Performance Optimization of Federated Learning over Wireless Networks

Mingzhe Chen[§,‖,*], Zhaohui Yang[†], Walid Saad[‡], Changchuan Yin[*], H. Vincent Poor[‖], and Shuguang Cui[§,¶]

[§]The Future Network of Intelligence Institute, The Chinese University of Hong Kong, Shenzhen, China,
Emails: {chenmingzhe,shuguangcui}@cuhk.edu.cn.
[‖]Department of Electrical Engineering, Princeton University, Princeton, NJ, USA, Email: poor@princeton.edu.
[*]Beijing Laboratory of Advanced Information Network, Beijing University of Posts and Telecommunications, Beijing, China 100876,
Email: ccyin@ieee.org.
[†]Centre for Telecommunications Research, Department of Informatics, Kings College London, WC2B 4BG, UK,
Email: yang.zhaohui@kcl.ac.uk.
[‡]Wireless@VT, Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA, Email: walids@vt.edu.
[¶]Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China.

***Abstract*—In this paper, the problem of training federated learning (FL) algorithms over a realistic wireless network is studied. In particular, in the considered model, wireless users perform an FL algorithm that trains their local FL models using their own data and send the trained local FL models to a base station (BS) that will generate a global FL model and send it back to the users. Since all training parameters are transmitted over wireless links, the quality of the training will be affected by wireless factors such as packet errors and availability of wireless resources. Meanwhile, due to the limited wireless bandwidth, the BS must select an appropriate subset of users to execute the FL learning algorithm so as to build a global FL model accurately. This joint learning, wireless resource allocation, and user selection problem is formulated as an optimization problem whose goal is to minimize an FL loss function that captures the performance of the FL algorithm. To address this problem, a closed-form expression for the expected convergence rate of the FL algorithm is first derived to quantify the impact of wireless factors on FL. Then, based on the expected convergence rate of the FL algorithm, the optimal transmit power for each user is derived, under a given user selection and uplink resource block (RB) allocation scheme. Finally, the user selection and uplink RB allocation is optimized so as to minimize the FL loss function. Simulation results show that the proposed joint federated learning and communication framework can reduce the FL loss function value by up to 10% and 16%, respectively, compared to 1) an optimal user selection algorithm with random resource allocation and 2) a random user selection and resource allocation algorithm.**

I. Introduction

Federated learning (FL) has been recently proposed by Google [1] as an effective approach to perform distributed machine learning tasks without relying on a centralized datacenter. FL is, in essence, a distributed machine learning algorithm that enables users to collaboratively learn a shared prediction model while keeping their collected data on their devices. However, to train an FL algorithm in a distributed manner, the users must transmit their training parameters over wireless links which can introduce training errors, due to the limited wireless resources and the inherent unreliability of wireless links.

This work was supported in part by CUHKSZ Presidential Postdoc Fellowship and grants NSFC-61629101, No. ZDSYS201707251409055, No. 2017ZT07X152, No. 2018B030338001, No. 2018YFB1800800, in part by the National Natural Science Foundation of China under Grants 61629101, 61671086, and 61871041, in part by Beijing Natural Science Foundation and Municipal Education Committee Joint Funding Project under Grant KZ201911232046, in part by the 111 Project under Grant B17007, and in part by the U.S. National Science Foundation under Grant CNS-1836802, CCF-0939370, and CCF-1513915.

Recently, a number of existing works such as in [1]–[5] have studied important problems related to the implementation of FL over wireless networks. In [1], the authors developed two update methods to reduce the uplink communication costs. The authors in [2] conducted an extensive empirical evaluation for five different FL models using four datasets. An echo state network-based FL algorithm is developed in [3] to optimize wireless virtual reality networks. In [4], the authors proposed a novel FL algorithm that can minimize the communication cost. The work in [5] exposed to potential of edge FL in wireless networks. While interesting, these prior works [1]–[5] assumed that wireless networks can readily integrate FL algorithms. However, in practice, due to the unreliability of the wireless channels and to the wireless resource limitations, FL algorithms will encounter training errors. For example, symbol errors introduced by the unreliable nature of the wireless channel and by resource limitations can impact the quality and correctness of the FL updates among users. Such errors will, in turn, affect the prediction performance of FL algorithms, as well as their convergence speed. Moreover, due to the wireless bandwidth limitations, the number of users can perform FL is limited; a design issue that is ignored in [1]–[4]. Hence, in practice, to effectively deploy FL over real-world wireless networks, it is necessary to investigate how the wireless factors affect the performance of FL algorithms.

The main contribution of this work is a novel framework for implementing FL algorithms over wireless networks by jointly considering FL and wireless metrics and factors. To our best knowledge, *this is the first work that provides a fundamental connection between the performance of FL algorithms and the underlying wireless network.* To this end, we propose a novel FL model in which cellular-connected wireless users transmit their locally trained FL models to a base station (BS) that generates the global FL model and sends it back to the users. For the considered FL model, the bandwidth of uplink is limited and, hence, the BS needs to select appropriate users to execute the FL algorithms so as to minimize the FL loss function. In addition, the impact of the packet errors on the parameter update process of the FL model is explicitly considered. To minimize training errors due to wireless links, we formulate a joint resource allocation and user selection problem for FL as an optimization problem whose goal is to minimize the value of the FL loss function while meeting the delay and energy consumption requirements of executing FL.

