Age-of-Information Minimization in Federated Learning based Networks with Non-IID Dataset

Kaidi Wang, *Member, IEEE,* Zhiguo Ding, *Fellow, IEEE,* Daniel K. C. So, *Senior Member, IEEE,* and Zhi Ding, *Fellow, IEEE*

***Abstract*—In this paper, a federated learning (FL) based system is investigated with non-independent and identically distributed (non-IID) dataset, where multiple devices participate in the global model aggregation through a limited number of sub-channels. By analyzing weight divergence and convergence rate, a new metric is proposed based on age-of-information (AoI), which incorporates latency and can provide an advanced device selection standard. After that, device selection, sub-channel assignment and resource allocation are jointly designed in an overall AoI minimization problem under the maximum energy consumption constraint. The formulated problem is decoupled into two sub-problems. After analyzing the feasibility, the resource allocation problem is transformed to a convex problem, and the closed-from solution is obtained based on KKT conditions. By introducing virtual sub-channels, device selection and sub-channel assignment are jointly solved by a matching based algorithm. Simulation results indicate that the proposed scheme is able to outperform all baselines in terms of both test accuracy and sum AoI, and the developed strategies can achieve significant improvements for all schemes.**

***Index Terms*—Age-of-information (AoI), device selection, federated learning (FL), resource allocation, sub-channel assignment**

I. Introduction

With the rapid development of mobile applications, massive amounts of data become available on edge devices [1]. In the conventional edge computing/learning schemes, data should be offloaded to a central server for processing, which consumes a lot of wireless communication resources [2]. Moreover, some data may contain privacy sensitive information and therefore cannot be collected in the practical deployment [3]. In this context, federated learning (FL), as a method of distributed learning, was proposed by Google and considered as a promising technique [4]. In FL, a neural network is shared between the server and all participating devices, where each device trains the model based on local data and transmits updates (e.g. weights or gradients) to the server [5]. Due to the fact that the raw data with larger size is not transmitted, FL can achieve higher privacy and communication efficiency compared to the conventional centralized learning (CL) [6]. However, since the performance of FL relies on periodic transmissions, it is necessary to design and optimize wireless communication networks accordingly [7].

Latency, as one of the important factors in determining the convergence time of FL [8], has been widely studied in existing works [9]–[16]. By defining latency as the time consumption of computation and communication, a latency minimization problem was formulated in [9], where the global optimal solution of resource allocation was obtained based on the bisection method. The authors of [10] focused on minimizing the time consumption of downlink and uplink transmissions. In order to guarantee the convergence, a binary convergence indicator was introduced to minimize the number of required communication rounds [10]. In [11], FL was considered in a cell-free massive multiple-input multiple-output (MIMO) scenario, in which multiple access points were introduced to play the role of relays. The formulated training time minimization problem was solved by a successive convex approximation based algorithm. In order to improve the convergence rate of FL, [12] formulated a global loss minimization problem under the maximum time consumption constraint. This problem was decoupled into two sub-problems, and the time consumption constraint was transformed to a latency minimization problem. In [13], FL was combined with edge computing for minimizing the latency, where devices can partially offload datasets to the server and train the global model based on the remaining data. In this case, the latency can be further reduced due to the decrease of local training time and the utilization of server idle time. The authors of [14] included other techniques to improve the efficiency of FL. Specifically, an intelligent reflecting surface (IRS) based FL system was designed, where two different transmission protocols, including frequency division multiple access (FDMA) and nonorthogonal multiple access (NOMA), were employed and compared. Some works combined latency minimization with other objectives in order to investigate the trade-off [15], [16]. Aiming to minimize the weighted sum of latency and global loss, pruning rate design and bandwidth allocation were jointly researched in [15]. With the help of the bisection method, an algorithm was developed to find the optimal solution. For studying the trade-off between latency and energy consumption, FL was considered in a vehicular network, where each device can offload part of the data to the

This work was supported in part by the UK EPSRC under grant number EP/W034522/1 and H2020 H2020-MSCA-RISE-2020 under grant number 101006411.

This material is also based upon work supported by the National Science Foundation under Grants 2009001 and 2029027.

Kaidi Wang and Daniel K. C. So are with the Department of Electrical and Electronic Engineering, The University of Manchester, Manchester, M13 9PL, UK (email: kaidi.wang@ieee.org, d.so@manchester.ac.uk).

Zhiguo Ding is with Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, UAE, and Department of Electrical and Electronic Engineering, University of Manchester, Manchester, UK (email: zhiguo.ding@manchester.ac.uk).

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

server for performing edge learning [16].

Due to the fact that FL generally involves a large number of devices, device selection is commonly considered to accommodate limited wireless resources or filter unimportant data [17]–[23]. In [17], it was illustrated that the performance of FL can be improved by selecting more devices in each communication round, and then, a device selection problem was formulated to maximize the number of selected devices in an over-the-air computation (AirComp) based FL system. By defining packet errors, the effect of wireless communication was included in the derived convergence rate [18]. In this work, device selection was achieved by solving a resource block allocation problem, where a Hungarian algorithm was employed to select devices based on the data size. A biased device selection scheme was proposed in [19], in which the server needs to estimate the local loss and select devices with high local loss. The results showed that the proposed device selection strategy can achieve significant improvements compared to random selection. The contribution based device selection was also studied in [20] and [21], where the contribution was defined as the increase of test accuracy and the decease of global loss, respectively. Specifically, on the basis of the conventional random device selection scheme, an additional phase was introduced to sort devices based on their contributions [20]. In [21], the transmitted local model updates were iteratively excluded to calculate the contribution, which was used to generate the probability of device selection in the next round. A reinforcement learning based algorithm was employed in [22] to select devices for participating the aggregation. It was indicated that the required communication rounds can be significantly reduced by the proposed solution in the case of non-independent and identically distributed (non-IID) dataset. The authors of [23] focused on the long-term optimization of FL, where device selection was investigated under a long-term energy consumption constraint. By pointing out that the later stage of learning is more sensitive to the number of selected devices, an algorithm was developed to achieve the long-term performance improvement.

Even though latency minimization and device selection have been extensively researched in the aforementioned works, their combination may pose new challenges. In particular, in order to reduce the latency, some devices with high channel quality may be consecutively selected in different communication rounds, which usually reduces the performance of learning and causes overfitting problems [24]–[26]. Moreover, with non-IID dataset, selecting any device consecutively leads to weight divergence towards a particular direction [27]. On the other hand, existing device selection strategies depend on either communication state or local data/model analysis. The former leads to a loss of learning performance, while the latter contradicts the motivation for utilizing FL, i.e., to preserve data privacy. Therefore, a new metric is required to balance latency and convergence performance while guiding device selection without analyzing local data/models. In this work, age-of-information (AoI) is introduced to explore the trade-off between latency and convergence. Different from the age-of-update (AoU) defined in [28] and [29], AoI in this work is a real number related to the time consumption of the previous communication rounds[1]. As a result, latency minimization and device selection can be jointly investigated in a formulated sum AoI minimization problem. Furthermore, in each communication round, all devices' AoI can be directly calculated and stored at the server, and hence, the system feedback overhead can be efficiently suppressed. The main contributions of this paper are listed below.

- A FL based network with non-IID dataset is studied, in which a subset of devices is selected to participate in the aggregation in each communication round. It is proved that in the non-IID case, any device selection strategy will produce an error, resulting in weight divergence and affecting the convergence rate. By revealing the error estimate with conventional random device selection, a novel AoI based device selection scheme is designed.
- By defining AoI as the idle time of each device, an overall AoI minimization problem is formulated under the constraints of maximum energy consumption and device selection. This problem is decoupled into two sub-problems, including an AoI minimization based sub-channel assignment problem and a latency minimization based resource allocation problem.
- The resource allocation problem is first proved to be infeasible under an extreme condition. Afterwards, this problem is transformed into a convex problem, and the closed-form solution is derived with the help of KKT conditions and the Lambert W function. For the problem of device selection and sub-channel assignment, a matching-based algorithm is developed, and the properties are analyzed.
- The considered FL network is simulated with MNIST and CIFAR-10 datasets. Simulation results indicate that the proposed scheme can achieve the best performance in terms of both test accuracy and sum AoI. The proposed resource allocation solution and sub-channel assignment algorithm are able to dynamically improve the performance under different parameter configurations.

II. System Model

Consider a FL based network with one server, N devices, and K sub-channels, where $N \geq K$. The non-IID dataset is employed, and all nodes are equipped with single-antennas. The collections of all devices and sub-channels are $\mathcal{N} = \{1, 2, \ldots, N\}$ and $\mathcal{K} = \{1, 2, \ldots, K\}$, respectively.

A. Training Model

In each communication round, the server broadcasts a global model to all devices, and then each device trains the received model based on local data and transmits local models to the server. Due to the limited number of available sub-channels, a subset of devices are selected in round t for aggregation,

[1]Note that in [29], AoU is considered as a weight factor for device selection and included in a global loss minimization problem. In this work, AoI minimization is investigated as an alternative to the conventional latency minimization problem to develop a fair device selection strategy. Furthermore, this paper provides a closed-form solution for latency minimization, which has advantages in terms of complexity and optimality, compared to the monotonic optimization based solution in [29].

denoted by $\mathcal{S}_t$, where $\mathcal{S}_t \subseteq \mathcal{N}$. The local loss and global loss can be respectively presented as follows:

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

and

$$\begin{aligned} F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t) &= \sum_{n\in\mathcal{S}_t}\frac{\beta_n}{\sum_{n\in\mathcal{S}_t}\beta_n}f_n(\mathbf{w}^{(\mathrm{t})}) \\ &= \frac{\sum_{n\in\mathcal{S}_t}\sum_{i=1}^{\beta_n}\ell(\mathbf{w}^{(\mathrm{t})};\mathbf{x}_{n,i},y_{n,i})}{\sum_{n\in\mathcal{S}_t}\beta_n}, \end{aligned} \tag{2}$$

where $\mathbf{w}^{(\mathrm{t})}$ is the global model in round t, β_n is the number of samples at device n, $\ell(\cdot)$ is a loss function, and $(\mathbf{x}_{n,i},y_{n,i})$ is the i-th sample of device n. The local model of device n in round t can be expressed as follows:

$$\mathbf{w}_n^{(\mathrm{t})} = \mathbf{w}^{(\mathrm{t})} - \frac{\lambda}{\beta_n}\sum_{i=1}^{\beta_n}\nabla\ell(\mathbf{w}^{(\mathrm{t})};\mathbf{x}_{n,i},y_{n,i}), \tag{3}$$

where λ is the learning rate. The server can update the global model based on federated averaging (FedAvg) [4] as follows:

$$\mathbf{w}^{(\mathrm{t}+1)} = \frac{\sum_{n\in\mathcal{S}_t}\beta_n\mathbf{w}_n^{(\mathrm{t})}}{\sum_{n\in\mathcal{S}_t}\beta_n} = \mathbf{w}^{(\mathrm{t})} - \lambda\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t). \tag{4}$$

B. Performance Analysis

In order to analyze the performance of the considered FL algorithm, the following assumptions are considered [30], [31]:

Assumption 1. *With respect to* $\mathbf{w}$*, the gradient* $\nabla f(\mathbf{w}^{(\mathrm{t})})$ *of* $f(\mathbf{w}^{(\mathrm{t})})$ *is uniformly Lipschitz continuous, which leads to*

$$\|\nabla F(\mathbf{w}^{(\mathrm{t}-1)},\mathcal{N}) - \nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\| \le L\|\mathbf{w}^{(\mathrm{t}-1)} - \mathbf{w}^{(\mathrm{t})}\|, \tag{5}$$

where L *is the Lipschitz constant, and* $\|\cdot\|$ *is the norm operator.*

Assumption 2. *Global loss function* $F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})$ *satisfies the Polyak-Lojasiewicz inequality with positive parameter* μ*, i.e.,*

$$\|\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\|^2 \ge 2\mu\left[F(\mathbf{w}^{(\mathrm{t})},\mathcal{N}) - F(\mathbf{w}^{*},\mathcal{N})\right]. \tag{6}$$

