Device Scheduling and Resource Allocation for Federated Learning under Delay and Energy Constraints

Wenqi Shi, Yuxuan Sun, *Member, IEEE*, Sheng Zhou, *Member, IEEE*, Zhisheng Niu, *Fellow, IEEE*

***Abstract*—Federated Learning (FL) is an emerging technique to enhance edge intelligence, where mobile devices train machine learning models collaboratively with their local data. Limited energy on devices and scarce wireless bandwidth can notably impact the convergence of FL over wireless networks, and thus device scheduling and resource allocation are critical. In this paper, we propose a joint device scheduling and resource allocation scheme to maximize the model accuracy under total training delay and device energy budgets. Since FL consists of multiple training rounds, there is an inherent trade-off between per-round delay, per-round energy consumption, and the total number of rounds. To find solution, we decouple the accuracy maximization problem into two sub-problems. First, given a scheduling policy, the bandwidth allocation and local computing frequency are jointly optimized to maximize the number of rounds that can be conducted. Then, a device scheduling policy is proposed to balance the trade-off between the per-round energy and delay cost and the number of rounds, with the ultimate goal of accuracy optimization. Experiments on various learning tasks and datasets show that the proposed scheme can greatly improve the convergence rate of resource-constrained FL.**

I. Introduction

Federated Learning (FL) is a promising technique to enable data-driven applications at the wireless edge while preserving data privacy, where mobile devices collaboratively train a machine learning (ML) model with their local data, and a central node such as a base station (BS) aggregates and broadcasts global model periodically [1]. To ensure the effectiveness and freshness of ML models in the time-varying environment, FL-based model training needs to be done in a *timely* manner using real-time data, which motivates us to *go beyond model accuracy* and consider *training delay* as an important metric [2]. Typical scenarios include content popularity prediction for edge caching, virtual and augmented reality, channel inference and beam selection, and spectrum management [3], [4].

However, it is quite challenging to meet the timeliness requirements of FL over wireless networks due to various kinds of resource constraints, including limited computing capability and battery capacity of mobile devices, and scarce wireless bandwidth. The key methodology is to jointly optimize communication and learning [5] via device scheduling and resource management [6]–[11], or revolutionizing communication protocols to enable over-the-air computation [12]–[14]. Among them, the training delay for communication is minimized by device selection and radio resource allocation [8], while computing and communication resource are jointly optimized in [10], [11] to improve energy efficiency.

Most existing papers set the total number of training rounds to a fixed number. However, when considering the total resource constraints for training, there is an inherent trade-off between per-round training cost and the total number of rounds that can be conducted before resource depletion. Scheduling more devices and allocating more resources in each round can bring greater improvement to model accuracy, while iterating more rounds is also beneficial. Therefore, a key problem is to balance these two terms to optimize the training performance. Our previous work [7] takes total training delay as a constraint, and proposes a joint device scheduling and bandwidth allocation scheme to maximize convergence rate.

In this work, the energy budgets of mobile devices are considered in additional to the total training delay, and a two-step device scheduling and resource allocation scheme is proposed to maximize the model accuracy. First, under a given scheduling policy, the bandwidth allocation and computing frequency are jointly optimized through convex optimization to maximize the total training rounds. Then, a device scheduling policy is proposed based on the convergence analysis, which balances the trade-off between per-round cost and total number of rounds. Experiments on MNIST and CIFAR-10 datasets show that the proposed scheme can greatly improve the convergence rate of resource-constrained FL.

II. System Model

Consider an FL system with one BS and M devices $\mathcal{M} = \{1, \ldots, M\}$. Each device has a local dataset $\mathcal{D}_i = \{\boldsymbol{x}_{i,d} \in \mathbb{R}^s, y_{i,d} \in \mathbb{R}\}_{d=1}^{D_i}$, where D_i is used to denote the size of $\mathcal{D}_i$. We use $\mathcal{D} = \bigcup_{i \in \mathcal{M}} \mathcal{D}_i$ to denote the whole dataset, where the total number of data samples is denoted by $D = \sum_{i \in \mathcal{M}} D_i$. The goal of FL is to find a model parameter $\boldsymbol{w}$, which can minimize a particular loss function on the whole dataset, i.e.

$$\min_{\boldsymbol{w}} \left\{ F(\boldsymbol{w}) \triangleq \frac{1}{D} \sum_{i \in \mathcal{M}} D_i F_i(\boldsymbol{w}) \right\}. \tag{1}$$

This work is sponsored in part by the National Key R&D Program of China 2018YFB1800804, the Nature Science Foundation of China (No. 62022049, No. 61871254, No. 61861136003), the China Postdoctoral Science Foundation No. 2020M680558, and Hitachi Ltd.

