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

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

[1]School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing, China

[2]Department of Electrical & Computer Engineering, George Mason University, Fairfax, VA, USA

E-mails: {yhuo,fanxin}@bjtu.edu.cn, {ywang56,ztian1}@gmu.edu

Abstract

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 theoretically analyze the efficacy of the proposed scheme by deriving a closed-form expression for the expected convergence rate of the FL over the air. Our theoretical results reveal 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 1-bit CS based FL over the air as 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 becoming increasingly costly for communications. Although the adoption of approximate data aggregation instead of exact data aggregation proposed by [1], [2] 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 [3]–[5]. In each round of iteration-based FL, starting from a common learning model received from the PS, local devices (workers) proceed to train the model with their own local data by updating their local model parameters, and then transmit their local updates to the PS. Next, all the collected local updates are averaged at the PS and then sent back to local workers for the next round local updates. 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, updates shared between local workers and the PS can be extremely large, e.g., the VGGNet architecture has approximately 138 million parameters. As a result, the pre-processing of updates has been considered in the literature to reduce the communication load per worker, such as *sparsification*, *quantization* and *communication censoring* schemes. Given the compressible nature of gradients where only a small percentage of entries have large values while the rest remain at relatively small values [6], *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 still 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]–[24] by allowing multiple workers to simultaneously transmit their updates over the same time-frequency resources and then applying an average-enabled computation-over-the-air principle [25]. 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 prompts the prosperity of analog aggregation based FL. In [17], a broadband analog aggregation scheme was designed for FL, 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 [18], a jointly optimization 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 [19] for FL over the air in multiple antenna systems, which 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 [20]. In [21], [22], the gradient sparsification, and a random linear projection for dimensionality reduction of large-size gradient in narrow-band channels was considered to reduce the communication requirements. The power allocation scheme in [21], [22] scales the power of the vectors containing the gradient information of different devices to satisfy the average power constraint.

Despite the prior work, some fundamental questions remain unanswered, which however prevent from achieving communication-efficient and high-performance FL with analog aggregation. Firstly, the quantitative relationship between FL and analog aggregation communication is not clear. 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 [19], [24]. 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]–[20]. However, gradient statistics in FL vary over both training iterations and feature dimensions, and are unknown a priori [26]. Thus, it is infeasible to design an optimal power control without prior knowledge of the local gradients at the PS, especially for the non-coding linear analog modulation in analog aggregation based FL. Thirdly, sparsification is introduced for communication efficiency

in analog aggregation based FL [21], [22] 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 FL over the air, 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 [27]–[29] into FL over the air 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:

- We propose an **o**ne-**b**it **CS** **a**nalog **a**ggregation (OBCSAA) for efficient FL. In our OBCSAA, we elaborately design a set of preprocessing, 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 performace tradeoff as a result of the aggregation errors caused by sparsification, dimension reduction, quantization, signal reconstruction and additive white gaussian noise, 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.

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 ideal case where FL is implemented by perfect aggregation over error-free wireless channels,

with much enhanced communication efficiency.

It is worth noting that, unlike its digital communication counterpart, the optimization design for FL over the air faces a much reduced degree of freedom due to analog aggregation, and is not yet well explored in the literature [17]–[22]. Compared to [17]–[19] 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]–[20] 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]–[20]. Compared to [21], [22], our work not only applies the 1-bit quantization after dimensionality reduction, but also provides a convergence analysis on 1-bit CS based FL over the air, which leads to a joint optimization of computation and communications.In short, our work is a holistic integration of gradient sparsification, dimensionality reduction, and quantization for efficient FL over the air.

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. Numerical results are presented in Section V, and conclusions are drawn in Section VI.

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 $\mathcal{D} = \bigcup_i \mathcal{D}_i$, where $\mathcal{D}_i = \{\mathbf{x}_{i,k}, \mathbf{y}_{i,k}\}_{k=1}^{K_i}$ is the local dataset and $K_i = |\mathcal{D}_i|$ is the number of data samples at the i-th worker, $i = 1, \ldots, U$. In $\mathcal{D}_i$, 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

$F(\mathbf{w};\mathcal{D})$ 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.,

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

where $F(\mathbf{w};\mathcal{D})=\frac{1}{K}\sum_{i=1}^{U}\sum_{k=1}^{K_i}f(\mathbf{w};\mathbf{x}_{i,k},\mathbf{y}_{i,k})$ is the summation of $K=\sum_{i=1}^{U}K_i$ sample-wise loss functions defined by the learning model.

To avoid directly uploading the raw local datasets to the PS for central training, the learning procedure in (**P1**) is conducted in a distributed manner by an iterative gradient-averaging algorithm [3], [30]. Specifically, at each iteration t, the gradient descent (GD)[1] is applied at local workers in parallel to minimize the local loss functions

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

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:

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

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:

$$\text{(Global gradient computing)}\quad \mathbf{g}=\frac{1}{K}\sum_{i=1}^{U}K_i\mathbf{g}_i, \tag{4}$$

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

$$\text{(Shared model updating)}\quad \mathbf{w}=\mathbf{w}-\alpha\mathbf{g}, \tag{5}$$

where α is the learning rate.

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

[1]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. Note that SGD works more computation-efficient at the cost of more iterations and hence more transmissions compared to GD.

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 [27]–[29] 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 [6]. The other is that FL is usually running in an averaged-based distributed learning mechanism. In our work, through gradient sparsification, the compression nature of CS allows to reduce the dimension 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.

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 κ-level sparsification denoted by

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

where $\mathtt{sparse}_{\kappa}(\cdot)$ 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 a top-κ sparsification strategy, i.e., elements with the largest κ magnitudes are retained while other elements are set to 0.

2) Dimension 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, 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[2], which results in an increased sparsity-level $\bar{\kappa}$ ($> \kappa$) for

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

the superposition 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 sparisty in the combined sparse gradient, i.e., $\kappa U > \bar{\kappa}$. In addition, $\mathbf{\Phi}$ is shared between the workers and the PS before transmissions.

