

On the Design of Communication Efficient Federated Learning over Wireless Networks

Richeng Jin, *Student Member, IEEE*, Xiaofan He, *Member, IEEE*, Huaiyu Dai, *Fellow, IEEE*

Abstract—Recently, federated learning (FL), as a promising distributed machine learning approach, has attracted lots of research efforts. In FL, the parameter server and the mobile devices share the training parameters over wireless links. As a result, reducing the communication overhead becomes one of the most critical challenges. Despite that there have been various communication-efficient machine learning algorithms in literature, few of the existing works consider their implementation over wireless networks. In this work, the idea of SignSGD is adopted and only the signs of the gradients are shared between the mobile devices and the parameter server. In addition, different from most of the existing works that consider Channel State Information (CSI) at both the transmitter side and the receiver side, only receiver side CSI is assumed. In such a case, an essential problem for the mobile devices is to select appropriate local processing and communication parameters. In particular, two tradeoffs are observed under a fixed total training time: (i) given the time for each communication round, the energy consumption versus the outage probability per communication round and (ii) given the energy consumption, the number of communication rounds versus the outage probability per communication round. Two optimization problems regarding the aforementioned two tradeoffs are formulated and solved. The first problem minimizes the energy consumption given the outage probability (and therefore the learning performance) requirement while the second problem optimizes the learning performance given the energy consumption requirement. Furthermore, the heterogeneous data distribution scenario is considered and a new algorithm that can deal with heterogeneous data distribution is proposed. Extensive simulations are performed to demonstrate the effectiveness of the proposed method.

Index Terms—Federated learning, wireless communications, communication efficiency, data heterogeneity.

I. INTRODUCTION

To train a machine learning model, traditional machine learning adopts a centralized approach in which the training data are aggregated on a single machine. On the one hand, such a centralized training approach is privacy-intrusive, especially when the data are collected by mobile devices and contain the owners' sensitive information (e.g., locations, user preference on websites, social media, etc.). On the other hand, transmitting all the collected data for mobile devices is impractical due to communication resource limitations. With such consideration, the concept of federated learning (FL), which enables training on a large corpus of decentralized data residing on mobile devices, is proposed in [1].

As a distributed training approach, FL adopts the parameter server paradigm in which most of the computation is offloaded to the mobile devices in a parallel manner and a parameter server is used to coordinate the training process. During each iteration, after receiving the FL model parameters from the server, the workers (i.e., mobile devices) train their local FL models using their local data and transmit the parameter updates to the server, which will aggregate the information from all the workers and send the aggregated results back. Since all the communications between the workers and the server are over wireless links, the learning performance depends on the wireless environments as well as the workers' communication resource constraints. There have been some works that study the communication aspects of FL [2]–[9]. Nonetheless, they either do not consider the existing strategies that have shown promising improvement in communication efficiency (e.g., gradient quantization [10]) or ignore the energy consumption of the workers and the impact of transmission errors as well as data heterogeneity. In addition, all these works assume perfect channel-state information (CSI) at both the server side and the worker side, which may not be reasonable in practice.

It is worth mentioning that in real-world FL applications over wireless networks, the communication time between the server and the workers is not negligible. Therefore, it becomes more critical to improve the learning performance with respect to the total training time instead of the number of rounds. With such consideration, the implementation of the FL algorithms given a fixed training time is considered in this work. In addition, the idea of SignSGD with majority vote [11] is adopted to improve the communication efficiency of the FL algorithm, in which only the signs of the parameter updates are shared between the server and the workers. The workers are assumed to transmit their parameter updates over flat-fading channels and CSI is only available at the receiver side. Channel capacity with outage is considered and each worker is supposed to determine its transmission rate and power. In such a case, the learning performance is determined by the number of communication rounds that the FL algorithm can be run and the outage probability per communication round. On the one hand, increasing the transmission power decreases the outage probability. On the other hand, a larger transmission power results in higher energy consumption. Similarly, increasing the transmission rate decreases the communication time and therefore increases the number of communication rounds given a fixed training time, while a larger transmission rate (with fixed transmission power) results in a higher outage probability per communication rounds. Such tradeoffs play essential roles

R. Jin and H. Dai are with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, USA, 27695 (e-mail: rjin2,hdai@ncsu.edu). X. He is with the Electronic Information School, Wuhan University, China (e-mail: xiaofanhe@whu.edu.cn).

in the implementation of the FL algorithm and therefore are the main focus of this work. More specifically, our main contributions are summarized as follows.

- Two tradeoffs in the implementation of SignSGD over wireless networks are identified. (i) Given the time for each communication round, the energy consumption versus the outage probability per communication round. (ii) Given the energy consumption, the number of communication rounds that the FL algorithm can be run given a fixed training time versus the outage probability at each communication round.
- Two optimization problems are formulated and solved. The first problem minimizes the energy consumption given the outage probability (and therefore the learning performance) requirement while the second problem optimizes the learning performance given the energy consumption requirement.
- In addition, the heterogeneous data distribution scenario is considered and a new algorithm that can deal with heterogeneous data distribution is proposed.
- Extensive simulations are performed to demonstrate the effectiveness of the proposed method.

The remainder of this work is organized as follows. Section III introduces the system model. Section II discusses the related works. Some analysis of the performance of SignSGD over wireless networks is provided in Section IV. The optimization problems are formulated in Section V and the corresponding solutions are presented in Section VI. Section VII extends the proposed method to the heterogeneous data distribution scenario. Section VIII presents the simulation results. Conclusions are presented in Section IX.

II. RELATED WORKS

To improve the communication efficiency of the distributed learning algorithms, various methods have been proposed, including quantization [10]–[15], sparsification [16]–[18] and subsampling [1], [19]. However, most of these works ignore the impact of wireless environments and the resource constraints of the mobile devices, which are of vital importance in the implementation of FL algorithms over real-world wireless networks.

Recently, there have been a number of existing works that study the implementation of FL algorithms over wireless networks. In [2], the weighted sum of the training time and the energy consumption is optimized by properly selecting the local computation parameters and the communication time allocated to each user. However, the formulation of the optimization problem and the proposed method rely on the assumption that the loss function is strongly convex. [3] also considers the energy consumption of the communications between the mobile devices and the server. The goal is to minimize the weighted sum of the energy consumption and the number of participated mobile devices by mobile device scheduling and effective bandwidth allocation. [4] considers a cell-free massive MIMO scenario and the training time is minimized by jointly optimizing the local computation and

the communication parameters. [5] empirically proposes a learning efficiency metric which is a function of the mini-batch size and the time of a communication round. Resource allocation and the mini-batch size are jointly optimized to maximize the learning efficiency. [6] takes the effect of packet transmission errors into consideration and analyzes its impact on the performance of FL. A joint bandwidth allocation and mobile device selection problem is formulated and solved to minimize a FL loss function that captures the performance of the FL algorithm. However, in these works, effective strategies for improving communication efficiency mentioned above are not considered. [7] adopts gradient quantization and proposes an one-bit broadband over-the-air aggregation scheme. The impact of wireless channel hostilities is analyzed. [8] and [9] propose to combine the quantization, sparsification and error compensation schemes, the energy consumption of the devices as well as the impact of transmission errors are ignored in these two works. Moreover, all these works assume CSI at both the transmitter side and the receiver side. In this work, we adopt the idea of SignSGD with majority vote [11] in the design of the communication system and consider flat-fading channels with receiver only CSI.