Since the non-IID data distribution is considered, the weight divergence is analyzed in this subsection, where CL is included as the benchmark [27]. The update of global models in CL is given by

$$\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t}+1)} = \mathbf{w}_{\mathrm{cen}}^{(\mathrm{t})} - \frac{\lambda}{\sum_{n\in\mathcal{N}}\beta_n}\sum_{n\in\mathcal{N}}\sum_{i=1}^{\beta_n}\nabla\ell(\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t})};\mathbf{x}_{n,i},y_{n,i}). \tag{7}$$

Theorem 1. *In any round* t*, the expected weight divergence between the considered FL and CL is bounded as follows:*

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{w}^{(\mathrm{t}+1)} - \mathbf{w}_{\mathrm{cen}}^{(\mathrm{t}+1)}\|\right] &\le (1+\lambda L)^t\mathbb{E}\left[\|\mathbf{w}^{(1)} - \mathbf{w}_{\mathrm{cen}}^{(1)}\|\right] \\ &+\lambda\sum_{i=1}^{t}(1+\lambda L)^{t-i}\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^{(i)},\mathcal{S}_i) - \nabla F(\mathbf{w}^{(i)},\mathcal{N})\right\|\right]. \end{aligned} \tag{8}$$

Proof: Refer to Appendix A. ∎

It is indicated by Theorem 1 that the divergence of global models is due to two terms. The first term $\|\mathbf{w}^{(1)} - \mathbf{w}_{\mathrm{cen}}^{(1)}\|$ is the weight divergence of the initial global model, which is amplified by $(1+\lambda L)^t$. The second term can be treated as the divergence of gradients in round i, which is caused by device selection, resulting in different data sizes and data distributions. Moreover, the effect caused by the second term is cumulative with the number of communication rounds, since it is amplified by $(1+\lambda L)^{t-i}$. That is, the impact on weight dievergence at the early stage of training is more significant. According to [32], by defining

$$\mathbf{e}^{(\mathrm{t})} \triangleq \nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t) - \nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N}), \tag{9}$$

its effect on convergence rate can be presented.

Theorem 2. *In the considered FL scenario, with an arbitrary set of devices* $\mathcal{S}_t \subseteq \mathcal{N}$*, the expected reduction of global loss in round* t *is bounded by*

$$\begin{aligned} \mathbb{E}\Big[F(\mathbf{w}^{(\mathrm{t}+1)},\mathcal{N}) - F(\mathbf{w}^{*})\Big] &\le \Big(1-\frac{\mu}{L}\Big)^t\mathbb{E}\Big[F(\mathbf{w}^{(1)},\mathcal{N}) - F(\mathbf{w}^{*})\Big] \\ &+\frac{1}{2L}\sum_{i=1}^{t}\Big(1-\frac{\mu}{L}\Big)^{t-i}\mathbb{E}\Big[\|\mathbf{e}^{(i)}\|^2\Big], \end{aligned} \tag{10}$$

where the learning rate satisfies $\lambda = \frac{1}{L}$.

Proof: Refer to Appendix B. ∎

It is indicated by Theorem 2 that the convergence rate of the considered FL is partially determined by $\mathbb{E}\big[\|\mathbf{e}^{(\mathrm{t})}\|^2\big]$. Therefore, in order to increase the convergence rate, $\mathbb{E}\big[\|\mathbf{e}^{(\mathrm{t})}\|^2\big]$ should be reduced[2]. Note that $\mathbb{E}\big[\|\mathbf{e}^{(\mathrm{t})}\|^2\big]$ is caused by device selection, which can be considered as one-stage cluster sampling without unequal sizes [33]. With random device selection, the following theorem can be obtained.

Theorem 3. *By utilizing the random device selection scheme, the convergence rate of the considered FL framework is decided by*

$$\mathbb{E}\Big[\|\mathbf{e}^{(\mathrm{t})}\|^2\Big] = \Big(1-\frac{K}{N}\Big)\frac{\sum_{n\in\mathcal{N}}\beta_n^2\big\|\nabla f_n(\mathbf{w}^{(\mathrm{t})}) - \nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\big\|^2}{K(N-1)(\frac{1}{N}\sum_{n\in\mathcal{N}}\beta_n)^2}. \tag{11}$$

Proof: Refer to Appendix C. ∎

It is indicated that the expression in Theorem 3 is based on the averaged gradients of all devices, i.e., $\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})$, which cannot be obtained due to the limitation of available subchannels. As mentioned in [33], by replacing $\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})$ with $\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)$, the approximate ratio estimator can be obtained as follows.

Corollary 1. *By utilizing the random device selection scheme, the approximate effect on the convergence rate is given by*

$$\mathbb{E}\Big[\|\mathbf{e}^{(\mathrm{t})}\|^2\Big] = \Big(1-\frac{K}{N}\Big)\frac{\sum_{n\in\mathcal{S}_t}\beta_n^2\big\|\nabla f_n(\mathbf{w}^{(\mathrm{t})}) - \nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)\big\|^2}{K(K-1)(\frac{1}{N}\sum_{n\in\mathcal{N}}\beta_n)^2}. \tag{12}$$

The above theorems and corollary indicate that the convergence rate of the proposed FL algorithm can be improved if

[2]It is worth pointing out that due to the fact that $\mu \le L$, $(1-\frac{\mu}{L}) \in [0,1)$, and hence, the effect of $\mathbb{E}\big[\|\mathbf{e}^{(\mathrm{t})}\|^2\big]$ diminishes with training. In other words, the impact of error $\mathbb{E}\big[\|\mathbf{e}^{(\mathrm{t})}\|^2\big]$ on convergence rate is more significant at the later stage of training.

1) the number of selected devices K is increased;
2) the deviation from the mean of the selected device, i.e., $\left\|\nabla f_n(\mathbf{w}^{(\mathrm{t})}) - \nabla F(\mathbf{w}^{(\mathrm{t})}, \mathcal{N})\right\|$, is decreased.

C. Signal Model

At the local training stage, the computational time consumption is given by

$$T_{k,n}^{\mathrm{cp}} = \frac{\mu \beta_n}{\tau_{k,n} C_n}, \tag{13}$$

where μ is the required central processing unit (CPU) cycles to train one sample, $\tau_{k,n} \in [0,1]$ is the computational resource allocation coefficient of device n assigned to sub-channel k, and C_n is the available CPU cycles at device n. Accordingly, the computational energy consumption can be presented as follows:

$$E_{k,n}^{\mathrm{cp}} = \kappa \mu \beta_n (\tau_{k,n} C_n)^2, \tag{14}$$

where κ is the power consumption coefficient of each CPU cycle. At the communication stage, the local models are transmitted to the server for aggregation. With the assigned sub-channel k, the achievable data rate of device n is

$$R_{k,n} = B \log_2(1 + \alpha_{k,n} P_n |h_{k,n}|^2), \tag{15}$$

where B is the bandwidth of each sub-channel, $\alpha_{k,n} \in [0,1]$ is the power allocation coefficient, P_n is the maximum transmit power of device n, and $|h_{k,n}|^2$ is the normalized channel gain of device n assigned to sub-channel k. Specifically, $|h_{k,n}|^2 = |\hat{h}_{k,n}|^2 \sigma^{-2}$, where $|\hat{h}_{k,n}|^2 = |g_n|^2 \eta d_n^{-\alpha}$ is the channel gain, σ^2 is the variance of noise, $g_n \sim \mathcal{CN}(0,1)$ is the small-scale fading coefficient, η is the frequency dependent factor, d_n is the distance between device n and the server, and α is the path loss exponent. The time consumption for communication can be expressed as follows:

$$T_{k,n}^{\mathrm{cm}} = \frac{D}{R_{k,n}}, \tag{16}$$

where D is the data size of local models. It is assumed that the data size of local models is the same for all devices. Based on the communication time, the energy consumption for communication is given by

$$E_{k,n}^{\mathrm{cm}} = \alpha_{k,n} P_n T_{k,n}^{\mathrm{cm}}. \tag{17}$$

For any device n assigned to sub-channel k, the time consumption in any round is

$$T_{k,n} = T_{k,n}^{\mathrm{cp}} + T_{k,n}^{\mathrm{cm}}, \tag{18}$$

and the time consumption of this communication round is determined by the most time-consuming device, as shown in follows:

$$T^{(\mathrm{t})} = \max_{n \in \mathcal{N}} \left\{ \sum_{k \in \mathcal{K}} \psi_{k,n}^{(\mathrm{t})} T_{k,n} \right\}, \tag{19}$$

where $\psi_{k,n}^{(\mathrm{t})} \in \{0,1\}$ is the sub-channel assignment indicator. Specifically, $\psi_{k,n}^{(\mathrm{t})} = 1$ indicates that device n is assigned to sub-channel k in round t, and $\psi_{k,n}^{(\mathrm{t})} = 0$ otherwise. In this communication round, the total energy consumption of any device n assigned to sub-channel k is

$$E_{k,n} = E_{k,n}^{\mathrm{cp}} + E_{k,n}^{\mathrm{cm}}. \tag{20}$$

III. Problem Formulation

In this section, the concept of AoI [34], [35] is introduced and an AoI minimization problem is considered. For any device n in round t, the AoI can be presented as follows:

$$A_n^{(\mathrm{t})} = \begin{cases} A_n^{(\mathrm{t}-1)} + T^{(\mathrm{t})}, & \text{if} \quad \sum_{k=1}^{K} \psi_{k,n}^{(\mathrm{t})} = 0, \\ 0, & \text{if} \quad \sum_{k=1}^{K} \psi_{k,n}^{(\mathrm{t})} = 1. \end{cases} \tag{21}$$

The above equation can be rewritten as follows:

$$A_n^{(\mathrm{t})} = \left(1 - \sum_{k=1}^{K} \psi_{k,n}^{(\mathrm{t})}\right) \left(A_n^{(\mathrm{t}-1)} + T^{(\mathrm{t})}\right). \tag{22}$$

Moreover, the AoI of all devices in the initial state is set to zero, i.e., $A_n^{(0)} = 0, \forall n \in \mathcal{N}$. The overall AoI minimization problem can be formulated as follows:

$$\min_{\boldsymbol{\psi}, \boldsymbol{\tau}, \boldsymbol{\alpha}} \quad \sum_{n=1}^{N} A_n^{(\mathrm{t})} \tag{23}$$

$$\text{s.t.} \quad E_{k,n} \le E_n^{\max}, \forall k \in \mathcal{K}, \forall n \in \mathcal{N}, \tag{23a}$$

$$\tau_{k,n} \in [0,1], \forall k \in \mathcal{K}, \forall n \in \mathcal{N}, \tag{23b}$$

$$\alpha_{k,n} \in [0,1], \forall k \in \mathcal{K}, \forall n \in \mathcal{N}, \tag{23c}$$

$$\psi_{k,n}^{(\mathrm{t})} \in \{0,1\}, \forall k \in \mathcal{K}, \forall n \in \mathcal{N}, \tag{23d}$$

$$\sum\nolimits_{n \in \mathcal{N}} \psi_{k,n}^{(\mathrm{t})} = 1, \forall k \in \mathcal{K}, \tag{23e}$$

$$\sum\nolimits_{k \in \mathcal{K}} \psi_{k,n}^{(\mathrm{t})} \in \{0,1\}, \forall n \in \mathcal{N}, \tag{23f}$$

where $\boldsymbol{\psi}$, $\boldsymbol{\tau}$ and $\boldsymbol{\alpha}$ are the collections of sub-channel assignment indicators, computational resource allocation coefficients and power allocation coefficients, respectively. In constraint (23a), the maximum energy consumption $E_n^{\max}$ is included in each communication round. The value ranges of $\tau_{k,n}$, $\alpha_{k,n}$ and $\psi_{k,n}^{(\mathrm{t})}$ are presented in constraints (23c)-(23d). Constraints (23e) and (23f) respectively show that any sub-channel can be occupied by one device and any device can be assigned to at most one sub-channel. As indicated by constraint (23f) that only part of devices are assigned to sub-channels, device selection is integrated in the sub-channel assignment problem. Compared to the conventional latency minimization problem, the inclusion of AoI can efficiently reduce the idle time of devices and thus increases the fairness of device selection. As a result, higher accuracy of FL can be achieved.

