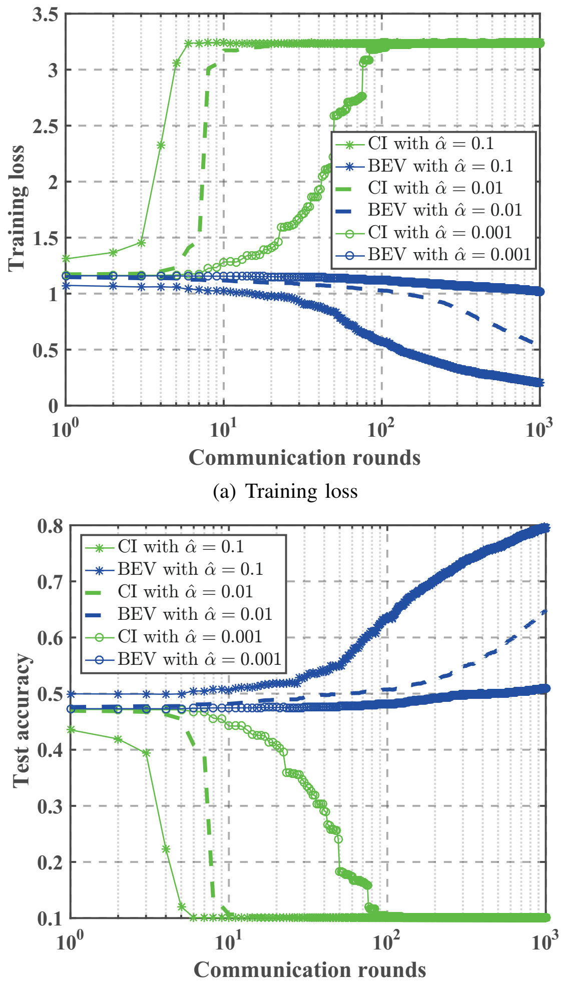
[1] http://yann.lecun.com/exdb/mnist/



(a) Training loss

(b) Test accuracy

Fig. 2: The performance of BEV and CI with a Byzantine attacker whose channel gain is the lowest.



(a) Training loss

(b) Test accuracy

Fig. 3: The performance of BEV and CI with a Byzantine attacker whose channel gain is the highest.

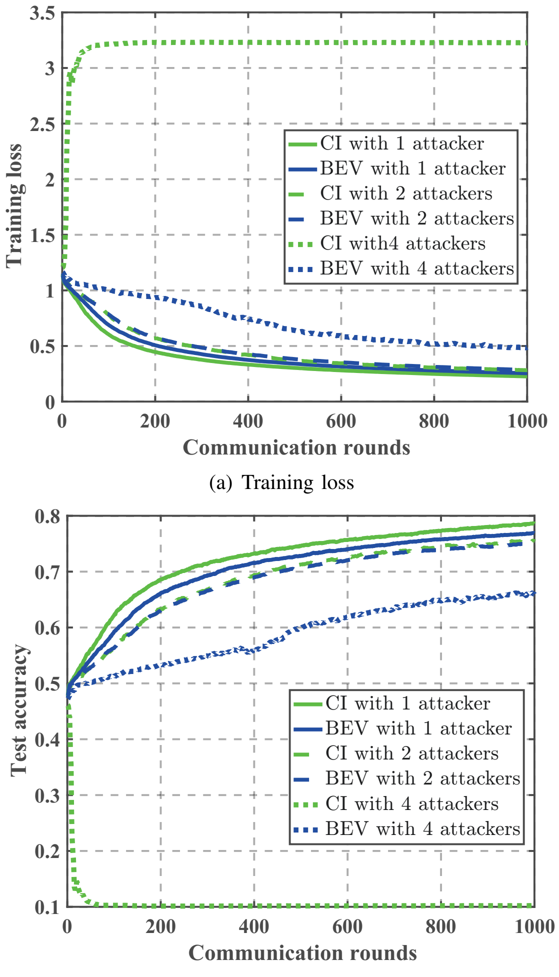
has the lowest channel gain among all local workers. It still adopts the strongest attack strategy to destroy FLOA. Since the Byzantine attack to FLOA is relatively weak, both BEV and CI can converge, if a proper learning rate $\hat{\alpha} = \frac{\bar{\alpha}}{L\sqrt{T}} = \frac{\Omega}{\omega}\alpha$ 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 for 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, when the convergence can be guaranteed, we prefer a large learning rate to achieve a fast convergence rate. Under a large learning rate, e.g., $\hat{\alpha} = 1$, our BEV works better 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 among all local workers. Thus, this is a strong attack. In this case of strong attacks, we compare the performance of BEV with CI under $\hat{\alpha} = \frac{\bar{\alpha}}{L\sqrt{T}} = \frac{\Omega}{\omega}\alpha$. 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 $\hat{\alpha}$ decreases, it is useful for CI to converge to the right direction, but it still cannot defend the attack after a few iterations. On the other hand, BEV can still converge, and hence is a better choice than CI in the presence of a strong attack. In addition, the convergence rate decreases as $\hat{\alpha}$ decreases. This implies that a larger learning rate is preferred 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. As we can see, when the number of Byzantine attackers is less than 4, both BEV and 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



(a) Training loss

(b) Test accuracy

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

slower rate. These results are consistent with our discussions in Remark 2 and Remark 4.

V. Conclusion

This paper studies the robustness of FL over the air (FLOA) against Byzantine attacks. We provide theoretical analysis on convergence performance of different transmission schemes. Our analytical results reveal the strongest attack that Byzantine attackers can impose to deter FLOA from converging to the correct direction. Our convergence analyses, corroborated by simulation results, delineate the convergence behavior of the CI and BEV schemes under various adversarial environments. Specifically, in the absence of any Byzantine attacker, CI has the performance comparable to the ideal error-free case, while BEV has 2% performance loss. In the weakest Byzantine attack, for a large learning rate, both CI and BEV can converge while BEV converges faster than CI. If there exists a strong Byzantine attacker, the convergence of CI cannot be guaranteed, but BEV can still converge. In practice, since it is impossible to determine the intensity of potential attacks, BEV is a better option to counter Byzantine attacks, 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).

