Convergence Acceleration in Wireless Federated Learning: A Stackelberg Game Approach

Kaidi Wang, *Member, IEEE*, Yi Ma, *Senior Member, IEEE*, Mahdi Boloursaz Mashhadi, *Senior Member, IEEE*, Chuan Heng Foh, *Senior Member, IEEE*, Rahim Tafazolli, *Senior Member, IEEE*, and Zhi Ding, *Fellow, IEEE*

***Abstract*—This paper studies issues that arise with respect to the joint optimization for convergence time in federated learning over wireless networks (FLOWN). We consider the criterion and protocol for selection of participating devices in FLOWN under the energy constraint and derive its impact on device selection. In order to improve the training efficiency, age-of-information (AoI) enables FLOWN to assess the freshness of gradient updates among participants. Aiming to speed up convergence, we jointly investigate global loss minimization and latency minimization in a Stackelberg game based framework. Specifically, we formulate global loss minimization as a leader-level problem for reducing the number of required rounds, and latency minimization as a follower-level problem to reduce time consumption of each round. By decoupling the follower-level problem into two sub-problems, including resource allocation and sub-channel assignment, we achieve an optimal strategy of the follower through monotonic optimization and matching theory. At the leader-level, we derive an upper bound of convergence rate and subsequently reformulate the global loss minimization problem and propose a new age-of-update (AoU) based device selection algorithm. Simulation results indicate the superior performance of the proposed AoU based device selection scheme in terms of the convergence rate, as well as efficient utilization of available sub-channels.**

***Index Terms*—Wireless federated learning, Stackelberg game, age-of-information, device selection, resource allocation, sub-channel assignment.**

I. Introduction

The rapid development of mobile devices and applications has ushered us into the fifth-generation (5G) era. Much of the network services in 5G and beyond is expected to address explosive growth and need of machine learning (ML) and data science [1]. In conventional centralized ML, a central server is equipped at the access point (AP) to collect all raw data for model training. However, due to the limited wireless resources and potential privacy issues, centralized ML is impractical for some scenarios [2]. In this context, federated learning (FL) is a framework for distributed ML algorithms to collaboratively train a central learning model while keeping the data locally [3]. Specifically, in FL, a global model is shared among multiple devices, and each device trains the received global model based on the local data and produces a local model [4]. Thereafter all local models are transmitted to the server via wireless communication networks to generate the updated global model [5]. Since raw data do not leave the device, and the model size is much smaller to share, such that there is less concern about data privacy and lower consumption of data network resources [6].

A. Related Works

Owing to its growing popularity, FL-related design and optimization in existing wireless communication architectures have attracted widespread attention, in which convergence time is regarded as an important performance metric. As indicated in [7], the convergence time is jointly determined by the number of communication rounds and the time consumption per round. The former is closely related to convergence rate, while the latter is normally defined as latency.

In view of the relationship between global loss and convergence rate, some works focused on the global loss minimization problem in order to reduce the number of required communication rounds [8]–[11]. In [8], a FL algorithm with multiple local training was designed. It considered the impact of local and global update rounds on the convergence bound, and developed an approximate solution of the global loss minimization problem. In [9], packet error rate was introduced to indicate whether the FL parameter transmission was successful or not. Specifically, the joint problem of user selection, resource block allocation, and power allocation was studied under delay and energy constraints [9]. Incorporating FL in a massive multiple-input-multiple-output (MIMO) scenario with energy harvesting, the authors of [10] included the consideration of user scheduling and power allocation in a global loss minimization problem. Adopting a model pruning scheme, in [11], the authors jointly optimized device selection, time slot allocation and pruning ratio in order to maximize the convergence rate with a latency constraint.

As another factor in determining the convergence time of FL, latency, including computation time and communication time, was extensively researched in previous works [12]–[15]. The authors of [12] designed a realistic wireless network for FL, where a limited number of users can be selected at each round for aggregation. By obtaining user selection and resource block allocation schemes, the convergence time

This work was supported in part by the U.K. Department for Science, Innovation, and Technology under Project TUDOR (Towards Ubiquitous 3D Open Resilient Network). This material is based upon work supported by the National Science Foundation under Grant No. 2332760.

K. Wang is with the Department of Electrical and Electronic Engineering, The University of Manchester, Manchester, M13 9PL, UK (email: kaidi.wang@ieee.org).

Y. Ma, M. Boloursaz Mashhadi, C. H. Foh, and R. Tafazolli are with 5GIC & 6GIC, Institute for Communication Systems (ICS), University of Surrey, Guildford, UK (email: y.ma@surrey.ac.uk; m.boloursazmashhadi@surrey.ac.uk; c.foh@surrey.ac.uk; r.tafazolli@surrey.ac.uk).

Z. Ding is with the Department of Electrical and Computer Engineering, University of California at Davis, Davis, CA 95616 USA (email: zding@ucdavis.edu).

of FL was minimized. Setting a local accuracy level at each device, the FL algorithm with multiple local update rounds was proposed in a cell-free massive MIMO scenario [13], in which time consumption for downlink transmission, uplink transmission, and computation was considered in the formulated training time minimization problem. A multi-task FL framework was studied in a multi-access edge computing (MEC) scenario, where edge nodes were included to accomplish different learning tasks [14]. In order to minimize the latency of each communication round, the optimal matching between edge nodes and end devices was obtained. In [15], the authors proposed a hybrid learning scheme, where part of data can be offloaded from devices to the server, while the remaining data was utilized for local training. It was demonstrated that the proposed scheme has the ability to reduce the total time consumption.

B. Motivation and Contribution

Although global loss minimization and latency minimization have been separately studied in existing works [8]–[15], the interaction between these two objectives remains unclear. Specifically, devices that have a significant constructive impact on training convergence may have poor channel conditions, thereby increasing the latency of the corresponding aggregation round. On the other hand, focusing on minimizing latency may cause devices with high channel gains to be repeatedly selected, leading to an increase in the global loss [16], [17]. Therefore, it is necessary to investigate this interaction and construct a dynamic trade-off. To this end, this work adopts Stackelberg game and presents a novel framework to jointly consider learning and communication in wireless FL systems, where the server and devices tend to minimize the global loss and latency, respectively. Unlike the conventional papers on global loss minimization that treat latency as a definite threshold [8]–[11], latency in this work can be flexibly adjusted to ensure the convergence rate. Compared to [18], an energy budget is included and its impact on device selection is analyzed, constructing a more practical and challenging scenario. On the other hand, inspired by the concept of age-of-information (AoI) [19], age-of-update (AoU) [20] is defined in this work as a metric to evaluate the staleness of model updates. In this context, a novel device selection method is designed to estimate the contribution of devices in each communication round without analyzing the model/gradient or transmitting additional information to the server. Different from [20] and [21] which target overall AoU/AoI minimization, AoU in this work is regarded as a weight to prioritize selecting devices with larger AoU. The main contributions of this paper are summarized as follows:

- A latency-sensitive FL scenario is considered, where multiple devices transmit parameters to the server over a limited number of sub-channels. In order to jointly minimize global loss and latency, a Stackelberg game based problem is formulated, where global loss minimization and latency minimization are considered as leader-level and follower-level problems, respectively. It is proved that the with the given sub-channels, some devices cannot transmit local models to the server due to the energy consumption constraint. Based on the analysis, the Stackelberg equilibrium of the formulated problems is established.
- The follower-level problem is divided into two sub-problems, including resource allocation and sub-channel assignment. Due to non-convexity and monotonicity, a monotonic optimization based solution is proposed for the resource allocation problem. Moreover, a matching based algorithm is developed to address the sub-channel assignment problem with the incomplete preference list, where the properties of the proposed algorithm are analyzed.
- In order to solve the leader-level problem, the upper bound of the convergence rate is derived, which indicates that the convergence rate can be improved by selecting devices with large data size. Therefore, the global loss minimization problem is reformulated as a weighted device selection problem. By ordering devices based on AoU and data size, a priority list is created, and an algorithm is designed to select devices by predicting sub-channel assignment and resource allocation.
- The simulation results on Modified National Institute of Standards and Technology (MNIST), Canadian Institute for Advanced Research (CIFAR-10) and Stanford Sentiment Treebank Version 2 (SST-2) databases are presented. It is indicated that the designed AoU based device selection scheme can improve the convergence rate and achieve the lowest global loss. Moreover, the proposed solutions for resource allocation and sub-channel assignment can efficiently utilize available sub-channel and dynamically adjust energy utilization in order to reduce latency.

C. Organization

The remainder of this paper is organized as follows. In Section II and Section III, the system model and problem formulation are described, respectively. The solution of latency minimization problem is presented in Section IV, and the solution of global loss minimization problem is obtained in Section V. Section VI demonstrates the simulation results. The conclusions are summarized in Section VII.

II. System Model

Consider an FL scenario where N wireless mobile devices collaboratively train a joint learning model. Each device is equipped with a single antenna and the FL process is orchestrated by a wireless server. In each communication round, the devices intend to train neural networks based on local data and then transmit parameters to the server for aggregation. Moreover, the limited communication resources are considered. Specifically, there are K available sub-channels, $K \leq N$, and each sub-channel is occupied by at most one device. Therefore, only a subset of devices can be selected for the global model aggregation in each communication round. The collections of all devices and sub-channels are denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$ and $\mathcal{K} = \{1, 2, \ldots, K\}$, respectively.

A. Computation Model

In each communication round, after receiving the global model, the selected devices need to train their respective local learning models with the equipped central processing units (CPUs). Based on the dynamic voltage and frequency scaling (DVFS) technique, the CPU core can be operated at different frequency levels, and hence, the consumed time and energy change accordingly [22]. For any device n assigned to sub-channel k, the computational time consumption is given by

$$T_{k,n}^{\mathrm{cp}}(\tau_{k,n}) = \frac{\mu\beta_n}{\tau_{k,n}C_n}, \tag{1}$$

where μ is a coefficient to denote the required CPU cycles for training one sample, β_n is the number of dataset samples at device n, $\tau_{k,n}$ is a designed proportion of computational capacity, and C_n is the CPU frequency of device n. Note that since the size of local data utilized in training does not change over the computing time, the test accuracy or loss reduction is not affected. The energy consumption for computation can be expressed as follows:

$$E_{k,n}^{\mathrm{cp}}(\tau_{k,n}) = \kappa_0\mu\beta_n(\tau_{k,n}C_n)^2, \tag{2}$$

where κ_0 is the power consumption coefficient per CPU cycle.

B. Communication Model

After training, local models are transmitted from selected devices to the server. For any device n, the achievable data rate at sub-channel k is given by

$$R_{k,n}(p_{k,n}) = B\log_2(1 + p_{k,n}|h_{k,n}|^2), \tag{3}$$

where B is the bandwidth, $p_{k,n}$ is the power allocation coefficient of device n, and $|h_{k,n}|^2$ is the normalized channel gain. Particularly, $|h_{k,n}|^2 = P_t|g_{k,n}|^2\eta d_n^{-a}\sigma^{-2}$, where P_t is the maximum transmit power in each sub-channel, $g_{k,n} \sim CN(0,1)$ the small-scale fading coefficient, η is the frequency dependent factor, d_n is the distance between device n and the server, a is the path loss exponent, and σ^2 is the variance of additive white Gaussian noise (AWGN). In the considered scenario, the small-scale fading coefficients vary with the communication rounds[1]. For simplicity, the index related to communication rounds is omitted from the notations. Based on the achievable data rate, the time consumption for communication can be presented by

$$T_{k,n}^{\mathrm{cm}}(p_{k,n}) = \frac{D(\boldsymbol{w}_n^{(\mathrm{t})})}{R_{k,n}(p_{k,n})}, \tag{4}$$

where $D(\boldsymbol{w}_n^{(\mathrm{t})})$ is the size of the local model $\boldsymbol{w}_n^{(\mathrm{t})}$ at device n in round t. It is assumed that the data size of local models is the same for all devices and rounds, i.e., $D(\boldsymbol{w}) = D(\boldsymbol{w}_n^{(\mathrm{t})}), \forall n, t$. The energy consumption for communication is given by

$$E_{k,n}^{\mathrm{cm}}(p_{k,n}) = p_{k,n}P_tT_{k,n}^{\mathrm{cm}}(p_{k,n}). \tag{5}$$

C. AoU based Device Selection

In any communication round t, a subset of devices, denoted by $\mathcal{N}_t$, is selected, i.e., $\mathcal{N}_t \subseteq \mathcal{N}$ and $|\mathcal{N}_t| \leq K$. The status of any device in round t can be represented by a binary variable $S_n^{(\mathrm{t})} \in \{0,1\}$, where $S_n^{(\mathrm{t})} = 1$ indicates device n is selected to participate in the aggregation in round t; $S_n^{(\mathrm{t})} = 0$ otherwise. The set of all device selection indicators is denoted by $\mathbf{S}^{(\mathrm{t})}$. Therefore, in any round, the set of selected devices $\mathcal{N}_t$ and the set of device selection indicators $\mathbf{S}^{(\mathrm{t})}$ can be mutually inferred.

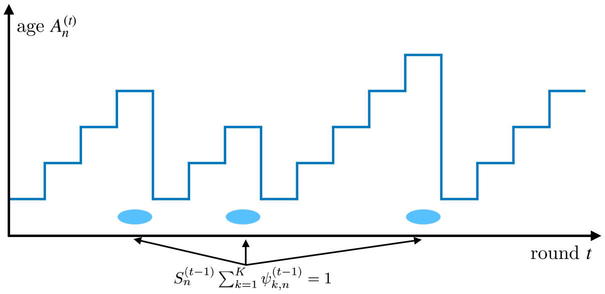


Fig. 1: An illustration of device's information age.

