Convergence-Latency-Aware Adaptive Modulation and Resource Allocation in RIS-Assisted Wireless Federated Learning

Liwei Wang, Wen Chen, Jun Li, Qingqing Wu, Ming Ding, Xusheng Zhu and Qiong Wu

arXiv:2607.19759v1 [cs.LG] 22 Jul 2026

***Abstract*—Federated learning (FL) over wireless networks suffers from significant training latency and degraded convergence due to unreliable wireless transmission, especially under blocked propagation environments. Although reconfigurable intelligent surfaces (RISs) can improve communication reliability, existing wireless FL studies rarely characterize the trade-off between learning convergence and communication delay under modulation-dependent transmission errors. In this paper, we consider a wireless FL system operating under RIS-assisted blocked-link propagation scenarios, and focus on adaptive modulation and sub-channel allocation for convergence-latency aware communication design. By characterizing the effect of symbol errors on uploaded local gradients, we derive a convergence-related upper bound that reveals the impact of symbol error rate (SER) on FL loss decay. Based on this result, we formulate a joint convergence-latency optimization problem, which is cast as a mixed-integer nonlinear programming (MINLP) problem, and solve it using a low-complexity hybrid alternating optimization framework. Extensive experiments on MNIST, CIFAR-10, and Speech Commands show that the proposed scheme consistently achieves faster convergence and higher test accuracy than existing adaptive communication schemes, especially in complex tasks and challenging wireless scenarios.**

***Index Terms*—Federated learning, adaptive modulation, reconfigurable intelligent surface (RIS), wireless resource allocation, low-latency communications.**

I. Introduction

With the proliferation of artificial intelligence (AI) applications in domains such as healthcare, finance, and autonomous systems, distributed machine learning frameworks have gained increasing traction [1]. Traditional machine learning approaches involve uploading all raw data to a central server, where centralized model training is then performed on the aggregated dataset. Nevertheless, the traditional centralized learning framework faces the challenge of data isolation, due to the increasing emphasis on privacy issues [2]. Federated learning (FL), in particular, has emerged as a privacy-preserving learning paradigm, enabling collaborative model training across distributed devices without exposing raw data [3]. FL typically necessitates a substantial number of communication iterations to achieve the target model accuracy, a phenomenon that becomes particularly pronounced when the training process involves a relatively large cohort of participating clients [4]. Consequently, the model performance and the latency induced by unreliable wireless transmissions and heterogeneous local computation capacities across clients emerge as a critical bottleneck in wireless FL systems [5]. Moreover, although computational efficiency may be optimized, the inherent long-range and blocked communication between edge devices and remote cloud servers inevitably incurs significant transmission delays and errors, thereby compromising client quality-of-service (QoS).

Latency is one of the most critical performance bottlenecks in wireless FL systems. The local computation delay depends on the data size and computational capability of participating devices, while the uplink transmission delay is influenced by channel conditions, modulation schemes, and allocated bandwidth. The downlink broadcasting delay, on the other hand, is typically smaller compared with the previous two items. Existing studies have investigated several strategies to reduce latency in wireless FL [6]–[12]. The authors in [6] jointly optimize the computation frequency and communication bandwidth while respecting the energy budget to minimize the total FL execution time. In [7], the authors address the latency-minimization problem in FL, explicitly accounting for potentially different privacy protections and data imbalance. Also, model quantization and compression techniques have been proposed to reduce the size of transmitted updates [8], thereby shortening transmission time. A novel FL aggregation mechanism is proposed in [9] to reduce the communication overhead, by selecting more valuable layers to upload by their feedback framework. The authors in [10], [11] propose a dynamic quantization framework that jointly minimizes per-client energy expenditure and improves uplink communication efficiency by adaptively tuning the quantization levels. [12] dynamically assigns heterogeneous quantization resolutions across clients and correspondingly re-weighted their scheduling probabilities, in order to match the resulting uplink heterogeneity and latency optimization. The above works broaden the scope of applying latency optimization in FL. However, these works often focus too much on communication delay and neglect the model accuracy problems caused thereby.

Unreliable communication is also a major challenge that FL faces in practical deployment. The core of FL is distributed

Liwei Wang, Wen Chen and Qingqing Wu are with the Broadband Access Network Laboratory, Shanghai Jiao Tong University, Minhang 200240, China (e-mail: wanglw2000; wenchen; qingqingwu@sjtu.edu.cn).

Jun Li is with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: jun.li@seu.edu.cn).

Ming Ding is with Data61, CSIRO, Sydney, NSW 2015, Australia (e-mail: ming.ding@data61.csiro.au).

Xusheng Zhu is with the Department of Electronic and Electrical Engineering, University College London, WC1E 6BT London, U.K. (e-mail: xusheng.zhu@ucl.ac.uk).

Qiong Wu is with the School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China (e-mail: qiongwu@jiangnan.edu.cn).

collaboration, however, unreliable communication will directly affect its learning efficiency and speed, as well as the quality and convergence of the model [13]. A lot of works have been done to alleviate the negative impacts caused by unreliable wireless communication in FL [14]–[18]. The authors in [14], [15] propose a novel FL framework in which the impact of the wireless packet transmission errors is considered to select clients for a better model performance. A FL framework, FL-GC, is presented in [16], where the server infers the local gradient information of clients with transmission failures or those not selected for participation, utilizing locally uploaded gradients from prior communication rounds. Also, recently reconfigurable intelligence surface (RIS) is regarded as one of the key technologies to address the unreliable issue in future wireless communication [19], [20]. A large amount of work has already applied the RIS to address the research issues related to unreliable communication in FL. The authors in [17] propose a novel performance-centric long-term design scheme that integrates multiple communication rounds, with the primary objective of minimizing the optimality gap associated with the loss function in AirComp via RISs. By converting SINR thresholds to SER to capture key performance constraints, [18] selects local models within an acceptable error range for global aggregation, enabling FL to incorporate more tolerable-error local models and gain performance improvements. Existing efforts have been devoted to achieve higher model accuracy at the cost of increased power consumption and latency, in an attempt to mitigate the impact of unreliable communication as much as possible.

Due to the limited resources of wireless FL, some works focus on achieving better model performance and higher learning efficiency through client scheduling. The authors in [21] and [22] formulate a joint optimization problem of resource allocation and client scheduling, to minimize the energy consumption in the social network scenario and the BS-client scenario respectively. Also, a joint bandwidth allocation and scheduling framework is built in [23] to achieve a good trade-off between the learning efficiency and latency. The work in [24] develops a client scheduling and resource allocation method to minimize the training delay in a novel client mobility network. Moreover, the client scheduling framework is delineated in [25] to achieve the optimal trade-off between channel quality and local data importance.

To sum up, existing studies have investigated latency reduction, communication reliability, and resource management in wireless federated learning. However, the trade-off between convergence and latency under modulation-dependent transmission errors is still not well understood. Motivated by this issue, we consider a wireless FL system where RISs are deployed to assist blocked NLoS transmissions, and focus on adaptive modulation and sub-channel allocation under modulation-dependent transmission errors. We characterize the impact of symbol errors on FL convergence, formulate a joint convergence–latency optimization problem, and develop a corresponding solution framework.

The main contributions of this paper are summarized as follows.

- We consider a wireless FL system under direct LoS and RIS-assisted blocked NLoS transmission scenarios, and establish a communication model that jointly characterizes uplink transmission latency and gradient distortion caused by symbol errors. The proposed model explicitly reveals the reliability-latency trade-off caused by adaptive modulation in wireless FL.
- We derive a convergence-related upper bound that links modulation-dependent symbol errors with the expected FL loss decay. The analysis quantifies how the SER affects gradient aggregation errors and further degrades the convergence behavior, thereby providing an analytical basis for convergence-latency aware modulation and resource allocation.
- We formulate the joint adaptive modulation and sub-channel allocation problem as a mixed-integer nonlinear programming (MINLP) problem and propose a low-complexity hybrid alternating-optimization framework to solve it efficiently. Specifically, continuous relaxation and Newton iterations are adopted for modulation optimization, while binary relaxation and KKT-based optimization are employed for sub-channel allocation. Experiments on MNIST, CIFAR-10, and Speech Commands demonstrate that the proposed design achieves faster convergence and higher accuracy than benchmark schemes.

II. System Model

In this section, we present the considered wireless FL system, where RISs are deployed only to assist blocked NLoS links. We then introduce the learning model, wireless transmission model, channel error model, and adaptive modulation mechanism.

A. FL Model

As shown in Fig. 1, we consider an RIS-assisted wireless FL system consisting of one parameter server (PS), K edge clients, and L RISs with R reflecting elements. For clients whose direct links to the PS are blocked, the uplink transmission is assisted by RIS reflection; otherwise, the direct LoS link is used. Each client $k \in \mathcal{K} = \{1, 2, \ldots, K\}$ owns a local dataset $\mathcal{D}_k$ with size $D_k = |\mathcal{D}_k|$, where each sample is denoted by (u_i, v_i).

