# Robust Federated Learning for Unreliable and Resource-limited Wireless Networks

Zhixiong Chen, *Student Member, IEEE*, Wenqiang Yi, *Member, IEEE*,
Yuanwei Liu, *Senior Member, IEEE*, and Arumugam Nallanathan, *Fellow, IEEE*

***Abstract*—Federated learning (FL) is an efficient and privacy-preserving distributed learning paradigm that enables massive edge devices to train machine learning models collaboratively. Although various communication schemes have been proposed to expedite the FL process in resource-limited wireless networks, the unreliable nature of wireless channels was less explored. In this work, we propose a novel FL framework, namely FL with gradient recycling (FL-GR), which recycles the historical gradients of unscheduled and transmission-failure devices to improve the learning performance of FL. To reduce the hardware requirements for implementing FL-GR in the practical network, we develop a memory-friendly FL-GR that is equivalent to FL-GR but requires low memory of the edge server. We then theoretically analyze how the wireless network parameters affect the convergence bound of FL-GR, revealing that minimizing the average square of local gradients' staleness (AS-GS) helps improve the learning performance. Based on this, we formulate a joint device scheduling, resource allocation and power control optimization problem to minimize the AS-GS for global loss minimization. To solve the problem, we first derive the optimal power control policy for devices and transform the AS-GS minimization problem into a bipartite graph matching problem. Through detailed analysis, we further transform the bipartite matching problem into an equivalent linear program which is convenient to solve. Extensive simulation results on three real-world datasets (i.e., MNIST, CIFAR-10, and CIFAR-100) verified the efficacy of the proposed methods. Compared to the FL algorithms without gradient recycling, FL-GR is able to achieve higher accuracy and fast convergence speed. In addition, the proposed device scheduling and resource allocation algorithm also outperforms the benchmarks in accuracy and convergence speed.**

***Index Terms*—Device scheduling, federated Learning, resource allocation, unreliable transmission**

## I. Introduction

To protect data privacy for wireless devices while jointly training machine learning models, federated learning (FL) has become a promising solution at the wireless mobile edge in 6G, which can realize collaborative learning among devices without revealing the original data [2]. However, implementing FL in practical wireless networks suffers from the limited wireless resource [3], which restricts the participating device number in the per-round learning process. Since the local datasets among devices are typically non-independent and identically distributed (non-IID), the limited participating devices may lead to biased model aggregation and greatly degrade the learning performance [4]. In addition, most FL algorithms assume an error-free wireless channel and ignore the unreliable nature of wireless communications [5]. Due to devices' constrained transmit power and bandwidth, it is hard to guarantee all the scheduled devices successfully transmit their parameters to the edge server [6]. This brings a new challenge for FL to enhance the robustness of the training process and mitigate the impact of erroneous transmission. An intuitive solution [7] is to discard the devices' parameter with errors, but it further reduces the participating device number and exacerbates the performance loss of FL. Thus, it is essential to develop innovative approaches for FL to address the scarcity of radio resources and the unreliability of wireless transmissions.

Zhixiong Chen, Wenqiang Yi, Yuanwei Liu, and Arumugam Nallanathan are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K. (emails: {zhixiong.chen, w.yi yuanwei.liu, a.nallanathan}@qmul.ac.uk).

Part of this work has been presented in International Wireless Communications and Mobile Computing (IWCMC), 2023 [1].

### A. Related Works

In wireless networks, FL is generally implemented by multiple devices coordinated by an edge server, in which devices are usually resource-limited in terms of wireless bandwidth, computing capability, and battery capacity. Thus, it is important to carefully design the device scheduling and wireless resource management policies that maximize the learning performance of FL. The existing device scheduling approaches mainly focused on selecting devices with the best channel condition [8], [9], most important parameters [10], [11], or both of them [12], [13] to accelerate the learning process. Specifically, the channel-aware device selection approach in [8] can maximize the sum of scheduled data samples under devices' long-term communication energy constraints. Based on the communication time analysis of FL, an optimal probabilistic scheduling policy has been proposed in [9] to reduce the training latency. By measuring the significance of devices with their gradient norm, a parameter importance-aware user selection scheme has been developed in [10] to minimize the convergence time of FL. In [11], prioritizing devices with rich and diverse datasets in the device scheduling policy has achieved higher accuracy and lower learning costs than random device scheduling. The joint channel condition and local model significance-aware device scheduling policy in [12] can provide better performance than scheduling policies based only on one of the two metrics, in which the norm of model update measures the model significance. By measuring the clients' potential contributions with the information entropy of their gradients, the joint channel and contribution-aware scheduling algorithm in [13] significantly improve the model accuracy and convergence

speed of FL. In addition to the device scheduling policy, wireless resource allocation is crucial to improve energy efficiency and reduce learning latency in FL. The joint time allocation, power control, and computation frequency scaling approach in [14] can substantially reduce the energy consumption of FL while satisfying latency requirements. In [15], a multi-dimensional control policy, including bandwidth allocation and workload partitioning, has been studied to improve the energy efficiency of FL. A joint communication and computation resource allocation scheme has been proposed in [16] to capture the trade-offs among convergence, wallclock time, and energy consumption. Besides the above device scheduling approaches, the over-the-air-based FL approaches in [17], [18] effectively address the scalability issues for limited wireless resources. Despite the effectiveness of the above approaches, they assumed that the parameter server always successfully receives the local models/gradients of all the scheduled devices and did not consider the unreliability of wireless transmissions.

To cope with unreliable communications between devices and the edge server, existing works focused on improving the successful transmission probabilities of devices by resource management [19]–[22] and retransmission mechanism design [23], [24], as well as compensate for the unsuccessful received devices' models by the past models [25], [26]. The device scheduling and resource allocation algorithm in [19] can maximize the expected number of devices with successful transmissions. A joint wireless resources and quantization bits allocation scheme has been developed in [20] to alleviate the effects of quantization errors and transmission outages on FL convergence performance. The joint device selection and resource allocation approach in [21] can effectively increase the successful information exchange probabilities over wireless networks and thus improve the learning performance of FL. The power allocation and gradient quantization scheme in [22] can improve the convergence speed of FL over a noisy wireless network. However, it only schedules a single device per iteration based on the channel condition. The retransmission protocol in [23] significantly increases the success probability of devices' model uploading, in which devices transmit their local model parameters multiple times, and the edge server uses the received signal with the highest signal-to-interference-plus-noise ratio (SINR) to recover the local models. Different from [23], the retransmission mechanism in [24] utilized the arithmetic mean of the received multiple-times signals from devices to update the global model, effectively reducing global model aggregation errors induced by channel fading in over-the-air FL. While demonstrably effective, the above approaches that maximize devices' successful transmission probabilities only aggregate the successfully uploaded devices' models and thus reduce the number of participants. In addition, the retransmission approaches may cause additional latency and energy consumption for FL. In the presence of decentralized FL systems, by reusing past local models, the robust decentralized stochastic gradient descent (SGD) approach proposed in [25] under transmission error situations can achieve the same asymptotic convergence rate as the vanilla decentralized SGD with perfect communications. It has been proved in [26] that the federated averaging (FedAvg) algorithm replacing error models with past local models in case of devices' model uploading error converges to the same global model parameters as the perfect FedAvg (without communication errors). However, the approaches in [25], [26] assumed that all devices participate in the per-round learning process and did not consider the design of wireless networks.

### *B. Motivations and Contributions*

Although the device scheduling schemes in [8]–[15] effectively cope with limited wireless resources, they have assumed ideal wireless channels with reliable and lossless transmissions between devices and the edge server, which may not always hold in practical wireless networks. The transmission designs for tackling unreliable channels in [19]–[24] remain to drop out the unsuccessful-transmission devices and reduce the number of participants of FL. In addition, the compensation approaches in [25], [26] did not consider wireless network design. They assumed all devices participate in the per-round learning process, which may be incompatible with the limited wireless resources. To mitigate the adverse impact of unreliable wireless channels and limited resources on FL, this work aims to jointly design the wireless network and learning mechanism to enhance the robustness of the training process and improve the learning performance of FL. Inspired by the success of using stale model parameters to accelerate the training process in asynchronous FL [27], we propose a novel FL framework, i.e., FL with gradient recycling (FL-GR), which recycles the latest historical local gradients received at the edge server to update the global model in each round. Note that unlike [25], [26] that utilize the past local models to replace the transmission-failure devices' model for global aggregation, this work recycles devices' gradients to update the global model and achieves a better learning performance which verified in our simulations. In addition, we investigate the effect of partial device participation and the staleness of local gradients on the convergence bound. It is worth mentioning that although FL-GR recycles the historical local gradients to update the global model, it differs from the asynchronous FL [27]. The asynchronous FL broadcasts the global model to all devices at the beginning of FL, while FL-GR only broadcasts the global model to the scheduled devices. In addition, the asynchronous FL updates the global model when receiving $H$ local models of total $K$ devices ($H < K$) that may be stale since they were updated at an older version of the global model, while FL-GR is a synchronous FL scheme and ensures the received local gradients are timely and not stale. Thus, the staleness of the devices' local gradients in asynchronous FL is larger than that in FL-GR in each round. Consequently, FL-GR achieves lower convergence error than asynchronous FL and obtains better learning performance [28]. The main contributions of this work are summarized as follows:

- To cope with limited resources and unreliable channels in wireless networks, we propose a novel FL framework, i.e., FL-GR, which recycles the historical gradients of unscheduled and transmission-failure devices for global model updates. This framework can achieve faster convergence speed and higher accuracy than the conventional

FL that only aggregates the successfully received local models. In addition, we formulate a joint device scheduling, resource block (RB) allocation, and power control problem to minimize the global loss, in which training latency and devices' energy consumption are considered.
- For the convenience of implementation in practical wireless networks, we propose a memory-friendly FL-GR that is equivalent to FL-GR, but with low memory space requirement of the edge server. Then, we theoretically analyze how the wireless network parameters affect the convergence bound of FL-GR. Based on the convergence bound, we define a new objective function, i.e., the average staleness of local gradients, and transform the global loss minimization problem into an explicit one for device scheduling, RB allocation, and power control.
- To solve the transformed problem, we first find the devices' optimal transmit power control policy under any given RB allocation policy. Then, we transform the original global loss minimization problem into a perfect bipartite matching problem. Through detailed analysis, we further transform the bipartite matching problem into equivalent linear programming whose optimal solution can be effectively solved with polynomial time complexity.
- We provide extensive experimental results on real-world datasets (i.e., MNIST, CIFAR-10, and CIFAR-100) with a typical non-IID setting to demonstrate the effectiveness of the proposed FL-GR and device scheduling algorithm. Compared to the FL algorithm without gradient recycling, FL-GR achieves higher learning accuracy and faster convergence speed. In addition, the proposed device scheduling algorithm outperforms the benchmarks in convergence speed and test accuracy.

### C. Organization and Notations

The rest of this paper is organized as follows: In Section II, we introduce the proposed FL-GR and system model, then formulate a global loss minimization problem. A memory-friendly implementation of FL-GR and the convergence analysis are illustrated in Section III. Section IV illustrates the proposed device scheduling, RB allocation, and power control algorithm that solves the global loss minimization problem. Section V verifies the effectiveness of FL-GR and the proposed device scheduling algorithm by simulations. The conclusion is drawn in Section VI. The main notations used in this paper are summarized in Table I.

TABLE I
NOTATION SUMMARY

| Notation | Definition |
|---|---|
| $\mathcal{K}$; $K$; | Set of devices; size of $\mathcal{K}$ |
| $\mathcal{M}$; $M$; | Set of resource blocks(RB); size of $\mathcal{M}$ |
| $\mathcal{D}_k$; $D_k$ | Local dataset of device $k$; size of $\mathcal{D}_k$ |
| $\mathcal{D}$; $D$ | Overall dataset in the system; size of $\mathcal{D}$ |
| $\boldsymbol{w}_{k,t}^{(l)}$; $\boldsymbol{w}_t$ | Local model of device $k$ in the $l$-th iteration of round $t$; global model in round $t$ |
| $F_k(\boldsymbol{w})$; $F(\boldsymbol{w})$ | Local loss function of device $k$; global loss function |
| $\eta$; $\lambda$ : | Learning rate; local iteration number |
| $f_k$; | CPU frequency of device $k$; |
| $\boldsymbol{G}_{k,t}$; | Latest successfully transmitted gradient of device $k$ in round $t$; |
| $p_{k,t}$; $p_{k,\max}$; | Transmit power of device $k$ in round $t$; maximum transmit power of device $k$; |
| $\tilde{\boldsymbol{g}}_{k,t}$ | Stochastic gradient of device $k$ in round $t$ |
| $\boldsymbol{p}_t$ | Power control policy of devices in round $t$ |
| $C_k$; $Q$ | Computation workload of one data sample at device $k$; data size of local gradient |
| $E_{k,\max}$; $\mathcal{T}_{\max}$ | Energy constraint of device $k$; maximum completion time for each round |
| $z_{k,t}^{(m)}$; $\boldsymbol{Z}_t$ | Allocation indicator of RB $m$ to device $k$ in round $t$; RB allocation policy for all devices |
| $\alpha_{k,t}$; $\boldsymbol{\alpha}_t$ | Scheduling indicator of device $k$ in round $t$; device scheduling vector in round $t$ |

## II. System Model and Learning Mechanism

In this work, we investigate an FL system under a noisy and resource-limited wireless network, where the unreliable property of the wireless uplinks, i.e., transmission error, is considered. To tackle the transmission error effect on FL performance, we propose a new FL framework in which the edge server recycles the historical latest received gradients of unscheduled and transmission-failure devices to accelerate the learning process. In addition, we characterize the learning costs of the proposed FL framework and formulate an optimization problem to minimize the global loss function.

### A. Federated Learning System

The considered FL system consisting of one edge server and $K$ devices indexed by $\mathcal{K} = \{1, 2, \cdots, K\}$. Each device $k$ ($k \in \mathcal{K}$) has a local dataset $\mathcal{D}_k$ with $D_k = |\mathcal{D}_k|$ data samples. Without loss of generality, we assume that there is no overlapping between local datasets from different devices, i.e., $\mathcal{D}_k \cap \mathcal{D}_h = \emptyset$ ($\forall k, h \in \mathcal{K}$). Thus, the entire dataset is denoted by $\mathcal{D} = \cup\{\mathcal{D}_k\}_{k=1}^{K}$ with a total number of samples $D = \sum_{k=1}^{K} D_k$. Given a data sample $(\boldsymbol{x}, y) \in \mathcal{D}$, where $\boldsymbol{x} \in \mathbb{R}^d$ is the $d$-dimensional input data vector, $y \in \mathbb{R}$ is the corresponding ground-truth label. Let $f(\boldsymbol{x}, y; \boldsymbol{w})$ denote the sample-wise loss function, which captures the error of the model parameter $\boldsymbol{w}$ on the input-output data pair $(\boldsymbol{x}, y)$. Thus, the local loss function of device $k$ that measures the model error on its local dataset is given by

$$F_k(\boldsymbol{w}) = \frac{1}{D_k} \sum\nolimits_{(\boldsymbol{x},y)\in\mathcal{D}_k} f(\boldsymbol{x}, y; \boldsymbol{w}). \tag{1}$$

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

$$F(\boldsymbol{w}) = \sum\nolimits_{k=1}^{K} p_k F_k(\boldsymbol{w}), \tag{2}$$

where $p_k$ is the weight of device $k$ such that $p_k \geq 0$ and $\sum_{k=1}^{K} p_k = 1$. Similar to many existing works, e.g., [9] and [15], we consider a balance size for local datasets and set $p_k = \frac{1}{K}, \forall k \in \mathcal{K}$.

The objective of the FL system is to train a global model, $\boldsymbol{w}$, so as to minimize the global loss, $F(\boldsymbol{w})$, on the whole dataset, $\mathcal{D}$. The optimization objective of FL can be expressed as $\min\limits_{\boldsymbol{w}} F(\boldsymbol{w})$.

