Adaptive Retransmission Design for Wireless Federated Edge Learning

XU Xinyi, LIU Shengli, YU Guanding
(Zhejiang University, Hangzhou 310027, China)

DOI: 10.12142/ZTECOM.202301002
https://kns.cnki.net/kcms/detail/34.1294.TN.20230220.1702.004.html, published online February 21, 2023
Manuscript received: 2022-10-27

Abstract: As a popular distributed machine learning framework, wireless federated edge learning (FEEL) can keep original data local, while uploading model training updates to protect privacy and prevent data silos. However, since wireless channels are usually unreliable, there is no guarantee that the model updates uploaded by local devices are correct, thus greatly degrading the performance of the wireless FEEL. Conventional retransmission schemes designed for wireless systems generally aim to maximize the system throughput or minimize the packet error rate, which is not suitable for the FEEL system. A novel retransmission scheme is proposed for the FEEL system to make a tradeoff between model training accuracy and retransmission latency. In the proposed scheme, a retransmission device selection criterion is first designed based on the channel condition, the number of local data, and the importance of model updates. In addition, we design the air interface signaling under this retransmission scheme to facilitate the implementation of the proposed scheme in practical scenarios. Finally, the effectiveness of the proposed retransmission scheme is validated through simulation experiments.

Keywords: federated edge learning; retransmission; unreliable communication; convergence rate; retransmission latency

Citation (IEEE Format): X. Y. Xu, S. L. Liu, and G. D. Yu, "Adaptive retransmission design for wireless federated edge learning," *ZTE Communications*, vol. 21, no. 1, pp. 3 – 14, Mar. 2023. doi: 10.12142/ZTECOM.202301002.

1 Introduction

With the construction of smart cities, a large number of Internet of Things devices, smartphones and other mobile devices have emerged from all aspects of our lives. The current society has entered the era of big data, and hundreds of millions of data are generated on mobile terminals every day[1 – 3], which poses novel challenges to both traditional centralized machine learning approaches and wireless communication techniques[4 – 5]. On the one hand, due to a large number of data, uploading all data to the cloud would result in a huge communication burden[6], and on the other hand, since the data contain user privacy, such as medical health and personal preferences, uploading raw data to the cloud would bring about the problem of privacy leakage[7 – 8].

To overcome the abovementioned challenges, a distributed machine learning framework named federated edge learning (FEEL) has been proposed recently[9 – 11]. Under FEEL, multiple distributed mobile devices use their locally dispersed data to jointly train a common machine learning model, rather than transferring raw data to a central node. The original data containing user privacy are stored on mobile devices, and only the intermediate data, such as gradients and parameters, are transmitted so that user privacy can be protected. In addition, FEEL shifts the model training process from the center to the local devices, thus making full use of distributed computing resources. Due to the advantages brought by the special architecture of FEEL, it has been intensively used in the fields of healthcare, computer vision, finance, etc.[12 – 15]

Recently, most research on FEEL assumes that communication links are reliable. For example, Ref. [16] considers the method of minimizing the transmitted energy under the delay constraint to improve the performance of FFEL. However, in practice, especially in wireless FEEL, channel transmission is generally unreliable due to random channel fading, shadowing, and noise. The accuracy of the intermediate data transmission during training cannot be guaranteed[17]. Retransmission is an important means to improve the accuracy of transmission in wireless communication systems, but with the cost of increasing the communication delay[18]. However, with the application of FEEL in medical and autonomous driving, it is more sensitive to the accuracy and delay of transmission[19]. This motivates us to investigate novel retransmission schemes for FEEL in this paper.

1.1 Related Work

There have been several studies considering the channel unreliability of wireless communications in distributed learning systems. In Ref. [20], the wireless channel in the FEEL system is modeled as an erasure channel and a scheme for this situation is proposed, which inherits the previous round of gradient when the packet is lost. Based on this, the authors further analyze the influence of coding rate on wireless

FEEL in Ref. [21]. In Ref. [22], a decentralized stochastic gradient descent method under the user datagram protocol (UDP) is proposed to reduce the impact of unreliable channels on decentralized federated learning. Moreover, an asynchronous decentralized stochastic gradient descent algorithm is proposed in Ref. [23] to reduce the impact of unreliable channels by performing asynchronous learning and reusing outdated gradients in device-to-device (D2D) networks. The authors in Ref. [24] have proposed an unbiased statistical reweighted aggregation scheme from the perspective of gradient aggregation, which comprehensively considers node fairness, unreliable parameter transmission, and resource constraints. In Ref. [25], a sparse federated learning framework is proposed, which compensates for the bias caused by unreliable communication through the similarity between local models, and adds local sparseness to reduce communication cost, which further improves performance. In Ref. [26], a federated learning framework is proposed, where the central server aggregates the global model according to the received parameters and the transmission correct probability, thereby reducing the impact of unreliable transmission. The authors in Ref. [27] further propose a decentralized D2D framework under unreliable channels, which reduces the impact of unreliable channels by jointly optimizing the transmission rate and bandwidth distribution.

From the perspective of wireless communication, retransmission has been applied to many current communication standards, including 5G and WiFi. So far, only a few works have studied the retransmission issue in distributed learning. Retransmission can improve the reliability of data packets, but it also reduces the timeliness of data. In some scenarios, it may even be considered to improve the timeliness of data at the cost of reduced reliability[28]. In Ref. [29], a Hybrid Automatic Repeat reQuest (HARQ) protocol suitable for multilayer cellular networks has been proposed, which can enhance error detection and correction in D2D communications. In Ref. [30], a retransmission scheme based on data importance is proposed for the edge learning system. The specific approach of this scheme is to make a tradeoff between the signal-to-noise ratio (SNR) and the uncertainty of the data, and correspondingly establish a threshold for retransmission.

1.2 Motivations and Contributions

As aforementioned, in wireless FEEL, devices upload gradients to the edge server through wireless channels, which is unreliable. This will affect the performance of model training. The goal of conventional retransmission schemes is to maximize the throughput of correctly transmitted data. However, the performance of FEEL with unreliable channels is limited by traditional retransmission since FEEL has different goals of learning accuracy and learning latency. In particular, the importance of data from different devices is different and generally contributes differently to the model training process. In addition, the communication cost introduced by retransmission of each device is also different due to various channel fading environments. The above factors need to be considered when developing a retransmission scheme for the edge learning system. The main contributions of this paper can be summarized as follows.

• We first propose a FEEL framework with unreliable channels, in which the gradients uploaded by the local devices are split into multiple packets, and the wireless channel exists the packet error rate (PER). Unreliable transmission leads to bias between the actual global gradient and the theoretical one, which is detrimental to model training.

• We mathematically analyze the effect of PER on the convergence rate and communication cost. To mitigate the impact of unreliable communications on learning performance, the retransmission device selection is optimized by making a tradeoff between convergence rate and communication cost.

• We derive the optimal solution to device retransmission selection, which greatly improves the model training performance. We also analyze the performance of the proposed retransmission selection scheme and develop a signaling protocol for retransmission.

• We employ a convolutional neural network (CNN) model of the CIFAR-10 and MNIST datasets to test the learning performance of our proposed retransmission selection scheme. Test results show that our proposed scheme outperforms several existing retransmission schemes.

The rest of the paper is organized as follows. In Section 2, we introduce the system model. In Section 3, the principle of retransmission design is introduced, and the corresponding protocol is proposed. In Section 4, we analyze the retransmission gain and cost and formulate the retransmission selection optimization problem. The retransmission selection is derived in Section 5. Finally, we draw the conclusions in Section 6.

