Joint User Association and Resource Allocation for Wireless Hierarchical Federated Learning With IID and Non-IID Data

Shengli Liu, Guanding Yu, *Senior Member, IEEE*, Xianfu Chen, *Member, IEEE*, and Mehdi Bennis, *Fellow, IEEE*

***Abstract*— In this work, *hierarchical federated learning* (HFL) over wireless multi-cell networks is proposed for large-scale model training while preserving data privacy. However, the imbalanced data distribution has a significant impact on the convergence rate and learning accuracy. In addition, a large learning latency is incurred due to the traffic load imbalance among *base stations* (BSs) and limited wireless resources. To cope with these challenges, we first provide an analysis of the model error and learning latency in wireless HFL. Then, joint user association and wireless resource allocation algorithms are investigated under *independent identically distributed* (IID) and non-IID training data, respectively. For the IID case, a learning latency aware strategy is designed to minimize the learning latency by optimizing user association and wireless resource allocation, where a mobile device selects the BS with the maximal uplink channel *signal-to-noise ratio* (SNR). For the non-IID case, the total data distribution distance and learning latency are jointly minimized to achieve the optimal user association and resource allocation. The results show that both data distribution and uplink channel SNR should be taken into consideration for user association in the non-IID case. Finally, the effectiveness of the proposed algorithms are demonstrated by the simulations.**

***Index Terms*— User association, hierarchical federated learning, non-IID, data distribution, learning latency.**

I. Introduction

A. Background

RECENTLY, artificial intelligence has played an importance role in many emerging applications, such as automatic driving, face and voice recognition, etc. With massive amounts of data, neural networks can be trained in a centralized way to support these applications [1]. As the data privacy and security concerns increase, *federated learning* (FL) has been proposed for distributed model training [2]. Under FL, clients exchange their model parameters with the edge server rather than uploading raw training data. Wireless FL has also been applied in various applications [3], [4], where mobile devices perform local model updates and transmit the model parameters to the *base station* (BS).

Since learning is based on single BS connectivity, which can be limited, calling for *hierarchical federated learning* (HFL) to fully exploit the training data of mobile devices connected to multiple BSs. After multiple local model updates at each BS, the cloud server aggregates the edge model from BSs to improve the learning performance. As a result, wireless HFL can achieve higher communication efficiency since mobile devices can obtain the global model without directly exchanging model parameters with the cloud server.

However, a large learning latency is still incurred due to the traffic load imbalance and limited wireless resource in wireless HFL. In general, mobile devices are able to communicate with multiple BSs. If more mobile devices are associated with a BS, less wireless resources are allocated to each mobile device, resulting in an increase in communication latency. Then, the global latency will increase owing to the edge model aggregation. In addition, the learning performance, i.e., convergence rate and learning accuracy, is affected by the data distribution imbalance, which may differ from device to device and the overall data distribution of the BS will be highly imbalanced with an improper user association. Therefore, it is critical to develop user association and resource allocation algorithms to improve the learning performance and reduce learning latency as well.

B. Related Works

There have been extensive efforts to analyze and improve the learning performance for both single-layer FL and HFL.

1) Single-Layer Federated Learning: Many existing works have been focused on single-layer FL. Regarding the computation and communication bottlenecks in a single-layer FL, various strategies have been designed to reduce the learning latency and improve the learning performance [5]. In [6]–[9], authors have applied network pruning and gradient compression to reduce latency for local model computation and model uploading. By exploiting the characteristics of the wireless

Manuscript received 18 October 2021; revised 29 January 2022; accepted 20 March 2022. Date of publication 4 April 2022; date of current version 11 October 2022. The work of Guanding Yu was supported by research grant under Grant GDNRC[2021]32. The work of Xianfu Chen was supported by the Zhejiang Laboratory Open Program under Grant 2021LC0AB06. The associate editor coordinating the review of this article and approving it for publication was S. Zhou. *(Corresponding author: Guanding Yu.)*

Shengli Liu is with the School of Information and Electrical Engineering, Zhejiang University City College, Hangzhou 310015, China (e-mail: victoryliu@zju.edu.cn).

Guanding Yu is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: yuguanding@zju.edu.cn).

Xianfu Chen is with the VTT Technical Research Centre of Finland, 90570 Oulu, Finland (e-mail: xianfu.chen@vtt.fi).

Mehdi Bennis is with the Centre for Wireless Communication, University of Oulu, 90540 Oulu, Finland (e-mail: mehdi.bennis@oulu.fi).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2022.3162595.

Digital Object Identifier 10.1109/TWC.2022.3162595

channel, over-the-air computation has been adopted for model aggregation in [10], [11] to improve the spectrum efficiency of wireless FL. Considering limited wireless resources, authors in [12] have proposed a solution based on selecting a part of devices for model aggregation during each iteration. Different scheduling policies have been designed in [13]–[15] to improve the convergence rate and reduce the communication latency. Moreover, the unreliable wireless channel has been taken into consideration [16]. The convergence rate with transmission error has been analyzed and the wireless resource allocation has been optimized to improve the learning performance.

With respect to the data distribution, many works have been devoted to analyzing the learning performance of single-layer FL under *non-independent identically distributed* (non-IID) training data. In [17], the effect of non-IID training data on the learning performance has been evaluated. By analyzing the model error compared with the model trained in a centralized way, the learning performance is determined by the data distribution distance between the client and the whole population. The convergence rate has been analyzed in [18] for FL with non-IID data. The authors in [19] have proposed a deep Q-learning based strategy to select a subset of devices in terms of uploaded weights in non-IID scenarios. By this means, the convergence can be speeded up as compared with the conventional *federated averaging* (FedAvg) algorithm.

2) Hierarchical Federated Learning: To train models on a large scale, HFL has been proposed on the basis of a single-layer FL. In [20], the authors have shown that the communication rounds can be reduced by hierarchical clustering of local updates on non-IID training data. The convergence of multi-level local *stochastic gradient descent* (SGD) on convex and non-convex objective functions has been analyzed for HFL in [21] and [22]. In [23] and [24], the user-edge assignment problem has been proposed for HFL with non-IID training data to improve the learning performance. By analyzing the effect of data distribution on the learning performance, the user-edge assignment has been optimized. However, the learning latency has not been considered in the above works.

For wireless HFL, the authors in [25] and [26] have deployed HFL in multi-layer wireless network to reduce the communication overhead and long latency compared with wireless single-layer FL. A client-edge-cloud HFL system has been proposed in [26], where the edge servers perform partial model aggregation. The proposed system achieves an enhanced learning performance in different data distribution scenarios. Similarly, fog learning has been proposed in [27] to distribute the multi-layer learning architecture over heterogenous wireless networks. A joint resource allocation and edge association problem has been formulated in [28] to improve both the communication efficiency and energy efficiency. Nevertheless, the effect of user association on the convergence rate or learning accuracy has not been considered.

C. Motivations and Contributions

Although there exist several works investigating wireless HFL, two technical challenges remain unsolved. *On the one hand, how to characterize the effect of data distribution and traffic load imbalance on the learning performance when examining HFL in wireless multi-cell networks?* For wireless HFL, the convergence rate, learning accuracy, and learning latency are all important performance indicators that depend on data distribution and traffic load imbalance. However, it is difficult to mathematically derive the effect of such analysis. *On the other hand, how to develop user association and resource allocation schemes to improve the learning performance?* Different from traditional cellular networks, user association and resource allocation depend on more factors, such as local computing power, channel state information, and data distribution. Thus, it is challenging to develop the optimal scheme for wireless HFL. To this end, in this paper, we analyze the learning performance, i.e., model error and learning latency, and derive the impact of user association and resource allocation on the learning performance. Moreover, the local computing power, data distribution, and channel state information are jointly accounted for when designing the optimal user association and resource allocation algorithms for wireless HFL in both the IID and non-IID cases. The main contributions from our work are summarized as follows.

- We study the problem of joint user association and wireless resource allocation in wireless HFL under both IID and non-IID cases, respectively. First, we analyze the learning performance, i.e., model error and learning latency, and characterize the impact of user association and resource allocation on the learning performance. On the one hand, the upper bound of model error is dependent on the data distribution and user association. On the other hand, the learning latency consisting of two parts (local-edge stage and edge-cloud stage), which is affected by the user association and resource allocation, is analyzed.
- For the IID case, the optimal user association and resource allocation are obtained by minimizing the learning latency. The results show that the optimal user association is the same as in a traditional multi-cell network, where mobile devices select the BS with the maximal uplink channel *signal-to-noise ratio* (SNR). In addition, wireless resources are allocated in accordance with both local computing power and uplink channel SNR, which is different from that in traditional throughput-oriented cellular networks.
- For the non-IID case, the weighted sum of total data distribution distance and learning latency is minimized to achieve the optimal user association and resource allocation. Different from the IID case, the solutions account for both data distribution distance and uplink channel SNR for the user association. The proposed user association policy is meaningful from the perspective of demonstrating the importance of rethinking user association for wireless HFL. Finally, numerical simulations are implemented to validate the effectiveness of the proposed algorithms.

D. Organization

The rest of the paper is organized as follows. In Section II, we introduce the system model of wireless HFL. Then, the

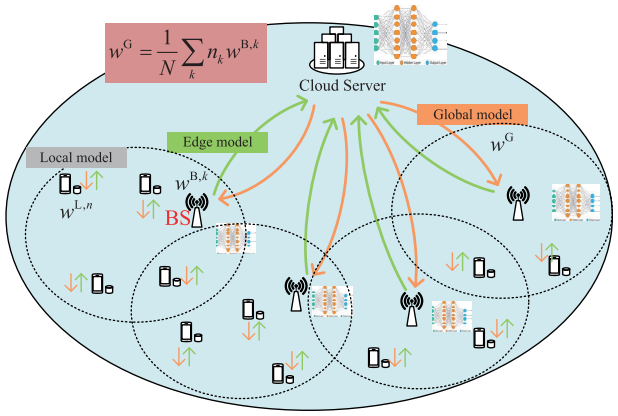


Fig. 1. The wireless HFL system model.

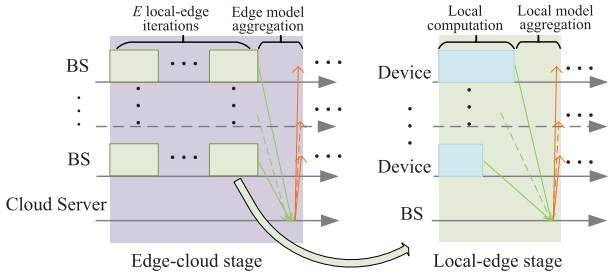


Fig. 2. Synchronous model aggregation in wireless HFL.

model error and learning latency are analyzed in Section III. For the IID case, the learning latency aware strategy is designed in Section IV. In Section V, we jointly minimize the total data distribution distance and learning latency to obtain the user association and resource allocation for the non-IID case. Experimental results are presented to verify the proposed algorithm in Section VI. Finally, Section VII concludes the whole paper.