III. SYSTEM MODEL

In this work, a wireless multi-user system consisting of one parameter server and a set \mathcal{M} of M workers is considered. In particular, each worker m stores a local dataset \mathcal{D}_m , which will be used for local training. The local dataset can be locally generated or collected through each worker's usage of mobile devices. Considering that the training of a prediction model, especially in deep learning, usually requires a large dataset, the goal of the workers is to cooperatively learn a machine learning model while keeping the local training data at their mobile devices.

A. Machine Learning Model

A typical federated optimization problem with M normal workers is considered. Formally, the goal is to minimize a finite-sum objective of the form

$$\min_{w \in \mathbb{R}^d} F(w) \quad \text{where} \quad F(w) \stackrel{\text{def}}{=} \frac{1}{M} \sum_{m=1}^M F_m(w). \quad (1)$$

For a machine learning problem, we have a sample space $I = X \times Y$, where X is a space of feature vectors and Y is a label space. Given the hypothesis space $\mathcal{W} \subseteq \mathbb{R}^d$, we define a loss function $l : \mathcal{W} \times I \rightarrow \mathbb{R}$ which measures the loss of prediction on the data point $(x, y) \in I$ made with the hypothesis vector $w \in \mathcal{W}$. In such a case, $F_m(w)$ is a local function defined by the local dataset of worker m and the hypothesis w . More specifically,

$$F_m(w) = \frac{1}{|\mathcal{D}_m|} \sum_{(x_n, y_n) \in \mathcal{D}_m} l(w; (x_n, y_n)), \quad (2)$$

where $|\mathcal{D}_m|$ is the size of worker m 's local dataset \mathcal{D}_m . The loss function $l(w; (x_n, y_n))$ depends on the learning tasks and the machine learning models.

To accommodate the requirement of communication efficiency in FL, we adopt the popular idea of gradient quantization as in SignSGD with majority vote [11], which is presented in Algorithm 1. At t -th communication round, each worker m computes the gradient $g_m^{(t)}$ based on its locally stored model weights $w^{(t)}$ and the local datasets \mathcal{D}_m . Then, instead of transmitting the gradient $g_m^{(t)}$ directly, the worker m transmits $\text{sign}(g_m^{(t)})$ to the parameter server, in which $\text{sign}(\cdot)$ is the sign function. After receiving the shared signs of the gradients from the workers, the parameter server performs aggregation using the majority vote rule and sends the aggregated result back to the workers. Finally, the workers update their local model weights using the aggregated result.

Algorithm 1 SignSGD with majority vote

1. Input: initial weight: $w^{(0)}$; number of workers: M ; learning rate: η .
2. for $t = 0, 1, \dots, T$ do
3. Each worker m obtains its gradient $g_m^{(t)} = \nabla F_m(w^{(t)})$ and transmits $\text{sign}(g_m^{(t)})$ to the parameter server.
4. The parameter server aggregates the shared information $\hat{g}_m^{(t)}, \forall m \in \mathcal{M}$ and sends $\tilde{g}^{(t)} = \text{sign}(\frac{1}{M} \sum_{m=1}^M \hat{g}_m^{(t)})$ back to the workers.
5. The workers update their local models

$$w^{(t+1)} = w^{(t)} - \eta \tilde{g}^{(t)}. \quad (3)$$

6. end for

B. Local Computation Model

In this work, we consider a similar local computation model as those in [2] and [6]. Let c_m and f_m denote the number of CPU cycles required for worker m to process per bit data and its CPU cycle frequency, respectively, which are assumed known to the parameter server. Then, the CPU energy consumption of worker m for one local iteration of computation is given by [20]

$$E_m^{\text{cmp}} = \frac{\alpha_m}{2} c_m D_m f_m^2, \quad (4)$$

in which $\frac{\alpha_m}{2}$ is the effective capacitance coefficient of worker m 's computing chip, D_m is the size of worker m 's training data per iteration (in bits). In addition, the computation time per local iteration of worker m is given by

$$T_m^{\text{cmp}} = \frac{c_m D_m}{f_m}. \quad (5)$$

C. Transmission Model

In this work, it is assumed that the workers transmit their local updates (i.e., the signs of the gradients) to the parameter server via the orthogonal frequency division multiple access (OFDMA), and does not interfere with each other. Given that the parameter server has more power and bandwidth compared to the mobile devices, the downlink transmission

time is ignored in this work for simplicity.¹ Moreover, similar to most of the existing literature (e.g., [2], [6]), it is assumed that the downlink transmissions are error-free.

For the uplink transmission, different from the existing works that consider perfect CSI at both the transmitter side and the receiver side, we consider flat-fading channels with receiver only CSI and the capacity with outage. We assume a discrete-time channel with stationary and ergodic time-varying gain $\sqrt{h_m}$ following Rayleigh distribution, and additive white Gaussian noise (AWGN) for each worker.

Capacity with outage is defined as the maximum rate that can be transmitted over a channel with some outage probability corresponding to the probability that the transmission cannot be decoded with negligible error probability [21]. Suppose that worker m transmits at a rate of $r_m = \log_2(1 + \gamma_{\min})$, in which γ_{\min} is some fixed minimum received SNR, the data can be correctly received if the instantaneous received SNR $\gamma_m = \frac{P_m h_m}{N_0 B_m}$ is greater than or equal to γ_{\min} , in which P_m is the transmission power of worker m ; N_0 is the noise power spectral density and B_m is the corresponding bandwidth. The probability of outage is thus $p_{\text{out}} = p(\gamma_m < \gamma_{\min})$. Particularly, for Rayleigh fading channel, we have

$$p_{\text{out}}(r_m) = 1 - e^{-\frac{(2^{r_m} - 1) N_0 B_m}{P_m}}. \quad (6)$$

The corresponding communication time and energy consumption are given by

$$T_m^{\text{com}} = \frac{s_m}{r_m B_m}, \quad E_m^{\text{com}} = \frac{P_m s_m}{r_m B_m} \quad (7)$$

in which s_m is the size of the transmitted data.²

For simplicity, it is assumed that for worker m , $\text{sign}(g_m^{(t)})$ is transmitted as a single packet in the uplink and the whole packet is decoded incorrectly when an outage happens.

IV. ANALYSIS OF THE PERFORMANCE OF ALGORITHM 1 OVER WIRELESS NETWORKS

Before diving into the details of the system design, we first analyze how the characteristics of wireless networks affect the performance of Algorithm 1. To facilitate the analysis, the following commonly adopted assumptions are made.

Assumption 1. (Lower bound). For all w and some constant F^* , we have objective value $F(w) \geq F^*$.

Assumption 2. (Smoothness). $\forall w_1, w_2$, we require for some non-negative constant L

$$F(w_1) \leq F(w_2) + \langle \nabla F(w_2), w_1 - w_2 \rangle + \frac{L}{2} \|w_1 - w_2\|_2^2, \quad (8)$$

¹Note that for a fixed transmission rate, the downlink transmission time is a constant which can be readily integrated to the first and the second constraints of the optimization problems (14) and (15), respectively, if needed.