W. Shi, Y. Sun, S. Zhou, and Z. Niu are with the Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: swq17@mails.tsinghua.edu.cn; {sunyuxuan, sheng.zhou, niuzhs}@tsinghua.edu.cn.

In this paper, we use empirical risk as the loss function, which can be defined as $F_i(\boldsymbol{w}) \triangleq \frac{1}{D_i} \sum_{\{\boldsymbol{x},y\}\in\mathcal{D}_i} f(\boldsymbol{w},\boldsymbol{x},y)$, and $f(\boldsymbol{w},\boldsymbol{x},y)$ captures the error of the model parameter $\boldsymbol{w}$ on the input-output data pair $\{\boldsymbol{x},y\}$.

FL uses an iterative procedure to solve (1). In particular, the k-th round contains the following 3 steps.

1) Global Model Broadcast: In the beginning of round k, a set of devices $\Pi_k \in \mathcal{M}$ are scheduled, and a Boolean variable $\pi_{i,k}$ is used to denote whether device i is in Π_k. Then the BS broadcasts global model $\boldsymbol{w}_{k-1}$ to all scheduled devices.

2) Local Model Update: Each scheduled device i initializes its local model of the current round by the received global model (i.e., $\boldsymbol{w}_{i,k}(0) \leftarrow \boldsymbol{w}_{k-1}$), and then uses stochastic gradient decent (SGD) to update the local model based on its local dataset as follows,

$$\boldsymbol{w}_{i,k}(j+1) = \boldsymbol{w}_{i,k}(j) - \eta\nabla F_i(\boldsymbol{w}_{i,k}(j)),\ j = 0,\ldots,\tau-1.$$

Here η is the learning rate, and the gradient $\nabla F_i(\boldsymbol{w}_{i,k}(j))$ is computed on $\mathcal{D}_{b,i}$, a randomly sampled mini-batch from $\mathcal{D}_i$ with batch size $d_i = |\mathcal{D}_{b,i}|$. The SGD is repeated for τ times and τ is considered as a fixed system parameter. After finishing all SGD steps, $\boldsymbol{w}_{i,k}(\tau)$ is uploaded to the BS. In the following part of the paper, we use $\boldsymbol{w}_{i,k}$ to denote $\boldsymbol{w}_{i,k}(\tau)$ unless otherwise specified.

3) Global Model Aggregation: After receiving $\boldsymbol{w}_{i,k}$ from all scheduled devices, the BS aggregates them as follows [15]:

$$\boldsymbol{w}_k = \frac{\sum_{i\in\Pi_k} D_i \boldsymbol{w}_{i,k}}{\sum_{i\in\Pi_k} D_i}. \tag{2}$$

A. Local Computation Model

Assume that computing gradient for each data sample requires c CPU cycles, and thus the total CPU cycles required for local model update is $c\tau d_i$. The CPU frequency of device i in the k-th round is denoted by $f_{i,k}$, which can be adjusted by the widely used Dynamic Voltage and Frequency Scaling (DVFS) technique [16]. Therefore, the delay and computation energy consumption of local model update is given by

$$t^{\text{cp}}_{i,k} = \frac{c\tau d_i}{f_{i,k}},\ \ 0 \le f_{i,k} \le f_{\max}, \tag{3}$$

$$e^{\text{cp}}_{i,k} = \kappa_i c\tau d_i f^2_{i,k}, \tag{4}$$

where κ_i is the effective switched capacitance that depends on the chip architecture of device i [10].

