2021 IEEE International Conference on Communications Workshops (ICC Workshops) | 978-1-7281-9441-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCWorkshops50388.2021.9473872

Communication-efficient Federated Learning Through 1-Bit Compressive Sensing and Analog Aggregation

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

***Abstract*—This paper studies communication-efficient federated learning (FL) over the air, which is based on 1-bit compressive sensing (CS) and analog aggregation transmissions. To analyze the impact of these two communication efficiency oriented technologies on FL, we derive a closed-form expression for the expected convergence rate of the FL algorithm. Our theoretical result implies that the communication efficiency comes at the expense of the performance degradation due to 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 on FL by an optimal scheduling and power scaling policy. An enumerated method is proposed to solve this non-convex problem, which is optimal but becomes computationally infeasible as the number of devices increases. Hence, we further propose a suboptimal solution based on the alternating direction method of multiplier to reduce the complexity when applied in 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 at network edges to a data center for data analysis and inference is becoming increasingly costly for communications. To reduce communication costs, the work in [1], [2] investigating approximate data aggregation instead of exact data aggregation has been theoretically and practically demonstrated to be effective. However, privacy concerns about sharing data also stymie centralized ML. As an alternative communication-efficient and privacy-preserving distributed ML scheme, federated learning (FL) is a prosperous paradigm that enables many edge devices (local workers) collaboratively to train a common learning model under the coordination of a parameter server (PS) in wireless networks [3]. Uploading model parameters instead of raw datasets, FL offers distinct advantages on protecting user privacy and leveraging distributed on-device computation compared to traditional centralized ML.

Since updates shared between local workers and the PS can be extremely large, the communication overhead becomes a main bottleneck of FL [4], [5]. To reduce the communication load per worker, the pre-processing of updates has been considered in the literature, such as *sparsification* [6], [7], *quantization* [8]–[10] and *communication censoring* [11], [12]. However, the communication overhead and transmission latency of FL over digital communication channels are still proportional to the number of active workers, and thus inefficient in large-scale environment. Fortunately, an analog aggregation model is recently proposed for communication-efficient FL by applying a computation over the air principle [13]. This benefits from the fact that FL only utilizes the averaged value of 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 [14]–[21].

Despite the effort of the prior work, some fundamental questions are unexplored to achieve a reliable and high-performance for analog aggregation based FL. Firstly, the specific relationship between FL and analog aggregation communication is not clear. Purely maximization of the number of participated workers is not necessarily optimal, which treats optimization on computation and communication one-after-another, e.g., [16], [21]. Secondly, to facilitate power control, most existing works assume that the signals to be transmit from local workers, i.e., local gradients, can be normalized with zero mean and unit variance, e.g., [14]–[17]. However, gradient statistics in FL vary over training iterations, and are unknown in advance [22]. Thus, it is infeasible to design an optimal power control without local gradients known at the PS in advance, due to the non-coding linear analog modulation in analog aggregation based FL. Thirdly, sparsification is considered in analog aggregation based FL [18], [19], which is a kind of lossy compression that may introduce aggregation errors, but the impact of these aggregation errors on FL is not yet clear let alone alleviating them.

Motivated by the above issues, in this paper, we study 1-bit compressive sensing (CS) for FL over the air, by developing a feasible worker selection and power control policy. To thoroughly improve communication-efficiency, to the best of our knowledge, this is the first work to introduce 1-bit CS [23]–[25] into FL over the air, 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 in the absence of any prior knowledge or assumptions on gradient statistics or

978-1-7281-9441-7/21/$31.00 ©2021 IEEE

specific distribution. More important, our work provides an important interpretation on the relationship between FL and analog aggregation with 1-bit CS techniques through a joint optimization of computation and communications. Our main contributions are outlined below:

- Building on 1-bit CS, we design an elaborate analog aggregation based FL, called OBCSAA. In our OBCSAA, workers first sparsify their local gradients and project the sparsified gradient vectors onto lower dimensional spaces. The lower dimensional vectors are quantized by 1-bit quantization, and then all the workers simultaneously transmit their preprocessing vectors to the PS in the same time-frequency resource. The PS reconstructs the averaged sparse gradients from its noisy observations.
- We derive a closed-form expression for the expected convergence rate of our OBCSAA, which not only interprets but also measures the impact of analog aggregation wireless communications and 1-bit CS on FL over the air.
- We formulate a joint optimization problem of computation and communication to optimize the worker selection and power control. To solve this non-convex optimization problem, we propose two solutions: the enumeration-based method and alternating direction method of multipliers (ADMM), which are applied to the scenarios where the number of participated workers are small and large, respectively.