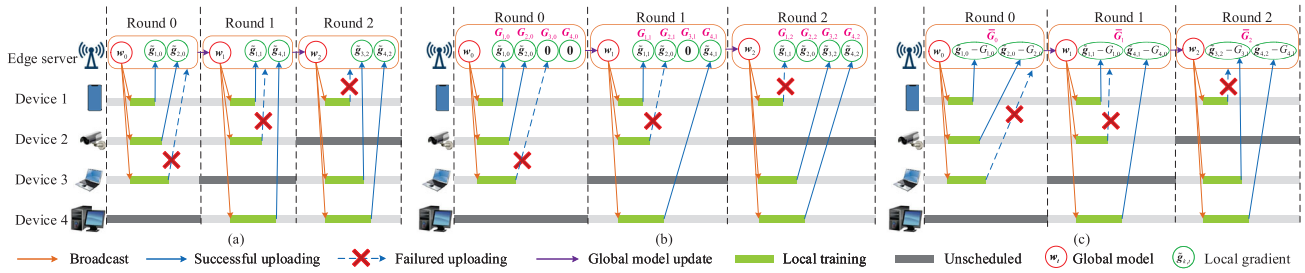



Fig. 1. Illustration of conventional FL framework and the proposed FL-GR: (a) Conventional FL only uses the current successfully received gradients from scheduled devices to update the global model. (b) The proposed FL-GR recycles the latest historical successful transmitted gradients of unscheduled and transmission-failure devices for the global model update. (c) Memory-friendly FL-GR.

### *B. FL with Gradient Recycling*

To address the unreliable transmissions and limited resources in the FL system, we propose a new FL framework, namely FL with gradient recycling (FL-GR), in which the edge server maintains a *gradient array* $\{\boldsymbol{G}_{k,t} : \forall k \in \mathcal{K}\}$ that caches the latest successfully received gradients for all devices and uses them for the global model update. Note that, at the beginning of The FL process, the gradient array is initialized as $\boldsymbol{G}_{k,t} = \mathbf{0}$ $(\forall k \in \mathcal{K})$. The learning process consists of $T$ global rounds and performs the following steps in each round $t$ $(t \in \{0, 1, \cdots, T-1\}$.

- **Step 1 (Global model broadcast)**: The edge server selects a subset of devices to participate in the current round training process and then broadcasts the latest global model, $\boldsymbol{w}_t$, to the selected devices. Let $\alpha_{k,t} \in \{0, 1\}$ to denote the scheduling indicator of device $k$, where $\alpha_{k,t} = 1$ indicates that device $k$ is scheduled in round $t$, $\alpha_{k,t} = 0$ otherwise. We use $\boldsymbol{\alpha}_t = \{\alpha_{1,t}, \alpha_{2,t}, \cdots, \alpha_{K,t}\}$ to represent the device scheduling decision in round $t$.
- **Step 2 (Local model training)**: After receiving the global model from the edge server, each scheduled device updates its local model by running $\lambda$ steps stochastic gradient descent (SGD) on its local dataset, according to

$$\boldsymbol{w}_{k,t}^{(l+1)} = \boldsymbol{w}_{k,t}^{(l)} - \eta \tilde{\nabla} F_k(\boldsymbol{w}_{k,t}^{(l)}), \forall l = 0, \cdots, \lambda - 1, \quad (3)$$

where $\boldsymbol{w}_{k,t}^{(l)}$ is the local model of device $k$ in the $l$-th local iteration in round $t$ with $\boldsymbol{w}_{k,t}^{(0)} = \boldsymbol{w}_t$, and $\eta > 0$ is the learning rate. In (3), the stochastic gradient $\tilde{\nabla} F_k(\boldsymbol{w}_{k,t}^{(l)})$ is given by

$$\tilde{\nabla} F_k(\boldsymbol{w}_{k,t}^{(l)}) = \frac{1}{L_b} \sum\nolimits_{(\boldsymbol{x},y)\in\mathcal{B}_{k,t}^{(l)}} \nabla f(\boldsymbol{x}, y; \boldsymbol{w}_{k,t}^{(l)}), \quad (4)$$

where $\mathcal{B}_{k,t}^{(l)}$ is a mini-batch data uniformly sampled from $\mathcal{D}_k$ with $L_b = |\mathcal{B}_{k,t}^{(l)}|$ data samples.
- **Step 3 (Local gradient uploading)**: After accomplishing local model training, each scheduled device $k$ $(k \in \mathcal{K})$ uploads its cumulative local stochastic gradient $\tilde{\boldsymbol{g}}_{k,t}$ to the edge server. $\tilde{\boldsymbol{g}}_{k,t}$ is given by

$$\tilde{\boldsymbol{g}}_{k,t} = \sum\nolimits_{l=0}^{\lambda-1} \tilde{\nabla} F_k(\boldsymbol{w}_{k,t}^{(l)}) = \frac{1}{\eta}\left(\boldsymbol{w}_t - \boldsymbol{w}_{k,t}^{(\lambda)}\right). \quad (5)$$

Due to the unreliable wireless channels, the local gradient may not be successfully transmitted to the edge server. Let $s_{k,t} \in \{0, 1\}$ denote the successful transmission indicator of device $k$ in round $t$, where $s_{k,t} = 1$ represents the uploaded information of device $k$ is successfully received at the edge server, $s_{k,t} = 0$ otherwise.

- **Step 4 (Global model update)**: After the edge server receives the local gradients from the scheduled devices, the edge server updates the gradient array as

$$\boldsymbol{G}_{k,t} = \begin{cases} \tilde{\boldsymbol{g}}_{k,t}, & \text{if } \alpha_{k,t}s_{k,t} = 1, \\ \boldsymbol{G}_{k,t-1}, & \text{otherwise}, \end{cases} \quad \forall k \in \mathcal{K}. \quad (6)$$

In (6), the edge server only refreshes the scheduled and successfully transmitted devices' gradient and maintains the latest historical successfully received gradients for unscheduled or transmission-failure devices. Then, the edge server updates the global model as

$$\begin{aligned} \boldsymbol{w}_{t+1} &= \boldsymbol{w}_t - \eta \frac{1}{K} \sum\nolimits_{k=1}^{K} \boldsymbol{G}_{k,t} \\ &= \boldsymbol{w}_t - \eta \frac{1}{K} \sum_{k=1}^{K} \left(\alpha_{k,t}s_{k,t}\tilde{\boldsymbol{g}}_{k,t} + (1-\alpha_{k,t}s_{k,t})\boldsymbol{G}_{k,t-1}\right). \end{aligned} \quad (7)$$

Note that, in (7), the edge server utilizes the successfully received gradient from scheduled devices in the current round and the historical latest received gradients of unscheduled or transmission-failure devices to update the global model. This differs from the existing works in [19]–[24] that only aggregate the scheduled and successfully transmitted devices' gradient to update the global model, i.e., $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \frac{\sum_{k=1}^{K} \alpha_{k,t}s_{k,t}\tilde{\boldsymbol{g}}_{k,t}}{\sum_{k=1}^{K} \alpha_{k,t}s_{k,t}}$.

For the proposed FL-GR, we have the following remark:

*Remark* 1. The recycling of historical local gradients in FL-GR has lower model aggregation error than the approaches in [25], [26] that reusing of historical models. For ease of comparison, we define the perfect updated global model based on all devices' local models as $\boldsymbol{w}_{t+1}^{*} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{w}_{k,t+1} = \boldsymbol{w}_t - \eta \frac{1}{K}\sum_{k=1}^{K} \tilde{\boldsymbol{g}}_{k,t}$. Note that, in [25], [26], the updated global model in round $(t+1)$ is $\boldsymbol{w}_{t+1}^{\mathrm{m}} = \frac{1}{K}\sum_{k=1}^{K}(\alpha_{k,t}s_{k,t}\boldsymbol{w}_{k,t+1} + (1-\alpha_{k,t}s_{k,t})\boldsymbol{w}_{k,t-\tau_{k,t}+1})$, where $\tau_{k,t}$ is the interval between the current round $t$ and the last round that device $k$ received global model. Thus, the aggregation model error of reusing local models in [25], [26] is given by

$$\begin{aligned} \Delta_m &= \left\|\boldsymbol{w}_{t+1}^{*} - \boldsymbol{w}_{t+1}^{\mathrm{m}}\right\|^2 \\ &= \left\| \frac{1}{K} \sum\nolimits_{k=1}^{K} (1 - \alpha_{k,t}s_{k,t}) \left(\boldsymbol{w}_{k,t+1} - \boldsymbol{w}_{k,t-\tau_{k,t}+1}\right) \right\|^2 \\ &= \left\| \frac{1}{K} \sum\nolimits_{k=1}^{K} (1 - \alpha_{k,t}s_{k,t})(\boldsymbol{w}_t - \eta\tilde{\boldsymbol{g}}_{k,t} \right. \end{aligned}$$

$$- \boldsymbol{w}_{k,t-\tau_{k,t}} + \eta \tilde{\boldsymbol{g}}_{k,t-\tau_{k,t}}) \Big\|^2.$$

The aggregation model error of reusing gradients is given by

$$\Delta_g = \Big\| \eta \frac{1}{K} \sum\nolimits_{k=1}^{K} (1 - \alpha_{k,t} s_{k,t}) \left( \tilde{\boldsymbol{g}}_{k,t-\tau_{k,t}} - \tilde{\boldsymbol{g}}_{k,t} \right) \Big\|^2$$

Based on the triangle inequality, we have $\Delta_m \geq \Delta_g$, i.e., the proposed approach that recycles historical local gradients has a smaller model aggregation error than the approaches in [25], [26] that reuse past local models. Thus, the proposed FL-GR outperforms the approaches in [25], [26], which is also verified in our simulations in Section V.

It is worth mentioning that without gradient recycling, our FL-GR degrades to the FedAvg algorithm [29]. For illustrating this, we rearrange the model update rule in (7) as

$$\begin{aligned} \boldsymbol{w}_{t+1} &= \boldsymbol{w}_t - \eta \frac{1}{K} \sum\nolimits_{k=1}^{K} \boldsymbol{G}_{k,t} = \frac{1}{K} \sum\nolimits_{k=1}^{K} (\boldsymbol{w}_t - \eta \boldsymbol{G}_{k,t}) \\ &= \frac{1}{K} \Big( \sum\nolimits_{k=1}^{K} \alpha_{k,t} s_{k,t} (\boldsymbol{w}_t - \eta \boldsymbol{G}_{k,t}) \\ &+ \sum\nolimits_{k=1}^{K} (1 - \alpha_{k,t} s_{k,t}) (\boldsymbol{w}_t - \eta \boldsymbol{G}_{k,t}) \Big) \\ &\overset{(a)}{=} \frac{1}{K} \Big( \sum_{k=1}^{K} \alpha_{k,t} s_{k,t} \boldsymbol{w}_{k,t+1} + \sum_{k=1}^{K} (1 - \alpha_{k,t} s_{k,t}) \boldsymbol{w}_t \Big), \end{aligned} \tag{8}$$

where (a) is due to without gradient recycling, the unsuccessful participating devices' gradients are $\mathbf{0}$. From (8), without gradient recycling, the global model is updated as averaging all devices' models, which includes the successful participating devices' updated models and the unsuccessful participating devices' models that are replaced with the current global model. Thus, without gradient recycling, FL-GR will degrade to FedAvg.

To better explain FL-GR, we illustrate the conventional FL framework and FL-GR in Fig. 1. Assume that one edge server and four devices are in the system to perform three rounds of the FL process. At the beginning of the FL, the edge server initials the global model $\boldsymbol{w}_0$ and the gradient array for all devices to $\mathbf{0}$. Take round 3 as an example, in which devices 1, 3, and 4 are scheduled to participate in the learning process, and device 1 cannot successfully transmit its gradient. The conventional FL in Fig. 1(a) only aggregates the successfully transmitted devices' gradients ($\tilde{\boldsymbol{g}}_{3,2}$ and $\tilde{\boldsymbol{g}}_{4,2}$) to update the global model, i.e., $\boldsymbol{w}_3 = \boldsymbol{w}_2 - \frac{1}{2}\eta(\tilde{\boldsymbol{g}}_{3,2} + \tilde{\boldsymbol{g}}_{4,2})$. However, the FL-GR in Fig. 1(b) utilizes the successfully received gradients ($\tilde{\boldsymbol{g}}_{3,2}$ and $\tilde{\boldsymbol{g}}_{4,2}$) and the historical received gradient of unscheduled or transmission failure devices ($\boldsymbol{G}_{1,2} = \tilde{\boldsymbol{g}}_{1,1}$ and $\boldsymbol{G}_{2,2} = \tilde{\boldsymbol{g}}_{2,0}$) to update the global model, i.e., $\boldsymbol{w}_2 = \boldsymbol{w}_1 - \frac{1}{4}\eta(\tilde{\boldsymbol{g}}_{1,1} + \tilde{\boldsymbol{g}}_{2,0} + \tilde{\boldsymbol{g}}_{3,2} + \tilde{\boldsymbol{g}}_{4,2})$.

### C. Computation model

Let $C_k$ denotes the number of CPU cycles required for device $k$ ($k \in \mathcal{K}$) to process one data sample, which can be measured offline as a priori knowledge. Let $f_k$ represents the computation capability (CPU cycles per second) of device $k$. Thus, the computational time of local training is given by

$$T_{k,t}^{C} = \frac{\lambda L_b C_k}{f_k}. \tag{9}$$

The corresponding energy consumption of device $k$ is

$$E_{k,t}^{C} = \kappa \lambda L_b C_k (f_k)^2, \tag{10}$$

where $\kappa$ is the energy coefficient of devices, which depends on the chip architecture.

Note that we have ignored the computation cost of global model update at the edge server and focused on resource-limited edge devices since the edge server usually has strong computation capabilities and is supplied by the grid power.

### D. Communication Model

In this work, we consider the orthogonal frequency division multiple access (OFDMA) with $M$ RBs indexed by $\mathcal{M} = \{1, 2, \cdots, M\}$ for devices to upload their local gradients. Each device can occupy one uplink RB in a communication round to upload its local gradient. Let $\boldsymbol{z}_{k,t} = (z_{k,t}^{(1)}, z_{k,t}^{(2)}, \cdots, z_{k,t}^{(M)})$ denote the RB allocation vector for device $k$ in round $t$, where $z_{k,t}^{(m)} \in \{0, 1\}$, $z_{k,t}^{(m)} = 1$ indicates that the $m$-th resource block is allocated to device $k$, and $z_{k,t}^{(m)} = 0$ otherwise. For ease of representation, we use $\boldsymbol{Z}_t = (\boldsymbol{z}_{1,t}, \boldsymbol{z}_{2,t}, \cdots, \boldsymbol{z}_{K,t})$ denote the RB allocation decision for all devices in round $t$. Denote $p_{k,t}$ as the transmit power of device $k$ in round $t$, its maximum value is $p_{k,\max}$. The channel gain from device $k$ to the edge server is modelled as $h_{k,t} = \rho_k(t) d_k^{-v}$, where $\rho_k(t)$ is the small-scale fading gain between device $k$ and the edge server, $d_k$ is the distance between device $k$ and the edge server, and $v$ being the path loss exponent. We consider Rayleigh fading, i.e., $\rho_k(t) \sim \exp(1)$, and it is independent and identically distributed across devices and rounds. Thus, the achievable transmit rate of device $k$ in round $t$ is

$$r_{k,t}(\boldsymbol{z}_{k,t}, p_{k,t}) = \sum_{m=1}^{M} z_{k,t}^{(m)} B \log_2 \left( 1 + \frac{p_{k,t} h_{k,t}}{I_m + B N_0} \right), \tag{11}$$

where $B$ is the bandwidth of each resource block, $N_0$ is the noise power spectral density, $I_m$ is the interference caused by the devices that are located in other service areas and use the same resource block [19]. It is noted that each device can only occupy at most one resource block, and each resource block can be accessed by at most one device. Thus, the RB allocation policy for devices should satisfy $\sum_{m=1}^{M} z_{k,t}^{(m)} \leq 1$ and $\sum_{k=1}^{K} z_{k,t}^{(m)} \leq 1$.

