Scheduling for Cellular Federated Edge Learning with Importance and Channel Awareness

Jinke Ren, Yinghui He, Dingzhu Wen, Guanding Yu,

Kaibin Huang, and Dongning Guo

Abstract

In cellular federated edge learning (FEEL), multiple edge devices holding local data jointly train a neural network by communicating learning updates with an access point without exchanging their data samples. With very limited communication resources, it is beneficial to schedule the most informative local learning updates. In this paper, a novel scheduling policy is proposed to exploit both diversity in multiuser channels and diversity in the "importance" of the edge devices' learning updates. First, a new probabilistic scheduling framework is developed to yield unbiased update aggregation in FEEL. The importance of a local learning update is measured by its gradient divergence. If one edge device is scheduled in each communication round, the scheduling policy is derived in closed form to achieve the optimal trade-off between channel quality and update importance. The probabilistic scheduling framework is then extended to allow scheduling multiple edge devices in each communication round. Numerical results obtained using popular models and learning datasets demonstrate that the proposed scheduling policy can achieve faster model convergence and higher learning accuracy than conventional scheduling policies that only exploit a single type of diversity.

Index Terms

Federated edge learning, scheduling, multiuser diversity, resource management, convergence analysis.

J. Ren is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China, and also with the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208 USA (e-mail: renjinke@zju.edu.cn).

Y. He and G. Yu are with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: 2014hyh@zju.edu.cn, yuguanding@zju.edu.cn).

D. Wen and K. Huang are with the Department of Electrical and Electronic Engineering at The University of Hong Kong, Hong Kong (e-mail: dzwen@eee.hku.hk, huangkb@eee.hku.hk).

D. Guo is with the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208 USA (e-mail: dGuo@northwestern.edu).

I. INTRODUCTION

As driven by the phenomenal growth in global mobile data traffic, *artifical intelligence* (AI) has revolutionized almost every branch of science and technology ranging from computer vision, natural language processing, to wireless networks [1]–[4]. The implementation of AI in wireless networks is foreseen as an innovative breakthrough for realizing an intelligent network. To enable rapid access to the large amount of real-time data generated by massive distributed edge devices, a new paradigm of computing, called *edge learning* has emerged by migrating learning from central clouds towards the edge of wireless networks [5]–[7]. Despite the proximity to data source in edge learning, it still faces with a critical challenge of privacy concern when transmitting raw data to the edge server. To overcome this deficiency, an innovative framework namely *federated edge learning* (FEEL) has been proposed, which features distributed learning at edge devices and update aggregation at the edge server [8]. By periodically reporting the local learning updates to the edge server for global aggregation, FEEL is capable of leveraging the heterogeneous data to attain an accurate learning model.

A major design objective in FEEL is to accelerate the training process and maintain certain learning accuracy under communication and computation resource constraints. Due to that communication suffers from impairment of wireless channels (e.g., channel fading, interference, and noise), it is often the main bottleneck for fast FEEL [9]–[12]. This fact calls for effective scheduling algorithms for *fast update acquisition* from highly distributed edge devices. Conventional scheduling principles focus on data rate [13] or quality-of-service [14] whereas assume that the transmitted data is equally important. However, this assumption is often inappropriate in FEEL since different local learning updates are of dissimilar significance to the model convergence [15]–[17]. This motivates researchers to design a new scheduling approach based on not only the channel states but also how important the local updates are for learning. Towards this end, we consider the scheduling problem in FEEL by jointly exploiting the diversity involved in both wireless channels and local learning updates.

A. Prior Work

Communication-efficient FEEL requires the joint design of learning algorithms and communication techniques. Recent years have seen much research interests from both industry and academia, covering key topics such as learning update compression [18]–[20], radio resource management [21]–[23], [25], [26], and over-the-air computation [27]–[29]. Specifically, the

authors of [18] proposed two update compression approaches based on random sparsification and probabilistic quantization to reduce the communication cost between edge server and edge devices. In [19], a sparse ternary compression framework was developed to adapt to the non-independent and identically distributed (non-IID) FEEL environment, which compresses the upstream and downstream via sparsification, ternarization, error accumulation, and Golomb encoding. Moreover, the deep gradient compression technology was employed in [20], where a hierarchical threshold selection policy was proposed to speed up the gradient sparsification process. On the other hand, the authors of [21] developed an efficient control algorithm to achieve the trade-off between local update and global aggregation during the learning process with limited communication resource budget. Further, a joint user selection and resource allocation policy towards minimizing the FEEL loss function was proposed in [22] by taking into account the packet error over wireless links. In addition, [23] presented an effective client scheduling and resource allocation policy for FEEL over wireless links with imperfect channel state information (CSI), and [24] proposed a joint user selection and resource allocation policy to minimize the FEEL convergence time based on probabilistic scheduling. Besides, the trade-off between learning time and energy consumption in FEEL was investigated in [25] and [26], where [25] developed a closed-form communication and computation resource allocation in a synchronous manner, and [26] designed an iterative algorithm for joint power control and resource allocation. In particular, a novel technique namely *over-the-air computation* is adopted for update aggregation in FEEL. Based on this, the authors of [27] considered joint device selection and beamforming design in FEEL. Moreover, a prominent broadband analog aggregation multiple-access scheme featuring the aggregation of analog modulated gradients was proposed in [28], where two communication-and-learning trade-offs were highlighted as well. To facilitate implementation, [28] was further extended in [29], which designed a digital broadband over-the-air aggregation scheme to evaluate the effect of wireless channel impairment on model convergence.

On the other hand, the local learning updates in FEEL are not equally important for model convergence and thus are valuable to be exploited for further performance improvement [30]. Some prior works incorporated the new feature of data importance for communication design in centralized edge learning systems, such as data retransmission [31] and user scheduling [32]. Specifically, a pioneering importance-aware automatic-repeat-request protocol was designed to balance the communication reliability and the data uncertainty [31]. Furthermore, the authors also proposed an importance-aware user scheduling scheme to improve the communication

efficiency in edge learning [32]. Inspired by this, we consider to exploit the update importance for scheduling design in FEEL, which has not been well studied in existing works.

B. Contributions and Organization

In this paper, we propose a novel importance- and channel-aware scheduling policy to improve the performance of FEEL. This work is most related to the prominent FEEL works of [33] in analyzing convergence behaviour using classical scheduling policies and [34] in combining channel state and update norm for scheduling. However, the scheduling policy in [33] is not specifically designed for FEEL while the convergence analysis is not provided in [34]. We go one further step than [33] and [34] in that we provide fundamental convergence analysis as well as develop an innovative scheduling for FEEL. Our main result demonstrates that the scheduling decision should elegantly balance the diversity in wireless channels and learning updates to achieve a satisfactory learning performance.

The main contributions of this work are summarized as follows.

- We first propose a novel probabilistic scheduling framework to yield unbiased gradient aggregation in FEEL. The one-round convergence rate is analyzed and the update importance is measured by the gradient divergence.

- We consider a simple strategy of scheduling one device in each communication round and devise an importance- and channel-aware scheduling policy. The corresponding convergence rate is also provided. More precisely, the scheduling probability increases linearly with both the data unbalanced indicator and the local gradient norm, whereas decreases sublinearly with the exponent of $-\frac{1}{2}$ when the local gradient upload latency is large.

- We further extend the probabilistic scheduling framework to the strategy of scheduling multiple devices in each communication round. An efficient scheduling without replacement algorithm is designed, followed by the closed-form expression of bandwidth allocation. Numerical results demonstrate gains achieved by the proposed scheduling scheme over conventional scheduling schemes.

The rest of the paper is organized as follows. Section II introduces the system model and motivates the learning mechanism. Section III investigates the simple strategy of scheduling one device in each communication round and formulates the probabilistic scheduling problem. The optimal solution and the convergence analysis are provided in Section IV. Section V discusses how to extend the probabilistic scheduling to the strategy of scheduling multiple devices in

each communication round. Section VI draws the numerical results, followed by conclusions in Section VII.

II. SYSTEM MODEL AND LEARNING MECHANISM
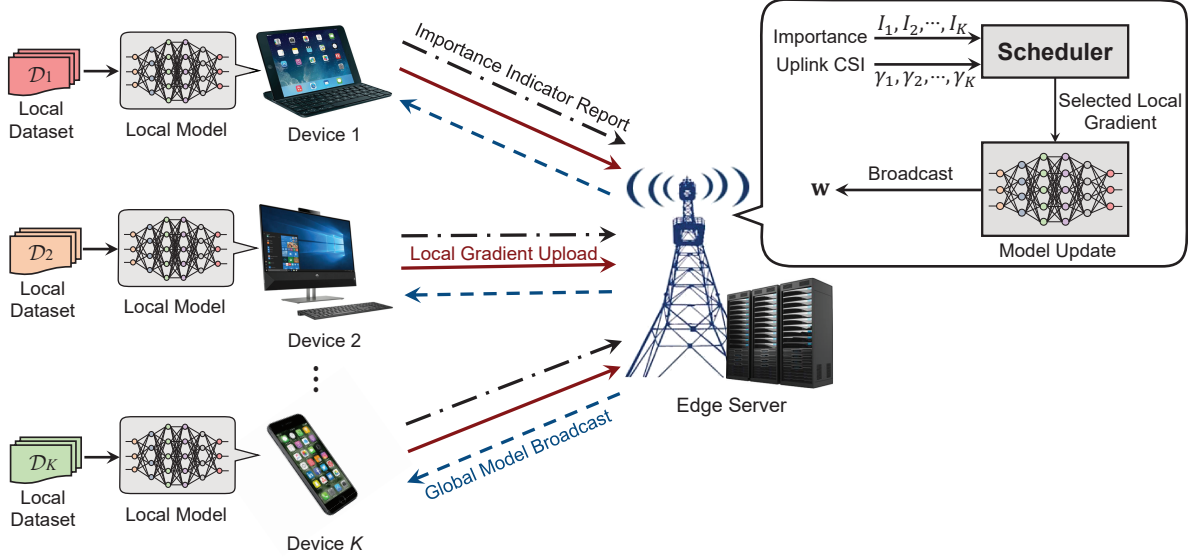
A. Federated Edge Learning System Model

Fig. 1. Federated edge learning system.