In order to improve the performance of FL, device selection should be determined by data quality. It is revealed in [19], [26] that staleness of gradient update may negatively impact learning outcomes. Therefore, the server tends to select the devices with fresher gradient update for aggregation. In particular, if a device has been skipped for several rounds, its gradient update is relatively informative, and the probability of the server selecting the device should increase. Conversely, if a device was selected in the previous round, the probability for its reselection should decrease. To this end, the concept of AoI is adopted to define AoU [20], [27]. In round t, any device n's AoU, denoted by $A_n^{(\mathrm{t})}$, is presented as follows:

$$A_n^{(\mathrm{t})} = \begin{cases} A_n^{(\mathrm{t-1})} + 1, & \text{if} \quad S_n^{(\mathrm{t-1})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t-1})} = 0, \\ 1, & \text{if} \quad S_n^{(\mathrm{t-1})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t-1})} = 1, \end{cases} \tag{6}$$

where $\psi_{k,n}^{(\mathrm{t})} \in \{0,1\}$ is the sub-channel assignment indicator. Particularly, $\psi_{k,n}^{(\mathrm{t})} = 1$ indicates that device n is assigned to sub-channel k in round t; $\psi_{k,n}^{(\mathrm{t})} = 0$ otherwise. According to the above definition, AoU represents the number of communication rounds since the last transmission, which is jointly decided by device selection and sub-channel assignment. As shown in Fig. 1, if device n was not selected in round $t-1$, i.e., $S_n^{(\mathrm{t-1})} = 0$, or it is not assigned to any sub-channel, i.e., $\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t-1})} = 0$, its AoU is incremented by 1; otherwise, its AoU is reset to 1. With this consideration, a large AoU implies more informative update, and therefore device selection should partially depend on AoU. In any round t, $\alpha_n^{(\mathrm{t})}$ defined below is used as a weight[2] to prioritize selecting devices with larger AoU for aggregation:

$$\alpha_n^{(\mathrm{t})} = \frac{A_n^{(\mathrm{t})}}{\sum_{i=1}^{N}A_i^{(\mathrm{t})}}. \tag{7}$$

D. Sub-Channel Assignment

The time consumption of each communication round, i.e., latency, also plays an important role in FL [7], [12], [28]. By

[1] Despite the location of devices is considered stationary in this paper, the proposed scheme can be extended to mobile scenarios, such as [23]–[25].

[2] In this work, the term "weight" refers to the weighting coefficient, and the neural network weights are represented by the term "model".

including computation and communication phases, the time consumption of any device n assigned to sub-channel k can be expressed as follows:

$$T_{k,n}(\tau_{k,n}, p_{k,n}) = T^{\mathrm{cp}}_{k,n}(\tau_{k,n}) + T^{\mathrm{cm}}_{k,n}(p_{k,n}), \tag{8}$$

where the time consumption for global model transmission from the server to devices is ignored as in [29], [30]. With the given set of selected devices $\mathcal{N}_t$, the latency in round t can be presented as follows:

$$T^{(\mathrm{t})} = \max_{n \in \mathcal{N}_t} \left\{ \sum_{k=1}^{K} \psi^{(\mathrm{t})}_{k,n} T_{k,n}(\tau_{k,n}, p_{k,n}) \right\}. \tag{9}$$

It indicates that the latency is affected by sub-channel assignment. Specifically, if any device is assigned to a sub-channel with an enhanced channel gain, the time consumption of this device decreases, and hence, the latency of this round may be reduced. Accordingly, the energy consumption of device n assigned to sub-channel k is given by

$$E_{k,n}(\tau_{k,n}, p_{k,n}) = E^{\mathrm{cp}}_{k,n}(\tau_{k,n}) + E^{\mathrm{cm}}_{k,n}(p_{k,n}). \tag{10}$$

III. Problem Formulation

This work focuses on minimizing the convergence time of FL, which is defined as the sum of latency across all communication rounds [7]. Therefore, the convergence time is determined by i) the number of communication rounds, and ii) the time consumption of each round. That is, the server can select devices with more data and larger AoU to reduce the number of required rounds, or devices with better channel conditions to reduce the latency in each round. However, devices with more data and larger AoU are often not preferred due to larger time and energy consumption, while devices with good channel conditions can be frequently selected, which usually causes smaller AoU. Therefore, there exists a trade-off between these two options. In order to comprehensively minimize convergence time, the number of required communication rounds and latency should be jointly minimized. This situation can be described by the Stackelberg game, where global loss minimization and latency minimization are considered as leader-level and follower-level problems, respectively.

1) Leader-Level Problem (Computation): In the considered FL algorithm, the local loss at device n is given by

$$f_n(\boldsymbol{w}^{(\mathrm{t})}) = \frac{1}{\beta_n} \sum_{i=1}^{\beta_n} \ell(\boldsymbol{w}^{(\mathrm{t})}; \boldsymbol{x}_{n,i}, y_{n,i}), \tag{11}$$

where $\boldsymbol{w}^{(\mathrm{t})}$ is the global model in round t, $\ell(\boldsymbol{w}^{(\mathrm{t})}; \boldsymbol{x}_{n,i}, y_{n,i})$ is a loss function, $\boldsymbol{x}_{n,i}$ is the i-th input data of device n, and $y_{n,i}$ is the corresponding label. In round t, the global loss can be presented below

$$F(\boldsymbol{w}^{(\mathrm{t})}) = \frac{\sum_{n=1}^{N} \sum_{i=1}^{\beta_n} \ell(\boldsymbol{w}^{(\mathrm{t})}; \boldsymbol{x}_{n,i}, y_{n,i})}{\sum_{n=1}^{N} \beta_n}. \tag{12}$$

By including device selection and sub-channel assignment, an AoU based global loss minimization problem is formulated as follows:

$$\min_{\mathbf{S}^{(\mathrm{t})}} \quad \frac{\sum_{n=1}^{N} \alpha^{(\mathrm{t})}_n S^{(\mathrm{t})}_n \sum_{k=1}^{K} \psi^{(\mathrm{t})}_{k,n} \sum_{i=1}^{\beta_n} \ell(\boldsymbol{w}^{(\mathrm{t})}; \boldsymbol{x}_{n,i}, y_{n,i})}{\sum_{n=1}^{N} \alpha^{(\mathrm{t})}_n S^{(\mathrm{t})}_n \sum_{k=1}^{K} \psi^{(\mathrm{t})}_{k,n} \beta_n}, \tag{13}$$

$$\text{s.t.} \quad S^{(\mathrm{t})}_n \in \{0,1\}, \forall n \in \mathcal{N}, \tag{13a}$$

$$\sum\nolimits_{n=1}^{N} S^{(\mathrm{t})}_n \le K. \tag{13b}$$

Constraint (13b) indicates that the number of selected devices in each round is not greater than the number of available sub-channels. Note that only device selection is considered in the computation phase, and the sub-channel assignment indicator is decided in the communication phase, even though it has an impact on the global loss.

2) Follower-Level Problem (Communication): Based on the $\mathbf{S}^{(\mathrm{t})}$ given by the leader-level problem, the set of selected devices $\mathcal{N}_t$ can be obtained. Then, sub-channel assignment and resource allocation can be implemented according to the set $\mathcal{N}_t$. In any round t, the latency minimization problem can be presented as follows:

$$\min_{\boldsymbol{\psi}^{(\mathrm{t})}, \boldsymbol{\tau}, \boldsymbol{p}} \quad \max_{n \in \mathcal{N}_t} \left\{ \sum_{k=1}^{K} \psi^{(\mathrm{t})}_{k,n} T_{k,n}(\tau_{k,n}, p_{k,n}) \right\}, \tag{14}$$

$$\text{s.t.} \quad E_{k,n}(\tau_{k,n}, p_{k,n}) \le E^{\max}_n, \tag{14a}$$

$$\tau_{k,n} \in [0,1], p_{k,n} \in [0,1], \tag{14b}$$

$$\psi^{(\mathrm{t})}_{k,n} \in \{0,1\}, \tag{14c}$$

$$\sum\nolimits_{n \in \mathcal{N}_t} \psi^{(\mathrm{t})}_{k,n} = 1, \forall k \in \mathcal{K}, \tag{14d}$$

$$\sum\nolimits_{k=1}^{K} \psi^{(\mathrm{t})}_{k,n} = 1, \forall n \in \mathcal{N}_t, \tag{14e}$$

where $\boldsymbol{\psi}^{(\mathrm{t})}$, $\boldsymbol{\tau}$, and $\boldsymbol{p}$ are the sets of all sub-channel assignment indicators, computational resource allocation coefficients, and power allocation coefficients, respectively. In constraint (14a), the maximum energy consumption $E^{\max}_n$ is included. In constraints (14b) and (14c), the value ranges of all optimization variables are defined. Constraints (14d) and (14e) indicate that each sub-channel can be occupied by one device, and each device can be assigned to one sub-channel, respectively.

3) Stackelberg Equilibrium: With any given solution of the leader-level problem, the formulated follower-level problem can be infeasible, as shown in follows:

Proposition 1. *With any device selection $S^{(\mathrm{t})}_n = 1$, problem (14) is infeasible if the following condition holds:*

$$\ln(2) P_t D(\boldsymbol{w}^{(\mathrm{t})}_n) \ge E^{\max}_n B |h_{k,n}|^2, \forall k \in \mathcal{K}. \tag{15}$$

Proof: Refer to Appendix A. ■

Proposition 1 indicates that the energy consumption constraint can affect device participation, thereby reducing learning performance [31]. The following remark can be obtained.

Remark 1. *In wireless FL scenarios, energy consumption constraints can restrict the transmission of local models individually, resulting in a decrease in the global loss.*

The above proposition and remark show that there exists an interaction between these two problems. That is, the selected devices may not be able to transmit local models to the server due to the poor channel conditions[3]. This interaction is consistent with the Stackelberg competition model, in which the leader and the follower tend to maximize their own utilities [32]. For solving the formulated Stackelberg game based problem, Stackelberg equilibrium [33] is introduced as follows:

Definition 1. *In the formulated Stackelberg game based problem, by respectively defining G_{L} and G_{F} as the objective functions of leader-level and follower-level problems, solution $(\mathbf{S}^{(\mathrm{t})*}, \boldsymbol{\psi}^{(\mathrm{t})*}, \boldsymbol{\tau}^*, \boldsymbol{p}^*)$ is the Stackelberg equilibrium if the following conditions hold:*

$$\begin{aligned} &G_{\mathrm{L}}(\mathbf{S}^{(\mathrm{t})*}, \boldsymbol{\psi}^{(\mathrm{t})*}, \boldsymbol{\tau}^*, \boldsymbol{p}^*) \leq G_{\mathrm{L}}(\mathbf{S}^{(\mathrm{t})}, \boldsymbol{\psi}^{(\mathrm{t})*}, \boldsymbol{\tau}^*, \boldsymbol{p}^*), \\ &G_{\mathrm{F}}(\mathbf{S}^{(\mathrm{t})*}, \boldsymbol{\psi}^{(\mathrm{t})*}, \boldsymbol{\tau}^*, \boldsymbol{p}^*) \leq G_{\mathrm{F}}(\mathbf{S}^{(\mathrm{t})*}, \boldsymbol{\psi}^{(\mathrm{t})}, \boldsymbol{\tau}, \boldsymbol{p}). \end{aligned} \tag{16}$$

In order to achieve the Stackelberg equilibrium, the leader should predict the possible solution of the follower-level problem, i.e., the feasibility of the selected devices, and then propose a strategy which can minimize the global loss after transmission. As a result, the solution of the follower-level problem should be obtained before the leader-level problem. Note that both problems are solved at the server, and the solutions of each communication round can be transmitted to all devices together with the global model. Since the server has powerful computing capabilities, the impact of this process on latency can be ignored.

IV. Solution of Follower-Level Problem

The formulated follower-level problem in (14) is a non-convex problem with binary constraints. In this section, the follower-level problem is decoupled into two sub-problems and solved iteratively. With the fixed sub-channel assignment, the sub-problem related to resource allocation is given by

$$\begin{aligned} \boldsymbol{\Gamma} = \min_{\boldsymbol{\tau}, \boldsymbol{p}} \quad & \left\{ T_{k,n}(\tau_{k,n}, p_{k,n}) | \forall n \in \mathcal{N}_t, \forall k \in \mathcal{K} \right\}, \\ \text{s.t.} \quad & \text{(14a), (14b)}, \end{aligned} \tag{17}$$

where $\boldsymbol{\Gamma}$ is a $K \times |\mathcal{N}_t|$ matrix containing the minimum time consumptions for all possible device and sub-channel combinations. With the given matrix $\boldsymbol{\Gamma}$, the sub-problem related to sub-channel assignment can be presented as follows:

$$\begin{aligned} \min_{\boldsymbol{\psi}^{(\mathrm{t})}} \quad & \max_{n \in \mathcal{N}_t} \left\{ \sum_{k=1}^{K} \psi_{k,n}^{(\mathrm{t})} \Gamma_{k,n} \right\}, \\ \text{s.t.} \quad & \text{(14c), (14d), (14e)}, \end{aligned} \tag{18}$$

where $\Gamma_{k,n}$ is one element of matrix $\boldsymbol{\Gamma}$. Note that there exists an infeasible combination if the condition in Proposition 1 holds. In this case, this combination will be marked as infeasible and avoided in problem (18).

