





Adaptive Hierarchical Federated Learning Over Wireless Networks

Bo Xu , Student Member, IEEE, Wenchao Xia , Member, IEEE, Wanli Wen , Member, IEEE, Pei Liu , Member, IEEE, Haitao Zhao , and Hongbo Zhu , Member, IEEE

I. INTRODUCTION

Abstract—Federated learning (FL) is promising in enabling large-scale model training by massive devices without exposing their local datasets. However, due to limited wireless resources, traditional cloud-based FL system suffers from the bottleneck of communication overhead in core network. Fueled by this issue, we consider a hierarchical FL system and formulate a joint problem of edge aggregation interval control and resource allocation to minimize the weighted sum of training loss and training latency. To quantify the learning performance, an upper bound of the average global gradient deviation, in terms of the edge aggregation interval, the training latency, and the number of successfully participating devices, is derived. Then an alternative problem is formulated, which can be decoupled into an edge aggregation interval control problem and a resource allocation problem, and solved by an iterative optimization algorithm. Specifically, given the resource allocation strategy, a relaxation and rounding method is proposed to optimize the edge aggregation interval. The problem of resource allocation including training time allocation and bandwidth allocation is solved separately based on the convex optimization theory. Simulation results show that the proposed algorithm, compared to the benchmarks, can achieve higher learning performance with lower training latency, and is capable of adaptively adjusting the edge aggregation interval and the resource allocation strategy according to the training process.

Index Terms—Hierarchical federated learning, adaptive aggregation, convergence analysis, resource allocation.

MACHINE learning (ML) techniques [1] have achieved unprecedented success in various fields, ranging from natural language processing to computer vision. Traditionally, ML model training can be conducted in a centralized manner [2]–[5], i.e., at a cloud server, which has powerful computation capability and collects a large number of training samples from the devices. However, due to limited wireless resources, uploading local datasets to the cloud server is too costly in terms of bandwidth. Besides, the edge devices may be unwilling to share their local datasets due to privacy concerns. As such, researchers have proposed a new distributed model training framework named federated learning (FL) [6], which is coordinated by a cloud server and leverages the computation capabilities of devices. A typical training process of FL contains multiple rounds, and in each round, devices perform local training based on local datasets, followed by a global aggregation at the cloud server which updates the global model. After that, the updated global model is broadcast to the devices for the next round of training. Since the cloud server acquires the local models from the devices, rather than their local datasets, FL promotes the training security [7].

Recent works have utilized FL to solve some communication issues, i.e., attack detection [8] and automatic modulation classification [9]. However, when implementing FL over wireless networks, training with limited computation and communication resources can increase the training latency and degrade the learning performance, especially when the learning task is complex and the size of local models is large. To solve this issue, existing works have proposed some approaches which can be roughly divided into four categories: data preparation, over-the-air computation, resource allocation, and device scheduling. Belonging to the first category, some technologies such as sparsification [10], [11], quantization [12], [13], and encoding [14] were adopted in FL to reduce the cost of model uploading. To further reduce the training latency, computation over-the-air approaches were proposed in [15], [16], which utilized the superposition property of the multiple access channels. In terms of resource allocation, the bandwidth, the local computing power, and the training latency were optimized in [17]–[19]. Besides, a robust design for FL was proposed in [20] to decline the effect of noise. Belonging to the last category, several scheduling criteria were proposed to improve the learning performance, such as the significance of local models [21], [22], the number of times that each device has been scheduled [23], [24], and the training latency of each

Manuscript received July 5, 2021; revised September 13, 2021 and October 27, 2021; accepted December 5, 2021. Date of publication December 15, 2021; date of current version February 14, 2022. The work of Bo Xu, Wenchao Xia, Haitao Zhao, and Hongbo Zhu was supported in part by the National Key Research and Development Program under Grant 2019YFB2103004, in part by the National Natural Science Foundation of China (NSFC) under Grant 92067201, in part by the Jiangsu Provincial Key Research and Development Program under Grant BE2020084-1, in part by the Natural Science Foundation on Frontier Leading Technology Basic Research Project of Jiangsu under Grant BK20212001, and in part by the Natural Science Research Project of Jiangsu Higher Education Institutions under Grants 21KJB510034 and 21KJB510027. The work of Wanli Wen was supported by the Natural Science Foundation of Chongqing China under Grant cstc2021jcyj-msxmX0458. The work of Pei Liu was supported by NSFC under Grant 62001336. The review of this article was coordinated by Dr. Xuanyu Cao. (Corresponding authors: Haitao Zhao; Wanli Wen.)

Bo Xu, Wenchao Xia, Haitao Zhao, and Hongbo Zhu are with the Key Laboratory of Wireless Communications, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu 210003, China, and also with the Engineering Research Center of Health Service System Based on Ubiquitous Wireless Networks, Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: 1018010321@njupt.edu.cn; xiawenchao@njupt.edu.cn; zhaoh@njupt.edu.cn; zhuhb@njupt.edu.cn).

Wanli Wen is with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China (e-mail: wanli_wen@cqu.edu.cn).

Pei Liu is with the School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China (e-mail: pei.liu@ieee.org).

Digital Object Identifier 10.1109/TVT.2021.3135541

device [25]. Besides, a lazily aggregation approach for device scheduling was proposed in [26], which detected the slowly-varying gradients and triggered the reuse of outdated gradients. Moreover, authors in [27]–[29] jointly considered resource allocation and device scheduling to reduce the training latency and improve the learning performance simultaneously. Note that the above works were mainly based on the cloud-based FL system, in which the devices' local models are directly transmitted to the cloud server. However, due to limited wireless resources and long transmission distance, the communication with the cloud server is unpredictable and unreliable, which can degrade the training efficiency as well as model accuracy.

Hierarchical FL system [30]–[35] is seen as a promising solution to solve this issue by leveraging the power of edge servers. Specifically, the edge servers can be deployed in the proximity of the devices, thus the communication distance is significantly shortened and more reliable communication can be achieved [36]. Specially, for the hierarchical FL system, two phrases of model aggregation are performed in each round, i.e., edge aggregation and global aggregation. The global aggregation is executed in the cloud server after multiple edge aggregations executed in the edge servers. Based on the hierarchical FL system, a joint strategy of resource allocation and device scheduling was proposed in [30], and a matching algorithm for the devices and the edge servers was proposed in [31]. To achieve the targeted learning performance, a hierarchical FL algorithm which tuned the number of training rounds was proposed in [32]. A hierarchical incentive mechanism design for FL was proposed in [33]. Recent work [37] combined FL with computation offloading, which enabled devices to offload partial local tasks to the edge servers. Moreover, authors in [32], [34], [35] analyzed the convergence upper bound of the hierarchical FL algorithm. However, to realize great benefits of a hierarchical FL system, although existing works have illustrated that reducing the edge aggregation interval can improve the convergence rate of the global model, the impact of wireless transmission latency was not considered. In particular, edge aggregation interval can be defined as the times of performing local iterations between two adjacent edge aggregations [35]. A similar concept called edge communication frequency was defined in [34], which is the reciprocal to the edge aggregation interval. Besides, a thoughtless resource allocation strategy may increase the training latency [17] and cause the failure of local model uploading [29]. Therefore, it is of significant importance to jointly consider the edge aggregation interval control and the resource allocation.

In this paper, we formulate a joint problem of edge aggregation interval control and resource allocation to minimize the weighted sum of training loss function and training latency by optimizing the edge aggregation interval and the resource allocation. The upper bound of the average global gradient deviation is analyzed, which captures the effects of the edge aggregation interval, the training time allocation, and the number of successfully participating devices. Then an alternative problem is formulated based on the derived upper bound, which can be decoupled into two sub-problems, i.e., edge aggregation interval control problem and resource allocation problem including bandwidth allocation and training time allocation.

For the first problem, an efficient relaxation and rounding method is proposed to find the solution of edge aggregation interval. The solution to the resource allocation problem is obtained by separately solving bandwidth allocation and training time allocation based on the convex optimization theory. Finally, the solution to the alternative problem can be achieved by iteratively solving these sub-problems until convergence.

The main contributions of this paper are summarized as follows.

- We consider an adaptive hierarchical FL system, in which the edge servers are deployed close to the devices in proximity and assist the cloud server to collect local models. Then a joint problem of edge aggregation interval control and resource allocation is formulated aiming at minimizing the weighted sum of training loss and latency.
- The upper bound of the average global gradient deviation is theoretically derived, which quantifies the effects of the edge aggregation interval, the training latency, and the number of the successfully participating devices. Using the obtained upper bound, an alternative problem is formulated, which can be decoupled into an edge aggregation interval control problem and a resource allocation problem including bandwidth allocation and training time allocation. To solve these problems, an iterative algorithm with low complexity is proposed. At each step of this algorithm, we derive new closed-form solutions of the edge aggregation interval, bandwidth allocation, and training time allocation.
- Our simulation results on different datasets show that the proposed algorithm, compared to the baselines, is capable of achieving higher learning performance with lower training latency because the proposed algorithm can adaptively adjust the edge aggregation interval and the resource allocation strategy during the training process.

The remainder of this paper is organized as follows. In Section II, we present the system model and formulate a joint problem of aggregation interval control and resource allocation. In Section III, we make the convergence analysis, and an alternative problem is formulated, which is solved by a proposed iterative optimization algorithm. In Section IV, we present the simulation results, and in Section V, we conclude the paper.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we first introduce the considered hierarchical FL system. Then, a joint problem of edge aggregation interval control and resource allocation is formulated. The key notations used in this paper are summarized in Table I.

A. Learning Model

As shown in Fig. 1, the considered hierarchical FL system consists of a set \mathcal{K} of K mobile devices, a set \mathcal{S} of S access points each connected to an edge server, and a cloud server. Due to the limited coverage of the access points, we denote the set of devices that are served by edge server s as \mathcal{V}_s with $\bigcup_{s=1}^S \mathcal{V}_s = \mathcal{K}$. Let $\mathcal{D}_k = \{(x_{k,i}, y_{k,i})\}_{i=1}^{|\mathcal{D}_k|}$ denote the local dataset at the k -th device, where $x_{k,i}$ denotes the i -th input training sample and $y_{k,i}$ is the labeled output of $x_{k,i}$. Here, the dataset of the devices

TABLE I
SUMMARY OF MAIN NOTATIONS

Notation	Description
\mathcal{K}, K	Set of all devices, size of \mathcal{K}
\mathcal{S}, S	Set of edge servers, size of \mathcal{S}
\mathcal{V}_s	Set of devices that are associated with edge server s
\mathcal{D}_k, D_k	The local dataset of device k , size of \mathcal{D}_k
$\tilde{\mathcal{D}}_s, \tilde{D}_s$	The dataset of device set \mathcal{V}_s , size of $\tilde{\mathcal{D}}_s$
$\bar{\mathcal{D}}, \bar{D}$	The whole dataset, size of $\bar{\mathcal{D}}$
R, G	Number of training rounds, number of performing local iterations per round
$\bar{F}(\mathbf{w}), \bar{F}_s(\mathbf{w}), F_k(\mathbf{w})$	Global loss function on dataset \mathcal{D} , edge loss function on dataset $\tilde{\mathcal{D}}_s$, local loss function on dataset \mathcal{D}_k
$r, t^{(r)}$	Index of training rounds, index of performing local iterations in round r .
$\bar{\mathbf{w}}^{(t^{(r)})}, \bar{\mathbf{w}}_s^{(t^{(r)})}, \mathbf{w}_k^{(t^{(r)})}$	Global model at local iteration $t^{(r)}$, edge model of edge server s at local iteration $t^{(r)}$, local model of device k at local iteration $t^{(r)}$
$\bar{\tau}_k^{c,r}(I^{(r)})$	Average computation latency of device k to perform $I^{(r)}$ local iterations in round r
$\bar{\tau}_k^{u,r}$	Average uploading latency of device k in round r
$\bar{\tau}^{(r)}$	Training latency in round r
$B, N_0, b_k^{(r)}$	System bandwidth, noise power, bandwidth allocated to device k in round r
L, ζ, ϵ	Convergence property parameters

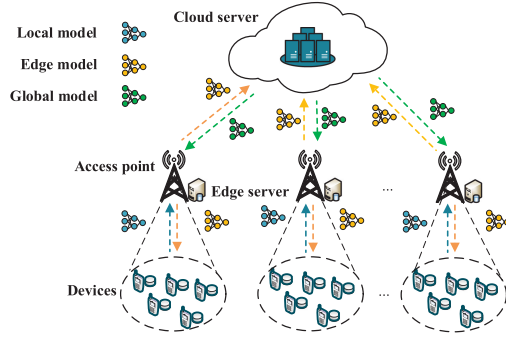


Fig. 1. The structure of the considered hierarchical FL system.

in \mathcal{V}_s is denoted as $\tilde{\mathcal{D}}_s = \bigcup_{k \in \mathcal{V}_s} \mathcal{D}_k$ with $|\tilde{\mathcal{D}}_s| = \tilde{D}_s$. Similarly, the dataset of all the devices is denoted as $\bar{\mathcal{D}} = \bigcup_{k \in \mathcal{K}} \mathcal{D}_k$ with $|\bar{\mathcal{D}}| = \bar{D}$. The goal of the FL training is to find a model parameter vector \mathbf{w} to minimize the global loss function $\bar{F}(\mathbf{w})$ on dataset $\bar{\mathcal{D}}$, i.e.,

$$\min_{\mathbf{w}} \bar{F}(\mathbf{w}) \triangleq \left\{ \frac{1}{\bar{D}} \sum_{k \in \mathcal{K}} \sum_{\{x_{k,i}, y_{k,i}\} \in \mathcal{D}_k} l(\mathbf{w}, x_{k,i}, y_{k,i}) \right\}, \quad (1)$$

where $l(\mathbf{w}, x_{k,i}, y_{k,i})$ is a sample-wise local function that captures the error of the model parameter \mathbf{w} on training sample $\{x_{k,i}, y_{k,i}\}$. Some common loss functions are listed in Table II.

