Federated Learning Over Wireless IoT Networks With Optimized Communication and Resources

Hao Chen, *Student Member, IEEE*, Shaocheng Huang, Deyou Zhang, Ming Xiao, *Senior Member, IEEE*, Mikael Skoglund, *Fellow, IEEE*, and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*—To leverage massive distributed data and computation resources, machine learning in the network edge is considered to be a promising technique, especially for large-scale model training. Federated learning (FL), as a paradigm of collaborative learning techniques, has obtained increasing research attention with the benefits of communication efficiency and improved data privacy. Due to the lossy communication channels and limited communication resources (e.g., bandwidth and power), it is of interest to investigate fast responding and accurate FL schemes over wireless systems. Hence, we investigate the problem of jointly optimized communication efficiency and resources for FL over wireless Internet of Things (IoT) networks. To reduce complexity, we divide the overall optimization problem into two subproblems, i.e., the client scheduling problem and the resource allocation problem. To reduce the communication costs for FL in wireless IoT networks, a new client scheduling policy is proposed by reusing stale local model parameters. To maximize successful information exchange over networks, a Lagrange multiplier method is first leveraged by decoupling variables, including power variables, bandwidth variables, and transmission indicators. Then, a linear-search-based power and bandwidth allocation method is developed. Given appropriate hyperparameters, we show that the proposed communication-efficient FL (CEFL) framework converges at a strong linear rate. Through extensive experiments, it is revealed that the proposed CEFL framework substantially boosts both the communication efficiency and learning performance of both training loss and test accuracy for FL over wireless IoT networks compared to a basic FL approach with uniform resource allocation.**

***Index Terms*—Communication efficiency, federated learning (FL), resource allocation, wireless Internet of Things (IoT) networks.**

Manuscript received 18 October 2021; revised 7 January 2022; accepted 7 February 2022. Date of publication 14 February 2022; date of current version 24 August 2022. This work was supported in part by the ERA-NET Smart Energy Systems SG+ 2017 Program through "SMART-MLA" Project under Grant 89029 (and SWEA number 42811-2), the FORMAS Project titled "Intelligent Energy Management in Smart Community with Distributed Machine Learning" under Grant 2021-00306, and the Swedish Research Council Project titled "Coding for Large-Scale Distributed Machine Learning" under Grant 2021-04772. The work of Deyou Zhang was supported by the Digital Futures Postdoc Fellowships. *(Corresponding authors: Ming Xiao; Deyou Zhang.)*

Hao Chen, Shaocheng Huang, Deyou Zhang, Ming Xiao, and Mikael Skoglund are with the School of Electrical Engineering and Computer Science, Royal Institute of Technology, 10044 Stockholm, Sweden (e-mail: haoch@kth.se; deanhsc@outlook.com; deyou@kth.se; mingx@kth.se; skoglund@kth.se).

H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Digital Object Identifier 10.1109/JIOT.2022.3151193

I. Introduction

With the development of various emerging smart applications (e.g., augmented reality/virtual reality, autonomous driving, and digital twin), the number of Internet of Things (IoT) devices has increased explosively and the massive data generated from these connected IoT devices have led to a surging demand for very high communication rates in future wireless communications, such as the projected sixth-generation (6G) mobile networks. It is envisioned in [1] that by 2025, the active number of IoT devices is expected to be over 75 billion. The massive amounts of data can bring diverse intelligent services due to the recent advances in artificial intelligence (AI) and large-scale machine learning (ML). However, the data originating from massive IoT devices are commonly generated and stored in a distributed manner over wireless networks for a wide range of networked AI applications, e.g., smart grids [2], remote health monitoring [3], Internet of Vehicles [4], etc. Due to the nature of limited wireless communication resources as well as privacy concerns, it is often inefficient or impractical to directly collect all raw data of devices at a central entity (e.g., the cloud). Alternatively, it is increasingly attractive to process data directly in edge clients for data analysis and inference by leveraging edge computing and intelligence with data kept locally.

In the regime of distributed machine learning, federated learning (FL), first coined by Google in 2016 [5], is a paradigm of distributed ML, which pushes the computation of AI applications into edge clients. Therefore, FL decouples the ability of ML from the need to reveal the data to a centralized location, which helps mitigate privacy and latency concerns. During the training process of FL, in which edge clients seek to train a common ML model, each client periodically transmits its locally derived model parameters to a central parameter server (PS). A set of global model parameters are updated in the PS according to aggregation strategies such as federated averaging algorithm (*FedAvg*) [3], and the PS then sends its updated global model parameters to clients for their local model updates. Compared with the traditional data-sharing-based collaborative learning, both communication efficiency and user data privacy are significantly improved in FL [6]. Since the ML parameters are frequently exchanged between the PS and the edge clients over a wireless network, the performance of FL is largely constrained by the properties of wireless communication networks, which can be unstable and may even fluctuate significantly over time because of the limited wireless resources (e.g., bandwidth) and unreliable wireless channels. Thus, this calls for a new design principle for FL from both learning and wireless communication perspectives.

A. Prior Work

Since the proposal of FL [5], there has been an increasing number of studies related to the implementation of FL over wireless networks [7]–[11]. Specifically, Bonawitz *et al.* [7] reported on a system design of FL algorithms in the domain of Android mobile phones and sketched the challenges and corresponding solutions.

Despite the advantages of FL in terms of communication overheads and user data privacy over the traditional data-sharing-based collaborative learning, the implementation of FL over wireless networks still suffers from bottlenecks. More specifically, since multiple communication rounds are required to reach a desired ML accuracy, especially when the number of participating clients is comparably large, the communication costs incurred by unreliable wireless transmission become nonnegligible in wireless FL systems. To reduce communication overhead in distributed ML, various learning algorithms have been proposed in recent years [12]–[20]. Among these efforts, one research direction is to reduce the communication footprint in the uploading phase to make the model training communication efficient. Typical approaches in this direction range from: 1) compressing the uploaded gradients via coding [12], [13], quantization [14], [15], and sparsification [16], [17]; 2) limiting the model sharing by only updating clients with significant training improvement [18], [19]; or 3) accelerating the training process by adopting a momentum method in the sparse update [20]. More specifically, a lazily aggregated gradient approach is proposed in [18] to skip unnecessary uploads, among which communication censoring schemes are developed to avoid less informative local updates so as to reduce the communication burden. It is also worth mentioning that the impact of network resources on the learning performance is not considered in any of those methods.

In addition to communication overhead, another series of work have focused on resource allocation in order to optimize the FL learning performance [21]–[36]. To improve wireless network efficiency, considerable research has been carried out and two main research directions play crucial roles, including admission control and device scheduling [23]–[27], [30]–[32] and resource scheduling management (e.g., spectrum and power) [26]–[29]. In [23], a new FedCS protocol is proposed to schedule as many devices as possible in a limited time frame. Another device scheduling policy is proposed in [24], among which the channel conditions and the significance of the local model updates are jointly considered. Nonetheless, these proposed policies are only evaluated via experiments and the convergence performance has not been theoretically analyzed. To characterize the performance of FL in wireless networks, an analytical model with regard to the FL convergence rate has been developed and the impacts have been evaluated by three different client scheduling policies, i.e., random scheduling, round robin, and proportional fair [25]. By building a connection between the wireless resource allocation and the FL learning performance, Chen *et al.* [26] and [27] proposed to optimize the user selection and power allocation to minimize the FL training loss. Despite of all these results, the existing methods often still involve high overhead both in computation and communication, especially for large-scale ML.

B. Contributions

Motivated by the above observations, we investigate FL with limited wireless resources. We study the problem of jointly optimizing resource and learning performance for reducing communication costs[1] and improving learning performance in wireless FL systems. Different from existing results, in what follows, we will study FL over wireless IoT networks from the aspect of communication efficiency and wireless resource optimization co-design. Particularly, a communication-efficient FL (CEFL) scheme is proposed for wireless FL systems jointly taking communication efficiency and resource optimization into account. The main contributions of our work can be summarized as follows.

1) We aim at communication-efficient FL over wireless IoT networks with limited resources. The joint optimization problem on communication efficiency and resource allocation is first formulated and then decoupled into a client scheduling subproblem and a resource allocation subproblem considering both bandwidth and power constraints.
2) To reduce the communication costs of FL in wireless IoT networks, a communication-efficient client scheduling policy is proposed by limiting communication exchanges and reusing stale local model parameters. To optimize the resource allocation at each communication round of FL training, the Lagrange multiplier method is leveraged to reformulate the resource optimization problem and an optimal solution based on the linear-search (LS) method is then derived.
3) We investigate the convergence and communication properties of the proposed CEFL algorithm both analytically and by simulation. Given a proper hyperparameter, we show that CEFL achieves a strong linear convergence rate and $O(\log [1/\epsilon])$ communication loads, where ϵ is the target accuracy. In addition, the relation between the learning performance and wireless resources, namely, bandwidth and power is theoretically analyzed. Experimental results also indicate that the proposed framework is communication efficient and resource optimized over wireless IoT networks. Our CEFL algorithm outperforms the vanilla FL approach both in communication overhead and training and test performance.

The remainder of this article is structured as follows. Section II describes the system model, while Section III discusses the design of our proposed FL algorithm optimized

[1]The communication cost is generic and can be the total data traffic load of exchanging model parameters among clients and the PS when achieving the target accuracy or iterating certain communication rounds.

TABLE I
SUMMARY OF MAIN NOTATION (AT COMMUNICATION ROUND t)

Notation	Description	Notation	Description
$\mathcal{N}$	Set of clients, $\mathcal{N}=\{1,...,N\}$	B^t	Total available bandwidth resource
$\mathcal{N}_e^t$	Set of scheduled clients	$\boldsymbol{B}^t$	Bandwidth allocation vector
$\mathcal{N}_c^t$	Set of in-active clients	$\boldsymbol{P}^t$	Power allocation vector
$\mathcal{D}_i$	Set of private data in client i	$P_i^{\max}$	Maximum power limit of client i
D_i	Number of training samples in client i	$P_i^{\min}$	Minimum power limit of client i
D	Total number of training data samples, $\sum_{i\in\mathcal{N}} D_i = D$	c_i^t	Transmission rate of client i
$\boldsymbol{w}^t$	Global FL model	$\boldsymbol{a}^t$	Transmission indicating vector
$\widehat{\boldsymbol{w}}_i^t$	Local copy of the local FL model of client i at PS	Γ^t	Time threshold
$\widetilde{\boldsymbol{w}}_i^t$	Local copy of the global FL model at client i	S	Transmitted packet size
$\boldsymbol{\xi}_{i,l}$	Training sample l of client i, $\mathcal{D}_i=\{\boldsymbol{\xi}_{i,1},...,\boldsymbol{\xi}_{i,D_i}\}$	τ_i^t	Communication time of client i

for the underlying wireless IoT network. In Section IV, we characterize the performance of our proposed framework over a wireless channel, which is validated via experiments in Section V. We conclude this article in Section VI and technical proofs are provided in the Appendix.