²Note that in the schemes where full precision gradients are transmitted, each worker is supposed to transmit 32 bits for each element in the gradient vectors. However, Algorithm 1 only requires 1 bit by transmitting the signs and therefore leads to an improvement of 32 times in communication time as well as communication energy consumption. In addition, s_m also depends on the machine learning model. For instance, in a softmax regression model for k -class classification tasks, $s_m = d \times k$.

where $\langle \cdot, \cdot \rangle$ is the standard inner product.

Given the above assumptions, the following result can be proved.

Theorem 1. Suppose that the model parameter at the beginning of t -th iteration is $w^{(t)}$, then by performing one iteration of Algorithm 1, we have

$$\mathbb{E}[F(w^{(t+1)})] \leq F(w^{(t)}) + \eta \|\nabla F(w^{(t)})\|_1 + \frac{L\eta^2 d}{2} - 2\eta \times \sum_{i=1}^d |\nabla F(w^{(t)})_i| P\left(\text{sign}\left(\sum_{m=1}^M \hat{g}_m^{(t)}\right)_i = \text{sign}(\nabla F(w^{(t)}))_i\right), \quad (9)$$

in which d is the dimension of the gradients; $\nabla F(w^{(t)})_i$ is the i -th entry of the gradient vector $\nabla F(w^{(t)})$ and $\text{sign}(\cdot)_i$ is the i -th entry of the vector after taking the sign operation. The expectation and the probability are over the dynamics of the wireless channels.

Proof. Please see Appendix A. \square

Note that given fixed $w^{(t)}$, the right-hand side of (9) depends on the probability of the signs of the aggregation result being the same as those of the true gradients (i.e., $P(\text{sign}(\sum_{m=1}^M \hat{g}_m^{(t)})_i = \text{sign}(\nabla F(w^{(t)}))_i)$). For the ease of discussion, we consider the i -th entry of the gradients and define a series of random variables $\{X_m\}_{m=1}^M$ given by

$$X_m = \begin{cases} 1, & \text{if } \text{sign}(\hat{g}_m^{(t)})_i \neq \text{sign}(\nabla F(w^{(t)}))_i, \\ 0, & \text{if } \text{sign}(\hat{g}_m^{(t)})_i = \text{sign}(\nabla F(w^{(t)}))_i. \end{cases} \quad (10)$$

X_m can be considered as the outcome of one Bernoulli trial with successful probability $P(X_m = 1)$. Let $Z = \sum_{m=1}^M X_m$, then it can be verified that

$$P\left(\text{sign}\left(\sum_{m=1}^M \hat{g}_m^{(t)}\right)_i = \text{sign}(\nabla F(w^{(t)}))_i\right) = P\left(Z < \frac{M}{2}\right). \quad (11)$$

In addition, Z follows the Poisson binomial distribution with mean $\mathbb{E}[Z] = \sum_{m=1}^M P(X_m = 1)$. Since Z is non-negative, the Markov's inequality gives

$$P(Z \geq M/2) \leq \frac{2\mathbb{E}[Z]}{M}, \quad (12)$$

and therefore

$$P(Z < M/2) = 1 - P(Z \geq M/2) \geq \frac{M - 2\mathbb{E}[Z]}{M}. \quad (13)$$

Note that $\mathbb{E}[Z]$ and $M - \mathbb{E}[Z]$ are the expected number of workers that share wrong and correct signs, respectively. The lower bound in (13) represents the difference between the ratios of workers that share the correct signs and that share the wrong signs.

In particular, let $p_m^{(t)}$ denote the probability of $\text{sign}(\hat{g}_m^{(t)})_i = \text{sign}(\nabla F(w^{(t)}))_i$ (i.e., the i -th entry of the gradient of worker m has the same sign as that of the true gradients $\nabla F(w^{(t)})$), then $P(X_m = 1) = p_m^{(t)} p_{out}(r_m) + (1 - p_m^{(t)})(1 - p_{out}(r_m))$.

When $p_m^{(t)} > 0.5$, minimizing $P(X_m = 1)$ is equivalent to minimizing $p_{out}(r_m)$. To this end, two tradeoffs can be observed. Firstly, it can be observed from (6) that given fixed bandwidth B_m and noise N_0 , the transmission rate r_m and transmission power P_m determine the outage probability. Increasing the transmission power P_m and decreasing the transmission rate r_m both decrease the outage probability. However, according to (7), a larger P_m and a smaller r_m result in higher communication energy consumption. In addition, given fixed time for each communication round (i.e., $T_m^{cmp} + T_m^{com}$), decreasing r_m increases the communication time T_m^{com} and therefore requires worker m to increasing the CPU frequency such that the local computation time can be reduced. As a result, the CPU energy consumption of worker m also increases. Therefore, there exists a tradeoff between the energy consumption of the workers and the learning performance. Secondly, given a fixed total training time, transmission power P_m and CPU frequency f_m , despite that decreasing the transmission rate r_m can decrease the outage probability during each iteration, it also increases the time for each communication round and therefore decreases the number of iterations that the FL algorithm can be run. Therefore, another tradeoff between the number of iterations and the outage probability per iteration can be observed.

V. PROBLEM FORMULATION

A. Energy Minimization Given Learning Performance Constraint

In order to obtain the tradeoff between the energy consumption of the workers and the learning performance, it is essential to know the minimum energy consumption of each worker to guarantee certain learning performance. Particularly, the learning performance is mainly determined by two parameters: the outage probability $p_{out,m}$ at each iteration and the number of total iterations that is inversely proportional to the time per communication round (denoted by T_l). Given $p_{out,m}$ and T_l , the goal of worker m is to minimize its energy consumption. The corresponding optimization problem is formulated as follows.

$$\begin{aligned} \min_{f_m, r_m, P_m} & \frac{\alpha_m}{2} c_m D_m f_m^2 + \frac{P_m s_m}{r_m B_m} \\ \text{s.t.} & \frac{c_m D_m}{f_m} + \frac{s_m}{r_m B_m} \leq T_l, \\ & P_{min,m} \leq P_m \leq P_{max,m}, \\ & f_{min,m} \leq f_m \leq f_{max,m}, \\ & 1 - e^{-\frac{(2^r m - 1) N_0 B_m}{P_m}} \leq p_{out,m}, \end{aligned} \quad (14)$$

in which the CPU frequency for local computation f_m , the transmission rate r_m and the transmission power P_m are the parameters to be optimized. The first constraint captures the delay requirement per communication round. The feasible regions of CPU frequency and transmission power of worker m are imposed by the second and the third constraints, respectively. The last constraint restricts the feasible range of the outage probability.

B. Learning Performance Optimization Given Energy Consumption Constraint

Recall that the performance of the FL algorithm mainly depends on the outage probability at each iteration and the total number of iterations. According to the discussion in Section IV, given a fixed total training time, transmission power P_m and CPU frequency f_m , minimizing the outage probability per iteration and maximizing the number of iterations are conflicting. Therefore, in this subsection, the goal is to find a good tradeoff such that the learning performance is optimized.