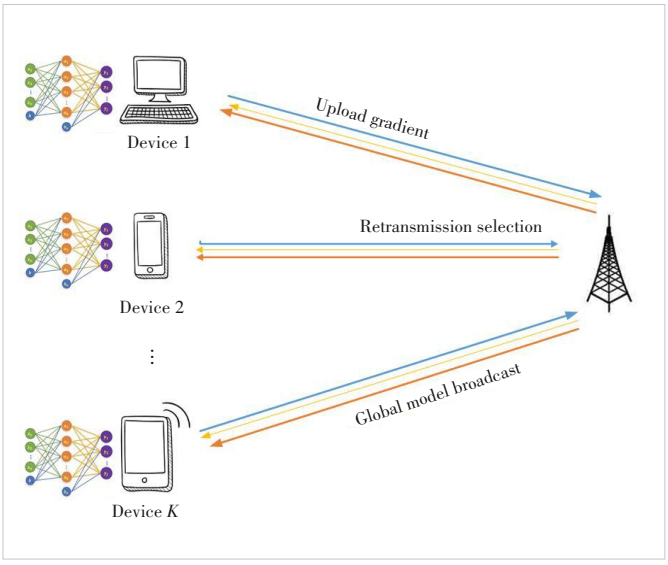
2 System Model

2.1 Machine Learning Model

As depicted in Fig. 1, we consider a FEEL system consisting of one edge server and K devices. Device k has n_k locally labeled data, and the total number of data in the entire system can be represented as $n = \sum_{k=1}^{K} n_k$. All devices only use their own data to jointly train a machine learning model $\boldsymbol{w}$ with the edge server, and the specific method is stochastic gradient descent (SGD). Considering the imbalance of data distribution, the global loss function can be written as:

$$L(\boldsymbol{w}) = \frac{1}{n}\sum_{k=1}^{K} n_k L_k(\boldsymbol{w}), \tag{1}$$

where $L_k(\boldsymbol{w})$ is the loss function of device k, and we have



▲Figure 1. Federated edge learning system

$$L_k(\boldsymbol{w}) = \frac{1}{n_k}\sum_{i=1}^{n_k} f(\boldsymbol{w}, x_{i,k}, y_{i,k}), \tag{2}$$

where $x_{i,k}$ represents the i-th training data of device k, $y_{i,k}$ represents the corresponding label, and $f(\cdot)$ represents the loss function of the training model. Some popular machine learning loss functions are summarized in Table 1.

The purpose of federated training is to find the optimal $\boldsymbol{w}^*$ that minimizes $L(\boldsymbol{w})$. FEEL is different from the traditional centralized machine learning framework. In the FEEL framework, all the original data are kept on local devices, and the training results are uploaded to the edge server. In the t-th round of training, the selected devices use the local data and the global model $\boldsymbol{w}^t$ received from the edge server to obtain the loss function $L_k(\boldsymbol{w}^t)$, and upload the gradient of $L_k(\boldsymbol{w}^t)$ to the edge server, which can be written as:

$$g_k^t = \nabla L_k(\boldsymbol{w}^t). \tag{3}$$

After receiving the uploaded gradients of all selected devices, the edge server decodes the data packets and aggregates the global gradient g^t as:

$$g^t = \frac{1}{n}\sum_{k=1}^{K} n_k g_k^t. \tag{4}$$

▼Table 1. Loss function for popular machine learning models

Learning Model	Loss Function $f(\boldsymbol{w}, x, y)$
Linear regression	$\frac{1}{2}\left\| y - \boldsymbol{w}^T x \right\|^2$
Least-squared support vector machine	$\frac{1}{2}\max\{0, 1 - y\boldsymbol{w}^T x\}^2$
Neural network	$\frac{1}{2}\left\| y - \phi(\boldsymbol{w}, x) \right\|$, where $\phi(\boldsymbol{w}, x)$ is the learning output

Then the edge server uses the global gradient g^t obtained by the aggregation to update the model, that is, $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta g^t$, where η is the learning ratio. After completing the update of the global model, the edge server broadcasts it to each device in the system. In this way, one round of iterative training of FEEL is completed.

2.2 Wireless Communication Model

In this paper, we utilize time division multiple access (TDMA) as the multiple access method. In a TDMA scenario, all devices use the same frequency band in different time slots and upload gradients to the edge server in turn. During one training iteration, it is assumed that the expected channel state information can be obtained by the channel estimation algorithms. Among the training iterations, the channel of the iteration differs from one another. The expected channel state information in each iteration is separately adopted for the performance analysis. Therefore, when a device uploads the gradients, it will occupy the full bandwidth, denoted by B. For ease of analysis, it is assumed that the wireless channel is static at each training gradient upload and changes in different rounds of training iterations. It is further assumed that the distances of all local devices to the edge server are known, and the small-scale fading is modeled as Rayleigh fading. Then, we can express the uploaded data rate of the device k as:

$$R_k = B\log_2\left(1 + \frac{P_k^U |h_k^U|^2}{N_0}\right), \tag{5}$$

where P_k^U is the transmit power of device k, h_k^U is the channel power gain between the device and the edge server, and N_0 is the noise power over the whole bandwidth B. We assume that each device is uploading and retransmitting data at the maximum available power. Note that this assumption fits many scenarios, such as LTE[31].

Since wireless channels are generally unreliable, channel errors need to be considered. It is assumed that the uploaded gradients of each device are divided into several packets, and each packet has redundant encoding for error detection. In this paper, the cyclic redundancy check (CRC) code is used to check for errors. Then the PER of device k can be expressed as:

$$p_k = 1 - \exp\left(-\frac{mBN_0}{P_k^U h_k^U}\right), \tag{6}$$

where m is the PER decision threshold[32].

Since the global model sent by the edge server to all devices is the same, the downlink channel can be modeled as a broadcast channel and a more robust encoding method can be used. In this paper, we consider that the channel error occurs only in the uplink channel, and assume that there is no channel error in the downlink channel. Let the channel bandwidth of the downlink channel be B_D, and denote γ as the smallest

SNR among all devices, and then the achievable downlink data rate is expressed as:

$$R_D = B_D \log_2(1 + \gamma). \tag{7}$$

3 Retransmission Protocol

In this section, we first introduce the principle of retransmission design in FEEL. Then, we propose a novel retransmission protocol and develop the corresponding processing modules for both devices and the edge server.

3.1 Principle of Retransmission Design

In FEEL, the edge server performs global model updates by periodically aggregating local gradients uploaded by devices. Therefore, the performance of the trained model depends on the quality of the gradients received by the edge server. However, unreliable gradient transmission may occur due to wireless channel impairments including interference, noise and shadowing. Therefore, it is predicted that the performance of model training is largely affected by channel impairments.

A common solution to unreliable transmission is retransmission. Conventionally, the purpose of retransmission is to ensure the reliability of data and at the same time maximize the system throughput. However, the main goal of FEEL is to maximize the training accuracy for a given training time. Therefore, a novel retransmission protocol is required for the FEEL system.

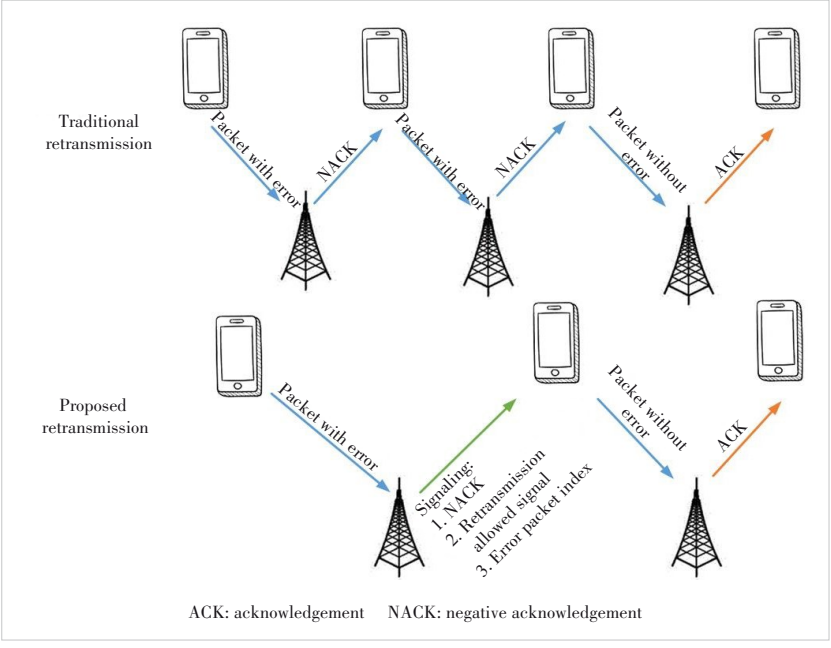
When designing the retransmission protocol for a FEEL system, one should consider both the training accuracy and the additional communication cost brought by retransmission. Retransmission can reduce erroneous packets so that the gradient updates received by the edge server deviate less from the ground-truth gradient, which can improve the convergence speed and the accuracy of model training. However, retransmission also increases the communication latency, resulting in an increase in training time. Therefore, we need to properly select the devices that need to be retransmitted and design appropriate signaling to make a fair tradeoff between learning accuracy and learning latency.