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

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

where $\mathcal{C}(\cdot)$ represents the overall effective operation including top-κ sparsification, CS compression, and 1-bit quantization.

4) Analog Aggregation Transmission: After the above collecting the compressive measurements in (7), all the workers 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 computing step in (4). Specifically, each local $\mathcal{C}(\mathbf{g}_{i,t})$ is multiplied with a pre-processing power control factor, denoted as $p_{i,t}$. Then, 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 additive white Gaussian noise (AWGN) vector, and $h_{i,t}$ denotes the channel coefficient between the i-th local worker and the PS at the t-th iteration[3].

Let $\beta_{i,t}$ denote the scheduling 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 pre-processing power control factor $p_{i,t}$ as

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

[3]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.

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 $c_{i,t}^s = \pm 1$ due to 1-bit quantization as the s-th element of $\mathcal{C}(\mathbf{g}_{i,t}) = [c_{i,t}^1, ..., c_{i,t}^s, ..., c_{i,t}^S]^T$. As we can see from (11), the power limitation is independent of the specific local gradient, which enables the optimization of power control to get rid of the prior knowledge on gradient or gradient statistics.

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

By such design of $p_{i,t}$, the received signal vector at the PS is 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

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

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

5) Reconstruction: After obtaining $\hat{\mathbf{y}}_t^{desired}$ from (13), the PS needs to further use a 1-bit CS reconstruction algorithm $\mathcal{C}^{-1}(\cdot)$ (e.g., binary iterative hard thresholding (BIHT) algorithm [28], fixed point continuation algorithms [31], basis pursuit algorithms [32] and other greedy matching pursuit algorithms [33]) to estimate the global gradient $\hat{\mathbf{g}}_t = \mathcal{C}^{-1}(\hat{\mathbf{y}}_t^{desired})$. Then the PS broadcasts the estimated $\hat{\mathbf{g}}_t$ to all the local workers for updating the shared model parameter as follows

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \hat{\mathbf{g}}_t. \tag{14}$$

Compared (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 [34], 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 [34]:

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

Assumption 3 (sample-wise gradient bounded): The sample-wise gradients at local workers are bounded by their global counterpart [35], [36]

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

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

Assumption 4 (local gradient bounded): The local gradients are bounded by [37]

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

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 describe 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*

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

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 1. **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.*

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

where

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

and $\mathbf{w}_t$ *converges to* $\mathbf{w}^*$.

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 [38]. That is, the FL algorithm achieves an 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 saddle point.

From **Theorem 1**, we have

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

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 a joint optimization of communication and computation next.

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). Minimizing B_t is equivalent to minimizing $R_t = 2LB_t$, i.e.,

$$\begin{aligned} R_t =& \frac{\sum_{i=1}^{U} K_i \rho_1 (1-\beta_{i,t})}{K} + C^2 \left(1 + (1+\delta)\frac{D-\kappa}{DS} G^2 + (\sum_{i=1}^{U} K_i \beta_{i,t} b_t)^{-2} \sigma^2 \right) \\ &+ \sum_{i=1}^{U} \beta_{i,t} (1+\delta) \frac{D-\kappa}{D} G^2. \end{aligned} \tag{24}$$

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. Such a joint optimization problem is formulated as

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

$$\text{s.t.} \quad \frac{\beta_{i,t}^2 K_i^2 b_t^2}{h_{i,t}^2} \leq P_i^{\text{Max}}, \tag{25b}$$

$$\beta_{i,t} \in \{0,1\}, i \in \{1,2,...,U\}. \tag{25c}$$

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 power scaling factor b_t and the 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 [39]. 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:

$\{P_i^{\text{Max}}, h_{i,t}, K_i\}_{i=1}^U$, $\boldsymbol{\Phi}$, 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 2. The enumeration-based method may be applicable for a small number of workers, e.g., $U \leq 10$; 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 much increased searching dimensions 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 increases linearly with the network size U.

The main idea is to decompose the hard 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

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

and

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

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**.

$$\begin{aligned}
\textbf{P3:} \quad \min_{b_t,\{r_{i,t},q_{i,t},\beta_{i,t}\}_{i=1}^{U}} \quad & Q_1(\mathbf{r}_t) + Q_2(\boldsymbol{\beta}_t) && \text{(28a)}\\
\text{s.t.} \quad & \left|\frac{K_i r_{i,t}}{h_{i,t}}\right|^2 \le P_i^{\text{Max}}, && \text{(28b)}\\
& r_{i,t} = \beta_{i,t} q_{i,t}, && \text{(28c)}\\
& q_{i,t} = b_t, && \text{(28d)}\\
& r_{i,t} > 0, b_t > 0, && \text{(28e)}\\
& \beta_{i,t} \in \{0,1\}, && \text{(28f)}\\
& i \in \{1,2,...,U\}. && \text{(28g)}
\end{aligned}$$

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

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

$$\begin{aligned}
\mathcal{L}(b_t, \boldsymbol{\beta}_t, \mathbf{r}_t, \mathbf{q}_t, \boldsymbol{\nu}_t, \boldsymbol{\xi}_t, \boldsymbol{\varsigma}_t) = & Q_1(\mathbf{r}_t) + Q_2(\boldsymbol{\beta}_t) + \sum_{i=1}^{U}\nu_{i,t}\left(\left|\frac{K_i r_{i,t}}{h_{i,t}}\right|^2 - P_i^{\text{Max}}\right) \\
& + \sum_{i=1}^{U}\xi_{i,t}(r_{i,t} - \beta_{i,t}q_{i,t}) + \frac{c}{2}\sum_{i=1}^{U}(r_{i,t} - \beta_{i,t}q_{i,t})^2 \\
& + \sum_{i=1}^{U}\varsigma_{i,t}(q_{i,t} - b_t) + \frac{c}{2}\sum_{i=1}^{U}(q_{i,t} - b_t)^2,
\end{aligned} \tag{29}$$