Let $Q$ denote the size of each gradient, i.e., the number of bits used to quantify the gradients. If device $k$ is scheduled to participate in the training process of round $t$, its transmission time is given by

$$T_{k,t}^{\mathrm{U}} = \frac{Q}{r_{k,t}(\boldsymbol{z}_{k,t}, p_{k,t})}. \tag{12}$$

The corresponding energy consumption of device $k$ for transmission is

$$E_{k,t}^{\mathrm{U}} = p_{k,t} T_{k,t}^{\mathrm{U}}. \tag{13}$$

### E. Successful Transmission Probability

In this work, we consider characterizing the unreliability of uplink transmissions of devices by the successful transmission

probability. Before studying the uplink success probability, we assume the downlink transmission is always successful, i.e., devices successfully receive the global model. It is worth mentioning that this assumption is valid since the edge server usually has more transmit power and can occupy more RBs for the global model broadcasting compared to devices.

Let $\gamma_{\text{th}}$ denotes the signal to interference plus noise ratio (SINR) threshold for successful data decoding. The successful transmission indicator of device $k$ in round $t$ is $s_{k,t} = \sum_{m=1}^{M} z_{k,t}^{(m)} \mathbb{1}\left(\text{SINR}_{k,t}^{(m)} \geq \gamma_{\text{th}}\right)$, where $\text{SINR}_{k,t}^{(m)} = \frac{p_{k,t}h_{k,t}}{I_m + BN_0}$ is the SINR of device $k$ in $m$-th RB. The successful transmission probability of device $k$ through $m$-th channel in round $t$ is given by

$$\begin{aligned}\Pr\left(\text{SINR}_{k,t}^{(m)} \geq \gamma_{\text{th}}\right) &= \Pr\left(\frac{p_{k,t}h_{k,t}}{I_m + BN_0} \geq \gamma_{\text{th}}\right) \\ &= \Pr\left(\rho_k(t) \geq \frac{\gamma_{\text{th}}(I_m + BN_0)}{p_{k,t} d_k^{-v}}\right) = e^{-\frac{\gamma_{\text{th}}(I_m + BN_0)}{p_{k,t} d_k^{-v}}}, \end{aligned} \tag{14}$$

where $e$ refers to the Euler's number.

### F. Problem Formulation

In this work, we aim to minimize the global loss function after $T$ training rounds in the resource-limited and unreliable wireless network. To this end, we formulate an optimization problem to jointly optimize device scheduling, RB allocation, and power control as follows:

$$\begin{aligned}
\mathcal{P}: \quad & \min_{\{\boldsymbol{\alpha}_t, \boldsymbol{Z}_t, \boldsymbol{p}_t\}_{t=0}^{T-1}} \mathbb{E}[F(\boldsymbol{w}_T)] && (15)\\
\text{s. t.} \quad & E_{k,t}^{\text{C}} + E_{k,t}^{\text{U}} \leq E_{k,\max}, \forall k \in \mathcal{K}, \forall t, && (15a)\\
& T_{k,t}^{\text{C}} + T_{k,t}^{\text{U}} \leq T_{\max}, \forall k \in \mathcal{K}, \forall t, && (15b)\\
& z_{k,t}^{(m)} \in \{0,1\}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, \forall t, && (15c)\\
& \sum\nolimits_{m=1}^{M} z_{k,t}^{(m)} \leq 1, \forall k \in \mathcal{K}, \forall t, && (15d)\\
& \sum\nolimits_{k=1}^{K} z_{k,t}^{(m)} \leq 1, \forall m \in \mathcal{M}, \forall t, && (15e)\\
& 0 \leq p_{k,t} \leq p_{k,\max}, \forall k \in \mathcal{K}, \forall t, && (15f)\\
& \alpha_{k,t} \in \{0,1\}, \forall k \in \mathcal{K}, \forall t, && (15g)
\end{aligned}$$

where (15a) stipulates that the energy consumption for each participating device $k$ ($k \in \mathcal{K}$) in each round cannot exceed its budget $E_{k,\max}$. $T_{\max}$ in (15b) is the maximum delay of one-round FL training. (15c), (15d) and (15e) correspond to the RB allocation restrictions, indicating that one device can occupy at most one RB for uplink transmission, and one RB can only be allocated to one device. (15f) is the devices' transmit power constraint. (15g) indicates which devices are scheduled in each round.

Solving problem $\mathcal{P}$ requires the explicit form of the global loss function related to the device scheduling, power control, and RB allocation policy. However, the evolution of machine learning models in the learning process is very complex. It is almost impossible to find an exact analytical expression of $\mathbb{E}[F(\boldsymbol{w}_T)]$ with respect to $\boldsymbol{\alpha}_t$, $\boldsymbol{Z}_t$, and $\boldsymbol{p}_t$. Thus, we turn to find an upper bound of $\mathbb{E}[F(\boldsymbol{w}_T)]$ in Section III-B and minimize it for the global loss minimization.

## III. Memory-friendly FL-GR and Convergence Analysis

In this section, to improve the implementation feasibility of the proposed FL-GR in practical wireless networks, we first propose a memory-friendly FL-GR that is equivalent to the proposed FL-GR in Section II-B but with low memory requirements of the edge server. Then, we theoretically analyze the convergence bound of FL-GR to reveal how the device scheduling, RB allocation, and power control policies affect its learning performance. Motivated by this, we define a new objective function, i.e., the average staleness of local gradients, to transform problem $\mathcal{P}$ into a tractable one for guiding the wireless network design.

### A. Memory-friendly FL-GR

It is worth mentioning that implementing the proposed FL-GR in Section II-B requires the edge server to maintain a huge array to cache the latest gradient information for each device. Thus, the cache size requirement of the edge server in FL-GR scales with the model size and the number of devices. This may restrict the scale of the wireless FL system since the server's memory may be exhausted when the number of devices is very large. To address this issue, we propose a **memory-friendly FL-GR** in which each device $k$ maintains a gradient array $\boldsymbol{G}_{k,t}$ to cache its previous latest gradient, and the edge server maintains a gradient array $\bar{\boldsymbol{G}}_t$ to cache local gradients' aggregation information. Then we replace step 3 and step 4 in Section II-B with the following steps:

- Replace Step 3 in Section II-B with: After all selected devices accomplish local model training, they upload the difference between their current and the previous latest cumulative gradient, i.e., $\tilde{\boldsymbol{g}}_{k,t} - \boldsymbol{G}_{k,t-1}$, to the edge server.
- Replace Step 4 in Section II-B with: The edge server updates $\bar{\boldsymbol{G}}_t$ as $\bar{\boldsymbol{G}}_t = \bar{\boldsymbol{G}}_{t-1} + \frac{1}{K}\sum_{k=1}^{K} \alpha_{k,t} s_{k,t}(\tilde{\boldsymbol{g}}_{k,t} - \boldsymbol{G}_{k,t-1})$, and all devices update their gradient array $\boldsymbol{G}_{k,t}$ according to (6). Then, the edge server updates the global model as $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \bar{\boldsymbol{G}}_t$.

By replacing step 3 and step 4 in Section II-B with the above two steps, the edge server distribute the memory requirement to the devices and form a memory-friendly FL-GR algorithm, as shown in Fig. 1(c). For this memory-friendly FL-GR algorithm, we have the following theorem.

**Theorem 1.** *The memory-friendly FL-GR which formed by replacing step 3 and step 4 in Section II-B with the above two steps is equivalent to the proposed FL-GR in Section II-B.*

*Proof.* We prove Theorem 1 by Mathematical induction. Firstly, the maintained gradient array $\bar{\boldsymbol{G}}_t$ at the edge server satisfies:

$$\begin{aligned}\bar{\boldsymbol{G}}_t &= \bar{\boldsymbol{G}}_{t-1} + \frac{1}{K}\sum\nolimits_{k=1}^{K} \alpha_{k,t} s_{k,t}(\tilde{\boldsymbol{g}}_{k,t} - \boldsymbol{G}_{k,t-1}) \\ &= \bar{\boldsymbol{G}}_{t-1} + \frac{1}{K}\sum\nolimits_{k=1}^{K} (\boldsymbol{G}_{k,t} - \boldsymbol{G}_{k,t-1}). \end{aligned} \tag{16}$$

Note that at the beginning of the learning process, the devices' gradient array $\boldsymbol{G}_{k,-1}$ and the server's gradient array $\bar{\boldsymbol{G}}_{-1}$ are all initialized with $\mathbf{0}$. Thus, when $t = 0$, we have

$$\bar{\boldsymbol{G}}_0 = \bar{\boldsymbol{G}}_{-1} + \frac{1}{K}\sum_{k=1}^{K}(\boldsymbol{G}_{k,0} - \boldsymbol{G}_{k,-1}) = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{G}_{k,0}. \tag{17}$$

**Algorithm 1** Memory-friendly Implementation of FL-GR

**Initialization:** The edge server initials its gradient array $\bar{\boldsymbol{G}}_{-1} = \boldsymbol{0}$ and the global model $\boldsymbol{w}_0$, each device $k$ ($k \in \mathcal{K}$) initial their gradient array as $\boldsymbol{G}_{k,-1} = \boldsymbol{0}$
**Server side:**
**for** $t = 0, 1, \cdots, T-1$ **do**
  Select a subset of devices and broadcast the latest global model $\boldsymbol{w}_t$ to them.
  **if** Receive the gradient information from the selected devices **then**
    Update the gradient array $\bar{\boldsymbol{G}}_t$ as $\bar{\boldsymbol{G}}_t = \bar{\boldsymbol{G}}_{t-1} + \frac{1}{K}\sum_{k=1}^{K} \alpha_{k,t} s_{k,t} (\tilde{\boldsymbol{g}}_{k,t} - \boldsymbol{G}_{k,t-1})$
    Update the global model according to $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \bar{\boldsymbol{G}}_t$.
  **else**
    $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t$
**Device side:**
**if** Device $k$ is scheduled **then**
  Receive the global model $\boldsymbol{w}_t$ from the edge server and initial $\boldsymbol{w}_{k,t}^{(0)} = \boldsymbol{w}_t$;
  **for** $l = 0, 1, \cdots, \lambda - 1$ **do**
    Update the local model according (3)
  Compute the cumulative stochastic gradient $\tilde{\boldsymbol{g}}_{k,t} = \frac{1}{\eta}\left(\boldsymbol{w}_t - \boldsymbol{w}_{k,t}^{(\lambda)}\right)$
  Upload the $\tilde{\boldsymbol{g}}_{k,t} - \boldsymbol{G}_{k,t-1}$ to the edge server.
  Update the gradient array $\boldsymbol{G}_{k,t}$ according to (6).

When $t = 1$,

$$\bar{\boldsymbol{G}}_1 = \bar{\boldsymbol{G}}_0 + \frac{1}{K}\sum\nolimits_{k=1}^{K}(\boldsymbol{G}_{k,1} - \boldsymbol{G}_{k,0}) = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{G}_{k,1} + \bar{\boldsymbol{G}}_0 - \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{G}_{k,0} = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{G}_{k,1}. \quad (18)$$

Similarly, we can conclude that for any $t$, $\bar{\boldsymbol{G}}_t = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{G}_{k,t}$ is established. Thus, the updated global model by this memory-friendly FL-GR is $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta\bar{\boldsymbol{G}}_t = \boldsymbol{w}_t - \eta\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{G}_{k,t}$, which is equivalent to update rule of the global model in (7) in Section II-B. □

According to Theorem 1, one can implement the above memory-friendly FL-GR to achieve an equivalent learning process with the proposed FL-GR in Section II in practical wireless networks. It is worth mentioning that the computation costs, communication costs, and the learned global model of these two algorithms are the same. The memory-efficient FL-GR only requires the edge server to maintain a single gradient array, whereas FL-GR necessitates the maintenance of all $K$ users' gradient arrays. For clarity, we summarize the detailed steps of memory-friendly FL-GR in Algorithm 1. In the following, we analyze the convergence performance of FL-GR and transform problem $\mathcal{P}$ into an tractable one for device scheduling, RB allocation, and power control.

### B. Convergence Analysis

For the simplicity of notation, we define the local full gradient on device $k$ in the $l$-th local iteration of the $t$-th round as $\nabla F_k(\boldsymbol{w}_{k,t}^{(l)}) = \frac{1}{D_k}\sum_{\boldsymbol{x}\in\mathcal{D}_k}\nabla f(\boldsymbol{x}, y; \boldsymbol{w}_{k,t}^{(l)})$. Let $F(\boldsymbol{w}^*)$ denote the loss function of the optimal global model $\boldsymbol{w}^*$, and $\tilde{\eta} = \eta\lambda$ as an auxiliary variable. In addition, it is worth mentioning that we recycle the latest historical gradients of the unscheduled and transmission-failure devices to update the global model. To identify the time information of devices' gradients, we define the staleness of device $k$'s local gradient as $\tau_{k,t}$, which evolves as

$$\tau_{k,t} = \begin{cases} \tau_{k,t-1} + 1, & \text{if } \alpha_{k,t}s_{k,t} = 0, \\ 0, & \text{if } \alpha_{k,t}s_{k,t} = 1, \end{cases} \quad \forall k \in \mathcal{K}. \quad (19)$$

Before starting the convergence analysis of FL-GR, we make the following standard assumptions for the local loss functions, i.e., $F_1(\boldsymbol{w}), F_2(\boldsymbol{w}), \cdots, F_K(\boldsymbol{w})$.

*Assumption* 1. All the local loss functions, $F_k(\boldsymbol{w})$ ($\forall k \in \mathcal{K}$), are $L$-smooth. That is, for all $\boldsymbol{v}$ and $\boldsymbol{w}$,

$$F_k(\boldsymbol{v}) \le F_k(\boldsymbol{w}) + \langle F_k(\boldsymbol{w}), \boldsymbol{v} - \boldsymbol{w}\rangle + \frac{L}{2}\|\boldsymbol{v} - \boldsymbol{w}\|^2. \quad (20)$$

*Assumption* 2. The stochastic gradient $\tilde{\nabla}F_k(\boldsymbol{w}_t)$ ($\forall k \in \mathcal{K}$) is an unbiased estimator of the full gradient $\nabla F_k(\boldsymbol{w}_t)$, i.e., $\mathbb{E}[\tilde{\nabla}F_k(\boldsymbol{w}_t)] = \nabla F_k(\boldsymbol{w}_t)$, and its variance is upper bounded by a constant $\sigma^2$, i.e., $\mathbb{E}\|\tilde{\nabla}F_k(\boldsymbol{w}_t) - \nabla F_k(\boldsymbol{w}_t)\|^2 \le \sigma^2$.

*Assumption* 3. The expected squared norm of devices' gradients is uniformly bounded by $G^2$, i.e., $\|\nabla F_k(\boldsymbol{w}_t)\|^2 \le G^2$, for all $k = 1, 2, \cdots, K$ and $t = 0, 1, \cdots, T-1$.

Assumption 1, 2, and 3 are standard and widely used in the FL literature for convergence analysis, e.g., [10], [12], [30]. These assumptions are satisfied by the loss functions of widely used learning models, e.g., support vector machines (SVM), Logistic regression, and most neural networks [31]. Particularly, a deep neural network defined by a composition of functions is a Lipschitz neural network if the functions in all layers are Lipschitz [32]. It has been proved in [32] and [33] that the convolution layer, linear layer, some nonlinear activation functions (e.g., Sigmoid, tanh, Leaky ReLU, and SoftPlus), and the widely used cross-entropy function have Lipschitz smooth gradients. That is, the loss functions of most neural networks that are consisted of Lipschitz layers are Lipschitz continuous.