II. System Model

As depicted in Fig. 1, we consider a wireless HFL scenario where multiple BSs and multiple mobile devices (e.g., mobile phone, laptop, and pad) collaboratively participate in training a neural network model, such as image classification or recognition. There exist overlapping areas among BSs, where mobile devices are uniformly distributed. Therefore, some mobile devices are able to access multiple BSs simultaneously. During model training, each device selects one BS to exchange the model parameters. Each BS sends the edge model to the cloud server for global model aggregation.

The number of classes in the classification or recognition task is C, for which we denote the set $\mathcal{C} = \{1, 2, \ldots, C\}$. Let $\mathcal{K} = \{1, 2, \ldots, K\}$ and $\mathcal{N} = \{1, 2, \ldots, N\}$ denote the sets of BS and mobile devices, respectively. There are N_k mobile devices located in the coverage of BS k, the set of which is denoted as $\mathcal{N}_k$. In addition, for the device n, the available BSs are denoted by a set $\mathcal{K}_n$. We adopt the indicator variable $a_{n,k} \in \{0, 1\}$ to represent the association between a BS k and a device n. If $a_{n,k} = 1$, the device n is served by the BS k to support the local model update, and vice versa. Each mobile device has the same amount of training data [29], denoted as $\mathcal{D}_n$. This assumption is made for the convenience of analysis. The proposed policies and algorithms are still applicable for these general cases where devices have different amounts of training data.

A. Hierarchical Federated Learning Model

For any neural network, model training aims to find an optimal function $\mathcal{H}_{\boldsymbol{w}} : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is the training data, $\mathcal{Y}$ is the ground-truth label, and $\boldsymbol{w}$ represents the model parameter. By minimizing the distance, $f\left(\mathcal{H}_{\boldsymbol{w}}\left(\mathcal{X}\right), \mathcal{Y}\right)$, between the network output $\mathcal{H}_{\boldsymbol{w}}\left(\mathcal{X}\right)$ and the label $\mathcal{Y}$, the optimal model $\boldsymbol{w}^*$ can be obtained. Therefore, the model training can be formulated as

$$\min_{\boldsymbol{w}} F\left(\boldsymbol{w}\right) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} f\left(\mathcal{H}_{\boldsymbol{w}}\left(x\right), y\right). \tag{1}$$

When the model is trained in a centralized way, SGD can be adopted. Similarly, FedSGD and FedAvg are proposed for FL [30]. Model parameters or gradients are exchanged between mobile devices and the edge server for model aggregation and update. Specifically, for FedSGD, the model aggregation is performed at each iteration. As for FedAvg, the model aggregation occurs after multiple local model updates. Without loss of generality, we consider a two-layer FL in this work, as in Fig. 1. The global model aggregation is divided into two parts: local-edge model aggregation and edge-cloud model aggregation, as shown in Fig. 2. Both FedSGD and FedAvg are adopted to obtain the global model. Since BSs are closer to mobile devices than the cloud server in general, the communication cost and transmission latency are low. Therefore, to facilitate the local model update, we use FedSGD to aggregate the local model at each iteration. Unlike local-edge aggregation, FedAvg is adopted for edge-cloud model aggregation to avoid frequent communication between the BS and core network. After E local iterations, all BSs upload the edge model to the cloud server. A similar HFL model is also adopted in [23], [25], [28].

1) Local-Edge Model Aggregation: For each local iteration, mobile devices should upload the local model parameters to its associated BS after one round local model update. Then, the BS aggregates the local model and broadcasts the edge model to its associated mobile devices.

Denote the local model of device n at the mE-th local iteration as $\boldsymbol{w}_{mE}^{\mathrm{L},n}$. Note that m represents the number of global iterations. Hence, the corresponding number of local iterations is mE. The local model update can be written as

$$\boldsymbol{w}_{mE+1}^{\mathrm{L},n} = \boldsymbol{w}_{mE}^{\mathrm{L},n} + \alpha \sum_{c} p^n\left(c\right) \nabla_{\boldsymbol{w}} \mathbb{E}_{x|y(c)} \left\{\log\left(\mathcal{H}_{\boldsymbol{w}_{mE}^{\mathrm{L},n}}\left(x\right)\right)\right\}, \tag{2}$$

where α is the learning rate, $p^n\left(c\right)$ is the data distribution of device n on the class c, and $\nabla_{\boldsymbol{w}} \mathbb{E}_{x|y(c)} \left\{\log\left(\mathcal{H}_{\boldsymbol{w}_{mE}^{\mathrm{L},n}}\left(x\right)\right)\right\}$ is the gradient calculated based on a batch of $\mathcal{D}_n$. Note that there are many ways to evaluate the data distribution, such as feature distribution and label distribution [31], [32]. In this

work, we mainly adopt the label distribution. Thus, the data distribution of device n is defined as the proportion of each class in the dataset $\mathcal{D}_n$, which is widely adopted in [17]–[19]. Specifically, denote D_n as the data volume owned by device n, and $d_{n,c}$ as the data volume of class c. Then, the data distribution can be written as

$$p^n(c) = \frac{d_{n,c}}{D_n}. \tag{3}$$

In this work, we concentrate on the popular learning task, i.e., classification task, which adopts cross entropy as the loss function in general. Other loss functions for regression problems, such as MSE, are not considered in this work. For the cross entropy with regularizer, our proposed algorithm can also be applicable. Moreover, in practical HFL where the gradient cannot be accurately derived, the learning performance degrades compared with the accurate case. However, the effect of data distribution remains unchanged and our proposal can still be applied. Denote the edge model of BS k at the $(mE+1)$-th iteration as $\boldsymbol{w}^{\mathrm{B},k}_{mE+1}$. Then, the edge model average aggregation can be written as

$$\boldsymbol{w}^{\mathrm{B},k}_{mE+1} = \frac{1}{n_k} \sum_{n \in \mathcal{N}_k} a_{n,k} \boldsymbol{w}^{\mathrm{L},n}_{mE+1}, \tag{4}$$

where n_k is the traffic load of BS k and $n_k = \sum_{n \in \mathcal{N}_k} a_{n,k}$. After that, the local model $\boldsymbol{w}^{\mathrm{L},n}_{mE+1}$ is updated with the edge model $\boldsymbol{w}^{\mathrm{B},k}_{mE+1}$.

2) Edge-Cloud Model Aggregation: Denote the global model at the $((m+1)E)$-th local iteration as $\boldsymbol{w}^{\mathrm{G}}_{(m+1)E}$. According to FedAvg, the global model aggregation can be expressed as

$$\boldsymbol{w}^{\mathrm{G}}_{(m+1)E} = \frac{1}{\sum_k n_k} \sum_k n_k \boldsymbol{w}^{\mathrm{B},k}_{(m+1)E}. \tag{5}$$

For the single-connectivity scenario where each mobile device can connect only one BS, $\sum_k n_k = N$. In summary, the one global iteration training process for HFL is presented in Algorithm 1.

Algorithm 1 Hierarchical Federated Learning Algorithm

From global iteration m to iteration $(m+1)$:

1: **for** local iteration $i = 1 : E$ **do**
2: Each device obtains the local model $\boldsymbol{w}^{\mathrm{L},n}_{mE+i}$ based on (2).
3: The local model $\boldsymbol{w}^{\mathrm{B},k}_{mE+i}$ can be aggregated by each BS according to (4).
4: Each device updates the local model with $\boldsymbol{w}^{\mathrm{B},k}_{mE+i}$.
5: **end for**
6: The global model $\boldsymbol{w}^{\mathrm{G}}_{(m+1)E}$ can be achieved by the cloud server based on (5).
7: The local and edge model are updated with $\boldsymbol{w}^{\mathrm{G}}_{(m+1)E}$.

B. Transmission Model

In wireless HFL, the model parameters are exchanged among mobile devices, BSs, and the cloud server. For local-edge model aggregation, the model parameters are transmitted over the wireless channels. We assume that BSs use multiple orthogonal narrowband channels. Thus, the interference among BSs is ignored.[1] Specifically, let B^{U}_k denote the uplink bandwidth planned for the BS k. In each BS, mobile devices can share the wireless channel for uploading local model with a multiple access mechanism. Without loss of generality, *orthogonal frequency-division multiple access* (OFDMA) is adopted in this paper. Therefore, the uplink data rate of device n associated with the BS k is

$$r^{\mathrm{U}}_{n,k} = a_{n,k} l_{n,k} B^{\mathrm{U}}_k R^{\mathrm{U}}_{n,k}, \tag{6}$$

where $l_{n,k}$ is the uplink bandwidth fraction allocated for the device n and $R^{\mathrm{U}}_{n,k}$ is the uplink spectrum efficiency. The efficiency is $R^{\mathrm{U}}_{n,k} = \log_2\left(1 + p^{\mathrm{U}}_n h^{\mathrm{U}}_{n,k}/N_0\right)$, where p^{U}_n is the transmission power of the device n, $h^{\mathrm{U}}_{n,k}$ is the uplink channel power gain of the device n, and N_0 is the channel noise variance [28], [34].

Regarding the edge model broadcasting, all available downlink channels can be used by the BS for broadcasting the edge model to its associated devices. Compared with the local model uploading, the latency for edge model broadcasting can be ignored since the BSs have adequate transmission power and broadcast bandwidth in general [28], [35], [36]. Hence, the analysis for the downlink data rate is omitted in this work.

For edge-cloud model aggregation, the BSs and cloud server exchange the global model over the wired or wireless backhaul. Denote r_k as the uplink data rate of the BS k. The cloud server can broadcast the global model to the BSs over all bandwidth of the backhaul. Owing to the sufficient transmission resource for global model broadcasting, the latency for global model broadcasting is neglected and the corresponding data rate analysis is omitted as well.

III. Learning Performance Analysis

In this section, we will analyze the learning performance, i.e., model error and learning latency, taking into account the effect of the imbalanced data distribution and traffic load. Note that the analysis on the model error and learning latency is applicable for both the IID and non-IID cases.

