Joint Optimization for Federated Learning Over the Air

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

ICC 2022 - IEEE International Conference on Communications | 978-1-5386-8347-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICC45855.2022.9838269

***Abstract*—In this paper, we focus on federated learning (FL) over the air based on analog aggregation transmission in realistic wireless networks. We first derive a closed-form expression for the expected convergence rate of FL over the air, which theoretically quantifies the impact of analog aggregation on FL. Based on that, we further develop a joint optimization model for accurate FL implementation, which allows a parameter server to select a subset of edge devices and determine an appropriate power scaling factor. Such a joint optimization of device selection and power control for FL over the air is then formulated as an mixed integer programming problem. Finally, we efficiently solve this problem via a simple finite-set search method. Simulation results show that the proposed solutions developed for wireless channels outperform a benchmark method, and could achieve comparable performance of the ideal case where FL is implemented over reliable and error-free wireless channels.**

***Index Terms*—Federated learning, analog aggregation, convergence analysis, joint optimization, worker scheduling, power scaling.**

I. Introduction

Recently, federated learning (FL) has been proposed as a well acknowledged approach for collaborative edge learning [1], [2]. In FL, edge devices (local workers) train local models from their own local data, and then transmit their local updates to a parameter server (PS). After aggregating these received local updates, the PS feeds back the averaged update to the local workers. These iterative updates between PS and workers, can be either model parameters for model averaging [1] or parameters' gradients for gradient averaging [2]. In this way, FL relieves communication overhead and protect user privacy compared to raw data sharing of traditional collaborative learning, specially when the local data is in large volume and privacy-sensitive. Existing work on FL focuses on FL algorithms given idealized link assumptions, but the impact of wireless environments on FL performance should be taken into account in the design of FL deployed in real wireless systems. Otherwise, the inherent characteristics of wireless links may introduce unwanted training errors that dramatically degrade the learning performance in terms of accuracy and convergence rate.

To solve this problem, research efforts have been spent on optimizing network resources used for transmitting model updates in FL [3]. 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) [4] provides a direct and efficient way for transmission of the averaged updates in FL, also known as analog aggregation based FL [5]–[10]. 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 [11], [12].

While there exist works of analog aggregation based FL [5]–[10], some of them mainly focus on designing transmission schemes without optimization [5]–[8], where they adopt preselected participating workers and fixed their power allocation. Although optimization issues are considered in [9], [10], the optimization is conducted on communication side alone, without an underlying connection to FL. Noticeably, while the optimization in existing works boils down to maximizing the number of selected workers, our theoretical results indicate that more workers is not necessary better over imperfect links and with limited communication resources. Thus, unlike these existing works, we seek to analyze the convergence behavior of analog aggregation based FL, which interprets the specific relationship between communications and FL in the paradigm of analog aggregation. Such a meaningful connection leads to a joint optimization framework for analog communications and FL. 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.
- 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. The optimization formulation

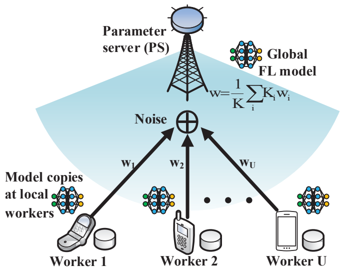

Fig. 1: FL via analog aggregation from wirelessly distributed data.

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.

II. System Model

As shown in Fig. 1, we consider a one-hop wireless network consisting of a single PS at a base station and U user devices as distributed local workers. Through FL, 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

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

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

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

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 [1], [13]

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

To solve **P1**, this paper adopts a model-averaging algorithm for FL [1], [13], where gradient descent is applied, and then the local model at the i-th local worker is updated as

$$\text{(Local model updating) } \mathbf{w}_i = \mathbf{w} - \frac{\alpha}{K_i} \sum_{k=1}^{K_i} \nabla f(\mathbf{w}; \mathbf{x}_{i,k}, \mathbf{y}_{i,k}), \quad (4)$$

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

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

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.

B. Analog Aggregation Transmission Model

To avoid heavy communication overhead, we adopt analog aggregation without coding, which allows multiple workers to simultaneously upload their updates to the PS over the same time-frequency resources. 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 $\mathbf{p}_{i,t} = [p_{i,t}^1, \cdots, p_{i,t}^d, \cdots, p_{i,t}^D]$ 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 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

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

where $h_{i,t}$ is the channel gain between the i-th worker and the PS at the t-th iteration[1], 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:

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

[1] 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.

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

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

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

$$\mathbf{w}_t = \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \mathbf{b}_t \right)^{\odot -1} \odot \mathbf{y}_t = \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t \right)^{\odot -1} \odot \mathbf{z}_t + \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \right)^{\odot -1} \sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \mathbf{w}_{i,t}, \tag{8}$$

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 (8) 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 (8).