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

$$\textbf{P4:} \quad \max_{\{\nu_{i,t}, \xi_{i,t}, \varsigma_{i,t}\}_{i=1}^{U}} \quad \mathcal{M}(\boldsymbol{\nu}_t, \boldsymbol{\xi}_t, \boldsymbol{\varsigma}_t) \tag{30a}$$

$$\text{s.t.} \quad \nu_{i,t} \geq 0, \ \xi_{i,t} \geq 0, \ \varsigma_{i,t} \geq 0, i \in \{1, 2, ..., U\}, \tag{30b}$$

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

$$\mathcal{M}(\boldsymbol{\nu}_t, \boldsymbol{\xi}_t, \boldsymbol{\varsigma}_t) = \min_{b_t, \{r_{i,t}, q_{i,t}, \beta_{i,t}\}_{i=1}^{U}} \quad \mathcal{L}(b_t, \mathbf{r}_t, \mathbf{q}_t, \boldsymbol{\beta}_t) \tag{31a}$$

$$\text{s.t.} \quad r_{i,t} > 0, b_t > 0, q_{i,t} > 0, \tag{31b}$$

$$\beta_{i,t} \in \{0, 1\}, i \in \{1, 2, ..., U\}. \tag{31c}$$

The ADMM technique [40] solves the dual problem **P4** by iteratively updating $\{\mathbf{r}_t, b_t\}$, $\{\mathbf{q}_t, \boldsymbol{\beta}_t\}$, and $\{\boldsymbol{\nu}_t, \boldsymbol{\xi}_t, \boldsymbol{\varsigma}_t\}$. 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 $\{\boldsymbol{\nu}_t^{\{l\}}, \boldsymbol{\xi}_t^{\{l\}}, \boldsymbol{\varsigma}_t^{\{l\}}\}$. Then, the update of the variables is sequentially performed at the $(l+1)$-th iteration as follows:

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

$$\{\mathbf{r}_t^{\{l+1\}}, b_t^{\{l+1\}}\} = \arg\min_{\mathbf{r}_t, b_t} \mathcal{L}(\mathbf{r}_t, b_t; \mathbf{p}_t^{\{l\}}, \boldsymbol{\beta}_t^{\{l\}}, \boldsymbol{\nu}_t^{\{l\}}, \boldsymbol{\xi}_t^{\{l\}}, \boldsymbol{\varsigma}_t^{\{l\}}). \tag{32}$$

Notice that (32) is a strictly convex problem, which can be easily solved to obtain the optimal solution, e.g., by using the projected Newton's method [39]. Since the complexity of solving this problem in (32) does not scale with U (i.e., $\mathcal{O}(1)$ complexity), thus the overall computational complexity of **Step 1** is $\mathcal{O}(1)$.

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

$$\{\mathbf{q}_t^{\{l+1\}}, \boldsymbol{\beta}_t^{\{l+1\}}\} = \arg\min_{\mathbf{q}_t, \boldsymbol{\beta}_t} \mathcal{L}(\mathbf{q}_t, \boldsymbol{\beta}_t; \mathbf{r}_t^{\{l+1\}}, b_t^{\{l+1\}}, \boldsymbol{\nu}_t^{\{l\}}, \boldsymbol{\xi}_t^{\{l\}}, \boldsymbol{\varsigma}_t^{\{l\}}). \tag{33}$$

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

$$\{q_{i,t}\}^{\{l+1\}} = \begin{cases} \arg\min_{q_{i,t}} \mathcal{L}\left(q_{i,t}, 0; \{\mathbf{r}_t\}^{\{l+1\}}, \{b_t\}^{\{l+1\}}, \{\nu_{i,t}\}^{\{l\}}, \{\xi_{i,t}\}^{\{l\}}, \{\varsigma_{i,t}\}^{\{l\}}\right), & \beta_{i,t} = 0, \\ \arg\min_{q_{i,t}} \mathcal{L}\left(q_{i,t}, 1; \{\mathbf{r}_t\}^{\{l+1\}}, \{b_t\}^{\{l+1\}}, \{\nu_{i,t}\}^{\{l\}}, \{\xi_{i,t}\}^{\{l\}}, \{\varsigma_{i,t}\}^{\{l\}}\right), & \beta_{i,t} = 1. \end{cases} \tag{34}$$

where

$$\mathcal{L}\left(q_{i,t},0;\{\mathbf{r}_t\}^{\{l+1\}},\{b_t\}^{\{l+1\}},\{\nu_{i,t}\}^{\{l\}},\{\xi_{i,t}\}^{\{l\}},\{\varsigma_{i,t}\}^{\{l\}}\right) \tag{35}$$
$$=\frac{K_i\rho_1}{K}+\{\xi_{i,t}\}^{\{l\}}\{r_{i,t}\}^{\{l+1\}}+\frac{c}{2}\left(\{r_{i,t}\}^{\{l+1\}}\right)^2+\varsigma_{i,t}\left(q_{i,t}-\{b_t\}^{\{l+1\}}\right)+\frac{c}{2}\left(q_{i,t}-\{b_t\}^{\{l+1\}}\right)^2,$$

and

$$\begin{aligned}&\mathcal{L}\left(q_{i,t},1;\{\mathbf{r}_t\}^{\{l+1\}},\{b_t\}^{\{l+1\}},\{\nu_{i,t}\}^{\{l\}},\{\xi_{i,t}\}^{\{l\}},\{\varsigma_{i,t}\}^{\{l\}}\right)\\&=(1+\delta)\frac{D-\kappa}{D}G^2+\{\xi_{i,t}\}^{\{l\}}\left(\{r_{i,t}\}^{\{l+1\}}-q_{i,t}\right)+\frac{c}{2}\left(\{r_{i,t}\}^{\{l+1\}}-q_{i,t}\right)^2\\&\quad+\varsigma_{i,t}\left(q_{i,t}-\{b_t\}^{\{l+1\}}\right)+\frac{c}{2}\left(q_{i,t}-\{b_t\}^{\{l+1\}}\right)^2.\end{aligned} \tag{36}$$