Appendix A
Proof of **Theorem 1**

Given the estimates of the global gradient in (7), we have the update rule for model parameters as follows

$$\begin{aligned}\mathbf{w}_t =&\mathbf{w}_{t-1} - \alpha\tilde{\mathbf{g}}_t\\ =&\mathbf{w}_{t-1} - \alpha\left(\sum_{m=1}^{M} p_{m,t}|h_{m,t}|\mathbf{g}_{m,t} + \epsilon_t\sum_{n=1}^{N}\hat{p}_{n,t}|h_{n,t}|\hat{\mathbf{g}}_{n,t}\right.\\ &\left.+\sum_{n=1}^{N} p_{n,t}|h_{n,t}|\bar{g}_t\mathbf{1} + \epsilon_t\mathbf{z}_t\right). \qquad (27)\end{aligned}$$

Substituting (27) to (12), we have

$$\begin{aligned}F(\mathbf{w}_t) \le& F(\mathbf{w}_{t-1}) + \mathbf{g}_t^T(\mathbf{w}_t - \mathbf{w}_{t-1}) + \frac{L}{2}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2\\ =&F(\mathbf{w}_{t-1}) - \alpha\mathbf{g}_t^T\left(\sum_{m=1}^{M} p_{m,t}|h_{m,t}|\mathbf{g}_{m,t}\right.\\ &\left.+\epsilon_t\sum_{n=1}^{N}\hat{p}_{n,t}|h_{n,t}|\hat{\mathbf{g}}_{n,t} + \sum_{n=1}^{N} p_{n,t}|h_{n,t}|\bar{g}_t\mathbf{1} + \epsilon_t\mathbf{z}_t\right)\\ &+\frac{\alpha^2 L}{2}\left\|\sum_{m=1}^{M} p_{m,t}|h_{m,t}|\mathbf{g}_{m,t} + \epsilon_t\sum_{n=1}^{N}\hat{p}_{n,t}|h_{n,t}|\hat{\mathbf{g}}_{n,t}\right.\\ &\left.+\sum_{n=1}^{N} p_{n,t}|h_{n,t}|\bar{g}_t\mathbf{1} + \epsilon_t\mathbf{z}_t\right\|^2. \qquad (28)\end{aligned}$$

Rewriting this inequality and taking the expectation, we have

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_{t-1})] \le& -\alpha\mathbf{g}_t^T\left(\sum_{m=1}^{M} p_{m,t}|h_{m,t}|\mathbf{g}_t\right.\\ &\left.+\sum_{n=1}^{N}\hat{p}_{n,t}|h_{n,t}|\hat{\mathbf{g}}_{n,t} + \sum_{n=1}^{N} p_{n,t}|h_{n,t}|\mathbb{E}[\bar{g}_t]\mathbf{1}\right)\\ &+\frac{\alpha^2 L}{2}\mathbb{E}\left[\left\|\sum_{m=1}^{M} p_{m,t}|h_{m,t}|\mathbf{g}_{m,t} + \epsilon_t\sum_{n=1}^{N}\hat{p}_{n,t}|h_{n,t}|\hat{\mathbf{g}}_{n,t}\right.\right.\\ &\left.\left.+\sum_{n=1}^{N} p_{n,t}|h_{n,t}|\bar{g}_t\mathbf{1} + \epsilon_t\mathbf{z}_t\right\|^2\right]. \qquad (29)\end{aligned}$$

Since $\mathbf{g}_t^T\mathbb{E}[\bar{g}_t]\mathbf{1} = \mathbf{g}_t^T\frac{\sum_{d=1}^D g_t^d}{D}\mathbf{1} = \frac{(\sum_{d=1}^D g_t^d)^2}{D} \geq 0$, we have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_{t-1})] \leq \\
&\quad -\alpha\left(\sum_{i=1}^M p_{i,t}|h_{i,t}|\|\mathbf{g}_t\|^2 + \sum_{n=1}^N \hat{p}_{n,t}|h_{n,t}|\mathbf{g}_t^T\hat{\mathbf{g}}_{n,t}\right) \\
&\quad + \frac{\alpha^2 L}{2}\mathbb{E}\left[\left\|\sum_{m=1}^M p_{m,t}|h_{m,t}|\mathbf{g}_{m,t} + \epsilon_t\sum_{n=1}^N \hat{p}_{n,t}|h_{n,t}|\hat{\mathbf{g}}_{n,t}\right.\right. \\
&\quad \left.\left. + \sum_{n=1}^N p_{n,t}|h_{n,t}|\bar{g}_t\mathbf{1} + \epsilon_t\mathbf{z}_t\right\|^2\right]. \qquad (30)
\end{aligned}$$

If $\mathbb{E}(F(\mathbf{w}_t) - F(\mathbf{w}_{t-1})) \leq 0$, the objective decreases monotonically, then FL converges in mean. As we can see from (30), if we set the learning rate to be small enough, then the second term on the right hand side of (30) diminishes, and convergence is ensured as long as

$$\sum_{i=1}^M p_{i,t}|h_{i,t}|\|\mathbf{g}_t\|^2 + \sum_{n=1}^N \hat{p}_{n,t}|h_{n,t}|\mathbf{g}_t^T\hat{\mathbf{g}}_{n,t} > 0. \qquad (31)$$

In order to break the convergence condition in (31), the N Byzantine attackers would seek to make $\mathbf{g}_t^T\hat{\mathbf{g}}_{n,t} < 0$ for any n. In fact, the best way for them is to send $\hat{\mathbf{g}}_{n,t} = -\mathbf{g}_{n,t}$ with their maximum power so as to make $\mathbb{E}[\mathbf{g}_t^T\hat{\mathbf{g}}_{n,t}] = -\|\mathbf{g}_t\|^2 < 0$.

Given the power constraint in (5), we have

$$\begin{aligned}
\mathbb{E}[\|\hat{p}_{n,t}\hat{\mathbf{g}}_{n,t}\|^2] &= \hat{p}_{n,t}^2\sum_{d=1}^D \mathbb{E}[(g_{n,t}^d)^2] \\
&= \hat{p}_{n,t}^2 D(\epsilon_t^2 + \bar{g}_t^2) \leq p_n^{\max}. \qquad (32)
\end{aligned}$$

As a result, the Byzantine attackers are supposed to send $\hat{\mathbf{g}}_{n,t} = -\mathbf{g}_{n,t}$ with their maximum power $\hat{p}_{n,t} = \sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}$.