C. Notation

We adopt the following notation in this article. We denote expectation by $\mathbb{E}[\,\cdot\,]$. $|\cdot|$ is the absolute value. $\|\cdot\|$ denotes the ℓ_2-norm of a vector. $|\mathcal{N}_e^t|$ represents the cardinality of set $\mathcal{N}_e^t$. $\lceil\cdot\rceil$ is the ceiling function. $\nabla f(\cdot)$ denotes the gradient of a function f. In addition, $\langle\cdot,\cdot\rangle$ denotes the inner product in a finite-dimensional Euclidean space. $\boldsymbol{w}^* \in \mathcal{X}$ denotes the optimal solution to (1), where $\mathcal{X}$ is the domain. In addition, we define $G_{\mathcal{X}} \triangleq \sup_{\boldsymbol{w}_x,\boldsymbol{w}_y\in\mathcal{X}} \|\boldsymbol{w}_x-\boldsymbol{w}_y\|$.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we describe the framework of FL over wireless multiclient systems. We will discuss the network model, learning model, communication model, and problem formulation. For ease of illustration, the notations used frequently in this article are summarized in Table I.

A. Network Model

As depicted in Fig. 1, we consider a general one-hop FL-supported wireless IoT network with a base station (BS) and N distributed clients denoted as the set $\mathcal{N}=\{1,\dots,N\}$. In this system, the BS directly connects to the PS, which is equipped with computational resources to provide communication and computation services to the clients. The clients represent IoT sensors gathering data for an FL task, such as mobile devices or organizations, which are communicated with the BS via wireless links. We assume that each client i collects measurement data and owns a fraction of labeled training samples, which is denoted as $\mathcal{D}_i=\{\boldsymbol{\xi}_{i,l}\}_{l=1}^{D_i}$ with $D_i=|\mathcal{D}_i|$ data samples and $\boldsymbol{\xi}_{i,l}$ representing the lth training sample at client i $\forall i\in\mathcal{N}$. The whole data set is thus denoted by $\mathcal{D}=\bigcup_{i\in\mathcal{N}}\mathcal{D}_i$ with the total number of data samples $D=\sum_{i=1}^{N} D_i$. We consider training an ML model of interest over this network (e.g., a classifier), where the PS and clients collaboratively build a shared model parameter for data analysis and inference by exchanging model parameters information while keeping all the data locally.

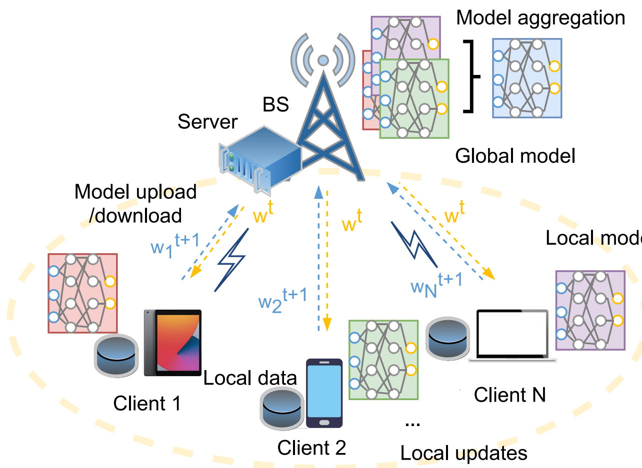


Fig. 1. Example of FL over the wireless IoT network with multiple clients and a BS.

B. Federated Learning Process

In the FL system, a global learning model of interest is trained in a distributed manner among geographically dispersed clients and then aggregated in a central server (i.e., PS). The goal of the training process is to find a model parameter $\boldsymbol{w}\in\mathbb{R}^d$ with the objective of minimizing a loss function $f(\boldsymbol{w})$ on the whole data set $\mathcal{D}$. The global learning objective of the network can be expressed as follows:

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) \triangleq \min_{\boldsymbol{w}} \frac{1}{D}\sum_{i=1}^{N}\sum_{l=1}^{D_i} F_i(\boldsymbol{w},\boldsymbol{\xi}_{i,l}) = \min_{\boldsymbol{w}} \sum_{i=1}^{N}\frac{D_i}{D} f_i(\boldsymbol{w}) \tag{1}$$

where the local loss function $f_i(\boldsymbol{w})$ of client i is defined as $f_i(\boldsymbol{w}) \triangleq (1/D_i)\sum_{l=1}^{D_i} F_i(\boldsymbol{w},\boldsymbol{\xi}_{i,l})$ and $F_i(\boldsymbol{w},\boldsymbol{\xi}_{i,l})$ characterizes the loss of the model parameter $\boldsymbol{w}$ on the training sample $\boldsymbol{\xi}_{i,l}$.

Our analysis is based on the widely used federated averaging (*FedAvg*) algorithm [5]. The whole training process is periodical with an arbitrary number of communication rounds (denoted as T), each of which has E local epochs. Then, the tth communication round is described by the following phases.

1) *Broadcasting Phase:* The PS (located in BS) wirelessly broadcasts the global model parameter $\boldsymbol{w}^t$ to all clients in the tth round.

2) *Local Updating Phase:* After receiving the global model parameter, each client $i \in \mathcal{N}$ trains its local model $\boldsymbol{w}_i^{t+1}$ by applying E epochs of the gradient descent (GD) method, i.e.,

$$\boldsymbol{w}_i^{t+1} = \boldsymbol{w}^t - \eta \nabla f_i\left(\boldsymbol{w}^t\right) \tag{2}$$

where η is the learning rate and $\nabla f_i(\boldsymbol{w}^t)$ is the gradient of local loss function. Then, client i uploads its updated local parameter $\boldsymbol{w}_i^{t+1}$ back to the PS. It is noted that alternative methods, such as stochastic GD (SGD), can also be used for local updates;

3) *Aggregating and Averaging Phase:* For general FL framework, once receiving all the local model parameters, the PS aggregates them and obtains an updated global model by

$$\boldsymbol{w}^{t+1} = \sum_{i=1}^{N} \frac{D_i}{D} \boldsymbol{w}_i^{t+1}. \tag{3}$$

The FL learning process implies that the FL model parameters are iteratively exchanged between the edge clients and the PS over wireless networks.

C. Communication Model

We consider FL over a wireless medium with limited bandwidth and power. After local training, clients upload their local FL models to the BS via frequency-division multiple access (FDMA). Therefore, the achievable rate of client i at the tth communication round is given by

$$c_i^t = B_i^t \log_2\left(1 + \frac{P_i^t \left|h_i^t\right|^2}{B_i^t N_0}\right) \tag{4}$$

where B_i^t and P_i^t are the allocated bandwidth and transmission power of client i, respectively. $|h_i^t|^2$ denotes the corresponding single-carrier block-fading channel gain, and N_0 denotes the noise power spectral density. For simplicity, it is also assumed that for client i, $\boldsymbol{w}_i^{t+1}$ is transmitted as a single packet in the uplink. Denoting by S the packet size of transmitted FL model as [26], the communication time from client i to the BS can be then given by

$$\tau_i^t = \frac{S}{c_i^t}. \tag{5}$$

Since the transmit power of the BS can be generally much higher than that of the client and the whole downlink bandwidth can be utilized to broadcast the global model $\boldsymbol{w}^t$, the latency of downlink transmission is ignored to simplify illustration [37]–[40]. Moreover, to capture the effect of random channel variations on the transmission of each local model parameter $\boldsymbol{w}_i^t$, we consider the current transmission failure if $\tau_i^t > \Gamma^t$ holds in a time duration Γ^t, and the corresponding outage probability is defined as

$$p_i^t = \Pr\left(\tau_i^t > \Gamma^t\right). \tag{6}$$

D. Problem Formulation

To achieve fast learning, the FL training process typically schedules as many clients as possible at each communication round [41]. However, it is undesirable for all clients engaged in learning to transmit their fresh local FL models to the PS especially when the updates are conveyed over a wireless medium with limited resources (e.g., transmit power and network bandwidth). Having more clients scheduled and uploading local models simultaneously can result in large overheads in communication, more unstable connections, and higher latency, which inevitably lead to learning task with less accuracy. To this end, we aim for an optimal solution of joint client scheduling and their associated resource allocation scheme in each communication round to pursue the best learning performance. By denoting the transmission indicating vector as $\boldsymbol{a}^t$, we formulate the following optimization problem with the objective of optimizing both communication and resource for FL over wireless IoT networks:

$$\text{(P-0)} \quad \max_{\boldsymbol{w}^t, \boldsymbol{a}^t, \boldsymbol{B}^t, \boldsymbol{P}^t} \sum_{i \in \mathcal{N}} a_i^t \tag{7}$$

$$\text{s.t. } f\left(\boldsymbol{w}^t\right) - f\left(\boldsymbol{w}^{t+1}\right) \geq \Delta^t \tag{7a}$$

$$P_i^{\min} \leq P_i^t \leq P_i^{\max} \ \forall i \in \mathcal{N} \tag{7b}$$

$$\sum_{i=1}^{N} B_i^t \leq B^t \ \forall i \in \mathcal{N} \tag{7c}$$

$$a_i^t = \begin{cases} 1, & \tau_i^t \leq \Gamma^t \\ 0, & \tau_i^t > \Gamma^t \ \forall i \in \mathcal{N} \end{cases} \tag{7d}$$

where the objective of (P-0) is to maximize the utilization of transmitted FL parameters (i.e., the number of successful transmission) while sustaining the learning performance. Constraint (7a) captures the FL learning convergence constraint by a parameter Δ^t, where $f(\boldsymbol{w}^t)$ is the global loss and $\boldsymbol{w}^t$ can be affected by $\boldsymbol{a}^t$, $\boldsymbol{B}^t$, and $\boldsymbol{P}^t$. Constraints (7b) and (7c) are the feasibility conditions on the power allocation of clients and the bandwidth limits, respectively [42], [43]. Constraint (7d) represents the successful transmission condition. Here, $a_i^t = 1$ represents the successful transmission of the fresh local model $\boldsymbol{w}_i^{t+1}$ from client i; otherwise, we have $a_i^t = 0$.

III. Federated Learning Algorithm Design Over Wireless IoT Networks

(P-0) is a nonconvex optimization problem and intractable due to the nonconvexity of its objective function and constraints (7a) and (7d). To solve problem (P-0), we decompose it into two subproblems, i.e.: 1) determining the client scheduling policy at each communication round and 2) deciding the optimal resource allocation scheme for the clients that have been selected from subproblem 1). We refer to the first subproblem as the *client scheduling* problem and the second subproblem as the *resource allocation* problem.

A. Client Scheduling

With the rapid development of integrated circuits, local computation time can be several orders of magnitude shorter than communication time between the clients and PS [44].

The vanilla FL framework can lead to large communication overheads (e.g., communication time) and it can be inefficient to sequentially update the trained models from all clients before global aggregating and averaging [23]. Accordingly, a subgroup of clients can be actively selected to transmit their local FL models simultaneously. Then, the communication efficiency can be improved and the communication latency can be reduced. To this end, the client scheduling policy plays a crucial role in the FL process especially when wireless resources are limited with. With the goal of reducing communication overheads per communication round, a communication-efficient client selection policy will be developed as follows.