As illustrated in Fig. 1, we consider an FEEL system comprising one edge server and K decentralized single-antenna edge devices, denoted by a set $\mathcal{K} = \{1, 2, \ldots, K\}$. Each device has a fraction of labeled data. Let $\mathcal{D}_k = ((\mathbf{x}_k^1, y_k^1), (\mathbf{x}_k^2, y_k^2), \cdots, (\mathbf{x}_k^{n_k}, y_k^{n_k}))$ denote the local dataset of device k, where $\mathbf{x}_k^i$ is its i-th training data sample, y_k^i is the corresponding ground-truth label, and n_k is the number of data that device k owns. In view of the heterogeneous data structure, we assume that local datasets are statistically independent across edge devices. Let $n = \sum_{k=1}^{K} n_k$. A shared learning model, denoted by $\mathbf{w}$, needs to be collaboratively trained across all edge devices with their distributed local datasets. To leverage the rich data over edge devices while preserving data privacy, all devices adopt the federated learning technique by iteratively reporting their locally computed gradient for global aggregation in lieu of raw data. Due to the gradient sparsity, the communication overhead can be significantly reduced by some gradient compression approaches. Therefore, we focus on the gradient-averaging implementation in the subsequent exposition. Moreover, only a small subset of devices can be chosen for every global

TABLE I
LOSS FUNCTION FOR POPULAR LEARNING MODELS

Learning Model	Loss Function $\ell\left(\mathbf{w}, \mathbf{x}, y\right)$
Least-squared Support Vector Machine (SVM)	$\frac{1}{2}\max\left\{0, 1 - y\mathbf{w}^{\top}\mathbf{x}\right\} + \frac{\lambda}{2}\left\|\mathbf{w}\right\|^{2}$, where λ is a regularization parameter and $\top$ represents the transpose operator.
Linear Regression	$\frac{1}{2}\left\|y - \mathbf{w}^{\top}\mathbf{x}\right\|^{2}$
Neural Network	$\frac{1}{2}\left\|y - f\left(\mathbf{x}; \mathbf{w}\right)\right\|^{2}$, where $f\left(\mathbf{x}; \mathbf{w}\right) = f_{N}\left(\cdots f_{2}\left(f_{1}\left(\mathbf{x}; \mathbf{w}_{1}\right); \mathbf{w}_{2}\right)\cdots; \mathbf{w}_{N}\right)$ is the learning output, N is the number of layers, and $f_{i}\left(\mathbf{x}, \mathbf{w}_{i}\right)$ is the i-th layer function conditioned on the weight matrix $\mathbf{w}_{i}$.

gradient aggregation because of the limitations in communication resources. To coordinate all edge devices, a scheduler is implemented at the edge server, which performs device selection and resource allocation.

B. Learning Model

In this work, we consider a general supervised machine learning task. We define an appropriate sample-wise loss function $\ell\left(\mathbf{w}, \mathbf{x}, y\right)$ to quantify the prediction error between the data sample $\mathbf{x}$ on learning model $\mathbf{w}$ and the ground-truth label y. Specifically, Table I lists the loss functions for several popular learning models, which will be employed in the experiment.

Without loss of generality, we assume that the learning model is deployed across all edge devices. Then, the local loss function of device k that measures the model error on its dataset $\mathcal{D}_{k}$ can be defined as

$$L_{k}\left(\mathbf{w}\right) = \frac{1}{n_{k}}\sum_{i=1}^{n_{k}}\ell\left(\mathbf{w}, \mathbf{x}_{k}^{i}, y_{k}^{i}\right), \quad \forall k. \tag{1}$$

Accordingly, the global loss function associated with all distributed local datasets is given by

$$L\left(\mathbf{w}\right) = \frac{1}{n}\sum_{k=1}^{K}n_{k}L_{k}\left(\mathbf{w}\right). \tag{2}$$

The main objective of the learning task is to seek an optimal model $\mathbf{w}^{*}$ that minimizes $L\left(\mathbf{w}\right)$. In addition, the ground-truth global gradient at $\mathbf{w}$ is defined as

$$\mathbf{g} = \nabla L\left(\mathbf{w}\right), \tag{3}$$

where ∇ represents the gradient operator.

C. Communication Model

We assume orthogonal frequency-division multiple access (OFDMA) is adopted for uplink channel access. In this case, the system bandwidth, denoted by B, is divided into multiple narrowband sub-channels that are allocated to all devices without interference. Define B_k as the bandwidth allocated to device k so that we have $\sum_{k=1}^{K} B_k \leq B$. Moreover, we assume frequency flat fading, and let γ_k denote the uplink signal-to-noise ratio (SNR) of device k, which is determined by its transmit power, uplink channel gain, and noise power. Then the achievable uplink data rate (in bit/s) for device k can be expressed as

$$r_k = B_k \log_2 \left(1 + \gamma_k\right), \quad \forall k. \tag{4}$$

For the downlink channel, we assume the edge server occupies the entire system bandwidth to broadcast the global model. Let γ denote the minimum downlink channel SNR among the devices, which depends on their downlink channel gains, noise power, and the transmit power of the edge server. Then the achievable downlink data rate is given by

$$r = B \log_2 \left(1 + \gamma\right). \tag{5}$$

D. Learning Mechanism

We now introduce the federated learning mechanism by describing the steps that the communication parties take, as illustrated in Fig. 2. The combination of the steps is referred to as a *communication round*.

- *Step 1 (Global Model Broadcast):* The edge server broadcasts the current learning model $\mathbf{w}^t$ to all devices.
- *Step 2 (Local Gradient Calculation):* Device k computes its local gradient as

$$\mathbf{g}_k^t = \nabla L_k \left(\mathbf{w}^t\right). \tag{6}$$

- *Step 3 (Importance Indicator Report):* As an option to be explained later, device k computes an importance indicator $I\left(\mathbf{g}_k^t\right)$ and reports the result to the edge server.
- *Step 4 (Device Selection and Resource Allocation):* Based on the devices' reports, the edge server selects a subset of devices, denoted by $\mathcal{S}^t$, determines the bandwidth allocation, and then solicits updates from the selected devices via some control channels.
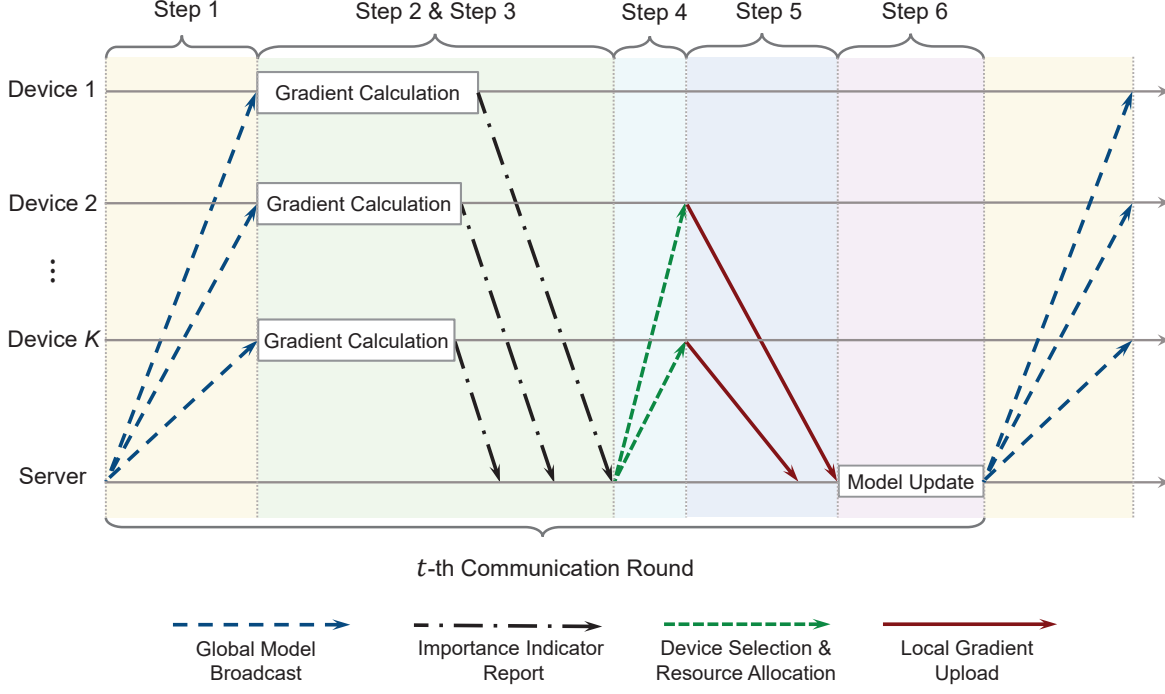
Fig. 2. Communication protocol for importance- and channel-aware federated edge learning.

- *Step 5 (Local Gradient Upload):* The selected devices upload their concrete local gradients to the edge server in the uplink multiaccess channel.[1]

- *Step 6 (Global Model Update):* After receiving all local gradients from the selected devices, the edge server computes the global gradient as

$$\mathbf{g}_{\mathcal{S}}^t = \frac{\sum_{k \in \mathcal{S}^t} n_k \mathbf{g}_k^t}{\sum_{k \in \mathcal{S}^t} n_k}. \tag{7}$$

Subsequently, the global model is updated to

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta^t \mathbf{g}_{\mathcal{S}}^t, \tag{8}$$

where $\eta^t > 0$ is the learning rate in the t-th communication round.

Starting from $t = 1$, the edge server and edge devices iterate the preceding steps until convergence.

---

[1]To facilitate theoretical analysis, we assume that the selected devices are always able to compute and upload their local gradients in the training duration.

III. Scheduling One Device in Each Round: System Design

In this section, we consider a simple strategy of scheduling one device in each communication round. We first propose a new design principle and then develop a probabilistic scheduling framework. Thereafter, we devise a concrete gradient importance indicator. Based on these, we formulate the scheduling problem for the trade-off between multiuser channel and update diversity.

A. Principle of Importance- and Channel-Aware Scheduling

As discussed earlier, the main objective of the FEEL system is to minimize the training time for attaining a desired learning accuracy. According to the federated learning mechanism, we can observe that the training time is determined by two factors, i.e., the one-round latency and the model variation in each round. Both factors are affected by the scheduling decision, making device scheduling a key problem. Specifically, the former mainly depends on the channel fading due to that the communication latency dominates the computation latency [7]. The latter is primarily decided by the selected local gradients based on the update equation (8). Since both factors are important, they should be balanced for joint scheduling design, leading to two types of multiuser diversity to be exploited in such a system. One is the independent fading in multiuser channels, namely *multiuser channel diversity*, while the other is the heterogeneous importance levels across multiuser local gradients, namely *multiuser update diversity*. Targeting active gradient acquisition in FEEL, we should simultaneously exploit the multiuser diversity in wireless channels and local gradients, called *channel-and-update diversity* to improve the learning performance.

