Energy and Spectrum Efficient Federated Learning via High-Precision Over-the-Air Computation

Liang Li, *Member*, *IEEE*, Chenpei Huang, *Student Member*, *IEEE*, Dian Shi, *Member*, *IEEE*, Hao Wang, *Member*, *IEEE*, Xiangwei Zhou, *Senior Member*, *IEEE*, Minglei Shu, *Member*, *IEEE*, and Miao Pan, *Senior Member*, *IEEE*

***Abstract*—Federated learning (FL) enables mobile devices to collaboratively learn a shared prediction model while keeping data locally. However, there are two major research challenges to practically deploy FL over mobile devices: (i) frequent wireless updates of huge size gradients v.s. limited spectrum resources, and (ii) energy-hungry FL communication and local computing during training v.s. battery-constrained mobile devices. To address those challenges, in this paper, we propose a novel multi-bit over-the-air computation (M-AirComp) approach for spectrum-efficient aggregation of local model updates in FL and further present an energy-efficient FL design for mobile devices. Specifically, a high-precision digital modulation scheme is designed and incorporated in the M-AirComp, allowing mobile devices to upload model updates at the selected positions simultaneously in the multi-access channel. Moreover, we theoretically analyze the convergence property of our FL algorithm. Guided by FL convergence analysis, we formulate a joint transmission probability and local computing control optimization, aiming to minimize the overall energy consumption (i.e., iterative local computing + multi-round communications) of mobile devices in FL. Extensive simulation results show that our proposed scheme outperforms existing ones in terms of spectrum utilization, energy efficiency, and learning accuracy.**

***Index Terms*—Federated learning, over-the-air computation, gradient quantization, energy efficiency.**

I. Introduction

With the development of mobile communications and Internet-of-Things (IoT) technologies, mobile devices with built-in sensors and Internet connectivity have proliferated huge volumes of data at the network edge. These data can be collected and analyzed to build increasingly complex machine learning models. To avoid raw-data sharing among the untrustworthy parties and leverage the ever-increasing computation capability of mobile devices, the emerging federated learning (FL) framework allows participating mobile devices to collaboratively train a machine learning model under the orchestration of a centralized server by just exchanging the local model updates with others via wireless communications. With such desirable properties, FL over mobile devices has inspired a wide utilization in a large variety of intelligent services, such as the keyword prediction [1], voice classifier [2], and e-health [3], etc.

Although only model updates instead of raw data are transferred between mobile devices and the FL server, such updates could contain hundreds of millions of parameters with complex neural networks. That makes the uplink transmissions from mobile devices to the FL server for model aggregation particularly challenging, resulting in a huge burden on both wireless networks and mobile devices. On the one hand, the spectrum resource that can be allocated to each device decreases proportionally as the number of devices increases, which hampers the scalability of FL to accommodate a large number of mobile devices with limited spectrum resources. On the other hand, transmitting a large volume of model updates periodically and executing heavy local on-device computations can quickly drain out the energy of battery-powered mobile devices. Such a mismatch restricts mobile devices or makes them reluctant to participate in FL.

Over-the-air computation (AirComp) provides a promising solution to address the aforementioned spectrum challenge by achieving scalable and efficient model update aggregation in FL. Unlike the conventional orthogonal multiple access techniques, where each user is restricted to its allocated spectrum band [4], AirComp allows all the users to utilize the whole spectrum for simultaneous transmission. By applying AirComp to FL, all the participating devices can transmit their model updates on the same channel. Due to the fact that multi-access channel (MAC) inherently yields an additive superposed signal, the signals of all the participating devices are aligned to obtain desired arithmetic computation results directly over the air, thus significantly improving the spectrum efficiency. However, most existing works employ the analogy modulation to design their over-the-air FL schemes, which is not compatible with commercial off-the-shelf digital mobile devices and thus hinders their deployment in current/future communication systems, such as LTE, 5G, Wi-Fi 6, and 6G, etc. Besides, most existing efforts focus on single-iteration transmission design for AirComp-based FL [5], [6], and the

L. Li is with the School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing, 100876, China (e-mail: liliang1127@bupt.edu.cn).