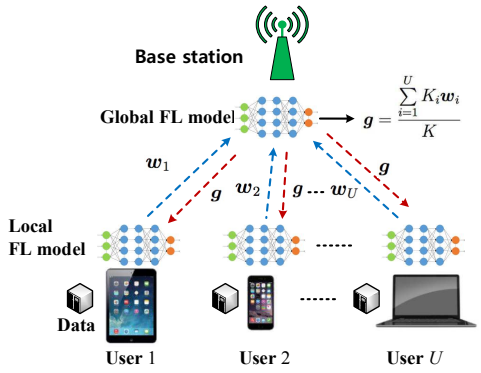

Fig. 1. The architecture of a wireless network that performs an FL algorithm.

Hence, our framework *jointly considers learning and wireless networking metrics*. To solve this problem, we first derive a closed-form expression for the expected convergence rate of the FL algorithm so as to build the relationship between the packet error rates and the performance of the FL algorithm. Based on this relationship, the optimization problem can be simplified as an mixed-integer nonlinear programming problem. To solve this simplified problem, we first find the optimal transmit power under given user selection and RB allocation. Then, we transform the original optimization problem into a bipartite matching problem. Finally, a Hungarian algorithm is used to find the optimal user selection and RB allocation.

The rest of this paper is organized as follows. The system model and problem formulation are described in Section II. The expected convergence rate of FL algorithms is studied in Section III. The optimal resource allocation and user selection are determined in Section IV. Simulation results are analyzed in Section V. Conclusions are drawn in Section VI.

II. System Model and Problem Formulation

Consider a cellular network having one base station (BS) and a set $\mathcal{U}$ of U users that cooperatively perform an FL algorithm for data analysis and inference (e.g., to generate radio maps using distributed datasets). FL enables the BS and the users to collaboratively learn a shared learning model while keeping all the training data at the device of each user. In an FL algorithm, each user will use its collected training data to train an FL model. Hereinafter, the FL model that is trained at the device of each user (using the data collected by the user itself) is called the *local FL model*. The BS is used to integrate the local FL models and generate a shared FL model. This shared FL model is used to improve the local FL model of each user so as to enable the users to collaboratively perform a learning task without training data transfer. Hereinafter, the FL model that is generated by the BS using the local FL models of the users is called the *global FL model*. As shown in Fig. 1, the uplink from the users to the BS is used to transmit the local FL model parameters while the downlink is used to transmit the global FL model.

A. Machine Learning Model

In our model, each user i collects a matrix $\boldsymbol{X}_i = [\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iK_i}]$ of input data where K_i is the number of the samples collected by each user i and each element $\boldsymbol{x}_{ik}$ is the FL algorithm's input vector. Let y_{ik} be the output of $\boldsymbol{x}_{ik}$. For simplicity, we consider an FL algorithm with a single output, which can be readily generalized to a case with multiple outputs [1]. The output data vector for training the FL algorithm of user i is $\boldsymbol{y}_i = [y_{i1}, \ldots, y_{iK_i}]$. We assume that the data collected by each user i is different from the other users, i.e., $(\boldsymbol{x}_i \neq \boldsymbol{x}_n, i \neq n, i, n \in \mathcal{U})$. We define a vector $\boldsymbol{w}_i$ to capture the parameters related to the local FL model that is trained by $\boldsymbol{X}_i$ and $\boldsymbol{y}_i$. $\boldsymbol{w}_i$ determines the local FL model of each user i. For example, in a linear regression learning algorithm, $\boldsymbol{x}_{ik}^{\mathrm{T}}\boldsymbol{w}_i$ represents the output and $\boldsymbol{w}_i$ determines the prediction accuracy. The training process of an FL algorithm is done in a way to solve:

$$\min_{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_U} \frac{1}{K}\sum_{i=1}^{U}\sum_{k=1}^{K_i} f\left(\boldsymbol{w}_i, \boldsymbol{x}_{ik}, y_{ik}\right), \tag{1}$$

$$\text{s. t. } \boldsymbol{w}_1 = \boldsymbol{w}_2 = \ldots = \boldsymbol{w}_U = \boldsymbol{g}, \quad \forall i \in \mathcal{U}, \tag{1a}$$

where $K = \sum_{i=1}^{U} K_i$ is total size of training data of all users. $\boldsymbol{g}$ is the global FL model generated by the BS and $f\left(\boldsymbol{w}_i, \boldsymbol{x}_{ik}, y_{ik}\right)$ is a loss function that captures the FL prediction accuracy. Different FL algorithms use different loss functions [6]. For example, for a linear regression FL, the loss function is $f\left(\boldsymbol{w}_i, \boldsymbol{x}_{ik}, y_{ik}\right) = \frac{1}{2}\left(\boldsymbol{x}_{ik}^{\mathrm{T}}\boldsymbol{w}_i - y_{ik}\right)^2$. (1a) is used to ensure that, once the FL algorithm converges, all of the users and the BS will share the same FL model. To solve (1), the BS will transmit the parameters $\boldsymbol{g}$ of the global FL model to its users so that they train their local FL models. Then, the users will transmit their local FL models to the BS to update the global FL model. The update of each user i's local FL model $\boldsymbol{w}_i$ depends on the global model $\boldsymbol{g}$ while the update of the global model $\boldsymbol{g}$ depends on all users' local FL models. The update of the local FL model $\boldsymbol{w}_i$ depends on the learning algorithm. For example, one can use gradient descent or randomized coordinate descent [1] to update the local FL model. The update of the global model $\boldsymbol{g}$ is given by [1]:

$$\boldsymbol{g} = \sum_{i=1}^{U} \frac{K_i \boldsymbol{w}_i}{K}. \tag{2}$$

Due to wireless transmissions, the local FL models received by the BS may contain erroneous symbols that affect the update of the global FL model and the FL learning performance. Hence, the wireless transmission delay will significantly affect the FL convergence. Hence, to deploy FL over wireless network, one must jointly consider the wireless and learning performance.