B. Probabilistic Scheduling

In the classical federated averaging protocol, all devices are equally likely to participate in each communication round regardless of variability in the amount of data they possess. In such a way, the aggregated global gradient is generally biased [35]. To address this issue, we now propose a new probabilistic scheduling framework.

Definition 1 (Probabilistic scheduling)**.** In the t-th communication round, the edge server selects a random device $X^t \in \{1, 2, \cdots, K\}$ for gradient uploading according to a scheduling distribution $\mathcal{P}^t = (p_1^t, p_2^t, \cdots, p_K^t)$, where $p_k^t = P(X^t = k)$ is the probability that device k

can be selected. Upon receipt of the gradient update $\mathbf{g}_{X^t}^t$, the edge server computes the global gradient as

$$\widehat{\mathbf{g}}^t = \frac{n_{X^t}}{n p_{X^t}^t} \mathbf{g}_{X^t}^t. \tag{9}$$

In this framework, probability p_k^t can be interpreted as the level of importance that device k can contribute to the global model convergence. We shall note that before global model update, the selected local gradient $\mathbf{g}_{X^t}^t$ needs to be scaled by a coefficient $\frac{n_{X^t}}{n}$ at the edge server. This coefficient well quantifies the unbalanced property in global data distribution and thus makes the global gradient $\widehat{\mathbf{g}}^t$ unbiased.

Lemma 1 (Unbiasedness). The expectation of the global gradient defined in (9) is equal to the ground-truth global gradient.

Proof: Using (2) and (3), the ground-truth global gradient in the t-th communication round is given by

$$\begin{aligned}
\mathbf{g}^t &= \frac{\partial L\left(\mathbf{w}^t\right)}{\partial \mathbf{w}^t} \\
&= \frac{1}{n} \sum_{k=1}^{K} n_k \frac{\partial L_k\left(\mathbf{w}^t\right)}{\partial \mathbf{w}^t} \\
&= \frac{1}{n} \sum_{k=1}^{K} n_k \mathbf{g}_k^t. \tag{10}
\end{aligned}$$

On the other hand, by taking expectation of $\widehat{\mathbf{g}}^t$ over the scheduling distribution $\mathcal{P}^t$, we have

$$\begin{aligned}
\mathbb{E}\left\{\widehat{\mathbf{g}}^t\right\} &= \sum_{k=1}^{K} p_k^t \frac{n_k}{n p_k^t} \mathbf{g}_k^t \\
&= \frac{1}{n} \sum_{k=1}^{K} n_k \mathbf{g}_k^t \\
&= \mathbf{g}^t, \tag{11}
\end{aligned}$$

where (11) is due to (10). Therefore, we can conclude that $\mathbb{E}\left\{\widehat{\mathbf{g}}^t\right\} = \mathbf{g}^t$, i.e., $\widehat{\mathbf{g}}^t$ is an unbiased estimate of the ground-truth global gradient.

The probabilistic scheduling yields unbiased gradient for all distributions as long as every device has a nonzero probability to update the server, including the typical uniform random scheduling in classical FL. However, different distributions lead to different performance in

general.

C. Gradient Importance Measurement

To develop a concrete metric to evaluate the importance of each local gradient, we first analyze the contribution that each local gradient can provide for global model convergence. For the purpose of analysis, we make the following assumption to the global loss function.

Assumption 1 (Lipschitz gradient continuity)**.** The gradient $\nabla L\left(\mathbf{w}\right)$ of the global loss function $L\left(\mathbf{w}\right)$ is Lipschitz continuous with a positive modulus ℓ, or equivalently, for any $\mathbf{u}$ and $\mathbf{v}$, it satisfies

$$\left\|\nabla L\left(\mathbf{u}\right) - \nabla L\left(\mathbf{v}\right)\right\| \leq \ell \left\|\mathbf{u} - \mathbf{v}\right\|, \tag{12}$$

where $\left\|\cdot\right\|$ represents the L_2 norm operator.

Then, we give the convergence result of the probabilistic scheduling, as presented in the following lemma. The detailed proof is provided in Appendix A.

Lemma 2 (One-round convergence rate)**.** When Assumptions 1 holds and let $\mathbf{w}^*$ denote the optimal global learning model. Then, the convergence rate of each communication round is given by

$$\mathbb{E}\left\{L\left(\mathbf{w}^{t+1}\right) - L\left(\mathbf{w}^*\right)\right\} \leq \mathbb{E}\left\{L\left(\mathbf{w}^t\right) - L\left(\mathbf{w}^*\right)\right\} - \eta^t \left(1 - \frac{1}{2}\ell\eta^t\right) \left\|\mathbf{g}^t\right\|^2$$
$$+ \frac{1}{2}\ell\left(\eta^t\right)^2 \mathbb{E}\left\{\left\|\widehat{\mathbf{g}}^t - \mathbf{g}^t\right\|^2\right\}. \tag{13}$$

Lemma 2 reveals that the expected gap between the global loss value and the optimal one is bounded by the aggregate of three terms. The first term on the right-hand side of (13) denotes the expected gap of the previous round. The second term is proportional to the squared norm of the ground-truth global gradient $\mathbf{g}^t$. The third term is proportional to the variance of the aggregated global gradient $\mathbb{E}\left\{\left\|\widehat{\mathbf{g}}^t - \mathbf{g}^t\right\|^2\right\}$. The first two terms are independent of the scheduling decision such that they can be viewed as two constants. The variance of the aggregated global gradient depends on the scheduling design and thus needs to be optimized. Note that the ground-truth global gradient reflects the steepest direction for the global loss decline. Therefore, the smaller the variance of aggregated global gradient is, the faster the global loss decreases.

The result in Lemma 2 provides a significant connection between gradient variance and model convergence rate. Inspired by this, we can take the difference between the scaled local gradient

and the ground-truth global gradient, called *gradient divergence* as a reasonable importance measurement, as described in the following definition.

Definition 2 (Gradient divergence)**.** Conditioned on the probabilistic scheduling framework, the importance indicator of each local gradient is defined as

$$I_k^t = \left\| \frac{n_k}{n p_k^t} \mathbf{g}_k^t - \mathbf{g}^t \right\|^2, \quad \forall k. \tag{14}$$

The gradient divergence reflects the deviation between each local gradient and the ground-truth global gradient. The smaller the local gradient divergence is, the more it can contribute to model convergence.

D. One-Round Latency Analysis

As mentioned earlier, the learning performance (training time) is determined not only by the gradient importance, but also by the one-round latency, making it an essential point to analyze. According to the previous federated learning mechanism, we can give the dominant kinds of latency in one communication round.

- *Global Model Broadcast Latency:* Let S denote the total number of learning parameters and q denote the quantitative bit number for each parameter. Then, the data size of the global model can be evaluated as $q \times S$ and the global model broadcast latency is given by

$$
\begin{aligned}
T^{\mathrm{B},t} &= \frac{qS}{r^t} \\
&= \frac{qS}{B \log_2 (1 + \gamma^t)}. \tag{15}
\end{aligned}
$$

- *Local Gradient Calculation Latency:* Define C as the number of floating-point operations used for performing the backpropagation algorithm with one data sample. Moreover, let f_k denote the computational capability (in floating-point operations per second) of device k. Accordingly, the local gradient calculation latency of device k can be expressed as

$$T_k^{\mathrm{C},t} = \frac{n_k C}{f_k}, \quad \forall k. \tag{16}$$

- *Local Gradient Upload Latency:* Due to the fact that each parameter has a counterpart gradient, the total number of elements in each local gradient equals S as well. For convenience, we also use q bits to quantize each gradient element. Besides, the selected device

can occupy the entire system bandwidth since we schedule one device in each round. Given that device k is scheduled, its local gradient upload latency can be expressed as

$$
\begin{aligned}
T_k^{\mathrm{U},t} &= \frac{qS}{r_k^t} \\
&= \frac{qS}{B \log_2\left(1 + \gamma_k^t\right)}, \quad \forall k.
\end{aligned}
$$
(17)

Assuming that the edge server is computationally powerful, we ignore the time for device scheduling and global model update. Moreover, the edge server cannot start scheduling until it completes the channel state estimation and gradient importance evaluation for all devices. Towards this end, the one-round latency for scheduling device k is given by

$$
T_k^t = T^{\mathrm{B},t} + \max_{i \in \mathcal{K}} \left\{ T_i^{\mathrm{C},t} \right\} + T_k^{\mathrm{U},t}, \quad \forall k.
$$
(18)

E. Problem Formulation

Thus far, the multiuser channel and update diversity have been evaluated by the one-round latency and the gradient divergence, respectively. Targeting training acceleration, it is desirable to schedule the device with the best channel state as well as the smallest gradient divergence. However, this case rarely occurs in practice, resulting in a trade-off between gradient divergence and one-round latency. Define $\rho \in [0, 1]$ as the weight coefficient that balances the gradient divergence and the one-round latency. [2] Then, the objective function can be defined as

$$
\mathbb{E}\left\{ \rho I_{X^t}^t + (1-\rho) T_{X^t}^t \right\} = \sum_{k=1}^{K} p_k^t \left(\rho \left\| \frac{n_k}{n p_k^t} \mathbf{g}_k^t - \mathbf{g}^t \right\|^2 + (1-\rho) T_k^t \right).
$$
(19)

Accordingly, the multiuser channel and update diversity trade-off problem can be formulated towards minimizing (19), as

$$
\mathscr{P}_1: \quad \underset{\left(p_1^t, \cdots, p_K^t\right)}{\text{minimize}} \quad \sum_{k=1}^{K} p_k^t \left(\rho \left\| \frac{n_k}{n p_k^t} \mathbf{g}_k^t - \mathbf{g}^t \right\|^2 + (1-\rho) T_k^t \right),
$$
(20a)

$$
\text{subject to} \quad \sum_{k=1}^{K} p_k^t = 1,
$$
(20b)

$$
p_k^t \geq 0, \ \forall k \in \mathcal{K}.
$$
(20c)

---

[2] The selection of ρ depends on the specific values of gradient divergence and one-round latency. A simple way is to choose a value of ρ to make the weighted gradient divergence and the weighted one-round latency in the same order.