For the *FedAvg* method in (1), in communication round t, after receiving the global model parameters $\boldsymbol{w}^t$ from the PS, every client $i \in \mathcal{N}$ updates its local parameter $\boldsymbol{w}_i^{t+1}$ via (2) for E epochs and is activated to feed updated $\boldsymbol{w}_i^{t+1}$ back to the PS. Instead of requesting fresh local model parameters from all clients, in (2), our client scheduling policy runs as follows. During each communication round t, the client with informative messages (i.e., $\boldsymbol{w}_i^t$) is enabled to upload its current new model parameters if the following selection criterion meets:

$$N^2\eta^2\left\|\nabla f_i\left(\boldsymbol{w}^t\right) - \nabla f_i\left(\widetilde{\boldsymbol{w}}_i^t\right)\right\|^2 \geq \sum_{k=1}^{K}\delta_k\left\|\boldsymbol{w}^{t+1-k} - \boldsymbol{w}^{t-k}\right\|^2 \quad (8)$$

where $\{\delta_k\}_{k=1}^K$ and K are predefined constants, and $\nabla f_i(\boldsymbol{w}^t) - \nabla f_i(\widetilde{\boldsymbol{w}}_i^t)$ is the gradient difference between two evaluations of $\nabla f_i(\boldsymbol{w})$ at current model parameter $\boldsymbol{w}^t$ and the previous round model parameter $\widetilde{\boldsymbol{w}}_i^t$. This condition compares the new local gradient to the stale copy at the client: only when the gradient difference is larger than the recent changes in $\boldsymbol{w}$, the new local model will be transmitted. Otherwise, the PS will reuse the stale copy at the PS. In addition, to avoid clients inactive for a long time, we force it to upload its local model parameters $\boldsymbol{w}_i^t$ to the PS if any client i has not been active for transmitting fresh model parameters during the past T_0 communication rounds. To this regard, we set a clock T_i, $i \in \mathcal{N}$ for each client i, counting the number of inactive communication rounds since last time it uploaded its local models. Thus, it always holds that

$$T_i \leq T_0 \ \forall i \in \mathcal{N}. \quad (9)$$

Once the fresh local model $\boldsymbol{w}_i^{t+1}$ in client i satisfies the above conditions (8) and (9), it will be uploaded to the PS, while the PS in BS will reuse the outdated local model parameters from the rest of clients.

We will prove in the next section that the proposed client scheduling policy-based algorithm can still converge in a linear rate and is communication efficient. Then, on the PS, the current copy of $\boldsymbol{w}_i$ from client i, denoted by $\widehat{\boldsymbol{w}}_i^{t+1}$, is updated as

$$\widehat{\boldsymbol{w}}_i^{t+1} := \begin{cases} \boldsymbol{w}_i^{t+1}, & a_i^t = 1 \\ \widehat{\boldsymbol{w}}_i^t, & a_i^t = 0 \end{cases} \quad (10)$$

where $\widehat{\boldsymbol{w}}_i^t$ is the local model of client i from previous rounds. Here, $a_i^t = 1$ implies that the server receives the fresh local model $\boldsymbol{w}_i^{t+1}$ from client i, otherwise, we have $a_i^t = 0$.

B. Power and Bandwidth Allocation

Once *client scheduling* is determined, the remaining subproblem is the bandwidth and power allocation among these scheduled clients. Given the set of scheduled clients $\mathcal{N}_e^t$ in the tth communication round, *resource allocation* subproblem can be formulated as follows:

$$\text{(P-1)} \quad \max_{\boldsymbol{a}^t, \boldsymbol{B}^t, \boldsymbol{P}^t} \sum_{i \in \mathcal{N}_e^t} a_i^t \quad (11)$$

$$\text{s.t.} \quad P_i^{\min} \leq P_i^t \leq P_i^{\max} \quad \forall i \in \mathcal{N}_e^t \quad (11a)$$

$$\sum_{i=1}^{N} B_i^t \leq B^t \quad \forall i \in \mathcal{N}_e^t \quad (11b)$$

$$a_i^t = \begin{cases} 1, & \tau_i^t \leq \Gamma^t \\ 0, & \tau_i^t > \Gamma^t \quad \forall i \in \mathcal{N}_e^t. \end{cases} \quad (11c)$$

Problem (P-1) is a mixed-integer nonlinear programming (MINLP) problem due to the binary variable $\{a_i^t\}$ and continuous variables $\{B_i^t\}$ and $\{P_i^t\}$. By introducing big-M constant for constraint (11c), problem (P-1) can be equivalently rewritten as

$$\text{(P-2)} \quad \min_{a_i^t \in \{0,1\}, \boldsymbol{B}^t, \boldsymbol{P}^t} \ -\sum_{i \in \mathcal{N}_e^t} a_i^t \quad (12)$$

$$\text{s.t.} \quad \text{(11a), (11b)}$$

$$\tau_i^t \leq \Gamma^t + M\left(1 - a_i^t\right) \quad (12a)$$

$$\tau_i^t \geq \Gamma^t - M a_i^t. \quad (12b)$$

By relaxing each binary variable $a_i^t \in \{0, 1\}$ to a continuous variable $\widetilde{a}_i^t \in [0, 1] \ \forall i \in \mathcal{N}_e^t$, we simplify (P-2) to a nonlinear programming problem, given by

$$\text{(P-3)} \quad \min_{\widetilde{a}_i^t \in [0,1], \boldsymbol{B}^t, \boldsymbol{P}^t} \ -\sum_{i \in \mathcal{N}_e^t} \widetilde{a}_i^t \quad (13)$$

$$\text{s.t.} \quad \text{(11a), (11b)}$$

$$\tau_i^t \leq \Gamma^t + M\left(1 - \widetilde{a}_i^t\right) \quad (13a)$$

$$\tau_i^t \geq \Gamma^t - M\widetilde{a}_i^t. \quad (13b)$$

(P-3) is still nonconvex, and we resort to the Karush–Kuhn–Tucker (KKT) conditions for building the relation between bandwidth and power allocation. More specifically, we first construct the associated Lagrangian function of (P-3) as follows:

$$\begin{aligned} \mathcal{L}_\rho\left(\widetilde{\boldsymbol{a}}^t, \boldsymbol{B}^t, \boldsymbol{P}^t\right) = &-\sum_{i \in \mathcal{N}_e^t} \widetilde{a}_i^t + \sum_{i \in \mathcal{N}_e^t} x_i\left(\widetilde{a}_i^t - 1\right) + \sum_{i \in \mathcal{N}_e^t} y_i\left(-\widetilde{a}_i^t\right) \\ &+ \mu\left(\sum_{i \in \mathcal{N}_e^t} B_i^t - B^t\right) + \sum_{i \in \mathcal{N}_e^t} p_i\left(P_i^{\min} - P_i^t\right) \\ &+ \sum_{i \in \mathcal{N}_e^t} l_i\left(P_i^t - P_i^{\max}\right) \\ &+ \sum_{i \in \mathcal{N}_e^t} \lambda_i\left(\frac{S}{c_i^t} - \Gamma^t + M\left(\widetilde{a}_i^t - 1\right)\right) \\ &+ \sum_{i \in \mathcal{N}_e^t} \upsilon_i\left(\Gamma^t - M\widetilde{a}_i^t - \frac{S}{c_i^t}\right) \end{aligned} \quad (14)$$

where $\{x_i, y_i, p_i, l_i, \lambda_i, \upsilon_i, \mu | i \in \mathcal{N}_e^t\}$ are nonnegative Lagrangian multipliers. The KKT conditions for (P-3) are written as

$$\frac{\partial \mathcal{L}}{\partial \widetilde{a}_i^t} = -1 + x_i - y_i + (\lambda_i - \upsilon_i)M = 0 \tag{15}$$

$$0 \leq \widetilde{a}_i \leq 1 \tag{16}$$

$$x_i(\widetilde{a}_i - 1) = 0 \tag{17}$$

$$y_i(-\widetilde{a}_i) = 0 \tag{18}$$

$$\lambda_i\left(\frac{S}{c_i} - \Gamma^t + M(\widetilde{a}_i - 1)\right) = 0 \tag{19}$$

$$\upsilon_i\left(\Gamma^t - M\widetilde{a}_i - \frac{S}{c_i}\right) = 0 \quad \forall i \in \mathcal{N}_e^t \tag{20}$$

where the solution pair of primal vectors and dual vectors is denoted as $(\widetilde{\boldsymbol{a}}^{\boldsymbol{*}}, \boldsymbol{B}^{\boldsymbol{*}}, \boldsymbol{P}^{\boldsymbol{*}})$ and $(\boldsymbol{x}^{\boldsymbol{*}}, \boldsymbol{y}^{\boldsymbol{*}}, \boldsymbol{p}^{\boldsymbol{*}}, \boldsymbol{l}^{\boldsymbol{*}}, \boldsymbol{\lambda}^{\boldsymbol{*}}, \boldsymbol{\upsilon}^{\boldsymbol{*}}, \mu^*)$. Then, according to (15)–(20), two lemmas can be obtained, as detailed as follows.

Lemma 1: Given the solution $(\widetilde{\boldsymbol{a}}^{\boldsymbol{*}}, \boldsymbol{B}^{\boldsymbol{*}}, \boldsymbol{P}^{\boldsymbol{*}})$ of (P-3), the relation among transmission indicator, bandwidth, and power allocation is given by

$$\frac{S}{c_i^*} = \Gamma^t + M\big(1 - \widetilde{a}_i^*\big) \quad \forall i \in \mathcal{N}_e^t \tag{21}$$

where c_i^* is defined by $c_i^* = B_i^* \log_2(1 + [(P_i^* |h_i^t|^2)/(B_i^* N_0)])$.

Proof: To prove this lemma, we split $\{\widetilde{a}_i^*\}$ into three cases and we will show that (21) is valid in every case. Specifically, Case 1: if $0 < \widetilde{a}_i^* < 1$ holds, it is easy to obtain $x_i^* = y_i^* = 0$, and $(\lambda_i^* - \upsilon_i^*)M = 1$, based on (15)–(20). Thus, $\lambda_i^* > 0, \upsilon_i^* = 0$, and $(S/c_i^*) = \Gamma^t + M(1 - \widetilde{a}_i^*)$ are achieved $\forall i \in \mathcal{N}_e^t$; Case 2: if $a_i^* = 0$ holds, we obtain $x_i^* = 0, (\lambda_i^* - \upsilon_i^*)M = 1 + y_i^*$, followed by $\lambda_i^* > 0, \upsilon_i^* = 0$, and $(S/c_i^*) = \Gamma^t + M \ \forall i \in \mathcal{N}_e^t$; and Case 3: if $a_i^* = 1$ holds, the following are implied as $y_i^* = 0,\ \upsilon_i^* = 0,\ x_i^* + M\lambda_i^* = 1, (S/c_i^*) \leq \Gamma^t$, and $\lambda_i^*([S/c_i^*] - \Gamma^t) = 0$. In this case, though $[S/c_i^*] < \Gamma^t$ with $\lambda_i^* = 0$ satisfies the KKT conditions, such a solution consumes more resources than condition of $[S/c_i^*] = \Gamma^t$ with $\lambda_i^* \geq 0$. Since we aim to achieve a resource optimized FL, we choose $[S/c_i^*] = \Gamma^t$, which proves (21) holds as well. This completes the proof. ■

Remark 1: (P-3) is a direct extension of the MINLP problem (P-1) or (P-2), and the proof of Lemma 1 shows that (21) also holds both in $\widetilde{a}_i^* = 0$ and $\widetilde{a}_i^* = 1$. Thus, the relation among transmission indicator, allocated bandwidth, and transmission power is also applicable for the initial problem (P-1) or (P-2), i.e.,

$$\frac{S}{c_i^*} = \Gamma^t + M\big(1 - a_i^*\big) \quad \forall i \in \mathcal{N}_e^t \tag{22}$$

although it is an MINLP problem. Particularly, for scheduled clients with successful transmission $a_i^* = 1$, we have $S/c_i^* = \Gamma^t$. In addition, with the allocated power and transmission indicator, bandwidth allocation can be obtained directly via (22).