3.2 Retransmission Protocol and Processing Module

In our proposed retransmission protocol, not all devices participate in retransmission, that is, retransmission selection is required. Considering the characteristics of FEEL, the device selection depends on not only the channel conditions but also the local data volume and the importance of the upload gradient. Gradient updates that have a more significant impact on global model training will be retransmitted with a larger probability. Moreover, the latency caused by retransmission should also be accounted for. In our proposed protocol, a device with a higher data rate is also more likely to be retransmitted because it brings less additional communication cost. In addition, the PER between the device and the edge server shall also be taken into account. Due to the robustness of model training, devices with a small PER would bring little performance gain when retransmitting. Also, for a device with a large PER, the reduction of the PER after retransmission is very limited, but it will cause a relatively large communication cost. Therefore, when the PER is too large or too small, the probability of the device being selected for retransmission is both small.

We also consider a new design of retransmission signaling, as shown in Fig. 2. Under the traditional retransmission scheme, after receiving an erroneous packet, the edge server only sends a negative acknowledgement (NACK) signal to the device, requiring the device to retransmit. Until the edge server successfully decodes the data packet, it sends an acknowledgement (ACK) signal to the device, and the device starts to transmit the next data packet. In our protocol, when an edge server receives a packet and detects an error using CRC codes, it sends a signal to the corresponding device that includes the information shown in Fig. 2.

In Fig. 2, NACK indicates that the packet is transmitted with an error, but unlike that in the traditional retransmission schemes, it does not indicate that the device needs to retransmit the packet. Whether to retransmit needs to be judged according to the retransmission selection algorithm. Retransmission allowed signal ν_k indicates whether the device is selected for retransmission, which is related to the channel conditions, the number of local data, and the importance of the gradient.



▲Figure 2. Retransmission signaling

Specifically, $\nu_k = 1$ indicates that the device k is selected for retransmission; otherwise $\nu_k = 0$ indicates no retransmission. When $\nu_k = 1$, it is equivalent to traditional NACK. Error packet index represents the gradient position contained in the transmission data packet. If it is selected for retransmission, the device can retransmit the gradient of the corresponding position.

According to the received signal, the device will determine whether the uploaded packet is transmitted correctly and whether it is allowed to retransmit. After that, it retransmits the particular data corresponding to the erroneous packet, as indicated by the edge server.

4 Retransmission Design

In this section, we first analyze the one-round convergence rate with unreliable channels. Then, we propose a new criterion to evaluate the gain of retransmission on learning performance. The retransmission cost is analyzed as well. Based on this, we formulate a mathematical optimization problem to make a tradeoff between retransmission gain and retransmission cost.

4.1 One-Round Convergence

Due to the PER, during one round of training, the global gradient obtained by the edge server using the received gradient is not equal to the theoretical gradient g^t in Eq. (4). Therefore, we define the actual global gradient obtained by the aggregation under the unreliable channel as $\hat{g}^t$, and we have:

$$\hat{g}^t = \frac{\sum_{k=1}^{K} n_k \hat{g}_k^t}{n}, \tag{8}$$

where $\hat{g}_k^t$ is the actual local gradient of device k received by the edge server. Therefore, when there exists PER, the model update is:

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \hat{g}^t = \boldsymbol{w}^t - \eta(g^t - o^t), \tag{9}$$

where o^t is the deviation of the global gradient introduced by unreliable transmission, and we have:

$$o^t = g^t - \frac{\sum_{k=1}^{K} n_k \hat{g}_k^t}{n}. \tag{10}$$

To facilitate mathematical analysis, we make the following assumption.

Assumption 1: (ℓ-smooth loss function) The global loss function is Lipschitz continuous with positive parameter ℓ, shown as:

$$\left\| g^{t+1} - g^t \right\| \leqslant \ell \left\| \boldsymbol{w}^{t+1} - \boldsymbol{w}^t \right\|. \tag{11}$$

Based on the above assumption, we can obtain the convergence rate of one round under an unreliable channel.

Theorem 1: When the learning rate $\eta = \frac{1}{\ell}$, the training loss function in one round can be written as:

$$\mathbb{E}\left\{L(\boldsymbol{w}^{t+1})\right\} \leqslant \mathbb{E}\left\{L(\boldsymbol{w}^t)\right\} - \frac{1}{2\ell}\mathbb{E}\left\{\left\| g^t \right\|^2\right\} + \frac{1}{2\ell}\mathbb{E}\left\{\left\| o^t \right\|^2\right\}. \tag{12}$$

See Appendix A for details.

From Eq. (12), it can be seen that the loss function is constrained by three terms. The first term $\mathbb{E}\left\{L(\boldsymbol{w}^t)\right\}$ represents the loss function of the previous training round, which is independent of unreliable transmissions. The second item $\frac{1}{2\ell}\mathbb{E}\left\{\left\| g^t \right\|^2\right\}$ is related to the theoretical gradient value of this round, which depends on the data in local devices, but is independent of PER and the retransmission scheme. The third term $\frac{1}{2\ell}\mathbb{E}\left\{\left\| o^t \right\|^2\right\}$ is the bias term introduced by channel errors, which will reduce the loss function, thus affecting the convergence speed. In order to reduce the influence of unreliable channels and improve training performance, we need to reduce channel interference. Therefore, we next analyze the impact of PER (p_k^t) on the gradient bias $\mathbb{E}\left\{\left\| o^t \right\|^2\right\}$. Since we focus on the retransmission design of each round, for the convenience of presentation, we ignore the superscript t that represents the number of training rounds in the following.

We first assume that the machine learning model has a total of D layers of neural networks, and the device divides the corresponding gradients into D packets during the uploading process. The d-th packet contains gradient updates for the d-th layer of the neural network, which is denoted as $g_{k,d}$. Let indicator $\rho_{k,d}$ denote whether the transmission of the d-th packet of device k is correct. That is, $\rho_{k,d} = 1$ indicates that there is no error in the transmission, which means that the edge server can decode and obtain the correct gradient $g_{k,d}$, and there is a probability of $P(\rho_{k,d} = 1) = 1 - p_k$. Similarly, we let $\rho_{k,d} = 0$ denote the occurrence of a transmission error with probability of $P(\rho_{k,d} = 0) = p_k$. After the edge server receives the packets, if the error is detected and retransmission is not considered, the corresponding gradient is set to zero, which can be written as:

$$\hat{g}_{k,d} = \begin{cases} g_{k,d}, \rho_{k,d} = 1 \\ 0, \rho_{k,d} = 0 \end{cases}. \tag{13}$$

Lemma 1: The impact of error transmission on learning performance can be expressed as the bias of gradients caused by packet transmission errors, which can be written as:

$$\mathbb{E}\left\{\left\| o \right\|^2\right\} \leqslant \frac{K}{n^2}\sum_{k=1}^{K} n_k^2 p_k^2 \bar{g}_k^2, \tag{14}$$

where $\bar{g}_k = \sum_{d=1}^{D} g_{k,d}$ denotes the sum of the gradient of device k. See Appendix B for details.