Hierarchical FL is an iterative approach and each round is assumed to include G steps of local iterations. In a certain round r , each device uploads its local model to the corresponding edge server for edge aggregation after every $I^{(r)}$ steps of local iterations. Besides, we assume G is a common multiple of edge aggregation interval $I^{(r)}$, i.e., $I^{(r)}|G = 0$.¹ Then all the edge servers upload their updated edge models to the cloud server

¹ $x|y$ (or $x \nmid y$) denotes that y is divided (or is not divided) by x , i.e., y is (or is not) an integer multiple of x .

TABLE II
COMMON LOSS FUNCTIONS FOR TRAINING

Model	Loss function
Linear regression	$\frac{1}{2} \ \mathbf{y}_k - \mathbf{w}^T x_k\ $
K-means	$\frac{1}{2} \min_i \ \mathbf{x}_k - \mathbf{w}_i\ $ with $\mathbf{w}_i \triangleq [\mathbf{w}_1^T, \mathbf{w}_2^T, \dots]^T$
Squared-SVM	$\frac{1}{2} \ \mathbf{w}\ ^2 + \frac{\rho}{2} \max\{0, 1 - \mathbf{y}_k \mathbf{w}^T x_k\}^2$ with constant ρ
Neural network	Cross-entropy [1]

for global aggregation at the end of each round. Due to limited communication and computation resources, some devices may fail to participate in training. Hence, we use $a_k^{(t^{(r)})} = 1$ denote that device k can finish $I^{(r)}$ steps of local iterations and upload its local model to the corresponding edge server successfully at the $t^{(r)}$ -th local iteration of round r , and $a_k^{(t^{(r)})} = 0$ otherwise. Then, the evolution of the updated local model $\mathbf{w}_k^{(t^{(r)+1})}$, $\forall k \in \mathcal{V}_s, \forall s \in \mathcal{S}$, takes the form [34]

$$\mathbf{w}_k^{(t^{(r)+1})} = \begin{cases} \mathbf{w}_k^{(t^{(r)})} - \gamma g_k(\mathbf{w}_k^{(t^{(r)})}), & \text{if } t^{(r)}|I^{(r)} \neq 0, \\ \frac{\sum_{k \in \mathcal{V}_s} D_k^{(t^{(r)})} (\mathbf{w}_k^{(t^{(r)})} - \gamma g_k(\mathbf{w}_k^{(t^{(r)})}))}{\tilde{D}_s^{(t^{(r)})}}, & \text{if } t^{(r)}|I^{(r)} = 0 \text{ and } t^{(r)}|G \neq 0, \\ \frac{\sum_{k \in \mathcal{K}} D_k^{(t^{(r)})} (\mathbf{w}_k^{(t^{(r)})} - \gamma g_k(\mathbf{w}_k^{(t^{(r)})}))}{\bar{D}^{(t^{(r)})}}, & \text{if } t^{(r)}|G = 0, \end{cases} \quad (2)$$

where $\gamma > 0$ is the learning rate, $g_k(\mathbf{w})$ is the local gradient of device k with model parameter \mathbf{w} , $D_k^{(t^{(r)})} = a_k^{(t^{(r)})} D_k$,

$\tilde{D}_s^{(t^{(r)})} = \sum_{k \in \mathcal{V}_s} a_k^{(t^{(r)})} D_k$, and $\bar{D}^{(t^{(r)})} = \sum_{k \in \mathcal{K}} a_k^{(t^{(r)})} D_k$. Moreover, to continue performing local iterations, each edge server s broadcasts the updated edge model to the devices in \mathcal{V}_s after performing the edge aggregation. Similarly, after performing global aggregation, the cloud server broadcasts the updated global model to all the devices for the next round of training.

B. Latency Model

In each round, the training latency contains the following three parts. The first part is the local computation latency and local model transmission latency of devices. The second part is the edge aggregation latency, edge model broadcasting latency, and edge model uploading latency of the edge servers. The last part is global aggregation and global model broadcasting latency of the cloud server. Since the edge and cloud servers are commonly rich in computation resource, we do not consider the computation latency for performing edge aggregation and global aggregation [17]. In addition, since the cloud and edge servers typically have high transmit power, we do not consider the transmission latency caused by them. Hence, in this paper, we mainly consider the latency caused by local computation and local model uploading.

1) *Local Computation Latency*: Due to the randomness of the local computation capability, we adapt the shifted exponential distribution to characterize the probability distribution of computation latency τ_k^c for device k to perform $I^{(r)}$ local iterations [29], i.e.,

$$\mathbb{P}(\tau_k^{c,r}(I^{(r)}) \leq \psi) = \begin{cases} 1 - e^{-\frac{\mu_k(\psi - I^{(r)} D_k q_k)}{I^{(r)} D_k}} & \text{if } \psi \geq I^{(r)} D_k q_k, \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

where $\mu_k > 0$ and $q_k > 0$ are the constants that indicate the fluctuation and the maximum of the computation capability, respectively, and $I^{(r)} D_k q_k$ is the minimal time consumed by device k to perform $I^{(r)}$ local iterations.

2) *Model Uploading Latency*: We assume the spectrum resource is divided into K orthogonal radio access channels, and each device can access to at most one channel. Then the system bandwidth is denoted by B , and the uplink rate of device k can be written as $d_k^{(r)} = b_k^{(r)} \log_2(1 + \frac{|h_k^{(r)}|^2 p_k}{N_0})$, where p_k is the transmit power of device k , $h_k^{(r)}$ is the average channel gain between device k and its associated edge server, $b_k^{(r)}$ is the allocated bandwidth of device k such that $\sum_{k=1}^K b_k^{(r)} = B$, and N_0 is the background noise. Then the uploading latency for device k to transmit the updated model to edge server s is denoted as $\tau_k^{u,r} = \frac{M}{d_k^{(r)}}$, where M is the size of local model.

C. Problem Formulation

A joint problem of edge aggregation interval control and resource allocation is considered in this work, which minimizes the weighted sum of the training loss and the training time allocation. For simplicity of expression, we first define $\mathcal{R} = \{1, 2, \dots, R\}$, $\mathcal{I} = \{I^{(1)}, I^{(2)}, \dots, I^{(R)}\}$, $\mathcal{T} =$

$\{\bar{\tau}^{(1)}, \bar{\tau}^{(2)}, \dots, \bar{\tau}^{(R)}\}$, and $\mathcal{B} = \{\mathcal{B}^{(1)}, \mathcal{B}^{(2)}, \dots, \mathcal{B}^{(R)}\}$ with $\mathcal{B}^{(r)} = \{b_1^{(r)}, b_2^{(r)}, \dots, b_K^{(r)}\}$. Then the optimization problem is formulated as

$$\min_{\mathcal{I}, \mathcal{T}, \mathcal{B}} \rho \bar{F}(\bar{\mathbf{w}}^{(T)}) + (1 - \rho) \sum_{r=1}^R \bar{\tau}^{(r)} \quad (4a)$$

$$\text{s.t. } I^{(r)} |G| = 0, \forall r \in \mathcal{R}, \quad (4b)$$

$$\sum_{k=1}^K b_k^{(r)} = B, \forall r \in \mathcal{R}, \quad (4c)$$

$$b_k^{(r)} \geq 0, \forall k \in \mathcal{K}, \forall r \in \mathcal{R}, \quad (4d)$$

$$\bar{\tau}^{(r)} \geq 0, \forall r \in \mathcal{R}, \quad (4e)$$

where $T = RG$ is the total number of local iterations, $\bar{\mathbf{w}}^{(T)}$ is the global model after performing T local iterations, and ρ is a weight to balance the training loss and the delay. It is difficult to solve problem (4) directly, because no exact closed-form expression of $\bar{F}(\bar{\mathbf{w}}^{(T)})$ with respect to the optimized variables \mathcal{I} , \mathcal{T} , and \mathcal{B} is available. Therefore, in the next section, we first make the convergence analysis to investigate the effects of these variables on the learning performance. Based on the convergence analysis results, then an alternative problem of edge aggregation interval control, training time allocation, and bandwidth allocation is formulated, which can be solved by an iterative optimization algorithm.

III. CONVERGENCE ANALYSIS AND ADAPTIVE OPTIMIZATION ALGORITHM

A. Convergence Analysis

For ease of convergence analysis, we remove the index r of $t^{(r)}$ and denote t ($1 \leq t \leq T$) as the number of performing local iterations. Besides, we denote the gradients on the dataset $\tilde{\mathcal{D}}_s$ and \mathcal{D} as $\nabla \tilde{F}_s(\mathbf{w})$ and $\nabla \bar{F}(\mathbf{w})$, respectively. Then we introduce the following assumptions.

Assumption 1: $g_k(\mathbf{w})$, $\nabla \bar{F}(\mathbf{w})$, and $\nabla \tilde{F}_s(\mathbf{w})$ are L -smooth, i.e., $\|g_k(\mathbf{w}) - g_k(\mathbf{w}')\| \leq L \|\mathbf{w} - \mathbf{w}'\|$, $\|\nabla \tilde{F}_s(\mathbf{w}) - \nabla \tilde{F}_s(\mathbf{w}')\| \leq L \|\mathbf{w} - \mathbf{w}'\|$, and $\|\nabla \bar{F}(\mathbf{w}) - \nabla \bar{F}(\mathbf{w}')\| \leq L \|\mathbf{w} - \mathbf{w}'\|$, $\forall \mathbf{w}, \mathbf{w}'$.

Assumption 2: The local gradients satisfies $\|g_k(\mathbf{w})\|^2 \leq \zeta^2$, $\forall \mathbf{w}$. The bounded edge divergence and global divergence satisfy $\|\nabla g_k(\mathbf{w}) - \nabla \tilde{F}_s(\mathbf{w})\|^2 \leq \epsilon^2$ and $\|\nabla \bar{F}(\mathbf{w}) - \nabla \tilde{F}_s(\mathbf{w})\|^2 \leq \epsilon^2$, $\forall \mathbf{w}$, respectively.

The above assumptions are widely used for the convergence analysis [27], [38], [39]. Besides, a more generalized metric named the average global gradient deviation $\frac{1}{T} \sum_{t=1}^T \mathbb{E} \|\nabla \bar{F}(\bar{\mathbf{w}}^{(t)})\|^2$ [34], [35] is proposed to approximate the training loss function $\bar{F}(\bar{\mathbf{w}}^{(T)})$. Note that compared with the upper bound of $\bar{F}(\bar{\mathbf{w}}^{(T)})$, the upper bound of $\frac{1}{T} \sum_{t=1}^T \mathbb{E} \|\nabla \bar{F}(\bar{\mathbf{w}}^{(t)})\|^2$ can reflect the similar effects of the optimized variables.