Lemma 2: Given the selected client $i (i \in \mathcal{N}_e^t)$ in communication round t, its communication time τ_i^t is a decreasing function of P_i^t, B_i^t, and $|h_i^t|^2$.

Proof: The first derivative and second derivative of c_i^t with respect to P_i^t can be both proven to be larger than zero. Thus, c_i^t is an increasing and convex function of P_i^t. According to (5), τ_i^t is then a decreasing function of the power P_i^t. The same procedure works for B_i^t and $|h_i^t|^2$, which proves Lemma 2. ■

Algorithm 1 LS Method for Resource Allocation

1: **Input**: $\mathcal{N}_e^t$, B^t, $\{h_i^t, P_i^{\min}, P_i^{\max} | \forall i \in \mathcal{N}_e^t\}$;
2: **Initialize**: $U_{nc}^t = \left\{i | i \in \mathcal{N}_e^t\right\}$, $U_c^t = \emptyset$;
3: construct the channel gain $|h_i^t|^2$ and sort it into a descending order as

$$|h_1^t|^2 \geq |h_2^t|^2 \geq \ldots \geq |h_i^t|^2 \geq \ldots \geq |h_{|\mathcal{N}_e^t|}^t|^2, \forall i \in \mathcal{N}_e^t; \tag{23}$$

4: **while** $\sum_{i \in U_c^t} B_i^t \leq B^t$ and $U_{nc}^t \neq \emptyset$ **do**
5: find client i whose H is maximum in U_{nc}^t as:

$$i = \arg\max_{i \in U_c^t} \{|h_i^t|^2\}; \tag{24}$$

6: set $a_i^t = 1$ and allocate power for client i as $P_i^t = P_i^{\max}$ and compute its required bandwidth B_i^t via (22);
7: assign

$$U_{nc}^t = U_{nc}^t / \{i\}, U_c^t = U_c^t \cup \{i\};$$

8: **end while**
9: **for** client $i \in U_{nc}^t$ **do**
10: set $a_i^t = 0$ and do not allocate network resources;
11: **end for**

Lemma 2 suggests that allocating larger transmission power P_i^t contributes to less communication time τ_i^t, which reduces the outage probability of transmission per communication round. In Algorithm 1, we summarize our proposed resource allocation approach, i.e., the LS algorithm, which is exactly based on these two lemmas. In step 3, we first sort the channel gains of the scheduled clients ($\forall i \in \mathcal{N}_e^t$) in a descending order, which is denoted as $H \triangleq |h_i^t|^2 \ \forall i \in \mathcal{N}_e^t$. Then, for client i with the maximum channel gain in the unallocated set $U_{nc}^t (i \in U_{nc}^t)$, we allocate its required power P_i^t and bandwidth B_i^t before categorizing it into the allocated set U_c^t via steps 4–8. The above steps are repeated until all scheduled clients are considered or all the available bandwidth and power resources are used up. It is noted that according to the proposed LS method, those clients in the unallocated set U_{nc}^t will not be allocated bandwidth or transmission power resources. Then, we give the performance analysis for the proposed LS method.

Theorem 1 (Optimal Solution): The proposed Algorithm 1 can provide an optimal solution for problem (P-1).

Proof: The objective of problem (P-1) is to maximize the number of successful transmissions. Suppose that the solution based on Algorithm 1 can serve K^* clients at most to successfully transmit their local FL models. For the convenience of explanation, based on (23), we assume K^* clients from $U_0^t = \{1, \ldots, K^*\}$ are selected and the transmission indicating vector can be denoted as $\boldsymbol{a}^* = \{\underbrace{1, \ldots, 1, \ldots, 1}_{K^*}, \underbrace{0, \ldots, 0}_{N-K^*}\}$.

With Lemmas 1 and 2, we obtain

$$P_k^t := \begin{cases} P_k^{\max}, & k \le K^* \\ P_i^{\min}, & \text{otherwise} \end{cases} \quad \sum_{k=1}^{K^*} B_k^t \le B^t; \quad \sum_{k=1}^{K^*+1} B_k^t > B^t \tag{25}$$

where $B_1^t \le \cdots \le B_{K^*}^t \le B_{K^*+1}^t \le \cdots \le B_N^t \ \forall k \in \mathcal{N}$. Obviously, to support K^* clients successfully transmit models, the resource allocation based on Algorithm 1 will consume the minimum bandwidth. We assume that there exists another allocation scheme in which (K^*+1) local fresh FL models can be successfully transmitted to PS. Denote the active clients by $U_1^t (U_1^t \subset \mathcal{N}_e^t, |U_1^t| = K^*+1)$, then the following holds:

$$P_l^{\min} \le P_l^t \le P_l^{\max} \quad \forall l \in U_1^t; \quad \sum_{l \in U_1^t, |U_1^t| = K^*+1} B_l^t \le B^t. \tag{26}$$

In (26), one possible solution can be $\{P_l' = P_l^{\max}, B_l' \le B_l^t \ \forall l \in U_1^t\}$ due to Lemma 2, in which $\sum_{l \in U_1^t} B_l' \le \sum_{l \in U_1^t} B_l^t$ holds. Then, we could have

$$B^t \ge \sum_{l \in U_1^t} B_l^t \ge \sum_{l \in U_1^t} B_l' \ge \sum_{k=1}^{K^*+1} B_k^t \tag{27}$$

where it is shown to lead to a contradiction with (25). Thus, the proposed Algorithm 1 can provide an optimal solution for (P-1), which proves Theorem 1. ■

Remark 2: It is worth mentioning that (P-1) can have multiple optimal solutions, while the one provided by Algorithm 1 consumes the least amount of bandwidth compared to the other solutions.

Finally, we summarize the proposed CEFL framework in Algorithm 2 for clarity. For each communication round t, the PS broadcasts the global FL model $\boldsymbol{w}^t$ to all selected clients. Each client trains its local model after receiving $\boldsymbol{w}^t$ and independently decides whether or not to upload its own fresh local model via criteria (8) and (9). Upon receiving the updated models from the scheduled clients, the PS updates the global model with (28). The above steps (i.e., steps 4–19) are repeated until a target test accuracy is achieved or the total number of communication rounds reaches the maximum permitted number T_{Total}. We highlight that Algorithm 2 gives a suboptimal solution to problem (P-0), while an optimal solution of subproblem 2) can be obtained via Algorithm 1 (i.e., the proposed LS method).

IV. Convergence and Communication Analysis

In this section, we will first provide the theoretical analysis on the convergence of the proposed CEFL algorithm. Then, we analyze the communication cost.

A. Convergence Analysis

To facilitate the convergence analysis, following [37]–[39], we assume that the BS-to-client transmission is error-free due to the rich power and bandwidth budget at the BS, and we evaluate the impact of noisy upload transmission on the convergence performance. In addition, local epoch $E = 1$ is considered for all clients to train a global FL model. In the following, before analyzing the convergence of the CEFL algorithm, we first provide the following sufficient conditions,

Algorithm 2 CEFL Over Wireless IoT Networks

1: **Input**: learning rate $\eta > 0$, and constants $\{\delta_k\}$;
2: **Initialize**: $\{\boldsymbol{w}^1, \widehat{\boldsymbol{w}}_i^1, \widetilde{\boldsymbol{w}}_i^1, \nabla f_i(\widetilde{\boldsymbol{w}}_i^1), |\forall i \in \mathcal{N}\}$;
3: **for** $t = 1, 2, \ldots, T_{Total}$ **do**
4: Server broadcasts $\boldsymbol{w}^t$ to all clients;
5: **Client Scheduling Policy:**
6: **for** each client $i \in \mathcal{N}$ **in parallel do**
7: Receive $\boldsymbol{w}^t$ and compute $\nabla f_i(\boldsymbol{w}^t)$;
8: Check condition (8) and (9);
9: **if** condition holds at client i **then**
10: Update $\boldsymbol{w}_i^{t+1}$ based on (2) and upload it;
11: Update $\widetilde{\boldsymbol{w}}_i^{t+1} = \boldsymbol{w}^t$;
12: **else**
13: Update $\widetilde{\boldsymbol{w}}_i^{t+1} = \widetilde{\boldsymbol{w}}_i^t$ and upload nothing;
14: **end if**
15: **end for**
16: **update** the scheduling clients set $\mathcal{N}_e^t$;
17: **Resource Allocation: call Algorithm 1;**
18: Server receives $\boldsymbol{w}_i^{t+1}$ and updates $\widehat{\boldsymbol{w}}_i^{t+1}$ via (10);
19: Server compute $\boldsymbol{w}^{t+1}$ by

$$\boldsymbol{w}^{t+1} := \sum_{i=1}^{N} \frac{D_i}{D} \widehat{\boldsymbol{w}}_i^{t+1}; \tag{28}$$

20: **until** the stopping criterion is satisfied.
21: **end for**

which are widely adopted in the analysis of decentralized optimization.

Assumption 1 (Smoothness): The global loss function $f(\boldsymbol{x})$ is L-smooth, i.e., for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \le L\|\boldsymbol{x} - \boldsymbol{y}\| \tag{29}$$

and this is equivalent to

$$f(\boldsymbol{x}) \le f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2. \tag{30}$$

Assumption 2 (Coercivity): The global loss function $f(\boldsymbol{x})$ is coercive over its feasible set $\mathcal{F}$, i.e., $f(\boldsymbol{x}) \to \infty$ if $\boldsymbol{x} \in \mathcal{F}$ and $\|\boldsymbol{x}\| \to \infty$. The global loss function $f(\boldsymbol{x})$ is lower bounded over $\boldsymbol{x} \in \mathcal{F}$.

Assumption 3 (Strong Convexity): The loss function $f(\boldsymbol{x})$ is μ*-strongly* convex, satisfying that

$$f(\boldsymbol{x}) \ge f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \tag{31}$$

and

$$2\mu\big(f(\boldsymbol{x}) - f(\boldsymbol{x}^*)\big) \le \|\nabla f(\boldsymbol{x})\|^2. \tag{32}$$

With these assumptions, we conclude the convergence properties of the CEFL algorithm as follows.