For both $\beta_{i,t}=0$ and $\beta_{i,t}=1$, (34) 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 (34) 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 (33) is given by $\{\mathbf{q}_t^{\{l+1\}},\boldsymbol{\beta}_t^{\{l+1\}}\}$. Notice that the complexity of solving each subproblem in (33) 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 $\{\boldsymbol{\nu}_t,\boldsymbol{\xi}_t,\boldsymbol{\varsigma}_t\}$, which is achieved by updating the multipliers as follows

$$\{\nu_{i,t}\}^{\{l+1\}}=\{\nu_{i,t}\}^{\{l\}}+c\left(\left|\frac{K_i\{r_{i,t}\}^{\{l+1\}}}{h_{i,t}}\right|^2-P_i^{\text{Max}}\right),\quad i=1,...,U, \tag{37}$$

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

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

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 [40]: 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 3. The proposed **Algorithm 2** is guaranteed to converge, because the dual problem **P4** is convex. Its convergence is insensitive to the step size c [41]. Due to the potential duality gap

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 (32);

3: Update $\{\mathbf{q}_t^{\{l+1\}}, \boldsymbol{\beta}_t^{\{l+1\}}\}$ by solving (33);

4: Update $\{\boldsymbol{\nu}_t^{\{l+1\}}, \boldsymbol{\xi}_t^{\{l+1\}}, \boldsymbol{\varsigma}_t^{\{l+1\}}\}$ by using (37), (38), and(39);

5: **Until** {the convergence threshold is satisfied or the maximum number of iterations is reached}.

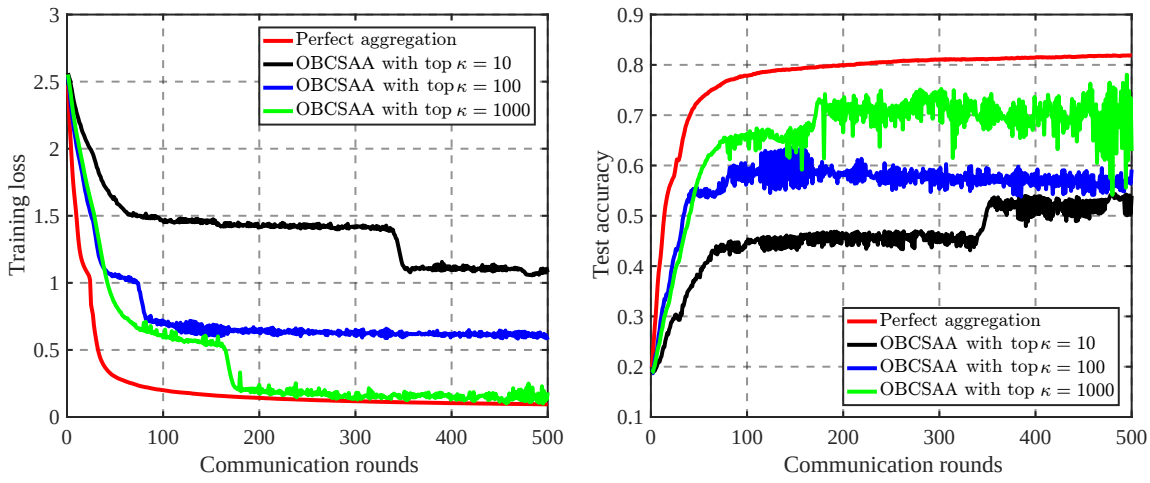
6: **return** $\{b_t^*, \boldsymbol{\beta}_t^*\}$.

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 4. We deduce that 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 $\mathcal{O}(U)$ is less sensitive to U than the complexity $\mathcal{O}(2^U)$ in the enumeration-based method.

V. 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 $U = 10$ workers, and set their maximum peak power to be $P_i^{\texttt{Max}} = P^{\texttt{Max}} = 10$ 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, i.e., $SNR = \frac{P^{\text{Max}}}{\sigma^2} = 5$ dB. We perform top $\kappa = 10$ sparsification, and the dimension of compressed local $\mathcal{C}(\mathbf{g}_i)$'s is set to $S = 1000$. The elements of the measurement matrix $\mathbf{\Phi}$ are generated from $\mathcal{N}(0, 1/S)$. The BIHT algorithm in [28] is selected for the signal reconstruction at the PS.



(a) Training loss

(b) Test accuracy

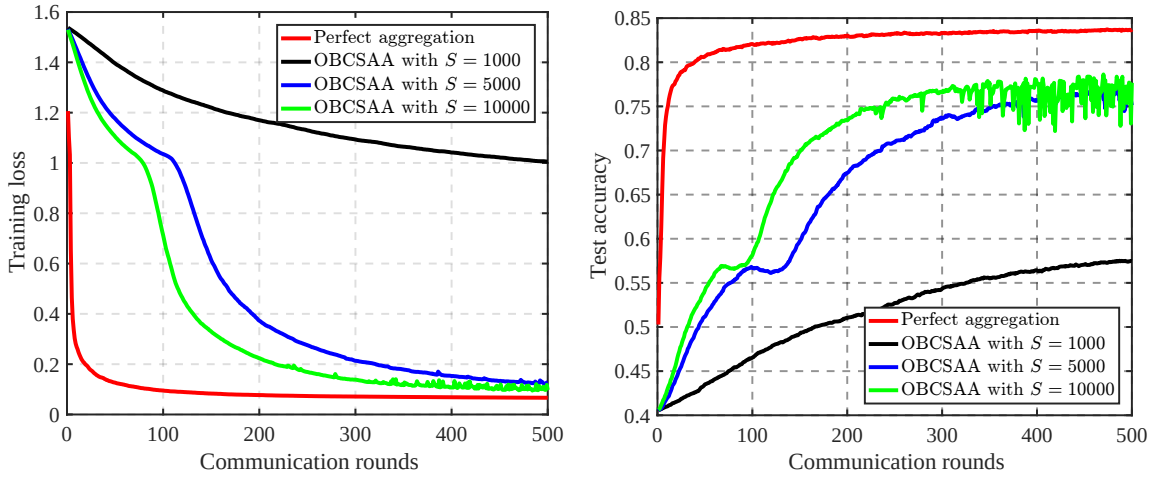
Fig. 1: The performance of our proposed OBCSAA under different sparsification operators compared to perfect aggregation without sparsification.