For a given model parameter $\boldsymbol{\theta}$, the local loss function for client k is given by

$$F_k(\boldsymbol{\theta}) = \frac{1}{D_k} \sum_{(\boldsymbol{u}_i, v_i) \in \mathcal{D}_k} f(\boldsymbol{\theta}; \boldsymbol{u}_i, v_i), \tag{1}$$

where $D_k = |\mathcal{D}_k|$ that represents the dataset size of the k-th client. $f(\boldsymbol{\theta}; \boldsymbol{u}_i, v_i)$ is the loss function, which is denoted by $f_k(\boldsymbol{\theta})$ hereafter for brevity. The loss function captures different performance of FL algorithm when applying different learning task. We assume that clients have unequal dataset size in the local training, and the global loss function with model parameter $\boldsymbol{\theta}$ is defined as

$$F(\boldsymbol{\theta}) = \frac{1}{D_{all}} \sum_{k=1}^{K} D_k F_k(\boldsymbol{\theta}), \tag{2}$$

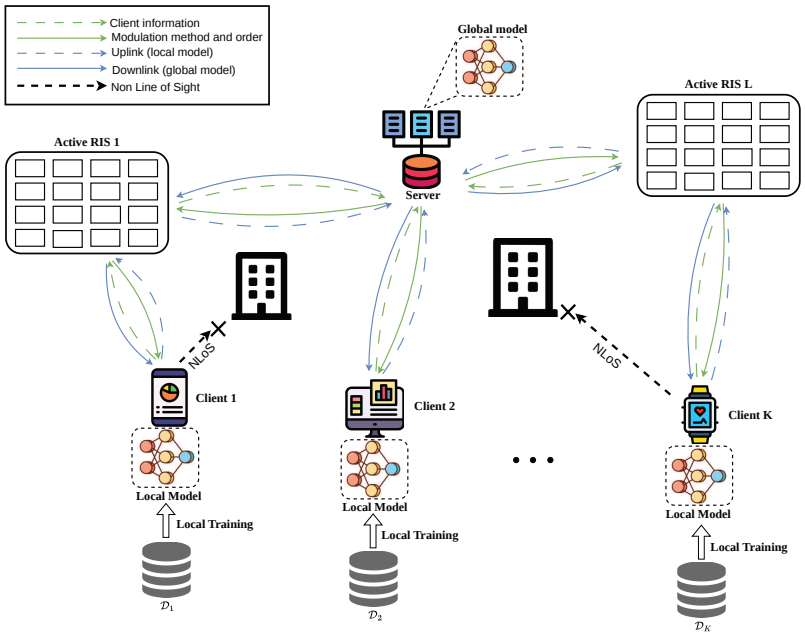


Fig. 1. System model of the RIS-assisted wireless FL system.

where $D_{all} = \sum_k D_k$. The global loss function is weighted average of the local loss function by the dataset size of each client. The FL process aims to optimize the model parameter $\boldsymbol{\theta}$ that minimizes $F(\boldsymbol{\theta})$, i.e.,

$$\boldsymbol{\theta}^* = \arg\min_{\theta} F(\boldsymbol{\theta}). \tag{3}$$

To facilitate cooperative training among edge devices, each client independently computes its local gradients by minimizing $F_k(\boldsymbol{\theta})$ in parallel, and subsequently, the server aggregates these local gradients to derive the global gradient In the t-th communication round, the k-th client evaluates the gradient of its local model with its corresponding local dataset $\mathcal{D}_k$, as expressed by

$$\boldsymbol{g}_k^t = \nabla F_k(\boldsymbol{\theta}^t) = \frac{1}{D_k} \sum_{(\boldsymbol{u}_i, v_i) \in \mathcal{D}_k} \nabla f_k(\boldsymbol{\theta}^t). \tag{4}$$

After receiving the updates of all clients, the PS aggregates all local gradients and computes the global gradient, which is updated as

$$\boldsymbol{g}^t = \nabla F(\boldsymbol{\theta}^t) = \frac{1}{D_{all}} \sum_{k=1}^{K} D_k \boldsymbol{g}_k^t. \tag{5}$$

Then based on the global gradient $\boldsymbol{g}^t$, the PS executes a global model update utilizing the gradient descent algorithm in order to derive an updated global model, which is given by

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta \boldsymbol{g}^t, \tag{6}$$

where η denotes the learning rate.

Due to wireless channel impairments, the uploaded gradients may be distorted by transmission errors. The impact of such errors on FL convergence will be analyzed in Section III.

B. Transmission Model

In our system, the global and local gradients are transmitted via the wireless channel with or without RIS assistance, and we do not consider the reflection time of the signal on RIS. It is considered that orthogonal frequency-division multiple access (OFDMA) is employed for accessing the uplink channel. In the case, the system bandwidth B is divided into N sub-channels, which are assigned to clients without interference. Define B_n as the bandwidth of the sub-channel n and we have $\sum_{n=1}^{N} B_n \leq B$. Since all clients are required to upload their local gradients in each communication round, each client should be allocated at least one sub-channel. Accordingly, we assume that the number of sub-channels satisfies $N \geq K$. Given that the downlink bandwidth is accessible to all clients and the entire system bandwidth can be applied to downlink transmission, we limit our analysis to channel errors occurring only during uplink transmission. We assume that the channel state information is able to be known by the channel estimation algorithm in each communication round.

1) Uplink Transmission: Considering the modulation order, the uplink data rate of client k can be expressed as

$$r_k^U(\boldsymbol{a}_k, \boldsymbol{M}_k) = \sum_{n=1}^{N} a_{k,n} B_n \log_2(M_{k,n}), \tag{7}$$

where $\boldsymbol{a}_k = [a_{k,1}, a_{k,2}, \cdots, a_{k,N}]$ is a sub-channel allocation vector, $a_{k,n} \in \{0, 1\}$ and $a_{k,n} = 1$ indicates that sub-channel n is allocated to client k, and $a_{k,n} = 0$, otherwise. $\boldsymbol{M}_k$ is the

modulation order vector and $M_{k,n}$ is the modulation order of the client k to transmit data on sub-channel n.

Then let Z denote the total number of learning parameters and ϕ denote the quantitative bit number for each parameter. Note that the amount of feedback client information data is small and the communication delay is negligible. Subsequently, the data size of the local gradient can be assessed as $\phi \times Z$, and the latency associated with uploading the local gradient is defined by

$$T_k^U = \frac{\phi Z}{r_k^U(\boldsymbol{a}_k, \boldsymbol{M}_k)} = \frac{\phi Z}{\sum_{n=1}^{N} a_{k,n} B_n \log_2(M_{k,n})}. \tag{8}$$

2) Downlink Transmission: For the downlink channel, we assume that all downlink bandwidth is occupied to broadcast the global model. With BPSK modulation, the achievable downlink data rate can be expressed as

$$r^D = B. \tag{9}$$

Meanwhile, due to the fact that each gradient has its corresponding model parameter, the total number of parameters in the global model to broadcast is Z as well. Then the downlink latency for the global model broadcasting is given by, with the ϕ-bits quantization,

$$T^D = \frac{\phi Z}{B}. \tag{10}$$

3) One Round Latency: In addition to the uplink and downlink transmission delays, we also consider the latency of local model updates. Let f_k(cycle/s) denote the computation capability of client k. Moreover, we define C as the total number of floating-point operations involved in backpropagation with one individual data sample. Then, with D_k, the local gradient calculation latency of the client k is

$$T_k^C = \frac{D_k C}{f_k}. \tag{11}$$

Due to the high computation power of the PS and the small amount of computation of the gradient weighted aggregation, we ignore the time for global model update. Moreover, the PS will not perform the aggregation process until all gradients from the clients are received. Accordingly, the one round latency can be expressed as

$$T = \max_k \{T_k^C + T_k^U\} + T^D. \tag{12}$$

C. Channel Error Model

In our paper, for both direct LoS and RIS-assisted blocked NLoS links, the SER depends on the adopted modulation scheme. Given that the downlink bandwidth is accessible to all clients and allows for the use of more robust channel coding, we focus solely on channel errors occurring during uplink transmission and use the nominal symbol-rate-limited throughput. There are two common modulation schemes that will be considered in this paper, MPSK and MQAM. We denote the signal-to-noise ratio (SNR) of client k on sub-channel n is $\gamma_{k,n} = \frac{p_{k,n}|h_{k,n}|^2}{N_0 B_n}$, where $p_{k,n}$ is the transmit power of client k on sub-channel n, $h_{k,n}$ is the corresponding channel coefficient, and $N_0 B_n$ denotes the noise power over sub-channel n. Then, we characterize the SER under two propagation conditions: direct LoS transmission and RIS-assisted blocked NLoS transmission.

1) LoS scenarios: In this case, we consider that there are no obstructions between the straight channel path from this client to the PS. Following [26], when MPSK is applied, we express the SER of client k on sub-channel n as

$$q_{k,n,M}^{\mathrm{LoS,PSK}} \leq 2Q\left(\sqrt{2\gamma_{k,n}\log_2(M_{k,n})}\sin\left(\frac{\pi}{M_{k,n}}\right)\right), \tag{13}$$

where $0 \leq q_{k,n,M}^{LoS,PSK} \leq 1$, $\forall k \in K$, $h_{k,n}$ is the channel gain between the client k and the PS on sub-channel n, and Q-function $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp\left(-\frac{u^2}{2}\right) du$ that denotes the tail function of the standard normal distribution.

Moreover, when MQAM is applied, we express the SER of client k on sub-channel n as

$$q_{k,n,M}^{\mathrm{LoS,QAM}} \leq 4Q\left(\sqrt{\frac{3\gamma_{k,n}\log_2(M_{k,n})}{M_{k,n}-1}}\right), \tag{14}$$

where $0 \leq q_{k,n,M}^{LoS,QAM} \leq 1$, $\forall k \in K$. Hence, we can express the SER of client k in LoS scenarios as

$$q_k^{LoS,PSK(QAM)} = \frac{\sum_{n=1}^{N} a_{k,n} q_{k,n,M}^{LoS,PSK(QAM)}}{\sum_{n=1}^{N} a_{k,n}}. \tag{15}$$

2) NLoS with RIS-assisted scenarios: In this scenario, we assume there are buildings blocking the channel between the client and the PS, which greatly weakens the transmission strength of the signal. We assume all NLoS clients transmit their signal via RISs, as MPSK is selected, then the SER of client k on sub-channel n is [19]

$$q_{k,n,M}^{NLoS,PSK} = \frac{1}{\pi}\int_0^{\frac{(M-1)\pi}{M}} \mathbb{M}_\gamma\left(\frac{-\sin^2\left(\frac{\pi}{M}\right)}{\sin^2\alpha}\right) d\alpha, \tag{16}$$

where $\mathbb{M}_\gamma$ is the moment generating function (MGF) as [26]

$$\mathbb{M}_\gamma(s) = \left(\frac{1}{1-\frac{sL(16-\pi^2)\gamma_{k,n}}{8}}\right)^{\frac{1}{2}} \exp\left(\frac{\frac{sL^2\pi^2\gamma_{k,n}}{16}}{1-\frac{sL(16-\pi^2)\gamma_{k,n}}{8}}\right). \tag{17}$$

Since $\mathbb{M}_\gamma$ is an increasing function of s, (16) can be upper bounded by setting $\alpha = \frac{M-1}{M}\pi$, which yields

$$q_{k,n,M}^{NLoS,PSK} \leq \frac{M-1}{M}\left(\frac{1}{1+\frac{\beta L(16-\pi^2)\gamma}{8}}\right)^{\frac{1}{2}} \exp\left(\frac{\frac{-\beta L^2\pi^2\gamma}{16}}{1+\frac{\beta L(16-\pi^2)\gamma}{8}}\right), \tag{18}$$

where $\beta = \frac{\sin^2(\pi/M)}{\sin^2((M-1)\pi/M)}$. From (18), we can see that with RIS-assisted, even though the SNR γ is relatively low, the average SER remains rather low.