Lemma 3: Suppose Assumptions 1 and 2 hold. Let $\{\boldsymbol{w}^t\}$ be the iterates generated by the *FedAvg* approach. If the learning rate satisfies $\eta = (1/L)$ and the outage probability is $p_i^t = 0$, the *FedAvg* update per communication round yields the following descent:

$$f\left(\boldsymbol{w}^{t+1}\right) \le f\left(\boldsymbol{w}^t\right) - \Delta_{\text{FedAvg}}^t \tag{33}$$

where $\Delta_{\text{FedAvg}}^t \triangleq (1/2L)\left\|\nabla f(\boldsymbol{w}^t)\right\|^2$.

Proof: The proof of Lemma 3 is similar to that in [18] and we omit it here due to space limitation. ■

Lemma 4: Suppose Assumptions 1 and 2 hold. Let $\{\boldsymbol{w}^t\}$ be the iterates generated by the CEFL approach. If the learning rate satisfies $\eta = (1/L)$ and the outage probability is $p_i^t = 0$, the CEFL update per communication round yields the following descent:

$$f\left(\boldsymbol{w}^{t+1}\right) \leq f\left(\boldsymbol{w}^t\right) - \Delta_{\text{CEFL}}^t \tag{34}$$

where Δ_{CEFL}^t is defined as

$$\Delta_{\text{CEFL}}^t \triangleq \frac{1}{2L}\left\|\nabla f\left(\boldsymbol{w}^t\right)\right\|^2 - \frac{1}{2L}\left\|\sum_{i\in\mathcal{N}_e^t}\frac{D_i}{D}\left[\nabla f_i\left(\widetilde{\boldsymbol{w}}_i^t\right) - \nabla f_i\left(\boldsymbol{w}^t\right)\right]\right\|^2. \tag{35}$$

Proof: The proof of Lemma 4 is similar to that in [18], and we omit it here due to space limitation. ■

Remark 3: With the above lemmas, the rationale of (8) follows next. Similar to the work in [18], the proposed client scheduling policy selects the fresh local models by assessing its contribution to the loss function decrease. To improve the communication efficiency of CEFL, each CEFL upload should bring more descent, i.e.,

$$\frac{\Delta_{\text{CEFL}}^t}{\left|\mathcal{N}_e^t\right|} \geq \frac{\Delta_{\text{FedAvg}}^t}{N}. \tag{36}$$

As stated in [18], $\nabla f(\boldsymbol{w}^t)$ can be approximated by recent gradients or weight differences since $f(\boldsymbol{w}^t)$ is L-smooth. Then, we define

$$\nabla f\left(\boldsymbol{w}^t\right) \approx \frac{1}{\eta^2}\sum_{k=1}^{K}\delta_k\left\|\boldsymbol{w}^{t+1-k} - \boldsymbol{w}^{t-k}\right\|^2 \tag{37}$$

where $\{\delta_k\}_{k=1}^K$ and K are constants. It is also noted that

$$\begin{aligned}&\left\|\sum_{i\in\mathcal{N}_e^t}\frac{D_i}{D}\left[\nabla f_i\left(\widetilde{\boldsymbol{w}}_i^t\right) - \nabla f_i\left(\boldsymbol{w}^t\right)\right]\right\|^2\\ &\quad\leq \left|\mathcal{N}_e^t\right|\sum_{i\in\mathcal{N}_e^t}\left\|\nabla f_i\left(\widetilde{\boldsymbol{w}}_i^t\right) - \nabla f_i\left(\boldsymbol{w}^t\right)\right\|^2.\end{aligned} \tag{38}$$

With (33)–(38), the condition (8) can be easily formed to decide if uploading fresh models.

Then, we conclude the convergence rate of the CEFL algorithm as follows.

Theorem 2 (Convergence Rate): Let Assumptions 1–3 hold and L and μ be defined therein, our proposed CEFL over wireless IoT networks in Algorithm 2 can achieve a strong expected linear rate, i.e.,

$$\mathbb{E}\left[f\left(\boldsymbol{w}^{t+1}\right) - f\left(\boldsymbol{w}^*\right)\right] \leq (1-\rho)\mathbb{E}\left[f\left(\boldsymbol{w}^t\right) - f\left(\boldsymbol{w}^*\right)\right] \tag{39}$$

where ρ is a positive constant satisfying the following condition:

$$\rho \leq \frac{\mu}{L}\sum_{i\in\mathcal{N}_e^t}\frac{D_i\left(1-p_i^t\right)}{D} \tag{40}$$

with p_i^t denoting the outage probability during uploading.

Proof: The proof is detailed in Appendix B. ■

Remark 4: Theorem 2 implies that conditioned on (40), the proposed CEFL still exhibits the same order of convergence rate as that of the original GD method even though some communications are skipped in CEFL. Theorem 2 also presents the impact of wireless factors on the convergence properties i.e., outage probability. Thus, with $p_i^t \to 0$, the proposed CEFL algorithm will converge in a strong convergence rate. Otherwise, the linear convergence rate does not hold any longer. Note that Theorem 2 provides a sufficient condition to guarantee the convergence speed of the proposed CEFL approach.

B. Communication Analysis

Next, we analyze the communication cost based on the linear convergence rate. In following analysis, communication of model parameters between the PS and the client is taken as 1 unit of communication.

Corollary 1: Assume that positive constant ρ meets (40). To realize an expected convergence of $f(\boldsymbol{w})$ under an accuracy threshold ϵ, i.e., $\mathbb{E}[f(\boldsymbol{w}^t) - f(\boldsymbol{w}^*)] \leq \epsilon$, the total number of communication rounds T_{total} for the CEFL algorithm is lower bounded by $T_{\text{total}} \geq \lceil\log_{1-\rho}[\epsilon/(\mathbb{E}[f(\boldsymbol{w}^1)])]\rceil$.

Proof: If we assume $\rho = (\mu/L)\sum_{i\in\mathcal{N}_e^t}[(D_i(1-p_i^t))/D]$, we have

$$\begin{aligned}\mathbb{E}\left[f\left(\boldsymbol{w}^{t+1}\right) - f\left(\boldsymbol{w}^*\right)\right] &\leq (1-\rho)\mathbb{E}\left[f\left(\boldsymbol{w}^t\right) - f\left(\boldsymbol{w}^*\right)\right]\\ &\leq (1-\rho)^2\mathbb{E}\left[f\left(\boldsymbol{w}^{t-1}\right) - f\boldsymbol{w}^*\right)\right]\\ &\quad\cdots\\ &\leq (1-\rho)^t\mathbb{E}\left[f\left(\boldsymbol{w}^1\right) - f\left(\boldsymbol{w}^*\right)\right].\end{aligned} \tag{41}$$

To achieve a predefined deviation defined by $f(\boldsymbol{w}^t) - f(\boldsymbol{w}^*) \leq \epsilon$, it is sufficient to have

$$(1-\rho)^t\mathbb{E}\left[f\left(\boldsymbol{w}^1\right) - f\left(\boldsymbol{w}^*\right)\right] \leq \epsilon. \tag{42}$$

Then, we can further derive

$$\begin{aligned}t &\geq \log_{1-\rho}\frac{\epsilon}{\mathbb{E}\left[f\left(\boldsymbol{w}^1\right) - f\left(\boldsymbol{w}^*\right)\right]}\\ &\geq \log_{1-\rho}\frac{\epsilon}{\mathbb{E}\left[f\left(\boldsymbol{w}^1\right)\right]}.\end{aligned} \tag{43}$$

Since the convergence round must be an integer, we get the result in Corollary 1. ■

Corollary 2 (Communication Cost): Let (40) holds, under the same conditions as Corollary 1, with deviation defined by $f(\boldsymbol{w}^t) - f(\boldsymbol{w}^*) \leq \epsilon$, the communication cost of the proposed CEFL algorithm is $O(\log[1/\epsilon])$.

Proof: With the formula of change of base of logarithms, Corollary 2 can be easily obtained. ■

Remark 5: Compared with N activated clients per communication round in the *FedAvg*-based FL algorithm, only a subset $\mathcal{N}_e^t$ of clients ($|\mathcal{N}_e^t| \leq N$) is active to upload fresh models in the CEFL-based approach. Corollary 2 shows that for the proposed CEFL algorithm, the total communication cost under an accuracy threshold ϵ is reduced to $O(\log[1/\epsilon])$.

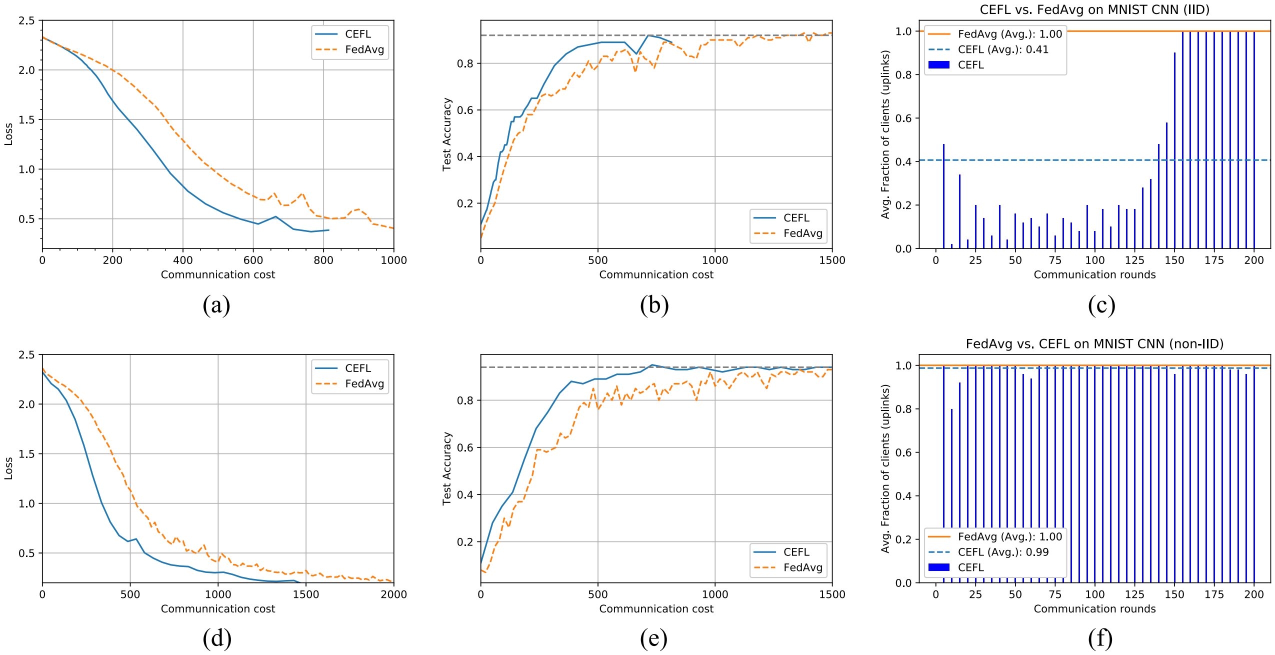


Fig. 2. Training neural network for classification on the MNIST data set: (a) communication overhead comparison (IID); (b) accuracy comparison (IID); (c) percentage of the involved clients (IID); (d) communication overhead comparison (non-IID); (e) accuracy comparison (non-IID); and (f) percentage of the involved clients (non-IID).