We consider the learning task of handwritten-digit recognition using the well-known MNIST dataset[4] 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 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. 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*. To satisfy RIP condition,

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



(a) Training loss

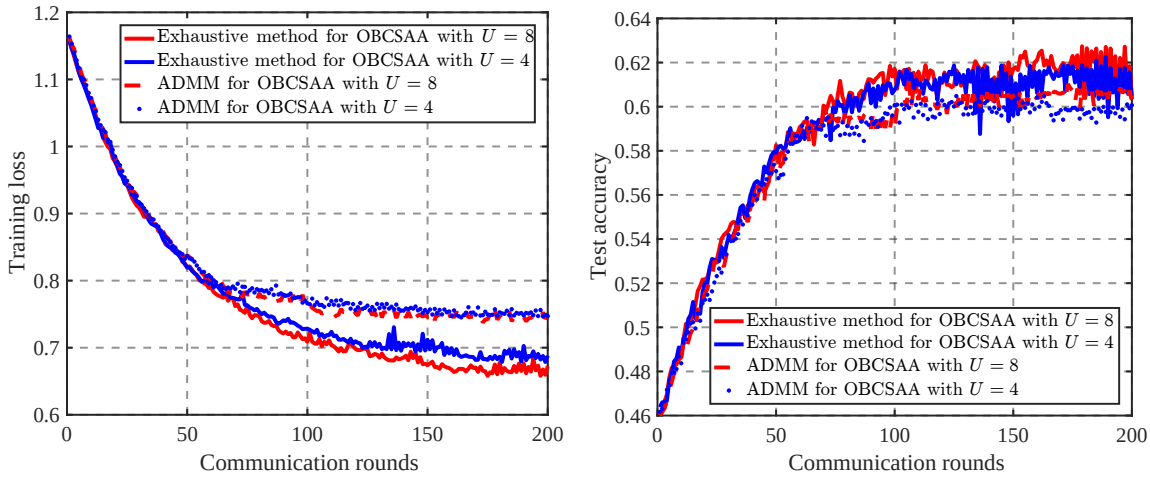
(b) Test accuracy

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

S is set to 10000. It is observed that our proposed OBCSAA can provide desired performance (which approaches to that of *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.

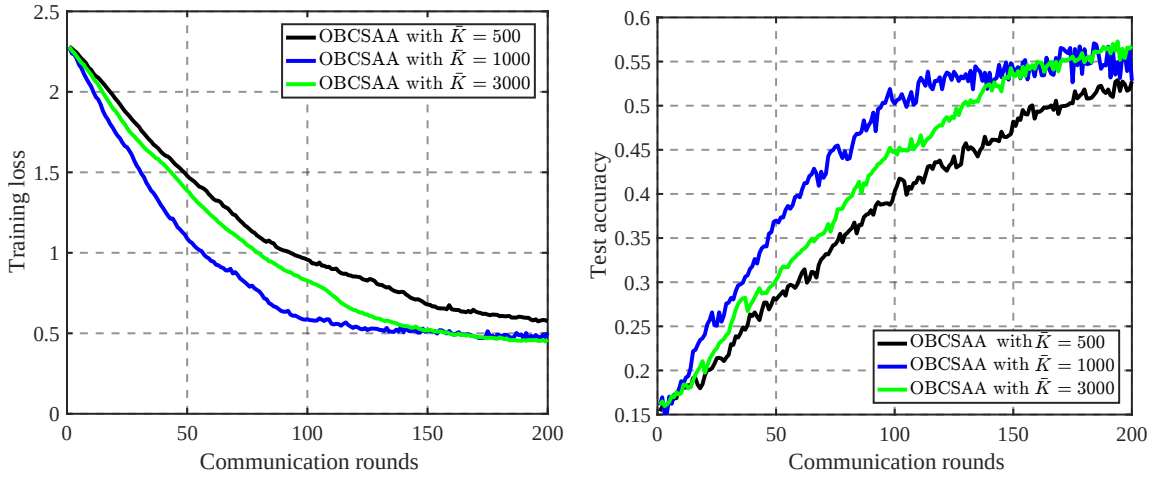
Fig. 2 shows the impact of the reduced dimension size S on the performance of our proposed OBCSAA under $\kappa = 1000$, where the performance increases as S increases. When S is large enough, performance barely increases. This is because that the larger S is, the more conducive to signal reconstruction. 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 10 percent lower than that of *perfect aggregation*. These results illustrates that our OBCSAA under appropriate parameters can greatly reduce the communication overhead and transmission latency while ensuring considerable FL performance.

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



(a) Training loss (b) Test accuracy

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

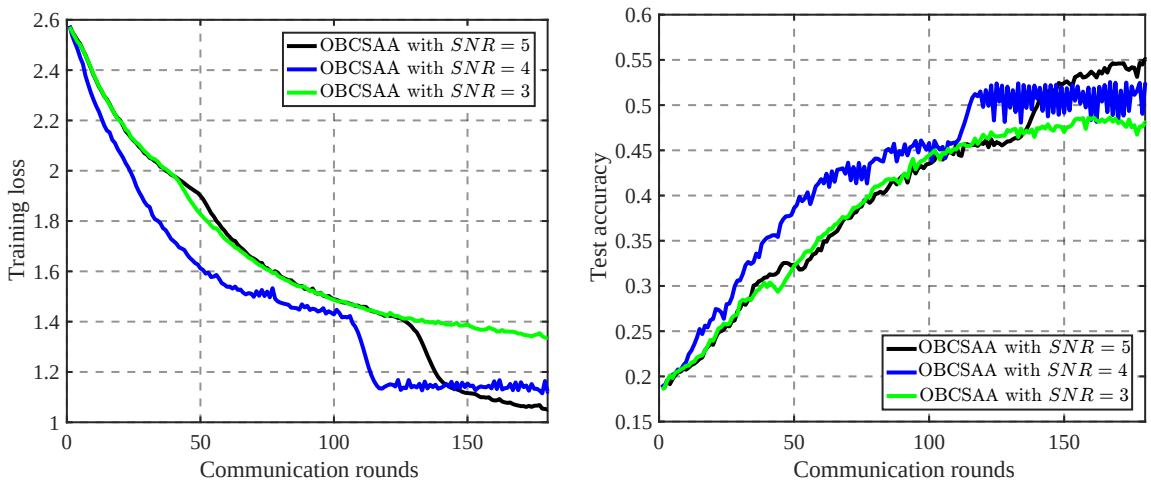


(a) Training loss (b) Test accuracy

Fig. 4: The performance of our proposed OBCSAA under different $\bar{K}$.

joint optimization scheme, which can alleviate the impact of aggregation errors on FL. Besides, we can see that the performance is higher, when the 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.