A. Model Error Analysis

Regarding the data distribution imbalance, the model error compared with the optimal model trained with IID data in a centralized way can be adopted to evaluate the convergence rate and learning accuracy in wireless HFL system [17], [18], [23]. Denote the optimal model at the (mE)-th iteration as $\boldsymbol{w}^*_{mE}$. Then, the model error can be expressed as $\left\|\boldsymbol{w}^{\mathrm{G}}_{mE} - \boldsymbol{w}^*_{mE}\right\|$. The optimal model can be also derived by

[1] We make this assumption to derive the optimal wireless resource allocation and user association strategy in closed-form. However, our work can be still extended into the more general scenario with frequency reuse by exploring these existing channel reuse and interference management techniques in multi-cell networks [33].

the gradient descent method, as

$$\boldsymbol{w}^*_{mE}=\boldsymbol{w}^*_{mE-1}+\alpha\sum_c p\left(c\right)\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}^*_{mE-1}}\left(x\right)\right)\right\}, \tag{7}$$

where $p\left(c\right)$ is the data distribution of IID training data. Since we adopt label distribution to evaluate the data distribution, having the IID training data means that each class has a uniform distribution for a typical classification task [17], [31]. In particular, $p\left(c\right)=\frac{1}{C}, c\in\mathcal{C}$. Therefore, $p\left(c\right)$ is determined by the specific learning task, without the need of aggregating the local data. The smaller the model error is, the better the learning performance will be. By analyzing the model error, the effect of user association and data distribution on the learning performance can be derived. To analyze the model error, the following assumptions are made.

- (Gradient smoothness) The gradient $\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}}\left(x\right)\right)\right\}$ is $L\left(c\right)$-Lipschitz smooth, which can be written as

$$\begin{aligned}&\left\|\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}_1}\left(x\right)\right)\right\}-\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}_2}\left(x\right)\right)\right\}\right\|\\&\quad\leq L\left(c\right)\left\|\boldsymbol{w}_1-\boldsymbol{w}_2\right\|,\end{aligned} \tag{8}$$

  where $L\left(c\right)$ is the Lipschitz constant.
- (Bounded gradient) The mE-th iteration gradient $\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}^{\mathrm{B},k}_{mE}}\left(x\right)\right)\right\}$ is bounded, which can be expressed as

$$\left\|\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}^{\mathrm{B},k}_{mE}}\left(x\right)\right)\right\}\right\|\leq A_{mE}. \tag{9}$$

Then, the upper bound of model error compared with the optimal model after one global iteration can be given in the following theorem.

Theorem 1: For the wireless HFL with N mobile devices, K BSs, C classes, and E synchronization iterations, the upper bound of the model error between the global model and the optimal model after one global iteration can be expressed as (10)*, shown at the bottom of the page, where $p\left(c\right)$ is the data distribution of IID training data.*

Proof: Please refer to Appendix A. ∎

Remark 1: From the theorem, the upper bound of the model error increases with the total data distribution distance from the IID training data, $\frac{1}{N}\sum_k\sum_c\left\|\sum_{n\in\mathcal{N}_k}a_{n,k}\left(p^n\left(c\right)-p\left(c\right)\right)\right\|$. Specifically, a large total data distribution distance causes a low convergence rate and learning accuracy. In addition, user association has a significant impact on the total data distribution distance. Due to the data distribution imbalance among mobile devices, a large total data distribution distance may be caused with an improper user association. Therefore, user association should be adjusted to alleviate the effect of data distribution imbalance by reducing the total data distribution distance. In particular, when the training data of mobile devices are IID, the total distribution distance is zero and user association will not affect the convergence rate and learning accuracy.

B. Learning Latency Analysis

Due to the limited wireless resources, the learning latency is the main bottleneck of wireless HFL. We adopt the synchronous model aggregation mechanism in both the local-edge aggregation stage and the edge-cloud aggregation stage, as described in Fig. 2.

1) Local-Edge Aggregation Stage: At this stage, the learning latency is composed of two parts: the latency for local model computation and the latency for local model uploading. Since the edge server equipped at the BS has in general rich computing resources, the latency for edge model aggregation is ignored in this paper.

- Local model computation: The latency for local model computation can be expressed as

$$t^{\mathrm{L}}_n=\frac{bd^{\mathrm{L}}}{f_n}, \tag{11}$$

  where b is the batchsize, d^{L} is the CPU cycle for one training data calculation, including both forward and back propagation, and f_n is the n-th device's computing power evaluated by the CPU frequency. Here, d^{L} is determined by the learning task and the computing power of mobile device.
- Local model uploading: Denote the data volume of the model parameters as M. Then, the latency for model uploading from device n to BS k can be expressed as

$$t^{\mathrm{U}}_{n,k}=\frac{M}{r^{\mathrm{U}}_{n,k}}. \tag{12}$$

Due to the synchronous model aggregation mechanism, the latency at the local-edge stage is determined by the maximum latency among all devices. Accordingly, the learning latency at the local-edge stage can be written as

$$t^{\mathrm{E}}=\max\left\{a_{n,k}\left(t^{\mathrm{L}}_n+t^{\mathrm{U}}_{n,k}\right)\right\}. \tag{13}$$

2) Edge-Cloud Aggregation Stage: The latency at this stage is mainly determined by the edge model uploading. Due to the high computing power of the cloud server, the latency for

$$\begin{aligned}\left\|\boldsymbol{w}^{\mathrm{G}}_{mE}-\boldsymbol{w}^*_{mE}\right\|\leq&\left(1+\alpha\sum p\left(c\right)L\left(c\right)\right)^E\left\|\boldsymbol{w}^{\mathrm{G}}_{(m-1)E}-\boldsymbol{w}^*_{(m-1)E}\right\|\\&+\alpha\left(\sum_{i=0}^{E-1}A_{mE-1-i}\left(1+\alpha\sum_c p\left(c\right)L\left(c\right)\right)^i\right)\underbrace{\frac{1}{N}\sum_k\sum_c\left\|\sum_{n\in\mathcal{N}_k}a_{n,k}\left(p^n\left(c\right)-p\left(c\right)\right)\right\|}_{\text{effect of data distribution imbalance}}.\end{aligned} \tag{10}$$

global model aggregation is neglected as well. The latency for the edge model uploading can be expressed as

$$t^{\mathrm{G}} = \max\left\{\frac{M}{r_k}\right\}. \tag{14}$$

To facilitate the analysis of learning latency at each iteration, we divide one global iteration into E local iterations. Therefore, the learning latency of one local iteration can be written as

$$T = t^{\mathrm{E}} + \frac{1}{E}t^{\mathrm{G}}. \tag{15}$$

A similar learning latency analysis method is also adopted in [25]. From the latency analysis, local uplink transmission is the main bottleneck. Due to the limited wireless resource for each BS, a large transmission latency will be caused if the traffic load is unevenly distributed across BSs. Thus, user association and wireless resource allocation should be carefully designed to mitigate the impact of traffic load imbalance, and eventually reduce the overall learning latency.

According to the learning performance analysis, the effect of user association and resource allocation on the learning performance differs from the IID case to the non-IID case. When the training data of mobile device is IID, user association and resource allocation only influence the learning latency without impacting the model error. However, both model error and learning latency are determined by user association and resource allocation under non-IID data.

IV. Optimal User Association and Resource Allocation: The IID Case

In this section, an optimization problem is first formulated for wireless HFL with IID data to improve the learning latency. Then, the optimal wireless resource allocation and user association algorithm is developed.

A. Problem Formulation

As analyzed above, improper user association and resource allocation will cause a large learning latency while not affecting the model error in the IID case. Therefore, the learning latency can be minimized to improve the learning performance of wireless HFL with IID data by jointly optimizing user association and resource allocation while the effect of HFL convergence is not considered in this problem. The problem can be mathematically formulated as

$$\min_{\left\{a_{n,k},l_{n,k},B_k^{\mathrm{U}}\right\}} T, \tag{16}$$

$$\text{subject to} \sum_{n\in\mathcal{N}_k} a_{n,k}l_{n,k} \le 1, \quad k\in\mathcal{K}, \tag{16a}$$

$$\sum_k B_k^{\mathrm{U}} \le B^{\mathrm{U}}, \tag{16b}$$

$$\sum_{k\in\mathcal{K}_n} a_{n,k} = 1, \tag{16c}$$

$$a_{n,k}\in\{0,1\}, \tag{16d}$$

where B^{U} is the total uplink bandwidth for all BSs, (16a) and (16b) are the constraints related to the resource allocation, and (16c) and (16d) are the constraints on user association. It is obvious that this problem is a *mixed integer nonlinear programming* (MINLP) problem. Thus, the optimal solution cannot be directly obtained in general. Fortunately, this problem can be decomposed into two subproblems by separating integer variables from continuous variables and by performing wireless resource allocation and user association, respectively.

B. Optimal Wireless Resource Allocation

Given the user association $a_{n,k}, n\in\mathcal{N}, k\in\mathcal{K}$, the problem (16) is transformed into minimizing the learning latency at the local-edge stage according to the learning latency analysis, which can be written as

$$\min_{\left\{l_{n,k},B_k^{\mathrm{U}},t^{\mathrm{E}}\right\}} t^{\mathrm{E}}, \tag{17}$$

$$\text{subject to (16a), (16b), and } a_{n,k}\left(\frac{bd^{\mathrm{L}}}{f_n}+\frac{M}{r_{n,k}^{\mathrm{U}}}\right)\le t^{\mathrm{E}}, \quad n\in\mathcal{N},\ k\in\mathcal{K}. \tag{17a}$$

As in Lemma 1 below, the above problem is convex.

Lemma 1: The problem in (17) *is a convex optimization problem.*

Proof: Please refer to Appendix B. ∎

Accordingly, the optimal resource allocation can be derived as in the following theorem.

Theorem 2: Given the user association $a_{n,k}, n\in\mathcal{N}, k\in\mathcal{K}$, the optimal resource allocation $l_{n,k}^*$ and B_k^{U*} can be written as

$$l_{n,k}^* = a_{n,k}\frac{M}{\left(t^{\mathrm{E}*}-\frac{bd^{\mathrm{L}}}{f_n}\right)B_k^{\mathrm{U}*}\log_2\left(1+\frac{p_n^{\mathrm{U}}h_{n,k}^{\mathrm{U}}}{N_0}\right)}, \tag{18}$$

and

$$B_k^{\mathrm{U}*} = \sum_{n\in\mathcal{N}_k} a_{n,k}\frac{M}{\left(t^{\mathrm{E}*}-\frac{bd^{\mathrm{L}}}{f_n}\right)\log_2\left(1+\frac{p_n^{\mathrm{U}}h_{n,k}^{\mathrm{U}}}{N_0}\right)}, \tag{19}$$

respectively, where $t^{\mathrm{E}*}$ is the optimal learning latency at the local-edge stage and can be obtained from

$$\sum_k\sum_n a_{n,k}\frac{M}{\left(t^{\mathrm{E}*}-\frac{bd^{\mathrm{L}}}{f_n}\right)\log_2\left(1+\frac{p_n^{\mathrm{U}}h_{n,k}^{\mathrm{U}}}{N_0}\right)} = B^{\mathrm{U}}. \tag{20}$$

Proof: Please refer to Appendix C. ∎

Remark 2: From Theorem 2, resource allocation for a given device is determined by both the local computing power and channel SNR. Mobile device with a high computing power or uplink channel SNR will be allocated with less wireless resources. Note that this resource allocation policy is different from that in traditional throughput-oriented cellular networks. Furthermore, to reduce learning latency, more wireless resources should be allocated to those devices with a low computing power or uplink channel SNR. In addition, user-balancing constraint is not necessary here since we assume that resource allocation is performed among different BSs. As a result, load balance can be achieved by resource allocation among BSs. For the BSs associated with more mobile devices, more uplink channels should be reserved for these BSs to reduce the local model uploading latency.

