Exploring Age-of-Information Weighting in Federated Learning under Data Heterogeneity

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

***Abstract*—This paper investigates wireless federated learning in data heterogeneous scenarios, where device selection usually leads to a degradation in learning performance. This paper is motivated by the fact that while training deep learning networks using federated stochastic gradient descent (FedSGD) on non-independent and identically distributed (non-IID) datasets, device selection can generate gradient errors that accumulate, leading to potential weight divergence, which is further exacerbated with low device participation. To mitigate weight divergence, an age-weighted FedSGD algorithm is designed in this paper to scale local gradients according to the previous device selection results. Furthermore, by revealing the relationship between device participation and latency, an energy consumption minimization problem is formulated accordingly, which consists of resource allocation and sub-channel assignment. By transforming the resource allocation problem into convex and utilizing KKT conditions, we derive the optimal resource allocation solution. Moreover, this paper develops a matching based algorithm to generate the enhanced sub-channel assignment. Simulation results indicate that i) age-weighted FedSGD is able to outperform conventional FedSGD in terms of convergence rate and achievable accuracy, and ii) the proposed resource allocation and sub-channel assignment strategies can significantly reduce energy consumption and improve learning performance by increasing device participation.**

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

I. Introduction

With the spread of computer chips, powerful computational capabilities become available at edge nodes, and therefore, the collected data can be directly utilized for learning tasks [2]. In this context, federated learning, as a promising technology for distributed learning, has attracted considerable attention from both academia and industry. In federated learning, a neural network is constructed by a central server and shared among all participating devices [3]. At each device, the received neural network is trained with local data and transmitted to the server for aggregation [4]. Compared with centralized learning that requires offloading raw data to the server, in federated learning, learning tasks are executed collaboratively without data sharing, and hence, privacy security can be improved [5]. Furthermore, since the size of the transmitted neural network is generally smaller than the size of original data, communication efficiency can be achieved [6]. However, since federated learning relies on periodic transmission, its performance is affected by wireless networks, and hence, the optimization of communications is recognized as an important research direction [7].

Due to the fact that federated learning usually involves a large number of devices for multiple rounds of training and transmission, device selection/sampling becomes a common method to implement this algorithm under limited bandwidth resources [8], [9]. Some existing works focused on addressing system heterogeneity by selecting devices based on hardware specifications and communication environment [10]–[13]. In [10], device selection and beamforming design were jointly considered in an over-the-air computation (AirComp) based federated learning framework, where an optimization problem was formulated to maximize the number of selected devices. By revealing the interaction between global loss and packet error rates, device selection was included to cope with the limited number of resource blocks [11]. Particularly, in this work, a device can be selected only if the latency and energy consumption constraints can be satisfied. Since the transmitted models can be severely damaged by noise in AirComp based federated learning, in [12], devices with weak channel conditions were ignored for aggregation as the transmit power is not sufficient to compensate for the effects of wireless communications. Recognizing the degradation of learning performance caused by low device availability, the authors of [13] proposed a device selection strategy based on achievable long-term participation rates to mitigate the impact of device selection variance on global model convergence.

In realistic scenarios of federated learning, non-independent and identically distributed (non-IID) data is unevenly distributed among devices, which brings challenges to device selection [14], [15]. Specifically, in system based device selection, the server tends to select devices with better channel conditions and/or powerful computational capacities, which may lead to a decline in learning performance on non-IID datasets [16]. To this end, by selecting devices that provide more contributions in the aggregation, some works jointly considered system heterogeneity and data heterogeneity [17]–[20]. In [17], channel conditions and local model updates

This work was supported by the UK EPSRC under grant number EP/P009719/2 and by H2020-MSCA-RISE-2020 under grant number 101006411.This material was 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 the Department of Electronic and Electrical Engineering, University of Manchester, UK, and the Department of Computer and Information Engineering, Khalifa University, UAE (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).

Part of this work has been published in IEEE International Conference on Communications Workshops, Denver, CO, USA, 2024 [1].

were studied, and four device selection polices were proposed based on different priorities. Simulation results demonstrated that jointly including both metrics can provide better learning performance than using either metric separately. In order to achieve the target global loss within less time consumption, the selection probabilities of devices in the classic random device selection scheme were optimized based on latency and gradient norms [18]. Considering that the device contribution is not only related to dataset size, a biased device selection scheme was developed in [19], in which the server transmits the global model to a set of candidate devices for evaluation, and then selects devices with larger local losses. In [20], age-of-information (AoI) was considered as a metric to improve the fairness of device selection. It was indicated that by minimizing the overall AoI of all devices, both learning performance and time consumption can be improved.

Since learning based device selection is performed on a set of available devices, its performance can be further improved by increasing device participation, which is determined by channel conditions, computational capacities, battery levels, etc. [13], [21]. Therefore, it is necessary to explore the optimal resource allocation based on these factors. Energy consumption, as an important criterion that can directly limit device participation, has been extensively researched in existing works [22]–[26]. In [22], a comprehensive energy consumption minimization problem was investigated in federated learning systems, where monotonicity analysis was utilized to obtain solutions. Wireless federated learning was also studied in eavesdropping scenarios, in which idle devices transmit jamming signals to improve the secrecy rate of the transmitting device [23]. In [24], energy harvesting and non-orthogonal multiple access (NOMA) were exploited to provide computing energy and facilitate uplink transmission, respectively. In these works, the bisection method was utilized for algorithm design [22]–[24]. The authors of [25] focused on studying long-term energy consumption minimization, where deep reinforcement learning was employed. In [26], NOMA schemes were adopted in a clustered federated learning system, where sub-channel assignment and power allocation were studied to further enhance device participation.

As aforementioned, with non-IID data, system based device selection leads to a decline in learning performance [10]–[13], while learning based device selection requires additional transmission and analysis for local models or gradients [17]–[19]. A novel method, namely age-weighted FedSGD, is proposed to mitigate the learning performance degradation caused by implementing device selection on non-IID datasets. This scheme can be employed in a variety of existing device selection strategies without extra information transmission and model/data analysis, and hence, it will not increase system overhead or cause privacy leakage. Different from existing studies [20], [27]–[29] that utilized AoI to guide device selection in federated learning, this paper aims to explore the role of AoI in federated learning and leverage it to mitigate the negative impact of adopting device selection in data heterogeneous scenarios. Moreover, to further improve learning performance by increasing device participation, an energy consumption minimization problem is jointly addressed through a low-complexity solution, thus avoiding the loss of optimality in previous works [22]–[24].

The main contributions can be summarized as follows:

- A wireless federated learning network with random device selection is investigated. It is proved that in conventional federated stochastic gradient descent (FedSGD), device selection with non-IID data results in an error in global gradients, which is accumulated and amplified during training, thereby increasing weight divergence.
- Based on the analyzed result, AoI is introduced to design age-weighted FedSGD, which can adjust the proportion of local gradients from selected devices in the global gradient. Moreover, it is indicated that low device participation can negatively affect weight divergence by changing the data distribution of the selected devices and lead to a decrease in convergence rate.
- To further mitigate weight divergence, an energy consumption minimization problem is formulated to increase device participation through enabling more devices to satisfy latency constraints. By decoupling the problem into two subproblems, KKT conditions and matching theory are utilized to develop the closed-form resource allocation solution and sub-channel assignment algorithm, respectively.
- Simulation results show that the proposed age-weighted FedSGD can significantly improve the performance of federated learning in the considered system, including convergence rate and achievable test accuracy. Moreover, KKT based resource allocation and matching based sub-channel assignment are able to minimize energy consumption and increase device participation.

II. System Model

Consider a wireless communication scenario for non-IID federated learning, where a server and N devices collaborate to execute a learning task through K sub-channels. All nodes are equipped with single-antennas. The collections of devices and sub-channels are represented by $\mathcal{N} = \{1, 2, \cdots, N\}$ and $\mathcal{K} = \{1, 2, \cdots, K\}$, respectively. It is assumed that the number of available sub-channels is less than the number of devices, and thus a subset of devices is randomly selected[1] to participate in the aggregation in each communication round, denoted by $\mathcal{S}_t$, where $|\mathcal{S}_t| \leq K < N$. In the considered federated learning algorithm, the local loss is given by

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

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

$$F(\mathbf{w}^{(\mathrm{t})}, \mathcal{S}_t) = \frac{\sum_{n \in \mathcal{S}_t} \beta_n f_n(\mathbf{w}^{(\mathrm{t})})}{\sum_{n \in \mathcal{S}_t} \beta_n}. \tag{2}$$

[1]Note that although this work considers classic random device selection, the proposed method can be utilized with multiple existing advanced device selection strategies.

The procedure of the considered federated learning algorithm follows FedSGD. In any communication round t, the following process is performed:

1) The server transmits global model $\mathbf{w}^{(\mathrm{t})}$ and device selection decision $\mathcal{S}_t$ to all devices.
2) Any device $n \in \mathcal{S}_t$ trains the received global model using all local samples, and transmits local gradients $\nabla f_n(\mathbf{w}^{(\mathrm{t})})$ to the server.
3) The server updates the global model as follows:

$$\begin{aligned}\mathbf{w}^{(\mathrm{t}+1)} &= \mathbf{w}^{(\mathrm{t})} - \lambda \frac{\sum_{n\in\mathcal{S}_t} \beta_n \nabla f_n(\mathbf{w}^{(\mathrm{t})})}{\sum_{n\in\mathcal{S}_t} \beta_n} \\ &= \mathbf{w}^{(\mathrm{t})} - \lambda \nabla F(\mathbf{w}^{(\mathrm{t})}, \mathcal{S}_t), \end{aligned} \tag{3}$$

where λ is the learning rate.

In this paper, the following assumptions are considered.

Assumption 1. *With respect to* $\mathbf{w}$, $\nabla F(\mathbf{w}, \mathcal{N})$ *is L-Lipschitz continuous, i.e.,*

$$\|\nabla F(\mathbf{w}^{(\mathrm{t}-1)}, \mathcal{N}) - \nabla F(\mathbf{w}^{(\mathrm{t})}, \mathcal{N})\| \leq L \|\mathbf{w}^{(\mathrm{t}-1)} - \mathbf{w}^{(\mathrm{t})}\|. \tag{4}$$

Assumption 2. *The global loss function* $F(\mathbf{w}^{(\mathrm{t})}, \mathcal{N})$ *satisfies the Polyak-Lojasiewicz inequality with positive parameter* μ*, as shown in follows:*

$$\|\nabla F(\mathbf{w}^{(\mathrm{t})}, \mathcal{N})\|^2 \geq 2\mu \left[F(\mathbf{w}^{(\mathrm{t})}, \mathcal{N}) - F(\mathbf{w}^{*}, \mathcal{N}) \right]. \tag{5}$$

It is worth noting that these assumptions can be satisfied by commonly adopted loss functions, and have been extensively considered in existing works on federated learning, such as [11], [12], [16]–[19].