Fig. 4 presents the impact of the number of data samples per worker $\bar{K}$ on our proposed OBCSAA. In this figure, the performance improves as $\bar{K}$ increases. When $\bar{K}$ is large enough, the



(a) Training loss

(b) Test accuracy

Fig. 5: The performance of our proposed OBCSAA under different the noise variance.

performance barely improves. This is because that as $\bar{K}$ increases, the PS has more data samples for training and hence has higher performance. As $\bar{K}$ continues to increase, the improvement on learning accuracy becomes trivial when the PS already has enough data samples for training.

In Fig. 5, we explore the performance of our proposed OBCSAA under different the noise variance., i.e., different SNR. As expected, as the noise variance increases, i.e., SNR decreases, the performance of our proposed OBCSAA decreases. This is because the larger noise variance is, the more errors would be introduced in the training procedure.

VI. 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 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.

ACKNOWLEDGMENTS

We are very grateful to all reviewers who have helped improve the quality of this paper. This work was partly supported by the National Natural Science Foundation of China (Grant Nos. 61871023 and 61931001), Beijing Natural Science Foundation (Grant No. 4202054), and the National Science Foundation of the US (Grant Nos. 1741338 and 1939553).

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

$$\mathbb{E}\|\mathbf{e}_{i,t}^s\|^2 = \mathbb{E}\|\tilde{\mathbf{g}}_{i,t} - \mathbf{g}_{i,t}\|^2 \leq (1+\delta)\frac{D-\kappa}{D}G^2,\ \ i = 1, ..., U. \tag{40}$$

Since $\mathbf{\Phi}$ satisfies the RIP condition [42],

$$(1-\delta)\|\mathbf{x}\|^2 \leq \|\mathbf{\Phi}\mathbf{x}\|^2 \leq (1+\delta)\|\mathbf{x}\|^2, \tag{41}$$

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

$$\begin{aligned}\mathbb{E}\|\mathbf{e}_{i,t}^q\|^2 &= \mathbb{E}\|\mathtt{sign}(\mathbf{\Phi}\tilde{\mathbf{g}}_{i,t}) - \mathbf{\Phi}\tilde{\mathbf{g}}_{i,t}\|^2 \\ &\leq \mathbb{E}(\|\mathtt{sign}(\mathbf{\Phi}\tilde{\mathbf{g}}_{i,t})\|^2 + \|\mathbf{\Phi}\tilde{\mathbf{g}}_{i,t}\|^2) \\ &\leq S + (1+\delta)\frac{D-\kappa}{D}G^2.\end{aligned} \tag{42}$$

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 [27]:

$$\hat{\mathbf{g}}_t = \arg\min_{\tilde{\mathbf{g}}_t} \|\tilde{\mathbf{g}}_t\|_1 \quad \mathtt{s.t.}\ \|\hat{\mathbf{y}}_t^{desired} - \mathbf{\Phi}\tilde{\mathbf{g}}_t\|^2 \leq \varepsilon_t \tag{43}$$

where ε_t is the norm-limited boundary, which is given by

$$\begin{aligned}
\mathbb{E}\|\hat{\mathbf{y}}_t^{desired} - \mathbf{\Phi}\tilde{\mathbf{g}}_t\|^2 =& \mathbb{E}\left\|\hat{\mathbf{y}}_t^{desired} - \frac{\sum_{i=1}^{U} K_i\beta_{i,t}(\mathbf{\Phi}\tilde{\mathbf{g}}_{i,t})}{\sum_{i=1}^{U} K_i\beta_{i,t}}\right\|^2 \\
=& \mathbb{E}\left\|\frac{\sum_{i=1}^{U} K_i\beta_{i,t}\mathbf{e}_{i,t}^q}{\sum_{i=1}^{U} K_i\beta_{i,t}} + \frac{\mathbf{z}_t}{\sum_{i=1}^{U} K_i\beta_{i,t}b_t}\right\|^2 \\
=& \mathbb{E}\left\|\mathbf{e}_{1,t}^q + \frac{\mathbf{z}_t}{\sum_{i=1}^{U} K_i\beta_{i,t}b_t}\right\|^2 \\
\leq& \mathbb{E}\|\mathbf{e}_{1,t}^q\|^2 + \mathbb{E}\left\|\frac{\mathbf{z}_t}{\sum_{i=1}^{U} K_i\beta_{i,t}b_t}\right\|^2 \\
\leq& S + (1+\delta)\frac{D-\kappa}{D}G^2 + \frac{S\sigma^2}{\left(\sum_{i=1}^{U} K_i\beta_{i,t}b_t\right)^2} \\
\doteq& \varepsilon_t.
\end{aligned} \tag{44}$$

