

# Wireless Federated Learning over Resource-Constrained Networks: Digital versus Analog Transmissions

Jiacheng Yao, *Graduate Student Member, IEEE*, Wei Xu, *Senior Member, IEEE*, Zhaohui Yang, *Member, IEEE*, Xiaohu You, *Fellow, IEEE*, Mehdi Bennis, *Fellow, IEEE*, and H. Vincent Poor, *Life Fellow, IEEE*

**Abstract**—To enable wireless federated learning (FL) in communication resource-constrained networks, two communication schemes, i.e., digital and analog ones, are effective solutions. In this paper, we quantitatively compare these two techniques, highlighting their essential differences as well as respectively suitable scenarios. We first examine both digital and analog transmission schemes, together with a unified and fair comparison framework under imbalanced device sampling, strict latency targets, and transmit power constraints. A universal convergence analysis under various imperfections is established for evaluating the performance of FL over wireless networks. These analytical results reveal that the fundamental difference between the digital and analog communications lies in whether communication and computation are jointly designed or not. The digital scheme decouples the communication design from FL computing tasks, making it difficult to support uplink transmission from massive devices with limited bandwidth and hence the performance is mainly communication-limited. In contrast, the analog communication allows over-the-air computation (AirComp) and achieves better spectrum utilization. However, the computation-oriented analog transmission reduces power efficiency, and its performance is sensitive to computation errors from imperfect channel state information (CSI). Furthermore, device sampling for both schemes are optimized and differences in sampling optimization are analyzed. Numerical results verify the theoretical analysis and affirm the superior performance of the sampling optimization.

**Index Terms**—Federated learning (FL), digital communication, over-the-air computation (AirComp), convergence analysis.

## I. INTRODUCTION

This work was supported in part by the National Key Research and Development Program under Grant 2020YFB1806608, and the Fundamental Research Funds for the Central Universities 2242022k60002 and 2242023K5003. The work of J. Yao was supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX24\_0407. The work of Z. Yang was supported by the Young Elite Scientists Sponsorship Program by CAST 2023QNRC001, Zhejiang Key R&D Program under Grant 2023C01021 and the Fundamental Research Funds for the Central Universities K2023QA0AL02. The work of M. Bennis was supported by the European Union's Horizon 2020 Research and Innovation project ADROIT6G project of the SNS-JU under Grant Agreement No. 101095363, and Project CENTRIC under Grant 101096379, through the Project SNS JU VERGE under Grant 101096034. The work of H. V. Poor was supported by the U.S. National Science Foundation under Grants CNS-2128448 and ECCS-2335876. Part of this work was presented in *IEEE ICC 2024* [1]. (*Corresponding author: Wei Xu*)

J. Yao, W. Xu and X. You are with the National Mobile Communications Research Laboratory (NCRL), Southeast University, Nanjing 210096, China, and also with Purple Mountain Laboratories, Nanjing 211111, China ({jcyao, wxu, xhyu}@seu.edu.cn).

Z. Yang is with the Zhejiang Lab, Hangzhou 311121, China, and also with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, Zhejiang 310027, China (yang\_zhaohui@zju.edu.cn).

M. Bennis is with the Center for Wireless Communications, Oulu University, Oulu 90014, Finland (e-mail: mehdi.bennis@oulu.fi).

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

THE dramatic development of data science has catalyzed significant advances in artificial intelligence (AI), which is driving innovation for anticipated sixth-generation (6G) mobile networks. The integration of AI and communication is envisioned to drive the shift from connected things to ubiquitous connected intelligence in wireless networks, supporting a large number of emerging intelligent applications [2]–[5]. Nonetheless, traditional centralized learning paradigms depend on extensive data transmission and considerable computational resources at cloud servers, which is challenging to implement in wireless networks. To better embrace AI, edge learning (EL) is viewed as a promising distributed learning technique that harnesses massive data and computational capacity available in edge devices distributed across wireless networks [6]–[8]. Distinguishing it from the traditional separate design for computation and communication, EL integrates the two and achieves efficient utilization of resources and improves performance through learning task-oriented communication design.

In particular, a key EL paradigm, namely federated learning (FL), has garnered significant attention from both academic and industrial circles, primarily due to its communication-efficient and privacy-enhancing characteristics [9], [10]. In FL, distributed edge devices utilize local datasets to collaboratively train a shared learning model with the assistance of a central parameter server (PS). By exchanging model parameters instead of raw data, the PS iteratively updates the global model until convergence. FL scheme minimizes the amount of transmitted data, as well as helping safeguard privacy and security. Recent studies have explored implementation of FL algorithms at wireless edge to support emerging AI applications [11]–[14]. However, limited communication resources has posed a significant bottleneck to the performance of wireless FL [15], [16]. One particular concern regards the uplink transmission process, where numerous participating devices need to transmit local updates to the PS, leading to a substantial increase in communication overhead and transmission latency [17]. Hence, the development of efficient uplink transmission is crucial to enable wireless FL.

To support data transmission in wireless FL, digital communication schemes have been widely considered in recent works, where local updates are quantized into finite bits and then transmitted to the PS via traditional frequency division multiple access (FDMA) and time division multiple access (TDMA) schemes. At the receiver, the PS relies on channel coding for error detection and correction, before model aggregation using the received local updates. In [12] and [18], the authors characterized the impact of packet errors on the convergence of FL, which enabled a task-oriented communication resource allocation scheme. The influence of

various finite-precision quantization schemes in uplink and downlink communications was considered in [19]. Building upon convergence analysis of the quantized FL, the quantization bits allocation was optimized in [20] and [21] to adapt channel diversity and requirements of the FL tasks. To further alleviate the communication bottleneck, one-bit quantization technique and reconfigurable intelligent surface (RIS) were used in [22] to reduce communication overhead and enhance communication reliability, respectively. Apart from resource allocation methods, modifications from the algorithmic perspective have been considered to combat unreliable transmissions. In [23], the authors proposed a user datagram protocol (UDP)-based robust training algorithm, which asymptotically achieved the same convergence rate as that with error-free communications. Moreover in [24], for replacing erroneous local updates, a global model reusing scheme, namely the GoMORE scheme, was devised to successfully mitigate the negative impacts of packet loss. Alternatively, another solution is to further squeeze the communication overhead, thus improving the convergence over resource-constrained networks. The model pruning in [25] was seen to be an effective way to compress the large-scale model into a smaller size, facilitating communication-efficient FL design.

In addition to these digital communication schemes, analog communication is an alternative communication-efficient way for deploying wireless FL. In particular, the local updates are amplitude-modulated and then simultaneously transmitted by reusing the available radio resource. Due to the superposition property of radio channels, the global model can be computed automatically over-the-air, which is therefore referred to as over-the-air computation (AirComp) [26]. Unlike the digital paradigm, analog communication pushes model aggregation from the PS to the air, which not only functionally but physically integrates the computation and communication. Benefiting from the over-the-air aggregation, the communication latency is substantially reduced and the spectrum utilization is much more efficient, leading to fast-convergent and communication-efficient FL. It was shown in [27] that the convergence rate of centralized learning remains approachable with this analog approach without power control and beamforming. Furthermore in [28], to combat deep fading, a novel truncated channel inversion scheme was proposed to exclude devices experiencing deep fades from the training process avoiding excessive energy consumption. Further insights into analog aggregation schemes were also discussed in the context of fundamental trade-offs between communication and learning. Besides, the impact of over-the-air aggregation errors on optimality gap was analyzed in [29] and [30] with power control optimization. Furthermore, the authors in [31] proposed an AirComp-based adaptive reweighing scheme for the aggregation, and jointly considered the power control and device selection design based on the derived optimality gap. To combat the additive noise, robust FL training methods were proposed in [32] for both the expectation-based and the worst-case noise models. Considering multi-antenna scenarios, the beamforming design at the receiver was optimized by solving a sparse and low-rank optimization problem in [33]. In practice, considering the lack of perfect channel state information (CSI)

for accurate power control, the work [34] investigated the impact of CSI uncertainty at the transmitter on FL convergence and revealed that CSI imperfection plays a key factor affecting the AirComp performance and convergence.

As mentioned above, by incorporating learning task-oriented resource allocation, both digital and analog transmissions are effective ways to fulfill the communication requirements of wireless FL [35]–[37]. In traditional communication for data transmission, digital communication schemes have been proven not only in theory but also in practice as dominantly outperforming analog communication techniques in almost all cases of interest. In communications for computation tasks, however, analog communication has shown to be exceptionally effective in some cases of resource-constrained networks [38]. Hence, it is of interest to comprehensively compare digital and analog transmissions for wireless FL. Several recent studies have compared the two communication paradigms from some specific perspectives, including communication latency [28], [39] and convergence performance [40], [41]. However, to the best of our knowledge, there is a lack of literature that presents a comprehensive and quantitative comparison between the two fundamental communication paradigms, especially under practical constraints. Also, there have been few attempts to elucidate the fundamental differences between digital and analog transmissions in the context of FL, which is crucial for its deployment and design.

Against this background, in this paper, we conduct a theoretical comparison between the digital and analog transmission schemes under practical constraints. The main contributions of this paper are summarized as follows.

- We propose a unified framework for digital and analog transmissions in wireless FL, and characterize the model aggregation distortion caused by wireless transmission schemes. Using this framework, a fair comparison is conducted under the consideration of a stringent transmission delay target and two types of transmit power budgets. We exploit optimality gap, defined by the gap between the optimal and actually achieved loss function value, to characterize the convergence behavior and establish a stringent upper bound of the optimality gap for precise analysis and optimization in the digital/analog transmission enabled wireless FL. It offers a precise characterization of the influence of wireless transmission imperfections on convergence in closed-form.
- Analytical results reveal that the digital transmission is hard to achieve satisfactory performance especially with limited radio resources due to orthogonal access and decoupled design. In contrast, the analog scheme exhibits a performance gain in terms of the optimality gap of the order of  $\frac{1}{N}$  with the increasing number of participating devices,  $N$ , and thereby achieving a higher level of efficiency in spectrum utilization. However, the introduction of computation goals in the analog communication process results in less efficient transmit power utilization, and the presence of CSI uncertainties inevitably comes with computational distortion, thus enlarging the optimality gap by the order of  $\frac{1}{\rho^2}$  with a decreasing level of channel estimation accuracy  $\rho$ .

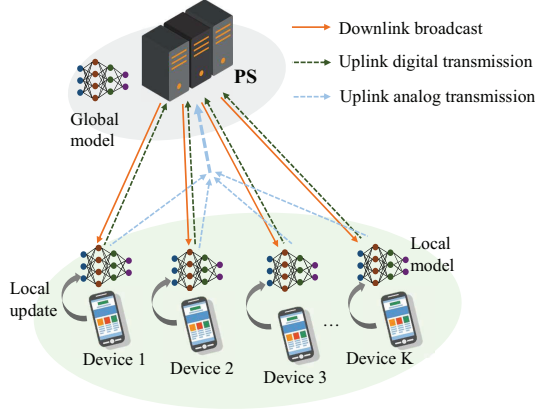


Fig. 1. The architecture of a typical wireless FL system.

- Based on the derived optimality gap, we formulate an inclusion probability optimization problem for effective device sampling in wireless FL. The optimization problems for both digital and analog cases are optimally solved by checking the Karush-Kuhn-Tucker (KKT) conditions and exploiting the Dinkelbach algorithm, respectively. Through the examination of optimal solutions, we identify the essential differences underlying the device sampling optimization for digital and analog transmissions.

Extensive numerical simulations are conducted to validate the derived analytical observations and the proposed sampling optimization. In particular, it is observed that the digital scheme has better power utilization, while the analog transmission is more spectrum-efficient.

The rest of this paper is organized as follows. In Section II, we describe the typical FL algorithm, with details of digital and analog transmissions, and propose a fair comparison framework. Section III provides some preliminaries for the convergence analysis. In Section IV, we analyze the convergence performance under different transmission schemes and offer engineering insights. Then, in Section V, we optimize the inclusion probabilities for both the digital and analog schemes. Simulation results and conclusions are given in Sections VI and VII, respectively.

**Notation:** Boldface lowercase (uppercase) letters represent vectors (matrices). The set of all real numbers is denoted by  $\mathbb{R}$ . Superscripts  $(\cdot)^T$  and  $(\cdot)^*$  stand for the transpose and conjugate operations, respectively. The operator  $\Re(\cdot)$  returns the real part of the input complex number. The operator  $\|\cdot\|$  takes the Euclidean norm of vectors. A circularly symmetric complex Gaussian distribution is denoted by  $\mathcal{CN}$ , and  $\mathbb{E}\{\cdot\}$  is the expectation operation.

## II. SYSTEM MODEL AND COMMUNICATION FRAMEWORK

We consider a typical wireless FL system as shown in Fig. 1, where  $K$  distributed devices are coordinated by a central PS to perform FL. The training procedure and transmission model are elaborated in the sequel.

### A. Federated Learning Model

In FL, the distributed devices collaboratively train a shared machine learning model via local computing based on their

local datasets and information exchange with the PS. Let  $\mathcal{D}_k$  denote the local dataset owned by the  $k$ -th device, which contains  $D_k = |\mathcal{D}_k|$  training samples. The goal of the FL algorithm is to find the optimal  $d$ -dimensional model parameter vector, denoted by  $\mathbf{w}^* \in \mathbb{R}^{d \times 1}$ , to minimize the global loss function  $F(\mathbf{w})$ , i.e.,

$$\begin{aligned} \mathbf{w}^* &= \arg \min_{\mathbf{w}} F(\mathbf{w}) = \arg \min_{\mathbf{w}} \frac{1}{D} \sum_{k=1}^K D_k F_k(\mathbf{w}) \\ &= \arg \min_{\mathbf{w}} \sum_{k=1}^K \alpha_k F_k(\mathbf{w}), \end{aligned} \quad (1)$$

where  $D \triangleq \sum_{k=1}^K D_k$ ,  $\alpha_k \triangleq \frac{D_k}{D}$  represents the aggregation weight for the  $k$ -th user, and  $F_k(\mathbf{w})$  is the local loss function at device  $k$  defined as

$$F_k(\mathbf{w}) = \frac{1}{D_k} \sum_{\mathbf{u} \in \mathcal{D}_k} \mathcal{L}(\mathbf{w}, \mathbf{u}), \quad (2)$$

where  $\mathbf{u}$  denotes a training sample selected from  $\mathcal{D}_k$ , and  $\mathcal{L}(\mathbf{w}, \mathbf{u})$  represents the sample-wise loss function with respect to  $\mathbf{u}$ . Due to the heterogeneity of the system, we note that local datasets at distinct devices are usually non-independent and non-identically distributed (non-IID), and the optimal model parameters in (1) are not necessarily the optimal for local datasets. Let  $\mathbf{w}_k^*$  denote the locally optimal model at device  $k$ , i.e.,  $\mathbf{w}_k^* = \arg \min_{\mathbf{w}} F_k(\mathbf{w})$ . It is usually different from the globally optimal  $\mathbf{w}^*$  unless the local dataset  $\mathcal{D}_k$  experiences the same distribution as the whole data population.