B. Wireless Transmission Model

We consider an Frequency Division Multiple Access (FDMA) scheme with total bandwidth B for local model uploading. $\gamma_{i,k}$ is used to denote the bandwidth allocation ratio to the i-th device in the k-th round, with $\sum_{i=1}^{M}\gamma_{i,k} = 1$ and $0 \le \gamma_{i,k} \le 1$. As a result, the achievable transmission rate (in bits/s) can be written as $r_{i,k} = \gamma_{i,k}B\log_2\left(1+\frac{p_i h^2_{i,k}}{\gamma_{i,k}BN_0}\right)$, where p_i denotes the transmit power of device i, $h_{i,k}$ denotes the channel gain, and N_0 is the noise power density. Let S be the size of $\boldsymbol{w}_{i,k}$ (in bits), we have the uploading delay

$$t^{\text{cm}}_{i,k} = \frac{S}{r_{i,k}}, \tag{5}$$

and the corresponding uploading energy consumption

$$e^{\text{cm}}_{i,k} = P_i t^{\text{cm}}_{i,k}, \tag{6}$$

where $P_i = \xi p_i + p_c$ is the power consumption of transmitter, ξ is the power amplifier coefficient, and p_c is the circuit power.

Regarding to the global model broadcasting, the delay can be ignored since the transmit power of BS is much higher than that of devices, and the whole downlink bandwidth is assumed solely used by the BS. The energy consumption can also be ignored since the power consumption of the receiver is relatively lower.

C. Problem Formulation

Since FL are usually resources constrained, we use T to denote the total training time budget, and use E_i to denote the total energy of device i. Suppose K rounds can be conducted, to satisfy the device energy constraint, we have

$$\sum_{k=1}^{K}\pi_{i,k}(e^{\text{cp}}_{i,k} + e^{\text{cm}}_{i,k}) \le E_i,\ \ \forall i \in \{1,\ldots,M\}. \tag{7}$$

And the total training delay of all K rounds can not exceed the total training delay budget T, i.e.

$$\sum_{k=1}^{K} t^{\text{round}}_k \le T, \tag{8}$$

where t^{round}_k is used to denoted the training delay of the k-th round. Because of the synchronous model aggregation of FL, t^{round}_k is determined by the slowest device among all the scheduled devices, i.e.,

$$\pi_{i,k}(t^{\text{cp}}_{i,k} + t^{\text{cm}}_{i,k}) \le t^{\text{round}}_k,\ \ \forall i \in \{1,\ldots,M\}. \tag{9}$$

Our goal is to maximize the model accuracy, which is usually achieved by minimizing the loss function in the machine learning community. Therefore, the overall optimization problem can be written as

$$\min_{K,\boldsymbol{\Pi}_{[K]},\boldsymbol{\gamma}_{[K]},\boldsymbol{f}_{[K]},\boldsymbol{t}^{\text{round}}_{[K]}} \quad F(\tilde{\boldsymbol{w}}) \tag{P1}$$

$$\text{s.t.} \quad (3)-(9). \tag{10}$$

Here, $\tilde{\boldsymbol{w}}$ is the best model parameter among all K rounds (i.e., $\tilde{\boldsymbol{w}} \triangleq \underset{\boldsymbol{w}\in\{\boldsymbol{w}_k|k=1,2,\ldots,K\}}{\arg\min} F(\boldsymbol{w})$), and the notation $\boldsymbol{\Pi}_{[K]} \triangleq [\Pi_1,\Pi_2,\ldots,\Pi_K]$ is used to denote the scheduling decisions in all K rounds, similar as $\boldsymbol{\gamma}_{[K]}$, $\boldsymbol{f}_{[K]}$, and $\boldsymbol{t}^{\text{round}}_{[K]}$.

III. Device Scheduling and Resource Allocation

Since problem (P1) is a complicated mix-integer non-linear programming problem, we decouple it into two sub-problems: a convex resource allocation sub-problem given the scheduling decision, and a device scheduling sub-problem.