First, the gradient bias term is affected by the PER p_k. The larger the PER of the device is, the larger the error term will be, and the smaller the loss function will decrease in one round. Second, the error term is affected by the number of local data on each device. The larger the number is, the more significant the impact of the device's PER on the entire model. Third, the error term is also affected by the gradient obtained from training. The larger the sum of uploaded gradients is, the larger the bias term would be introduced. Finally, since the global gradient is obtained by aggregating the uploaded gradients of selected devices, the bias term can be expressed as the sum of the bias introduced by each device due to unreliable transmission. Through the above analysis, we can obtain the convergence rate of one round in the presence of transmission errors as:

$$\mathbb{E}\left\{L(\boldsymbol{w}^{t+1})\right\} \leqslant \mathbb{E}\left\{L(\boldsymbol{w}^{t})\right\} - \frac{1}{2\ell}\mathbb{E}\left\{\left\| g^{t} \right\|^2\right\} + \frac{K}{2\ell n^2}\sum_{k=1}^{K} n_k^2 p_k^2 \bar{g}_k^2. \tag{15}$$

4.2 Gain of Retransmission

Next, we analyze the learning performance gain brought by retransmission. Define the PER of device k after the retransmission selection as q_k, which can be written as:

$$q_k = p_k\left(1 - \nu_k(1 - p_k)\right), \tag{16}$$

where p_k is the probability that an error occurs in one transmission, and $\nu_k(1 - p_k)$ represents the probability that device k is selected for retransmission and there is no error in the retransmission. Based on Eq. (14), considering the retransmission, the impact of PER on the convergence can be expressed as:

$$\mathbb{E}\left\{\left\| o_r \right\|^2\right\} \leqslant \frac{K}{n^2}\sum_{k=1}^{K} n_k^2 q_k^2 \bar{g}_k^2, \tag{17}$$

where o_r represents the bias between the theoretical gradients and the actual gradients after retransmission.

The PER of the device selected for retransmission will be reduced after retransmission, and its impact on learning performance will also be reduced. Therefore, we can present the following definition to analyze the gain which is achieved by retransmission.

Definition 1: We define the gain of retransmission as the difference between the bias of global gradients before and after retransmission on the learning performance, which can be written as

$$\Omega = \frac{K}{n^2}\sum_{k=1}^{K} n_k^2 p_k^2 \bar{g}_k^2 - \frac{K}{n^2}\sum_{k=1}^{K} n_k^2 q_k^2 \bar{g}_k^2 = \sum_{k=1}^{K} \Omega_k, \tag{18}$$

where Ω_k is the gain of retransmission of device k. Since the whole system can be regarded as a collection of all devices, we have:

$$\Omega_k = \frac{K}{n^2} n_k^2 \bar{g}_k^2\left(p_k^2 - q_k^2\right). \tag{19}$$

Eq. (19) reveals that the retransmission gain of the device is related to the number of local data, the value of the gradient update, and the reduction of the PER before and after retransmission. A larger data volume and gradient value of the device will bring a larger gain of retransmission to the learning performance. This solution can also be applied to dynamic wireless channels, just changing the retransmission PER to the actual PER.

4.3 Cost of Retransmission

Although device retransmission will bring gains to the learning performance, retransmission will also increase communication latency due to the additional resource required by retransmission. Therefore, we give the definition of the cost of retransmission as follows.

Definition 2: The cost of retransmission of device k is defined as the increase in latency introduced by retransmission, which can be expressed as

$$C_k = \frac{qNp_k}{R_k}\nu_k, \tag{20}$$

where q is the number of quantization bits and N is the total number of parameters.

4.4 Problem Formulation

Until now we have analyzed the gain and cost of retransmission. Retransmission will bring a gain in learning performance but increase additional communication costs. Therefore, we need to consider the tradeoff between cost and gain when developing a retransmission scheme. Our goal is to maximize retransmission gain while minimizing retransmission cost. We define $\beta \in [0,1]$ as a factor for the tradeoff between retransmission gain and retransmission cost, and the following retransmission gain-cost tradeoff problem can be established.

$$\text{P1: } \min_{\nu_k} \sum_{k=1}^{K}\left(-\beta\Omega_k + (1-\beta)C_k\right), \tag{21}$$

subject to

$$\nu_k \in \{0,1\}, \forall k. \tag{21a}$$

Eq. (21a) represents the retransmission indicator limitation. When β is close to 0, it means that the main goal is to reduce

the latency when retransmission is selected. When β is close to 1, it means that improving the convergence rate is the main goal.

5 Retransmission Optimization and Theoretical Analysis

In this section, we first give a retransmission selection strategy based on P1. Then, we analyze the effect of PER on retransmission selection.

5.1 Optimal Solution

By inserting Eqs. (16), (19), and (2) into Eq. (21), and relaxing the $\{0,1\}$ variable ν_k to $[0,1]$, P1 can be formulated as:

$$\text{P2: } \min_{\nu_k} \sum_{k=1}^{K} -\beta \frac{K}{n^2} n_k^2 \bar{g}_k^2 \left(p_k^2 - \left(p_k - \nu_k p_k (1-p_k) \right)^2 \right) + (1-\beta) \frac{qNp_k}{R_k} \nu_k, \tag{22}$$

subject to

$$\nu_k \in [0,1], \forall k. \tag{22a}$$

Eq. (22) consists of two parts: the first part is related to federated learning (FL) training loss, and the second part is related to FL one-round training latency. This is a classical convex optimization problem, and the optimal solution can be obtained through the Karush-Kuhn-Tucker (KKT) condition.

Theorem 2: The retransmission selection policy can be expressed as:

$$\nu_k^* = \left[\frac{1}{1-p_k} - \frac{(1-\beta)qNn^2}{2\beta K n_k^2 \bar{g}_k^2 p_k^2 (1-p_k)^2 R_k} \right]_0^1, \forall k, \tag{23}$$

where $[X]_0^1 = \min\{1, \max\{X, 0\}\}$. See Appendix C for further details.

Theorem 2 reveals that the retransmission indicator is a value bounded by 0 and 1, which is related to the local data volume, gradient value, data rate, and the PER of the device. Specifically, the probability of being selected for retransmission ν_k^* increases with the data number n_k and the gradient value $\bar{g}_k$ in the order of $-\frac{1}{2}$. This is because with a large number of device data and gradient values, the learning performance gain obtained by retransmission is also large. Also, ν_k^* increases with the data rate R_k in the order of -1. Since the data rate is large, the communication cost of retransmission will be small, and the probability of the device being selected for retransmission will increase. The impact of the device PER on the retransmission selection will be analyzed in the next section.

Since the obtained ν_k^* is the optimal solution after relaxation, we need to consider how to convert it into a $\{0,1\}$ variable for retransmission selection. We give two strategies. The first is to perform threshold processing on ν_k^*, with 0.5 as the limit. If $\nu_k^* \geqslant 0.5$, it means retransmission, and if $\nu_k^* < 0.5$, it will not be retransmitted. The second is to sort all devices from large to small according to the value of ν_k^*, and select the largest proportion $M\%$ of devices of ν_k^* for retransmission. The choice of M reflects the tradeoff between model accuracy and training latency.

5.2 Theoretical Analysis

In this section, we will analyze the impact of PER on the retransmission indicator. We first define:

$$m_k = \frac{(1-\beta)qNn^2}{2\beta K n_k^2 \bar{g}_k^2 R_k}. \tag{24}$$

From Eq. (24), m_k is related to the number of local data, gradient value and data rate, but is irrelevant to the PER. When the local data volume, the gradient value, and the uploaded data rate of device k are large, device k is more important in the retransmission design, and m_k is correspondingly small. Therefore, m_k reflects the contribution of the gradient of device k to the global model training, as well as the state of its channel. And m_k is always greater than 0. Moreover, the importance of device decreases as m_k increases. Then, in order to analyze the influence of p_k on the retransmission indicator ν_k^*, we define the following function:

$$f(p_k) = \frac{1}{1-p_k} - \frac{(1-\beta)qNn^2}{2\beta K n_k^2 \bar{g}_k^2 p_k^2 (1-p_k)^2 R_k} = \frac{1}{1-p_k} - \frac{m_k}{p_k^2 (1-p_k)^2}, \tag{25}$$

where $f(p_k)$ is a strictly unimodal function with $p_k \in [0,1]$. See Appendix D for details.