A. Joint Optimization of Computational Resource Allocation and Power Allocation

In problem (17), the time consumption of any combination only depends on its corresponding resource allocation. Therefore, this problem can be divided into multiple sub-problems to obtain all elements of matrix $\boldsymbol{\Gamma}$. The feasibility of combinations is verified by utilizing Proposition 1. For any feasible combination $\Gamma_{k,n}$, the sub-problem is given by

$$\min_{\tau_{k,n}, p_{k,n}} \quad T_{k,n}^{\mathrm{cp}}(\tau_{k,n}) + T_{k,n}^{\mathrm{cm}}(p_{k,n}) \tag{19}$$

$$\text{s.t.} \quad E_{k,n}^{\mathrm{cp}}(\tau_{k,n}) + E_{k,n}^{\mathrm{cm}}(p_{k,n}) \leq E_n^{\max}, \tag{19a}$$

$$\tau_{k,n} \in [0,1], p_{k,n} \in [0,1]. \tag{19b}$$

Since problem (19) is non-convex, traditional optimization methods, such as convex optimization, cannot be directly employed. In this case, monotonic optimization is introduced. In order to utilize monotonic optimization, the monotonicity of problem (19) should be analyzed. According to [34], the following proposition can be obtained.

Proposition 2. *For any device n assigned to sub-channel k, with computational resource allocation coefficient $\tau_{k,n}$ and power allocation coefficient $p_{k,n}$, the time consumption $T_{k,n}(\tau_{k,n}, p_{k,n})$ is a decreasing function, while the energy consumption $E_{k,n}(\tau_{k,n}, p_{k,n})$ is an increasing function.*

Proof: Refer to Appendix B. ■

Proposition 2 indicates the following remark.

Remark 2. *In the considered FL framework, minimizing the latency leads to the maximized energy consumption.*

According to Proposition 2, it is indicated that the objective function and all constraints in problem (19) are monotonic. Next, this problem is transformed to the canonical formulation, as shown in follows:

$$\max_{\mathbf{z}_{k,n}} \quad f(\mathbf{z}_{k,n}) \tag{20}$$

$$\text{s.t.} \quad \mathbf{z}_{k,n} \in \mathcal{G}, \tag{20a}$$

where $\mathbf{z}_{k,n} = \{\tau_{k,n}, p_{k,n}\}$, $\mathcal{G} = \{\mathbf{z}_{k,n} \in [\mathbf{0}, \mathbf{1}], g(\mathbf{z}_{k,n}) \leq 0\}$,

$$f(\mathbf{z}_{k,n}) = -\frac{\mu \beta_n}{\tau_{k,n} C_n} - \frac{D(\boldsymbol{w}_n^{(\mathrm{t})})}{B \log_2(1 + p_{k,n}|h_{k,n}|^2)}, \tag{21}$$

and

$$g(\mathbf{z}_{k,n}) = \kappa_0 \mu \beta_n (\tau_{k,n} C_n)^2 + \frac{p_{k,n} P_t D(\boldsymbol{w}_n^{(\mathrm{t})})}{B \log_2(1 + p_{k,n}|h_{k,n}|^2)} - E_n^{\max}. \tag{22}$$

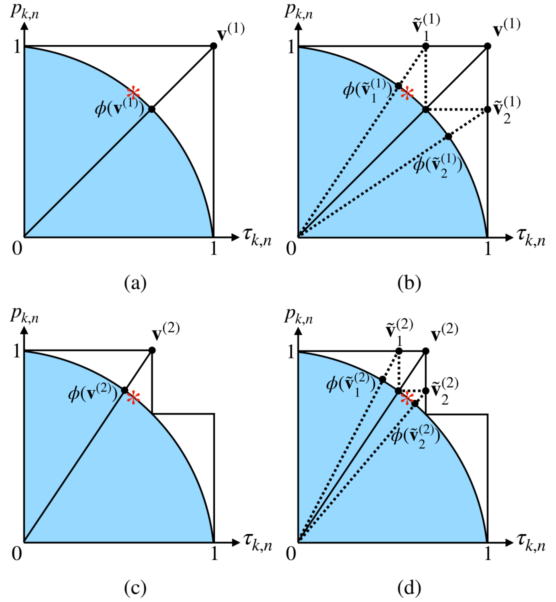
In problem (20), the objective function $f(\mathbf{z}_{k,n})$ is monotonically increasing with $\mathbf{z}_{k,n}$, and then the optimal solution is located on the boundary of the feasible set $\mathcal{G}$. However, due to the non-convexity, the expression of the boundary cannot be directly derived. In this case, a polyblock outer approximation algorithm is proposed to approach the feasible set $\mathcal{G}$ by constructing polyblock $\mathcal{P}$ [35], as shown in **Algorithm 1**.

As shown in Fig. 2(a), by setting the first vertex $\mathbf{v}^{(1)} \in \mathcal{V}^{(1)}$, the initial polyblock $\mathcal{P}^{(1)}$ can be constructed as the box $[\mathbf{0}, \mathbf{1}]$,

[3] Note that transmit power P_t does not affect the feasibility as the normalized channel condition $|h_{k,n}|^2$ also contains the transmit power.

Algorithm 1 Polyblock Outer Approximation Algorithm

```
1:  Initialize vertex set $\mathcal{V}^{(1)} = \{\mathbf{v}^{(1)}\}$, where $\mathbf{v}^{(1)} = \{1, 1\}$.
2:  Initialize polyblock $\mathcal{P}^{(1)}$ with vertex set $\mathcal{V}^{(1)}$.
3:  Set $\epsilon$ and $\theta = 1$.
4:  if $|f(\boldsymbol{\phi}(\mathbf{v}^{(\theta)})) - f(\boldsymbol{\phi}(\mathbf{v}^{(\theta-1)}))| > \epsilon$ then
5:      Obtain $\boldsymbol{\phi}(\mathbf{v}^{(\theta)})$ from Eq. (29).
6:      Calculate vertices $\tilde{\mathbf{v}}_1^{(\theta)}$ and $\tilde{\mathbf{v}}_2^{(\theta)}$ as follows:
            $\tilde{\mathbf{v}}_i^{(\theta)} = \mathbf{v}^{(\theta)} - (v_i^{(\theta)} - \phi_i(\mathbf{v}^{(\theta)}))\mathbf{e}_i, \forall i \in \{1, 2\}.$
7:      Update vertex set $\mathcal{V}^{(\theta+1)}$ as follows:
            $\mathcal{V}^{(\theta+1)} = \{\mathcal{V}^{(\theta)} \backslash \mathbf{v}^{(\theta)}\} \cup \{\tilde{\mathbf{v}}_1^{(\theta)}, \tilde{\mathbf{v}}_2^{(\theta)}\}.$
8:      Construct polyblock $\mathcal{P}^{(\theta+1)}$ with vertex set $\mathcal{V}^{(\theta+1)}$.
9:      Find vertex $\mathbf{v}^{(\theta+1)}$ from $\mathcal{V}^{(\theta+1)}$, where
            $\mathbf{v}^{(\theta+1)} = \text{argmax}\{f(\boldsymbol{\phi}(\mathbf{v})) | \mathbf{v} \in \mathcal{V}^{(\theta+1)}\}.$
10:     Set $\theta = \theta + 1$.
11: end if
12: Set $\mathbf{z}_{k,n}^* = \boldsymbol{\phi}(\mathbf{v}^{(\theta)})$.
```


$p_{k,n}$
1
$\mathbf{v}^{(1)}$
$\phi(\mathbf{v}^{(1)})$
0
1
$\tau_{k,n}$
$p_{k,n}$
1
$\tilde{\mathbf{v}}_1^{(1)}$
$\mathbf{v}^{(1)}$
$\phi(\tilde{\mathbf{v}}_1^{(1)})$
$\tilde{\mathbf{v}}_2^{(1)}$
$\phi(\tilde{\mathbf{v}}_2^{(1)})$
0
1
$\tau_{k,n}$
$p_{k,n}$
1
$\mathbf{v}^{(2)}$
$\phi(\mathbf{v}^{(2)})$
0
1
$\tau_{k,n}$
$p_{k,n}$
1
$\tilde{\mathbf{v}}_1^{(2)}$
$\mathbf{v}^{(2)}$
$\phi(\tilde{\mathbf{v}}_1^{(2)})$
$\tilde{\mathbf{v}}_2^{(2)}$
$\phi(\tilde{\mathbf{v}}_2^{(2)})$
0
1
$\tau_{k,n}$

(a) (b)

(c) (d)

Fig. 2: An illustration of Algorithm 1. The blue area is the feasible set $\mathcal{G}$, and the red star is the optimal point.

which contains the feasible set[4]. The projection of $\mathbf{v}^{(1)}$ on the upper boundary of feasible set $\mathcal{G}$ is calculated, denoted by $\boldsymbol{\phi}(\mathbf{v}^{(1)})$. Based on point $\boldsymbol{\phi}(\mathbf{v}^{(1)})$, two new vertices $\tilde{\mathbf{v}}_1^{(1)}$ and $\tilde{\mathbf{v}}_2^{(1)}$ can be obtained to replace $\mathbf{v}^{(1)}$, as shown in Fig. 2(b). The new vertices are calculated as follows:

$$\tilde{\mathbf{v}}_i^{(1)} = \mathbf{v}^{(1)} - (v_i^{(1)} - \phi_i(\mathbf{v}^{(1)}))\mathbf{e}_i, \forall i \in \{1, 2\}, \tag{23}$$

where $v_i^{(1)}$ is the i-th element of $\mathbf{v}^{(1)}$, $\phi_i(\mathbf{v}^{(1)})$ is the i-th element of $\boldsymbol{\phi}(\mathbf{v}^{(1)})$, and $\mathbf{e}_i$ is the i-th unit vector. As shown in Fig. 2(c), a new polyblock $\mathcal{P}^{(2)}$ is obtained, where $\mathcal{G} \subset \mathcal{P}^{(2)} \subset \mathcal{P}^{(1)}$. The vertex set of $\mathcal{P}^{(2)}$ is updated as follows:

$$\mathcal{V}^{(2)} = \{\mathcal{V}^{(1)} \backslash \mathbf{v}^{(1)}\} \cup \{\tilde{\mathbf{v}}_1^{(1)}, \tilde{\mathbf{v}}_2^{(1)}\}. \tag{24}$$

After that, the optimal vertex is selected from vertex set $\mathcal{V}^{(2)}$ and denoted by $\mathbf{v}^{(2)}$, which satisfies the following condition:

$$\mathbf{v}^{(2)} = \text{argmax}\{f(\boldsymbol{\phi}(\mathbf{v})) | \mathbf{v} \in \mathcal{V}^{(2)}\}. \tag{25}$$

The projection of the optimal vertex can achieve the maximum value of (21). For example, Fig. 2(c) shows the case that $\tilde{\mathbf{v}}_1^{(1)}$ is selected as the optimal vertex $\mathbf{v}^{(2)}$. This process is repeated, and a smaller polyblock $\mathcal{P}^{(3)} \subset \mathcal{P}^{(2)}$ is constructed based on vertex $\mathbf{v}^{(2)}$, as shown in Fig. 2(d). This algorithm is completed if the following condition is satisfied:

$$|f(\boldsymbol{\phi}(\mathbf{v}^{(\theta)})) - f(\boldsymbol{\phi}(\mathbf{v}^{(\theta-1)}))| \leq \epsilon, \tag{26}$$

where ϵ is the error tolerance. The output $\mathbf{z}_{k,n}^* = \boldsymbol{\phi}(\mathbf{v}^{(\theta)})$ is the optimal solution of problem (20). Moreover, the projection of any vertex $\mathbf{v}^{(\theta)}$ satisfies the following condition:

$$\boldsymbol{\phi}(\mathbf{v}^{(\theta)}) = \zeta \mathbf{v}^{(\theta)}, \tag{27}$$

where $\zeta \in (0, 1)$ is a ratio coefficient obtained as follows:

$$\begin{aligned} \zeta &= \max\{\tilde{\zeta} \in (0,1) | \tilde{\zeta}\mathbf{v}^{(\theta)} \in \mathcal{G}\} \\ &= \max\{\tilde{\zeta} \in (0,1) | g(\tilde{\zeta}\mathbf{v}^{(\theta)}) \leq 0, \tilde{\zeta}\mathbf{v}^{(\theta)} \in [\mathbf{0}, \mathbf{1}]\}. \end{aligned} \tag{28}$$

Since $g(\tilde{\zeta}\mathbf{v}^{(\theta)})$ is a monotonic increasing function, ζ satisfies $g(\zeta \mathbf{v}^{(\theta)}) = 0$, which can be written as follows:

$$\kappa_0 \mu \beta_n (\zeta v_1^{(\theta)} C_n)^2 + \frac{\zeta v_2^{(\theta)} P_t D(\boldsymbol{w}_n^{(\mathrm{t})})}{B \log_2(1 + \zeta v_2^{(\theta)} |h_{k,n}|^2)} = E_n^{\max}. \tag{29}$$

The above nonlinear equation can be solved by multiple numerical solvers, such as fsolve in MATLAB or Python. It is worth pointing out that during **Algorithm 1**, with any given vertex $\mathbf{v}^{(\theta)} = \{v_1^{(\theta)}, v_2^{(\theta)}\}$, where $\theta \geq 2$, condition $\zeta \in (0, 1)$ is always satisfied. When $\theta = 1$, $\zeta = 1$ may be obtained with initial vertex $\mathbf{v}^{(1)}$, which means that the feasible set $\mathcal{G}$ includes the first vertex. In other words, the energy consumption constraint $g(\mathbf{z}_{k,n}) < 0$ can be satisfied with any $\tau_{k,n}$ and $p_{k,n}$ in $[0, 1]$. In this case, the optimal solution is obtained by $\mathbf{z}_{k,n}^* = \mathbf{v}^{(1)}$.