V. Simulation Results and Analysis

In this section, we evaluate the performance of the proposed CEFL approach under real data sets.

A. Simulation Setup

For simulations, we consider a typical single-cell wireless IoT network that consists of $N = 10$ edge clients and a BS located at its center similar to the example network shown in Fig. 1. We assume that the BS has two ring-shaped boundary regions. The inner and outer boundaries have radii of 10 and 500 m, respectively. The N clients are uniformly and randomly distributed between the two boundaries, where the distance (in meter) between client i and the BS is denoted as d_i. The wireless channels from each client to the BS follow i.i.d. Rayleigh fading with the total allowed bandwidth $B =$ 20 MHz, and the channel h_i^t is modeled as

$$h_i^t = \sqrt{L(d_i)}o_i^t \tag{44}$$

where $o_i^t \sim \mathcal{CN}(0, \sigma^2)$ is the small-scale fading coefficient of the link between client i and BS, and $L(d_i) = \beta_0(d_i)^{-\alpha}$ is the distance-dependent pathloss with exponent α and coefficient β_0 [33]. β_0 is a frequency-dependent constant, which is set as $(c/[4\pi f_c])^2$ with $c = 3 \times 10^8$ m/s and the carrier frequency $f_c = 3$ GHz. Then, the p_i^t can be formulated as

$$p_i^t = Pr\left(\tau_i > \Gamma^t\right) = 1 - \exp\left(-\frac{Q_i^t}{P_i^t}\right) \tag{45}$$

where $Q_i^t = [(B_i^t N_0)/(L(d_i)\sigma^2)](2^{[S/(B_i^t\Gamma^t)]} - 1)$. Unless specifically stated otherwise, other parameters are given in Table II, following the studies in [21], [27] and [45]. To investigate the performance especially in communication efficiency, we compare our CEFL approach against the vanilla FL approach [5], over wireless networks, among which the *FedAvg* method is adopted. For the target model, we consider a convolutional neural network (CNN) architecture which has two 5×5 convolutional layers (the first with 10 channels, the second with 20 channels, each followed by a ReLU function), each followed by a 2×2 max-pooling layer, a fully connected layer with 500 units and a ReLU function, and a final softmax output layer. The communication cost per link for exchanging model parameters between the PS and the client is 1 unit. Here, 1 unit means a fixed-length information sequence with a reasonable length. Our model was simulated by Tensorflow in Python 3.7 and all experiments were carried out on the environment with the following hardware specifications: CPU Intel Core i5 @2.3 GHz; RAM 16 GB.

TABLE II
Simulation Parameters [46]

Parameter	Value	Parameter	Value
f_c	3 GHz	B	20 MHz
α	2.9	$P^{\max}$	20 dBm
N_0	-174 dBm/Hz	$P^{\min}$	0 dBm

B. Simulation Results

We evaluate the performance of the proposed approach via the MNIST data set for handwritten digits classification [47]. The MNIST data set has 60 000 training images and 10 000 testing images of the ten digits. We adopt the common assumption that each client is connected with an equal amount of training data samples and the local training samples are nonoverlapping with each other [15], [20]. Besides, different data distributions of training samples are considered, both i.i.d. case and non-i.i.d. case. For the i.i.d. case, the original data set is first uniformly partitioned into N pieces and each client is assigned one piece. While for the non-i.i.d. case, the original training data set is first partitioned into N pieces according to the label order, and each piece is then randomly partitioned into two shards (i.e., $2N$ shards in total). Finally, each of N clients is assigned two shards with different label distributions. There are multiple existing works studying how to mitigate the impact of different data distributions on the learning performance of decentralized machine learning [48]–[53]. In this work, since we mainly focus on investigating the optimized communication and resources of the proposed CEFL framework, the challenge of statistical heterogeneity across the clients is important and out of the scope of this article.

The performance of the proposed CEFL approach for classification on the MNIST data set is evaluated in Fig. 2 by training a CNN model over cumulative communication overhead. We first report the results for i.i.d. data partition case. The corresponding experimental results about the training loss, test accuracy, and the utilization of clients are shown in Fig. 2(a)–(c), respectively. It is observed that under the same total communication overhead, the proposed CEFL approach performs better than the vanilla FL method (*FedAvg*). This is because that less informative messages (i.e., local FL models) from clients are restricted to upload to the PS in the CEFL approach while all messages are transmitted to the PS for updating in the *FedAvg* method at each communication round. We use an intuitive explanation as shown in Fig. 2(c) to showcase the effectiveness of CEFL on selectively uploaded local models. In Fig. 2(c), one blue stick refers to the percentage of participated clients to upload local models at each communication round. In the initial communication rounds, communication events happen sparsely, while during the late communication rounds (the 150th–200th), almost all messages (i.e., local FL models) from clients are critical and selected to transmit in our proposed policy. This is due to the fact that the learned global FL model is a generalized representation of all clients. The experimental result also implies that the proposed client scheduling policy works better during the first half of the communication rounds. Similar performance results have also been observed in Fig. 2(d)–(f), where non-i.i.d. case on the MNIST training data set is taken into account. It is worth noting that compared with that of the i.i.d. case, the percentage of participated clients is much higher in the non-i.i.d. case.

Furthermore, we evaluate the impact of the resource allocation scheme on the learning performance in Fig. 3. For performance comparison, we also implement the equal resource allocation approach of [27] as a benchmark. For this allocation scheme, in each communication round t, both the transmission power and its bandwidth of each client are identical, i.e., $B_i^t = B/N, P_i^t = P^{\max}\ \forall i \in \mathcal{N}$. In Fig. 3(a), we first show how the resource allocation scheme affects the convergence behavior of the global FL model training in terms of the value of the training loss. As the communication cost increases, the training losses of the considered algorithms decrease at different rates, while the proposed CEFL framework consisting of the new client scheduling policy and the LS-based resource allocation (denoted as "CEFL-Opt") achieves the lowest loss. Fig. 3(b) also represents that the proposed LS-based CEFL algorithm achieves the highest test accuracy among all schemes under the same communication budget on MNIST. This is reasonable since both client selection and resource allocation are taken into account in the proposed CEFL framework so as to reduce the effect of wireless transmission errors in FL.

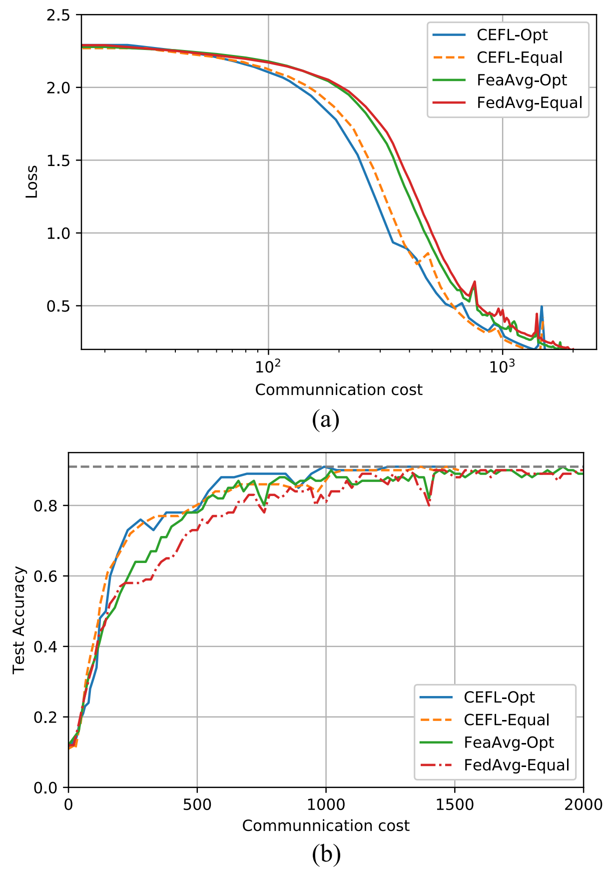


Fig. 3. Impact of the resource allocation scheme on the convergence of the proposed CEFL on the MNIST data set: (a) training loss comparison and (b) test accuracy comparison.

VI. Conclusion

We have studied the joint optimization problem of communication and resources with FL over wireless IoT networks, in which both client selection and resource allocation are considered. A CEFL framework was proposed combined with a new client scheduling policy and an LS-based allocation method, respectively. We showed that the presented LS approach was able to provide an optimal solution for bandwidth and power allocation and the convergence and communication properties of the proposed CEFL algorithm were also theoretically analyzed. Extensive experimental results revealed that the proposed CEFL algorithm outperforms the state-of-the-art baseline method both in communication overheads and learning performance under different data distributions. Besides, the proposed CEFL framework can effectively schedule clients according to both the learned model parameter characteristics and wireless channel dynamics.

APPENDIX A
SUPPORTING LEMMA 5

Before proving Theorem 1, we first introduce the following Lemma 5.

Lemma 5: Suppose that the iterates $\{\boldsymbol{w}^t\}$ of problem (1) are generated by full GD over wireless IoT networks: $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \boldsymbol{g}^t$ with $\boldsymbol{g}^t \triangleq \nabla f(\boldsymbol{w}^t) + \boldsymbol{e}^t$ and learning rate $\eta = (1/L)$ and the error $\boldsymbol{e}^t$ meets

$$0 \le \mathbb{E}\left[\left\|\boldsymbol{e}^t\right\|^2\right] \le 2L\left(\frac{\mu}{L} - \rho\right)\mathbb{E}\left[f\left(\boldsymbol{w}^t\right) - f\left(\boldsymbol{w}^*\right)\right] \tag{46}$$

where $\rho \le (\mu/L)$ is a positive constant. Under Assumptions 1–3, the following inequality holds

$$\mathbb{E}\left[f\left(\boldsymbol{w}^{t+1}\right) - f\left(\boldsymbol{w}^*\right)\right] \le (1-\rho)\mathbb{E}\left[f\left(\boldsymbol{w}^t\right) - f\left(\boldsymbol{w}^*\right)\right]. \tag{47}$$

Proof: Using Assumption 1, we can derive

$$\begin{aligned} f\left(\boldsymbol{w}^{t+1}\right) &\le f\left(\boldsymbol{w}^t\right) + \left\langle \boldsymbol{w}^{t+1} - \boldsymbol{w}^t, \nabla f\left(\boldsymbol{w}^t\right)\right\rangle + \frac{L}{2}\left\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\right\|^2 \\ &= f\left(\boldsymbol{w}^t\right) - \left\langle \eta\left(\nabla f\left(\boldsymbol{w}^t\right) + \boldsymbol{e}^t\right), \nabla f\left(\boldsymbol{w}^t\right)\right\rangle \\ &\quad + \frac{L}{2}\left\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\right\|^2 \\ &= f\left(\boldsymbol{w}^t\right) - \left\langle \frac{1}{L}\left(\nabla f\left(\boldsymbol{w}^t\right) + \boldsymbol{e}^t\right), \nabla f\left(\boldsymbol{w}^t\right)\right\rangle \\ &\quad + \frac{L}{2}\left\|\frac{1}{L}\left(\nabla f\left(\boldsymbol{w}^t\right) + \boldsymbol{e}^t\right)\right\|^2 \\ &= f\left(\boldsymbol{w}^t\right) - \frac{1}{2L}\left\|\nabla f\left(\boldsymbol{w}^t\right)\right\|^2 + \frac{1}{2L}\left\|\boldsymbol{e}^t\right\|^2. \end{aligned} \tag{48}$$