Theorem 1: Given the optimal global model $\bar{\mathbf{w}}^*$ and the learning rate $0 < \gamma \leq \frac{1}{L}$, the upper bound of the average global

gradient deviation is given as follows

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \mathbb{E} \|\nabla \bar{F}(\bar{\mathbf{w}}^{(t)})\|^2 &\leq \frac{2}{T\gamma} \left[\bar{F}(\bar{\mathbf{w}}^{(1)}) - \bar{F}(\bar{\mathbf{w}}^{(*)}) \right] \\ &+ \frac{1}{T} \sum_{t=1}^T \left(\frac{3L^2S}{\bar{D}^2} \sum_{s \in \mathcal{S}} \tilde{D}_s^2 \mathbb{E} \|\bar{\mathbf{w}}^{(t)} - \tilde{\mathbf{w}}_s^{(t)}\|^2 \right. \\ &\left. + \frac{3L^2K}{\bar{D}^2} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} D_k^2 \mathbb{E} \|\tilde{\mathbf{w}}_s^{(t)} - \mathbf{w}_k^{(t)}\|^2 + \frac{12\zeta^2 \mathbb{E}(\bar{D} - \bar{D}^{(t)})}{\bar{D}} \right). \end{aligned} \quad (5)$$

Proof: See Appendix A for reference.

From *Theorem 1*, we can observe that the upper bound of $\frac{1}{T} \sum_{t=1}^T \mathbb{E} \|\nabla F(\bar{\mathbf{w}}^{(t)})\|^2$ decreases as $D^{(t)}$ increases, which is determined by the number of the successfully participating devices. However, the effects of the edge aggregation interval is still not clear. Thus, we introduce the following theorem.

Theorem 2: Suppose $0 \leq \gamma \leq \min\{\frac{1}{L}, \sqrt{\frac{\bar{D}^2}{9G^2L^2(\bar{D}^2+S \sum_{s \in \mathcal{S}} \tilde{D}_s^2)}}\}$, $\forall s \in \mathcal{S}$, and let

$$\begin{aligned} \Gamma_{k,s}(\tilde{I}^{(t)}) &= \frac{72\gamma^2 G^2 L^2 S \zeta^2 \Phi_1 (\bar{D} D_s + \sum_{i \in \mathcal{S}} \tilde{D}_i^2)}{\bar{D}^3} + \frac{12\zeta^2}{\bar{D}} \\ &+ \left(\frac{18\gamma^2 G^2 L^4 S \Phi_1 (|\mathcal{V}_s| \bar{D}^2 + K \sum_{i \in \mathcal{S}} \tilde{D}_i^2)}{\bar{D}^4} + \frac{3L^2 K}{\bar{D}^2} \right) \\ &\frac{12\gamma^2 \zeta^2 \Phi_2 \sum_{j \in \mathcal{V}_s} D_j^2 (\tilde{I}^{(t)})^2}{\tilde{D}_s}, \end{aligned} \quad (6)$$

and

$$\begin{aligned} \Upsilon_{k,s}(\tilde{I}^{(t)}) &= \left(\frac{18\gamma^2 G^2 L^4 S \Phi_1 (|\mathcal{V}_s| \bar{D}^2 + K \sum_{i \in \mathcal{S}} \tilde{D}_i^2)}{\bar{D}^4} + \frac{3L^2 K}{\bar{D}^2} \right) \\ &\frac{6\gamma^2 \epsilon^2 D_k^2 \Phi_2 (\tilde{D}_s^2 + |\mathcal{V}_s| \sum_{j \in \mathcal{V}_s} D_j^2) (\tilde{I}^{(t)})^2}{\tilde{D}_s^2}, \end{aligned} \quad (7)$$

where $\Phi_1 = \frac{\bar{D}^2}{\bar{D}^2 - 9\gamma^2 G^2 L^2 (\bar{D}^2 + S \sum_{i \in \mathcal{S}} \tilde{D}_i^2)}$, and $\Phi_2 = \max_{s \in \mathcal{S}} \frac{\tilde{D}_s^2}{\bar{D}_s^2 - 6\gamma^2 G^2 L^2 (\bar{D}_s^2 + |\mathcal{V}_s| \sum_{j \in \mathcal{V}_s} D_j^2)}$.

Then the upper bound of $\frac{1}{T} \sum_{t=1}^T \mathbb{E} \|\nabla F(\bar{\mathbf{w}}^{(t)})\|^2$ takes the form

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \mathbb{E} \|\nabla \bar{F}(\bar{\mathbf{w}}^{(t)})\|^2 &\leq \frac{2}{T\gamma} \left[\bar{F}(\bar{\mathbf{w}}^{(1)}) - \bar{F}(\bar{\mathbf{w}}^{(*)}) \right] \\ &+ \frac{1}{T} \sum_{t=1}^T \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \left(\Gamma_{k,s}(\tilde{I}^{(t)}) \mathbb{E}(D_k - D_k^{(t)}) + \Upsilon_{k,s}(\tilde{I}^{(t)}) \right. \\ &\left. + \frac{27\gamma^2 G^2 L^2 S \epsilon^2 \Phi_1 \tilde{D}_s^2}{|\mathcal{V}_s| \bar{D}^2} \right), \end{aligned} \quad (8)$$

where $\tilde{I}^{(t)}$ is equivalent to $I^{(r)}$, when $(r-1)G+1 \leq t \leq rG$.

Proof: See Appendix B for reference.

From *Theorem 2*, we notice that the upper bound of $\frac{1}{T} \sum_{t=1}^T \mathbb{E} \|\nabla \bar{F}(\bar{\mathbf{w}}^{(t)})\|^2$ increases with the convergence property parameters L , ζ , and ϵ , respectively. Moreover, the upper bound decreases with the number of the successfully participating devices and the edge aggregation interval. However, if the time allocated for training is given, the decreased edge aggregation interval can decrease the time reserved for performing local iterations, which results in more devices failing to finish the local computation and model transmission on time. Hence performing the edge aggregation interval control can achieve a balance between the learning performance and the training latency. In addition, if the allocated training time is sufficiently large and all devices can be successfully associated to the corresponding edge servers, we can set the learning rate $\gamma^{(t)} \propto \frac{1}{\sqrt{t}}$. When T is sufficiently large, we can derive that $\Phi_1 \rightarrow 1$ and $\Phi_2 \rightarrow 1$. Then, a simpler upper bound of $\frac{1}{T} \sum_{t=1}^T \mathbb{E} \|\nabla \bar{F}(\bar{\mathbf{w}}^{(t)})\|^2$ is $\mathcal{O}(\frac{1}{\sqrt{T}})$ [35], which is consistent with the results obtained by centralized gradient descent with a non-convex training loss function [40]. Furthermore, since our goal is to minimize the weighted sum of the training loss function and the training latency, by leveraging the upper bound of $\frac{1}{T} \sum_{t=1}^T \mathbb{E} \|\nabla \bar{F}(\bar{\mathbf{w}}^{(t)})\|^2$, the problem (4) can be approximated as the following alternative problem in each round r , i.e.,

$$\begin{aligned} \min_{I^{(r)}, \mathcal{B}^{(r)}, \bar{\tau}^{(r)}} & \rho \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \left(\Gamma_{k,s}(I^{(r)}) D_k e^{-\frac{\mu_k}{I^{(r)} D_k} \left(\frac{I^{(r)} \bar{\tau}^{(r)}}{G} - \tau_k^{u,r} - I^{(r)} D_k q_k \right)} \right. \\ & \left. + \Upsilon_{k,s}(I^{(r)}) \right) + (1-\rho) \bar{\tau}^{(r)} \end{aligned} \quad (9a)$$

$$\text{s.t.} \quad (4b) - (4e). \quad (9b)$$

Problem (9) is a combinatorial optimization problem and is still hard to solve. To find solutions to problem (9), we decompose problem (9) into multiple sub-problems with separated objectives: an edge aggregation interval control problem with fixed training time allocation, and a resource allocation problem including bandwidth allocation and training time allocation with given edge aggregation interval.

B. Proposed Algorithm

1) *Optimization of $I^{(r)}$:* Given $\mathcal{B}^{(r)}$ and $\bar{\tau}^{(r)}$, the edge aggregation control problem for the devices in \mathcal{V}_s is equivalent to

$$\begin{aligned} \min_{I^{(r)}} & \rho \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \left(\Gamma_{k,s}(I^{(r)}) D_k e^{-\frac{\mu_k}{I^{(r)} D_k} \left(\frac{I^{(r)} \bar{\tau}^{(r)}}{G} - \tau_k^{u,r} - I^{(r)} D_k q_k \right)} \right. \\ & \left. + \Upsilon_{k,s}(I^{(r)}) \right) \end{aligned} \quad (10a)$$

$$\text{s.t.} \quad (4b). \quad (10b)$$

Rather than using a traversal method, we adopt a more efficient method named relaxation and rounding [41], in which we first

relax the integer constraint (4b) as $1 \leq I^{(r)} \leq G$. Then the relaxed problem is convex, and can be solved by using the following theorem.

Theorem 3: The optimal solution $I^{r,*}$ in the relaxed problem (10) satisfies

$$I^{r,*} = \begin{cases} 1, & \text{if } \frac{\partial \Lambda(I^{(r)})}{\partial I^{(r)}} \Big|_{I^{(r)}=1} > 0, \\ G, & \text{if } \frac{\partial \Lambda(I^{(r)})}{\partial I^{(r)}} \Big|_{I^{(r)}=G} < 0, \\ \hat{I}^{r,*}, & \text{otherwise,} \end{cases} \quad (11)$$

where $\Lambda(\cdot)$ is the objective function of (9) and $\hat{I}^{r,*}$ is the solution of $\frac{\partial \Lambda(I^{(r)})}{\partial I^{(r)}} = 0$.

Proof: See Appendix C for reference.

2) *Optimization of $\mathcal{B}^{(r)}$ and $\bar{\tau}^{(r)}$:* Given $I^{(r)}$, we first optimize $\mathcal{B}^{(r)}$ with fixed $\bar{\tau}^{(r)}$, then we optimize $\mathcal{B}^{(r)}$ according to the obtained $\bar{\tau}^{(r)}$.

The bandwidth allocation problem is equivalent to

$$\min_{\mathcal{B}^{(r)}} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \left(\Gamma_{k,s}(I^{(r)}) D_k e^{-\frac{\mu_k}{I^{(r)} D_k} \left(\frac{I^{(r)} \bar{\tau}^{(r)}}{G} - \frac{M}{b_k^{(r)} \log_2 \left(1 + \frac{h_k^{(r)} p_k}{N_0} \right)} - I^{(r)} D_k q_k \right)} \right) \quad (12a)$$

$$\text{s.t. (4c), (4d).} \quad (12b)$$

Besides, the training latency control problem is equivalent to

$$\min_{\bar{\tau}^{(r)}} \rho \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \left(\Gamma_{k,s}(I^{(r)}) D_k e^{-\frac{\mu_k}{I^{(r)} D_k} \left(\frac{I^{(r)} \bar{\tau}^{(r)}}{G} - \tau_k^{u,r} - I^{(r)} D_k q_k \right)} \right) + (1 - \rho) \bar{\tau}^{(r)} \quad (13a)$$

$$\text{s.t. (4e).} \quad (13b)$$

We can solve problem (12) using the follow theorem.

Theorem 4: The optimal solution $b_k^{r,*}$, $\forall k \in \mathcal{V}_s$, of problem (12) is

$$b_k^{r,*} = \frac{\mu_k M}{2 I^{(r)} D_k \log_2 \left(1 + \frac{h_k^{(r)} p_k}{N_0} \right) \mathcal{W} \left(\frac{\nu^* \mu_k M e^{\frac{\mu_k}{I^{(r)} D_k} \left(\frac{I^{(r)} \bar{\tau}^{(r)}}{G} - I^{(r)} D_k q_k \right)}}{4 \Gamma_{k,s}(I^{(r)}) D_k^2 I^{(r)} \log_2 \left(1 + \frac{h_k^{(r)} p_k}{N_0} \right)} \right)}, \quad (14)$$

where $\mathcal{W}(\cdot)$ is the Lambert \mathcal{W} function and ν^* is evaluated according to the constraint (4c).

Proof: See Appendix D for reference.

Then we solve problem (13) using the follow theorem.

Theorem 5: The optimal solution $\bar{\tau}^{*,r}$ of problem (13) satisfies

$$\bar{\tau}^{*,r} = \begin{cases} 0 & \text{if } \frac{\partial \Lambda(\bar{\tau}^{(r)})}{\partial \bar{\tau}^{(r)}} \Big|_{\bar{\tau}^{(r)}=0} > 0, \\ \hat{\tau}^{*,r} & \text{otherwise,} \end{cases} \quad (15)$$

where $\hat{\tau}^{*,r}$ is the solution of $\frac{\partial \Lambda(\bar{\tau}^{(r)})}{\partial \bar{\tau}^{(r)}} = 0$.