IV. SCHEDULING OPTIMIZATION AND CONVERGENCE ANALYSIS

In this section, we first develop the optimal scheduling policy to Problem $\mathscr{P}_1$. Based on this, we will provide the convergence performance of the FEEL system.

A. Optimal Scheduling Policy

The main challenge in solving Problem $\mathscr{P}_1$ is that the ground-truth global gradient cannot be obtained by neither the edge server nor edge devices, making the gradient divergence hard to calculate in practice. To make progress, we first rewrite the first term in the objective function (19) as

$$\mathbb{E}\left\{\rho I_{X^t}^t\right\} = \sum_{k=1}^{K} p_k^t \left(\rho \left\| \frac{n_k}{np_k^t}\mathbf{g}_k^t - \mathbf{g}^t \right\|^2\right)$$

$$= \sum_{k=1}^{K} \left[\frac{\rho}{p_k^t}\left(\frac{n_k}{n}\right)^2 \|\mathbf{g}_k^t\|^2 \right] - \rho \|\mathbf{g}^t\|^2, \tag{21}$$

where (21) is due to the unbiasedness property of the aggregated global gradient. In particular, the ground-truth global gradient, the global model broadcast latency, and the local gradient calculation latency are independent of the scheduling decision so that they can be excluded from the optimization problem. Accordingly, Problem $\mathscr{P}_1$ can be equivalently transformed into

$$\mathscr{P}_2: \ \minimize_{\left(p_1^t, \cdots, p_K^t\right)} \ \sum_{k=1}^{K} \left[\frac{\rho}{p_k^t}\left(\frac{n_k}{n}\right)^2 \|\mathbf{g}_k^t\|^2 + (1-\rho)p_k^t T_k^{\mathrm{U},t} \right], \tag{22}$$

$$\text{subject to} \quad \text{(20b) and (20c).}$$

To solve Problem $\mathscr{P}_2$, we first establish the following lemma, which is proved in Appendix B.

Lemma 3. Given a general optimization problem with positive coefficients a_k, b_k, and c_k as follows

$$\mathscr{P}_3: \ \minimize_{(x_1,\cdots,x_K)} \ \sum_{k=1}^{K} a_k \frac{1}{x_k} + b_k x_k, \tag{23a}$$

$$\text{subject to} \quad \sum_{k=1}^{K} c_k x_k = d, \tag{23b}$$

$$x_k \geq 0, \ \forall k \in \mathcal{K}. \tag{23c}$$

The optimal solution can be expressed as

$$x_k^* = \sqrt{\frac{a_k}{b_k + \lambda^* c_k}}, \quad \forall k. \tag{24}$$

where λ^* is the optimal value satisfying $\sum_{k=1}^{K} c_k x_k^* = d$.

With comprehensive comparison, we can observe that the structures of problems $\mathscr{P}_2$ and $\mathscr{P}_3$ are identical such that the result in (24) is applicable in solving $\mathscr{P}_2$. Towards this end, we can obtain the optimal scheduling policy, as characterized in the following theorem.

Theorem 1 (Importance- and channel-aware scheduling). Considering the strategy of scheduling one device in each communication round, the optimal probability that the edge server selects device k to report its local gradient is given by

$$p_k^{t*} = \frac{n_k}{n} \left\| \mathbf{g}_k^t \right\| \sqrt{\frac{\rho}{(1-\rho)T_k^{\mathrm{U},t} + \lambda^{t*}}}, \quad \forall k, \tag{25}$$

where λ^{t*} is the optimal value of the Lagrangian multiplier that guarantees $\sum_{k=1}^{K} p_k^{t*} = 1$ and can be obtained by the one-dimensional searching algorithm.

Theorem 1 reveals that the optimal scheduling decision is mainly determined by three factors: the data unbalanced indicator $\frac{n_k}{n}$, the local gradient norm $\|\mathbf{g}_k^t\|$, and the local gradient upload latency $T_k^{\mathrm{U},t}$. More precisely, for given value of λ^{t*}, the scheduling probability increases linearly with both the data unbalanced indicator and the local gradient norm, whereas decreases sublinearly with the exponent of $-\frac{1}{2}$ when the local gradient upload latency is large. The data unbalanced indicator reflects the data size proportion of each local dataset in the global dataset. Therefore, a larger data unbalanced indicator implies that the local gradient contains more data information such that it can contribute more to the model convergence. On the other hand, it is desirable to schedule the device with a larger gradient norm as it includes more non-zero elements and can further promote the model update. Moreover, the result suggests us to assign a small probability to the device with bad channel state. Also, scheduling a device with bad channel state will incur straggler issue, which should be avoided in the scheduling design.

From Theorem 1, we can observe that the importance level of each local gradient is well characterized by two scalars: its gradient norm and local dataset size. Moreover, both scalars can be easily derived at local devices while transmitting them to the scheduler will not incur excessive delay. This result drives us to use the product of these two scalars as the gradient

importance measurement. On the other hand, the scheduler is able to evaluate each local gradient upload latency via dynamical channel state estimation. Towards this end, both channel and update diversity can be jointly exploited for practical scheduling design.

B. Convergence Analysis

We now investigate the convergence behaviour of the learning algorithm under the proposed scheduling solution. To facilitate analysis, we further make the following assumption to the global loss function.

Assumption 2 (Strong convexity)**.** The global loss function $L(\mathbf{w})$ is strongly convex with a positive parameter μ such that for any $\mathbf{u}$ and $\mathbf{v}$, it follows

$$L(\mathbf{u}) \geq L(\mathbf{v}) + \nabla L(\mathbf{v})^{\top}(\mathbf{u} - \mathbf{v}) + \frac{\mu}{2}\|\mathbf{u} - \mathbf{v}\|^2. \tag{26}$$

This standard assumption is satisfied by many popular learning models, such as the least-squared SVM and the linear regression in Table I. In Section VI, we will conduct experiments to demonstrate that our analytical results also work well for some popular loss functions which do not satisfy this assumption. Then, by applying the result of Theorem 1 into Lemma 2, we can derive the convergence upper bound of each communication round, as shown in the following theorem. The detailed proof is presented in Appendix C.

Theorem 2 (Convergence upper bound)**.** Conditioned on the scheduling policy in Theorem 1, the convergence upper bound in the t-th communication round can be given by

$$\mathbb{E}\left\{L\left(\mathbf{w}^{t+1}\right) - L\left(\mathbf{w}^*\right)\right\} \leq \prod_{i=1}^{t}\left(1 - 2\mu\eta^i\right)\mathbb{E}\left\{L\left(\mathbf{w}^1\right) - L\left(\mathbf{w}^*\right)\right\}$$
$$+ \frac{\ell}{2n}\sum_{i=1}^{t} A^i\left(\eta^i\right)^2 \sum_{k=1}^{K} n_k \left\|\mathbf{g}_k^i\right\| \sqrt{\frac{(1-\rho)T_k^{\mathrm{U},i} + \lambda^{i*}}{\rho}}, \tag{27}$$

where $A^i = \prod_{j=i+1}^{t}\left(1 - 2\mu\eta^j\right)$ is a weight coefficient.

Theorem 2 indicates that the expected gap between the global loss value and the optimal one is bounded by the aggregate of two terms: one is the expected gap in the initial round and the other is the cumulate impact of the proposed scheduling policy on model convergence. The former will converge to zero as t increases when the learning rate satisfies $0 < \eta^i < \frac{1}{2\mu}$, $\forall i$.

Meanwhile, the weight coefficient A^i will also decrease, making the latter approach to zero as well. Therefore, the learning model will finally converge to the optimum under the proposed scheduling policy.

The convergence upper bound in Theorem 2 holds for any learning rate that satisfies $0 < \eta^i < \frac{1}{2\mu}$ $\forall i$. To explicitly characterize the convergence performance, we diminish the learning rate as the training proceeds such that the convergence rate of the learning algorithm can be given as follows. The detailed proof is provided in Appendix D.

Corollary 1 (Learning convergence rate)**.** Let T denote the number of communication rounds, and $\eta^t = \frac{\chi}{t + \nu}$ denote the learning rate for given $\chi > \frac{1}{2\mu}$ and $\nu > 0$. Then the convergence rate of the learning algorithm is given by

$$\mathbb{E}\left\{L\left(\mathbf{w}^T\right) - L\left(\mathbf{w}^*\right)\right\} \leq \frac{\zeta}{T + \nu}, \tag{28}$$

where $\zeta = \max\left\{\frac{\ell G^2 \chi^2}{2\left(2\mu\chi - 1\right)}, \left(1 + \nu\right)\mathbb{E}\left\{L\left(\mathbf{w}^1\right) - L\left(\mathbf{w}^*\right)\right\}\right\}$ and $G = \max\|\widehat{\mathbf{g}}^t\|$ is the largest gradient norm across the communication rounds.

According to Corollary 1, we can conclude that the increment of total communication rounds leads to the convergence of the learning algorithm. Moreover, the convergence rate is asymptotically $\mathcal{O}\left(\frac{1}{T}\right)$.

V. SCHEDULING MULTIPLE DEVICES IN EACH ROUND

In this section, we investigate the strategy of scheduling multiple devices in each communication round. We first develop a low-complexity algorithm for scheduling design and then propose an optimal bandwidth allocation for practical implementation.

A. Scheduling Design

The probabilistic scheduling framework is still applicable in the strategy of scheduling multiple devices in each communication round by combining the *sampling without replacement* approach [36]. Let M denote the number of devices that need to be scheduled in each round. Then the probabilistic scheduling framework can be defined as follows.

Definition 3 (Probabilistic scheduling without replacement)**.** In the t-th communication round, the edge server selects a random device sequence $\mathcal{Y}^t = (Y_1^t, Y_2^t, \cdots, Y_M^t)$ without replacement

according to the scheduling distribution $\mathcal{P}^t = (p_1^t, p_2^t, \cdots, p_K^t)$, where $Y_m^t \in \{1, 2, \cdots, K\}$ is a random variable representing the index of the m-th scheduled device. Upon receipt of the gradient updates $\mathbf{g}_{Y_m^t}^t$, the edge server computes the global gradient as

$$\widehat{\mathbf{g}}_{\mathcal{Y}}^t = \frac{1}{Mn} \sum_{m=1}^{M} \frac{n_{Y_m^t}}{q_{Y_m^t}^t} \mathbf{g}_{Y_m^t}^t, \tag{29}$$

where $q_{Y_m^t}^t = \frac{p_{Y_m^t}^t}{1 - \sum_{j=1}^{m-1} p_{Y_j^t}^t}$ is the probability of scheduling device Y_m^t conditioned on the selected device sequence $(Y_1^t, \cdots, Y_{m-1}^t)$.