C. Optimal User Association

With the optimal wireless resource allocation, the problem in (17) can be rewritten as

$$\min_{\{t^{\mathrm{E}*},a_{n,k}\}} t^{\mathrm{E}*}, \tag{21}$$

subject to (16c), (16d), and

$$\sum_k \sum_n a_{n,k} \frac{M}{\left(t^{\mathrm{E}*} - \frac{bd^{\mathrm{L}}}{f_n}\right) \log_2\left(1 + \frac{p_n^{\mathrm{U}} h_{n,k}^{\mathrm{U}}}{N_0}\right)} \leq B^{\mathrm{U}}. \tag{21a}$$

Note that the problem in (21) is equivalent to the problem in (16) since we just replace the optimal resource allocation with $a_{n,k}$ and $t^{\mathrm{E}*}$. The optimal user association can be achieved by the following theorem.

Theorem 3: For wireless HFL with IID training data, the optimal user association of mobile device n is

$$k_n^* = \arg\max_{k \in \mathcal{K}_n} \left\{ \log_2\left(1 + \frac{p_n^{\mathrm{U}} h_{n,k}^{\mathrm{U}}}{N_0}\right) \right\}. \tag{22}$$

Proof: Based on (16c) and (16d), (21a) can be rewritten as

$$\sum_n \frac{M}{\left(t^{\mathrm{E}*} - \frac{bd^{\mathrm{L}}}{f_n}\right) \sum_k a_{n,k} \log_2\left(1 + \frac{p_n^{\mathrm{U}} h_{n,k}^{\mathrm{U}}}{N_0}\right)} \leq B^{\mathrm{U}}. \tag{23}$$

To obtain the minimal learning latency $t^{\mathrm{E}*}$, $\sum_k a_{n,k} \log_2\left(1 + \frac{p_n^{\mathrm{U}} h_{n,k}^{\mathrm{U}}}{N_0}\right)$ should be the maximum. Therefore, the device n should be associated with the BS with the highest uplink channel SNR, which ends the proof. ∎

Remark 3: From Theorem 3, the optimal user association strategy for wireless HFL with IID training data is the same with that in traditional wireless networks. The uplink channel SNR is the most important factor for the user association. Different from the scheduling problem in wireless single-layer FL, the local computing power doesn't need to be considered in the user association.

The optimal user association and wireless resource allocation for HFL with IID training data is concluded in Algorithm 2. Note that different from the traditional iterative algorithm, we decompose the problem by replacing the optimal resource allocation with $a_{n,k}$ and $t^{\mathrm{E}*}$. Therefore, the optimal solutions can be directly obtained without multiple iterations between these two subproblems and remain the same with that of original problem in (16).

V. Optimal User Association and Resource Allocation: The Non-IID Case

In this section, an optimization problem is formulated to improve the learning performance by considering both the total data distribution distance and learning latency. After that, an algorithm for optimal wireless resource allocation and user association is developed.

A. Problem Formulation

According to the previous analysis, both the upper bound of model error and the learning latency with non-IID training data are affected by user association and resource allocation. That is, the learning performance depends on both the upper bound of model error and the learning latency. Therefore, the weighted sum of the total data distribution distance and the learning latency can be adopted to evaluate the learning performance, expressed as,

$$I = \beta \frac{1}{N} \sum_k \sum_c \left\| \sum_{n \in \mathcal{N}_k} a_{n,k} \left(p^n(c) - p(c)\right) \right\| + (1 - \beta) T, \tag{24}$$

where β is the weight coefficient to balance the importance of the total data distribution distance and the learning latency.[2] Note that the smaller the I is, the better the learning performance will be. To improve learning performance in wireless HFL system, we should minimize I by jointly optimizing user association and resource allocation. Therefore, the optimization problem to achieve the optimal user association and resource allocation can be formulated as

$$\min_{\{a_{n,k}, l_{n,k}, B_k^U\}} I, \tag{25}$$

subject to (16a), (16b), (16c), and (16d). This problem is also an MINLP problem and NP-hard. Thus, it is hard to find its optimal solution. Since the total data distribution distance is not affected by the wireless resource allocation, this problem can be decomposed into two subproblems: wireless resource allocation and user association.

Algorithm 2 Latency Aware User Association Algorithm for the IID Scenario

Input: $h_{n,k}^{\mathrm{U}}$, p_n^{U}, N_0.

Output: User association, $a_{n,k}^*$, wireless resource allocation, $l_{n,k}^*$ and $B_k^{\mathrm{U}*}$.

1: Calculate the uplink channel SNR $\frac{p_n^{\mathrm{U}} h_{n,k}^{\mathrm{U}}}{N_0}$.
2: Obtain the optimal user association $a_{n,k}^*$ according to Theorem 3.
3: Determine the wireless resource allocation $l_{n,k}^*$ and $B_k^{\mathrm{U}*}$ as (18) and (19).

B. Wireless Resource Allocation and User Association

Given the user association $a_{n,k}$, the problem for wireless resource allocation in HFL is the same under both IID and non-IID data. Therefore, the optimal wireless resource allocation can be achieved according to Theorem 2.

With the optimal resource allocation, the problem in (25) can be rewritten as

$$\min_{\{a_{n,k}, t^{\mathrm{E}'}\}} \beta \frac{1}{N} \left(\sum_k \sum_c \left\| \sum_{n \in \mathcal{N}_k} a_{n,k} \left(p^n(c) - p(c)\right) \right\| \right) + (1 - \beta) \left(t^{\mathrm{E}'} + \frac{1}{E} t^{\mathrm{G}} \right), \tag{26}$$

[2]We suggest that the weight coefficient β should be set to ensure that the total data distribution distance and the learning latency are within the same order of magnitude to achieve a better learning performance, as shown in simulations.

subject to (16c), (16d), and

$$\sum_{n}\sum_{k} a_{n,k}\frac{M}{\left(t^{\mathrm{E}'}-\frac{bd^{\mathrm{L}}}{f_n}\right)\log_2\left(1+\frac{p_n^{\mathrm{U}}h_{n,k}^{\mathrm{U}}}{N_0}\right)}\leq B^{\mathrm{U}}. \tag{26a}$$

Here, $t^{\mathrm{E}'}$ is the optimal local-edge latency obtained by optimizing the wireless resource allocation under the given $a_{n,k}$. Similarly, this problem is equivalent to the problem in (25) by substituting $a_{n,k}$ and $t^{\mathrm{E}'}$ for $l_{n,k}$ and B_k^U, respectively. It is still very hard to solve this problem due to the binary indicators, $a_{n,k}$. However, compared with the problem (25), there are fewer optimization variables since the variables about wireless resource allocation are replaced with the local-edge latency, and thereby the computational complexity is reduced. By further observation, the optimal user association can be achieved under the given $t^{\mathrm{E}'}$. By introducing a new variable $q_{k,c}$ to remove the norm operation in (26), the problem under the given $t^{\mathrm{E}'}$ can be rewritten as

$$\min_{\{a_{n,k},q_{k,c}\}}\left\{\beta\frac{1}{N}\sum_{k}\sum_{c}q_{k,c}+(1-\beta)\left(t^{\mathrm{E}'}+\frac{1}{E}t^{\mathrm{G}}\right)\right\}, \tag{27}$$

subject to (26a), (16c), (16d), and

$$q_{k,c}\geq\sum_{n}a_{n,k}\left(p^n(c)-p(c)\right),\quad k\in\mathcal{K},\ c\in\mathcal{C}, \tag{27a}$$

$$q_{k,c}\geq\sum_{n}a_{n,k}\left(p(c)-p^n(c)\right),\quad k\in\mathcal{K},\ c\in\mathcal{C}. \tag{27b}$$

The problem in (27) is a *mixed integer linear programming* (MILP) problem. The optimal solution to user association under the given optimal resource allocation can be achieved by the traditional algorithm, i.e., the branch-and-bound algorithm. A low-complexity algorithm, such as interior point method, can also be used to achieve the sub-optimal user association by variable relaxation.

To obtain the insightful conclusions about the user association, the problem in (27) can be transformed with Lagrangian relaxation into (28), shown at the bottom of the page, subject to (16c) and (16d), where $\lambda_{k,c}$, $\mu_{k,c}$, and γ are nonnegative Lagrangian multipliers related to the constraints (27a), (27b), and (26a), respectively. From the Lagrangian function, the conclusion about the optimal user association can be directly obtained in the following lemma.

Lemma 2: For wireless HFL with non-IID data, the optimal user association of the device n under the given wireless resource allocation satisfies (29)*, shown at the bottom of the page, where $\lambda_{k,c}^*$, $\mu_{k,c}^*$, and γ^* are the corresponding optimal Lagrangian multipliers.*

Proof: Please refer to Appendix D. ■

Remark 4: From Lemma 2, user association is determined by both the uplink channel SNR and data distribution in the non-IID case. $\sum_{c}\left(\lambda_{k,c}^*-\mu_{k,c}^*\right)\left(p^n(c)-p(c)\right)$ can be interpreted as the weighted data distribution distance when device n is associated with BS k. A small weighted data distribution distance means that the device n covers more data required by the BS k or the data distribution of this device is close to IID. Therefore, from the perspective of data distribution, the probability that the device n associates with the BS k decreases with the increase of the weighted data distribution distance. Different from the traditional multi-cell networks and wireless HFL under the IID case, the device should select the BS with a high uplink channel SNR and a small weighted data distribution distance to improve the learning performance. This proposed policy is meaningful and can demonstrate the importance of rethinking user association for wireless HFL.

To achieve optimal Lagrangian multipliers, $\lambda_{k,c}^*$, $\mu_{k,c}^*$, and γ^*, the primal-dual method can be applied [34], [37], [38]. The detailed procedures can be summarized in Algorithm 3.