Proof: See Appendix E for reference.

Algorithm 1: Iterative Algorithm for Resource Allocation.

- 1: **Initialize:** Set feasible solutions of $I^{(r)}$, $\mathcal{B}^{(r)}$, and $\bar{\tau}^{(r)}$.
 - 2: **repeat:**
 - 3: With given $\mathcal{B}^{(r)}$ and $\bar{\tau}^{(r)}$, obtain the optimal solution of $I^{(r)}$ according to **Theorem 3**.
 - 4: With given $I^{(r)}$, obtain the solutions of $\mathcal{B}^{(r)}$ and $\bar{\tau}^{(r)}$ according to **Theorem 4** and **Theorem 5**, respectively.
 - 5: **Until** Objective value of problem (9) converges.
 - 6: Evaluate $G^{(r)}$.
-

Since the relaxed problem of (10), problem (12), and problem (13) are all convex, problem (9) can be solved by an iterative optimization algorithm in which each step is non-increasing. To achieve a practicable solution to problem (9), i.e., an integer solution of $I^{(r)}$ satisfying the constraint (4b), we introduce the following update rule of $I^{(r)}$ and G . Specifically, we first obtain $I^{r,*}$ according to Theorem 3. Then we evaluate the actual number of performing local iterations $G^{(r)} = \lfloor \frac{G}{I^{r,*}} \rfloor I^{r,*}$. The complete procedure is shown in **Algorithm 1**. Note that the relaxed problem of (10) can be solved by using the bisection method, and the time complexity is $\mathcal{O}(\log_2 \frac{1}{\beta_1})$, where β_1 is the accuracy of the bisection method. Similarly, we can notice that b_k^* and $\hat{\tau}^{*,r}$ in problem (12) and (13) can also be solved by using the bisection method, and the time complexity is $\mathcal{O}(K \log_2 \frac{1}{\beta_2})$ and $\mathcal{O}(\log_2 \frac{1}{\beta_3})$, respectively, where β_2 and β_3 are the accuracy of the bisection method.

C. The Whole Policy

As shown in **Algorithm 2**, we give the complete procedure of the proposed adaptive hierarchical FL algorithm. In steps 1-2, local models w_k and the convergence property parameters such as L , ζ , ϵ , and ϵ are initialized. Besides, at the beginning of each round, i.e., in step 4, the edge aggregation interval and the training latency can be optimized according to **Algorithm 1**. Then the actual number of performing local iterations is determined. After that, steps 5-22 are consistent with the learning model designed in Section II. In addition, as shown in steps 14, 15, and 19, to perform the next round of training, we adapt the method proposed in [29] to estimate the convergence property parameters L , ζ , and ϵ , respectively. Specially, to reduce the complexity of **Algorithm 1**, the convergence property parameters are estimated at the end of each training round instead of after each edge aggregation. Besides, since these parameters are the scalars, we do not consider the transmission latency of these parameters. As a whole, compared to the existing hierarchical FL algorithm, the additional computational complexity of the proposed algorithm mainly consists of two parts. The first part is the additional computation complexity to solve the problem (9), which is $\mathcal{O}(\varpi(\log_2 \frac{1}{\beta_1} + K \log_2 \frac{1}{\beta_2} + \log_2 \frac{1}{\beta_3}))$, where ϖ is the total number of iterations in **Algorithm 1**. The second part additional computation complexity to evaluate the convergence property parameters, which is $\mathcal{O}(K)$. Since the cloud server has the powerful computation capability to solve the problem (9),

Algorithm 2: The Proposed Adaptive Hierarchical FL Algorithm.

```

1: Initialize  $\mathbf{w}_k^{(1)}, \forall k \in \mathcal{K}$ , as a random vector.
2: Initialize  $L^{(1)}, \zeta^{(1)}, \epsilon^{(1)}$ , and  $\epsilon^{(1)}$ .
3: for  $r = 1, 2, \dots, R$  do
4:   Evaluate  $I^{(r)}, \mathcal{B}^{(r)}, \bar{\tau}^{(r)}$ , and  $G^{(r)}$ .
5:   for
      $t = \sum_{j=1}^{r-1} G^{(j)} + 1, \sum_{j=1}^{r-1} G^{(j)} + 2, \dots, \sum_{j=1}^r G^{(j)}$ 
     do
6:     foreach server  $s = 1, 2, \dots, S$  in parallel do
7:       Each device in  $\mathcal{V}_s$  updates its local model.
8:       if  $-\sum_{j=1}^{r-1} G^{(j)} | I^{(r)} = 0$  then
9:         Devices transmit their local models to
           the corresponding edge servers.
10:      Edge server  $s$  performs the edge model
        aggregation.
11:      if  $-\sum_{j=1}^{r-1} G^{(j)} | G^{(r)} \neq 0$  then
12:        Edge server  $s$  broadcasts its updated edge
          model
          to the devices in  $\mathcal{V}_s$ .
13:      else
14:        Edge server  $s$  estimates
           $\nabla \tilde{F}_s(\bar{\mathbf{w}}(\sum_{j=1}^{r-1} G^{(j)})) =$ 

$$\frac{\sum_{k \in \mathcal{V}_s} a_k^{(t)} D_k g_k(\bar{\mathbf{w}}(\sum_{j=1}^{r-1} G^{(j)}))}{\bar{D}_s^{(t)}}, \text{ where}$$


$$g_k(\bar{\mathbf{w}}(\sum_{j=1}^{r-1} G^{(j)})) = \frac{\bar{\mathbf{w}}(\sum_{j=1}^{r-1} G^{(j)}) - \mathbf{w}_k^{(t+1)}}{G^{(r)} \gamma}.$$

15:        Each device  $k \in \mathcal{V}_s$ , estimates
           $L_k^{(r+1)} = \frac{\|g_k(\bar{\mathbf{w}}(\sum_{j=1}^{r-1} G^{(j)})) - g_k(\mathbf{w}_k^{(t)})\|}{\|\bar{\mathbf{w}}(\sum_{j=1}^{r-1} G^{(j)}) - \mathbf{w}_k^{(t)}\|}$ , and
           $\epsilon_k^{(r+1)} =$ 

$$\|\nabla \tilde{F}_s(\bar{\mathbf{w}}(\sum_{j=1}^{r-1} G^{(j)})) - g_k(\bar{\mathbf{w}}(\sum_{j=1}^{r-1} G^{(j)}))\|.$$

16:        The edge server  $s$  transmits its updated local
          model
          to the cloud server.
17:      end if, end if
18:    end for, end for, end for
19:    Cloud server evaluates  $L^{(r+1)} = \frac{\sum_{k \in \mathcal{K}} a_k^{(t)} D_k L_k^{(r+1)}}{\sum_{k \in \mathcal{K}} a_k^{(t)} D_k}$ ,
      
$$\zeta^{(r+1)} = \frac{\sum_{k \in \mathcal{K}} a_k^{(t)} \|g_k(\bar{\mathbf{w}}(\sum_{j=0}^{r-1} G^{(j)}))\|}{\sum_{k \in \mathcal{K}} a_k^{(t)} D_k}, \text{ and}$$


$$\epsilon^{(r+1)} = \frac{\sum_{k \in \mathcal{K}} a_k^{(t)} D_k \epsilon_k^{(r+1)}}{\sum_{k \in \mathcal{K}} a_k^{(t)} D_k}.$$

20:    Cloud server performs the global model aggregation.
21:    Cloud server broadcasts the updated global model to
      all devices.
22:  end for

```

and the cost to evaluate the convergence property parameters is very low, the overhead in the above two parts can be negligible.

IV. SIMULATION RESULTS

A. Experiment Settings

Unless otherwise specified, we assume that the considered hierarchical FL system consists of $K = 50$ devices and $S = 5$

edge servers which are uniformly deployed in a disc with a radius of 500 m, and a cloud server at the center of the disc. The path loss model is given as $L[\text{dB}] = 128.1 + 37.6 \log_{10} d_{[\text{km}]}$, and the standard deviation of the log-normal shadowing fading is 8 dB [42]. We set the total bandwidth $B = 20$ MHz, and the background noise power $N_0 = 10^{-19}$. To simplify, the transmit power of all devices is set to $p_1 = \dots = p_K = 20$ dBm [29]. We set the parameters of the computation latency model $q_k = 0.1$ ms/samples and $\mu_k = \frac{1}{q_k}$, respectively [29]. The maximal number of clusters is set as We evaluate the performance of the proposed algorithm on the well-known MNIST dataset [43] which has 60,000 training images and 10,000 test images with 10 types of labels. Different training data distributions, i.e., i.i.d. case and non-i.i.d. cases, are considered in this paper. Specifically, in the i.i.d. case, the training samples of the MNIST dataset are randomly partitioned into 50 pieces and each device is assigned a piece. While for the non-i.i.d. cases, the training samples are first partitioned into 10 pieces according to their labels. Let parameter v denote the non-i.i.d. level [29]. Specially, a smaller v represents a higher non-i.i.d. level. Then each piece with the same label is randomly partitioned into $5v$ shards, which means we have $50v$ shards in total. After that, each device is assigned v shards with different labels. Besides, we apply a standard multilayer perceptron model for training, which has one hidden layer of 128 hidden nodes and finally a output layer. The batch size is set as 32, the optimizer is SGD, the training loss function is Cross-entropy, and the learning rate is set as 0.05. The number of training rounds R and the number of performing local iterations per round G are both set as 100.

Furthermore, to show the effectiveness of the proposed algorithm, we introduce five baseline algorithms for comparison. The first one is the common mini-batch SGD with partial participation (denoted by CP) in which the edge aggregation interval is set as $I^{(r)} = 1$. The second one is common mini-batch SGD with full participation (denoted by CF) in which the edge aggregation interval is set as $I^{(r)} = 1$ and all devices can participate in training successfully. The third one is fixed edge aggregation interval with partial participation (denoted by FP) in which the edge aggregation interval is set as $I^{(r)} = 100$. The fourth one is fixed edge aggregation interval with full participation (denoted by FF), in which the edge aggregation interval is set as $I^{(r)} = 100$ and all devices can participate in training successfully. The last one is the adaptive edge aggregation interval with equal bandwidth allocation (denoted by AE).

B. Evaluation of the Convergence Property Parameters

As shown in Fig. 2, since the convergence property parameters L , ζ , ϵ , and ϵ directly affect the learning performance and the optimization algorithm design, we first show the trends of these parameters v.s. the number of training rounds under different training data distributions. Specially, if no devices can successfully participate in training, the evaluated parameters are set as 0. In Figs. 2(a), we can notice that the training loss function in the non-i.i.d. cases is less smooth and has higher value of L than that in the i.i.d. case. This is because when the local datasets of different devices are non-i.i.d., given the same

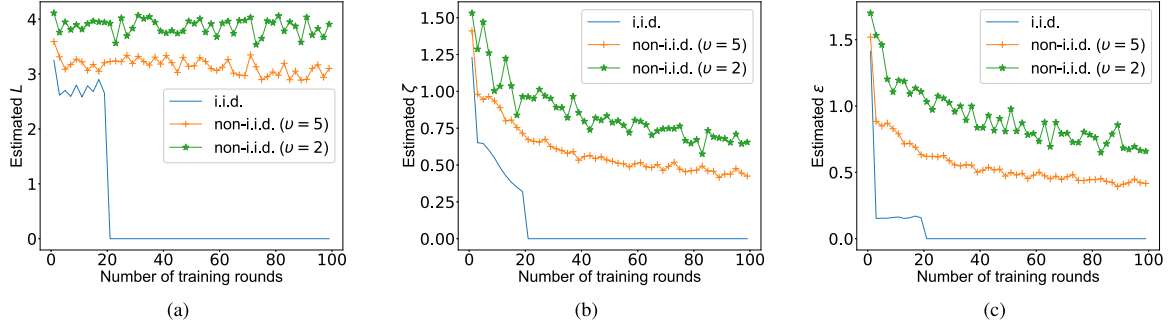


Fig. 2. Evaluations of the convergence property parameters on MNIST with different training data distributions.