Before illustrating the details of convergence bound, we introduce two lemmas based on the above assumptions to assist our convergence analysis.

**Lemma 1.** *Let Assumption 1, 2, and 3 hold, the learning rate satisfy* $\eta \le \frac{1}{2\lambda L}$*, the drift of the local model from the global model after* $l$ *iterations is bounded as*

$$\mathbb{E}\left\|\boldsymbol{w}_{k,t}^{(l)} - \boldsymbol{w}_t\right\|^2 \le \frac{4(\lambda-1)\tilde{\eta}^2}{\lambda}\left(2G^2 + \frac{\sigma^2}{\lambda}\right). \quad (21)$$

*Proof.* See Appendix A. □

**Lemma 2.** *Let Assumption 1, 2, and 3 hold, the learning rate satisfy* $\eta \le \frac{1}{2\lambda L}$*, the difference between the global models in two different rounds, i.e.,* $t$ *and* $t'$ *(*$t \ge t'$*), is bounded as*

$$\mathbb{E}\|\boldsymbol{w}_t - \boldsymbol{w}_{t'}\|^2 \le 3\tilde{\eta}^2(t-t')^2\Big((1 + \frac{8\tilde{\eta}^2L^2(\lambda-1)}{\lambda})G^2 + (1 + \frac{4\tilde{\eta}^2L^2(\lambda-1)}{\lambda^2})\sigma^2\Big). \quad (22)$$

*Proof.* See Appendix B. □

Based on Lemma 1 and Lemma 2, we derive the one-round convergence bound of the proposed FL-GR in Theorem 2 as follows:

**Theorem 2.** *Let Assumption 1, 2, and 3 hold, the learning rate satisfy* $\eta \leq \frac{1}{2\lambda L}$*, the one-round convergence bound is given by*

$$\mathbb{E}\left[F(\boldsymbol{w}_{t+1})-F(\boldsymbol{w}_t)\right] \leq \left(-\frac{1}{2}\tilde{\eta}+3L\tilde{\eta}\right)\left\|\nabla F(\boldsymbol{w}_t)\right\|^2 + c_1 + \frac{c}{K}\sum_{k=1}^{K}(\tau_{k,t-1}+1)^2\Big(1-\alpha_{k,t}\sum_{m=1}^{M} z_{k,t}^{(m)}\Pr(\mathrm{SINR}_{k,t}^{(m)}\geq\gamma_{\mathrm{th}})\Big), \tag{23}$$

*where* $c=\frac{9}{8}(\tilde{\eta}+1)\left((1+\frac{2(\lambda-1)}{\lambda})G^2+(1+\frac{(\lambda-1)}{\lambda^2})\sigma^2\right)$*, and* $c_1=\frac{(\tilde{\eta}+3\tilde{\eta}L)(\lambda-1)}{\lambda}(2G^2+\frac{\sigma^2}{\lambda})+\frac{(\tilde{\eta}+1)\sigma^2}{2}$.

*Proof.* See Appendix C. □

According to Theorem 2, the summation of the square of local gradients' staleness, i.e., the last term on the right-hand side (RHS) of (23), is a critical factor that negatively affects the learning convergence rate. By increasing the number of scheduled devices and their successful transmission probabilities, the expected staleness of local gradients would be reduced and thus accelerate the learning process. Due to the limited wireless resources, one should carefully design the device scheduling, RB allocation, and power control to improve the number of devices with successful transmission while satisfying their energy and delay constraints.

Based on Theorem 2, the convergence performance of FL-GR after $T$ training rounds is given by the following corollary.

**Corollary 1.** *Let the assumptions in Theorem 2 hold, the expected gap between the global loss after* $T$ *training rounds and the optimal loss is bounded by*

$$\mathbb{E}\left[F(\boldsymbol{w}_T)-F(\boldsymbol{w}^*)\right] \leq (1-\tilde{\eta}L+6\tilde{\eta}L^2)^T\mathbb{E}\left[F(\boldsymbol{w}_0)-F(\boldsymbol{w}^*)\right] + c_1\frac{1-(1-\tilde{\eta}L+6\tilde{\eta}L^2)^T}{\tilde{\eta}L-6\tilde{\eta}L^2} + c\sum_{t=1}^{T-1}(1-\tilde{\eta}L+6\tilde{\eta}L^2)^{T-1-t}\frac{1}{K}\sum_{k=1}^{K}(\tau_{k,t-1}+1)^2 \times\Big(1-\alpha_{k,t}\sum_{m=1}^{M} z_{k,t}^{(m)}\Pr(\mathrm{SINR}_{k,t}^{(m)}\geq\gamma_{\mathrm{th}})\Big). \tag{24}$$

*Proof.* See Appendix D. □

From Corollary 1, the expected gap between the global loss after $T$ rounds and the optimal loss is bounded by three terms: 1) the initial gap between the global loss and the optimal loss. 2) a constant term related to the system hyperparameters caused by multiple local iterations ($\lambda > 1$) and stochastic gradient error. 3) the cumulative staleness of local gradients over $T$ training rounds. The first two terms determined by the system hyperparameters and the initial global model are unrelated to device scheduling, RB allocation, and power control policies. The last term is highly related to the wireless network design, which indicates that an out-of-date local gradient may degrade the learning performance. To minimize the global loss function and improve the learning performance, one should carefully design the device scheduling, RB allocation, and power control policy to minimize the average staleness of local gradients (last term on the RHS of (24)) for preventing the over stale local gradients. For the global loss optimization, we have the following remark:

*Remark* 2. It is worth mentioning that similar to many existing works, e.g., [16], [34], the available devices and wireless resources in problem $\mathcal{P}$ are independent across different rounds. Thus, the convergence bound in (24) can be minimized by directly minimizing the average staleness of local gradients in each round, i.e., $\frac{1}{K}\sum_{k=1}^{K}(\tau_{k,t-1}+1)^2\Big(1-\alpha_{k,t}\sum_{m=1}^{M} z_{k,t}^{(m)}\Pr\left(\mathrm{SINR}_{k,t}^{(m)}\geq\gamma_{\mathrm{th}}\right)\Big)$. Inspired by this, we define a new objective function based on Theorem 2 and Corollary 1, i.e., $\frac{1}{K}\sum_{k=1}^{K}(\tau_{k,t-1}+1)^2\Big(1-\alpha_{k,t}\sum_{m=1}^{M} z_{k,t}^{(m)}\Pr\left(\mathrm{SINR}_{k,t}^{(m)}\geq\gamma_{\mathrm{th}}\right)\Big)$, which directly minimizes the upper bound on $\mathbb{E}[F(\boldsymbol{w}_{t+1})-F(\boldsymbol{w}_t)]$ in each round and achieves the minimization of the $T$-rounds convergence bound in (24).

## IV. Optimal Device Scheduling, Resource Allocation, and Power Control

In this section, we propose an effective device scheduling, RB allocation, and power control algorithm that solves problem $\mathcal{P}$. Towards this end, we first transform problem $\mathcal{P}$ into a tractable one based on the convergence analysis in Section III-B. Then, we solve the optimal power control and RB allocation policies in an effective manner.

### A. Problem Transformation

The convergence analysis results in Theorem 2 and Corollary 1 reveal how the wireless network design affects the learning performance of FL-GR. According to Remark 2, we transform problem $\mathcal{P}$ into minimize $\frac{1}{K}\sum_{k=1}^{K}(\tau_{k,t-1}+1)^2\Big(1-\alpha_{k,t}\sum_{m=1}^{M} z_{k,t}^{(m)}\Pr\left(\mathrm{SINR}_{k,t}^{(m)}\geq\gamma_{\mathrm{th}}\right)\Big)$ in each round through device scheduling, RB allocation, and power control policies. Since $\alpha_{k,t}=\sum_{m=1}^{M} z_{k,t}^{(m)}\in\{0,1\}$, we have $\alpha_{k,t}\sum_{m=1}^{M} z_{k,t}^{(m)}\Pr(\mathrm{SINR}_{k,t}^{(m)}\geq\gamma_{\mathrm{th}}) = \sum_{m=1}^{M} z_{k,t}^{(m)}\Pr(\mathrm{SINR}_{k,t}^{(m)}\geq\gamma_{\mathrm{th}})$. That is, when the RB allocation decision is given, the device scheduling policy can be directly computed by $\alpha_{k,t}=\sum_{m=1}^{M} z_{k,t}^{(m)}$ $(\forall k\in\mathcal{K})$. Therefore, we transform problem $\mathcal{P}$ into minimizing the average square of local gradients' staleness in each round as follows:

$$\widehat{\mathcal{P}}: \quad \min_{\boldsymbol{Z}_t,\boldsymbol{p}_t} \ \mathcal{R}(\boldsymbol{Z}_t,\boldsymbol{p}_t) \quad \text{s. t.} \ \ (15a)-(15f). \tag{25}$$

where $\mathcal{R}(\boldsymbol{Z}_t,\boldsymbol{p}_t) = \frac{1}{K}\sum_{k=1}^{K}(\tau_{k,t-1}+1)^2\Big(1-\sum_{m=1}^{M} z_{k,t}^{(m)}\Pr(\mathrm{SINR}_{k,t}^{(m)}\geq\gamma_{\mathrm{th}})\Big)$. Problem $\widehat{\mathcal{P}}$ is a non-convex optimization problem which is difficult to solve. In the following, we derive the optimal power control policy for each device under any given RB allocation decision and transform problem $\widehat{\mathcal{P}}$ into an equivalent linear programming problem that can be effectively addressed.

### B. Optimal Power Control

For any given RB allocation policy $\boldsymbol{Z}_t$, it is straightforward to see that the power control policies of devices do not affect each other and independently contribute to the objective

function. Therefore, the power control policy for each device can be solely optimized by itself. With given RB allocation policy $\boldsymbol{Z}_t$, we decompose the power control optimization problem for each device $k$ ($k \in \mathcal{K}$) as follows:

$$\mathcal{P}_1: \quad \min_{p_{k,t}} \ h_k(p_{k,t}) \tag{26}$$
$$\text{s. t.} \quad (15a), (15f).$$

where

$$h_k(p_{k,t}) = -(\tau_{k,t-1}+1)^2 \sum_{m=1}^{M} z_{k,t}^{(m)} e^{-\frac{\gamma_{\text{th}}(I_m+BN_0)}{p_{k,t} d_k^{-v}}}. \tag{27}$$

Problem $\mathcal{P}_1$ is a non-convex optimization problem. To solve the optimal power control policy, below we analyze the properties of the objective function and constraints of problem $\mathcal{P}_1$. Firstly, the first-order partial derivative of the objective function with respect to $p_{k,t}$ is given by

$$\frac{\partial h_k(p_{k,t})}{\partial p_{k,t}} = -(\tau_{k,t-1}+1)^2 \times \sum_{m=1}^{M} z_{k,t}^{(m)} e^{-\frac{\gamma_{\text{th}}(I_m+BN_0)}{p_{k,t} d_k^{-v}}} \frac{\gamma_{\text{th}}(I_m+BN_0)}{p_{k,t}^2 d_k^{-v}}, \forall k \in \mathcal{K}. \tag{28}$$

It is straightforward to see that $\frac{\partial h_k(p_{k,t})}{\partial p_{k,t}} < 0$ since $p_{k,t} > 0$. That is, the objective function $h_k(p_{k,t})$ is a monotonically decreasing function with the transmit power $p_{k,t}$ ($\forall k \in \mathcal{K}$). Thus, the optimal transmit power for each device is its maximum available power. Based on constraint (15a), the energy consumption of gradient information uploading should satisfy $E_{k,t}^{\text{U}} \le E_{k,\max} - E_{k,t}^{\text{C}}$. In addition, the first-order partial derivative of $E_{k,t}^{\text{U}}$ with respect to $p_{k,t}$ is given by

$$\frac{\partial E_{k,t}^{\text{U}}}{\partial p_{k,t}} = \frac{\sum_{m=1}^{M} \frac{z_{k,t}^{(m)} QB}{(1+\gamma_{k,t})\ln 2} \left( (1+\gamma_{k,t})\ln(1+\gamma_{k,t}) - \gamma_{k,t} \right)}{\left( \sum_{m=1}^{M} z_{k,t}^{(m)} B \log_2(1+\gamma_{k,t}) \right)^2}, \tag{29}$$

where $\gamma_{k,t} = \frac{p_{k,t} h_{k,t}}{I_m + BN_0}$. Since $\ln(1+x) > \frac{x}{1+x}$ for $x > 0$, we have $\frac{\partial E_{k,t}^{\text{U}}}{\partial p_{k,t}} > 0$. Therefore, $E_{k,t}^{\text{U}}$ is monotonically increases with $p_{k,t}$. Hence, the transmit power of device $k$ should satisfies $p_{k,t} \le p_{k,t}^{\text{E}}$, where $p_{k,t}^{\text{E}}$ satisfy $\frac{p_{k,t}^{\text{E}} Q}{r_{k,t}(\boldsymbol{z}_{k,t}, p_{k,t}^{\text{E}})} = E_{k,\max} - \kappa \lambda L_b C_k f_k^2$. Combining with (15f), the optimal power control policy for device $k$ is

$$p_{k,t}^* = \min\{p_{k,t}^{\text{E}}, p_{k,\max}\}, \forall k \in \mathcal{K}, \tag{30}$$

where $p_{k,t}^{\text{E}}$ satisfy $\frac{p_{k,t}^{\text{E}} Q}{r_{k,t}(\boldsymbol{z}_{k,t}, p_{k,t}^{\text{E}})} = E_{k,\max} - \kappa \lambda L_b C_k f_k^2$.

### C. Optimal Resource Block Allocation

Up to now, we can compute the optimal power control policy for each device $k$ ($k \in \mathcal{K}$) with any allocated RB $m$ ($m \in \mathcal{M}$) based on (30), denoted by $p_{k,t}^*(m)$. Thus, we compute the optimal power control policy for all devices in all RBs (i.e., $\{p_{k,t}^*(m) : \forall m \in \mathcal{M}, \forall k \in \mathcal{K}\}$) and substitute them into problem $\widehat{\mathcal{P}}$ to simplify it as the following RB allocation problem.

$$\mathcal{P}_2: \quad \max_{\boldsymbol{Z}_t} \frac{1}{K} \sum_{k=1}^{K} \sum_{m=1}^{M} z_{k,t}^{(m)} (\tau_{k,t-1}+1)^2 e^{-\frac{\gamma_{\text{th}}(I_m+BN_0)}{p_{k,t}^*(m) d_k^{-v}}} \tag{31}$$
$$\text{s.t. } (15c), (15d), (15e),$$
$$\frac{\lambda L_b C_k}{f_k} + \frac{Q}{r_{k,t}(z_{k,t}, p_{k,t}^*(m))} \le T_{k,\max}, \forall k, m. \tag{31a}$$

Problem $\mathcal{P}_2$ is a typical non-linear integer programming problem which is difficult to solve. Below we transform it into a maximum weight perfect bipartite matching problem and find its optimal solution within polynomial time. The bipartite matching problem is to find a matching (i.e., a set of edges chosen such that no two edges share an endpoint.) with the maximum weight for the bipartite graph, where the weight is the summation of all the edges in the matching [35].