Moreover, by utilizing the above-mentioned MGF, we are also able to derive the SER for square of client k on sub-channel n for MQAM constellations [19]

$$\begin{aligned} & q_{k,n,M}^{NLoS,QAM} \\ =& \frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)\int_0^{\pi/2}\mathbb{M}_\gamma\left(\frac{-3}{2(M-1)\sin^2\alpha}\right)d\alpha \\ & -\frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)^2\int_0^{\pi/4}\mathbb{M}_\gamma\left(\frac{-3}{2(M-1)\sin^2\alpha}\right)d\alpha. \quad (19) \end{aligned}$$

Similarly, we can obtain an upper bound for (19) by setting $\alpha=\frac{\pi}{2}$ in the first term and $\alpha=\frac{\pi}{4}$ in the second term respectively, which yields

$$\begin{aligned} q_{k,n,M}^{NLoS,QAM} \le & 2\left(1-\frac{1}{\sqrt{M}}\right)\mathbb{M}_\gamma\left(\frac{-3}{2(M-1)}\right) \\ & -\left(1-\frac{1}{\sqrt{M}}\right)^2\mathbb{M}_\gamma\left(\frac{-6}{M-1}\right). \quad (20) \end{aligned}$$

Then we can also express the SER of client k in NLoS scenarios as

$$q_k^{NLoS,PSK(QAM)}=\frac{\sum_{n=1}^N a_{k,n}q_{k,n,M}^{NLoS,PSK(QAM)}}{\sum_{n=1}^N a_{k,n}}. \quad (21)$$

III. Performance Analysis and Problem Formulation

The modulation order affects both the uplink transmission latency and the SER-induced gradient distortion, which motivates the convergence-latency analysis.

In this section, we will first analyze the convergence behavior and how the channel error affects the performance of FL. To obtain a tractable convergence-related characterization of the impact of wireless transmission errors, we make several assumptions on the loss function $F(\boldsymbol{\theta})$. Then, the generalized optimality gap, which is suitable for generic wireless networks, is derived to characterize the learning efficiency between two arbitrary communication rounds. And the obtained optimality gap sheds light on how the imperfect gradient updates affect the convergence rate of FL. Next, we focus on the effect of RIS-assisted system channel error rate on FL convergence performance and formulate the corresponding optimization problem with adaptive modulation and sub-channel allocation mechanism.

A. Performance Analysis

For analysis, we make the following assumptions.

Assumption 1(ℓ-smooth): The global loss function $F(\boldsymbol{\theta})$ is smooth, at any point $\boldsymbol{\theta}\in\mathbb{R}^d$, with positive constant $\ell>0$, i.e.,

$$\left\|\nabla F(\boldsymbol{\theta})-\nabla F(\boldsymbol{\theta}')\right\|\le\ell\left\|\boldsymbol{\theta}-\boldsymbol{\theta}'\right\|, \quad \forall\boldsymbol{\theta},\boldsymbol{\theta}'\in\mathbb{R}^d, \quad (22)$$

where $\|\cdot\|$ is the *L2*-norm operator.

Assumption 2(convex): We assume that $F(\boldsymbol{\theta})$ is convex on $\boldsymbol{\theta}$, such that

$$F(\boldsymbol{\theta})\ge F\left(\boldsymbol{\theta}'\right)+\nabla F\left(\boldsymbol{\theta}'\right)^\top\left(\boldsymbol{\theta}-\boldsymbol{\theta}'\right), \quad \forall\boldsymbol{\theta},\boldsymbol{\theta}'\in\mathbb{R}^d, \quad (23)$$

where $\top$ stands for matrix transpose.

Due to the channel errors, we assume that there is a theoretical global gradient g^t obtained from the weighted average of the PS and the actual global gradient $\widehat{g}^t$ via the channel. With *Assumption 1,2*, it is easy to get

$$F(\boldsymbol{\theta})\le F\left(\boldsymbol{\theta}'\right)+\nabla F\left(\boldsymbol{\theta}'\right)^\top\left(\boldsymbol{\theta}-\boldsymbol{\theta}'\right)+\frac{\ell}{2}\left\|\boldsymbol{\theta}-\boldsymbol{\theta}'\right\|^2. \quad (24)$$

Considering $\widehat{g}^t$ with channel errors, based on (6), we can get $\boldsymbol{\theta}^{t+1}=\boldsymbol{\theta}^t-\eta\widehat{\boldsymbol{g}}^t$. Then replacing $\boldsymbol{\theta}$, $\boldsymbol{\theta}'$ with $\boldsymbol{\theta}^{t+1}$, $\boldsymbol{\theta}^t$ and applying it into (24), we have

$$F\left(\boldsymbol{\theta}^{t+1}\right)\le F\left(\boldsymbol{\theta}^t\right)+\left(\boldsymbol{g}^t\right)^\top\left(-\eta\widehat{\boldsymbol{g}}^t\right)+\frac{\ell\eta^2}{2}\left\|\widehat{\boldsymbol{g}}^t\right\|^2. \quad (25)$$

Then we take expectations on both sides of the inequality, it comes that

$$\begin{aligned} & \mathbb{E}\left\{F\left(\boldsymbol{\theta}^{t+1}\right)\right\} \\ & =\mathbb{E}\left\{F\left(\boldsymbol{\theta}^t\right)\right\}-\mathbb{E}\Big\{\left(\eta-\ell\eta^2\right)\left(\boldsymbol{g}^t-\widehat{\boldsymbol{g}}^t\right)^\top\boldsymbol{g}^t \\ & \qquad -\left(\eta-\frac{\ell\eta^2}{2}\right)\left\|\boldsymbol{g}^t\right\|^2+\frac{\ell\eta^2}{2}\left\|\boldsymbol{g}^t-\widehat{\boldsymbol{g}}^t\right\|^2\Big\} \\ & \overset{(a)}{=}\mathbb{E}\left\{F\left(\boldsymbol{\theta}^t\right)\right\}+\frac{1}{2\ell}\mathbb{E}\Big\{\left\|\boldsymbol{g}^t-\widehat{\boldsymbol{g}}^t\right\|^2\Big\}-\frac{1}{2\ell}\left\|\boldsymbol{g}^t\right\|^2, \quad (26) \end{aligned}$$

where (a) is because we set the learning rate $\eta=\frac{1}{\ell}$. Then incorporating (5) and subtracting $\mathbb{E}\{F(\boldsymbol{\theta}^*)\}$ in both sides of the above inequality, where $\boldsymbol{\theta}^*$ is the optimal global FL model, we have the lower bound as

$$\begin{aligned} & \mathbb{E}\left\{F\left(\boldsymbol{\theta}^{t+1}\right)-F\left(\boldsymbol{\theta}^*\right)\right\} \\ & \le\mathbb{E}\left\{F\left(\boldsymbol{\theta}^t\right)-F\left(\boldsymbol{\theta}^*\right)\right\}-\frac{1}{2\ell D_{all}^2}\sum_{k=1}^K D_k^2\big(\left\|\boldsymbol{g}_k^t\right\|^2 \\ & \qquad -\mathbb{E}\Big\{\left\|\boldsymbol{g}_k^t-\widehat{\boldsymbol{g}}_k^t\right\|^2\Big\}\big). \quad (27) \end{aligned}$$

The above inequality indicates that the anticipated difference between the global loss function value and the optimal loss is constrained by the sum of three components. The first term on the right side of (27) represents the expected difference from the prior communication round. And the second term is directly related to the squared norm of the true local gradient $\boldsymbol{g}_k^t$ with the influence of its dataset size D_k. We can see that the first two terms are independent of the modulation allocating scheme. The third term is proportional to the gradient error of the aggregated global gradient, which is contingent upon the modulation order allocating design and necessitates optimization. Consequently, a reduction in the error of the aggregated global gradient will lead to a more rapid decrease in the global loss. Next, we investigate the relationship between SER and the rate of loss decay.