To effectively handle the optimization problem in (1), an FL algorithm performs the model training in an iterative manner. Specifically, the  $m$ -th round of the FL algorithm consists of the following steps.

- 1) *Model Broadcasting:* The PS broadcasts the latest global model  $\mathbf{w}_m$  to all devices.
- 2) *Local Computing:* After receiving  $\mathbf{w}_m$ , each device exploits its local dataset to compute the local gradient as

$$\mathbf{g}_m^k \triangleq \nabla F_k(\mathbf{w}_m) = \frac{1}{D_k} \sum_{\mathbf{u} \in \mathcal{D}_k} \nabla \mathcal{L}(\mathbf{w}_m, \mathbf{u}), \quad \forall k. \quad (3)$$

- 3) *Local Update Uploading:* Each device reports its local gradient to the PS.
- 4) *Model Aggregation:* Upon receiving all local gradients, the PS updates the global model according to

$$\mathbf{w}_{m+1} = \mathbf{w}_m - \eta \mathbf{g}_m, \quad (4)$$

where  $\eta$  is the learning rate and  $\mathbf{g}_m$  is given by

$$\mathbf{g}_m \triangleq \sum_{k=1}^K \alpha_k \mathbf{g}_m^k. \quad (5)$$

The above steps iterate until a convergence condition is met.

Considering the potentially massive number of devices and limited resources in practice, only a subset of devices can participate in each round of the training. Let  $\mathcal{S}_m$  denote the set of activated devices selected in the  $m$ -th communication round and  $N = |\mathcal{S}_m|$  be the number of participating devices per round. Due to imbalanced dataset sizes and data heterogeneity,

we assume that the PS performs non-uniform device sampling without replacement to select the participating devices per round. Specifically, the devices are randomly selected one by one from the remaining unselected device set. Once the number of selected devices reaches  $N$ , the sampling process terminates. Denote the inclusion probability of the device  $k$  as  $r_k$ , which represents the probability of device  $k$  being sampled per round and satisfies  $r_k \leq 1, \forall k$ , and  $\sum_{k=1}^K r_k = N$ . Due to the non-IID nature of the data, misaligned inclusion probability may bias the global model away from the local optimum, thereby decelerating the convergence and causing performance loss. Hence, in the following sections, we focus on the performance evaluation under fixed inclusion probabilities and characterize the impact of device sampling for wireless FL.

Also, in wireless FL, the parameter transmission in Steps 1) and 3) relies on wireless communication between the PS and devices, which comes with additional imperfection in the model training procedure. Considering a sufficient power budget at the PS, the downlink transmission is usually assumed error-free [12]. Otherwise, for uplink transmission with limited communication resources, additional errors are inevitable. Efficient transmission and resource allocation schemes need to be designed to alleviate this impact of wireless environment.

### B. Uplink Transmission Method

We rely on the wireless uplink transmission to provide an estimation of the actual gradient in (5). Assume that the total uplink bandwidth  $B$  can be divided into up to  $M$  subbands, which supports orthogonal access for  $M$  devices. Without loss of generality, a frequency non-selective block fading channel model is adopted, where the wireless channels remain unchanged within a communication round. Let  $\bar{h}_k = d_k^{-\frac{\alpha}{2}} h_k$  be the channel between the  $k$ -th device and the PS, where  $d_k$  denotes the distance between the PS and device  $k$ ,  $\alpha$  represents the large-scale path loss exponent, and  $h_k$  represents the small-scale fading of the channel. Assume that the channels are independent Rayleigh fadings, i.e.,  $h_k \sim \mathcal{CN}(0, 1)$ . In practice, perfect estimation of the small-scale fading of the channel is usually not available. Let  $\hat{h}_k$  denote the estimated channel at device  $k$ . Then, we model the CSI imperfection of the small-scale fading as

$$h_k = \rho \hat{h}_k + \sqrt{1 - \rho^2} v_k, \quad (6)$$

where  $\rho \in (0, 1]$  is the correlation coefficient between  $h_k$  and  $\hat{h}_k$  to reflect the level of channel estimation accuracy, and  $v_k \sim \mathcal{CN}(0, 1)$  is the channel estimation error independent of  $\hat{h}_k$ . In the following, we introduce two typical uplink transmission schemes, i.e., digital and analog transmissions.

1) *Digital Transmission Model*: In the digital transmission, the  $N$  selected devices first quantize their local updates into a finite number of  $b$  bits and then simultaneously transmit the quantized local updates to the PS. Specifically, we assume that the local update  $\mathbf{g}_m^k$  is quantized by the stochastic quantization method in [20]. Denote the maximum and the minimum values of the modulus among all parameters in  $\mathbf{g}_m^k$  by  $g_{m,\max}^k$  and  $g_{m,\min}^k$ , respectively. Then, the interval  $[g_{m,\min}^k, g_{m,\max}^k]$  is divided evenly into  $2^b - 1$  quantization intervals. The uniformly

distributed knobs are denoted by  $\tau_i = g_{m,\min}^k + \frac{g_{m,\max}^k - g_{m,\min}^k}{2^b - 1} i$  for  $i = 0, \dots, 2^b - 1$ . Given  $|x| \in [\tau_i, \tau_{i+1})$ , the quantization function  $\mathcal{Q}(x)$  is expressed as

$$\mathcal{Q}(x) = \begin{cases} \text{sign}(x)\tau_i & \text{w.p. } \frac{\tau_{i+1} - |x|}{\tau_{i+1} - \tau_i}, \\ \text{sign}(x)\tau_{i+1} & \text{w.p. } \frac{|x| - \tau_i}{\tau_{i+1} - \tau_i}, \end{cases} \quad (7)$$

where  $\text{sign}(\cdot)$  represents the signum function and “w.p.” represents “with probability.” Exploiting the quantization function in (7), the local update  $\mathbf{g}_m^k$  is quantized as  $\mathcal{Q}(\mathbf{g}_m^k) \triangleq [\mathcal{Q}(g_{m,1}^k), \dots, \mathcal{Q}(g_{m,d}^k)]^T$ , which is transmitted to the PS. Note that the exact value of  $g_{m,\max}^k$  and  $g_{m,\min}^k$  need to be transmitted to the PS with sufficient precision to support effective recovery. Hence, the total number of bits needed for transmitting amounts to

$$b_{\text{total}} = d(b + 1) + q, \quad (8)$$

where  $q$  is the number of bits used to represent  $g_{m,\max}^k$  and  $g_{m,\min}^k$ , and the additional one bit is the sign bit.

During the uplink FL parameter report, transmission errors are inevitable due to the channel dynamics and limited communication resources. Without loss of generality, we adopt the typical FDMA technique as an example. Assume that  $M \geq N$  and hence each device can occupy different subbands equally to avoid interference with each other.<sup>1</sup> Then, the channel capacity of device  $k$  can be evaluated as

$$C_k = B_k \log_2 \left( 1 + \frac{P_k |\bar{h}_k|^2}{B_k N_0} \right), \quad (9)$$

where  $B_k$  is the bandwidth allocated to device  $k$  and it is set to  $\frac{B}{N}$ ,  $P_k$  is the transmit power at device  $k$ , and  $N_0$  is the noise power density.

The transmission delay under the digital transmission is primarily influenced by stragglers, which refer to devices with poor channel conditions. To avoid the uncontrolled severe delay brought by stragglers, we assume that all the devices transmit the local updates at a fixed rate rather than a dynamic one based on instantaneous signal-to-noise ratio (SNR) levels. Hence, the use of a fixed-rate transmission acts as a truncation mechanism for stragglers. Additionally, for devices experiencing favorable channel conditions, it is more beneficial to transmit at a lower rate with enhanced transmission reliability. The target transmission rate is denoted by  $R = \frac{B}{N} \log_2(1 + \theta)$ , where  $\theta$  is a chosen constant. According to [12], the transmission is assumed error-free if the transmission rate is no larger than the channel capacity. Hence, the probability of successful transmission at device  $k$  is calculated as

$$p_k = \Pr \{R \leq C_k\} = \exp \left( -\frac{BN_0\theta}{2NP_k d_k^{-\alpha}} \right). \quad (10)$$

At the PS, a cyclic redundancy check (CRC) mechanism is applied to check the detected data such that erroneous local

<sup>1</sup>We generally assume orthogonal access between different devices and refrain from specifying the particular multiple access design. Hence, the following analysis can be safely extended to orthogonal access scenarios like TDMA and orthogonal frequency division multiple access (OFDMA).

updates can be excluded from the model aggregation. Finally, the obtained estimate of the desired gradient in (5) is given by

$$\hat{\mathbf{g}}_{m,D} = \sum_{k=1}^K \frac{\chi_k \alpha_k \xi_{k,D}}{r_k} \mathcal{Q}(\mathbf{g}_m^k), \quad (11)$$

where  $\chi_k$  is an indicator variable for the device selection, and  $\xi_{k,D}$  represents distortion brought by packet loss. To be concrete,  $\chi_k$  is 1 if  $k \in \mathcal{S}_m$  and otherwise  $\chi_k$  is 0. Considering the definition of the inclusion probability, we have  $\mathbb{E}[\chi_k] = r_k \leq 1$ , which decreases the desired expected aggregation coefficient for unbiased gradient estimation. In order to compensate for the impact of partial participation, we multiply the coefficient  $\frac{1}{r_k}$  in (11), such that  $\frac{1}{r_k} \mathbb{E}[\chi_k] = 1$ . Analogously, the distortion  $\xi_{k,D}$  is characterized by the probability in (10) as

$$\xi_{k,D} = \begin{cases} \frac{1}{p_k} & \text{w.p. } p_k, \\ 0 & \text{w.p. } 1 - p_k, \end{cases} \quad (12)$$

to ensure  $\mathbb{E}[\xi_{k,D}] = 1$ . With the gradient estimate in (11), the global model updated at the  $(m+1)$ -th round equals to

$$\tilde{\mathbf{w}}_{m+1} = \tilde{\mathbf{w}}_m - \eta \hat{\mathbf{g}}_{m,D}, \quad (13)$$

where  $\tilde{\mathbf{w}}_m$  denotes the model obtained at the previous round.

2) *Analog Transmission Model:* In the analog transmission with AirComp, selected devices simultaneously upload the uncoded analog signals of local gradients to the PS by fully reusing the time-frequency resource. A weighted summation of the local updates in (5) can be achieved by exploiting channel pre-equalization and the waveform superposition nature of the wireless channel. In this study, we consider that the total bandwidth is constrained for fair comparison and all subbands are utilized for the transmission of identical parameters. This is because the uncoded nature of the analog transmission diminishes its robustness, rendering it more vulnerable to interference and even the malicious attacks.<sup>2</sup> Specifically, the received signal at the PS is expressed as

$$\mathbf{y} = \sum_{k=1}^K \chi_k \bar{h}_k \beta_k \mathbf{g}_m^k + \mathbf{z}_m, \quad (14)$$

where  $\beta_k$  is the pre-processing factor at device  $k$ , and  $\mathbf{z}_m$  is additive white Gaussian noise following  $\mathcal{CN}(\mathbf{0}, B N_0 \mathbf{I})$ . To accurately estimate the desired gradient in (5), the pre-processing factor  $\beta_k$  should be adapted to the channel coefficient  $\bar{h}_k$ . Unlike the digital transmission, CSI is needed at the transmitter for the analog transmission. Channel pre-equalization is performed based on the CSI available at each device. For simplicity, we adopt the typical truncated channel inversion scheme to combat deep fades [28]. It is expressed as

$$\beta_k = \begin{cases} \frac{\zeta \lambda \alpha_k d_k^{\frac{\alpha}{2}} \bar{h}_k^*}{r_k |\bar{h}_k|^2} & |\hat{h}_k|^2 \geq \gamma_{\text{th}}, \\ 0 & |\hat{h}_k|^2 < \gamma_{\text{th}}, \end{cases} \quad (15)$$

<sup>2</sup>The derived results directly extend to the case of dividing bandwidth for distinct parameter transmission in broadband scenarios [28].

TABLE I  
MAIN DIFFERENCES BETWEEN THE TWO PARADIGMS

Paradigms	Gradient estimation	Transmission delay	Power budget
Digital	(11)	(21)	(23)
Analog	(16)	(24)	(25), (26)

where  $\gamma_{\text{th}}$  is a predetermined power-cutoff threshold,  $\zeta$  is a scaling factor for ensuring the transmit power constraint, and compensation coefficient  $\lambda$  is selected to alleviate the impact of imperfect CSI [34]. Through the pre-processing in (15), we aim to eliminate the influence of the uneven channel fading  $\bar{h}_k$ , and the inclusion probability  $p_k$ , thereby ensuring the unbiased gradient estimation.