Lemma 4 (Unbiasedness). The expectation of the global gradient defined in (29) is equal to the ground-truth global gradient.

Proof: The proof can be conducted by proving the unbiasedness of each selected local gradient. Assume that we have selected $m-1$ devices, denoted by a sequence $(k_1^t, \cdots, k_{m-1}^t)$. Then the scheduling probabilities for these devices become zero in the m-th selection step. Accordingly, the conditional scheduling distribution for selecting the next device is given by

$$q_k^t = \begin{cases} 0, & \text{if } k \in (k_1^t, \cdots, k_{m-1}^t), \\ \dfrac{p_k^t}{1 - \sum_{j=1}^{m-1} p_{k_j^t}^t}, & \text{otherwise.} \end{cases} \tag{30}$$

Consequently, the edge server randomly selects a device Y_m^t and computes the scaled local gradient by

$$\widehat{\mathbf{g}}_{Y_m^t}^t = \frac{n_{Y_m^t}}{n q_{Y_m^t}^t} \mathbf{g}_{Y_m^t}^t. \tag{31}$$

Taking expectation of $\widehat{\mathbf{g}}_{Y_m^t}^t$ over the conditional scheduling distribution, we have

$$\begin{aligned} \mathbb{E}\left\{\widehat{\mathbf{g}}_{Y_m^t}^t\right\} &= \sum_{k=1}^{K} q_k^t \frac{n_k}{n q_k^t} \mathbf{g}_k^t \\ &= \frac{1}{n} \sum_{k=1}^{K} n_k \mathbf{g}_k^t \\ &= \mathbf{g}^t. \end{aligned} \tag{32}$$

Therefore, each selected local gradient is unbiased, making their average defined in (29) an unbiased estimate of the ground-truth global gradient. This completes the proof.

Based on the preceding discussion, we now present an efficient scheduling algorithm, as described in Algorithm 1. To facilitate practical implementation, we first assume that the entire bandwidth is available to each device and then sequentially select one device from the candidate device set in each iteration until reaching the required number of devices. Although it sacrifices the optimality but will significantly reduce the computational complexity. Specifically, the total number of iterations is M. Moreover, let ϵ denote the maximum tolerance for the one-dimensional search for λ^{t*}. Then, the computational complexity of Algorithm 1 is $\mathcal{O}\left(\log \frac{1}{\epsilon} + M\right)$, which can be easily implemented in practical systems.

Algorithm 1: Importance- and Channel-Aware Scheduling at Edge Server

---

Input: Uplink SNR γ_k^t, local gradient norm $\|\mathbf{g}_k^t\|$, local dataset size n_k, scheduled device number M.

Output: Scheduling device sequence $\mathcal{Y}^t$.

1 Initialize $\mathcal{Y}^t = \emptyset$, $\mathcal{Q}^t = \emptyset$, $m = 1$.
2 Compute the optimal scheduling distribution $\mathcal{P}^{t*} = (p_1^{t*}, p_2^{t*}, \cdots, p_K^{t*})$ according to (25), involving the one-dimensional search for λ^{t*}.
3 Let $\mathcal{Q}^t \leftarrow \mathcal{P}^{t*}$.
4 **repeat**
5 $\quad$ Randomly select a device k_m^t based on $\mathcal{Q}^t$.
6 $\quad$ $\mathcal{Y}^t \leftarrow \mathcal{Y}^t + \{k_m^t\}$.
7 $\quad$ Update $\mathcal{Q}^t$ according to (30).
8 $\quad$ $m \leftarrow m + 1$.
9 **until** $m > M$;

B. Bandwidth Allocation

Given a typical device sequence $\mathcal{Y}^t = (k_1^t, k_2^t, \cdots, k_M^t)$ for gradient aggregation, another concern related is the corresponding bandwidth allocation towards minimizing the one-round latency. Let $B_{k_m^t}^t$ denote the bandwidth allocated to the device k_m^t in the t-th communication round. Then, the local gradient upload latency can be expressed as

$$
\begin{aligned}
T_{k_m^t}^{\mathrm{U},t} &= \frac{qS}{r_{k_m^t}^t} \\
&= \frac{qS}{B_{k_m^t}^t \log_2\left(1 + \gamma_{k_m^t}^t\right)}, \quad \forall k_m^t \in \mathcal{Y}_t.
\end{aligned}
\tag{33}
$$

On the other hand, the global model broadcast latency and the local gradient calculation latency are both independent of the bandwidth allocation. Thus, the bandwidth allocation should be

optimized towards minimizing the overall gradient upload latency, as

$$\mathscr{P}_4: \quad \underset{\left(B_{k_1^t}^t, \cdots, B_{k_M^t}^t\right)}{\text{minimize}} \quad \underset{k_m^t \in \mathcal{Y}^t}{\text{maximize}} \left\{ \frac{qS}{B_{k_m^t}^t \log_2\left(1 + \gamma_{k_m^t}^t\right)} \right\}, \tag{34a}$$

$$\text{subject to} \quad \sum_{k_m^t \in \mathcal{Y}^t} B_{k_m^t}^t \leq B, \tag{34b}$$

$$B_{k_m^t}^t \geq 0, \ \forall k_m^t \in \mathcal{Y}^t. \tag{34c}$$

It is easy to observe that $\mathscr{P}_4$ is convex such that it can be solved by the KKT conditions [38]. For ease of notation, define $R_{k_m^t}^t = \log_2\left(1 + \gamma_{k_m^t}^t\right)$ as the instantaneous uplink data rate of device k_m^t with unit bandwidth. Let $\left(B_{k_1^t}^{t*}, B_{k_2^t}^{t*}, \cdots, B_{k_M^t}^{t*}\right)$ denote the optimal solution to $\mathscr{P}_4$. Then, by parametric algorithm and simple mathematical calculation, we can find that the optimal bandwidth allocation is achieved when the local gradient upload latency for each device is equalized. Therefore, the optimal bandwidth allocation can be presented in the following theorem.

Theorem 3 (Bandwidth allocation). Given the device sequence $\mathcal{Y}^t = (k_1^t, k_2^t, \cdots, k_M^t)$ to report their local gradients in the t-th communication round, the optimal bandwidth allocation is given by

$$B_{k_m^t}^{t*} = \frac{B}{R_{k_m^t}^t} \left(\sum_{m=1}^M \frac{1}{R_{k_m^t}^t}\right)^{-1}, \quad \forall k_m^t \in \mathcal{Y}_t. \tag{35}$$

The result in Theorem 3 is rather intuitive that more bandwidth should be allocated to the devices with worse channel states to achieve the smallest one-round latency. By this means, the communication latency of each scheduled device can be equalized and thus facilitate the update synchronization needed for gradient aggregation.

VI. NUMERICAL RESULTS

In this section, we conduct experiments to validate the theoretical analysis and test the performance of the proposed algorithms. All codes are implemented in python 3.6 and are run on a Linux server equipped with four NVIDIA GeForce GTX 1080 Ti GPUs.

A. Experiment Settings

The default experiment settings are given as follows unless specified otherwise. We consider a small-cell network with one edge server and $K = 30$ edge devices. The cell radius is 500 m.

The edge server is located at the center of the network and each device is uniformly distributed within the coverage. The path loss between each device and the edge server is generated by $128.1 + 37.6\log_{10}(d)$ (in dB) according to the LTE standard [37], where d is the device-to-server distance in kilometer. The channel noise power density is -174 dBm/Hz. The transmit powers of each device and the edge server are set as 24 dBm and 46 dBm, respectively. The system bandwidth is $B = 1$ MHz. The average quantitative bit number for each gradient and parameter element is $q = 16$ bits.

For exposition, we consider the learning task of training classifiers. Two prevalent learning models of least-squared SVM and convolutional neural network (CNN) are employed for implementation.[3] To better simulate the mobile data distribution in FEEL, we consider the non-IID data partitioning way as follows. For the SVM model, we choose two typical classes of "airplane" and "automobile" in the well-known dataset CIFAR-10 for classification, where each device owns 330 data samples in one class. For the CNN model, we use the popular MNIST dataset that consists of 10 categories ranging from digit "0" to "9" and a total of 60,000 labeled training data samples. We first sort all data samples by their digit labels, divide them into 60 shards of size 1,000, and then assign each device with two shards. By this means, each device obtains the data samples with only two types of digits, making the data distribution over devices a pathological non-IID manner. Moreover, the learning step-sizes for the SVM model and CNN model are set as $\eta_{\mathrm{SVM}} = 0.0001$ and $\eta_{\mathrm{CNN}} = 0.005$, respectively.

To demonstrate the effectiveness of the proposed scheduling policy, two baseline policies namely *channel-aware scheduling* and *importance-aware scheduling* are also implemented in the following experiments. Specifically, the coefficient is $\rho = 0$ in the former such that the scheduling decision depends merely on the channel state. Meanwhile, the coefficient is $\rho = 1$ in the latter such that the scheduling decision is solely decided by the update importance.