We first consider that only one symbol will make an error in wireless transmission and we denote $g_{k,z}^t$ as the theoretical local gradient of z-th parameter of client k and $\widehat{g}_{k,z}^t$ as the actual global gradient via the channel. Based on the lemma in

[27], the expectation of $\widehat{g}_{k,z}^t$ is derived as

$$\begin{aligned}
&\mathbb{E}\left\{\widehat{g}_{k,z}^t \middle| g_{k,z}^t\right\} \\
=&g_{k,z}^t + \frac{1}{\sigma}\{\underbrace{q_k(1-q_k)^{\sigma-1}\frac{range_g}{2^\phi-1}\sum_{i=\phi-\log_2 M}^{\phi-1}(1-2g_{k,z}^{t(i)})2^i}_{\text{bias by the first symbol error}} \\
&+\cdots+\underbrace{q_k(1-q_k)^{\sigma-1}\frac{range_g}{2^\phi-1}\sum_{i=0}^{\phi-(\sigma-1)\log_2 M-1}(1-2g_{k,z}^{t(i)})2^i}_{\text{bias by the last symbol error}}\} \\
=&g_{k,z}^t + \frac{range_g\cdot q_k(1-q_k)^{\sigma-1}}{\sigma(2^\phi-1)}\sum_{i=0}^{\phi-1}(1-2g_{k,z}^{t(i)})2^i, \qquad (28)
\end{aligned}$$

where $\sigma = \lceil \phi/\log_2 M_{k,n} \rceil$ represents the number of symbols contained in a parameter and each symbol contains $\log_2 M_{k,n}$ bits. For a given modulation order, σ is fixed when characterizing the symbol-error-induced gradient distortion. $range_g \triangleq \max(g) - \min(g)$ and we denote $g_{k,z}^{t(i)} \in \{0,1\}$ as the i-th digit. It is emphasized that the second component on the right of (28) represents a non-zero expected bias, which arises due to the presence of random bit errors. Hence there is $\mathbb{E}\{(1-2g_{k,z}^{t(i)})^2(2^i)^2\} = (2^i)^2$. Let ϵ be the bias for the z-th parameter of the k-th client, i.e., $\epsilon = g_{k,z}^t - \widehat{g}_{k,z}^t$, then we get the mean square bias as follows

$$\Delta \triangleq \mathbb{E}\{\,\|\epsilon\|^2\,\} = q_k(1-q_k)^{\sigma-1}\frac{range_g^2(4^\phi-1)}{3(2^\phi-1)^2}. \qquad (29)$$

The above formula calculates the bias in case of a single symbol error in each parameter. It is known that errors in symbols are independent of each other, with n symbol errors existing, we have

$$\begin{aligned}
&\mathbb{E}\left\{\left\|g_{k,z}^t - \widehat{g}_{k,z}^t\right\|^2\right\} \\
=&\mathbb{E}\left\{\left\|\sum_{i=1}^n \epsilon_i\right\|^2\right\} \overset{(b)}{\leq} n\mathbb{E}\left\{\sum_{i=1}^n \|\epsilon_i\|^2\right\} = n^2\Delta, \qquad (30)
\end{aligned}$$

where ϵ_i represents the mean square error of the i-th wrong symbol. The (b) is deduced from Jensen's inequality. Applying the probability of existing n symbol errors in a parameter (σ symbols), we can continue to deduce that the average error of the gradient is

$$\begin{aligned}
&\mathbb{E}\left\{\left\|g_{k,z}^t - \widehat{g}_{k,z}^t\right\|^2\right\} \\
\leq&\sum_{n=0}^{\sigma}\binom{\sigma}{n}(1-q_k)^{\sigma-n}{q_k}^n n^2\Delta \\
=&\sum_{n=0}^{\sigma}\binom{\sigma}{n}(1-q_k)^{\sigma-n}{q_k}^n n(n-1)\Delta \\
&+\sum_{n=0}^{\sigma}\binom{\sigma}{n}(1-q_k)^{\sigma-n}{q_k}^n n\Delta. \qquad (31)
\end{aligned}$$

We make $n(n-1) = 2\binom{n}{2}$ and $n = \binom{n}{1}$. With basic computational properties of combinatorial numbers, i.e., $\binom{a}{b}\binom{b}{c} = \binom{a-c}{b-c}\binom{a}{c}$, we have

$$\begin{aligned}
&\mathbb{E}\left\{\left\|g_{k,z}^t - \widehat{g}_{k,z}^t\right\|^2\right\} \\
\leq&\sigma(\sigma-1)\Delta\sum_{n=2}^{\sigma}\binom{\sigma-2}{n-2}(1-q_k)^{\sigma-n}{q_k}^n \\
&+\sigma\Delta\sum_{n=1}^{\sigma}\binom{\sigma-1}{n-1}(1-q_k)^{\sigma-n}{q_k}^n \\
\overset{(c)}{=}&\sigma\Delta q_k\left[(\sigma-1)\,q_k+1\right], \qquad (32)
\end{aligned}$$

where (c) is deduced from binomial theorem. Next, we conduct the calculation on the level of gradient vectors and we can get the expectation of the average gradient error for the k-th client as follows

$$\begin{aligned}
\mathbb{E}\left\{\left\|g_k^t - \hat{g}_k^t\right\|^2\right\} &= \sum_z \mathbb{E}\left\{\left\|g_{k,z}^t - \hat{g}_{k,z}^t\right\|^2\right\} \\
&\leq \sigma Z\frac{\text{range}_g^2(4^\phi-1)}{3(2^\phi-1)^2}\left((\sigma-1)q_k^3+q_k^2\right) \\
&\leq \xi\left\|g_k^t\right\|^2\left((\sigma-1)q_k^3+q_k^2\right), \qquad (33)
\end{aligned}$$

where $\xi = \frac{\sigma Z(4^\phi-1)}{3(2^\phi-1)^2}$, which is determined by the quantitative bit number and the total number of parameters. Substituting (33) into (27), we have the following theorem as

Theorem 1: Given the learning rate $\eta = \frac{1}{\ell}$, the SER q_k of client k, the dataset size D_{all} of clients participated in local training, the dataset size D_k of client k and the optimal global FL model $\boldsymbol{\theta}^*$,with the assumptions, the bound of $\mathbb{E}\left\{F\left(\boldsymbol{\theta}^{t+1}\right) - F\left(\boldsymbol{\theta}^*\right)\right\}$ can be given by

$$\begin{aligned}
&\mathbb{E}\left\{F\left(\boldsymbol{\theta}^{t+1}\right) - F\left(\boldsymbol{\theta}^*\right)\right\} \leq \mathbb{E}\left\{F\left(\boldsymbol{\theta}^{t}\right) - F\left(\boldsymbol{\theta}^*\right)\right\} \\
&\underbrace{-\frac{1}{2\ell D_{all}^2}\sum_{k=1}^K\left(D_k\|\boldsymbol{g}_k^t\|\right)^2\left\{1-\xi\left[(\sigma-1)\,q_k^3+q_k^2\right]\right\}}_{\text{Impact of wireless factors on FL convergence, defined as } \delta \text{ in III.B}}. \qquad (34)
\end{aligned}$$

Compared to (27), we can see that the annotated term, the gradient error of the aggregated global gradient, is denoted by q_k. Specifically, an increase in q_k leads to a greater discrepancy between the theoretical global gradient $\boldsymbol{g}^t$ and the actual global gradient $\widehat{\boldsymbol{g}}^t$, leading to a reduction in the efficacy of loss decay. Therefore, a higher modulation order may reduce uplink latency but can also increase gradient distortion and slow down FL convergence.

This result provides the analytical basis for the convergence-latency aware adaptive modulation and resource allocation design in the following subsection. It should be noted that the bound is not claimed to prove convergence for non-convex neural networks; rather, it motivates a surrogate communication metric that penalises SER-induced gradient distortion. For non-convex neural networks, the bound should be interpreted as indicating the relative impact of modulation-induced gradient distortion rather than as a strict convergence guarantee.

B. Problem Formulation

From (34), we have seen that the efficacy of loss decay is mainly related to client gradient modulus, client data volume and SER. Therefore, we define the loss decay as δ. The larger δ is, the faster the loss function decreases, which is capable of assessing the rate of convergence while ensuring

the minimal decay of global loss. Moreover, there is a positive correlation among $D_k\|\boldsymbol{g}_k^t\|^2$ and δ, which is in line with common sense, i.e., a larger data volume and a greater gradient value contribute more significantly to the model update. In our *Section II.C*, we have presented the theory that the modulation order has an impact on both the transmission rate and SER. A higher modulation order can achieve a higher data transmission rate, but it also leads to greater signal distortion and a higher SER. Allocating different sub-channels to clients can result in varying SER and transmission delay. Therefore, we construct an objective function with the aim of achieving a faster model convergence rate with the smallest possible delay, thereby ensuring the effectiveness of the model. In each communication round, we denote

$$\begin{aligned} obj =& \delta - \lambda T \\ =& \frac{1}{2\ell D_{all}^2} \sum_{k=1}^{K} \left(D_k \left\|\boldsymbol{g}_k^t\right\|\right)^2 \left\{1 - \xi \left[(\sigma-1)\, q_k^3 + q_k^2\right]\right\} \\ & - \lambda \left\{ \max_k \{T_k^C + T_k^U\} + T^D \right\}, \end{aligned} \tag{35}$$

where $\lambda \in [0, \lambda_{max})$ is a minimal weight coefficient, which serves to slightly punish the delay, ensuring that the solution with a smaller delay is preferred among solutions with similar convergence speeds. Consequently, the joint optimization of modulation selection and resource allocation is imperative, with the enhancement of (35) serving as the key design objective.

According to what we described in *Section II*, the interplay between modulation selection and spectrum allocation has been identified as a critical factor influencing the key performance metrics of our FL system. To address these technical challenges, our optimization strategy focuses on identifying optimised combinations of modulation schemes and spectrum resource allocation patterns that improves the convergence-latency objective, particularly under the constraints of unreliable wireless communication channels. The fundamental optimization challenge lies in achieving an optimal balance between system convergence rate and learning latency, which is formulated as

$$\mathcal{P}: \max_{\boldsymbol{a}_k, \boldsymbol{M}_k} obj = \delta - \lambda T, \tag{36}$$

$$\text{s.t.} \quad a_{k,n} \in \{0,1\}, \forall k \in \mathcal{K}, \forall n, \tag{36a}$$

$$\sum_{k=1}^{K} a_{k,n} \leq 1, \forall n, \tag{36b}$$

$$\sum_{n=1}^{N} a_{k,n} \geq 1, \quad \forall k \in \mathcal{K}, \tag{36c}$$

$$q_k \leq q_{\max}, \forall k \in \mathcal{K}, \tag{36d}$$

$$M_{k,n} = 2^j, j \in \mathbb{Z}_{>0} | (MPSK), \tag{36e}$$

$$M_{k,n} = 4^j, j \in \mathbb{Z}_{>0} | (MQAM), \tag{36f}$$

Remark 1: The SER threshold $q_{\max}$ in (36d) is selected to ensure a positive convergence-related loss decay. According to Theorem 1, it is sufficient to choose $q_{\max}$ such that $1 - \xi\left((\sigma_{\max}-1)q_{\max}^3 + q_{\max}^2\right) > 0$, where $\sigma_{\max}$ denotes the largest possible value of $\sigma_{k,n}$ over the considered modulation set. Under this condition, the convergence-related term in Theorem 1 remains positive for all feasible clients.

