1-Bit Compressive Sensing for Efficient Federated Learning Over the Air

Xin Fan, *Graduate Student Member, IEEE* Yue Wang, *Member, IEEE*, Yan Huo, *Senior Member, IEEE*, and Zhi Tian, *Fellow, IEEE*

***Abstract*—For distributed learning among collaborative users, this paper develops and analyzes a communication-efficient scheme for federated learning (FL) over the air, which incorporates 1-bit compressive sensing (CS) into analog-aggregation transmissions. To facilitate design parameter optimization, we analyze the efficacy of the proposed scheme by deriving a closed-form expression for the expected convergence rate. Our theoretical results unveil the tradeoff between convergence performance and communication efficiency as a result of the aggregation errors caused by sparsification, dimension reduction, quantization, signal reconstruction and noise. Then, we formulate a joint optimization problem to mitigate the impact of these aggregation errors through joint optimal design of worker scheduling and power scaling policy. An enumeration-based method is proposed to solve this non-convex problem, which is optimal but becomes computationally infeasible as the number of devices increases. For scalable computing, we resort to the alternating direction method of multipliers (ADMM) technique to develop an efficient implementation that is suitable for large-scale networks. Simulation results show that our proposed 1-bit CS based FL over the air achieves comparable performance to the ideal case where conventional FL without compression and quantification is applied over error-free aggregation, at much reduced communication overhead and transmission latency.**

***Index Terms*—Federated learning, analog aggregation, 1-bit compressive sensing, convergence analysis, joint optimization.**

I. Introduction

Centralized machine learning (ML) that collects distribute data from edge devices (local workers) to a parameter server (PS) for data analysis and inference is costly in communications for big data applications. Although the adoption of approximate data aggregation instead of exact data aggregation proposed by [2], [3] can effectively reduce communication costs, the privacy issues exposed by data collection cannot be ignored. As an alternative, federated learning (FL) is a promising paradigm that enables many local workers to collaboratively train a common learning model under the coordination of a PS in wireless networks [1], [4], [5]. In FL, local devices (workers) and the PS exchange individually updated model parameters in each iteration, until convergence. Without exchanging raw datasets during the iterations between the PS and local workers, FL offers distinct advantages on protecting user privacy and leveraging distributed on-device computation compared to traditional learning at a centralized data center.

In FL, model updates shared between local workers and the PS can be extremely large, e.g., the VGGNet architecture has approximately 138 million model parameters [6]. As a result, model compression has been considered in the literature to reduce the communication load per worker, such as *sparsification*, *quantization* and *communication censoring* schemes. *sparsification* schemes only keep the large values of local updates to reduce the communication load [7], [8]. *Quantization* is to compress the continuous-valued update information to a few finite bits so that it can be effectively communicated over a digital channel [9]–[11]. *Communication censoring* is to evaluate the importance of each update in order to avoid less informative transmissions [12]–[16]. All these useful strategies are investigated predominantly for FL with digital communications.

However, the communication overhead and transmission latency of FL over digital communication channels are proportional to the number of active workers, and thus cannot be applicable in large-scale environments. To overcome this problem, an analog aggregation model is recently proposed for FL [17]–[28] by allowing multiple workers to simultaneously transmit their updates over the same time-frequency resources and then applying a computation-over-the-air principle [29]. It benefits from the fact that FL only relies on the averaged value of distributed local updates rather than their individual values. Exploiting the waveform superposition property of a wireless multiple access channel (MAC), analog aggregation automatically enables to directly obtain the averaged updates required by FL, which propels the prosperity of analog aggregation based FL over the air (FLOA). In [17], a broadband analog aggregation scheme was designed for FLOA, in which a set of tradeoffs between communication and learning are derived for broadband power control and device scheduling, where the learning metric is set as the fraction of scheduled devices. In [21], a power control policy was proposed to minimize the mean square error (MSE) of the aggregated signal. Similarly, a joint design of device scheduling and beamforming was presented in [18] for FLOA in multiple antenna systems, which

Manuscript received 13 July 2021; revised 26 December 2021 and 24 May 2022; accepted 20 September 2022. This work was partly supported by Beijing Natural Science Foundation (Grant #4202054),the National Natural Science Foundation of China (Grants #61871023 and #61931001), the US National Science Foundation (Grants #1939553, #2003211, #2128596, #2136202 and #2231209), and the Virginia Research Investment Fund (Commonwealth Cyber Initiative Grant #223996). Part of this paper has been presented at the 2021 IEEE International Conference on Communications Workshops (ICC Workshops), Montreal, Canada, Jun. 14-23, 2021 [1]. (Corresponding author: Yan Huo.)

X. Fan and Y. 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).

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

aims to maximize the number of selected workers under the given MSE requirements. Based on one-bit gradient quantization, a digital version of broadband over-the-air aggregation was proposed, and the effects of wireless channel hostilities on the convergence rate was analyzed in [19]. In [23], [24], the gradient sparsification, and a random linear projection for dimensionality reduction of large-size gradient in narrowband channels was considered to reduce the communication requirements. The power allocation scheme in [23], [24] scales the power of the vectors containing the gradient information of different devices to satisfy the average power constraint. In [26], [28], a robust power control scheme for FLOA against Byzantine attacks.

Despite the prior work, some fundamental questions remain unanswered, which however prevent from achieving communication-efficient and high-performance FLOA. Firstly, the quantitative relationship between FL and analog aggregation communication is not clear [20], [22]. Simple maximization of the number of participated workers is learning-agnostic and hence not necessarily optimal, which decouples the optimization of computation and communication, e.g., the works in [18], [27]. Secondly, to facilitate power control, most existing works are developed based on a strong assumption that the signals to be transmitted from local workers, i.e., local gradients, can be normalized to have zero mean and unit variance [17]–[19], [21]. However, gradient statistics in FL vary over both training iterations and feature dimensions, and are unknown a priori [30]. Thus, it is infeasible to design an optimal power control without prior knowledge of the local gradients at the PS, especially for the uncoded linear analog modulation in FLOA. Thirdly, sparsification is introduced for communication efficiency in FLOA [23], [24] as a means of lossy compression of local gradients, which may introduce aggregation errors, but the impact of these aggregation errors on FL is not yet clear, let alone how to alleviate their side effects.

To solve the aforementioned issues, in this paper, we introduce 1-bit compressive sensing (CS) for efficient FLOA, by developing an optimized practical worker selection and power control policy. To the best of our knowledge, this is the first work to introduce 1-bit CS [31]–[33] into FLOA for high communication efficiency, where both the dimension of local gradients and the number of quantization bits can be reduced significantly. Further, thanks to the 1-bit quantization, our power control becomes feasible since it hinges on the quantized values of known magnitude, without relying on any prior knowledge or assumptions on gradient statistics or specific distribution. More importantly, our work provides an essential interpretation on the relationship between FL and analog aggregation with 1-bit CS techniques to enable joint optimization of computation and communications. Our main contributions are outlined below[1]:

- We propose a **o**ne-**b**it **CS** **a**nalog **a**ggregation (OBCSAA) technique for efficient FL. In our OBCSAA, we elaborately design a set of compression, analog aggregation transmission, signal reconstruction solutions to achieve communication-efficient FL.
- We derive a closed-form expression for the expected convergence rate of our OBCSAA. This closed-form expression measures the performance tradeoff as a result of the aggregation errors caused by sparsification, dimension reduction, quantization, signal reconstruction and additive white gaussian noise (AWGN), which provides a fresh perspective to design analog wireless systems.
- Guided by the theoretical results, we formulate a joint optimization problem of computation and communication to optimize the worker selection and power control. Given the practical limitation on allowable peak transmit power and available bandwidth, this optimization problem aims to mitigate the aggregation errors. To solve this non-convex optimization problem, we propose two solutions: the enumeration-based method and the alternating direction method of multipliers (ADMM) approach for the scenarios of small networks and large networks, respectively.

It is worth noting that, there exist related works that apply 1-bit CS to FL [34], analyze the convergence of FL algorithms with information compression methods [35]–[41] and optimize the convergence errors [38]. However, they are originally designed for the vanilla FL over digital-communication links, which cannot be directly applied to FLOA settings of this work. In fact, there is a big difference between vanilla FL and FLOA from the aspects of local-gradient aggregation and global-gradient computation. In vanilla FL over digital communications, the knowledge of local updates at individual workers can be extracted and leveraged for resource optimization and algorithm design, which is however inaccessible in FLOA due to the analog aggregation nature. This major difference as well as the physical-layer aspects of wireless connections gives rise to technical challenges to deal with in the approach design and system optimization for FLOA, including the design for 1-bit CS, optimizing the communication resource allocation and transmission scheduling in practical implementation, as they are not yet fully considered and well explored by the current literature [17]–[19], [21], [23], [24]. Compared to [17], [18], [21] that consider the fraction of scheduled devices as the learning metric which separates communication and computation, our learning metric is learning convergence with respect to CS and communication factors, which hence provides the exact relationship between communication and computation. Different from [17]–[19], [21] developed on the assumption that the local updates have to follow independent and identically distributed (IID) with zero mean and unit variance, our work adopts 1-bit CS, which enables to achieve power control for individual workers even without any gradient statistical information required by [17]–[19], [21]. Compared to [23], [24], our work not only applies the 1-bit quantization after dimensionality reduction, but also provides convergence analysis on 1-bit CS based FLOA, which leads to joint optimization of computation and communications. Through signal reconstruction, we recover the sum of sparse original local gradients for training, which

[1]Part of this paper has been presented at the 2021 IEEE International Conference on Communications Workshops (ICC Workshops), Montreal, Canada, Jun. 14-23, 2021 [1].

is different from the existing work [19] that uses quantized local gradients. In short, our work is a holistic integration of gradient sparsification, dimensionality reduction, quantization and signal reconstruction for efficient FLOA.

We evaluate the proposed OBCSAA in solving image classification problems on the MNIST dataset. Simulation results show that our proposed OBCSAA achieves comparable performance to the existing work [19] and the ideal case where FL is implemented by perfect aggregation over error-free wireless channels, with much enhanced communication efficiency.

The rest of this paper is organized as follows. The system model of 1-bit compressive sensing for FL over the air is presented in Section II. The closed-form expression of the expected convergence rate is derived in Section III to quantify the impact of the aggregation errors on FL. A joint optimization problem of communication and FL to optimize worker selection and power control are studied in Section IV. In Section V, we provide discussions on the convergence analysis and algorithm design guideline of our work in the stochastic gradient descent (SGD) case. Numerical results are presented in Section VI, and conclusions are drawn in Section VII.

II. System Model

We consider a wireless FL system consisting of a single PS and U local workers. Exploiting wireless analog aggregation transmissions with 1-bit CS, the PS and all local workers collaboratively train a shared learning model.

A. FL Model

Suppose that the union of all training datasets is denoted as $\mathfrak{D} = \bigcup_i \mathfrak{D}_i$, where $\mathfrak{D}_i = \{\mathbf{x}_{i,k}, \mathbf{y}_{i,k}\}_{k=1}^{K_i}$ is the local dataset and $K_i = |\mathfrak{D}_i|$ is the number of data samples at the i-th worker, $i = 1, \ldots, U$. In $\mathfrak{D}_i$, the k-th data sample and its label are denoted as $\mathbf{x}_{i,k}$ and $\mathbf{y}_{i,k}$, $k = 1, 2, ..., K_i$, respectively. The objective of the training procedure is to minimize the global loss function $F(\mathbf{w}; \mathfrak{D})$ of the global shared learning model parameterized by $\mathbf{w} = [w^1, \ldots, w^D] \in \mathcal{R}^D$ of the dimension D, i.e.,

$$\textbf{P1:} \quad \mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathcal{R}^D} \quad F(\mathbf{w}; \mathfrak{D}), \tag{1}$$

where $F(\mathbf{w}; \mathfrak{D}) = \frac{1}{K} \sum_{i=1}^{U} \sum_{k=1}^{K_i} f(\mathbf{w}; \mathbf{x}_{i,k}, \mathbf{y}_{i,k})$ is the sum of $K = \sum_{i=1}^{U} K_i$ sample-wise loss functions defined by the learning model. Note that the solution $\mathbf{w}^*$ can be a local optimum or a saddle point in the non-convex case, or the global optimum in the convex case.

To avoid directly uploading the raw local datasets to the PS for centralized training, the learning procedure in (**P1**) is conducted in a distributed manner by an iterative gradient-averaging algorithm [4], [42]. Specifically, at each iteration t, a first-order algorithm, such as the gradient descent (GD)[2], is applied at local workers in parallel to minimize the local loss functions

$$F_i(\mathbf{w}_i; \mathfrak{D}_i) = \frac{1}{K_i} \sum_{k=1}^{K_i} f(\mathbf{w}_i; \mathbf{x}_{i,k}, \mathbf{y}_{i,k}), \quad i = 1, ..., U, \tag{2}$$

where $\mathbf{w}_i = [w_i^1, \ldots, w_i^D] \in \mathcal{R}^D$ is the local model parameter. Each local worker updates its local gradient from the received global learning model given its own local dataset:

$$\mathbf{g}_i = \frac{1}{K_i} \sum_{k=1}^{K_i} \nabla f(\mathbf{w}_i; \mathbf{x}_{i,k}, \mathbf{y}_{i,k}), \quad i = 1, ..., U, \tag{3}$$

where $\nabla f(\mathbf{w}_i; \mathbf{x}_{i,k}, \mathbf{y}_{i,k})$ is the gradient of $f(\mathbf{w}_i; \mathbf{x}_{i,k}, \mathbf{y}_{i,k})$ with respect to $\mathbf{w}_i$.