B. Transmission Model

For uplink, we assume that an orthogonal frequency-division multiple access (OFDMA) technique in which each user occupies one RB. The uplink rate of user i transmitting its local FL parameters to the BS is given by:

$$c_i^{\mathrm{U}}\left(\boldsymbol{r}_i, P_i\right) = \sum_{n=1}^{R} r_{i,n} B^{\mathrm{U}} \log_2\left(1 + \frac{P_i h_i}{\sum\limits_{i' \in \mathcal{U}_n'} P_{i'} h_{i'} + B^{\mathrm{U}} N_0}\right), \tag{3}$$

where $\boldsymbol{r}_i = [r_{i,1}, \ldots, r_{i,R}]$ is an RB allocation vector with R being the total number of RBs, $r_{i,n} \in \{0, 1\}$ and $\sum_{n=1}^{R} r_{i,n} = 1$; $r_{i,n} = 1$ indicates that RB n is allocated to user i, and $r_{i,n} = 0$,

otherwise; $\mathcal{U}_n'$ represents the set of users that are located at the other service areas and transmit data over RB n; B^{U} is the bandwidth of each RB and P_i is the transmit power of user i; h_i is the channel gain between user i and the BS; N_0 is the noise power spectral density; $\sum_{i' \in \mathcal{U}_n'} P_{i'} h_{i'}$ is the interference caused by the users that are located at the other service areas and use the same RB. Note that, although we ignore the optimization of resource allocation for the users located at the other service areas, we must consider the interference caused by the users using the same RBs, since this interference may significantly affect the packet error rates and the performance of FL.

Similarly, the downlink data rate of the BS transmitting the parameters of global FL model to each user i is given by:

$$c_i^{\mathrm{D}} = B^{\mathrm{D}} \log_2 \left(1 + \frac{P_B h_i}{\sum_{j \in \mathcal{B}'} P_B h_{ij} + B^{\mathrm{D}} N_0} \right), \tag{4}$$

where B^{D} is the bandwidth that the BS used to broadcast the global FL model to each user i; P_B is the transmit power of the BS; $\mathcal{B}'$ is the set of other BSs that cause interference to the BS that performs the FL algorithm; h_{ij} is the channel gain between user i and BS j.

The transmission delay between user i and the BS over both uplink and downlink will be:

$$l_i^{\mathrm{U}}(\boldsymbol{r}_i, P_i) = \frac{Z(\boldsymbol{w}_i)}{c_i^{\mathrm{U}}(\boldsymbol{r}_i, P_i)}, \; l_i^{\mathrm{D}}\left(B_i^{\mathrm{D}}\right) = \frac{Z(\boldsymbol{g})}{c_i^{\mathrm{D}}(B_i^{\mathrm{D}})}, \tag{5}$$

where function $Z(\boldsymbol{w}_i)$ is the data size of $\boldsymbol{w}_i$ and $Z(\boldsymbol{g})$ is the data size of the parameters related to the global FL model. Here, $Z(\boldsymbol{w}_i)$ and $Z(\boldsymbol{g})$ are determined by FL algorithm type.

C. Packet Error Rates

For simplicity, each local FL model $\boldsymbol{w}_i$ is considered as a single packet for uplink transmission. A cyclic redundancy check (CRC) mechanism [7] is used to check the data errors in the received local FL models at the BS. In particular, $C(\boldsymbol{w}_i) = 0$ indicates that the local FL model received by the BS contains data errors; otherwise, we have $C(\boldsymbol{w}_i) = 1$. The packet error rate experienced by the transmission of each local FL model $\boldsymbol{w}_i$ to the BS is [8]:

$$q_i(\boldsymbol{r}_i, P_i) = \sum_{n=1}^{R} r_{i,n} q_{i,n}, \tag{6}$$

where $q_{i,n} = \left(1 - \exp\left(-\frac{m\left(\sum_{i' \in \mathcal{U}_n'} P_{i'} h_{i'} + B^{\mathrm{U}} N_0\right)}{P_i h_i}\right)\right)$ is the packet error rate over RB n with m being a waterfall threshold. We assume that, when the received local FL model contains errors, the BS will not use it for the update of the global FL model. We also assume that the BS will not ask the corresponding users to resend the local FL models when the received local FL models contain data errors. As a result, the global FL model in (2) can be given by:

$$\boldsymbol{g}(\boldsymbol{a}, \boldsymbol{P}, \boldsymbol{R}) = \frac{\sum_{i=1}^{U} K_i a_i \boldsymbol{w}_i C(\boldsymbol{w}_i)}{\sum_{i=1}^{U} K_i a_i C(\boldsymbol{w}_i)}, \tag{7}$$

where $\boldsymbol{a} = [a_1, \ldots, a_U]$ is the vector of the user selection index with $a_i = 1$ indicating that user i performs the FL algorithm and $a_i = 0$, otherwise, $\boldsymbol{R} = [\boldsymbol{r}_1, \cdots, \boldsymbol{r}_U]$, $\boldsymbol{P} = [P_1, \cdots, P_U]$, $\sum_{i=1}^{U} K_i a_i C(\boldsymbol{w}_i)$ is the total number of training data samples, which depends on the user selection vector $\boldsymbol{a}$ and packet transmission $C(\boldsymbol{w}_i)$, $K_i \boldsymbol{w}_i C(\boldsymbol{w}_i) = 0$ indicates that the local FL model of user i contains data errors and, hence, the BS will not use it to generate the global FL model, $\boldsymbol{g}(\boldsymbol{a}, \boldsymbol{P}, \boldsymbol{R})$ is the global FL model that explicitly incorporates the effect of wireless transmission.