According to [36], the complexity of **Algorithm 1** is dominated by step 9 and the number of iterations. Specifically, step 9 selects an optimal vertex from the vertex candidate set, which contains $\theta + 1$ vertices at the θ-th iteration. Since the complexity in solving (20) to obtain each vertex is $\mathcal{O}(2)$, the complexity of step 9 is $\mathcal{O}(2(\theta + 1))$. With the given number of iterations T_c, the total complexity of **Algorithm 1** can be expressed as $\mathcal{O}(2T_c(T_c + 1))$.

B. Matching based Sub-Channel Assignment

Based on **Algorithm 1**, the minimum time consumption of all selected devices in all sub-channels is obtained in matrix $\boldsymbol{\Gamma}$. In this subsection, a matching based algorithm is proposed to solve the binary integer programming problem in (18), where matrix $\boldsymbol{\Gamma}$ is utilized to construct the preference list. In problem (18), the set of selected devices, i.e., $\mathcal{N}_t$, is provided by the leader-level problem. As revealed in Section V, to minimize global loss, the number of selected devices is maximized in each round, i.e., $|\mathcal{N}_t| = K$. Therefore, $\mathcal{N}_t$ and $\mathcal{K}$ are two disjoint sets with the same size, and this scenario

[4] Note that the initial vertex, i.e., $\tau_{k,n} = 1$ and $p_{k,n} = 1$, may be infeasible for problem (20). However, the optimal solution obtained from Algorithm 1 is the projection of the vertex, which is always included in the feasible set $\mathcal{G}$.

can be considered as a one-to-one matching Ψ from $\mathcal{N}_t$ to $\mathcal{K}$. Furthermore, since some combinations in matrix $\mathbf{\Gamma}$ are marked as infeasible, the player in this matching may have an incomplete preference list [37], [38]. In this case, the elements in $\mathbf{\Gamma}$ cannot be directly employed, and the utility of device n assigned to sub-channel k in matching Ψ is defined below

$$U_n(\Psi) = \begin{cases} U_{\max}, & \text{if } \Gamma_{k,n} \text{ is infeasible}, \\ \Gamma_{k,n}, & \text{otherwise}, \end{cases} \tag{30}$$

where $U_{\max}$ is a large constant indicating that the assignment of device n and sub-channel k is infeasible. In the one-to-one matching, the utility of any sub-channel is equal to the utility of the occupied device, i.e., $U_k(\Psi) = U_{\Psi(k)}(\Psi), \forall k \in \mathcal{K}$. By calculating the utility, the preference list of any player can be established. Since that problem (18) is a minimization problem, the preference of device n can be defined below

$$(k, \Psi) \prec_n (k', \Psi') \Rightarrow U_n(\Psi) > U_n(\Psi'). \tag{31}$$

The above function indicates that device n is willing to be assigned to sub-channel k' in matching Ψ', rather than k in matching Ψ, since its utility can be strictly decreased by switching from k to k'.

Since all players are matched, the considered case follows the concept of two-sided exchange matchings [39]. In this case, if any device intends to be assigned to a sub-channel, it needs to exchange with the device occupying this sub-channel, instead of directly joining this sub-channel. A notation $\Psi_{n'}^{n}$ is introduced to represent the case where two devices n and n' are swapped in matching Ψ, which is defined as follows:

$$\Psi_{n'}^{n}(m) = \begin{cases} \Psi(n') = k', & m = n, \\ \Psi(n) = k, & m = n', \end{cases} \tag{32}$$

where $\Psi_{n'}^{n}$ only replaces two pairs defined by $\Psi(n) = k$ and $\Psi(n') = k'$ into $\Psi(n) = k'$ and $\Psi(n') = k$, respectively. It is indicated that any swap operation involves four players, and hence, it should be approved by the involved players. However, in a matching with the incomplete preference list, there may be sub-channels that are unacceptable for some devices. In order to prioritize feasible combinations, the approval of sub-channels is removed, and the swap operation is approved by the swap-blocking pair (n, n'), which is defined below [40]:

Definition 2. *A swap-blocking pair (n, n') is confirmed if and only if the following conditions hold*

1) *$U_n(\Psi_{n'}^{n}) \leq U_n(\Psi)$ and $U_{n'}(\Psi_{n'}^{n}) \leq U_{n'}(\Psi)$;*
2) *At least one inequality above is strict.*

By searching swap-blocking pairs, a matching based sub-channel assignment algorithm is presented in **Algorithm 2**. The proposed algorithm can be started from any initial matching. During the execution of the algorithm, an active device n is selected and sequential swap is attempted with all other devices. If a swap-blocking pair (n, n') is formed, the new matching $\Psi_{n'}^{n}$ is recorded, and the algorithm continues. The main loop of **Algorithm 2** executes repeatedly. When the last device has searched for all other devices, the first device becomes the active device again. The algorithm ends if no swap blocking pair can be found in a full round of the main loop. At this stage, a stable matching is established, which satisfies the following definition [39], [40]:

Algorithm 2 Sub-Channel Assignment Algorithm

```
Initialization:
Initialize initial matching $\Psi$ by randomly pairing all devices and sub-channels.
Main Loop:
for $n \in \mathcal{N}_t$ do
    Device $n$ makes a proposal to exchange with device $n' \in \mathcal{N}_t$, where $n \neq n'$.
    if $(n, n')$ is a swap-blocking pair then
        Matching $\Psi_{n'}^{n}$ is approved.
        Devices $n$ and $n'$ exchange sub-channels.
        Set $\Psi = \Psi_{n'}^{n}$
    end if
end for
The main loop is repeated until no swap-blocking pair can be found in a complete round.
```

Definition 3. *A matching Ψ is two-sided exchange-stable (2ES) if and only if there is no further swap-blocking pair.*

Note that some devices may not be assigned to feasible sub-channels in the stable matching, i.e., their utilities are equal to $U_{\max}$. In this case, these devices cannot transmit local models to the server, and the corresponding sub-channel assignment indicators should be set to zero in the leader-level problem. In the following, the complexity, convergence and stability of the proposed matching based sub-channel assignment algorithm are analyzed.

1) Complexity: By considering the worst case, the computational complexity of the proposed matching based algorithm can be expressed as $\mathcal{O}(CK^2)$, where C is the number of main loops. In the worst case, K devices can play the role of active devices to search $K-1$ devices, and thus $K(K-1)$ times of calculations should be performed in one loop. With the given number of main loops C, the computational complexity of the proposed algorithm is obtained.

2) Convergence: Assuming $\Psi^{(\mathrm{a})}$ and $\Psi^{(\mathrm{b})}$ are two adjacent matching in **Algorithm 2**, i.e., $\Psi^{(\mathrm{a})} \rightarrow \Psi^{(\mathrm{b})}$, where $a \neq b$, one swap-blocking pair is found among this transformation. According to Definition 2, at least one device can achieve less utility, and the utilities of other devices cannot be increased. Hence, the matching transformation cannot be reversed. With the finite number of devices and sub-channels, the number of possible matchings is finite and equal to the Bell number [41]. Therefore, from any initial matching, the proposed algorithm is guaranteed to converge to a stable matching.

3) Stability: Based on Definition 3, any final matching obtained from the matching based sub-channel assignment algorithm is 2ES. Specifically, if the final matching obtained from Algorithm 2 is not 2ES, there exists at least one swap-blocking pair, which can further reduce the sum utility of all devices. However, it contradicts the conditions for completing the proposed algorithm.

V. Solution of Leader-Level Problem

To solve the formulated global loss minimization problem in (13), the loss function and local data must be available at the server, which is impractical and contradicts the motivation to utilize FL. In this section, by analyzing the impact of device selection on the expected convergence rate, such information can be detached from the leader-level problem. By employing the gradient descent method, with global model $\boldsymbol{w}^{(\mathrm{t})}$, device n can update the local model as follows:

$$\boldsymbol{w}_n^{(\mathrm{t})} = \boldsymbol{w}^{(\mathrm{t})} - \frac{\lambda}{\beta_n}\sum_{i=1}^{\beta_n} \nabla \ell(\boldsymbol{w}^{(\mathrm{t})}; \boldsymbol{x}_{n,i}, y_{n,i}), \tag{33}$$

where λ is the learning rate. After that, the selected devices transmit the updated local models to the server for aggregation. By including device selection and sub-channel assignment, the aggregated global model in round t is given by[5]

$$\begin{aligned}\boldsymbol{w}^{(\mathrm{t+1})} =& \frac{\sum_{n=1}^{N} S_n^{(\mathrm{t})} \sum_{k=1}^{K} \psi_{k,n}^{(\mathrm{t})} \beta_n \boldsymbol{w}_n^{(\mathrm{t})}}{\sum_{n=1}^{N} S_n^{(\mathrm{t})} \sum_{k=1}^{K} \psi_{k,n}^{(\mathrm{t})} \beta_n} \\ =& \boldsymbol{w}^{(\mathrm{t})} - \frac{\lambda \sum_{n=1}^{N} S_n^{(\mathrm{t})} \sum_{k=1}^{K} \psi_{k,n}^{(\mathrm{t})} \sum_{i=1}^{\beta_n} \nabla \ell(\boldsymbol{w}^{(\mathrm{t})}; \boldsymbol{x}_{n,i}, y_{n,i})}{\sum_{n=1}^{N} S_n^{(\mathrm{t})} \sum_{k=1}^{K} \psi_{k,n}^{(\mathrm{t})} \beta_n} \\ =& \boldsymbol{w}^{(\mathrm{t})} - \lambda[\nabla F(\boldsymbol{w}^{(\mathrm{t})}) - \hat{\boldsymbol{w}}^{(\mathrm{t})}], \end{aligned} \tag{34}$$

where

$$\begin{aligned}&\hat{\boldsymbol{w}}^{(\mathrm{t})} \\ \triangleq& \nabla F(\boldsymbol{w}^{(\mathrm{t})}) - \frac{\sum_{n=1}^{N} S_n^{(\mathrm{t})} \sum_{k=1}^{K} \psi_{k,n}^{(\mathrm{t})} \sum_{i=1}^{\beta_n} \nabla \ell(\boldsymbol{w}^{(\mathrm{t})}; \boldsymbol{x}_{n,i}, y_{n,i})}{\sum_{n=1}^{N} S_n^{(\mathrm{t})} \sum_{k=1}^{K} \psi_{k,n}^{(\mathrm{t})} \beta_n}. \end{aligned} \tag{35}$$

In order to derive the expected convergence rate, the following assumptions are considered [8], [9], [23], [46]:

1) With respect to the global model $\boldsymbol{w}^{(\mathrm{t})}$, $\nabla F(\boldsymbol{w}^{(\mathrm{t})})$ is uniformly Lipschitz continuous, i.e.,

$$\|\nabla F(\boldsymbol{w}^{(\mathrm{t+1})}) - \nabla F(\boldsymbol{w}^{(\mathrm{t})})\| \le L\|\boldsymbol{w}^{(\mathrm{t+1})} - \boldsymbol{w}^{(\mathrm{t})}\|, \tag{36}$$

where L is a positive parameter.

2) $F(\boldsymbol{w}^{(\mathrm{t})})$ is strongly convex with a positive constant μ, i.e.,

$$\begin{aligned} F(\boldsymbol{w}^{(\mathrm{t+1})}) \ge& F(\boldsymbol{w}^{(\mathrm{t})}) + (\boldsymbol{w}^{(\mathrm{t+1})} - \boldsymbol{w}^{(\mathrm{t})})^{\top} \nabla F(\boldsymbol{w}^{(\mathrm{t})}) \\ &+ \frac{\mu}{2}\|\boldsymbol{w}^{(\mathrm{t+1})} - \boldsymbol{w}^{(\mathrm{t})}\|. \end{aligned} \tag{37}$$

3) $F(\boldsymbol{w}^{(\mathrm{t})})$ is twice-continuously differentiable, as below:

$$\mu \boldsymbol{I} \preceq \nabla^2 F(\boldsymbol{w}^{(\mathrm{t})}) \preceq L\boldsymbol{I}. \tag{38}$$

4) The following inequality is satisfied with any device n and sample i:

$$\|\nabla \ell(\boldsymbol{w}^{(\mathrm{t})}); \boldsymbol{x}_{n,i}, y_{n,i})\|^2 \le \rho \|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2, \tag{39}$$

where ρ is a non-negative constant.

It is worth pointing out that the above assumptions are commonly considered in the FL related optimization, and can be satisfied by widely adopted loss functions. Based on these assumptions, the following proposition yields the effect of device selection on convergence rate.