To transform $\mathcal{P}_2$ into a bipartite matching problem, we construct a complete and balanced bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \mathcal{K} \cup \overline{\mathcal{M}}$ is the vertex set, and $\mathcal{E}$ is the set of edges that connect the vertices in $\mathcal{K}$ and $\overline{\mathcal{M}}$. In $\mathcal{G}$, each vertex $k$ in $\mathcal{K}$ corresponds a device $k$. $\overline{\mathcal{M}} = \mathcal{M} \cup \mathcal{M}_v$ is an extended set of $\mathcal{M}$, where each vertex $m$ in $\mathcal{M}$ corresponds to RB $m$. $\mathcal{M}_v$ is the virtual vertex set used to construct a balanced bipartite graph $\mathcal{G}$, which makes the size of $\overline{\mathcal{M}}$ equal to the size of $\mathcal{K}$, i.e., $|\overline{\mathcal{M}}| = |\mathcal{K}|$. The weight of edges in $\mathcal{G}$ is given by

$$\Delta_{k,m} = \begin{cases} (\tau_{k,t-1}+1)^2 e^{-\frac{\gamma_{\text{th}}(I_m+BN_0)}{p_{k,t}^*(m) d_k^{-v}}}, & \text{if(31a)}, k \in \mathcal{K}, m \in \mathcal{M}, \\ 0, & \text{else.} \end{cases} \tag{32}$$

Note that this work assumes that the number of devices exceeds the number of RBs. When the number of RBs exceeds the number of devices, we can introduce a virtual device set $\mathcal{K}_v$ such that the $|\mathcal{K}| + |\mathcal{K}_v| = |\mathcal{M}|$, and construct a similar graph to the case of $|\mathcal{K}| > |\mathcal{M}|$.

According to the above-defined bipartite graph $\mathcal{G}$, we transform $\mathcal{P}_2$ into a maximum weight perfect bipartite matching problem, which aims to find a perfect matching $\mathcal{H}$ of $\mathcal{G}$ maximizing $\sum_{e \in \mathcal{H}} \Delta_{k,m}$. Let $\theta_{k,m} \in \{0,1\}$ be the edge connecting vertex $k$ ($k \in \mathcal{K}$) and vertex $m$ ($m \in \overline{\mathcal{M}}$), where $\theta_{k,m} = 1$ denote that RB $m$ is allocated to device $k$, and $\theta_{k,m} = 0$ otherwise. For the sake of presentation, we use $\boldsymbol{\theta}_k = \{\theta_{k,1}, \cdots, \theta_{k,|\overline{\mathcal{M}}|}\}$ to denote the connection indicator of device $k$ to all the RBs. Hence, we formulate the bipartite matching problem as the following optimization problem.

$$\widehat{\mathcal{P}_2}: \quad \max_{\{\boldsymbol{\theta}_k\}_{k=1}^K} \sum\nolimits_{k=1}^{K} \sum\nolimits_{m=1}^{|\overline{\mathcal{M}}|} \theta_{k,m} \Delta_{k,m} \tag{33}$$
$$\text{s. t.} \quad \sum\nolimits_{m=1}^{|\overline{\mathcal{M}}|} \theta_{k,m} = 1, \forall k \in \mathcal{K}, \tag{33a}$$
$$\sum\nolimits_{k=1}^{K} \theta_{k,m} = 1, \forall m \in \overline{\mathcal{M}}, \tag{33b}$$
$$\theta_{k,m} \in \{0,1\}, \forall k \in \mathcal{K}, \forall m \in \overline{\mathcal{M}}. \tag{33c}$$

It is worth mentioning that any solution to problem $\widehat{\mathcal{P}_2}$ corresponds to a perfect matching of graph $\mathcal{G}$. However, problem $\widehat{\mathcal{P}_2}$ is a linear integer programming, which is still difficult to solve. By relaxing the integrality constraint (33c), we can

obtain the following linear programming problem:

$$\widetilde{\mathcal{P}_2}: \max_{\{\boldsymbol{\theta}_k\}_{k=1}^K} \sum\nolimits_{k=1}^{K} \sum\nolimits_{m=1}^{|\overline{\mathcal{M}}|} \theta_{k,m}\Delta_{k,m} \quad (34)$$

$$\text{s. t.} \quad (33a), (33b),$$

$$0 \le \theta_{k,m} \le 1, \forall k \in \mathcal{K}, \forall m \in \overline{\mathcal{M}}. \quad (34a)$$

Problem $\widetilde{\mathcal{P}_2}$ is the linear programming relaxation of problem $\widehat{\mathcal{P}_2}$, which can be solved by using the current matrix multiplication time algorithm [36] with time complexity of $\mathcal{O}((K^{2+1/6})^2)$ since it has $K^2$ variables (i.e., $\theta_{k,m} : k \in \mathcal{K}, m \in \overline{\mathcal{M}}$). Note that in problem $\widetilde{\mathcal{P}_2}$, each row in the coefficient matrix corresponding to (33a) and constraint (33b) only contains a '1'. This implements that each square submatrix of this coefficient matrix has a determinant equal to 0, +1, or -1. Thus, this coefficient matrix is a totally unimodular matrix. Based on [35], the optimal solution of problem $\widetilde{\mathcal{P}_2}$ is an integer solution which is equal to the optimal solution of problem $\widehat{\mathcal{P}_2}$. That is, the optimal solution of $\widehat{\mathcal{P}_2}$ can be obtained by directly solving problem $\widetilde{\mathcal{P}_2}$. In the above analysis, we first transform problem $\widehat{\mathcal{P}}$ into an equivalent maximum weight perfect bipartite matching problem, i.e., problem $\widehat{\mathcal{P}_2}$. Then, we further transform problem $\widehat{\mathcal{P}_2}$ into its equivalent linear programming $\widetilde{\mathcal{P}_2}$. It is worth mentioning that these are two equivalent transformations and do not change the optimality of problem $\widehat{\mathcal{P}}$. Thus, the optimal solution of problem $\widehat{\mathcal{P}}$ can be addressed by first solving the optimal solution of problem $\widetilde{\mathcal{P}_2}$. When the optimal solution of problem $\widetilde{\mathcal{P}_2}$ is found, the optimal RB allocation is determined. Furthermore, the optimal device scheduling policy can be computed by $\alpha_{k,t}^* = \sum_{m=1}^{M} z_{k,t}^{(m),*}$ $(\forall k \in \mathcal{K})$, and the optimal transmit power of each device can be determined by (30).

According to the above analysis, we can solve problem $\widehat{\mathcal{P}}$ in an effective manner to obtain the optimal device scheduling, power control, and RB allocation policies. For clarity, we summarize the detailed steps of solving problem $\widehat{\mathcal{P}}$ in Algorithm 2. Firstly, Algorithm 2 requires computing all devices' optimal power control policies in all RBs according to (30), which requires computing $K \times M$ times of power control policy and has a time complexity of $\mathcal{O}(KM)$. Then, we construct a bipartite graph to transform problem $\widehat{\mathcal{P}}$ into a maximum weight perfect bipartite matching problem, i.e., $\widehat{\mathcal{P}_2}$. This step requires calculating the successful transmission probabilities for all devices in all RBs and judging whether the devices' delay satisfies the latency constraint. The time complexity of this step is $\mathcal{O}(2KM)$. Finally, we transform problem $\widehat{\mathcal{P}_2}$ into equivalent linear programming (i.e., $\widetilde{\mathcal{P}_2}$) and utilize the current matrix multiplication time algorithm [36] to solve its optimal solution for obtaining the RB allocation policy. After that, we find the optimal power control of scheduling devices based on the RB allocation policy and compute the device scheduling policies as $\alpha_{k,t}^* = \sum_{m=1}^{M} z_{k,t}^{(m),*}$ $(\forall k \in \mathcal{K})$. Thus, the overall time complexity of Algorithm 2 is $\mathcal{O}(3KM + (K^{2+1/6})^2)$.

Algorithm 2 requires computing the optimal power control policy and successful transmission probabilities for all devices in all RBs, which has a time complexity of $\mathcal{O}(2KM)$. Then, we construct a bipartite graph and solve the corresponding linear programming, and the time complexity is $\mathcal{O}((K^{2+1/6})^2)$. Thus, the overall time complexity of Algorithm 2 is $\mathcal{O}(2KM + (K^{2+1/6})^2)$.

**Algorithm 2** Optimal Device Scheduling, Power control, and RB allocation

1: Compute the optimal power control policy for each device in all RBs according to (30)
2: Compute the successful transmission probabilities for all devices with all RB, i.e., $\Pr(\text{SINR}_{k,t}^{(m)} \ge \gamma_{\text{th}})$ $(\forall k \in \mathcal{K}, \forall m \in \mathcal{M})$
3: Construct a bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and compute the weight of each edge in $\mathcal{E}$ according to (32)
4: Construct the linear programming problem $\widetilde{\mathcal{P}_2}$
5: Solve problem $\widetilde{\mathcal{P}_2}$ and obtain the optimal bipartite perfect matching $\{\boldsymbol{\theta}_k\}_{k=1}^K$
6: Compute the optimal RB allocation policy $\boldsymbol{Z}_t^* = \{z_{k,t}^{(m),*} : k \in \mathcal{K}, m \in \mathcal{M}\}$, where $z_{k,t}^{(m),*} = \theta_{k,m}$
7: Compute the optimal device scheduling policy $\boldsymbol{\alpha}_t^* = \{\alpha_{k,t}^* : \forall k \in \mathcal{K}\}$, where $\alpha_{k,t}^* = \sum_{m=1}^{M} z_{k,t}^{(m),*}$ $(\forall k \in \mathcal{K})$
8: Return the optimal device scheduling policy $\boldsymbol{\alpha}_t^*$, RB allocation policy $\boldsymbol{Z}_t^*$, and power control policy $\boldsymbol{p}_t^*$

## V. Numerical Results

In this section, we evaluate the performance of our proposed FL-GR and the device scheduling algorithm.

TABLE II
Network Architecture for the Classification Model

| Dataset | Model Name | Model Architecture |
|---|---|---|
| MNIST | MLP | F: [784, 128, 10] |
| CIFAR-10 | CNN | C: [6, M, 16, M] |
| | | F: [1600, 256, 64, 10] |
| CIFAR-100 | VGG-11 | C: VGG-11 feature extractor [37] |
| | | F: [512, 256, 100] |

### A. Simulation Settings

For the simulations, we consider a cellular network with a coverage radius of 500m, in which one base station is located at its centre and $K$ devices are randomly distributed. The CPU frequency of each device is randomly selected from $\{0.8, 1.0, 1.2, 1.4\}$ GHz. We evaluate the proposed algorithm under three classification learning tasks, i.e., the handwritten digits classification task on the MNIST dataset, as well as the image classification tasks on the CIFAR-10 and CIFAR-100 datasets. The network architectures used for the learning tasks on these three datasets are summarized in Table II, where 'F' denotes the fully connected module, 'C' denotes the convolution module, 'M' denotes the $2 \times 2$ max-pooling layer, and the number indicates the number of neurons in fully connected layers or filters in convolution layers. Particularly, for the CNN used on the CIFAR-10 dataset, the size of convolution kernels are all set to be $5 \times 5$. The input and hidden layers in all the learning models are all activated by the ReLU function. For all three datasets, we adopt a typical heterogeneous data-splitting method that is widely used in the existing FL works, e.g., [16], [34]. We first classify the training data samples according to their labels, then split the data samples in each class into $2K/N$ shards ($N = 10$ on MNIST and CIFAR-10; $N = 100$ on CIFAR-100), and finally randomly distribute two shards of data samples to each device. That is, each device has a data distribution corresponding to at most 2 classes. That is,

TABLE III
SYSTEM PARAMETERS

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| $K$ | 100 | $M$ | 10 |
| $B$ | 1MHz | $N_0$ | -174dBm |
| $v$ | 2 | $\gamma_{\text{th}}$ | 0dB |
| $\kappa$ | $5 \times 10^{-27}$ | $\eta$ | 0.05 |
| $\tau$ | 5 | $L_b$ | 64 |
| $p_{k,\max}(\forall k)$ | 30mW | $I_m(\forall m \in \mathcal{M})$ | $[10^2, 10^5]BN_0$ |
| $Q$(CNN) | 1,962,016 | $C_k$(CNN) | 326,338,5 |
| $E_{k,\max}$(CNN) | 1.0J | $Q$(VGG-11) | 305,685,860 |
| $C_k$(VGG-11) | 76,421,465 | $E_{k,\max}$(VGG-11) | 230J |
| $T_{\max}$(CNN) | 0.3s | $T_{\max}$(VGG-11) | 35s |

TABLE IV
REQUIRED ROUNDS TO REACH A TARGET ACCURACY

| Dataset | $S$ | Target Accuracy | Proposed | Best Baseline | Saved Time |
|---|---|---|---|---|---|
| MNIST | 5 | 85% | 48 | 80(MC) | 40% |
| | 10 | 90% | 50 | 233(MC) | 78.5% |
| CIFAR-10 | 5 | 50% | 128 | 318(W/GR) | 59.7% |
| | 10 | 55% | 147 | 376(W/GR) | 60.9% |
| CIFAR-100 | 5 | 65% | 457 | 640(W/GR) | 28.6% |
| | 10 | 70% | 509 | 760(W/GR) | 33% |

each device has a data distribution corresponding to at most 2 classes. For all models, a momentum of 0.9 is adopted and cross-entropy is adopted as the loss function. In addition, each parameter of these models is quantitated as 16 bits [15]. For all devices in the system, each CPU cycle can process 4 FLOPs. Thus, the required CPU cycles to process one data sample are equal to the number of FLOPs of its model divided by 4. The parameters chosen in the simulations are based on the parameter settings of a typical wireless FL system [10], [14], [19], [38], [39]. If not specified, the default system settings are listed in Table III.

### B. Effectiveness of Gradient Recycling

To evaluate the effectiveness of the proposed FL-GR, we compare it with the following benchmarks in terms of test accuracy under different numbers of successful transmission devices (denoted as $S$) per round. Note that we do not consider the wireless resource limitations and unreliable channels in this subsection. The edge server randomly selects a subset of $S$ devices to undertake the local training process, and FL-GR recycles the historical gradients of unscheduled devices for the global model update. 1) Without gradient recycling (W/GR): In each round, the edge server only aggregates the successfully received gradients from the scheduled devices to update the global model. This scheme is widely used in existing literature, e.g., [19], [21]–[24]. 2) FedProx [40]: FedProx utilizes a proximal term to limit the impact of local updates for improving model performance under heterogeneous data distributions among devices. 3) Model compensation (MC) [25], [26]: In each round, the edge server uses the successfully received local models and the past local models of transmission-failure or unscheduled devices for global model aggregation.