C. Proposed Algorithm and Computational Complexity Analysis

In line with the above analysis, the user association under the given local-edge latency $t^{\mathrm{E}'}$ can be obtained by the branch-and-bound algorithm or interior point method. For the global solutions of (26), one-dimensional search method can be applied in the interval $[t_{\min}^{\mathrm{E}'}, t_{\max}^{\mathrm{E}'}]$. Herein, $t_{\min}^{\mathrm{E}'}$ and $t_{\max}^{\mathrm{E}'}$ are the minimal and maximal local-edge latency and can be obtained by max-SNR based and min-SNR based user association, respectively. The optimal resource allocation can be directly obtained according to Theorem 2, without multiple

$$\begin{aligned}L\left(\lambda_{k,c},\mu_{k,c},\gamma,q_{k,c},a_{n,k}\right)=&\ \beta\frac{1}{N}\sum_{k}\sum_{c}q_{k,c}\left(1-\lambda_{k,c}-\mu_{k,c}\right)-\gamma B^{\mathrm{U}}+(1-\beta)\left(t^{\mathrm{E}'}+\frac{1}{E}t^{\mathrm{G}}\right)\\&+\sum_{k}\sum_{n}a_{n,k}\left(\gamma\frac{M}{\left(t^{\mathrm{E}'}-\frac{bd^{\mathrm{L}}}{f_n}\right)\log_2\left(1+\frac{p_n^{\mathrm{U}}h_{n,k}^{\mathrm{U}}}{N_0}\right)}+\sum_{c}\left(\lambda_{k,c}-\mu_{k,c}\right)\left(p^n(c)-p(c)\right)\right).\end{aligned} \tag{28}$$

$$k_n^*=\arg\min_{k\in\mathcal{K}_n}\left\{\gamma^*\frac{M}{\left(t^{\mathrm{E}'}-\frac{bd^{\mathrm{L}}}{f_n}\right)\log_2\left(1+\frac{p_n^{\mathrm{U}}h_{n,k}^{\mathrm{U}}}{N_0}\right)}+\sum_{c}\left(\lambda_{k,c}^*-\mu_{k,c}^*\right)\left(p^n(c)-p(c)\right)\right\}. \tag{29}$$

Algorithm 3 Primal-Dual Algorithm for Optimal User Association

Input: $t^{\mathrm{E}'}$, $p^n(c)$, $p(c)$, and other system parameters.
Output: User association, $a^*_{n,k}$, Lagrangian multipliers, $\lambda^*_{k,c}$, $\mu^*_{k,c}$, and γ^*.
1: Initialize $\lambda^{(0)}_{k,c}$, $\mu^{(0)}_{k,c}$, $\gamma^{(0)}$, $s=1$, and the step size η_λ, η_μ, and η_γ.
2: **repeat**
3: Obtain $a^*_{n,k}$ and $q^*_{k,c}$ according to (44) and (45), respectively.
4: Update Lagrangian multipliers as $\lambda^{(s)}_{k,c} = \lambda^{(s-1)}_{k,c} + \eta_\lambda \frac{\partial L}{\partial \lambda_{k,c}}$, $\mu^{(s)}_{k,c} = \mu^{(s-1)}_{k,c} + \eta_\mu \frac{\partial L}{\partial \mu_{k,c}}$, and $\gamma^{(s)} = \gamma^{(s-1)} + \eta_\gamma \frac{\partial L}{\partial \gamma}$, respectively.
5: $s = s+1$.
6: **until** Convergence
7: Obtain the optimal user association $a^*_{n,k}$ and Lagrangian multipliers, $\lambda^*_{k,c}$, $\mu^*_{k,c}$, and γ^*.

iterations between these two subproblems. Accordingly, the joint user association and resource allocation algorithm can be summarized in Algorithm 4. Note that the algorithm should be performed at each iteration due to the dynamic wireless environment and computing power of mobile devices, as well as the algorithm for the IID case.

Algorithm 4 Joint User Association and Resource Allocation Algorithm for the Non-IID Case

Input: The interval $[t^{\mathrm{E}'}_{\min}, t^{\mathrm{E}'}_{\max}]$ and other system parameters.
Output: User association, $a^*_{n,k}$, wireless resource allocation, $l^*_{n,k}$ and $B^{\mathrm{U}*}_k$.
1: Initialize the best objective value I^* with an enough large value.
2: Initialize the optimal user association $a^*_{n,k}$.
3: Initialize $t^{\mathrm{E}'} = t^{\mathrm{E}'}_{\max}$.
4: **repeat**
5: Obtain the optimal user association $a_{n,k}$ by solving the problem (27) under $t^{\mathrm{E}'}$.
6: Obtain the objective value I and $t^{\mathrm{E}'}$ with $a_{n,k}$.
7: Set $t^{\mathrm{E}'} = t^{\mathrm{E}'} - \delta$. /* δ is a very small positive value. */
8: **if** $I < I^*$ **then**
9: $I^* = I$.
10: $a^*_{n,k} = a_{n,k}$.
11: **end if**
12: **until** $t^{\mathrm{E}'} \le t^{\mathrm{E}'}_{\min}$
13: Obtain the optimal user association $a^*_{n,k}$.
14: Obtain the optimal wireless allocation $l^*_{n,k}$ and $B^{\mathrm{U}*}_k$ according to Theorem 2 with $a^*_{n,k}$.

Regarding the branch-and-bound algorithm adopted to obtain the optimal user association under the given local-edge latency, the computational complexity is exponential in general, thereby cannot be implemented in practice. The interior point method [39] can be used to achieve the sub-optimal solution to (27) with a polynomial computational complexity, $\mathcal{O}\left((KC+NK)^{3.5} J\right)$, where J is the number of quantization bits. With the interior point method, a comparable learning performance can also be obtained as demonstrated in the simulations. Denote S as the number of required steps in the one-dimensional search algorithm. Then, the computational complexity for solving the problem (26) is $\mathcal{O}\left(S(KC+NK)^{3.5} J\right)$.

TABLE I
SIMULATION PARAMETERS

Parameters	Values
Path loss model	$128.1 + 37.6\log_{10}(d)$
BS coverage radius	1500m
Transmission power of device	28dBm
Noise power density	-174dBm/Hz
Number of devices	30
Number of BSs	4
Total bandwidth of uplink channel, B^{U}	50MHz
The number of local updates, E	5
The uplink data rate of BS, r	20 Mbps
Batchsize	64
The weight coefficient, β	0.9

VI. EXPERIMENTS

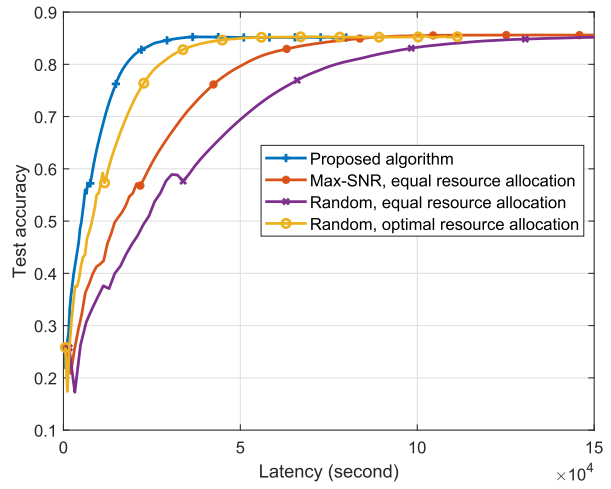
In the simulations, we consider 4 BSs and 30 mobile devices randomly distributed in system coverage. These devices have one or multiple candidate BSs. Both large-scale fading and small-scale fading are considered for the channels between the mobile devices and BSs. The path loss model of large-scale fading is $128.1 + 37.6\log_{10}(d)$, and the small-scale fading follows Rayleigh distribution. Two classic neural networks, i.e., ResNet18 and MobileNet, and two classic datasets, i.e., CIFAR10 and CIFAR100, are adopted to demonstrate the effectiveness of the proposed policies, where 50,000 training data and 10,000 test data are included in these datasets. The computing power of mobile devices is randomly distributed in the interval [2GHz, 3GHz]. Other major parameter values used in experiments are listed in Table I.

A. IID Case

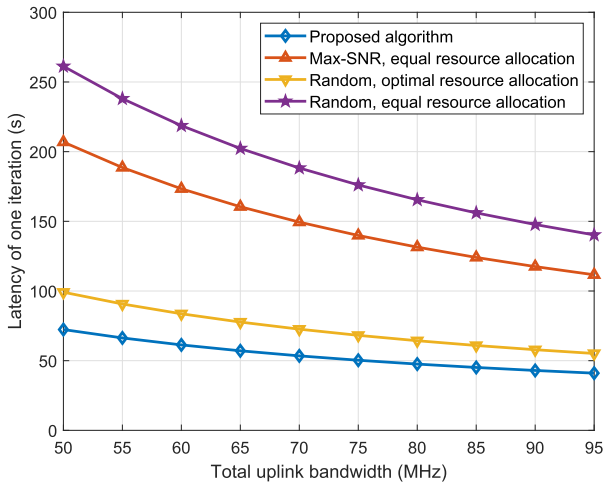
In this subsection, the learning performances, i.e., convergence rate, learning accuracy, and learning latency, are demonstrated to validate the effectiveness of Algorithm 2 with ResNet18 and CIFAR10. The training data are uniformly allocated to all devices, ensuring that the training data of each device are IID. Three baselines are simulated for performance comparisons.

- *Max-SNR, equal resource allocation*: User association is performed based on max-SNR while the wireless resources are equally allocated to all devices and BSs.
- *Random, optimal resource allocation*: The user association is randomly performed while the wireless resources are optimally allocated.
- *Random, equal resource allocation*: The user association is randomly performed and the wireless resources are equally allocated.

Fig. 3 depicts the performance comparison among different user association and resource allocation strategies in the IID



(a) Convergence rate.



(b) Latency.

Fig. 3. Performance comparison of different user association and resource allocation strategies in the IID case.

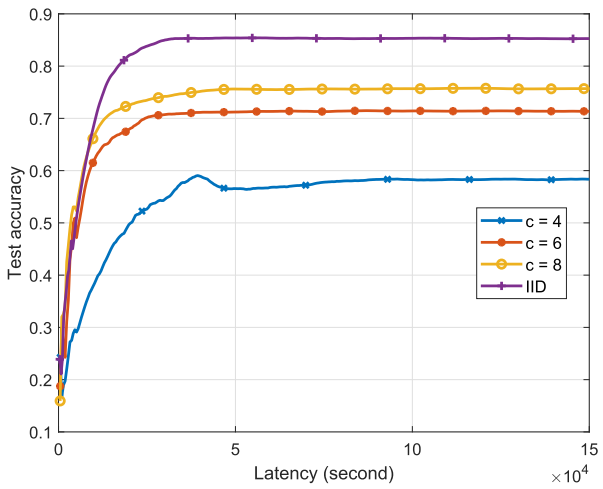


Fig. 4. Effect of data distribution on test accuracy.