The problem can be characterized as a discrete optimization problem, which involves the joint optimization of wireless resource allocation ($a_{k,n}$) and modulation order ($M_{k,n}$). In the following section, we will specifically elaborate on how we will solve the optimization problem (36).

IV. Problem Optimization

The optimization problem is a mixed-integer nonlinear programming (MINLP) problem due to the binary sub-channel allocation variables, discrete modulation orders, nonlinear SER functions, and the max-latency term. To obtain a tractable solution, we adopt a hybrid alternating optimization framework, where different optimization techniques are employed for different subproblems. Specifically, given the sub-channel allocation, the modulation-order subproblem is solved via continuous relaxation and Newton iterations with discrete projection. Given the modulation order, the sub-channel allocation is optimized through binary relaxation followed by KKT-based optimization. The two subproblems are solved alternately until convergence.

A. Modulation Order Optimization

The objective function for modulation order optimization is given by

$$obj\left(M_{k,n}\right) = \delta\left(M_{k,n}\right) - \lambda T\left(M_{k,n}\right), \tag{37}$$

where $\delta\left(M_{k,n}\right)$ is the loss decay function, reflecting the impact of gradient errors on model convergence, and $T\left(M_{k,n}\right)$ is the communication latency due to the modulation scheme. To simplify the analysis of communication error impact, we consider the uplink SER expression, which is explained in Appendix A.

Since the modulation order is selected from a finite discrete set, the modulation order optimization subproblem is combinatorial in nature. To obtain a low-complexity solution, we first solve a continuous relaxation, where $M_{k,n}$ is allowed to vary within $[M_{\min}, M_{\max}]$. For the relaxed problem, the first-order stationary condition is given by

$$f(M_{k,n}) \triangleq \frac{\partial obj(M_{k,n})}{\partial M_{k,n}} = 0. \tag{38}$$

The continuous solution is obtained by the Newton-Raphson method as

$$M_{k,n}^{(i+1)} = \Pi_{[M_{\min}, M_{\max}]} \left(M_{k,n}^{(i)} - \frac{f(M_{k,n}^{(i)})}{f'(M_{k,n}^{(i)})} \right), \tag{39}$$

where $f'(M_{k,n})$ denotes the derivative of $f(M_{k,n})$, and $\Pi_{[M_{\min}, M_{\max}]}(\cdot)$ is the projection operator that guarantees the feasibility of the relaxed solution. The obtained solution is regarded as a stationary point of the relaxed modulation subproblem. Since $\sigma_{k,n} = \lceil \phi / \log_2(M_{k,n}) \rceil$ is a piecewise constant function, it is treated as locally fixed during each Newton update and recalculated after each iteration.

After convergence, the obtained stationary solution $\widetilde{M_{k,n}}$ is projected onto the discrete modulation set. To avoid infeasible high-order modulation, we do not simply select the nearest

discrete value. Instead, the neighboring discrete modulation orders around $\widetilde{M}_{k,n}$ are evaluated, and the final modulation order is selected as

$$M_{k,n}^{\star} = \arg \max_{M \in \mathcal{M}_{k,n}^{\mathrm{feas}} \cap \mathcal{N}(\widetilde{M}_{k,n})} obj(M), \tag{40}$$

where $\mathcal{N}(\widetilde{M}_{k,n})$ denotes the neighboring discrete modulation orders of $\widetilde{M}_{k,n}$, and

$$\mathcal{M}_{k,n}^{\mathrm{feas}} = \{M \in \mathcal{M} : q_{k,n}(M) \le q_{\max}\}. \tag{41}$$

If no feasible modulation order exists in $\mathcal{N}(\widetilde{M}_{k,n})$, the modulation order is gradually reduced until the SER constraint is satisfied.

B. Sub-Channel Allocation Optimization

Given the optimized modulation orders, we then optimize the sub-channel allocation policy. Since the original sub-channel allocation variable $a_{k,n}$ is binary, the resulting problem is a combinatorial optimization problem. To obtain a tractable solution, we first relax the binary constraint as $0 \le a_{k,n} \le 1, \quad \forall k \in \mathcal{K}, \ \forall n \in \mathcal{N}$.

After solving the relaxed problem, a binary recovery operation is performed to obtain the final sub-channel allocation.

For notational simplicity, we define the convergence-related utility of allocating sub-channel n to client k as

$$\alpha_{k,n} = \frac{D_k^2 \|g_k^t\|^2}{2\ell D_{\mathrm{all}}^2} \left[1 - \xi \left((\sigma_{k,n} - 1) q_{k,n}^3 + q_{k,n}^2\right)\right], \tag{42}$$

where $q_{k,n}$ denotes the SER of client k on sub-channel n under the selected modulation order, and $\sigma_{k,n} = \lceil \phi / \log_2(M_{k,n}) \rceil$. Moreover, the effective uplink transmission rate of client k is given by

$$R_k(\mathbf{a}_k) = \sum_{n=1}^{N} a_{k,n} B_n \log_2(M_{k,n}). \tag{43}$$

Accordingly, the uplink latency of client k can be rewritten as

$$T_k^U(\mathbf{a}_k) = \frac{\phi Z}{R_k(\mathbf{a}_k)}. \tag{44}$$

To handle the non-smooth maximum latency term, we introduce an auxiliary variable τ satisfying

$$T_k^C + \frac{\phi Z}{R_k(\mathbf{a}_k)} \le \tau, \quad \forall k \in \mathcal{K}. \tag{45}$$

Then, the relaxed sub-channel allocation problem can be formulated as

$$\mathbf{P}_{\mathrm{RA}}: \quad \max_{\mathbf{a}, \tau} \quad \sum_{k=1}^{K} \sum_{n=1}^{N} a_{k,n} \alpha_{k,n} - \lambda \tau \tag{46}$$

$$\text{s.t.} \quad T_k^C + \frac{\phi Z}{\sum_{n=1}^{N} a_{k,n} B_n \log_2(M_{k,n})} \le \tau, \quad \forall k, \tag{46a}$$

$$\sum_{k=1}^{K} a_{k,n} \le 1, \quad \forall n, \tag{46b}$$

$$\sum_{n=1}^{N} a_{k,n} \ge 1, \quad \forall k, \tag{46c}$$

$$0 \le a_{k,n} \le 1, \quad \forall k, n. \tag{46d}$$

Algorithm 1 Joint Adaptive Modulation and Sub-Channel Allocation

Require: Client set $\mathcal{K}$, sub-channel set $\mathcal{N}$, channel information, dataset sizes $\{D_k\}$, local gradient norms $\{\|g_k^t\|\}$, modulation set $\mathcal{M}$, SER threshold $q_{\max}$, trade-off parameter λ, maximum iteration number $I_{\max}$, and convergence tolerance ϵ.

Ensure: Sub-channel allocation $\mathbf{a}^{\star}$ and modulation order $\mathbf{M}^{\star}$.

1: Initialize a feasible sub-channel allocation $\mathbf{a}^{(0)}$ and set $i = 0$.
2: Compute the initial objective value $obj^{(0)}$.
3: **repeat**
4: Given $\mathbf{a}^{(i)}$, obtain the modulation order $\mathbf{M}^{(i+1)}$ according to the modulation order optimization method in Section IV-A.
5: Given $M^{(i+1)}$, solve the relaxed sub-channel allocation problem in Section IV-B.
6: Recover $a^{(i+1)}$ and correct infeasible assignments to satisfy the allocation and SER constraints.
7: Compute the objective value $obj^{(i+1)}$ according to (35).
8: Set $i = i + 1$.
9: **until** $|obj^{(i)} - obj^{(i-1)}| \le \epsilon$ or $i \ge I_{\max}$
10: Set $\mathbf{a}^{\star} = \mathbf{a}^{(i)}$ and $\mathbf{M}^{\star} = \mathbf{M}^{(i)}$.
11: **return** $\mathbf{a}^{\star}$ and $\mathbf{M}^{\star}$.

Problem $\mathbf{P}_{\mathrm{RA}}$ is a continuous relaxation of the original sub-channel allocation problem. Compared with the original formulation, the auxiliary variable τ transforms the maximum latency term into a set of smooth inequality constraints, which enables the application of KKT-based analysis.

The optimality conditions of $\mathbf{P}_{\mathrm{RA}}$ are derived in Appendix B. Based on the obtained relaxed solution $\tilde{a}_{k,n}$, we recover the binary allocation by assigning each sub-channel to the client with the largest relaxed allocation value, i.e.,

$$a_{k,n}^{\star} = \begin{cases} 1, & k = \arg\max\limits_{j \in \mathcal{K}} \tilde{a}_{j,n}, \\ 0, & \text{otherwise.} \end{cases} \tag{47}$$

After binary recovery, the SER constraint is checked for each client, and infeasible assignments are corrected by sub-channel reassignment or modulation-order reduction. Since the KKT analysis is performed for the relaxed problem, the recovered binary solution is a low-complexity feasible solution rather than a globally optimal solution to the original mixed-integer problem.

Algorithm 1 summarizes the overall procedure, whose complexity is $O(I(KNR_N + C_{\mathrm{RA}}))$, where I, R_N, and C_{RA} denote the numbers of alternating iterations, Newton iterations, and the complexity of the relaxed allocation subproblem, respectively.

V. Experimental Results

In this section, we evaluate the performance of the proposed adaptive modulation and resource allocation scheme through extensive simulations. The proposed method is compared with several benchmark schemes to demonstrate its effectiveness.