C. Huang, D. Shi and M. Pan are with the Electrical and Computer Engineering Department, University of Houston, TX, 77004, USA (e-mail: chuang25@uh.edu, dshi3@uh.edu, mpan2@uh.edu).

H. Wang is with the Division of Computer Science and Engineering, Louisiana State University, Baton Rouge, LA, 70803, USA (e-mail: haowang@lsu.edu).

X. Zhou is with the Division of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA, 70803, USA (e-mail: xwzhou@lsu.edu).

M. Shu is with the Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), Jinan, 250353, China (e-mail: shuml@sdas.org).

The work of L. Li was supported in part by National Natural Science Foundation of China under grants 62201071. The work of C. Huang, D. Shi, and M. Pan was supported in part by the US National Science Foundation under grant CNS-2107057. *(Corresponding author: Minglei Shu.)*

impacts of AirComp on the long-term federated training performance, especially the convergence rate, are widely overlooked.

In this work, we design a multi-bit Aircomp (M-AirComp) FL scheme, named ESOAFL, whose merits are two-fold: i) It is compatible with the most common Quadrature Amplitude Modulation (QAM) transmitter where gradient quantization is incorporated to facilitate the digital modulation, so that one do not need to modify the modulation protocols manufactured within commercial off-the-shelf mobile devices for the periodical gradient transmission. ii) It filters the FL participants with good channel conditions based on a well-controlled transmission probability to transmit the updated gradients, which helps save the transmission energy compared with other AirComp-based FL schemes. We analyze the convergence property of our ESOAFL algorithm and derive the number of communication rounds needed for achieving the convergence. Guided by the theoretical results, we model the energy consumption of all the FL devices from the long-term learning perspective, where wireless communication (i.e., "talking") and local computing (i.e., "working") are two main focuses. To make the ESOAFL battery-friendly to the participating mobile devices, a joint transmission probability and local computing control scheme is developed to balance "talking" and "working" during performing the ESOAFL, with the goal of energy consumption minimization. Our salient contributions are summarized as follows.

- We propose an energy and spectrum efficient M-AirComp FL (ESOAFL) scheme where the updated gradients of every FL participant are quantized into high-precision bitstreams, adapting to the digital modulation settings. To facilitate the M-AirComp, a transmission control policy is integrated in ESOAFL to only allow the FL participants with good channel conditions for FL model aggregation by introducing a tunable parameter, i.e., transmission probability.
- We theoretically analyze the convergence property of our ESOAFL to characterize the impacts of the M-AirComp on FL. Guided by it, the transmission probability and local computing iterations are jointly optimized from the long-term learning perspective, aiming to achieve energy-efficient federated training on mobile devices over spectrum-constrained wireless networks.
- We conduct extensive simulations to verify the superiority of the ESOAFL compared to several baselines, under varying learning models, training datasets, and network settings. It shows that our ESOAFL scheme can improve spectral efficiency dozens of times and save at least half of the energy consumption.

The remainder of the paper is organized as follows. Section II provides some preliminaries of AirComp and FL. Section III presents our M-AirComp design and the ESOAFL scheme. Section IV gives the theoretical analysis of ESOAFL and elaborates on the joint transmission probability and local computing control approach. Numerical simulations are provided in Section V, and VI reviews related works. Section VII finally concludes the paper and provides future work.

II. Preliminaries of FL and AirComp FL

A. Preliminaries of FL

We consider a federated learning system consisting of K participating users carrying mobile devices, where each user $k \in \{1, 2, \ldots, K\}$ has its own dataset, denoted by $\mathcal{D}_k$. The goal of FL is to collaborate the users to perform a unified optimization task, formally written as:

$$\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) \triangleq \frac{1}{K}\sum_{k=1}^{K} f_k(\mathbf{w}), \tag{1}$$

where f_k is the local loss function corresponding to user k, and d is the dimension of the model parameters.