We evaluate the proposed OBCSAA in solving an image classification problem 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, while greatly improves the communication efficiency.

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. Then 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 a 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 achieved in the distributed setting by an iterative gradient-averaging algorithm [26], [27].

At each iteration t, the gradient descent (GD) is applied at local workers in parallel[1] to minimize the local loss functions

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

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

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

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 achieved.

B. Analog Aggregation Transmission Model

In order to reduce the transmission overhead and speed up communication time, we propose to apply 1-bit CS [23]–[25] in FL over the air, which allows to reduce the dimension of the transmitted vectors and enables all the local workers to simultaneously use the same time-frequency resources to transmit their updates to the PS.

1) Sparsification: Before transmission at the t-th iteration, all the workers set all but the κ elements of their local $\mathbf{g}_{i,t}$'s to 0, resulting κ-th 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. In our paper, we perform the top-κ sparsification, i.e., elements with the highest κ magnitudes are retained, other elements are set to 0.

2) Dimension Reduction: All workers employ the same measurement matrix $\mathbf{\Phi} \in \mathbb{R}^{S \times D}$ that is a random Gaussian sub-Nyquist sampling matrix. To satisfy the Restricted Isometry Property (RIP) [28], $U\kappa \leq S \ll D$ and each entry of $\mathbf{\Phi}$ i.i.d. follows $\mathcal{N}(0, \sigma_{sp}^2)$, where $U\kappa$ is the upper bound of sparisty in the combined sparse gradient. In addition, $\mathbf{\Phi}$ is shared between the workers and the PS before transmissions.

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

$$\mathcal{C}(\mathbf{g}_{i,t}) = \texttt{sign}(\mathbf{\Phi}\texttt{sparse}_\kappa(\mathbf{g}_{i,t}) = \texttt{sign}(\mathbf{\Phi}\tilde{\mathbf{g}}_{i,t}), \tag{7}$$

where $\mathcal{C}(\cdot)$ represents the entire compressive operation.

4) Analog Aggregation Transmission: After the above compressive operation, all the workers transmit their compressed local $\mathcal{C}(\mathbf{g}_{i,t})$'s in an analog fashion, so that they are aggregated over the air at the PS to implement the global gradient

[1] In this work, we take the basic gradient descent as an example, which can be extend to the stochastic gradient descent (SGD) by using a mini-batch at each worker for training. Note that SGD needs more iterations and hence more transmissions compared to GD.

computing step in (4). Each local gradient $\mathcal{C}(\mathbf{g}_{i,t})$ is multiplied with a pre-processing power control factor, denoted as $p_{i,t}$. 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 from the PS and the i-th local worker at the t-th iteration[2].

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

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

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 $\mathcal{C}(\mathbf{g}_{i,t}) = [c_{i,t}^1, ..., c_{i,t}^s, ..., c_{i,t}^S]^T$.

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

$$\hat{\mathbf{y}}_t^{desired} = \frac{\mathbf{y}_t}{\sum_{i=1}^{U} K_i \beta_{i,t} b_t} = \mathbf{y}_t^{desired} + \frac{\mathbf{z}_t}{\sum_{i=1}^{U} K_i \beta_{i,t} b_t}, \tag{13}$$

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 [24] and other greedy matching pursuit algorithms [29]) 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 FL, due to analog aggregation transmissions and 1-bit CS.

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 [30], 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.

Assumption 2 (twice-continuously differentiable): $F(\mathbf{w})$ is twice-continuously differentiable. Accordingly, the eigenvalues of the Hessian matrix of $F(\mathbf{w})$ are bounded by [30]:

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

Assumption 3 (sample-wise gradient bounded): The sample-wise gradients at local workers are bounded by their global counterpart [31], [32]

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

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

Assumption 4 (local gradient bounded): The local gradients are bounded by [33]

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

where G is positive constant.

B. Convergence Analysis

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.*

$$\frac{1}{T} \sum_{t=1}^{T} \| \nabla F(\mathbf{w}_{t-1}) \|^2 \leq \frac{2L}{T(1-\rho_2)} \mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)] + \frac{2L}{T(1-\rho_2)} \sum_{t=1}^{T} B_t, \tag{19}$$

where $\mathbf{w}_t$ converges to $\mathbf{w}^*$, $B_t = \frac{\sum_{i=1}^{U} K_i \rho_1 (1-\beta_{i,t})}{2LK \sum_{i=1}^{U} K_i \beta_{i,t}} + \sum_{i=1}^{U} \beta_{i,t} (1+\delta) \frac{D-\kappa}{2LD} G^2 + \frac{C^2}{2L} \left(1 + (1+\delta) \frac{D-\kappa}{SD} G^2 + \frac{\sigma^2}{\left(\sum_{i=1}^{U} K_i \beta_{i,t} b_t \right)^2} \right)$, $C = \frac{2\varpi}{1-\varrho}$, $\varpi = \frac{2\sqrt{1+\delta}}{\sqrt{1-\delta}}$, $\varrho = \frac{\sqrt{2}\delta}{1-\delta}$, and $0 < \delta < 1$ is the constant in the RIP condition.