Subtracting $f(\boldsymbol{w}^*)$ from both sides of (48), it gives

$$\begin{aligned} f\left(\boldsymbol{w}^{t+1}\right) - f\left(\boldsymbol{w}^*\right) &\le f\left(\boldsymbol{w}^t\right) - f\left(\boldsymbol{w}^*\right) - \frac{1}{2L}\left\|\nabla f\left(\boldsymbol{w}^t\right)\right\|^2 + \frac{1}{2L}\left\|\boldsymbol{e}^t\right\|^2 \\ &\overset{(a)}{\le} f\left(\boldsymbol{w}^t\right) - f\left(\boldsymbol{w}^*\right) - \frac{2\mu}{2L}\left[f\left(\boldsymbol{w}^t\right) - f\left(\boldsymbol{w}^*\right)\right] \\ &\quad + \frac{1}{2L}\left\|e^t\right\|^2 \\ &= \left(1 - \frac{\mu}{L}\right)\left[f\left(\boldsymbol{w}^t\right) - f\left(\boldsymbol{w}^*\right)\right] + \frac{1}{2L}\left\|\boldsymbol{e}^t\right\|^2 \end{aligned} \tag{49}$$

where (a) holds due to (32). Taking expectations of (49), we can derive

$$\begin{aligned} \mathbb{E}\left[\left(\boldsymbol{w}^{t+1}\right) - f\left(\boldsymbol{w}^*\right)\right] &\le \left(1 - \frac{\mu}{L}\right)\mathbb{E}\left[f\left(\boldsymbol{w}^t\right) - f\left(\boldsymbol{w}^*\right)\right] \\ &\quad + \frac{1}{2L}\mathbb{E}\left[\left\|e^t\right\|^2\right] \\ &\overset{(b)}{\le} \left(1 - \frac{\mu}{L} + \frac{\mu}{L} - \rho\right)\mathbb{E}\left[f\left(\boldsymbol{w}^t\right) - f\left(\boldsymbol{w}^*\right)\right] \\ &= (1-\rho)\mathbb{E}\left[f\left(\boldsymbol{w}^t\right) - f\left(\boldsymbol{w}^*\right)\right] \end{aligned} \tag{50}$$

where (b) uses the condition (46). This completes the proof of Lemma 5. ■

APPENDIX B
PROOF OF THEOREM 2

Proof: Based on the result given in Lemma 5, we present the proof for Theorem 1 in detail. Combining (2) with (28), we have

$$\boldsymbol{w}^{t+1} - \boldsymbol{w}^t = -\eta\left(\nabla f\left(\boldsymbol{w}^t\right) + \boldsymbol{e}^t\right) \tag{51}$$

where $\boldsymbol{e}^t$ is gradient deviation caused by the PS at communication round t that it uses an old copy of the local FL model from client $i \in \mathcal{N}$ when the newly local FL model cannot be successfully received. Let $\mathcal{N}_e^t$ and $\mathcal{N}_c^t$ be the sets of clients that *do* and *do not* communicate with the PS, respectively. In particular, $\boldsymbol{e}^t$ can be expressed as (52), shown at the bottom of the page, where (c) holds because of $\nabla f(\boldsymbol{w}^t) = (\sum_{i\in\mathcal{N}_e^t} D_i \nabla f_i(\boldsymbol{w}^t) + \sum_{i\in\mathcal{N}_c^t} D_i \nabla f_i(\boldsymbol{w}^t))/D$. Thus, with $\eta^t = (1/L)$, we have

$$\begin{aligned} \boldsymbol{w}^{t+1} - \boldsymbol{w}^t = -\frac{1}{L}\Bigg(&\nabla f\left(\boldsymbol{w}^t\right) + \sum_{i\in\mathcal{N}_e^t}\frac{D_i\left(1-a_i^t\right)}{D}\left(\nabla f_i\left(\widetilde{\boldsymbol{w}}_i^t\right) - \nabla f_i\left(\boldsymbol{w}^t\right)\right) \\ &+ \sum_{i\in\mathcal{N}_c^t}\frac{D_i}{D}\left(\nabla f_i\left(\widetilde{\boldsymbol{w}}_i^t\right) - \nabla f_i\left(\boldsymbol{w}^t\right)\right)\Bigg). \end{aligned} \tag{53}$$

Then, by taking expectations and norms in both sides of (52), we have

$$\begin{aligned} \mathbb{E}\left[\left\|\boldsymbol{e}^t\right\|^2\right] &= \mathbb{E}\Bigg[\Bigg\|\sum_{i\in\mathcal{N}_e^t}\frac{D_i\left(1-a_i^t\right)}{D}\left(\nabla f_i\left(\widehat{\boldsymbol{w}}_i^t\right) - \nabla f_i\left(\boldsymbol{w}^t\right)\right) \\ &\qquad + \sum_{i\in\mathcal{N}_c^t}\frac{D_i}{D}\left(\nabla f_i\left(\widehat{\boldsymbol{w}}_i^t\right) - \nabla f_i\left(\boldsymbol{w}^t\right)\right)\Bigg\|^2\Bigg] \\ &\le \mathbb{E}\Bigg[\Bigg\|\sum_{i\in\mathcal{N}_e^t}\frac{D_i\left(1-a_i^t\right)}{D}\left\|\nabla f_i\left(\widehat{\boldsymbol{w}}_i^t\right) - \nabla f_i\left(\boldsymbol{w}^t\right)\right\| \\ &\qquad + \sum_{i\in\mathcal{N}_c^t}\frac{D_i}{D}\left\|\nabla f_i\left(\widehat{\boldsymbol{w}}_i^t\right) - \nabla f_i\left(\boldsymbol{w}^t\right)\right\|\Bigg\|^2\Bigg] \\ &\overset{(e)}{\le} \mathbb{E}\Bigg[\Bigg\|\sum_{i\in\mathcal{N}_e^t}\frac{D_i\left(1-a_i^t\right)LG_{\mathcal{X}}}{D} + \sum_{i\in\mathcal{N}_c^t}\frac{D_iLG_{\mathcal{X}}}{D}\Bigg\|^2\Bigg] \\ &= \mathbb{E}\Bigg[\Bigg\|LG_{\mathcal{X}}\Bigg(\sum_{i\in\mathcal{N}}\frac{D_i}{D} - \sum_{i\in\mathcal{N}_e^t}\frac{D_ia_i^t}{D}\Bigg)\Bigg\|^2\Bigg] \end{aligned}$$

$$\begin{aligned} e^t &= -\nabla f\left(\boldsymbol{w}^t\right) + \frac{\sum_{i\in\mathcal{N}_e^t} D_i a_i^t \nabla f_i\left(\boldsymbol{w}^t\right) + \sum_{i\in\mathcal{N}_e^t} D_i\left(1-a_i^t\right)\nabla f_i\left(\widetilde{\boldsymbol{w}}_i^t\right) + \sum_{i\in\mathcal{N}_c^t} D_i \nabla f_i\left(\widetilde{\boldsymbol{w}}_i^t\right)}{D} \\ &\overset{(c)}{=} \frac{-\sum_{i\in\mathcal{N}_e^t} D_i\left(1-a_i^t\right)\nabla f_i\left(\boldsymbol{w}^t\right) + \sum_{i\in\mathcal{N}_e^t} D_i\left(1-a_i^t\right)\nabla f_i\left(\widetilde{\boldsymbol{w}}_i^t\right) + \sum_{i\in\mathcal{N}_c^t} D_i\nabla f_i\left(\widetilde{\boldsymbol{w}}_i^t\right) - \sum_{i\in\mathcal{N}_c^t} D_i \nabla f_i\left(\boldsymbol{w}^t\right)}{D} \\ &= \sum_{i\in\mathcal{N}_e^t}\frac{D_i\left(1-a_i^t\right)}{D}\left(\nabla f_i\left(\widetilde{\boldsymbol{w}}_i^t\right) - \nabla f_i\left(\boldsymbol{w}^t\right)\right) + \sum_{i\in\mathcal{N}_c^t}\frac{D_i}{D}\left(\nabla f_i\left(\widetilde{\boldsymbol{w}}_i^t\right) - \nabla f_i\left(\boldsymbol{w}^t\right)\right) \end{aligned} \tag{52}$$

$$\overset{(f)}{\leq} L^2G_{\mathcal{X}}^2\left(1-\sum_{i\in\mathcal{N}_e^t}\frac{D_i\mathbb{E}\left[a_i^t\right]}{D}\right)$$
$$\overset{(g)}{=} L^2G_{\mathcal{X}}^2\left(1-\sum_{i\in\mathcal{N}_e^t}\frac{D_i\left(1-p_i^t\right)}{D}\right) \tag{54}$$

where (e) is due to Assumption 1, (f) is because of $\sum_{i\in\mathcal{N}} D_i = D$, and (g) is based on $\mathbb{E}[a_i^t] = 1-p_i^t$, respectively. According to Lemma 5, if let (46) always holds, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{e}^t\right\|^2\right] &\leq L^2G_{\mathcal{X}}^2\left(1-\sum_{i\in\mathcal{N}_e^t}\frac{D_i\left(1-p_i^t\right)}{D}\right)\\ &\leq 2L\left(\frac{\mu}{L}-\rho\right)\mathbb{E}\left[f\left(\boldsymbol{w}^t\right)-f\left(\boldsymbol{w}^*\right)\right].\end{aligned} \tag{55}$$

That is, to guarantee (47), the available ρ must satisfy

$$\rho \leq \frac{\mu}{L} - \frac{LG_{\mathcal{X}}^2\left(1-\sum_{i\in\mathcal{N}_e^t}\frac{D_i(1-p_i^t)}{D}\right)}{2\mathbb{E}\left[f(\boldsymbol{w}^t)-f(\boldsymbol{w}^*)\right]}. \tag{56}$$

According to (29) in Assumption 1 and (32) in Assumption 3, we have

$$\mathbb{E}\left[f\left(\boldsymbol{w}^t\right)-f\left(\boldsymbol{w}^*\right)\right] \leq \frac{\left\|\nabla f\left(\boldsymbol{w}^t\right)\right\|^2}{2\mu} \leq \frac{L^2G_{\mathcal{X}}^2}{2\mu}. \tag{57}$$

Then, (56) can be derived as

$$\begin{aligned}\rho &\leq \frac{\mu}{L} - \frac{\mu\left(1-\sum_{i\in\mathcal{N}_e^t}\frac{D_i(1-p_i^t)}{D}\right)}{L}\\ &= \frac{\mu}{L}\sum_{i\in\mathcal{N}_e^t}\frac{D_i\left(1-p_i^t\right)}{D}.\end{aligned} \tag{58}$$

This completes the proof of Theorem 2. ■

REFERENCES

[1] "Internet of Things (IoT) Connected Devices Installed Base Worldwide From 2015 to 2025." 2016. [Online]. Available: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/

[2] S. Ciavarella, J.-Y. Joo, and S. Silvestri, "Managing contingencies in smart grids via the Internet of Things," *IEEE Trans. Smart Grid*, vol. 7, no. 4, pp. 2134–2141, Jul. 2016.