B. Scheduling One Device in Each Round

1) Performance Comparison in SVM: Fig. 3(a) depicts the learning performance of the proposed scheduling policy as compared with the two baseline policies using the SVM model. From this figure, we can observe that at the initial stage, the proposed policy with $\rho = 5 \times 10^{-6}$ can

---

[3]The CNN has 6 layers including two 5×5 convolution layers with ReLu activation (the first layer with 32 channels and the second layer with 64 channels), each followed with a 2×2 max pooling layer, a fully connected layer with 512 units and ReLu activation, and a softmax output layer.
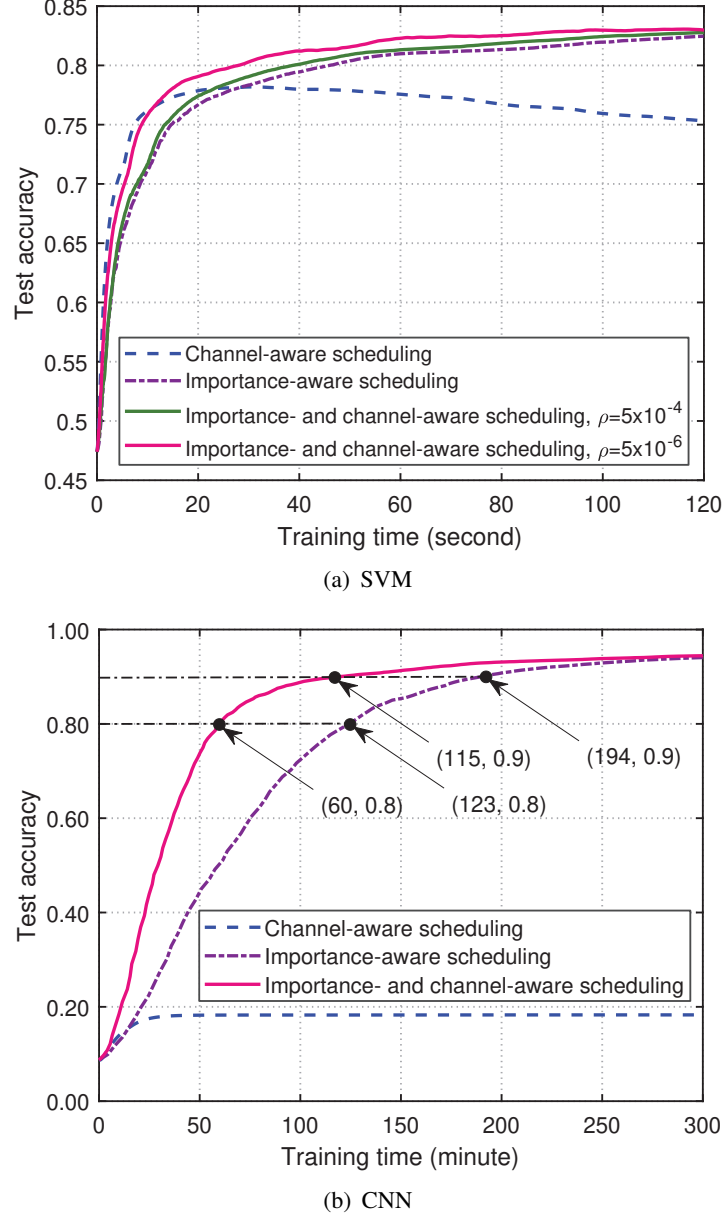
(a) SVM

(b) CNN

Fig. 3. Performance comparison among different scheduling policies in the strategy of scheduling one device in each communication round.

achieve comparable convergence speed as that of the channel-aware scheduling but converges faster than the importance-aware scheduling. It is because that the learning updates of different devices play a similar role in model convergence at the beginning of the training process, making channel state the dominant factor for learning performance improvement. Nevertheless, the proposed policy can outperform the channel-aware scheduling as the training proceeds. The underlaying reason is that the gain by exploiting channel diversity is saturated such that

the improvement primarily comes from the update diversity. In particular, the channel-aware scheduling suffers from a visible performance degradation after achieving an accuracy peak. As we can imagine, part of devices with poor communication channels cannot participate in the training process when scheduling decision merely depends on the channel state. This inevitably incurs the data deficiency and will result in overfitting issue. On the other hand, the proposed policy always outperforms the importance-aware scheduling policy since it well balances the channel state and update importance in the training duration. Last, we can find that the weight coefficient ρ has a great impact on model convergence. In specific, an overlarge ρ may lead to a marginal performance gain in that the proposed policy is hardly to exploit the channel diversity across edge devices. This suggests that the coefficient should be carefully picked to balance the channel state and update importance and thereby improving the learning performance.

2) Performance Comparison in CNN: Fig. 3(b) illustrates the performance comparison among the three scheduling policies using the CNN model. The weight coefficient for the proposed scheduling policy is set as $\rho = 5 \times 10^{-3}$. Similar trends as in the SVM model are observed, and the proposed policy is found to consistently surpass the baseline policies, which confirms the robustness and stability of the theoretical results against the model structure. More precisely, when the target accuracy is 0.8, the required training time is 60 minutes for the importance- and channel-aware scheduling while the importance-aware scheduling takes 123 minutes to achieve the same accuracy. Therefore, the proposed policy can save more than half time to achieve the target accuracy comparing against the importance-aware scheduling. The performance gain becomes more remarkable if the target accuracy is 0.9. In contrast, the channel-aware scheduling is incapable of achieving the target accuracy due to the lack of data information from the devices with poor channel conditions. In particular, the performance gap between the channel-aware scheduling and importance- and channel-aware scheduling is more evident than that in the SVM model. It attributes to the fact that the update diversity in the multiclass classification scenario is larger than that in the binary classification scenario. These observations further demonstrate the superiority of the proposed scheduling policy to deal with the non-IID data in distributed FEEL system.

C. Scheduling Multiple Devices in Each Round

1) Performance Comparison: The learning performance of the three scheduling policies are evaluated in the strategy of scheduling multiple devices in each communication round. Both
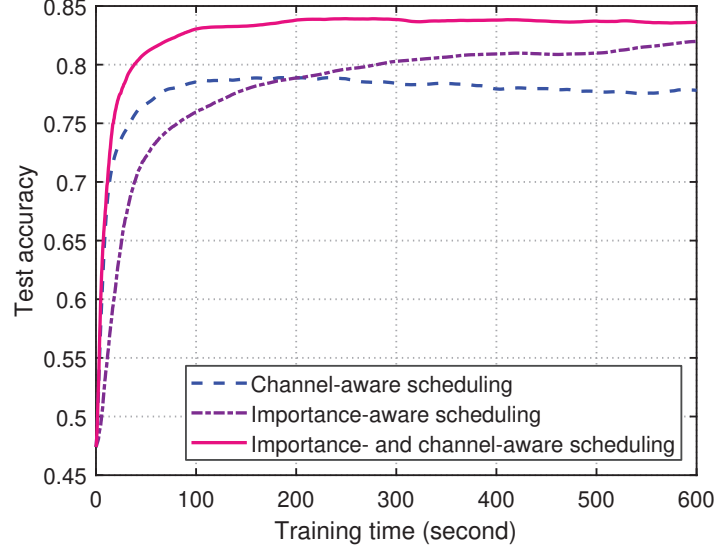
(a) SVM
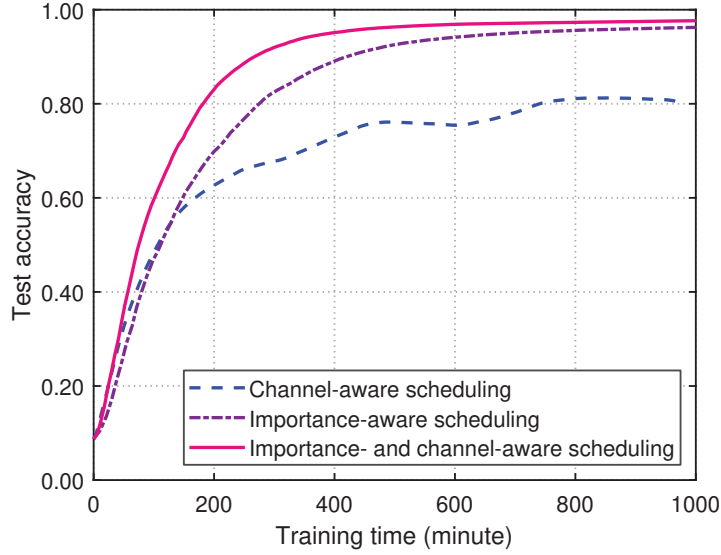
(b) CNN

Fig. 4. Performance comparison among different scheduling policies in the strategy of scheduling multiple devices in each communication round.

the models of SVM and CNN are experimented. For SVM, we assume that $M = 10$ devices can be scheduled in each round and the coefficient is set as $\rho = 5 \times 10^{-6}$. For CNN, we assume that $M = 3$ devices can be scheduled in each round and the coefficient is set as $\rho = 1 \times 10^{-3}$. The results are displayed in Figs. 4(a) and 4(b), respectively. It can be observed from all plots that the proposed policy achieves the fastest convergence speed along with the highest learning accuracy throughout the entire training duration. This result is rather intuitive since

the channel-aware scheduling overlooks the update importance whereas the importance-aware scheduling is unconscious of the channel condition. It also confirms the applicability of the proposed sequential scheduling without replacement algorithm, even though it cannot guarantee the optimal device scheduling. On the other hand, the performance improvement in this strategy is more conspicuous than that in the strategy of scheduling one device in each communication round, which substantiates the theoretical gain brought by the intelligent trade-off between quality and quantity in update aggregation.
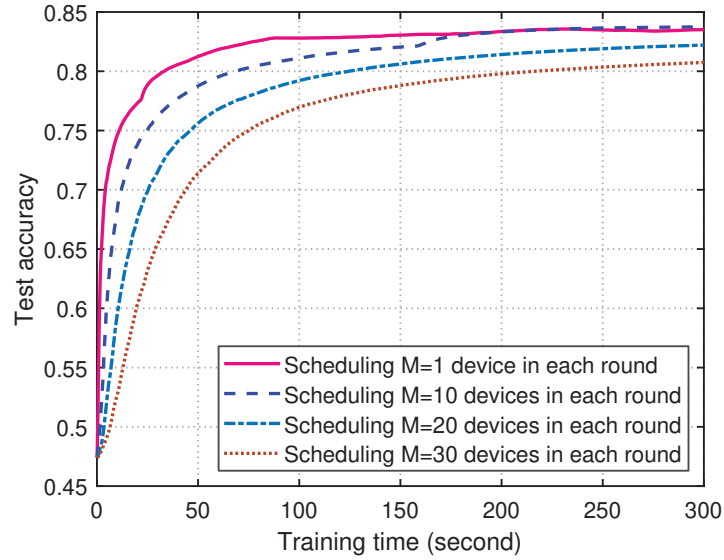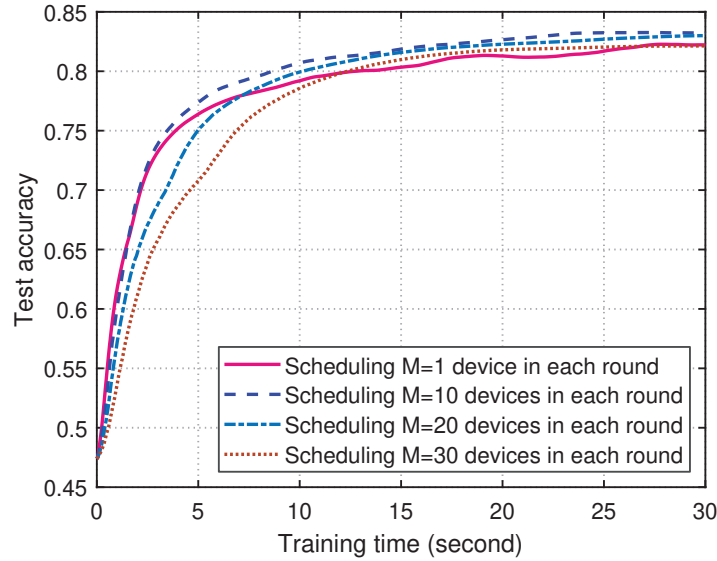

(a) $B = 1$ MHz

(b) $B = 20$ MHz

Fig. 5. Performance comparison between two strategies.