Then the local gradients are sent to the PS, which are aggregated as the global gradient:

$$\mathbf{g} = \frac{1}{K} \sum_{i=1}^{U} K_i \mathbf{g}_i, \tag{4}$$

and the global gradient $\mathbf{g}$ is sent back to the local workers, which is then used to update the shared model as

$$\mathbf{w} = \mathbf{w} - \alpha \mathbf{g}, \tag{5}$$

where α is the learning rate.

The FL implements (3), (4) and (5) iteratively[3], until it converges or the maximum number of iterations is reached.

B. Analog Aggregation Transmission Model

In the scenarios of FL applied over large-scale networks and for training a high-dimensional model parameters, the transmissions between the PS and local workers consume a lot of communication resources and cause training latency. Meanwhile, due to the transmit power and bandwidth limitations posed by practical wireless communications, the digital communication approach of transmitting and reconstructing all the gradient entries one-by-one in an individual manner is an overkill. Thus, in order to reduce the transmission overhead and speed up communication time, we propose to apply 1-bit compressive sensing [31]–[33] in FL over the air, which is motivated by two facts. One is that the gradients involved in large-size learning problems usually turn out to be compressible with only a small number of entries having significant values [7], [43], [44]. The other is that FL is usually running in an average-based distributed learning mechanism. In our work, through gradient sparsification, the compression nature of CS allows to reduce the dimensionality of the transmitted gradient vectors. Meanwhile, analog aggregation enables all local workers to simultaneously use the same time-frequency resources to transmit their updates to the PS. Further, the 1-bit quantization not only minimizes the quantization overhead, but also circumvents the unrealistic requirement on known distribution of local gradients. The procedure of the proposed 1-bit CS method for FL is elaborated next.

[2]In this work, we take the basic gradient descent as an example, which can be extended to the stochastic gradient descent (SGD) by using a mini-batch at each worker for training, as shown in Section V. Note that SGD works more computation-efficient at the cost of more iterations and hence more transmissions compared to GD.

[3]Practical implementation involves updating over a mini-batch of samples per iterations and multiple rounds of iterations before communications. However, they do not affect the conceptual development in this paper.

1) Sparsification: Before transmission at the t-th iteration, all local workers set all but the κ elements of their local $\mathbf{g}_{i,t}$'s to 0, resulting in κ-level sparsification denoted by

$$\tilde{\mathbf{g}}_{i,t} = \texttt{sparse}_{\kappa}(\mathbf{g}_{i,t}), \tag{6}$$

where $\texttt{sparse}_{\kappa}(\cdot)$ is a sparsification operation of a vector such that $\tilde{\mathbf{g}}_{i,t}$ is of length D and sparsity order κ. In our paper, we perform the top-κ sparsification strategy as an example, i.e., elements with the largest κ magnitudes are retained while other elements are set to 0.

2) Dimensionality Reduction: To transmit the non-zero entries of their sparsified local gradient vectors, the workers need to transmit the indices and values of the non-zero entries to the PS separately, which results in additional data transmissions. To avoid this overhead, we use a similar method as in [24] that all workers employ the same measurement matrix $\mathbf{\Phi} \in \mathbb{R}^{S\times D}$ ($S \ll D$) that is a random Gaussian matrix. Note that the specific κ-nonzero indices of sparse gradients after the top-κ sparsification are usually different worker by worker[4], which results in an increased sparsity-level $\bar{\kappa}$ ($> \kappa$) for the superposed gradient signal. For reliable reconstruction of the compressed gradients, it is desired that the restricted isometry property (RIP) condition be met, that is, $\kappa U \leq S \ll D$ and each entry of $\mathbf{\Phi}$ i.i.d. follows $\mathcal{N}(0, \sigma_{sp}^2)$, where κU is the upper bound of the sparsity of the combined sparse gradient, i.e., $\kappa U > \bar{\kappa}$. In addition, $\mathbf{\Phi}$ is shared between the workers and the PS before transmission.

3) Quantization: Next, 1-bit quantization is applied to $\mathbf{\Phi}\tilde{\mathbf{g}}_{i,t}$'s, so that the resulting compressed local gradient $\mathcal{C}(\mathbf{g}_{i,t})$ at each worker is given by

$$\begin{aligned}\mathcal{C}(\mathbf{g}_{i,t}) &= \texttt{sign}(\mathbf{\Phi}\texttt{sparse}_{\kappa}(\mathbf{g}_{i,t})) \\ &= \texttt{sign}(\mathbf{\Phi}\tilde{\mathbf{g}}_{i,t}), \quad i = 1, ..., U,\end{aligned} \tag{7}$$

where $\mathcal{C}(\cdot)$ represents the overall effective operation including top-κ sparsification, CS compression, and 1-bit quantization. We denote $\mathcal{C}(\mathbf{g}_{i,t}) = [c_{i,t}^1, ..., c_{i,t}^s, ..., c_{i,t}^S]^T$, where $c_{i,t}^s = \pm 1$ due to 1-bit quantization.

4) Analog Aggregation Transmission: After collecting the compressive measurements as in (7), all the workers, subject to power control, transmit their local $\mathcal{C}(\mathbf{g}_{i,t})$'s in an analog fashion, which are aggregated over the air at the PS to implement the global gradient updating step in (4). Specifically, each local $\mathcal{C}(\mathbf{g}_{i,t})$ at worker i is multiplied by a power control factor $p_{i,t}$, and then sent to the PS over the air. When all participating users transmit synchronously, the received signal vector at the PS is given by

$$\mathbf{y}_t = \sum_{i=1}^{U} h_{i,t} p_{i,t} \mathcal{C}(\mathbf{g}_{i,t}) + \mathbf{z}_t, \tag{8}$$

where $\mathbf{z}_t \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is AWGN vector, and $h_{i,t}$ denotes the channel coefficient between the i-th local worker and the PS at the t-th iteration[5].

Power control includes both worker selection and transmit power scaling. Let $\beta_{i,t}$ denote the worker selection indicator, i.e., $\beta_{i,t} = 1$ indicates that the i-th worker at the t-th iteration is scheduled to the FL algorithm, and $\beta_{i,t} = 0$, otherwise. To implement the averaging gradient step in (4), the signal vector of interest at the PS at the t-th iteration is given by

$$\mathbf{y}_t^{desired} = \frac{\sum_{i=1}^{U} K_i \beta_{i,t} \mathcal{C}(\mathbf{g}_{i,t})}{\sum_{i=1}^{U} K_i \beta_{i,t}}. \tag{9}$$

To obtain the signal vector of interest, we design the power control factor $p_{i,t}$ as

$$p_{i,t} = \frac{\beta_{i,t} K_i b_t}{h_{i,t}}, \tag{10}$$

where b_t is a power scaling factor. Through this power scaling, the transmit power at the i-th local worker satisfies the power limitation P_i^{Max} as

$$|p_{i,t} c_{i,t}^s|^2 = \left(\frac{\beta_{i,t} K_i b_t}{h_{i,t}} c_{i,t}^s\right)^2 = \frac{\beta_{i,t}^2 K_i^2 b_t^2}{h_{i,t}^2} \leq P_i^{\text{Max}}, \tag{11}$$

where $c_{i,t}^s$ is eliminated, due to $c_{i,t}^s = \pm 1$.

Remark 1. As we can see from (11), the power is independent of the specific local gradient, which enables the optimization of power control in the absence of any the prior knowledge of the gradients or gradient statistics. This is one of our motivation to incorporate 1-bit CS into FL over the air.

After applying the power control $p_{i,t}$ and substituting (10) into (8), the received signal vector of (8) can be rewritten as

$$\mathbf{y}_t = \sum_{i=1}^{U} K_i b_t \beta_{i,t} \mathcal{C}(\mathbf{g}_{i,t}) + \mathbf{z}_t. \tag{12}$$

Upon receiving $\mathbf{y}_t$, the PS estimates the signal vector of interest via a post-processing operation as[6]

$$\begin{aligned}\hat{\mathbf{y}}_t^{desired} =& (\sum_{i=1}^{U} K_i \beta_{i,t} b_t)^{-1} \mathbf{y}_t \\ =& (\sum_{i=1}^{U} K_i \beta_{i,t})^{-1} \sum_{i=1}^{U} K_i \beta_{i,t} \mathcal{C}(\mathbf{g}_{i,t}) + (\sum_{i=1}^{U} K_i \beta_{i,t} b_t)^{-1} \mathbf{z}_t \\ =& \mathbf{y}_t^{desired} + \frac{\mathbf{z}_t}{\sum_{i=1}^{U} K_i \beta_{i,t} b_t},\end{aligned} \tag{13}$$

where $(\sum_{i=1}^{U} K_i \beta_{i,t} b_t)^{-1}$ is the post-processing factor.

[4]When distributed workers have i.i.d. data, their κ-nonzero indices turn to appear with large overlapping.

[5]In this paper, we consider block fading channels, where the channel state information (CSI) remains unchanged within each iteration in FL, but may independently vary from one iteration to another. We assume that the CSI is perfectly known at both the PS and local workers so that the channel phase offset can be compensated at the local workers before they transmit their gradient updates.

[6]It is noted that this channel inversion method can be improved by adopting truncated channel values in the policy, which leads to better learning performance for FL over deep fading channels [17].

5) Reconstruction: After obtaining $\hat{\mathbf{y}}_t^{desired}$ from (13), the PS needs to use a CS reconstruction algorithm $\mathcal{C}^{-1}(\cdot)$ to estimate the global gradient $\hat{\mathbf{g}}_t = \mathcal{C}^{-1}(\hat{\mathbf{y}}_t^{desired})$. Many options for $\mathcal{C}^{-1}(\cdot)$ are available, such as the binary iterative hard thresholding (BIHT) algorithm [32], the fixed point continuation algorithms [45], the basis pursuit algorithms [46] and other greedy matching pursuit algorithms [47]. Then the PS broadcasts the estimated $\hat{\mathbf{g}}_t$ to all the local workers for updating the shared model parameter as follows

$$\mathbf{w}_{i,t+1} = \mathbf{w}_{i,t} - \alpha\hat{\mathbf{g}}_t, \quad i = 1, 2, ..., U. \tag{14}$$

Comparing (14) and (5), aggregation errors may be introduced in 1-bit CS based FL over the air, due to analog aggregation transmissions, top-κ sparsification, CS compression, and 1-bit quantization.

III. The Convergence Analysis

In this section, we study the effect of analog aggregation transmissions and 1-bit CS on FL over the air, by analyzing its convergence behavior.

A. Basic Assumptions

To facilitate the convergence analysis, we make the following standard assumptions on the loss function and gradients.

Assumption 1 (Lipschitz continuity, smoothness): The gradient $\nabla F(\mathbf{w})$ of the loss function $F(\mathbf{w})$ is L-Lipschitz [48], that is,

$$\|\nabla F(\mathbf{w}_{t+1}) - \nabla F(\mathbf{w}_t)\| \leq L\|\mathbf{w}_{t+1} - \mathbf{w}_t\|, \tag{15}$$

where L is a non-negative Lipschitz constant for the continuously differentiable function $F(\cdot)$.

Assumption 2 (twice-continuously differentiable): The function $F(\mathbf{w})$ is twice-continuously differentiable and L-smoothness. Accordingly, the eigenvalues of the Hessian matrix of $F(\mathbf{w})$ are bounded by [48]:

$$\nabla^2 F(\mathbf{w}_t) \preceq L\mathbf{I}. \tag{16}$$

Assumption 3 (sample-wise gradient bounded): The sample-wise gradients at local workers are bounded by their global counterpart [49], [50]

$$\| \nabla f(\mathbf{w}_t) \|^2 \leq \rho_1 + \rho_2 \| \nabla F(\mathbf{w}_t) \|^2, \tag{17}$$

where $\rho_1 \geq 0$ and $0 \leq \rho_2 < 1$.

Assumption 4 (local gradient bounded): The local gradients are bounded by [51]

$$\|\mathbf{g}_{i,t}\|^2 \leq G^2, \forall i, t, \tag{18}$$

where G is positive constant.

B. Convergence Analysis

We first analyze the total error between the recovered averaged gradient in (14) and the ideal one in (5), including the errors caused by sparsification, quantization, AWGN and reconstruction algorithms. Based on the above **Assumption 4**, we derive the following **Lemma 1** to delineate the total error.