At the receiver, the PS scales the real part of  $\mathbf{y}$  in (14) with  $\frac{1}{\zeta}$  and obtain an estimate of the actual gradient in (5). It yields

$$\hat{\mathbf{g}}_{m,A} = \sum_{k=1}^K \frac{\chi_k \alpha_k \xi_{k,A}}{r_k} \mathbf{g}_m^k + \bar{\mathbf{z}}_m, \quad (16)$$

where  $\bar{\mathbf{z}}_m \triangleq \frac{\Re\{\mathbf{z}_m\}}{\zeta}$  is the equivalent noise, and  $\xi_{k,A}$  denotes the distortion brought by the analog transmission with imperfect CSI. It follows

$$\xi_{k,A} = \begin{cases} \frac{\lambda \Re\{\bar{h}_k^* \hat{h}_k\}}{|\hat{h}_k|^2} & \text{w.p. } e^{-\gamma_{\text{th}}}, \\ 0 & \text{w.p. } 1 - e^{-\gamma_{\text{th}}}. \end{cases} \quad (17)$$

Similarly, the global model at the  $(m+1)$ -th round under the analog transmission is updated as

$$\tilde{\mathbf{w}}_{m+1} = \tilde{\mathbf{w}}_m - \eta \hat{\mathbf{g}}_{m,A}. \quad (18)$$

### C. A Unified Framework for Wireless FL Comparison

To minimize the optimality gap brought by imperfect uplink transmission, the overall FL task-oriented optimization over the wireless networks can be formulated as

$$\begin{aligned} & \text{minimize} && \mathbb{E}[F(\mathbf{w}_{m+1})] - F(\mathbf{w}^*) \\ & \text{subject to} && C_1 : T \leq T_{\max}, \\ & && C_2 : P_k \leq P_{\max}, \quad \forall k, \end{aligned} \quad (19)$$

where the expectation is taken over channel dynamics,  $T$  represents uplink transmission delay per round,  $T_{\max}$  and  $P_{\max}$  denotes the maximum transmission delay target and the transmit power unit, respectively. Constraint  $C_1$  and  $C_2$  respectively represent the maximum transmission delay and maximum transmit power constraint in practice. Apart from the maximum power budget, another typical transmit power constraint is the average power budget [28], i.e.,

$$\bar{C}_2 : \mathbb{E}[P_k] \leq P_{\text{ave}}, \quad \forall k, \quad (20)$$

where  $P_{\text{ave}}$  denotes the average power budget and limits the energy consumption during the uplink transmission process.

For fair comparison between the two transmission paradigms, we measure the achievable objective value of the problem in (19) under the same transmission delay target and transmit power budget. Specific constraints for the two transmission paradigms are listed as follows, summarized in Table I.

For digital transmission, the transmission delay per communication round is calculated as

$$T_D = \frac{b_{\text{total}}}{R} = \frac{Nd(b+1)}{B \log_2(1+\theta)}, \quad (21)$$

where the evaluation holds with a sufficiently large model size  $d$ . Hence, constraint  $C_1$  is reformulated as

$$\frac{Nd(b+1)}{B \log_2(1+\theta)} \leq T_{\max} \Rightarrow \theta \geq 2^{\frac{Nd(b+1)}{BT_{\max}}} - 1. \quad (22)$$

For constraint  $C_2$ , due to its interference-free characteristic, full power transmission is optimal and hence the constraint is reformulated by

$$P_k = P_{\max}, \quad \forall k. \quad (23)$$

Also, with the average transmit power budget, we assume invariant transmit power over different communication rounds and have  $P_k = P_{\text{ave}}, \quad \forall k$ .

For analog transmission, according to [39, Eq. (16)], the per-round delay follows

$$T_A = \frac{dM}{B}, \quad (24)$$

which is a constant. For feasibility, we assume that the target  $T_{\max}$  cannot be smaller than  $T_A$ . The maximum power constraint  $C_2$  is rewritten as

$$\max_{m,k} \left\{ \|\beta_k \mathbf{g}_m^k\|^2 \right\} \leq P_{\max}, \quad (25)$$

for the analog transmission. Unlike the digital transmission, it is impossible to fully utilize the maximum power in analog transmission due to the need for channel pre-equalization. On the other hand, the average power constraint  $\bar{C}_2$  follows

$$\mathbb{E} \left[ \|\beta_k \mathbf{g}_m^k\|^2 \right] \leq P_{\text{ave}}, \quad (26)$$

where the expectation is taken over the wireless channel dynamics and different communication rounds.

### III. PRELIMINARIES

To pave the way for performance analysis, this section provides necessary assumptions and lemmas about the learning algorithms and the transmission paradigms, which will be useful in the next section.

#### A. Assumptions for Learning Algorithms

To begin with, we make several common assumptions on the loss functions, which are widely used in FL studies like [12], [29], [42].

*Assumption 1:* The local loss functions  $F_k(\cdot)$  are  $\mu$ -strongly convex for all devices, that is

$$F_k(\mathbf{w}) \geq F_k(\mathbf{v}) + \nabla F_k(\mathbf{v})^T (\mathbf{w} - \mathbf{v}) + \frac{\mu}{2} \|\mathbf{w} - \mathbf{v}\|^2. \quad (27)$$

*Assumption 2:* The local loss functions  $F_k(\cdot)$  are differentiable and have  $L$ -Lipschitz gradients, which follows

$$\|\nabla F_k(\mathbf{w}) - \nabla F_k(\mathbf{v})\| \leq L \|\mathbf{w} - \mathbf{v}\|, \quad (28)$$

and it is equivalent to

$$F_k(\mathbf{w}) \leq F_k(\mathbf{v}) + \nabla F_k(\mathbf{v})^T (\mathbf{w} - \mathbf{v}) + \frac{L}{2} \|\mathbf{w} - \mathbf{v}\|^2. \quad (29)$$

*Assumption 3:* In most practical applications, it is safe to assume that the sample-wise gradient is always upper bounded by a finite constant  $\gamma$ , i.e.,

$$\|\nabla \mathcal{L}(\mathbf{w}, \mathbf{u})\| \leq \gamma. \quad (30)$$

*Assumption 4:* The distance between the locally optimal model,  $\mathbf{w}_k^*$ , and the globally optimal model,  $\mathbf{w}^*$ , is uniformly bounded by a finite constant  $\delta$ , i.e.,

$$\|\mathbf{w}_k^* - \mathbf{w}^*\| \leq \delta. \quad (31)$$

#### B. Preliminary Lemmas

We present lemmas regarding the strong convexity and Lipschitz smooth properties of the global loss function.

*Lemma 1:* With  $\mu$ -strongly convex and  $L$ -smooth local loss functions, the global loss function  $F(\cdot)$  is also  $\mu$ -strongly convex and  $L$ -smooth.

*Proof:* Recalling the definition of  $F(\cdot)$  in (1), with Assumptions 1-2, it is easily verified that any linear combination of  $\mu$ -strongly convex and  $L$ -smooth local loss functions also satisfies (27) and (29). The proof completes.  $\square$

We then provide the following lemma regarding the imperfection in digital and analog transmission paradigms.

*Lemma 2:* Under the stochastic quantization and the proposed digital aggregation in (11),  $\hat{\mathbf{g}}_{m,D}$  is an unbiased estimate of the actual gradient in (5). For the considered analog paradigm in (16), by choosing  $\lambda = \frac{e^{\gamma_{\text{th}}}}{\rho}$ , the gradient estimate  $\hat{\mathbf{g}}_{m,A}$  is also unbiased.

*Proof:* Please refer to Appendix A.  $\square$

Although both the digital and analog transmissions achieve unbiased gradient estimations, there are fundamental differences in the distortion between the two paradigms. For the digital transmission, the distortion mainly lies in the gradients themselves, i.e., gradient quantization errors. On the other hand, due to the integration of communication and computation in AirComp, the analog transmission additionally suffers from distortion in coefficient aggregation, i.e., computation error, which is due to the CSI imperfection. This essential difference further discriminates the performances of digital and analog transmissions, which are elaborated in the next section.

### IV. COMPARISON WITH CONVERGENCE ANALYSIS

In this section, we analyze the convergence performance under the digital and analog transmissions with the practical constraints for wireless FL. Based on the derived results, we further conduct quantitative comparisons between the two paradigms from various perspectives of view.

$$\mathbb{E}[F(\tilde{\mathbf{w}}_{m+1})] - F(\mathbf{w}^*) \leq \frac{L}{2} (1 - \eta\mu + 2\eta^2 L^2 g_D(\mathbf{r}, b))^{m+1} \mathbb{E}[\|\tilde{\mathbf{w}}_0 - \mathbf{w}^*\|^2] + \frac{\eta(L\phi(b) + 2L^3\delta^2)g_D(\mathbf{r}, b)}{2\mu - 4\eta L^2 g_D(\mathbf{r}, b)}, \quad (32)$$

$$\mathbb{E}[F(\tilde{\mathbf{w}}_{m+1})] - F(\mathbf{w}^*) \leq \frac{L}{2} (1 - \eta\mu + 2\eta^2 L^2 g_A(\mathbf{r}, \gamma_{th}))^{m+1} \mathbb{E}[\|\tilde{\mathbf{w}}_0 - \mathbf{w}^*\|^2] + \frac{\eta(L\varphi(\mathbf{r}, \gamma_{th}) + 2L^3\delta^2 g_A(\mathbf{r}, \gamma_{th}))}{2\mu - 4\eta L^2 g_A(\mathbf{r}, \gamma_{th})}. \quad (33)$$

#### A. Convergence under the Maximum Power Budget

We characterize the convergence performance under different transmission paradigms in the following theorems.

*Theorem 1 (Digital Transmission):* For a fixed learning rate satisfying  $\eta \leq \frac{\mu}{2L^2 g_D(\mathbf{r}, b)}$ , the optimality gap of the distributed gradient update in the  $(m+1)$ -th iteration of the digital transmission is equal to (32) at the top of the next page, where  $\phi(b)$  is a constant defined in Appendix B regarding the quantization errors,  $\mathbf{r} \triangleq [r_1, \dots, r_K]^T$ , and  $g_D(\mathbf{r}, b) \triangleq \sum_{k=1}^K \frac{\alpha_k}{p_k r_k}$ .

*Proof:* Please refer to Appendix B.  $\square$

*Theorem 2 (Analog Transmission):* For a fixed learning rate satisfying  $\eta \leq \frac{\mu}{2L^2 g_A(\mathbf{r}, \gamma_{th})}$ , the optimality gap of the distributed gradient update in the  $(m+1)$ -th iteration of the analog transmission is equal to (33) at the top of the next page, where  $g_A(\mathbf{r}, \gamma_{th}) \triangleq \sum_{k=1}^K \frac{\alpha_k}{r_k} \left( e^{\gamma_{th}} + \frac{(1-\rho^2)E_1(\gamma_{th})e^{2\gamma_{th}}}{2\rho^2} \right) - 1$ ,  $E_1(x) \triangleq \int_x^\infty \frac{e^{-t}}{t} dt$ , and  $\varphi(\mathbf{r}, \gamma_{th}) \triangleq \frac{BN_0\gamma^2 e^{2\gamma_{th}}}{2P_{\max}\rho^2\gamma_{th}} \max_k \left\{ \frac{\alpha_k^2}{r_k^2} d_k^\alpha \right\}$ .

*Proof:* Please refer to Appendix C.  $\square$

From *Theorems 1-2*, we find that the convergence rate mainly depends on the choice of the learning rate  $\eta$ , while the imperfections in transmission also have a certain impact. We conclude the following immediate observations on the convergence rate.

*Remark 1:* As observed in (32) and (33), the convergence performance of an FL algorithm is negatively related to  $g_D(\mathbf{r}, b)$  for digital transmission and to  $g_A(\mathbf{r}, \gamma_{th})$  for analog transmission. We refer to  $g_D(\mathbf{r}, b)$  and  $g_A(\mathbf{r}, \gamma_{th})$  as the *virtual sum weight* for the digital and analog transmissions, respectively, which reflects the degree of hindrance to the convergence imposed by unequal sampling and vulnerable wireless communication. Under the ideal case, with full device participation and no transmission outage, the *virtual sum weight* equals to 1, otherwise it is amplified by the imperfect characteristics. It is interesting to note that, for devices with more data samples, i.e., larger  $\alpha_k$ , the impact of imperfections is exaggerated.

*Remark 2:* Comparing  $g_D(\mathbf{r}, b)$  and  $g_A(\mathbf{r}, \gamma_{th})$ , it can be seen that the vulnerability of digital transmission introduces additional heterogeneity, i.e., varying  $p_k$ , which does not exist in the analog paradigm. This is because outage probability in the digital case is determined by channel conditions and varying across different devices. On the other hand, due to the uniform truncation threshold, all participating devices enjoy the same truncation probability in the analog transmission. Hence, in design of inclusion probabilities  $\mathbf{r}$  for the digital case, we need to adapt the inclusion probabilities to both

dataset size and channel condition. By contrast, in the case of analog transmission, only the heterogeneity of the dataset size needs to be considered.

According to *Theorems 1-2*, we are ready to derive the optimality gap after convergence for further evaluation in the following corollary, which reflects the ultimately achievable performance of the wireless FL.