Note that signSGD converges with a rate of $O(\frac{1}{\sqrt{T}})$ [11], in which T is the total number of iterations. Therefore, in this work, the objective is to maximize $\sqrt{T}(M - 2\mathbb{E}[Z])/M$, in which \sqrt{T} captures the impact of the number of iterations and $(M - 2\mathbb{E}[Z])/M$ captures the learning performance improvement at each iteration (c.f. (13)). In addition, according to the discussion in Section III, $\mathbb{E}[Z] = \sum_{m=1}^M p_m^{(t)} p_{out}(r_m) + (1 - p_m^{(t)})(1 - p_{out}(r_m))$, in which $p_m^{(t)}$ is determined by the local dataset of worker m and therefore unknown to the server. To facilitate the discussion, we assume that $p_m^{(t)} = 1, \forall m, t$.³ Given fixed total training time, since the number of iterations is inversely proportional to the time consumption of each iteration, the optimization problem is formulated as follows.

$$\begin{aligned} & \max_{T_l, r_m} \frac{M - 2 \sum_{m=1}^M p_{out}(r_m)}{\sqrt{T_l}} \\ \text{s.t. } & \frac{\alpha_m}{2} c_m D_m f_m^2 + \frac{P_m s_m}{r_m B_m} \leq E_m, \forall m, \\ & \max_m \left\{ \frac{c_m D_m}{f_m} + \frac{s_m}{r_m B_m} \right\} \leq T_l, \end{aligned} \quad (15)$$

in which the communication round time T_l and the transmission rate r_m are the parameters to be optimized. The first constraint captures the energy consumption requirement for each worker m and the second constraint captures the delay requirement for each iteration.

Furthermore, we assume that the workers transmit with high SNR and therefore we have

$$p_{out}(r_m) \approx \frac{(2^{r_m} - 1) N_0 B_m}{P_m}. \quad (16)$$

VI. OPTIMIZATION OF SYSTEM PARAMETERS FOR FEDERATED LEARNING

A. Energy Minimization Given Outage Probability Constraint

We note that the optimization problem (14) is not always feasible. In particular, according to the delay requirement T_l , it is required that $r_m \geq \frac{s_m}{(T_l - \frac{c_m D_m}{f_{max,m}}) B_m}$. Combining it with the power constraint and plugging them into (6) yields

$$p_{out}(r_m) \geq 1 - e^{-\frac{\frac{s_m}{(T_l - \frac{c_m D_m}{f_{max,m}}) B_m} - 1}{\frac{P_{max,m}}{N_0 B_m}}}, \quad (17)$$

³Note that when all the workers have the same dataset, $g_m^{(t)} = \nabla F(w^{(t)}), \forall m$. In the homogeneous data distribution setting, $g_m^{(t)}$ can be considered as a noisy version of $\nabla F(w^{(t)})$. As long as the noise is not too large (e.g., when the local datasets are large enough), this assumption is approximately true. This is verified in our simulation results.

which may contradict the last condition in (14). Therefore, two scenarios are considered.

1) *The optimization problem (14) is infeasible:* In this case, we assume that $P_m = P_{max,m}$, $f_m = f_{max,m}$ and $r_m = \frac{s_m}{(T_l - \frac{c_m D_m}{f_{max,m}}) B_m}$.

Remark 1. We note that T_l and $p_{out}(r_m)$ are the two most important parameters that determine the performance of the FL algorithm. When the optimization problem (14) is infeasible, the delay requirement and the outage probability requirement cannot be satisfied simultaneously for worker m . Since the communication round time is supposed to be determined by the slowest worker (the straggler), we assume that each worker tries its best to reduce its outage probability while accommodating the delay requirement.

2) *The optimization problem (14) is feasible:* For the ease of presentation, we define $r_m^{(1)} = \log_2 \left(-\frac{P_{min,m} \ln(1 - p_{out,m})}{N_0 B_m} + 1 \right)$, $r_m^{(2)} = \log_2 \left(-\frac{P_{max,m} \ln(1 - p_{out,m})}{N_0 B_m} + 1 \right)$, and $r_m^{(3)} = \frac{s_m}{B_m (T_l - \frac{c_m D_m}{f_{max,m}})}$.

Lemma 1. Given any $r_m^{(1)} \leq r_m \leq r_m^{(2)}$, the optimal transmission power P_m^* is given by

$$P_m^* = -\frac{N_0 B_m (2^{r_m} - 1)}{\ln(1 - p_{out,m})}. \quad (18)$$

Given any $r_m \geq r_m^{(3)}$, the optimal CPU frequency for local computation is given by

$$f_m^* = \max \left\{ \frac{c_m D_m}{T_l - \frac{s_m}{r_m B_m}}, f_{min,m} \right\}. \quad (19)$$

Proof. Please see Appendix B. \square

With Lemma 1 at hand, the optimization problem (14) can be reformulated as follows.

$$\begin{aligned} & \min_{r_m} \frac{\alpha_m c_m D_m}{2} z_m^2(r_m) - \frac{N_0 s_m (2^{r_m} - 1)}{\ln(1 - p_{out,m}) r_m} \\ \text{s.t. } & \max\{r_m^{(1)}, r_m^{(3)}\} \leq r_m \leq r_m^{(2)}, \end{aligned} \quad (20)$$

in which $z_m(r_m) = \max\left\{ \frac{c_m D_m}{T_l - \frac{s_m}{r_m B_m}}, f_{min,m} \right\}$.

It can be verified that the objective in (20) is convex and therefore, the widely used subgradient methods [22] can be adopted to solve the optimization problem (20).

B. Learning Performance Optimization Given Energy Consumption Constraint

Lemma 2. In the optimization problem (15), given any fixed T_l , the optimal transmission rate of worker m is given by

$$r_m^* = \max \left\{ \frac{P_m s_m}{B_m (E_m - \frac{\alpha_m}{2} c_m D_m f_m^2)}, \frac{s_m f_m}{B_m f_m T_l - B_m c_m D_m} \right\}. \quad (21)$$

Proof. Please see Appendix C. \square

Let $\mathcal{U} = \{m | \frac{P_m s_m}{B_m (E_m - \frac{\alpha_m}{2} c_m D_m f_m^2)} \geq \frac{s_m f_m}{B_m f_m T_l - B_m c_m D_m}\}$. According to Lemma 2, the workers can be divided into two groups. The optimal transmission rates of the workers in the

first group (i.e., \mathcal{U}) is limited by their energy consumption upper limit E_m while those of the workers in the second group is limited by the communication round time T_l which is subject to design. Further define the following two functions.

$$g(x) = \frac{2 \sum_{m \in \mathcal{U}} \left(2^{\frac{P_m s_m}{B_m (E_m - \frac{\alpha_m}{2} c_m D_m f_m^2)} - 1} \right) N_0 B_m}{P_m \sqrt{x}} \quad (22)$$

$$+ \frac{2 \sum_{m \notin \mathcal{U}} \left(2^{\frac{s_m f_m}{B_m c_m D_m} - 1} \right) N_0 B_m}{P_m \sqrt{x}},$$

$$h(x) = \frac{M}{\sqrt{x}}. \quad (23)$$

Based on Lemma 2, the optimization problem (15) can be reformulated as follows.

$$\begin{aligned} \min_{T_l} \quad & g(T_l) - h(T_l) \\ \text{s.t.} \quad & T_l \geq \max_m \left\{ \frac{c_m D_m}{f_m} \right\}. \end{aligned} \quad (24)$$

It can be verified that both $g(x)$ and $h(x)$ are convex functions of x . Therefore, (24) is a difference of convex programming problem, which can be solved by the DCA algorithm [23].