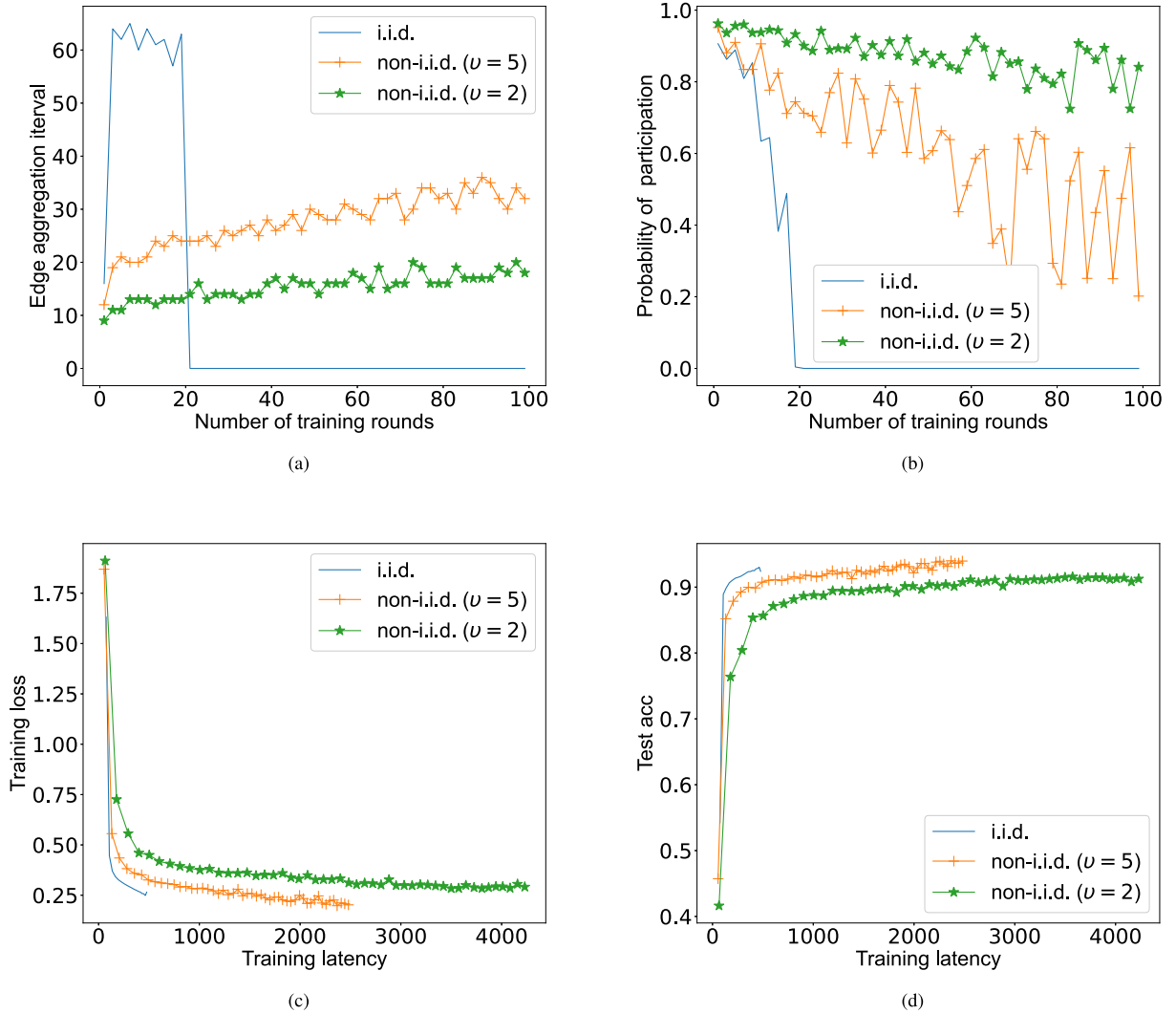


Fig. 3. Performance of the proposed algorithm on MNIST with different training data distributions. (a) the edge aggregation interval v.s. the number of training rounds. (b) the participation probability v.s. the number of training rounds. (c) the training loss v.s. the training latency. (d) the test accuracy v.s. the training latency.

global model, the local gradients or the local models among the devices vary greatly. Besides, as shown in Figs. 2(b) and 2(c), we can observe the similar results with respect to ζ and ϵ . Furthermore, it is finding that all of the three parameters have a slow trend. This is because during the training process, the training loss function continues to decrease and tend to convergence.

C. Performance Under Different Training Data Distributions

In Fig. 3, we show the performance of the proposed algorithm with different training data distributions. Figs. 3(a) and 3(b) show the edge aggregation interval $I^{(r)}$ and the probability of participation $\frac{\sum_{k=1}^K \mathbb{P}(a_k^{(t(r))}=1)}{K}$ v.s. the number of training rounds. Besides, Figs. 3(c) and 3(d) show the training loss and

TABLE III
AVERAGE EDGE AGGREGATION INTERVAL AND AVERAGE PARTICIPATING PROBABILITY OF DIFFERENT ALGORITHMS ON DIFFERENT DATASETS

Algorithms	Datasets	Average edge aggregation interval	Average participating probability
Proposed	MNIST	15.1	88.3%
Proposed	CIFAR-10	7.3	97.3%
CP	MNIST	1	83.6%
CP	CIFAR-10	1	90.2%
FP	MNIST	100	91.2%
FP	CIFAR-10	100	98.9%
AE	MNIST	16.5	88.4%
AE	CIFAR-10	7.4	94.4%

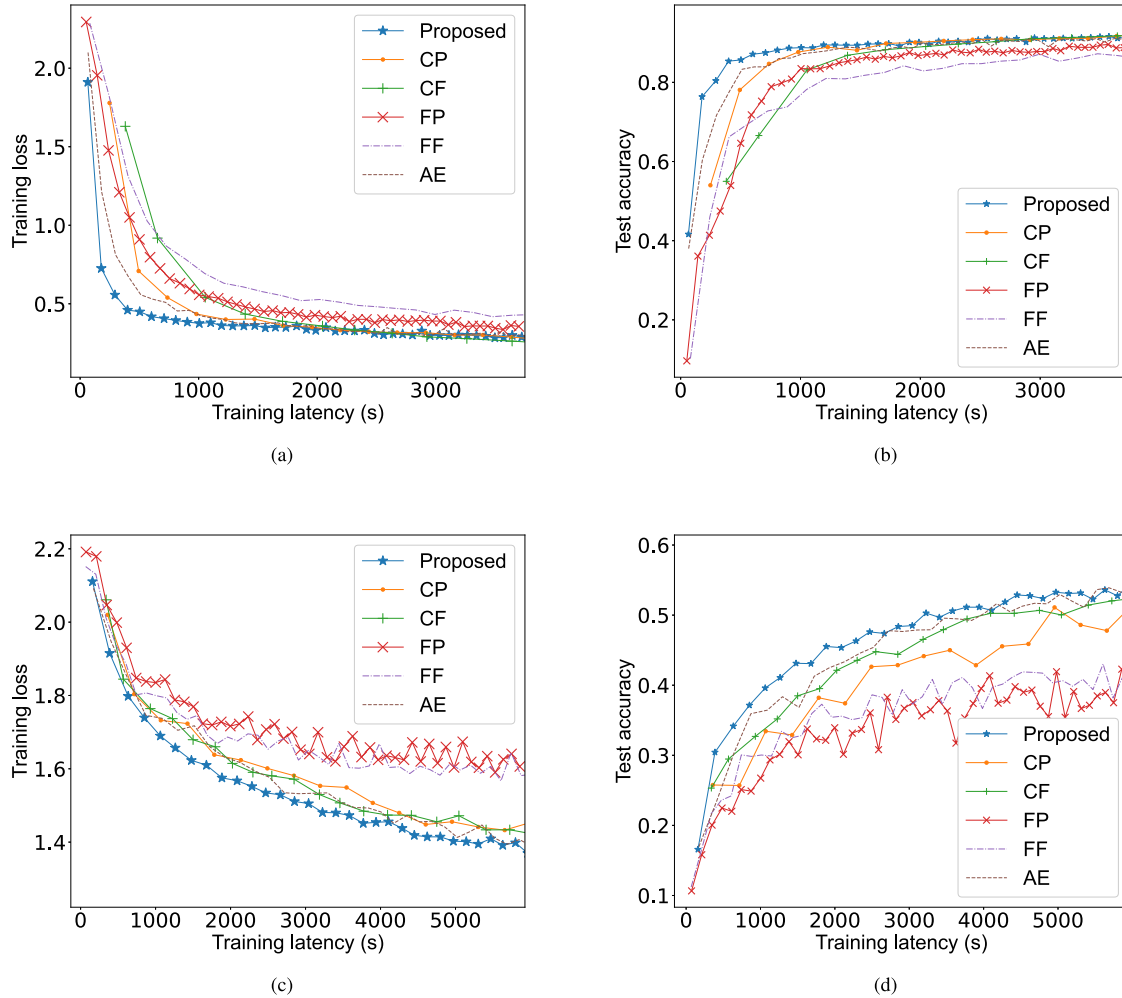


Fig. 4. Comparison among different algorithms on MNIST and CIFAR-10 in the non-i.i.d case with $v = 2$. (a) the training loss v.s. the training latency on MNIST. (b) the test accuracy v.s. the training latency on MNIST. (c) the training loss v.s. the training latency on CIFAR-10. (d) the test accuracy v.s. the training latency on CIFAR-10.

test accuracy v.s. the training latency. From Fig. 3(a) and 3(b), we can find that the edge aggregation interval decreases and the average probability of participation increases with the non-i.i.d. level, respectively. The reason is that there is a trade-off between the learning performance and the training latency. Specifically, in the non-i.i.d. cases, the weight of the learning performance is increased. Hence, to improve the learning performance, we can reduce the edge aggregation interval and let more devices

successfully participate in training at the cost of larger training latency. It is interesting to notice that the optimized edge aggregation interval in the i.i.d. case is more close to the number G of local iterations per round. This is because when the local datasets of different devices are i.i.d., devices have similar local data distributions and can achieve higher learning performance without performing too many edge aggregations. Besides, from Figs. 3(c) and 3(d), we can observe that after 100 rounds of

training, the proposed algorithm in the i.i.d. case can converge with the lowest training latency. The above results validate that the proposed algorithm can adaptively adjust the edge aggregation interval and the resource allocation strategy according to the non-i.i.d. level during the training process.

D. Comparison of Different Algorithms

Furthermore, we compare the proposed algorithm with the five other baseline algorithms in the non-i.i.d case with $v = 2$. Besides, we consider a more challenging learning task based on the CIFAR-10 dataset [44] for image classification, which has 50,000 training images and 10,000 test images. For this task, we use a CNN model with two 3×3 convolution layers with respective 16 and 32 channels, also followed with 2×2 max pooling, one fully connected layers with 120 units, and finally a softmax output layer. The batch size is 32, the optimizer is SGD, and the learning rate is 0.02. Table III summarizes the average edge aggregation interval $\frac{\sum_{r=1}^R I^{(r)}}{R}$ and the average participating probability $\frac{\sum_{r=1}^R \sum_{k=1}^K \mathbb{P}(a_k^{(t(r))}=1)}{RK}$ of different algorithms on MNIST and CIFAR-10 datasets. It is shown that when dealing with the task of CIFAR-10, the proposed algorithm needs to train with smaller edge aggregation interval and require more devices to participate in training. Moreover, Fig. 4 shows that the proposed algorithm can achieve lower training loss and higher test accuracy with lower training latency compared to the baseline algorithms on the both two datasets. The advantage of the proposed algorithm is twofold: firstly, the proposed algorithm always outperforms the baselines since it well balances the edge aggregation interval and the time consumption in training process. Besides, in the proposed algorithm only a part of devices can successfully participate in the training, which can avoid the straggler issue caused by bad channel states. Hence we can conclude that the proposed algorithm can simultaneously improve the learning performance and reduce the training latency.

V. CONCLUSION

In this paper, we formulated a joint problem of edge aggregation interval control and training time allocation, aiming at minimizing the weight sum of training loss and training latency. According to the convergence analysis results, which quantified the effects of the edge aggregation interval, the training latency, and the number of successfully participating devices, was derived. We then reformulated an alternative problem, which can be decoupled into three sub-problems and solved iteratively. The simulation results have shown that the proposed algorithm, compared to state of the art benchmarks, by jointly performing edge aggregation interval control and resource allocation, can achieve higher learning performance with lower training latency.