Due to the presence of integer variables, the formulated AoI minimization problem is difficult to transform into a convex problem. Therefore, this problem is decoupled into two sub-problems, including the sub-channel assignment problem and the resource allocation problem. The sub-channel assignment problem is presented as follows:

$$\min_{\boldsymbol{\psi}} \quad \sum_{n=1}^{N} A_n^{(\mathrm{t})} \tag{24}$$

$$\text{s.t.} \quad \text{(23d), (23e), (23f).}$$

It is worth pointing out that the above problem also includes device selection. Specifically, if any device is assigned to a sub-channel, it is selected in this communication round. In the resource allocation problem, the sub-channel assignment indicator $\boldsymbol{\psi}$ is fixed. By removing the constant part, the

AoI minimization problem can be transformed to a latency minimization problem as

$$\min_{\boldsymbol{\tau},\boldsymbol{\alpha}} \quad \max\{T_{k,n}|\forall n\in\mathcal{S}_t\} \tag{25}$$
$$\text{s.t.} \quad \text{(23a), (23b), (23c).}$$

IV. KKT Conditions based Resource Allocation

In resource allocation problem (25), the selected devices and the corresponding sub-channels are given. Moreover, the resource allocation coefficients of any device, i.e., $\tau_{k,n}$ and $\alpha_{k,n}$, are independent of other devices. Therefore, problem (25) can be further divided into K sub-problems, where the resource allocation problem of device n assigned to sub-channel k is given by

$$\min_{\tau_{k,n},\alpha_{k,n}} \frac{\mu\beta_n}{\tau_{k,n}C_n}+\frac{D}{B\log_2(1+\alpha_{k,n}P_n|h_{k,n}|^2)} \tag{26}$$
$$\text{s.t.} \quad \kappa\mu\beta_n(\tau_{k,n}C_n)^2+\frac{\alpha_{k,n}P_nD}{B\log_2(1+\alpha_{k,n}P_n|h_{k,n}|^2)}\le E_n^{\max}, \tag{26a}$$
$$0\le\tau_{k,n}\le 1, \tag{26b}$$
$$0\le\alpha_{k,n}\le 1. \tag{26c}$$

Note that the above problem is infeasible in the extreme case, as shown in the following.

Proposition 1. *The resource allocation problem in (26) is infeasible if the following condition is satisfied:*

$$\ln(2)D\ge E_n^{\max}B|h_{k,n}|^2. \tag{27}$$

Proof: Refer to Appendix D. ∎

Moreover, problem (26) is non-convex due to the objective function and constraint (26a). In this case, $T_{k,n}^{\mathrm{cm}}$ is introduced to transform the problem. From (15) and (16), the following equation can be obtained:

$$\alpha_{k,n}=\frac{2^{\frac{D}{BT_{k,n}^{\mathrm{cm}}}}-1}{P_n|h_{k,n}|^2}, \tag{28}$$

and constraint (26c) becomes

$$\begin{cases} 1-2^{\frac{D}{BT_{k,n}^{\mathrm{cm}}}}\le 0, \\ 2^{\frac{D}{BT_{k,n}^{\mathrm{cm}}}}-1-P_n|h_{k,n}|^2\le 0. \end{cases} \tag{29}$$

At this stage, problem (26) can be transformed as follows:

$$\min_{\tau_{k,n},T_{k,n}^{\mathrm{cm}}} \quad \frac{\mu\beta_n}{\tau_{k,n}C_n}+T_{k,n}^{\mathrm{cm}} \tag{30}$$
$$\text{s.t.} \quad \kappa\mu\beta_n(\tau_{k,n}C_n)^2+T_{k,n}^{\mathrm{cm}}\frac{2^{\frac{D}{BT_{k,n}^{\mathrm{cm}}}}-1}{|h_{k,n}|^2}\le E_n^{\max}, \tag{30a}$$
$$\text{(26b), (29).}$$

Note that the above problem is equivalent to problem (26). Therefore, the optimality of the resource allocation problem is not affected by this transformation. Problem (30) is convex and satisfies Slater's condition, which means KKT conditions can be adopted to obtain the optimal solution. The Lagrangian function of problem (30) is given by

$$\begin{aligned}\mathcal{L}=&\frac{\mu\beta_n}{\tau_{k,n}C_n}+T_{k,n}^{\mathrm{cm}}+\lambda_1\left[\kappa\mu\beta_n(\tau_{k,n}C_n)^2+T_{k,n}^{\mathrm{cm}}\frac{2^{\frac{D}{BT_{k,n}^{\mathrm{cm}}}}-1}{|h_{k,n}|^2}-E_n^{\max}\right]\\&+\lambda_2(-\tau_{k,n})+\lambda_3(\tau_{k,n}-1)+\lambda_4(1-2^{\frac{D}{BT_{k,n}^{\mathrm{cm}}}})\\&+\lambda_5(2^{\frac{D}{BT_{k,n}^{\mathrm{cm}}}}-1-P_n|h_{k,n}|^2),\end{aligned} \tag{31}$$

where $\lambda_i,\forall i\in\{1,2,3,4,5\}$ is the Lagrangian multiplier for the corresponding constraint. The partial derivatives of function (31) with respect to $\tau_{k,n}$ and $T_{k,n}^{\mathrm{cm}}$ are

$$\frac{\partial\mathcal{L}}{\partial\tau_{k,n}}=-\frac{\mu\beta_n}{(\tau_{k,n})^2C_n}+2\lambda_1\kappa\mu\beta_n\tau_{k,n}C_n^2-\lambda_2+\lambda_3, \tag{32}$$

and

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial T_{k,n}^{\mathrm{cm}}}=&\lambda_1\left(\frac{2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}-1}{|h_{k,n}|^2}-\frac{\ln(2)D2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}}{B|h_{k,n}|^2T_{k,n}^{\mathrm{cm}*}}\right)\\&+\frac{\ln(2)\lambda_4D2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}}{B(T_{k,n}^{\mathrm{cm}*})^2}-\frac{\ln(2)\lambda_5D2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}}{B(T_{k,n}^{\mathrm{cm}*})^2}+1,\end{aligned} \tag{33}$$

respectively. The optimal solutions $\tau_{k,n}^*$ and $T_{k,n}^{\mathrm{cm}*}$ should satisfy the following equations:

$$-\frac{\mu\beta_n}{(\tau_{k,n}^*)^2C_n}+2\lambda_1\kappa\mu\beta_n\tau_{k,n}^*C_n^2-\lambda_2+\lambda_3=0, \tag{34}$$

and

$$\begin{aligned}&\lambda_1B(T_{k,n}^{\mathrm{cm}*})^2(2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}-1)-\ln(2)\lambda_1DT_{k,n}^{\mathrm{cm}*}2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}\\&+\ln(2)(\lambda_4-\lambda_5)|h_{k,n}|^2D2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}+|h_{k,n}|^2B(T_{k,n}^{\mathrm{cm}*})^2=0.\end{aligned} \tag{35}$$

Moreover, the following conditions can be obtained:

$$\begin{cases} \kappa\mu\beta_n(\tau_{k,n}C_n)^2+T_{k,n}^{\mathrm{cm}*}\frac{2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}-1}{|h_{k,n}|^2}-E_n^{\max}\le 0, & \text{(36a)}\\ -\tau_{k,n}^*\le 0,\ \tau_{k,n}^*-1\le 0, & \text{(36b)}\\ 1-2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}\le 0,\ 2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}-P_n|h_{k,n}|^2-1\le 0, & \text{(36c)}\\ \lambda_1\left[\kappa\mu\beta_n(\tau_{k,n}^*C_n)^2+T_{k,n}^{\mathrm{cm}*}\frac{2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}-1}{|h_{k,n}|^2}-E_n^{\max}\right]=0, & \text{(36d)}\\ \lambda_2\tau_{k,n}^*=0,\ \lambda_3(\tau_{k,n}^*-1)=0, & \text{(36e)}\\ \lambda_4[1-2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}]=0,\ \lambda_5[2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}-P_n|h_{k,n}|^2-1]=0, & \text{(36f)}\\ \lambda_i\ge 0,\forall i\in\{1,2,3,4,5\}. & \text{(36g)}\end{cases}$$

By analyzing the KKT conditions, the optimal solution of problem (30) can be presented as follows:

Proposition 2. *By defining the following variables:*

$$\begin{cases} A_1\triangleq 2\kappa\left[\frac{E_n^{\max}}{\kappa\mu\beta_n}-\frac{P_nD}{\kappa\mu\beta_nB\log_2(1+P_n|h_{k,n}|^2)}\right]^{\frac{3}{2}},\\ A_2\triangleq|h_{k,n}|^2(E_n^{\max}-\kappa\mu\beta_nC_n^2),\\ A_0\triangleq-\frac{A_2\ln(2^{\frac{D}{B}})}{A_2W_{-1}(-\ln(2^{\frac{D}{A_2B}})/2^{\frac{D}{A_2B}})+\ln(2^{\frac{D}{B}})}, \end{cases} \tag{37}$$

four solutions can be presented below.

1) If the following condition is satisfied

$$\kappa\mu\beta_nC_n^2+\frac{P_nD}{B\log_2(1+P_n|h_{k,n}|^2)}\le E_n^{\max}, \tag{38}$$

the optimal solution is

$$\begin{cases} \tau_{k,n}^{*}=1, \\ T_{k,n}^{\mathrm{cm}*}=\dfrac{D}{B\log_2(1+P_n|h_{k,n}|^2)}. \end{cases} \tag{39}$$

2) In the case that

$$0<E_n^{\max}-\frac{P_nD}{B\log_2(1+P_n|h_{k,n}|^2)}\le\kappa\mu\beta_nC_n^2, \tag{40}$$

and

$$A_1|h_{k,n}|^2+P_n|h_{k,n}|^2>(1+P_n|h_{k,n}|^2)\ln(1+P_n|h_{k,n}|^2), \tag{41}$$

the optimal solution is

$$\begin{cases} \tau_{k,n}^{*}=\dfrac{1}{C_n}\left[\dfrac{E_n^{\max}}{\kappa\mu\beta_n}-\dfrac{P_nD}{\kappa\mu\beta_nB\log_2(1+P_n|h_{k,n}|^2)}\right]^{\frac{1}{2}}, \\ T_{k,n}^{\mathrm{cm}*}=\dfrac{D}{B\log_2(1+P_n|h_{k,n}|^2)}. \end{cases} \tag{42}$$

3) If the following inequalities hold:

$$A_0\ge\frac{D}{B\log_2(1+P_n|h_{k,n}|^2)}, \tag{43}$$

and

$$\ln(2)D2^{\frac{D}{A_0B}}-A_2B>2\kappa C_n^3A_0B|h_{k,n}|^2, \tag{44}$$

the optimal solution is

$$\begin{cases} \tau_{k,n}^{*}=1, \\ T_{k,n}^{\mathrm{cm}*}=A_0. \end{cases} \tag{45}$$

4) Otherwise, the optimal solution can be obtained by solving the following equations:

$$\begin{cases} 2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}-1-\frac{\ln(2)D}{BT_{k,n}^{\mathrm{cm}*}}2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}+2\kappa(\tau_{k,n}^{*})^3C_n^3|h_{k,n}|^2=0, \\ \kappa\mu\beta_n(\tau_{k,n}^{*}C_n)^2+T_{k,n}^{\mathrm{cm}*}\frac{2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}-1}{|h_{k,n}|^2}-E_n^{\max}=0. \end{cases} \tag{46}$$

where $\tau_{k,n}^{*}\in(0,1]$ *and* $T_{k,n}^{\mathrm{cm}*}\in(0,\frac{D}{B\log_2(1+P_n|h_{k,n}|^2)}]$.

Proof: Refer to Appendix E. ∎

Based on $T_{k,n}^{\mathrm{cm}*}$, the optimal power allocation coefficient $\alpha_{k,n}^{*}$ can be calculated from (28), and the decoupled resource allocation problem in (25) is solved.