Appendix B
Proof of **Theorem 2**

Given the estimates of the global gradient in (7), the power allocation policy in (10), and the strongest attacks in **Theorem 1**, we have the update rule for model parameters as follows

$$\begin{aligned}
\mathbf{w}_t =& \mathbf{w}_{t-1} - \alpha\tilde{\mathbf{g}}_t \\
=& \mathbf{w}_{t-1} - \alpha\left(\sum_{m=1}^M p_{m,t}|h_{m,t}|\mathbf{g}_{m,t} + \epsilon_t\sum_{n=1}^N \hat{p}_{n,t}|h_{n,t}|\hat{\mathbf{g}}_{n,t}\right. \\
&\left. + \sum_{n=1}^N p_{n,t}|h_{n,t}|\bar{g}_t\mathbf{1} + \epsilon_t\mathbf{z}_t\right) \\
=& \mathbf{w}_{t-1} - \alpha\left(\sum_{m=1}^M b_0\mathbf{g}_{m,t}\right. \\
&\left. -\epsilon_t\sum_{n=1}^N\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t} + \sum_{n=1}^N b_0\bar{g}_t\mathbf{1} + \epsilon_t\mathbf{z}_t\right). \\
& \qquad (33)
\end{aligned}$$

Substituting (33) to (12), we get

$$\begin{aligned}
F(\mathbf{w}_t) &\leq F(\mathbf{w}_{t-1}) + \mathbf{g}_t^T(\mathbf{w}_t - \mathbf{w}_{t-1}) + \frac{L}{2}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2 \\
&= F(\mathbf{w}_{t-1}) - \alpha\mathbf{g}_t^T\left(\sum_{m=1}^M b_0\mathbf{g}_{m,t}\right. \\
&\quad \left.-\epsilon_t\sum_{n=1}^N\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t} + \sum_{n=1}^N b_0\bar{g}_t\mathbf{1} + \epsilon_t\mathbf{z}_t\right) \\
&\quad + \frac{\alpha^2 L}{2}\left\|\sum_{m=1}^M b_0\mathbf{g}_{m,t} - \epsilon_t\sum_{n=1}^N\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t}\right. \\
&\quad \left. + \sum_{n=1}^N b_0\bar{g}_t\mathbf{1} + \epsilon_t\mathbf{z}_t\right\|^2. \qquad (34)
\end{aligned}$$

Rewriting this inequality and taking the expectation, we get

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_{t-1})] \leq -\alpha\mathbf{g}_t^T\left(\sum_{m=1}^M b_0\mathbf{g}_t\right. \\
&\qquad \left.-\epsilon_t\sum_{n=1}^N\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}\mathbb{E}[|h_{n,t}|]\mathbf{g}_t + \sum_{n=1}^N b_0\mathbb{E}[\bar{g}_t]\mathbf{1}\right) \\
&\qquad + \frac{\alpha^2 L}{2}\mathbb{E}\left[\left\|\sum_{m=1}^M b_0\mathbf{g}_{m,t} + \sum_{n=1}^N b_0\bar{g}_t\mathbf{1}\right.\right. \\
&\qquad \left.\left. -\epsilon_t\sum_{n=1}^N\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t} + \epsilon_t\mathbf{z}_t\right\|^2\right], \qquad (35)
\end{aligned}$$

where $\mathbb{E}[|h_{i,t}|] = \sigma_i\sqrt{\frac{\pi}{2}}$, because of the Rayleigh distributed $|h_{i,t}|$.

Since $\mathbf{g}_t^T\mathbb{E}[\bar{g}_t]\mathbf{1} = \mathbf{g}_t^T\frac{\sum_{d=1}^D g_t^d}{D}\mathbf{1} = \frac{(\sum_{d=1}^D g_t^d)^2}{D} \geq 0$, we have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_{t-1})] \leq -\alpha\left(Mb_0\right. \\
&\quad \left.-\epsilon_t\sum_{n=1}^N\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}\sigma_n\sqrt{\frac{\pi}{2}}\right)\|\mathbf{g}_t\|^2 \\
&\quad + \frac{\alpha^2 L}{2}\mathbb{E}\left[\left\|\sum_{m=1}^M b_0\mathbf{g}_{m,t} - \epsilon_t\sum_{n=1}^N\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t}\right.\right. \\
&\quad \left.\left. + \sum_{n=1}^N b_0\bar{g}_t\mathbf{1} + \epsilon_t\mathbf{z}_t\right\|^2\right] \\
&\quad \leq -\alpha\left(Mb_0 - \sum_{n=1}^N\sqrt{\frac{p_n^{\max}}{D}}\sigma_n\sqrt{\frac{\pi}{2}}\right)\|\mathbf{g}_t\|^2 \\
&\quad + \frac{\alpha^2 L}{2}\mathbb{E}\left[\left\|\sum_{m=1}^M b_0\mathbf{g}_{m,t} - \epsilon_t\sum_{n=1}^N\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t}\right.\right. \\
&\quad \left.\left. + \sum_{n=1}^N b_0\bar{g}_t\mathbf{1} + \epsilon_t\mathbf{z}_t\right\|^2\right]. \qquad (36)
\end{aligned}$$