VII. EXTENDING TO THE HETEROGENEOUS DATA DISTRIBUTION SCENARIO

In the previous discussion, it is assumed that $p_m^{(t)} > 0.5$ and therefore minimizing $P(X_m = 1)$ is equivalent to minimizing $p_{out}(r_m)$. This assumption holds with a high probability in the homogeneous data distribution scenario. Nonetheless, in the heterogeneous data distribution scenario, such an assumption may not hold.

Example 1. Suppose that the i -th coordinate of worker m 's gradient is given by

$$\nabla F_m(w^{(t)})_i = \begin{cases} -1, & \text{if } 1 \leq m \leq M-1, \\ M, & \text{if } m = M. \end{cases} \quad (25)$$

It can be easily verified that

$$\text{sign} \left(\sum_{m=1}^M \nabla F_m(w^{(t)}) \right)_i \neq \text{sign}(\nabla F(w^{(t)}))_i, \quad (26)$$

which leads to wrong aggregation.

Remark 2. In the homogeneous data distribution scenario, the local datasets (and therefore the gradients) of the workers are drawn from the same distribution. As a result, the probability of wrong aggregation as in Example 1 is small. In the ideal scenario in which all the workers have the same local dataset, all the workers have the same gradient and therefore the probability of wrong aggregation is 0.

Nonetheless, in the heterogeneous data distribution scenario, the probability of wrong aggregation as in Example 1 depends on the data distribution of the workers. For instance, if $F_m(x) = x^2 + 2, \forall 1 \leq m \leq M-1$ and $F_M(x) = x^2 - 2M$,

the probability of wrong aggregation is 1, which prevents the convergence of Algorithm 1.

With such consideration, Algorithm 2 is proposed in this work. In particular, compared to Algorithm 1, there is a pre-processing step (i.e., step 3) in Algorithm 2. Taking $M = 3$ in Example 1 as an example, it can be shown that

$$P(X_m = 1) = \begin{cases} \frac{1}{2} + b, & \text{if } 1 \leq m \leq 2, \\ \frac{1}{2} - 3b, & \text{if } m = 3. \end{cases} \quad (27)$$

Therefore,

$$\begin{aligned} P\left(Z < \frac{3}{2}\right) &= P\left(\sum_{m=1}^3 X_m = 1\right) + P\left(\sum_{m=1}^3 X_m = 0\right) \\ &= 2\left(\frac{1}{2} + b\right)\left(\frac{1}{2} - b\right)\left(\frac{1}{2} + 3b\right) + \left(\frac{1}{2} - b\right)^2\left(\frac{1}{2} - 3b\right) \\ &\quad + \left(\frac{1}{2} - b\right)^2\left(\frac{1}{2} + 3b\right) \\ &= \frac{1}{2} + \frac{1}{2}b - 6b^3. \end{aligned} \quad (28)$$

It can be verified that when $0 \leq b \leq \frac{1}{\sqrt{12}}$, $P(Z < 3/2) > \frac{1}{2}$. That being said, the probability of correct aggregation is strictly larger than $\frac{1}{2}$ when b is small enough. For more general scenarios where $b \leq \frac{1-2p_{out}(r_m)}{2|\nabla F_m(w^{(t)})_i|}$, the following Lemma 2 can be proved.

Algorithm 2 SIGNSGD with majority vote over wireless networks

1. Input: initial weight: w_0 ; number of workers: M ; learning rate: η ; the outage probability of worker m : $p_{out}(r_m)$; some positive constant: b .
2. for $t = 0, 1, \dots, T$ do
3. Each worker m obtains its gradient $\nabla F_m(w^{(t)})$ and does the following pre-processing

$$(g_m^{(t)})_i = \begin{cases} -\text{sign}(\nabla F_m(w^{(t)}))_i, & \text{with probability } p_m^i, \\ \text{sign}(\nabla F_m(w^{(t)}))_i, & \text{with probability } 1 - p_m^i, \end{cases} \quad (29)$$

where $p_m^i = \frac{\frac{1}{2} - p_{out}(r_m) - b|\nabla F_m(w^{(t)})_i|}{1 - 2p_{out}(r_m)}$; $b \leq \frac{1 - 2p_{out}(r_m)}{2|\nabla F_m(w^{(t)})_i|}$.

4. Each worker m obtains its gradient $g_m^{(t)} = \nabla F_m(w^{(t)})$ and transmits $\text{sign}(g_m^{(t)})$ to the parameter server.
5. The parameter server aggregates the shared information $\hat{g}_m^{(t)}, \forall m \in \mathcal{M}$ and sends $\tilde{g}^{(t)} = \text{sign}(\frac{1}{M} \sum_{m=1}^M \hat{g}_m^{(t)})$ back to the workers.
6. The workers update their local models

$$w^{(t+1)} = w^{(t)} - \eta \tilde{g}^{(t)}. \quad (30)$$

7. end for

Theorem 2. At each iteration, there exists a constant b such that when $p_{out}(r_m) \leq \min_i \{\frac{1}{2} - b|\nabla F_m(w^{(t)})_i|\}$, in which

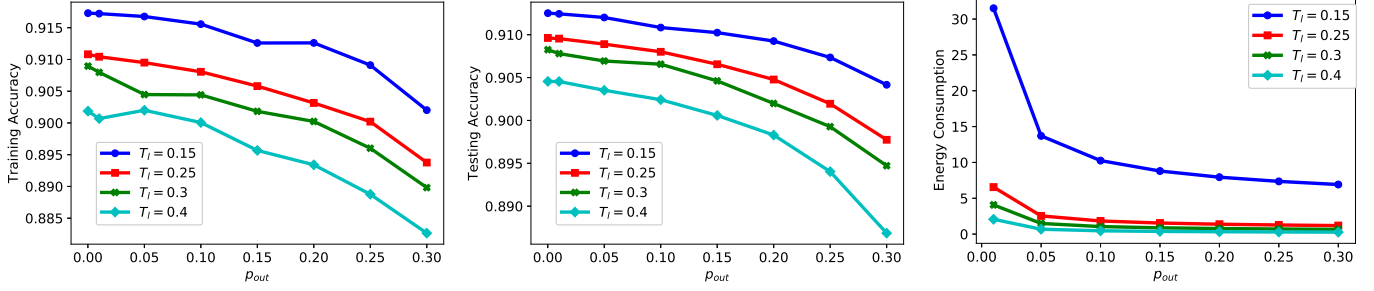


Fig. 1: The Impact of Outage Probability in the Homogeneous Data Distribution Scenario

$\nabla F_m(w^{(t)})_i$ is the i -th entry of the gradient $\nabla F_m(w^{(t)})$, the expected number of workers that share the wrong signs is given by

$$\mathbb{E}[Z] = \sum_{m=1}^M P(X_m = 1) = \frac{M}{2} - bM|\nabla F(w^{(t)})_i|. \quad (31)$$

Proof. Please see Appendix D. \square

Remark 3. Theorem 2 indicates that for any $|\nabla F(w^{(t)})_i| > 0$, $\mathbb{E}[Z] < \frac{M}{2}$. As we discussed in Section IV, Z follows the Poisson binomial distribution. Lyapunov Central Limit Theorem tells us that the distribution of Z can be approximated by a normal distribution with mean $\mathbb{E}[Z]$ when the number of workers is large enough (usually in the order of tens). Therefore, it can be easily obtained that $P(Z < \frac{M}{2}) > \frac{1}{2}$, based on which the convergence of Algorithm 2 can be proved [24].