III. 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 [3], [14], [15].

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,

$$\|\nabla F(\mathbf{w}_{t+1}) - \nabla F(\mathbf{w}_t)\| \leq L \|\mathbf{w}_{t+1} - \mathbf{w}_t\|, \quad \forall \mathbf{w}_t, \mathbf{w}_{t+1}, \tag{9}$$

where L is a positive constant.

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

$$F(\mathbf{w}_{t+1}) \geq F(\mathbf{w}_t) + (\mathbf{w}_{t+1} - \mathbf{w}_t)^T \nabla F(\mathbf{w}_t) + \frac{\mu}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2, \quad \forall \mathbf{w}_t, \mathbf{w}_{t+1}. \tag{10}$$

Assumption 3 (bounded local gradients): The sample-wised local gradients at local workers are bounded by their global counterpart [14], [15]

$$\|\nabla f(\mathbf{w}_t)\|^2 \leq \rho_1 + \rho_2 \|\nabla F(\mathbf{w}_t)\|^2, \tag{11}$$

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

According to [2], [16], 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 the model updating rule for $\mathbf{w}_t$ of the FL-over-the-air scheme is given by (8), $\forall t$. Given the learning rate $\alpha = \frac{1}{L}$, 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*

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \leq \underbrace{\sum_{i=1}^{t-1} \prod_{j=1}^{i} A_{t+1-j} B_{t-i} + B_t}_{\Delta_t} + \prod_{j=1}^{t} A_j \mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)], \tag{12}$$

where $A_t = 1 - \frac{\mu}{L} + \rho_2 \sum_{d=1}^{D} (\frac{K}{\sum_{i=1}^{U} K_i \beta_{i,t}^d} - 1)$ and $B_t = \frac{\rho_1}{2L} \sum_{d=1}^{D} (\frac{K}{\sum_{i=1}^{U} K_i \beta_{i,t}^d} - 1) + \|(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t)^{\odot -1}\|^2 \frac{L\sigma^2}{2}$.

Proof. All the proofs, which are omitted in this paper due to the page limit, can be found in our journal version at [17]: https://arxiv.org/pdf/2104.03490.pdf □

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 learning rate $\alpha = \frac{1}{L}$, the convergence at the T-th iteration is given by*

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla F(\mathbf{w}_{t-1})\|^2 \leq \frac{2L}{T(1 - \rho_2 D(\frac{K}{K_{min}} - 1))} \mathbb{E}[F(\mathbf{w}_0)] - F(\mathbf{w}^*)] + \frac{2L \sum_{t=1}^{T} B_t}{T(1 - \rho_2 D(\frac{K}{K_{min}} - 1))}. \tag{13}$$

Proof. Please refer to our journal version [17]. □

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

$$\min_{0,1,...,T} \mathbb{E}[\|\nabla F(\mathbf{w}_{t-1})\|^2] \leq \frac{1}{T} \sum_{t=1}^{T} \|\nabla F(\mathbf{w}_{t-1})\|^2 \overset{T \to \infty}{\leq} \underbrace{\frac{2L \sum_{t=1}^{T} B_t}{T(1 - \rho_2 D(\frac{K}{K_{min}} - 1))}}_{\triangle_T^{NC}}, \tag{14}$$

which guarantees convergence of the FL algorithm to a stationary point [13]. Similarly, the performance gap at the step t for non-convex cases is given by $\triangle_t^{NC} = \frac{2L \sum_{t=1}^{T} B_t}{T(1-\rho_2 D(\frac{K}{K_{min}} -1))}$.

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. 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, given 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*

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \leq \underbrace{\sum_{i=1}^{t-1}\prod_{j=1}^{i} A_{t+1-j}^{SGD} B_{t-i}^{SGD} + B_t^{SGD}}_{\triangle_t^{SGD}} + \prod_{j=1}^{t} A_j^{SGD}\mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)], \quad (15)$$

where $A_t^{SGD} = 1 - \frac{\mu}{L} + \rho_2(\sum_{d=1}^{D}(\frac{(\sum_{i=1}^{U} K_b)^2 - 2K(\sum_{i=1}^{U} K_b)}{K^2} + \frac{(\sum_{i=1}^{U} K_b)}{\sum_{i=1}^{U} K_b \beta_{i,t}^d}) + \frac{(\sum_{i=1}^{U}(K_i - K_b))^2}{K^2})$ *and* $B_t^{SGD} = \frac{\rho_1}{2L}(\sum_{d=1}^{D}(\frac{(\sum_{i=1}^{U} K_b)^2 - 2K(\sum_{i=1}^{U} K_b)}{K^2} + \frac{(\sum_{i=1}^{U} K_b)}{\sum_{i=1}^{U} K_b \beta_{i,t}^d}) + \frac{(\sum_{i=1}^{U}(K_i - K_b))^2}{K^2}) + \left\|\left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t\right)^{\odot -1}\right\|^2 \frac{L\sigma^2}{2}.$