Let $r \in \{1, 2, \ldots, R\}$ be the index of FL global communication round, and H be the number of local training iterations executed between every two consecutive global communication rounds. Moreover, we define $\mathbf{w}^r$ as the global model at the r-th communication round and define $\mathbf{w}_k^{r,h}$ as the local model of user k at the h-th local iteration in the r-th communication round. Then the local training process of user k in the r-th communication round is given by:

$$\mathbf{w}_k^{r,h+1} = \mathbf{w}_k^{r,h} - \eta \nabla F_k(\mathbf{w}_k^{r,h}) \text{ for } h = 0, 1, ..., H-1, \tag{2}$$

where $\nabla F_k(\mathbf{w}_k^{r,h})$ is a stochastic gradient of function $f(\cdot)$ with a random batch-size data, and η is the local learning rate. Here, $\nabla F_k(\mathbf{w}_k^{r,h})$ is an unbiased estimation of $\nabla f_k(\mathbf{w}_k^{r,h})$, i.e., $\mathbb{E}_{\xi\sim\mathcal{D}_k}\left[\nabla F_k(\mathbf{w}) \mid \xi\right] = \nabla f_k(\mathbf{w})$, where ξ represents the randomness like the batch-size index. After finishing the local training, every participating user uploads its local model updates to the server for global aggregation, i.e., $\eta\sum_{h=0}^{H-1}\nabla F_k(\mathbf{w}_k^{r,h})$, and the server then broadcasts the most recent global model to initiate a new round of local training. The above process is repeated until the global model converges.

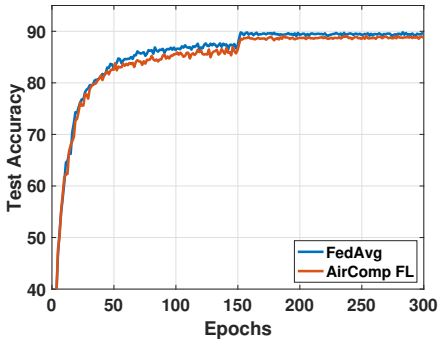
B. Preliminaries of AirComp FL

During the FL process, all the users have to transmit their local updates to the server for global aggregation, which may result in severe transmission congestion and consume significant communication resources, especially in cases with massive participating users. As one of the advanced wireless techniques, over-the-air computation (AirComp) enables all the users to simultaneously transmit the local gradients over the same wireless medium without spectrum allocation and naturally aggregates the local updates during the signal propagation, which exhibits great potentials to improve the spectrum utilization.

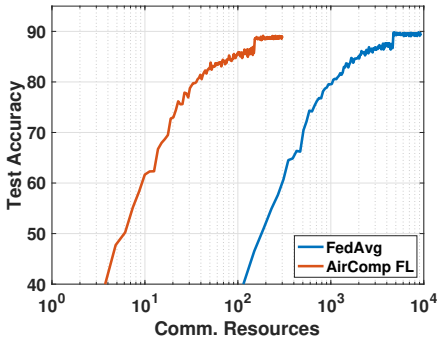
Let $\mathcal{X} := \{x_1, x_2, ..., x_K\}$ and $\tilde{y}$ denote the input set and the output objective of the AirComp operation, respectively. Here, x_k is the gradients to be transmitted by user k, and $\tilde{y}$ is the global aggregation result received at the server with AirComp. Generally, an AirComp-based wireless communication system adopts precoding and amplification at transmitters, while receivers often have equalization blocks for signal detection. Therefore, AirComp computes the aggregated objective as

$$\tilde{y} := \text{Air}\left(\mathcal{X}\right) = \frac{a}{K}\left[\sum_{k=0}^{K} h_k p_k x_k + n\right], \tag{3}$$

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3287549



(a) Test accuracy vs. epochs



(b) Test accuracy vs. comm. resources

Fig. 1. Over-the-Air federated learning (AirComp FL).

where $h_k \in \mathbb{C}$ is the channel coefficient between user k and the server, and $n \sim N(0, \sigma_z^2)$ is the additive white Gaussian noise (AWGN) at the receiver. The Tx-scaling factor $p_k \in \mathbb{C}$, a.k.a. power control policy, compensates the phase shift posed by the channel and amplifies the transmit signal. The goal of the Tx-scaling is to ensure that each participating user contributes equally at the receiving antenna and the superposed signal is proportional to the ideal summation, which is defined as the average operation over the input set without AirComp, i.e., $y := \frac{1}{K}\sum_{k=1}^{K} x_k$. Accordingly, the Rx-scaling factor $a \in \mathbb{R}$ acts as an equalizer and recovers the sampled analog result to its expected value.