APPENDIX

A. Proof of Theorem 1

Although edge model $\tilde{\mathbf{w}}_s^{(t)}$ and global model $\bar{\mathbf{w}}^{(t)}$ are only updated at the stages of edge aggregation and cloud aggregation, respectively, we can use them for analysis and assume that they

can be observed at any local iteration t . When $t = (r - 1)G + c$, where $r \in \{1, 2, \dots, R\}$ and $c \in \{1, 2, \dots, G\}$, we have

$$\begin{aligned} \mathbb{E} \bar{F}(\bar{\mathbf{w}}^{(t+1)}) &= \mathbb{E} \bar{F} \left(\bar{\mathbf{w}}^{(t)} - \gamma \frac{1}{\bar{D}(t)} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} a_k^{(t)} D_k g_k(\mathbf{w}_k^{(t)}) \right) \\ &\stackrel{(a)}{\leq} \mathbb{E} \bar{F}(\bar{\mathbf{w}}^{(t)}) - \gamma \mathbb{E} \left\langle \nabla \bar{F}(\bar{\mathbf{w}}^{(t)}), \frac{1}{\bar{D}(t)} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} a_k^{(t)} D_k g_k(\mathbf{w}_k^{(t)}) \right\rangle \\ &\quad + \frac{\gamma^2 L}{2} \mathbb{E} \left\| \frac{1}{\bar{D}(t)} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} a_k^{(t)} D_k g_k(\mathbf{w}_k^{(t)}) \right\|^2 \\ &\stackrel{(b)}{\leq} \mathbb{E} F(\bar{\mathbf{w}}^{(t)}) + \frac{\gamma}{2} \left(\mathbb{E} \left\| \nabla \bar{F}(\bar{\mathbf{w}}^{(t)}) \right\|^2 \right. \\ &\quad \left. - \frac{1}{\bar{D}(t)} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} a_k^{(t)} D_k g_k(\mathbf{w}_k^{(t)}) \right\|^2 - \mathbb{E} \left\| \nabla \bar{F}(\bar{\mathbf{w}}^{(t)}) \right\|^2 \right), \end{aligned} \quad (16)$$

where (a) holds due to the proposition of Lipschitz smooth [38], and (b) holds because we suppose $\gamma \leq \frac{1}{L}$. Then we can obtain

$$\begin{aligned} &\mathbb{E} \left\| \nabla \bar{F}(\bar{\mathbf{w}}^{(t)}) - \frac{1}{\bar{D}(t)} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} a_k^{(t)} D_k g_k(\mathbf{w}_k^{(t)}) \right\|^2 \\ &\leq \mathbb{E} \left\| \nabla \bar{F}(\bar{\mathbf{w}}^{(t)}) - \frac{1}{\bar{D}} \sum_{s \in \mathcal{S}} \tilde{D}_s \nabla \tilde{F}_s(\tilde{\mathbf{w}}_s^{(t)}) \right\|^2 \\ &\quad + \frac{1}{\bar{D}} \sum_{s \in \mathcal{S}} \tilde{D}_s \nabla \tilde{F}_s(\tilde{\mathbf{w}}_s^{(t)}) - \frac{1}{\bar{D}} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} D_k g_k(\mathbf{w}_k^{(t)}) \\ &\quad + \frac{1}{\bar{D}} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} D_k g_k(\mathbf{w}_k^{(t)}) - \frac{1}{\bar{D}(t)} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} a_k^{(t)} D_k g_k(\mathbf{w}_k^{(t)}) \Big\|^2 \\ &\stackrel{(a)}{\leq} \frac{3L^2 S}{\bar{D}^2} \sum_{s \in \mathcal{S}} \tilde{D}_s^2 \mathbb{E} \|\bar{\mathbf{w}}^{(t)} - \tilde{\mathbf{w}}_s^{(t)}\|^2 \\ &\quad + \frac{3L^2 K}{\bar{D}^2} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} D_k^2 \mathbb{E} \|\tilde{\mathbf{w}}_s^{(t)} - \mathbf{w}_k^{(t)}\|^2 + \frac{12\zeta^2 \mathbb{E}(\bar{D} - \bar{D}(t))}{\bar{D}}. \end{aligned} \quad (17)$$

where (a) holds because of Assumption 2 and Theorem 1 in [27]. Then substituting (17) into (16) yields

$$\begin{aligned} &\frac{1}{T} \sum_{t=1}^T \mathbb{E} \|\nabla \bar{F}(\bar{\mathbf{w}}^{(t)})\|^2 \leq \frac{2}{T\gamma} [\bar{F}(\bar{\mathbf{w}}^{(1)}) - \bar{F}(\bar{\mathbf{w}}^{(*)})] \\ &\quad + \frac{1}{T} \sum_{t=1}^T \left(\frac{3L^2 S}{\bar{D}^2} \sum_{s \in \mathcal{S}} \tilde{D}_s^2 \mathbb{E} \|\bar{\mathbf{w}}^{(t)} - \tilde{\mathbf{w}}_s^{(t)}\|^2 \right. \\ &\quad \left. + \frac{3L^2 K}{\bar{D}^2} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} D_k^2 \mathbb{E} \|\tilde{\mathbf{w}}_s^{(t)} - \mathbf{w}_k^{(t)}\|^2 + \frac{12\zeta^2 \mathbb{E}(\bar{D} - \bar{D}(t))}{\bar{D}} \right), \end{aligned} \quad (18)$$

where \mathbf{w}^* is the optimal global mode. Then we complete the proof.

B. Proof of Theorem 2

We first investigate how the training process affects $\frac{1}{T} \sum_{t=1}^T \sum_{s \in \mathcal{S}} \tilde{D}_s^2 \mathbb{E} \|\bar{\mathbf{w}}^{(t)} - \tilde{\mathbf{w}}_s^{(t)}\|^2$ and $\frac{1}{T} \sum_{t=1}^T \sum_{k \in \mathcal{V}_s} D_k^2 \mathbb{E} \|\tilde{\mathbf{w}}_s^{(t)} - \mathbf{w}_k^{(t)}\|^2$. When $t = (r-1)G + c$, where $r \in \{1, 2, \dots, R\}$ and $c \in \{1, 2, \dots, G\}$, we can obtain

$$\begin{aligned} & \mathbb{E} \|\bar{\mathbf{w}}^{(t)} - \tilde{\mathbf{w}}_s^{(t)}\|^2 \\ & \leq 3\gamma^2 \mathbb{E} \left\| \sum_{\alpha=(r-1)G}^{t-1} \left(\frac{1}{\tilde{D}_s^{(\alpha)}} \sum_{k \in \mathcal{V}_s} D_k^{(\alpha)} g_k(\mathbf{w}_k^{(\alpha)}) - \nabla \tilde{F}_s(\tilde{\mathbf{w}}_s^{(\alpha)}) \right) \right\|^2 \\ & + 3\gamma^2 \mathbb{E} \left\| \sum_{\alpha=(r-1)G}^{t-1} \left(\frac{1}{\bar{D}} \sum_{s \in \mathcal{S}} \tilde{D}_s \nabla \tilde{F}_s(\tilde{\mathbf{w}}_s^{(\alpha)}) \right. \right. \\ & \left. \left. - \frac{1}{\bar{D}^{(\alpha)}} \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{V}_i} D_j^{(\alpha)} g_j(\mathbf{w}_j^{(\alpha)}) \right) \right\|^2 \\ & + 3\gamma^2 \mathbb{E} \left\| \sum_{\alpha=(r-1)G}^{t-1} \left(\nabla \tilde{F}_s(\tilde{\mathbf{w}}_s^{(\alpha)}) - \frac{1}{\bar{D}} \sum_{i \in \mathcal{S}} \tilde{D}_i \nabla \tilde{F}_i(\tilde{\mathbf{w}}_i^{(\alpha)}) \right) \right\|^2. \end{aligned} \quad (19)$$

For the first and the second term in the right hand side of (19), similar to the proof in (17), we can obtain (20), shown at the bottom of the page, and

$$\begin{aligned} & 3\gamma^2 \mathbb{E} \left\| \sum_{\alpha=(r-1)G+1}^{t-1} \left(\frac{1}{\bar{D}} \sum_{i \in \mathcal{S}} \tilde{D}_i \nabla \tilde{F}_i(\tilde{\mathbf{w}}_i^{(\alpha)}) \right. \right. \\ & \left. \left. - \frac{1}{\bar{D}^{(\alpha)}} \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{V}_i} a_j^{(\alpha)} D_j g_j(\mathbf{w}_j^{(\alpha)}) \right) \right\|^2 \end{aligned}$$

$$\begin{aligned} & \leq 6\gamma^2 G \sum_{\alpha=(r-1)G+1}^{rG} \left(\frac{L^2 K}{\bar{D}^2} \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{V}_i} D_j^2 \mathbb{E} \|\tilde{\mathbf{w}}_i^{(\alpha)} - \mathbf{w}_j^{(\alpha)}\|^2 \right. \\ & \left. + \frac{4\zeta^2}{\bar{D}} \mathbb{E}(\bar{D} - \bar{D}^{(\alpha)}) \right). \end{aligned} \quad (21)$$

Besides, for the third term in the right hand side of (19), we can derive that

$$\begin{aligned} & 3\gamma^2 \mathbb{E} \left\| \sum_{\alpha=(r-1)G+1}^{t-1} \left(\nabla \tilde{F}_s(\tilde{\mathbf{w}}_s^{(\alpha)}) - \frac{1}{\bar{D}} \sum_{i \in \mathcal{S}} \tilde{D}_i \nabla \tilde{F}_i(\tilde{\mathbf{w}}_i^{(\alpha)}) \right) \right\|^2 \\ & \leq 9\gamma^2 G^2 \epsilon^2 + 9\gamma^2 G L^2 \sum_{\alpha=(r-1)G+1}^{rG} \left(\mathbb{E} \|\bar{\mathbf{w}}^{(\alpha)} - \tilde{\mathbf{w}}_s^{(\alpha)}\|^2 \right. \\ & \left. + \frac{S}{\bar{D}^2} \sum_{i \in \mathcal{S}} \tilde{D}_i^2 \mathbb{E} \|\bar{\mathbf{w}}^{(\alpha)} - \tilde{\mathbf{w}}_i^{(\alpha)}\|^2 \right). \end{aligned} \quad (22)$$

With the results of (20), (21), and (22), we can obtain (23), shown at the bottom of the page.

When $0 \leq \gamma \leq \sqrt{\frac{\bar{D}^2}{9G^2 L^2 (\bar{D}^2 + S \sum_{s \in \mathcal{S}} \tilde{D}_s^2)}}$, we can obtain

$$\begin{aligned} & \frac{1}{T} \sum_{t=1}^T \sum_{s \in \mathcal{S}} \tilde{D}_s^2 \mathbb{E} \|\bar{\mathbf{w}}^{(t)} - \tilde{\mathbf{w}}_s^{(t)}\|^2 \\ & \leq \frac{1}{T} \sum_{t=1}^T \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \frac{24\gamma^2 G^2 \zeta^2 \Phi_1 (\bar{D} D_s + \sum_{i \in \mathcal{S}} \tilde{D}_i^2) \mathbb{E}(D_k - D_k^{(t)})}{\bar{D}} \\ & + \frac{1}{T} \sum_{t=1}^T \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \end{aligned}$$

$$\begin{aligned} & 3\gamma^2 \mathbb{E} \left\| \sum_{\alpha=(r-1)G+1}^{t-1} \left(\frac{1}{\tilde{D}_s^{(\alpha)}} \sum_{k \in \mathcal{V}_s} a_k^{(\alpha)} D_k g_k(\mathbf{w}_k^{(\alpha)}) - \nabla \tilde{F}_s(\tilde{\mathbf{w}}_s^{(\alpha)}) \right) \right\|^2 \leq 6\gamma^2 G \sum_{\alpha=(r-1)G+1}^{rG} \left(\frac{4\zeta^2}{\tilde{D}_s} \mathbb{E}(\tilde{D}_s - \tilde{D}_s^{(\alpha)}) \right. \\ & \left. + \frac{L^2 |\mathcal{V}_s|}{\tilde{D}_s^2} \sum_{k \in \mathcal{V}_s} D_k^2 \mathbb{E} \|\tilde{\mathbf{w}}_s^{(\alpha)} - \mathbf{w}_k^{(\alpha)}\|^2 \right), \end{aligned} \quad (20)$$