[3] R. K. Pathinarupothi, P. Durga, and E. S. Rangan, "IoT-based smart edge for global health: Remote monitoring with severity detection and alerts transmission," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2449–2462, Apr. 2019.

[4] B. Ji *et al.*, "Survey on the Internet of Vehicles: Network architectures and applications," *IEEE Commun. Stand. Mag.*, vol. 4, no. 1, pp. 34–41, Mar. 2020.

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

[6] K. Wei *et al.*, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3454–3469, 2020.

[7] K. Bonawitz *et al.*, "Towards federated learning at scale: System design," in *Proc. Conf. Mach. Learn. Syst.*, 2019, pp. 374–388.

[8] 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.

[9] X. Wang, C. Wang, X. Li, V. C. M. Leung, and T. Taleb, "Federated deep reinforcement learning for Internet of Things with decentralized cooperative edge caching," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9441–9455, Oct. 2020.

[10] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang, and M. Guizani, "Reliable federated learning for mobile networks," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 72–80, Apr. 2020.

[11] Y. Liu, X. Yuan, Z. Xiong, J. Kang, X. Wang, and D. Niyato, "Federated learning for 6G communications: Challenges, methods, and future directions," *China Commun.*, vol. 17, no. 9, pp. 105–118, Sep. 2020.

[12] 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. 2021.

[13] H. Chen, Y. Ye, M. Xiao, M. Skoglund, and H. V. Poor, "Coded stochastic ADMM for decentralized consensus optimization with edge computing," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5360–5373, Apr. 2021.

[14] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization," in *Proc. 23th Int. Conf. Artif. Intell. Stat.*, 2020, pp. 2021–2031.

[15] J. Mills, J. Hu, and G. Min, "Communication-efficient federated learning for wireless edge intelligence in IoT," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 5986–5994, Jul. 2020.

[16] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, May 2020.

[17] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-i.i.d. data," *IEEE Trans. Neural Netw. Learn. Syst*, vol. 31, no. 9, pp. 3400–3413, Sep. 2020.

[18] T. Chen, G. B. Giannakis, T. Sun, and W. Yin, "LAG: Lazily aggregated gradient for communication-efficient distributed learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5050–5060.

[19] H. S. Ghadikolaei, S. Stich, and M. Jaggi, "LENA: Communication-efficient distributed learning with self-triggered gradient uploads," in *Proc. IEEE 24th Int. Conf. Artif. Intell. Stat.*, vol. 130, Apr. 2021, pp. 3943–3951.

[20] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2018, pp. 1–14.

[21] 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.

[22] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, "Energy-efficient radio resource allocation for federated edge learning," in *Proc. IEEE 54th Int. Conf. Commun. (ICC)*, Dublin, Ireland, 2020, pp. 1–6.

[23] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *Proc. IEEE 53rd Int. Conf. Commun.*, Shanghai, China, 2019, pp. 1–7.

[24] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Update aware device scheduling for federated learning at the wireless edge," in *Proc. IEEE Int. Symp. Inf. Theory*, Los Angeles, CA, USA, 2020, pp. 2598–2603.

[25] 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.

[26] 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.

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

[28] W. Y. B. Lim, J. S. Ng, Z. Xiong, D. Niyato, C. Miao, and D. I. Kim, "Dynamic edge association and resource allocation in self-organizing hierarchical federated learning networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3640–3653, Dec. 2021.

[29] W. Y. B. Lim *et al.*, "Decentralized edge intelligence: A dynamic resource allocation framework for hierarchical federated learning," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 3, pp. 536–550, Mar. 2022.

[30] 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.

[31] X. Lyu, C. Ren, W. Ni, H. Tian, R. P. Liu, and E. Dutkiewicz, "Optimal online data partitioning for geo-distributed machine learning in edge of wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2393–2406, Oct. 2019.

[32] M. M. Amiri, S. R. Kulkarni, and H. V. Poor, "Federated learning with downlink device selection," 2021, *arXiv:2107.03510*.

[33] 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. 5, pp. 3394–3409, Mar. 2021.

[34] J. Kang, Z. Xiong, D. Niyato, S. Xie, and J. Zhang, "Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10700–10714, Dec. 2019.

[35] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication-efficient federated learning," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 118, no. 17, 2021, Art. no. e2024789118.

[36] K. Wei *et al.*, "Low-latency federated learning over wireless channels with differential privacy," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 290–307, Jan. 2022.

[37] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE Conf. Comput. Commun. (IEEE INFOCOM)*, Paris, France, 2019, pp. 1387–1395.

[38] C. T. Dinh *et al.*, "Federated learning over wireless networks: Convergence analysis and resource allocation," *IEEE/ACM Trans. Netw.*, vol. 29, no. 1, pp. 398–409, Feb. 2021.

[39] 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.

[40] S. Mumtaz, H. Lundqvist, K. M. S. Huq, J. Rodriguez, and A. Radwan, "Smart direct-LTE communication: An energy saving perspective," *Ad Hoc Netw.*, vol. 13, pp. 296–311, Feb. 2014.

[41] 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.

[42] M. Ali, S. Qaisar, M. Naeem, and S. Mumtaz, "Energy efficient resource allocation in D2D-assisted heterogeneous networks with relays," *IEEE Access*, vol. 4, pp. 4902–4911, 2016.

[43] S. Goudarzi, N. Kama, M. H. Anisi, S. Zeadally, and S. Mumtaz, "Data collection using unmanned aerial vehicles for Internet of Things platforms," *Comput. Electr. Eng.*, vol. 75, pp. 1–15, May 2019.

[44] G. Lan, S. Lee, and Y. Zhou, "Communication-efficient algorithms for decentralized and stochastic optimization," *Math. Program.*, vol. 180, no. 1, pp. 237–284, 2020.

[45] S. Wang *et al.*, "Federated learning for task and resource allocation in wireless high-altitude balloon networks," *IEEE Internet Things J.*, vol. 8, no. 24, pp. 17460–17475, Dec. 2021.

[46] L. Wang, K.-K. Wong, R. W. Heath, and J. Yuan, "Wireless powered dense cellular networks: How many small cells do we need?" *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2010–2024, Sep. 2017.

[47] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[48] K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons, "The non-IID data quagmire of decentralized machine learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4387–4398.

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

[50] V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar, "Federated multi-task learning," 2017, *arXiv:1705.10467*.

[51] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Int. Conf. Mach. Learn. Syst.*, vol. 2, 2020, pp. 429–450.

[52] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-IID features via local batch normalization," 2021, *arXiv:2102.07623*.

[53] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," 2020, *arXiv:2002.06440*.

Hao Chen (Student Member, IEEE) received the B.Sc. degree in communication engineering and the M.Sc. degree in electronic and communication engineering from Soochow University, Suzhou, China, in 2014 and 2017, respectively. He is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden.

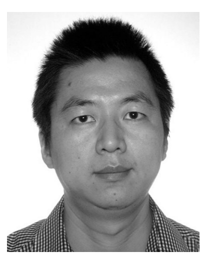
His current research interests include distributed machine learning, distributed optimization, and edge computing.

Shaocheng Huang received the Ph.D. degree from the Division of Information Science and Engineering, KTH Royal Institute of Technology, Stockholm, Sweden, in December 2020.

His research interests include massive MIMO, millimeter-wave communications, machine learning, and blind signal processing.

Deyou Zhang received the B.S. and M.S. degrees from Harbin Institute of Technology, Harbin, China, in 2012 and 2014, respectively, and the Ph.D. degree from the University of Sydney, Sydney, NSW, Australia, in 2020.

He is currently working as a Postdoctoral Fellow with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. His research interests include millimeter-wave communications, wireless positioning, and federated learning.



Ming Xiao (Senior Member, IEEE) received the bachelor's and master's degrees in engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 1997 and 2002, respectively, and the Ph.D. degree from the Chalmers University of Technology, Gothenburg, Sweden, in November 2007.

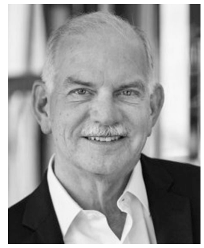
From 1997 to 1999, he worked as a Network and Software Engineer with ChinaTelecom, Beijing, China. From 2000 to 2002, he also held a position with the SiChuan Communications Administration, Chengdu, China. Since 2007, he has been with the Department of Information Science and Engineering, School of Electrical Engineering and Computer Science, Royal Institute of Technology, Stockholm, Sweden, where he is currently an Associate Professor.

Dr. Xiao was an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS from 2012 to 2017 and IEEE WIRELESS COMMUNICATIONS LETTERS from 2012 to 2016. He has been a Senior Editor for IEEE COMMUNICATIONS LETTERS since January 2015, an Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS since 2018, and an Area Editor for IEEE OPEN JOURNAL OF THE COMMUNICATION SOCIETY since 2019. He was the Lead Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Special Issue on Millimeter Wave Communications for Future Mobile Networks in 2017.

Mikael Skoglund (Fellow, IEEE) received the Ph.D. degree from the Chalmers University of Technology, Gothenburg, Sweden, in 1997.

In 1997, he joined the Royal Institute of Technology (KTH), Stockholm, Sweden, where he was appointed to the Chair of Communication Theory in 2003. At KTH, he heads the Division of Information Science and Engineering and the Department of Intelligent Systems. He has authored and coauthored more than 185 journal and 400 conference papers. He has worked on problems in source-channel coding, coding and transmission for wireless communications, Shannon theory, information and control, and statistical signal processing.

Dr. Skoglund has served on numerous technical program committees for IEEE sponsored conferences, he was the General Co-Chair for IEEE ITW 2019, and he is serving as the TPC Co-Chair for IEEE ISIT in 2022. From 2003 to 2008, he was an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS. From 2008 to 2012, he was on the Editorial Board for the IEEE TRANSACTIONS ON INFORMATION THEORY and starting in the Fall of 2021, he joined it once more.



H. Vincent Poor (Life Fellow, IEEE) received the Ph.D. degree in EECS from Princeton University, Princeton, NJ, USA, in 1977.

From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign, Champaign, IL, USA. Since 1990, he has been on the faculty, where he is currently the Michael Henry Strater University Professor, Princeton University. From 2006 to 2016, he served as the Dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory, machine learning and network science, and their applications in wireless networks, and energy systems and related fields. Among his publications in these areas is the forthcoming book *Machine Learning and Wireless Communications* (Cambridge University Press).

Dr. Poor received the IEEE Alexander Graham Bell Medal in 2017. He is a member of the National Academy of Engineering and the National Academy of Sciences and is a Foreign Member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies.