Joint Optimization of Communications and Federated Learning Over the Air

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

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

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

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

Abstract

Federated learning (FL) is an attractive paradigm for making use of rich distributed data while protecting data privacy. Nonetheless, nonideal communication links and limited transmission resources may hinder the implementation of fast and accurate FL. In this paper, we study joint optimization of communications and FL based on analog aggregation transmission in realistic wireless networks. We first derive closed-form expressions for the expected convergence rate of FL over the air, which theoretically quantify the impact of analog aggregation on FL. Based on the analytical results, we develop a joint optimization model for accurate FL implementation, which allows a parameter server to select a subset of workers and determine an appropriate power scaling factor. Since the practical setting of FL over the air encounters unobservable parameters, we reformulate the joint optimization of worker selection and power allocation using controlled approximation. Finally, we efficiently solve the resulting mixed-integer programming problem via a simple yet optimal finite-set search method by reducing the search space. Simulation results show that the proposed solutions developed for realistic wireless analog channels outperform a benchmark method, and achieve comparable performance of the ideal case where FL is implemented over noise-free wireless channels.

Index Terms

Federated learning, analog aggregation, convergence analysis, joint optimization, worker scheduling, power scaling.

I. INTRODUCTION

In recent years, with the development of IoT and social networks, huge amounts of data have been generated at the edge of networks [1]. To obtain useful information from big data, machine learning has been widely applied to deal with complex models and tasks in emerging data-driven applications, such as autonomous driving, virtual and augmented reality [2]. Standard machine learning is usually developed under a centralized architecture, where each node located at the edge sends its collected data to a central node for centralized data processing. However, with the exponential growth of the volume of local data and the increasing concerns on user data privacy, it is neither practical nor safe to directly transmit the data of local devices to a central node due to the limited communication and processing capability as well as the lack of privacy protection. As such, distributed machine learning is well motivated to overcome these issues.

In the regime of distributed machine learning, federated learning (FL) has been proposed as a well noted approach for collaborative learning [3]. In FL, local workers train local models from their own data, and then transmit their local updates to a parameter server (PS). The PS aggregates these received local updates and sends the averaged update back to the local workers. These iterative updates between the PS and workers, can be either model parameters or their gradients, for model averaging [4] and gradient averaging [5], respectively. In this way, FL relieves communication overheads and protects user privacy compared to traditional data-sharing based collaborative learning, especially when the local data is in large volume and privacy-sensitive. Existing research on FL mostly focuses on FL algorithms under idealized link assumptions. However, the impacts of wireless environments on FL performance should be taken into account in the design of FL deployed in practical wireless systems. Otherwise, such impacts may introduce unwanted training errors that dramatically degrade the learning performance in terms of accuracy and convergence rate [6].

To solve this problem, research efforts have been spent on optimizing network resources used for transmitting model updates in FL [7], [8]. These works of FL over wireless networks adopt digital communications, using a transmission-then-aggregation policy. Unfortunately, the communication overhead and transmission latency become large as the number of active workers increases. On the other hand, it is worth noting that FL aims for global aggregation and hence only utilizes the averaged updates of distributed workers rather than the individual local updates from workers. Alternatively, the nature of waveform superposition in wireless multiple access

channel (MAC) [9]–[12] provides a direct and efficient way for transmission of the averaged updates in FL, also known as analog aggregation based FL [13]–[18]. As a joint transmission-and-aggregation policy, analog aggregation transmission enables all the participating workers to simultaneously upload their local model updates to the PS over the same time-frequency resources as long as the aggregated waveform represents the averaged updates, thus substantially reducing the overhead of wireless communication for FL.

The research on analog aggregation based FL is still at early stage, leaving some fundamental questions unexplored, such as its convergence behavior and design of efficient algorithms. Given the limited transmit power and communication bandwidth at user devices, users may have to contend for communication resources when transmitting their local updates to the PS. It gives rise to the need for an efficient transmission paradigm, along with network resource allocation in terms of worker selection and transmit power control. All these practical issues motivate our work to study FL from the perspectives of both wireless communications and machine learning. In this paper, we quantify the impact of analog aggregation on the convergence behavior and performance of FL. Such quantitative results are essential in guiding the joint optimization of communication and computing resources. This paper aims at a comprehensive study on problem formulation, solution development, and algorithm implementation for the joint design and optimization of wireless communication and FL. Our key contributions are summarized as follows:

- We derive closed-form expressions for the expected convergence rate of FL over the air in the cases of convex and non-convex, respectively, which not only interprets but also quantifies the impact of wireless communications on the convergence and accuracy of FL over the air. Also, full-size gradient descent (GD) and mini-batched statistical gradient descent (SGD) methods are both considered in this work. These closed-form expressions unveil a fundamental connection between analog wireless communication and FL with analog aggregation, which provides a fresh perspective to measure how the parameter design of analog wireless systems affects the performance of FL over the air.

- Based on the closed-form theoretical results, we formulate a joint optimization problem of learning, worker selection, and power control, with a goal of minimizing the global FL loss function given limited transmit power and bandwidth. The optimization formulation turns out to be universal for the convex and non-convex cases with GD and SGD. Further, for practical implementation of the joint optimization problem in the presence of some

unobservable parameters, we develop an alternative reformulation that approximates the original unattainable problem as a feasible optimization problem under the operational constraints of analog aggregation.

- To efficiently solve the approximate problem, we identity a tight solution space by exploring the relationship between the number of workers and the power scaling. Thanks to the reduced search space, we propose a simple discrete enumeration method to efficiently find the globally optimal solution.

We evaluate the proposed joint optimization scheme for FL with analog aggregation in solving linear regression and image classification problems, respectively. Simulation results show that our proposed FL is superior to the benchmark scheme that uses random worker selection and power control, and achieves comparable performance to the ideal case where FL is implemented over noise-free wireless channels.

The remainder of this paper is organized as follows. Related work is presented in Section II. The system model for FL over the air and the associated joint communication and learning optimization formulation are presented in Section III. Section IV derives the closed-form expressions of the expected convergence rate of the FL over the air as the foundation for algorithm design and performance analysis. Section V provides a framework of joint optimization of communication and FL, and develops the corresponding algorithms. Section VI presents numerical results, followed by conclusions in Section VII.

II. RELATED WORK

This section reviews the literature and highlights the novelty of this paper with respect to related works.

To achieve communication efficiency in distributed learning, most of the existing strategies focus on digital communications, which may involve the preprocessing of transmitted updates or the management of wireless resources. For example, a popular line of work is to reduce the communication load per worker by compression of the updates under the assumptions of ideal communication links, such as exploiting coding schemes [19], utilizing the sparsity of updates [20], employing quantization of the updates [21], and avoiding less informative local updates via communication censoring schemes [22]–[26]. Another line of work is to support FL through communication resource management, such as worker scheduling schemes to maximize the number of participating workers [27], joint optimization of resource allocation and worker
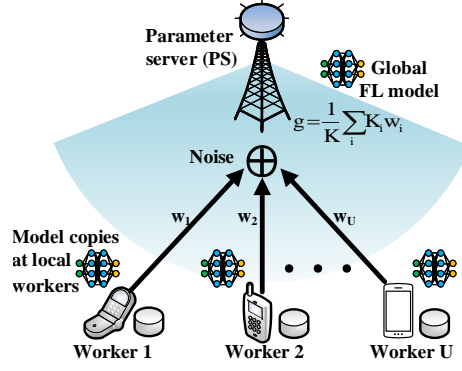
Fig. 1: Federated learning via analog aggregation from wirelessly distributed data.

scheduling [7], and communication and computation resource allocation and scheduling for cell-free networks [8].

There are some pioneering works on analog aggregation based FL [13]–[18], most of which focus on designing transmission schemes [13]–[16]. They adopt preselected participating workers and fix their power allocation without further optimization along FL iterations. The optimization issues are considered in [17], [18], but they are mainly conducted on the communication side alone, without an underlying connection to FL. When communication-based metrics are used, the optimization in existing works often suggests to maximize the number of selected workers that participate FL, but our theoretical results indicate that selecting more workers is not necessarily better over imperfect links or under limited communication resources. Thus, unlike these existing works, we seek to analyze the convergence behavior of analog aggregation based FL, which provides a fresh angle to interpret the specific relationship between communications and FL in the paradigm of analog aggregation. Such a connection leads to this work on a joint optimization framework for analog communications and FL, in which the work selection and power allocation decisions are optimized during the iterative FL process.

III. SYSTEM MODEL

As shown in Fig. 1, we consider a one-hop wireless network consisting of a single parameter server (PS) at a base station and U user devices as distributed local workers. Through federated learning, the PS and all workers collaborate to train a common model for supervised learning and data inference, without sharing local data.

A. FL Model

Let $\mathcal{D}_i = \{\mathbf{x}_{i,k}, \mathbf{y}_{i,k}\}_{k=1}^{K_i}$ denote the local dataset at the i-th worker, $i = 1, \ldots, U$, where $\mathbf{x}_{i,k}$ is the input data vector, $\mathbf{y}_{i,k}$ is the labeled output vector, $k = 1, 2, ..., K_i$, and $K_i = |\mathcal{D}_i|$ is the number of data samples available at the i-th worker. With $K = \sum_{i=1}^{U} K_i$ samples in total, these U workers seek to collectively train a learning model parameterized by a global model parameter $\mathbf{w} = [w^1, \ldots, w^D] \in \mathcal{R}^D$ of dimension D, by minimizing the following loss function

$$\text{(Global loss function)} \quad F(\mathbf{w}; \mathcal{D}) = \frac{1}{K} \sum_{i=1}^{U} \sum_{k=1}^{K_i} f(\mathbf{w}; \mathbf{x}_{i,k}, \mathbf{y}_{i,k}), \tag{1}$$

where the global loss function $F(\mathbf{w}; \mathcal{D})$ is a summation of K data-dependent components, each component $f(\mathbf{w}; \mathbf{x}_{i,k}, \mathbf{y}_{i,k})$ is a sample-wise local function that quantifies the model prediction error of the same data model parameterized by the shared model parameter $\mathbf{w}$, and $\mathcal{D} = \bigcup_i \mathcal{D}_i$.

In distributed learning, each worker trains a local model $\mathbf{w}_i$ from its local data $\mathcal{D}_i$, which can be viewed as a local copy of the global model $\mathbf{w}$. That is, the local loss function is

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

where $\mathbf{w}_i = [w_i^1, \ldots, w_i^D] \in \mathcal{R}^D$ is the local model parameter. Through collaboration, it is desired to achieve $\mathbf{w}_i = \mathbf{w} = \mathbf{w}^*, \forall i$, so that all workers reach the globally optimal model $\mathbf{w}^*$. Such a distributed learning can be formulated via consensus optimization as [4], [28]

$$\textbf{P1:} \quad \min_{\mathbf{w}} \quad \frac{1}{K} \sum_{i=1}^{U} \sum_{k=1}^{K_i} f(\mathbf{w}_i; \mathbf{x}_{i,k}, y_{i,k}). \tag{3a}$$

To solve **P1**, this paper adopts a model-averaging algorithm for FL [4], [28]. It is essentially an iterative process consisting of both computing and communication steps at each iteration. Specifically, in each communication round, the PS broadcasts the current $\mathbf{w}$ to all workers. Then, the i-th worker uses a learning algorithm to update its $\mathbf{w}_i$ by minimizing its local data-dependent loss function in (2). In this work, gradient descent[1] is applied, in which the local model at the i-th local worker is updated as

$$\text{(Local model updating)} \quad \mathbf{w}_i = \mathbf{w} - \alpha \nabla F_i(\mathbf{w}_i; \mathcal{D}_i)$$

$$= \mathbf{w} - \frac{\alpha}{K_i} \sum_{k=1}^{K_i} \nabla f(\mathbf{w}; \mathbf{x}_{i,k}, \mathbf{y}_{i,k}), \tag{4}$$

---

[1]In this work, we take the basic gradient descent as an example, while the proposed methodology can be extended to mini-batch gradient descent as well.