A. Weight Divergence in Conventional FedSGD

Since device selection is implemented in a federated learning algorithm using non-IID datasets, the data distribution of the selected devices may be different from the global data distribution, and therefore, the weight divergence issue may occur [16], [20]. That is, the divergence between the weights obtained from the considered federated learning framework and the desired weights obtained from centralized learning increases with training. To evaluate weight divergence, complete device selection is included as the baseline, where all devices are selected in each communication round. The update of the true global model[2] in this case is given by

$$\mathbf{w}_{\mathrm{T}}^{(\mathrm{t}+1)} = \mathbf{w}_{\mathrm{T}}^{(\mathrm{t})} - \lambda \nabla F(\mathbf{w}_{\mathrm{T}}^{(\mathrm{t})}, \mathcal{N}), \tag{6}$$

where

$$F(\mathbf{w}_{\mathrm{T}}^{(\mathrm{t})}, \mathcal{N}) = \frac{\sum_{n\in\mathcal{N}} \beta_n f_n(\mathbf{w}_{\mathrm{T}}^{(\mathrm{t})})}{\sum_{n\in\mathcal{N}} \beta_n}. \tag{7}$$

Note that the complete device selection scheme can be treated as centralized learning, since in the considered federated learning algorithm, all local data is utilized for training and the number of local epochs is one. Based on the definition of the true global model, the following theorem can be obtained.

Theorem 1. *Defining the error caused by device selection as the difference in the global loss gradient between random device selection and complete device selection, i.e.,*

$$\mathbf{e}^{(\mathrm{t})} \triangleq \nabla F(\mathbf{w}^{(\mathrm{t})}, \mathcal{S}_t) - \nabla F(\mathbf{w}^{(\mathrm{t})}, \mathcal{N}), \tag{8}$$

the weight divergence in the considered federated learning framework is bounded by:

$$\begin{aligned}\left\| \mathbf{w}^{(\mathrm{t}+1)} - \mathbf{w}_{\mathrm{T}}^{(\mathrm{t}+1)} \right\| \leq (1+\lambda L)^t \left\| \mathbf{w}^{(1)} - \mathbf{w}_{\mathrm{T}}^{(1)} \right\| + \lambda \left\| \sum_{i=1}^{t} \mathbf{e}^{(i)} \right\| \\ + \lambda^2 L \sum_{j=1}^{t-1} (1+\lambda L)^{j-1} \left\| \sum_{i=1}^{t-j} \mathbf{e}^{(i)} \right\|. \end{aligned} \tag{9}$$

Proof: Refer to Appendix A. ∎

Theorem 1 indicates that in the considered federated learning framework, device selection leads to weight divergence, and this effect can be described as the error of the global loss gradient in each communication round. Based on Theorem 1, the following remarks can be obtained.

Remark 1. *By introducing error* $\mathbf{e}^{(\mathrm{t})}$*, the update of the global model in the considered federated learning algorithm can be viewed as complete device selection with the error, as follows:*

$$\mathbf{w}^{(\mathrm{t}+1)} = \mathbf{w}^{(\mathrm{t})} - \lambda \nabla F(\mathbf{w}^{(\mathrm{t})}, \mathcal{N}) - \lambda \mathbf{e}^{(\mathrm{t})}. \tag{10}$$

Remark 2. *In the considered federated learning algorithm, the weight divergence is mainly caused by two parts, including the difference between initial global models, i.e.,* $\|\mathbf{w}^{(1)} - \mathbf{w}_{\mathrm{T}}^{(1)}\|$*, and the accumulated error, i.e.,* $\|\sum_i \mathbf{e}^{(i)}\|$.

Remark 3. *The impact of the accumulated error from previous rounds is amplified with training, since* $1 + \lambda L > 1$*. That is, the impact of errors in the early stages of training plays a major role in weight divergence.*

Remark 4. *When utilizing different initial global models, even if complete device selection is applied, i.e., the error is zero, large weight divergence may still be encountered.*

According to Theorem 1, the influence of the accumulated errors until any round is amplified in subsequent training, which implies that the device selection results have different impacts depending on communication rounds. However, in conventional FedSGD, this difference is not reflected. It is also indicated by Theorem 1 that the weight divergence can be mitigated by reducing $\|\sum_{i=1}^{t} \mathbf{e}^{(i)}\|$. To this end, conventional methods, such as importance sampling, focus on reducing $\|\mathbf{e}^{(\mathrm{t})}\|$ in each communication round, which requires analyzing local gradients and is therefore difficult to implement in realistic scenarios due to high-complexity or privacy issues [9], [30]. Inspired by the fact that the errors are accumulated, this paper introduces a weighting factor[3] to scale the error of the current round according to the accumulated error from previous rounds, thereby reducing weight divergence.

[2]It is worth emphasizing that the true global model, i.e., $\mathbf{w}_{\mathrm{T}}^{(\mathrm{t})}$, is introduced as a comparison with the actual global model obtained in the considered federated learning algorithm, i.e., $\mathbf{w}^{(\mathrm{t})}$. Since the true global model is not available in practical scenarios, it is only utilized in this subsection to facilitate analysis.

[3]In this paper, the terms "weighting factor" and "weights" refer to the local gradient adjustment coefficient and neural network parameters, respectively.

Specifically, by treating $\mathbf{e}^{(\mathrm{t})}$ and $\sum_{i=1}^{t-1}\mathbf{e}^{(i)}$ as two vectors, if the elements in $\mathbf{e}^{(\mathrm{t})}$ are close to the corresponding elements in $\sum_{i=1}^{t-1}\mathbf{e}^{(i)}$, a small weighting factor is adopted to reduce these elements; otherwise, these elements are amplified. As a result, the accumulated error $\sum_{i=1}^{t-1}\mathbf{e}^{(i)}$ can be compensated by using $\mathbf{e}^{(\mathrm{t})}$, and the value of $\|\sum_{i=1}^{t}\mathbf{e}^{(i)}\|$ can be reduced.

B. Age-weighted FedSGD

Based on the definition of error in (8), the weighting factor should be applied to $\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)$, since $\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})$ is unknown in practical training. In other words, the weighting factor should be designed according to the difference in device selection between communication rounds. In particular, in the considered system, some devices may need to wait several rounds before participating in the aggregation. In this case, AoI is introduced to record recent device selection status and generate weighting factors [27], [28]. For device n, its AoI in round t is defined as follows:

$$A_n^{(\mathrm{t})}=\begin{cases}1, & \text{if} \quad n\in\mathcal{S}_{t-1},\\ A_n^{(\mathrm{t}-1)}+1, & \text{if} \quad n\notin\mathcal{S}_{t-1}.\end{cases} \tag{11}$$

The above equation indicates that if device n is selected in last round, its AoI becomes 1; otherwise, it increases by 1. Based on this definition, the age-weighted FedSGD is proposed with the following weighting factor[4]:

$$\omega_n^{(\mathrm{t})}=\frac{A_n^{(\mathrm{t})}|\mathcal{S}_t|}{\sum_{i\in\mathcal{S}_t}A_i^{(\mathrm{t})}}, \tag{12}$$

where $|\mathcal{S}_t|$ is included for normalization. At the server, the global model is updated based on the age-weighted local gradients, as follows:

$$\mathbf{w}^{(\mathrm{t}+1)}=\mathbf{w}^{(\mathrm{t})}-\lambda\nabla G(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t), \tag{13}$$

where

$$G(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)=\frac{\sum_{n\in\mathcal{S}_t}\omega_n^{(\mathrm{t})}\beta_n f_n(\mathbf{w}^{(\mathrm{t})})}{\sum_{n\in\mathcal{S}_t}\beta_n}. \tag{14}$$

Note that the AoI of all devices can be counted at the server, and thus the proposed scheme does not require additional information transmission. Moreover, by including the AoI based weighting factor in the updates of local models, the proposed approach can also be utilized in federated averaging (FedAvg), where the AoI of all devices can be transmitted together with the global model.

As aforementioned, in non-IID scenarios, device selection results lead to various issues depending on the communication rounds, and age-weighted FedSGD exploits AoI to mitigate such differences. Specifically, an AoI based weighting factor $\omega_n^{(\mathrm{t})}$ is incorporated to scale the influence of devices based on the previous device selection results. For example, since the impact of $\mathbf{e}^{(\mathrm{t}-1)}$ is amplified in round t, a large weighting factor is added to device n, where $n\in\mathcal{S}_t\cap\{\mathcal{N}\backslash\mathcal{S}_{t-1}\}$. For the same reason, a small weighting factor is added to device n, where $n\in\mathcal{S}_t\cap\mathcal{S}_{t-1}$. Fig. 1 further explains age-weighted FedSGD, where 2 devices are selected from 4 in each round. Compared to conventional FedSGD, global model $\mathbf{w}^{(\mathrm{t}+2)}$ in age-weighted FedSGD is closer to the local model of device C, because the gradients used in the model update are weighted according to the device selection result in round $t+1$. Similarly, in round $t+3$, device D dominates the aggregation. As a result, the distance between $\mathbf{w}^{(\mathrm{t}+3)}$ and true global model $\mathbf{w}_{\mathrm{T}}^{(\mathrm{t}+2)}$ is reduced by utilizing age-weighted FedSGD.

[4]The AoI based weighting factor presented in (12) is for simplicity. It is observed that learning performance is highly sensitive to AoI, and other expressions of the weighting factor may provide further improvements.

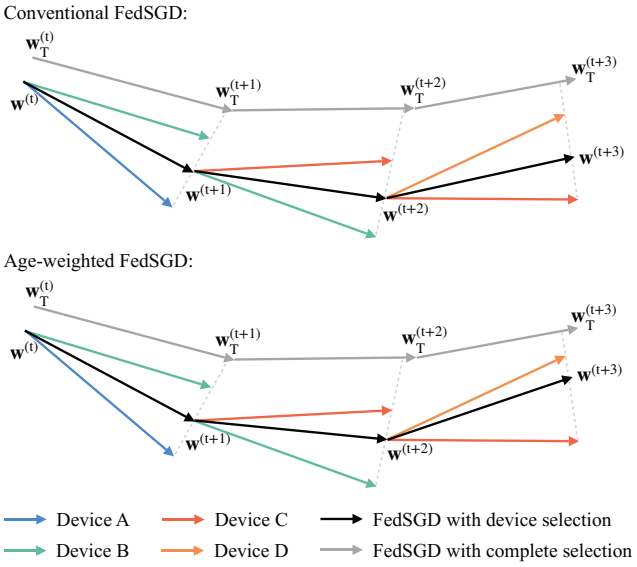


Fig. 1: An illustration of weight divergence for federated learning with conventional FedSGD and age-weighted FedSGD.