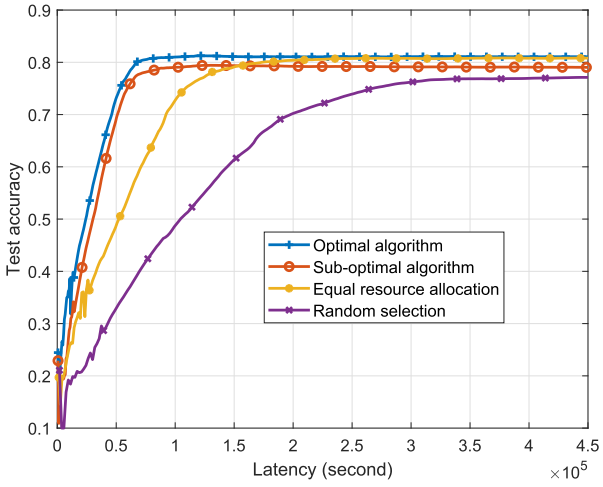


Fig. 5. The optimization algorithm performance comparison.

case. From Fig. 3(a), the proposed algorithm can achieve a larger convergence rate compared to the baselines. It can be explained by that both user association and resource allocation can reduce the learning latency, thereby improving the convergence rate. In addition, the max-SNR based user association is the optimal solution in the IID case. Moreover, wireless resources should also be allocated according to the device computing power and channel state information. Moreover, all algorithms enjoy the same learning accuracy. The reason is that the user association does not affect the model error in the IID case. As for Fig. 3(b), the learning latency at one iteration decreases with the total uplink bandwidth. Since more wireless resources are allocated to mobile devices, the local communication latency can be reduced. With more wireless resources, the effect of user association and resource allocation becomes small since wireless communication is no longer the bottleneck.

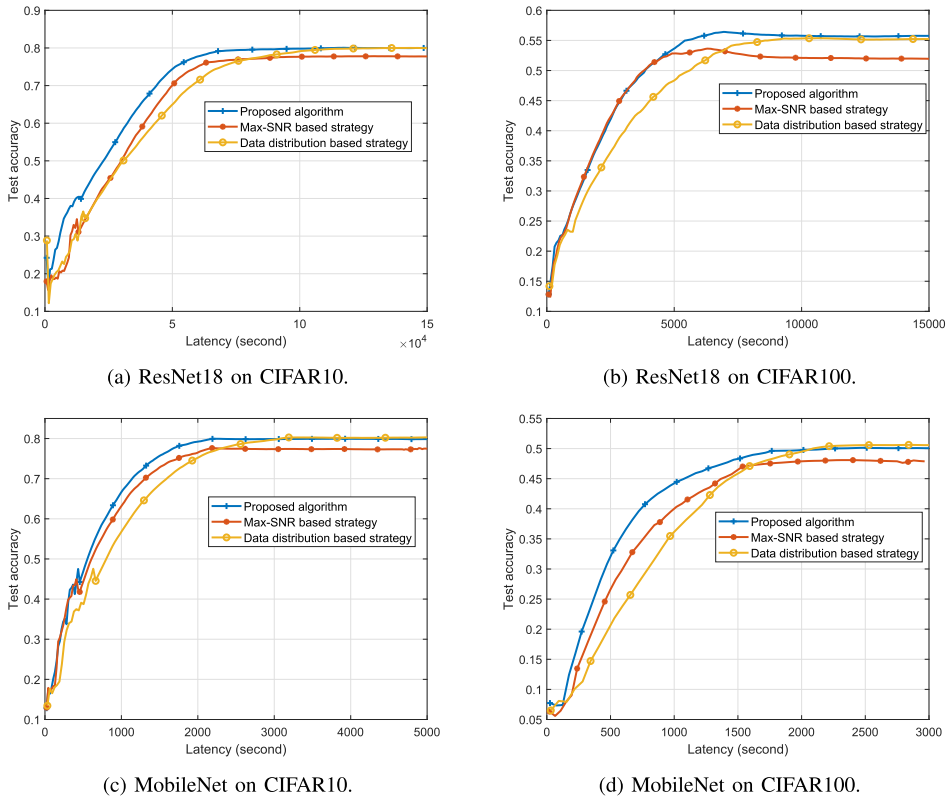
B. Non-IID Case

This subsection demonstrates the performance of the proposed algorithm for the non-IID case. In this case, the training data are equally allocated to all devices, but each device only has partial classes of data.

1) Data Distribution: Fig. 4 presents the impact of data distribution on the learning performance with ResNet18 and CIFAR10. In the simulations, mobile devices uniformly select c classes from the training data. Therefore, the number of selected classes represents the data distribution distance from the IID training data. The data distribution distance becomes smaller when mobile devices select more classes. From the figure, both convergence rate and learning accuracy increase with the number of selected classes. The reason is that the model error compared with the optimal model trained with IID data decreases as the data distribution distance decreases. From this result, the necessity of reducing the data distribution distance using user association is corroborated.

2) Optimality: We examine the optimality of the proposed Algorithm 4 with ResNet18 and CIFAR10. In addition to the proposed optimal algorithm, three baselines are also implemented for the performance comparison.

- *Sub-optimal algorithm*: The interior point method is applied to obtain the user association by variable relaxation and rounding. In addition, the wireless resources are optimally allocated.



(a) ResNet18 on CIFAR10.

(b) ResNet18 on CIFAR100.

(c) MobileNet on CIFAR10.

(d) MobileNet on CIFAR100.

Fig. 6. Performance comparison of different user association strategies in the non-IID case.

- *Equal resource allocation*: The user association is obtained by joint data distribution and learning latency. However, the wireless resources are equally allocated.
- *Random selection*: The devices randomly select a BS and the wireless resources are equally allocated.

Fig. 5 illustrates the optimality of the proposed algorithm. From the curves, the proposed algorithm outperforms three baseline algorithms from both aspects of the convergence rate and learning accuracy. With the sub-optimal algorithm, the comparable learning performance can be achieved while significantly reducing the computational complexity. Therefore, the sub-optimal algorithm is more practical than the proposed optimal algorithm. In addition, the optimal wireless resource allocation can also significantly improve the learning performance in the non-IID case. Since the local computing power and channel state information differ from device to device in general, the latency for local communication can be reduced by the resource allocation.

3) User Association Strategy: In this part, three user association strategies are tested with two models (ResNet18 and MobileNet) and two datasets (CIFAR10 and CIFAR100), respectively. They are the proposed algorithm, the max-SNR based strategy, and the data distribution based strategy.

- *Max-SNR based strategy*: Max-SNR based strategy is the optimal solution in the IID case. Mobile devices select the BS with the highest uplink channel SNR. Meanwhile, the optimal wireless resource allocation is executed for mobile devices and BSs.
- *Data distribution based strategy*: The user association is obtained by only minimizing the data distribution distance while the uplink channel SNR is not considered. Moreover, the wireless resources are optimally allocated.

Note that Max-SNR based strategy is the algorithm without considering data distribution according to the analysis in IID case, corresponding to [28]. The authors in [28] consider both the learning latency and energy consumption. For comparison, we ignore the energy consumption in this simulation. Data distribution based strategy is the algorithm without considering SNR, corresponding to [23].

Fig. 6 describes the learning performance comparison of different user association strategies in the non-IID case. From the figure, the proposed algorithm can achieve a higher convergence rate and learning accuracy compared with the max-SNR based strategy. The reason behind is that the proposed algorithm has a smaller model error compared with the optimal model, thereby increasing the convergence rate and learning accuracy. Moreover, as compared against the data distribution based strategy, a faster convergence rate and a comparable learning accuracy can be obtained by the proposed algorithm. It is because that the proposed algorithm can reduce the learning latency while maintaining a small model error by jointly minimizing the total data distribution distance and learning latency. Therefore, different from the traditional cellular networks and HFL, both the uplink channel SNR and data distribution should be incorporated into user association in wireless HFL.

4) The Weight Coefficient β: Fig. 7 shows the impact of the weight coefficient selection on the learning performance with ResNet18 and CIFAR10. From the figure, the larger the β is, the closer the proposed algorithm is to the data distribution

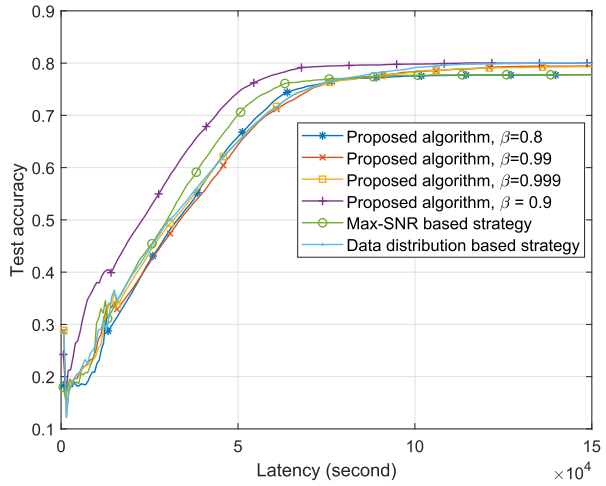


Fig. 7. Effect of the weight β on the learning performance.

based strategy. Conversely, similar learning performance to the max-SNR strategy is achieved under a small β. In other words, the proposed algorithm reduces to the max-SNR based strategy under a sufficiently small β and the data distribution based strategy under a sufficiently large β. The results tell that properly setting β can well balance the total data distribution distance and the learning latency. To achieve a better learning performance, the weight coefficient β should be set to ensure that the total data distribution distance and the learning latency are within an order of magnitude.

VII. Conclusion

In this paper, we investigate user association and resource allocation in wireless HFL taking into account the imbalanced data distribution and traffic load. First, we derive the model error and learning latency and analyze the effect of the data distribution, user association, and wireless resource allocation on the learning performance. Then, the user association and resource allocation are optimized for both the IID and non-IID cases to improve the learning performance. For the IID case, max-SNR based user association achieves the optimal learning latency. Meanwhile, the wireless resources should be optimally allocated according to both local computing power and uplink channel SNR. For the non-IID case, both the data distribution distance and learning latency are minimized to improve the learning performance. Different from the IID case, the optimal user association is determined by both the data distribution and uplink channel SNR. Furthermore, wireless resources are optimally allocated for mobile devices according to both local computing power and uplink channel SNR. Finally, extensive simulation results demonstrate the effectiveness of the proposed algorithms.