$$\begin{aligned} & \frac{1}{T} \sum_{t=1}^T \sum_{s \in \mathcal{S}} \tilde{D}_s^2 \mathbb{E} \|\bar{\mathbf{w}}^{(t)} - \tilde{\mathbf{w}}_s^{(t)}\|^2 = \frac{1}{T} \sum_{r=1}^R \sum_{a=(r-1)G+1}^{rG} \sum_{s \in \mathcal{S}} \tilde{D}_s^2 \mathbb{E} \|\bar{\mathbf{w}}^{(a)} - \tilde{\mathbf{w}}_s^{(a)}\|^2, \\ & \leq \frac{1}{T} \sum_{t=1}^T \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \frac{24\gamma^2 G^2 \zeta^2 (\bar{D} D_s + \sum_{i \in \mathcal{S}} \tilde{D}_i^2) \mathbb{E}(D_k - D_k^{(t)})}{\bar{D}} \\ & + \frac{1}{T} \sum_{t=1}^T \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \frac{6\gamma^2 G^2 L^2 D_k^2 (|\mathcal{V}_s| \bar{D}^2 + K \sum_{i \in \mathcal{S}} \tilde{D}_i^2) \mathbb{E} \|\tilde{\mathbf{w}}_s^{(t)} - \mathbf{w}_k^{(t)}\|^2}{\bar{D}^2} \\ & + \frac{1}{T} \sum_{t=1}^T \sum_{s \in \mathcal{S}} \left(\frac{9\gamma^2 G^2 L^2 \tilde{D}_s^2 (\bar{D}^2 + S \sum_{i \in \mathcal{S}} \tilde{D}_i^2) \mathbb{E} \|\bar{\mathbf{w}}^{(t)} - \tilde{\mathbf{w}}_s^{(t)}\|^2}{\bar{D}^2} \right) + \sum_{s \in \mathcal{S}} 9\gamma^2 G^2 \epsilon^2 \tilde{D}_s^2. \end{aligned} \quad (23)$$

$$\frac{6\gamma^2 G^2 L^2 D_k^2 \Phi_1 (|\mathcal{V}_s| \bar{D}^2 + K \sum_{i \in \mathcal{S}} \bar{D}_i^2) \mathbb{E} \|\tilde{\mathbf{w}}_s^{(t)} - \mathbf{w}_k^{(t)}\|^2}{\bar{D}^2} + \sum_{s \in \mathcal{S}} 9\gamma^2 G^2 \epsilon^2 \Phi_1 \bar{D}_s^2, \quad (24)$$

$$\text{where } \Phi_1 = \frac{\bar{D}^2}{\bar{D}^2 - 9\gamma^2 G^2 L^2 (\bar{D}^2 + S \sum_{i \in \mathcal{S}} \bar{D}_i^2)}.$$

With respect to $\mathbb{E} \|\tilde{\mathbf{w}}_s^{(t)} - \mathbf{w}_k^{(t)}\|^2$, suppose $t = (r-1)G + b^{(r)}I^{(r)} + d$, where $b^{(r)} \in \{0, 1, \dots, \frac{G}{I^{(r)}} - 1\}$, and $d \in \{1, 2, \dots, I^{(r)}\}$. Similar to the proof in (19), we can obtain

$$\begin{aligned} & \mathbb{E} \|\tilde{\mathbf{w}}_s^{(t)} - \mathbf{w}_k^{(t)}\|^2 \\ & \leq 3\gamma^2 \mathbb{E} \left\| \sum_{\alpha=(r-1)G+b^{(r)}I^{(r)}+1}^{t-1} \left(g_k(\mathbf{w}_k^{(\alpha)}) - \nabla \tilde{F}_s(\tilde{\mathbf{w}}_s^{(\alpha)}) \right) \right\|^2 \\ & + 3\gamma^2 \mathbb{E} \left\| \sum_{\alpha=(r-1)G+b^{(r)}I^{(r)}+1}^{t-1} \left(\nabla \tilde{F}_s(\tilde{\mathbf{w}}_s^{(\alpha)}) \right. \right. \\ & \left. \left. - \frac{1}{\bar{D}_s} \sum_{j \in \mathcal{V}_s} D_j g_j(\mathbf{w}_j^{(\alpha)}) \right) \right\|^2 \\ & + 3\gamma^2 \mathbb{E} \left\| \sum_{\alpha=(r-1)G+b^{(r)}I^{(r)}+1}^{t-1} \left(\frac{1}{\bar{D}_s} \sum_{j \in \mathcal{V}_s} D_j g_j(\mathbf{w}_j^{(\alpha)}) \right. \right. \\ & \left. \left. - \frac{1}{\bar{D}_s^{(\alpha)}} \sum_{j \in \mathcal{V}_s} a_j^{(\alpha)} D_j g_j(\mathbf{w}_j^{(\alpha)}) \right) \right\|^2. \end{aligned} \quad (25)$$

Similar to the proof in (20), (21), and (22), we have

$$\begin{aligned} & 3\gamma^2 \mathbb{E} \left\| \sum_{\alpha=b^{(r)}I^{(r)}}^{t-1} \left(g_k(\mathbf{w}_k^{(\alpha)}) - \nabla \tilde{F}_s(\tilde{\mathbf{w}}_s^{(\alpha)}) \right) \right\|^2 \leq 6\gamma^2 I^{(r)} \\ & \sum_{\alpha=(r-1)G+b^{(r)}I^{(r)}+1}^{(r-1)G+b^{(r)}I^{(r)}+I^{(r)}} \left(L^2 \mathbb{E} \|\tilde{\mathbf{w}}_s^{(\alpha)} - \mathbf{w}_k^{(\alpha)}\|^2 + \epsilon^2 \right), \end{aligned} \quad (26)$$

$$\begin{aligned} & 3\gamma^2 \mathbb{E} \left\| \sum_{\alpha=b^{(r)}I^{(r)}}^{t-1} \left(\nabla \tilde{F}_s(\tilde{\mathbf{w}}_s^{(\alpha)}) - \frac{1}{\bar{D}_s} \sum_{j \in \mathcal{V}_s} D_j g_j(\mathbf{w}_j^{(\alpha)}) \right) \right\|^2 \\ & \leq \frac{6\gamma^2 |\mathcal{V}_s| I^{(r)}}{\bar{D}_s^2} \sum_{\alpha=(r-1)G+b^{(r)}I^{(r)}+1}^{(r-1)G+b^{(r)}I^{(r)}+I^{(r)}} \sum_{j \in \mathcal{V}_s} \\ & D_j^2 \left(L^2 \mathbb{E} \|\tilde{\mathbf{w}}_s^{(\alpha)} - \mathbf{w}_j^{(t)}\|^2 + \epsilon^2 \right), \end{aligned} \quad (27)$$

and

$$\begin{aligned} & 3\gamma^2 \mathbb{E} \left\| \sum_{\alpha=b^{(r)}I^{(r)}}^{t-1} \left(\frac{1}{\bar{D}_s} \sum_{j \in \mathcal{V}_s} D_j g_j(\mathbf{w}_j^{(\alpha)}) \right. \right. \\ & \left. \left. - \frac{1}{\bar{D}_s^{(\alpha)}} \sum_{j \in \mathcal{V}_s} a_j^{(\alpha)} D_j g_j(\mathbf{w}_j^{(\alpha)}) \right) \right\|^2 \\ & \leq \frac{12\gamma^2 \zeta^2 I^{(r)}}{\bar{D}_s} \sum_{\alpha=(r-1)G+b^{(r)}I^{(r)}+1}^{(r-1)G+b^{(r)}I^{(r)}+I^{(r)}} \mathbb{E}(\bar{D}_s - \bar{D}_s^{(\alpha)}). \end{aligned} \quad (28)$$

With the results of (26), (27), and (28), we can obtain

$$\begin{aligned} & \frac{1}{T} \sum_{t=1}^T \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} D_k^2 \mathbb{E} \|\tilde{\mathbf{w}}_s^{(t)} - \mathbf{w}_k^{(t)}\|^2 \\ & = \frac{1}{T} \sum_{r=1}^R \sum_{b^{(r)}=0}^{\frac{G}{I^{(r)}}-1} \sum_{\alpha=(r-1)G+b^{(r)}I^{(r)}+1}^{(r-1)G+b^{(r)}I^{(r)}+I^{(r)}} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} D_k^2 \mathbb{E} \|\tilde{\mathbf{w}}_s^{(t)} - \mathbf{w}_k^{(t)}\|^2 \\ & \leq \frac{1}{T} \sum_{t=1}^T \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \frac{12\gamma^2 \zeta^2 \sum_{j \in \mathcal{V}_s} D_j^2 (\tilde{I}^{(t)})^2 \mathbb{E}(D_k - D_k^{(t)})}{\bar{D}_s} \\ & + \frac{1}{T} \sum_{t=1}^T \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \frac{6\gamma^2 D_k^2 (\bar{D}_s^2 + |\mathcal{V}_s| \sum_{j \in \mathcal{V}_s} D_j^2) (\tilde{I}^{(t)})^2}{\bar{D}_s^2} \\ & \left(L^2 \mathbb{E} \|\tilde{\mathbf{w}}_s^{(t)} - \mathbf{w}_k^{(t)}\|^2 + \epsilon^2 \right), \end{aligned} \quad (29)$$

where $\tilde{I}^{(t)}$ is equivalent to $I^{(r)}$, when $(r-1)G + 1 \leq t \leq rG$.

Since $\tilde{I}^{(t)} \leq G$, suppose $0 < \gamma \leq \sqrt{\frac{\bar{D}_s^2}{6G^2 L^2 (\bar{D}_s^2 + |\mathcal{V}_s| \sum_{k \in \mathcal{V}_s} D_k^2)}}$, $\forall s \in \mathcal{S}$, we can obtain

$$\begin{aligned} & \frac{1}{T} \sum_{t=1}^T \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} D_k^2 \mathbb{E} \|\tilde{\mathbf{w}}_s^{(t)} - \mathbf{w}_k^{(t)}\|^2 \leq \\ & \frac{1}{T} \sum_{t=1}^T \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \frac{12\gamma^2 \zeta^2 \Phi_2 \sum_{j \in \mathcal{V}_s} D_j^2 (\tilde{I}^{(t)})^2 \mathbb{E}(D_k - D_k^{(t)})}{\bar{D}_s} \\ & + \frac{1}{T} \sum_{t=1}^T \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \frac{6\gamma^2 \epsilon^2 D_k^2 \Phi_2 (\bar{D}_s^2 + |\mathcal{V}_s| \sum_{j \in \mathcal{V}_s} D_j^2) (\tilde{I}^{(t)})^2}{\bar{D}_s^2}. \end{aligned} \quad (30)$$

By substituting (24) and (30) into (5), we can complete the proof.

C. Proof of Theorem 3

First of all, by checking the second-order derivation of objective function $\Lambda(I^{(r)})$, we have $\frac{\partial^2 \Lambda(I^{(r)})}{\partial^2 I^{(r)}} > 0$. Then the first-order derivative of objective function $\Lambda(I^{(r)})$ can be denoted as

$$\begin{aligned} & \frac{\partial \Lambda(I^{(r)})}{\partial I^{(r)}} \\ & = \rho \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \left(\left(\frac{D_k \partial \Gamma_{k,s}(I^{(r)})}{\partial I^{(r)}} - \frac{\mu_k \tau_k^{u,r} \Gamma_{k,s}(I^{(r)})}{(I^{(r)})^2} \right) \right. \\ & \left. - \frac{\mu_k}{I^{(r)} D_k} \left(\frac{I^{(r)} \bar{\pi}^{(r)}}{G} - \tau_k^{u,r} - I^{(r)} D_k q_k \right) + \frac{\partial \Upsilon_{k,s}(I^{(r)})}{\partial I^{(r)}} \right). \end{aligned} \quad (31)$$

Since the right hand side of (31) is an increasing function of $I^{(r)}$, if $\frac{\partial \Lambda(I^{(r)})}{\partial I^{(r)}} = 0$ has no solution in $[1, G]$ and $\Lambda(1) > 0$, the optimal $I^{r,*}$ is 1. Besides, if $\frac{\partial \Lambda(I^{(r)})}{\partial I^{(r)}} = 0$ has no solution in $[1, G]$ and $\Lambda(G) < 0$, the optimal $I^{r,*}$ is G . Furthermore, if $\Lambda(I^{(r)}) = 0$ has solutions in the optimal $I^{r,*}$ is the solution of $\frac{\partial \Lambda(I^{(r)})}{\partial I^{(r)}} = 0$. Hence we complete the proof.