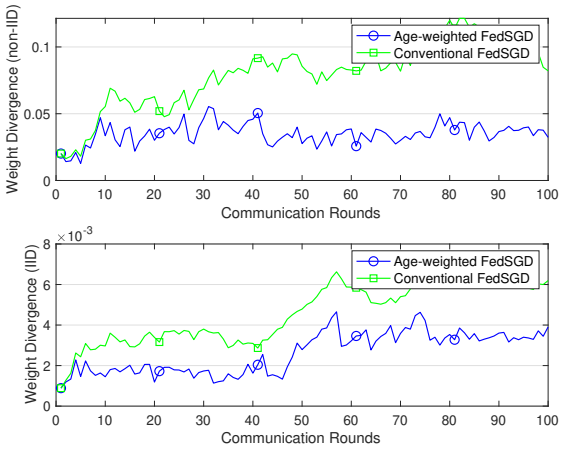


Fig. 2: An empirical result to validate the weight divergence issue on the balanced MNIST dataset. $N=10$ and $K=5$.

The performance of age-weighted FedSGD and conventional FedSGD is compared in Fig. 2, where weight divergence is calculated by $\|\mathbf{w}^{(\mathrm{t})}-\mathbf{w}_{\mathrm{T}}^{(\mathrm{t})}\|$. It can be observed that on non-IID datasets, weight divergence in conventional FedSGD increases with training, which confirms Theorem 1. As mentioned in Remark 3, the impact of error is amplified with training, leading to increasingly serious weight divergence. Therefore, the resulting performance degradation cannot be overcome by increasing the number of training rounds. It

is also indicated that age-weighted FedSGD can efficiently reduce weight divergence and control it to a certain level. Furthermore, it is worth pointing out that age-weighted FedSGD is still valid with IID data, although the weight divergence issue is not severe in this case.

C. Impact of Device Participation

This subsection focuses on the impact of device participation on the considered federated learning algorithm. Since weight divergence is caused by the changes in data distribution [16], a task specification is required for analysis. In this subsection, the commonly considered multi-class classification problem is studied. For other tasks, the same conclusion can be obtained in a similar way. Consider a C-class classification problem with compact space $\mathcal{X}$ and label space $\mathcal{Y}=\mathcal{C}$, where $\mathcal{C}=\{1,2,\ldots,C\}$. The data distribution of device n is defined as follows:

$$P_n=\left[\frac{\sum_{i=1}^{\beta_n}\mathbb{1}_{y_{n,i}=c}}{\beta_n}\middle|\forall c\in\mathcal{C}\right], \tag{15}$$

By adopting the cross-entropy loss, the local loss is given by

$$\begin{aligned} f_n(\mathbf{w}^{(\mathrm{t})}) &= \mathbb{E}_{\mathbf{x},y\sim P_n}\left[\sum_{c\in\mathcal{C}}\mathbb{1}_{y=c}\log f_c(\boldsymbol{x},\mathbf{w}^{(\mathrm{t})})\right] \\ &= \sum_{c\in\mathcal{C}}P_n(y=c)\mathbb{E}_{\boldsymbol{x}|y=c}\left[\log f_c(\boldsymbol{x},\mathbf{w}^{(\mathrm{t})})\right], \end{aligned} \tag{16}$$

where $f_c(\boldsymbol{x},\mathbf{w}^{(\mathrm{t})})$ indicates the probability for class c, and the sample $(\boldsymbol{x},y)$ follows data distribution P_n. In this case, by defining $P_{\mathcal{S}_t}$ and $P_{\mathcal{N}}$ as the data distributions of the selected devices and all devices, i.e.,

$$P_{\mathcal{S}_t}=\left[\frac{\sum_{n\in\mathcal{S}_t}\sum_{i=1}^{\beta_n}\mathbb{1}_{y_{n,i}=c}}{\sum_{n\in\mathcal{S}_t}\beta_n}\middle|\forall c\in\mathcal{C}\right]=\frac{\sum_{n\in\mathcal{S}_t}\beta_nP_n}{\sum_{n\in\mathcal{S}_t}\beta_n}, \tag{17}$$

and

$$P_{\mathcal{N}}=\left[\frac{\sum_{n\in\mathcal{N}}\sum_{i=1}^{\beta_n}\mathbb{1}_{y_{n,i}=c}}{\sum_{n\in\mathcal{N}}\beta_n}\middle|\forall c\in\mathcal{C}\right]=\frac{\sum_{n\in\mathcal{N}}\beta_nP_n}{\sum_{n\in\mathcal{N}}\beta_n}, \tag{18}$$

the error in (8) can be expressed as

$$\mathbf{e}^{(\mathrm{t})}=\sum_{c\in\mathcal{C}}[P_{\mathcal{S}_t}(y=c)-P_{\mathcal{N}}(y=c)]\nabla\mathbb{E}_{\boldsymbol{x}|y=c}\left[\log f_c(\boldsymbol{x},\mathbf{w}^{(\mathrm{t})})\right]. \tag{19}$$

Based on this equation, the following remarks can be obtained.

Remark 5. *In multi-class classification problems, the error caused by device selection is mainly determined by the distance between the data distribution of the selected devices $P_{\mathcal{S}_t}$ and the global data distribution $P_{\mathcal{N}}$, and this impact is affected by the gradient $\nabla\mathbb{E}_{\boldsymbol{x}|y=c}\left[\log f_c(\boldsymbol{x},\mathbf{w}^{(\mathrm{t})})\right]$.*

Remark 6. *In the considered federated learning framework, if the data distribution of the selected devices is the same as the global data distribution, the weight divergence issue can be avoided.*

Equation (19) indicates that in the multi-class classification problem, the error is partially decided by the term $P_{\mathcal{S}_t}(y=c)-P_{\mathcal{N}}(y=c)$. According to the definitions in (17) and (18), this case can be viewed as a ratio estimation, where $\mathbb{E}[P_{\mathcal{S}_t}(y=c)-P_{\mathcal{N}}(y=c)]=0$. However, the ratio estimation is biased, and it is indicated that the bias can be reduced if sampling size $|\mathcal{S}_t|$ is large [20], [31]. In other words, the weight divergence issue can be mitigated if device participation increases. In Fig. 3, the impact of device participation on weight divergence is demonstrated, where 20% of selected devices becomes unavailable in the case of low device participation. This result shows that although both cases have the same trend, weight divergence is more severe and more unstable at lower device participation.

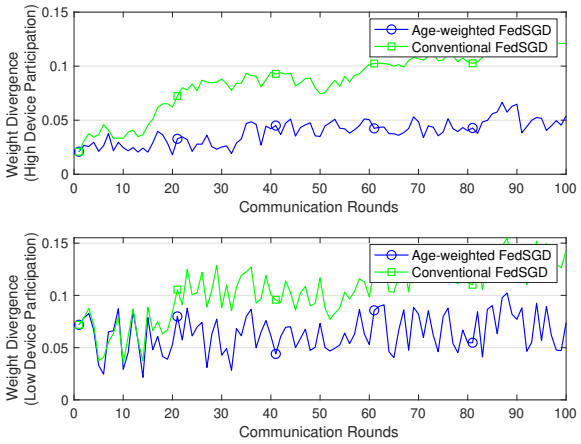


Fig. 3: An empirical result to show the impact of device participation on weight divergence on the balanced MNIST dataset. $N=10$ and $K=5$.

In order to further explore the impact of device participation, the convergence rate of age-weighted FedSGD is analyzed. For age-weighted FedSGD, the difference in global gradients between random device selection and complete device selection is defined as follows:

$$\mathbf{g}^{(\mathrm{t})}\triangleq\nabla G(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)-\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N}). \tag{20}$$

Since all devices are selected in complete device selection, $A_n^{(\mathrm{t})}=1,\forall n,t$ always holds, and hence, the AoI based weighting factor satisfies $\omega_n^{(\mathrm{t})}=1$. In this case, the following equation can be obtained:

$$\nabla G(\mathbf{w}^{(\mathrm{t})},\mathcal{N})=\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N}). \tag{21}$$

The above equation indicates that age-weighted FedSGD can be utilized for complete device selection without any impact. Based on (20) and (21), the expected convergence rate of age-weighted FedSGD can be obtained.

Theorem 2. *With age-weighted FedSGD, the expected reduction of global loss in round t is bounded by*

$$\begin{aligned} &\mathbb{E}\left[F(\mathbf{w}^{(\mathrm{t}+1)},\mathcal{N})-F(\mathbf{w}^{*})\right] \\ &\leq\left(1-\frac{\mu}{L}\right)^{t}\mathbb{E}\left[F(\mathbf{w}^{(1)},\mathcal{N})-F(\mathbf{w}^{*})\right]+\frac{1}{2L}\sum_{i=1}^{t}\left(1-\frac{\mu}{L}\right)^{t-i}\mathbb{E}\left[\|\mathbf{g}^{(i)}\|^2\right], \end{aligned} \tag{22}$$

where

$$\mathbb{E}\left[\|\mathbf{g}^{(\mathrm{t})}\|^2\right] \tag{23}$$
$$= \left(1-\frac{|\mathcal{S}_t|}{N}\right)\frac{\sum_{n\in\mathcal{N}}\beta_n^2\left\|\omega_n^{(\mathrm{t})}\nabla f_n(\mathbf{w}^{(\mathrm{t})})-\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})\right\|^2}{|\mathcal{S}_t|(N-1)(\frac{1}{N}\sum_{n\in\mathcal{N}}\beta_n)^2}.$$

Proof: Refer to Appendix B. ∎

Theorem 2 indicates that with age-weighted FedSGD, the convergence rate of the considered federated learning algorithm can be improved by increasing device participation.

D. Local Training and Transmission

As described in the previous subsection, increasing device participation can further improve learning performance, including mitigating weight divergence and accelerating convergence rate. However, in practical federated learning scenarios, selected devices may not be able to participate in the aggregation due to latency or energy consumption limitations. Since these two metrics are jointly determined by the parameters in local training and transmission phases, such as computing power, data size, transmit power, channel gain, etc., there exists a trade-off between them [28], [32]. In this case, resource allocation can be leveraged to satisfy energy consumption or latency conditions, thereby increasing device participation.