V. Matching based Device Selection and Sub-Channel Assignment

Since the number of devices may exceed the number of sub-channels, a matching based algorithm is developed to jointly solve the device selection and sub-channel assignment problem. In this case, a device is selected if it is assigned to a valid sub-channel, which yields a feasible solution to the resource allocation problem. To this end, $N-K$ virtual sub-channels are introduced in this section, where the channel gains of all devices in virtual sub-channels are zero. As a result, the collection of all sub-channels, including physical and virtual sub-channels, can be denoted by $\mathcal{M}=\{1,2,\ldots,K,K+1,\ldots,N\}$.

A. Design of Matching-based Algorithm

By introducing virtual sub-channels, $\mathcal{N}$ and $\mathcal{M}$ become two disjoint sets with the same size, and therefore, a one-to-one matching can be constructed, as shown in follows:

Definition 1. *Given two disjoint sets $\mathcal{N}$ and $\mathcal{M}$, a one-to-one matching Φ is a mapping from $\mathcal{N}$ to $\mathcal{M}$, such that*
1) $\Phi(n)\in\mathcal{M},\forall n\in\mathcal{N},\ \Phi(k)\in\mathcal{N},\forall k\in\mathcal{M}$;
2) $|\Phi(n)|=1,\forall n\in\mathcal{N},\ |\Phi(k)|=1,\forall k\in\mathcal{M}$;
3) $n=\Phi(k)\Rightarrow\Phi(n)=k$.

In the above definition, the details of the considered matching is described. Condition *1)* indicates that any player in one set is matched with one player in the other set. Condition *2)* indicates that any player is only matched with one player. Condition *3)* indicates that the matching of device n and sub-channel k can be inferred from each other. The above definition implies that the considered matching is also a swap matching. That is, if any device intends to be matched to a sub-channel, it should exchange with the device occupying this sub-channel. The swap matching is defined as follows:

Definition 2. *From any matching Φ, a swap matching $\Phi_{n'}^{n}$ is obtained by*

$$\Phi_{n'}^{n}=\{\Phi\backslash\{(k,n),(k',n')\}\}\cup\{(k',n),(k,n')\}, \tag{47}$$

where $\Phi(k)=n,\ \Phi(k')=n',\ \Phi_{n'}^{n}(k)=n',\ \Phi_{n'}^{n}(k')=n$.

Based on the objective function of problem (24), the utility of any device n in matching Φ can be presented as follows:

$$U_n(\Phi)=\left(1-\sum_{k\in\mathcal{K}}\psi_{k,n}^{(\mathrm{t})}\right)\left(A_n^{(\mathrm{t}-1)}+\max_{n\in\mathcal{N}}\left\{\sum_{k\in\mathcal{K}}\psi_{k,n}^{(\mathrm{t})}T_{k,n}\right\}\right). \tag{48}$$

It is indicated by the above equation that the utility of any device n is zero if this device is assigned to a physical sub-channel. On the other hand, if a device is assigned to a virtual sub-channel, it is not selected in this communication round, and hence, its utility is entirely determined by other devices. Due to the fact that each sub-channel is occupied by one device, the utility of any sub-channel is decided by the occupied device, i.e., $U_k(\Phi)=U_{\Phi(k)}(\Phi),\forall k\in\mathcal{M}$. In this section, unselfish players are considered because the exchange operations between devices will be strictly restricted with the conventional selfish players. Specifically, any device assigned to a physical sub-channel cannot be swapped with a device assigned to a virtual sub-channel, as its utility will be increased from zero to a positive value. As a result, a preference list can be constructed. For any player $i\in\mathcal{N}\cup\mathcal{M}$, it prefers matching $\Phi_n^{n'}$ over matching Φ if

$$\Phi\prec_i\Phi_{n'}^{n}\Leftrightarrow U_n(\Phi)+U_{n'}(\Phi)>U_n(\Phi_{n'}^{n})+U_{n'}(\Phi_{n'}^{n}). \tag{49}$$

The transformation of matching from Ψ to $\Phi_n^{n'}$ should be approved by all involved players, including the exchanged devices and the occupied sub-channels. Due to the fact that the utility of any sub-channel is equal to the utility of the occupied device, the approval of sub-channels can be omitted. Furthermore, the server tends to select more devices in each communication round, thus, it will avoid assigning any device

Algorithm 1 Matching based Algorithm

Initialization:
Randomly match all devices and sub-channels.
Main Loop:
for $n \in \mathcal{N}$ **do**
  Device n searches device $n' \in \mathcal{N}$, where $n \neq n'$.
  if $\Phi \prec_i \Phi^n_{n'}, \forall i \in \mathcal{N}$ **then**
    Matching $\Phi^n_{n'}$ is approved.
    Devices n and n' exchange sub-channels.
    Set $\Phi = \Phi^n_{n'}$.
  end if
end for

to a physical sub-channel where it is not feasible, as defined in Proposition 1. If all involved players agree on swap matching $\Phi^{n'}_n$, then (n, n') becomes a swap-blocking pair. The definition is presented below.

Definition 3. *A swap-blocking pair (n, n') holds if and only if the following conditions can be satisfied:*
1) $U_n(\Phi) + U_{n'}(\Phi) > U_n(\Phi^n_{n'}) + U_{n'}(\Phi^n_{n'})$;
2) $\ln(2)D < E^{\max}_j B|h_{i,j}|^2, \forall j \in \{n, n'\}$, *if* $i = \Phi^n_{n'}(j) \in \mathcal{K}$.

Based on the definition of swap-blocking pairs, the matching based device selection and sub-channel assignment algorithm can be proposed in **Algorithm 1**. During the proposed algorithm, devices are sequentially selected to search for other devices with the same sequence, where the optimal resource allocation in Proposition 2 is adopted. If a swap-blocking pair (n, n') is obtained, the current matching is transformed to $\Phi^n_{n'}$. If no swap-blocking pair can be constructed within a complete cycle of the main loop, the algorithm ends and the final matching becomes the solution of problem (24).

B. Properties Analysis

This section focuses on analyzing the properties of the proposed algorithm, including complexity, convergence, and stability.

1) Complexity: The complexity of the proposed algorithm can be presented as follows:

Proposition 3. *Given a number of main loops C, the computational complexity of the matching based algorithm is $\mathcal{O}(CN^2)$.*

Proof: In the worst case, each device needs to search all other devices, and then $N(N-1)$ times of calculations are implemented in one loop. With the given number of main loops C, the complexity of the proposed algorithm approximates to $\mathcal{O}(CN^2)$. ■

2) Convergence: The convergence of the proposed algorithm can be presented as follows.

Proposition 4. *From any initial matching, the matching based algorithm can always converge to a stable matching.*

Proof: In the proposed algorithm, the matching can be transformed only if a swap-blocking pair is found. Based on the definition of swap-blocking pairs, the sum utility of all involved players should be strictly reduced, while the utilities of uninvolved players remain the same or decrease. Therefore, the sum utility of all devices is strictly decreasing with matching transformations. With the finite devices and sub-channels, the potential matchings are also finite, and hence, a final matching is always obtained. ■

3) Stability: The stability of the proposed algorithm can be analyzed according to the following definition:

Definition 4. *A matching is two-sided exchange-stable if and only if there is no swap-blocking pair.*

The stability of the proposed algorithm is shown below.

Proposition 5. *The resulting final matching from Algorithm 1 is always two-sided exchange-stable.*

Proof: Assuming that the final matching obtained by Algorithm 1 is not two-sided exchange-stable, there exists at least one swap-blocking pair, which can further reduce the sum utility. This case contradicts the convergence proposition and therefore cannot hold. This proposition can thus be proved. ■

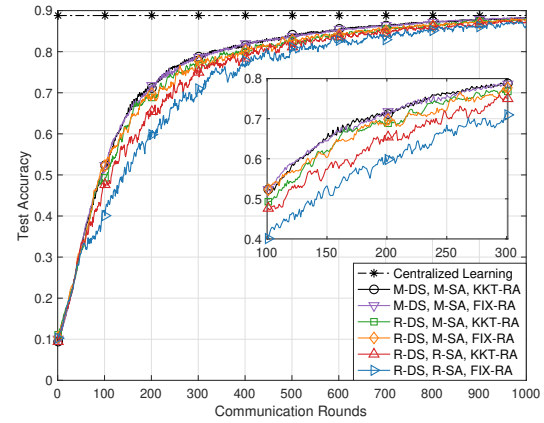
VI. Simulation Results

In the simulation, N devices are randomly distributed within a radius centered at the server. During the learning process, the positions of all devices remain the same, while the small-scale fading varies from round to round. In order to mitigate the effect of device location on learning results, in Fig. 1 to Fig. 3, all schemes adopt the same set of generated locations. By selecting K devices randomly without replacement, random device selection (R-DS) is included as the benchmark for matching based device selection (M-DS). In terms of sub-channel assignment, the matching based sub-channel assignment (M-SA) approach in [29] is incorporated to assign the selected devices into available sub-channels, and its initial state is regarded as random sub-channel allocation (R-SA). Moreover, fixed resource allocation (FIX-RA) is also considered as the baseline for KKT based resource allocation (KKT-RA), in which all devices' computational resource allocation coefficients and power allocation coefficients are set to 0.5. This simulation is based on unbalanced non-IID MNIST and CIFAR-10 datasets. That is, the devices have different data sizes and samples from each device have the limited number of labels. In particular, each device has one label in MNIST dataset and five labels in CIFAR-10 dataset. For MNIST dataset, the neural network is built with two ReLu hidden layers (128 and 256 neurons) and a Softmax output layer. For CIFAR-10 dataset, a multi-layer convolutional neural network (CNN) is constructed by stacking six 3×3 Conv2D layers, a ReLu layer and a Softmax output layer. Specifically, the Conve2D layers contain two 32-filter Conv2D layers, two 64-filter Conv2D layers, and two 128-filter Conv2D layers, where ever two layers followed by a 2×2 max pooling layer and a 0.25 dropout layer. The ReLu layer includes 128 neurons, followed by a 0.5 dropout layer. Furthermore, the simulations on MNIST and CIFAR-10 datasets utilize full-batch and mini-batch, respectively, where the epoch is 1, and the batch size for mini-batch is 128. The main parameters of the simulation are shown in Table I.

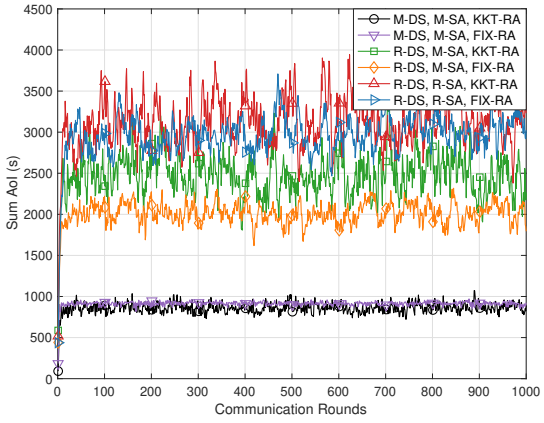
Fig. 1 presents the performance of FL under different schemes and the corresponding sum AoI. In order to show

TABLE I: Table of Parameters

Number of Devices	$N = 20$
Radius of the disc	500 m
Carrier frequency	$f = 1$ GHz
AWGN noise power	$\sigma^2 = -174$ dBm
Path loss exponent	$a = 3.76$
Bandwidth for each sub-channel	$B = 1$ MHz
Power consumption coefficient	$\kappa_0 = 10^{-28}$
CPU cycles for each bit of tasks	$\mu = 10^7$
Model size	$D = 1$ Mbit
Learning rate (MNIST/CIFAR-10)	$\lambda = 0.01/0.001$
Optimizer (MNIST/CIFAR-10)	SGD/Adam



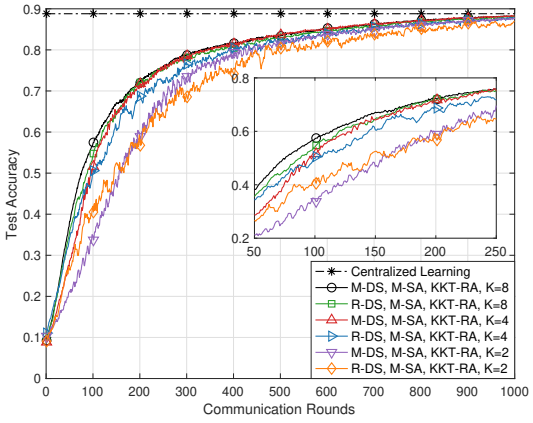
(a) Test accuracy.