C. Preliminary Experiments on AirComp FL

To demonstrate the spectrum-efficient benefit of AirComp FL, we conduct the preliminary experiments on AirComp FL and the classic FedAvg without AirComp, as shown in Fig. 1. Here, 10 users are considered to participate in an FL task and collaboratively train a ResNet-20 model on the CIFAR-10 dataset. Both the communication bandwidth and the number of training epochs are set to be the same for these two schemes. Taking test accuracy as the measure, Fig. 1(a) depicts the convergence performance of the training process, while Fig. 1(b) displays the communication resource consumption during the training. It shows that, compared with FedAvg, AirComp FL only requires a little more or even the same number of data epochs to achieve the target accuracy. This impies that AirComp operation imposes negligible impacts on the convergence rate of FL. Meanwhile,

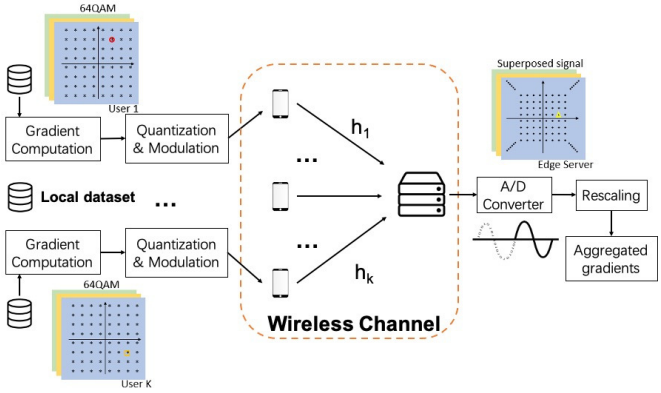


Fig. 2. The design of multi-bit Over-the-Air computation.

the communication resource consumption of the AirComp FL is much less than that of FedAvg, since the latter forces the users to use orthogonal channels for interference avoidance instead of performing concurrent transmission over the same spectrum like AirComp FL does. Note here that we use the normalized communication resources for Fig. 1(b) illustration and assume one unit communication resource is consumed in each communication round in AirComp FL.

III. The Design of M-AirComp and M-AirComp-based FL

A. M-AirComp Design

Different from the most existing AirComp methods with an analogy modulation scheme, we establish a digital modulation scheme for the AirComp to cater for the commercial transmit devices and design a multi-bit over-the-Air computation scheme (M-AirComp). To this end, the Rx-scaling factor a performs as a digital domain equalizer, and the division operation in Eq. (3) to calculate the arithmetic average is also in the digital domain. In order to eliminate the burden of redesigning the modulation scheme, we tend to integrate the gradient quantization to the most common Quadrature Amplitude Modulation (QAM) in LTE, 5G, and Wi-Fi 6 standard [7]. Instead of transmitting arbitrary values, gradients to transmit are clipped and quantized as Multiple Amplitude Shift Keying (MASK) symbols, so as to be compatible with modern digital devices. Two MASK-modulated gradients can be transmitted orthogonally using in-phase (I) and quadrature (Q) channel simultaneously. We notice that it is equivalent to mapping two separate gradients onto a symbol from the square M^2 QAM constellation. Here, we limit M between 2 to 2^b. For example, when b is set as 3, the user will use 64QAM to transmit two gradients, as shown in Fig. 2. In this way, altering the value M at the transmitter according to the estimated channel gain allows full digital data transmission while preserving b-bit resolution.