For any selected device n assigned to sub-channel k, it trains the global model based on all local samples, and hence, the computing time can be expressed as follows:

$$T_{k,n}^{\mathrm{cp}} = \frac{\mu\beta_n}{\tau_{k,n}C_n}, \tag{24}$$

where μ is the required number of cycles to train each sample, $\tau_{k,n}$ is the computing resource allocation coefficient, and C_n is the computational capacity of device n. According to [20], [22], the corresponding energy consumption for local training is given by

$$E_{k,n}^{\mathrm{cp}} = \kappa\mu\beta_n(\tau_{k,n}C_n)^2, \tag{25}$$

where κ is the power consumption coefficient of each central processing unit (CPU) cycle. After local training, the local gradient is sent to the server through the assigned sub-channel at the following data rate:

$$R_{k,n} = B\log_2(1+\alpha_{k,n}P_n|h_{k,n}|^2), \tag{26}$$

where B is the allocated bandwidth of each sub-channel, $\alpha_{k,n}$ is the power allocation coefficient, P_n is the maximum transmit power, $|h_{k,n}|^2 = \eta|g_n|^2 d_n^{-a}\sigma^{-2}$ is the normalized channel gain, η is the frequency dependent factor, g_n is the small-scale fading coefficient, d_n is the distance between device n and the server, α is the path loss exponent, and σ^2 is the noise power. The communication time of device n assigned to sub-channel k can be expressed as follows:

$$T_{k,n}^{\mathrm{cm}} = \frac{D}{R_{k,n}}, \tag{27}$$

where D is the size of the local gradient for each device. The energy consumption for transmission is given by

$$E_{k,n}^{\mathrm{cm}} = \alpha_{k,n}P_nT_{k,n}^{\mathrm{cm}}. \tag{28}$$

III. Problem Formulation

In federated learning algorithms, an aggregation deadline is usually considered, which ensures that the server updates the global model at a certain point in time. Therefore, in this work, latency is regarded as a key metric that determines device participation, and an energy consumption minimization problem is formulated under the maximum time consumption constraint. The problem is shown as follows:

$$\min_{\boldsymbol{\tau},\boldsymbol{\alpha},\boldsymbol{\psi}} \quad \sum_{n\in\mathcal{S}_t}\sum_{k\in\mathcal{K}}\psi_{k,n}(E_{k,n}^{\mathrm{cp}}+E_{k,n}^{\mathrm{cm}}) \tag{29a}$$
$$\text{s.t.} \quad T_{k,n}^{\mathrm{cp}}+T_{k,n}^{\mathrm{cm}} \le T_n^{\max}, \forall k\in\mathcal{K}, \forall n\in\mathcal{S}_t, \tag{29b}$$
$$\tau_{k,n}\in[0,1], \forall k\in\mathcal{K}, \forall n\in\mathcal{S}_t, \tag{29c}$$
$$\alpha_{k,n}\in[0,1], \forall k\in\mathcal{K}, \forall n\in\mathcal{S}_t, \tag{29d}$$
$$\psi_{k,n}^{(\mathrm{t})}\in\{0,1\}, \forall k\in\mathcal{K}, \forall n\in\mathcal{S}_t, \tag{29e}$$
$$\sum\nolimits_{n\in\mathcal{S}_t}\psi_{k,n}^{(\mathrm{t})}\in\{0,1\}, \forall k\in\mathcal{K}, \tag{29f}$$
$$\sum\nolimits_{k\in\mathcal{K}}\psi_{k,n}^{(\mathrm{t})}\in\{0,1\}, \forall n\in\mathcal{S}_t, \tag{29g}$$

where $\boldsymbol{\tau}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\psi}$ are the sets of all computing resource allocation coefficients, power allocation coefficients, and sub-channel assignment indicators, respectively. In constraint (29b), $T_n^{\max}$ denotes the maximum time consumption of each communication round. Constraints (29c) and (29d) indicate that computing resource allocation coefficients and power allocation coefficients range from 0 to 1. Constraints (29e), (29f) and (29g) represent that the sub-channel assignment indicator is a binary variable, any sub-channel can be occupied by at most one device, and any device can be assigned to at most one sub-channel, respectively. In particular, in problem (29), resource allocation and sub-channel assignment are performed with the given set of selected devices.

Due to the fact that the formulated problem is a mixed integer linear programming problem, it is decoupled into two sub-problems and solved iteratively. With the fixed sub-channel assignment, the resource allocation problem can be presented as follows:

$$\min_{\boldsymbol{\tau},\boldsymbol{\alpha}} \quad \sum_{n\in\mathcal{S}_t}\sum_{k\in\mathcal{K}}E_{k,n}^{\mathrm{cp}}+E_{k,n}^{\mathrm{cm}} \tag{30a}$$
$$\text{s.t.} \quad \text{(29b), (29c), and (29d).}$$

By removing the constraints related to resource allocation, the sub-channel assignment problem is shown in follows:

$$\min_{\boldsymbol{\psi}} \quad \sum_{n\in\mathcal{S}_t}\sum_{k\in\mathcal{K}}\psi_{k,n}(E_{k,n}^{\mathrm{cp}}+E_{k,n}^{\mathrm{cm}}) \tag{31a}$$
$$\text{s.t.} \quad \text{(29e), (29f), and (29g).}$$

IV. Joint Optimization of Computational Resource Allocation and Power Allocation

Since the adjustment of resource allocation coefficients for any device cannot affect other devices, the resource allocation problem in (30) is divided into K sub-problems and solved independently. The resource allocation problem for device n assigned to sub-channel k is given by

$$\min_{\tau_{k,n},\alpha_{k,n}} \ \kappa\mu\beta_n(\tau_{k,n}C_n)^2+\frac{\alpha_{k,n}P_nD}{B\log_2(1+\alpha_{k,n}P_n|h_{k,n}|^2)} \quad (32a)$$

$$\text{s.t.} \ \ \frac{\mu\beta_n}{\tau_{k,n}C_n}+\frac{D}{B\log_2(1+\alpha_{k,n}P_n|h_{k,n}|^2)} \le T_n^{\max}, \quad (32b)$$

$$\tau_{k,n} \in [0,1], \quad (32c)$$

$$\alpha_{k,n} \in [0,1]. \quad (32d)$$

Note that the above problem is infeasible if any constraint is not satisfied. Hence, the following remark can be drawn.

Remark 7. *For any device n assigned to sub-channel k, its local gradient cannot be transmitted if*

$$\frac{\mu\beta_n}{C_n}+\frac{D}{B\log_2(1+P_n|h_{k,n}|^2)} > T_n^{\max}. \quad (33)$$

That is, the selected devices may not be able to transmit local gradients to the server with the given time limitation, even if all resources are utilized. On the other hand, if this condition does not hold, the training and transmission tasks can alway be completed, which means $\tau_{k,n} > 0$ and $\alpha_{k,n} > 0$. Therefore, x_1 and x_2 are introduced to replace the optimization variables, where $x_1 = 1/\tau_{k,n}$ and $x_2 = 1/[B\log_2(1+\alpha_{k,n}P_n|h_{k,n}|^2)]$. Problem (32) can be equivalently transformed as follows:

$$\min_{\mathbf{x}} \ \ \frac{\kappa\mu\beta_nC_n^2}{x_1^2}+x_2\frac{(2^{\frac{1}{x_2B}}-1)D}{|h_{k,n}|^2}, \quad (34a)$$

$$\text{s.t.} \ \ \mu\beta_nC_n^{-1}x_1+Dx_2 \le T_n^{\max}, \quad (34b)$$

$$x_1 \ge 1, \quad (34c)$$

$$x_2 \ge \frac{1}{B\log_2(1+P_n|h_{k,n}|^2)}, \quad (34d)$$

where $\mathbf{x} = \{x_1, x_2\}$. It can be proved that the above problem is convex and satisfies Slater's condition, and hence, KKT conditions are utilized to derive the optimal solution [33]. By introducing the Lagrangian multiplier λ_i for the inequality constraints, the Lagrangian function is given by

$$L(\mathbf{x})=\frac{\kappa\mu\beta_nC_n^2}{x_1^2}+x_2\frac{(2^{\frac{1}{x_2B}}-1)D}{|h_{k,n}|^2}+\lambda_1\left(\frac{\mu\beta_n}{C_n}x_1+Dx_2-T_n^{\max}\right)$$
$$+\lambda_2(1-x_1)+\lambda_3\left[\frac{1}{B\log_2(1+P_n|h_{k,n}|^2)}-x_2\right]. \quad (35)$$

Based on the Lagrangian function, the optimal solution of problem (34) can be presented below.

Proposition 1. *In case of $\mu\beta_nC_n^{-1}+D\upsilon_1 \le T_n^{\max}$, by defining*

$$\begin{cases} \upsilon_1 \triangleq \dfrac{1}{B\log_2(1+P_n|h_{k,n}|^2)}, \\ \upsilon_2 \triangleq \dfrac{1}{B(T_n^{\max}-\mu\beta_nC_n^{-1})}, \end{cases} \quad (36)$$

the optimal solution of problem (34) is given by

1) $x_1^* = 1$ and $x_2^* = \upsilon_1$, if the following condition holds:

$$\mu\beta_nC_n^{-1}+D\upsilon_1 = T_n^{\max}. \quad (37)$$

2) $x_1^* = 1$ and $x_2^* = (T_n^{\max}-\mu\beta_nC_n^{-1})/D$, if

$$\begin{cases} \mu\beta_nC_n^{-1}+D\upsilon_1 < T_n^{\max}, \\ D\upsilon_2\ln(2)2^{D\upsilon_2}-2^{D\upsilon_2}+1-2\kappa C_n^3|h_{k,n}|^2 > 0. \end{cases} \quad (38)$$

3) $x_1^* = (T_n^{\max}-D\upsilon_1)C_n(\mu\beta_n)^{-1}$ and $x_2^* = \upsilon_1$ if

$$\begin{cases} \mu\beta_nC_n^{-1}+D\upsilon_1 < T_n^{\max}, \\ 2^{\frac{1}{B\upsilon_1}}-1-\dfrac{1}{B\upsilon_1}\ln(2)2^{\frac{1}{B\upsilon_1}}+\dfrac{2\kappa(\mu\beta_n)^3|h_{k,n}|^2}{(T_n^{\max}-D\upsilon_1)^3} > 0. \end{cases} \quad (39)$$

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

$$\begin{cases} \dfrac{2\kappa C_n^3}{(x_1^*)^3} = \dfrac{\ln(2)2^{\frac{1}{Bx_2^*}}}{B|h_{k,n}|^2x_2^*}-\dfrac{2^{\frac{1}{Bx_2^*}}-1}{|h_{k,n}|^2}, \\ \mu\beta_nC_n^{-1}x_1^*+Dx_2^*-T_n^{\max}=0. \end{cases} \quad (40)$$

Proof: Refer to Appendix C. ∎

According to the above proposition, the optimal solution of problem (32) can be obtained as follows:

$$\begin{cases} \tau_{k,n}^* = 1/x_1^*, \\ \alpha_{k,n}^* = \dfrac{2^{\frac{1}{Bx_2^*}}-1}{P_n|h_{k,n}|^2}, \end{cases} \quad (41)$$

and the formulated problem in (30) is jointly solved.

V. Matching based Sub-Channel Assignment

In this section, the formulated sub-channel assignment problem in (31) is solved with the given resource allocation solutions. Specifically, the optimal resource allocation for all devices assigned to all sub-channels can be obtained in Section IV, and therefore, this solution is treated as a preference list to construct a matching based sub-channel assignment algorithm. Note that some combinations of devices and sub-channels may not be feasible due to inability to satisfy the maximum time consumption constraint, and the proposed algorithm may tend to assign devices to the corresponding infeasible sub-channels to achieve lower energy consumption. In order to avoid this case, a large value is assigned as the energy consumption of these infeasible combinations, and any combination with this energy consumption will be removed from the final matching.