Lemma 1. *The total error $\mathbf{e}_t = \hat{\mathbf{g}}_t - \mathbf{g}_t$ at the t-th iteration in FL is bounded by*

$$\mathbb{E}\|\mathbf{e}_t\|^2 = \mathbb{E}(\|\hat{\mathbf{g}}_t - \mathbf{g}_t\|^2) \leq 2\sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{D}G^2 + 2C^2\left(1 + (1+\delta)\frac{D-\kappa}{SD}G^2 + \frac{\sigma^2}{\left(\sum_{i=1}^{U}K_i\beta_{i,t}b_t\right)^2}\right), \tag{19}$$

where $0 < \delta < 1$ is the constant in the RIP condition, $C = \frac{2\varpi}{1-\varrho}$, $\varpi = \frac{2\sqrt{1+\delta}}{\sqrt{1-\delta}}$ and $\varrho = \frac{\sqrt{2}\delta}{1-\delta}$.

Proof. The proof of **Lemma 1** is provide in Appendix A. □

Remark 2. **Lemma 1** indicates that a larger κ leads to a smaller error, which suggests that sparsification is applied at the expense of accuracy. And a larger S leads to a smaller error because of less compression.

Next we present the main theorem for the expected convergence rate of the 1-bit CS based FL over the air with analog aggregation, as in **Theorem 1**.

Theorem 1. *Given the power scaling factor b_t, worker selection vectors $\beta_{i,t}$, and the learning rate $\alpha = \frac{1}{L}$, we have the following convergence rate at the T-th iteration.*

$$\frac{1}{T}\sum_{t=1}^{T}\| \nabla F(\mathbf{w}_{t-1}) \|^2 \leq \frac{2L}{T(1-2\rho_2(U+K))}\mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)] + \frac{2L}{T(1-2\rho_2(U+K))}\sum_{t=1}^{T}B_t, \tag{20}$$

where

$$B_t = \frac{\rho_1(U+K)\sum_{i=1}^{U}K_i(1-\beta_{i,t})}{LK} + 2\sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{LD}G^2 + \frac{2C^2}{L}\left(1 + (1+\delta)\frac{D-\kappa}{SD}G^2 + \frac{\sigma^2}{\left(\sum_{i=1}^{U}K_i\beta_{i,t}b_t\right)^2}\right), \tag{21}$$

*and $\mathbf{w}^*$ is a feasible solution to the problem **P1**.*

Proof. The proof of **Theorem 1** is provide in Appendix B. □

In **Theorem 1**, the expected gradient norm is used as an indicator of convergence [52]–[54]. That is, the FL algorithm achieves a τ-suboptimal solution if:

$$\frac{1}{T}\sum_{t=1}^{T}\| \nabla F(\mathbf{w}_{t-1}) \|^2 \leq \tau, \tag{22}$$

which guarantees the convergence of the algorithm to a stationary point. If the objective function $F(\mathbf{w})$ is non-convex, then FL may converge to a local minimum or a saddle point.

From **Theorem 1**, we have

$$\frac{1}{T}\sum_{t=1}^{T}\| \nabla F(\mathbf{w}_{t-1}) \|^2 \leq \frac{2L}{T(1-2\rho_2(U+K))}\mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)] + \frac{2L}{T(1-2\rho_2(U+K))}\sum_{t=1}^{T}B_t \overset{T\to\infty}{\longrightarrow} \frac{2L}{T(1-2\rho_2(U+K))}\sum_{t=1}^{T}B_t. \tag{23}$$

The error floor at convergence is given by (23). Obviously, minimizing this error floor can improve the convergence performance of FL. Capitalizing on this theoretical result, we provide joint optimization of communication and computation next.

Remark 3. As we can see from **Theorem 1**, the maximization of the number of participating workers is not necessarily the optimal solution to achieving the best convergence performance. This is because, the more the number of participating workers, the larger the learning errors. Thus, it is necessary to select the participating workers properly, instead of simply maximizing the number of participating workers.

Remark 4. If we do not consider sparsification and the signals transmitted by local workers can be perfectly recovered, then $\kappa = D$ holds, reducing the second term of the right hand of (21) to be 0. When more participating workers selected, more terms of the worker selection vector become $\beta_{i,t} = 1$ for selected worker i, which results in a smaller B_t in (21) and a lower error floor in (23). Thus, it is desired to maximize the number of participating workers to mitigate the noise effect. However, the selection of participating workers needs to satisfy their individual power constraints. Thus, a proper power scaling factor b_t is needed.

Remark 5. From (21) and (23), we can see that a larger b_t leads to more reduction of the negative impact of AWGN on learning performance. However, The larger b_t, the less number of local workers can participate FL. Thus, we need to jointly optimize $\beta_{i,t}$ and b_t.

Remark 6. Obviously, the larger the sparsity ratio (i.e., the larger κ), the better the convergence performance. Also, the error floor decreases as the compressed dimension size S increases. However, the efficiency of communication decreases as κ and S increase. Thus, we need to balance learning performance and communication efficiency, and then set the values of these two variables κ and S for the desired tradeoff in practical demands.

Remark 7. In the non-i.i.d. case, the factor ρ_1 in **Assumption 3** would become larger. As shown in (23), a larger ρ_1 leads to a larger error floor, resulting in worse learning performance.

IV. Minimization of the Error Floor for Federated Learning Algorithm

In this section, we formulate a joint optimization problem to minimize the error floor in (23) for 1-bit CS based FL over the air. In solving such a problem, we first develop an optimal solution via discrete programming, and then propose a computationally scalable ADMM-based suboptimal solution for large-scale wireless networks.

A. Joint Optimization Problem Formulation

In the deployment of FL over the air, the error floor in (23) is accumulated over iterations, resulting a performance gap between $F(\mathbf{w}_{t-1})$ and $F(\mathbf{w}^*)$. Thus, we design an online policy to minimize this gap at each iteration, which amounts to iteratively minimizing B_t under the constraint of transmit power limitation in (11).

At each iteration t, the PS aims to determine the power scaling factor b_t and the scheduling indicator $\boldsymbol{\beta}_t = [\beta_{1,t}, \beta_{2,t}, ..., \beta_{U,t}]$ in order to minimize R_t, for given values of the factors (i.e., C, S, and κ) related to 1-bit CS. Discarding irrelevant terms, minimizing B_t is equivalent to minimizing

$$R_t = \frac{\rho_1(U+K)\sum_{i=1}^{U} K_i(1-\beta_{i,t})}{K} + 2C^2\Big(\sum_{i=1}^{U} K_i\beta_{i,t}b_t\Big)^{-2}\sigma^2 + 2\sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{D}G^2. \tag{24}$$

The joint optimization problem is thus formulated as

$$\textbf{P2:} \quad \min_{b_t, \boldsymbol{\beta}_t} \quad R_t \tag{25a}$$

$$\text{s.t.} \quad \frac{\beta_{i,t}^2 K_i^2 b_t^2}{h_{i,t}^2} \le P_i^{\text{Max}}, \ \beta_{i,t} \in \{0,1\}, i = 1, 2, ..., U. \tag{25b}$$

To implement the optimization algorithm for solving the above problem at the PS, the number of local samples and the maximum power at each worker should be sent to the PS by the local workers before the process of FL. These numbers are fixed and never change in the iterative process.

B. Optimal Solution via Discrete Programming

As a mixed integer programming (MIP), **P2** is non-convex and challenging to solve due to the coupling of the real-valued power scaling factor b_t and the binary-valued scheduling indicator $\boldsymbol{\beta}_t$. Note that once $\boldsymbol{\beta}_t$ is given, the problem **P2** reduces to a convex problem, where the optimal power scaling b_t can be efficiently solved using off-the-shelf optimization algorithms, e.g., interior point method [55]. Accordingly, a straightforward method is to enumerate all the 2^U possibilities of $\boldsymbol{\beta}_t$ and output the one that yields the lowest objective value. This enumeration-based method is summarized in **Algorithm 1**.

Algorithm 1 Optimal solution via the enumeration-based method

Initialization:
$\{P_i^{\text{Max}}, h_{i,t}, K_i\}_{i=1}^U$, $\mathbf{\Phi}$, G, κ.
Ensure:
The optimal solution $\{b_t^*, \boldsymbol{\beta}_t^*\}$.
1: **Repeat**
2: Select $\boldsymbol{\beta}_t$ from its possibility;
3: Given $\boldsymbol{\beta}_t$, solve **P2** to find $\{b_t\}$;
4: If the objective value is lower under this $\{b_t, \boldsymbol{\beta}_t\}$, then update $\{b_t^*, \boldsymbol{\beta}_t^*\}$;
5: **Until** {all the possible of $\boldsymbol{\beta}_t$ are enumerated}
6: **return** $\{b_t^*, \boldsymbol{\beta}_t^*\}$.

Remark 8. The enumeration-based method may be applicable for a small number of workers, e.g., $U \le 10$; however, it quickly becomes computationally infeasible as U increases.

C. ADMM-based Suboptimal Solution

The enumeration-based method proposed in the last subsection is simple to implement, because the computation involves basic function evaluations only. However, large-scale networks with a large network size U makes it susceptible to high computational complexity. To address the problem, we propose an ADMM-based algorithm to jointly optimize the local worker selection and power control. As we will show later, the proposed ADMM-based approach has a computational complexity that scales linearly in U.

The main idea is to decompose the combinatorial optimization **P2** into U parallel smaller integer programming problems. Nonetheless, conventional decomposition techniques, such as dual decomposition, cannot be directly applied to **P2** due to the coupled variables $\{b_t, \boldsymbol{\beta}_t\}$ and the constraint (25b) among the workers. To eliminate these coupling factors, we first introduce an auxiliary vector $\mathbf{r}_t = [r_{1,t}, r_{2,t}, ..., r_{U,t}]$ and define two auxiliary functions as

$$Q_1(\mathbf{r}_t) = 2C^2(\sum_{i=1}^{U} K_i r_{i,t})^{-2}\sigma^2, \tag{26}$$

and

$$Q_2(\boldsymbol{\beta}_t) = \frac{\rho_1(U+K)\sum_{i=1}^{U} K_i(1-\beta_{i,t})}{K} + 2\sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{D}G^2. \tag{27}$$

Then we introduce another auxiliary vector $\mathbf{q}_t = [q_{1,t}, q_{2,t}, ..., q_{U,t}]$ and reformulate **P2** as the following **P3**.

$$\textbf{P3:} \min_{b_t, \{r_{i,t}, q_{i,t}, \beta_{i,t}\}_{i=1}^{U}} \quad Q_1(\mathbf{r}_t) + Q_2(\boldsymbol{\beta}_t) + \sum_{i=1}^{U} Q_{3,i}(r_{i,t}) \tag{28a}$$

$$\text{s.t.} \quad r_{i,t} = \beta_{i,t} q_{i,t}, \ i = 1, 2, ..., U, \tag{28b}$$

$$q_{i,t} = b_t, \ i = 1, 2, ..., U, \tag{28c}$$

$$r_{i,t} > 0, b_t > 0, \ i = 1, 2, ..., U, \tag{28d}$$

$$\beta_{i,t} \in \{0, 1\}, \ i = 1, 2, ..., U, \tag{28e}$$

where

$$Q_{3,i}(r_{i,t}) = \begin{cases} 0, & r_{i,t} \in \left\{ r_{i,t} \middle| \left| \frac{K_i r_{i,t}}{h_{i,t}} \right|^2 \le P_i^{\text{Max}} \right\}, \\ \infty, & \texttt{otherwise}. \end{cases} \tag{29}$$

Here, the constraints (28b) and (28c) are introduced to decouple $\beta_{i,t}$ and b_t while guaranteeing that **P3** and **P2** are equivalent.