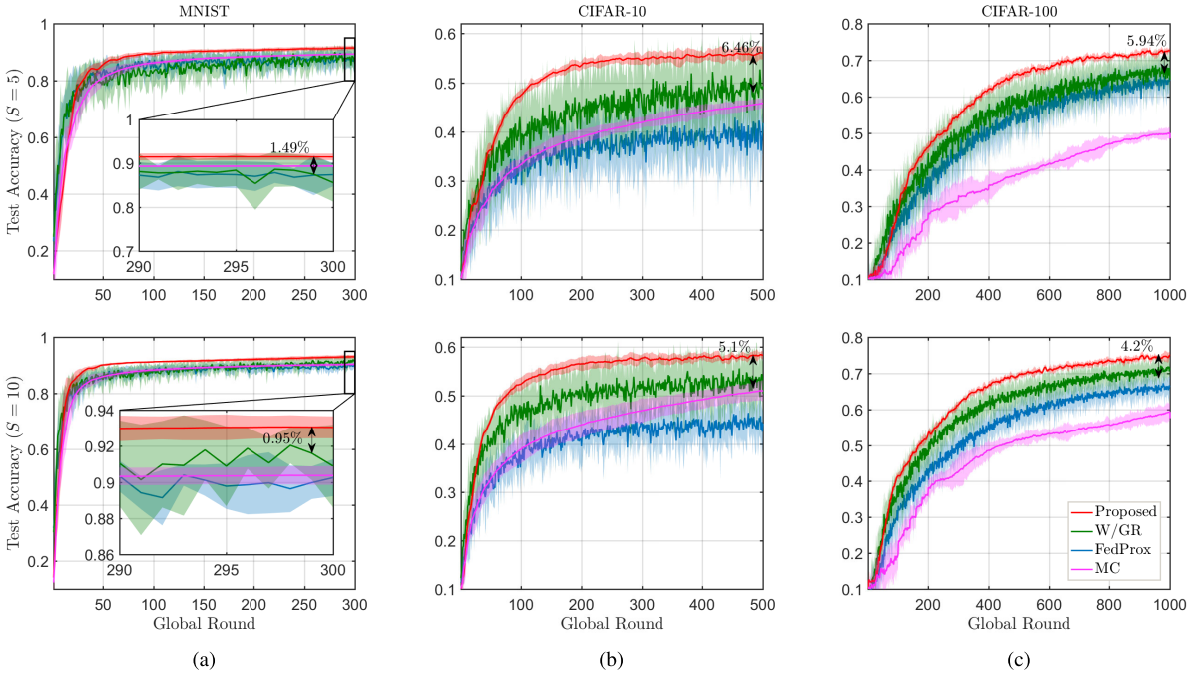
Fig. 2 shows that FL-GR outperforms the benchmarks on all three datasets. In addition, the learning performance of all FL algorithms improved with the increasing number of successful-transmission devices, i.e., the test accuracy of $S = 10$ is greater than that of $S = 5$ for all FL algorithms. Specifically, from the results on the MNIST dataset in Fig. 2(a), when $S = 5$ devices successfully transmitted their gradients to the edge server in each round, FL-GR achieved a 1.49% accuracy improvement compared to the FL algorithms without gradient recycling. Although FL-GR only achieves a slight performance gain (i.e., 0.95%) when $S = 10$, its learning process is more stable than the benchmarks. Fig. 2(b) and Fig. 2(c) evaluate the learning performance of FL-GR on CIFAR-10 and CIFAR-100, respectively, drawing a similar conclusion to the MNIST dataset. In particular, we can observe a more distinct accuracy boosting of FL-GR on these two complicated datasets than the MNIST dataset. From Fig. 2(b), FL-GR obtains 6.46% and 5.1% accuracy improvement when $S = 5$ and $S = 10$, respectively. Fig. 2(c) shows that FL-GR boosts 5.94% and 4.2% accuracy when $S = 5$ and $S = 10$, respectively.

In Table IV, we present the number of optimization rounds necessary to achieve a target accuracy for both the proposed approach and the top-performing baseline algorithm. Specifically, on the CIFAR-100 dataset, when $S = 5$, FL-GR spends only 457 rounds to achieve 65% accuracy, while W/GR (the best benchmark) requires 640 rounds. That is, FL-GR can reduce by 28.6% training time to obtain 65% test accuracy compared to the benchmarks. When $S = 10$, FL-GR is able to save 33% training time to achieve 70% test accuracy compared to the benchmarks.

It is worth mentioning that the simulation results in Fig. 2 show that the proposed FL-GR outperforms the model compensation approach in [25], [26]. This verified the analysis result in Remark 1, i.e., FL-GR recycles the historical local gradients has a smaller model aggregation error than the model compensation approach in [25], [26] that reusing past local models. Thus, FL-GR outperforms the model compensation approach. In addition, although the simulations in [25], [26] demonstrated that the model compensation approach outperforms the W/GR method under full participation and small transmission error rates, our simulation results on the CIFAR-10 and CIFAR-100 datasets show that it does not perform better than W/GR under small successful participation ratios (i.e., $S = 5$ and $S = 10$ correspond to 5% and 10% successful participation ratio, respectively).

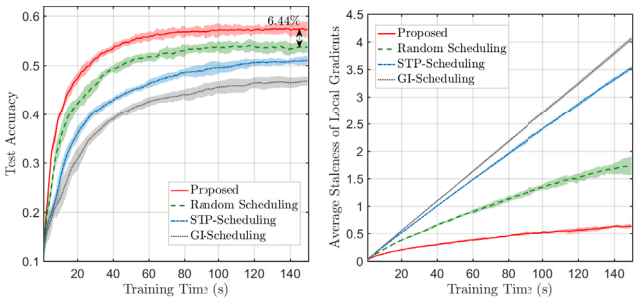
### C. Comparison of Device Scheduling Policies

In this subsection, we compare the proposed device scheduling algorithm to the following scheduling policies: 1) Random scheduling: In each round, the edge server randomly selects a subset of devices and their corresponding RBs that satisfy the constraint (15a)-(15f). 2) Gradient importance-aware scheduling (GI-Scheduling): The edge server selects a subset of devices with the maximum gradient norm and satisfies the constraint (15a)-(15f) in each round. 3) Successful transmission probability-aware scheduling (STP-Scheduling): The edge server selects a subset of devices with the maximum successful transmission probabilities and satisfies the constraint (15a)-(15f). Note that all the device scheduling approaches in this subsection serve the proposed FL-GR to schedule devices instead of other learning frameworks. In fact, random scheduling



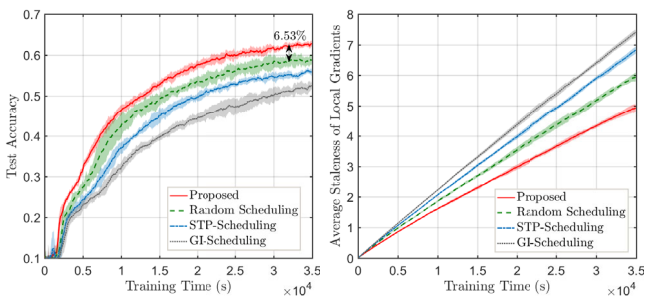

(a) (b) (c)

Fig. 2. Comparison of learning performance of different FL algorithms: (a) on MNIST dataset; (b) on CIFAR-10 dataset; (c) on CIFAR-100 dataset.




(a) (b)

Fig. 3. Comparison of learning performance for different device scheduling algorithms on the CIFAR-10 dataset.



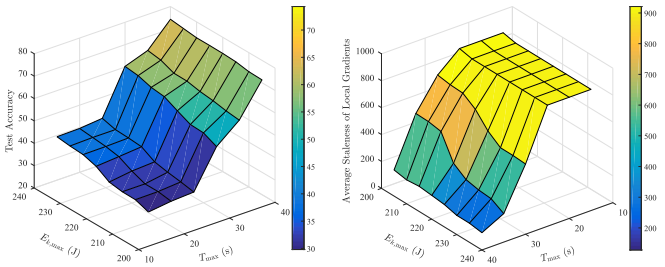

(a) (b)

Fig. 4. Comparison of learning performance for different device scheduling algorithms on the CIFAR-100 dataset.

is equivalent to randomly selecting a perfect matching in the constructed bipartite graph of the proposed device scheduling algorithm instead of the maximum weight perfect matching to schedule devices. GI-Scheduling and STP-Scheduling both construct a similar bipartite graph to the proposed scheduling algorithm and find the maximum weight perfect matching corresponding to their device scheduling policy. In the graph of GI-Scheduling, the weight of each edge is equal to the gradient norm times the successful transmission probability. In the STP-Scheduling, the weight of each edge is the successful transmission probability.

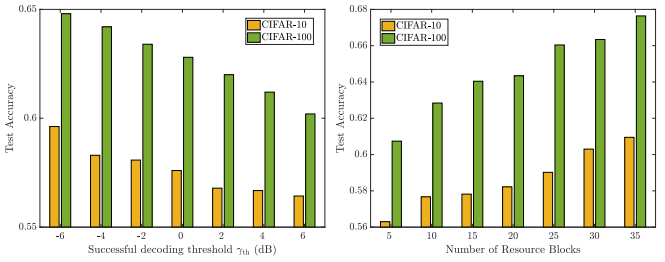
Fig. 3 shows the learning performance of different device scheduling algorithms on the CIFAR-10 dataset. From Fig. 3(a), we can see that the proposed device scheduling algorithm performs better than the other three device scheduling approaches in terms of convergence speed and final test accuracy. Specifically, the proposed device scheduling algorithm achieves around 6.44% accuracy improvement compare to the random scheduling approach. Fig. 3(b) presents the average staleness of local gradients for all four device scheduling algorithms. It is observed that the proposed algorithm possesses the lowest staleness of local gradients. In addition, for the three benchmarks, the device scheduling algorithm with lower staleness obtains higher accuracy and faster convergence speed. This verified our theoretical analysis results in Theorem 2 and Corollary 1, which suggests scheduling the devices with large staleness to reduce the average square staleness of local gradients in each round.

A similar comparison is made on the CIFAR-100 dataset in Fig. 4. We can observe the same conclusion with the simulation on the CIFAR-10 dataset. Specifically, The proposed device scheduling algorithm boosts 6.53% accuracy and possesses the lowest staleness of local gradients compared to the three benchmarks. This simulation further verifies the effectiveness of our convergence analysis in Theorem 2 and Corollary 1. In addition, it is worth mentioning that the simulation results on both CIFAR-10 and CIFAR-100 show that random scheduling performs better than STP-scheduling




(a) (b)

Fig. 5. Impacts of energy and delay constraints on the learning performance on CIFAR-100 dataset.




(a) (b)

Fig. 6. Impacts of wireless parameters on the learning performance: (a) Successful decoding threshold; (b) Number of RBs.

and GI-scheduling when they serve FL-GR. This is because STP-scheduling and GI-scheduling induce higher average staleness of local gradients, as shown in Fig. 3(b) and Fig. 4(b). However, when these scheduling approaches serve for FedAvg, some existing works, e.g., [12], [19], have shown that random scheduling performs worse than STP-scheduling and GI-scheduling.

### D. Impact of Wireless Parameters

This section analyzes the impacts of wireless network parameters on the learning performance of the proposed FL-GR, including energy constraint, delay constraint, the number of RBs, and the successful decode threshold. Note that in this section, the test accuracy on CIFAR-10 and CIFAR-100 is achieved after 150 and $3.5 \times 10^4$ seconds of training, respectively. Based on our simulation results in Section V-C, FL-GR can converge within the pre-setting training time.

In Fig. 5, we test the impacts of energy and delay constraints on the final test accuracy of the proposed FL-GR on CIFAR-100 datasets. From Fig. 5(a), with the increase in energy and delay budgets, we can see that FL-GR achieves higher test accuracy. The reason is that the large energy and delay budgets can increase the number of successful participating devices and reduce the average staleness of local gradients, as shown in Fig. 5(b). When the energy and delay budgets are small, the devices with long time delays and large energy consumption may not satisfy the delay and energy consumption constraints or cannot successfully upload their gradient information to the edge server. Thus, the number of successful participants has been restricted, which results in the high average staleness of local gradients and low test accuracy.

Fig. 6 evaluates the impacts of wireless network parameters on the learning performance of the proposed FL-GR on both CIFAR-10 and CIFAR-100 datasets. From Fig. 6(a), we can see the test accuracy of FL-GR on CIFAR-10 and CIFAR-100 decrease with the increase of $\gamma_{\rm th}$. This is because the large $\gamma_{\rm th}$ reduces the successful transmission probabilities of devices, decreasing the number of successful participants. Hence, the average staleness of local gradients increased. Based on our convergence analysis results, the learning performance of FL-GR will decrease with the increase of $\gamma_{\rm th}$. Fig. 6(b) shows how the number of RBs affects the learning performance of FL-GR. It is observed that the test accuracy on both CIFAR-10 and CIFAR-100 increases with the increase in the number of RBs. The reason is the rise in RBs improves the number of successful participants in each round and thus reduces the average staleness of local gradients.

## VI. Conclusion

In this work, we have developed a novel FL framework, namely FL-GR, that recycles devices' historical gradients to update the global model in the learning process. This framework efficiently copes with the scarcity of radio resources and the unreliability of wireless communications in practical wireless networks. To improve the learning performance of FL-GR, we have formulated an optimization problem to minimize global loss through device scheduling, RB allocation, and power control. To solve this problem, we have investigated the convergence bound of FL-GR and transformed the global loss minimization problem into a tractable one. Then, we derived the optimal power control for any given RB allocation policy and further transformed the global loss minimization problem into an equivalent linear programming problem, which can be solved efficiently. Simulation results on three real-world datasets (i.e., MNIST, CIFAR-10, and CIFAR-100) have shown that the proposed FL-GR achieves higher accuracy and faster convergence speed compared to the FL algorithms without gradient recycling. In addition, the proposed device scheduling algorithm outperforms the existing algorithm in accuracy and convergence speed.

## Appendix

### A. Proof of Lemma 1

For $\lambda = 1$, the bound trivially holds since $\boldsymbol{w}_{k,t}^{(0)} = \boldsymbol{w}_t$. For $\lambda \geq 2$, we have

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{w}_{k,t}^{(l)} - \boldsymbol{w}_t\|^2 = \mathbb{E}\|\boldsymbol{w}_{k,t}^{(l-1)} - \boldsymbol{w}_t - \eta\tilde{\nabla}F_k(\boldsymbol{w}_{k,t}^{(l-1)})\|^2 \\
&\overset{(a)}{=} \mathbb{E}\|\boldsymbol{w}_{k,t}^{(l-1)} - \boldsymbol{w}_t - \eta\nabla F_k(\boldsymbol{w}_{k,t}^{(l-1)})\|^2 \\
&+ \eta^2\mathbb{E}\|\nabla F_k(\boldsymbol{w}_{k,t}^{(l-1)}) - \tilde{\nabla}F_k(\boldsymbol{w}_{k,t}^{(l-1)})\|^2 \\
&\overset{(b)}{\leq} \mathbb{E}\|\boldsymbol{w}_{k,t}^{(l-1)} - \boldsymbol{w}_t - \eta\nabla F_k(\boldsymbol{w}_{k,t}^{(l-1)})\|^2 + \eta^2\sigma^2 \\
&= \mathbb{E}\|\boldsymbol{w}_{k,t}^{(l-1)} - \boldsymbol{w}_t\|^2 + \eta^2\mathbb{E}\|\nabla F_k(\boldsymbol{w}_{k,t}^{(l-1)})\|^2 \\
&- 2\mathbb{E}\Big\langle\frac{1}{\sqrt{\lambda-1}}(\boldsymbol{w}_{k,t}^{(l-1)} - \boldsymbol{w}_t), \eta\sqrt{\lambda-1}\nabla F_k(\boldsymbol{w}_{k,t}^{(l-1)})\Big\rangle + \eta^2\sigma^2 \\
&\overset{(c)}{\leq} (1 + \frac{1}{\lambda-1})\mathbb{E}\|\boldsymbol{w}_{k,t}^{(l-1)} - \boldsymbol{w}_t\|^2
\end{aligned}$$

$$+\lambda\eta^2\mathbb{E}\|\nabla F_k(\boldsymbol{w}_{k,t}^{(l-1)})\|^2+\eta^2\sigma^2$$
$$\overset{(d)}{\leq}(1+\frac{1}{\lambda-1})\mathbb{E}\|\boldsymbol{w}_{k,t}^{(l-1)}-\boldsymbol{w}_t\|^2+2\lambda\eta^2\|\nabla F_k(\boldsymbol{w}_t)\|^2$$
$$+2\lambda\eta^2\mathbb{E}\|\nabla F_k(\boldsymbol{w}_{k,t}^{(l-1)})-\nabla F_k(\boldsymbol{w}_t)\|^2+\eta^2\sigma^2$$
$$\overset{(e)}{\leq}(1+\frac{1}{\lambda-1}+2\lambda\eta^2L^2)\mathbb{E}\|\boldsymbol{w}_{k,t}^{(l-1)}-\boldsymbol{w}_t\|^2$$
$$+2\lambda\eta^2\|\nabla F_k(\boldsymbol{w}_t)\|^2+\eta^2\sigma^2, \tag{35}$$

where (a) is derived by adding and subtracting $\nabla F_k(\boldsymbol{w}_{k,t}^{(l-1)})$ into $\tilde{\nabla}F_k(\boldsymbol{w}_{k,t}^{(l-1)})$ and using the unbiased stochastic gradient in Assumption 2, (b) is due to the bounded variance of stochastic gradient in Assumption 2, (c) comes from the triangle inequality, (d) is derived by adding and subtracting $\nabla F_k(\boldsymbol{w}_t)$ into $\nabla F_k(\boldsymbol{w}_{k,t}^{(l-1)})$ and using the triangle inequality, (e) is due to the $L$-smooth of local loss functions in Assumption 1. Let $\eta\leq\frac{1}{2\lambda L}$, we have $2\lambda\eta^2L^2\leq\frac{1}{2\lambda}\leq\frac{1}{2(\lambda-1)}$. Thus, we have

$$\underbrace{\mathbb{E}\|\boldsymbol{w}_{k,t}^{(l)}-\boldsymbol{w}_t\|^2}_{y_l}\leq\underbrace{\Big(1+\frac{3}{2(\lambda-1)}\Big)}_{c_1}\underbrace{\mathbb{E}\|\boldsymbol{w}_{k,t}^{(l-1)}-\boldsymbol{w}_t\|^2}_{y_{l-1}}$$
$$+\underbrace{2\lambda\eta^2\left\|\nabla F_k(\boldsymbol{w}_t)\right\|^2+\eta^2\sigma^2}_{c_2}. \tag{36}$$