A. Design of Matching based Algorithm

At this stage, with the preference list setting, all devices in $\mathcal{S}_t$ can be assigned to sub-channels, and thus problem (31) can be considered as a one-to-one matching Ψ from $\mathcal{S}_t$ to $\mathcal{K}$, where $\mathcal{S}_t$ and $\mathcal{K}$ are two disjoint sets with the same size. In the resource allocation problem, it is indicated that the energy consumption of any device n or sub-channel k in matching Ψ is independent of other players, and therefore, the utility of any player can be defined as follows:

$$U_i(\Psi) = E_{k,n}^{\text{cp}} + E_{k,n}^{\text{cm}}, \forall i \in \{n,k\}. \quad (42)$$

Due to the fact that the device and sub-channel in a combination have the same utility, the intent of sub-channels can be omitted. Moreover, since each device is assigned to one sub-channel and each sub-channel is occupied by one device, if a device tends to establish a new matching, it needs to exchange with another device instead of joining the combination directly. That is, the considered matching is a swap matching, defined as follows:

Algorithm 1 Matching based Algorithm

Initialization:
Randomly match all players in $\mathcal{S}_t$ and $\mathcal{K}$ to obtain Ψ.
Set $\Psi_\alpha = 0$ and $\Psi_\beta = 1$.
Main Loop:
if $\Psi_\alpha \neq \Psi_\beta$ **then**
  Set $\Psi_\alpha = \Psi$.
  for $n \in \mathcal{S}_t$ **do**
    Device n searches device $n' \in \mathcal{S}_t$, where $n \neq n'$.
    if (n, n') is a swap-blocking pair **then**
      Devices n and n' exchange sub-channels.
      Matching $\Psi^n_{n'}$ is obtained.
      Set $\Psi = \Psi^n_{n'}$.
    end if
  end for
  Set $\Psi_\beta = \Psi$.
end if

Definition 1. *From matching Ψ with $\Psi(n) = k$ and $\Psi(n') = k'$, a swap matching $\Psi^{n'}_n$ represents an exchange of devices n and n', i.e.,*

$$\Psi^{n'}_n = \Psi \backslash \{\{k, n\}, \{k', n'\}\} \cup \{\{k, n'\}, \{k', n\}\}. \tag{43}$$

As defined above, a swap matching means that two devices exchange their assigned sub-channels. Note that the motivation to form a swap matching is the reduction in energy consumption, which can be presented as follows:

$$\Psi \preceq_i \Psi^{n'}_n \Leftrightarrow U_i(\Psi) \leq U_i(\Psi^{n'}_n), \forall i \in \{n, n'\}, \tag{44}$$

where $\Psi \preceq_i \Psi^{n'}_n$ indicates device i prefers $\Psi^{n'}_n$ to Ψ. Moreover, symbol $\prec_i$ is also introduced to represent the strict preference of device i. The swap matching should be approved by all involved players, in which the utility of any player increases or remains unchanged. In this case, devices n and n' becomes a swap-blocking pair (n, n'), defined as follows:

Definition 2. *(n, n') is a swap-blocking pair if and only if $\Psi \prec_i \Psi^{n'}_n, \exists i \in \{n, n'\}$ and $\Psi \preceq_i \Psi^{n'}_n, \forall i \in \{n, n'\}$.*

Based on the definition of the swap-blocking pair, a matching based sub-channel assignment algorithm is presented in Algorithm 1. In this algorithm, an initial matching is firstly obtained by randomly assigning all devices into all sub-channels. Afterwards, each device in turn operates on the remaining devices in order to find the swap-blocking pair. If any two devices can form a swap-blocking pair, their sub-channels are exchanged and the new matching is recorded. This algorithm is repeated until no new swap-blocking pair can be found in a complete cycle. Based on the finial matching provided by Algorithm 1, the solution of sub-channel assignment problem (31) can be obtained by removing all infeasible combinations.

B. Properties Analysis

In this subsection, the properties of the proposed matching based sub-channel assignment algorithm, including complexity, convergence, and stability, are analyzed.

1) Complexity: The computational complexity of the proposed algorithm is $\mathcal{O}(CK^2)$, where C is the number of cycles. Specifically, during a complete cycle of the main loop, each device needs to test the viability of creating swap-blocking pairs with all other devices, and hence, for all K devices, $K(K-1)$ times of calculations should be performed. With the given number of cycles C, the computational complexity can be expressed as $CK(K-1)$.

2) Convergence: From any initial matching, the proposed algorithm is guaranteed to converge to a final matching without swap-blocking pairs. It can be observed from Algorithm 1 that the matching is transformed due to the construction of swap-blocking pairs. Suppose Ψ_a and Ψ_b are two adjacent matching, where $a \neq b$, then there exists a swap-blocking pair in the transformation from Ψ_a to Ψ_b. Based on the definition of a swap-blocking pair, it indicates that the utility of at least one device is strictly reduced while the utility of the other device is not increased. Moreover, for the devices that are not involved, their utility remains the same. As a result, the sum utility, or sum energy consumption, is strictly decreased with this transformation. With the given devices and sub-channels, there is a lower bound on the energy consumption, and therefore, the proposed algorithm can always converge to a final matching.

3) Stability: The proposed matching based sub-channel assignment algorithm is able to provide a two-side exchange stable solution, which is defined as follows:

Definition 3. *A matching is two-side exchange stable if and only if no swap-blocking pair can be formed.*

According to the above definition, the final matching obtained by the proposed algorithm is alway two-side exchange stable, since the convergence analysis proves that there are no swap-blocking pairs in the final matching.

VI. Simulation Results

A. Simulation Settings and Benchmark Methods

The simulation results are presented to demonstrate the performance of age-weighted FedSGD and proposed solutions. In this simulation, the devices are randomly deployed in a disc with radius R, while the server is located in the center. The learning rate λ is 0.01, bandwidth B is 1 MHz, noise power σ^2 is -174 dBm, path loss exponent a is 3.76, power consumption coefficient κ is 10^{-29}, and cycles coefficient μ is 10^6. To evaluate the learning performance, MNIST, CIFAR-10, and CIFAR-100 datasets are adopted with an SGD optimizer. For MNIST digit recognition tasks, a simple neural network is built with a 128-neuron ReLu hidden layer and a softmax output layer. For CIFAR-10 image classification tasks, a neural network is constructed by stacking a 32-filter 4×4 2D convolution (Conv2D) layers, a 2×2 max pooling layer, a 128-neuron ReLu hidden layer and a softmax output layer. For CIFAR-100 image classification tasks,, the neural network is constructed with a 128-filter 4×4 Conv2D layers, a 2×2 max pooling layer, a 256-neuron ReLu hidden layer and a softmax output layer. To compare the performance of the proposed KKT based resource allocation (denoted by KRA) and matching based sub-channel assignment (denoted

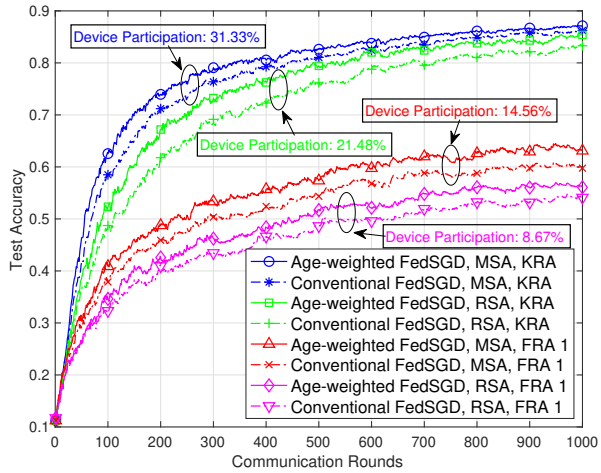


Fig. 4: The convergence performance on the unbalanced MNIST dataset. $N = 10$, $K = 4$, $T_n^{\max} = 5$ s, $P_n = 10$ dBm, $C_n = 1$ GHz, $R = 200$ m, and $D = 10$ Mbits.

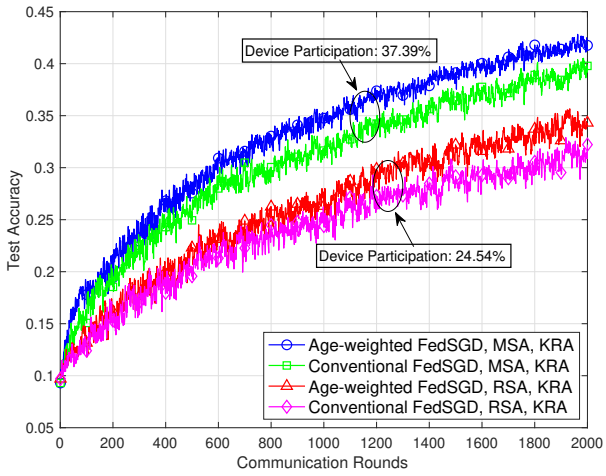


Fig. 5: The convergence performance on the balanced CIFAR-10 dataset. $N = 10$, $K = 5$, $T_n^{\max} = 10$ s, $P_n = 10$ dBm, $C_n = 1$ GHz, $R = 200$ m, and $D = 15$ Mbits.

by MSA), fixed resource allocation (denoted by FRA) and random sub-channel assignment (denoted by RSA) are respectively included as the baseline, where resource allocation coefficients are set as $\tau_{k,n} = \alpha_{k,n} = 0.5, \forall k, n$ with FRA 1, and $\tau_{k,n} = \alpha_{k,n} = 1, \forall k, n$ with FRA 2. Furthermore, the device participation is calculated by $\bar{S}/N$, where $\bar{S}$ is the arithmetic mean of $\{S_t | \forall t\}$ and $S_t = |\mathcal{S}_t|$.

B. Federated Learning Performance

In Fig. 4, the MNIST digit recognition task with unbalanced non-IID data is adopted, where 9000 training samples are randomly distributed to all devices, and the samples of each device belong to 1 or 2 classes. It demonstrates that with the same device selection results, the proposed age-weighted FedSGD is able to outperform conventional FedSGD with any sub-channel assignment and resource allocation schemes. Compared to RSA and FRA, MSA and KRA can significantly

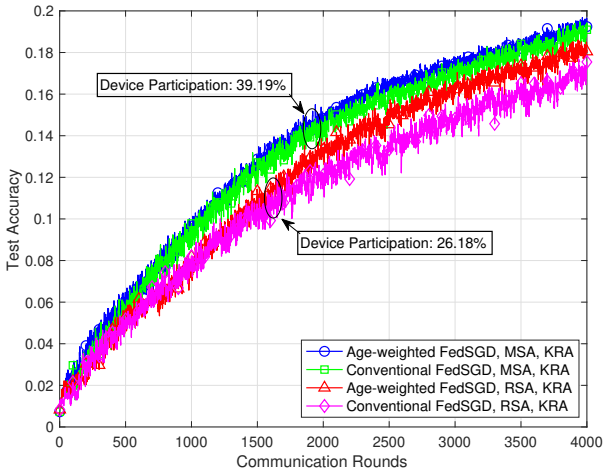


Fig. 6: The convergence performance on the balanced CIFAR-100 dataset. $N = 50$, $K = 20$, $T_n^{\max} = 10$ s, $P_n = 10$ dBm, $C_n = 1$ GHz, $R = 200$ m, and $D = 20$ Mbits.