Theorem 2 reveals that the optimal retransmission indicator first increases and then decreases with p_k. Therefore, there exists an optimal p_k^* that maximizes $f(p_k)$. This result is rather intuitive, which shows that there is a tradeoff between retransmission gain and cost. For the device with a low PER, due to the robustness of neural networks, retransmission has little gain in learning performance, but will increase communication cost. Therefore, its probability of being selected for retransmission is relatively low. For the device with a relatively high PER, there will still be a high PER after retransmission. Thus, the gain in model training performance is not large. Also, the retransmission cost is large, and the probability of being selected is low. Note that devices with intermediate PER can improve the accuracy of gradient data after retransmission, and will not bring reused data or additional deviation.

6 Numerical Result

In this section, we conduct extensive experiments to verify the effectiveness of the proposed retransmission scheme.

6.1 Simulation Settings

Assume that the coverage area of the edge server is 1.5 km, and there are K (K=10) mobile devices that are randomly distributed across the cellular network. The transmit power of each device is 28 dBm, and the transmit power of the edge server is 33 dBm. Then, the noise power spectral density is −174 dBm/Hz and the PER decision threshold m = 0.2 dB. Since in the TDMA scenario, all devices occupy one channel to upload gradients. The uplink channel takes into account large-scale fading, given by $128.1 + 37.6\log(d)$, where d represents the distance between the device and the edge server in kilometers. We also consider small-scale fading of the channel, specifically represented by Rayleigh fading. All devices and the edge server jointly train a CNN model. We choose CIFAR-10 and MNIST as datasets. CIFAR-10 consists of 50 000 training images and 10 000 testing images. And MNIST consists of 55 000 training images and 5 000 testing images. The datasets are both non-identically and independently distributed (non-IID) and divided into 10 categories. Also, we choose the learning ratio η = 0.05. We quantize each element of the uploaded gradient with 16 bits. All elements of each layer are treated as one packet, and a 32-bit CRC code is added. Other major parameters are listed in Table 2.

6.2 Performance of Proposed Retransmission Scheme

Based on the previous theoretical analysis, the proposed algorithm can make a tradeoff between reducing the gradient aggregation bias caused by unreliable transmission and controlling the transmission delay, thereby accelerating the model convergence. We use the global training loss and global test accuracy to evaluate the learning performance of the whole learning system. In the simulation of this section, the discretization method for the retransmission factor ν_k^* is to take 0.5 as the threshold. That is, the selection indicator is set to 0 if ν_k^* is less than 0.5 and set to 1 if it is larger than 0.5.

The comparison algorithms in Fig. 3 are shown as follows.

• Without PER: The wireless channel is ideal and PER-free, meaning that all gradients can be transmitted correctly.

• Without retransmission: There is PER in the uplink channel, but retransmission is not considered. If the uploaded data packet is judged to be incorrect, it will be set to zero and the packet will be discarded.

• Existing retransmission schemes: Using the existing retransmission scheme based on the transmission result. The devices retransmit the erroneous data packets after receiving the NACK.

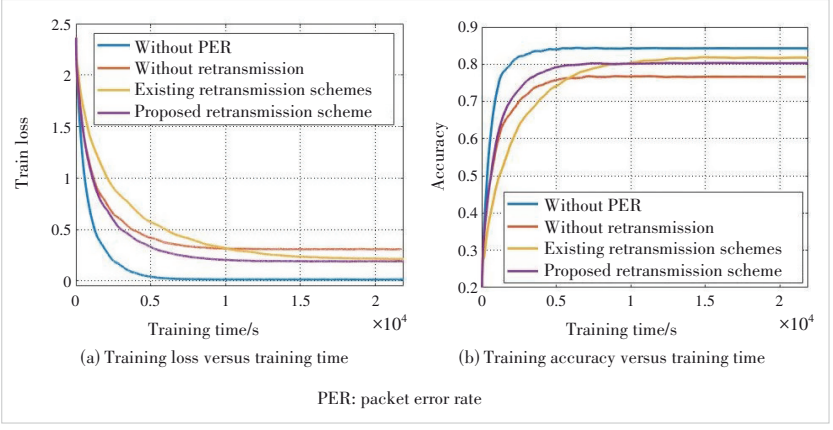
• The proposed retransmission scheme: Using the scheme proposed in this paper, we made the retransmission selection according to the device's local data, gradient data, and PER.

We first perform simulations under the CIFAR-10 dataset. The curves of training loss and test accuracy versus training time under different retransmission schemes are shown in Fig. 3. As can be seen from the figure, when transmitting on a reliable channel, no retransmission is required. At this time, the model training can reach convergence in a very short time with a high model accuracy. When the channels are unreliable and retransmission is not performed, the performance of model training will be greatly degraded. When retransmission is not performed, model training can reach convergence very fast, but the accuracy of the final model is pretty low. As a result, when there is no retransmission, the communication cost is relatively small. Although multiple rounds of training are required, one round of training latency is short, so the overall latency is short. However, due to the large bias between the received gradient and the local gradient, the performance of the final trained model is not satisfactory, which also confirms the necessity of retransmission. It can also be seen that, in the existing retransmission scheme, although the accuracy of the final model is high, it takes much longer time to converge. This is because the existing retransmission scheme aims to maximize the throughput, without considering selecting retransmission devices, or the importance of uploading gradients for model training. Due to a large number of transmitted gradient data and participating training de-

▼**Table 2. Simulation parameters**

Parameters	Values
Path loss model	$128.1 + 37.6\log(d)$
Transmission power of the edge server	33 dBm
Transmission power of device	28 dBm
Additive white Gaussian noise power	−174 dBm/Hz
Bandwidth of downlink	10 MHz
Quantization bit of each element	16
Number of devices	10
Bandwidth of uplink	10 MHz
CRC code	32

CRC: cyclic redundancy check



(a) Training loss versus training time (b) Training accuracy versus training time

PER: packet error rate

▲**Figure 3. Performance comparison between transmission schemes under CIFAR-10**

vices, the wireless FEEL system needs to spend a lot of time to achieve model convergence without retransmission selection. Therefore, the existing retransmission schemes cannot exhibit good performance under the FEEL system. As shown in Fig. 3, the retransmission scheme proposed in this paper can make the model training converge in a short time, and achieve high accuracy at the same time. The reason is that the influence of different gradients has been considered in the retransmission. This scheme can maximize the retransmission gain, reduce the influence of channel errors, and improve the performance of model training by selecting proper retransmission devices. In order to further illustrate the effectiveness of our proposed scheme, we increase the number of devices to 20 for simulation, and the results are shown in Fig. 4.

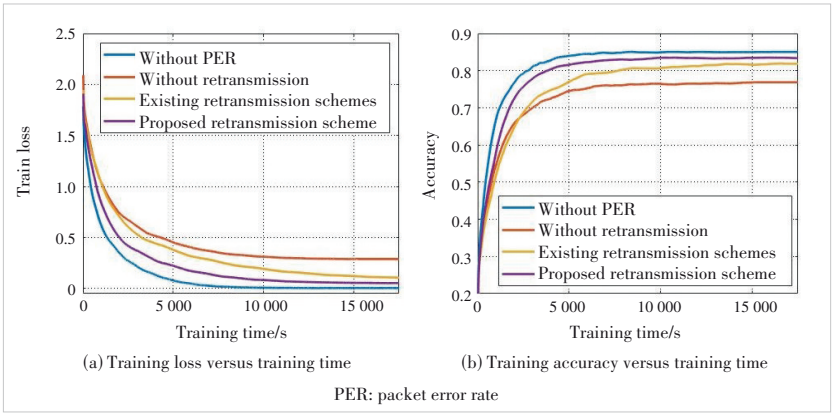
6.3 Performance with Difference Retransmission Ratios

When selecting $M\%$ of devices for retransmission in each round of transmission, the choice of parameter M may reflect the tradeoff between model accuracy and training latency in our proposed retransmission scheme.