Algorithm 3 Device Selection Algorithm

```
Initialization:
Generate list $\mathcal{Q}^{(\mathrm{t})}$ based on (43).
Initialize set $\mathcal{N}_t$ by selecting the first $K$ devices from $\mathcal{Q}^{(\mathrm{t})}$.
Main Loop:
Obtain sub-channel assignment from Algorithm 2.
if $\sum_{n\in\mathcal{N}_t}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})} < K$ and $(N) \notin \mathcal{N}_t$ then
    for $n \in \mathcal{N}_t$ do
        if $\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})} = 0$ then
            Remove device $n$ from set $\mathcal{N}_t$.
            Add the next unselected device from list $\mathcal{Q}^{(\mathrm{t})}$.
        end if
    end for
end if
```

Proposition 3. *In the case that device selection and sub-channel assignment satisfy $S_n^{(\mathrm{t})} = \sum_{k=1}^{K} \psi_{k,n}^{(\mathrm{t})}, \forall n \in \mathcal{N}$, with the learning rate $\lambda = 1/L$ and the optimal global model $\mathbf{w}^*$, the upper bound of convergence rate in round t is*

$$\begin{aligned} \mathbb{E}[F(\boldsymbol{w}^{(\mathrm{t+1})}) - F(\boldsymbol{w}^*)] \le& \left(1 - \frac{\mu}{L}\right)^t \mathbb{E}[F(\boldsymbol{w}^{(1)}) - F(\boldsymbol{w}^*)] \\ &+ \frac{2\rho}{L}\sum_{i=1}^{t}\left(1-\frac{\mu}{L}\right)^{t-i} \frac{\|\nabla F(\boldsymbol{w}^{(\mathrm{i})})\|^2}{\sum_{n=1}^{N}\beta_n}\sum_{n=1}^{N}\beta_n\left(1 - S_n^{(\mathrm{i})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{i})}\right). \end{aligned} \tag{40}$$

Proof: Refer to Appendix C. ■

It is indicated that the convergence rate is bounded by two terms. The first term, i.e., $\left(1-\frac{\mu}{L}\right)^t \mathbb{E}[F(\boldsymbol{w}^{(1)}) - F(\boldsymbol{w}^*)]$, is the expected gap between the first global loss and the optimal global loss in round t. The second term is related to device selection, where the status of devices in round i, i.e., $S_n^{(\mathrm{i})}$, is included. Proposition 3 indicates the expected gap between the global loss in round t and the optimal loss. In this case, the global loss minimization problem in (13) can be achieved by minimizing the expected gap. Since that the first term does not include device selection, it can be treated as a constant. Meanwhile, terms $\frac{2\rho}{L}$ and $\frac{\|\nabla F(\boldsymbol{w}^{(\mathrm{i})})\|^2}{\sum_{n=1}^{N}\beta_n}$ are positive and not effected by device selection, and hence, they can be removed. By incorporating the AoU based weight $\alpha_n^{(\mathrm{t})}$ into the device selection indicator $S_n^{(\mathrm{t})}$, the objective function of the leader-level problem (13) can be reformulated as follows:

$$\min_{\mathbf{S}^{(\mathrm{t})}} \sum_{n=1}^{N}\beta_n\left(1 - \alpha_n^{(\mathrm{t})} S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\right) \tag{41}$$

By removing the constant term $\sum_{n=1}^{N}\beta_n$, the leader-level problem (13) can be equivalently transformed as follows:

$$\begin{aligned} \max_{\mathbf{S}^{(\mathrm{t})}} \quad & \sum_{n=1}^{N}\alpha_n^{(\mathrm{t})}\beta_n S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})} \\ \text{s.t.} \quad & \text{(13a), (13b).} \end{aligned} \tag{42}$$

Problem (42) can be treated as a weighted device selection problem, where AoU and data size play the role of weight

[5] Although we consider federated averaging (FedAvg) for aggregation in this paper, an extension of the results to other federated optimization algorithms [42]–[45] is straightforward.

factors. That is, the server tends to select the devices with large AoU and/or data size. In this case, the server can order and select devices based on this weight. In any round t, all devices are sorted in a list $\mathcal{Q}^{(\mathrm{t})}$, which satisfies the following condition:

$$\alpha_{(1)}^{(\mathrm{t})}\beta_{(1)} \geq \alpha_{(2)}^{(\mathrm{t})}\beta_{(2)} \geq \cdots \geq \alpha_{(N)}^{(\mathrm{t})}\beta_{(N)}, \tag{43}$$

where (1) and (N) denote the devices with the highest and lowest priority, respectively. Moreover, it is indicated by problem (42) that the server tends to select more devices, and hence, constraint (13b) can be rewritten as $\sum_{n=1}^{N} S_n^{(\mathrm{t})} = K$. According to the list $\mathcal{Q}^{(\mathrm{t})}$, problem (42) can be solved by **Algorithm 3**. In the proposed device selection algorithm, K devices with the highest priority are selected in the initialization phase. During **Algorithm 3**, any device not assigned to a sub-channel will be replaced, until all selected devices have been assigned to sub-channels, or all devices in list $\mathcal{Q}^{(\mathrm{t})}$ have been adopted. In the worst case, all devices in list $\mathcal{Q}^{(\mathrm{t})}$ are traversed, and therefore the complexity of **Algorithm 3** can be expressed as $\mathcal{O}(N)$.

VI. Simulation Results

In this section, the performance of the proposed solution is simulated and demonstrated. In this simulation, multiple devices are randomly distributed in a disc with radius R, and the server is located in the center of the disc. The experiments includes three datasets (MNIST, CIFAR-10, and SST-2) with different models[6]. The imbalanced independent identically distributed (IID) data distribution is adopted, where a factor $c_n \in [1, 10]$ is randomly generated for all devices, and training samples (500 for MNIST, 50000 for CIFAR-10, and 67349 for SST-2) are shuffled and partitioned across devices based on fraction $c_n / \sum_{i \in \mathcal{N}} c_i$. Unlike [8]–[15] which focus on minimizing either global loss or latency, this paper studies the interaction between global loss minimization and latency minimization. Thereby, a direct comparison with the existing schemes in these works is unfair. Alternatively, the proposed scheme is benchmarked against the following conventional device selection schemes:

- *AoU based DS*: The server selects the top K devices in (43).
- *Random DS*: The server selects K devices randomly.
- *Cluster based DS*: All devices are randomly allocated to N/K clusters such that the clusters are selected in rotation.
- *Fixed DS*: The same K devices are selected in all rounds.

In the above schemes, the proposed solutions are adopted, including monotonic optimization based resource allocation (MO-RA) and matching based sub-channel assignment (M-SA). Moreover, fixed resource allocation (FIX-RA) and random sub-channel assignment (R-SA) are incorporated as benchmarks, where $\tau_{k,n} = p_{k,n} = 0.5, \forall k, n$ is set in FIX-RA. The parameters of the simulation are shown in Table I.

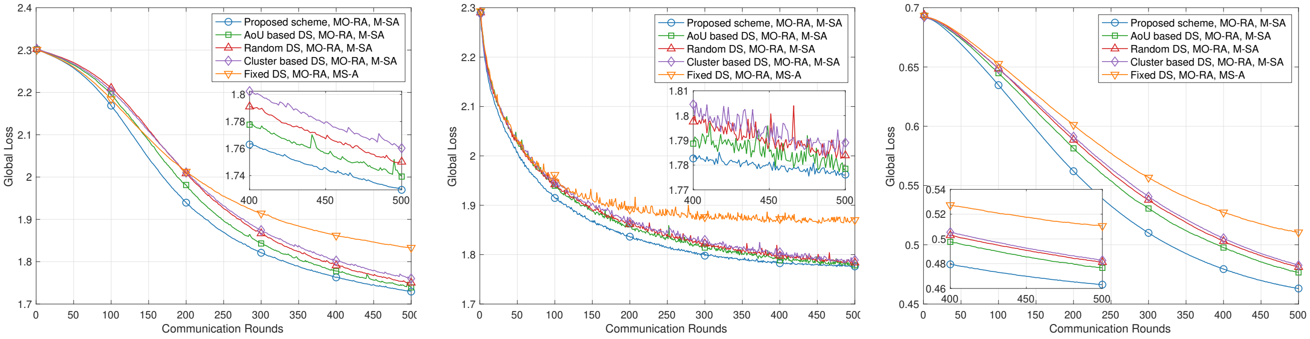
TABLE I: Table of Parameters

Carrier frequency f	1 GHz
AWGN noise power σ^2	-174 dBm
Path loss exponent a	3.76
Bandwidth for each sub-channel B	1 MHz
Power consumption coefficient κ_0	10^{-28}
CPU cycles for each bit of tasks μ	10^7
Computational capacity C_n	1 GHz
Error tolerance ϵ	0.01
Model size $D(\boldsymbol{w})$ (MNIST/CIFAR-10/SST-2)	$1/5/5$ Mbit
Maximum energy $E_n^{\max}$ (MNIST/CIFAR-10/SST-2)	0.02/0.1/0.1 joule
Learning rate λ (MNIST/CIFAR-10/SST-2)	0.01/0.001/0.01
Batch size (MNIST/CIFAR-10/SST-2)	32/512/128
Optimizer (MNIST/CIFAR-10/SST-2)	SGD/Adam/SGD

In Fig. 3, the global loss achieved by different schemes is presented. Given the same number of communication rounds, the proposed scheme can achieve the lowest global loss on all datasets, reflecting the derived convergence rate in Proposition 3. This improvement benefits from two approaches, including AoU based weights and efficient utilization of sub-channels. The former is able to select devices that can provide more contributions in each communication round, while the latter ensures the maximization of the number of selected devices. Moreover, it can be observed that without Algorithm 3, AoU based DS can still outperform other schemes, which confirms that AoU based weights have the capability to improve learning efficiency. In terms of the random DS and cluster based DS, the devices have the same probability to be selected, and hence, the similar global loss is achieved by these two schemes. For the fixed DS, due to the fact that the size of training data in this case is less than that in other schemes, its performance is the worst. Compared to the MNIST dataset, the CIFAR-10 image classification task is more complex, and hence, the differences between the schemes are not that obvious, as shown in Fig. 3(b). However, the performance of these schemes is still consistent with Fig. 3(a). It is worth noting that by employing the CIFAR-10 dataset, the data size and model size is increased, and therefore, the number of devices that can satisfy the energy constraint is reduced. This issue is severe when utilizing AoU based DS, as the server tends to select devices that are generally more difficult to meet this condition, resulting in poor performance. This effect is mitigated when a simpler task is adopted, as shown in Fig. 3(c). Since there are only two labels on the SST-2 dataset, the differences between schemes become significant, and the advantages of the proposed scheme in global loss are clearly demonstrated.

Fig. 4 shows the performance of the proposed optimization solutions, where different resource allocation and sub-channel assignment approaches are incorporated into the proposed device selection scheme. It is clear that the proposed scheme can achieve the best performance with the proposed solutions, i.e., MO-RA and M-SA. This is because in this case, the server can select devices who can provide more contributions in the aggregation without replacing them with lower priority devices in (43). When other solutions are utilized, some

[6] For MNIST digit recognition tasks, a multi-layer perceptron neural network is built, where two ReLu hidden layers with 128 and 256 neurons follows by a softmax output layer. For CIFAR-10 image classification tasks, a convolutional neural network (CNN) is constructed with two 3x3 convolution layers (one with 32 filters and one with 64 filters, each followed by a 2x2 max pooling layer), a 128-neuron ReLu hidden layer, and a softmax output layer. For SST-2 text classification tasks, a tokenizer with a vocabulary size of $4,000$ is included, and the neural network is built with a 128-neuron ReLu hidden layer and a sigmoid output layer.

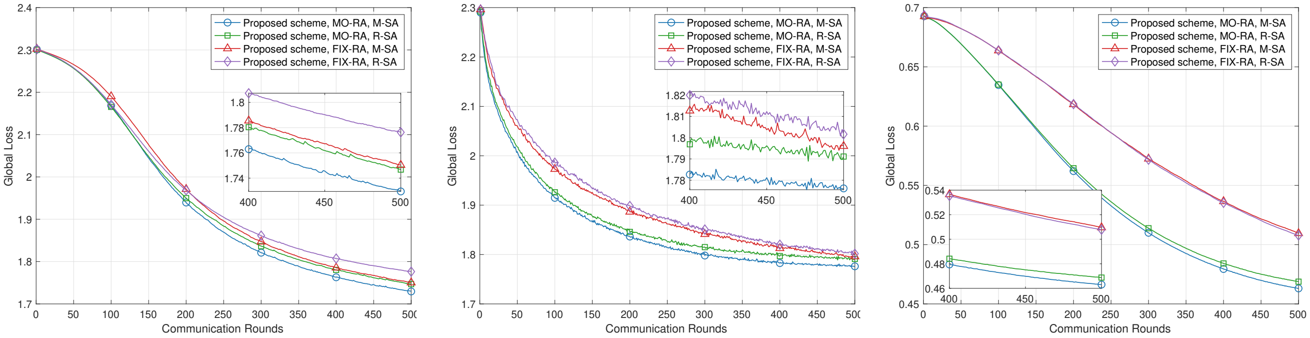


(a) MNIST

(b) CIFAR-10

(c) SST-2

Fig. 3: The performance of the proposed scheme. $N = 20$, $K = 4$, $P_t = 10$ dBm, and $R = 500$ m.