A. Resource Allocation Given Scheduling Decisions

Consider the k-th round, given scheduling decision Π_k, remaining training time budget T_k, and remaining device energy $E_{i,k}$, the resource allocation sub-problem is as follows:

$$\min_{\boldsymbol{\gamma}_k, \boldsymbol{f}_k, t_k^{\text{round}}} \quad t_k^{\text{round}} \tag{P2}$$

$$\text{s.t.} \quad t_{i,k}^{\text{cp}} + t_{i,k}^{\text{cm}} \leq t_k^{\text{round}}, \ \forall i \in \Pi_k, \tag{11}$$

$$e_{i,k}^{\text{cp}} + e_{i,k}^{\text{cm}} \leq E_{i,k} \frac{t_k^{\text{round}} M}{T_k \|\Pi_k\|}, \ \forall i \in \Pi_k, \tag{12}$$

$$0 \leq f_{i,k} \leq f_{\max}, \tag{13}$$

$$\sum_{i=1}^{M} \gamma_{i,k} = 1, \ \ \gamma_{i,k} \geq 0. \tag{14}$$

Here $\|\Pi_k\|$ is used to denote the number of scheduled devices. Since the model accuracy increases with the number of training rounds before the FL converges, it is natural to set minimizing the round delay as the objective, which is equivalent to maximizing the number of rounds. Constraint (11) is the round delay constraint. To capture the device energy constraint, we assume a random device scheduling policy, with the number of scheduled devices being $\|\Pi_k\|$, will be applied in the remaining training procedure. Hence, the expected times of each device being scheduled are the expected number of rounds $\frac{T_k}{t_k^{\text{round}}}$ multiplied by the probability of being scheduled $\frac{\|\Pi_k\|}{M}$. Assume that the remaining device energy is equally divided into each time of being scheduled, and thus the energy consumption of the current round is constrained by $e_{i,k}^{\text{cp}} + e_{i,k}^{\text{cm}} \leq E_{i,k} / \frac{T_k \|\Pi_k\|}{t_k^{\text{round}} M}$, which gives (12).

Theorem 1. *Problem* (P2) *is a convex problem. If the feasible region is not empty, then the optimal solution must satisfy one of the followings:*

(1) If $0 < f_{i,k}^* < f_{max}$*, then*

$$\begin{cases} \frac{c\tau d_i}{f_{i,k}^*} + G_{i,k}(\gamma_{i,k}^*) = t_k^{round*}, \\ \kappa_i c \tau d_i {f_{i,k}^*}^2 + P_i G_{i,k}(\gamma_{i,k}^*) = t_k^{round*} \frac{E_{i,k} M}{T_k \|\Pi_k\|}, \end{cases} \tag{15}$$

where $G_{i,k}(\gamma) \triangleq \frac{S}{\gamma B log_2\left(1 + \frac{p_i h_{i,k}^2}{\gamma B N_0}\right)}$.

(2) If $f_{i,k}^* = f_{max}$*, then*

$$\frac{c\tau d_i}{f_{max}} + G_{i,k}(\gamma_{i,k}^*) = t_k^{round*}. \tag{16}$$

Proof. By taking the first and second derivative of the denominator of $G_{i,k}(\gamma)$, we can prove that it is monotonically increasing and concave. Therefore, $G_{i,k}(\gamma)$ is monotonically decreasing and convex [17], and hence problem (P2) is convex.

Due to the limited space, we provide an intuitive explanation to Theorem 1 instead of formally solving the KKT conditions. First suppose $f_{\max}$ is large enough, then to maximize the number of rounds can be conducted, the optimal resource allocation policy needs to use up all time and energy. Therefore, the equality in (11), (12) holds, which gives the first case in Theorem 1. On the other hand, if $f_{\max}$ is not large enough, meaning that device energy is abundant compared with the training time budget. Therefore, $f_{i,k}^*$ is limited by $f_{\max}$, and the equality in (11) holds, and thus the second case in Theorem 1 can be derived by combining them. □

Algorithm 1 Bisection Search Alg. for (P2)