TABLE I
MAIN SIMULATION PARAMETERS

Parameter	Value
Total bandwidth B	10 MHz
Number of RIS reflecting elements R	16
Noise power spectral density N_0	-174 dBm/Hz
Path-loss model $PL[dB]$	$128.1 + 37.6\log(d)$
Trade-off parameter λ	10^{-1}
Number of clients K	$\{10, 20, 30, 40, 50\}$
Average SNR (dB)	$\{10, 15, 20, 25\}$
Dirichlet parameter ρ	0.6
Average number of baseline experiments	5

A. Experiment Setup

We consider a RIS-assisted wireless federated learning system consisting of one parameter server and K edge clients. The system bandwidth is divided into N orthogonal subchannels using OFDMA. We set the channel path loss exponent model to PL, where $PL[dB] = 128.1 + 37.6\log(d)$ with d representing the distance between clients and the PS in kilometer. The LoS condition is modeled as a distance-dependent random process following the 3GPP TR 36.828 specification [28]. For the link with distance d, the LoS probability is given by $P_{\text{LoS}}(d) = \min\left(\frac{a}{d}, 1\right)\left(1 - e^{-d/b}\right) + e^{-d/b}$, where a and b are scenario-dependent constants. Unless otherwise specified, the remaining simulation parameters are summarized in Table I.

The local datasets are uniformly and randomly partitioned among participating clients, resulting in IID and Non-IID data distribution. We examine the result of our proposed algorithm, using the following neural networks and datasets.

- **MLP with MNIST.** The MNIST consists of 70,000 grayscale handwritten digit images of size 28×28, including 60,000 training samples and 10,000 test samples across 10 classes (digits 0–9).
- **CNN with CIFAR-10.** The CIFAR-10 contains 60,000 color images of size 32×32 from 10 object categories, with 50,000 training samples and 10,000 test samples. Each image consists of three RGB channels.
- **CNN with Speech Commands.** The Google Speech Commands Dataset consists of one-second audio recordings of spoken keywords sampled at 16 kHz, collected from a large number of speakers. We adopt a standard 12-class classification setting, including 10 target commands along with additional "unknown" and "silence" classes.

We additionally consider a Dirichlet non-IID partition on CIFAR-10. Specifically, the class proportions of each client are sampled from a Dirichlet distribution with concentration parameter $\rho = 0.6$. In the following experiment, the comparison between Proposed (adaptive modulation-only scheme) and Proposed+resource allocation (RA) serves as an ablation study to evaluate the additional gain brought by resource allocation. Also, to evaluate the performance, we set some baselines to compare with the proposed scheme:

- **BPSK, QPSK, 16QAM.** Fixed-order modulation method.
- **SDPR [18].** A SINR-based selection mechanism, SDPR, which selects clients based on their SINRs to allow more clients with minor errors to participate in FL.
- **No-RIS.** The RIS is removed from the considered system, and blocked or weak NLoS links suffer from more severe attenuation and higher transmission errors, leading to degraded FL performance.

B. Cross-Dataset Performance Analysis

To comprehensively evaluate the effectiveness of the proposed scheme, we further compare its performance across multiple learning tasks with different levels of complexity, including MNIST, CIFAR-10, and Speech Commands.

From the results in Fig. 2, Fig. 3, Fig. 4, several important observations can be drawn. First, all schemes exhibit a similar convergence pattern across different datasets, where the test accuracy increases rapidly at the early stage and gradually saturates as the communication latency accumulates. For example, on the MNIST dataset, all schemes reach above 95% accuracy within approximately 20 s, while on CIFAR-10 and Speech Commands, a greater latency is required to achieve stable convergence due to increased task complexity.

Second, as the task becomes more complex, the performance gap among different schemes becomes more evident. Taking the IID data as an example, on MNIST, the final accuracy gap between the Proposed+RA scheme and the best baseline is relatively small (around 0.3%). In contrast, on CIFAR-10, the Proposed+RA scheme achieves approximately 85.8%, outperforming the SDPR [18] method (around 85.5%) and fixed-modulation schemes (around 85.1%) by a noticeable margin. This gap further enlarges in the Speech Commands task, where Proposed+RA reaches about 92.5%, compared with 92% for the Proposed scheme, 91.5% for SDPR [18], 91.4% for QPSK and 91% for 16QAM.

Third, fixed-modulation schemes exhibit distinct limitations under complex tasks. Although QPSK and 16QAM achieve competitive performance on MNIST, their degradation becomes more obvious on CIFAR-10 and Speech Commands. In particular, 16QAM consistently yields the lowest accuracy due to its higher symbol error rate, while QPSK provides better reliability but lower spectral efficiency. For instance, on the Speech Commands, the accuracy gap between QPSK and 16QAM is about 0.4%, reflecting the trade-off between reliability and transmission rate. Also, in Fig. 3, compared with the IID setting, all schemes exhibit slower convergence and lower test accuracy under the Non-IID setting. This is because heterogeneous local data distributions increase the inconsistency among local updates, which makes global aggregation more sensitive to both gradient bias and transmission errors. As a result, the performance gap among different communication strategies becomes more pronounced.

In contrast, the proposed adaptive modulation scheme consistently outperforms all fixed-modulation baselines across all datasets. The advantage becomes increasingly significant as task complexity grows. For example, the performance gain of the Proposed scheme over fixed-modulation baselines is marginal on MNIST but increases to approximately 0.5% on CIFAR-10 and further to around 1% on Speech Commands. This trend indicates that adaptive modulation is particularly

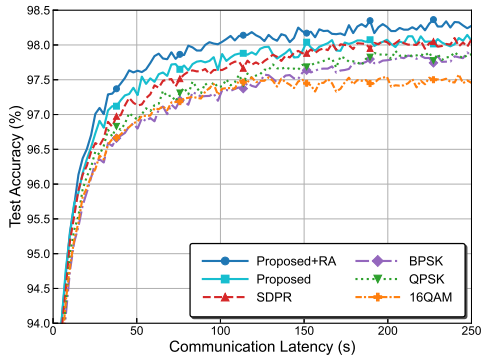
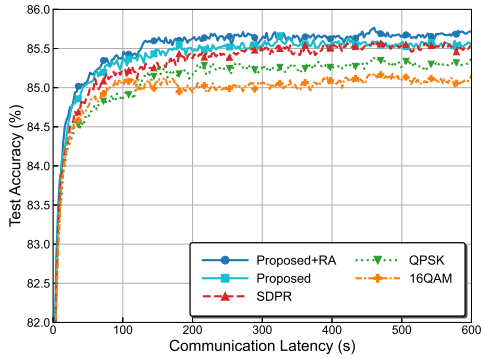
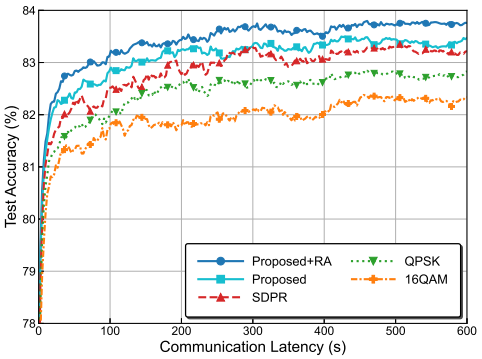


Fig. 2. Performance on MNIST.



(a) IID setting



(b) Dirichlet Non-IID setting

Fig. 3. Performance on CIFAR-10 under different data distributions.

effective in scenarios where gradient distortion has a stronger impact on model convergence. Moreover, Proposed+RA further improves over the adaptive modulation-only scheme by about 0.3% on CIFAR-10 and 0.5% on Speech Commands, indicating that resource allocation becomes more beneficial when task complexity and communication contention increase. These results are consistent with Theorem 1, since reducing SER-induced gradient distortion improves the convergence-related loss decay factor.

C. System-Level Performance Analysis

To further evaluate the performance of our proposed scheme under dynamic wireless environments, we investigate its behavior with respect to the number of participating clients and the channel quality (SNR), as shown in Fig. 5 and Fig. 6.

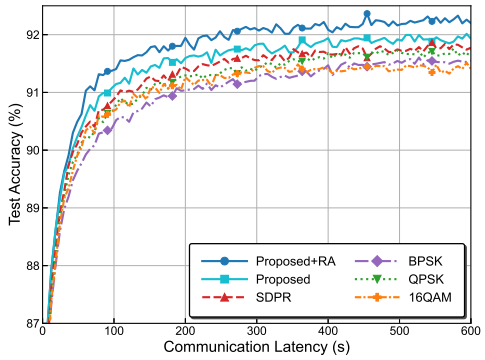


Fig. 4. Performance on Speech Commands.

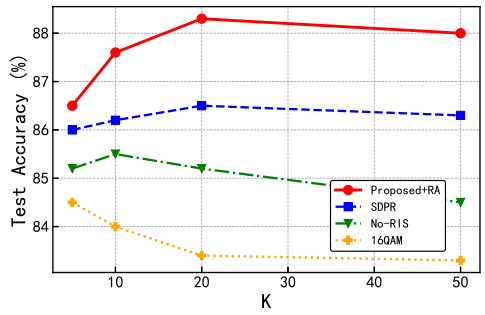


Fig. 5. Performance versus the number of clients.

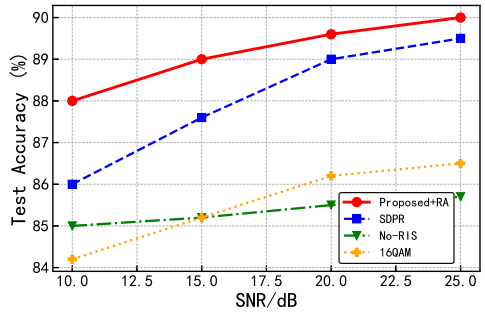


Fig. 6. Performance versus the average SNR.

Fig. 5 illustrates the impact of the number of participating clients on learning performance. Increasing the number of clients has two opposite effects: it provides more training data diversity and improves aggregation quality, but also introduces heavier communication load and more severe resource contention. The Proposed+RA scheme consistently achieves the best performance, with the accuracy increasing from about 86.5% at $K = 10$ to 88.2% at $K = 20$, and remains around 88% at $K = 50$, while SDPR [18] stays around 86.3%, the No-RIS baseline decreases from 85% to 84.5%, and fixed 16QAM drops from 84.3% to 83.3%. This demonstrates the superior scalability and error tolerance of our proposed scheme. In addition, the No-RIS baseline degrades more obviously as the client number increases, since more blocked or weak links lead to more severe error accumulation.