Proof. Please refer to our journal version [17]. □

From **Theorem 3**, the cumulative performance gap of FL after the t-th iteration for the SGD case is bounded by $\triangle_t^{SGD} = \sum_{i=1}^{t-1}\prod_{j=1}^{i} A_{t+1-j}^{SGD} B_{t-i}^{SGD} + B_t^{SGD}$.

IV. 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. 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

$$\triangle_t = B_t + A_t \triangle_{t-1}, \quad (16)$$

$$\triangle_t^{NC} = B_t, \quad (17)$$

$$\triangle_t^{SGD} = B_t^{SGD} + A_t^{SGD} \triangle_{t-1}^{SGD}. \quad (18)$$

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 (16), (17) and (18) as the objective to minimize, which is given by

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

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

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

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 $\beta_{i,t}$ and b_t at the t-th iteration, the PS carries out a joint optimization problem formulated as follows:

$$\textbf{P2:} \quad \min_{\{b_t, \beta_{i,t}\}_{i=1}^{U}} R_t \quad (19a)$$

$$\text{s.t.} \left|\frac{\beta_{i,t} K_i b_t}{h_{i,t}} w_{i,t}\right|^2 \leq P_i^{\max}, \quad (19b)$$

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

where K_i should be K_b in (19b) for the SGD case.

However, in (19b), 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 $\mathbf{w}_{t-1} \approx \frac{1}{U}\sum_{i=1}^{U} \mathbf{w}_{i,t}$. According to FL, 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 [13], [18].

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

$$|w_{t-1} - w_{i,t}| \leq \eta, \quad (20)$$

where $\eta \geq 0$ is related to the learning rate α that satisfies the following condition[2]

$$\eta \geq \max\left\{\left\{\left|\frac{\alpha}{K_i}\sum_{k=1}^{K_i} \nabla f(w, \mathbf{x}_{i,k}, \mathbf{y}_{i,k})\right|\right\}_{i=1}^{U}\right\}. \quad (21)$$

Under **Assumption 4**, we reformulate the original optimization problem **(P2)** into the following problem (**P3**), by replacing $w_{i,t}$ in (19b) by its approximation:

$$\textbf{P3:} \quad \min_{\{b_t, \beta_{i,t}\}_{i=1}^{U}} R_t \quad (22a)$$

$$\text{s.t.} \left|\frac{\beta_{i,t} K_i b_t}{h_{i,t}}\right|^2 (|w_{t-1}| + \eta)^2 \leq P_i^{\max}, \quad (22b)$$

$$\beta_{i,t} \in \{0, 1\}, i \in \{1, 2, ..., U\}, \quad (22c)$$

[2] 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 $\eta \geq \max\{\{|\alpha \mathbb{E}_{\mathcal{D}_i}[\nabla f(w, \mathbf{x}_{i,k}, \mathbf{y}_{i,k})]|\}_{i=1}^{U}\}$

where the power constraint (22b) is constructed based on the fact that $|\frac{\beta_{i,t}K_i b_t}{h_{i,t}}w_{i,t}|^2 = |\frac{\beta_{i,t}K_i b_t}{h_{i,t}}|^2|w_{i,t}|^2 \leq |\frac{\beta_{i,t}K_i b_t}{h_{i,t}}|^2(|w_{t-1}|+\eta)^2$.

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 (22b) and (22c), irrespective of the objective function (22a). Hence, it holds universally for any R_t, R_t^{NC}, and R_t^{SGD}.

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*

$$\mathcal{S}=\left\{\left\{\left(b_t^{(k)},\beta_{i,t}^{(k)}\right)\right\}_{k=1}^{U}\middle| b_t^{(k)}=\left|\frac{\sqrt{P_k^{\max}}h_{k,t}}{K_k(|w_{t-1}|+\eta)}\right|,\right.$$
$$\left.\boldsymbol{\beta}_t^{(k)}(b_t^{(k)})=\left[\beta_{1,t}^{(k)},\dots,\beta_{U,t}^{(k)}\right], k=1,\dots,U\right\}, \tag{23}$$

where $\boldsymbol{\beta}_t^{(k)}$ is a function of $b_t^{(k)}$, in the form $\beta_{i,t}^{(k)} = H(P_i^{\max} - |\frac{K_i b_t^{(k)}(|w_{t-1}|+\eta)}{h_{i,t}}|)$ and $H(x)$ is the Heaviside step function, i.e., $H(x)=1$ for $x>0$, and $H(x)=0$ otherwise.