In this case, the reconstruction error norm is bounded by

$$\|\hat{\mathbf{g}}_t - \tilde{\mathbf{g}}_t\|^2 \leq \frac{C^2}{S}\varepsilon_t, \tag{45}$$

where C is the constant depending on the properties of the measurement matrix $\mathbf{\Phi}$ but not on the signal [43]. According to the **Theorem 1.2** in [42], if $\mathbf{\Phi}$ has $\delta \leq \sqrt{2}-1$, C can be given by

$$C = \frac{2\varpi}{1-\varrho}, \tag{46}$$

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 (45) 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

$$\begin{aligned}
\mathbb{E}\|\mathbf{e}_t\|^2 =& \mathbb{E}(\|\hat{\mathbf{g}}_t - \mathbf{g}_t\|^2) = \mathbb{E}(\|\hat{\mathbf{g}}_t - (\tilde{\mathbf{g}}_t + \mathbf{e}_t^s)\|^2) \leq \mathbb{E}(\|\hat{\mathbf{g}}_t - \tilde{\mathbf{g}}_t\|^2 + \|\mathbf{e}_t^s\|^2) \\
&\leq \frac{C^2}{S}\varepsilon_t + \sum_{i=1}^{U}\beta_{i,t}(1+\delta)\frac{D-\kappa}{D}G^2 \\
&= C^2\left(1 + (1+\delta)\frac{D-\kappa}{SD}G^2 + \frac{\sigma^2}{\left(\sum_{i=1}^{U} K_i\beta_{i,t}b_t\right)^2}\right) + \sum_{i=1}^{U}\beta_{i,t}(1+\delta)\frac{D-\kappa}{D}G^2,
\end{aligned} \tag{47}$$

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

$$\begin{aligned} F(\mathbf{w}_t) &= F(\mathbf{w}_{t-1}) + (\mathbf{w}_t - \mathbf{w}_{t-1})^T \nabla F(\mathbf{w}_{t-1}) + \frac{1}{2}(\mathbf{w}_t - \mathbf{w}_{t-1})^T \nabla^2 F(\mathbf{w}_{t-1})(\mathbf{w}_t - \mathbf{w}_{t-1}) \\ &\overset{(a)}{\leq} F(\mathbf{w}_{t-1}) + (\mathbf{w}_t - \mathbf{w}_{t-1})^T \nabla F(\mathbf{w}_{t-1}) + \frac{L}{2} \parallel \mathbf{w}_t - \mathbf{w}_{t-1} \parallel^2, \end{aligned} \tag{48}$$

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

After recovering the desired $\hat{\mathbf{g}}_t$ from the received signal by solving (43), then the common model is updated by

$$\begin{aligned} \mathbf{w}_t =& \mathbf{w}_{t-1} - \alpha \hat{\mathbf{g}}_t \\ =& \mathbf{w}_{t-1} - \alpha(\nabla F(\mathbf{w}_{t-1}) - \mathbf{o}), \end{aligned} \tag{49}$$

where

$$\mathbf{o} = \nabla F(\mathbf{w}_{t-1}) - \hat{\mathbf{g}}_t. \tag{50}$$

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 (48) can be expressed as

$$\begin{aligned} \mathbb{E}[F(\mathbf{w}_t)] \leq& \mathbb{E}\Big[F(\mathbf{w}_{t-1}) - \alpha(\nabla F(\mathbf{w}_{t-1}) - \mathbf{o})^T \nabla F(\mathbf{w}_{t-1}) + \frac{L\alpha^2}{2} \parallel \nabla F(\mathbf{w}_{t-1}) - \mathbf{o} \parallel^2 \Big] \\ \overset{(b)}{=}& \mathbb{E}[F(\mathbf{w}_{t-1})] - \frac{1}{2L} \parallel \nabla F(\mathbf{w}_{t-1}) \parallel^2 + \frac{1}{2L} \mathbb{E}[\parallel \mathbf{o} \parallel^2], \end{aligned} \tag{51}$$

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

$$\frac{L\alpha^2}{2} \parallel \nabla F(\mathbf{w}_{t-1}) - \mathbf{o} \parallel^2 = \frac{1}{2L} \parallel \nabla F(\mathbf{w}_{t-1}) \parallel^2 - \frac{1}{L} \mathbf{o}^T \nabla F(\mathbf{w}_{t-1}) + \frac{1}{2L} \parallel \mathbf{o} \parallel^2 . \tag{52}$$

According to (47), $\|\mathbf{e}_t\|^2 \leq \frac{C^2}{S}\varepsilon_t + \sum_{i=1}^{U}\beta_{i,t}(1+\delta)\frac{D-\kappa}{D}G^2$. Then we derive $\mathbb{E}[\| \mathbf{o} \|^2]$ as follows

$$\begin{aligned}
\mathbb{E}[\| \mathbf{o} \|^2] =&\mathbb{E}[\|\nabla F(\mathbf{w}_{t-1}) - \hat{\mathbf{g}}_t\|] \\
=&\mathbb{E}[\|\nabla F(\mathbf{w}_{t-1}) - \mathbf{g}_t - \mathbf{e}_t\|] \\
=&\mathbb{E}\Bigg[\Bigg\| \frac{\sum_{i=1}^{U}\sum_{k=1}^{K_i}\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})}{K} \\
&- (\sum_{i=1}^{U}K_i\beta_{i,t})^{-1}\sum_{i=1}^{U}\sum_{k=1}^{K_i}\beta_{i,t}\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k}) - \mathbf{e}_t\Bigg\|^2\Bigg] \\
\leq&\mathbb{E}\left[\left\| \frac{\sum_{i=1}^{U}\sum_{k=1}^{K_i}\nabla f(\mathbf{w}_{t-1},\mathbf{x}_{i,k},\mathbf{y}_{i,k})(1-\beta_{i,t})}{K} - \mathbf{e}_t\right\|^2\right],
\end{aligned} \tag{53}$$

Applying the triangle inequality of norms: $\|\mathbf{X}+\mathbf{Y}\| \leq \|\mathbf{X}\| + \|\mathbf{Y}\|$, and the submultiplicative property of norms: $\|\mathbf{XY}\| \leq \|\mathbf{X}\|\|\mathbf{Y}\|$, we further derive (53) as follows

$$\begin{aligned}
\mathbb{E}[\| \mathbf{o} \|^2] \leq&\mathbb{E}\left[\frac{\sum_{i=1}^{U}\sum_{k=1}^{K_i}\| \nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})(1-\beta_{i,t}) \|^2}{K}\right] + \mathbb{E}[\|\mathbf{e}_{t-1}\|^2] \\
\leq&\mathbb{E}\left[\frac{\sum_{i=1}^{U}\sum_{k=1}^{K_i}\| \nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k}) \|^2 (1-\beta_{i,t})^2}{K}\right] + \mathbb{E}[\|\mathbf{e}_{t-1}\|^2] \\
\leq&\frac{\sum_{i=1}^{U}\sum_{k=1}^{K_i}\| \nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k}) \|^2 (1-\beta_{i,t})}{K} + \frac{C^2}{S}\varepsilon_t + \sum_{i=1}^{U}\beta_{i,t}(1+\delta)\frac{D-\kappa}{D}G^2.
\end{aligned} \tag{54}$$