By introducing multipliers, $\xi_{i,t} \ge 0$'s and $\varsigma_{i,t} \ge 0$'s to the constraints in (28b) and (28c), we can write a partial augmented Lagrangian of **P3** as

$$\begin{aligned} &\mathcal{L}(b_t, \boldsymbol{\beta}_t, \mathbf{r}_t, \mathbf{q}_t, \boldsymbol{\xi}_t, \boldsymbol{\varsigma}_t) \\ =&Q_1(\mathbf{r}_t) + Q_2(\boldsymbol{\beta}_t) + \sum_{i=1}^{U} Q_{3,i}(r_{i,t}) \\ &+ \sum_{i=1}^{U} \xi_{i,t}(r_{i,t} - \beta_{i,t} q_{i,t}) + \frac{c}{2}\sum_{i=1}^{U}(r_{i,t} - \beta_{i,t} q_{i,t})^2 \\ &+ \sum_{i=1}^{U} \varsigma_{i,t}(q_{i,t} - b_t) + \frac{c}{2}\sum_{i=1}^{U}(q_{i,t} - b_t)^2, \end{aligned} \tag{30}$$

where $\boldsymbol{\xi}_t = [\xi_{1,t}, \xi_{2,t}, ..., \xi_{U,t}]$, $\boldsymbol{\varsigma}_t = [\varsigma_{1,t}, \varsigma_{2,t}, ..., \varsigma_{U,t}]$, and $c > 0$ is a fixed step size. The corresponding dual problem is

$$\textbf{P4:} \max_{\{\xi_{i,t}, \varsigma_{i,t}\}_{i=1}^{U}} \quad \mathcal{M}(\boldsymbol{\xi}_t, \boldsymbol{\varsigma}_t) \tag{31a}$$

$$\text{s.t.} \quad \xi_{i,t} \ge 0, \ \varsigma_{i,t} \ge 0, \quad i = 1, 2, ..., U, \tag{31b}$$

where $\mathcal{M}(\boldsymbol{\xi}_t, \boldsymbol{\varsigma}_t)$ is the dual function, which is given by

$$\mathcal{M}(\boldsymbol{\xi}_t, \boldsymbol{\varsigma}_t) = \min_{b_t, \{r_{i,t}, q_{i,t}, \beta_{i,t}\}_{i=1}^{U}} \quad \mathcal{L}(b_t, \mathbf{r}_t, \mathbf{q}_t, \boldsymbol{\beta}_t) \tag{32a}$$

$$\text{s.t.} \quad r_{i,t} > 0, b_t > 0, q_{i,t} > 0, \beta_{i,t} \in \{0, 1\}, i = 1, 2, ..., U. \tag{32b}$$

The ADMM technique [56] solves the dual problem **P4** by iteratively updating $\{\mathbf{r}_t, b_t\}$, $\{\mathbf{q}_t, \boldsymbol{\beta}_t\}$, and $\{\boldsymbol{\xi}_t, \boldsymbol{\varsigma}_t\}$. We denote the values at the l-th iteration as $\{\mathbf{r}_t^{\{l\}}, b_t^{\{l\}}\}$, $\{\mathbf{q}_t^{\{l\}}, \boldsymbol{\beta}_t^{\{l\}}\}$, and $\{\boldsymbol{\xi}_t^{\{l\}}, \boldsymbol{\varsigma}_t^{\{l\}}\}$. Then, the variables are sequentially updated at the $(l+1)$-th iteration as follows:

1) Step 1: Given $\{\mathbf{q}_t^{\{l\}}, \boldsymbol{\beta}_t^{\{l\}}\}$, and $\{\boldsymbol{\xi}_t^{\{l\}}, \boldsymbol{\varsigma}_t^{\{l\}}\}$, we first minimize $\mathcal{L}$ with respect to $\{\mathbf{r}_t, b_t\}$, where

$$\{\mathbf{r}_t^{\{l+1\}}, b_t^{\{l+1\}}\} = \arg\min_{\mathbf{r}_t, b_t} \mathcal{L}(\mathbf{r}_t, b_t). \tag{33}$$

Notice that (33) is a convex problem, which can be solved to obtain the optimal solution, e.g., by using the projected Newton's method [55]. On the other hand, this optimization problem in (33) can also be decomposed into $U+2$ parallel convex subproblems. Specifically, let $Q_1(r_t) = 2C^2(r_t)^{-2}\sigma^2$ in place of Q_1 in (26), and $r_t = \sum_{i=1}^{U} K_i r_{i,t}$ be an extra constraint. Then (33) can be solved by solving $U+2$ parallel subproblems (due to $U+2$ optimization variables, i.e., $r_t, r_{1,t}, ..., r_{U,t}, b_t$). Since the complexity of solving these $U+2$ subproblems does not scale with U, thus the overall computational complexity of **Step 1** is $\mathcal{O}(U)$.

2) Step 2: Given $\{\mathbf{r}_t^{\{l+1\}}, b_t^{\{l+1\}}\}$, and $\{\boldsymbol{\xi}_t^{\{l\}}, \boldsymbol{\varsigma}_t^{\{l\}}\}$, we then minimize $\mathcal{L}$ with respect to $\{\mathbf{q}_t, \boldsymbol{\beta}_t\}$, where

$$\{\mathbf{q}_t^{\{l+1\}}, \boldsymbol{\beta}_t^{\{l+1\}}\} = \arg\min_{\mathbf{q}_t, \boldsymbol{\beta}_t} \mathcal{L}(\mathbf{q}_t, \boldsymbol{\beta}_t). \tag{34}$$

This optimization can be decomposed into U parallel subproblems. In each subproblem (e.g., the i-th subproblem), by considering $\beta_{i,t} = 0$ and $\beta_{i,t} = 1$, respectively, the i-th subproblem is expressed as

$$\{q_{i,t}\}^{\{l+1\}} = \begin{cases} \arg\min_{q_{i,t}} \mathcal{L}(q_{i,t}, 0), & \beta_{i,t} = 0, \\ \arg\min_{q_{i,t}} \mathcal{L}(q_{i,t}, 1), & \beta_{i,t} = 1, \end{cases} \tag{35}$$

where

$$\mathcal{L}\left(q_{i,t},0\right)=\frac{\rho_1(U+K)K_i}{K}+\{\xi_{i,t}\}^{\{l\}}\{r_{i,t}\}^{\{l+1\}}+\frac{c}{2}\left(\{r_{i,t}\}^{\{l+1\}}\right)^2+\varsigma_{i,t}\left(q_{i,t}-\{b_t\}^{\{l+1\}}\right)+\frac{c}{2}\left(q_{i,t}-\{b_t\}^{\{l+1\}}\right)^2, \tag{36}$$

and

$$\mathcal{L}\left(q_{i,t},1\right)=(1+\delta)\frac{D-\kappa}{D}G^2+\frac{c}{2}\left(q_{i,t}-\{b_t\}^{\{l+1\}}\right)^2+\{\xi_{i,t}\}^{\{l\}}\left(\{r_{i,t}\}^{\{l+1\}}-q_{i,t}\right)+\varsigma_{i,t}\left(q_{i,t}-\{b_t\}^{\{l+1\}}\right)+\frac{c}{2}\left(\{r_{i,t}\}^{\{l+1\}}-q_{i,t}\right)^2. \tag{37}$$

For both $\beta_{i,t}=0$ and $\beta_{i,t}=1$, (35) solves a strictly convex problem, and hence is easy to obtain the optimal solution. Accordingly, we can simply select between $\beta_{i,t}=0$ or $\beta_{i,t}=1$ that yields a smaller objective value in (35) as $\{\beta_{i,t}\}^{\{l+1\}}$, and the corresponding optimal solution of $\{q_{i,t}\}^{\{l+1\}}$. After solving the U parallel subproblems, the optimal solution to (34) is given by $\{\mathbf{q}_t^{\{l+1\}},\boldsymbol{\beta}_t^{\{l+1\}}\}$. Notice that the complexity of solving each subproblem in (34) scales with U, and thus the overall computational complexity of **Step 2** is $\mathcal{O}(U)$.

3) Step 3: Finally, given $\{\mathbf{r}_t^{\{l+1\}},b_t^{\{l+1\}}\}$ and $\{\mathbf{q}_t^{\{l+1\}},\boldsymbol{\beta}_t^{\{l+1\}}\}$, we maximize $\mathcal{L}$ with respect to $\{\boldsymbol{\xi}_t,\boldsymbol{\varsigma}_t\}$, which is achieved by updating the multipliers as follows

$$\{\xi_{i,t}\}^{\{l+1\}}=\{\xi_{i,t}\}^{\{l\}}+c\Big(\{r_{i,t}\}^{\{l+1\}}-\{\beta_{i,t}\}^{\{l+1\}}\{q_{i,t}\}^{\{l+1\}}\Big),\quad i=1,...,U, \tag{38}$$

$$\{\varsigma_{i,t}\}^{\{l+1\}}=\{\varsigma_{i,t}\}^{\{l\}}+c\left(\{q_{i,t}\}^{\{l+1\}}-\{b_t\}^{\{l+1\}}\right),\quad i=1,...,U. \tag{39}$$

Obviously, the computational complexity of **Step 3** is $\mathcal{O}(U)$ as well.

The ADMM method implements the above **Steps 1** to **3** iteratively until meeting a specified stopping criterion. In general, the stopping criterion is specified by two thresholds [56]: an absolute tolerance (e.g., $\sum_{i=1}^{U}|\{q_{i,t}\}^{\{l+1\}}-\{b_t\}^{\{l+1\}}|$) and a relative tolerance (e.g., $|\{b_t\}^{\{l+1\}}-\{b_t\}^{\{l\}}|$). The pseudo-code of the ADMM based method solving (**P3**) is summarized in **Algorithm 2**.

Remark 9. The proposed **Algorithm 2** is guaranteed to converge, because the dual problem **P4** is convex. Its convergence is insensitive to the step size c [57]. Due to the potential duality gap of non-convex problems, **Algorithm 2** may not exactly converge to the primal optimal solution to **P3**, Thus, the dual optimal solution $\{b_t^*,\boldsymbol{\beta}_t^*\}$ is an approximate solution to **P3**.

Remark 10. In this work, the complexity is evaluated with response to the number of workers U, that is, how it scales when U increases asymptotically. In this sense, the computational complexity of one ADMM iteration (including the 3 steps) is $\mathcal{O}(U)$, because the highest complexity of these three steps is $\mathcal{O}(U)$. This complexity $\mathcal{O}(U)$ scales linearly in U rather than an exponential complexity $\mathcal{O}(2^U)$ in the enumeration-based method.

Algorithm 2 ADMM-based suboptimal solution

Initialization:
$\{P_i^{\text{Max}},h_{i,t},K_i\}_{i=1}^U$, $\mathbf{\Phi}$, G, κ.
Ensure:
The optimal solution $\{b_t^*,\boldsymbol{\beta}_t^*\}$.
1: **Repeat**
2: Update $\{\mathbf{r}_t^{\{l+1\}},b_t^{\{l+1\}}\}$ by solving (33);
3: Update $\{\mathbf{q}_t^{\{l+1\}},\boldsymbol{\beta}_t^{\{l+1\}}\}$ by solving (34);
4: Update $\{\boldsymbol{\xi}_t^{\{l+1\}},\boldsymbol{\varsigma}_t^{\{l+1\}}\}$ by using (38), and(39);
5: **Until** {the convergence threshold is satisfied or the maximum number of iterations is reached}.
6: **return** $\{b_t^*,\boldsymbol{\beta}_t^*\}$.

V. The SGD Case

In this section, we focus on the convergence behavior and algorithm design principle of our work in the SGD case, i.e., using the mini-batch SGD algorithm to update learning models. Here, we provide the expected convergence rate of the 1-bit CS based FL over the air with analog aggregation for the mini-batch gradient descent with a constant mini-batch size K_b, while the results directly apply to the standard SGD by setting $K_b=1$. **Theorem 2** summarizes the convergence behavior of the 1-bit CS based FL over the air with analog aggregation for the SGD case.

Theorem 2. *Given the power scaling factor b_t, worker selection vectors $\beta_{i,t}$, and the learning rate $\alpha=\frac{1}{L}$, we have the following convergence rate for the SGD case at the T-th iteration.*

$$\frac{1}{T}\sum_{t=1}^{T}\parallel\nabla F(\mathbf{w}_{t-1})\parallel^2\leq\frac{2LK^2}{T(K^2-2\rho_2(\vartheta+2UK^2))}\mathbb{E}[F(\mathbf{w}_0)-F(\mathbf{w}^*)]+\frac{2LK^2}{T(K^2-2\rho_2(\vartheta+2UK^2))}\sum_{t=1}^{T}B_t^{sgd}, \tag{40}$$

where

$$B_t^{sgd}=\frac{\rho_1\left(\vartheta+2UK^2\left(\sum_{i=1}^{U}\beta_{i,t}\right)^{-1}\right)}{LK^2}+\frac{2}{L}\left(\frac{C^2}{S}\varepsilon_t+\sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{D}G^2\right), \tag{41}$$

and $\vartheta=2U^2K_b^2+K-4KUK_b-UK_b$.

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

Per **Theorem 2**, we minimize B_t^{sgd} to reduce the performance gap at each iteration, which is equivalent to minimizing

$$R_t^{\mathtt{sgd}} = U\rho_1 \left(\sum_{i=1}^{U} \beta_{i,t}\right)^{-1} + C^2 (\sum_{i=1}^{U} K_b \beta_{i,t} b_t)^{-2} \sigma^2 + \sum_{i=1}^{U} \beta_{i,t} \frac{D-\kappa}{D} G^2. \quad (42)$$

For given values of the factors (i.e., C, S, and κ) related to 1-bit CS, the joint optimization problem for the SGD case is thus formulated as

$$\textbf{P5:} \quad \min_{b_t, \boldsymbol{\beta}_t} \quad R_t^{\mathtt{sgd}} \quad (43a)$$

$$\text{s.t.} \quad \frac{\beta_{i,t}^2 K_b^2 b_t^2}{h_{i,t}^2} \leq P_i^{\text{Max}}, \ \beta_{i,t} \in \{0,1\}, i = 1, 2, ..., U. \quad (43b)$$

To solve the above optimization problem **P5**, we can also apply the enumeration-based method and the ADMM-based method as described in Section IV. For the ADMM-based method, the defined two auxiliary functions are formulated as

$$Q_1^{\mathtt{sgd}}(\mathbf{r}_t) = U\rho_1 \left(\sum_{i=1}^{U} r_{i,t}\right)^{-1} + C^2 (\sum_{i=1}^{U} K_i r_{i,t})^{-2} \sigma^2, \quad (44)$$

and

$$Q_2^{\mathtt{sgd}}(\boldsymbol{\beta}_t) = \sum_{i=1}^{U} \beta_{i,t} \frac{D-\kappa}{D} G^2. \quad (45)$$

Then, the rest of the solving procedure can be developed in a way similar to the GD case as in Section IV.C, which we thus omit here.