*Corollary 1:* With sufficient iterations, the optimality gap achieved by digital and analog transmissions, respectively, converges to

$$G_D = \frac{\eta(L\phi(b) + 2L^3\delta^2)g_D(\mathbf{r}, b)}{2\mu - 4\eta L^2 g_D(\mathbf{r}, b)}, \quad (34)$$

$$G_A = \frac{\eta(L\varphi(\mathbf{r}, \gamma_{th}) + 2L^3\delta^2 g_A(\mathbf{r}, \gamma_{th}))}{2\mu - 4\eta L^2 g_A(\mathbf{r}, \gamma_{th})}. \quad (35)$$

*Proof:* Consider the digital transmission scenario with a sufficient number of iterations. We have

$$\begin{aligned} & \lim_{m \rightarrow \infty} \mathbb{E}[F(\tilde{\mathbf{w}}_{m+1})] - F(\mathbf{w}^*) \\ & \leq \lim_{m \rightarrow \infty} \frac{L}{2} (1 - \eta\mu + 2\eta^2 L^2 g_D(\mathbf{r}, b))^{m+1} \mathbb{E}[\|\tilde{\mathbf{w}}_0 - \mathbf{w}^*\|^2] \\ & \quad + \frac{\eta(L\phi(b) + 2L^3\delta^2)g_D(\mathbf{r}, b)}{2\mu - 4\eta L^2 g_D(\mathbf{r}, b)} \\ & \stackrel{(a)}{=} \frac{\eta(L\phi(b) + 2L^3\delta^2)g_D(\mathbf{r}, b)}{2\mu - 4\eta L^2 g_D(\mathbf{r}, b)} = G_D, \end{aligned} \quad (36)$$

where the inequality is obtained through *Theorem 1* and the equality in (a) is due to the fact that  $\eta < \frac{\mu}{2L^2 g_D(\mathbf{r}, b)}$ , i.e.,  $(1 - \eta\mu + 2\eta^2 L^2 g_D(\mathbf{r}, b)) < 1$ . Hence, the achieved optimality gap at convergence is bounded by  $G_D$ . As for the analog transmission, the proof is almost the same and is omitted here for simplicity.  $\square$

From *Corollary 1*, we further compare the two typical paradigms from the following perspectives and conclude insightful remarks that are instructive for the deployment of FL in wireless networks. As a summary, we list main comparison results in Table II. For the sake of simplicity in analysis, without loss of generality, we drop the unbalance of the datasets and assume uniform inclusion probabilities, i.e.,  $\alpha_k = \frac{1}{K}$ , and  $r_k = \frac{N}{K}$ ,  $\forall k$ , which does not cause any essential changes. Also we set that  $T_{\max} = T_A$ . Note that the learning rate is assumed to be sufficiently small and hence the convergence rate remains the same for all cases.

*1) Impact of Transmit Power:* At low SNR levels, the achievable optimality gap under the digital transmission,  $G_D$ , vanishes as  $\mathcal{O}(\exp(\varepsilon/P_{\max}))$  with the maximum transmit power budget  $P_{\max}$ , where  $\varepsilon \triangleq \max_k \left\{ \frac{BN_0\theta}{2Nd_k^\alpha} \right\}$ . At high

TABLE II  
MAIN COMPARISON RESULTS WITH RESPECT TO OPTIMALITY GAP

Paradigms	Transmit power budget, $P$		Device number, $N$	Imperfect CSI, $\rho$
	Low SNR	High SNR		
Digital	$\mathcal{O}(\exp(\frac{\varepsilon}{P})) \searrow$	$\rightarrow G_D^\infty$	$\mathcal{O}(\frac{1}{N} \exp(\varepsilon_1 2^{\varepsilon_2 N}/N)) \nearrow$	$/$
Analog	$\mathcal{O}(\frac{1}{P}) \searrow$	$\rightarrow G_A^\infty$	$\mathcal{O}(\frac{1}{N}) \searrow$	$\mathcal{O}(\frac{1}{\rho^2}) \nearrow$

\* The upward arrow indicates amplification at a certain order, while the downward arrow has the opposite meaning. The horizontal arrow indicates that it ultimately tends towards a fixed value.

SNR regime, i.e.,  $P_{\max} \rightarrow \infty$ , the successful transmission probability  $p_k \rightarrow 1, \forall k$  and  $G_D$  tends to

$$G_D^\infty \triangleq \lim_{P_{\max} \rightarrow \infty} G_D = \frac{\eta(L\phi(b) + 2L^3\delta^2)K}{2\mu N - 4\eta L^2 K}. \quad (37)$$

On the other hand, the decay rate for  $G_A$  is equal to  $\mathcal{O}(1/P_{\max})$  with low SNR values and the high SNR-limiting value is

$$G_A^\infty \triangleq \lim_{P_{\max} \rightarrow \infty} G_A = \frac{2\eta L^3 \delta^2 (Kc - N)}{2\mu N - 4\eta L^2 (Kc - N)}, \quad (38)$$

where  $c \triangleq e^{\gamma_{\text{th}}} + \frac{(1-\rho^2)E_1(\gamma_{\text{th}})e^{2\gamma_{\text{th}}}}{2\rho^2}$ .

*Remark 3:* As SNR increases, the optimality gap for the analog case mainly comes from the non-IID datasets while the impact of the noise asymptotically diminishes. For the digital case, however, quantization errors additionally impose an impact. Under the analog transmission, the negative impact of non-IID datasets is enlarged due to imperfect AirComp. Imperfect CSI results in mismatched channel inversion in AirComp, rendering perfect computation of weighted sum impossible. Moreover, the performance degradation brought by imperfect CSI in the analog transmission cannot be mitigated by occupying more resources. Conversely, in the digital transmission, the convergence performance can be improved by occupying additional resources for increasing the number of quantization bits.

*2) Impact of Device Number:* With the increasing number of participating devices,  $N$ , the virtual sum rate for the analog transmission,  $g_A(\mathbf{r}, \gamma_{\text{th}})$ , decreases at a rate of  $\frac{1}{N}$ , i.e., a faster convergence rate is achieved. As for the optimality gap, the impact of non-IID datasets asymptotically dominates  $G_A$  and the decay rate is equal to  $\mathcal{O}(1/N)$ . Due to the involvement of more devices, a more accurate global gradient is obtained at the PS, which in turn facilitates the FL convergence and leads to better performance. Meanwhile, since different devices involved in the AirComp share the same time-frequency resource, an increase in access devices causes no deterioration of the AirComp performance, fully capturing the performance gain from more participating devices.

On the other hand, for the digital case, convergence performance does not necessarily monotonically change with  $N$ . Although more participating devices do bring performance gains, it also leads to a significant deterioration of the transmission performance considering that limited communication resources are divided among additional users. Thus the convergence is compromised between communication reliability and the computation accuracy for wireless

FL. Specifically, the optimality gap,  $G_D$ , enlarges with a rate of  $\mathcal{O}(\frac{1}{N} \exp(\varepsilon_1 2^{\varepsilon_2 N}/N))$  with sufficiently large  $N$ , where  $\varepsilon_1 = \frac{BN_0}{2Pd_K^{-\alpha}}$  and  $\varepsilon_2 = \frac{b+1}{M}$ .

*Remark 4:* Benefiting from the characteristics of AirComp, more participating devices in the analog transmission always lead to performance improvement regardless of other parameters. Hence, allowing all active devices to participate in the FL training is the best choice for analog transmission. By contrast, in the digital transmission, it is necessary to seek a balance between the transmission performance and diversity gain through an optimization of  $N$ .

*3) Impact of Imperfect CSI:* The imperfect CSI at the transmitter only affects the performance of analog transmission, which deteriorates at the order of  $\frac{1}{\rho^2}$ . Due to imperfect CSI, the aggregation computation and the truncation decision in AirComp are contaminated, thus leading to a mismatch in the model aggregation and the impact of noise amplification.

*Remark 5:* After incorporating computation capabilities into the analog case, the emergence of computation error as a new source of error has positioned computational accuracy as a crucial factor affecting the convergence performance. It is concluded that CSI is a key factor affecting the performance gain brought by AirComp. Moreover, the truncation threshold  $\gamma_{\text{th}}$  should be optimized to adapt different levels of channel estimation accuracy. It can be effectively solved via bisection search in [34].

*4) Impact of the Number of Quantization Bits:* In the digital transmission, the number of quantization bits,  $b$ , also influences the FL performance in the following implicit ways. By selecting the minimum feasible  $\theta = 2^{\frac{Nd(b+1)}{BT_{\max}}} - 1$  in (22), the achievable optimality gap  $G_D$  is rewritten as

$$\begin{aligned} G_D &\approx \frac{\eta}{2\mu} \left( \frac{L\Delta^2}{(2^b - 1)^2} + 2L^3\delta^2 \right) g_D(\mathbf{r}, b) \\ &= \frac{\eta}{2\mu} \left( \frac{L\Delta^2}{(2^b - 1)^2} + 2L^3\delta^2 \right) \\ &\quad \times \left( \sum_{k=1}^K \frac{\alpha_k}{r_k} \exp \left( \frac{BN_0 \left( 2^{\frac{Nd(b+1)}{BT_{\max}}} - 1 \right)}{2NP_k d_k^{-\alpha}} \right) \right), \quad (39) \end{aligned}$$

where the approximation is obtained in region of  $\eta \ll \frac{\mu}{2L^2 g_D(\mathbf{r}, b)}$ . It is found that as  $b$  increases,  $G_D$  tends to first decrease and then increase. This is due to the diminishing quantization error term  $\phi(b)$  with an increasing quantization accuracy and finally  $G_D$  is dominated by the impact of packet loss. Therefore, it is necessary to optimize of the integer



variable  $b$  to pursue better convergence performance, which can be solved by a low-complexity exhaustive search method.

### B. Convergence Analysis under the Average Power Budget

We consider the convergence with the average transmit power budget. For the digital transmission, by replacing  $P_{\max}$  with  $P_{\text{ave}}$ , we derive the similar results as *Theorem 1* and is omitted here due to page limit. As for the analog transmission, we have the following corollary.

*Corollary 2:* For a fixed learning rate satisfying  $\eta \leq \frac{\mu}{2L^2 g_A(\mathbf{r}; \gamma_{\text{th}})}$ , the optimality gap of the distributed gradient update in the  $(m+1)$ -th iteration under the analog transmission follows

$$\begin{aligned} \mathbb{E}[F(\tilde{\mathbf{w}}_{m+1})] - F(\mathbf{w}^*) &\leq \frac{L}{2} (1 - \eta\mu + 2\eta^2 L^2 g_A(\mathbf{r}, \gamma_{\text{th}}))^{m+1} \mathbb{E}[\|\tilde{\mathbf{w}}_0 - \mathbf{w}^*\|^2] \\ &\quad + \frac{\eta (L\varphi_{\text{ave}}(\mathbf{r}, \gamma_{\text{th}}) + 2L^3 \delta^2 g_A(\mathbf{r}, \gamma_{\text{th}}))}{2\mu - 4\eta L^2 g_A(\mathbf{r}, \gamma_{\text{th}})}, \end{aligned} \quad (40)$$

where  $\varphi_{\text{ave}}(\mathbf{r}, \gamma_{\text{th}}) \triangleq \frac{BN_0 \gamma_{\text{th}}^2 e^{2\gamma_{\text{th}}} E_1(\gamma_{\text{th}})}{2P_{\text{ave}} \rho^2} \max_k \left\{ \frac{\alpha_k^2}{r_k^2} d_k^\alpha \right\}$ . The optimality gap with sufficient iterations follows

$$G_{A,\text{ave}} = \frac{\eta (L\varphi_{\text{ave}}(\mathbf{r}, \gamma_{\text{th}}) + 2L^3 \delta^2 g_A(\mathbf{r}, \gamma_{\text{th}}))}{2\mu - 4\eta L^2 g_A(\mathbf{r}, \gamma_{\text{th}})}. \quad (41)$$

*Proof:* please refer to Appendix D.  $\square$

*Remark 6:* It is worth noting that  $E_1(\gamma_{\text{th}}) < \frac{1}{\gamma_{\text{th}}}$  when  $\gamma_{\text{th}} > 0$ . Compared with the maximum transmit power budget, a smaller optimality gap for the analog transmission is achieved with the average power budget. Due to the need for channel alignment in AirComp, the performance is dominantly limited by the device with the worst channel condition. Furthermore, the strict peak power constraint amplifies the impact of worst-case channel conditions, resulting in looser convergence performance compared to the long-term constraint.

To summarize, while the analog AirComp improves the spectrum utilization compared to the digital paradigm, it faces challenges in fully utilizing the power resource, particularly with strict peak power constraints. Conversely, orthogonal access in digital transmission is not suitable for scenarios with massive access due to the limitations in spectrum resources.

### C. Discussions on Scenarios with Advanced System Designs

To facilitate performance analysis, we introduce assumptions regarding the system design, including multiple access, parameter quantization, and power control methods. Subsequently, we delve into the implications of advanced system designs on the FL performance and comparison.

In the digital transmission, the FL performance can primarily be improved from two aspects, namely enhancing transmission reliability and optimizing resource utilization. Specifically, advanced transmissions strategies help minimize transmission errors and packet losses due to channel fading. Furthermore, if other resource allocation methods, such as the model compression design and device scheduling strategies,

are exploited toward the FL tasks, they prioritize crucial parameter/device transmissions and thus lifting the resource utilization. On the other hand, in the analog transmission, the FL performance through AirComp is primarily influenced by the over-the-air computational accuracy. Optimized transceiver and power control designs help mitigate the negative impact of channel fading on the FL performance.

While further optimization of system designs enhances performance, it is essential to note that the performance limits for the digital and analog transmissions remains unchanged. As observed in the above analytical results, in the digital transmission paradigm, due to the decoupling of the communication and computation processes, the number of bits that can be accurately transmitted with the limited resources is determined, which places an upper bound of the FL performance. In contrast, within the analog transmissions, the receiver does not aim to recover information from individual sources but instead prioritizes the precision of computation results derived from the over-the-air superimposed signals, thereby making computational accuracy a decisive role. Hence, the performance limit of the analog transmission is contingent upon the channel estimation accuracy and additive noise level.

## V. DEVICE SAMPLING OPTIMIZATION

Based on the derived results in Section IV, we are able to further establish an optimization design of the device sampling for the wireless FL to improve the convergence.

### A. Digital Transmission

By direct inspection of (34), the optimality gap  $G_D$  monotonically decreases with a decreasing virtual sum weight. Hence, the device sampling optimization problem with the digital transmission is formulated as

$$\begin{aligned} \underset{\mathbf{r}}{\text{minimize}} \quad & g_D(\mathbf{r}, b) = \sum_{k=1}^K \frac{\alpha_k}{p_k r_k} \\ \text{subject to} \quad & \sum_{k=1}^K r_k = N, \quad r_k \leq 1, \quad k = 1, 2, \dots, K, \end{aligned} \quad (42)$$

which is a convex problem. By exploiting the KKT conditions, we obtain the optimal inclusion probability as

$$r_k^* = \min \left\{ \sqrt{\frac{\alpha_k}{\nu p_k}}, 1 \right\}, \quad (43)$$

where  $\nu$  is the Lagrangian multiplier and it is selected to satisfy  $\sum_{k=1}^K r_k^* = N$ . Note that the value of  $\sum_{k=1}^K r_k^*$  varies monotonically with  $\nu$  and thus we can rely on a bisection-based search method [13] to get the optimal solution of problem (42).