Fig. 6 shows the learning performance under different average SNR conditions. As the SNR increases, all schemes improve because better channel quality reduces symbol errors

in uplink model transmission. However, the Proposed+RA scheme remains the best-performing method over the entire SNR range. At 10 dB, it achieves an accuracy of about 88%, compared with 86% for SDPR [18], 85% for No-RIS, and 84% for fixed 16QAM. At 25 dB, its accuracy further improves to nearly 90%, while SDPR [18] reaches about 89.5%, and the No-RIS and 16QAM baselines remain lower at around 85.7% and 86.5%, respectively. It is also observed that SDPR [18] gradually approaches the Proposed+RA as the SNR increases. This is because SDPR [18] relies on SNR-based client selection and RIS-assisted communication enhancement. Under better channel conditions, more clients satisfy the requirement and participate in aggregation with fewer transmission errors. Nevertheless, our proposed method still maintains a consistent advantage by jointly optimizing adaptive modulation and resource allocation, which enables a better reliability-latency trade-off and more efficient use of wireless resources.

VI. Conclusion

This paper studied convergence-latency aware adaptive modulation and sub-channel allocation for wireless FL under direct LoS and RIS-assisted blocked NLoS transmission scenarios. By modeling local computation and communication latency, we established a convergence-related bound that explicitly captures the effect of modulation-dependent symbol errors on FL loss decay. Based on this result, we formulated a joint convergence-latency optimization problem and solved it using a low-complexity hybrid alternating optimization framework.

Simulation results on MNIST, CIFAR-10, and Speech Commands showed that our proposed scheme consistently outperforms benchmark methods in both convergence speed and final accuracy. The performance gain is particularly significant in complex tasks, dense-client settings, and low-SNR scenarios, where transmission errors and latency become the main bottlenecks. These results confirm the effectiveness of jointly optimizing communication reliability and resource efficiency for wireless FL.

Appendix A: Detailed Derivation of Modulation Order Optimization

A.1 Problem Formulation

We begin with the optimization problem of modulation order $M_{k,n}$ defined as

$$\max_{M_{k,n}} \delta(M_{k,n}) - \lambda T(M_{k,n}). \tag{48}$$

The objective function can be explicitly written as

$$\begin{aligned} obj =& \frac{1}{2\ell D_{all}^2} \sum_{k=1}^{K} D_k^2 \|g_k^t\|^2 \left[1 - \xi\left((\sigma-1)q_{k,n}^3 + q_{k,n}^2\right)\right] \\ & - \lambda \left(\frac{\phi Z}{B_n \log_2(M_{k,n})}\right). \end{aligned} \tag{49}$$

A.2 SER Approximation and $G(M_{k,n})$ Definition

The SER for *M*-ary modulation is often expressed in terms of the Q-function, which involves an integral form that complicates gradient-based optimization. To circumvent this issue, we approximate the Q-function using the following exponential expression [29]

$$Q(x) \approx \frac{1}{12} exp\left(-\frac{x^2}{2}\right) + \frac{1}{4} exp\left(-\frac{2x^2}{3}\right), \tag{50}$$

which provides a tight and differentiable approximation suitable for optimization purposes. Based on this approximation, taking the LoS scene as an example, the SER can be rewritten as a function of $M_{k,n}$

$$q_{k,n}(M_{k,n}) \approx \frac{1}{6} e^{-G(M_{k,n})} + \frac{1}{2} e^{-\frac{4}{3}G(M_{k,n})}, \tag{51}$$

where the function $G(M_{k,n})$ is defined as

$$G(M_{k,n}) = \frac{h_{k,n}^2 E_{k,n} \log_2(M_{k,n}) \sin^2\left(\frac{\pi}{M_{k,n}}\right)}{N_0}. \tag{52}$$

The following derivation uses the LoS-MPSK case as an example. For other channel conditions and modulation formats, the same procedure applies by replacing $q_{k,n}(M_{k,n})$ with the corresponding SER expression in Section II-C.

A.3 Objective Function Differentiation

We begin by calculating the derivative of the objective function with respect to $M_{k,n}$

$$\begin{aligned} \frac{\partial \text{obj}}{\partial M_{k,n}} =& -\frac{D_k^2 \|g_k^t\|^2 \xi}{2\ell D_{all}^2} \left[3(\sigma-1)q_{k,n}^2 + 2q_{k,n}\right] \frac{\partial q_{k,n}}{\partial M_{k,n}} \\ & - \lambda \frac{\partial T}{\partial M_{k,n}} = 0. \end{aligned} \tag{53}$$

A.4 Derivation of $\frac{\partial q_{k,n}}{\partial M_{k,n}}$ and $\frac{\partial T}{\partial M_{k,n}}$

Substituting the approximation

$$\frac{\partial q_{k,n}}{\partial M_{k,n}} = \left(-\frac{1}{6} e^{-G(M_{k,n})} - \frac{2}{3} e^{-\frac{4}{3}G(M_{k,n})}\right) \frac{\partial G(M_{k,n})}{\partial M_{k,n}}. \tag{54}$$

The derivative of $G(M_{k,n})$ with respect to $M_{k,n}$ is calculated as

$$\begin{aligned} & \frac{\partial G(M_{k,n})}{\partial M_{k,n}} \\ =& \frac{h_k^2 E_k}{N_0 \ln(2) M_{k,n}} \left(\sin^2 \frac{\pi}{M_{k,n}} - \frac{\pi \log_2(M_{k,n}) \sin \frac{2\pi}{M_{k,n}}}{M_{k,n}}\right). \end{aligned} \tag{55}$$

The latency $T(M_{k,n})$ is defined as

$$T(M_{k,n}) = \frac{\phi Z}{B_n \log_2(M_{k,n})}. \tag{56}$$

Thus, its derivative is

$$\frac{\partial T}{\partial M_{k,n}} = -\frac{\phi Z}{B_n M_{k,n} (\log_2(M_{k,n}))^2 \ln(2)}. \tag{57}$$

A.5 Substituting and Simplifying

Substituting the derivative expressions into the objective derivative, we obtain

$$\begin{aligned}&\frac{\lambda\phi Z}{B_n M_{k,n}(\log_2(M_{k,n}))^2\ln(2)}\\ =&\frac{D_k^2\|g_k^t\|^2\xi}{2\ell D_{all}^2}\left[3(\sigma-1)q_{k,n}^2+2q_{k,n}\right]\cdot\\ &\left(-\frac{1}{6}e^{-G(M_{k,n})}-\frac{2}{3}e^{-\frac{4}{3}G(M_{k,n})}\right)\frac{\partial G(M_{k,n})}{\partial M_{k,n}}.\end{aligned}\tag{58}$$

This is the final implicit equation for the optimization of modulation order. Then we solve it using the Newton-Raphson method. The algorithm for solving the relaxed modulation order problem is as follows

- Step 1: Initialize $M_{k,n}^{(0)}$ within $[M_{\min}, M_{\max}]$ and set the convergence threshold ϵ.
- Step 2: Compute $f(M_{k,n}^{(i)}) = \partial obj(M_{k,n}^{(i)})/\partial M_{k,n}^{(i)}$.
- Step 3: Update the modulation order by

$$M_{k,n}^{(i+1)} = \Pi_{[M_{\min},M_{\max}]}\left(M_{k,n}^{(i)} - \frac{f(M_{k,n}^{(i)})}{f'(M_{k,n}^{(i)})}\right).$$

- Step 4: Recalculate $\sigma_{k,n}$ and repeat Steps 2–3 until $|M_{k,n}^{(i+1)} - M_{k,n}^{(i)}| < \epsilon$.
- Step 5: Project the continuous solution onto the feasible discrete modulation set and select the modulation order that maximizes the original objective while satisfying the SER constraint.

Appendix B: Detailed Derivation of KKT Conditions for sub-channel Allocation

In this appendix, we derive the KKT conditions for the relaxed sub-channel allocation problem. Given the modulation orders, the relaxed resource allocation problem is formulated as

$$\begin{aligned}\max_{\mathbf{a},\tau}\quad &\sum_{k=1}^{K}\sum_{n=1}^{N}a_{k,n}\alpha_{k,n}-\lambda\tau\\ \text{s.t.}\quad &T_k^C+\frac{\phi Z}{\sum_{n=1}^{N}a_{k,n}B_n\log_2(M_{k,n})}-\tau\le 0,\quad\forall k,\\ &\sum_{k=1}^{K}a_{k,n}-1\le 0,\quad\forall n,\\ &1-\sum_{n=1}^{N}a_{k,n}\le 0,\quad\forall k,\\ &-a_{k,n}\le 0,\quad\forall k,n,\\ &a_{k,n}-1\le 0,\quad\forall k,n.\end{aligned}\tag{59}$$

For brevity, we define $c_{k,n} = B_n\log_2(M_{k,n}), R_k = \sum_{n=1}^{N}a_{k,n}c_{k,n}$. Then, the latency constraint of client k can be written as $T_k^C + \frac{\phi Z}{R_k} - \tau \le 0$.

To derive the KKT conditions, we convert the maximization problem into an equivalent minimization problem, which is

$$\min_{\mathbf{a},\tau} -\sum_{k=1}^{K}\sum_{n=1}^{N}a_{k,n}\alpha_{k,n}+\lambda\tau.\tag{60}$$

Let $\rho_k \ge 0$ be the Lagrange multiplier associated with the latency constraint of client k, $\mu_n \ge 0$ be the multiplier associated with the sub-channel exclusiveness constraint, $\underline{\nu}_{k,n} \ge 0$ be the multiplier associated with $-a_{k,n} \le 0$, and $\overline{\nu}_{k,n} \ge 0$ be the multiplier associated with $a_{k,n} - 1 \le 0$. The Lagrangian function is given by

$$\begin{aligned}\mathcal{L} =& -\sum_{k=1}^{K}\sum_{n=1}^{N}a_{k,n}\alpha_{k,n}+\lambda\tau\\ &+\sum_{k=1}^{K}\rho_k\left(T_k^C+\frac{\phi Z}{R_k}-\tau\right)\\ &+\sum_{n=1}^{N}\mu_n\left(\sum_{k=1}^{K}a_{k,n}-1\right)+\sum_{k=1}^{K}\omega_k\left(1-\sum_{n=1}^{N}a_{k,n}\right)\\ &-\sum_{k=1}^{K}\sum_{n=1}^{N}\underline{\nu}_{k,n}a_{k,n}+\sum_{k=1}^{K}\sum_{n=1}^{N}\overline{\nu}_{k,n}(a_{k,n}-1).\end{aligned}\tag{61}$$

The KKT conditions are given as follows.