2) Scheduled Device Number: Given the alternative strategies of scheduling one and multiple devices in each communication round, a relevant concern is the proper selection between them. For exposition, we take the SVM model as an example and analyze the impact of the communication resource on model convergence. The curves of learning accuracy versus the training time with different scheduled device numbers are presented in Figs. 5(a) and 5(b), for small bandwidth $B = 1$ MHz and large bandwidth $B = 20$ MHz, respectively. In Fig. 5(a), we can observe that the scheme of scheduling one device in each round can achieve faster convergence speed than other schemes without sacrificing the learning accuracy. In contrast, the scheme of scheduling 10 devices in each round is superior to other schemes in terms of convergence speed when the system bandwidth is 20 MHz. This result has a profound and refreshing implication that the device number selection criterion should adapt to the wireless environment. Upon a certain guarantee on communication resource, scheduling multiple devices in each round can reduce the global gradient variance and thus accelerate the training process. This is also aligned with the previous discussions that the communication efficiency is the primary concern in current FEEL systems.

VII. CONCLUSION

In this paper, we have proposed a new importance- and channel-aware scheduling policy for update aggregation in FEEL. We have developed a probabilistic scheduling framework to guarantee the unbiasedness of the aggregated global gradient as well as to speed up convergence. An optimal scheduling policy has been designed to incorporate the channel state and update importance for decision making. We have also provided the concrete convergence analysis. Moreover, the probabilistic scheduling framework has been extended to the strategy of scheduling multiple devices in each communication round and some practical implementation issues have been highlighted as well. Comprehensive experiments using real datasets substantiate the performance gain of the proposed scheduling policy as compared with the benchmark policies.

APPENDIX A
PROOF OF LEMMA 2

First, we define an auxiliary function as

$$G\left(\mathbf{w}\right) = \frac{\ell}{2}\mathbf{w}^{\top}\mathbf{w} - L\left(\mathbf{w}\right). \tag{36}$$

By taking the second-order partial derivative of $G(\mathbf{w})$ on $\mathbf{w}$, we have

$$\frac{\partial^2 G(\mathbf{w})}{\partial \mathbf{w}^2} = \ell - \frac{\partial^2 L(\mathbf{w})}{\partial \mathbf{w}^2} \overset{(a)}{\geq} 0, \tag{37}$$

where (a) is due to (12). Therefore, $G(\mathbf{w})$ is convex on $\mathbf{w}$, leading to the following sufficient condition as

$$G(\mathbf{u}) \geq G(\mathbf{v}) + \nabla G(\mathbf{v})^\top (\mathbf{u} - \mathbf{v}), \tag{38}$$

which is equivalent to

$$L(\mathbf{u}) \leq L(\mathbf{v}) + \nabla L(\mathbf{v})^\top (\mathbf{u} - \mathbf{v}) + \frac{\ell}{2} \|\mathbf{u} - \mathbf{v}\|^2. \tag{39}$$

On the other hand, the global model update at the edge server follows

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta^t \widehat{\mathbf{g}}^t. \tag{40}$$

Let $\mathbf{u} = \mathbf{w}^{t+1}$ and $\mathbf{v} = \mathbf{w}^t$ such that $\mathbf{u} - \mathbf{v} = -\eta^t \widehat{\mathbf{g}}^t$. Applying it into (39), we have

$$L\left(\mathbf{w}^{t+1}\right) \leq L\left(\mathbf{w}^t\right) + \left(\mathbf{g}^t\right)^\top \left(-\eta^t \widehat{\mathbf{g}}^t\right) + \frac{\ell}{2} \left\|-\eta^t \widehat{\mathbf{g}}^t\right\|^2. \tag{41}$$

Taking expectation in both sides of (41), it follows that

$$\begin{aligned}
\mathbb{E}\left\{L\left(\mathbf{w}^{t+1}\right)\right\} &\leq \mathbb{E}\left\{L\left(\mathbf{w}^t\right)\right\} - \eta^t \left(\mathbf{g}^t\right)^\top \mathbb{E}\left\{\widehat{\mathbf{g}}^t\right\} + \frac{\ell}{2}\left(\eta^t\right)^2 \mathbb{E}\left\{\left\|\widehat{\mathbf{g}}^t\right\|^2\right\} \\
&= \mathbb{E}\left\{L\left(\mathbf{w}^t\right)\right\} - \eta^t \left(\mathbf{g}^t\right)^\top \mathbb{E}\left\{\widehat{\mathbf{g}}^t\right\} + \frac{\ell}{2}\left(\eta^t\right)^2 \left[\left(\mathbb{E}\left\{\widehat{\mathbf{g}}^t\right\}\right)^2 + \mathbb{V}\left\{\widehat{\mathbf{g}}^t\right\}\right] \\
&\overset{(b)}{=} \mathbb{E}\left\{L\left(\mathbf{w}^t\right)\right\} - \eta^t \left(1 - \frac{\ell}{2}\eta^t\right) \left\|\mathbf{g}^t\right\|^2 + \frac{1}{2}\ell\left(\eta^t\right)^2 \mathbb{V}\left\{\widehat{\mathbf{g}}^t\right\},
\end{aligned} \tag{42}$$

where $\mathbb{V}\left\{\widehat{\mathbf{g}}^t\right\} = \mathbb{E}\left\{\left\|\widehat{\mathbf{g}}^t - \mathbf{g}^t\right\|^2\right\}$ is the variance of the global gradient and (b) is because $\widehat{\mathbf{g}}^t$ is an unbiased estimate of the ground-truth global gradient. Then subtracting $\mathbb{E}\left\{L\left(\mathbf{w}^*\right)\right\}$ in both sides of (42), we can obtain the one-round convergence rate in Lemma 2.

APPENDIX B
PROOF OF LEMMA 3

The objective function (23a) is convex because $\frac{1}{x}$ and x are both convex on $(0, +\infty)$. Moreover, the constraints (23b) and (23c) are linear. Thus, $\mathscr{P}_3$ is convex and can be solved by the

KarushKuhnTucker (KKT) conditions [38]. The partial Lagrange function is defined as

$$\mathscr{L} = \sum_{k=1}^{K} a_k \frac{1}{x_k} + b_k x_k + \lambda \left(\sum_{k=1}^{K} c_k x_k - d \right),\tag{43}$$

where $\lambda \geq 0$ is the Lagrange multiplier associated with the constraint (23b). Let $\{x_1^*, x_2^*, \cdots, x_K^*\}$ denote the optimal solution to $\mathscr{P}_3$. Then applying KKT conditions leads to the following necessary and sufficient conditions

$$\frac{\partial \mathscr{L}}{\partial x_k^*} = -\frac{a_k}{(x_k^*)^2} + b_k + \lambda^* c_k = \begin{cases} \geq 0, & x_k^* = 0 \\ = 0, & x_k^* > 0 \end{cases}, \quad \forall k,\tag{44}$$

$$\sum_{k=1}^{K} c_k x_k^* - d = 0.\tag{45}$$

By solving these equations, we can obtain the solution in Lemma 3.

APPENDIX C

PROOF OF THEOREM 2

To prove Theorem 2, we first establish the following lemma.

Lemma 5. The squared norm of the ground-truth global gradient in the t-th communication round satisfies

$$\left\| \mathbf{g}^t \right\|^2 \geq 2\mu \left(L\left(\mathbf{w}^t\right) - L\left(\mathbf{w}^*\right) \right).\tag{46}$$

Proof: Since the global loss function is strongly convex with the parameter μ, it follows

$$L\left(\mathbf{w}^{t+1}\right) \geq L\left(\mathbf{w}^t\right) + \left(\mathbf{g}^t\right)^\top \left(\mathbf{w}^{t+1} - \mathbf{w}^t\right) + \frac{\mu}{2} \left\| \mathbf{w}^{t+1} - \mathbf{w}^t \right\|^2.\tag{47}$$

Minimizing both sides of (47) with respect to $\mathbf{w}^{t+1}$, we have

$$\min_{\mathbf{w}^{t+1}} L\left(\mathbf{w}^{t+1}\right) \geq \min_{\mathbf{w}^{t+1}} \left[L\left(\mathbf{w}^t\right) + \left(\mathbf{g}^t\right)^\top \left(\mathbf{w}^{t+1} - \mathbf{w}^t\right) + \frac{\mu}{2} \left\| \mathbf{w}^{t+1} - \mathbf{w}^t \right\|^2 \right].\tag{48}$$

The minimization of the left-hand side of (48) is achieved when $\mathbf{w}^{t+1} = \mathbf{w}^*$ while the right-hand side of (48) is minimized when $\mathbf{w}^{t+1} = \mathbf{w}^t - \frac{1}{\mu}\mathbf{g}^t$ [39]. Thus, it leads to

$$L\left(\mathbf{w}^*\right) \geq L\left(\mathbf{w}^t\right) - \frac{1}{2\mu} \left\| \mathbf{g}^t \right\|^2,\tag{49}$$

which is equivalent to the desired result in (46).