From Fig. 4, when M is too small, e.g., 20% or 40%, both the convergence rate and final model accuracy become low. This is because the impact of channel error is strong when the number of selected retransmission devices is small. When M is too big, e.g., 80%, the convergence speed is low and the final accuracy has no significant advantage. This is because retransmission will increase the latency, and some devices are not of high importance, resulting in limited retransmission gain.

6.4 Performance Comparison Under Other Datasets

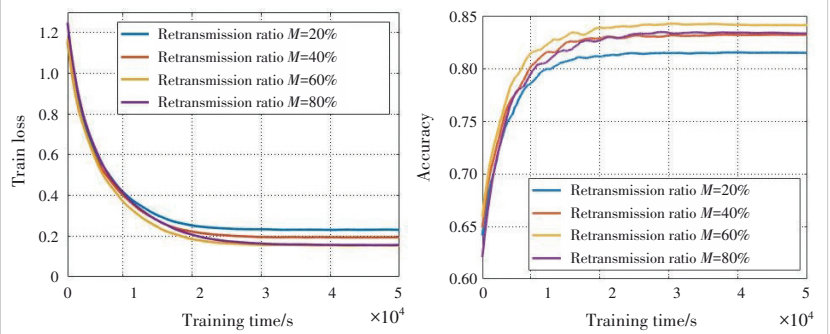
To verify the broad effectiveness of our proposed scheme, we change the training dataset to MNIST for further simulations. MNIST consists of 0 – 9 numbers handwritten by different people. The curves of training loss and test accuracy are shown in Fig. 6. After the dataset is changed, the effect of channel unreliability on model training and the performance improvement of our proposed scheme can still be seen. From Fig. 7, the proportion M of retransmission devices still affects performance, which further proves the necessity of retransmission device selection.



(a) Training loss versus training time (b) Training accuracy versus training time

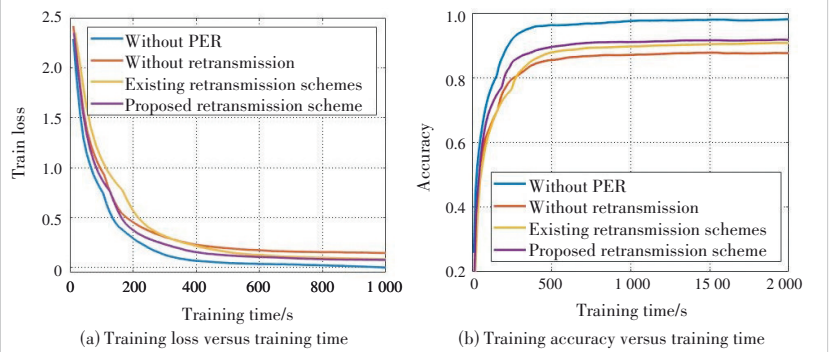
PER: packet error rate

▲ **Figure 4. Performance comparison between different retransmission schemes under CIFAR-10 with device number K=20**



(a) Training loss versus training time (b) Training accuracy versus training time

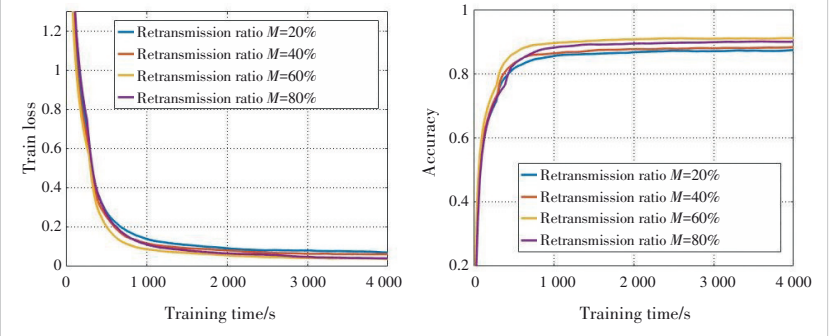
▲ **Figure 5. Performance comparison between different M under CIFAR-10**



(a) Training loss versus training time (b) Training accuracy versus training time

PER: packet error rate

▲ **Figure 6. Performance comparison between transmission schemes under MNIST**



(a) Training loss versus training time (b) Training accuracy versus training time

▲ **Figure 7. Performance comparison between different M under MNIST**

7 Conclusions

In this paper, we mainly study the retransmission design for FEEL under unreliable channels. We first analyze the impact of unreliable transmission on the training performance of the FEEL model, and derive the relation between the loss function and the channel PER in one round. Based on this, we analyze the gain to the convergence rate brought by device retransmission, as well as the communication cost introduced. Then, we propose a retransmission selection scheme for FEEL with unreliable channels, which can make a tradeoff between the training accuracy and the transmission latency. It comprehensively considers the channel conditions, the number of local data, and the importance of updates. We also present the air interface signaling and retransmission protocol design under the proposed retransmission selection scheme. Finally, the effectiveness of the proposed retransmission scheme is verified by extensive simulation experiments. The results show that our proposal can effectively reduce the impact of unreliable wireless channels on the training of the FEEL model, and is superior to the existing retransmission schemes.

Appendix A

Proof of Theorem 1

We first use the second-order Taylor expansion of $L\left(\boldsymbol{w}^{t+1}\right)$ to get

$$L\left(\boldsymbol{w}^{t+1}\right)=L\left(\boldsymbol{w}^{t}\right)+\left(\boldsymbol{w}^{t+1}-\boldsymbol{w}^{t}\right)\nabla L\left(\boldsymbol{w}^{t}\right)+\frac{1}{2}\left(\boldsymbol{w}^{t+1}-\boldsymbol{w}^{t}\right)^{T}\nabla^{2}L\left(\boldsymbol{w}^{t}\right)\left(\boldsymbol{w}^{t+1}-\boldsymbol{w}^{t}\right). \tag{26}$$

Based on Eq. (11) in Assumption 1, we can get

$$L\left(\mathbf{w}^{t+1}\right)\leqslant L\left(\mathbf{w}^{t}\right)+\left(\mathbf{w}^{t+1}-\mathbf{w}^{t}\right)g^{t}+\frac{1}{2}\beta\left\|\mathbf{w}^{t+1}-\mathbf{w}^{t}\right\|^{2}. \tag{27}$$

By taking expectation over both sides, it follows

$$\mathbb{E}\left\{L\left(\mathbf{w}^{t+1}\right)\right\}\leqslant\mathbb{E}\left\{L\left(\mathbf{w}^{t}\right)\right\}+\mathbb{E}\left\{-\eta\left(g^{t}-o^{t}\right)^{T}g^{t}\right\}+\frac{1}{2}\beta\eta^{2}\mathbb{E}\left\{\left\|g^{t}-o^{t}\right\|^{2}\right\}. \tag{28}$$

To remove the cross-term, we fix $\eta=\frac{1}{\beta}$. Then it follows

$$\begin{aligned}&\mathbb{E}\left\{L\left(\mathbf{w}^{t+1}\right)\right\}\leqslant\mathbb{E}\left\{L\left(\mathbf{w}^{t}\right)\right\}-\frac{1}{\beta}\mathbb{E}\left\{\left(g^{t}-o^{t}\right)^{T}g^{t}\right\}+\frac{1}{2\beta}\mathbb{E}\left\{\left\|g^{t}-o^{t}\right\|^{2}\right\}=\\&\mathbb{E}\left\{L\left(\mathbf{w}^{t}\right)\right\}-\frac{1}{2\beta}\mathbb{E}\left\{\left(g^{t}-o^{t}\right)^{T}g^{t}\right\}+\frac{1}{2\beta}\mathbb{E}\left\{\left(g^{t}-o^{t}\right)^{T}o^{t}\right\}=\\&\mathbb{E}\left\{L\left(\mathbf{w}^{t}\right)\right\}-\frac{1}{2\beta}\mathbb{E}\left\{\left(g^{t}-o^{t}\right)^{T}\left(g^{t}+o^{t}\right)\right\}=\mathbb{E}\left\{L\left(\mathbf{w}^{t}\right)\right\}-\frac{1}{2\beta}\mathbb{E}\left\{\left\|g^{t}\right\|^{2}\right\}+\frac{1}{2\beta}\mathbb{E}\left\{\left\|o^{t}\right\|^{2}\right\}.\end{aligned} \tag{29}$$

Thus, we have completed the proof of Theorem 1.