B.1 Primal Feasibility

The primal variables should satisfy

$$T_k^C+\frac{\phi Z}{R_k}-\tau\le 0,\quad\forall k,\tag{62}$$

$$\sum_{k=1}^{K}a_{k,n}\le 1,\quad\forall n,\tag{63}$$

$$\sum_{n=1}^{N}a_{k,n}\ge 1,\quad\forall k,\tag{64}$$

$$0\le a_{k,n}\le 1,\quad\forall k,n.\tag{65}$$

B.2 Dual Feasibility

The Lagrange multipliers should satisfy

$$\rho_k\ge 0,\quad \mu_n\ge 0,\quad \omega_k\ge 0,\quad \underline{\nu}_{k,n}\ge 0,\quad \overline{\nu}_{k,n}\ge 0.$$

B.3 Complementary Slackness

The complementary slackness conditions are

$$\rho_k\left(T_k^C+\frac{\phi Z}{R_k}-\tau\right)=0,\quad\forall k,\tag{66}$$

$$\mu_n\left(\sum_{k=1}^{K}a_{k,n}-1\right)=0,\quad\forall n,\tag{67}$$

$$\omega_k\left(1-\sum_{n=1}^{N}a_{k,n}\right)=0,\quad\forall k,\tag{68}$$

$$\underline{\nu}_{k,n}a_{k,n}=0,\quad\forall k,n,\tag{69}$$

$$\overline{\nu}_{k,n}(a_{k,n}-1)=0,\quad\forall k,n.\tag{70}$$

B.4 Stationarity

Taking the derivative of the Lagrangian with respect to τ, we have

$$\frac{\partial\mathcal{L}}{\partial\tau}=\lambda-\sum_{k=1}^{K}\rho_k=0,\quad\sum_{k=1}^{K}\rho_k=\lambda.\tag{71}$$

Next, taking the derivative of the Lagrangian with respect to $a_{k,n}$ gives

$$\frac{\partial\mathcal{L}}{\partial a_{k,n}}=-\alpha_{k,n}-\rho_k\frac{\phi Z c_{k,n}}{R_k^2}+\mu_n-\omega_k-\underline{\nu}_{k,n}+\overline{\nu}_{k,n}=0,\tag{72}$$

where $c_{k,n} = B_n \log_2(M_{k,n})$ and $R_k = \sum_{n=1}^{N} a_{k,n} c_{k,n}$.

Therefore, the stationarity condition can be rewritten as

$$\alpha_{k,n} + \rho_k \frac{\phi Z c_{k,n}}{R_k^2} + \omega_k = \mu_n - \underline{\nu}_{k,n} + \overline{\nu}_{k,n}. \tag{73}$$

For an interior solution satisfying $0 < a_{k,n} < 1$, we have

$$\underline{\nu}_{k,n} = 0, \quad \overline{\nu}_{k,n} = 0. \tag{74}$$

Thus, the following relation holds

$$\alpha_{k,n} + \rho_k \frac{\phi Z B_n \log_2(M_{k,n})}{R_k^2} + \omega_k = \mu_n. \tag{75}$$

This condition indicates that the relaxed allocation balances the convergence-related utility, the latency reduction gain, and the minimum-allocation requirement for each client.

B.5 Binary Recovery

The KKT conditions provide the optimality characterization of the relaxed sub-channel allocation problem. Since the original problem requires binary sub-channel allocation, the relaxed solution $\tilde{a}_{k,n}$ is mapped back to a binary solution by

$$a_{k,n}^{\star} = \begin{cases} 1, & k = \arg\max\limits_{j \in \mathcal{K}} \tilde{a}_{j,n}, \\ 0, & \text{otherwise}. \end{cases} \tag{76}$$

After the binary recovery, the SER constraint is checked. If the recovered solution violates the SER constraint of any client, the corresponding sub-channel is reassigned to the next feasible client or the modulation order is adjusted to a lower level. This step guarantees that the final solution satisfies the reliability constraint of the original problem.

REFERENCES

[1] Y. Jing, J. Wang, C. Jiang, and Y. Zhan, "Satellite MEC with federated learning: Architectures, technologies and challenges," *IEEE Network*, vol. 36, no. 5, pp. 106–112, 2022.

[2] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–19, 2019.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2017, pp. 1273–1282.

[4] D. C. Nguyen, P. Cheng, M. Ding, D. Lopez-Perez, P. N. Pathirana, J. Li, A. Seneviratne, Y. Li, and H. V. Poor, "Enabling AI in future wireless networks: A data life cycle perspective," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 553–595, 2021.

[5] X. Deng, J. Li, C. Ma, K. Wei, L. Shi, M. Ding, and W. Chen, "Low-latency federated learning with DNN partition in distributed industrial IoT networks," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 3, pp. 755–775, 2023.

[6] O. Marnissi, H. El Hammouti, and E. H. Bergou, "Latency minimization in heterogeneous federated learning through joint compression and resource allocation," in *Proc. IEEE 100th Veh. Technol. Conf. (VTC-Fall)*, 2024, pp. 1–5.

[7] K. Wei, J. Li, C. Ma, M. Ding, C. Chen, S. Jin, Z. Han, and H. V. Poor, "Low-latency federated learning over wireless channels with differential privacy," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 290–307, 2022.

[8] Z. Chen, W. Yi, H. Shin, and A. Nallanathan, "Adaptive model pruning for communication and computation efficient wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7582–7598, 2024.

[9] L. Wang, J. Li, W. Chen, Q. Wu, and M. Ding, "Communication-efficient model aggregation with layer divergence feedback in federated learning," *IEEE Commun. Lett.*, vol. 28, no. 10, pp. 2293–2297, 2024.

[10] X. Han, W. Chen, J. Li, M. Ding, Q. Wu, K. Wei, X. Deng, and Z. Mei, "Energy-efficient wireless federated learning via doubly adaptive quantization," *IEEE Trans. Cogn. Commun. Netw.*, vol. 11, no. 4, pp. 2216–2231, 2025.

[11] J.-S. Tan, S. Yang, H.-Y. Zhai, S. Zhao, Z. Feng, and Q. Bi, "Federated learning with doubly-adaptive quantization in unreliable wireless networks: Convergence analysis and low-latency design," *IEEE Internet Things J.*, vol. 13, no. 1, pp. 449–462, 2026.

[12] S. Chen, L. Li, G. Wang, M. Pang, and C. Shen, "Federated learning with heterogeneous quantization bit allocation and aggregation for internet of things," *IEEE Internet Things J.*, vol. 11, no. 2, pp. 3132–3143, 2024.

[13] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3579–3605, 2021.

[14] 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.

[15] Y. Jing, Y. Qu, T. Wu, C. Dong, S. Guo, and Q. Wu, "Air-ground integrated online federated learning under unreliable communication," *IEEE Trans. Cogn. Commun. Netw.*, vol. 11, no. 6, pp. 4110–4125, 2025.

[16] Z. Chen, W. Yi, Y. H. Kim, and A. Nallanathan, "Gradient compensation enabled federated learning for unreliable wireless links," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, 2024, pp. 3697–3702.

[17] Y. Zhao, Q. Wu, W. Chen, C. Wu, and H. V. Poor, "Performance-oriented design for intelligent reflecting surface-assisted federated learning," *IEEE Trans. Commun.*, vol. 71, no. 9, pp. 5228–5243, 2023.

[18] P. Sun, E. Liu, W. Ni, R. Wang, Z. Xing, B. Li, and A. Jamalipour, "Reconfigurable intelligent surface-assisted wireless federated learning with imperfect aggregation," *IEEE Trans. Commun.*, vol. 73, no. 2, pp. 1058–1071, 2025.

[19] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116 753–116 773, 2019.

[20] X. Zhu, Q. Wu, and W. Chen, "On the performance of RIS-aided spatial modulation for downlink transmission," *IEEE Trans. Wireless Commun.*, vol. 23, no. 11, pp. 16 203–16 217, 2024.

[21] Z. Chen, W. Chen, J. Li, Q. Wu, M. Ding, X. Han, X. Deng, and L. Wang, "Hierarchical federated learning for social network with mobility," *IEEE Trans. Cogn. Commun. Netw.*, vol. 12, pp. 2888–2903, 2026.

[22] X. Han, J. Li, W. Chen, Z. Mei, K. Wei, M. Ding, and H. V. Poor, "Analysis and optimization of wireless federated learning with data heterogeneity," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7728–7744, 2024.

[23] W. Shi, S. Zhou, and Z. Niu, "Device scheduling with fast convergence for wireless federated learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[24] K. Fan, W. Chen, J. Li, X. Deng, X. Han, and M. Ding, "Mobility-aware joint user scheduling and resource allocation for low latency federated learning," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, 2023, pp. 1–6.

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

[26] J. G. Proakis and M. Salehi, *Digital Communications, 5th ed.* New York, NY, USA: McGraw-Hill, 2008.

[27] L. Qu, Y. Mao, S. Song, and C.-Y. Tsui, "Energy-efficient channel decoding for wireless federated learning: Convergence analysis and adaptive design," *IEEE Trans. Wireless Commun.*, vol. 23, no. 11, pp. 17 222–17 235, 2024.

[28] 3GPP, "Further enhancements to LTE TDD interference management and traffic adaptation," 3GPP, Sophia Antipolis, France, Tech. Rep. TR 36.828, 2012.

[29] X. Zhu, W. Chen, Z. Li, Q. Wu, Z. Zhang, K. Wang, and J. Li, "RIS-aided spatial scattering modulation for mmWave MIMO transmissions," *IEEE Trans. Commun.*, vol. 71, no. 12, pp. 7378–7392, 2023.