Remark 11. From $R_t^{\mathtt{sgd}}$ in (42), we can see that the larger K_b, the lower $R_t^{\mathtt{sgd}}$, i.e., the better learning performance.

VI. Simulation Results And Evaluation

In the simulations, we evaluate the performance of the proposed 1-bit CS based FL over the air for an image classification task. The simulation settings are given as follows unless specified otherwise. We consider that the FL system has U workers, which is varying from 4 to 20 in our simulations, and set their maximum peak power to be $P_i^{\text{Max}} = 10 + i$ mW for any $i \in [1, U]$. The wireless channels between the workers and the PS are modeled as i.i.d. Rayleigh fading, by generating $h_{i,t}$'s from an normal distribution $\mathcal{N}(0,1)$ for different i and t. Without loss of the generality, the variance of AWGN at PS is set to be $\sigma^2 = 10^{-4}$ mW. Unless otherwise specified, we perform top $\kappa = 1000$ sparsification, and the dimension of compressed local $\mathcal{C}(\mathbf{g}_i)$'s is set to $S = 10000$ for 10 workers. The elements of the measurement matrix $\mathbf{\Phi}$ are generated from $\mathcal{N}(0, 1/S)$. The BIHT algorithm in [32] is selected as an example for the signal reconstruction at the PS.

We consider the learning task of handwritten-digit recognition using the well-known MNIST dataset[7] 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 are available for training a learning model. 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 cross entropy as the loss function, and rectified linear unit (ReLU) as the activation function. The total number of parameters in the MLP is $D = 50890$. The learning rate α is set as 0.1. 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 $i \in [1, U]$.

For comparison, we use a benchmark where the transmission of local gradient updates is always reliable and error-free to achieve perfect aggregation, i.e., overlooking the influence of the wireless channel. This benchmark is an ideal case, which is named as *perfect aggregation*. Also, we compare our proposed scheme with the existing analog aggregation based work (named as *OBDA*) in [19] that adopts the idea of signSGD [58]. For performance evaluation, we provide the results of training loss and test accuracy versus communication rounds under different parameter settings as follows.

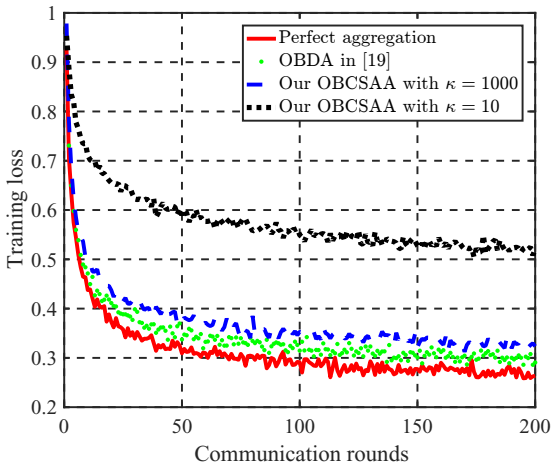
In Fig. 1, we first explore the impact of different sparsification operators on our proposed OBCSAA by evaluating the training loss and test accuracy of the MLP. It is observed that our proposed OBCSAA can provide desired performance (which approaches to that of *OBDA* and *perfect aggregation*), with a degree of sparsification, e.g., $\kappa = 1000$, where the sparsity ratio is $1000/50890$. As κ increases, when all FL algorithms converge, the training loss decreases and the test accuracy increases. This is because that the larger κ is, the less gradient update information loses per communication round.

Fig. 2 shows the impact of the reduced dimension size S on the performance of our proposed OBCSAA under $\kappa = 1000$, where the performance increases as S increases. When S is large enough, performance barely increases. This is because that the larger S is, the more conducive to signal reconstruction. When S is large enough, the optimal performance of the reconstruction algorithm is achieved. In fact, the larger S is, the more communication resources are needed. Thus, there is a tradeoff between FL performance and communication efficiency. Compared with the traditional uncompressed FL adopting digital communications, our proposed OBCSAA under $S = 5000$ and $\kappa = 1000$ occupies only one channel and $\frac{5000}{50890}$ transmission time, while the performance is less than 5 and 10 percent lower than that of *OBDA* and *perfect aggregation*. These results illustrates that our OBCSAA under appropriate parameters can greatly reduce the communication overhead and transmission latency while ensuring comparable FL performance to the idealized and orthogonal transmission cases.

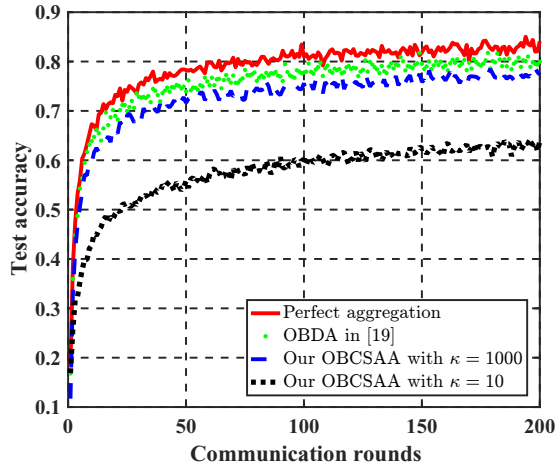
The performance of the proposed enumeration-based method and ADMM for OBCSAA under different U are compared in Fig. 3, where the enumeration-based method has better performance compared to ADMM. This results precisely demonstrate the effectiveness of our joint optimization scheme, which can alleviate the impact of aggregation errors on FL. Besides, we can see that the performance is higher, when the

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

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2022.3209190

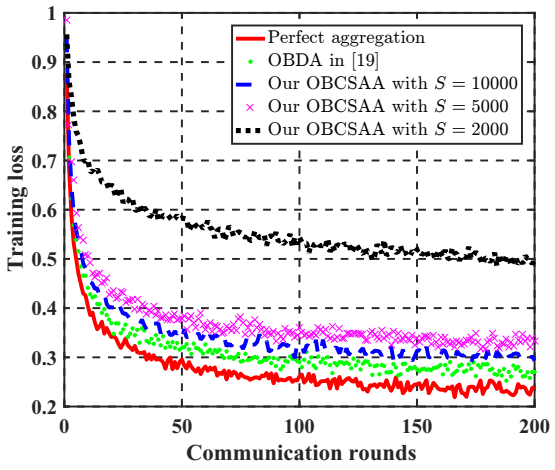


(a) Training loss

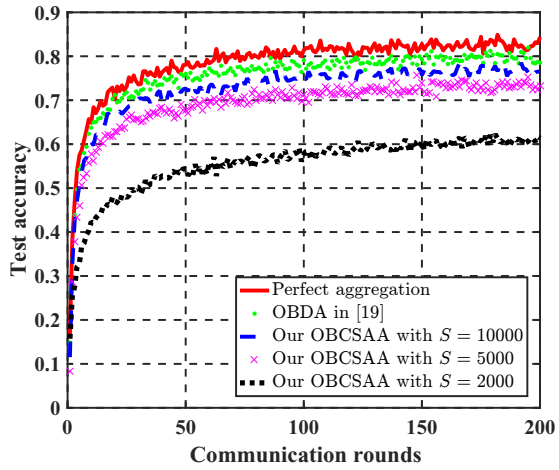


(b) Test accuracy

Fig. 1: The performance of our proposed OBCSAA under different sparsification operators compared to perfect aggregation without sparsification.



(a) Training loss



(b) Test accuracy

Fig. 2: The performance of our proposed OBCSAA under different S.

total number of local workers U is larger. This is because an increase in the number of workers leads to an increased volume of data available for the FL algorithm and more workers with high channel gain can be selected.

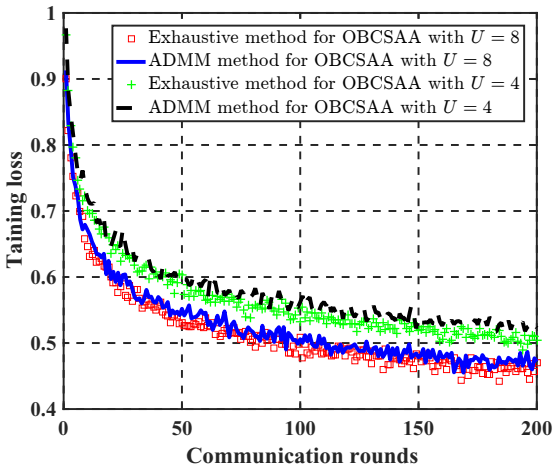
In Fig. 4, we evaluate the test accuracy of our proposed OBCSAA compared with benchmarks for the SGD case with varying total numbers of local workers, where we fix the mini-batch size $K_b = 64$. As we can see from Fig. 4, the performances of all the schemes are increasing as the total number of local workers increases. While the total number of local workers continues to increase, the performance improvement for all methods shrinks and flattens out eventually. This is because the data samples turn to be sufficient for accurate training when U exceeds a certain level.

In Fig. 5, we further compare the communication efficiency together with the learning accuracy achieved by our proposed OBCSAA with that of an existing analog compression scheme [23], called CA-DSGD, which reflects the tradeoff between the communication cost and the learning accuracy. The communication cost is denoted as $\frac{S}{D}$, where the larger S indicates the higher communication cost. According to Fig. 5, to obtain a desired learning accuracy, the communication cost consumed by our proposed OBCSAA is less than that of CA-DSGD. In other words, given the same communication cost, our proposed OBCSAA can achieve a higher learning accuracy than CA-DSGD. Thanks to the joint optimization of power control and worker selection, our OBCSAA achieves a better tradeoff between learning and communication than CA-DSGD where only a fixed power allocation is adopted without worker selection.

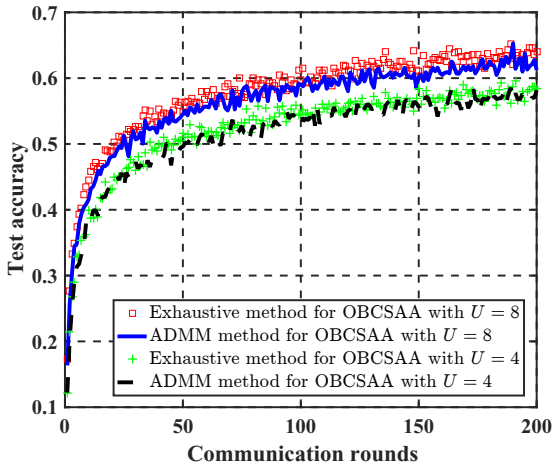
VII. Conclusion

This paper studies a communication-efficient FL based on 1-bit CS and analog aggregation transmissions. A closed-form expression is derived for the expected convergence rate of the FL algorithm. This theoretical result reveals the tradeoff between convergence performance and communication efficiency as a result of the aggregation errors caused by sparsification, dimension reduction, quantization, signal reconstruction and noise. Guided by this revelation, a joint optimization problem of communication and learning is developed to mitigate aggregation errors, which results in an optimal worker selection and power control. An enumeration-based method and an ADMM

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2022.3209190



(a) Training loss



(b) Test accuracy

Fig. 3: The performance of joint optimization solving methods for our proposed OBCSAA under different U.

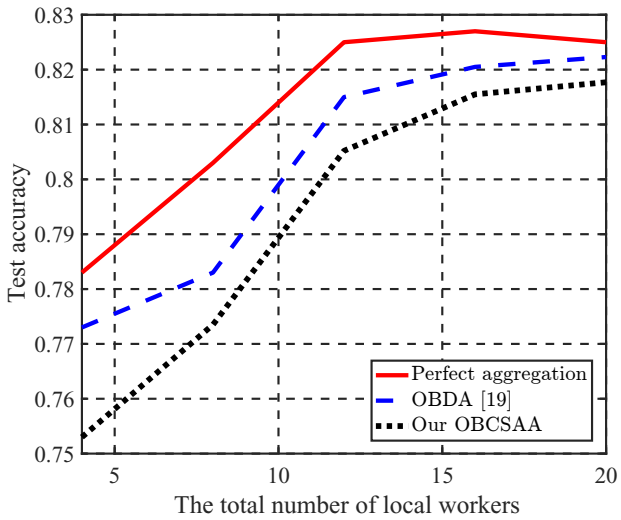


Fig. 4: The performance comparison of different methods under various total numbers of local workers in the SGD case with $\kappa = 1000$ and $S = 10000$.