Applying (17) in **Assumption 3** to (54), we further derive the following result as

$$\mathbb{E}[\| \mathbf{o} \|^2] \leq \frac{1}{K}\sum_{i=1}^{U}K_i(\rho_1 + \rho_2 \| \nabla F(\mathbf{w}_{t-1}) \|^2)(1-\beta_{i,t}) + \frac{C^2}{S}\varepsilon_t + \sum_{i=1}^{U}\beta_{i,t}(1+\delta)\frac{D-\kappa}{D}G^2. \tag{55}$$

Substituting (55) to (51), we have:

$$\begin{aligned}
\mathbb{E}[F(\mathbf{w}_t)] \leq&\frac{1}{2L}\left(\frac{1}{K}\sum_{i=1}^{U}K_i(\rho_1 + \rho_2 \| \nabla F(\mathbf{w}_{t-1}) \|^2)(1-\beta_{i,t}) + \frac{C^2}{S}\varepsilon_t + \sum_{i=1}^{U}\beta_{i,t}(1+\delta)\frac{D-\kappa}{D}G^2\right) \\
&+ \mathbb{E}[F(\mathbf{w}_{t-1})] - \frac{1}{2L}\| \nabla F(\mathbf{w}_{t-1}) \|^2 \\
=&\mathbb{E}[F(\mathbf{w}_{t-1})] + \left(\frac{\sum_{i=1}^{U}K_i\rho_2(1-\beta_{i,t})}{2LK} - \frac{1}{2L}\right)\| \nabla F(\mathbf{w}_{t-1}) \|^2 \\
&+ \frac{\sum_{i=1}^{U}K_i\rho_1(1-\beta_{i,t})}{2LK} + \frac{1}{2L}(\frac{C^2}{S}\varepsilon_t + \sum_{i=1}^{U}\beta_{i,t}(1+\delta)\frac{D-\kappa}{D}G^2).
\end{aligned} \tag{56}$$

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

$$\mathbb{E}[F(\mathbf{w}_t)-F(\mathbf{w}_0)] \leq -\sum_{t=1}^{T} A_t \parallel \nabla F(\mathbf{w}_{t-1}) \parallel^2 + \sum_{t=1}^{T} B_t, \tag{57}$$

where

$$A_t = \frac{1}{2L} - \frac{\sum_{i=1}^{U} K_i \rho_2 (1-\beta_{i,t})}{2LK}, \tag{58}$$

$$B_t = \frac{\sum_{i=1}^{U} K_i \rho_1 (1-\beta_{i,t})}{2LK} + \frac{1}{2L}(\frac{C^2}{S}\varepsilon_t + \sum_{i=1}^{U} \beta_{i,t}(1+\delta)\frac{D-\kappa}{D}G^2). \tag{59}$$

The inequality (57) can be also written as

$$\sum_{t=1}^{T} A_t \parallel \nabla F(\mathbf{w}_{t-1}) \parallel^2 \leq \mathbb{E}[F(\mathbf{w}_0)-F(\mathbf{w}_t)] + \sum_{t=1}^{T} B_t \leq \mathbb{E}[F(\mathbf{w}_0)-F(\mathbf{w}^*)] + \sum_{t=1}^{T} B_t. \tag{60}$$

Since $\frac{1-\rho_2}{2L} \leq A_t \leq \frac{1}{2L}$, we have

$$\frac{1}{T}\sum_{t=1}^{T} \frac{1-\rho_2}{2L} \parallel \nabla F(\mathbf{w}_{t-1}) \parallel^2 \leq \frac{1}{T}\sum_{t=1}^{T} A_t \parallel \nabla F(\mathbf{w}_{t-1}) \parallel^2 \leq \frac{1}{T}\mathbb{E}[F(\mathbf{w}_0)-F(\mathbf{w}^*)] + \frac{1}{T}\sum_{t=1}^{T} B_t. \tag{61}$$

As a result, we get

$$\frac{1}{T}\sum_{t=1}^{T} \parallel \nabla F(\mathbf{w}_{t-1}) \parallel^2 \leq \frac{2L}{T(1-\rho_2)}\mathbb{E}[F(\mathbf{w}_0)-F(\mathbf{w}^*)] + \frac{2L}{T(1-\rho_2)}\sum_{t=1}^{T} B_t. \tag{62}$$

The proof is completed. □

REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

[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] 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.

[7] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," *arXiv preprint arXiv:1704.05021*, 2017.

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

[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 T. Zhi, "Coke: Communication-censored kernel learning for decentralized non-parametric learning," *arXiv preprint arXiv:2001.10133*, 2020.

[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] 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.

[19] 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.

[20] G. Zhu, Y. Du, D. Gunduz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *arXiv preprint arXiv:2001.05713*, 2020.

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

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

[23] M. M. Amiri, T. M. Duman, and D. Gündüz, "Collaborative machine learning at the wireless edge with blind transmitters," *arXiv preprint arXiv:1907.03909*, 2019.

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

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

[26] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning in fading channels," *arXiv preprint arXiv:2003.02089*, 2020.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] S. Bubeck, "Convex optimization: Algorithms and complexity," *arXiv preprint arXiv:1405.4980*, 2014.

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

[36] 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.

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

[38] J. Wang and G. Joshi, "Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms," *arXiv preprint arXiv:1808.07576*, 2018.

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

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

[41] 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.

[42] 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.

[43] 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.