Using the triangle inequality of norms and Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{m=1}^{M} b_0\mathbf{g}_{m,t} - \epsilon_t\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t}\right.\right.\\
&\qquad\left.\left.+\sum_{n=1}^{N} b_0\bar{g}_t\mathbf{1}+\epsilon_t\mathbf{z}_t\right\|^2\right]\\
&=\mathbb{E}\left[\left\|\sum_{m=1}^{M} b_0\mathbf{g}_{m,t} - \epsilon_t\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t}\right.\right.\\
&\qquad\left.\left.+\sum_{n=1}^{N} b_0\bar{g}_t\mathbf{1}\right\|^2\right]+\mathbb{E}[\|\epsilon_t\mathbf{z}_t\|^2]\\
&\le\mathbb{E}\left[\left(\sum_{m=1}^{M}\|b_0\mathbf{g}_{m,t}\|+\sum_{n=1}^{N}\left\|\epsilon_t\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t}\right\|\right.\right.\\
&\qquad\left.\left.+\sum_{n=1}^{N}\|b_0\bar{g}_t\mathbf{1}\|\right)^2\right]+\epsilon^2z^2\\
&\le\mathbb{E}\left[(U+N)\left(\sum_{m=1}^{M}b_0^2\|\mathbf{g}_{m,t}\|^2+\sum_{n=1}^{N}b_0^2\|\bar{g}_t\mathbf{1}\|^2\right.\right.\\
&\qquad\left.\left.+\sum_{n=1}^{N}\frac{\epsilon_t^2p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}|h_{n,t}|^2\|\mathbf{g}_{n,t}\|^2\right)\right]+\epsilon^2z^2\\
&=(U+N)\left(\sum_{m=1}^{M}b_0^2\mathbb{E}[\|\mathbf{g}_{m,t}\|^2]+\sum_{n=1}^{N}b_0^2D\left(\frac{\sum_{d=1}^{D}g_t^d}{D}\right)^2\right.\\
&\qquad\left.+\sum_{n=1}^{N}\frac{\epsilon_t^2p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}\mathbb{E}[|h_{n,t}|^2]\mathbb{E}[\|\mathbf{g}_{n,t}\|^2]\right)+\epsilon^2z^2\\
&\le(U+N)\left(\sum_{m=1}^{M}b_0^2(\|\mathbf{g}_t\|^2+\delta^2)+\sum_{n=1}^{N}b_0^2\|\mathbf{g}_t\|^2\right.\\
&\qquad\left.+\sum_{n=1}^{N}\frac{\epsilon_t^2p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}2\sigma_n^2(\|\mathbf{g}_t\|^2+\delta^2)\right)+\epsilon^2z^2\\
&\le(U+N)\left(\left(Ub_0^2+\sum_{n=1}^{N}\frac{2\sigma_n^2p_n^{\max}}{D}\right)\|\mathbf{g}_t\|^2\right.\\
&\qquad\left.+\left(Mb_0^2+\sum_{n=1}^{N}\frac{2\sigma_n^2p_n^{\max}}{D}\right)\delta^2\right)+\epsilon^2z^2. \qquad (37)
\end{aligned}$$

Substituting (37) to (36), we get

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t)-F(\mathbf{w}_{t-1})]\le-\alpha\left(Mb_0-\sum_{n=1}^{N}\sqrt{\frac{\pi\sigma_n^2p_n^{\max}}{2D}}\right)\|\mathbf{g}_t\|^2\\
&+\frac{\alpha^2L}{2}\left((U+N)\left(\left(Ub_0^2+\sum_{n=1}^{N}\frac{2\sigma_n^2p_n^{\max}}{D}\right)\|\mathbf{g}_t\|^2\right.\right.\\
&+\left.\left.\left(Mb_0^2+\sum_{n=1}^{N}\frac{2\sigma_n^2p_n^{\max}}{D}\right)\delta^2\right)+\epsilon^2z^2\right)\\
&\le\left(\frac{\alpha^2L}{2}\Omega_{CI}-\alpha\omega_{CI}\right)\|\mathbf{g}_t\|^2+\frac{\alpha^2L}{2}(\Omega_{CI}\delta^2+\epsilon^2z^2), \qquad (38)
\end{aligned}$$

where

$$\omega_{CI}=Mb_0-\sum_{n=1}^{N}\sqrt{\frac{\pi\sigma_n^2p_n^{\max}}{2D}}, \qquad (39)$$

$$\Omega_{CI}=(U+N)\left(Ub_0^2+\sum_{n=1}^{N}\frac{2\sigma_n^2p_n^{\max}}{D}\right). \qquad (40)$$

If $\mathbb{E}(F(\mathbf{w}_t)-F(\mathbf{w}_{t-1}))\le 0$, the objective decreases monotonically, then FL converges in mean. Thus, to ensure the convergence, we have the following convergence condition

$$\frac{\alpha^2L}{2}\Omega_{CI}-\alpha\omega_{CI}<0. \qquad (41)$$

Now extend the expectation over randomness in the trajectory, and perform a telescoping sum over the T iterations:

$$\begin{aligned}
F(\mathbf{w}_0)-F(\mathbf{w}^*)&\ge F(\mathbf{w}_0)-\mathbb{E}[F(\mathbf{w}_T)]\\
&=\mathbb{E}\left[\sum_{t=1}^{T}(F(\mathbf{w}_{t-1})-F(\mathbf{w}_t))\right]\\
&\ge\mathbb{E}\left[\sum_{t=1}^{T}\left(\left(\alpha\omega_{CI}-\frac{\alpha^2L}{2}\Omega_{CI}\right)\|\mathbf{g}_t\|^2\right.\right.\\
&\qquad\left.\left.-\frac{\alpha^2L}{2}(\Omega_{CI}\delta^2+\epsilon^2z^2)\right)\right]. \qquad (42)
\end{aligned}$$

We can rearrange this inequality to yield the rate:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\left(\left(\alpha\omega_{CI}-\frac{\alpha^2L}{2}\Omega_{CI}\right)\|\mathbf{g}_t\|^2\right)\right]\\
&\quad\le F(\mathbf{w}_0)-F(\mathbf{w}^*)+\frac{\alpha^2L}{2}T(\Omega_{CI}\delta^2+\epsilon^2z^2). \qquad (43)
\end{aligned}$$

If FL converges, the condition (41) holds, yielding $\alpha\omega_{CI}-\frac{\alpha^2L}{2}\Omega_{CI}>0$, and then we get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{T}\|\mathbf{g}_t\|^2\right]\le&\frac{1}{T(\alpha\omega_{CI}-\frac{\alpha^2L}{2}\Omega_{CI})}\left(F(\mathbf{w}_0)-F(\mathbf{w}^*)\right.\\
&\left.+\frac{\alpha^2L}{2}T(\Omega_{CI}\delta^2+\epsilon^2z^2)\right). \qquad (44)
\end{aligned}$$