D. Proof of Theorem 4

By checking the second-order derivation of objective function (12a), we can see that problem (12) is a convex function, which can be solved by using the Karush-Kuhn-Tucker (KKT) conditions, and the corresponding Lagrange function is

$$\begin{aligned} \mathcal{L}(\mathcal{B}^{(r)}, \nu) = & \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \Gamma_{k,s}(I^{(r)}) D_k \\ & - \frac{\mu_k}{I^{(r)} D_k} \left(\frac{I^{(r)} \bar{\tau}^{(r)}}{G} - \frac{M}{b_k^{(r)} \log_2 \left(1 + \frac{h_k^{(r)} p_k}{N_0} \right)} - I^{(r)} D_k q_k \right) \\ & + \nu \left(\sum_{k \in \mathcal{K}} b_k^{(r)} - B \right), \end{aligned} \quad (32)$$

where $\nu > 0$ is the Lagrange multiplier with constraint (4c). The first order of (32) with respect to $b_k^{(r)}$, $\forall k \in \mathcal{V}_s$, is

$$\begin{aligned} \frac{\partial \mathcal{L}(\mathcal{B}^{(r)}, \nu)}{\partial b_k^{(r)}} = & - \frac{\Gamma_{k,s}(I^{(r)}) \mu_k M}{(b_k^{(r)})^2 I^{(r)} \log_2 \left(1 + \frac{h_k^{(r)} p_k}{N_0} \right)} \\ & - \frac{\mu_k}{I^{(r)} D_k} \left(\frac{I^{(r)} \bar{\tau}^{(r)}}{G} - \frac{M}{b_k^{(r)} \log_2 \left(1 + \frac{h_k^{(r)} p_k}{N_0} \right)} - I^{(r)} D_k q_k \right) + \nu. \end{aligned} \quad (33)$$

By solving (33), we can complete the proof.

E Proof of Theorem 5

By checking the second-order derivation of objective function (13a), we can see that problem (13) is a convex function. The first-order derivative of objective function is denoted as

$$\begin{aligned} \frac{\partial \Lambda(\bar{\tau}^{(r)})}{\partial \bar{\tau}^{(r)}} = & -\rho \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{V}_s} \frac{\Gamma_{k,s}(I^{(r)}) \mu_k}{G} \\ & - \frac{\mu_k}{I^{(r)} D_k} \left(\frac{I^{(r)} \bar{\tau}^{(r)}}{G} - \tau_k^{u,r} - I^{(r)} D_k q_k \right) + (1 - \rho). \end{aligned} \quad (34)$$

Since the right hand side of (34) is an increasing function of $\bar{\tau}^{(r)}$. If $\frac{\partial \Lambda(\bar{\tau}^{(r)})}{\partial \bar{\tau}^{(r)}} = 0$ does not have solution in $(0, +\infty)$, the optimal $\bar{\tau}^{*,r}$ is 0. Otherwise, the optimal $\bar{\tau}^{*,r}$ is the solution of $\frac{\partial \Lambda(\bar{\tau}^{(r)})}{\partial \bar{\tau}^{(r)}} = 0$. Then we complete the proof.

REFERENCES

- [1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [2] K. Lee, M. Lam, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran, "Speeding up distributed machine learning using codes," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1514–1529, Mar. 2018.
- [3] S. Wang *et al.*, "When edge meets learning: Adaptive control for resource-constrained distributed machine learning," in *Proc. IEEE Conf. Comput. Commun.*, Honolulu, HI, USA, 2018, pp. 63–71.
- [4] W. Xia, G. Zheng, Y. Zhu, J. Zhang, J. Wang, and A. P. Petropulu, "A deep learning framework for optimization of MISO downlink beamforming," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1866–1880, Mar. 2020.
- [5] W. Xia, G. Zheng, K.-K. Wong, and H. Zhu, "Model-driven beamforming neural networks," *IEEE Wireless Commun.*, vol. 27, no. 1, pp. 68–75, Feb. 2020.
- [6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Statist.*, Fort Lauderdale, FL, USA, 2017, pp. 1–10.
- [7] G. Gui, M. Liu, F. Tang, N. Kato, and F. Adachi, "6G: Opening new horizons for integration of comfort, security, and intelligence," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 126–132, Oct. 2020.
- [8] S. I. Popoola, R. Ande, B. Adebisi, G. Gui, M. Hammoudeh, and O. Jogunola, "Federated deep learning for zero-day botnet attack detection in IoT edge devices," *IEEE Internet Things J.*, to be published, doi: [10.1109/JIOT.2021.3100755](https://doi.org/10.1109/JIOT.2021.3100755).
- [9] Y. Wang, G. Gui, H. Gacanin, B. Adebisi, H. Sari, and F. Adachi, "Federated learning for automatic modulation classification under class imbalance and varying noise condition," *IEEE Trans. Cogn. Commun. Netw.*, to be published, doi: [10.1109/TCCN.2021.3089738](https://doi.org/10.1109/TCCN.2021.3089738).
- [10] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, Copenhagen, Denmark, 2017, pp. 440–445.
- [11] H. Sun, S. Li, F. R. Yu, Q. Qi, J. Wang, and J. Liao, "Toward communication-efficient federated learning in the Internet of Things with edge computing," *IEEE Internet Things J.*, vol. 7, no. 11, pp. 11053–11067, Nov. 2020.
- [12] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "UVe-QFed: Universal vector quantization for federated learning," *IEEE Trans. Signal Process.*, vol. 69, no. 1, pp. 500–514, Dec. 2020, doi: [10.1109/TSP.2020.3046971](https://doi.org/10.1109/TSP.2020.3046971).
- [13] Y. Du, S. Yang, and K. Huang, "High-dimensional stochastic gradient quantization for communication-efficient edge learning," *IEEE Trans. Signal Process.*, vol. 68, no. 1, pp. 2128–2142, Mar. 2020.
- [14] S. Prakash *et al.*, "Coded computing for low-latency federated learning over wireless edge networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 233–250, Jan. 2020.
- [15] M. M. Amiri and D. Gndz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, vol. 68, no. 1, pp. 2155–2169, Mar. 2020.
- [16] G. Zhu, Y. Du, D. Gndz, 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.
- [17] 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.
- [18] V.-D. Nguyen, S. K. Sharma, T. X. Vu, S. Chatzinotas, and B. Ottersten, "Efficient federated learning algorithm for resource allocation in wireless IoT networks," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2276–2288, Mar. 2021.
- [19] J. Yao and N. Ansari, "Enhancing federated learning in fog-aided IoT by CPU frequency and wireless power control," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3438–3445, Mar. 2021.
- [20] F. Ang, L. Chen, N. Zhao, Y. Chen, W. Wang, and F. R. Yu, "Robust federated learning with noisy communication," *IEEE Trans. Wireless Commun.*, vol. 68, no. 6, pp. 3452–3464, Jun. 2020.
- [21] J. Ren, Y. He, D. Wen, G. Yu, K. Huang, and D. Guo, "Scheduling in cellular federated edge learning with importance and channel awareness," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7690–7703, Nov. 2020.
- [22] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Update aware device scheduling for federated learning at the wireless edge," in *Proc. IEEE Int. Symp. Inf. Theor. Proc.*, Los Angeles, CA, USA, 2020, pp. 2598–2603.
- [23] W. Xia, T. Q. S. Quek, K. Guo, W. Wen, H. H. Yang, and H. Zhu, "Multi-armed bandit based client scheduling for federated learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7108–7123, Nov. 2020.
- [24] N. Yoshida, T. Nishio, M. Morikura, and K. Yamamoto, "MAB-based client selection for federated learning with uncertain resources in mobile networks," in *Proc. IEEE Glob. Commun. Conf. Workshops*, Taipei, Taiwan, 2020, pp. 1–6.
- [25] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *Proc. IEEE Int. Conf. Commun.*, Shanghai, China, 2019, pp. 1–7.
- [26] T. Chen, G. Giannakis, T. Sun, and W. Yin, "LAG: Lazily aggregated gradient for communication-efficient distributed learning," in *Proc. ACM*

- Adv. Neural. Inf. Proces. Syst.*, Montreal, QC, Canada, 2018, pp. 5050–5060.
- [27] 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.
 - [28] Y. He, J. Ren, G. Yu, and J. Yuan, “Importance-aware data selection and resource allocation in federated edge learning system,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13593–13 605, Nov. 2020.
 - [29] W. Shi, S. Zhou, Z. Niu, M. Jiang, and L. Geng, “Joint device scheduling and resource allocation for latency constrained wireless federated learning,” *IEEE Wireless Commun.*, vol. 20, no. 1, pp. 453–467, Jan. 2020.
 - [30] S. Luo, X. Chen, Q. Wu, Z. Zhou, and S. Yu, “HFED: 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.
 - [31] S. R. Pandey et al., “Edge-assisted democratized learning towards federated analytics,” *IEEE Internet Things J.*, vol. 9, no. 1, pp. 572–588, Jan. 2022.
 - [32] S. Hosseinalipour et al., “Multi-stage hybrid federated learning over large-scale d2d-enabled fog networks,” 2020, *arXiv preprint arXiv:2007.09511*.
 - [33] W. Y. B. Lim et al., “Hierarchical incentive mechanism design for federated machine learning in mobile networks,” *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9575–9588, Oct. 2020.
 - [34] L. Liu, J. Zhang, S. H. 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.
 - [35] J. Wang, S. Wang, R.-R. Chen, and M. Ji, “Local averaging helps: Hierarchical federated learning and convergence analysis,” 2020, *arXiv: 2010.12998*.
 - [36] S. Mao, S. Leng, S. Maharjan, and Y. Zhang, “Energy efficiency and delay tradeoff for wireless powered mobile-edge computing systems with multi-access schemes,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1855–1867, Mar. 2020.
 - [37] Z. Ji, L. Chen, N. Zhao, Y. Chen, G. Wei, and F. R. Yu, “Computation offloading for edge-assisted federated learning,” *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9330–9344, Sep. 2021.
 - [38] S. Wang et al., “Adaptive federated learning in resource constrained edge computing systems,” *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, Jun. 2019.
 - [39] D. Liu and O. Simeone, “Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control,” *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 170–185, Jan. 2021.
 - [40] H. Yu, R. Jin, and S. Yang, “On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization,” in *Proc. Int. Conf. Mach. Learn.*, Long Beach, CA, USA, 2019, pp. 12431–12467.
 - [41] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, “Energy-efficient radio resource allocation for federated edge learning,” in *Proc. IEEE Int. Conf. Commun.(ICC)*, 2020, pp. 1–6.
 - [42] T. X. Tran and D. Pompili, “Joint task offloading and resource allocation for multi-server mobile-edge computing networks,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 856–868, Jan. 2019.
 - [43] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” in *Proc. IEEE Proc. IRE.*, vol. 86, no. 11, pp. 2278–2324, 1998.
 - [44] A. Krizhevsky, V. Nair, and G. Hinton, “The CIFAR-10 dataset,” 2014. [Online]. Available: <http://www.cs.toronto.edu/kriz/cifar.html>



NJUPT in 2018.

Bo Xu (Student Member, IEEE) received the B.S. degree in 2018 from the Nanjing University of Posts and Telecommunications (NJUPT), Nanjing, China, where he is currently working toward the Ph.D. degree (successive postgraduate and doctoral programs) in communication and information system. His research interests include mobile edge computing, Big Data, and distributed learning. His current research interests include edge intelligence, edge computing, and federated learning. He was the recipient of the First-Class Scholarship and Special Freshman Scholarship from



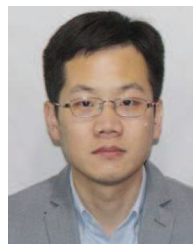
university of Posts and Telecommunications. His current research interests include edge intelligence, edge computing, cloud radio access networks, and massive MIMO. He was the recipient of the Best Paper Award at IEEE GLOBECOM 2016.



the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China. His research interests include green communications, mobile edge computing and caching, and federated learning.



and a master's Supervisor. He performs research on massive MIMO, low-cost ADC, random access, and physical layer security.



Haitao Zhao received the M.S. and the Ph.D. degrees (Hons.) in signal and information processing from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2008 and 2011, respectively. He is currently a Professor with the Nanjing University of Posts and Telecommunications. His current research interests include wireless multimedia modeling, capacity prediction, and wireless network coding.



Hongbo Zhu (Member, IEEE) received the B.S. degree in communications engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, and the Ph.D. degree in information and communications engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 1982 and 1996, respectively. He is currently a Professor with the Nanjing University of Posts and Telecommunications. He is also the Head of the Coordination Innovative Center of IoT Technology and Application (Jiangsu), which is the first governmental authorized Coordination Innovative Center of IoT in China. He is a referee or expert in multiple national organizations and committees. He has authored or coauthored more than 200 technical papers published in various journals and conferences. He is leading a big group and multiple funds on IoT and wireless communications with current focus on architecture and enabling technologies for Internet of Things. His research interests include mobile communications, wireless communication theory, and electromagnetic compatibility.