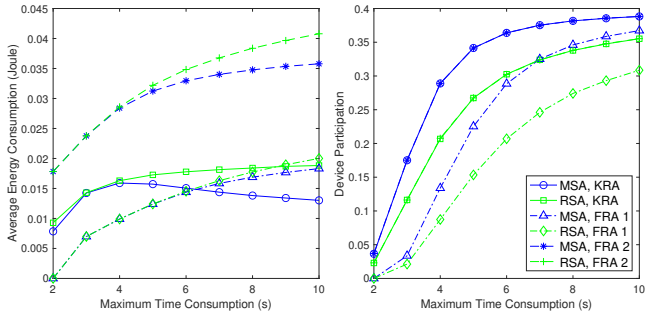


Fig. 7: The impact of the maximum time consumption. $N = 10$, $K = 4$, $P_n = 10$ dBm, $C_n = 1$ GHz, $R = 200$ m, and $D = 10$ Mbits.

increase device participation, thereby improving the learning performance, which verifies Theorem 2. Moreover, since the number of selected devices in each communication round is obviously reduced with FRA, there exists a large gap between FRA and KRA in achievable test accuracy.

The CIFAR-10 image classification task is employed in Fig. 5, in which each device has 5000 samples with the unique label. It demonstrates that by utilizing the proposed sub-channel assignment strategy, device participation is increased by 52%. As a result, for both age-weight FedSGD and conventional FedSGD, the learning performance is improved, which confirms Theorem 2. Moreover, it can be found from Fig. 5 that with the same device selection results, age-weighted FedSGD can achieve faster convergence rate compared to conventional FedSGD.

The balanced CIFAR-100 dataset is adopted in Fig. 6, in which each device has 1000 2-class samples. Due to the fact that the Non-IID degree in this figure is decreased compared to that in Fig. 4 and Fig. 5, the improvement of age-weighted FedSGD with MSA is not significant, but it can still improve the learning performance, especially in the later stage of training. With RSA, the advantage of age-weighted FedSGD is obvious, and this is because the number of selected devices

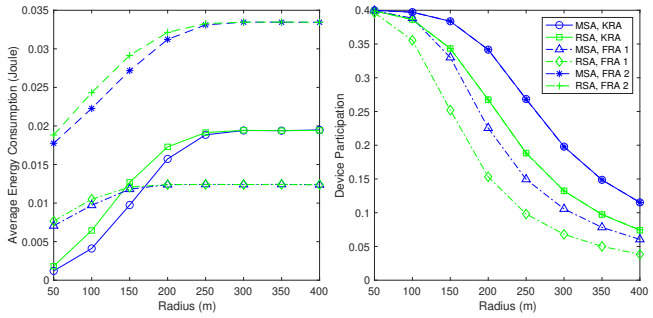


Fig. 8: The impact of channel conditions. $N = 10$, $K = 4$, $T_n^{\max} = 5$ s, $P_n = 10$ dBm, $C_n = 1$ GHz, and $D = 10$ Mbits.

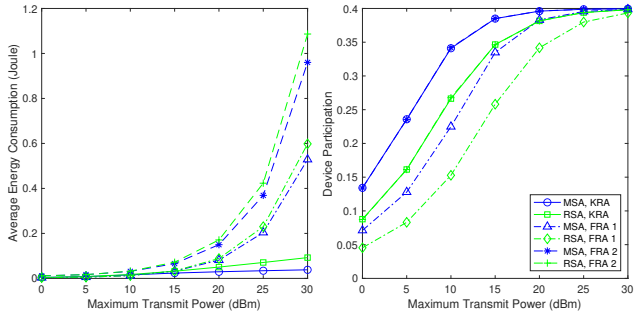


Fig. 9: The impact of the maximum transmit power. $N = 10$, $K = 4$, $T_n^{\max} = 5$ s, $C_n = 1$ GHz, $R = 200$ m, and $D = 10$ Mbits.

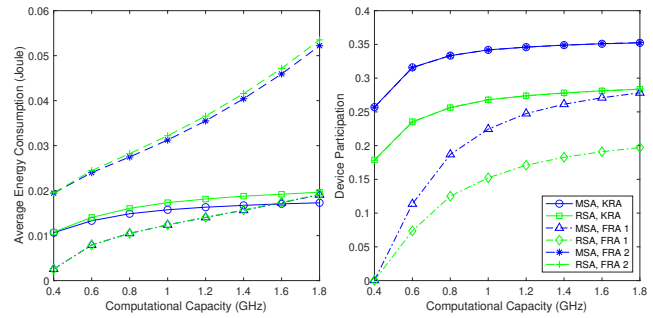


Fig. 10: The impact of the computational capacity. $N = 10$, $K = 4$, $T_n^{\max} = 5$ s, $P_n = 10$ dBm, $R = 200$ m, and $D = 10$ Mbits.

in each round is significantly reduced. It is worth noting that the considered federated learning algorithm is able to continue to converge. However, in order to clearly demonstrate the difference, only the first several rounds are included for demonstration. In addition, under the adopted simulation parameters, the baseline FRA 1 fails to converge on CIFAR-10 and CIFAR-100 datasets due to low device participation, and is therefore not included in Fig. 5 and Fig. 6.

C. System Performance

The performance of the proposed solutions is shown in Fig. 7 to Fig. 9, where the maximum time consumption, radius, maximum transmit power, and computational capacity are respectively included to show its impact on the average energy consumption and device participation. It can be observed

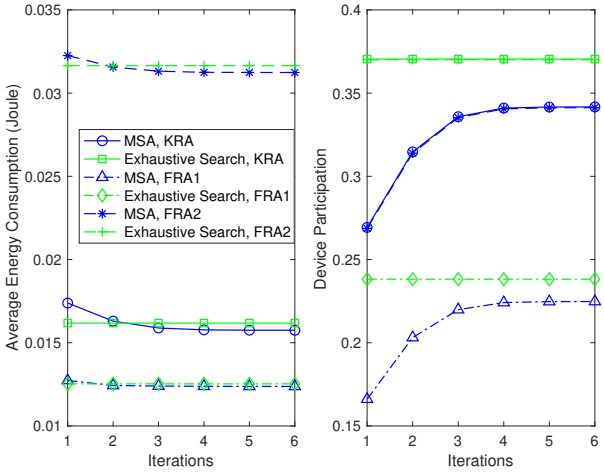


Fig. 11: The convergence of the matching based sub-channel assignment algorithm. $N = 10$, $K = 4$, $T_n^{\max} = 5$ s, $P_n = 10$ dBm, $C_n = 1$ GHz, $R = 200$ m, and $D = 10$ Mbits.

that compared to FRA 2, the proposed KKT based resource allocation solution can achieve less energy consumption while ensuring device participation. Compared to FRA 1, KRA can reduce energy consumption and increase device participation, which explains the improved learning performance in Fig. 4 to Fig. 6. Moreover, Fig. 7 indicates that there is a trade-off between energy consumption and device participation. Specifically, when the maximum time consumption increases from 2 s to 4 s, the energy consumption is raised in order to significantly increase device participation, from 0.4 to 3. When the maximum time consumption is greater than 4 s, device participation can be slowly increased, but the average energy consumption is reduced. In Fig. 8, the increase in radius can be regarded as the deterioration of channel conditions, and then it can be observed that there is an upper boundary in energy consumption. That is, when the radius is greater than 250 m, the average energy consumption remains at a fixed level, even though device participation continues to decrease. Similarly, an upper bound in device participation can be found in Fig. 9. When the maximum transmit power is equal to 30 dBm, the maximum device participation is achieve by all schemes, and the proposed solutions still consume minimum energy. Compared to the maximum transmit power, the impact of computational capacity is not significant, but the overall trend is the same, i.e., the energy consumption and device participation increase monotonically, as shown in Fig. 10.

The convergence of the proposed sub-channel assignment algorithm is demonstrated in Fig. 11, where exhaustive search is included as the benchmark. In this figure, the exhaustive search is set up to find the combination that maximizes device participation while guaranteeing a low level of average energy consumption. Therefore, although the average energy consumption obtained by the exhaustive search is slightly larger than that of the proposed matching based algorithm, the device participation is significantly increased. It can be observed that the proposed sub-channel assignment algorithm can achieve approximately 92% performance of the global optimum within

4 iterations. Compared to exhaustive search with complexity $\mathcal{O}(K!)$, it can be considered as a low complexity sub-optimal algorithm. Furthermore, the figure verifies the properties of the proposed algorithm, including convergence and stability.

D. Discussion

The simulation results indicate that the proposed age-weighted FedSGD algorithm can significantly reduce weight divergence in data heterogeneity scenarios, thereby improving the performance of federated learning. It also explains the principle that introducing AoI in federated learning can improve learning efficiency in existing works. Moreover, the proposed solutions for resource allocation and sub-channel assignment can further improve learning performance by increasing device participation. We note that the proposed schemes can be implemented without extra information transmission, and the computation can be performed at the server equipped with sufficient computational resources and energy. Therefore, the increase in computational complexity brought by adopting these schemes can be neglected. Furthermore, as the proposed scheme can achieve higher accuracy with fewer communication rounds, it is well suited for time- or energy-sensitive scenarios, where neural network training needs to be completed with a given number of communication rounds.

VII. Conclusions

This paper investigates a wireless federated learning framework on non-IID datasets, where random device selection is exploited due to the limited number of sub-channels. By exploring the issue of conventional FedSGD in weight divergence, age-weighted FedSGD is designed to adjust the proportion of local gradients according to the previous state of devices. To further improve the learning performance, an energy consumption minimization problem is formulated, where the resource allocation solution and the sub-channel assignment algorithm are developed based on KKT conditions and matching theory, respectively. The superiority of designed age-weighted FedSGD and the effectiveness of the proposed resource allocation and sub-channel assignment strategies are demonstrated in the simulation results. For future works, to adapt to more advanced artificial intelligence scenarios, it is necessary to conduct further analysis based on weaker assumptions. Furthermore, adopting real-world scenarios and datasets to validate the proposed scheme is also an important research direction.