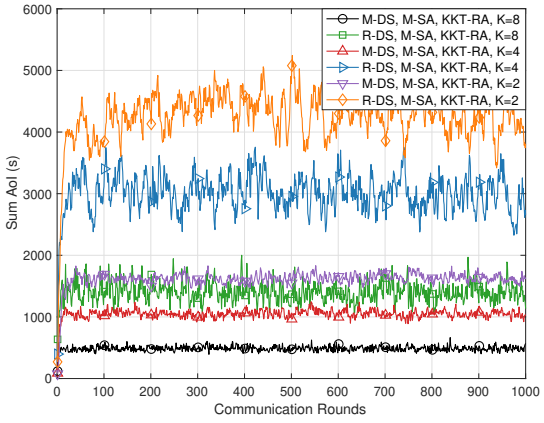
(b) Sum AoI.

Fig. 1: The performance of FL on MNIST dataset. $K = 5$, $P_t = 15$ dBm, $C_n = 0.5$ GHz, and $E_n^{\max} = 0.1$ Joule.

a stable result, each curve is an average of 10 simulations. It is indicated by Fig. 1(a) that the proposed solution, i.e., M-DS and M-SA with KKT-RA, can achieve the best performance, including the fastest convergence rate and the highest test accuracy. By comparing with i) M-DS and M-SA with FIX-RA and ii) R-DS and M-SA with KKT-RA, it shows that in terms of the test accuracy, the impact of utilizing AoI minimization based device selection is significant, while the effect of optimal resource allocation is not obvious. Moreover, by employing M-SA and KKT-RA, the number of selected devices in each round can be increased, and therefore, the performance of R-DS can be improved. Fig. 1(b) shows that the proposed scheme is able to significantly reduce the overall AoI. Furthermore, M-SA and KKT-RA can be utilized independently or jointly to



(a) Test accuracy.



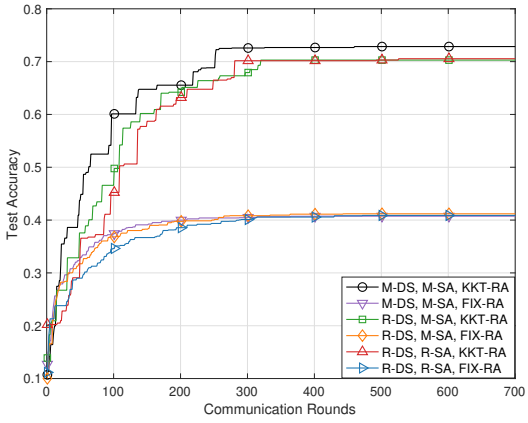
(b) Sum AoI.

Fig. 2: The impact of the number of available sub-channels. $P_t = 15$ dBm, $C_n = 0.5$ GHz, and $E_n^{\max} = 0.1$ Joule.

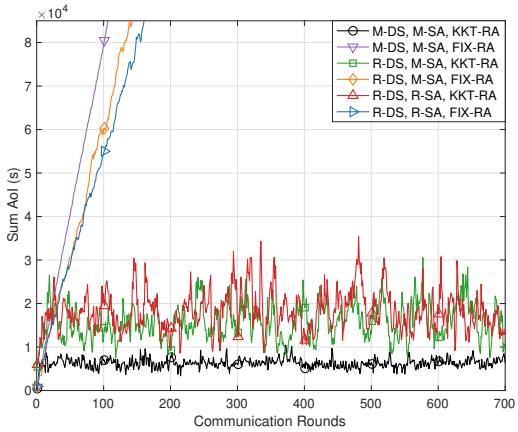
reduce the sum AoI.

The impact of the number of sub-channels is shown in Fig. 2. It is demonstrated by Fig. 2(a) that the convergence rate of the considered FL can be improved by increasing the number of available sub-channels, which confirms the conclusion in Theorem 3 and Corollary 1. Fig. 2(b) shows that the overall AoI can be reduced when more devices are selected in each communication round, and the proposed scheme is able to outperform the corresponding benchmark with any available sub-channels.

The proposed scheme is simulated with CIFAR-10 dataset, as shown in Fig. 3. In this simulation, the maximum test accuracy is plotted to avoid overcrowding. Due to the fact that the data size of each device with CIFAR-10 dataset is greater than that with MNIST dataset, KKT based resource allocation plays a more important role in Fig. 3. It can be found from Fig. 3(a) that the highest test accuracy of the schemes with fixed resource allocation is 41% while the lowest test accuracy of schemes with KKT based resource allocation is 70%. The reason is shown in Fig. 3(b). That is, with the fixed resource allocation, some devices cannot successfully transmit signals to the server, and hence, their AoI is monotonically increasing. Moreover, the proposed scheme can still achieve the best performance, where the test accuracy can be 73% and



(a) Test accuracy.



(b) Sum AoI.

Fig. 3: The performance of FL on CIFAR-10 dataset. $K = 5$, $P_t = 15$ dBm, $C_n = 0.5$ GHz, and $E_n^{\max} = 0.2$ Joule.

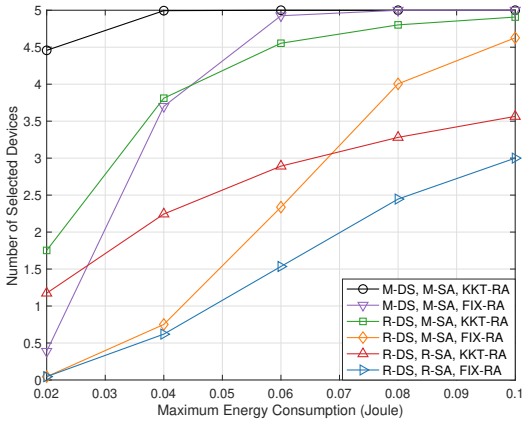


Fig. 4: The number of selected devices versus the maximum energy. $K = 5$, $P_t = 15$ dBm, and $C_n = 0.5$ GHz.

the overall AoI is around 7000 seconds.

The relationship between the maximum energy consumption and the number of selected devices is presented in Fig. 4. Basically, for all schemes, the number of selected devices can be increased with the increasing maximum energy consumption. Moreover, this figure explains why the proposed scheme can achieve the highest test accuracy. That is, by utilizing

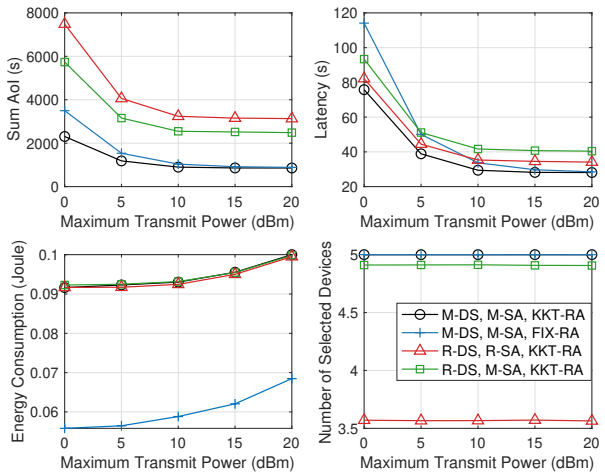


Fig. 5: The impact of the maximum transmit power. $K = 5$, $C_n = 0.5$ GHz, and $E_n^{\max} = 0.1$ Joule.

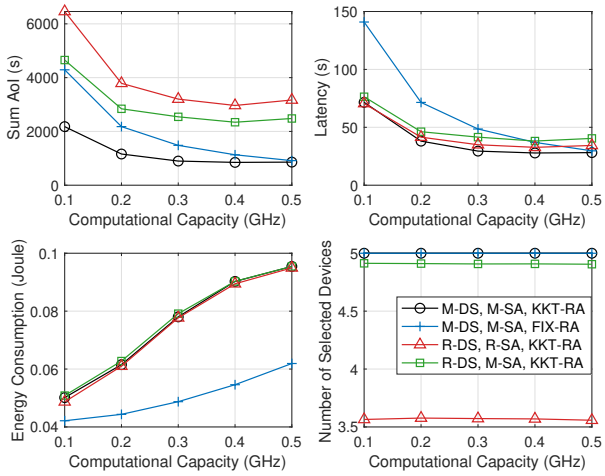


Fig. 6: The impact of the computational capacity. $K = 5$, $P_t = 15$ dBm, and $E_n^{\max} = 0.1$ Joule.

the KKT based resource allocation and matching based device selection and sub-channel assignment, the number of feasible devices is increased. As a result, the server can select devices from a larger subset in order to increase the convergence rate. Furthermore, this figure confirms the conclusion in Proposition 1 that $E_n^{\max}$ can effect the feasibility of the resource allocation problem.

The impact of the maximum transmit power and computational capacity is shown in Fig. 5 and Fig. 6, respectively. With increasing transmit power or computational capacity, the time consumption for transmission or computation would decrease, and hence, the overall AoI would drop due to the reduction in latency. It is worth emphasizing that the proposed scheme can also achieve the minimal latency compared to R-DS, and the developed KKT-RA has the ability to reduce latency. Meanwhile, the energy consumption is monotonically raising with the transmit power and computational capacity. In terms of the proposed scheme, it is able to achieve the minimum sum AoI, because the KKT based resource allocation solution and the matching based device selection and sub-channel assignment

algorithm can efficiently utilize the given energy and sub-channels, respectively. Furthermore, it is also indicated that the maximum transmit power and computational capacity will not affect the number of selected devices.

VII. Conclusions

This paper investigated FL with non-IID dataset over wireless networks. Due to the limited number of available sub-channels, a subset of devices is selected to participate in aggregation in each communication round. Based on the analysis of weight divergence and convergence rate, the disadvantage of existing device selection strategies is revealed, and then an overall AoI minimization problem is designed to establish a new metric. It is indicated that the AoI minimization based device selection is able to reduce the latency and improve convergence performance without analyzing the local data/models. According to the KKT conditions and matching theory, the formulated AoI minimization problem is solved. Simulation results show that the AoI minimization based device selection scheme can improve the performance of FL with MNIST and CIFAR-10 datasets, and the proposed solution can efficiently utilize the given energy and sub-channels. An important direction for future research is to explore this model in multi-antenna scenarios and perform corresponding beamforming designs.

Appendix A: Proof of Theorem 1

The expected weight divergence between FL and CL can be expressed as follows:

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{w}^{(\mathrm{t}+1)}-\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t}+1)}\|\right]\\&=\mathbb{E}\left[\left\|\mathbf{w}^{(\mathrm{t})}-\lambda\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)-\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t})}+\lambda\nabla F(\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t})},\mathcal{N})\right\|\right] \quad (50)\\&\leq\mathbb{E}\left[\|\mathbf{w}^{(\mathrm{t})}-\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t})}\|\right]+\lambda\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)-\nabla F(\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t})},\mathcal{N})\right\|\right].\end{aligned}$$

By adding and subtracting $\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})$, it becomes

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{w}^{(\mathrm{t}+1)}-\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t}+1)}\|\right]\\&=\mathbb{E}\left[\|\mathbf{w}^{(\mathrm{t})}-\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t})}\|\right]+\lambda\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)-\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\right.\right.\\&\quad\left.\left.+\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})-\nabla F(\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t})},\mathcal{N})\right\|\right]\\&\leq\mathbb{E}\left[\|\mathbf{w}^{(\mathrm{t})}-\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t})}\|\right]+\lambda\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)-\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\right\|\right]\\&\quad+\lambda\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})-\nabla F(\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t})},\mathcal{N})\right\|\right]. \quad (51)\end{aligned}$$

Based on Assumption 1, the weight divergence can be transformed as follows:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{w}^{(\mathrm{t}+1)}-\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t}+1)}\|\right]&\leq(1+\lambda L)\mathbb{E}\left[\|\mathbf{w}^{(\mathrm{t})}-\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t})}\|\right] \quad (52)\\&\quad+\lambda\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)-\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\right\|\right].\end{aligned}$$

Similarly, the weight divergence in round t is given by

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{w}^{(\mathrm{t})}-\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t})}\|\right]&\leq(1+\lambda L)\mathbb{E}\left[\|\mathbf{w}^{(\mathrm{t}-1)}-\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t}-1)}\|\right] \quad (53)\\&\quad+\lambda\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^{(\mathrm{t}-1)},\mathcal{S}_{t-1})-\nabla F(\mathbf{w}^{(\mathrm{t}-1)},\mathcal{N})\right\|\right].\end{aligned}$$

As a result, the following inequality can be obtained:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{w}^{(\mathrm{t}+1)}-\mathbf{w}_{\mathrm{cen}}^{(\mathrm{t}+1)}\|\right]&\leq(1+\lambda L)^t\mathbb{E}\left[\|\mathbf{w}^{(1)}-\mathbf{w}_{\mathrm{cen}}^{(1)}\|\right] \quad (54)\\&\quad+\lambda\sum_{i=1}^{t}(1+\lambda L)^{t-i}\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^{(i)},\mathcal{S}_i)-\nabla F(\mathbf{w}^{(i)},\mathcal{N})\right\|\right],\end{aligned}$$

and the proof is completed.