D. Energy Consumption Model

In our network, the energy consumption of each user consists of the energy needed for two purposes: a) Transmission of the local FL model and b) Training of the local FL model. The energy consumption of each user i is given by [9]:

$$e_i(\boldsymbol{r}_i, P_i) = \varsigma \omega_i \vartheta^2 Z(\boldsymbol{X}_i) + P_i l_i^{\mathrm{U}}(\boldsymbol{r}_i, P_i), \tag{8}$$

where ϑ is the frequency of the central processing unit (CPU) clock of each user i, ω_i is the number of CPU cycles required for computing per bit data of user i, and ς is the energy consumption coefficient depending on the chip of each user i's device [9]. In (8), $\varsigma \omega_i \vartheta^2 Z(\boldsymbol{X}_i)$ is the energy consumption of user i training the local FL model at its own device and $P_i l_i^{\mathrm{U}}(\boldsymbol{r}_i, P_i)$ represents the energy consumption of local FL model transmission from user i to the BS.

E. Problem Formulation

To jointly design the wireless network and the FL algorithm, we now formulate an optimization problem whose goal is to minimize the loss function of an FL algorithm by optimizing the various wireless parameters, as follows:

$$\min_{\boldsymbol{a}, \boldsymbol{P}, \boldsymbol{R}} \frac{1}{K} \sum_{i=1}^{U} \sum_{k=1}^{K_i} f\left(\boldsymbol{g}(\boldsymbol{a}, \boldsymbol{P}, \boldsymbol{R}), \boldsymbol{x}_{ik}, y_{ik}\right) \tag{9}$$

$$\text{s. t. } a_i, r_{i,n} \in \{0, 1\}, \quad \forall i \in \mathcal{U}, n = 1, \ldots, R, \tag{9a}$$

$$\sum_{n=1}^{R} r_{i,n} = a_i, \;\; \forall i \in \mathcal{U}, \tag{9b}$$

$$l_i^{\mathrm{U}}(\boldsymbol{r}_i, P_i) + l_i^{\mathrm{D}} \leq \gamma_{\mathrm{T}}, \;\; \forall i \in \mathcal{U}, \tag{9c}$$

$$e_i(\boldsymbol{r}_i, P_i) \leq \gamma_{\mathrm{E}}, \;\; \forall i \in \mathcal{U}, \tag{9d}$$

$$\sum_{i \in \mathcal{U}} r_{i,n} \leq 1, \;\; \forall n = 1, \ldots, R, \tag{9e}$$

$$0 \leq P_i \leq P_{\max}, \quad \forall i \in \mathcal{U}, \tag{9f}$$

where γ_{T} is the delay requirement for implementing the FL algorithm, γ_{E} is the energy consumption of the FL algorithm, and B is the total downlink bandwidth. (9a) and (9b) indicates that each user can occupy only one RB for uplink data transmission. (9c) is the delay needed to execute the FL algorithm. (9d) is the energy consumption requirement of performing an FL algorithm. (9e) indicates that each uplink RB can be allocated to at most one user. (9f) is a maximum transmit power constraint.

From (6) and (7), we see that the transmit power and resource allocation determine the packet error rate, thus affecting the update of the global FL models. In consequence, the loss

function of the FL algorithm in (9) depends on the resource allocation and transmit power. Moreover, (9c) shows that, in order to perform an FL algorithm, the users must satisfy a specific delay requirement. In particular, in an FL algorithm, the BS must wait to receive the local model of each user before updating its global FL model. Hence, transmission delay plays a key role in the FL performance. In a practical FL algorithm, it is desirable that all users transmit their local FL models to the BS simultaneously. From (9d), we see that to perform the FL algorithm, a given user must have enough energy to transmit and update the local FL model throughout the FL iterative process. If this given user does not have enough energy, the BS cannot choose this user to participate in the FL process. In consequence, to implement an FL algorithm, the wireless network must provide low energy consumption and latency, and high reliability data transmission.

III. Analysis of Global FL Model

To solve (9), we first need to analyze how the packet error rate affects the performance of the FL. To find the relationship between the packet error rates and the FL performance, we must first analyze the expected convergence rate of the FL.

In the studied network, the users adopt a standard gradient descent method to update their local FL models as done in [1]. Therefore, during the training process, the update of user i's local FL model $\boldsymbol{w}_i$ at time t is given by:

$$\boldsymbol{w}_{i,t+1} = \boldsymbol{g}_t\left(\boldsymbol{a}, \boldsymbol{P}, \boldsymbol{R}\right) - \frac{\lambda}{K_i}\sum_{k=1}^{K_i}\nabla f\left(\boldsymbol{g}_t\left(\boldsymbol{a}, \boldsymbol{P}, \boldsymbol{R}\right), \boldsymbol{x}_{ik}, y_{ik}\right), \tag{10}$$

where λ is the learning rate and $\nabla f\left(\boldsymbol{g}_t\left(\boldsymbol{a}, \boldsymbol{P}, \boldsymbol{R}\right), \boldsymbol{x}_{ik}, y_{ik}\right)$ is the gradient of $f\left(\boldsymbol{g}_t\left(\boldsymbol{a}, \boldsymbol{P}, \boldsymbol{R}\right), \boldsymbol{x}_{ik}, y_{ik}\right)$ with respect to $\boldsymbol{g}_t\left(\boldsymbol{a}, \boldsymbol{P}, \boldsymbol{R}\right)$.