Proof. Please see our journal version [17]. □

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

$$\textbf{P4:} \quad \min_{(b_t,\boldsymbol{\beta}_t)\in\mathcal{S}} R_t = R_t\left(b_t,\boldsymbol{\beta}_t\right). \tag{24}$$

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 (23).

Putting together, we propose a jo**i**nt optimizatio**n** for **FL** **o**ver **t**he **a**ir (INFLOTA), 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.

Remark 1. (**Complexity**) Our INFLOTA 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**.

V. 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.

In the considered network, we set $U=20$, $P_i^{\max}=P^{\max}=10$ mW for any $i\in[1,U]$, and $\sigma^2=10^{-4}$ mW. The wireless channel gain $h_{i,t}$ 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.

A. Linear regression experiments

In linear regression experiments, the synthetic data used to train FL is generated randomly from $[0,1]$. The input x and the output y follow the function $y=-2x+1+n\times 0.4$ where $n\sim\mathcal{N}(0,1)$. 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. 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 communications. *Random policy* considers the influence of wireless communication but without any optimization. Thus, its performance is the 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.

In Fig. 3, we show how wireless transmission affects the convergence behavior of FL in terms the value of the loss function and the global FL model remains unchanged which shows that FL 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 shows that the channel noise does not affect the convergence of the FL algorithm but it affects the value that the FL algorithm converges to.

B. Evaluation on the MNIST dataset

In order to evaluate the performance of our proposed INFLOTA in realistic application scenarios with real data, we

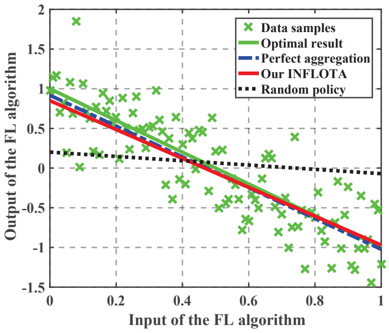

Fig. 2: An example of implementing FL for linear regression.

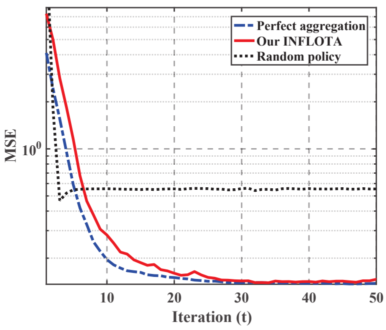

Fig. 3: MSE as the number of iteration varies.

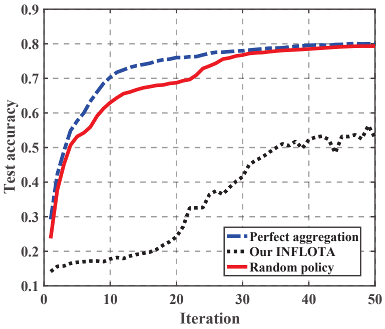

Fig. 4: The test accuracy as the iteration varies.

train a multilayer perceptron (MLP) on the MNIST dataset[3] with a 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 test accuracy versus the iteration index t in Fig. 4. As we can see, our proposed INFLOTA outperforms *Random policy*, and achieves comparable performance as *Perfect aggregation*.

VI. Conclusion

In this paper, we have studied the joint optimization of communications and FL over the air with analog aggregation. 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 rates, 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 FL. 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

This work was partly supported by the National Natural Science Foundation of China (Grants 61871023 and 61931001), Beijing Natural Science Foundation (Grant 4202054), the National Science Foundation of the US (Grants 1741338, 1939553, 2003211, 2128596 and 2136202), and VRIF-CCI (Grant 223996).

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

References

[1] 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.

[2] 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.

[3] 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.

[4] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3498–3516, 2007.

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

[6] 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.

[7] 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.

[8] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Bev-sgd: Best effort voting sgd for analog aggregation based federated learning against byzantine attackers," *arXiv preprint arXiv:2110.09660*, 2021.

[9] 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.

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

[11] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Communication-efficient federated learning through 1-bit compressive sensing and analog aggregation," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2021, pp. 1–6.

[12] ——, "1-bit compressive sensing for efficient federated learning over the air," *arXiv preprint arXiv:2103.16055*, 2021.

[13] 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.

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

[15] 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.

[16] 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.

[17] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Joint optimization of communications and federated learning over the air," *arXiv preprint arXiv:2104.03490*, 2021.

[18] 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.