Appendix B: Proof of Theorem 2

According to [36], the following inequality can be derived with Assumption 1:

$$\begin{aligned}&\frac{1}{2L}\left\|\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})-\nabla F(\mathbf{w}^{(\mathrm{t}+1)},\mathcal{N})\right\|^2\\&\leq F(\mathbf{w}^{(\mathrm{t}+1)},\mathcal{N})-F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})+[\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})^\top(\mathbf{w}^{(\mathrm{t})}-\mathbf{w}^{(\mathrm{t}+1)})]\\&\leq\frac{L}{2}\|\mathbf{w}^{(\mathrm{t})}-\mathbf{w}^{(\mathrm{t}+1)}\|^2. \quad (55)\end{aligned}$$

From (4) and (9), the following equation can be obtained:

$$\mathbf{w}^{(\mathrm{t})}-\mathbf{w}^{(\mathrm{t}+1)}=\lambda\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)=\lambda\left[\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})+\mathbf{e}^{(\mathrm{t})}\right], \quad (56)$$

and then (55) can be transformed as follows:

$$\begin{aligned}&F(\mathbf{w}^{(\mathrm{t}+1)},\mathcal{N})\\&\leq F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})-\lambda\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})^\top\left[\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})+\mathbf{e}^{(\mathrm{t})}\right]\\&\quad+\frac{\lambda^2L}{2}\left\|\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})+\mathbf{e}^{(\mathrm{t})}\right\|^2 \quad (57)\\&=F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})-\lambda\|\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\|^2-\lambda\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})^\top\mathbf{e}^{(\mathrm{t})}\\&\quad+\frac{\lambda^2L}{2}\left[\|\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\|^2+\|\mathbf{e}^{(\mathrm{t})}\|^2+2\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})^\top\mathbf{e}^{(\mathrm{t})}\right].\end{aligned}$$

With the given learning rate $\lambda=\frac{1}{L}$, the above inequality can be rewritten as follows:

$$F(\mathbf{w}^{(\mathrm{t}+1)},\mathcal{N})\leq F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})-\frac{1}{2L}\|\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\|^2+\frac{1}{2L}\|\mathbf{e}^{(\mathrm{t})}\|^2. \quad (58)$$

By subtracting $F(\mathbf{w}^*)$ and taking expectation on both sides, the following inequality can be obtained:

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{w}^{(\mathrm{t}+1)},\mathcal{N})-F(\mathbf{w}^*)\right]&\leq\mathbb{E}\left[F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})-F(\mathbf{w}^*)\right] \quad (59)\\&\quad-\frac{1}{2L}\mathbb{E}\left[\|\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\|^2\right]+\frac{1}{2L}\mathbb{E}\left[\|\mathbf{e}^{(\mathrm{t})}\|^2\right].\end{aligned}$$

Based on Assumption 2, the above inequality becomes

$$\begin{aligned}&\mathbb{E}\left[F(\mathbf{w}^{(\mathrm{t}+1)},\mathcal{N})-F(\mathbf{w}^*)\right] \quad (60)\\&\leq\left(1-\frac{\mu}{L}\right)\mathbb{E}\left[F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})-F(\mathbf{w}^*)\right]+\frac{1}{2L}\mathbb{E}\left[\|\mathbf{e}^{(\mathrm{t})}\|^2\right].\end{aligned}$$

The upper bound of the convergence rate is given by

$$\begin{aligned}&\mathbb{E}\left[F(\mathbf{w}^{(\mathrm{t}+1)},\mathcal{N})-F(\mathbf{w}^*)\right]\\&\leq\left(1-\frac{\mu}{L}\right)^2\mathbb{E}\left[F(\mathbf{w}^{(\mathrm{t}-1)},\mathcal{N})-F(\mathbf{w}^*)\right]\\&\quad+\left(1-\frac{\mu}{L}\right)\frac{1}{2L}\mathbb{E}\left[\|\mathbf{e}^{(\mathrm{t}-1)}\|^2\right]+\frac{1}{2L}\mathbb{E}\left[\|\mathbf{e}^{(\mathrm{t})}\|^2\right]\\&\leq\left(1-\frac{\mu}{L}\right)^t\mathbb{E}\left[F(\mathbf{w}^{(1)},\mathcal{N})-F(\mathbf{w}^*)\right]\\&\quad+\frac{1}{2L}\sum_{i=1}^{t}\left(1-\frac{\mu}{L}\right)^{t-i}\mathbb{E}\left[\|\mathbf{e}^{(i)}\|^2\right], \quad (61)\end{aligned}$$

and the proof is completed.

APPENDIX C: PROOF OF THEOREM 3

To prove this theorem, a binary variable $x_n^{(\mathrm{t})}$ is defined, where $x_n^{(\mathrm{t})}=1$ indicates device n is selected in round t, i.e., $n\in\mathcal{S}_t$; $x_n^{(\mathrm{t})}=0$ otherwise. In any communication round t with the given set of selected devices $\mathcal{S}_t$, the averaged gradients $\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)$ can be rewritten as follows:

$$\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)=\frac{\sum_{n\in\mathcal{S}_t}\beta_n\nabla f_n(\mathbf{w}^{(\mathrm{t})})}{\sum_{n\in\mathcal{S}_t}\beta_n}=\frac{\frac{1}{K}\sum_{n\in\mathcal{N}}x_n^{(\mathrm{t})}\beta_n\nabla f_n(\mathbf{w}^{(\mathrm{t})})}{\frac{1}{K}\sum_{n\in\mathcal{N}}x_n^{(\mathrm{t})}\beta_n}\triangleq\frac{\bar{y}_{\mathcal{S}_t}}{\bar{x}_{\mathcal{S}_t}}. \quad (62)$$

The above equation can be considered as a ratio estimation, where $\bar{y}_{\mathcal{S}_t}$ is positively correlated with $\bar{x}_{\mathcal{S}_t}$. The probability of selecting device n is obtained from $\binom{N}{K}$, as follows:

$$\mathbb{E}[x_n^{(\mathrm{t})}]=P(x_n^{(\mathrm{t})}=1)=\frac{K}{N}. \quad (63)$$

Similarly, $\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})$ can be rewritten as follows:

$$\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})=\frac{\frac{1}{N}\sum_{n\in\mathcal{N}}\beta_n\nabla f_n(\mathbf{w}^{(\mathrm{t})})}{\frac{1}{N}\sum_{n\in\mathcal{N}}\beta_n}\triangleq\frac{\bar{y}_{\mathcal{N}}}{\bar{x}_{\mathcal{N}}}. \quad (64)$$

Hence, the following equation can be obtained:

$$\begin{aligned}&\left\|\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)-\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\right\|\\=&\left\|\frac{\bar{y}_{\mathcal{S}_t}}{\bar{x}_{\mathcal{S}_t}}-\frac{\bar{y}_{\mathcal{N}}}{\bar{x}_{\mathcal{N}}}\right\|=\frac{1}{\bar{x}_{\mathcal{S}_t}}\left\|\bar{y}_{\mathcal{S}_t}-\bar{x}_{\mathcal{S}_t}\frac{\bar{y}_{\mathcal{N}}}{\bar{x}_{\mathcal{N}}}\right\|. \quad (65)\end{aligned}$$

As a result, $\mathbb{E}\left[\|\mathbf{e}^{(\mathrm{t})}\|^2\right]$ can be expressed as follows:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{e}^{(\mathrm{t})}\|^2\right]&=\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)-\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\right\|^2\right]\\&=\frac{1}{(\frac{1}{N}\sum_{n\in\mathcal{N}}\beta_n)^2}\mathbb{E}\left[\left\|\bar{y}_{\mathcal{S}_t}-\bar{x}_{\mathcal{S}_t}\frac{\bar{y}_{\mathcal{N}}}{\bar{x}_{\mathcal{N}}}\right\|^2\right]. \quad (66)\end{aligned}$$

Moreover, the following transformation can be derived:

$$\begin{aligned}&\mathbb{E}\left[\left\|\bar{y}_{\mathcal{S}_t}-\bar{x}_{\mathcal{S}_t}\frac{\bar{y}_{\mathcal{N}}}{\bar{x}_{\mathcal{N}}}\right\|^2\right]=\mathbb{E}\left[\left\|\bar{y}_{\mathcal{S}_t}-\bar{x}_{\mathcal{S}_t}\frac{\bar{y}_{\mathcal{N}}}{\bar{x}_{\mathcal{N}}}-\bar{y}_{\mathcal{N}}+\frac{\bar{y}_{\mathcal{N}}}{\bar{x}_{\mathcal{N}}}\bar{x}_{\mathcal{N}}\right\|^2\right]\\=&\mathbb{E}\left[\left(\left\|\bar{y}_{\mathcal{S}_t}-\bar{x}_{\mathcal{S}_t}\frac{\bar{y}_{\mathcal{N}}}{\bar{x}_{\mathcal{N}}}\right\|-\mathbb{E}\left[\left\|\bar{y}_{\mathcal{S}_t}-\bar{x}_{\mathcal{S}_t}\frac{\bar{y}_{\mathcal{N}}}{\bar{x}_{\mathcal{N}}}\right\|\right]\right)^2\right]\\=&\mathbb{V}\left[\left\|\bar{y}_{\mathcal{S}_t}-\bar{x}_{\mathcal{S}_t}\frac{\bar{y}_{\mathcal{N}}}{\bar{x}_{\mathcal{N}}}\right\|\right]\\=&\mathbb{V}\left[\frac{1}{K}\sum_{n\in\mathcal{N}}x_n^{(\mathrm{t})}\beta_n\left\|\nabla f_n(\mathbf{w}^{(\mathrm{t})})-\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\right\|\right]. \quad (67)\end{aligned}$$

By defining $\mathcal{A}_n\triangleq\left\|\beta_n\nabla f_n(\mathbf{w}^{(\mathrm{t})})-\beta_n\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\right\|$, the above equation can be transformed as follows:

$$\begin{aligned}&\mathbb{V}\left[\left\|\bar{y}_{\mathcal{S}_t}-\bar{x}_{\mathcal{S}_t}\frac{\bar{y}_{\mathcal{N}}}{\bar{x}_{\mathcal{N}}}\right\|\right]\\=&\frac{1}{K^2}\mathrm{Cov}\left(\sum_{n\in\mathcal{N}}x_n^{(\mathrm{t})}\mathcal{A}_n,\sum_{m\in\mathcal{N}}x_m^{(\mathrm{t})}\mathcal{A}_m\right)\\=&\frac{1}{K^2}\sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{N}}\mathcal{A}_n\mathcal{A}_m\mathrm{Cov}\left(x_n^{(\mathrm{t})}x_m^{(\mathrm{t})}\right) \quad (68)\\=&\frac{1}{K^2}\left[\sum_{n\in\mathcal{N}}\mathcal{A}_n^2\mathbb{V}(x_n^{(\mathrm{t})})+\sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{N},m\neq n}\mathcal{A}_n\mathcal{A}_m\mathrm{Cov}\left(x_n^{(\mathrm{t})}x_m^{(\mathrm{t})}\right)\right].\end{aligned}$$