We assume that $F\left(\boldsymbol{g}\right) = \frac{1}{K}\sum_{i=1}^{U}\sum_{k=1}^{K_i} f\left(\boldsymbol{g}, \boldsymbol{x}_{ik}, y_{ik}\right)$ where $\boldsymbol{g}$ is short for $\boldsymbol{g}\left(\boldsymbol{a}, \boldsymbol{P}, \boldsymbol{R}\right)$. Based on (10), the update of global FL model $\boldsymbol{g}$ at time t can be given by:

$$\boldsymbol{g}_{t+1} = \boldsymbol{g}_t - \lambda\left(\nabla F(\boldsymbol{g}_t) - \boldsymbol{o}\right), \tag{11}$$

where $\boldsymbol{o} = \nabla F(\boldsymbol{g}_t) - \frac{\sum_{i=1}^{U} K_i a_i \boldsymbol{w}_i C(\boldsymbol{w}_i)}{\sum_{i=1}^{U} K_i a_i C(\boldsymbol{w}_i)}$ with

$$C\left(\boldsymbol{w}_i\right) = \begin{cases} 1, & \text{with probability } 1 - q_i\left(\boldsymbol{r}_i, P_i\right), \\ 0, & \text{with probability } q_i\left(\boldsymbol{r}_i, P_i\right). \end{cases} \tag{12}$$

We also assume that the FL algorithm converges to an optimal global FL model $\boldsymbol{g}^*$ after learning steps. To derive the expected convergence rate of the FL algorithms, we also make the following assumptions: a) $\|\nabla F(\boldsymbol{g}_{t+1}) - \nabla F(\boldsymbol{g}_t)\| \leq L\|\boldsymbol{g}_{t+1} - \boldsymbol{g}_t\|$ where L is a positive constant and $\|\boldsymbol{g}_{t+1} - \boldsymbol{g}_t\|$ is the norm of $\boldsymbol{g}_{t+1} - \boldsymbol{g}_t$, b) $F(\boldsymbol{g}_{t+1}) \geq F(\boldsymbol{g}_t) + (\boldsymbol{g}_{t+1} - \boldsymbol{g}_t)^T\nabla F(\boldsymbol{g}_t) + \frac{\mu}{2}\|\boldsymbol{g}_{t+1} - \boldsymbol{g}_t\|^2$, c) $\mu\boldsymbol{I} \preceq \nabla^2 F(\boldsymbol{g}) \preceq L\boldsymbol{I}$, and d) $\|\nabla f\left(\boldsymbol{g}_t, \boldsymbol{x}_{ik}, y_{ik}\right)\|^2 \leq \zeta_1 + \zeta_2\nabla\|F\left(\boldsymbol{g}_t\right)\|^2$ with $\zeta_1 \geq 0$ and $\zeta_2 \geq 1$ These assumptions can be easily satisfied by the general FL loss functions such as linear or logistic loss functions. The expected convergence rate of the FL algorithms can be obtained by the following theorem.

Theorem 1. Given the transmit power vector $\boldsymbol{P}$, RB allocation matrix $\boldsymbol{R}$, user selection vector $\boldsymbol{a}$, optimal global FL model $\boldsymbol{g}^*$, and the learning rate $\lambda = \frac{1}{L}$, the upper bound of $\mathbb{E}[F(\boldsymbol{g}_{t+1}) - F(\boldsymbol{g}^*)]$ can be given by:

$$\mathbb{E}[F(\boldsymbol{g}_{t+1}) - F(\boldsymbol{g}^*)] \leq \frac{\zeta_1}{2LK}\sum_{i=1}^{U} K_i q_i\left(\boldsymbol{r}_i, P_i\right)\frac{1 - A^t}{1 - A} + A^t\mathbb{E}(F(\boldsymbol{g}_0) - F(\boldsymbol{g}^*)), \tag{13}$$

where $A = 1 - \frac{2\mu}{L} + \frac{\mu\zeta_2}{LK}\sum_{i=1}^{U} K_i q_i\left(\boldsymbol{r}_i, P_i\right)$.

Proof. Due to space limitations, the proof is omitted. □

From Theorem 1, we see that, when the learning rate λ is a constant ($\lambda = \frac{1}{L}$), the FL algorithm that considers the effect of the packet error rates will finally converge as t increases. However, a gap, $\frac{\zeta_1}{2LK}\sum_{i=1}^{U} K_i\left(1 - a_i + a_i q_i\left(\boldsymbol{r}_i, P_i\right)\right)$, exists between $\mathbb{E}[F(\boldsymbol{g}_t)]$ and $\mathbb{E}[F(\boldsymbol{g}^*)]$. This gap is caused by the packet errors and user selection. As the packet error rate decreases, the gap between $\mathbb{E}[F(\boldsymbol{g}_t)]$ and $\mathbb{E}[F(\boldsymbol{g}^*)]$ decreases. Meanwhile, as the number of users that implement the FL algorithm increases, the gap also decreases. Moreover, as the packet error rate decreases, the value of A also decreases, which indicates that the convergence speed of the FL algorithm improves. Hence, it is necessary to optimize resource allocation, user selection, and transmit power for the implementation of FL algorithms.