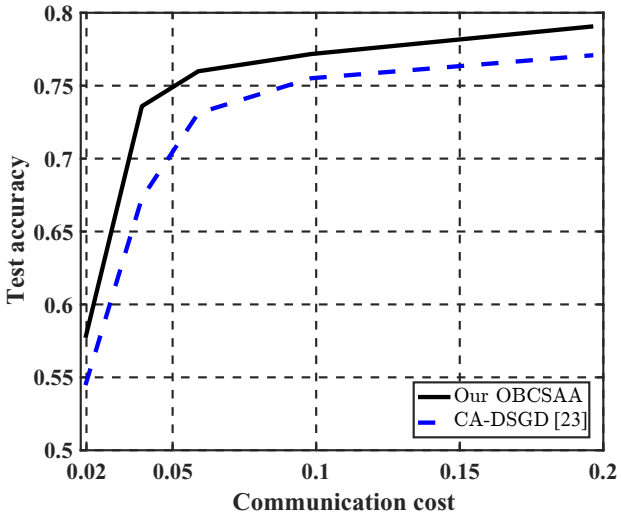


Fig. 5: The tradeoff between learning accuracy and communication cost.

method are proposed to solve this challenging non-convex problem, which can obtain the optimal solution for small-scale networks and sub-optimal solution for large-scale networks, respectively. Simulation results show that our proposed FL can greatly improve communication efficiency while ensuring desired learning performance.

APPENDIX A
PROOF OF LEMMA 1

Proof. Under the **Assumption 4**, the sparsification error $\mathbf{e}_{i,t}^s \in \mathbb{R}^D, \forall i,t$ satisfies

$$\mathbb{E}\|\mathbf{e}_{i,t}^s\|^2 = \mathbb{E}\|\tilde{\mathbf{g}}_{i,t} - \mathbf{g}_{i,t}\|^2 \leq \frac{D-\kappa}{D} G^2. \tag{46}$$

Since $\mathbf{\Phi}$ satisfies the RIP condition [59],

$$(1-\delta)\|\mathbf{x}\|^2 \leq \|\mathbf{\Phi}\mathbf{x}\|^2 \leq (1+\delta)\|\mathbf{x}\|^2, \tag{47}$$

where $\mathbf{x}$ is a k-sparse vector, then the quantization error $\mathbf{e}_{i,t}^q \in \mathbb{R}^S$ is derived as

$$\begin{aligned}\mathbb{E}\|\mathbf{e}_{i,t}^q\|^2 &= \mathbb{E}\|\mathtt{sign}(\mathbf{\Phi}\tilde{\mathbf{g}}_{i,t}) - \mathbf{\Phi}\tilde{\mathbf{g}}_{i,t}\|^2 \\ &\leq \mathbb{E}(\|\mathtt{sign}(\mathbf{\Phi}\tilde{\mathbf{g}}_{i,t})\|^2 + \|\mathbf{\Phi}\tilde{\mathbf{g}}_{i,t}\|^2) \\ &\leq S + (1+\delta)\frac{D-\kappa}{D} G^2. \end{aligned} \tag{48}$$

When the PS obtains $\hat{\mathbf{y}}_t^{desired}$ in (13), it reconstructs the signal $\hat{\mathbf{g}}_t$, in the presence of norm-limited measurement error $\mathbf{e}_t^r$. It has been shown that robust reconstruction can be achieved by solving [31]:

$$\hat{\mathbf{g}}_t = \arg\min_{\tilde{\mathbf{g}}_t} \|\tilde{\mathbf{g}}_t\|_1 \quad \mathtt{s.t.}\ \|\hat{\mathbf{y}}_t^{desired} - \mathbf{\Phi}\tilde{\mathbf{g}}_t\|^2 \leq \varepsilon_t. \tag{49}$$

where ε_t is a small value that bounds the norm of the reconstruction error, and is typically set empirically [31]. We opt to set ε_t in a more principled way by the expected norm, as follows:

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{y}}_t^{desired}-\mathbf{\Phi}\tilde{\mathbf{g}}_t\|^2&=\mathbb{E}\left\|\hat{\mathbf{y}}_t^{desired}-\frac{\sum_{i=1}^U K_i\beta_{i,t}(\mathbf{\Phi}\tilde{\mathbf{g}}_{i,t})}{\sum_{i=1}^U K_i\beta_{i,t}}\right\|^2\\&=\mathbb{E}\left\|\frac{\sum_{i=1}^U K_i\beta_{i,t}\mathbf{e}_{i,t}^q}{\sum_{i=1}^U K_i\beta_{i,t}}+\frac{\mathbf{z}_t}{\sum_{i=1}^U K_i\beta_{i,t}b_t}\right\|^2\\&=\mathbb{E}\left\|\mathbf{e}_{1,t}^q+\frac{\mathbf{z}_t}{\sum_{i=1}^U K_i\beta_{i,t}b_t}\right\|^2\\&\le\mathbb{E}\|\mathbf{e}_{1,t}^q\|^2+\mathbb{E}\left\|\frac{\mathbf{z}_t}{\sum_{i=1}^U K_i\beta_{i,t}b_t}\right\|^2\\&\le S+(1+\delta)\frac{D-\kappa}{D}G^2+\frac{S\sigma^2}{\left(\sum_{i=1}^U K_i\beta_{i,t}b_t\right)^2}\\&\doteq\varepsilon_t.\end{aligned}\tag{50}$$

In this case, the reconstruction error norm is bounded by

$$\|\hat{\mathbf{g}}_t-\tilde{\mathbf{g}}_t\|^2\le\frac{C^2}{S}\varepsilon_t,\tag{51}$$

where C is the constant depending on the properties of the measurement matrix $\mathbf{\Phi}$ but not on the signal [60]. According to the **Theorem 1.2** in [59], if $\mathbf{\Phi}$ has $\delta\le\sqrt{2}-1$, C can be given by

$$C=\frac{2\varpi}{1-\varrho},\tag{52}$$

where $\varpi=\frac{2\sqrt{1+\delta}}{\sqrt{1-\delta}}$ and $\varrho=\frac{\sqrt{2}\delta}{1-\delta}$.

It is noted that $\tilde{\mathbf{g}}_t$ in (51) is the desired sparse global gradient after the worker selection. As a result, the total error at the t-th iteration in FL is given by

$$\begin{aligned}&\mathbb{E}\|\mathbf{e}_t\|^2=\mathbb{E}(\|\hat{\mathbf{g}}_t-\mathbf{g}_t\|^2)=\mathbb{E}(\|\hat{\mathbf{g}}_t-(\tilde{\mathbf{g}}_t+\mathbf{e}_t^s)\|^2)\\&\le\mathbb{E}(\|\hat{\mathbf{g}}_t-\tilde{\mathbf{g}}_t\|+\|\mathbf{e}_t^s\|)^2\le\mathbb{E}(2\|\hat{\mathbf{g}}_t-\tilde{\mathbf{g}}_t\|^2+2\|\mathbf{e}_t^s\|^2)\\&\le 2\frac{C^2}{S}\varepsilon_t+2\sum_{i=1}^U\beta_{i,t}\frac{D-\kappa}{D}G^2\\&=2C^2\left(1+(1+\delta)\frac{D-\kappa}{SD}G^2+\frac{\sigma^2}{\left(\sum_{i=1}^U K_i\beta_{i,t}b_t\right)^2}\right)\\&+2\sum_{i=1}^U\beta_{i,t}\frac{D-\kappa}{D}G^2,\end{aligned}\tag{53}$$

where $\mathbf{e}_t^s=\sum_{i=1}^U\beta_{i,t}\mathbf{e}_{i,t}^s$. □

APPENDIX B
PROOF OF THEOREM 1

Proof. To prove **Theorem 1**, we first rewrite $F(\mathbf{w}_t)$ as the expression of its second-order Taylor expansion, which is given by

$$\begin{aligned}&F(\mathbf{w}_t)=F(\mathbf{w}_{t-1})+(\mathbf{w}_t-\mathbf{w}_{t-1})^T\nabla F(\mathbf{w}_{t-1})\\&+\frac{1}{2}(\mathbf{w}_t-\mathbf{w}_{t-1})^T\nabla^2F(\mathbf{w}_{t-1})(\mathbf{w}_t-\mathbf{w}_{t-1})\\&\overset{(a)}{\le}F(\mathbf{w}_{t-1})+(\mathbf{w}_t-\mathbf{w}_{t-1})^T\nabla F(\mathbf{w}_{t-1})+\frac{L}{2}\parallel\mathbf{w}_t-\mathbf{w}_{t-1}\parallel^2,\end{aligned}\tag{54}$$

where **Assumption 2** is applied in the step (a).

After recovering the desired $\hat{\mathbf{g}}_t$ from the received signal by solving (49), then the common model is updated by

$$\begin{aligned}\mathbf{w}_t&=\mathbf{w}_{t-1}-\alpha\hat{\mathbf{g}}_t\\&=\mathbf{w}_{t-1}-\alpha(\nabla F(\mathbf{w}_{t-1})-\mathbf{o}),\end{aligned}\tag{55}$$

where

$$\mathbf{o}=\nabla F(\mathbf{w}_{t-1})-\hat{\mathbf{g}}_t.\tag{56}$$

Given the learning rate $\alpha=\frac{1}{L}$ (a special setting for simpler expression without losing the generality), then the expected optimization function of $\mathbb{E}[F(\mathbf{w}_t)]$ from (54) can be expressed as

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}_t)]\le&\mathbb{E}\Big[F(\mathbf{w}_{t-1})-\alpha(\nabla F(\mathbf{w}_{t-1})-\mathbf{o})^T\nabla F(\mathbf{w}_{t-1})\\&+\frac{L\alpha^2}{2}\parallel\nabla F(\mathbf{w}_{t-1})-\mathbf{o}\parallel^2\Big]\\\overset{(b)}{=}&\mathbb{E}[F(\mathbf{w}_{t-1})]-\frac{1}{2L}\parallel\nabla F(\mathbf{w}_{t-1})\parallel^2+\frac{1}{2L}\mathbb{E}[\parallel\mathbf{o}\parallel^2],\end{aligned}\tag{57}$$

where the step (b) is derived from the fact that

$$\begin{aligned}\frac{L\alpha^2}{2}\parallel\nabla F(\mathbf{w}_{t-1})-\mathbf{o}\parallel^2=&\frac{1}{2L}\parallel\nabla F(\mathbf{w}_{t-1})\parallel^2\\&-\frac{1}{L}\mathbf{o}^T\nabla F(\mathbf{w}_{t-1})+\frac{1}{2L}\parallel\mathbf{o}\parallel^2.\end{aligned}\tag{58}$$

According to (53), it has $\|\mathbf{e}_t\|^2\le\frac{2C^2}{S}\varepsilon_t+2\sum_{i=1}^U\beta_{i,t}\frac{D-\kappa}{D}G^2$. Then we derive $\mathbb{E}[\parallel\mathbf{o}\parallel^2]$ as follows

$$\begin{aligned}&\mathbb{E}[\parallel\mathbf{o}\parallel^2]=\mathbb{E}[\|\nabla F(\mathbf{w}_{t-1})-\hat{\mathbf{g}}_t\|]\\&=\mathbb{E}[\|\nabla F(\mathbf{w}_{t-1})-\mathbf{g}_t-\mathbf{e}_t\|]\\&=\mathbb{E}\left[\left\|\frac{\sum_{i=1}^U\sum_{k=1}^{K_i}\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})}{K}\right.\right.\\&\left.\left.-(\sum_{i=1}^U K_i\beta_{i,t})^{-1}\sum_{i=1}^U\sum_{k=1}^{K_i}\beta_{i,t}\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})-\mathbf{e}_t\right\|^2\right]\\&\le\mathbb{E}\left[\left\|\sum_{i=1}^U(\frac{1}{K}-\frac{\beta_{i,t}}{\sum_{i=1}^U K_i\beta_{i,t}})\sum_{k=1}^{K_i}\nabla f(\mathbf{w}_{t-1},\mathbf{x}_{i,k},\mathbf{y}_{i,k})-\mathbf{e}_t\right\|^2\right].\end{aligned}\tag{59}$$

Applying the triangle inequality of norms: $\|\mathbf{X}+\mathbf{Y}\|\le\|\mathbf{X}\|+\|\mathbf{Y}\|$, the submultiplicative property of norms: $\|\mathbf{XY}\|\le\|\mathbf{X}\|\|\mathbf{Y}\|$, and the Jensen's inequality: $(\sum_{i=1}^n a_i)^2\le$

$n\sum_{i=1}^{n} a_i^2$, we further derive (59) as follows

$$\begin{aligned}
&\mathbb{E}[\| \mathbf{o} \|^2] \leq \mathbb{E}[2\|\mathbf{e}_{t-1}\|^2] \\
&+ \mathbb{E}\left[2\left\| \sum_{i=1}^{U}(\frac{1}{K} - \frac{\beta_{i,t}}{\sum_{i=1}^{U} K_i\beta_{i,t}}) \sum_{k=1}^{K_i} \nabla f(\mathbf{w}_{t-1}, \mathbf{x}_{i,k}, \mathbf{y}_{i,k})\right\|^2\right] \\
&\leq 2\mathbb{E}[\|\mathbf{e}_{t-1}\|^2] + \mathbb{E}\Bigg[2(U+K)\sum_{i=1}^{U}(\frac{1}{K} \\
&- \frac{\beta_{i,t}}{\sum_{i=1}^{U} K_i\beta_{i,t}})^2 \sum_{k=1}^{K_i} \|\nabla f(\mathbf{w}_{t-1}, \mathbf{x}_{i,k}, \mathbf{y}_{i,k})\|^2\Bigg] \\
&\leq \frac{4C^2}{S}\varepsilon_t + 4\sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{D}G^2 + 2(U+K)\sum_{i=1}^{U}(\frac{1}{K} \\
&- \frac{\beta_{i,t}}{\sum_{i=1}^{U} K_i\beta_{i,t}})^2 \sum_{k=1}^{K_i} \|\nabla f(\mathbf{w}_{t-1}, \mathbf{x}_{i,k}, \mathbf{y}_{i,k})\|^2. \qquad (60)
\end{aligned}$$