The following equation can be obtained:

$$\mathbb{E}[(x_n^{(\mathrm{t})})^2]=\frac{K}{N}. \quad (69)$$

The variance of $x_n^{(\mathrm{t})}$ is given by

$$\mathbb{V}(x_n^{(\mathrm{t})})=\mathbb{E}\left[(x_n^{(\mathrm{t})})^2\right]-\left(\mathbb{E}\left[x_n^{(\mathrm{t})}\right]\right)^2=\frac{K}{N}\left(1-\frac{K}{N}\right). \quad (70)$$

Moreover, the probability of selecting device m after selecting device n is given by $\binom{N-1}{K-1}$, i.e.,

$$\mathbb{E}\left[x_n^{(\mathrm{t})}x_m^{(\mathrm{t})}\right]=\left(\frac{K-1}{N-1}\right)\left(\frac{K}{N}\right), \quad (71)$$

and then the covariance of $x_n^{(\mathrm{t})}x_m^{(\mathrm{t})}$ is given by

$$\begin{aligned}\mathrm{Cov}\left(x_n^{(\mathrm{t})}x_m^{(\mathrm{t})}\right)&=\mathbb{E}\left[x_n^{(\mathrm{t})}x_m^{(\mathrm{t})}\right]-\mathbb{E}\left[x_n^{(\mathrm{t})}\right]\mathbb{E}\left[x_m^{(\mathrm{t})}\right]\\&=-\frac{1}{N-1}\left(1-\frac{K}{N}\right)\left(\frac{K}{N}\right). \quad (72)\end{aligned}$$

Therefore, $\mathbb{E}\left[\|\mathbf{e}^{(\mathrm{t})}\|^2\right]$ can be rewritten as follows:

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{e}^{(\mathrm{t})}\|^2\right]\\=&\left(1-\frac{K}{N}\right)\frac{(N-1)\sum_{n\in\mathcal{N}}\mathcal{A}_n^2-\sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{N},m\neq n}\mathcal{A}_n\mathcal{A}_m}{KN(N-1)(\frac{1}{N}\sum_{n\in\mathcal{N}}\beta_n)^2}\\=&\left(1-\frac{K}{N}\right)\frac{(N-1)\sum_{n\in\mathcal{N}}\mathcal{A}_n^2-\left\|\sum_{n\in\mathcal{N}}\mathcal{A}_n\right\|^2+\sum_{n\in\mathcal{N}}\mathcal{A}_n^2}{KN(N-1)(\frac{1}{N}\sum_{n\in\mathcal{N}}\beta_n)^2}\\=&\left(1-\frac{K}{N}\right)\frac{N\sum_{n\in\mathcal{N}}\mathcal{A}_n^2-\left\|\sum_{n\in\mathcal{N}}\mathcal{A}_n\right\|^2}{KN(N-1)(\frac{1}{N}\sum_{n\in\mathcal{N}}\beta_n)^2}\\=&\left(1-\frac{K}{N}\right)\frac{\sum_{n\in\mathcal{N}}\beta_n^2\left\|\nabla f_n(\mathbf{w}^{(\mathrm{t})})-\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\right\|^2}{K(N-1)(\frac{1}{N}\sum_{n\in\mathcal{N}}\beta_n)^2}, \quad (73)\end{aligned}$$

and the proof is completed.

APPENDIX D: PROOF OF PROPOSITION 1

It is indicated by (26a) that the optimal resource allocation coefficients $\tau_{k,n}^*$ and $\alpha_{k,n}^*$ satisfy the following condition:

$$\kappa\mu\beta_n(\tau_{k,n}^*C_n)^2+\frac{\alpha_{k,n}^*P_nD}{B\log_2(1+\alpha_{k,n}^*P_n|h_{k,n}|^2)}\leq E_n^{\max}. \quad (74)$$

The following inequality can be obtained:

$$\frac{\alpha_{k,n}^*P_nD}{B\log_2(1+\alpha_{k,n}^*P_n|h_{k,n}|^2)}<E_n^{\max}. \quad (75)$$

Due to the fact that $\ln(1+\alpha_{k,n}^* P_n|h_{k,n}|^2) \leq \alpha_{k,n}^* P_n|h_{k,n}|^2$, the following inequality always holds:

$$\frac{\alpha_{k,n}^* P_n D}{B\log_2(1+\alpha_{k,n}^* P_n|h_{k,n}|^2)} \geq \frac{\ln(2)D}{B|h_{k,n}|^2}. \quad (76)$$

Therefore, the problem is infeasible if

$$\frac{\ln(2)D}{B|h_{k,n}|^2} \geq E_n^{\max}, \quad (77)$$

and the proof is completed.

Appendix E: Proof of Proposition 2

In resource allocation problem (30), device n is assigned to sub-channel k, and hence, $1-2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}} \neq 0$ and $-\tau_{k,n}^* \neq 0$ always hold. Based on (36e) and (36f), $\lambda_2 = 0$ and $\lambda_4 = 0$ can be obtained. As a result, (32) and (33) can be respectively rewritten as follows:

$$-\frac{\mu\beta_n}{(\tau_{k,n}^*)^2 C_n} + 2\lambda_1\kappa\mu\beta_n\tau_{k,n}^* C_n^2 + \lambda_3 = 0, \quad (78)$$

and

$$\lambda_1 B(T_{k,n}^{\mathrm{cm}*})^2(2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}-1) - \ln(2)\lambda_1 D T_{k,n}^{\mathrm{cm}*} 2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}} \quad (79)$$
$$-\ln(2)\lambda_5|h_{k,n}|^2 D 2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}} + |h_{k,n}|^2 B(T_{k,n}^{\mathrm{cm}*})^2 = 0.$$

Based on the above equations and the conditions in (36), four possible solutions can be presented.

1) If $\lambda_3 > 0$ and $\lambda_5 > 0$, the optimal solution can be obtained from (36e) and (36f) as

$$\tau_{k,n}^* = 1, \quad (80)$$

and

$$T_{k,n}^{\mathrm{cm}*} = \frac{D}{B\log_2(1+P_n|h_{k,n}|^2)}. \quad (81)$$

In this case, based on (36a), the following inequality holds:

$$\kappa\mu\beta_n C_n^2 + \frac{P_n D}{B\log_2(1+P_n|h_{k,n}|^2)} \leq E_n^{\max}. \quad (82)$$

2) If $\lambda_3 = 0$ and $\lambda_5 > 0$, (81) holds. From (78), the following condition can be obtained:

$$\lambda_1 = \frac{1}{2\kappa(\tau_{k,n}^*)^3 C_n^3} > 0. \quad (83)$$

Note that the above inequality always holds. In this case, the following equation is obtained from (36d):

$$\kappa\mu\beta_n(\tau_{k,n}^* C_n)^2 + \frac{P_n D}{B\log_2(1+P_n|h_{k,n}|^2)} - E_n^{\max} = 0, \quad (84)$$

and the expression of $\tau_{k,n}^*$ is given by

$$\tau_{k,n}^* = \frac{1}{C_n}\left[\frac{E_n^{\max}}{\kappa\mu\beta_n} - \frac{P_n D}{\kappa\mu\beta_n B\log_2(1+P_n|h_{k,n}|^2)}\right]^{\frac{1}{2}}. \quad (85)$$

From (36b), the following inequality should be satisfied:

$$0 < E_n^{\max} - \frac{P_n D}{B\log_2(1+P_n|h_{k,n}|^2)} \leq \kappa\mu\beta_n C_n^2, \quad (86)$$

where the inequality in (83) always holds under the above inequality. Moreover, $\lambda_5 > 0$ should be satisfied. By introducing $T_{k,n}^{\mathrm{cm}*}$ from (81), (79) can be transformed to

$$\begin{aligned}&\ln(2)\lambda_5|h_{k,n}|^2 B(1+P_n|h_{k,n}|^2)[\log_2(1+P_n|h_{k,n}|^2)]^2\\ =&D|h_{k,n}|^2 + \lambda_1 D P_n|h_{k,n}|^2\\ &-\ln(2)\lambda_1 D(1+P_n|h_{k,n}|^2)\log_2(1+P_n|h_{k,n}|^2). \end{aligned} \quad (87)$$

From (83) and (85), the above equation can be transformed to

$$\begin{aligned}\lambda_5 =& \frac{A_1|h_{k,n}|^2}{A_1\ln(2)|h_{k,n}|^2 B(1+P_n|h_{k,n}|^2)[\log_2(1+P_n|h_{k,n}|^2)]^2}\\ &+\frac{P_n|h_{k,n}|^2 - \ln(2)(1+P_n|h_{k,n}|^2)\log_2(1+P_n|h_{k,n}|^2)}{A_1\ln(2)|h_{k,n}|^2 B(1+P_n|h_{k,n}|^2)[\log_2(1+P_n|h_{k,n}|^2)]^2}, \end{aligned} \quad (88)$$

where A_1 is defined in (37). Inequality $\lambda_5 > 0$ can be transformed as follows:

$$A_1|h_{k,n}|^2 + P_n|h_{k,n}|^2 > (1+P_n|h_{k,n}|^2)\ln(1+P_n|h_{k,n}|^2). \quad (89)$$

3) If $\lambda_3 > 0$ and $\lambda_5 = 0$, (80) holds, and the following condition can be obtained from (79):

$$\lambda_1 = \frac{B|h_{k,n}|^2 T_{k,n}^{\mathrm{cm}*}}{\ln(2)D2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}} - BT_{k,n}^{\mathrm{cm}*}(2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}-1)} > 0. \quad (90)$$

The above inequality indicates that

$$\ln(2)D2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}} - BT_{k,n}^{\mathrm{cm}*}(2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}-1) > 0, \quad (91)$$

and it can be rewritten as follows:

$$(1+\alpha_{k,n}^* P_n|h_{k,n|^2})\ln(1+\alpha_{k,n}^* P_n|h_{k,n|^2}) > \alpha_{k,n}^* P_n|h_{k,n|^2}. \quad (92)$$

It is indicated by Taylor series that this condition always holds. From (36d), the following equation should be satisfied:

$$\kappa\mu\beta_n C_n^2 + T_{k,n}^{\mathrm{cm}*}\frac{2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}}-1}{|h_{k,n}|^2} - E_n^{\max} = 0. \quad (93)$$

By defining $y^* \triangleq 1/T_{k,n}^{\mathrm{cm}*}$ and A_2 as in (37), (93) can be rewritten as follows:

$$\begin{aligned}&2^{\frac{D}{B}y^*} = A_2 y^* + 1\\ \overset{(a)}{\Rightarrow}&(y^*+1/A_2)2^{-\frac{D}{B}y^*} = 1/A_2\\ \overset{(b)}{\Rightarrow}&\left(-y^* - \frac{1}{A_2}\right)(2^{\frac{D}{B}})^{(-y^*-\frac{1}{A_2})} = -\frac{(2^{\frac{D}{B}})^{-\frac{1}{A_2}}}{A_2}\\ \overset{(c)}{\Rightarrow}&\ln(2^{\frac{D}{B}})\left(-y^* - \frac{1}{A_2}\right)e^{\ln(2^{\frac{D}{B}})(-y^*-\frac{1}{A})} = -\frac{\ln(2^{\frac{D}{B}})(2^{\frac{D}{B}})^{-\frac{1}{A_2}}}{A_2}\\ \Rightarrow&\ln(2^{\frac{D}{B}})\left(-y^* - \frac{1}{A_2}\right)e^{\ln(2^{\frac{D}{B}})(-y^*-\frac{1}{A_2})} = -\frac{\ln(2^{\frac{D}{A_2 B}})}{2^{\frac{D}{A_2 B}}}, \end{aligned} \quad (94)$$

where in (a), (b) and (c), the equation is multiplied by $2^{-\frac{D}{B}y^*}/A_2$, $-(2^{\frac{D}{B}})^{-\frac{1}{A_2}}$ and $\ln(2^{\frac{D}{B}})$, respectively. At this stage, the Lambert W function can be utilized to obtain the closed-form solution. By defining $x \triangleq 2^{\frac{D}{A_2 B}}$, where $x > 0$, the right side of the above equation can be written as $-\ln(x)/x$, and the minimum value, i.e., $-1/e$, can be determined through

derivatives. Since $-1/e \leq -\ln(2^{\frac{D}{A_2B}})/(2^{\frac{D}{A_2B}}) < 0$, there exist two real value solutions, as shown in follows:

$$\begin{cases} \ln(2^{\frac{D}{B}})(-y^*-1/A_2) = W_0\left(-\ln(2^{\frac{D}{A_2B}})/2^{\frac{D}{A_2B}}\right), \\ \ln(2^{\frac{D}{B}})(-y^*-1/A_2) = W_{-1}\left(-\ln(2^{\frac{D}{A_2B}})/2^{\frac{D}{A_2B}}\right). \end{cases} \tag{95}$$