Intuitively, according to Theorem 2, the performance of Algorithm 2 depends on both b and $\nabla F(w^{(t)})$. However, since $\nabla F(w^{(t)})$ is unknown to the server, optimizing the learning performance of Algorithm 2 is highly non-trivial and left as our future work. In this work, we mainly consider the energy minimization problem given fixed b and $p_{out}(r_m)$. In this case, by setting $p_{out,m} = p_{out}(r_m)$, the energy consumption can be minimized by solving (14).

Remark 4. It can be observed from (31) that the expected number of workers that share wrong signs is independent of $p_{out}(r_m)$. In the meantime, according to (14), the feasible region of the energy consumption minimization problem with a smaller $p_{out}(r_m)$ is a subset of that with a larger $p_{out}(r_m)$. As a result, it is optimal to select $p_{out}(r_m) = \min_i \{\frac{1}{2} - b|\nabla F_m(w^{(t)})_i|\}$ at communication round t .

VIII. SIMULATION RESULTS

In this section, we examine the performance of the proposed methods through extensive simulations. We implement a softmax regression model on the well-known MNIST dataset that consists of 10 categories ranging from digit “0” to “9” and a total of 60,000 training samples and 10,000 testing samples. Therefore, the size of updates for each worker is $s_m = 7850$ bits per communication round. It is assumed that there are 10 workers that collaboratively train a global model given a

total training time of 50 seconds. For all the workers, we set $\alpha_m = 2 \times 10^{-28}$; $c_m = 20$ cycles/bit; $D_m = 5 \times 10^6$ bits; $f_{min,m} = 0.3$ GHz; $f_{max,m} = 2$ GHz; $s_m = 7850$ bits; $P_{min,m} = 0$; $P_{max,m} = 1$ W; $N_0 = 10^{-8}$ W/Hz; $B_m = 15$ kHz. In the homogeneous data distribution scenario, each worker randomly samples 2000 training samples from the training dataset. In the heterogeneous data distribution scenario, the whole training dataset is divided into 10 subsets, each containing the training data for one label. Each worker randomly samples 2000 training samples from one of the subsets.

A. Energy Minimization Given Learning Performance Constraint in the Homogeneous Data Distribution Scenario

In this subsection, the impact of the outage probability and communication round time in (14) is examined. We set the same outage probability constraints for all the workers, i.e., $p_{out,m} = p_{out}, \forall m$. The three figures in Fig. 1 shows the training accuracy, testing accuracy and the per worker energy consumption of Algorithm 1 with different p_{out} and T_l , respectively. It can be observed that as p_{out} and T_l increase, the energy consumption decreases. This is because the feasible region of (14) corresponding to a smaller p_{out} and T_l is a subset of that of (14) corresponding to a larger p_{out} and T_l . On the other hand, both the training accuracy and the testing accuracy decrease as p_{out} and T_l increase, which validates the existence of the tradeoff between the energy consumption and the learning performance.

B. Learning Performance Optimization Given Energy Consumption Constraint in the Homogeneous Data Distribution Scenario

In this subsection, we examine the impact of the transmission power P_m and the communication round time T_l . The energy consumption upper limit is set as $E_m = 100$ J. Fig. 2 shows the performance of Algorithm 1 with different P_m and T_l . For the solid curves, the transmission rate r_m ’s are given by (21) while the configurations of the marked points are given by the solution of (15). It can be shown that as T_l increases, the learning performance of Algorithm 1 first increases and then decreases. According to (21), when T_l increases, r_m decreases and therefore the outage probability $p_{out,m}$ also decreases. However, in the meantime, as T_l increases, the

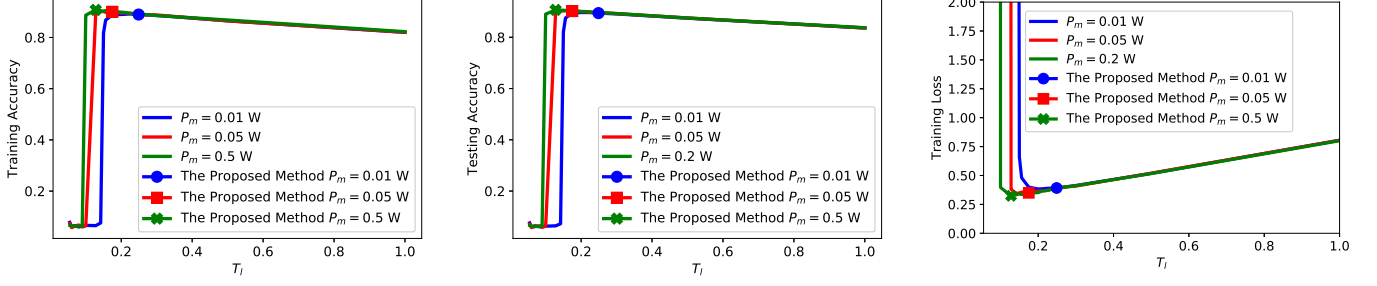


Fig. 2: The Impact of T_l in the Homogeneous Data Distribution Scenario

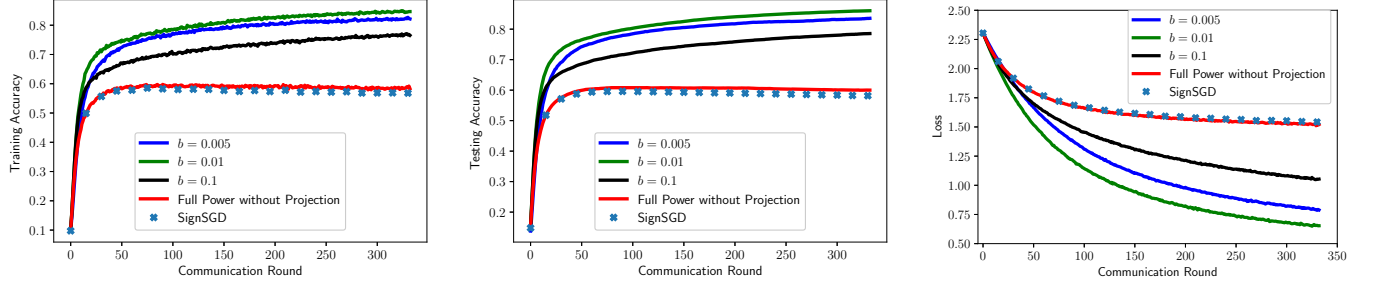


Fig. 3: The Performance of Algorithm 2 in the Heterogeneous Data Distribution Scenario

TABLE I: Average Energy Consumption of the Workers

b	0.005	0.01	0.1	Full Power without Projection
Energy Consumption (J)	25.05	42.24	46.62	46.62

number of communication rounds decreases given the fixed training time. As a result, when the outage probability has a larger impact on the learning performance, increasing T_l results in better performance. When T_l is larger than a certain critical value, the number of communication rounds plays a more important role and therefore increasing T_l leads to worse performance. In addition, such a critical value decreases as the transmission power increases. This is because for the same critical $p_{out,m}$, a larger P_m corresponds to a larger r_m and therefore a smaller T_l . Furthermore, it can be observed from Fig. 2 that the proposed method works close to the optimal operation point for all the examined scenarios, which validates its effectiveness.