IV. Optimization of Prediction Errors for Federated Learning Algorithm

In this section, we minimize the loss function of the FL algorithms. From Theorem 1, we see that, to minimize the loss function in (9), we need to only minimize the gap, $\frac{\zeta_1}{2LK}\sum_{i=1}^{U} K_i\left(1 - a_i + a_i q_i\left(\boldsymbol{r}_i, P_i\right)\right)\frac{1-A^t}{1-A}$. When $A \geq 1$, the FL algorithms will not converge. In consequence, here, we only consider the minimization of the FL loss function when $A < 1$. Hence, as t is large enough, which indicates that the FL algorithm converges, we have $A^t = 0$. The gap can be rewritten as follows:

$$\frac{\frac{\zeta_1}{2LK}\sum_{i=1}^{U} K_i\left(1 - a_i + a_i q_i\left(\boldsymbol{r}_i, P_i\right)\right)}{\frac{2\mu}{L} - \frac{\mu\zeta_2}{LK}\sum_{i=1}^{U} K_i\left(1 - a_i + a_i q_i\left(\boldsymbol{r}_i, P_i\right)\right)}. \tag{14}$$

From (14), we see that minimizing (14) only needs to minimize $\sum_{i=1}^{U} K_i\left(1 - a_i + a_i q_i\left(\boldsymbol{r}_i, P_i\right)\right)$. Meanwhile, since $a_i = \sum_{n=1}^{R} r_{i,n}$ and $q_i\left(\boldsymbol{r}_i, P_i\right) = \sum_{n=1}^{R} r_{i,n} q_{i,n}$, we have, when $a_i = 1$, $q_i\left(\boldsymbol{r}_i, P_i\right) \leq 0$ and when $a_i = 0$, $q_i\left(\boldsymbol{r}_i, P_i\right) = 0$. In consequence, $a_i q_i\left(\boldsymbol{r}_i, P_i\right) = q_i\left(\boldsymbol{r}_i, P_i\right)$. The problem in (9) can be simplified as follows:

$$\min_{\boldsymbol{P}, \boldsymbol{R}} \sum_{i=1}^{U} K_i\left(1 - \sum_{n=1}^{R} r_{i,n} + q_i\left(\boldsymbol{r}_i, P_i\right)\right) \tag{15}$$

s. t. (9c)-(9f).

$$r_{i,n} \in \{0,1\}, \quad \forall i \in \mathcal{U}, n = 1, \ldots, R, \tag{15a}$$

$$\sum_{n=1}^{R} r_{i,n} \leq 1, \ \ \forall i \in \mathcal{U}, \tag{15b}$$

In (15), the BS allocating a RB to each user i indicates that each user i is associated with the BS. In consequence, the user selection index a_i can be represented by the RB allocation index $\sum_{n=1}^{R} r_{i,n}$. Next, we first find the optimal transmit power for each user given the uplink RB allocation matrix $\boldsymbol{R}$. Then, we find the uplink RB allocation to minimize the FL loss function.

A. Optimal Transmit Power

The optimal transmit power of each user i can be determined by the following lemma.

Lemma 1. Given the uplink RB allocation vector $\boldsymbol{r}_i$ of each user i, the optimal transmit power of each user i, P_i^* is:

$$P_i^* = \min\left\{P_{\max}, P_{i,\gamma_{\mathrm{E}}}\right\}, \tag{16}$$

where $P_{i,\gamma_{\mathrm{E}}}$ satisfies the equality $\varsigma\omega_i\vartheta^2 Z\left(\boldsymbol{X}_i\right) + \frac{P_{i,\gamma_{\mathrm{E}}} Z(\boldsymbol{w}_i)}{c_i^{\mathrm{U}}\left(\boldsymbol{r}_i, P_{i,\gamma_{\mathrm{E}}}\right)} = \gamma_{\mathrm{E}}$.

Proof. To prove Lemma 1, we first prove that $e_i\left(\boldsymbol{r}_i, P_i\right)$ is an increasing function of P_i. Based on (3) and (8), we have:

$$e_i\left(\boldsymbol{r}_i, P_i\right) = \varsigma\omega_i\vartheta^2 Z\left(\boldsymbol{X}_i\right) + \frac{P_i}{\sum\limits_{n=1}^{R} r_{i,n} B^{\mathrm{U}} \log_2\left(1 + \kappa_{i,n} P_i\right)}, \tag{17}$$

where $\kappa_{i,n} = \frac{h_i}{\sum\limits_{i' \in \mathcal{U}'_n} P_{i'} h_{i'} + B^{\mathrm{U}} N_0}$. The first-derivative of $e_i\left(\boldsymbol{r}_i, P_i\right)$ with respect to P_i is given by:

$$\frac{\partial e_i\left(\boldsymbol{r}_i, P_i\right)}{\partial P_i} = \frac{\left(\ln 2\right) \sum\limits_{n=1}^{R} \frac{r_{i,n}}{1+\kappa_{i,n}P_i}\left(\left(1+\kappa_{i,n}P_i\right)\ln\left(1+\kappa_{i,n}P_i\right) - \kappa_{i,n}P_i\right)}{\left(\sum\limits_{n=1}^{R} r_{i,n} B^{\mathrm{U}} \ln\left(1+\kappa_{i,n}P_i\right)\right)^2}. \tag{18}$$