1: Give a big enough t_{up}, initialize $t_{\text{low}} = 0$, $t = t_{\text{up}}$, and success = False
2: **while** NOT success **do**
3: Solve $f_{i,k}^*$ and $\gamma_{i,k}^*$ by substituting $t_k^{\text{round}*} = t$ into Theorem 1
4: Compute the summation of required bandwidth allocation ratio $s = \sum_i \gamma_{i,k}^*$
5: **if** $1 - \varepsilon \leq s \leq 1$ **then**
6: Obtain the solution with accuracy level ε, set success = True
7: **else if** $0 < s < 1 - \varepsilon$ **then**
8: Bandwidth is surplus, set $t_{\text{up}} = t$, $t = \frac{t + t_{\text{low}}}{2}$.
9: **else**
10: Bandwidth is deficient, set $t_{\text{low}} = t$, $t = \frac{t + t_{\text{up}}}{2}$.
11: **end if**
12: **end while**

Since it is hard to analytically solve (P2) because of the non-linear function $G_{i,k}(\cdot)$, we provide Alg. 1 to find numerical solution based on Theorem 1 and (14). The basic idea is that given any potential value of $t_k^{\text{round}*}$, $f_{i,k}^*$ can be first derived by eliminating $G_{i,k}(\gamma_{i,k}^*)$ and solving (15), and then substituting it into Theorem 1 gives the value of $G_{i,k}(\gamma_{i,k}^*)$, which can be further used to derive $\gamma_{i,k}^*$ based on bisection search (step 3). After that, we can search for the optimal $t_k^{\text{round}*}$ by checking whether constraint (14) is satisfied (step 5, 6). Thanks to the monotonicity of $G_{i,k}$, the searching region can be halved according to whether the bandwidth is surplus or deficient for the current $t_k^{\text{round}*}$(step 7-10), corresponding to a low-complexity bisection search algorithm.

B. Device Scheduling

The device scheduling sub-problem is as follows:

$$\min_{K, \boldsymbol{\Pi}_{[K]}} \quad F(\tilde{\boldsymbol{w}}) \tag{P3}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} t_k^{\text{round}} \leq T, \tag{17}$$

$$\sum_{k=1}^{K} \pi_{i,k} (e_{i,k}^{\text{cp}} + e_{i,k}^{\text{cm}}) \leq E_i. \tag{18}$$

There are two main difficulties in solving (P3). First, it is hard if not impossible to derive an analytical expression of the objective (i.e. $F(\tilde{\boldsymbol{w}})$) w.r.t. the number of rounds and the scheduling policy. Therefore, we utilize the FL convergence analysis result in our previous work [7].

Theorem 2. *(Theorem 3 in [7]) Assume the loss functions of all devices satisfy:*

(1) $F_i(\boldsymbol{w})$ *is* ρ*-Lipschitz (i.e.,* $\|F_i(\boldsymbol{w}) - F_i(\boldsymbol{w}')\| \leq \rho \|\boldsymbol{w} - \boldsymbol{w}'\|$, $\forall \boldsymbol{w}, \boldsymbol{w}'$*.),* β*-smooth (i.e.,* $\|\nabla F_i(\boldsymbol{w}) - \nabla F_i(\boldsymbol{w}')\| \leq \beta \|\boldsymbol{w} - \boldsymbol{w}'\|$, $\forall \boldsymbol{w}, \boldsymbol{w}'$*), and convex. (2) For any* i *and*

$\boldsymbol{w}$, the difference between the local gradient and the global gradient can be bounded by $\|\nabla F_i(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\| \leq \delta_i$.

When $\eta \leq \frac{1}{\beta}$ and randomly scheduling $\|\Pi\|$ devices in each round, the difference between $F(\tilde{\boldsymbol{w}})$ and the oracle minimum of global loss function $F(\boldsymbol{w}^*)$ satisfies:

$$\mathbb{E}\left\{\frac{1}{F(\tilde{\boldsymbol{w}}) - F(\boldsymbol{w}^*)}\right\} \geq \frac{1}{\epsilon_0(\|\Pi\|, K) + \rho h(\tau) + B(\|\Pi\|)}, \tag{19}$$