Let $\alpha=\frac{\omega_{CI}}{L\Omega_{CI}\sqrt{T}}\bar{\alpha}$, where $\bar{\alpha}<2\sqrt{T}$ is a positive constant, and then we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{T}\|\mathbf{g}_t\|^2\right] \le \frac{1}{T\left(\bar{\alpha}\frac{\omega_{CI}^2}{L\Omega_{CI}\sqrt{T}}-\frac{\bar{\alpha}^2\omega_{CI}^2}{2LT\Omega_{CI}}\right)}\left(F(\mathbf{w}_0)\right.\\
&\qquad \left.-F(\mathbf{w}^*)+\frac{\bar{\alpha}^2\omega_{CI}^2}{2L\Omega_{CI}}\left(\delta^2+\frac{1}{\Omega_{CI}}\epsilon^2 z^2\right)\right)\\
&=\frac{1}{T(\frac{\bar{\alpha}}{\sqrt{T}}-\frac{\bar{\alpha}^2}{2T})}\left(\frac{L\Omega_{CI}}{\omega_{CI}^2}(F(\mathbf{w}_0)-F(\mathbf{w}^*))\right.\\
&\qquad \left.+\frac{\bar{\alpha}^2}{2}\left(\delta^2+\frac{1}{\Omega_{CI}}\epsilon^2 z^2\right)\right)\\
&\le\frac{1}{T\frac{\bar{\alpha}}{2\sqrt{T}}}\left(\frac{L\Omega_{CI}}{\omega_{CI}^2}(F(\mathbf{w}_0)-F(\mathbf{w}^*))\right.\\
&\qquad \left.+\frac{\bar{\alpha}^2}{2}\left(\delta^2+\frac{1}{\Omega_{CI}}\epsilon^2 z^2\right)\right)\\
&=\frac{1}{\sqrt{T}\bar{\alpha}}\left(\frac{2L\Omega_{CI}}{\omega_{CI}^2}(F(\mathbf{w}_0)-F(\mathbf{w}^*))\right.\\
&\qquad \left.+\bar{\alpha}^2\left(\delta^2+\frac{1}{\Omega_{CI}}\epsilon^2 z^2\right)\right). \qquad (45)
\end{aligned}$$

Appendix C
Proof of Theorem 3

Given the estimates of the global gradient in (7), the power allocation policy in (11), and the strongest attacks in **Theorem 1**, we get the update rule for model parameters as follows

$$\begin{aligned}
&\mathbf{w}_t=\mathbf{w}_{t-1}-\alpha\tilde{\mathbf{g}}_t\\
&=\mathbf{w}_{t-1}-\alpha\left(\sum_{m=1}^{M}p_{m,t}|h_{m,t}|\mathbf{g}_{m,t}+\epsilon_t\sum_{n=1}^{N}\hat{p}_{n,t}|h_{n,t}|\hat{\mathbf{g}}_{n,t}\right.\\
&\left.+\sum_{n=1}^{N}p_{n,t}|h_{n,t}|\bar{g}_t\mathbf{1}+\epsilon_t\mathbf{z}_t\right)\\
&=\mathbf{w}_{t-1}-\alpha\left(\sum_{m=1}^{M}\sqrt{\frac{p_m^{\max}}{D}}|h_{m,t}|\mathbf{g}_{m,t}+\epsilon_t\mathbf{z}_t\right.\\
&\left.-\epsilon_t\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t}+\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D}}|h_{n,t}|\bar{g}_t\mathbf{1}\right). \qquad (46)
\end{aligned}$$

Substituting (46) to (12), we get

$$\begin{aligned}
&F(\mathbf{w}_t)\le F(\mathbf{w}_{t-1})+\mathbf{g}_t^T(\mathbf{w}_t-\mathbf{w}_{t-1})+\frac{L}{2}\|\mathbf{w}_t-\mathbf{w}_{t-1}\|^2\\
&=F(\mathbf{w}_{t-1})-\alpha\mathbf{g}_t^T\left(\sum_{m=1}^{M}\sqrt{\frac{p_m^{\max}}{D}}|h_{m,t}|\mathbf{g}_{m,t}\right.\\
&-\epsilon_t\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t}+\epsilon_t\mathbf{z}_t\\
&\left.+\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D}}|h_{n,t}|\bar{g}_t\mathbf{1}\right)+\frac{\alpha^2L}{2}\left\|\sum_{m=1}^{M}\sqrt{\frac{p_m^{\max}}{D}}|h_{m,t}|\mathbf{g}_{m,t}\right.\\
&-\epsilon_t\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t}\\
&\left.+\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D}}|h_{n,t}|\bar{g}_t\mathbf{1}+\epsilon_t\mathbf{z}_t\right\|^2. \qquad (47)
\end{aligned}$$

Rearranging this inequality and taking the expectation, we get

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t)-F(\mathbf{w}_{t-1})]\le-\alpha\mathbf{g}_t^T\left(\sum_{m=1}^{M}\sqrt{\frac{p_m^{\max}}{D}}\mathbb{E}[|h_{m,t}|\mathbf{g}_{m,t}]\right.\\
&-\epsilon_t\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}\mathbb{E}[|h_{n,t}|\mathbf{g}_{n,t}]\\
&\left.+\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D}}\mathbb{E}[|h_{n,t}|\bar{g}_t\mathbf{1}]+\mathbb{E}[\epsilon_t\mathbf{z}_t]\right)\\
&+\frac{\alpha^2L}{2}\mathbb{E}\left[\left\|\sum_{m=1}^{M}\sqrt{\frac{p_m^{\max}}{D}}|h_{m,t}|\mathbf{g}_{m,t}+\epsilon_t\mathbf{z}_t\right.\right.\\
&\left.\left.-\epsilon_t\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t}+\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D}}|h_{n,t}|\bar{g}_t\mathbf{1}\right\|^2\right]\\
&\le-\alpha\left(\sum_{i=1}^{M}\sqrt{\frac{p_i^{\max}}{D}}\sigma_i\sqrt{\frac{\pi}{2}}\right.\\
&\left.-\epsilon_t\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}\sigma_n\sqrt{\frac{\pi}{2}}\right)\|\mathbf{g}_t\|^2\\
&+\frac{\alpha^2L}{2}\mathbb{E}\left[\left\|\sum_{m=1}^{M}\sqrt{\frac{p_m^{\max}}{D}}|h_{m,t}|\mathbf{g}_{m,t}+\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D}}|h_{n,t}|\bar{g}_t\mathbf{1}\right.\right.\\
&\left.\left.-\epsilon_t\sum_{n=1}^{N}\sqrt{\epsilon_t^2\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t}+\epsilon_t\mathbf{z}_t\right\|^2\right]\\
&\le-\alpha\sqrt{\frac{\pi}{2}}\left(\sum_{i=1}^{M}\sqrt{\frac{p_i^{\max}}{D}}\sigma_i-\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D}}\sigma_n\right)\|\mathbf{g}_t\|^2\\
&+\frac{\alpha^2L}{2}\mathbb{E}\left[\left\|\sum_{m=1}^{M}\sqrt{\frac{p_m^{\max}}{D}}|h_{m,t}|\mathbf{g}_{m,t}+\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D}}|h_{n,t}|\bar{g}_t\mathbf{1}\right.\right.\\
&\left.\left.-\epsilon_t\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t}+\epsilon_t\mathbf{z}_t\right\|^2\right]. \qquad (48)
\end{aligned}$$