Appendix A
Proof of Theorem 1

According to the definition of $\boldsymbol{w}_{mE}^{\mathrm{G}}$ and $\boldsymbol{w}_{mE}^{*}$, we have (30), shown at the bottom of the page, where $p^k(c)$ is the data distribution of BS k and (a) holds since the local-edge aggregation is performed at every local iteration, which has been proved in [23]. Then, the model error can be further written as

$$\begin{aligned}
&\left\|\boldsymbol{w}_{mE}^{\mathrm{G}}-\boldsymbol{w}_{mE}^{*}\right\| \\
&\leq \left\|\sum_k \frac{n_k}{\sum_k n_k}\boldsymbol{w}_{mE-1}^{\mathrm{B},k}-\boldsymbol{w}_{mE-1}^{*}\right\| \\
&\quad +\alpha\|\sum_k \frac{n_k}{\sum_k n_k}\sum_c p^k(c)\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}_{mE-1}^{\mathrm{B},k}}(x)\right)\right\} \\
&\quad -\sum_c p(c)\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}_{mE-1}^{*}}(x)\right)\right\}\| \\
&\leq \sum_k \frac{n_k}{\sum_k n_k}\left\|\boldsymbol{w}_{mE-1}^{\mathrm{B},k}-\boldsymbol{w}_{mE-1}^{*}\right\| \\
&\quad +\|\alpha\sum_k \frac{n_k}{\sum_k n_k}\sum_c p^k(c)\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}_{mE-1}^{\mathrm{B},k}}(x)\right)\right\} \\
&\quad -\alpha\sum_k \frac{n_k}{\sum_k n_k}\sum_c p(c)\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}_{mE-1}^{\mathrm{B},k}}(x)\right)\right\}\| \\
&\quad +\|\alpha\sum_k \frac{n_k}{\sum_k n_k}\sum_c p(c)\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}_{mE-1}^{\mathrm{B},k}}(x)\right)\right\} \\
&\quad -\alpha\sum_k \frac{n_k}{\sum_k n_k}\sum_c p(c)\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}_{mE-1}^{*}}(x)\right)\right\}\|.
\end{aligned} \tag{31}$$

With the assumptions on the gradient smoothness and bounded gradient, the model error can be further written as

$$\begin{aligned}
&\left\|\boldsymbol{w}_{mE}^{\mathrm{G}}-\boldsymbol{w}_{mE}^{*}\right\| \\
&\leq \sum_k \frac{n_k}{\sum_k n_k}\left(1+\alpha\sum_c p(c)L(c)\right)\left\|\boldsymbol{w}_{mE-1}^{\mathrm{B},k}-\boldsymbol{w}_{mE-1}^{*}\right\| \\
&\quad +\alpha A_{mE-1}\sum_k \frac{n_k}{\sum_k n_k}\sum_c \left\|p^k(c)-p(c)\right\|.
\end{aligned} \tag{32}$$

$$\begin{aligned}
\left\|\boldsymbol{w}_{mE}^{\mathrm{G}}-\boldsymbol{w}_{mE}^{*}\right\| &= \left\|\sum_k \frac{n_k}{\sum_k n_k}\boldsymbol{w}_{mE}^{\mathrm{B},k}-\boldsymbol{w}_{mE}^{*}\right\| \\
&\overset{(a)}{=} \|\sum_k \frac{n_k}{\sum_k n_k}\left(\boldsymbol{w}_{mE-1}^{\mathrm{B},k}+\alpha\sum_c p^k(c)\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}_{mE-1}^{\mathrm{B},k}}(x)\right)\right\}\right) \\
&\quad -\left(\boldsymbol{w}_{mE-1}^{*}+\alpha\sum_c p(c)\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}_{mE-1}^{*}}(x)\right)\right\}\right)\|.
\end{aligned} \tag{30}$$

For the first term in (32), we have

$$\begin{aligned}
&\left\|\boldsymbol{w}_{mE-1}^{\mathrm{B},k}-\boldsymbol{w}_{mE-1}^{*}\right\| \\
&\quad\leq\left\|\boldsymbol{w}_{mE-2}^{\mathrm{B},k}-\boldsymbol{w}_{mE-2}^{*}\right\| \\
&\qquad+\alpha\|\sum_{c}p^{k}\left(c\right)\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}_{mE-2}^{\mathrm{B},k}}\left(x\right)\right)\right\} \\
&\qquad-\sum_{c}p\left(c\right)\nabla_{\boldsymbol{w}}\mathbb{E}_{x|y(c)}\left\{\log\left(\mathcal{H}_{\boldsymbol{w}_{mE-2}^{*}}\left(x\right)\right)\right\}\| \\
&\quad\leq\left(1+\alpha\sum_{c}p\left(c\right)L\left(c\right)\right)\left\|\boldsymbol{w}_{mE-2}^{\mathrm{B},k}-\boldsymbol{w}_{mE-2}^{*}\right\| \\
&\qquad+\alpha A_{mE-2}\sum_{c}\left\|p^{k}\left(c\right)-p\left(c\right)\right\|. \qquad (33)
\end{aligned}$$

By substituting (33) to the first term in (32) until the $(m-1)\,E$-th iteration, the model error after one global iteration can be written as (34), which ends the proof.

Appendix B
Proof of Lemma 1

Define $u_{n,k}=l_{n,k}B_{k}^{\mathrm{U}}$. Then the uplink data rate $r_{n,k}^{\mathrm{U}}$ and the constraints (16a), (16b) can be rewritten as

$$r_{n,k}^{\mathrm{U}}=u_{n,k}\log_{2}\left(1+\frac{p_{n}^{\mathrm{U}}h_{n,k}^{\mathrm{U}}}{N_{0}}\right), \qquad (35)$$

and

$$\sum_{n}\sum_{k}a_{n,k}u_{n,k}\leq B^{\mathrm{U}}, \qquad (36)$$

respectively. Accordingly, the problem (17) becomes convex since $f\left(x\right)=\frac{1}{x},x>0$ is convex, which ends the proof.

Appendix C
Proof of Theorem 2

According to the Lagrangian multiplier method, the Lagrangian function of the problem (17) can be constructed as

$$\begin{aligned}
&L\left(u_{n,k},\eta_{n,k},\nu,t^{\mathrm{E}},u_{n,k}\right)=t^{\mathrm{E}} \\
&\quad+\sum_{n}\sum_{k}\eta_{n,k}\left(a_{n,k}\left(\frac{bd^{\mathrm{L}}}{f_{n}}+\frac{M}{u_{n,k}\log_{2}\left(1+\frac{p_{n}^{\mathrm{U}}h_{n,k}^{\mathrm{U}}}{N_{0}}\right)}\right)-t^{\mathrm{E}}\right) \\
&\quad+\nu\left(\sum_{n}\sum_{k}a_{n,k}u_{n,k}-B^{\mathrm{U}}\right), \qquad (37)
\end{aligned}$$

where $\eta_{n,k}$ and ν are Lagrangian multipliers related to the constraints (17a) and (36), respectively. Then, the corresponding *Karush-Kuhn-Tucker* (KKT) conditions are

$$\frac{\partial L}{\partial t^{\mathrm{E}}}=1-\sum_{n}\sum_{k}\eta_{n,k}=0,\quad n\in\mathcal{N},\ k\in\mathcal{K}, \qquad (38)$$

$$\frac{\partial L}{\partial u_{n,k}}=a_{n,k}\nu-a_{n,k}\eta_{n,k}\frac{M}{u_{n,k}^{2}\log_{2}\left(1+\frac{p_{n}^{\mathrm{U}}h_{n,k}^{\mathrm{U}}}{N_{0}}\right)}=0,\quad n\in\mathcal{N},\quad k\in\mathcal{K}, \qquad (39)$$

$$\eta_{n,k}\left(a_{n,k}\left(\frac{bd^{\mathrm{L}}}{f_{n}}+\frac{M}{u_{n,k}\log_{2}\left(1+\frac{p_{n}^{\mathrm{U}}h_{n,k}^{\mathrm{U}}}{N_{0}}\right)}\right)-t^{\mathrm{E}}\right)=0,\quad n\in\mathcal{N},\quad k\in\mathcal{K}, \qquad (40)$$

$$\nu\left(\sum_{n}\sum_{k}a_{n,k}u_{n,k}-B^{\mathrm{U}}\right)=0, \qquad (41)$$

$$\eta_{n,k}\geq 0,\quad \nu\geq 0. \qquad (42)$$

The condition in (38) shows that there exists $\eta_{n,k}$ that satisfies $\eta_{n,k}\neq 0$. With this conclusion, $\nu\neq 0$ can be derived according to (39), thereby $\eta_{n,k}\neq 0,\forall\left\{n,k|a_{n,k}=1\right\}$. Then, we have $a_{n,k}\left(\frac{bd^{\mathrm{L}}}{f_{n}}+\frac{M}{u_{n,k}\log_{2}\left(1+\frac{p_{n}^{\mathrm{U}}h_{n,k}^{\mathrm{U}}}{N_{0}}\right)}\right)=t^{\mathrm{E}},\forall\left\{n,k|a_{n,k}=1\right\}$ and $\sum_{n}\sum_{k}a_{n,k}u_{n,k}=B^{\mathrm{U}}$ according to the conditions in (40)

$$\begin{aligned}
\left\|\boldsymbol{w}_{mE}^{\mathrm{G}}-\boldsymbol{w}_{mE}^{*}\right\|&\leq\left(1+\alpha\sum p\left(c\right)L\left(c\right)\right)^{E}\left\|\boldsymbol{w}_{(m-1)E}^{\mathrm{G}}-\boldsymbol{w}_{(m-1)E}^{*}\right\| \\
&\quad+\alpha\left(\sum_{i=0}^{E-1}A_{mE-1-i}\left(1+\alpha\sum_{c}p\left(c\right)L\left(c\right)\right)^{i}\right)\underbrace{\frac{1}{N}\sum_{k}\sum_{c}\left\|\sum_{n\in\mathcal{N}_{k}}a_{n,k}\left(p^{n}\left(c\right)-p\left(c\right)\right)\right\|}_{\text{effect of data distribution imbalance}}. \qquad (34)
\end{aligned}$$

$$\begin{aligned}
\min_{\{q_{k,c},a_{n,k}\}}L\left(\lambda_{k,c},\mu_{k,c},\gamma\right)&=\beta\frac{1}{N}\sum_{k}\sum_{c}q_{k,c}\left(1-\lambda_{k,c}-\mu_{k,c}\right)-\gamma B^{\mathrm{U}}+\left(1-\beta\right)\left(t^{\mathrm{E}'}+\frac{1}{E}t^{\mathrm{G}}\right) \\
&\quad+\sum_{k}\sum_{n}a_{n,k}\left(\gamma\frac{M}{\left(t^{\mathrm{E}'}-\frac{bd^{\mathrm{L}}}{f_{n}}\right)\log_{2}\left(1+\frac{p_{n}^{\mathrm{U}}h_{n,k}^{\mathrm{U}}}{N_{0}}\right)}+\sum_{c}\left(\lambda_{k,c}-\mu_{k,c}\right)\left(p^{n}\left(c\right)-p\left(c\right)\right)\right). \qquad (43)
\end{aligned}$$

$$\text{subject to}\ \sum_{k\in\mathcal{K}_{n}}a_{n,k}=1, \qquad (43a)$$

$$0\leq a_{n,k}\leq 1. \qquad (43b)$$

$$a_{n,k}^{*}=\begin{cases}1, & k=\arg\min\limits_{k}\left(\gamma\dfrac{M}{\left(t^{\mathrm{E}\prime}-\frac{bd^{\mathrm{L}}}{f_n}\right)\log_2\left(1+\frac{p_n^{\mathrm{U}}h_{n,k}^{\mathrm{U}}}{N_0}\right)}+\sum\limits_{c}\left(\lambda_{k,c}-\mu_{k,c}\right)\left(p^n\left(c\right)-p\left(c\right)\right)\right),\\ 0, & \text{otherwise.}\end{cases} \tag{44}$$

$$q_{k,c}^{*}=\begin{cases}\min\left(\sum\limits_{n}a_{n,k}^{*}\left(p^n\left(c\right)-p\left(c\right)\right),\sum\limits_{n}a_{n,k}^{*}\left(p\left(c\right)-p^n\left(c\right)\right)\right), & 1-\lambda_{k,c}-\mu_{k,c}\geq 0,\\ \max\left(\sum\limits_{n}a_{n,k}^{*}\left(p^n\left(c\right)-p\left(c\right)\right),\sum\limits_{n}a_{n,k}^{*}\left(p\left(c\right)-p^n\left(c\right)\right)\right), & 1-\lambda_{k,c}-\mu_{k,c}<0.\end{cases} \tag{45}$$

and (41). Accordingly, the expression in (20) can be obtained. With the condition $\sum\limits_{k}a_{n,k}u_{n,k}=B_k^{\mathrm{U}}$, the optimal solutions of $l_{n,k}$ and B_k^{U} can also be derived as in (18) and (19), which ends the proof.