The proof below shows that only one solution is feasible. Since that $-1/e \leq -\ln(2^{\frac{D}{A_2B}})/(2^{\frac{D}{A_2B}}) < 0$, it can be obtained that $-1 \leq W_0(-\ln(2^{\frac{D}{A_2B}})/2^{\frac{D}{A_2B}}) < 0$, and then, the following inequality holds:

$$-1 \leq \ln(2^{\frac{D}{B}})\left(-y^* - \frac{1}{A_2}\right) < 0. \tag{96}$$

Depending on the relationship between A_2 and $\ln(2^{\frac{D}{B}})$, different inequalities can be obtained, as shown in follows:

$$\begin{cases} T_{k,n}^{\mathrm{cm}*} \geq \frac{A_2 \ln(2^{\frac{D}{B}})}{A_2 - \ln(2^{\frac{D}{B}})}, & \text{if} \quad A_2 > \ln(2^{\frac{D}{B}}), \\ T_{k,n}^{\mathrm{cm}*} \leq \frac{A_2 \ln(2^{\frac{D}{B}})}{A_2 - \ln(2^{\frac{D}{B}})}, & \text{if} \quad A_2 < \ln(2^{\frac{D}{B}}). \end{cases} \tag{97}$$

From (90), the following condition should be satisfied:

$$\ln(2) D 2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}} - BT_{k,n}^{\mathrm{cm}*}(2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}} - 1) > 0. \tag{98}$$

Based on (93), the above inequality can be rewritten as

$$\ln(2) D 2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}} - A_2 B > 0 \Rightarrow \frac{D}{B \log_2(A_2/\ln(2^{\frac{D}{B}}))} > T_{k,n}^{\mathrm{cm}*}. \tag{99}$$

By combining the first inequality of (97) and (99), the following inequality can be obtained:

$$\frac{D}{B \log_2(A_2/\ln(2^{\frac{D}{B}}))} > \frac{A_2 \ln(2^{\frac{D}{B}})}{A_2 - \ln(2^{\frac{D}{B}})}, \tag{100}$$

and this inequality can be transformed to

$$\ln(2^{\frac{D}{B}})/A_2 - 1 < \ln\left(\ln(2^{\frac{D}{B}})/A_2\right). \tag{101}$$

Since $\ln(2^{\frac{D}{B}})/A_2 > 0$, the above inequality cannot hold. That is, when $A_2 > \ln(2^{\frac{D}{B}})$, this solution contradicts condition $\lambda_1 > 0$. Moreover, the second inequality of (97) also cannot hold since $T_{k,n}^{\mathrm{cm}*}$ is greater than zero. As a result, the solution based on $W_0(-\ln(2^{\frac{D}{A_2B}})/2^{\frac{D}{A_2B}})$ can be removed, and the closed-form expression of $T_{k,n}^{\mathrm{cm}*}$ is given by

$$T_{k,n}^{\mathrm{cm}*} = -\frac{A_2 \ln(2^{\frac{D}{B}})}{A_2 W_{-1}(-\ln(2^{\frac{D}{A_2B}})/2^{\frac{D}{A_2B}}) + \ln(2^{\frac{D}{B}})}. \tag{102}$$

Based on (36c), the value range of $T_{k,n}^{\mathrm{cm}*}$ is given by

$$T_{k,n}^{\mathrm{cm}*} \geq \frac{D}{B \log_2(1 + P_n |h_{k,n}|^2)}. \tag{103}$$

Moreover, since $\lambda_3 > 0$, from (78), the following condition is included:

$$1 - 2\lambda_1 \kappa C_n^3 > 0. \tag{104}$$

From (90), this condition can be transformed to

$$\ln(2) D 2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}} - A_2 B > 2\kappa C_n^3 B |h_{k,n}|^2 T_{k,n}^{\mathrm{cm}*}. \tag{105}$$

4) If $\lambda_3 = 0$ and $\lambda_5 = 0$, (83), (90) and (36d) hold. In this case, the following equations can be obtained:

$$\begin{cases} 2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}} - 1 - \frac{\ln(2) D}{BT_{k,n}^{\mathrm{cm}*}} 2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}} + 2\kappa (\tau_{k,n}^*)^3 C_n^3 |h_{k,n}|^2 = 0, \\ \kappa \mu \beta_n (\tau_{k,n}^* C_n)^2 + T_{k,n}^{\mathrm{cm}*} \frac{2^{\frac{D}{BT_{k,n}^{\mathrm{cm}*}}} - 1}{|h_{k,n}|^2} - E_n^{\max} = 0. \end{cases} \tag{106}$$

The expressions of $\alpha_{k,n}^*$ and $T_{k,n}^{\mathrm{cm}*}$ cannot be directly presented, but the above functions can be simply solved by utilizing numerical solvers. In this case, the value ranges of $\alpha_{k,n}^*$ and $T_{k,n}^{\mathrm{cm}*}$ should be included. This proposition is proved.

REFERENCES

[1] Z. Yang, M. Chen, K.-K. Wong, H. V. Poor, and S. Cui, "Federated learning for 6G: Applications, challenges, and opportunities," *Engineering*, vol. 8, pp. 33–41, 2022.

[2] 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.

[3] 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.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

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

[7] 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.

[8] 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.

[9] Z. Yang, M. Chen, W. Saad, C. S. Hong, M. Shikh-Bahaei, H. V. Poor, and S. Cui, "Delay minimization for federated learning over wireless communication networks," *arXiv preprint arXiv:2007.03462*, 2020.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] S. Mao, L. Liu, N. Zhang, J. Hu, K. Yang, F. R. Yu, and V. C. M. Leung, "Intelligent reflecting surface-assisted low-latency federated learning over wireless networks," *IEEE Internet Things J.*, vol. 10, no. 2, pp. 1223–1235, 2023.

[15] J. Ren, W. Ni, and H. Tian, "Toward communication-learning trade-off for federated learning at the network edge," *IEEE Commun. Lett.*, vol. 26, no. 8, pp. 1858–1862, 2022.

[16] S. S. Shinde, A. Bozorgchenani, D. Tarchi, and Q. Ni, "On the design of federated learning in latency and energy constrained computation offloading operations in vehicular edge computing systems," *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 2041–2057, 2022.

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

[18] 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.

[19] Y. J. Cho, J. Wang, and G. Joshi, "Client selection in federated learning: Convergence analysis and power-of-choice selection strategies," *arXiv preprint arXiv:2010.01243*, 2020.

[20] S. R. Pandey, L. D. Nguyen, and P. Popovski, "A contribution-based device selection scheme in federated learning," *IEEE Commun. Lett.*, vol. 26, no. 9, pp. 2057–2061, 2022.

[21] H. Wu and P. Wang, "Node selection toward faster convergence for federated learning on non-IID data," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 5, pp. 3099–3111, 2022.

[22] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing federated learning on non-IID data with reinforcement learning," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, 2020, pp. 1698–1707.

[23] J. Xu and H. Wang, "Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1188–1200, 2021.

[24] C. Castera, J. Bolte, C. Févotte, and E. Pauwels, "Second-order step-size tuning of sgd for non-convex optimization," *Neural Processing Letters*, vol. 54, no. 3, pp. 1727–1752, 2022.

[25] 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.

[26] 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.

[27] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[28] 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.

[29] K. Wang, Y. Ma, M. B. Mashhadi, C. H. Foh, R. Tafazolli, and Z. Ding, "Age of information in federated learning over wireless networks," *arXiv preprint arXiv:2209.06623*, 2022.

[30] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms I: Fundamentals*. Springer science & business media, 2013, vol. 305.

[31] B. T. Polyak, "Gradient methods for the minimisation of functionals," *USSR Computational Mathematics and Mathematical Physics*, vol. 3, no. 4, pp. 864–878, 1963.

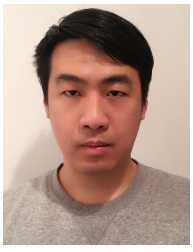
[32] X. Peng, L. Li, and F.-Y. Wang, "Accelerating minibatch stochastic gradient descent using typicality sampling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4649–4659, 2020.

[33] S. L. Lohr, *Sampling: design and analysis*. CRC press, 2021.

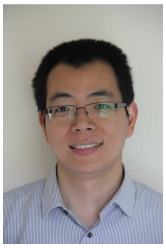
[34] R. D. Yates, "The age of information in networks: Moments, distributions, and sampling," *IEEE Trans. Inf. Theory*, vol. 66, no. 9, pp. 5712–5728, 2020.

[35] 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.

[36] Y. Nesterov, "Introductory lectures on convex programming," 1998.



Kaidi Wang (S'16-M'20) 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, mobile edge computing, and federated learning.



Zhiguo Ding (S'03-M'05-F'20) received his B.Eng from the Beijing University of Posts and Telecommunications in 2000, and the Ph.D degree from Imperial College London in 2005. He is currently a Professor in Communications at Khalifa University, and has also been affiliated with the University of Manchester and Princeton University.

Dr. Ding's research interests are 6G networks, multiple access, energy harvesting networks and statistical signal processing. He is serving as an Area Editor for the *IEEE Transactions on Wireless Communications*, and *IEEE Open Journal of the Communications Society*, an Editor for *IEEE Transactions on Vehicular Technology*, and was an Editor for *IEEE Wireless Communication Letters*, *IEEE Transactions on Communications*, *IEEE Communication Letters* from 2013 to 2016. He recently received the EU Marie Curie Fellowship 2012-2014, the Top IEEE TVT Editor 2017, IEEE Heinrich Hertz Award 2018, IEEE Jack Neubauer Memorial Award 2018, IEEE Best Signal Processing Letter Award 2018, Friedrich Wilhelm Bessel Research Award 2020, and IEEE SPCC Technical Recognition Award 2021. He is a Fellow of the IEEE, a Distinguished Lecturer of IEEE ComSoc, and a Web of Science Highly Cited Researcher in two categories 2022.

Daniel K. C. So (S'96-M'03-SM'14) received the BEng (Hons) degree in Electrical and Electronics Engineering from the University of Auckland, New Zealand, and the PhD degree in Electrical and Electronics Engineering from the Hong Kong University of Science and Technology (HKUST). He joined the University of Manchester in 2003 and is now a Professor. He is the Discipline Head of Education and Deputy Head of Department in the Department of Electrical & Electronic Engineering.

His research interests include green communications, NOMA, beyond 5G and 6G networks, machine learning and federated learning, RIS, heterogeneous networks, SWIPT, and massive MIMO. He is currently serving as a Senior Editor of IEEE Wireless Communication Letters after being an Editor from 2016-2020. He served as an Editor of IEEE Transactions on Wireless Communications between 2017-2023. He is the Lead Guest Editor for a special issue in IEEE Transactions on Green Communications and Networking. He also served as a symposium co-chair of IEEE ICC 2019 and 2025, and Globecom 2020, and track co-chair for IEEE Vehicular Technology Conference (VTC) Spring 2016, 2017, 2018, 2021 and 2022, and VTC Fall 2023. He is also the chair of the Special Interest Group on Green Cellular Networks within the IEEE ComSoc Green Communications and Computing Technical Committee since 2020.

Zhi Ding (S'88-M'90-SM'95-F'03) 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.