By telescoping the above inequation, we have $y_l\leq c_2\frac{1-c_1^l}{1-c_1}\leq c_2\frac{c_1^{\lambda-1}-1}{c_1-1}$. That is,

$$\mathbb{E}\|\boldsymbol{w}_{k,t}^{(l)}-\boldsymbol{w}_t\|^2\leq(2\lambda\eta^2\left\|\nabla F_k(\boldsymbol{w}_t)\right\|^2+\eta^2\sigma^2)\times\frac{\Big(1+\frac{3}{2(\lambda-1)}\Big)^{\lambda-1}-1}{\frac{3}{2(\lambda-1)}}. \tag{37}$$

In (37), we have $(1+\frac{3}{2(\lambda-1)})^{\lambda-1}=(1+\frac{3}{2(\lambda-1)})^{\frac{2(\lambda-1)}{3}\frac{3}{2}}\leq e^{\frac{3}{2}}\leq 5$ and $\frac{2(\lambda-1)}{3}\leq(\lambda-1)$. Thus,

$$\mathbb{E}\|\boldsymbol{w}_{k,t}^{(l)}-\boldsymbol{w}_t\|^2\leq 4(\lambda-1)\left(2\lambda\eta^2\left\|\nabla F_k(\boldsymbol{w}_t)\right\|^2+\eta^2\sigma^2\right)$$
$$\overset{(a)}{\leq}4(\lambda-1)(2\lambda\eta^2G^2+\eta^2\sigma^2), \tag{38}$$

where (a) follows Assumption 3. Let $\eta=\frac{\tilde{\eta}}{\lambda}$, the proof is completed.

### B. Proof of Lemma 2

For any two round $t$ and $t'$ that satisfies $t\geq t'$, we have

$$\mathbb{E}\|\boldsymbol{w}_t-\boldsymbol{w}_{t'}\|^2=\mathbb{E}\Big\|\sum\nolimits_{j=t'}^{t-1}(\boldsymbol{w}_{j+1}-\boldsymbol{w}_j)\Big\|^2$$
$$=\tilde{\eta}^2\mathbb{E}\Big\|\sum\nolimits_{j=t'}^{t-1}\frac{1}{\lambda K}\sum\nolimits_{k=1}^{K}\sum\nolimits_{l=0}^{\lambda-1}\tilde{\nabla}F_k(\boldsymbol{w}_{k,j-\tau_{k,j}}^{(l)})\Big\|^2$$
$$\overset{(a)}{\leq}\tilde{\eta}^2(t-t')\sum\nolimits_{j=t'}^{t-1}\mathbb{E}\Big\|\frac{1}{\lambda K}\sum\nolimits_{k=1}^{K}\sum\nolimits_{l=0}^{\lambda-1}\tilde{\nabla}F_k(\boldsymbol{w}_{k,j-\tau_{k,j}}^{(l)})\Big\|^2$$
$$\overset{(b)}{\leq}3\tilde{\eta}^2(t-t')\sum\nolimits_{j=t'}^{t-1}\frac{1}{\lambda K}\sum\nolimits_{k=1}^{K}\sum_{l=0}^{\lambda-1}\mathbb{E}\Big\|\tilde{\nabla}F_k(\boldsymbol{w}_{k,j-\tau_{k,j}}^{(l)})$$
$$-\nabla F_k(\boldsymbol{w}_{k,j-\tau_{k,j}}^{(l)})\Big\|^2$$
$$+3\tilde{\eta}^2(t-t')\sum\nolimits_{j=t'}^{t-1}\frac{1}{\lambda K}\sum\nolimits_{k=1}^{K}\sum\nolimits_{l=0}^{\lambda-1}\mathbb{E}\Big\|\nabla F_k(\boldsymbol{w}_{k,j-\tau_{k,j}}^{(l)})$$
$$-\nabla F_k(\boldsymbol{w}_{j-\tau_{k,j}})\Big\|^2$$
$$+3\tilde{\eta}^2(t-t')\sum\nolimits_{j=t'}^{t-1}\frac{1}{K}\sum\nolimits_{k=1}^{K}\mathbb{E}\Big\|\nabla F_k\left(\boldsymbol{w}_{j-\tau_{k,j}}\right)\Big\|^2$$
$$\overset{(c)}{\leq}3\tilde{\eta}^2(t-t')\sum_{j=t'}^{t-1}\frac{1}{\lambda K}\sum_{k=1}^{K}\sum_{l=0}^{\lambda-1}L^2\mathbb{E}\Big\|\boldsymbol{w}_{k,j-\tau_{k,j}}^{(l)}-\boldsymbol{w}_{j-\tau_{k,j}}\Big\|^2$$
$$+3\tilde{\eta}^2(t-t')^2(\sigma^2+G^2), \tag{39}$$

where (a) is due to Jensen's inequality, (b) is derived by adding and subtracting both $\nabla F_k(\boldsymbol{w}_{k,j-\tau_{k,j}}^{l})$ and $\nabla F_k(\boldsymbol{w}_{j-\tau_{k,j}})$ into $\tilde{\nabla}F_k(\boldsymbol{w}_{k,j-\tau_{k,j}}^{l})$, then using the triangle inequality and Jensen's inequality, (c) comes from the Assumption 1, 2, and 3. According to Lemma 1, we have

$$\mathbb{E}\|\boldsymbol{w}_{k,j-\tau_{k,j}}^{(l)}-\boldsymbol{w}_{j-\tau_{k,j}}\|^2\leq\frac{4(\lambda-1)\tilde{\eta}^2}{\lambda}\Big(2G^2+\frac{\sigma^2}{\lambda}\Big). \tag{40}$$

Substituting (40) into (39), the proof is completed.

### C. Proof of Theorem 2

By using the $L$-smooth of the loss functions, we have

$$F(\boldsymbol{w}_{t+1})-F(\boldsymbol{w}_t)\leq\langle\nabla F(\boldsymbol{w}_t),\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\rangle+\frac{L}{2}\left\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\right\|^2. \tag{41}$$

Thus, we have

$$\mathbb{E}\left[F(\boldsymbol{w}_{t+1})-F(\boldsymbol{w}_t)\right]$$
$$\leq-\tilde{\eta}\mathbb{E}\Big\langle\nabla F(\boldsymbol{w}_t),\frac{1}{K\lambda}\sum\nolimits_{k=1}^{K}\sum\nolimits_{l=0}^{\lambda-1}\tilde{\nabla}F_k(\boldsymbol{w}_{k,t-\tau_{k,t}}^{(l)})\Big\rangle$$
$$+\frac{L\tilde{\eta}}{2}\mathbb{E}\Big\|\frac{1}{K\lambda}\sum\nolimits_{k=1}^{K}\sum\nolimits_{l=0}^{\lambda-1}\tilde{\nabla}F_k(\boldsymbol{w}_{k,t-\tau_{t,k}}^{(l)})\Big\|^2$$
$$=\underbrace{-\tilde{\eta}\mathbb{E}\Big\langle\nabla F(\boldsymbol{w}_t),\mathbf{x}_{t-\tau_{t,k}}\Big\rangle}_{A_1}$$
$$+\underbrace{-\tilde{\eta}\mathbb{E}\Big\langle\nabla F(\boldsymbol{w}_t),\frac{1}{K\lambda}\sum\nolimits_{k=1}^{K}\sum\nolimits_{l=0}^{\lambda-1}\nabla F_k(\boldsymbol{w}_{k,t-\tau_{t,k}}^{(l)})\Big\rangle}_{A_2}$$
$$+\underbrace{\frac{L\tilde{\eta}}{2}\mathbb{E}\Big\|\frac{1}{K\lambda}\sum\nolimits_{k=1}^{K}\sum\nolimits_{l=0}^{\lambda-1}\tilde{\nabla}F_k(\boldsymbol{w}_{k,t-\tau_{t,k}}^{(l)})\Big\|^2}_{A_3}, \tag{42}$$

where $\mathbf{x}_{t-\tau_{t,k}}=\frac{1}{K\lambda}\sum_{k=1}^{K}\sum_{l=0}^{\lambda-1}(\tilde{\nabla}F_k(\boldsymbol{w}_{k,t-\tau_{t,k}}^{(l)})-\nabla F_k(\boldsymbol{w}_{k,t-\tau_{t,k}}^{(l)}))$. Below we focus on bounding the three terms in (42). Due to reuse of noisy gradient, the stochastic gradient noise $\tilde{\nabla}F_k(\boldsymbol{w}_{k,t-\tau_{t,k}}^{(l)})-\nabla F_k(\boldsymbol{w}_{k,t-\tau_{t,k}}^{(l)})$ is correlated with $\boldsymbol{w}_t$. Thus, $A_1\neq 0$. For $A_1$, we have

$$A_1=\underbrace{-\tilde{\eta}\mathbb{E}\Big\langle\nabla F(\boldsymbol{w}_t)-\nabla F(\boldsymbol{w}_{t-\tau_{t,k}}),\mathbf{x}_{t-\tau_{t,k}}\Big\rangle}_{B_1}$$
$$\underbrace{-\tilde{\eta}\mathbb{E}\Big\langle\nabla F(\boldsymbol{w}_{t-\tau_{t,k}}),\mathbf{x}_{t-\tau_{t,k}}\Big\rangle}_{B_2}. \tag{43}$$

Since $\boldsymbol{w}_{t-\tau_{t,k}}$ is independent with $\tilde{\nabla}F_k(\boldsymbol{w}_{k,t-\tau_{t,k}}^{(l)})-\nabla F_k(\boldsymbol{w}_{k,t-\tau_{t,k}}^{(l)})$, we have $B_2=0$. For $B_1$, we have:

$$B_1=-\tilde{\eta}\frac{1}{K\lambda}\sum\nolimits_{k=1}^{K}\sum\nolimits_{l=0}^{\lambda-1}\mathbb{E}\Big\langle\nabla F(\boldsymbol{w}_t)-\nabla F(\boldsymbol{w}_{t-\tau_{t,k}}),$$

$$\tilde{\nabla}F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}}) - \nabla F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})\Big\rangle$$
$$\overset{(a)}{\leq} \frac{1}{2}\tilde{\eta}\frac{1}{K}\sum\nolimits_{k=1}^{K}\mathbb{E}\big\|\nabla F(\boldsymbol{w}_t)-\nabla F(\boldsymbol{w}_{t-\tau_{t,k}})\big\|^2$$
$$+\frac{1}{2}\tilde{\eta}\frac{1}{K\lambda}\sum_{k=1}^{K}\sum_{l=0}^{\lambda-1}\mathbb{E}\Big\|\tilde{\nabla}F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})-\nabla F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})\Big\|^2$$
$$\overset{(b)}{\leq} \frac{1}{2}\tilde{\eta}\frac{1}{K}\sum\nolimits_{k=1}^{K}\mathbb{E}\big\|\nabla F(\boldsymbol{w}_t) - \nabla F(\boldsymbol{w}_{t-\tau_{t,k}})\big\|^2 + \frac{1}{2}\tilde{\eta}\sigma^2$$
$$\overset{(c)}{\leq} \frac{1}{2}\tilde{\eta}L^2\frac{1}{K}\sum\nolimits_{k=1}^{K}\mathbb{E}\left\|\boldsymbol{w}_t - \boldsymbol{w}_{t-\tau_{t,k}}\right\|^2 + \frac{1}{2}\tilde{\eta}\sigma^2, \quad (44)$$

where (a) follows the triangle inequality, (b) is due to Assumption 2, (c) comes from the $L$-smooth of the loss function. According to the above analysis, we have

$$A_1 \leq \frac{1}{2}\tilde{\eta}L^2\frac{1}{K}\sum\nolimits_{k=1}^{K}\mathbb{E}\left\|\boldsymbol{w}_t - \boldsymbol{w}_{t-\tau_{t,k}}\right\|^2 + \frac{1}{2}\tilde{\eta}\sigma^2. \quad (45)$$

For $A_2$, we have

$$A_2 = -\tilde{\eta}\mathbb{E}\Big\langle\nabla F(\boldsymbol{w}_t), \frac{1}{K\lambda}\sum\nolimits_{k=1}^{K}\sum\nolimits_{l=0}^{\lambda-1}\nabla F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})\Big\rangle$$
$$\leq -\frac{1}{2}\tilde{\eta}\|\nabla F(\boldsymbol{w}_t)\|^2$$
$$+\frac{1}{2}\tilde{\eta}\mathbb{E}\Big\|\frac{1}{K\lambda}\sum_{k=1}^{K}\sum_{l=0}^{\lambda-1}\Big(\nabla F_k(\boldsymbol{w}_t) - \nabla F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})\Big)\Big\|^2$$
$$\leq -\frac{1}{2}\tilde{\eta}\|\nabla F(\boldsymbol{w}_t)\|^2$$
$$+\tilde{\eta}\mathbb{E}\Big\|\frac{1}{K\lambda}\sum_{k=1}^{K}\sum_{l=0}^{\lambda-1}\big(\nabla F_k(\boldsymbol{w}_t) - \nabla F_k(\boldsymbol{w}_{t-\tau_{t,k}})\big)\Big\|^2$$
$$+\tilde{\eta}\mathbb{E}\Big\|\frac{1}{K\lambda}\sum_{k=1}^{K}\sum_{l=0}^{\lambda-1}\Big(\nabla F_k(\boldsymbol{w}_{t-\tau_{t,k}}) - \nabla F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})\Big)\Big\|^2$$
$$\leq -\frac{1}{2}\tilde{\eta}\|\nabla F(\boldsymbol{w}_t)\|^2 + \tilde{\eta}\frac{1}{K}\sum\nolimits_{k=1}^{K}L^2\mathbb{E}\left\|\boldsymbol{w}_t - \boldsymbol{w}_{t-\tau_{t,k}}\right\|^2$$
$$+\tilde{\eta}\frac{1}{K\lambda}\sum\nolimits_{k=1}^{K}\sum\nolimits_{l=0}^{\lambda-1}L^2\mathbb{E}\Big\|\boldsymbol{w}_{t-\tau_{t,k}} - \boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}}\Big\|^2. \quad (46)$$