Using the triangle inequality of norms and Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}\Bigg[\Bigg\|\sum_{m=1}^{M}\sqrt{\frac{p_m^{\max}}{D}}|h_{m,t}|\mathbf{g}_{m,t}+\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D}}|h_{n,t}|\bar{g}_t\mathbf{1}\\
&-\epsilon_t\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t}+\epsilon_t\mathbf{z}_t\Bigg\|^2\Bigg]\\
&=\mathbb{E}\Bigg[\Bigg\|\sum_{m=1}^{M}\sqrt{\frac{p_m^{\max}}{D}}|h_{m,t}|\mathbf{g}_{m,t}+\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D}}|h_{n,t}|\bar{g}_t\mathbf{1}\\
&-\epsilon_t\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\mathbf{g}_{n,t}\Bigg\|^2\Bigg]+\mathbb{E}[\|\epsilon_t\mathbf{z}_t\|^2]\\
&\le\mathbb{E}\Bigg[\Bigg(\sum_{m=1}^{M}\sqrt{\frac{p_m^{\max}}{D}}|h_{m,t}|\|\mathbf{g}_{m,t}\|+\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D}}|h_{n,t}|\|\bar{g}_t\mathbf{1}\|\\
&+\epsilon_t\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}}|h_{n,t}|\|\mathbf{g}_{n,t}\|\Bigg)^2\Bigg]+\epsilon^2z^2\\
&\le\mathbb{E}\Bigg[(U+N)\Bigg(\sum_{m=1}^{M}\frac{p_m^{\max}}{D}|h_{m,t}|^2\|\mathbf{g}_{m,t}\|^2\\
&+\sum_{n=1}^{N}\frac{\epsilon_t^2p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}|h_{n,t}|^2\|\mathbf{g}_{n,t}\|^2\\
&+\sum_{n=1}^{N}\frac{p_n^{\max}}{D}|h_{n,t}|^2\|\mathbf{g}_t\|^2\Bigg)\Bigg]+\epsilon^2z^2\\
&\le(U+N)\Bigg(\sum_{i=1}^{U}\frac{p_i^{\max}}{D}2\sigma_i^2\|\mathbf{g}_t\|^2+\sum_{m=1}^{M}\frac{p_m^{\max}}{D}2\sigma_m^2\delta^2\\
&+\sum_{n=1}^{N}\frac{\epsilon_t^2p_n^{\max}}{D(\epsilon_t^2+\bar{g}_t^2)}2\sigma_n^2\delta^2\Bigg)+\epsilon^2z^2\\
&\le(U+N)\Bigg(\sum_{i=1}^{U}\frac{p_i^{\max}}{D}2\sigma_i^2\|\mathbf{g}_t\|^2+\sum_{i=1}^{U}\frac{p_i^{\max}}{D}2\sigma_i^2\delta^2\Bigg)+\epsilon^2z^2.
\end{aligned}\tag{49}$$

Substituting (49) to (48), we get

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t)-F(\mathbf{w}_{t-1})]\\
&\le-\alpha\sqrt{\frac{\pi}{2}}\Bigg(\sum_{i=1}^{M}\sqrt{\frac{p_i^{\max}}{D}}\sigma_i-\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}}{D}}\sigma_n\Bigg)\|\mathbf{g}_t\|^2\\
&+\frac{\alpha^2L}{2}\Bigg((U+N)\Bigg(\sum_{i=1}^{U}\frac{p_i^{\max}}{D}2\sigma_i^2\|\mathbf{g}_t\|^2\\
&+\sum_{i=1}^{U}\frac{p_i^{\max}}{D}2\sigma_i^2\delta^2\Bigg)+\epsilon^2z^2\Bigg)\\
&=\Big(\frac{\alpha^2L}{2}\Omega_{BEV}-\alpha\omega_{BEV}\Big)\|\mathbf{g}_t\|^2+\frac{\alpha^2L}{2}(\Omega_{BEV}\delta^2+\epsilon^2z^2),
\end{aligned}\tag{50}$$

where

$$\omega_{BEV}=\sum_{i=1}^{M}\sqrt{\frac{p_i^{\max}\pi}{2D}}\sigma_i-\sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}\pi}{2D}}\sigma_n,\tag{51}$$

$$\Omega_{BEV}=(U+N)\sum_{i=1}^{U}\frac{2\sigma_i^2p_i^{\max}}{D}.\tag{52}$$

If $\mathbb{E}(F(\mathbf{w}_t)-F(\mathbf{w}_{t-1}))\le 0$, the objective decreases monotonically, then FL converges in mean. Thus, to ensure the convergence, we have the following convergence condition

$$\frac{\alpha^2L}{2}\Omega_{BEV}-\alpha\omega_{BEV}<0.\tag{53}$$

Now extend the expectation over randomness in the trajectory, and perform a telescoping sum over the T iterations:

$$\begin{aligned}
&F(\mathbf{w}_0)-F(\mathbf{w}^*)\ge F(\mathbf{w}_0)-\mathbb{E}[F(\mathbf{w}_T)]\\
&\qquad=\mathbb{E}\Bigg[\sum_{t=1}^{T}(F(\mathbf{w}_{t-1})-F(\mathbf{w}_t))\Bigg]\\
&\qquad\ge\mathbb{E}\Bigg[\sum_{t=1}^{T}\Bigg(\Big(\alpha\omega_{BEV}-\frac{\alpha^2L}{2}\Omega_{BEV}\Big)\|\mathbf{g}_t\|^2\\
&\qquad-\frac{\alpha^2L}{2}(\Omega_{BEV}\delta^2+\epsilon^2z^2)\Bigg)\Bigg].
\end{aligned}\tag{54}$$