C. Energy Minimization Given Learning Performance Constraint in the Heterogeneous Data Distribution Scenario

In this subsection, the performance of Algorithm 2 is examined. The outage probability $p_{out,m}$'s are set according to Theorem 2. Fig. 3 shows the performance of Algorithm 2 for different b when $T_l = 0.15$. In the “Full Power without Projection” counterpart, we show the performance of Algorithm 1 and set $P_m = P_{max,m}$ and $f_m = f_{max,m}$, and $r_m = s_m / (T_l B_m - c_m D_m B_m / f_m)$, i.e., the outage probability $p_{out,m}$ is minimized given that the communication round time

T_l is satisfied. For the “SignSGD” baseline, it is assumed that the communication between the workers and the parameter server is perfect (i.e., the outage probabilities are zero). It can be observed that Algorithm 2 outperforms the “Full Power without Projection” and “SignSGD” counterparts for all the examined b 's. More specifically, when $b = 0.01$, Stochastic-Sign SGD gives an improvement of around 30% in testing accuracy.

Table I shows the corresponding average energy consumption of the workers. It can be observed that when $b = 0.1$, the energy consumption of Algorithm 2 is the same as that of “Full Power without Projection”. In this case, the outage probability requirement and the communication round time requirement cannot be satisfied simultaneously. Therefore, the workers are operating with $P_m = P_{max,m}$ and $f_m = f_{max,m}$. In this case, the only difference between Algorithm 2 and “Full Power without Projection” is the pre-processing step (i.e., step 3) in Algorithm 2. This indicates that the pre-processing step along gives an improvement of around 20% in testing accuracy. Moreover, Table. I shows that the average energy consumption increases as b increases. This is because, as b increases, the outage probability $p_{out,m}$ decreases. As a result, similar to the results in the homogeneous data distribution scenario, the average energy consumption of the workers increases. However, it can be observed that different from the homogeneous data distribution scenario, increasing b (and therefore decreasing the outage probability) does not necessarily improve the learning performance in the heterogeneous data distribution scenario. For instance, Algorithm 2 with $b = 0.01$ performs around 10% better than $b = 0.1$ in testing accuracy. This indicates

that compared with Algorithm 1 and “SignSGD”, Algorithm 2 improves the learning performance while saves energy by selecting an appropriate b .

IX. CONCLUSIONS

In this work, the implementation of FL algorithms over wireless networks is studied. In particular, the tradeoff between the energy consumption and the learning performance and the tradeoff between the number of iterations that the FL algorithm can be run given a fixed training time and the outage probability at each communication round are identified. Two optimization problems are formulated and solved for appropriate local processing and communication parameter configuration, each corresponding to one tradeoff. Furthermore, since SignSGD fails to converge in the heterogeneous data distribution scenario, a new FL algorithm that can deal with data heterogeneity across workers is proposed and the corresponding energy minimization problem is solved. It is shown that the proposed algorithm improves the learning performance with less energy consumption for the workers. The simulation results demonstrate the effectiveness of the proposed method.

APPENDIX A PROOF OF THEOREM 1

Proof. According to Assumption 2, we have

$$\begin{aligned}
& F(w^{(t+1)}) - F(w^{(t)}) \\
& \leq \langle \nabla F(w^{(t)}), w^{(t+1)} - w^{(t)} \rangle + \frac{L}{2} \|w^{(t+1)} - w^{(t)}\|^2 \\
& = -\eta \langle \nabla F(w^{(t)}), \text{sign}(\sum_{m=1}^M \hat{g}_m^{(t)}) \rangle + \frac{L}{2} \|\eta \text{sign}(\sum_{m=1}^M \hat{g}_m^{(t)})\|^2 \\
& = -\eta \langle \nabla F(w^{(t)}), \text{sign}(\sum_{m=1}^M \hat{g}_m^{(t)}) \rangle + \frac{L\eta^2 d}{2} \\
& = \eta \|\nabla F(w^{(t)})\|_1 + \frac{L\eta^2 d}{2} - 2\eta \sum_{i=1}^d |\nabla F(w^{(t)})_i| \times \\
& \quad \mathbb{1}_{\text{sign}(\sum_{m=1}^M \hat{g}_m^{(t)})_i = \text{sign}(\nabla F(w^{(t)})_i)},
\end{aligned} \tag{32}$$

in which $\nabla F(w^{(t)})_i$ is the i -th entry of the vector $\nabla F(w^{(t)})$. Taking expectation on both sides yields

$$\begin{aligned}
\mathbb{E}[F(w^{(t+1)})] & \leq F(w^{(t)}) + \eta \|\nabla F(w^{(t)})\|_1 + \frac{L\eta^2 d}{2} \\
& \quad - 2\eta \sum_{i=1}^d |\nabla F(w^{(t)})_i| P(\text{sign}(\sum_{m=1}^M \hat{g}_m^{(t)})_i = \text{sign}(\nabla F(w^{(t)})_i)).
\end{aligned} \tag{33}$$

□

APPENDIX B PROOF OF LEMMA 1

Proof. According to the constraint

$$1 - e^{-\frac{(2^{r_m} - 1)N_0 B_m}{P_m}} \leq p_{out,m}, \tag{34}$$

it can be obtained that

$$P_m \geq -\frac{N_0 B_m (2^{r_m} - 1)}{\ln(1 - p_{out,m})}. \tag{35}$$

Since the objective function $\frac{\alpha_m}{2} c_m D_m f_m^2 + \frac{P_m s_m}{r_m B_m}$ is an increasing function of P_m , we have

$$P_m^* = -\frac{N_0 B_m (2^{r_m} - 1)}{\ln(1 - p_{out,m})}. \tag{36}$$

According to the constraint

$$\frac{c_m D_m}{f_m} + \frac{s_m}{r_m B_m} \leq T_l, \tag{37}$$

we have

$$f_m \geq \frac{c_m D_m}{T_l - \frac{s_m}{r_m B_m}}. \tag{38}$$

In addition, the objective function is an increasing function of f_m . Therefore,

$$f_m^* = \max \left\{ \frac{c_m D_m}{T_l - \frac{s_m}{r_m B_m}}, f_{min,m} \right\} \tag{39}$$

□

APPENDIX C PROOF OF LEMMA 2

Proof. According to the constraint

$$\frac{\alpha_m}{2} c_m D_m f_m^2 + \frac{P_m s_m}{r_m B_m} \leq E_m, \tag{40}$$

we have

$$r_m \geq \frac{P_m s_m}{B_m (E_m - \frac{\alpha_m}{2} c_m D_m f_m^2)}. \tag{41}$$