For $A_3$, we have

$$A_3 = \frac{L\tilde{\eta}}{2}\mathbb{E}\Big\|\frac{1}{K\lambda}\sum\nolimits_{k=1}^{K}\sum\nolimits_{l=0}^{\lambda-1}\tilde{\nabla}F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})\Big\|^2$$
$$\overset{(a)}{\leq} L\tilde{\eta}\frac{1}{K\lambda}\sum_{k=1}^{K}\sum_{l=0}^{\lambda-1}\mathbb{E}\Big\|\tilde{\nabla}F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}}) - \nabla F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})\Big\|^2$$
$$+ L\tilde{\eta}\mathbb{E}\Big\|\frac{1}{K\lambda}\sum\nolimits_{k=1}^{K}\sum\nolimits_{l=0}^{\lambda-1}\nabla F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})\Big\|^2$$
$$\overset{(b)}{\leq} L\tilde{\eta}\sigma^2 + L\tilde{\eta}\mathbb{E}\Big\|\frac{1}{K\lambda}\sum\nolimits_{k=1}^{K}\sum\nolimits_{l=0}^{\lambda-1}\nabla F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})\Big\|^2, \quad (47)$$

where (a) is derived by adding and subtracting $\nabla F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})$ into $\tilde{\nabla}F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})$, then using the triangle inequality and Jensen's inequality, (b) is due to the bounded variance of stochastic gradient in Assumption 2. Below we bound $\Big\|\frac{1}{K\lambda}\sum_{k=1}^{K}\sum_{l=0}^{\lambda-1}\nabla F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})\Big\|^2$ in (47) as:

$$\mathbb{E}\Big\|\frac{1}{K\lambda}\sum\nolimits_{k=1}^{K}\sum\nolimits_{l=0}^{\lambda-1}\nabla F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})\Big\|^2$$
$$\overset{(a)}{\leq} \frac{3}{K\lambda}\sum_{k=1}^{K}\sum_{l=0}^{\lambda-1}\mathbb{E}\Big\|\nabla F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})-\nabla F_k(\boldsymbol{w}_{t-\tau_{t,k}})\Big\|^2$$
$$+\frac{3}{K}\sum_{k=1}^{K}\mathbb{E}\big\|\nabla F_k(\boldsymbol{w}_{t-\tau_{t,k}})-\nabla F_k(\boldsymbol{w}_t)\big\|^2+3\|\nabla F(\boldsymbol{w}_t)\|^2$$
$$\overset{(b)}{\leq} 3\frac{1}{K\lambda}\sum_{k=1}^{K}\sum_{l=0}^{\lambda-1}L^2\mathbb{E}\Big\|\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}}-\boldsymbol{w}_{t-\tau_{t,k}}\Big\|^2$$
$$+3\frac{1}{K}\sum\nolimits_{k=1}^{K}L^2\mathbb{E}\left\|\boldsymbol{w}_{t-\tau_{t,k}}-\boldsymbol{w}_t\right\|^2+3\|\nabla F(\boldsymbol{w}_t)\|^2, \quad (48)$$

where (a) is derived by adding and subtracting $\nabla F_k(\boldsymbol{w}_{t-\tau_{t,k}})$ and $\nabla F(\boldsymbol{w}_t)$ into $\nabla F_k(\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}})$, then using the triangle inequality and Jensen's inequality, (b) is due to Assumption 1. Substituting (48) into (47), we have

$$A_3 \leq L\tilde{\eta}\sigma^2 + 3L\tilde{\eta}\frac{1}{K\lambda}\sum_{k=1}^{K}\sum_{l=0}^{\lambda-1}L^2\mathbb{E}\|\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}} - \boldsymbol{w}_{t-\tau_{t,k}}\|^2$$
$$+3L\tilde{\eta}\frac{1}{K}\sum_{k=1}^{K}L^2\mathbb{E}\left\|\boldsymbol{w}_{t-\tau_{t,k}}-\boldsymbol{w}_t\right\|^2+3L\tilde{\eta}\left\|\nabla F(\boldsymbol{w}_t)\right\|^2. \quad (49)$$

Substituting (45), (46), and (49) into (42), we have:

$$\mathbb{E}\left[F(\boldsymbol{w}_{t+1}) - F(\boldsymbol{w}_t)\right] \leq \Big(-\frac{1}{2}\tilde{\eta} + 3L\tilde{\eta}\Big)\|\nabla F(\boldsymbol{w}_t)\|^2$$
$$+(\frac{1}{2}+L)\tilde{\eta}\sigma^2+(\frac{3}{2}+3L)\tilde{\eta}L^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\boldsymbol{w}_t-\boldsymbol{w}_{t-\tau_{t,k}}\right\|^2$$
$$+(1+3L)\tilde{\eta}L^2\frac{1}{K\lambda}\sum_{k=1}^{K}\sum_{l=0}^{\lambda-1}\mathbb{E}\|\boldsymbol{w}_{t-\tau_{t,k}}-\boldsymbol{w}^{(l)}_{k,t-\tau_{t,k}}\|^2. \quad (50)$$

According to Lemma 1, Lemma 2, and $\tilde{\eta} \leq \frac{1}{2L}$ we have

$$\mathbb{E}\left[F(\boldsymbol{w}_{t+1}) - F(\boldsymbol{w}_t)\right] \leq (-\frac{1}{2}\tilde{\eta} + 3L\tilde{\eta})\left\|\nabla F(\boldsymbol{w}_t)\right\|^2$$
$$+ c_1 + c\mathbb{E}\Big[\frac{1}{K}\sum\nolimits_{k=1}^{K}\tau_{k,t}^2\Big]. \quad (51)$$

where $c_1 = \frac{(\tilde{\eta}+3\tilde{\eta}L)(\lambda-1)}{\lambda}(2G^2 + \frac{\sigma^2}{\lambda}) + \frac{(\tilde{\eta}+1)\sigma^2}{2}$, $c = \frac{9}{8}(\tilde{\eta}+1)\Big((1+\frac{2(\lambda-1)}{\lambda})G^2 + (1+\frac{(\lambda-1)}{\lambda^2})\sigma^2\Big)$. According to the evolution of devices' staleness in (19), we have $\tau_{k,t} = (1-\alpha_{k,t}s_{k,t})(\tau_{k,t-1}+1)$. Note that $\alpha_{k,t}s_{k,t} \in \{0,1\}$, which induces $(1-\alpha_{k,t}s_{k,t})^2 = (1-\alpha_{k,t}s_{k,t})$. Thus, we have

$$\mathbb{E}\Big[\frac{1}{K}\sum\nolimits_{k=1}^{K}\tau_{k,t}^2\Big] = \mathbb{E}\Big[\frac{1}{K}\sum_{k=1}^{K}(1-\alpha_{k,t}s_{k,t})^2(\tau_{k,t-1}+1)^2\Big]$$
$$= \mathbb{E}\Big[\frac{1}{K}\sum\nolimits_{k=1}^{K}\left(1-\alpha_{k,t}s_{k,t}\right)(\tau_{k,t-1}+1)^2\Big]$$
$$= \frac{1}{K}\sum_{k=1}^{K}(\tau_{k,t-1}+1)^2\Big(1-\alpha_{k,t}\sum_{m=1}^{M}z_{k,t}^{(m)}\Pr(\mathrm{SINR}_{k,t}^{(m)} \geq \gamma_{\mathrm{th}})\Big), \quad (52)$$

where the last inequality is due to $\mathbb{E}\left[s_{k,t}\right] = \sum_{m=1}^{M}z_{k,t}^{(m)}\Pr(\mathrm{SINR}_{k,t}^{(m)} \geq \gamma_{\mathrm{th}})$. By substituting (52) into (51), the proof is completed.

### D. Proof of Corollary 1

To prove Corollary 1, we first prove a key property of smooth functions. Let $F(\boldsymbol{w}^*)$ denote the optimal global loss, i.e., $F(\boldsymbol{w}^*) \leq F(\boldsymbol{w}), \forall \boldsymbol{w}$. Based on the $L$-smooth of $F(\boldsymbol{w})$,

$$F(\boldsymbol{w}^*) \leq F\Big(\boldsymbol{w} - \frac{1}{L}\nabla F(\boldsymbol{w})\Big)$$
$$\leq F(\boldsymbol{w}) - \langle\nabla F(\boldsymbol{w}), \frac{1}{L}\nabla F(\boldsymbol{w})\rangle + \frac{1}{2L}\|\nabla F(\boldsymbol{w})\|^2$$

$$= F(\boldsymbol{w}) - \frac{1}{2L}\|\nabla F(\boldsymbol{w})\|^2. \quad (53)$$

By rearranging the above inequality, the global loss function $F(\boldsymbol{w})$ with $L$-smooth satisfies

$$\|\nabla F(\boldsymbol{w})\|^2 \leq 2L\left(F(\boldsymbol{w}) - F(\boldsymbol{w}^*)\right). \quad (54)$$

Let $F(\boldsymbol{w}_{t+1})$ and $F(\boldsymbol{w}_t)$ in (51) subtract $F(\boldsymbol{w}^*)$, then utilizing the property of $L$-smooth in (54), we have

$$\begin{aligned}\mathbb{E}\left[F(\boldsymbol{w}_{t+1}) - F(\boldsymbol{w}^*)\right] &\leq (1-\tilde{\eta}L+6\tilde{\eta}L^2)\mathbb{E}\left[F(\boldsymbol{w}_t) - F(\boldsymbol{w}^*)\right] \\ &+ c\frac{1}{K}\sum_{k=1}^{K}(\tau_{k,t-1}+1)^2\Big(1-\alpha_{k,t}\sum_{m=1}^{M} z_{k,t}^{(m)}\Pr(\mathrm{SINR}_{k,t}^{(m)} \geq \gamma_{\mathrm{th}})\Big) \\ &+ \frac{(\tilde{\eta}+3\tilde{\eta}L)(\lambda-1)}{\lambda}(2G^2+\frac{\sigma^2}{\lambda}) + \frac{(\tilde{\eta}+1)\sigma^2}{2}, \quad (55)\end{aligned}$$

By telescoping the above inequality, the proof is completed.

## REFERENCES


[1] Z. Chen, W. Yi, Y. Liu, and A. Nallanathan, "Convergence analysis for wireless federated learning with gradient recycling," in *Proc.Int. Wireless Commun. and Mobile Computing (IWCMC)*, 2023, pp. 1232–1237.

[2] H. Hellström, J. M. B. da Silva Jr, M. M. Amiri, M. Chen, V. Fodor, H. V. Poor, C. Fischione, *et al.*, "Wireless for machine learning: A survey," *Foundations and Trends® in Signal Processing*, vol. 15, no. 4, pp. 290–399, 2022.

[3] Z. Yang, M. Chen, K.-K. Wong, H. V. Poor, and S. Cui, "Federated learning for 6G: Applications, challenges, and opportunities," *Engineering*, 2021.

[4] Z. Chen, W. Yi, Y. Liu, and A. Nallanathan, "Knowledge-aided federated learning for energy-limited wireless networks," *IEEE Trans. Commun.*, vol. 71, no. 6, pp. 3368–3386, 2023.

[5] L. U. Khan, W. Saad, Z. Han, E. Hossain, and C. S. Hong, "Federated learning for internet of things: Recent advances, taxonomy, and open challenges," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1759–1799, 2021.

[6] Y. Xi, A. Burr, J. Wei, and D. Grace, "A general upper bound to evaluate packet error rate over quasi-static fading channels," *IEEE Trans. Wireless Commun.*, vol. 10, no. 5, pp. 1373–1377, 2011.

[7] Z. Chen, W. Yi, A. Nallanathan, and G. Y. Li, "Federated learning for energy-limited wireless networks: A partial model aggregation approach," *arXiv preprint arXiv:2204.09746*, 2022.

[8] J. Xu and H. Wang, "Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1188–1200, 2021.

[9] M. Zhang, G. Zhu, S. Wang, J. Jiang, Q. Liao, C. Zhong, and S. Cui, "Communication-efficient federated edge learning via optimal probabilistic device scheduling," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8536–8551, 2022.

[10] 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, 2021.

[11] A. Taïk, Z. Mlika, and S. Cherkaoui, "Data-aware device scheduling for federated edge learning," *IEEE Trans. Cognitive Commun. and Networking*, vol. 8, no. 1, pp. 408–421, 2022.

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

[13] J. Leng, Z. Lin, M. Ding, P. Wang, D. Smith, and B. Vucetic, "Client scheduling in wireless federated learning based on channel and learning qualities," *IEEE Wireless Commun. Letters*, vol. 11, no. 4, pp. 732–735, 2022.

[14] 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, 2021.

[15] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, "Energy-efficient resource management for federated edge learning with cpu-gpu heterogeneous computing," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 7947–7962, 2021.

[16] S. Yue, J. Ren, J. Xin, D. Zhang, Y. Zhang, and W. Zhuang, "Efficient federated meta-learning over multi-access wireless networks," *IEEE J. Sel, Areas Commun.*, vol. 40, no. 5, pp. 1556–1570, 2022.

[17] 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, 2020.

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

[19] 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, 2021.

[20] Y. Wang, Y. Xu, Q. Shi, and T.-H. Chang, "Quantized federated learning under transmission delay and outage constraints," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 323–341, 2022.

[21] H. Chen, S. Huang, D. Zhang, M. Xiao, M. Skoglund, and H. V. Poor, "Federated learning over wireless iot networks with optimized communication and resources," *IEEE Internet Things J.*, vol. 9, no. 17, pp. 16 592–16 605, 2022.

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

[23] M. Salehi and E. Hossain, "Federated learning in unreliable and resource-constrained cellular wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5136–5151, 2021.

[24] H. Hellström, V. Fodor, and C. Fischione, "Over-the-air federated learning with retransmissions," in *Proc SPAWC*, 2021, pp. 291–295.

[25] H. Ye, L. Liang, and G. Y. Li, "Decentralized federated learning with unreliable communications," *IEEE J. Sel. Topics in Signal Processing*, pp. 1–1, 2022.

[26] M. Shirvanimoghaddam, A. Salari, Y. Gao, and A. Guha, "Federated learning with erroneous communication links," *IEEE Commun. Letters*, vol. 26, no. 6, pp. 1293–1297, 2022.

[27] Q. Ma, Y. Xu, H. Xu, Z. Jiang, L. Huang, and H. Huang, "Fedsa: A semi-asynchronous federated learning mechanism in heterogeneous edge computing," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3654–3672, 2021.

[28] S. Dutta, J. Wang, and G. Joshi, "Slow and stale gradients can win the race," *IEEE J. Sel. Areas Info. Theory*, vol. 2, no. 3, pp. 1012–1024, 2021.

[29] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. AISTATS*, 20–22, Apr. 2017.

[30] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-IID data," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[31] A. Dedieu, "Improved error rates for sparse (group) learning with lipschitz loss functions," *arXiv preprint arXiv:1910.08880*, 2019.

[32] E. Abbasnejad, J. Shi, and A. van den Hengel, "Deep Lipschitz networks and dudley GANs," 2018. [Online]. Available: https://openreview.net/forum?id=rkw-_jlb0W

[33] A. Virmaux and K. Scaman, "Lipschitz regularity of deep neural networks: analysis and efficient estimation," in *Proc. Neural Information Processing Systems (NeurIPS)*, 2018.

[34] Y. Sun, S. Zhou, Z. Niu, and D. Gündüz, "Dynamic scheduling for over-the-air federated edge learning with energy constraints," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 227–242, 2022.

[35] A. Schrijver *et al.*, *Combinatorial optimization: polyhedra and efficiency*. Berlin Germany:Springer-Velag, 2003.

[36] M. B. Cohen, Y. T. Lee, and Z. Song, "Solving linear programs in the current matrix multiplication time," *Journal of the ACM (JACM)*, vol. 68, no. 1, pp. 1–39, 2021.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[38] Z. Chen, W. Yi, A. S. Alam, and A. Nallanathan, "Dynamic task software caching-assisted computation offloading for multi-access edge computing," *IEEE Trans. Commun.*, vol. 70, no. 10, pp. 6950–6965, 2022.

[39] Z. Chen, W. Yi, and A. Nallanathan, "Exploring representativity in device scheduling for wireless federated learning," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2023.

[40] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc Machine Learning and Systems*, 2020.