Applying (17) in **Assumption 3** to (60), we further derive the following result as

$$\begin{aligned}
&\mathbb{E}[\| \mathbf{o} \|^2] \leq 2(U+K)\sum_{i=1}^{U}(\frac{1}{K} - \frac{\beta_{i,t}}{\sum_{i=1}^{U} K_i\beta_{i,t}})^2 K_i(\rho_1 \\
&+ \rho_2\|\nabla F(\mathbf{w}_{t-1})\|^2) + \frac{4C^2}{S}\varepsilon_t + 4\sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{D}G^2 \\
&= 2(U+K)(\frac{1}{\sum_{i=1}^{U} K_i\beta_{i,t}} - \frac{1}{K})(\rho_1 + \rho_2\|\nabla F(\mathbf{w}_{t-1})\|^2) \\
&+ \frac{4C^2}{S}\varepsilon_t + \sum_{i=1}^{U} 4\beta_{i,t}\frac{D-\kappa}{D}G^2 \\
&\leq \frac{2(U+K)\sum_{i=1}^{U} K_i(1-\beta_{i,t})}{K}(\rho_1 + \rho_2\|\nabla F(\mathbf{w}_{t-1})\|^2) \\
&+ \frac{4C^2}{S}\varepsilon_t + 4\sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{D}G^2. \qquad (61)
\end{aligned}$$

Substituting (61) to (57), we have:

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t)] \leq \mathbb{E}[F(\mathbf{w}_{t-1})] - \frac{1}{2L} + \\
&\frac{2(U+K)\sum_{i=1}^{U} K_i(1-\beta_{i,t})}{2LK}(\rho_1 + \rho_2\|\nabla F(\mathbf{w}_{t-1})\|^2) \\
&+\frac{4C^2\varepsilon_t}{2LS} + 4\sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{2LD}G^2 \parallel \nabla F(\mathbf{w}_{t-1}) \parallel^2 \\
&=\mathbb{E}[F(\mathbf{w}_{t-1})] + \frac{\rho_1(U+K)\sum_{i=1}^{U} K_i(1-\beta_{i,t})}{LK} \\
&+\left(\frac{\rho_2(U+K)\sum_{i=1}^{U} K_i(1-\beta_{i,t})}{LK} - \frac{1}{2L}\right) \parallel \nabla F(\mathbf{w}_{t-1}) \parallel^2 \\
&+\frac{2}{L}(\frac{C^2}{S}\varepsilon_t + \sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{D}G^2). \qquad (62)
\end{aligned}$$

Summing up the above inequalities from $t=1$ to $t=T$, we get

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_0)] \leq -\sum_{t=1}^{T} A_t \parallel \nabla F(\mathbf{w}_{t-1}) \parallel^2 + \sum_{t=1}^{T} B_t, \qquad (63)$$

where

$$A_t = \frac{1}{2L} - \frac{\rho_2(U+K)\sum_{i=1}^{U} K_i(1-\beta_{i,t})}{LK}, \qquad (64)$$

$$\begin{aligned}
B_t = &\frac{\rho_1(U+K)\sum_{i=1}^{U} K_i(1-\beta_{i,t})}{LK} \\
&+ \frac{2}{L}\left(\frac{C^2}{S}\varepsilon_t + \sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{D}G^2\right). \qquad (65)
\end{aligned}$$

The inequality (63) can be also written as

$$\begin{aligned}
\sum_{t=1}^{T} A_t \parallel \nabla F(\mathbf{w}_{t-1}) \parallel^2 &\leq \mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}_t)] + \sum_{t=1}^{T} B_t \\
&\leq \mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)] + \sum_{t=1}^{T} B_t. \qquad (66)
\end{aligned}$$

Since $\frac{1}{2L} - \frac{\rho_2(U+K)}{L} \leq A_t \leq \frac{1}{2L}$, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}(\frac{1}{2L} - \frac{\rho_2(U+K)}{L}) \parallel \nabla F(\mathbf{w}_{t-1}) \parallel^2 \\
&\leq \frac{1}{T}\sum_{t=1}^{T} A_t \parallel \nabla F(\mathbf{w}_{t-1}) \parallel^2 \\
&\leq \frac{1}{T}\mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)] + \frac{1}{T}\sum_{t=1}^{T} B_t. \qquad (67)
\end{aligned}$$

As a result, we get

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \parallel \nabla F(\mathbf{w}_{t-1}) \parallel^2 &\leq \frac{2L}{T(1-2\rho_2(U+K))}\sum_{t=1}^{T} B_t \\
&+ \frac{2L}{T(1-2\rho_2(U+K))}\mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)]. \qquad (68)
\end{aligned}$$

The proof is completed. □

APPENDIX C
PROOF OF THEOREM 2

Proof. For the SGD method, the local gradient of the i-th worker is calculated at the t-th iteration as

$$\mathbf{g}_{i,t} = \mathbb{E}_{\mathfrak{D}_i}\left[\frac{\sum_{k=1}^{K_b} \nabla f(\mathbf{w}_{t-1}, \mathbf{x}_{i,k}, \mathbf{y}_{i,k})}{K_b}\right], \quad i = 1, 2, ..., U, \qquad (69)$$

where $\mathbb{E}_{\mathfrak{D}_i}[\cdot]$ is the expectation, which represents that the i-th worker randomly chooses K_b samples from its local dataset $\mathfrak{D}_i$ to compute the local gradient.

Given (69), the desired signal vector at the PS at the t-th iteration in (9) for the GD case is now expressed as

$$\mathbf{y}_t^{desired} = \frac{\sum_{i=1}^{U} K_b\beta_{i,t}\mathcal{C}(\mathbf{g}_{i,t})}{\sum_{i=1}^{U} K_b\beta_{i,t}} = \frac{\sum_{i=1}^{U}\beta_{i,t}\mathcal{C}(\mathbf{g}_{i,t})}{\sum_{i=1}^{U}\beta_{i,t}}. \qquad (70)$$

Accordingly, the estimation of the signal vector of interest via a post-processing operation in (13) is written as

$$\begin{aligned}\hat{\mathbf{y}}_t^{desired} &= (\sum_{i=1}^{U} K_b\beta_{i,t}b_t)^{-1}\mathbf{y}_t \\ &=(\sum_{i=1}^{U} K_b\beta_{i,t})^{-1}\sum_{i=1}^{U} K_b\beta_{i,t}\mathcal{C}(\mathbf{g}_{i,t}) + (\sum_{i=1}^{U} K_b\beta_{i,t}b_t)^{-1}\mathbf{z}_t \\ &=\mathbf{y}_t^{desired} + \frac{\mathbf{z}_t}{\sum_{i=1}^{U} K_b\beta_{i,t}b_t},\end{aligned} \tag{71}$$

where $(\sum_{i=1}^{U} K_b\beta_{i,t}b_t)^{-1}$ works as the post-processing factor.

Then, we get the norm boundary $\varepsilon_t^{\mathrm{sgd}}$ defined in (49) as

$$\begin{aligned}\varepsilon_t^{\mathrm{sgd}} &\doteq \mathbb{E}\|\hat{\mathbf{y}}_t^{desired} - \mathbf{\Phi}\tilde{\mathbf{g}}_t\|^2 \\ &=\mathbb{E}\left\|\hat{\mathbf{y}}_t^{desired} - \frac{\sum_{i=1}^{U} K_b\beta_{i,t}(\mathbf{\Phi}\tilde{\mathbf{g}}_{i,t})}{\sum_{i=1}^{U} K_b\beta_{i,t}}\right\|^2 \\ &=\mathbb{E}\left\|\frac{\sum_{i=1}^{U}\beta_{i,t}\mathbf{e}_{i,t}^q}{\sum_{i=1}^{U}\beta_{i,t}} + \frac{\mathbf{z}_t}{\sum_{i=1}^{U} K_b\beta_{i,t}b_t}\right\|^2 \\ &=\mathbb{E}\left\|\mathbf{e}_{1,t}^q + \frac{\mathbf{z}_t}{\sum_{i=1}^{U} K_b\beta_{i,t}b_t}\right\|^2 \\ &\le\mathbb{E}\|\mathbf{e}_{1,t}^q\|^2 + \mathbb{E}\left\|\frac{\mathbf{z}_t}{\sum_{i=1}^{U} K_b\beta_{i,t}b_t}\right\|^2 \\ &\le S + (1+\delta)\frac{D-\kappa}{D}G^2 + \frac{S\sigma^2}{\left(\sum_{i=1}^{U} K_b\beta_{i,t}b_t\right)^2}.\end{aligned} \tag{72}\tag{73}$$

Thus, the total error at the t-th iteration in FL is bounded by

$$\begin{aligned}&\mathbb{E}\|\mathbf{e}_t\|^2 = \mathbb{E}(\|\hat{\mathbf{g}}_t - \mathbf{g}_t\|^2) = \mathbb{E}(\|\hat{\mathbf{g}}_t - (\tilde{\mathbf{g}}_t + \mathbf{e}_t^s)\|^2) \\ &\le \mathbb{E}(2\|\hat{\mathbf{g}}_t - \tilde{\mathbf{g}}_t\|^2 + 2\|\mathbf{e}_t^s\|^2) \\ &\le \frac{2C^2}{S}\varepsilon_t^{\mathrm{sgd}} + 2\sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{D}G^2 \\ &= 2C^2\left(1 + (1+\delta)\frac{D-\kappa}{SD}G^2 + \frac{\sigma^2}{\left(\sum_{i=1}^{U} K_b\beta_{i,t}b_t\right)^2}\right) \\ &+ 2\sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{D}G^2.\end{aligned} \tag{74}$$

Let $\mathcal{N}_{i,t}$ denote the set of the samples that are not chosen by the i-th worker at the t-th iteration, $\mathbb{E}[\|\mathbf{o}\|^2]$ can be derived as

$$\begin{aligned}&\mathbb{E}[\|\ \mathbf{o}\ \|^2] = \mathbb{E}[\|\nabla F(\mathbf{w}_{t-1}) - \hat{\mathbf{g}}_t\|] \\ &=\mathbb{E}[\|\nabla F(\mathbf{w}_{t-1}) - \mathbf{g}_t - \mathbf{e}_t\|] \\ &=\mathbb{E}\Bigg[\Bigg\|\frac{\sum_{i=1}^{U}\sum_{k=1}^{K_i}\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})}{K} \\ &\quad - (\sum_{i=1}^{U} K_b\beta_{i,t})^{-1}\sum_{i=1}^{U}\beta_{i,t}\mathbb{E}_{\mathfrak{D}_i}\left[\sum_{k=1}^{K_b}\nabla f(\mathbf{w}_{t-1},\mathbf{x}_{i,k},\mathbf{y}_{i,k})\right] - \mathbf{e}_t\Bigg\|^2\Bigg] \\ &\le\mathbb{E}\Bigg[\Bigg\|\sum_{i=1}^{U}(\frac{1}{K} - \frac{\beta_{i,t}}{\sum_{i=1}^{U} K_i\beta_{i,t}})\mathbb{E}_{\overline{\mathcal{N}}_{i,t}}\left[\sum_{k\in\overline{\mathcal{N}}_{i,t}}\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})\right] \\ &\quad + \frac{\sum_{i=1}^{U}\mathbb{E}[\sum_{k\in\mathcal{N}_{i,t}}\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})]}{K} - \mathbf{e}_t\Bigg\|^2\Bigg] \\ &\quad\le \mathbb{E}\Bigg[2\Bigg\|\sum_{i=1}^{U}(\frac{1}{K} - \frac{\beta_{i,t}}{\sum_{i=1}^{U} K_i\beta_{i,t}})\mathbb{E}_{\overline{\mathcal{N}}_{i,t}}\left[\sum_{k\in\overline{\mathcal{N}}_{i,t}}\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})\right] \\ &\quad + \frac{\sum_{i=1}^{U}\mathbb{E}[\sum_{k\in\mathcal{N}_{i,t}}\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})]}{K}\Bigg\|^2\Bigg] + 2\|\mathbf{e}_t\|^2, \\ &\le\left(4\sum_{i=1}^{U} K_b\right)\sum_{i=1}^{U}\Bigg(\frac{1}{K} \\ &\quad - \beta_{i,t}\left(\sum_{i=1}^{U} K_b\beta_{i,t}\right)^{-1}\Bigg)^2\mathbb{E}_{\overline{\mathcal{N}}_{i,t}}\left[\sum_{k\in\overline{\mathcal{N}}_{i,t}}\|\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})\|^2\right] \\ &\quad + 2\frac{\|\sum_{i=1}^{U}\mathbb{E}[\sum_{k\in\mathcal{N}_{i,t}}\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})]\|^2}{K^2} + 2\|\mathbf{e}_t\|^2.\end{aligned} \tag{75}$$