*Remark 7:* The optimal inclusion probability is positively correlated with the local dataset size while it behaves conversely correlated with the successful transmission probability. In other words, a device with a larger dataset is deemed more important for model training, thereby deserving a sampling bias. Conversely, devices with lower successful transmission probabilities contribute less to the model training process,

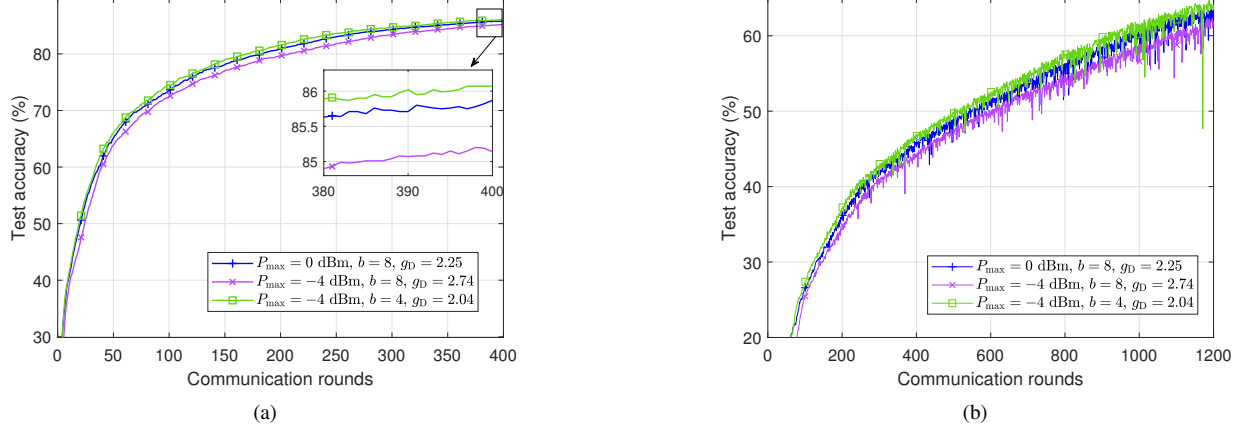


Fig. 2. Convergence performance under digital transmission: (a) MNIST dataset, (b) CIFAR-10 dataset.

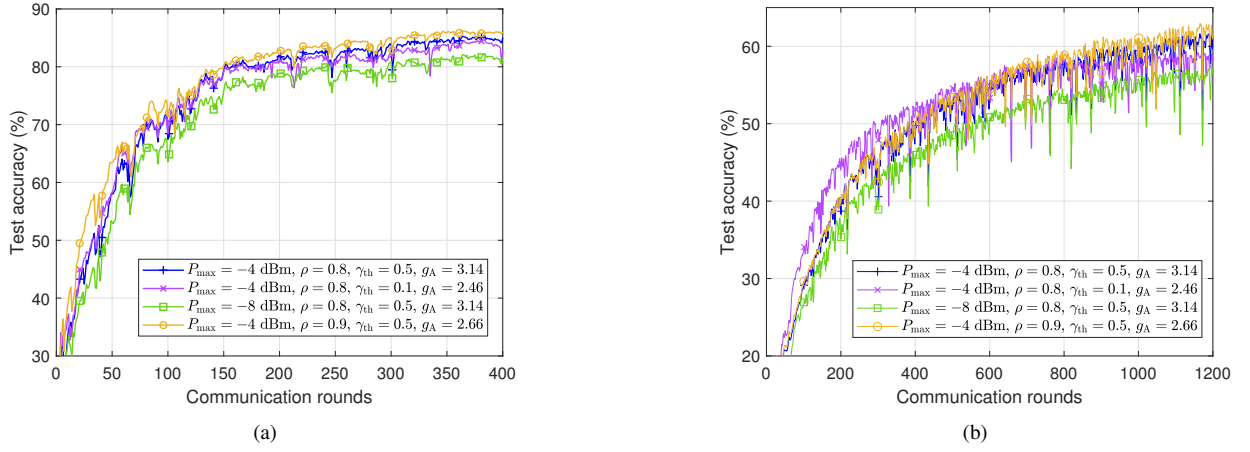


Fig. 3. Convergence performance under analog transmission: (a) MNIST dataset, (b) CIFAR-10 dataset.

requiring more frequent sampling to compensate. Thus, the goal of our inclusion probability optimization is to address the imbalances in the dataset size, and the heterogeneity introduced by uneven channel fading. It ensures fair and effective participation among diverse devices.

Moreover, note that the influence of quantization error and data heterogeneity are equally amplified by  $g_D(\mathbf{r}, b)$ . It indicates that the optimization of inclusion probabilities  $\mathbf{r}$  cannot adequately adapt to varying local data distributions.

### B. Analog Transmission

As for the analog transmission, the device sampling optimization is expressed as

$$\begin{aligned} & \underset{\mathbf{r}}{\text{minimize}} && \frac{\varphi(\mathbf{r}, \gamma_{\text{th}}) + 2L^2\delta^2 g_A(\mathbf{r}, \gamma_{\text{th}})}{2\mu - 4\eta L^2 g_A(\mathbf{r}, \gamma_{\text{th}})} \\ & \text{subject to} && \sum_{k=1}^K r_k = N, \quad r_k \leq 1, \quad k = 1, 2, \dots, K. \end{aligned} \quad (44)$$

Note that under the average transmit power budget, (41) only differs from the objective value in the constant term, and hence we will not discuss it separately. Considering the intractable fractional form of the objective function in (44), we rely on the well-known Dinkelbach algorithm for reformulation [43],

[44]. According to the definition of  $\varphi(\mathbf{r}, \gamma_{\text{th}})$  and  $g_A(\mathbf{r}, \gamma_{\text{th}})$  in (35), it is easy to check that the denominator of the objective function in (44) is concave and the numerator is convex. Hence, the iterative Dinkelbach algorithm guarantees to converge to the global optimum of (44). Concretely, in the  $t$ -th iteration, we reformulate the problem in (44) as

$$\begin{aligned} & \underset{\mathbf{r}}{\text{minimize}} && \varphi(\mathbf{r}, \gamma_{\text{th}}) + (2L^2\delta^2 + 4\eta L^2\zeta^{(t-1)})g_A(\mathbf{r}, \gamma_{\text{th}}) \\ & \text{subject to} && \sum_{k=1}^K r_k = N, \quad r_k \leq 1, \quad k = 1, 2, \dots, K. \end{aligned} \quad (45)$$

where  $\zeta^{(t-1)}$  is a constant determined in the previous round. Note that the problem in (45) is convex and thus can be solved by numerical convex program solvers, e.g., CVX tools [45]. After obtaining the optimal  $\mathbf{r}^{(t)}$  of the  $t$ -th subproblem in (45), the auxiliary constant is updated as

$$\zeta^{(t)} = \frac{\varphi(\mathbf{r}^{(t)}, \gamma_{\text{th}}) + 2L^2\delta^2 g_A(\mathbf{r}^{(t)}, \gamma_{\text{th}})}{2\mu - 4\eta L^2 g_A(\mathbf{r}^{(t)}, \gamma_{\text{th}})}. \quad (46)$$

Iterating the above steps until convergence, we obtain the optimal  $\mathbf{r}$  of the problem in (44).

*Remark 8:* Unlike the digital transmission case, the device sampling optimization is committed to seeking a trade-off between the equivalent noise power  $\varphi(\mathbf{r}, \gamma_{\text{th}})$  and virtual sum

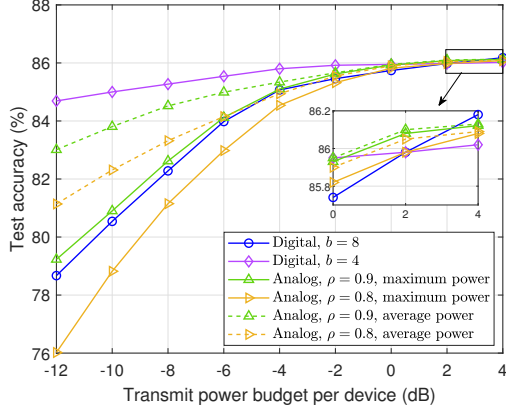


Fig. 4. Test accuracy versus transmit power budget.

weight  $g_A(\mathbf{r}, \gamma_{th})$ , and the parameter  $\delta$  functions as a weighting factor to facilitate the optimal trade-off. At high SNR regimes or with extremely uneven local data distributions, the noise term is comparably ignorable and hence the optimality gap is dominated by  $g_A$ . Hence, the optimization of  $\mathbf{r}$  is isolated from specific channel conditions and only needs to match the size of local datasets.

## VI. NUMERICAL RESULTS

In this section, we provide simulation results to verify the performance analysis and the inclusion probability optimization. We deploy  $K = 20$  edge devices uniformly distributed in a square area with radius 500 m and a PS at the center of the square area. The most popular MNIST dataset and CIFAR-10 dataset are exploited for the FL performance evaluation. The MNIST dataset contains 10 classes of handwritten digits ranging from 0 to 9 and we train a multi-layer perceptron (MLP) with  $d = 23,860$  parameters via the wireless FL algorithm for classification purposes. Moreover, the CIFAR-10 dataset includes 10 classes with labels 0-9 and we train a convolutional neural network (CNN) with  $d = 60,000$  parameters. The trained CNN contains two convolutional layers and three fully connected layers. Max pooling operation is conducted following each convolutional layer and the activation function is ReLU. Different edge devices own different data samples, and each local dataset has up to two types of data samples to capture the non-IID characteristic.

Unless otherwise specified, the other parameters are set as: the number of participating devices  $N = 10$ , the bandwidth,  $B = 1$  MHz, the path loss exponent,  $\alpha = 3$ , the noise power  $N_0 = -80$  dBm/Hz, the maximum transmit power budget,  $P_{max} = 0$  dB, the number of quantization bits,  $b = 8$ , the truncation threshold,  $\gamma_{th} = 0.5$ , the delay target  $T_{max}$  is equal to  $T_A$  in (24), and the learning rate  $\eta = 0.01$ . We set  $L = 8$  and  $\mu = 2$ , which fall within the existing typical range of values in [46], [47]. Additionally, the parameter  $\delta$ , serving as an upper bound of  $\|\mathbf{w}_k^* - \mathbf{w}^*\|^2$ , is estimated through simulation tests.

### A. Convergence Performance

In Figs. 2 and 3, we depict the convergence performance for the digital and analog transmission. As shown in Fig.

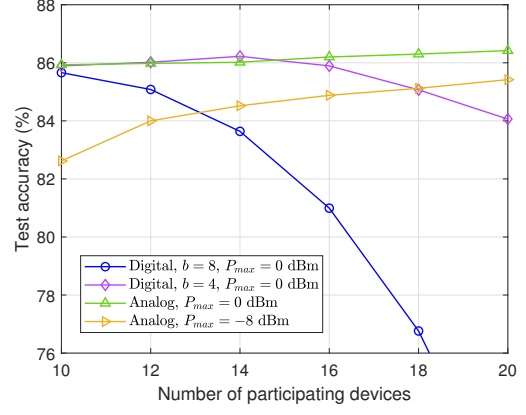


Fig. 5. Test accuracy versus the number of participating devices.

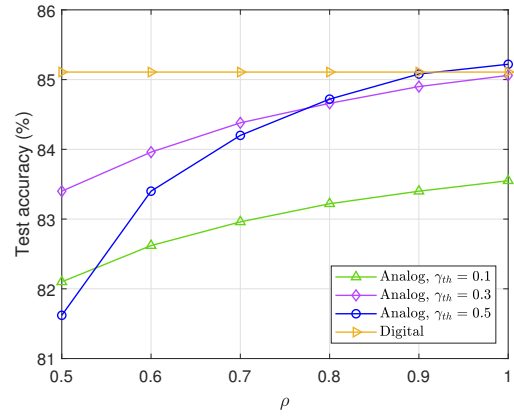


Fig. 6. Test accuracy versus the accuracy level of channel estimation accuracy.

2, we observe that the convergence rate and optimality gap under digital transmission exhibit a negative correlation with the virtual sum weight, aligning with our theoretical analysis. Moreover, the convergence behavior remains consistent with the analytical results despite the complexity of the classification task, thereby validating the accuracy of the theoretical analysis.

For the analog case depicted in Fig. 3, consistent with the analytical findings, we notice that the convergence rate is negatively correlated with the virtual sum weight  $g_A$ , which is determined by  $\rho$  and  $\gamma_{th}$ . On the other hand, transmit power only affects the achievable optimality gap after convergence. This is because changes in transmit power only affect the equivalent power of the additive noise. Additionally, modifications in  $\rho$  and  $\gamma_{th}$  affect the distortion of the aggregation coefficient, which in turn influences the computation error. Furthermore, the increased complexity of FL tasks renders fluctuations in the performance curve more sensitive to noise. Consequently, in the analog communication, the superimposed white Gaussian noise is significantly severer than quantization errors observed in the digital transmission, thus leading to more pronounced fluctuations in convergence performance. It implies that for more complex learning tasks, it becomes imperative to further reduce the variance of gradient estimation to mitigate excessive fluctuations and their adverse impacts on convergence.