Proof. Given the page limit, please see our journal version: https://arxiv.org/abs/2103.16055. □

In **Theorem 1**, the expected gradient norm is used as an indicator of convergence [34]. That is, the FL algorithm achieves an τ-suboptimal solution if:

$$\frac{1}{T} \sum_{t=1}^{T} \| \nabla F(\mathbf{w}_{t-1}) \|^2 \leq \tau. \tag{20}$$

From **Theorem 1**, we have

$$\frac{1}{T} \sum_{t=1}^{T} \| \nabla F(\mathbf{w}_{t-1}) \|^2 \leq \frac{2L}{T(1-\rho_2)} \mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)] + \frac{2L}{T(1-\rho_2)} \sum_{t=1}^{T} B_t \overset{T \to \infty}{\longrightarrow} \frac{2L}{T(1-\rho_2)} \sum_{t=1}^{T} B_t. \tag{21}$$

The error floor at convergence is given by (21). Obviously, minimizing this error floor can improve the convergence performance of FL. To this end, we provide a joint optimization of communication and computation next.

[2] 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 over iterations. We assume that the CSI is perfectly known at both the PS and local workers.

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 (21) 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 feasible 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 (21) is accumulated over iterations, result 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 accounts 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.,

$$R_t = \frac{\sum_{i=1}^{U} K_i \rho_1 (1-\beta_{i,t})}{K\sum_{i=1}^{U} K_i \beta_{i,t}} + C^2(1+(1+\delta)\frac{D-\kappa}{DS}G^2 + (\sum_{i=1}^{U} K_i \beta_{i,t} b_t)^{-2}\sigma^2) + \sum_{i=1}^{U}\beta_{i,t}(1+\delta)\frac{D-\kappa}{D}G^2. \tag{22}$$

Given the factors (i.e., C, S, and κ) related to 1-bit CS fixed, the joint optimization problem carried out at the PS during the t-th iteration to determine the power scaling factor b_t and the scheduling indicator $\boldsymbol{\beta}_t = [\beta_{1,t}, \beta_{2,t}, ..., \beta_{U,t}]$ is formulated as

$$\begin{aligned} \textbf{P2:} \quad & \min_{b_t, \boldsymbol{\beta}_t} \quad R_t && \text{(23a)}\\ & \text{s.t.} \quad \frac{\beta_{i,t}^2 K_i^2 b_t^2}{h_{i,t}^2} \leq P_i^{\text{Max}}, && \text{(23b)}\\ & \qquad \beta_{i,t} \in \{0,1\}, i \in \{1,2,...,U\}. && \text{(23c)} \end{aligned}$$

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 [35]. 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.

Remark 1. The enumeration-based method may be applicable for a small number of workers, e.g., $U \leq 10$, however, quickly becomes computationally infeasible as U increases.

C. ADMM-based Suboptimal Solution

Since the enumeration-based method susceptible to high computational complexity, we propose an ADMM-based algorithm to jointly optimize the local worker selection and power control. 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 coupling variables $\{b_t, \boldsymbol{\beta}_t\}$ and constraint (23b) among the workers. To eliminate these coupling factors, we first introduce an artificial vector $\mathbf{r}_t = [r_{1,t}, r_{2,t}, ..., r_{U,t}]$ and define two auxiliary functions as

$$Q_1(\mathbf{r}_t) = C^2(\sum_{i=1}^{U} K_i r_{i,t})^{-2}\sigma^2 \tag{24}$$

and

$$Q_2(\boldsymbol{\beta}_t) = \frac{\sum_{i=1}^{U} K_i \rho_1 (1-\beta_{i,t})}{K\sum_{i=1}^{U} K_i \beta_{i,t}} + C^2(1+(1+\delta)\frac{D-\kappa}{SD}G^2) + \sum_{i=1}^{U}\beta_{i,t}(1+\delta)\frac{D-\kappa}{D}G^2. \tag{25}$$

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

$$\begin{aligned} \textbf{P3:} \quad & \min_{b_t, \{r_{i,t}, q_{i,t}, \beta_{i,t}\}_{i=1}^{U}} \quad Q_1(\mathbf{r}_t) + Q_2(\boldsymbol{\beta}_t) && \text{(26a)}\\ & \text{s.t.} \quad \left|\frac{K_i r_{i,t}}{h_{i,t}}\right|^2 \leq P_i^{\text{Max}}, && \text{(26b)}\\ & \qquad r_{i,t} = \beta_{i,t} q_{i,t}, && \text{(26c)}\\ & \qquad q_{i,t} = b_t, && \text{(26d)}\\ & \qquad r_{i,t} > 0, b_t > 0, && \text{(26e)}\\ & \qquad \beta_{i,t} \in \{0,1\}, && \text{(26f)}\\ & \qquad i \in \{1,2,...,U\}, && \text{(26g)} \end{aligned}$$

where the constraints (26c) and (26d) are introduced for the decomposition of the coupling factors $\beta_{i,t}$ and b_t while guaranteeing that **P3** and **P2** are equivalent.

By introducing multipliers $\nu_{i,t} \geq 0$'s, $\xi_{i,t} \geq 0$'s and $\varsigma_{i,t} \geq 0$'s to the constraints in (26b), (26c) and (26d), we can write a partial augmented Lagrangian of **P3** as

$$\begin{aligned} &\mathcal{L}(b_t, \boldsymbol{\beta}_t, \mathbf{r}_t, \mathbf{q}_t, \boldsymbol{\nu}_t, \boldsymbol{\xi}_t, \boldsymbol{\varsigma}_t)\\ &= Q_1(\mathbf{r}_t) + Q_2(\boldsymbol{\beta}_t) + \sum_{i=1}^{U}\nu_{i,t}\Big(\left|\frac{K_i r_{i,t}}{h_{i,t}}\right|^2 - P_i^{\text{Max}}\Big)\\ &+ \sum_{i=1}^{U}\xi_{i,t}(r_{i,t} - \beta_{i,t} q_{i,t}) + \frac{c}{2}\sum_{i=1}^{U}(r_{i,t} - \beta_{i,t} q_{i,t})^2\\ &+ \sum_{i=1}^{U}\varsigma_{i,t}(q_{i,t} - b_t) + \frac{c}{2}\sum_{i=1}^{U}(q_{i,t} - b_t)^2, \end{aligned} \tag{27}$$

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

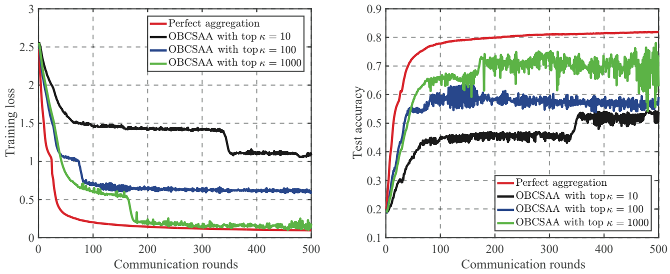
$$\begin{aligned} \textbf{P4:} \quad & \max_{\{\nu_{i,t}, \xi_{i,t}, \varsigma_{i,t}\}_{i=1}^{U}} \quad \mathcal{M}(\boldsymbol{\nu}_t, \boldsymbol{\xi}_t, \boldsymbol{\varsigma}_t) && \text{(28a)}\\ & \text{s.t.} \quad \nu_{i,t} \geq 0, \ \xi_{i,t} \geq 0, \ \varsigma_{i,t} \geq 0, && \text{(28b)}\\ & \qquad i \in \{1,2,...,U\}, && \text{(28c)} \end{aligned}$$

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

$$\begin{aligned} & \min_{b_t, \{r_{i,t}, q_{i,t}, \beta_{i,t}\}_{i=1}^{U}} \quad \mathcal{L}(b_t, \mathbf{r}_t, \mathbf{q}_t, \boldsymbol{\beta}_t) && \text{(29a)}\\ & \text{s.t.} \quad r_{i,t} > 0, b_t > 0, q_{i,t} > 0, && \text{(29b)}\\ & \qquad \beta_{i,t} \in \{0,1\}, i \in \{1,2,...,U\}. && \text{(29c)} \end{aligned}$$

The ADMM technique [36] solves the dual problem **P4** by iteratively updating $\{\mathbf{r}_t, b_t\}$, $\{\mathbf{q}_t, \boldsymbol{\beta}_t\}$, and $\{\boldsymbol{\nu}_t, \boldsymbol{\xi}_t, \boldsymbol{\varsigma}_t\}$. Due to the page limit, we will detail the update of the variables sequentially performed at each ADMM iteration in our journal version.

Remark 2. We can infer that the computational complexity of one ADMM iteration (including the 3 updating steps) is $\mathcal{O}(U)$, because the highest complexity of these three updating 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.



(a) Training loss (b) Test accuracy

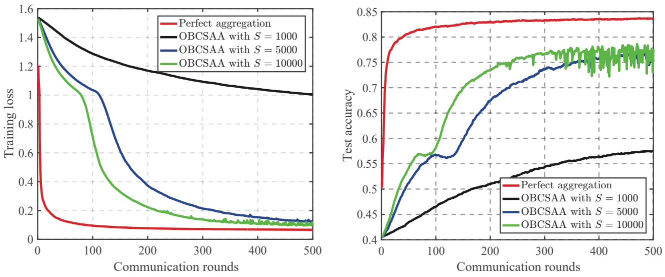
Fig. 1: The performance of our proposed OBCSAA under different sparsification operators.

V. Simulation Results And Analysis

In the simulations, we evaluate the performance of the proposed one 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 $P_i^{\text{Max}} = P^{\text{Max}} = 10$ mW for any $i \in [1, U]$. The wireless channels from the PS to the workers are modeled as i.i.d. Rayleigh fading, so that $h_{i,t}$'s are generated 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 S is set to 1000. The elements of the measurement matrix $\mathbf{\Phi}$ are generated from $\mathcal{N}(0, 1/S)$. The BIHT algorithm in [24] is selected for the signal reconstruction at the PS.

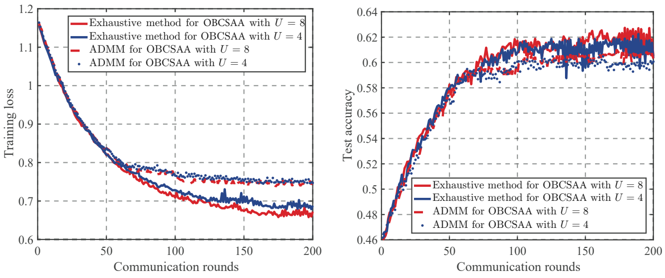
We consider the learning task of handwritten-digit recognition using the MNIST dataset[3]. In the MNIST dataset, a total of 60000 labeled training data samples and 10000 test samples are available for training a learning model. In our experiments, we train a multilayer perceptron (MLP) with a 784-neuron 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 training samples and distribute them to all local workers as their local data, i.e., $K_i = \bar{K} = 3000$, for any $i \in [1, U]$.

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 communication 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, S is set to 10000. It is observed that our proposed OBCSAA can still guarantee considerable performance (perform closely to *perfect aggregation*), despite a fairly high degree of sparsity, e.g., $\kappa = 1000$, and 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.



(a) Training loss (b) Test accuracy

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



(a) Training loss (b) Test accuracy

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

Fig. 2 shows how the reduced dimension size S affect the performance of our proposed OBCSAA under $\kappa = 1000$. As we can see, 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. On the other hand, 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 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.

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

VI. Conclusion

In this paper, we propose a communication-efficient FL based on 1-bit CS and analog aggregation transmissions. To quantify the impact of the sparsity, compression, and reconstruction of 1-bit compressive sensing on FL over the air, we derive a closed-form expression for the expected convergence rate of the FL algorithm. This theoretical result indicates that the application of 1-bit CS and analog aggregation transmissions leads to a performance degradation due to aggregation errors. To mitigate this performance degradation, we formulate a joint optimization problem of communication and learning, which is to find an optimal solution for worker selection and power control. An enumeration-based method and an ADMM-based method are proposed to solve this challenging non-convex problem, which are suitable for obtaining the optimal solution for small-scale networks and approximate solution for large-scale networks, respectively. Simulation results show that our proposed FL can greatly improve communication efficiency while ensuring desired learning performance.

Acknowledgments

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).

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] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[4] H. B. McMahan, E. Moore, D. Ramage, S. Hampson *et al.*, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

[5] M. I. Jordan, J. D. Lee, and Y. Yang, "Communication-efficient distributed statistical inference," *Journal of the American Statistical Association*, vol. 114, no. 526, pp. 668–681, 2019.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] K. Bryan and T. Leise, "Making do with less: An introduction to compressed sensing," *Siam Review*, vol. 55, no. 3, pp. 547–566, 2013.

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

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

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

[32] 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.

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

[34] 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.

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

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