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

When local updating is completed, each worker transmits its updated parameter $\mathbf{w}_i$ to the PS via wireless uplinks to update the global $\mathbf{w}$ as

$$\text{(Global model updating)} \quad \mathbf{w} = \frac{\sum_{i=1}^{U} K_i \mathbf{w}_i}{K}. \tag{5}$$

Then, the PS broadcasts $\mathbf{w}$ in (5) to all participating workers as their initial value in the next round. The FL implements the local model-updating in (4) and the global model-averaging in (5) iteratively, until convergence. It has been shown that this FL algorithm converges to the globally optimal solution of the original problem in **P1** under the conditions that F is a convex function and the data transmission between the PS and workers is error-free [4], [28].

Note that the implementation steps in (4) and (5) only concern the computational aspect of FL, by assuming perfect communications for both the global $\mathbf{w}$ and local $\mathbf{w}_i$ between the PS and all workers. However, the communication impacts on FL performance should not be ignored. Especially in practical wireless network environments, certain errors are inevitably introduced during transmissions of the updates due to the imperfect characteristics of wireless channels.

B. Analog Aggregation Transmission Model

To avoid heavy communication overhead and save transmission bandwidth of FL over wireless channels, we adopt analog aggregation without coding, which allows multiple workers to simultaneously upload their local model updates to the PS over the same time-frequency resources. All workers transmit their local $\mathbf{w}_i$'s in an analog form with perfect time synchronization among them[2]. In this way, the local updates $\mathbf{w}_i$'s are aggregated over the air to implement the global model updating step in (5). Such an analog aggregation is conducted in an entry-wise manner. That is, the d-th entries w_i^d from all workers, $i = 1, ..., U$, are aggregated to compute w^d in (5), for any $d \in [1, D]$.

Let $\mathbf{p}_{i,t} = [p_{i,t}^1, \ldots, p_{i,t}^d, \ldots, p_{i,t}^D]$ denote the power control vector of worker i at the t-th iteration. Noticeably, the choice of $\mathbf{p}_{i,t}$ in FL over the air should be made not only to effectively

---

[2]The implementation of time synchronization and the impact of imperfect synchronization are beyond the scope of this work. Interested readers are referred to [12], [29].

implement the aggregation rule in (5), but also to properly accommodate the need for network resource allocation. Accordingly, we set the power control policy as

$$p_{i,t}^d = \frac{\beta_{i,t}^d K_i b_t^d}{h_{i,t}^d}, \tag{6}$$

where $h_{i,t}$ is the channel gain between the i-th worker and the PS at the t-th iteration[3], b_t^d is the power scaling factor, and $\beta_{i,t}^d$ is a transmission scheduling indicator. That is, $\beta_{i,t}^d = 1$ means that the d-th entry of the local model parameter $\mathbf{w}_{i,t}$ of the i-th worker is scheduled to contribute to the FL algorithm at the t-th iteration, and $\beta_{i,t}^d = 0$, otherwise. Through power scaling, the transmit power used for uploading the d-th entry from the i-th worker should not exceed a maximum power limit $P_i^{d,\max} = P_i^{\max}$ for any d, as follows:

$$|p_{i,t}^d w_{i,t}^d|^2 = \left| \frac{\beta_{i,t}^d K_i b_t^d}{h_{i,t}^d} w_{i,t}^d \right|^2 \leq P_i^{\max}. \tag{7}$$

At the PS side, the received signal at the t-th iteration can be written as

$$\mathbf{y}_t = \sum_{i=1}^U \mathbf{p}_{i,t} \odot \mathbf{w}_{i,t} \odot \mathbf{h}_{i,t} + \mathbf{z}_t$$
$$= \sum_{i=1}^U K_i \mathbf{b}_t \odot \boldsymbol{\beta}_{i,t} \odot \mathbf{w}_{i,t} + \mathbf{z}_t, \tag{8}$$

where $\odot$ represents Hadamard product, $\mathbf{h}_{i,t} = [h_{i,t}^1, h_{i,t}^2, ..., h_{i,t}^D]$, $\boldsymbol{\beta}_{i,t} = [\beta_{i,t}^1, \beta_{i,t}^2, .., \beta_{i,t}^D]$, $\mathbf{b}_t = [b_t^1, b_t^2, ..., b_t^D]$, and $\mathbf{z}_t \sim \mathcal{CN}(0, \sigma^2 \mathbf{I})$ is additive white Gaussian noise (AWGN).

Given the received $\mathbf{y}_t$, the PS estimates $\mathbf{w}_t$ via a post-processing operation as

$$\mathbf{w}_t = \left(\sum_{i=1}^U K_i \boldsymbol{\beta}_{i,t} \odot \mathbf{b}_t \right)^{\odot -1} \odot \mathbf{y}_t$$
$$= \left(\sum_{i=1}^U K_i \boldsymbol{\beta}_{i,t} \right)^{\odot -1} \sum_{i=1}^U K_i \boldsymbol{\beta}_{i,t} \odot \mathbf{w}_{i,t} + \left(\sum_{i=1}^U K_i \boldsymbol{\beta}_{i,t} \odot \mathbf{b}_t \right)^{\odot -1} \odot \mathbf{z}_t, \tag{9}$$

where $(\sum_{i=1}^U K_i \boldsymbol{\beta}_{i,t} \odot \mathbf{b}_t)^{\odot -1}$ is a properly chosen scaling vector to produce equal weighting for participating $\mathbf{w}_i$'s in (9) as desired in (5), and $(\mathbf{X})^{\odot -1}$ represents the inverse Hadamard operation of $\mathbf{X}$ that calculates its entry-wise reciprocal. Noticeably, in order to implement the averaging of (5) in FL over the air, such a post-processing operation requires $\mathbf{b}_t$ to be the same for all workers for given t and d, which allows to eliminate $\mathbf{b}_t$ from the first term in (9).

---

[3]In this paper, we assume the channel state information (CSI) to be constant within each iteration, but may vary over iterations. We also assume that the CSI is perfectly known at the PS, and leave the imperfect CSI case in future work.

Comparing (9) with (5), there exist differences between $\mathbf{w}_t$ and $\mathbf{w}$ due to the effect of wireless communications. This work aims to mitigate such a gap through optimizing the worker selection $\boldsymbol{\beta}_{i,t}$ and power scaling $\mathbf{b}_t$ for FL over the air. To this end, our next step is to unveil an important but unexplored foundation, i.e., how wireless communications affect the convergence behavior of FL over the air.

IV. THE CONVERGENCE ANALYSIS OF FL WITH ANALOG AGGREGATION

In this section, we study the effect of analog aggregation transmission on FL over the air, by analyzing its convergence behavior for both the convex and the non-convex cases. To average the effects of instantaneous SNRs, we derive the expected convergence rate of FL over the air, which quantifies the impact of wireless communications on FL using analog aggregation transmissions.

A. Convex Case

We first make the following assumptions that are commonly adopted in the optimization literature [7], [30]–[34].

Assumption 1 (Lipschitz continuity, smoothness): The gradient $\nabla F(\mathbf{w})$ of the loss function $F(\mathbf{w})$ is uniformly Lipschitz continuous with respect to $\mathbf{w}$, that is,

$$\|\nabla F(\mathbf{w}_{t+1}) - \nabla F(\mathbf{w}_t)\| \leq L\|\mathbf{w}_{t+1} - \mathbf{w}_t\|, \quad \forall \mathbf{w}_t, \mathbf{w}_{t+1}, \tag{10}$$

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

Assumption 2 (strongly convex): $\nabla F(\mathbf{w})$ is strongly convex with a positive parameter μ, obeying

$$F(\mathbf{w}_{t+1}) \geq F(\mathbf{w}_t) + (\mathbf{w}_{t+1} - \mathbf{w}_t)^T \nabla F(\mathbf{w}_t) + \frac{\mu}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2, \quad \forall \mathbf{w}_t, \mathbf{w}_{t+1}. \tag{11}$$

Assumption 3 (bounded local gradients): The sample-wised local gradients at local workers are bounded by their global counterpart [32], [33]

$$\|\nabla f(\mathbf{w}_t)\|^2 \leq \rho_1 + \rho_2 \|\nabla F(\mathbf{w}_t)\|^2, \tag{12}$$

where $\rho_1, \rho_2 \geq 0$.

According to [5], [35], the FL algorithm applied over ideal wireless channels is able to solve **P1** and converges to an optimal $\mathbf{w}^*$. In the presence of wireless transmission errors, we derive the expected convergence rate of the FL over the air with analog aggregation, as in **Theorem 1**.

Theorem 1. *Adopt **Assumptions 1-3**, and denote the globally optimal learning model in* (3) *as* $\mathbf{w}^*$. *The model updating rule for* $\mathbf{w}_t$ *of the FL-over-the-air scheme is given by* (9), $\forall t$. *Given the transmit power scaling factors* $\boldsymbol{b}_t$, *worker selection vectors* $\boldsymbol{\beta}_{i,t}$, *and setting the learning rate to be* $\alpha = \frac{1}{L}$, *the expected performance gap* $\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)]$ *of* $\mathbf{w}_t$ *at the* t-*th iteration is given by*

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \leq B_t + A_t \mathbb{E}[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)], \tag{13}$$

with

$$A_t = 1 - \frac{\mu}{L} + \rho_2 \sum_{d=1}^{D} \left(\frac{K}{\sum_{i=1}^{U} K_i \beta_{i,t}^d} - 1 \right), \tag{14}$$

$$B_t = \frac{\rho_1}{2L} \sum_{d=1}^{D} \left(\frac{K}{\sum_{i=1}^{U} K_i \beta_{i,t}^d} - 1 \right) + \left\| \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t \right)^{\odot -1} \right\|^2 \frac{L\sigma^2}{2}, \tag{15}$$

where the expectation is over the channel AWGN of zero mean and variance σ^2.

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

Based on **Theorem 1**, we further derive the cumulative performance gap resulted from wireless communications and the worker selection of the whole FL process, summarized by the following **Lemma 1**.

Lemma 1. *Given an initial global model* $\mathbf{w}_0$, *the cumulative performance gap* $\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)]$ *of FL after* t *iterations is bounded by*

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \leq \underbrace{\sum_{i=1}^{t-1} \prod_{j=1}^{i} A_{t+1-j} B_{t-i} + B_t}_{\Delta_t} + \prod_{j=1}^{t} A_j \mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)]. \tag{16}$$

Proof. Given the expected performance gap at the t-th iteration in (13), we carry out recursions as follows:

$$\begin{aligned}
\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] &\leq B_t + A_t \mathbb{E}[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)] \\
&\leq B_t + A_t \Big(B_{t-1} + A_{t-1} \mathbb{E}[F(\mathbf{w}_{t-2}) - F(\mathbf{w}^*)] \Big) \\
&\leq \dots \\
&\leq \sum_{i=1}^{t-1} \prod_{j=1}^{i} A_{t+1-j} B_{t-i} + B_t + \prod_{j=1}^{t} A_j \mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)]. \tag{17}
\end{aligned}$$

$\square$

.