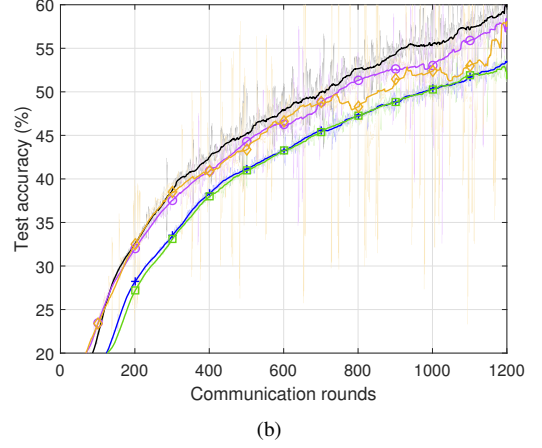
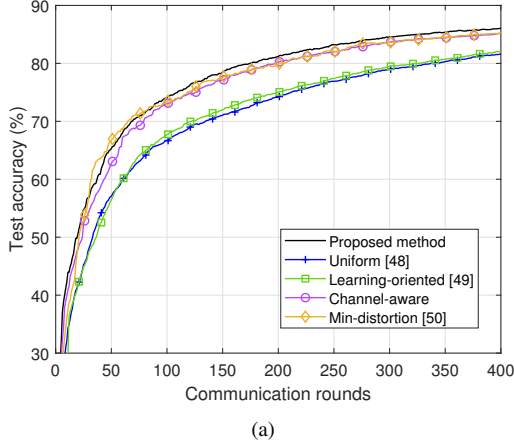


Fig. 7. Convergence performance with different inclusion probabilities and digital transmission: (a) MNIST dataset, (b) CIFAR-10 dataset.

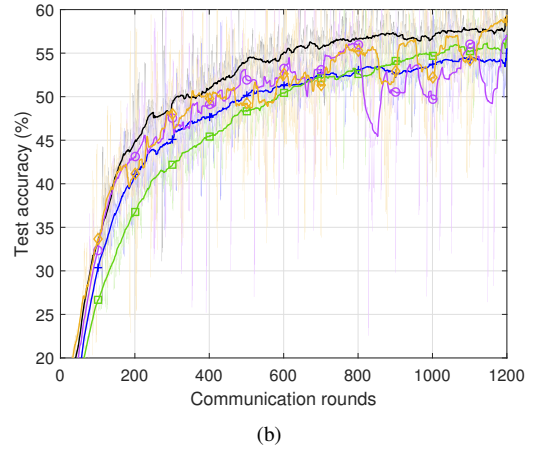
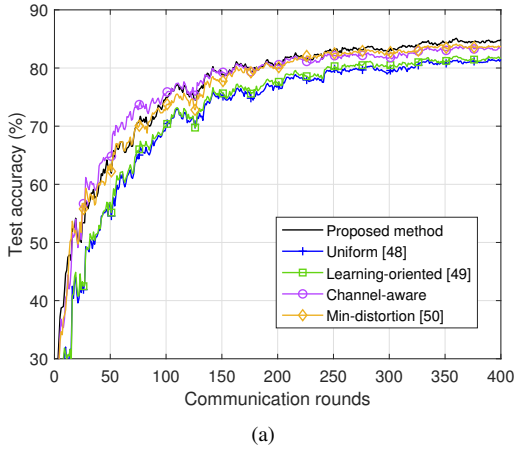


Fig. 8. Convergence performance with different inclusion probabilities and analog transmission: (a) MNIST dataset, (b) CIFAR-10 dataset.

### B. Impact of Transmit Power Budget

In Fig. 4, we show the test accuracy versus different transmit power budgets. It is observed that the digital transmission scheme outperforms the analog scheme, particularly with high SNR levels. In such cases, employing more quantization bits yields the best performance. Conversely, for low SNR levels, reducing the quantization bits leads to marginal performance loss, highlighting the flexibility of the digital schemes by selecting different quantization accuracies. On the other hand, the analog scheme faces significant performance limitations, particularly with the maximum transmit power budget and less CSI, due to the stringent requirements of channel inversion. Therefore, in terms of power utilization, the digital scheme is more efficient than the analog counterpart.

### C. Impact of Participating Device Numbers

Fig. 5 illustrates the test accuracy versus the number of participating devices. We note that for the analog transmission, the test accuracy gradually increases as  $N$  increases. In contrast, although the performance in digital case may be improved initially, it eventually decline rapidly as each device can only occupy a limited amount of resources, making it unable to support high-rate transmission. Consequently, the results suggest that for digital transmission, the selection

of  $N$  requires further optimization according to the actual conditions, with a preference for fewer devices.

### D. Impact of Channel Estimation Accuracy

In Fig. 6, we present the impact of channel estimation accuracy on the analog case. It is evident that better performance can be achieved with more accurate CSI. Additionally, we observe that smaller truncation thresholds are more suitable for larger  $\rho$ , while larger truncation thresholds are preferred for smaller  $\rho$ . This is because higher CSI uncertainties have a significant impact on truncation choices, necessitating looser truncation conditions to reduce incorrect choices.

### E. Impact of Different Inclusion Probabilities

In Figs. 7 and 8, we depict the convergence performance with different inclusion probabilities. For comparison, we consider the following baselines for comparison. For the sake of fairness, all schemes refrain from utilizing specific information on instantaneous CSI and gradients.

- Uniform [48]: The inclusion probabilities are uniformly assigned the same value, i.e.,  $p_k = \frac{1}{K}$ .
- Learning-oriented [49]: From the perspective of learning algorithms, the probability is set to be proportional to the size of the local datasets, i.e.,  $p_k \propto \alpha_k$ .

- Channel-aware: From the perspective of wireless channels, the probability is set to be proportional to the large-scale path loss, i.e.,  $p_k \propto d_k^{-\frac{\alpha}{2}}$ .
- Min-distortion [50]: To minimize the communication distortion in the analog transmission, the probability is set to be proportional to  $\alpha_k d_k^{\frac{\alpha}{2}}$  by considering both the local datasets and channel conditions.

As shown in Fig. 7, the proposed method consistently outperforms the aforementioned baseline methods across all levels. The first two baselines neglect the influence of the wireless transmission process, resulting in performance degradation. The sampling method based on channel conditions tends to select devices with better channels, effectively reducing packet loss rates and yielding significant performance improvements. However, due to its oversight of imbalanced size of local datasets, its final performance remains inferior to our proposed method. The fourth baseline, tailored for the analog transmission scenarios, partially accounts for the impact of local datasets and wireless channels but lacks optimality, leading to limited performance gains.

As for the analog transmission case in Fig. 8, we note that although the performance of the optimized probability is superior, the performance gain compared to the other baselines is not significant. This limit arises from the reliance on constants  $L$ ,  $\mu$ , and  $\delta$  in the optimization problem (44), which are challenging to determine accurately in practice, thus affecting the final performance. Similarly, akin to the digital transmission, the sampling method based on channel conditions effectively mitigates the negative impact of the imperfect wireless transmission. However, its disregard for data characteristics results in suboptimal performance, particularly in the complex classification tasks on CIFAR-10 dataset, leading to significant performance fluctuations. Furthermore, the baseline method of minimizing computational distortion overlooks the impact of data heterogeneity, thus impeding its ability to achieve satisfactory performance.

## VII. CONCLUSION

In this paper, we have provided a detailed comparison between digital and analog transmission enabled wireless FL. To this end, we considered general transmission designs for both schemes and conducted a fair comparison between them. Then, we analyzed the convergence behavior of wireless FL in terms of the convergence rate and optimality gap under digital and analog cases, and compared the convergence performance from multiple perspectives. It was found that digital transmission is more suitable for scenarios with sufficient radio resources and CSI uncertainties. On the other hand, analog transmission is suitable when there are massive numbers of participating devices. Next, we addressed sampling optimization for both cases, and further developed insights for optimization, which are useful for practical deployment. Finally, experimental results illuminated the analytical results and the sampling strategies. Additionally, an explicit and precise characterization of data heterogeneity and targeted system designs with theoretical guarantees should be of our interest in the future work.

## APPENDIX A PROOF OF LEMMA 2

For the digital case, according to [19, Lemma 5], we first conclude that the quantized gradients  $\mathcal{Q}(\mathbf{g}_m^k)$  is unbiased, i.e.,

$$\mathbb{E}[\mathcal{Q}(\mathbf{g}_m^k)] = \mathbf{g}_m^k. \quad (47)$$

Combining with the fact that  $\mathbb{E}[\xi_{k,D}] = 1$  in (12), we have

$$\begin{aligned} \mathbb{E}[\hat{\mathbf{g}}_{m,D}] &\stackrel{(a)}{=} \sum_{k=1}^K \alpha_k \mathbb{E}\left[\frac{\chi_k}{r_k}\right] \mathbb{E}[\xi_{k,D}] \mathbb{E}[\mathcal{Q}(\mathbf{g}_m^k)] \\ &= \sum_{k=1}^K \alpha_k \mathbf{g}_m^k = \mathbf{g}_m, \end{aligned} \quad (48)$$

where (a) comes from the definition of  $\hat{\mathbf{g}}_{m,D}$  and the independence among device sampling, small-scale fading and stochastic quantization.

As for the analog transmission, by exploiting [34, Lemma 1], we have  $\mathbb{E}[\xi_{k,A}] = 1$ . Combining with the statistical characteristic of  $\chi_k$  and  $\bar{\mathbf{z}}_m$  and following the same procedures in (48), we get the desired conclusion, i.e.,  $\mathbb{E}[\hat{\mathbf{g}}_{m,A}] = \mathbf{g}_m$ . The proof completes.

## APPENDIX B PROOF OF THEOREM 1

To begin with, we define an auxiliary variable as

$$\hat{\mathbf{w}}_{m+1} = \tilde{\mathbf{w}}_m - \eta \mathbf{g}_m, \quad (49)$$

which represents the model obtained at  $(m+1)$ -th round via ideal communication and full participation. Then, by exploiting Assumption 2 and the fact that  $\nabla F(\mathbf{w}^*) = 0$ , we have

$$\begin{aligned} \mathbb{E}[F(\tilde{\mathbf{w}}_{m+1})] - F(\mathbf{w}^*) &\leq \frac{L}{2} \mathbb{E}[\|\tilde{\mathbf{w}}_{m+1} - \mathbf{w}^*\|^2] \\ &\stackrel{(a)}{=} \frac{L}{2} \left( \underbrace{\mathbb{E}[\|\tilde{\mathbf{w}}_{m+1} - \hat{\mathbf{w}}_{m+1}\|^2]}_{A_1} + \underbrace{\mathbb{E}[\|\hat{\mathbf{w}}_{m+1} - \mathbf{w}^*\|^2]}_{A_2} \right), \end{aligned} \quad (50)$$

where (a) is due to the fact that  $\hat{\mathbf{g}}_{m,D}$  is an unbiased estimate of  $\mathbf{g}_m$ . For the term  $A_1$ , it is bounded by

$$\begin{aligned} A_1 &= \eta^2 \mathbb{E}[\|\hat{\mathbf{g}}_{m,D} - \mathbf{g}_m\|^2] \\ &= \eta^2 \mathbb{E}\left[\left\|\sum_{k=1}^K \frac{\chi_k \alpha_k \xi_{k,D}}{r_k} \mathcal{Q}(\mathbf{g}_m^k) - \sum_{k=1}^K \alpha_k \mathbf{g}_m^k\right\|^2\right] \\ &\stackrel{(a)}{=} \eta^2 \mathbb{E}\left[\left\|\sum_{k=1}^K \alpha_k \left(\frac{\chi_k \xi_{k,D}}{r_k} \mathcal{Q}(\mathbf{g}_m^k) - \sum_{i=1}^K \alpha_i \mathbf{g}_m^i\right)\right\|^2\right] \\ &\stackrel{(b)}{\leq} \eta^2 \sum_{k=1}^K \alpha_k \mathbb{E}\left[\left\|\frac{\chi_k \xi_{k,D}}{r_k} \mathcal{Q}(\mathbf{g}_m^k) - \sum_{i=1}^K \alpha_i \mathbf{g}_m^i\right\|^2\right] \\ &= \eta^2 \sum_{k=1}^K \alpha_k \mathbb{E}\left[\left\|\left(\frac{\chi_k \xi_{k,D}}{r_k} \mathcal{Q}(\mathbf{g}_m^k) - \mathbf{g}_m^k\right)\right\|^2\right] \end{aligned}$$

$$\begin{aligned}
& + \left( \mathbf{g}_m^k - \sum_{i=1}^K \alpha_i \mathbf{g}_m^i \right) \Big\| \Big\|^2 \Big] \\
& \stackrel{(c)}{=} \underbrace{\eta^2 \sum_{k=1}^K \alpha_k \mathbb{E} \left[ \left\| \frac{\chi_k \xi_{k,D}}{r_k} \mathcal{Q}(\mathbf{g}_m^k) - \mathbf{g}_m^k \right\|^2 \right]}_{B_1} \\
& + \underbrace{\eta^2 \sum_{k=1}^K \alpha_k \mathbb{E} \left[ \left\| \mathbf{g}_m^k - \sum_{i=1}^K \alpha_i \mathbf{g}_m^i \right\|^2 \right]}_{B_2}, \quad (51)
\end{aligned}$$

where (a) is because  $\sum_{k=1}^K \alpha_k = 1$ , (b) exploits the convexity of  $\|\cdot\|^2$ , and (c) is due to the fact that  $\mathbb{E} \left[ \frac{\chi_k \xi_{k,D}}{r_k} \mathcal{Q}(\mathbf{g}_m^k) \right] = \mathbf{g}_m^k$ . According to [20], the variance of quantization error is bounded as

$$\begin{aligned}
\mathbb{E} \left[ \left\| \mathcal{Q}(\mathbf{g}_m^k) - \mathbf{g}_m^k \right\|^2 \right] & \leq \frac{d}{4} \left( \frac{g_{m,\max}^k - g_{m,\min}^k}{2^b - 1} \right)^2 \\
& \leq \frac{\Delta^2}{(2^b - 1)^2} \triangleq \phi(b) \quad (52)
\end{aligned}$$

where  $\Delta^2$  is defined as a uniform upper bound of  $\frac{d}{4} (g_{m,\max}^k - g_{m,\min}^k)^2$ ,  $\forall m, k$ . Then,  $B_1$  is bounded by