(a) MNIST

(b) CIFAR-10

(c) SST-2

Fig. 4: The performance of the proposed scheme using different solutions. $N = 20$, $K = 4$, $P_t = 10$ dBm, and $R = 500$ m.

devices may no longer meet the maximum energy consumption constraint and thus be replaced by devices with lower priority in (43), resulting in poor learning performance. This result also corroborates our conclusion that the contribution of a device in aggregation is proportional to its AoU and data size. In Fig. 4(a), an obvious feature is that the proposed scheme using FIX-RA and R-SA performs well in the early stage of training (from round 1 to round 100), but poorly in the later stage. This is due to the fact that in this case, the number of devices that can satisfy the maximum energy consumption requirement is very small, and thus device selection is performed on a smaller subset. For simple tasks, such as MNIST with IID data distribution, repeatedly using smaller training datasets can achieve faster convergence in the early stages, however, this can lead to overfitting problems and poor final results. Moreover, it is worth pointing out that compared to sub-channel assignment, resource allocation plays a more important role in minimizing the global loss, and this trend is more significant when the model size is larger, as shown in Fig. 4(b) and Fig. 4(c).

In Fig. 5 and Fig. 6, the impact of the number of devices and radius on global loss is presented, respectively. In Fig. 5, since the size of the total training data is fixed, an increase in the number of devices means less training data per device. That is, when the number of selected devices, i.e., K, is fixed, as the number of devices rises, the amount of training data utilized in each communication round reduces, resulting in an increase in the global loss. With the proposed optimization solutions, devices that can provide more contributions can be selected, and therefore the learning performance is improved. In Fig. 6, the increase in radius can be understood as a deterioration in channel conditions. According to Proposition 1, in this case, more devices become unavailable and cannot participate in the aggregation. As a result, the achievable global loss increases with the radius. By employing the proposed optimization solutions, the negative impact caused by channel degradation can be alleviated to a certain extent, thereby narrowing the gap in global loss.

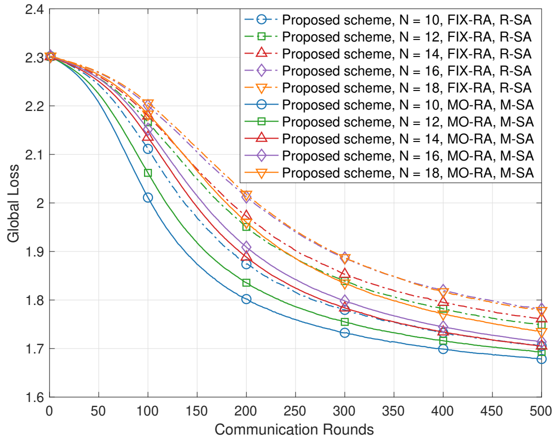


Fig. 5: The impact of the number of devices with the MNIST dataset. $K = 4$, $P_t = 10$ dBm, and $R = 500$ m.

In the considered Stackelberg game based framework, the number of selected devices varies between schemes, and

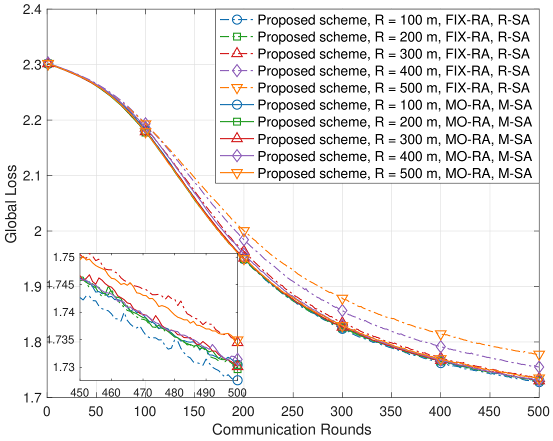


Fig. 6: The impact of the radius with the MNIST dataset. $N = 20$, $K = 4$, and $P_t = 10$ dBm.

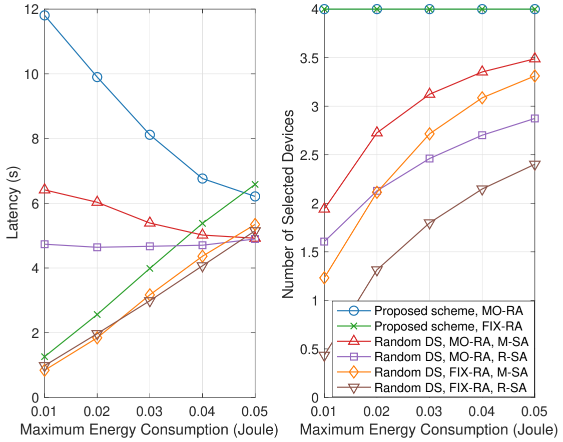


Fig. 8: The impact of the maximum energy consumption. $K = 4$, and $P_t = 10$ dBm.

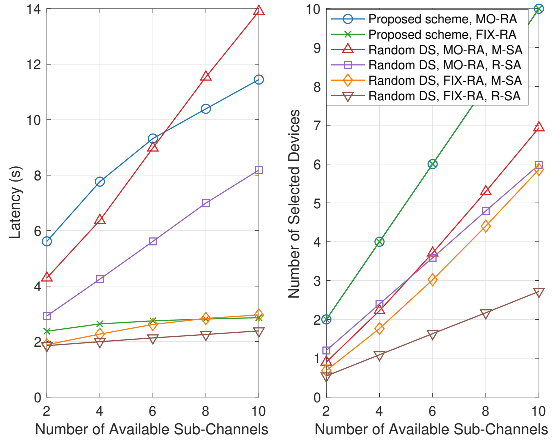


Fig. 7: The impact of the number of available sub-channels. $P_t = 10$ dBm, and $E_n^{\max} = 0.02$ joule.

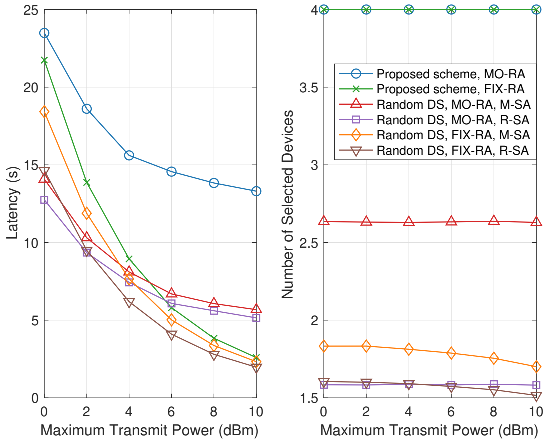


Fig. 9: The impact of the maximum transmit power. $K = 4$, and $E_n^{\max} = 0.02$ joule.

therefore, the latency cannot be compared independently. It can be seen from Fig. 7 that the proposed scheme can efficiently utilize all available sub-channels, while the latency of each communication round is also increased accordingly. That is because the proposed scheme needs to guarantee the performance of training through the leader level problem. This explains why the achievable global loss of the proposed scheme is lower than others. It is also indicated that the proposed RA and SA algorithms can increase the number of selected devices with random DS.

The effect of energy limitation is demonstrated in Fig. 8, and the results confirm Proposition 1. According to Proposition 1, as the maximum energy consumption increases, the device participation increases. As a result, by employing random DS, the number of selected devices in each communication round is increased, and meanwhile the latency of each round also increases. On the other hand, it can be observed that the proposed algorithm, MO-RA, has the ability to dynamically adjust the computational resource allocation coefficients and power allocation coefficients. Therefore, the latency can be reduced with this solution.

As shown in Fig. 9, the latency decreases with the increasing transmit power, since the achievable data rate can be increased. With FIX-RA, the number of selected devices is reduced when the transmit power is greater than 6 dBm, because the fixed power allocation coefficient can no longer satisfy the energy consumption constraint. For MO-RA, the power allocation coefficient is optimized, and thus the number of selected devices is not affected. However, the decrease in latency has been slowed down accordingly.

VII. Conclusions

This paper investigates FL in a practical wireless communication scenario, where limited sub-channels and energy consumption are considered. With the goal of minimizing convergence time, we jointly formulate global loss minimization and latency minimization based on the Stackelberg game. We present a monotonic optimization based algorithm to find the optimal solution of computational resource allocation and power allocation. Based on the obtained solution of resource allocation, we further develop a matching based algorithm with incomplete preference list to solve the sub-channel assignment problem. For the global loss minimization problem, we incorporate AoU based weights and derive an upper bound on convergence rate, which presents a priority for selecting devices. Simulation results demonstrate significant reduction of the global loss, establishing that our developed solutions can efficiently utilize available sub-channels and reduce latency. It is worth highlighting that the AoU based device selection can be extended to mobile environments as a practical and promising research direction.

Appendix A: Proof of Proposition 1

Suppose that device n is selected and assigned to sub-channel k in round t, i.e., $S_n^{(\mathrm{t})} = 1$ and $\psi_{k,n}^{(\mathrm{t})} = 1$, based on constraint (14a), the following condition will be satisfied by the optimal solutions $\tau_{k,n}^*$ and $p_{k,n}^*$:

$$E_{k,n}^{\mathrm{cp}}(\tau_{k,n}^*) + E_{k,n}^{\mathrm{cm}}(p_{k,n}^*) \leq E_n^{\max}. \tag{44}$$

Due to the fact that the energy consumption for computation is strictly greater than zero, i.e., $E_{k,n}^{\mathrm{cp}}(\tau_{k,n}^*) > 0$, the following inequality can be obtained:

$$E_{k,n}^{\mathrm{cm}}(p_{k,n}^*) < E_n^{\max}, \tag{45}$$

which can be rewritten as follows:

$$\frac{p_{k,n}^* P_t D(\boldsymbol{w}_n^{(\mathrm{t})})}{B\log_2(1+p_{k,n}^*|h_{k,n}|^2)} < E_n^{\max}. \tag{46}$$

Since $p_{k,n}^*|h_{k,n}|^2 > 0$, $p_{k,n}^*|h_{k,n}|^2 > \ln(1 + p_{k,n}^*|h_{k,n}|^2)$ holds, the following inequality can be obtained:

$$\frac{\ln(2)P_t D(\boldsymbol{w}_n^{(\mathrm{t})})}{B|h_{k,n}|^2} < \frac{p_{k,n}^* P_t D(\boldsymbol{w}_n^{(\mathrm{t})})}{B\log_2(1+p_{k,n}^*|h_{k,n}|^2)}. \tag{47}$$

Therefore, the following condition must be satisfied:

$$\frac{\ln(2)P_t D(\boldsymbol{w}_n^{(\mathrm{t})})}{B|h_{k,n}|^2} < E_n^{\max}, \tag{48}$$

and this proposition is proved. ■

Appendix B: Proof of Proposition 2

It is obvious that the time consumption of any device n in sub-channel k is monotonically decreasing with $\tau_{k,n}$ and $p_{k,n}$, and hence, the proof of this part is omitted. In terms of the energy consumption, as shown in constraint (19a), the monotonicity can be proved by two parts. The first term, $E_{k,n}^{\mathrm{cp}}(\tau_{k,n})$, is the energy consumption for computation, which is monotonically increasing with the computational resource allocation coefficient. The second term is the energy consumption for communication, which can be presented by

$$E_{k,n}^{\mathrm{cm}}(p_{k,n}) = \frac{p_{k,n} P_t D(\boldsymbol{w}_n^{(\mathrm{t})})}{B\log_2(1+p_{k,n}|h_{k,n}|^2)}. \tag{49}$$

The derivative of the above function is

$$\begin{aligned}\frac{\partial E_{k,n}^{\mathrm{cm}}(p_{k,n})}{\partial p_{k,n}} = & \frac{\ln(2)P_t D(\boldsymbol{w}_n^{(\mathrm{t})})}{B(1+p_{k,n}|h_{k,n}|^2)\ln^2(1+p_{k,n}|h_{k,n}|^2)} \\ & \times[(1+p_{k,n}|h_{k,n}|^2)\ln(1+p_{k,n}|h_{k,n}|^2)-p_{k,n}|h_{k,n}|^2].\end{aligned} \tag{50}$$

Since the first term of the above function is always greater than zero, the monotonicity of function $E_{k,n}^{\mathrm{cm}}(p_{k,n})$ depends on the term in the square bracket. Particularly, $E_{k,n}^{\mathrm{cm}}(p_{k,n})$ is an increasing function if the following inequality holds:

$$(1+p_{k,n}|h_{k,n}|^2)\ln(1+p_{k,n}|h_{k,n}|^2)-p_{k,n}|h_{k,n}|^2 \geq 0. \tag{51}$$

Suppose $\theta_{k,n} \triangleq \frac{1}{p_{k,n}|h_{k,n}|^2+1}$, the above inequality is equivalent to the following inequality:

$$\frac{1}{\theta_{k,n}}(\theta_{k,n} - 1 - \ln\theta_{k,n}) \geq 0. \tag{52}$$

Since $\theta_{k,n} > 0$, $\ln\theta_{k,n} \leq \theta_{k,n} - 1$ holds, and then the above inequality is always satisfied, which indicates that (49) is an increasing function. As a result, the total energy consumption is an increasing function, and this proposition is proved. ■

Appendix C: Proof of Proposition 3

In order to prove this proposition, an auxiliary function is defined as follows:

$$G(\boldsymbol{w}) = \frac{L}{2}\boldsymbol{w}^\top\boldsymbol{w} - F(\boldsymbol{w}). \tag{53}$$

With respect to $\boldsymbol{w}$, the second-order partial derivative of the above function is given by

$$\frac{\partial^2 G(\boldsymbol{w})}{\partial\boldsymbol{w}^2} = L - \frac{\partial^2 F(\boldsymbol{w})}{\boldsymbol{w}^2}. \tag{54}$$