Appendix A: Proof of Theorem 1

According to (3) and (6), the weight divergence between random device selection and complete device selection can be expressed as follows:

$$\begin{aligned}&\left\|\mathbf{w}^{(\mathrm{t}+1)}-\mathbf{w}_{\mathrm{T}}^{(\mathrm{t}+1)}\right\|\\&=\left\|\mathbf{w}^{(\mathrm{t})}-\mathbf{w}_{\mathrm{T}}^{(\mathrm{t})}-\lambda\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)+\lambda\nabla F(\mathbf{w}_{\mathrm{T}}^{(\mathrm{t})},\mathcal{N})\right\|.\end{aligned}\tag{45}$$

From (8), $\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)=\mathbf{e}^{(\mathrm{t})}+\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})$ can be obtained, and the above equation can be transformed as follows:

$$\begin{aligned}&\left\|\mathbf{w}^{(\mathrm{t}+1)}-\mathbf{w}_{\mathrm{T}}^{(\mathrm{t}+1)}\right\|\\&=\left\|\mathbf{w}^{(\mathrm{t})}-\mathbf{w}_{\mathrm{T}}^{(\mathrm{t})}-\lambda\mathbf{e}^{(\mathrm{t})}-\lambda\left[\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})-\nabla F(\mathbf{w}_{\mathrm{T}}^{(\mathrm{t})},\mathcal{N})\right]\right\|\\&=\left\|\mathbf{w}^{(1)}-\mathbf{w}_{\mathrm{T}}^{(1)}-\lambda\sum_{i=1}^{t}\mathbf{e}^{(i)}-\lambda\sum_{i=1}^{t}\left[\nabla F(\mathbf{w}^{(i)},\mathcal{N})-\nabla F(\mathbf{w}_{\mathrm{T}}^{(i)},\mathcal{N})\right]\right\|\\&\le\left\|\mathbf{w}^{(1)}-\mathbf{w}_{\mathrm{T}}^{(1)}\right\|+\lambda\left\|\sum_{i=1}^{t}\mathbf{e}^{(i)}\right\|\\&\quad+\lambda\sum_{i=1}^{t}\left\|\nabla F(\mathbf{w}^{(i)},\mathcal{N})-\nabla F(\mathbf{w}_{\mathrm{T}}^{(i)},\mathcal{N})\right\|.\end{aligned}\tag{46}$$

Based on Assumption 1, the following inequality holds:

$$\begin{aligned}&\left\|\mathbf{w}^{(\mathrm{t}+1)}-\mathbf{w}_{\mathrm{T}}^{(\mathrm{t}+1)}\right\|\\&\le\left\|\mathbf{w}^{(1)}-\mathbf{w}_{\mathrm{T}}^{(1)}\right\|+\lambda\left\|\sum_{i=1}^{t}\mathbf{e}^{(i)}\right\|+\lambda L\sum_{i=1}^{t}\left\|\mathbf{w}^{(i)}-\mathbf{w}_{\mathrm{T}}^{(i)}\right\|.\end{aligned}\tag{47}$$

With the similar way, the following inequality can be obtained:

$$\begin{aligned}&\lambda L\left\|\mathbf{w}^{(\mathrm{t})}-\mathbf{w}_{\mathrm{T}}^{(\mathrm{t})}\right\|\\&\le\lambda L\left\|\mathbf{w}^{(1)}-\mathbf{w}_{\mathrm{T}}^{(1)}\right\|+\lambda^2L\left\|\sum_{i=1}^{t-1}\mathbf{e}^{(i)}\right\|+(\lambda L)^2\sum_{i=1}^{t-1}\left\|\mathbf{w}^{(i)}-\mathbf{w}_{\mathrm{T}}^{(i)}\right\|.\end{aligned}\tag{48}$$

By substituting the above inequality, (47) can be rewritten as follows:

$$\begin{aligned}&\left\|\mathbf{w}^{(\mathrm{t}+1)}-\mathbf{w}_{\mathrm{T}}^{(\mathrm{t}+1)}\right\|\\&\le(1+\lambda L)\left\|\mathbf{w}^{(1)}-\mathbf{w}_{\mathrm{T}}^{(1)}\right\|+\lambda\left\|\sum_{i=1}^{t}\mathbf{e}^{(i)}\right\|+\lambda^2L\left\|\sum_{i=1}^{t-1}\mathbf{e}^{(i)}\right\|\\&\quad+\lambda L(1+\lambda L)\sum_{i=1}^{t-1}\left\|\mathbf{w}^{(i)}-\mathbf{w}_{\mathrm{T}}^{(i)}\right\|.\end{aligned}\tag{49}$$

Similarly, $\lambda L(1+\lambda L)\left\|\mathbf{w}^{(\mathrm{t}-1)}-\mathbf{w}_{\mathrm{T}}^{(\mathrm{t}-1)}\right\|$ in the above inequality can be transformed to

$$\begin{aligned}&\lambda L(1+\lambda L)\left\|\mathbf{w}^{(\mathrm{t}-1)}-\mathbf{w}_{\mathrm{T}}^{(\mathrm{t}-1)}\right\|\\&\le\lambda L(1+\lambda L)\left\|\mathbf{w}^{(1)}-\mathbf{w}_{\mathrm{T}}^{(1)}\right\|+\lambda^2L(1+\lambda L)\left\|\sum_{i=1}^{t-2}\mathbf{e}^{(i)}\right\|\\&\quad+(\lambda L)^2(1+\lambda L)\sum_{i=1}^{t-2}\left\|\mathbf{w}^{(i)}-\mathbf{w}_{\mathrm{T}}^{(i)}\right\|,\end{aligned}\tag{50}$$

and included in (49) to obtain the following inequality:

$$\begin{aligned}&\left\|\mathbf{w}^{(\mathrm{t}+1)}-\mathbf{w}_{\mathrm{T}}^{(\mathrm{t}+1)}\right\|\\&\le(1+\lambda L)^2\left\|\mathbf{w}^{(1)}-\mathbf{w}_{\mathrm{T}}^{(1)}\right\|+\lambda\left\|\sum_{i=1}^{t}\mathbf{e}^{(i)}\right\|+\lambda^2L\left\|\sum_{i=1}^{t-1}\mathbf{e}^{(i)}\right\|\\&\quad+\lambda^2L(1+\lambda L)\left\|\sum_{i=1}^{t-2}\mathbf{e}^{(i)}\right\|+\lambda L(1+\lambda L)^2\sum_{i=1}^{t-2}\left\|\mathbf{w}^{(i)}-\mathbf{w}_{\mathrm{T}}^{(i)}\right\|.\end{aligned}\tag{51}$$

By induction, the weight divergence between random device selection and complete device selection can be presented as follows:

$$\begin{aligned}&\left\|\mathbf{w}^{(\mathrm{t}+1)}-\mathbf{w}_{\mathrm{T}}^{(\mathrm{t}+1)}\right\|\\ \leq&(1+\lambda L)^t\left\|\mathbf{w}^{(1)}-\mathbf{w}_{\mathrm{T}}^{(1)}\right\|+\lambda\left\|\sum_{i=1}^{t}\mathbf{e}^{(i)}\right\|+\lambda^2L\left\|\sum_{i=1}^{t-1}\mathbf{e}^{(i)}\right\| \quad (52)\\ &+\lambda^2L(1+\lambda L)\left\|\sum_{i=1}^{t-2}\mathbf{e}^{(i)}\right\|+\lambda^2L(1+\lambda L)^2\left\|\sum_{i=1}^{t-3}\mathbf{e}^{(i)}\right\|+\cdots.\end{aligned}$$

The above inequality can be rewritten as follows:

$$\begin{aligned}\left\|\mathbf{w}^{(\mathrm{t}+1)}-\mathbf{w}_{\mathrm{T}}^{(\mathrm{t}+1)}\right\|\leq&(1+\lambda L)^t\left\|\mathbf{w}^{(1)}-\mathbf{w}_{\mathrm{T}}^{(1)}\right\|+\lambda\left\|\sum_{i=1}^{t}\mathbf{e}^{(i)}\right\|\\ &+\lambda^2L\sum_{j=1}^{t-1}(1+\lambda L)^{j-1}\left\|\sum_{i=1}^{t-j}\mathbf{e}^{(i)}\right\|,\end{aligned} \quad (53)$$

and the proof is completed. ■

Appendix B: Proof of Theorem 2

Based on (20) and (21), the upper bound of the convergence rate can be proved through a similar approach to that in [20]. In order to derive the expression of $\mathbb{E}\left[\|\mathbf{g}^{(i)}\|^2\right]$, the gradient of the global loss, i.e., $\nabla G(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)$, can be viewed as ratio estimation, as follows:

$$\nabla G(\mathbf{w}^{(\mathrm{t})},\mathcal{S}_t)=\frac{\frac{1}{|\mathcal{S}_t|}\sum_{n\in\mathcal{N}}x_n^{(\mathrm{t})}\omega_n^{(\mathrm{t})}\beta_n\nabla f_n(\mathbf{w}^{(\mathrm{t})})}{\frac{1}{|\mathcal{S}_t|}\sum_{n\in\mathcal{N}}x_n^{(\mathrm{t})}\beta_n}\triangleq\frac{\bar{y}_{\mathcal{S}_t}}{\bar{x}_{\mathcal{S}_t}}, \quad (54)$$

where $x_n^{(\mathrm{t})}$ is a binary variable to indicate the device selection result of device n in round t. In particular, $x_n^{(\mathrm{t})}=1$ indicates that device n is selected in round t, i.e., $n\in\mathcal{S}_t$; $x_n^{(\mathrm{t})}=0$ otherwise. Since random device selection is adopted, in any communication round t, the probability of selecting device n from $\mathcal{N}$ is given by

$$\mathbb{E}[x_n^{(\mathrm{t})}]=P(x_n^{(\mathrm{t})}=1)=\frac{|\mathcal{S}_t|}{N}. \quad (55)$$

Similarly, the gradient of the global loss with complete device selection, i.e., $\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})$, can be expressed as follows:

$$\nabla F(\mathbf{w}^{(\mathrm{t})},\mathcal{N})=\frac{\frac{1}{N}\sum_{n\in\mathcal{N}}\beta_n\nabla f_n(\mathbf{w}^{(\mathrm{t})})}{\frac{1}{N}\sum_{n\in\mathcal{N}}\beta_n}\triangleq\frac{\bar{y}_{\mathcal{N}}}{\bar{x}_{\mathcal{N}}}. \quad (56)$$

At this stage, the following equations can be obtained:

$$\mathbb{E}[\bar{x}_{\mathcal{S}_t}]=\bar{x}_{\mathcal{N}}, \quad (57)$$

and

$$\mathbb{E}[\bar{y}_{\mathcal{S}_t}]=\bar{y}_{\mathcal{N}}. \quad (58)$$

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

$$\mathbb{E}\left[\|\mathbf{g}^{(\mathrm{t})}\|^2\right]=\frac{1}{(\frac{1}{N}\sum_{n\in\mathcal{N}}\beta_n)^2}\mathbb{E}\left[\left\|\bar{y}_{\mathcal{S}_t}-\bar{x}_{\mathcal{S}_t}\frac{\bar{y}_{\mathcal{N}}}{\bar{x}_{\mathcal{N}}}\right\|^2\right]. \quad (59)$$

The rest of the proof can be obtained by referring to [20]. ■

Appendix C: Proof of Proposition 1

By calculating the partial derivatives of (35) and setting them to zero, the following equations can be obtained:

$$\begin{cases}-\dfrac{2\kappa\mu\beta_nC_n^2}{(x_1^*)^3}+\lambda_1\dfrac{\mu\beta_n}{C_n}-\lambda_2=0, & (60a)\\ -\dfrac{\ln(2)D2^{\frac{1}{Bx_2^*}}}{|h_{k,n}|^2Bx_2^*}+\dfrac{D(2^{\frac{1}{Bx_2^*}}-1)}{|h_{k,n}|^2}+\lambda_1D-\lambda_3=0. & (60b)\end{cases}$$

Moreover, the following conditions should be satisfied:

$$\begin{cases}\mu\beta_nC_n^{-1}x_1^*+Dx_2^*-T_n^{\max}\leq0, & (61a)\\ 1-x_1^*\leq0, & (61b)\\ \upsilon_1-x_2^*\leq0, & (61c)\\ \lambda_1(\mu\beta_nC_n^{-1}x_1^*+Dx_2^*-T_n^{\max})=0, & (61d)\\ \lambda_2(1-x_1^*)=0, & (61e)\\ \lambda_3(\upsilon_1-x_2^*)=0, & (61f)\\ \lambda_i\geq0,\forall i\in\{1,2,3\}, & (61g)\end{cases}$$

where υ_1 is defined in (36). If $\lambda_1=0$, the following equation can be obtained from (60a):

$$-2\kappa\mu\beta_nC_n^2(x_1^*)^{-3}=\lambda_2. \quad (62)$$

Since $\lambda_2\geq0$, the above function conflicts with (61b), and hence, $\lambda_1>0$ always holds. Therefore, it can be obtained from (61d) that the following equation is always satisfied:

$$\mu\beta_nC_n^{-1}x_1^*+Dx_2^*-T_n^{\max}=0. \quad (63)$$

At this stage, based on the different values of λ_2 and λ_3, four cases need to be discussed.