Since $\frac{\partial e_i(\boldsymbol{r}_i, P_i)}{\partial P_i}$ is always positive, $e_i\left(\boldsymbol{r}_i, P_i\right)$ is a monotonically increasing function. Contradiction is used to prove Lemma 1. We assume that P_i' ($P_i' \neq P_i^*$) is the optimal transmit power of user i. In (9d), $e_i\left(\boldsymbol{r}_i^*, P_{i,\gamma_{\mathrm{E}}}\right)$ is a monotonically increasing function of P_i. Hence, as $P_i' > P_i^*$, $e_i\left(\boldsymbol{r}_i^*, P_i'\right) > \gamma_{\mathrm{E}}$, which does not meet the constraint (9f). From (6), we see that, the packer error rates decrease as the transmit power increases. Thus, as $P_i' < P_i^*$, we have $q_i\left(\boldsymbol{r}_i, P_i^*\right) \leqslant q_i\left(\boldsymbol{r}_i, P_i'\right)$. In consequence, as $P_i' < P_i^*$, P_i' cannot minimize the function in (15). Hence, we have $P_i' = P_i^*$. This completes the proof. □

From Lemma 1, we see that the optimal transmit power depends on the size of the collected data $Z\left(\boldsymbol{X}_i\right)$, the size of the local FL model $Z\left(\boldsymbol{w}_i\right)$, and the interference in each RB. In particular, as the size of the collected data and local FL model increases, each user must spend more energy for training FL model and, hence, the energy that can be used for data transmission decreases. In consequence, the value of the FL loss function increases.

B. Optimal Uplink Resource Block Allocation

Based on Lemma 1 and (6), the optimization problem in (15) can be simplified as follows:

$$\min_{\boldsymbol{R}} \sum_{i=1}^{U} K_i \left(1 - \sum_{n=1}^{R} r_{i,n} + \sum_{n=1}^{R} r_{i,n} q_{i,n}\right) \tag{19}$$

s. t. (9a), (9b), and (9e),

$$l_i^{\mathrm{U}}\left(\boldsymbol{r}_i, P_i^*\right) + l_i^{\mathrm{D}} \leq \gamma_{\mathrm{T}}, \ \ \forall i \in \mathcal{U}, \tag{19a}$$

$$e_i\left(\boldsymbol{r}_i, P_i^*\right) \leq \gamma_{\mathrm{E}}, \ \ \forall i \in \mathcal{U}. \tag{19b}$$

Obviously, the objective function (19) is an integer linear programming problem, which can be solved by a bipartite matching algorithm. Compared to traditional convex optimization algorithms, using bipartite matching to solve problem (19) does not require calculating the gradients of each variable nor dynamically adjusting the step size for convergence.

To use the bipartite matching algorithm for solving (19), we first transform the optimization problem into a bipartite matching problem. We construct a bipartite graph $\mathcal{A} = (\mathcal{U} \times \mathcal{R}, \mathcal{E})$ in which, $\mathcal{R}$ is the set of RBs that can be allocated to each user, each vertex in $\mathcal{U}$ represents a user and each vertex in $\mathcal{R}$ represents an RB, $\mathcal{E}$ is the set of edges that connect to the vertices from each set $\mathcal{U}$ and $\mathcal{R}$. Let $\vartheta_{in} \in \mathcal{E}$ be the edge connecting vertex i in $\mathcal{U}$ and vertex n in $\mathcal{R}$ with $\vartheta_{in} \in \{0,1\}$, where $\vartheta_{in} = 1$ indicates that RB n is allocated to user i, otherwise, we have $\vartheta_{in} = 0$. Let matching $\mathcal{T}$ be a subset of edges in $\mathcal{E}$, in which no two edges share a common vertex in $\mathcal{R}$, such that each RB n can only be allocated to one user (constraint (9e) is satisfied). Nevertheless, in $\mathcal{T}$, all of the edges associated with a vertex $i \in \mathcal{U}$ will not share a common vertex $n \in \mathcal{R}$, such that each user i can occupy only one RB (constraint (9b) is satisfied). The weight of edge ϑ_{in} is given by:

$$\psi_{in} = \begin{cases} K_i\left(q_{i,n} - 1\right), l_i^{\mathrm{U}} + l_i^{\mathrm{D}} \leq \gamma_{\mathrm{T}} \text{ and } e_i \leq \gamma_{\mathrm{E}}, \\ +\infty, \qquad\qquad \text{otherwise}, \end{cases} \tag{20}$$

where l_i^{U} and e_i is short for $l_i^{\mathrm{U}}\left(r_{i,n}, P_i^*\right)$ and $e_i\left(r_{i,n}, P_i^*\right)$. From (20), we see that when RB n is allocated to user i, if the delay and energy requirements cannot be meet, $\psi_{in} = +\infty$, which indicates that RB n will not be allocated to user i. The purpose of this bipartite matching problem is to find an optimal matching set $\mathcal{T}^*$ that can minimize the weights of the edges in $\mathcal{T}^*$. A standard Hungarian algorithm [10] can be adopted to find the optimal matching set $\mathcal{T}^*$. When the optimal matching set is found, the optimal RB allocation is determined.