Since $F(\boldsymbol{w})$ satisfies the uniformly Lipschitz condition, the above function is always greater than or equal to zero, and hence, $G(\boldsymbol{w})$ is a convex function. By utilizing the second-order Taylor series expansion, the following inequality can be obtained:

$$G(\boldsymbol{w}^{(\mathrm{t}+1)}) \geq G(\boldsymbol{w}^{(\mathrm{t})}) + (\boldsymbol{w}^{(\mathrm{t}+1)} - \boldsymbol{w}^{(\mathrm{t})})^\top \nabla G(\boldsymbol{w}^{(\mathrm{t})}). \tag{55}$$

From (53), the above inequality can be equivalently transformed as follows:

$$\begin{aligned}F(\boldsymbol{w}^{(\mathrm{t}+1)}) \leq & F(\boldsymbol{w}^{(\mathrm{t})}) + (\boldsymbol{w}^{(\mathrm{t}+1)} - \boldsymbol{w}^{(\mathrm{t})})^\top \nabla F(\boldsymbol{w}^{(\mathrm{t})}) \\ & + \frac{L}{2}\|\boldsymbol{w}^{(\mathrm{t}+1)} - \boldsymbol{w}^{(\mathrm{t})}\|^2.\end{aligned} \tag{56}$$

Based on (34), the following inequality can be obtained:

$$\begin{aligned}F(\boldsymbol{w}^{(\mathrm{t}+1)}) \leq & F(\boldsymbol{w}^{(\mathrm{t})}) - \lambda[\nabla F(\boldsymbol{w}^{(\mathrm{t})}) - \hat{\boldsymbol{w}}^{(\mathrm{t})}]^\top \nabla F(\boldsymbol{w}^{(\mathrm{t})}) \\ & + \frac{\lambda^2 L}{2}\|\nabla F(\boldsymbol{w}^{(\mathrm{t})}) - \hat{\boldsymbol{w}}^{(\mathrm{t})}\|^2.\end{aligned} \tag{57}$$

When $\lambda = 1/L$, the above inequality can be transformed as follows:

$$\begin{aligned}& \mathbb{E}[F(\boldsymbol{w}^{(\mathrm{t}+1)})] \\ \leq & \mathbb{E}\Big\{F(\boldsymbol{w}^{(\mathrm{t})}) - \lambda[\nabla F(\boldsymbol{w}^{(\mathrm{t})}) - \hat{\boldsymbol{w}}^{(\mathrm{t})}]^\top \nabla F(\boldsymbol{w}^{(\mathrm{t})}) \\ & + \frac{\lambda^2 L}{2}\|\nabla F(\boldsymbol{w}^{(\mathrm{t})}) - \hat{\boldsymbol{w}}^{(\mathrm{t})}\|^2\Big\} \\ = & \mathbb{E}\Big[F(\boldsymbol{w}^{(\mathrm{t})}) - \frac{1}{L}\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2 + \frac{1}{L}(\hat{\boldsymbol{w}}^{(\mathrm{t})})^\top \nabla F(\boldsymbol{w}^{(\mathrm{t})}) \\ & + \frac{1}{2L}\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2 - \frac{1}{L}(\hat{\boldsymbol{w}}^{(\mathrm{t})})^\top \nabla F(\boldsymbol{w}^{(\mathrm{t})}) + \frac{1}{2L}\|\hat{\boldsymbol{w}}^{(\mathrm{t})}\|^2\Big] \\ = & \mathbb{E}[F(\boldsymbol{w}^{(\mathrm{t})})] - \frac{1}{2L}\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2 + \frac{1}{2L}\mathbb{E}(\|\hat{\boldsymbol{w}}^{(\mathrm{t})}\|^2).\end{aligned} \tag{58}$$

By defining

$$\begin{cases} f(\mathcal{N}_t,\boldsymbol{\psi}^{(\mathrm{t})},\boldsymbol{w}^{(\mathrm{t})}) \triangleq \sum\limits_{n\in\mathcal{N}_t}\sum\limits_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\sum\limits_{i=1}^{\beta_n}\nabla\ell(\boldsymbol{w}^{(\mathrm{t})};\boldsymbol{x}_{n,i},y_{n,i}), \\ g(\mathcal{N}_t,\boldsymbol{\psi}^{(\mathrm{t})},\boldsymbol{w}^{(\mathrm{t})}) \triangleq \sum\limits_{n\in\mathcal{N}_t}\sum\limits_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\sum\limits_{i=1}^{\beta_n}\|\nabla\ell(\boldsymbol{w}^{(\mathrm{t})};\boldsymbol{x}_{n,i},y_{n,i})\|, \end{cases} \tag{59}$$

based on condition $S_n^{(\mathrm{t})} = \sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})} = 1, \forall n \in \mathcal{N}_t$, the following equation can be derived:

$$\begin{aligned}\mathbb{E}(\|\hat{\boldsymbol{w}}^{(\mathrm{t})}\|^2) &= \mathbb{E}\left[\left\|\nabla F(\boldsymbol{w}^{(\mathrm{t})}) - \frac{f(\mathcal{N}_t,\boldsymbol{\psi}^{(\mathrm{t})},\boldsymbol{w}^{(\mathrm{t})})}{\sum_{n=1}^{N}S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\beta_n}\right\|^2\right] \\ &= \mathbb{E}\left[\left\|\frac{f(\mathcal{N}_t,\boldsymbol{\psi}^{(\mathrm{t})},\boldsymbol{w}^{(\mathrm{t})})}{\sum_{n=1}^{N}\beta_n} - \frac{f(\mathcal{N}_t,\boldsymbol{\psi}^{(\mathrm{t})},\boldsymbol{w}^{(\mathrm{t})})}{\sum_{n=1}^{N}S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\beta_n}\right.\right. \\ &\quad \left.\left. + \frac{\sum_{n\in\mathcal{N}\backslash\mathcal{N}_t}\sum_{i=1}^{\beta_n}\nabla\ell(\boldsymbol{w}^{(\mathrm{t})};\boldsymbol{x}_{n,i},y_{n,i})}{\sum_{n=1}^{N}\beta_n}\right\|^2\right],\end{aligned} \tag{60}$$

where $\mathcal{N}\backslash\mathcal{N}_t$ is the collection of unselected devices in round t. According to the triangle-inequality, the above equation can be transformed as follows:

$$\begin{aligned}\mathbb{E}(\|\hat{\boldsymbol{w}}^{(\mathrm{t})}\|^2) \leq \mathbb{E}\Bigg\{&\frac{\left[\sum_{n=1}^{N}\beta_n\left(1-S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\right)\right]g(\mathcal{N}_t,\boldsymbol{\psi}^{(\mathrm{t})},\boldsymbol{w}^{(\mathrm{t})})}{(\sum_{n=1}^{N}\beta_n)(\sum_{n=1}^{N}S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\beta_n)} \\ &+ \frac{\sum_{n\in\mathcal{N}\backslash\mathcal{N}_t}\sum_{i=1}^{\beta_n}\|\nabla\ell(\boldsymbol{w}^{(\mathrm{t})};\boldsymbol{x}_{n,i},y_{n,i})\|}{\sum_{n=1}^{N}\beta_n}\Bigg\}^2.\end{aligned} \tag{61}$$

According to (39), inequalities

$$g(\mathcal{N}_t,\boldsymbol{\psi}^{(\mathrm{t})},\boldsymbol{w}^{(\mathrm{t})}) \leq \sum_{n=1}^{N}S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\beta_n\sqrt{\rho\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2}, \tag{62}$$

and

$$\begin{aligned}&\sum_{n\in\mathcal{N}\backslash\mathcal{N}_t}\sum_{i=1}^{\beta_n}\|\nabla\ell(\boldsymbol{w}^{(\mathrm{t})};\boldsymbol{x}_{n,i},y_{n,i})\| \\ &\leq \left[\sum_{n=1}^{N}\beta_n\left(1-S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\right)\right]\sqrt{\rho\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2},\end{aligned} \tag{63}$$

can be obtained. Therefore, (61) can be rewritten as follows:

$$\begin{aligned}\mathbb{E}(\|\hat{\boldsymbol{w}}^{(\mathrm{t})}\|^2) &\leq \mathbb{E}\Bigg\{\frac{\left[\sum_{n=1}^{N}\beta_n\left(1-S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\right)\right]\sqrt{\rho\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2}}{\sum_{n=1}^{N}\beta_n} \\ &\quad + \frac{\left[\sum_{n=1}^{N}\beta_n\left(1-S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\right)\right]\sqrt{\rho\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2}}{\sum_{n=1}^{N}\beta_n}\Bigg\}^2 \\ &= \frac{4\rho\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2}{\left(\sum_{n=1}^{N}\beta_n\right)^2}\mathbb{E}\left[\sum_{n=1}^{N}\beta_n\left(1-S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\right)\right]^2.\end{aligned} \tag{64}$$

Due to the fact that the number of all devices' samples is not less than the number of unselected devices' samples, i.e.,

$$\sum_{n=1}^{N}\beta_n \geq \sum_{n=1}^{N}\beta_n\left(1-S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\right) \geq 0, \tag{65}$$

the following inequality can be obtained from (64):

$$\mathbb{E}(\|\hat{\boldsymbol{w}}^{(\mathrm{t})}\|^2) \leq \frac{4\rho\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2}{\sum_{n=1}^{N}\beta_n}\mathbb{E}\left[\sum_{n=1}^{N}\beta_n\left(1-S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\right)\right]. \tag{66}$$

As a result, (58) can be rewritten as follows:

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{w}^{(\mathrm{t+1})})] &\leq \mathbb{E}[F(\boldsymbol{w}^{(\mathrm{t})})] - \frac{1}{2L}\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2 \\ &\quad + \frac{2\rho\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2}{L\sum_{n=1}^{N}\beta_n}\sum_{n=1}^{N}\beta_n\left(1-S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\right).\end{aligned} \tag{67}$$

By subtracting $\mathbb{E}[F(\boldsymbol{w}^*)]$ in both sides of the above function, the following inequality can be obtained

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{w}^{(\mathrm{t+1})}) - F(\boldsymbol{w}^*)] &\leq \mathbb{E}[F(\boldsymbol{w}^{(\mathrm{t})}) - F(\boldsymbol{w}^*)] - \frac{1}{2L}\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2 \\ &\quad + \frac{2\rho\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2}{L\sum_{n=1}^{N}\beta_n}\sum_{n=1}^{N}\beta_n\left(1-S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\right).\end{aligned} \tag{68}$$

According to [47], the following inequality can be obtained from (37) and (38):

$$\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2 \geq 2\mu[F(\boldsymbol{w}^{(\mathrm{t})}) - F(\boldsymbol{w}^*)]. \tag{69}$$

Hence, (68) can be rewritten as follows:

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{w}^{(\mathrm{t+1})}) - F(\boldsymbol{w}^*)] &\leq \left(1 - \frac{\mu}{L}\right)\mathbb{E}[F(\boldsymbol{w}^{(\mathrm{t})}) - F(\boldsymbol{w}^*)] \\ &\quad + \frac{2\rho\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2}{L\sum_{n=1}^{N}\beta_n}\sum_{n=1}^{N}\beta_n\left(1-S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\right).\end{aligned} \tag{70}$$

Similarly, the convergence of successive rounds is given by

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{w}^{(\mathrm{t+1})}) - F(\boldsymbol{w}^*)] &\leq \left(1-\frac{\mu}{L}\right)^2\mathbb{E}[F(\boldsymbol{w}^{(\mathrm{t-1})}) - F(\boldsymbol{w}^*)] \\ &\quad + \left(1-\frac{\mu}{L}\right)\frac{2\rho\|\nabla F(\boldsymbol{w}^{(\mathrm{t-1})})\|^2}{L\sum_{n=1}^{N}\beta_n}\sum_{n=1}^{N}\beta_n\left(1-S_n^{(\mathrm{t-1})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t-1})}\right) \\ &\quad + \frac{2\rho\|\nabla F(\boldsymbol{w}^{(\mathrm{t})})\|^2}{L\sum_{n=1}^{N}\beta_n}\sum_{n=1}^{N}\beta_n\left(1-S_n^{(\mathrm{t})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{t})}\right).\end{aligned} \tag{71}$$

As a result, we derive the upper bound of the convergence rate as

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{w}^{(\mathrm{t+1})}) - F(\boldsymbol{w}^*)] &\leq \left(1-\frac{\mu}{L}\right)^t\mathbb{E}[F(\boldsymbol{w}^{(1)}) - F(\boldsymbol{w}^*)] \\ &\quad + \frac{2\rho}{L}\sum_{i=1}^{t}\left(1-\frac{\mu}{L}\right)^{t-i}\frac{\|\nabla F(\boldsymbol{w}^{(\mathrm{i})})\|^2}{\sum_{n=1}^{N}\beta_n}\sum_{n=1}^{N}\beta_n\left(1-S_n^{(\mathrm{i})}\sum_{k=1}^{K}\psi_{k,n}^{(\mathrm{i})}\right),\end{aligned} \tag{72}$$

which completes the proof. ■

References

[1] S. Savazzi, M. Nicoli, M. Bennis, S. Kianoush, and L. Barbieri, "Opportunities of federated learning in connected, cooperative, and automated industrial systems," *IEEE Commun. Mag.*, vol. 59, no. 2, pp. 16–21, 2021.