Appendix D
Proof of Lemma 2

As analyzed in Section V.B, we transform the problem in (27) into (28) by the Lagrangian relaxation method, as (43), shown at the bottom of the previous page. Under the given $\lambda_{k,c}$, $\mu_{k,c}$, and γ, optimal solutions of $a_{n,k}^{*}$ and $q_{k,c}^{*}$ can be easily obtained as (44) and (45), shown at the top of the page.

Therefore, with optimal Lagrangian multipliers, $\lambda_{k,c}^{*}$, $\mu_{k,c}^{*}$, and γ^{*}, Lemma 2 can be derived from (44), which ends the proof.

References

[1] O. I. Abiodun *et al.*, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, Nov. 2018, Art. no. e00938.

[2] J. Konecný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," 2016, *arXiv:1610.02527*.

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

[4] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, Jan. 2020.

[5] J. Park *et al.*, "Communication-efficient and distributed learning over wireless networks: Principles and applications," *Proc. IEEE*, vol. 109, no. 5, pp. 796–819, May 2021.

[6] Y. Jiang *et al.*, "Model pruning enables efficient federated learning on edge devices," 2019, *arXiv:1909.12326*.

[7] S. Liu, G. Yu, R. Yin, and J. Yuan, "Adaptive network pruning for wireless federated learning," *IEEE Wireless Commun. Lett.*, vol. 10, no. 7, pp. 1572–1576, Jul. 2021.

[8] H. Sun, X. Ma, and R. Q. Hu, "Adaptive federated learning with gradient compression in uplink NOMA," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 16325–16329, Dec. 2020.

[9] S. Liu, G. Yu, R. Yin, J. Yuan, and F. Qu, "Adaptive batchsize selection and gradient compression for wireless federated learning," in *Proc. IEEE Global Commun. Conf.*, Taipei, Taiwan, Dec. 2020, pp. 1–6.

[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] G. Zhu, Y. Du, D. Gunduz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2120–2135, Mar. 2021.

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

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

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

[15] M. M. Wadu, S. Samarakoon, and M. Bennis, "Joint client scheduling and resource allocation under channel uncertainty in federated learning," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5962–5974, Sep. 2021.

[16] 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.

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

[18] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-IID data," in *Proc. Int. Conf. Learn. Represent.*, Addis Ababa, Ethiopia, Apr. 2020, pp. 1–26.

[19] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing federated learning on non-IID data with reinforcement learning," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Toronto, ON, Canada, Jul. 2020, pp. 1698–1707.

[20] C. Briggs, Z. Fan, and P. Andras, "Federated learning with hierarchical clustering of local updates to improve training on non-IID data," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Glasgow, U.K., Jul. 2020, pp. 1–9.

[21] T. Castiglia, A. Das, and S. Patterson, "Multi-level local SGD: Distributed SGD for heterogeneous hierarchical networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2021, pp. 1–36.

[22] J. Wang, S. Wang, R.-R. Chen, and M. Ji, "Demystifying why local aggregation helps: Convergence analysis of hierarchical SGD," 2020, *arXiv:2010.12998*.

[23] N. Mhaisen, A. Awad, A. Mohamed, A. Erbad, and M. Guizani, "Optimal user-edge assignment in hierarchical federated learning based on statistical properties and network topology constraints," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 1, pp. 55–66, Jan./Feb. 2022.

[24] M. Duan, D. Liu, X. Chen, R. Liu, and Y. Tan, "Self-balancing federated learning with global imbalanced data in mobile systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 1, pp. 59–71, Jul. 2021.

[25] M. S. H. Abad, E. Ozfatura, D. Gunduz, and O. Ercetin, "Hierarchical federated learning ACROSS heterogeneous cellular networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 8866–8870.

[26] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Dublin, Ireland, Jun. 2020, pp. 1–6.

[27] S. Hosseinalipour, C. G. Brinton, V. Aggarwal, H. Dai, and M. Chiang, "From federated to fog learning: Distributed machine learning over heterogeneous wireless networks," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 41–47, Dec. 2020.

[28] S. Luo, X. Chen, Q. Wu, Z. Zhou, and S. Yu, "HFEL: Joint edge association and resource allocation for cost-efficient hierarchical federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6535–6548, Oct. 2020.

[29] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, Jan. 2020.

[30] H. B. McMahan *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Statist.*, Orlando, FL, USA, Apr. 2017, pp. 1273–1282.

[31] P. Kairouz *et al.*, "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, nos. 1–2, pp. 1–210, Jun. 2021.

[32] K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons, "The non-IID data quagmire of decentralized machine learning," in *Proc. Int. Conf. Mach. Learn.*, vol. 119, Nov. 2020, pp. 4387–4398.
[33] A. S. Hamza, S. S. Khalifa, H. S. Hamza, and K. Elsayed, "A survey on inter-cell interference coordination techniques in OFDMA-based cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 1642–1670, 4th Quart., 2013.
[34] D. Wen, M. Bennis, and K. Huang, "Joint parameter-and-bandwidth allocation for improving the efficiency of partitioned edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8272–8286, Sep. 2020.
[35] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1935–1949, Mar. 2021.
[36] 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, Jan. 2021.
[37] M. L. Fisher, "The Lagrangian relaxation method for solving integer programming problems," *Manage. Sci.*, vol. 27, no. 1, pp. 1–18, 1981.
[38] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.
[39] M. J. Todd, "A low complexity interior-point algorithm for linear programming," *SIAM J. Optim.*, vol. 2, no. 2, pp. 198–209, May 1992.

Shengli Liu received the B.S.E. degree in information engineering from Soochow University, Suzhou, China, in 2017, and the Ph.D. degree from the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China, in 2022. In 2021, he was a Visiting Research Scholar with the Centre for Wireless Communication, University of Oulu, Finland, and the VTT Technical Research Centre of Finland, Finland. He is currently a Post-Doctoral Researcher at the School of Information and Electrical Engineering, Zhejiang University City College, Hangzhou, and the College of Information Science and Electronic Engineering, Zhejiang University. His current research interests include machine learning and federated learning.

Guanding Yu (Senior Member, IEEE) received the B.E. and Ph.D. degrees in communication engineering from Zhejiang University, Hangzhou, China, in 2001 and 2006, respectively.

From 2013 to 2015, he was a Visiting Professor at the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. In 2006, he joined Zhejiang University, where he is currently a Professor with the College of Information Science and Electronic Engineering. His research interests include 5G communications and networks, integrated sensing and communications, mobile edge computing/learning, and machine learning for wireless networks. He received the 2016 IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award. He has served as a Guest Editor for *IEEE Communications Magazine* Special Issue on Full-Duplex Communications, an Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Series on Green Communications and Networking and Series on Machine Learning in Communications and Networks, IEEE WIRELESS COMMUNICATIONS LETTERS, IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, and IEEE ACCESS, and a Lead Guest Editor for *IEEE Wireless Communications* Magazine Special Issue on LTE in Unlicensed Spectrum. He regularly sits on the Technical Program Committee (TPC) Boards of prominent IEEE Conferences, such as ICC, GLOBECOM, and VTC. He has served as a Symposium Co-Chair for IEEE Globecom 2019 and the Track Chair for IEEE VTC 2019'Fall.

Xianfu Chen (Member, IEEE) received the Ph.D. degree (Hons.) in signal and information processing from the Department of Information Science and Electronic Engineering (ISEE), Zhejiang University, Hangzhou, China, in March 2012. Since April 2012, he has been with the VTT Technical Research Centre of Finland, Oulu, Finland, where he is currently a Senior Scientist. His research interests include wireless communications and networking, with emphasis on human-level and artificial intelligence for resource awareness in next-generation communication networks. He was a recipient of the 2021 IEEE Communications Society Outstanding Paper Award and the 2021 IEEE Internet of Things Journal Best Paper Award. He received the Exemplary Reviewer Certificate of the IEEE TRANSACTIONS ON COMMUNICATIONS in 2021. He serves as an Editor for IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING and *Microwave and Wireless Communications*, an Academic Editor for *Wireless Communications and Mobile Computing*, and an Associate Editor for *China Communications*. He has served as a member of the First Editorial Board for the *Journal of Communications and Information Networks* and the Guest Editor for several international journals, including *IEEE Wireless Communications* magazine. He has also served as the Track Co-Chair and a TPC Member for a number of IEEE ComSoc Flagship Conferences. He is the Vice Chair of IEEE Special Interest Group on Big Data with Computational Intelligence and IEEE Special Interest Group on AI Empowered Internet of Vehicles.

Mehdi Bennis (Fellow, IEEE) is currently a Full (Tenured) Professor at the Centre for Wireless Communications, University of Oulu, Finland, and the Academy of Finland Research Fellow and the Head of the Intelligent Connectivity and Networks/Systems Group (ICON). He has published more than 200 research papers in international conferences, journals, and book chapters. His main research interests include radio resource management, heterogeneous networks, game theory, and distributed machine learning in 5G networks and beyond. He was a recipient of several prestigious awards, including the 2015 Fred W. Ellersick Prize from the IEEE Communications Society, the 2016 Best Tutorial Prize from the IEEE Communications Society, the 2017 EURASIP Best paper Award for the Journal of Wireless Communications and Networks, the all-University of Oulu Award for research, the 2019 IEEE ComSoc Radio Communications Committee Early Achievement Award, and the 2020 Clarviate Highly Cited Researcher by the Web of Science. He is an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS (TCOM) and a Specialty Chief Editor of *Data Science for Communications in the Frontiers in Communications and Networks* journal.