where $\epsilon_0(\|\Pi\|, K) \triangleq \frac{1+\sqrt{1+4\eta\varphi K^2\tau(\rho h(\tau)+B(\|\Pi\|))}}{2\eta\varphi K\tau}$, $\varphi \triangleq \omega\left(1-\frac{\beta\eta}{2}\right)$, $\omega \triangleq \min_k \frac{1}{\|\boldsymbol{w}_k - \boldsymbol{w}^*\|}$, $h(x) \triangleq \frac{\delta}{\beta}((\eta\beta+1)^x - 1) - \eta\delta x$, $\delta \triangleq \frac{\sum_i D_i\delta_i}{D}$, $B(\|\Pi\|) \triangleq \frac{M-\|\Pi\|}{\|\Pi\|} \cdot \frac{\beta\sum_{i=1}^M\sum_{j=1}^M\left(D_i^2D_j^2\left(g_i^2(\tau)+g_j^2(\tau)\right)\right)}{2M(M-1)D_{\min}^2D^2}$, $D_{\min} \triangleq \min_{i\in\mathcal{M}} D_i$, *and* $g_i(x) \triangleq \frac{\delta_i}{\beta}((\eta\beta+1)^x - 1)$.

Theorem 2 quantifies the trade-off between the per-round resource consumption (related to the number of scheduled devices $\|\Pi\|$) and the number of rounds (i.e. K), and thus can be used to approximate the objective of (P3).

Second, since channel states $h_{i,k}$ of future rounds are unknown, and the scheduling policies of different rounds couple with each other due to the total training delay and device energy constraints, solving (P3) optimally is extremely hard. To find a solution, we solve (P3) in each round myopically.

As a result, the myopic problem in the k-th round with the approximated objective is given by

$$\min_{\Pi_k} \quad \epsilon_0(\|\Pi_k\|, \hat{K}) + \rho h(\tau) + B(\|\Pi_k\|) \tag{P4}$$

$$\text{s.t.} \quad \hat{K} = k - 1 + \left\lfloor \frac{T_k}{t_k^{\text{round}*}(\Pi_k)} \right\rfloor, \tag{20}$$

where T_k is the remaining training time budget, $t_k^{\text{round}*}(\Pi_k)$ is the optimal round delay given by Alg. 1, and thus $\left\lfloor \frac{T_k}{t_k^{\text{round}*}(\Pi_k)} \right\rfloor$ is the maximum expected rounds can be conducted with remaining resources, and $\hat{K}$ is the maximum expected total rounds.

Since (P4) is still a combinatorial problem, we provide a greedy algorithm to schedule devices. In Alg. 2, the optimal round delay of scheduling each unscheduled device is given by Alg. 1, based on which we can choose the device with the minimum round delay (step 3). Then the objective of (P4) is estimated (step 4). We will iteratively add a device into the scheduled device set until the objective begins to increase (step 5-8), which means that scheduling more devices will decrease the convergence rate of FL according to Theorem 2.

Algorithm 2 Greedy Scheduling Algorithm

1: Initialize $\Pi_k = \emptyset$, and $O = \infty$
2: **while** $|\mathcal{M}| > 0$ **do**
3: Greedy scheduling: $x = \arg\min_{i\in\mathcal{M}} \left(t_k^{\text{round}*}(\Pi_k \cup \{i\})\right)$.
4: Estimate $\hat{K} = k - 1 + \left\lfloor \frac{T_k}{t_k^{\text{round}*}(\Pi_k\cup\{x\})} \right\rfloor$, and $O' = \epsilon_0(\|\Pi_k\|+1, \hat{K}) + \rho h(\tau) + B(\|\Pi_k\|+1)$
5: **if** $O' > O$ **then**
6: **break**
7: **end if**
8: Update $\mathcal{M} \leftarrow \mathcal{M} \setminus \{x\}$, $\Pi_k \leftarrow \Pi_k \cup \{x\}$, and $O \leftarrow O'$
9: **end while**