According to the constraint

$$\frac{c_m D_m}{f_m} + \frac{s_m}{r_m B_m} \leq T_l, \tag{42}$$

we have

$$r_m \geq \frac{s_m f_m}{B_m f_m T_l - B_m c_m D_m}. \tag{43}$$

In addition, it can be shown that the objective function $\frac{M-2 \sum_{m=1}^M p_{out}(r_m)}{\sqrt{T_l}}$ is a decreasing function of r_m . Therefore,

$$r_m^* = \max \left\{ \frac{P_m s_m}{B_m (E_m - \frac{\alpha_m}{2} c_m D_m f_m^2)}, \frac{s_m f_m}{B_m f_m T_l - B_m c_m D_m} \right\}. \tag{44}$$

□

APPENDIX D
PROOF OF LEMMA 2

Proof. First of all, we further define a series of random variables $\{\hat{X}_m\}_{m=1}^M$ given by

$$\hat{X}_m = \begin{cases} 1, & \text{if } \text{sign}(\hat{g}_m^{(t)})_i \neq \text{sign}(\nabla F_m(w^{(t)}))_i, \\ 0, & \text{if } \text{sign}(\hat{g}_m^{(t)})_i = \text{sign}(\nabla F_m(w^{(t)}))_i. \end{cases} \quad (45)$$

It can be verified that

$$\begin{aligned} P(\hat{X}_m = 1) &= p_m^i p_{out}(r_m) + (1 - p_m^i)(1 - p_{out}(r_m)) \\ &= \frac{1}{2} - b|\nabla F_m(w^{(t)})_i|. \end{aligned} \quad (46)$$

Then we consider the following scenarios:

A. Scenario 1: $\text{sign}(\nabla F(w^{(t)})) = 1$.

In this case, according to the definition of X_m given by (10),

$$\begin{aligned} P(X_m = 1) &= P(\hat{X}_m = 1)\mathbb{1}_{\nabla F_m(w^{(t)})_i > 0} \\ &\quad + P(\hat{X}_m = 0)\mathbb{1}_{\nabla F_m(w^{(t)})_i < 0} \\ &= \frac{1}{2} - b\nabla F_m(w^{(t)})_i. \end{aligned} \quad (47)$$

Therefore,

$$\begin{aligned} \sum_{m=1}^M P(X_m = 1) &= \frac{M}{2} - b \sum_{m=1}^M \nabla F_m(w^{(t)})_i \\ &= \frac{M}{2} - bM \nabla F(w^{(t)})_i. \end{aligned} \quad (48)$$

B. Scenario 2: $\text{sign}(\nabla F(w^{(t)})) = -1$.

In this case, according to the definition of X_m given by (10),

$$\begin{aligned} P(X_m = 1) &= P(\hat{X}_m = 1)\mathbb{1}_{\nabla F_m(w^{(t)})_i < 0} \\ &\quad + P(\hat{X}_m = 0)\mathbb{1}_{\nabla F_m(w^{(t)})_i > 0} \\ &= \frac{1}{2} + b\nabla F_m(w^{(t)})_i. \end{aligned} \quad (49)$$

Therefore,

$$\begin{aligned} \sum_{m=1}^M P(X_m = 1) &= \frac{M}{2} + b \sum_{m=1}^M \nabla F_m(w^{(t)})_i \\ &= \frac{M}{2} + bM \nabla F(w^{(t)})_i. \end{aligned} \quad (50)$$

Combining the above two scenarios, it can be verified that

$$\sum_{m=1}^M P(X_m = 1) = \frac{M}{2} - bM|\nabla F(w^{(t)})_i|, \quad (51)$$

which completes the proof. \square

REFERENCES

- [1] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.
- [2] N. H. Tran, W. Bao, A. Zomaya, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *INFOCOM*. IEEE, 2019, pp. 1387–1395.
- [3] Q. Zeng, Y. Du, K. K. Leung, and K. Huang, "Energy-efficient radio resource allocation for federated edge learning," *arXiv preprint arXiv:1907.06040*, 2019.
- [4] T. T. Vu, D. T. Ngo, N. H. Tran, H. Q. Ngo, M. N. Dao, and R. H. Middleton, "Cell-free massive MIMO for wireless federated learning," *arXiv preprint arXiv:1909.12567*, 2019.
- [5] J. Ren, G. Yu, and G. Ding, "Accelerating DNN training in wireless federated edge learning system," *arXiv preprint arXiv:1905.09712*, 2019.
- [6] 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," *arXiv preprint arXiv:1909.07972*, 2019.
- [7] 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.
- [8] M. M. Amiri and D. Gunduz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *arXiv preprint arXiv:1901.00844*, 2019.
- [9] —, "Federated learning over wireless fading channels," *arXiv preprint arXiv:1907.09769*, 2019.
- [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] J. Bernstein, Y. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," *arXiv preprint arXiv:1802.04434*, 2018.
- [12] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," in *Advances in Neural Information Processing Systems*, 2017, pp. 1709–1720.
- [13] J. Wu, W. Huang, J. Huang, and T. Zhang, "Error compensated quantized sgd and its applications to large-scale distributed optimization," *arXiv preprint arXiv:1806.08054*, 2018.
- [14] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "Terngrad: Ternary gradients to reduce communication in distributed deep learning," in *Advances in neural information processing systems*, 2017, pp. 1509–1519.
- [15] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. McMahan, "cpsgd: Communication-efficient and differentially-private distributed sgd," in *Advances in Neural Information Processing Systems*, 2018, pp. 7564–7575.
- [16] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Sparse binary compression: Towards distributed deep learning with minimal communication," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.
- [17] —, "Robust and communication-efficient federated learning from non-iid data," *arXiv preprint arXiv:1903.02891*, 2019.
- [18] H. Wang, S. Sievert, S. Liu, Z. Charles, D. Papailiopoulos, and S. Wright, "Atom: Communication-efficient learning via atomic sparsification," in *Advances in Neural Information Processing Systems*, 2018, pp. 9850–9861.
- [19] S. Caldas, J. Konečný, H. B. McMahan, and A. Talwalkar, "Expanding the reach of federated learning by reducing client resource requirements," *arXiv preprint arXiv:1812.07210*, 2018.
- [20] T. D. Burd and R. W. Brodersen, "Processor design for portable systems," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 13, no. 2-3, pp. 203–221, 1996.
- [21] A. Goldsmith, *Wireless communications*. Cambridge university press, 2005.
- [22] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," *lecture notes of EE392o, Stanford University, Autumn Quarter*, vol. 2004, pp. 2004–2005, 2003.
- [23] P. D. Tao and L. T. H. An, "Convex analysis approach to DC programming: theory, algorithms and applications," *Acta mathematica vietnamica*, vol. 22, no. 1, pp. 289–355, 1997.

- [24] R. Jin, Y. Huang, X. He, H. Dai, and T. Wu, “Stochastic-Sign SGD for federated learning with theoretical guarantees,” *arXiv preprint arXiv:2002.10940*, 2020.