Lemma 1 reveals that the FL algorithm converges asymptotically in t under mild conditions, as stated in **Proposition 1**.

Proposition 1. *Given the learning rate $\alpha = \frac{1}{L}$, the convergence of the FL algorithm is guaranteed with $\lim_{t\to\infty} \mathbf{w}_t = \mathbf{w}^*$, as long as ρ_2 in (12) satisfies the following condition:*

$$0 < \rho_2 < \frac{\mu}{(\frac{K}{K_{min}} - 1)DL}, \tag{18}$$

where $K_{min} = \min\{K_i\}_{i=1}^{U}$.

Proof. When $A_t < 1$, $\forall t$, it is evident that $\lim_{t\to\infty} \prod_{j=1}^{t+1} A_j = 0$. From **Lemma 1**, to guarantee the convergence, a sufficient condition is to ensure $A_{max} \triangleq \max\{A_t, t = 1, 2...\} < 1$. Given (14), it holds that

$$A_t = 1 - \frac{\mu}{L} + \rho_2 \sum_{d=1}^{D} \left(\frac{K}{\sum_{i=1}^{U} K_i \beta_{i,t}^d} - 1 \right)$$

$$\leq 1 - \frac{\mu}{L} + \rho_2 \sum_{d=1}^{D} \left(\frac{K}{K_{min}} - 1 \right), \tag{19}$$

where $K_{min} = \min\{K_i\}_{i=1}^{U}$. When all workers have the same amount of data, i.e., $K_i = \frac{K}{U}, \forall i$, then $A_t \leq 1 - \frac{\mu}{L} + \rho_2 D(U - 1)$.

To ensure $A_{max} < 1$, we have

$$A_{max} \leq 1 - \frac{\mu}{L} + \rho_2 \sum_{d=1}^{D} \left(\frac{K}{K_{min}} - 1 \right) < 1. \tag{20}$$

From (20), it holds that $\rho_2 < \frac{\mu}{(\frac{K}{K_{min}}-1)DL}$. On the other hand, $\rho_2 > 0$, according to (12) in **Assumption 3**. As a result, we have $0 < \rho_2 < \frac{\mu}{(\frac{K}{K_{min}}-1)DL}$, which completes the proof. $\square$

Proposition 1 indicates that the convergence behavior of the FL algorithm depends on both the learning-related parameters, i.e., μ, L, ρ_1, ρ_2, and communication-related parameters, including $\boldsymbol{\beta}$, $\mathbf{b}$ and σ^2. Interestingly, the channel noise σ^2 and $\mathbf{b}$ do not affect A_t, and hence they do not affect the convergence of the FL algorithm but determine the steady state that the FL algorithm converges to.

Lemma 1 also provides the expected convergence rate of an FL algorithm when the transmission link is error-free. In this ideal case, it offers the fastest convergence rate achievable by the FL algorithm, which is derived by the following **Lemma 2**.

Lemma 2. *Consider a resource-unconstrained and error-free mode where the effects of wireless channels, as well as that of noise, are already mitigated or fully compensated. Given the optimal global* $\mathbf{w}^*$ *and the learned* $\mathbf{w}_t$ *in* (9)*, the upper bound of* $\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)]$ *for the FL over the air is given by*

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \leq (1 - \frac{\mu}{L})^t \mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)]. \tag{21}$$

Proof. Without channel noise or worker selection (that is, all workers participate the FL and deliver their data perfectly), we have $\sigma^2 = 0$ and $\sum_{d=1}^{D} \left(\frac{K}{\sum_{i=1}^{U} K_i \beta_{i,t}^d} - 1 \right) = 0$. Then, in (14) and (15), we have $B_t = B = 0$ and $A_t = A = 1 - \frac{\mu}{L}$, $\forall t$. As a result, (16) is reduced to (21). $\square$

It is worth noting that **Lemma 2** provides the convergence rate for the ideal case, which assumes that the impacts of wireless communications, including noise, channel and constrained resources, are all mitigated to result in error-free transmission. According to (16) in the realistic case, the trajectory of $\mathbb{E}[F(\mathbf{w}_{t+1})]$ exhibits jump discontinuity with a gap term Δ_t at each step t, as defined in (16):

$$\Delta_t = \sum_{i=1}^{t-1} \prod_{j=1}^{i} A_{t+1-j} B_{t-i} + B_t.$$

This gap reflects the impact of wireless communication factors on FL, by means of the worker selection, transmit power scaling and AWGN. Intuitively, this gap diminishes as the number of selected workers increases, which reduces the value of A_t. Meanwhile, as the power scaling factor $\boldsymbol{b}_t$ increases, B_t is decreased, which leads to a reduction of the gap as well. Hence, it is necessary to optimize transmit power scaling factors and worker selection in order to minimize the gap in (16) for the implementation of FL algorithms over a realistic wireless network.

B. Non-convex Case

When the loss function $F(w)$ is nonconvex, such as in the case of convolutional neural networks, we derive the convergence rate of the FL over the air with analog aggregation for the nonconvex case without **Assumption 2**, which is summarized in **Theorem 2**.

Theorem 2. *Under the **Assumptions 1** and **3** for the non-convex case, given the transmit power scaling factors $\boldsymbol{b}_t$, worker selection vectors $\boldsymbol{\beta}_{i,t}$, the optimal global FL model $\mathbf{w}^*$, and the learning rate $\alpha = \frac{1}{L}$, the convergence at the T-th iteration is given by*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_{t-1})\|^2 \leq \frac{2L}{T(1 - \rho_2 D(\frac{K}{K_{min}} - 1))}\mathbb{E}[F(\mathbf{w}_0)] - F(\mathbf{w}^*)]$$
$$+ \frac{2L\sum_{t=1}^{T}B_t}{T(1 - \rho_2 D(\frac{K}{K_{min}} - 1))}. \tag{22}$$

Proof. Please refer to Appendix B. □

As we can see from **Theorem 2**, when T is large enough, we have

$$\min_{0,1,\ldots,T}\mathbb{E}[\|\nabla F(\mathbf{w}_{t-1})\|^2] \leq \frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_{t-1})\|^2 \overset{T\to\infty}{\leq} \underbrace{\frac{2L\sum_{t=1}^{T}B_t}{T(1 - \rho_2 D(\frac{K}{K_{min}} - 1))}}_{\triangle_T^{NC}}, \tag{23}$$

which guarantees convergence of the FL algorithm to a stationary point [28], [36]. Similarly, the performance gap at the step t for non-convex cases is given by

$$\triangle_t^{NC} = \frac{2L\sum_{t=1}^{T}B_t}{T(1 - \rho_2 D(\frac{K}{K_{min}} - 1))}. \tag{24}$$

Note that the non-convex case and the convex shares the same sufficient condition for convergence as (18) in **Proposition 1**.

C. Stochastic gradient descent

Our work can be extended to stochastic versions of gradient descent (SGD) as well. Here, we provide convergence analysis for mini-batch gradient descent with a constant mini-batch size K_b, while the results directly apply to the standard SGD by setting $K_b = 1$. Theorem 3 summarizes the convergence behavior of SGD for the strongly convex case.

Theorem 3. *Under the **Assumptions 1, 2** and **3** for the convex case, and given the transmit power scaling factors $\boldsymbol{b}_t$, worker selection vectors $\boldsymbol{\beta}_{i,t}$, the optimal global FL model $\mathbf{w}^*$, the learning rate $\alpha = \frac{1}{L}$ and the mini-batch size K_b, the convergence behavior of the SGD implementation of FL over the air is given by*

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \leq \underbrace{\sum_{i=1}^{t-1}\prod_{j=1}^{i}A_{t+1-j}^{SGD}B_{t-i}^{SGD} + B_t^{SGD}}_{\Delta_t^{SGD}} + \prod_{j=1}^{t}A_j^{SGD}\mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)], \tag{25}$$

where

$$A_t^{SGD} = 1 - \frac{\mu}{L} + \rho_2 \left(\sum_{d=1}^{D} \left(\frac{(\sum_{i=1}^{U} K_b)^2 - 2K(\sum_{i=1}^{U} K_b)}{K^2} + \frac{(\sum_{i=1}^{U} K_b)}{\sum_{i=1}^{U} K_b \beta_{i,t}^d} \right) \right.$$
$$\left. + \frac{(\sum_{i=1}^{U} (K_i - K_b))^2}{K^2} \right), \tag{26}$$

$$B_t^{SGD} = \frac{\rho_1}{2L} \left(\sum_{d=1}^{D} \left(\frac{(\sum_{i=1}^{U} K_b)^2 - 2K(\sum_{i=1}^{U} K_b)}{K^2} + \frac{(\sum_{i=1}^{U} K_b)}{\sum_{i=1}^{U} K_b \beta_{i,t}^d} \right) \right.$$
$$\left. + \frac{(\sum_{i=1}^{U} (K_i - K_b))^2}{K^2} \right) + \left\| \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t \right)^{\odot - 1} \right\|^2 \frac{L\sigma^2}{2}. \tag{27}$$

Proof. Please refer to Appendix C. $\qquad\square$

From **Theorem 3**, the cumulative performance gap of FL after the t-th iteration for the SGD case is bounded by

$$\triangle_t^{SGD} = \sum_{i=1}^{t-1} \prod_{j=1}^{i} A_{t+1-j}^{SGD} B_{t-i}^{SGD} + B_t^{SGD}. \tag{28}$$

Remark 1. If K_b is set to be K_i, then **Theorem 3** for SGD is the same as **Theorem 1** for GD. In addition, since the common mini-batch size is no larger than the minimum local data size, i.e., $K_b \le K_{min} \le \frac{K}{U}$, both A_t^{SGD} in (26) and B_t^{SGD} in (27) decrease as K_b increases, which leads to a smaller Δ_t^{SGD} in (25). In other words, FL has a better convergence performance with a larger K_b. On the other hand, such an improvement on performance is achieved at the cost of high computation load at the local workers in each communication round, which reflects the tradeoff between training performance and computational complexity in SGD.

Similarly, we can also derive the mild convergence condition for SGD, given by the following **Proposition 2**.

Proposition 2. *Given the learning rate $\alpha = \frac{1}{L}$, the convergence of the FL algorithm for SGD cases is guaranteed with $\lim_{t \to \infty} \mathbf{w}_t = \mathbf{w}^*$, as long as ρ_2 in (12) satisfies the following condition:*

$$0 < \rho_2 < \frac{\mu}{\left(\frac{2UK_b}{K} + \frac{U^2 K_b^2}{K^2} + DU - \frac{2DUK_b}{K} + \frac{DU^2 K_b^2}{K^2} \right) L}. \tag{29}$$

Proof. Similar to the GD case, to guarantee the convergence, a sufficient condition is still to ensure $A_{max}^{SGD} \triangleq \max\{A_t^{SGD}, t = 1, 2...\} < 1$. It holds that,

$$A_t^{SGD} = 1 - \frac{\mu}{L} + \rho_2 \left(\frac{(K - UK_b)^2}{K^2} + \sum_{d=1}^{D} \left(\frac{U^2 K_b^2 - 2KUK_b}{K^2} + \frac{U}{\sum_{i=1}^{U} \beta_{i,t}^d} \right) \right)$$

$$\leq 1 - \frac{\mu}{L} + \rho_2 \left(\frac{(K - UK_b)^2}{K^2} + \frac{DU^2 K_b^2 - 2DKUK_b}{K^2} + DU \right), \tag{30}$$

which comes from the fact that $\sum_{i=1}^{U} \beta_{i,t}^d \geq 1$.

Thus, we have

$$A_{max}^{SGD} \leq 1 - \frac{\mu}{L} + \rho_2 \left(\frac{(K - UK_b)^2 + DU^2 K_b^2 - 2DKUK_b + DUK^2}{K^2} \right) < 1. \tag{31}$$

From (31), we can get $\rho_2 < \frac{\mu}{(1 - \frac{2UK_b}{K} + \frac{U^2 K_b^2}{K^2} + DU - \frac{2DUK_b}{K} + \frac{DU^2 K_b^2}{K^2})L}$. Considering $\rho_2 > 0$ in **Assumption 3**, we get (29) as a result. $\qquad\square$

V. PERFORMANCE OPTIMIZATION FOR FEDERATED LEARNING OVER THE AIR

In this section, we first formulate a joint optimization problem to reduce the gap for FL over the air, which turns out to be applicable for both the convex and non-convex cases, using either GD or SGD implementations. To make it applicable in practice in the presence of some unobservable parameters at the PS, we reformulate it to an approximate problem by imposing a conservative power constraint. To efficiently solve such an approximate problem, we first identify a tight solution space and then develop an optimal solution via discrete programming.

A. Problem Formulation for Joint Optimization

Since we are concerned with convergence accuracy, our optimization problem boils down to minimizing the performance gap for different cases (i.e., $\triangle_t$, $\triangle_t^{NC}$, and $\triangle_t^{SGD}$) at each iteration under the corresponding convergence conditions (i.e., **Proposition 1** and **Proposition 2**).

We recognize that solving **P1** amounts to iteratively minimizing those gap $\triangle_t$, $\triangle_t^{NC}$, and $\triangle_t^{SGD}$ under the transmit power constraint in (7). At the t-th iteration, the objective functions for those three cases are given by

$$\triangle_t = B_t + A_t \triangle_{t-1}, \tag{32}$$

$$\triangle_t^{NC} = B_t, \tag{33}$$

$$\triangle_t^{SGD} = B_t^{SGD} + A_t^{SGD} \triangle_{t-1}^{SGD}. \tag{34}$$

where $\triangle_0 = 0$ and $\triangle_0^{SGD} = 0$. Note that when the optimization is executed at the t-th iteration, $\triangle_{t-1}$ and $\triangle_{t-1}^{SGD}$ can be treated as constants.

Considering the entry-wise transmission for analog aggregation, we remove irrelevant items and extract the component of the d-th entry from those gap in (32), (33) and (34) as the objective to minimize, which is given by

$$R_t[d] = \frac{L\sigma^2}{2\left(\sum_{i=1}^{U} \beta_{i,t}^d K_i b_t^d\right)^2} + \frac{K\rho_1 + 2KL\rho_2\triangle_{t-1}}{2L\sum_{i=1}^{U} K_i \beta_{i,t}^d}, \quad \forall d, \tag{35}$$

$$R_t^{NC}[d] = \frac{L\sigma^2}{2\left(\sum_{i=1}^{U} \beta_{i,t}^d K_i b_t^d\right)^2} + \frac{K\rho_1}{2L\sum_{i=1}^{U} K_i \beta_{i,t}^d}, \quad \forall d, \tag{36}$$

$$R_t^{SGD}[d] = \frac{L\sigma^2}{2\left(\sum_{i=1}^{U} \beta_{i,t}^d K_i b_t^d\right)^2} + \frac{U(\rho_1 + 2L\rho_2\triangle_{t-1})}{2L\sum_{i=1}^{U} K_i \beta_{i,t}^d}, \quad \forall d. \tag{37}$$

Since all entries indexed by d are separable with respect to the design parameters, we perform entry-wise optimization by considering $\mathbf{w}_t$ and $\mathbf{w}_{i,t}$ one entry at a time, where the superscript d and the index of different cases are omitted hereafter. To determine the worker selection vector $\beta_{i,t}$ and the power scaling factor b_t at the t-th iteration, the PS carries out a joint optimization problem formulated as follows:

$$\textbf{P2:} \quad \min_{\{b_t, \beta_{i,t}\}_{i=1}^{U}} R_t \tag{38a}$$

$$\text{s.t.} \quad \left| \frac{\beta_{i,t} K_i b_t}{h_{i,t}} w_{i,t} \right|^2 \leq P_i^{\max}, \tag{38b}$$

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

where K_i should be K_b in (38b) for the SGD case.

However, in (38b), the knowledge of $\{w_{i,t}\}_{i=1}^{U}$ is needed but is unavailable to the PS due to analog aggregation.

To overcome this issue, we reformulate a practical optimization problem via an approximation that $\mathbf{w}_{t-1} \approx \frac{1}{U}\sum_{i=1}^{U} \mathbf{w}_{i,t}$, in light of the consensus constraint in **P1**. According to (47) in the proof of **Theorem 1**, each local parameter $\mathbf{w}_{i,t}$ is updated from the broadcast $\mathbf{w}_{t-1}$ along the direction of the averaged gradient over its local data $\frac{\alpha}{K_i}\sum_{k=1}^{K_i} \nabla f(\mathbf{w}_{t-1}; \mathbf{x}_{i,k}, \mathbf{y}_{i,k})$. Hence, it is reasonable to make the following common assumption on bounded local gradients, considering that the local gradients can be controlled by adjusting the learning rate or through simple clipping [21], [28], [37], [38].

Assumption 4 (bounded local gradients): The gap between the global parameter w_{t-1} and the local parameter update $w_{i,t}, \forall i, t$ is bounded by

$$|w_{t-1} - w_{i,t}| \leq \eta, \tag{39}$$

where $\eta \geq 0$ is related to the learning rate α that satisfies the following condition[4]

$$\eta \geq \max \left\{ \left\{ \left| \frac{\alpha}{K_i} \sum_{k=1}^{K_i} \nabla f(w, \mathbf{x}_{i,k}, \mathbf{y}_{i,k}) \right| \right\}_{i=1}^{U} \right\}. \tag{40}$$

Under **Assumption 4**, we reformulate the original optimization problem **(P2)** into the following problem **(P3)**, by replacing $w_{i,t}$ in (38b) by its approximation:

$$\textbf{P3:} \quad \min_{\{b_t, \beta_{i,t}\}_{i=1}^{U}} R_t \tag{41a}$$

$$\text{s.t.} \quad \left| \frac{\beta_{i,t} K_i b_t}{h_{i,t}} \right|^2 (|w_{t-1}| + \eta)^2 \leq P_i^{\max}, \tag{41b}$$

$$\beta_{i,t} \in \{0, 1\}, i \in \{1, 2, ..., U\}, \tag{41c}$$

where the power constraint (41b) is constructed based on the fact that

$$\left| \frac{\beta_{i,t} K_i b_t}{h_{i,t}} w_{i,t} \right|^2 = \left| \frac{\beta_{i,t} K_i b_t}{h_{i,t}} \right|^2 |w_{i,t}|^2$$

$$\leq \left| \frac{\beta_{i,t} K_i b_t}{h_{i,t}} \right|^2 (|w_{t-1}| + \eta)^2. \tag{42}$$

Since w_{t-1} is always available at the PS, **P3** becomes a feasible formulation for adoption in practice. Next, we develop the optimal solution to **P3**.

B. Optimal Solution to *P3* via Discrete Programming

At first glance, a direct solution to **P3** leads to a mixed integer programming (MIP), which unfortunately incurs high complexity. To solve **P3** in an efficient manner, we develop a simple solution by identifying a tight search space without loss of optimality. The tight search space, given in the following **Theorem 4**, is a result of the constraints in (41b) and (41c), irrespective of the objective function (41a). Hence, it holds universally for any R_t, such as those in (35)-(37).

---

[4]This implies the value range of η. In practice, η can take $|w_{t-1} - w_{t-2}|$. In addition, for the SGD case, we have $\eta \geq \max\{\{|\alpha \mathbb{E}_{\mathcal{D}_i}[\nabla f(w, \mathbf{x}_{i,k}, \mathbf{y}_{i,k})]|\}_{i=1}^{U}\}$

Theorem 4. *When all the required parameters in* **P3** *i.e.,* $\{P_i^{\max}, w_{t-1}, h_{i,t}, K_i, \eta\}_{i=1}^{U}$, *are available at the PS, the solution space of* $(b_t, \beta_{i,t})$ *in* **P3** *can be reduced to the following tight search space without loss of optimality as*

$$
\mathcal{S} = \left\{ \left\{ \left(b_t^{(k)}, \beta_{i,t}^{(k)} \right) \right\}_{k=1}^{U} \middle| b_t^{(k)} = \left| \frac{\sqrt{P_k^{\max}} h_{k,t}}{K_k(|w_{t-1}| + \eta)} \right|, \boldsymbol{\beta}_t^{(k)}(b_t^{(k)}) = \left[\beta_{1,t}^{(k)}, \dots, \beta_{U,t}^{(k)} \right], k = 1, \dots, U \right\},
$$
(43)

where $\boldsymbol{\beta}_t^{(k)}$ *is a function of* $b_t^{(k)}$, *in the form:*

$$
\beta_{i,t}^{(k)} = H\left(P_i^{\max} - \left| \frac{K_i b_t^{(k)}(|w_{t-1}| + \eta)}{h_{i,t}} \right| \right)
$$
(44)

and $H(x)$ *is the Heaviside step function, i.e.,* $H(x) = 1$ *for* $x > 0$, *and* $H(x) = 0$ *otherwise.*

Proof. Please see Appendix B. $\qquad\square$

Thanks to **Theorem 4**, we equivalently transform **P3** from a MIP into a discrete programming (DP) problem **P4** as follows

$$
\textbf{P4:} \qquad \min_{(b_t, \boldsymbol{\beta}_t) \in \mathcal{S}} R_t = R_t(b_t, \boldsymbol{\beta}_t)
$$
(45)

According to **P4**, the objective R_t can only take on U possible values corresponding to the U feasible values of b_t; meanwhile, given each b_t, the value of $\boldsymbol{\beta}_t$ is uniquely determined. Hence the minimum R_t can be obtained via line search over the U feasible points $(b_t, \boldsymbol{\beta}_t)$ in (43). Note that the feasible points in (43) are determined by the channel gains, power limits and data sizes of the U workers. Hence, the optimal transmission policy decided by **P4** reflects the tradeoff among workers in terms of their channel quality, available power resource, and data quality.

It is worth noting that the solution b_t^* of **P4** may exceed the maximum value allowed at a worker, due to the approximation introduced in **Assumption 4**. To strictly comply to the power constraint, each worker needs to take the following bounding step when sending its local parameters:

1) if $\left| \frac{K_i b_t^* w_{i,t}}{h_{i,t}} \right|^2 \leq P_i^{\max}$, then the i-th worker sends $\frac{K_i b_t^* w_{i,t}}{h_{i,t}}$;

2) otherwise, it sends $\sqrt{P_i^{\max}} \mathrm{sgn}(w_{i,t})$, where $\mathrm{sgn}(\cdot)$ is the sign function.

Putting together, we propose a joint optimization for **FL** over the air (INFLOTA) as summarized in **Algorithm 1**, which is a dynamic scheduling and power scaling policy. By using different R_t, our INFLOTA can be adjust to all the considered cases including the convex and non-convex, using either GD or SGD implementations.

Algorithm 1 The implementation of INFLOTA

Given:

System parameters $\{P_i^{\max}, K_i, \eta\}_{i=1}^{U}$;

1: The PS initializes $\{\mathbf{w}_0, b_1^*, \boldsymbol{\beta}_1^*\}$ and broadcasts them to all the local workers.

2: **for** $t = 1 : T$ **do**

3:    **At the workers:**

4:      **Computation:** Iteratively update the local model via (4) where $\mathbf{w} = \mathbf{w}_{t-1}$ is from the PS;

5:      **Communication:** Upon receiving $(b_t, \boldsymbol{\beta}_t)$, send $\mathrm{sgn}(w_{i,t}) \min\left(\frac{K_i b_t |w_{i,t}|}{h_{i,t}}, \sqrt{P_i^{\mathrm{Max}}}\right)$ to the PS, if $\beta_{i,t} = 1, \forall i, d$;

6:    **At the PS:**

7:      Calculate the global model $\mathbf{w}_t$ via (9) from $\mathbf{y}_t$ that is aggregated from local workers;

8:      **for** $d = 1 : D$ **do**

9:        Calculate $\mathcal{S}$ from (43), which yields U feasible points $\{(b_{t+1}^{(k)}, \boldsymbol{\beta}_{t+1}^{(k)})\}_{k=1}^{U}$;

10:        Solve **P4** in (45) by using a line search over the U feasible points to find the optimal $\{b_{t+1}^*, \boldsymbol{\beta}_{t+1}^*\}$ for given d and t;

11:      **end for**

12:      Send w_t and $(b_{t+1}, \boldsymbol{\beta}_{t+1})$ (including all D optimal $\{b_{t+1}^*, \boldsymbol{\beta}_{t+1}^*\}$) to all the workers;

13: **end for**

Remark 2. (**Optimality**) **P3** is equivalently reformulated as **P4**, which is solved by a search method with much reduced computational complexity thanks to the reduced search space. Comparing **P3** to **P2**, the constraint (41b) of **P3** is more restrictive than the constraint in (38b) of **P2**. Since **P3** reduces the feasible domain of **P2**, the solution to **P3** cannot be superior to that of **P2**. Therefore, the optimal solution of **P3** is an upper bound of **P2**, i.e., R_t calculated by solving **P3** is larger than the actual one.

Remark 3. (**Complexity**) **Algorithm 1** provides a holistic solution for implementation of the overall FL at both the PS and workers sides. Its computational complexity is mainly determined by that of the optimization step in **P4**. The complexity order of the optimization step is low at $\mathcal{O}(U)$, since the search space is reduced to U points only via **P4**.

Remark 4. (**Implementation**) To implement the FL over the air in **Algorithm 1**, the PS must

know the CSI, the number of the data samples and the maximum transmit power of all local workers. This information can be obtained by the PS when local workers initially connect to it. Before the implementation of **Algorithm 1**, the PS must first broadcast the global model information to the workers. Noticeably, taking **P4** into the implementation of FL, some workers need to send $\text{sgn}(w_{i,t}) \min \left(\frac{K_i b_t |w_{i,t}|}{h_{i,t}}, \sqrt{P_i^{\text{Max}}} \right)$ to meet the requirement on the maximum transmit power. Such a bounding method can be viewed as a quantization measure, which does not affect the convergence [39].

VI. SIMULATION RESULTS AND ANALYSIS

In the simulations, we evaluate the performance of the proposed INFLOTA for both linear regression and image classification tasks, which are based on a synthetic dataset and the MNIST dataset, respectively.

The considered network has $U = 20$ workers, whose maximum power is set to be $P_i^{\text{max}} = P^{\text{max}} = 10$ mW for any $i \in [1, U]$. The receiver noise power at PS is set to be $\sigma^2 = 10^{-4}$ mW, i.e., $SNR = \frac{P^{\text{max}}}{\sigma^2} = 5$ dB. The wireless channel gain between the workers and the PS are generated from a Rayleigh fading model. Here, $h_{i,t}$ is generated from an exponential distribution with unit mean for different i and t.

We use two baseline methods for comparison: a) an FL algorithm that assumes idealized wireless transmissions with error-free links to achieve perfect aggregation, and b) an FL algorithm that randomly determines the power scalar and user selection. They are named as *Perfect aggregation* and *Random policy*, respectively. In *Random policy*, the probability of each worker being selected is 50% and the power scalar is generated from an exponential distribution with unit mean.

A. *Linear regression experiments*

In linear regression experiments, the synthetic data used to train the FL algorithm is generated randomly from $[0, 1]$. The input x and the output y follow the function $y = -2x + 1 + n \times 0.4$ where n follows a Gaussian distribution $\mathcal{N}(0, 1)$. The FL algorithm is used to model the relationship between x and y.

Since linear regression only involves two parameters, we train a simple two-layer neural network, with one neuron in each layer, without activation functions, which is the convex case.
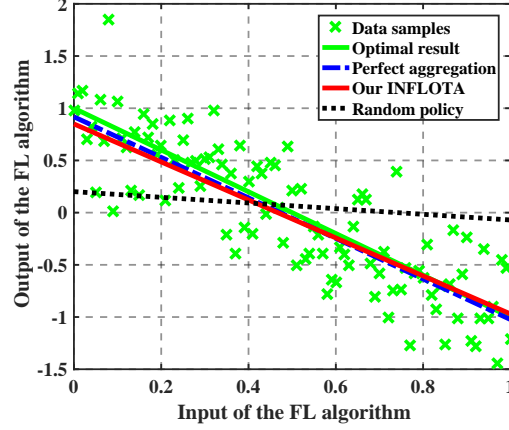
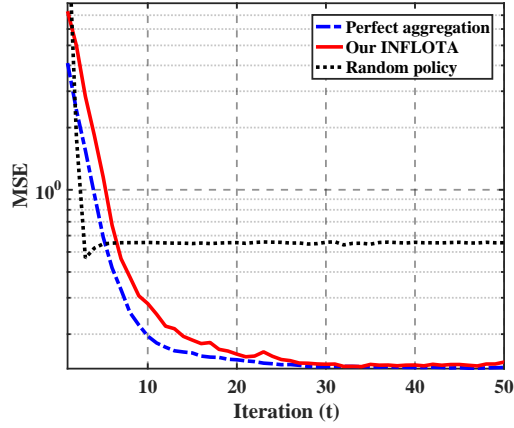Fig. 2: An example of implementing FL for linear regression.

Fig. 3: MSE as the number of iteration varies.

The loss function is the MSE of the model prediction $\hat{y}$ and the labeled true output y. The learning rate is set to 0.01.

Fig. 2 shows an example of using FL for linear regression. The optimal result of a linear regression is $y = -2x + 1$, because the original data generation function is $y = -2x + 1 + 0.4n$. In Fig. 2, we can see that the most accurate approximation is achieved by *Perfect aggregation*, which is the ideal case without considering the influence of wireless communication. *Random policy* considers the influence of wireless communication but without any optimization. Thus, its performance is worst. Our proposed INFLOTA performs closely to the ideal case, which jointly considers the learning and the influence of wireless communication. This is because that our proposed INFLOTA can optimize worker selection and power control so as to reduce the effect of wireless transmission errors on FL.
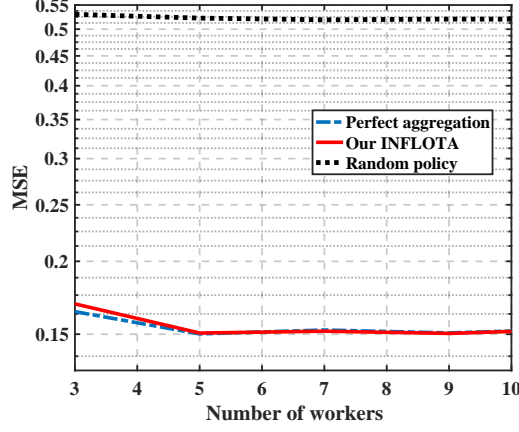
Fig. 4: MSE as the number of workers varies.

In Fig. 3, we show how wireless transmission affects the convergence behavior of the global FL model training in terms the value of the loss function and the global FL model remains unchanged which shows that the global FL model converges. As we can see, as the number of iterations increases, the MSE values of all the considered learning algorithms decrease at different rates, and eventually flatten out to reach their steady state. All schemes converge, but to different steady state values. This behavior corroborates the results in **Lemma 1** and **Proposition 1** that the channel noise does not affect the convergence of the FL algorithm but it affects the value that the FL algorithm converges to.

Fig. 4 shows how MSE varies with the total number of workers U. In general, the MSE performance of all considered FL schemes decreases as U increases. This is due to the fact that an increase in the number of workers leads to an increased volume of data available for the FL training and hence improved accuracy of the estimated model parameters. Moreover, as the number of workers increases, the effect of wireless transmission on the global FL model accuracy starts to diminish. This is because the data samples may be already enough for accurate training when U exceeds a certain level.

In Fig. 5, we present how the MSE changes with the average number of samples per worker $\bar{K} = K/U$. The number of data samples per worker fluctuates around the average number, i.e., we set $K_i = \text{round}(\text{uniform}[\bar{K} - 5, \bar{K} + 5])$. As $\bar{K}$ increases, all of the considered learning algorithms have more data samples available for training, and hence the MSE of all of considered FL algorithms decrease in Fig. 5. As the average data samples per worker continues to increase, the MSE improvement slows down and eventually saturates. This is because that as the data
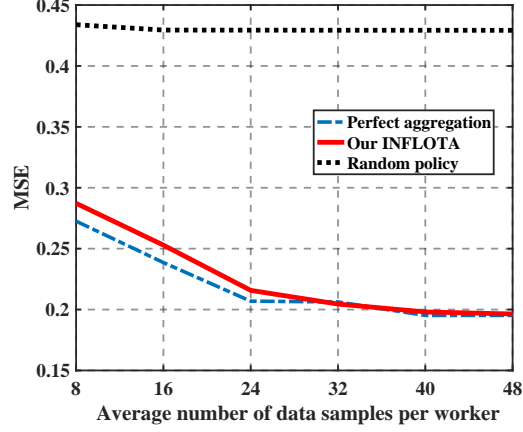
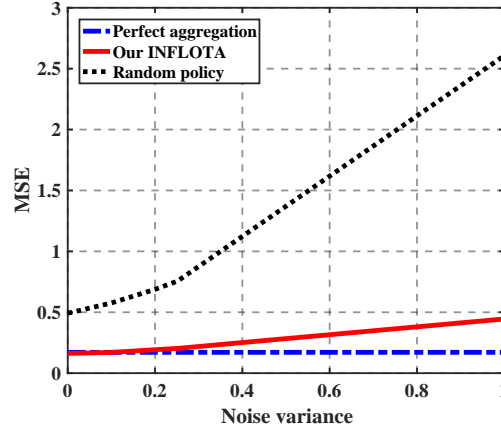Fig. 5: MSE as the number of data samples per worker varies.

Fig. 6: MSE as the noise variance varies.

samples per worker continues to increase, the data samples are enough for training the FL model.

Fig. 6 presents how the AWGN received by the PS affects the MSE. We can see that as the noise variance increases, the MSE values of all of considered FL algorithms increase, except for *Perfect aggregation*. When the noise variance is small (e.g., less than 10^{-1}), it has little effect on the performance of FL algorithms.

B. Evaluation on the MNIST dataset

In order to evaluate the performance of our proposed INFLOTA in realistic application scenarios with real data, we train a multilayer perceptron (MLP) on the MNIST dataset[5] with a

---

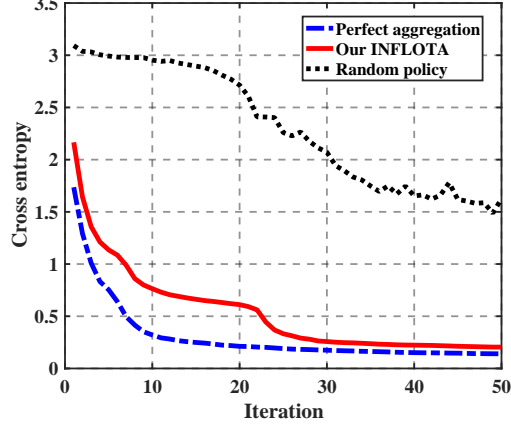[5]http://yann.lecun.com/exdb/mnist/

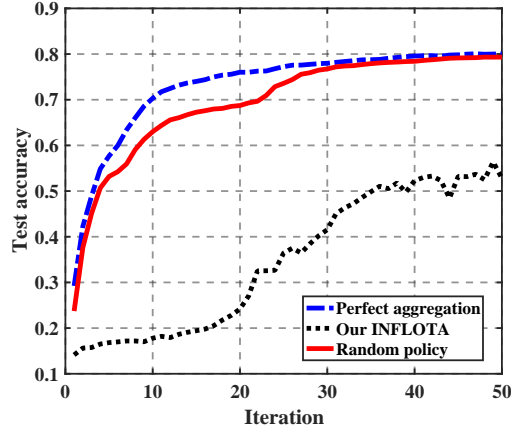Fig. 7: Cross entropy as the number of iteration varies.

Fig. 8: Test accuracy as the number of iteration varies.

784-neuron input layer, a 64-neuron hidden layer, and a 10-neuron softmax output layer, which is a non-convex case. 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 50890. The learning rate α is set as 0.1. In MNIST dataset, there are 60000 training samples and 10000 test samples. We randomly take out $500 - 1000$ training samples and distribute them to 20 local workers as their local data. Then the three trained FL are tested with 10000 test samples. We provide the results of cross entropy and test accuracy versus the iteration index t in Fig. 7 and Fig. 8, respectively. Since the MNIST dataset is designed for handwritten digit identification, the test accuracy presents the identification accuracy. As we can see, our proposed INFLOTA outperforms *Random policy*, and achieves comparable performance as *Perfect aggregation*.

VII. CONCLUSION

In this paper, we have studied the joint optimization of communications and FL over the air with analog aggregation, in which both worker selection and transmit power control are considered under the constraints of limited communication resources. Under the convex and non-convex cases with either the GD or SGD implementations, we respectively derive closed-form expressions for the expected convergence rate of the FL algorithm, which can quantify the impact of resource-constrained wireless communications on FL under the analog aggregation paradigm. Through analyzing the expected convergence rate, we have proposed a joint optimization scheme of worker selection and power control, which can mitigate the impact of wireless communications on the convergence and performance of the FL algorithm. More significantly, our joint optimization scheme is applicable for both the convex and non-convex cases, using either GD or SGD implementations. Simulation results show that the proposed optimization scheme is effective in mitigating the impact of wireless communications on FL.

ACKNOWLEDGMENTS

APPENDIX A

PROOF OF **THEOREM 1**

Theorem 1 considers the full GD method for convex problems. Following the proof for the gradient methods with noise in [33], we first present the inequality implied by the **Assumption 1**, as follows

$$F(\mathbf{w}_t) \leq F(\mathbf{w}_{t-1}) + (\mathbf{w}_t - \mathbf{w}_{t-1})^T \nabla F(\mathbf{w}_{t-1}) + \frac{L}{2}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2. \tag{46}$$

Employing a standard full GD method, the i-th worker updates its local FL model parameter $\mathbf{w}_{i,t}$ at the t-th iteration by

$$\mathbf{w}_{i,t} = \mathbf{w}_{t-1} - \frac{\alpha}{K_i}\sum_{k=1}^{K_i}\nabla f(\mathbf{w}_{t-1}, \mathbf{x}_{i,k}, \mathbf{y}_{i,k}), \quad i = 1, 2, ..., U. \tag{47}$$

Substituting (47) to (9), we have

$$
\begin{aligned}
\mathbf{w}_t =& \mathbf{w}_{t-1} + \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t \right)^{\odot -1} \odot \mathbf{z}_t \\
& - \alpha \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \right)^{\odot -1} \odot \sum_{i=1}^{U} \sum_{k=1}^{K_i} \boldsymbol{\beta}_{i,t} \odot \nabla f(\mathbf{w}_{t-1}; \mathbf{x}_{i,k}, y_{i,k}) \\
=& \mathbf{w}_{t-1} - \alpha (\nabla F(\mathbf{w}_{t-1}) - \mathbf{o}),
\end{aligned}
\tag{48}
$$

where

$$
\begin{aligned}
\mathbf{o} =& \nabla F(\mathbf{w}_{t-1}) + \left(\alpha \sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t \right)^{\odot -1} \odot \mathbf{z}_t \\
& - \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \right)^{\odot -1} \odot \sum_{i=1}^{U} \sum_{k=1}^{K_i} \boldsymbol{\beta}_{i,t} \odot \nabla f(\mathbf{w}_{t-1}; \mathbf{x}_{i,k}, y_{i,k}).
\end{aligned}
\tag{49}
$$

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

$$
\begin{aligned}
\mathbb{E}[F(\mathbf{w}_t)] \leq & \mathbb{E}\left[F(\mathbf{w}_{t-1}) - \alpha (\nabla F(\mathbf{w}_{t-1}) - \mathbf{o})^T \nabla F(\mathbf{w}_{t-1}) + \frac{L\alpha^2}{2} \|\nabla F(\mathbf{w}_{t-1}) - \mathbf{o}\|^2 \right] \\
& \overset{(a)}{=} \mathbb{E}[F(\mathbf{w}_{t-1})] - \frac{1}{2L} \|\nabla F(\mathbf{w}_{t-1})\|^2 + \frac{1}{2L} \mathbb{E}[\|\mathbf{o}\|^2],
\end{aligned}
\tag{50}
$$

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

$$
\frac{L\alpha^2}{2} \|\nabla F(\mathbf{w}_{t-1}) - \mathbf{o}\|^2 = \frac{1}{2L} \|\nabla F(\mathbf{w}_{t-1})\|^2 - \frac{1}{L} \mathbf{o}^T \nabla F(\mathbf{w}_{t-1}) + \frac{1}{2L} \|\mathbf{o}\|^2.
\tag{51}
$$

$\mathbb{E}[\|\mathbf{o}\|^2]$ can be derived as follows

$$
\begin{aligned}
\mathbb{E}[\|\mathbf{o}\|^2] =& \mathbb{E}\left[\left\|\nabla F(\mathbf{w}_{t-1}) + \left(\alpha \sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t\right)^{\odot-1} \odot \mathbf{z}_t \right.\right. \\
& \left.\left. - \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t}\right)^{\odot-1} \odot \sum_{i=1}^{U}\sum_{k=1}^{K_i} \boldsymbol{\beta}_{i,t} \odot \nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})\right\|^2\right] \\
=& \mathbb{E}\left[\left\|\frac{\sum_{i=1}^{U}\sum_{k=1}^{K_i} \nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})}{K} + \left(\alpha \sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t\right)^{\odot-1} \odot \mathbf{z}_t \right.\right. \\
& \left.\left. - \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t}\right)^{\odot-1} \odot \sum_{i=1}^{U}\sum_{k=1}^{K_i} \boldsymbol{\beta}_{i,t} \odot \nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})\right\|^2\right] \\
=& \mathbb{E}\left[\left\|\sum_{i=1}^{U}\left(\frac{\mathbf{1}}{K} - \boldsymbol{\beta}_{i,t} \odot \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t}\right)^{\odot-1}\right) \odot \sum_{k=1}^{K_i} \nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})\right.\right. \\
& \left.\left. + \left(\alpha \sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t\right)^{\odot-1} \odot \mathbf{z}_t\right\|^2\right],
\end{aligned}
\tag{52}
$$

where $\mathbf{1}$ is the all-1 vector of length D. The dimension of $\mathbf{1}$ is the same length as that of $\boldsymbol{\beta}_{i,t}$.

Employing the triangle inequality of norms $\|\mathbf{X} + \mathbf{Y}\| \leq \|\mathbf{X}\| + \|\mathbf{Y}\|$, the submultiplicative property of norms $\|\mathbf{X}\mathbf{Y}\| \leq \|\mathbf{X}\|\|\mathbf{Y}\|$, and the Jensen's inequality, (52) can be further derived as follows

$$
\begin{aligned}
\mathbb{E}[\|\mathbf{o}\|^2] \leq& \mathbb{E}\left[\left\|\sum_{i=1}^{U}\left(\frac{\mathbf{1}}{K} - \boldsymbol{\beta}_{i,t} \odot \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t}\right)^{\odot-1}\right) \odot \sum_{k=1}^{K_i} \nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})\right\|^2\right] \\
& + \mathbb{E}\left[\left\|\left(\alpha \sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t\right)^{\odot-1} \odot \mathbf{z}_t\right\|^2\right] \\
\leq& \mathbb{E}\left[K \sum_{i=1}^{U}\left\|\frac{\mathbf{1}}{K} - \boldsymbol{\beta}_{i,t} \odot \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t}\right)^{\odot-1}\right\|^2 \sum_{k=1}^{K_i} \|\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})\|^2\right] \\
& + \mathbb{E}\left[\left\|\left(\alpha \sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t\right)^{\odot-1} \odot \mathbf{z}_t\right\|^2\right] \\
\leq& K \sum_{i=1}^{U}\left\|\frac{\mathbf{1}}{K} - \boldsymbol{\beta}_{i,t} \odot \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t}\right)^{\odot-1}\right\|^2 \sum_{k=1}^{K_i} \|\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})\|^2 \\
& + \left\|\left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t\right)^{\odot-1}\right\|^2 \sigma^2 L^2.
\end{aligned}
\tag{53}
$$

Applying (12) in **Assumption 3** to (53) leads to

$$
\begin{aligned}
\mathbb{E}[\|\mathbf{o}\|^2] \leq & K \sum_{i=1}^{U} \left\| \frac{1}{K} - \boldsymbol{\beta}_{i,t} \odot \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \right)^{\odot -1} \right\|^2 K_i(\rho_1 + \rho_2 \|\nabla F(\mathbf{w}_{t-1})\|^2) \\
& + \left\| \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t \right)^{\odot -1} \right\|^2 \sigma^2 L^2 \\
= & K \sum_{i=1}^{U} \sum_{d=1}^{D} \left(\frac{1}{K} - \frac{\beta_{i,t}^d}{\sum_{i=1}^{U} K_i \beta_{i,t}^d} \right)^2 K_i(\rho_1 + \rho_2 \|\nabla F(\mathbf{w}_{t-1})\|^2) \\
& + \left\| (\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t)^{\odot -1} \right\|^2 \sigma^2 L^2 \\
= & K \sum_{i=1}^{U} \sum_{d=1}^{D} \left(\frac{1}{K^2} - \frac{2}{K} \frac{\beta_{i,t}^d}{\sum_{i=1}^{U} K_i \beta_{i,t}^d} + \frac{(\beta_{i,t}^d)^2}{(\sum_{i=1}^{U} K_i \beta_{i,t}^d)^2} \right) K_i(\rho_1 + \rho_2 \|\nabla F(\mathbf{w}_{t-1})\|^2) \\
& + \left\| (\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t)^{\odot -1} \right\|^2 \sigma^2 L^2 \\
= & \sum_{d=1}^{D} \left(\frac{K}{\sum_{i=1}^{U} K_i \beta_{i,t}^d} - 1 \right) (\rho_1 + \rho_2 \|\nabla F(\mathbf{w}_{t-1})\|^2) + \left\| (\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t)^{\odot -1} \right\|^2 \sigma^2 L^2.
\end{aligned}
\tag{54}
$$

Substituting (54) to (50), we have:

$$
\begin{aligned}
\mathbb{E}[F(\mathbf{w}_t)] \leq & \frac{1}{2L} \left(\sum_{d=1}^{D} \left(\frac{K}{\sum_{i=1}^{U} K_i \beta_{i,t}^d} - 1 \right) (\rho_1 + \rho_2 \|\nabla F(\mathbf{w}_{t-1})\|^2) \right. \\
& \left. + \left\| (\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t)^{\odot -1} \right\|^2 \sigma^2 L^2 \right) + \mathbb{E}[F(\mathbf{w}_{t-1})] - \frac{1}{2L} \|\nabla F(\mathbf{w}_{t-1})\|^2.
\end{aligned}
\tag{55}
$$

Subtract $\mathbb{E}[F(\mathbf{w}^*)]$ from both sides of (55), we have:

$$
\begin{aligned}
\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \leq & \mathbb{E}[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)] - \frac{1}{2L} \|\nabla F(\mathbf{w}_{t-1})\|^2 \\
& + \frac{1}{2L} \left(\sum_{d=1}^{D} \left(\frac{K}{\sum_{i=1}^{U} K_i \beta_{i,t}^d} - 1 \right) (\rho_1 + \rho_2 \|\nabla F(\mathbf{w}_{t-1})\|^2) + \left\| (\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t)^{\odot -1} \right\|^2 \sigma^2 L^2 \right).
\end{aligned}
\tag{56}
$$

To minimize both sides of (11), we have

$$
\min_{\mathbf{w}_t} F(\mathbf{w}_t) \geq \min_{\mathbf{w}_t} (F(\mathbf{w}_{t-1}) + (\mathbf{w}_t - \mathbf{w}_{t-1})^T \nabla F(\mathbf{w}_{t-1}) + \frac{\mu}{2} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2).
\tag{57}
$$

The minimization of the left-hand side is achieved by $\mathbf{w}_t = \mathbf{w}^*$, while the minimization of the right-hand side is achieved by $\mathbf{w}_t = \mathbf{w}_{t-1} - \frac{1}{\mu}\nabla F(\mathbf{w}_{t-1})$. Thus, we have

$$F(\mathbf{w}^*) \geq F(\mathbf{w}_{t-1}) - \frac{1}{2\mu}\|\nabla F(\mathbf{w}_{t-1})\|^2. \tag{58}$$

Then

$$\|\nabla F(\mathbf{w}_{t-1})\|^2 \geq 2\mu(F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)). \tag{59}$$

Substituting (59) to (56), we get

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \leq (1 - \frac{\mu}{L})\mathbb{E}[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)]$$

$$+ \frac{1}{2L}\left(\sum_{d=1}^{D}\left(\frac{K}{\sum_{i=1}^{U} K_i\beta_{i,t}^d} - 1\right)(\rho_1 + \rho_2\|\nabla F(\mathbf{w}_{t-1})\|^2) + \left\|(\sum_{i=1}^{U} K_i\boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t)^{\odot-1}\right\|^2 \sigma^2 L^2\right)$$

$$= (1 - \frac{\mu}{L})\mathbb{E}[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)] + \frac{\rho_2}{2L}\left(\sum_{d=1}^{D}\left(\frac{K}{\sum_{i=1}^{U} K_i\beta_{i,t}^d} - 1\right)\right)\|\nabla F(\mathbf{w}_{t-1})\|^2$$

$$+ \frac{\rho_1}{2L}\sum_{d=1}^{D}\left(\frac{K}{\sum_{i=1}^{U} K_i\beta_{i,t}^d} - 1\right) + \left\|(\sum_{i=1}^{U} K_i\boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t)^{\odot-1}\right\|^2 \frac{L\sigma^2}{2}. \tag{60}$$

Next, in the same way that (59) is derived, to minimize both sides of (46), we have

$$\|\nabla F(\mathbf{w}_{t-1})\|^2 \leq 2L(F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)). \tag{61}$$

Substituting (61) to (60), we get

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \leq \left(1 - \frac{\mu}{L} + \rho_2\sum_{d=1}^{D}\left(\frac{K}{\sum_{i=1}^{U} K_i\beta_{i,t}^d} - 1\right)\right)\mathbb{E}[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)]$$

$$+ \frac{\rho_1}{2L}\sum_{d=1}^{D}\left(\frac{K}{\sum_{i=1}^{U} K_i\beta_{i,t}^d} - 1\right) + \left\|(\sum_{i=1}^{U} K_i\boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t)^{\odot-1}\right\|^2 \frac{L\sigma^2}{2}$$

$$= B_t + A_t\mathbb{E}[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)], \tag{62}$$

where

$$A_t = 1 - \frac{\mu}{L} + \rho_2\sum_{d=1}^{D}\left(\frac{K}{\sum_{i=1}^{U} K_i\beta_{i,t}^d} - 1\right), \tag{63}$$

$$B_t = \frac{\rho_1}{2L}\sum_{d=1}^{D}\left(\frac{K}{\sum_{i=1}^{U} K_i\beta_{i,t}^d} - 1\right) + \left\|(\sum_{i=1}^{U} K_i\boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t)^{\odot-1}\right\|^2 \frac{L\sigma^2}{2}. \tag{64}$$

The proof is completed.

APPENDIX B

PROOF OF **THEOREM 2**

Theorem 2 considers the full GD method for non-convex problems. The proof of **Theorem 2** follows that of **Theorem 1** until (55). From (55), we have

$$
\mathbb{E}[F(\mathbf{w}_t)] \leq \mathbb{E}[F(\mathbf{w}_{t-1})] - \frac{1}{2L}\left(1 - \rho_2 \sum_{d=1}^{D}\left(\frac{K}{\sum_{i=1}^{U} K_i \beta_{i,t}^d} - 1\right)\right)\|\nabla F(\mathbf{w}_{t-1})\|^2
$$
$$
+ \frac{\rho_1}{2L}\sum_{d=1}^{D}\left(\frac{K}{\sum_{i=1}^{U} K_i \beta_{i,t}^d} - 1\right) + \left\|(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t)^{\odot -1}\right\|^2 \frac{L\sigma^2}{2}
$$
$$
= \mathbb{E}[F(\mathbf{w}_{t-1})] - \frac{2 - A_t - \frac{\mu}{L}}{2L}\|\nabla F(\mathbf{w}_{t-1})\|^2 + B_t. \tag{65}
$$

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

$$
\mathbb{E}[F(\mathbf{w}_t)] - \mathbb{E}[F(\mathbf{w}_0)] \leq -\sum_{t=1}^{T}\frac{2 - A_t - \frac{\mu}{L}}{2L}\|\nabla F(\mathbf{w}_{t-1})\|^2 + \sum_{t=1}^{T} B_t, \tag{66}
$$

which leads to

$$
\sum_{t=1}^{T}\frac{2 - A_t - \frac{\mu}{L}}{2L}\|\nabla F(\mathbf{w}_{t-1})\|^2 \leq \mathbb{E}[F(\mathbf{w}_0)] - \mathbb{E}[F(\mathbf{w}_t)] + \sum_{t=1}^{T} B_t
$$
$$
\leq \mathbb{E}[F(\mathbf{w}_0)] - \mathbb{E}[F(\mathbf{w}^*)] + \sum_{t=1}^{T} B_t. \tag{67}
$$

Recalling **Proposition 1**, we have

$$
0 \leq \frac{1 - \rho_2 D(\frac{K}{K_{min}} - 1)}{2L} \leq \frac{2 - A_t - \frac{\mu}{L}}{2L} \leq \frac{1}{2L}, \quad \forall t. \tag{68}
$$

Substituting (68) to (67), we get

$$
\frac{1}{T}\sum_{t=1}^{T}\frac{1 - \rho_2 D(\frac{K}{K_{min}} - 1)}{2L}\|\nabla F(\mathbf{w}_{t-1})\|^2 \leq \frac{1}{T}\sum_{t=1}^{T}\frac{2 - A_t - \frac{\mu}{L}}{2L}\|\nabla F(\mathbf{w}_{t-1})\|^2
$$
$$
\leq \frac{1}{T}(\mathbb{E}[F(\mathbf{w}_0)] - \mathbb{E}[F(\mathbf{w}^*)]) + \frac{1}{T}\sum_{t=1}^{T} B_t. \tag{69}
$$

As a result, we have the conclusion in **Theorem 2**,

$$
\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_{t-1})\|^2 \leq \frac{2L}{T(1 - \rho_2 D(\frac{K}{K_{min}} - 1))}\mathbb{E}[F(\mathbf{w}_0)] - \mathbb{E}[F(\mathbf{w}^*)]
$$
$$
+ \frac{2L\sum_{t=1}^{T} B_t}{T(1 - \rho_2 D(\frac{K}{K_{min}} - 1))}. \tag{70}
$$

APPENDIX C

PROOF OF **THEOREM 3**

Exploiting the SGD method, the local parameter of the i-th worker is updated at the t-th iteration by

$$\mathbf{w}_{i,t} = \mathbf{w}_{t-1} - \alpha \mathbb{E}_{\mathcal{D}_i} \left[\frac{\sum_{k=1}^{K_b} \nabla f(\mathbf{w}_{t-1}, \mathbf{x}_{i,k}, \mathbf{y}_{i,k})}{K_b} \right], \quad i = 1, 2, ..., U, \tag{71}$$

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

Substituting (71) to (9), we reach a averaged gradient estimate as

$$
\begin{aligned}
\mathbf{w}_t =& \mathbf{w}_{t-1} + \left(\sum_{i=1}^{U} K_b \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t \right)^{\odot -1} \odot \mathbf{z}_t \\
& - \alpha \left(\sum_{i=1}^{U} K_b \boldsymbol{\beta}_{i,t} \right)^{\odot -1} \odot \sum_{i=1}^{U} \left(K_b \boldsymbol{\beta}_{i,t} \odot \mathbb{E}_{\mathcal{D}_i} \left[\frac{\sum_{k=1}^{K_b} \nabla f(\mathbf{w}_{t-1}, \mathbf{x}_{i,k}, \mathbf{y}_{i,k})}{K_b} \right] \right) \\
=& \mathbf{w}_{t-1} - \alpha (\nabla F(\mathbf{w}_{t-1}) - \mathbf{o}),
\end{aligned}
\tag{72}
$$

where

$$
\begin{aligned}
\mathbf{o} =& \nabla F(\mathbf{w}_{t-1}) + \left(\alpha \sum_{i=1}^{U} K_b \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t \right)^{\odot -1} \odot \mathbf{z}_t \\
& - \left(\sum_{i=1}^{U} K_b \boldsymbol{\beta}_{i,t} \right)^{\odot -1} \odot \sum_{i=1}^{U} \left(\boldsymbol{\beta}_{i,t} \odot \mathbb{E}_{\mathcal{D}_i} \left[\sum_{k=1}^{K_b} \nabla f(\mathbf{w}_{t-1}, \mathbf{x}_{i,k}, \mathbf{y}_{i,k}) \right] \right).
\end{aligned}
\tag{73}
$$