1) If $\lambda_2>0$ and $\lambda_3>0$, $x_1^*=1$ and $x_2^*=\upsilon_1$ can be obtained from (61e) and (61f), respectively. In this case, the following condition can be derived from (63):

$$\mu\beta_nC_n^{-1}+D\upsilon_1=T_n^{\max}. \quad (64)$$

2) If $\lambda_2>0$ and $\lambda_3=0$, $x_1^*=1$, and it can be obtained from (63) that $x_2^*=(T_n^{\max}-\mu\beta_nC_n^{-1})/D$. In this case, (61c) should be considered, and the condition becomes

$$\mu\beta_nC_n^{-1}+D\upsilon_1<T_n^{\max}. \quad (65)$$

Note that the case $\lambda_3>0$ has been considered, and thus the equality condition is removed. Since $\lambda_3=0$, the following inequality can be obtained from (60b):

$$\lambda_1=\frac{\ln(2)2^{\frac{1}{Bx_2^*}}}{|h_{k,n}|^2Bx_2^*}-\frac{2^{\frac{1}{Bx_2^*}}-1}{|h_{k,n}|^2}>0. \quad (66)$$

Substituting the above equation into (60a), it becomes

$$\left(\frac{\ln(2)2^{\frac{1}{Bx_2^*}}}{|h_{k,n}|^2Bx_2^*}-\frac{2^{\frac{1}{Bx_2^*}}-1}{|h_{k,n}|^2}\right)\frac{\mu\beta_n}{C_n}-2\kappa\mu\beta_nC_n^2=\lambda_2>0, \quad (67)$$

and the following inequality can be obtained:

$$\frac{\ln(2)2^{\frac{1}{Bx_2^*}}}{|h_{k,n}|^2Bx_2^*}-\frac{2^{\frac{1}{Bx_2^*}}-1}{|h_{k,n}|^2}-2\kappa C_n^3>0. \quad (68)$$

It is indicated that the above inequality includes (66), and thus condition (66) can be omitted. By including the expression of x_2^*, the condition in this case is given by

$$D\upsilon_2 \ln(2) 2^{D\upsilon_2} - 2^{D\upsilon_2} + 1 - 2\kappa C_n^3 |h_{k,n}|^2 > 0. \tag{69}$$

3) If $\lambda_2 = 0$ and $\lambda_3 > 0$, by including $x_2^* = \upsilon_1$ to (63), the expression of x_1^* can be presented as follows:

$$x_1^* = (T_n^{\max} - D\upsilon_1) C_n (\mu \beta_n)^{-1}, \tag{70}$$

and (61b) can be rewritten as (65), where the equality condition is removed as it holds for the case $\lambda_2 > 0$. From (60a), the following condition can be obtained:

$$\lambda_1 = 2\kappa C_n^3 (x_1^*)^{-3} > 0, \tag{71}$$

which can be rewritten as follows:

$$D\upsilon_1 < T_n^{\max}. \tag{72}$$

The above inequality is always satisfied with inequality (65). Moreover, the equation in (71) can be substituted into (60b), as shown in follows:

$$-\frac{D\ln(2) 2^{\frac{1}{Bx_2^*}}}{|h_{k,n}|^2 B x_2^*} + \frac{D(2^{\frac{1}{Bx_2^*}} - 1)}{|h_{k,n}|^2} + \frac{2\kappa C_n^3 D}{(x_1^*)^3} = \lambda_3 > 0. \tag{73}$$

By including the obtained solutions of x_1^* and x_2^*, it becomes

$$2^{\frac{1}{B\upsilon_1}} - 1 - \frac{1}{B\upsilon_1} \ln(2) 2^{\frac{1}{B\upsilon_1}} + \frac{2\kappa(\mu\beta_n)^3 |h_{k,n}|^2}{(T_n^{\max} - D\upsilon_1)^3} > 0. \tag{74}$$

4) If $\lambda_2 = 0$ and $\lambda_3 = 0$, the following equation can be obtained from (60a) and (60b):

$$\frac{2\kappa C_n^3}{(x_1^*)^3} = \frac{\ln(2) 2^{\frac{1}{Bx_2^*}}}{B|h_{k,n}|^2 x_2^*} - \frac{2^{\frac{1}{Bx_2^*}} - 1}{|h_{k,n}|^2}. \tag{75}$$

Moreover, by including (63), the solution can be obtained by solving the following equations:

$$\begin{cases} \dfrac{2\kappa C_n^3}{(x_1^*)^3} = \dfrac{\ln(2) 2^{\frac{1}{Bx_2^*}}}{B|h_{k,n}|^2 x_2^*} - \dfrac{2^{\frac{1}{Bx_2^*}} - 1}{|h_{k,n}|^2}, \\ \mu\beta_n C_n^{-1} x_1^* + D x_2^* - T_n^{\max} = 0. \end{cases} \tag{76}$$

This proposition is proved. ■

REFERENCES

[1] K. Wang, Z. Ding, D. K. C. So, and Z. Ding, "Energy efficient federated learning with age-weighted FedSGD," in *2024 IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2024, pp. 457–462.

[2] 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, Dec. 2020.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Stat.* PMLR, Apr. 2017, pp. 1273–1282.

[4] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[5] 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, Feb. 2021.

[6] 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, Dec. 2021.

[7] 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, Jan. 2022.

[8] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *ICC 2019 - 2019 IEEE Int. Conf. Commun. (ICC)*, 2019, pp. 1–7.

[9] L. Fu, H. Zhang, G. Gao, M. Zhang, and X. Liu, "Client selection in federated learning: Principles, challenges, and opportunities," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 21 811–21 819, Dec. 2023.

[10] 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, Mar. 2020.

[11] 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, pp. 269–283, Jan. 2021.

[12] W. Guo, R. Li, C. Huang, X. Qin, K. Shen, and W. Zhang, "Joint device selection and power control for wireless federated learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2395–2410, Aug. 2022.

[13] M. Ribero, H. Vikalo, and G. de Veciana, "Federated learning under intermittent client availability and time-varying communication constraints," *IEEE J. Sel. Top. Signal Process.*, vol. 17, no. 1, pp. 98–111, Jan. 2023.

[14] 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.

[15] T. Gafni, N. Shlezinger, K. Cohen, Y. C. Eldar, and H. V. Poor, "Federated learning: A signal processing perspective," *IEEE Signal Process. Mag.*, vol. 39, no. 3, pp. 14–41, May 2022.

[16] 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.

[17] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Convergence of update aware device scheduling for federated learning at the wireless edge," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3643–3658, June 2021.

[18] B. Luo, W. Xiao, S. Wang, J. Huang, and L. Tassiulas, "Tackling system and statistical heterogeneity for federated learning with adaptive client sampling," in *IEEE INFOCOM 2022 - IEEE Conf. Comput. Commun.*, 2022, pp. 1739–1748.

[19] Y. J. Cho, J. Wang, and G. Joshi, "Towards understanding biased client selection in federated learning," in *Proc. 25th Int. Conf. Artif. Intell. Stat.* PMLR, Mar. 2022, pp. 10 351–10 375.

[20] 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.*, vol. 23, no. 8, pp. 8939–8953, Aug. 2024.

[21] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[22] 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, Mar. 2021.

[23] T. Wang, N. Huang, Y. Wu, and T. Q. S. Quek, "Energy-efficient wireless federated learning: A secrecy oriented design via sequential artificial jamming," *IEEE Trans. Veh. Technol.*, vol. 72, no. 5, pp. 6412–6427, May 2023.

[24] M. Alishahi, P. Fortier, W. Hao, X. Li, and M. Zeng, "Energy minimization for wireless-powered federated learning network with NOMA," *IEEE Wireless Commun. Lett.*, vol. 12, no. 5, pp. 833–837, May 2023.

[25] Y. Ren, C. Wu, and D. K. So, "Joint edge association and aggregation frequency for energy-efficient hierarchical federated learning by deep reinforcement learning," in *ICC 2023 - IEEE Int. Conf. Commun. (ICC)*, 2023, pp. 3639–3645.

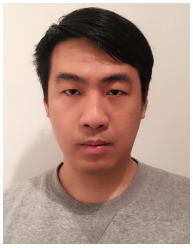
[26] Y. Lin, K. Wang, and Z. Ding, "Rethinking clustered federated learning in NOMA enhanced wireless networks," vol. 23, no. 11, pp. 16 875–16 890, Nov. 2024.

[27] 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 Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 8743–8747.

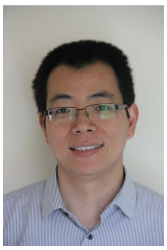
[28] K. Wang, Y. Ma, M. B. Mashhadi, C. H. Foh, R. Tafazolli, and Z. Ding, "Convergence acceleration in wireless federated learning: A stackelberg game approach," *IEEE Trans. Veh. Technol.*, pp. 1–15, Sept. 2024.

[29] B. Wu, F. Fang, and X. Wang, "Joint age-based client selection and resource allocation for communication-efficient federated learning over NOMA networks," *IEEE Trans. Commun.*, vol. 72, no. 1, pp. 179–192, Jan. 2024.

[30] E. Rizk, S. Vlaski, and A. H. Sayed, "Federated learning under importance sampling," *IEEE Trans. Signal Process.*, vol. 70, pp. 5381–5396, Sept. 2022.
[31] S. L. Lohr, *Sampling: design and analysis*. CRC press, 2021.
[32] K. Wang, Z. Ding, D. K. C. So, and G. K. Karagiannidis, "Stackelberg game of energy consumption and latency in MEC systems with NOMA," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2191–2206, Apr. 2021.
[33] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.



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, IEEE SPCC Technical Recognition Award 2021, and IEEE VTS Best Magazine Paper Award 2023. 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 was 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.