[2] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3579–3605, 2021.

[3] Z. Qin, G. Y. Li, and H. Ye, "Federated learning and wireless communications," *IEEE Wireless Commun.*, vol. 28, no. 5, pp. 134–140, 2021.

[4] W. Xia, W. Wen, K.-K. Wong, T. Q. Quek, J. Zhang, and H. Zhu, "Federated-learning-based client scheduling for low-latency wireless communications," *IEEE Wireless Commun.*, vol. 28, no. 2, pp. 32–38, 2021.

[5] M. Chen, H. V. Poor, W. Saad, and S. Cui, "Wireless communications for collaborative federated learning," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 48–54, 2020.

[6] L. Qiao, Z. Gao, M. B. Mashhadi, and D. Gündüz, "Massive digital over-the-air computation for communication-efficient federated edge learning," *arXiv preprint arXiv:2405.15969*, 2024.

[7] 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.

[8] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, 2019.

[9] 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 Trans. Wireless Commun.*, vol. 20, no. 1, 2021.

[10] R. Hamdi, M. Chen, A. B. Said, M. Qaraqe, and H. V. Poor, "Federated learning over energy harvesting wireless networks," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 92–103, 2022.

[11] S. Liu, G. Yu, R. Yin, J. Yuan, L. Shen, and C. Liu, "Joint model pruning and device selection for communication-efficient federated edge learning," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 231–244, 2022.

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

[13] T. T. Vu, D. T. Ngo, N. H. Tran, H. Q. Ngo, M. N. Dao, and R. H. Middleton, "Cell-free massive MIMO for wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6377–6392, 2020.

[14] D. Chen, C. S. Hong, L. Wang, Y. Zha, Y. Zhang, X. Liu, and Z. Han, "Matching-theory-based low-latency scheme for multitask federated learning in mec networks," *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11 415–11 426, 2021.

[15] Z. Ji, L. Chen, N. Zhao, Y. Chen, G. Wei, and F. R. Yu, "Computation offloading for edge-assisted federated learning," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9330–9344, 2021.

[16] H. T. Nguyen, H. V. Poor, and M. Chiang, "Contextual model aggregation for fast and robust federated learning in edge computing," *arXiv preprint arXiv:2203.12738*, 2022.

[17] H. Imani, J. Anderson, and T. El-Ghazawi, "isample: Intelligent client sampling in federated learning," in *2022 IEEE 6th International Conference on Fog and Edge Computing (ICFEC)*, 2022, pp. 58–65.

[18] 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, 2021.

[19] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1183–1210, 2021.

[20] H. H. Yang, A. Arafa, T. Q. S. Quek, and H. Vincent Poor, "Age-based scheduling policy for federated learning in mobile edge networks," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8743–8747.

[21] K. Wang, Z. Ding, D. K. C. So, and Z. Ding, "Age-of-information minimization in federated learning based networks with Non-IID dataset," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2024.

[22] K. Wang, F. Fang, D. B. d. Costa, and Z. Ding, "Sub-channel scheduling, task assignment, and power allocation for OMA-based and NOMA-based MEC systems," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2692–2708, 2021.

[23] M. B. Mashhadi, M. Mahdavimoghadam, R. Tafazolli, and W. Saad, "Collaborative learning with a drone orchestrator," *IEEE Trans. Veh. Technol.*, vol. 73, no. 1, pp. 637–650, 2024.

[24] S. Zhang, J. Li, L. Shi, M. Ding, D. C. Nguyen, W. Tan, J. Weng, and Z. Han, "Federated learning in intelligent transportation systems: Recent applications and open problems," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 5, pp. 3259–3285, 2024.

[25] R. Zhang, H. Wang, B. Li, X. Cheng, and L. Yang, "A survey on federated learning in intelligent transportation systems," *arXiv preprint arXiv:2403.07444*, 2024.

[26] W. Dai, Y. Zhou, N. Dong, H. Zhang, and E. Xing, "Toward understanding the impact of staleness in distributed machine learning," in *International Conference on Learning Representations*, 2019.

[27] R. Talak, S. Karaman, and E. Modiano, "Improving age of information in wireless networks with perfect channel state information," *IEEE/ACM Transactions on Networking*, vol. 28, no. 4, pp. 1765–1778, 2020.

[28] J. Ren, Y. He, D. Wen, G. Yu, K. Huang, and D. Guo, "Scheduling for cellular federated edge learning with importance and channel awareness," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7690–7703, 2020.

[29] D. Chen, C. S. Hong, L. Wang, Y. Zha, Y. Zhang, X. Liu, and Z. Han, "Matching-theory-based low-latency scheme for multitask federated learning in MEC networks," *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11 415–11 426, 2021.

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

[31] K. Wang, Z. Ding, D. K. So, and Z. Ding, "Exploring age-of-information weighting in federated learning under data heterogeneity," *arXiv preprint arXiv:2405.15978*, 2024.

[32] Z. Han, *Game theory in wireless and communication networks: theory, models, and applications*. Cambridge University Press, 2012.

[33] T. Basar and G. J. Olsder, *Dynamic noncooperative game theory*. Siam, 1999, vol. 23.

[34] H. Tuy, "Monotonic optimization: Problems and solution approaches," *SIAM Journal on Optimization*, vol. 11, no. 2, pp. 464–494, 2000.

[35] Y. J. A. Zhang, L. Qian, J. Huang *et al.*, "Monotonic optimization in communication and networking systems," *Foundations and Trends® in Networking*, vol. 7, no. 1, pp. 1–75, 2013.

[36] M. Hua, L. Yang, Q. Wu, and A. L. Swindlehurst, "3D UAV trajectory and communication design for simultaneous uplink and downlink transmission," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5908–5923, 2020.

[37] D. Gusfield and R. W. Irving, *The stable marriage problem: structure and algorithms*. MIT press, 1989.

[38] K. Iwama, D. Manlove, S. Miyazaki, and Y. Morita, "Stable marriage with incomplete lists and ties," in *Automata, Languages and Programming: 26th International Colloquium, ICALP'99 Prague, Czech Republic, July 11–15, 1999 Proceedings*. Springer, 2002, pp. 443–452.

[39] A. Roth and M. Sotomayor, "Two-sided matching," *Handbook of game theory with economic applications*, vol. 1, pp. 485–541, 1992.

[40] E. Bodine-Baron, C. Lee, A. Chong, B. Hassibi, and A. Wierman, "Peer effects and stability in matching markets," in *International Symposium on Algorithmic Game Theory*. Springer, 2011, pp. 117–129.

[41] D. Ray, *A game-theoretic perspective on coalition formation*. Oxford University Press, 2007.

[42] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in neural information processing systems*, vol. 33, pp. 7611–7623, 2020.

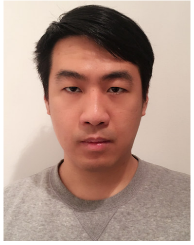
[43] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[44] J. Hamer, M. Mohri, and A. T. Suresh, "Fedboost: A communication-efficient algorithm for federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 3973–3983.

[45] D. Wen, K.-J. Jeon, and K. Huang, "Federated dropout—a simple approach for enabling federated learning on resource constrained devices," *IEEE Wireless Commun. Lett.*, vol. 11, no. 5, pp. 923–927, 2022.

[46] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Distributed federated learning for ultra-reliable low-latency vehicular communications," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 1146–1159, 2020.

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



Kaidi Wang (Member, IEEE) received the MS degree in communications and signal processing from Newcastle University in 2014, and the PhD degree in wireless communication from the University of Manchester in 2020. He is a research associate in the Department of Electrical and Electronic Engineering, the University of Manchester. From 2021 to 2023, he has been a research fellow of Wireless Communications at the Institute for Communication Systems, home of 5GIC and 6GIC at the University of Surrey. His current research interests include non-orthogonal multiple access, near-field communications, mobile edge computing, and federated learning.

Yi Ma (Senior Member, IEEE) is a Chair Professor within the Institute for Communication Systems (ICS), University of Surrey, Guildford, U.K. He has authored or co-authored 200+ peer-reviewed IEEE journals and conference papers in the areas of deep learning, cooperative communications, cognitive radios, interference utilization, cooperative localization, radio resource allocation, multiple-input multiple-output, estimation, synchronization, and modulation and detection techniques. He holds 10 international patents in the areas of spectrum sensing and signal modulation and detection. He has served as the Tutorial Chair for EuroWireless2013, PIMRC2014, and CAMAD2015. He was the Founder of the Crowd-Net Workshop in conjunction with ICC'15, ICC'16, and ICC'17. He is the Co-Chair of the Signal Processing for Communications Symposium in ICC'19.

Mahdi Boloursaz Mashhadi (Senior Member, IEEE) is a Lecturer at the 5G/6G Innovation Centre (5G/6GIC) at the Institute for Communication Systems (ICS), University of Surrey (UoS), UK. Prior to joining ICS, he was a postdoctoral research associate at the Intelligent Systems and Networks (ISN) Research Group, Imperial College London, 2019-2021. He received B.S., M.S., and Ph.D. degrees in mobile telecommunications from the Sharif University of Technology (SUT), Tehran, Iran, in 2011, 2013, and 2018, respectively. He was a visiting research associate with the University of Central Florida, Orlando, USA, in 2018, and Queen's University, Ontario, Canada, in 2017. He has more than 40 peer reviewed publications and patents in the areas of wireless communications, machine learning, and signal processing. He received the Best Paper Award from the IEEE EWDTS 2012 conference, and the Exemplary Reviewer Award from the IEEE ComSoc in 2021 and 2022. He has served as a panel judge for the International Telecommunication Union (ITU) to evaluate innovative submissions on applications of AI/ML in 5G and beyond wireless networks since 2021. He is an associate editor for the Springer Nature Wireless Personal Communications Journal.

Chuan Heng Foh (Senior Member, IEEE) received the M.Sc. degree from Monash University, Melbourne, VIC, Australia, in 1999, and the Ph.D. degree from the University of Melbourne, Melbourne, in 2002. After the Ph.D. degree, he spent six months as a Lecturer with Monash University. In December 2002, he joined Nanyang Technological University, Singapore, as an Assistant Professor until 2012. He is currently a Senior Lecturer with the University of Surrey, Guildford, U.K. He has authored or coauthored more than 180 refereed papers in international journals and conferences. His research interests include protocol design, machine learning application, and performance analysis of various computer networks, including wireless local area networks, mobile ad-hoc and sensor networks, vehicular networks, the Internet of Things, 5G/6G networks, and open RAN. He served as the Vice Chair (Europe/Africa) for IEEE TCGCC in 2015 and 2017. He is currently the Vice-Chair of the IEEE VTS Ad Hoc Committee on Mission Critical Communications.

Rahim Tafazolli (Senior Member, IEEE) is Regius Professor of Electronic Engineering, Professor of Mobile and Satellite Communications, Founder and Director of 5GIC, 6GIC and ICS (Institute for Communication System) at the University of Surrey. He has over 40 years of experience in digital communications research and teaching. He has authored and co-authored 1,000+ research publications and is regularly invited to deliver keynote talks and distinguished lectures to international conferences and workshops. He was the leader of study on "grand challenges in IoT" (Internet of Things) in the UK, 2011-2012, for RCUK (Research Council UK) and the UK TSB (Technology Strategy Board). He is the Editor of two books on Technologies for Wireless Future (Wiley) vol. 1, in 2004 and vol. 2, in 2006. He holds Fellowship of Royal Academy of Engineering (FREng), Institute of Engineering and Technology (FIET) as well as that of Wireless World Research Forum. He was also awarded the 28th KIA Laureate Award- 2015 for his contribution to communications technology.

Zhi Ding (Fellow, IEEE) is with the Department of Electrical and Computer Engineering at the University of California, Davis, where he holds the position of distinguished professor. He received his Ph.D. degree in Electrical Engineering from Cornell University in 1990. From 1990 to 2000, he was a faculty member of Auburn University and later, University of Iowa. Prof. Ding joined the College of Engineering at UC Davis in 2000. His major research interests and expertise cover the areas of wireless networking, communications, signal processing, multimedia, and learning. Prof. Ding supervised over 30 PhD dissertations since joining UC Davis. His research team of enthusiastic researchers works very closely with industry to solve practical problems and contributes to technological advances. His team has collaborated with researchers around the world and welcomes self-motivated young talents as new members.

Prof. Ding is a Fellow of IEEE and has served as the Chief Information Officer and Chief Marketing Officer of the IEEE Communications Society. He was associate editor for IEEE Transactions on Signal Processing from 1994-1997, 2001-2004, and associate editor of IEEE Signal Processing Letters 2002-2005. He was a member of technical committee on Statistical Signal and Array Processing and member of technical committee on Signal Processing for Communications (1994-2003). Dr. Ding was the General Chair of the 2016 IEEE International Conference on Acoustics, Speech, and Signal Processing and the Technical Program Chair of the 2006 IEEE Globecom. He was also an *IEEE Distinguished Lecturer* (Circuits and Systems Society, 2004-06, Communications Society, 2008-09). He served on as IEEE Transactions on Wireless Communications Steering Committee Member (2007-2009) and its Chair (2009-2010). Dr. Ding is a coauthor of the textbook: *Modern Digital and Analog Communication Systems*, 5th edition, Oxford University Press, 2019. Prof. Ding received the IEEE Communication Society's WTC Award in 2012 and the IEEE Communication Society's Education Award in 2020.