We can rearrange this inequality to yield the rate:

$$\begin{aligned}
&\mathbb{E}\Bigg[\sum_{t=1}^{T}\Big(\alpha\omega_{BEV}-\frac{\alpha^2L}{2}\Omega_{BEV}\Big)\|\mathbf{g}_t\|^2\Bigg]\\
&\quad\le F(\mathbf{w}_0)-F(\mathbf{w}^*)+\frac{\alpha^2L}{2}\sum_{t=1}^{T}(\Omega_{BEV}\delta^2+\epsilon^2z^2).
\end{aligned}\tag{55}$$

If FL converges, $\alpha\omega_{BEV}-\frac{\alpha^2L}{2}\Omega_{BEV}>0$, and then we get

$$\begin{aligned}
&\mathbb{E}\Bigg[\sum_{t=1}^{T}\frac{1}{T}\|\mathbf{g}_t\|^2\Bigg]\\
&\le\frac{F(\mathbf{w}_0)-F(\mathbf{w}^*)+\frac{\alpha^2L}{2}\sum_{t=1}^{T}(\Omega_{BEV}\delta^2+\epsilon^2z^2)}{T(\alpha\omega_{BEV}-\frac{\alpha^2L}{2}\Omega_{BEV})}.
\end{aligned}\tag{56}$$

Let $\alpha=\frac{\omega_{BEV}}{L\Omega_{BEV}\sqrt{T}}\bar{\alpha}$, where $\bar{\alpha}<2\sqrt{T}$ is a positive constant, and then we have

$$\begin{aligned}
&\mathbb{E}\Bigg[\sum_{t=1}^{T}\frac{1}{T}\|\mathbf{g}_t\|^2\Bigg]\\
&\le\frac{F(\mathbf{w}_0)-F(\mathbf{w}^*)+\frac{\bar{\alpha}^2\omega_{BEV}^2}{2L\Omega_{BEV}}\Big(\delta^2+\frac{1}{\Omega_{BEV}}\epsilon^2z^2\Big)}{T\Big(\bar{\alpha}\frac{\omega_{BEV}^2}{L\Omega_{BEV}\sqrt{T}}-\frac{\bar{\alpha}^2\omega_{BEV}^2}{2LT\Omega_{BEV}}\Big)}\\
&=\frac{\frac{L\Omega_{BEV}}{\omega_{BEV}^2}(F(\mathbf{w}_0)-F(\mathbf{w}^*))+\frac{\bar{\alpha}^2}{2}\Big(\delta^2+\frac{1}{\Omega_{BEV}}\epsilon^2z^2\Big)}{T(\frac{\bar{\alpha}}{\sqrt{T}}-\frac{\bar{\alpha}^2}{2T})}\\
&\le\frac{\frac{L\Omega_{BEV}}{\omega_{BEV}^2}(F(\mathbf{w}_0)-F(\mathbf{w}^*))+\frac{\bar{\alpha}^2}{2}\Big(\delta^2+\frac{1}{\Omega_{BEV}}\epsilon^2z^2\Big)}{T\frac{\bar{\alpha}}{2\sqrt{T}}}\\
&=\frac{\frac{2L\Omega_{BEV}}{\bar{\alpha}\omega_{BEV}^2}(F(\mathbf{w}_0)-F(\mathbf{w}^*))+\bar{\alpha}\Big(\delta^2+\frac{1}{\Omega_{BEV}}\epsilon^2z^2\Big)}{\sqrt{T}}.
\end{aligned}\tag{57}$$

REFERENCES

[1] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Bev-sgd: Best effort voting sgd against byzantine attacks for analog aggregation based federated learning over the air," in *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2022, pp. 1–6.

[2] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE communications magazine*, vol. 58, no. 1, pp. 19–25, 2020.

[3] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[4] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[5] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[6] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[7] 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.

[8] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 440–445.

[9] Y. Lin, S. Han, H. Mao, Y. Wang, and B. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *International Conference on Learning Representations*, 2018.

[10] 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.

[11] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[12] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," in *Advances in Neural Information Processing Systems*, 2017, pp. 1709–1720.

[13] Y. Liu, W. Xu, G. Wu, Z. Tian, and Q. Ling, "Communication-censored admm for decentralized consensus optimization," *IEEE Transactions on Signal Processing*, vol. 67, no. 10, pp. 2565–2579, 2019.

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

[15] T. Chen, G. Giannakis, T. Sun, and W. Yin, "Lag: Lazily aggregated gradient for communication-efficient distributed learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 5050–5060.

[16] P. Xu, Z. Tian, and Y. Wang, "An energy-efficient distributed average consensus scheme via infrequent communication," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2018, pp. 648–652.

[17] P. Xu, Y. Wang, X. Chen, and T. Zhi, "Coke: Communication-censored kernel learning for decentralized non-parametric learning," *arXiv preprint arXiv:2001.10133*, 2020.

[18] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Joint optimization for federated learning over the air," in *2022 IEEE International Conference on Communications (ICC 2022)*. IEEE, 2022, pp. 1–6.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] ——, "1-bit compressive sensing for efficient federated learning over the air," *arXiv preprint arXiv:2103.16055*, 2021.

[26] 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.

[27] ——, "Federated learning over wireless fading channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3546–3557, 2020.

[28] M. M. Amiri, T. M. Duman, and D. Gündüz, "Collaborative machine learning at the wireless edge with blind transmitters," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2019, pp. 1–5.