$$\begin{aligned}
B_1 & = \sum_{k=1}^K \alpha_k \mathbb{E} \left[ \left\| \left( \frac{\chi_k \xi_{k,D}}{r_k} \mathcal{Q}(\mathbf{g}_m^k) - \frac{\chi_k \xi_{k,D}}{r_k} \mathbf{g}_m^k \right) \right. \right. \\
& \quad \left. \left. + \left( \frac{\chi_k \xi_{k,D}}{r_k} \mathbf{g}_m^k - \mathbf{g}_m^k \right) \right\|^2 \right] \\
& = \sum_{k=1}^K \alpha_k \mathbb{E} \left[ \left( \frac{\chi_k \xi_{k,D}}{r_k} \right)^2 \right] \mathbb{E} \left[ \left\| \mathcal{Q}(\mathbf{g}_m^k) - \mathbf{g}_m^k \right\|^2 \right] \\
& \quad + \sum_{k=1}^K \alpha_k \mathbb{E} \left[ \left( \frac{\chi_k \xi_{k,D}}{r_k} - 1 \right)^2 \right] \mathbb{E} \left[ \left\| \mathbf{g}_m^k \right\|^2 \right] \\
& \stackrel{(a)}{\leq} \sum_{k=1}^K \frac{\phi(b) \alpha_k}{p_k r_k} + \sum_{k=1}^K \alpha_k \left( \frac{1}{p_k r_k} - 1 \right) \mathbb{E} \left[ \left\| \nabla F_k(\tilde{\mathbf{w}}_m) \right\|^2 \right], \quad (53)
\end{aligned}$$

where (a) uses  $\mathbb{E} \left[ \left( \frac{\chi_k \xi_{k,D}}{r_k} \right)^2 \right] = \frac{1}{p_k r_k}$  and  $\mathbb{E} \left[ \left( \frac{\chi_k \xi_{k,D}}{r_k} - 1 \right)^2 \right] = \frac{1}{p_k r_k} - 1$ . Next, by expanding the square term, we reformulate  $B_2$  as

$$\begin{aligned}
B_2 & = \sum_{k=1}^K \alpha_k \mathbb{E} \left[ \left\| \nabla F_k(\tilde{\mathbf{w}}_m) - \sum_{i=1}^K \alpha_i \nabla F_i(\tilde{\mathbf{w}}_m) \right\|^2 \right] \\
& = \sum_{k=1}^K \alpha_k \left( \mathbb{E} \left[ \left\| \nabla F_k(\tilde{\mathbf{w}}_m) \right\|^2 \right] + \mathbb{E} \left[ \left\| \sum_{i=1}^K \alpha_i \nabla F_i(\tilde{\mathbf{w}}_m) \right\|^2 \right] \right. \\
& \quad \left. - 2 \mathbb{E} \left[ \nabla F_k(\tilde{\mathbf{w}}_m)^T \left( \sum_{i=1}^K \alpha_i \nabla F_i(\tilde{\mathbf{w}}_m) \right) \right] \right) \\
& = \sum_{k=1}^K \alpha_k \mathbb{E} \left[ \left\| \nabla F_k(\tilde{\mathbf{w}}_m) \right\|^2 \right] - \mathbb{E} \left[ \left\| \sum_{i=1}^K \alpha_i \nabla F_i(\tilde{\mathbf{w}}_m) \right\|^2 \right]
\end{aligned}$$

$$= \sum_{k=1}^K \alpha_k \mathbb{E} \left[ \left\| \nabla F_k(\tilde{\mathbf{w}}_m) \right\|^2 \right] - \mathbb{E} \left[ \left\| \nabla F(\tilde{\mathbf{w}}_m) \right\|^2 \right]. \quad (54)$$

Then for  $A_2$ , we have

$$\begin{aligned}
A_2 & = \mathbb{E} \left[ \left\| \tilde{\mathbf{w}}_m - \mathbf{w}^* - \eta \nabla F(\tilde{\mathbf{w}}_m) \right\|^2 \right] \\
& = \mathbb{E} \left[ \left\| \tilde{\mathbf{w}}_m - \mathbf{w}^* \right\|^2 \right] - 2\eta \mathbb{E} \left[ (\tilde{\mathbf{w}}_m - \mathbf{w}^*)^T \nabla F(\tilde{\mathbf{w}}_m) \right] \\
& \quad + \eta^2 \mathbb{E} \left[ \left\| \nabla F(\tilde{\mathbf{w}}_m) \right\|^2 \right] \\
& \stackrel{(a)}{\leq} (1 - \eta\mu) \mathbb{E} \left[ \left\| \tilde{\mathbf{w}}_m - \mathbf{w}^* \right\|^2 \right] + 2\eta \mathbb{E} \left[ F(\mathbf{w}^*) - F(\tilde{\mathbf{w}}) \right] \\
& \quad + \eta^2 \mathbb{E} \left[ \left\| \nabla F(\tilde{\mathbf{w}}_m) \right\|^2 \right] \\
& \stackrel{(b)}{\leq} (1 - \eta\mu) \mathbb{E} \left[ \left\| \tilde{\mathbf{w}}_m - \mathbf{w}^* \right\|^2 \right] + \eta^2 \mathbb{E} \left[ \left\| \nabla F(\tilde{\mathbf{w}}_m) \right\|^2 \right], \quad (55)
\end{aligned}$$

where the inequality in (a) is due to *Assumption 1*, and (b) is due to the fact that  $F(\mathbf{w}^*) - F(\mathbf{w}) \leq 0$  for  $\forall \mathbf{w} \in \mathbb{R}^d$ .

Combining all the results in (51)-(55), it yields

$$\begin{aligned}
& \mathbb{E} \left[ \left\| \tilde{\mathbf{w}}_{m+1} - \mathbf{w}^* \right\|^2 \right] \\
& \leq (1 - \eta\mu) \mathbb{E} \left[ \left\| \tilde{\mathbf{w}}_m - \mathbf{w}^* \right\|^2 \right] \\
& \quad + \sum_{k=1}^K \frac{\eta^2 \alpha_k}{p_k r_k} \mathbb{E} \left[ \left\| \nabla F_k(\tilde{\mathbf{w}}_m) \right\|^2 \right] + \sum_{k=1}^K \frac{\eta^2 \alpha_k \phi(b)}{p_k r_k}. \quad (56)
\end{aligned}$$

We further rewrite the second term in the right hand side (RHS) of (56) as

$$\begin{aligned}
& \mathbb{E} \left[ \left\| \nabla F_k(\tilde{\mathbf{w}}_m) \right\|^2 \right] \\
& \stackrel{(a)}{=} \mathbb{E} \left[ \left\| \nabla F_k(\tilde{\mathbf{w}}_m) - \nabla F_k(\mathbf{w}_k^*) \right\|^2 \right] \\
& \stackrel{(b)}{\leq} L^2 \mathbb{E} \left[ \left\| \tilde{\mathbf{w}}_m - \mathbf{w}_k^* \right\|^2 \right] = L^2 \mathbb{E} \left[ \left\| \tilde{\mathbf{w}}_m - \mathbf{w}^* + \mathbf{w}^* - \mathbf{w}_k^* \right\|^2 \right] \\
& \stackrel{(c)}{\leq} 2L^2 \mathbb{E} \left[ \left\| \tilde{\mathbf{w}}_m - \mathbf{w}^* \right\|^2 \right] + 2L^2 \delta^2, \quad (57)
\end{aligned}$$

where (a) comes from  $\nabla F_k(\mathbf{w}_k^*) = \mathbf{0}$ , (b) exploits *Assumption 2*, and (c) uses *Assumption 3* and the inequality  $\|\mathbf{a} + \mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$ . By defining  $g_D(\mathbf{r}, b) = \sum_{k=1}^K \frac{\alpha_k}{p_k r_k}$ , we conclude that

$$\begin{aligned}
& \mathbb{E} \left[ \left\| \tilde{\mathbf{w}}_{m+1} - \mathbf{w}^* \right\|^2 \right] \\
& \leq (1 - \eta\mu + 2\eta^2 L^2 g_D(\mathbf{r}, b)) \mathbb{E} \left[ \left\| \tilde{\mathbf{w}}_m - \mathbf{w}^* \right\|^2 \right] \\
& \quad + \eta^2 (\phi(b) + 2L^2 \delta^2) g_D(\mathbf{r}, b) \\
& \leq (1 - \eta\mu + 2\eta^2 L^2 g_D(\mathbf{r}, b))^{m+1} \mathbb{E} \left[ \left\| \tilde{\mathbf{w}}_0 - \mathbf{w}^* \right\|^2 \right] \\
& \quad + \frac{\eta (\phi(b) + 2L^2 \delta^2) g_D(\mathbf{r}, b)}{\mu - 2\eta L^2 g_D(\mathbf{r}, b)}. \quad (58)
\end{aligned}$$

Plugging (58) into (50), we obtain the convergence result and complete the proof.

## APPENDIX C PROOF OF THEOREM 2

As for the analog transmission, the main difference from the digital transmission lies in the term  $B_1$  in (51). With the analog case,  $B_1$  is expressed as

$$B_1 = \sum_{k=1}^K \alpha_k \mathbb{E} \left[ \left\| \frac{\chi_k \xi_{k,A}}{r_k} \mathbf{g}_m^k + \bar{\mathbf{z}}_m - \mathbf{g}_m^k \right\|^2 \right]$$

$$\begin{aligned}
&= \sum_{k=1}^K \alpha_k \mathbb{E} \left[ \left\| \left( \frac{\chi_k \xi_{k,A}}{r_k} - 1 \right) \mathbf{g}_m^k \right\|^2 \right] + \mathbb{E} [\|\bar{\mathbf{z}}_m\|^2] \\
&= \sum_{k=1}^K \alpha_k \mathbb{E} \left[ \left( \frac{\chi_k \xi_{k,A}}{r_k} - 1 \right)^2 \right] \mathbb{E} [\|\nabla F_k(\tilde{\mathbf{w}}_m)\|^2] \\
&\quad + \mathbb{E} [\|\bar{\mathbf{z}}_m\|^2]. \tag{59}
\end{aligned}$$

For the equivalent noise term, recalling that  $\bar{\mathbf{z}}_m = \frac{\Re\{\mathbf{z}_m\}}{\zeta}$ , we first derive the scaling factor  $\zeta$ . Constrained by the transmit power budget in (25), the scaling factor  $\zeta$  must satisfy

$$\max_{k \in \mathcal{S}_m} \left\{ \frac{\zeta^2 e^{2\gamma_{\text{th}}} \alpha_k^2 d_k^\alpha}{\rho^2 r_k^2 |\hat{h}_k|^2} \mathbb{E} [\|\mathbf{g}_m^k\|^2] \right\} \leq P_{\max}. \tag{60}$$

Based on *Assumption 3* and the definition in (3), we can conclude that

$$\|\mathbf{g}_m^k\| \leq \frac{1}{D_k} \sum_{\mathbf{u} \in \mathcal{D}_k} \|\nabla \mathcal{L}(\mathbf{w}_m, \mathbf{u})\| \leq \gamma. \tag{61}$$

Note that for all the activated devices, we have  $|\hat{h}_k|^2 \geq \gamma_{\text{th}}$ . Hence, we select the feasible  $\zeta$  as

$$\zeta = \frac{\rho \sqrt{P_{\max} \gamma_{\text{th}}}}{\gamma e^{\gamma_{\text{th}}}} \min_k \left\{ \frac{r_k}{\alpha_k} d_k^{-\frac{\alpha}{2}} \right\}. \tag{62}$$

Then, we have

$$\mathbb{E} [\|\bar{\mathbf{z}}_m\|^2] = \frac{B N_0 \gamma^2 e^{2\gamma_{\text{th}}}}{2 P_{\max} \rho^2 \gamma_{\text{th}}} \max_k \left\{ \frac{\alpha_k^2}{r_k^2} d_k^\alpha \right\}. \tag{63}$$

Next, the variance of the coefficient distortion,  $\frac{\chi_k \xi_{k,D}}{r_k}$ , is calculated as

$$\begin{aligned}
&\mathbb{E} \left[ \left( \frac{\chi_k \xi_{k,A}}{r_k} - 1 \right)^2 \right] \stackrel{(a)}{=} r_k \mathbb{E} \left[ \left( \frac{\xi_{k,A}}{r_k} - 1 \right)^2 \right] + 1 - r_k \\
&= r_k \left( \mathbb{E} \left[ \left( \frac{\Re\{h_k^* \hat{h}_k\} e^{\gamma_{\text{th}}}}{|\hat{h}_k|^2 \rho r_k} - 1 \right)^2 \middle| |\hat{h}_k|^2 \geq \gamma_{\text{th}} \right] \right. \\
&\quad \times \Pr \{ |\hat{h}_k|^2 \geq \gamma_{\text{th}} \} + \Pr \{ |\hat{h}_k|^2 < \gamma_{\text{th}} \} \left. \right) + 1 - r_k \\
&= r_k e^{-\gamma_{\text{th}}} \mathbb{E} \left[ \left( \frac{\Re\{h_k^* \hat{h}_k\} e^{\gamma_{\text{th}}}}{|\hat{h}_k|^2 \rho r_k} - 1 \right)^2 \middle| |\hat{h}_k|^2 \geq \gamma_{\text{th}} \right] \\
&\quad + 1 - r_k e^{-\gamma_{\text{th}}} \\
&\stackrel{(b)}{=} \left( e^{\gamma_{\text{th}}} + \frac{(1 - \rho^2) \mathbb{E}_1(\gamma_{\text{th}}) e^{2\gamma_{\text{th}}}}{2 \rho^2} \right) \frac{1}{r_k} - 1, \tag{64}
\end{aligned}$$

where (a) exploits the independence of  $\chi_k$  and  $\xi_{k,A}$ , (b) is due to [34, Eq. (25)]. Substituting (64) into (59) and combining the results in (54) and (55), we complete the proof.