Appendix B

Proof of Lemma 1

First, the bias term can be expressed as the difference between the ground-truth gradient and the aggregated gradient, which can be expressed as

$$\begin{aligned}&\mathbb{E}\left\{\left\|o\right\|^{2}\right\}=\mathbb{E}\left\{\left\|g^{t}-\hat{g}^{t}\right\|^{2}\right\}=\\&\mathbb{E}\left\{\left\|\frac{\sum_{k=1}^{K}n_{k}\sum_{d=1}^{D}g_{k,i}}{n}-\frac{\sum_{k=1}^{K}n_{k}\sum_{d=1}^{D}\hat{g}_{k,i}}{n}\right\|^{2}\right\}=\\&\mathbb{E}\left\{\left\|\frac{\sum_{k=1}^{K}n_{k}\sum_{d=1}^{D}\left(1-\rho_{k,d}\right)g_{k,d}}{n}\right\|^{2}\right\}.\end{aligned} \tag{30}$$

By opening it with the sum of squares formula and substituting the probability of the indicator $\rho_{k,d}$, $P(\rho_{k,d}=0)=p_k$ and $P(\rho_{k,d}=1)=1-p_k$, we can get

$$\begin{aligned}&\mathbb{E}\left\{\left\|o\right\|^{2}\right\}=\frac{1}{n^{2}}\mathbb{E}\left\{\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\sum_{d_1=1}^{D}\sum_{d_2=1}^{D}n_{k_1}\left(1-\rho_{k_1,d_1}\right)g_{k_1,d_1}n_{k_2}\left(1-\rho_{k_2,d_2}\right)g_{k_2,d_2}\right\}=\\&\frac{1}{n^{2}}\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\sum_{d_1=1}^{D}\sum_{d_2=1}^{D}n_{k_1}p_{k_1}g_{k_1,d_1}n_{k_2}p_{k_2}g_{k_2,d_2}=\\&\frac{1}{n^{2}}\left(\sum_{k=1}^{K}n_{k}p_{k}\sum_{d=1}^{D}g_{k,d}\right)^{2}\leqslant\frac{K}{n^{2}}\sum_{k=1}^{K}n_{k}^{2}p_{k}^{2}\left(\sum_{d=1}^{D}g_{k,d}\right)^{2}.\end{aligned} \tag{31}$$

Denoting $\bar{g}_k=\sum_{d=1}^{D}g_{k,d}$, we can obtain the solution in Lemma 1.

Appendix C

Proof of Theorem 2

First, we take the first-order and second-order differentials of the objective function, and get

$$\begin{aligned}&\frac{\partial\sum_{k=1}^{K}\left(-\beta\Omega_{k}+\left(1-\beta\right)C_{k}\right)}{\partial\nu_{k}}=\\&-\frac{2\beta K}{n^{2}}n_{k}^{2}\bar{g}_{k}^{2}p_{k}^{2}\left(1-p_{k}-\nu_{k}\left(1-p_{k}\right)^{2}\right)+\left(1-\beta\right)\frac{qNp_{k}}{R_{k}},\end{aligned} \tag{32}$$

$$\frac{\partial^{2}\sum_{k=1}^{K}\left(-\beta\Omega_{k}+\left(1-\beta\right)C_{k}\right)}{\partial\nu_{k}^{2}}=\frac{2\beta K}{n^{2}}n_{k}^{2}\bar{g}_{k}^{2}p_{k}^{2}\left(1-p_{k}\right)^{2}\geqslant 0. \tag{33}$$

So the objective function of P2 is convex. In addition, Eq. (22a) is a linear constraint. Therefore, we can conclude that P2 is convex and we can use the KKT condition to find the optimal solution. We define the Lagrangian function $\mathcal{L}$ under the inequality constraints, as

$$\mathcal{L} = \sum_{k=1}^{K} \beta \frac{K}{n^2} n_k^2 \bar{g}_k^2 \left(p_k^2 - \left(p_k - \nu_k p_k (1 - p_k) \right)^2 \right) + (1-\beta) \frac{qNp_k}{R_k} \nu_k + \sum_{k=1}^{K} \mu_k (-\nu_k) + \sum_{k=1}^{K} \lambda_k (\nu_k - 1), \tag{34}$$

where $\mu_k \geqslant 0$ and $\lambda_k \geqslant 0$, which are both constraint coefficients of Eq. (22a). Let ν_k^* represent the optimal solution of P2. Then using the KKT condition, we can get

$$\frac{\partial \mathcal{L}}{\partial \nu_k^*} = -\frac{2\beta K}{n^2} n_k^2 \bar{g}_k^2 p_k^2 \left(1 - p_k - \nu_k^* (1 - p_k)^2 \right) + (1-\beta) \frac{qNp_k}{R_k} - \mu_k + \lambda_k, \forall k, \tag{35}$$

$$\mu_k (-\nu_k^*) = 0, \forall k, \tag{36}$$

$$\lambda_k (\nu_k^* - 1) = 0, \forall k. \tag{37}$$

By solving the above equations, we can get the optimal solution, as shown in Theorem 2.

Appendix D

Proof of Theorem 3

Taking the partial derivative of $f(p_k)$ over p_k, it follows

$$\frac{\partial f(p_k)}{\partial p_k} = \frac{p_k^3 (1 - p_k) + 2m_k (1 - 2p_k)}{p_k^3 (1 - p_k)}. \tag{38}$$

Then we define $h(p_k) = p_k^3 (1 - p_k) + 2m_k (1 - 2p_k)$. Taking the first-order and second-order differentials of $h(p_k)$, we have:

$$\frac{\partial h(p_k)}{\partial p_k} = 3p_k^2 - 4p_k^3 - 4m_k, \frac{\partial^2 h(p_k)}{\partial p_k^2} = 6p_k (1 - 2p_k). \tag{39}$$

Let $\frac{\partial^2 h(p_k)}{\partial p_k^2} = 0$, we have $\frac{\partial h(p_k)}{\partial p_k}$ that increases on (0, 0. 5) and decreases on (0. 5, 1). There is a unique p_k^* so that

$$h(p_k) \begin{cases} < 0, p_k \in (p_k^*, 0) \\ = 0, p_k = p_k^* \\ > 0, p_k \in (p_k^*, 1). \end{cases} \tag{40}$$

where p_k^* is related to m_k. And since $m_k > 0$, $p_k^* \in (0,1)$.

Therefore, we can prove that $f(p_k)$ increases on $(0, p_k^*)$ and decreases on $(p_k^*,1)$.