V. Simulation Results and Analysis

For our simulations, we consider a circular network area having a radius $r = 500$ m with one BS at its center servicing $U = 20$ uniformly distributed users. The other parameters used in simulations are listed in Table I. The 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

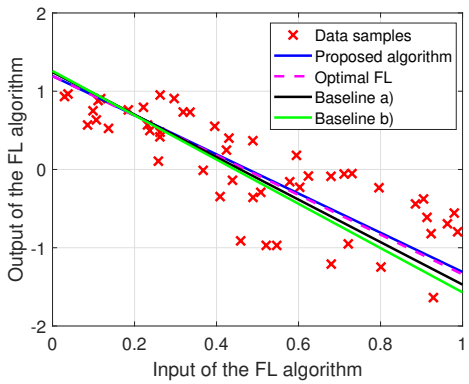

Fig. 2. An example of implementing FL algorithms for function approximation.

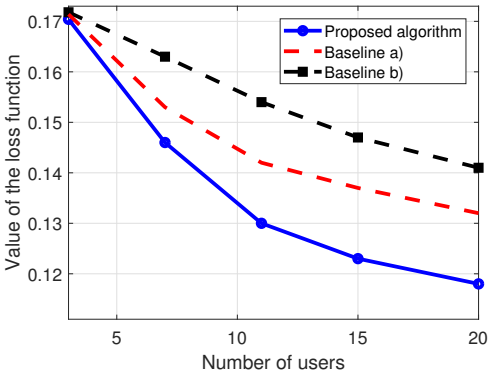

Fig. 3. Value of the loss function as the number of iteration varies.

algorithm is used as the linear regression approach to model the relationship between x and y. For comparison purposes, we use two baselines: a) an FL algorithm that optimizes user selection with random resource allocation and b) an FL algorithm that randomly determines user selection and resource allocation. *Baseline a)* is actually an FL algorithm without consideration of wireless factors (which can be seen as a version of the original FL algorithm in [1]). *Baseline b)* is an FL algorithm without consideration of wireless factors and optimizing FL performance.

Fig. 2 shows an example of using FL algorithms for linear regression. In this figure, optimal FL indicates the packet error rates of all users are zero. From Fig. 2, we can see that the proposed FL algorithm can fit the data samples more accurately than baselines a) and b). This is due to the fact that the proposed FL algorithm jointly considers the learning and wireless factors and, hence, it can optimize user selection and resource allocation to reduce the effect of wireless transmission errors on training FL algorithm and improve the performance of the FL algorithm.

Fig. 3 shows how the value of the FL loss function changes as the total number of users varies. In this figure, an appropriate subset of users is selected to perform the FL algorithm. From Fig. 3, we see that, as the number of users increases, the value of the loss function decreases. Moreover, as the number of users increases, the effect of packet errors on the global FL model decreases. This is due to the fact that an increase in the number of users leads to more data available for the FL algorithm training and, hence, improving the accuracy of approximation of the gradient of the loss function. Fig. 3 also

TABLE I
SYSTEM PARAMETERS

Parameter	Value	Parameter	Value
α	2	N_0	-174 dBm
P_B	1 W	B	20 MHz
M	64	B^{U}	150 kHz
σ_i	1	$P_{\max}$	0.01 W
f	10^9	K_i	[12,10,8,4,2]
ς	10^{-27}	γ_{T}	100 ms
ω_i	40	γ_{E}	0.02 J

shows that the proposed algorithm reduces the loss function by, respectively, up to 10% and 16% compared to baselines a) and b). The 10% reduction of the loss function stems from the fact that the proposed algorithm optimizes the resource allocation. The 16% reduction stems from the fact that the proposed algorithm joint considers learning and wireless effects and, hence, it can optimize the user selection and resource allocation to reduce the FL loss function.

VI. Conclusion

In this paper, we have developed a novel framework that enables the implementation of FL algorithms over wireless networks. We have formulated an optimization problem that jointly considers user selection and resource allocation for the minimization of the value of FL loss function. To solve this problem, we have derived a closed-form expression of the expected convergence rate of the FL algorithm that considers the wireless factors. Based on the derived expected convergence rate, the optimal transmit power is determined given the user selection and uplink RB allocation. Then, the Hungarian algorithm is used to find the optimal user selection and RB allocation so as to minimize the FL loss function. Simulation results have shown that the FL algorithm that considers the wireless factors yields significant improvements in performance compared to the existing implementation of the FL algorithm that does not account for wireless factors.

References

[1] 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.
[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.
[3] M. Chen, O. Semiari, W. Saad, X. Liu, and C. Yin, "Federated echo state learning for minimizing breaks in presence in wireless virtual reality networks," *arXiv preprint arXiv:1812.01202*, 2018.
[4] J. Konečnỳ, B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," *arXiv preprint arXiv:1511.03575*, 2015.
[5] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, to appear, 2019.
[6] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Communications Surveys Tutorials*, to appear, 2019.
[7] B. Li, H. Shen, and D. Tse, "An adaptive successive cancellation list decoder for polar codes with cyclic redundancy check," *IEEE Communications Letters*, vol. 16, no. 12, pp. 2044–2047, December 2012.
[8] Y. Xi, A. Burr, J. Wei, and D. Grace, "A general upper bound to evaluate packet error rate over quasi-static fading channels," *IEEE Transactions on Wireless Communications*, vol. 10, no. 5, pp. 1373–1377, May 2011.
[9] Z. Yang, C. Pan, K. Wang, and M. Shikh-Bahaei, "Energy efficient resource allocation in UAV-enabled mobile edge computing networks," *IEEE Trans. Wireless Commun.*, to appear, 2019.
[10] R. Jonker and T. Volgenant, "Improving the hungarian assignment algorithm," *Operations Research Letters*, vol. 5, no. 4, pp. 171–175, 1986.