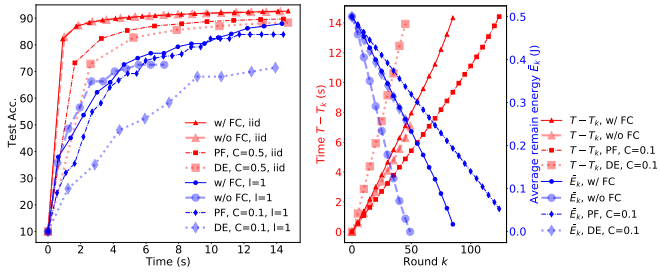
IV. Experiment

We consider an FL system with $M = 20$ devices, uniformly located in a cell with a radius of 600 meters, and a BS located in the center [7]. The path loss model is $128.1 + 37.6\log_{10}d$ (d is the distance between device and BS, in km) [10]. The wireless bandwidth is $B = 20$ MHz, the device transmit power is set to be $p_i = 5$ dBm and $P_i = 10$ dBm, while the spectrum density of noise is $N_0 = -179$ dBm/Hz. The effective switched capacitance of mobile devices is $\kappa_i = 10^{-28}$ [10], and the maximum computing frequency is $f_{\max} = 2$ GHz.

The proposed algorithm is evaluated on two well-known datasets, MNIST [18] and CIFAR-10 [19]. Different training data distributions are considered, namely i.i.d. and non-i.i.d.. For i.i.d. dataset, the whole training dataset is randomly shuffled and partitioned into 20 shards, and each device is assigned with a shard. While for non-i.i.d. dataset, all training samples are sorted according to the label and partitioned into $20l$ shards, and then each device is assigned with l shards. Here l captures the non-i.i.d. level, where a smaller l means a higher non-i.i.d. level. We apply a multilayer perceptron (MLP) model with one hidden layer of 64 nodes for MNIST, and a convolutional neural network (CNN) model with the standard LeNet-5 architecture [18] for CIFAR-10. Each scheduled device performs $\tau = 5$ times of SGD step for local model update, where the mini-batch size is $d_i = 128$ and the learning rate is $\eta = 0.01$. According to the architecture of learning models, we have $S = 204$ kB, $c = 2.03 \times 10^5$ for MNIST, and $S = 246$ kB, $c = 2.6 \times 10^6$ for CIFAR-10, respectively. For the delay and energy constraints, we set $T = 15$ s, $E_i = 0.5$ J for MNIST, and $T = 2000$ s, $E_i = 50$ J for CIFAR-10, respectively.

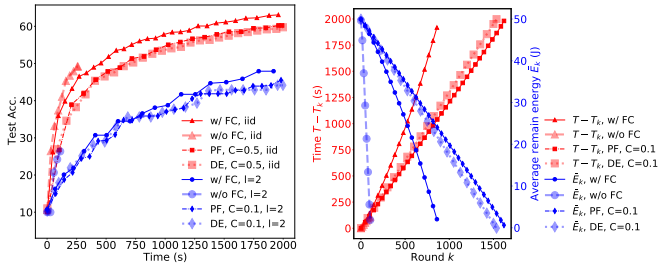
In Fig. 1 and Fig. 2, we compare the performance of the proposed scheme with 3 baselines. The first baseline is the joint device scheduling and bandwidth allocation scheme in [7]. Since [7] does not consider local CPU frequency control to reduce energy consumption, we set the local CPU frequency $f_i = f_{\max}$, and use w/o FC to denote the first baseline. On the contrary, we use w/ FC to denote the proposed scheme. To evaluate the effects of the proposed device scheduling policy, we compare it with the proportional fair policy proposed in [6] (denoted by PF) that schedules the devices with best channel conditions, and a policy that schedules the devices with most remaining energy (denoted by DE), both with Alg.1 for local CPU frequency control and bandwidth allocation. For PF and DE, we use C to denote the fraction of scheduled devices in each round, where C is set to be 0.1 and 0.5.

Fig. 1. (a) shows the test accuracy v.s. time on MNIST, where results of i.i.d. dataset is in red and results of non-i.i.d. dataset with $l = 1$ is in blue. We notice that the proposed scheme outperforms PF and DE, by improving the accuracy from 89.7 %, 88.4 % to 92.7 % on the i.i.d. dataset, and from



(a) (b)

Fig. 1. Comparison between different schemes on MNIST dataset. (a) Test accuracy v.s. time. (b) Training time and average remain device energy v.s. round on the *non-i.i.d. dataset* ($l = 1$). Results are averaged over 5 independent trails.



(a) (b)

Fig. 2. Comparison between different schemes on CIFAR-10 dataset. (a) Test accuracy v.s. time. (b) Training time and average remain device energy v.s. round on the *non-i.i.d. dataset* ($l = 2$). Results are averaged over 5 independent trails.