Based on Lemma 5, we can prove Theorem 2 as follows. Combining (46) with the one-round convergence rate in (42), we can derive that

$$
\begin{aligned}
&\mathbb{E}\left\{L\left(\mathbf{w}^{t+1}\right) - L\left(\mathbf{w}^*\right)\right\} \\
&\leq \mathbb{E}\left\{L\left(\mathbf{w}^t\right) - L\left(\mathbf{w}^*\right)\right\} - 2\mu\eta^t \mathbb{E}\left\{\left(L\left(\mathbf{w}^t\right) - L\left(\mathbf{w}^*\right)\right)\right\} + \frac{\ell}{2}\left(\eta^t\right)^2 \mathbb{E}\left\{\left\|\widehat{\mathbf{g}}^t\right\|^2\right\} \\
&= \left(1 - 2\mu\eta^t\right) \mathbb{E}\left\{L\left(\mathbf{w}^t\right) - L\left(\mathbf{w}^*\right)\right\} + \frac{\ell}{2}\left(\eta^t\right)^2 \mathbb{E}\left\{\left\|\widehat{\mathbf{g}}^t\right\|^2\right\} \\
&\stackrel{\text{recursively}}{\leq} \prod_{i=1}^{t}\left(1 - 2\mu\eta^i\right) \mathbb{E}\left\{L\left(\mathbf{w}^1\right) - L\left(\mathbf{w}^*\right)\right\} + \frac{\ell}{2}\sum_{i=1}^{t} A^i \left(\eta^i\right)^2 \mathbb{E}\left\{\left\|\widehat{\mathbf{g}}^i\right\|^2\right\},
\end{aligned} \tag{50}
$$

where $A^i = \prod_{j=i+1}^{t}\left(1 - 2\mu\eta^j\right)$. Then applying the optimal scheduling policy into $\mathbb{E}\left\{\left\|\widehat{\mathbf{g}}^i\right\|^2\right\}$, we can obtain the desired result in Theorem 2, which ends the proof.

APPENDIX D

PROOF OF COROLLARY 1

We will prove Corollary 1 by induction. First, according to (50), the following inequality holds in each communication round, as

$$
\mathbb{E}\left\{L\left(\mathbf{w}^{t+1}\right) - L\left(\mathbf{w}^*\right)\right\} \leq \left(1 - 2\mu\eta^t\right) \mathbb{E}\left\{L\left(\mathbf{w}^t\right) - L\left(\mathbf{w}^*\right)\right\} + \frac{\ell}{2}\left(\eta^t\right)^2 \mathbb{E}\left\{\left\|\widehat{\mathbf{g}}^t\right\|^2\right\}, \tag{51}
$$

For the diminishing learning rate $\eta^t = \dfrac{\chi}{t + \nu}$ where $\chi > \dfrac{1}{2\mu}$ and $\nu > 0$, we have

- For $t = 1$, it satisfies

$$
\begin{aligned}
\left.\frac{\zeta}{t + \nu}\right|_{t=1} &= \frac{\zeta}{1 + \nu} \\
&\stackrel{(c)}{\geq} \frac{(1 + \nu)\,\mathbb{E}\left\{L\left(\mathbf{w}^1\right) - L\left(\mathbf{w}^*\right)\right\}}{1 + \nu} \\
&= \mathbb{E}\left\{L\left(\mathbf{w}^1\right) - L\left(\mathbf{w}^*\right)\right\},
\end{aligned} \tag{52}
$$

where (c) is derived from the definition of ζ. Therefore, the desired result holds when $t = 1$.

- Assume that the desired result holds for some $t = N \geq 1$, i.e., $\mathbb{E}\left\{L\left(\mathbf{w}^N\right) - L\left(\mathbf{w}^*\right)\right\} \leq \dfrac{\zeta}{N + \nu}$. Define $G = \max\|\widehat{\mathbf{g}}^t\|$ as the upper bound of the global gradient norm. Then when

$t = N + 1$, it follows

$$\mathbb{E}\left\{ L\left(\mathbf{w}^{N+1}\right) - L\left(\mathbf{w}^*\right)\right\}$$

$$\leq \left(1 - 2\mu\eta^N\right)\mathbb{E}\left\{ L\left(\mathbf{w}^N\right) - L\left(\mathbf{w}^*\right)\right\} + \frac{\ell}{2}\left(\eta^N\right)^2\mathbb{E}\left\{\left\|\widehat{\mathbf{g}}^N\right\|^2\right\}$$

$$\leq \left(1 - 2\mu\frac{\chi}{N+\nu}\right)\frac{\zeta}{N+\nu} + \frac{\ell}{2}\left(\frac{\chi}{N+\nu}\right)^2 G^2$$

$$= \frac{\zeta\left(N+\nu-1\right) + \dfrac{1}{2}\ell G^2\chi^2 - \left(2\mu\chi - 1\right)\zeta}{\left(N+\nu\right)^2}$$

$$\leq \frac{\zeta\left(N+\nu-1\right) + \dfrac{1}{2}\ell G^2\chi^2 - \left(2\mu\chi - 1\right)\dfrac{\ell G^2\chi^2}{2\left(2\mu\chi - 1\right)}}{\left(N+\nu\right)^2}$$

$$= \frac{\zeta\left(N+\nu-1\right)}{\left(N+\nu\right)^2}$$

$$\leq \frac{\zeta}{N+1+\nu}. \tag{53}$$

Hence, the desired result holds for $t = N + 1$ if it holds for $t = N$. Towards this end, we can conclude that $\mathbb{E}\left\{ L\left(\mathbf{w}^T\right) - L\left(\mathbf{w}^*\right)\right\} \leq \dfrac{\zeta}{T+\nu}$ holds for $\forall T \in \mathbb{Z}^+$, which completes the proof.

References

[1] Q. Mao, F. Hu, and Q. Hao, "Deep learning for intelligent wireless networks: A comprehensive survey," *IEEE Commun. Surv. Tut.*, vol. 20, no. 4, pp. 2595-2621, Fourthquarter 2018.

[2] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *to appear in IEEE Network,* 2019.

[3] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surv. Tut.*, vol. 21, no. 4, pp. 3039-3071, Fourthquarter 2019.

[4] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Commun. Surv. Tut.*, vol. 21, no. 3, pp. 2224-2287, Thirdquarter 2019.

[5] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning," *IEEE Network*, vol. 33, no. 5, pp. 156-165, Sep. 2019.

[6] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proceedings of the IEEE,* vol. 107, no. 11, pp. 2204-2239, Nov. 2019.

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

[8] 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. Int. Conf. Artif. Intell. Statist. (AISTATS)*, Fort Lauderdale, FL, USA, Apr. 2017, pp. 1-10.

[9]   T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *Proc. IEEE Int. Conf. Commun. (ICC),* Shanghai, China, May 2019, pp. 17.

[10] J. Ren, G. Yu, and G. Ding, "Accelerating DNN training in wireless federated edge learning system," *arXiv preprint arXiv:1905.09712*, 2019.

[11] M. S. H. Abad, E. Ozfatura, D. Gündüz, and O. Ercetin, "Hierarchical federated learning across heterogeneous cellular networks," *arXiv preprint arXiv:1909.02362*, 2019.

[12] Z. Zhao, C. Feng, H. H. Yang, and X. Luo, "Federated-learning-enabled intelligent fog radio access networks: Fundamental theory, key techniques, and future trends," *IEEE Wireless Commun. Mag.*, vol. 27, no. 2, pp. 22-28, Apr. 2020.

[13] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2726-2737, Nov. 2005.

[14] M. Ergen, S. Coleri, and P. Varaiya, "QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems," *IEEE Trans. Broadcast.*, vol. 49, no. 4, pp. 362-370, Dec. 2003.

[15] D. Wen, X. Li, Q. Zeng, J. Ren, and K. Huang, "An overview of data-importance aware radio resource management for edge machine learning", *J. Commun. Inform. Netw,* vol. 4, no. 4, pp. 1-14, Dec. 2019.

[16] 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. (NeurIPS)*, Montreal, QC, Canada, Dec. 2018, pp. 5050-5060.

[17] H. H. Yang, A. Arafa, T. Q. S. Quek, and H. V. Poor, "Age-based scheduling policy for federated learning in mobile edge networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 8743-8747.

[18] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *Proc. of NeurIPS Workshop on Private Multi-Party Machine Learning*, Barcelona, Spain, Dec. 2016, pp. 1-5.

[19] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Sparse binary compression: Towards distributed deep learning with minimal communication," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, Budapest, Hungary, Jul. 2019, pp. 1-8.

[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. Joint Conf. Neural Networks (IJCNN)*, Budapest, Hungary, Jul. 2019, pp. 1-8.

[21] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.,* vol. 37, no. 6, pp. 1205-1221, Jun. 2019.

[22] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "Performance optimization of federated learning over wireless networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Waikoloa, HI, USA, 2019, pp. 1-6.

[23] M. Manimel Wadu, S. Samarakoon, and M. Bennis, "Federated learning under channel uncertainty: Joint client scheduling and resource allocation," *arXiv preprint arXiv:2002.00802*, 2020.

[24] M. Chen, H. V. Poor, W. Saad, and S Cui, "Convergence time optimization for federated learning over wireless networks," *arXiv preprint arXiv:2001.07845*, 2020.

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

[26] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *arXiv preprint arXiv:1911.02417*, 2019.

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

[28] 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.

[29] G. Zhu, Y. Du, D. Gündüz, K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *arXiv preprint arXiv:2001.05713*, 2020.

[30] P. Zhao and T. Zhang, "Stochastic optimization with importance sampling for regularized loss minimization," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, Jul. 2015, pp. 19.

[31] D. Liu, G. Zhu, J. Zhang, and K. Huang, "Wireless data acquisition for edge learning: Importance-aware retransmission," in *Proc. IEEE Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, Cannes, France, Jul. 2019, pp. 15.

[32] D. Liu, G. Zhu, J. Zhang, and K. Huang, "Data-importance aware user scheduling for communication-efficient edge machine learning," *arXiv preprint arXiv:1910.02214*, 2019.

[33] 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.

[34] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Update aware device scheduling for federated learning at the wireless edge," *arXiv preprint arXiv:2001.10402*, 2020.

[35] K.Bonawitz, *et al.* "Towards federated learning at scale: System design," *arXiv preprint arXiv:1902.01046*, 2019.

[36] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," *arXiv preprint arXiv:1907.02189*, 2019.

[37] *Radio Frequency (RF) System Scenarios*, document 3GPP TR 25.942, v.14.0.0, 2017.

[38] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[39] M. P. Friedlander and M. Schmidt, "Hybrid deterministic-stochastic methods for data fitting," *SIAM J. Sci. Comput,* vol. 34, no. 3, pp. A1380-A1405, May 2012.