References

[1] ZHANG T, GAO L, HE C Y, et al. Federated learning for the Internet of Things: applications, challenges, and opportunities [J]. IEEE Internet of Things magazine, 2022, 5(1): 24 – 29. DOI: 10.1109/IOTM.004.2100182
[2] GUO F X, YU F R, ZHANG H L, et al. Enabling massive IoT toward 6G: a comprehensive survey [J]. IEEE Internet of Things journal, 2021, 8(15): 11891 – 11915. DOI: 10.1109/JIOT.2021.3063686
[3] MOHAMMADI F G, SHENAVARMASOULEH F, ARABNIA H R. Applications of machine learning in healthcare and Internet of Things (IOT): a comprehensive review [EB/OL]. [2022-10-10]. https://arxiv.org/abs/2202.02868
[4] VERBRAEKEN J, WOLTING M, KATZY J, et al. A survey on distributed machine learning [J]. ACM computing surveys, 2021, 53(2): 1 – 33. DOI: 10.1145/3377454
[5] MAJEED I A, KAUSHIK S, BARDHAN A, et al. Comparative assessment of federated and centralized machine learning [EB/OL]. [2022-10-10]. https://arxiv.org/abs/2202.01529
[6] GUPTA R, ALAM T. Survey on federated-learning approaches in distributed environment [J]. Wireless personal communications, 2022, 125(2): 1631 – 1652. DOI: 10.1007/s11277-022-09624-y
[7] JIANG Y L, ZHANG K, QIAN Y, et al. Anonymous and efficient authentication scheme for privacy-preserving distributed learning [J]. IEEE transactions on information forensics and security, 2022, 17: 2227 – 2240. DOI: 10.1109/TIFS.2022.3181848
[8] TRELEAVEN P, SMIETANKA M, PITHADIA H. Federated learning: the pioneering distributed machine learning and privacy-preserving data technology [J]. Computer, 2022, 55(4): 20 – 29. DOI: 10.1109/MC.2021.3052390
[9] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: challenges, methods, and future directions [J]. IEEE signal processing magazine, 2020, 37 (3): 50 – 60. DOI: 10.1109/MSP.2020.2975749
[10] LIU J, HUANG J Z, ZHOU Y, et al. From distributed machine learning to federated learning: A survey [J]. Knowledge and information systems, 2022, 64(4): 885 – 917. DOI: 10.1007/s10115-022-01664-x
[11] ALEDHARI M, RAZZAK R, PARIZI R M, et al. Federated learning: a survey on enabling technologies, protocols, and applications [J]. IEEE access: practical innovations, open solutions, 2020, 8: 140699 – 140725. DOI: 10.1109/access.2020.3013541
[12] ABREHA H G, HAYAJNEH M, SERHANI M A. Federated learning in edge computing: a systematic survey [J]. Sensor, 2022, 22(2): 450. DOI: 10.3390/s22020450
[13] LIM W Y B, LUONG N C, HOANG D T, et al. Federated learning in mobile edge networks: a comprehensive survey [J]. IEEE communications surveys & tutorials, 2020, 22(3): 2031 – 2063. DOI: 10.1109/COMST.2020.2986024
[14] NGUYEN D C, PHAM Q V, PATHIRANA P N, et al. Federated learning for smart healthcare: a survey [J]. ACM computing surveys, 2023, 55(3): 1 – 37. DOI: 10.1145/3501296
[15] ZHENG Z H, ZHOU Y Z, SUN Y L, et al. Applications of federated learning in smart cities: Recent advances, taxonomy, and open challenges [J]. Connection science, 2022, 34(1): 1 – 28. DOI: 10.1080/09540091.2021.1936455
[16] YANG Z H, CHEN M Z, SAAD W, et al. Energy efficient federated learning over wireless communication networks [J]. IEEE transactions on wireless communications, 2021, 20(3): 1935 – 1949. DOI: 10.1109/TWC.2020.3037554
[17] CHEN M Z, YANG Z H, SAAD W, et al. A joint learning and communications framework for federated learning over wireless networks [J]. IEEE transactions on wireless communications, 2021, 20(1): 269 – 283. DOI: 10.1109/TWC.2020.3024629
[18] NADEEM F, LI Y H, VUCETIC B, et al. Analysis and optimization of HARQ

for URLLC [C]//IEEE Globecom Workshops. IEEE, 2022: 1 – 6. DOI: 10.1109/GCWkshps52748.2021.9682028
[19] JIANG P W, WEN C K, JIN S, et al. Deep source-channel coding for sentence semantic transmission with HARQ [J]. IEEE transactions on communications, 2022, 70(8): 5225 – 5240. DOI: 10.1109/TCOMM.2022.3180997
[20] SHIRVANIMOGHADDAM M, SALARI A, GAO Y F, et al. Federated learning with erroneous communication links [J]. IEEE communications letters, 2022, 26(6): 1293 – 1297. DOI: 10.1109/LCOMM.2022.3167094
[21] SALARI A, SHIRVANIMOGHADDAM M, VUCETIC B, et al. Rate-convergence tradeoff of federated learning over wireless channel [EB/OL]. [2022-10-10]. https://arxiv.org/abs/2205.04672
[22] YE H, LIANG L, LI G Y. Decentralized federated learning with unreliable communications [J]. IEEE journal of selected topics in signal processing, 2022, 16(3): 487 – 500. DOI: 10.1109/JSTSP.2022.3152445
[23] JEONG E, ZECCHIN M, KOUNTOURIS M. Asynchronous decentralized learning over unreliable wireless networks [EB/OL]. [2022-10-10]. https://arxiv.org/abs/2202.00955
[24] LI Z D, ZHOU Y J, WU D P, et al. Fairness-aware federated learning with unreliable links in resource-constrained Internet of Things [J]. IEEE Internet of Things journal, 2022, 9(18): 17359 – 17371. DOI: 10.1109/JIOT.2022.3156046
[25] MAO Y Z, ZHAO Z H, YANG M L, et al. SAFARI: sparsity enabled federated learning with limited and unreliable communications [EB/OL]. [2022-10-10]. https://arxiv.org/abs/2204.02321
[26] SALEHI M, HOSSAIN E. Federated learning in unreliable and resource-constrained cellular wireless networks [J]. IEEE transactions on communications, 2021, 69(8): 5136 – 5151. DOI: 10.1109/TCOMM.2021.3081746
[27] JIANG Z H, YU G D, CAI Y L, et al. Decentralized edge learning via unreliable device-to-device communications [J]. IEEE transactions on wireless communications, 2022, 21(11): 9041 – 9055. DOI: 10.1109/TWC.2022.3172147
[28] NADEEM F, LI Y H, VUCETIC B, et al. HARQ optimization for real-time remote estimation in wireless networked control [EB/OL]. [2022-10-10]. https://arxiv.org/abs/2201.05838
[29] SHAH S W H, RAHMAN M M U, MIAN A N, et al. Effective capacity analysis of HARQ-enabled D2D communication in multi-tier cellular networks [J]. IEEE transactions on vehicular technology, 2021, 70(9): 9144 – 9159. DOI: 10.1109/TVT.2021.3100675
[30] LIU D Z, ZHU G X, ZENG Q S, et al. Wireless data acquisition for edge learning: data-importance aware retransmission [J]. IEEE transactions on wireless communications, 2021, 20(1): 406 – 420. DOI: 10.1109/TWC.2020.3024980
[31] SIMONSSON A, FURUSKAR A. Uplink power control in LTE - overview and performance, subtitle: principles and benefits of utilizing rather than compensating for SINR variations [C]//IEEE 68th Vehicular Technology Conference. IEEE, 2008: 1 – 5. DOI: 10.1109/VETECF.2008.317
[32] XI Y, BURR A, WEI J B, et al. A general upper bound to evaluate packet error rate over quasi-static fading channels [J]. IEEE transactions on wireless communications, 2011, 10(5): 1373 – 1377. DOI: 10.1109/TWC.2011.012411.100787

Biographies

XU Xinyi received her BE degree in communication engineering from Zhejiang University, China in 2021. Now she is working towards her MS degree with the College of Information Science and Electronic Engineering, Zhejiang University. Her research interest focuses on federated learning.

LIU Shengli received his BS degree in information engineering from Soochow University, China in 2017, and his PhD degree from the College of Information Science and Electronic Engineering, Zhejiang University, China in 2022. He currently holds a post-doctoral position at the College of Information Science and Electronic Engineering, Zhejiang University. In 2021, he was a Visiting Research Scholar with the Centre for Wireless Communication, University of Oulu, Finland and the VTT Technical Research Centre of Finland. His current research interests mainly include machine learning and federated learning.

YU Guanding (yuguanding@zju.edu.cn) received his BE and PhD degrees in communication engineering from Zhejiang University, China in 2001 and 2006, respectively. He joined Zhejiang University in 2006 and is now a professor with the College of Information and Electronic Engineering. From 2013 to 2015, he was also a visiting professor at the School of Electrical and Computer Engineering, Georgia Institute of Technology, USA. His research interests include 5G communications and networks, mobile edge computing, and machine learning for wireless networks.