[29] Y. Sun, S. Zhou, and D. Gündüz, "Energy-aware analog aggregation for federated learning with redundant data," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–7.

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

[31] M. Gastpar, "Uncoded transmission is exactly optimal for a simple gaussian "sensor" network," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 5247–5251, 2008.

[32] O. Abari, H. Rahul, D. Katabi, and M. Pant, "Airshare: Distributed coherent transmission made seamless," in *2015 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2015, pp. 1742–1750.

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

[34] S. Wang, Y. Hong, R. Wang, Q. Hao, Y.-C. Wu, and D. W. K. Ng, "Edge federated learning via unit-modulus over-the-air computation (extended version)," *arXiv preprint arXiv:2101.12051*, 2021.

[35] C. Xu, S. Liu, Z. Yang, Y. Huang, and K.-K. Wong, "Learning rate optimization for federated learning exploiting over-the-air computation," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3742–3756, 2021.

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

[37] 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.

[38] 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.

[39] Y. Dong, J. Cheng, M. J. Hossain, and V. C. Leung, "Secure distributed on-device learning networks with byzantine adversaries," *IEEE Network*, vol. 33, no. 6, pp. 180–187, 2019.

[40] G. Damaskinos, E. M. El Mhamdi, R. Guerraoui, A. H. A. Guirguis, and S. L. A. Rouault, "Aggregathor: Byzantine machine learning via robust gradient aggregation," in *The Conference on Systems and Machine Learning (SysML), 2019*, no. CONF, 2019.

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

[42] Z. Wu, Q. Ling, T. Chen, and G. B. Giannakis, "Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks," *IEEE Transactions on Signal Processing*, vol. 68, pp. 4583–4596, 2020.

[43] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 2, pp. 1–25, 2017.

[44] S. Huang, Y. Zhou, T. Wang, and Y. Shi, "Byzantine-resilient federated machine learning via over-the-air computation," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2021, pp. 1–6.

[45] 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.

[46] R. Guerraoui, S. Rouault *et al.*, "The hidden vulnerability of distributed learning in byzantium," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3521–3530.

[47] E.-M. El-Mhamdi, R. Guerraoui, and S. Rouault, "Fast and robust distributed learning in high dimension," in *2020 International Symposium on Reliable Distributed Systems (SRDS)*. IEEE, 2020, pp. 71–80.

[48] C. Xie, O. Koyejo, and I. Gupta, "Zeno: Byzantine-suspicious stochastic gradient descent," *arXiv preprint arXiv:1805.10032*, vol. 24, 2018.

[49] C. Xie, S. Koyejo, and I. Gupta, "Zeno++: Robust fully asynchronous sgd," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 495–10 503.

[50] S. Xia, J. Zhu, Y. Yang, Y. Zhou, Y. Shi, and W. Chen, "Fast convergence algorithm for analog federated learning," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.

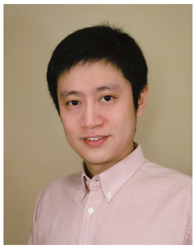
[51] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 170–185, 2020.

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

[53] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signsgd: Compressed optimisation for non-convex problems," in *International Conference on Machine Learning*. PMLR, 2018, pp. 560–569.

[54] J. Wang and G. Joshi, "Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms," in *ICML Workshop on Coding Theory for Machine Learning*, 2019.

Xin Fan (Student Member, IEEE) received his B.E. degree and M.E. degree from School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China, in 2016 and 2018, respectively. He is currently a Ph.D. student in Beijing Jiaotong University from 2018, and a visiting Ph.D. student in the Electrical and Computer Engineering Department of George Mason University, Fairfax, VA, USA, from 2020. His current research interests lie in the areas of wireless communications, security and privacy, optimization, statistical signal processing, blockchain, and machine learning.



Yue Wang (Senior Member, IEEE) received the Ph.D. degree in communication and information system from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China, in 2011. He is currently a Research Assistant Professor with Electrical and Computer Engineering Department, George Mason University, Fairfax, VA, USA, where he was a Postdoctoral Researcher. Prior to that, he was a Senior Engineer with Huawei Technologies Co., Ltd., China. From 2009 to 2011, he was a Visiting Ph.D. Student with Electrical and Computer Engineering Department, Michigan Technological University, Houghton, MI, USA. His general interests include signal processing, wireless communications, machine learning, and their applications in cyber physical systems. His specific research focuses on compressive sensing, massive MIMO, millimeter-wave communications, cognitive radios, DoA estimation, high-dimensional data analysis, and distributed optimization and learning.

Yan Huo (Senior Member, IEEE) received the B.E. and Ph.D. degrees in communication and information system from Beijing Jiaotong University, Beijing, China, in 2004 and 2009, respectively. He was a visiting scholar with the Department of Computer Science, The George Washington University, from 2015 to 2016. He is currently a Professor with the School of Electronics and Information Engineering, Beijing Jiaotong University. His current research interests include wireless communications, physical layer security, privacy protection, and edge computing. He has served as an associate editor for the IEEE Access and a Reviewer for a number of journals, including the IEEE Wireless Communications, the IEEE Internet of Things Journal, the IEEE Transactions on Wireless Communications, the IEEE Transactions on Vehicular Technology, and the IEEE Transactions on Mobile Computing.

Zhi Tian (Fellow, IEEE) is a Professor in the Electrical and Computer Engineering Department of George Mason University, Fairfax, VA, USA. Previously, she was on the faculty of Michigan Technological University from 2000 to 2014. She served as a Program Director at the US National Science Foundation from 2012 to 2014. Her research interest lies in the areas of wireless communications, statistical signal processing, and machine learning. Current research focuses on massive MIMO, millimeter-wave communications, and distributed network optimization and learning. She was an IEEE Distinguished Lecturer for both the IEEE Communications Society and the IEEE Vehicular Technology Society. She served as Associate Editor for IEEE Transactions on Wireless Communications and IEEE Transactions on Signal Processing. She is a Member-of-Large of the IEEE Signal Processing Society (2019-2021). She received the IEEE Communications Society TCCN Publication Award in 2018.