83.9 %, 71.5 % to 88.0 % on the non-i.i.d. dataset, respectively. The reason is that the proposed scheme can adaptively decide how many devices to schedule based on the data distribution, while PF and DE with a fixed fraction of scheduled devices can hardly adapt to different data distributions. Further, the proposed scheme takes both channel condition and remaining energy into account, while either PF or DE only considers one of them. Fig. 1. (b) shows the elapsed training time and the average remaining device energy v.s. round on the *non-i.i.d. dataset with* $l = 1$ of all schemes. Since the w/o FC scheme uses the maximum local CPU frequency for local model updating, it consumes more energy than the proposed w/ FC scheme, and exhausts all device energy in only 50 rounds, leading to a lower model accuracy as shown in Fig. 1. (a). Fig. 1. (b) also shows that since PF chooses the devices with best channel conditions, it consumes less time and energy compared with DE, leading to a faster convergence.

Fig. 2. (a) and (b) show the results on CIFAR-10, where the proposed scheme also outperforms all baselines. However, due to the relatively heavier computing workload of updating the CNN model, the improvement of choosing the devices with best channel conditions is reduced, and thus the convergence performance of PF and DE is similar.

V. Conclusions

We have studied the device scheduling and resource allocation problem for FL over wireless networks, with the goal of maximizing the model accuracy under total training delay and device energy budgets. Our proposed scheme is adaptive to various resource constraints, and can automatically find the balance between the per-round training cost and the total number of training rounds to optimize the training performance. Experiments on various datasets and learning models have demonstrated the performance improvement of the proposed scheme, compared with state-of-the-art schemes. In the future, the heterogeneity in channel conditions, computing capabilities, and data distributions can be further considered.

References

[1] K. Bonawitz *et al.*, "Towards federated learning at scale: System design," *arXiv preprint arXiv:1902.01046*, 2019.

[2] Y. Sun, W. Shi, X. Huang, S. Zhou, and Z. Niu, "Edge learning with timeliness constraints: Challenges and solutions," *IEEE Communications Magazine*, vol. 58, no. 12, pp. 27–33, 2020.

[3] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 46–51, 2020.

[4] Z. Yang, M. Chen, K.-K. Wong, H. V. Poor, and S. Cui, "Federated learning for 6G: Applications, challenges, and opportunities," *arXiv preprint arXiv:2101.01338*, 2021.

[5] D. Gündüz, D. B. Kurka, M. Jankowski, M. M. Amiri, E. Ozfatura, and S. Sreekumar, "Communicate to learn at the edge," *IEEE Communications Magazine*, vol. 58, no. 12, pp. 14–19, 2020.

[6] H. H. Yang, Z. Liu, T. Q. S. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Transactions on Communications*, vol. 68, no. 1, pp. 317–333, 2020.

[7] W. Shi, S. Zhou, Z. Niu, M. Jiang, and L. Geng, "Joint device scheduling and resource allocation for latency constrained wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 453–467, 2020.

[8] M. Chen, H. V. Poor, W. Saad, and S. Cui, "Convergence time optimization for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2457–2471, 2021.

[9] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, "Energy-efficient radio resource allocation for federated edge learning," in *2020 IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, pp. 1–6, 2020.

[10] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1935–1949, 2021.

[11] X. Mo and J. Xu, "Energy-efficient federated edge learning with joint communication and computation design," *Journal of Communications and Information Networks*, vol. 6, no. 2, pp. 110–124, 2021.

[12] M. Mohammadi 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.

[13] 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, 2020.

[14] Y. Sun, S. Zhou, and D. Gündüz, "Energy-aware analog aggregation for federated learning with redundant data," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pp. 1–7, IEEE, 2020.

[15] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

[16] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolić, *Digital integrated circuits: a design perspective*, vol. 7. Pearson education Upper Saddle River, NJ, 2003.

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

[18] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[19] A. Krizhevsky, V. Nair, and G. Hinton, "The CIFAR-10 dataset," *online: http://www. cs. toronto. edu/kriz/cifar. html*, 2014.