Let $\mathcal{N}_{i,t}$ denote the set of the samples that are not chosen by the i-th worker at the t-th

iteration, $\mathbb{E}[\|\mathbf{o}\|^2]$ can be derived as follows

$$
\begin{aligned}
\mathbb{E}[\|\mathbf{o}\|^2] =&\, \mathbb{E}\left[\left\|\frac{\sum_{i=1}^{U}\sum_{k=1}^{K_i}\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})}{K} + \left(\alpha\sum_{i=1}^{U}K_b\boldsymbol{\beta}_{i,t}\odot\boldsymbol{b}_t\right)^{\odot-1}\odot\mathbf{z}_t\right.\right. \\
&\left.\left. - \left(\sum_{i=1}^{U}K_b\boldsymbol{\beta}_{i,t}\right)^{\odot-1}\odot\sum_{i=1}^{U}\left(\boldsymbol{\beta}_{i,t}\odot\mathbb{E}_{\mathcal{D}_i}\left[\sum_{k=1}^{K_b}\nabla f(\mathbf{w}_{t-1},\mathbf{x}_{i,k},\mathbf{y}_{i,k})\right]\right)\right\|^2\right] \\
=&\,\mathbb{E}\left[\left\|\sum_{i=1}^{U}\left(\frac{\mathbf{1}}{K}-\boldsymbol{\beta}_{i,t}\odot\left(\sum_{i=1}^{U}K_b\boldsymbol{\beta}_{i,t}\right)^{\odot-1}\right)\odot\mathbb{E}_{\overline{\mathcal{N}}_{i,t}}\left[\sum_{k\in\overline{\mathcal{N}}_{i,t}}\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})\right]\right.\right. \\
&\left.\left. + \frac{\sum_{i=1}^{U}\mathbb{E}[\sum_{k\in\mathcal{N}_{i,t}}\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})]}{K} + \left(\alpha\sum_{i=1}^{U}K_b\boldsymbol{\beta}_{i,t}\odot\boldsymbol{b}_t\right)^{\odot-1}\odot\mathbf{z}_t\right\|^2\right], \\
\leq&\,\left(\sum_{i=1}^{U}K_b\right)\sum_{i=1}^{U}\left\|\frac{\mathbf{1}}{K}-\boldsymbol{\beta}_{i,t}\odot\left(\sum_{i=1}^{U}K_b\boldsymbol{\beta}_{i,t}\right)^{\odot-1}\right\|^2\mathbb{E}_{\overline{\mathcal{N}}_{i,t}}\left[\sum_{k\in\overline{\mathcal{N}}_{i,t}}\|\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})\|^2\right] \\
&\, + \frac{\|\sum_{i=1}^{U}\mathbb{E}[\sum_{k\in\mathcal{N}_{i,t}}\nabla f(\mathbf{w}_{t-1};\mathbf{x}_{i,k},\mathbf{y}_{i,k})]\|^2}{K^2} + \left\|\left(\sum_{i=1}^{U}K_b\boldsymbol{\beta}_{i,t}\odot\boldsymbol{b}_t\right)^{\odot-1}\right\|^2\sigma^2 L^2.
\end{aligned}
\tag{74}
$$

Applying **Assumption 3**, we get

$$
\begin{aligned}
\mathbb{E}[\|\mathbf{o}\|^2] \leq&\, \left(\sum_{i=1}^{U}K_b\right)\sum_{d=1}^{D}\left(\frac{\left(\sum_{i=1}^{U}K_b\right)-2K}{K^2}+\frac{1}{\sum_{i=1}^{U}K_b\beta_{i,t}^d}\right)(\rho_1+\rho_2\|\nabla F(\mathbf{w}_{t-1})\|^2)+ \\
&\, \frac{(\sum_{i=1}^{U}(K_i-K_b))^2}{K^2}(\rho_1+\rho_2\|\nabla F(\mathbf{w}_{t-1})\|^2) + \left\|\left(\sum_{i=1}^{U}K_b\boldsymbol{\beta}_{i,t}\odot\boldsymbol{b}_t\right)^{\odot-1}\right\|^2\sigma^2 L^2.
\end{aligned}
\tag{75}
$$

Substituting (75) into (50), we have

$$
\begin{aligned}
\mathbb{E}[F(\mathbf{w}_t)] \leq&\, \frac{1}{2L}\left(\left\|\left(\sum_{i=1}^{U}K_b\boldsymbol{\beta}_{i,t}\odot\boldsymbol{b}_t\right)^{\odot-1}\right\|^2\sigma^2 L^2\right. \\
&\, + \left(\sum_{d=1}^{D}\left(\frac{\left(\sum_{i=1}^{U}K_b\right)^2-2K\left(\sum_{i=1}^{U}K_b\right)}{K^2}+\frac{\left(\sum_{i=1}^{U}K_b\right)}{\sum_{i=1}^{U}K_b\beta_{i,t}^d}\right)\right. \\
&\, \left.\left. + \frac{(\sum_{i=1}^{U}(K_i-K_b))^2}{K^2}\right)\right)(\rho_1+\rho_2\|\nabla F(\mathbf{w}_{t-1})\|^2)\bigg) \\
&\, + \mathbb{E}[F(\mathbf{w}_{t-1})] - \frac{1}{2L}\|\nabla F(\mathbf{w}_{t-1})\|^2.
\end{aligned}
\tag{76}
$$

Subtracting $\mathbb{E}[F(\mathbf{w}^*)]$ from both sides of (76), and applying (59) and (61), we get

$$
\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \leq B_t^{SGD} + A_t^{SGD}\mathbb{E}[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)],
\tag{77}
$$

where

$$A_t^{SGD} = 1 - \frac{\mu}{L} + \rho_2 \left(\sum_{d=1}^{D} \left(\frac{(\sum_{i=1}^{U} K_b)^2 - 2K(\sum_{i=1}^{U} K_b)}{K^2} + \frac{(\sum_{i=1}^{U} K_b)}{\sum_{i=1}^{U} K_b \beta_{i,t}^d} \right) \right.$$
$$\left. + \frac{(\sum_{i=1}^{U}(K_i - K_b))^2}{K^2} \right), \tag{78}$$

$$B_t^{SGD} = \frac{\rho_1}{2L} \left(\sum_{d=1}^{D} \left(\frac{(\sum_{i=1}^{U} K_b)^2 - 2K(\sum_{i=1}^{U} K_b)}{K^2} + \frac{(\sum_{i=1}^{U} K_b)}{\sum_{i=1}^{U} K_b \beta_{i,t}^d} \right) \right.$$
$$\left. + \frac{(\sum_{i=1}^{U}(K_i - K_b))^2}{K^2} \right) + \left\| \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t \right)^{\odot -1} \right\|^2 \frac{L\sigma^2}{2}. \tag{79}$$

Applying (77) recursively, we have

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \le \sum_{i=1}^{t-1} \prod_{j=1}^{i} A_{t+1-j}^{SGD} B_{t-i}^{SGD} + B_t^{SGD} + \prod_{j=1}^{t} A_j^{SGD} \mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)]. \tag{80}$$

which completes the proof.

APPENDIX D

PROOF OF **THEOREM 4**

To minimize R_t, it can be seen from (35), (36) and (37) that we should maximize the number of the selected workers and the transmit power scaling factor in the t-th iteration. Thus, the selected workers should send their parameters at their maximum power. In order to reach the desired parameter aggregation at the PS as in (5), each worker needs to use the same transmit power scaling factor b_t, which is a parameter that needs to be optimized (b_t determines the worker selection). According to (35), (36) and (37), a larger b_t leads to a smaller R_t. On the other hand, (41b) indicates that a larger b_t results in less workers is selected, which then results in an increase of R_t.

Rewriting (38b) and replacing $|w_{i,t}|$ with $(|w_{t-1}| + \eta)$, we obtain the maximum acceptable b_t of the i-th worker as

$$b_{i,t}^{\max} = \left| \frac{\sqrt{P_i^{\max}} h_{i,t}}{K_i(|w_{t-1}| + \eta)} \right|. \tag{81}$$

Accordingly, b_t should be chosen from $\{b_{i,t}^{\max}\}_{i=1}^{U}$. Once b_t is determined, $\boldsymbol{\beta}_t$ can be determined by verifying whether the transmit power meets the condition in (7). As a result, we obtain a reduced solution space of the optimization problem **P2** as

$$\mathcal{S} = \left\{ \{(b_t^{(k)}, \beta_{i,t}^{(k)})\}_{k=1}^{U} \middle| b_t^{(k)} = b_{k,t}^{\max}, \boldsymbol{\beta}_t^{(k)}(b_t^{(k)}) = [\beta_{1,t}^{(k)}, \dots, \beta_{U,t}^{(k)}], k = 1, \dots, U \right\}, \tag{82}$$

with

$$\beta_{U,t}^{(k)} = H\left(P_U^{\max} - \left| \frac{K_U b_t^{(k)}(|w_{t-1}| + \eta)}{h_{U,t}} \right| \right) \tag{83}$$

where

$$H(x) = \begin{cases} 1, & x > 0, \\ 0, & x \leq 0. \end{cases} \tag{84}$$

is the Heaviside step function.

REFERENCES

[1] M. Chiang and T. Zhang, "Fog and iot: An overview of research opportunities," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 854–864, 2016.

[2] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proceedings of the IEEE*, vol. 107, no. 11, pp. 2204–2239, 2019.

[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] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

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

[7] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, 2020.

[8] 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," *IEEE Transactions on Wireless Communications*, 2020.

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

[10] L. Chen, N. Zhao, Y. Chen, F. R. Yu, and G. Wei, "Over-the-air computation for iot networks: Computing multiple functions with antenna arrays," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 5296–5306, 2018.

[11] M. Goldenbaum, H. Boche, and S. Stańczak, "Harnessing interference for analog function computation in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 61, no. 20, pp. 4893–4906, 2013.

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

[13] 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.

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

[15] 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.

[16] 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.

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

[18] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2022–2035, 2020.

[19] M. Ye and E. Abbe, "Communication-computation efficient gradient coding," *arXiv preprint arXiv:1802.03475*, 2018.

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

[21] 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.

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

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

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] M. Goldenbaum and S. Stanczak, "Robust analog function computation via wireless multiple-access channels," *IEEE Transactions on Communications*, vol. 61, no. 9, pp. 3863–3877, 2013.

[30] O. Shamir, N. Srebro, and T. Zhang, "Communication-efficient distributed optimization using an approximate newton-type method," in *International conference on machine learning*, 2014, pp. 1000–1008.

[31] S. Magnússon, H. Shokri-Ghadikolaei, and N. Li, "On maintaining linear convergence of distributed learning and optimization under limited communication," *IEEE Transactions on Signal Processing*, 2020.

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

[33] 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.

[34] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, "The convergence of sparsified gradient methods," in *Advances in Neural Information Processing Systems*, 2018, pp. 5973–5983.

[35] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1835–1854, 2016.

[36] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *Siam Review*, vol. 60, no. 2, pp. 223–311, 2018.

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

[38] H. Tang, C. Yu, X. Lian, T. Zhang, and J. Liu, "Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6155–6165.

[39] 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.