#### APPENDIX D PROOF OF COROLLARY 2

To begin with, the expectation of  $|\beta_k|^2$  is calculated as

$$\begin{aligned}
\mathbb{E} [|\beta_k|^2] &= \frac{\zeta^2 \lambda^2 \alpha_k d_k^\alpha}{r_k^2} \mathbb{E} \left[ \frac{1}{|\hat{h}_k|^2} \right] \\
&\stackrel{(a)}{=} \frac{\zeta^2 \lambda^2 \alpha_k d_k^\alpha}{r_k^2} \mathbb{E}_1(\gamma_{\text{th}}), \tag{65}
\end{aligned}$$

where (a) comes from the fact that  $|\hat{h}_k|^2$  follows an exponential distribution and the integral  $\int_{\gamma_{\text{th}}}^{\infty} \frac{1}{x} e^{-x} dx = \mathbb{E}_1(\gamma_{\text{th}})$ . Substituting (61) into (26), we get a feasible  $\zeta$  as

$$\zeta = \frac{\rho \sqrt{P_{\text{ave}}}}{\gamma e^{\gamma_{\text{th}}} \sqrt{\mathbb{E}_1(\gamma_{\text{th}})}} \min_k \left\{ \frac{r_k}{\alpha_k} d_k^{-\frac{\alpha}{2}} \right\}. \tag{66}$$

Then, following the same steps as in Appendix C, we complete the proof.

#### REFERENCES

- [1] J. Yao, W. Xu, Z. Yang, X. You, M. Bennis, and H. V. Poor, "Digital versus analog transmissions for federated learning over wireless networks," accepted by *Proc. IEEE Int. Conf. Commun. (ICC)*, 2024. Available: <https://arxiv.org/abs/2402.09657>.
- [2] Y. C. Eldar, A. Goldsmith, D. Gundüz, and H. V. Poor, *Machine Learning and Wireless Communications*. Cambridge, UK: Cambridge University Press, 2022.
- [3] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May/Jun. 2020.
- [4] W. Shi *et al.*, "Intelligent reflection enabling technologies for integrated and green Internet-of-Everything beyond 5G: Communication, sensing, and security," *IEEE Wireless Commun.*, vol. 30, no. 2, pp. 147–154, Apr. 2023.
- [5] Z. He *et al.*, "Unlocking potentials of near-field propagation: ELAA-empowered integrated sensing and communication," 2024, arXiv:2404.18587.
- [6] W. Xu *et al.*, "Edge learning for B5G networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 1, pp. 9–39, Jan. 2023.
- [7] G. Zhu *et al.*, "Pushing AI to wireless network edge: An overview on integrated sensing, communication, and computation towards 6G," *Sci. China Inf. Sci.*, vol. 66, no. pp. 130301:1–19, Feb. 2023.
- [8] M. Chen *et al.*, "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3579–3605, Dec. 2021.
- [9] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2017, pp. 1273–1282.
- [10] Y. Yang, Z. Zhang, and Q. Yang, "Communication-efficient federated learning with binary neural networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3836–3850, Dec. 2021.
- [11] Y. Guo, R. Zhao, S. Lai, L. Fan, X. Lei, and G. K. Karagiannidis, "Distributed machine learning for multiuser mobile edge computing systems," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 3, pp. 460–473, Apr. 2022.
- [12] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, Jan. 2021.
- [13] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Babaei, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1935–1949, Mar. 2021.
- [14] M. M. Amiri and D. Gundüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, May 2020.
- [15] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.
- [16] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication-efficient federated learning," *Proc. Nat. Acad. Sci.*, vol. 118, no. 17, 2021.
- [17] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "UveQFed: Universal vector quantization for federated learning," *IEEE Trans. Signal Process.*, vol. 69, pp. 500–514, 2021.
- [18] M. Salehi and E. Hossain, "Federated learning in unreliable and resource-constrained cellular wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5136–5151, Aug. 2021.
- [19] S. Zheng, C. Shen and X. Chen, "Design and analysis of uplink and downlink communications for federated learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2150–2167, Jul. 2021.



- [20] Y. Wang, Y. Xu, Q. Shi, and T.-H. Chang, "Quantized federated learning under transmission delay and outage constraints," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 323–341, Jan. 2022.
- [21] M. Lan, Q. Ling, S. Xiao, and W. Zhang, "Quantization bits allocation for wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 8336–8351, Nov. 2023.
- [22] H. Li, R. Wang, W. Zhang, and J. Wu, "One bit aggregation for federated edge learning with reconfigurable intelligent surface: Analysis and optimization," *IEEE Trans. Wireless Commun.*, vol. 22, no. 2, pp. 872–888, Feb. 2023.
- [23] H. Ye, L. Liang, and G. Y. Li, "Decentralized federated learning with unreliable communications," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 3, pp. 487–500, Apr. 2022.
- [24] J. Yao, Z. Yang, W. Xu, M. Chen, and D. Niyato, "GoMORE: Global model reuse for resource-constrained wireless federated learning," *IEEE Wireless Lett.*, vol. 12, no. 9, pp. 1543–1547, Sept. 2023.
- [25] S. Liu, G. Yu, R. Yin, J. Yuan, L. Shen, and C. Liu, "Joint model pruning and device selection for communication-efficient federated edge learning," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 231–244, Jan. 2022.
- [26] H. H. Yang, Z. Chen, T. Q. S. Quek, and H. V. Poor, "Revisiting analog over-the-air machine learning: The blessing and curse of interference," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 3, pp. 406–419, Apr. 2022.
- [27] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Signal Process.*, vol. 68, pp. 2897–2911, Apr. 2020.
- [28] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, Jan. 2020.
- [29] X. Cao, G. Zhu, J. Xu, and S. Cui, "Transmission power control for over-the-air federated averaging at network edge," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1571–1586, May 2022.
- [30] X. Yu, B. Xiao, W. Ni, and X. Wang, "Optimal adaptive power control for over-the-air federated edge learning under fading channels," *IEEE Trans. Commun.*, vol. 71, no. 9, pp. 5199–5213, Sept. 2023.
- [31] W. Guo *et al.*, "Joint device selection and power control for wireless federated learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2395–2410, Aug. 2022.
- [32] F. Ang *et al.*, "Robust federated learning with noisy communication," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3452–3464, Jun. 2020.
- [33] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2022–2035, Mar. 2020.
- [34] J. Yao, Z. Yang, W. Xu, D. Niyato, and X. You, "Imperfect CSI: A key factor of uncertainty to over-the-air federated learning," *IEEE Wireless Lett.*, vol. 12, no. 12, pp. 2273–2277, Dec. 2023.
- [35] K. Guo, Z. Chen, H. H. Yang, and T. Q. S. Quek, "Dynamic scheduling for heterogeneous federated learning in private 5G edge networks," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 1, pp. 26–40, Jan. 2022.
- [36] X. Wei and C. Shen, "Federated learning over noisy channels: Convergence analysis and design examples," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 1253–1268, Jun. 2022.
- [37] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 170–185, Jan. 2021.
- [38] X. Li, G. Zhu, Y. Gong, and K. Huang, "Wirelessly powered data aggregation for IoT via over-the-air function computation: Beamforming and power control," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3437–3452, Jul. 2019.
- [39] Z. Lin, X. Li, V. K. N. Lau, Y. Gong, and K. Huang, "Deploying federated learning in large-scale cellular networks: Spatial convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1542–1556, Mar. 2022.
- [40] M. Mohammadi Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, vol. 68, pp. 2155–2169, 2020.
- [41] H. Xing, O. Simeone, and S. Bi, "Federated learning over wireless device-to-device networks: Algorithms and convergence analysis," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3723–3741, Dec. 2021.
- [42] E. Rizk, S. Vlaski, and A. H. Sayed, "Federated learning under importance sampling," *IEEE Trans. Signal Process.*, vol. 70, pp. 5381–5396, 2022.
- [43] W. Dinkelbach, "On nonlinear fractional programming," *Manage. Sci.*, vol. 133, no. 7, pp. 492–498, Mar. 1967.
- [44] Z. He *et al.*, "Energy efficient beamforming optimization for integrated sensing and communication," *IEEE Wireless Commun. Lett.*, vol. 11, no. 7, pp. 1374–1378, Jul. 2022.
- [45] M. Grant and S. Boyd. (2016). CVX: MATLAB Software for Disciplined Convex Programming. [Online]. Available: <http://cvxr.com/cvx>
- [46] J. Wang, Y. Mao, T. Wang, and Y. Shi, "Green federated learning over cloud-RAN with limited fronthaul capacity and quantized neural networks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 5, pp. 4300–4314, May 2024.
- [47] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Convergence of federated learning over a noisy downlink," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1422–1437, Mar. 2022.
- [48] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-iid data," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2019.
- [49] T. Li *et al.*, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, 2020, pp. 429–450.
- [50] Y. Sun, Z. Lin, Y. Mao, S. Jin, and J. Zhang, "Channel and gradient-importance aware device scheduling for over-the-air federated learning," *IEEE Trans. Wireless Commun. (Early Access)*. 2023.



**Jiacheng Yao** (Graduate Student Member, IEEE) received the B.S. degree in information science and engineering from Southeast University, Nanjing, China, in 2022. He is currently working towards a Ph.D. degree in information and communication engineering at the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China. His current research includes integrated communication and computation. He was a recipient of the Best Paper Award from IEEE ICC 2024.



**Wei Xu** (S'07-M'09-SM'15) received his B.Sc. degree in electrical engineering and his M.S. and Ph.D. degrees in communication and information engineering from Southeast University, Nanjing, China in 2003, 2006, and 2009, respectively. Between 2009 and 2010, he was a Post-Doctoral Research Fellow at the University of Victoria, Canada. He was an Adjunct Professor of the University of Victoria in Canada from 2017 to 2020, and a Distinguished Visiting Fellow of the Royal Academy of Engineering, U.K. in 2019. He is currently a Professor at Southeast University. His research interests include information theory, signal processing, and machine learning for wireless communications.

He received the Science and Technology Award for Young Scholars of the China Institute of Communications in 2018, the Science and Technology Award of the Chinese Institute of Electronics (Second Prize) in 2019, the National Natural Science Foundation of China for Outstanding Young Scholars in 2020, the IEEE Communications Society Heinrich Hertz Award in 2023, and the Best Paper Awards at IEEE Globecom 2014, IEEE ICC 2014, ISWCS 2018, WCSP 2017, 2021, and ICC 2024. He served as an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS from 2018 to 2023. He also served for IEEE COMMUNICATIONS LETTERS as firstly an Editor from 2012 to 2017, then a Senior Editor from 2020 to 2023, and is now an Area Editor. He is a Fellow of IET.





**Zhaohui Yang** (Member, IEEE) received the Ph.D. degree from Southeast University, Nanjing, China, in 2018. From 2018 to 2020, he was a Post-Doctoral Research Associate at the Center for Telecommunications Research, Department of Informatics, King's College London, U.K. From 2020 to 2022, he was a Research Fellow at the Department of Electronic and Electrical Engineering, University College London, U.K. He is currently a ZJU Young Professor with the Zhejiang Key Laboratory of Information Processing Communication and Networking, College of Information Science and Electronic Engineering, Zhejiang University. His research interests include joint communication, sensing, and computation, federated learning, and semantic communication. He received the 2023 IEEE Marconi Prize Paper Award, 2023 IEEE Katherine Johnson Young Author Paper Award, 2023 IEEE ICCCN best paper award, and 2024 IEEE ICC best paper award. He was the Co-Chair for international workshops with more than ten times including IEEE ICC, IEEE GLOBECOM, IEEE WCNC, IEEE TGCN, IEEE CL, IEEE TMLCN. He has served as a Guest Editor for several journals including IEEE Journal on Selected Areas in Communications.



**Xiaohu You** (Fellow, IEEE) received the M.S. and Ph.D. degrees in electrical engineering from Southeast University, Nanjing, China, in 1985 and 1988, respectively. Since 1990, he has been with the National Mobile Communications Research Laboratory, Southeast University, where he is currently the Director and a Professor. From 1999 to 2002, he was a Principal Expert of the C3G Project, responsible for organizing China's 3G Mobile Communications Research and Development Activities. From 2001 to 2006, he was a Principal Expert of the China

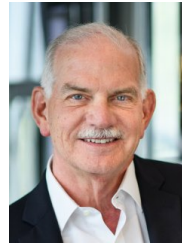
National 863 Beyond 3G FuTURE Project. Since 2013, he has been a Principal Investigator of the China National 863 5G Project. He has contributed over 200 IEEE journal articles and two books in the areas of adaptive signal processing and neural networks, and their applications to communication systems. His research interests include mobile communication systems, and signal processing and its applications.

Dr. You was selected as an IEEE Fellow for his contributions to the development of mobile communications in China in 2011. He was a recipient of the National 1st Class Invention Prize in 2011. He has served as the General Chair of the IEEE Wireless Communications and Networking Conference 2013, the IEEE Vehicular Technology Conference 2016, and the IEEE International Conference on Communications 2019. He is the Secretary-General of the FuTURE Forum and the Vice-Chair of the China IMT-2020 Promotion Group and the China National Mega Project on New Generation Mobile Network.



**Mehdi Bennis** (Fellow, IEEE) is a Professor at the Centre for Wireless Communications, University of Oulu, Finland, IEEE Fellow and head of the intelligent connectivity and networks/systems group (ICON). He has published more than 400 research papers in international conferences, journals and book chapters. He has been the recipient of several prestigious awards including the 2015 Fred W. Ellersick Prize from the IEEE Communications Society, the 2016 Best Tutorial Prize from the IEEE Communications Society, the 2017 EURASIP Best

paper Award for the Journal of Wireless Communications and Networks, the all-University of Oulu award for research, the 2019 IEEE ComSoc Radio Communications Committee Early Achievement Award and the 2020-2023 Clarivate Highly Cited Researcher by the Web of Science.



**H. Vincent Poor** (S'72, M'77, SM'82, F'87) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990 he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor. During 2006 to 2016, he served as the dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge.

His research interests are in the areas of information theory, machine learning and network science, and their applications in wireless networks, energy systems and related fields. Among his publications in these areas is the recent book *Machine Learning and Wireless Communications*. (Cambridge University Press, 2022). Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences and is a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the IEEE Alexander Graham Bell Medal in 2017.