Applying (17) in **Assumption 3** to (75), we further derive the following result as

$$\begin{aligned}&\mathbb{E}[\|\ \mathbf{o}\ \|^2] \le 4UK_b\sum_{i=1}^{U}\left(\frac{1}{K} - \beta_{i,t}\left(\sum_{i=1}^{U} K_b\beta_{i,t}\right)^{-1}\right)^2 K_b(\rho_1 \\ &+ \rho_2\|\nabla F(\mathbf{w}_{t-1})\|^2) \\ &+ 2\frac{\sum_{i=1}^{U}(K_i - K_b)(\rho_1 + \rho_2\|\nabla F(\mathbf{w}_{t-1})\|^2)}{K^2} + 2\|\mathbf{e}_t\|^2 \\ &\le 4UK_b\sum_{i=1}^{U}\left(\frac{K_b}{K^2} - 2K_b\frac{1}{K}\beta_{i,t}\left(\sum_{i=1}^{U} K_b\beta_{i,t}\right)^{-1}\right. \\ &+ K_b\beta_{i,t}\left.\left(\sum_{i=1}^{U} K_b\beta_{i,t}\right)^{-2}\right)(\rho_1 + \rho_2\|\nabla F(\mathbf{w}_{t-1})\|^2) \\ &+ 2\frac{\sum_{i=1}^{U}(K_i - K_b)(\rho_1 + \rho_2\|\nabla F(\mathbf{w}_{t-1})\|^2)}{K^2} + 2\|\mathbf{e}_t\|^2 \\ &\le 4UK_b\left(\frac{UK_b}{K^2} - 2\frac{1}{K} + \left(\sum_{i=1}^{U} K_b\beta_{i,t}\right)^{-1}\right)(\rho_1 \\ &+ \rho_2\|\nabla F(\mathbf{w}_{t-1})\|^2) + 2\frac{(K - UK_b)(\rho_1 + \rho_2\|\nabla F(\mathbf{w}_{t-1})\|^2)}{K^2} \\ &+ \frac{4C^2}{S}\varepsilon_t^{\mathrm{sgd}} + 4\sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{D}G^2.\end{aligned} \tag{76}$$

Substituting (76) to (57), we have:

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t)] \leq \frac{\vartheta + 2UK^2\left(\sum_{i=1}^{U}\beta_{i,t}\right)^{-1}}{LK^2}(\rho_1+\rho_2\|\nabla F(\mathbf{w}_{t-1})\|^2)\\
&+\frac{4C^2\varepsilon_t^{\mathrm{sgd}}}{2LS}+4\sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{2LD}G^2\\
&\quad+\mathbb{E}[F(\mathbf{w}_{t-1})]-\frac{1}{2L}\parallel\nabla F(\mathbf{w}_{t-1})\parallel^2\\
&=\mathbb{E}[F(\mathbf{w}_{t-1})]+\frac{\rho_1\left(\vartheta+2UK^2\left(\sum_{i=1}^{U}\beta_{i,t}\right)^{-1}\right)}{LK^2}\\
&+\left(\frac{\rho_2\left(\vartheta+2UK^2\left(\sum_{i=1}^{U}\beta_{i,t}\right)^{-1}\right)}{LK^2}-\frac{1}{2L}\right)\parallel\nabla F(\mathbf{w}_{t-1})\parallel^2\\
&+\frac{2}{L}(\frac{C^2}{S}\varepsilon_t+\sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{D}G^2). \qquad (77)
\end{aligned}$$

where $\vartheta = 2U^2K_b^2 + K - 4KUK_b - UK_b$.

Summing up the above inequalities from $t=1$ to $t=T$, we get

$$\mathbb{E}[F(\mathbf{w}_t)-F(\mathbf{w}_0)] \leq -\sum_{t=1}^{T}A_t^{\mathrm{sgd}}\parallel\nabla F(\mathbf{w}_{t-1})\parallel^2+\sum_{t=1}^{T}B_t^{\mathrm{sgd}}, \qquad (78)$$

where

$$A_t^{\mathrm{sgd}} = \frac{1}{2L}-\frac{\rho_2\left(\vartheta+2UK^2\left(\sum_{i=1}^{U}\beta_{i,t}\right)^{-1}\right)}{LK^2}, \qquad (79)$$

$$\begin{aligned}
B_t^{\mathrm{sgd}} &= \frac{\rho_1\left(\vartheta+2UK^2\left(\sum_{i=1}^{U}\beta_{i,t}\right)^{-1}\right)}{LK^2}\\
&\quad+\frac{2}{L}\left(\frac{C^2}{S}\varepsilon_t+\sum_{i=1}^{U}\beta_{i,t}\frac{D-\kappa}{D}G^2\right). \qquad (80)
\end{aligned}$$

The inequality (78) can be rewritten as

$$\begin{aligned}
\sum_{t=1}^{T}A_t^{\mathrm{sgd}}\parallel\nabla F(\mathbf{w}_{t-1})\parallel^2 &\leq \mathbb{E}[F(\mathbf{w}_0)-F(\mathbf{w}_t)]+\sum_{t=1}^{T}B_t^{\mathrm{sgd}}\\
&\leq \mathbb{E}[F(\mathbf{w}_0)-F(\mathbf{w}^*)]+\sum_{t=1}^{T}B_t^{\mathrm{sgd}}. \qquad (81)
\end{aligned}$$

Since $\frac{1}{2L}-\frac{\rho_2(\vartheta+2UK^2)}{LK^2} \leq A_t^{\mathrm{sgd}} \leq \frac{1}{2L}$, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2L}-\frac{\rho_2(\vartheta+2UK^2)}{LK^2}\right)\parallel\nabla F(\mathbf{w}_{t-1})\parallel^2\\
&\leq\frac{1}{T}\sum_{t=1}^{T}A_t^{\mathrm{sgd}}\parallel\nabla F(\mathbf{w}_{t-1})\parallel^2\\
&\leq\frac{1}{T}\mathbb{E}[F(\mathbf{w}_0)-F(\mathbf{w}^*)]+\frac{1}{T}\sum_{t=1}^{T}B_t^{\mathrm{sgd}}. \qquad (82)
\end{aligned}$$

As a result, we get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\parallel\nabla F(\mathbf{w}_{t-1})\parallel^2\\
&\leq\frac{2LK^2}{T(K^2-2\rho_2(\vartheta+2UK^2))}\mathbb{E}[F(\mathbf{w}_0)-F(\mathbf{w}^*)]\\
&+\frac{2LK^2}{T(K^2-2\rho_2(\vartheta+2UK^2))}\sum_{t=1}^{T}B_t^{\mathrm{sgd}}. \qquad (83)
\end{aligned}$$

The proof is completed. □

REFERENCES

[1] 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.

[2] Z. He, Z. Cai, S. Cheng, and X. Wang, "Approximate aggregation for tracking quantiles and range countings in wireless sensor networks," *Theoretical Computer Science*, vol. 607, pp. 381–390, 2015.

[3] J. Li, S. Cheng, Z. Cai, J. Yu, C. Wang, and Y. Li, "Approximate holistic aggregation in wireless sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 13, no. 2, pp. 1–24, 2017.

[4] 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.

[5] 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.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] 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.

[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. 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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*, vol. 20, no. 3, pp. 2120–2135, 2020.

[20] 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, pp. 1–6.

[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] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Joint optimization of communications and federated learning over the air," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 4434–4449, 2022.

[23] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3546–3557, 2020.

[24] ——, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2155–2169, 2020.

[25] 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.

[26] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Best effort voting power control for byzantine-resilient federated learning over the air," in *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2022, pp. 1–6.

[27] 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.

[28] 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," *IEEE Internet of Things Journal*, pp. 1–1, 2022.

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

[30] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning," *IEEE Transactions on Wireless Communications*, vol. 20, no. 8, pp. 5115–5128, 2021.

[31] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *2008 42nd Annual Conference on Information Sciences and Systems*. IEEE, 2008, pp. 16–21.

[32] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2082–2102, 2013.

[33] D.-Q. Dai, L. Shen, Y. Xu, and N. Zhang, "Noisy 1-bit compressive sensing: models and algorithms," *Applied and Computational Harmonic Analysis*, vol. 40, no. 1, pp. 1–32, 2016.

[34] C. Li, G. Li, and P. K. Varshney, "Communication-efficient federated learning based on compressed sensing," *IEEE Internet of Things Journal*, vol. 8, no. 20, pp. 15 531–15 541, 2021.

[35] A. Abdi and F. Fekri, "Quantized compressive sampling of stochastic gradients for efficient communication in distributed deep learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3105–3112.

[36] P. Jiang and G. Agrawal, "A linear speedup analysis of distributed deep learning with sparse and quantized communication," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[37] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "Uveqfed: Universal vector quantization for federated learning," *IEEE Transactions on Signal Processing*, vol. 69, pp. 500–514, 2020.

[38] Y. Li, Y. Cui, and V. Lau, "Optimization-based genqsgd for federated edge learning," *arXiv preprint arXiv:2110.12987*, 2021.

[39] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2021–2031.

[40] J. Wangni, J. Wang, J. Liu, and T. Zhang, "Gradient sparsification for communication-efficient distributed optimization," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[41] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Sparse binary compression: Towards distributed deep learning with minimal communication," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[42] 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.

[43] N. Ström, "Sparse connection and pruning in large dynamic artificial neural networks," in *Fifth European Conference on Speech Communication and Technology*. Citeseer, 1997.

[44] ——, "Scalable distributed dnn training using commodity gpu cloud computing," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[45] E. T. Hale, W. Yin, and Y. Zhang, "A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing," *CAAM TR07-07, Rice University*, vol. 43, p. 44, 2007.

[46] A. Moshtaghpour, L. Jacques, V. Cambareri, K. Degraux, and C. De Vleeschouwer, "Consistent basis pursuit for signal and matrix estimates in quantized compressed sensing," *IEEE signal processing letters*, vol. 23, no. 1, pp. 25–29, 2015.

[47] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[48] S. Bubeck, "Convex optimization: Algorithms and complexity," *Foundations and Trends in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[49] D. P. Bertsekas, J. N. Tsitsiklis, and J. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[50] M. P. Friedlander and M. Schmidt, "Hybrid deterministic-stochastic methods for data fitting," *SIAM Journal on Scientific Computing*, vol. 34, no. 3, pp. A1380–A1405, 2012.

[51] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified sgd with memory," in *Advances in Neural Information Processing Systems*, 2018, pp. 4447–4458.

[52] 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.

[53] X. Lian, Y. Huang, Y. Li, and J. Liu, "Asynchronous parallel stochastic gradient for nonconvex optimization," *Advances in Neural Information Processing Systems*, vol. 28, pp. 2737–2745, 2015.

[54] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Transactions on signal processing*, vol. 66, no. 11, pp. 2834–2848, 2018.

[55] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[56] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

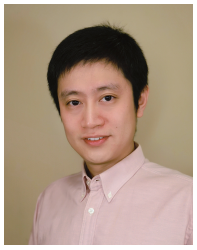
[57] E. Ghadimi, A. Teixeira, I. Shames, and M. Johansson, "Optimal parameter selection for the alternating direction method of multipliers (admm): quadratic problems," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 644–658, 2014.

[58] 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.

[59] E. J. Candes *et al.*, "The restricted isometry property and its implications for compressed sensing," *Comptes rendus mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008.

[60] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 8, pp. 1207–1223, 2006.

Xin Fan (S'22) 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 was a visiting Ph.D. student in the Electrical and Computer Engineering Department of George Mason University, Fairfax, VA, USA, from 2020 to 2022. He is currently a Ph.D. student in Beijing Jiaotong University from 2018. His current research interests lie in the areas of wireless communications, machine learning, security and privacy, optimization, statistical signal processing, and blockchain.



Yue Wang (M'11, SM'22) received his Ph.D. degree in communication and information system from Beijing University of Posts and Telecommunications, China. Currently, he is a Research Assistant Professor with the Electrical and Computer Engineering Department of George Mason University, Fairfax, VA, USA. Previously, he was a Senior Research Engineer with Huawei Technologies Co., Ltd., China. His general interests lie in the areas of signal processing, wireless communications, artificial intelligence, and their applications in cyber physical systems. His current research focuses on compressed sensing, massive MIMO, millimeter-wave communications, cognitive radios, Internet of Things, DoA estimation, high-dimensional data analysis, and distributed optimization and learning.

Yan Huo (M'12, SM'19) 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 focuses on wireless communications, security and privacy, and the Internet of Things. It involves building and simulating prototype systems and conducting real experiments and measurements. He has served as an Associate Editor for the IEEE ACCESS and a Reviewer for a number of journals, including the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the IEEE WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON MOBILE COMPUTING, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS.

Zhi Tian (M'00, SM'06, F'13) 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 was the Chair of the IEEE Signal Processing Society Big Data Special Interest Group, and a Member-of-Large of the IEEE Signal Processing Society (2019-2021). She received the IEEE Communications Society TCCN Publication Award in 2018.