Assume that the server equips with a high-resolution analog-to-digital converter (ADC) (e.g., 16-bit). While receiving, multiple QAM symbols superpose at the sampling instance, which can be viewed from (a part of) a higher-order rectangular QAM constellation diagram (when the number of mobile devices is odd) or a zero-centered constellation diagram (when the

number of users is even). Since the biggest possible value after aggregation can be obtained from user feedback, we can utilize this value as the ADC reference voltage. In order to alleviate the detection complexity, we directly use the quantized samples followed by Rx-scaling defined in Eq. (3) in the digital domain. In this way, the transmission module is implemented in a digital manner, which enables the M-AirComp to have better compatibility compared with traditional AirComp. The process is also illustrated in Fig. 2. This result can be viewed as the desired computational result added by quantization error and channel noise, whose impacts on federated learning performance are analyzed in the following section.

In the transmission process, every device is subject to an average transmitting power budget, i.e., P^0. The transmission power constraint is given by

$$\mathbb{E}[|p_k|^2] \leq P^0, \forall k, \tag{4}$$

where the expectation is taken over the distribution of random channel coefficients. Recall that gradient parameters transmitted by different devices are received with identical amplitudes for implementing gradient aggregation via AirComp, which can be achieved by inverting the channels via power control. In practice, some devices facing severe signal fading may not completely align their amplitude due to the power limit, i.e., the Tx-scaling factor $p_k \in \mathbb{C}$ cannot be infinitely enlarged to meet the amplitude alignment requirement. This work adopts an energy efficient power control policy that performs channel-inversion-based power control only for the users with desired channel gain. The users with poor channel conditions are not allowed to transmit, i.e. its transmit power is set to be zero. Let g_{th} be the channel gain threshold for possible transmission, and the power control policy p_k for any user k can be represented as:

$$p_k = \begin{cases} \frac{\sqrt{\varrho} h_k^\dagger}{|h_k|^2}, & |h_k|^2 \geq g_{\text{th}} \\ 0, & |h_k|^2 < g_{\text{th}} \end{cases}. \tag{5}$$

Here, ϱ is a scaling factor to guarantee the desired SNR, which determines the receiving power of the gradient update from each user; h_k represents the channel coefficient and its conjugate is denoted by $h_k^\dagger$. Under the above power control policy, only users facing channel gain larger than g_{th} can be allowed to transmit their updated gradients. Note that the threshold g_{th} can be adjusted to control the gradient transmissions. With the power constraint in Eq. (4), we have $|h_k|^2 = \varrho/|p_k|^2 \geq \varrho/P^0$. It means that the threshold g_{th} can be set as an arbitrary value larger than the minimum value $g_{\text{th}}^{\min} := \frac{\varrho}{P^0}$. Specifically, in a certain communication environment, the greater the threshold g_{th} we set, the more the users are allowed to upload their updated gradients. By varying the threshold g_{th}, our M-AirComp design has the potential to only involve the users with good channel conditions, which allows to lower the transmit power of the edge devices and thereby benefits in energy-saving. We define ρ as the average transmission probability that the users' channel gain is above the power-cutoff threshold g_{th}, which reflects the participation degree of the FL users. Note here that any threshold g_{th} will correspond to a transmission probability ρ. Assume that the channel coefficient is Rayleigh distributed, i.e., $h_k \sim CN(0, \sqrt{\lambda})$ and thus the channel gain $g_k = |h_k|^2$ follows an exponential distribution. The transmission probability ρ corresponding to the threshold g_{th} can be calculated as:

$$\rho = \Pr(g_k \geq g_{\text{th}}) = \int_{g_{\text{th}}}^{\infty} \lambda e^{-\lambda x} \mathrm{d}x = e^{-\lambda g_{\text{th}}}. \tag{6}$$

With the transmission probability ρ, the Rx-scaling factor a will be set as $\frac{1}{\sqrt{\varrho}\rho}$ to rescale the received signal. By substituting $g_{\text{th}}^{\min} := \frac{\varrho}{P^0}$ into Eq. (6), we have the highest transmission probability $\rho^{\max}$ as $\rho^{\max} = e^{-\lambda g_{\text{th}}^{\min}} = e^{-\lambda \frac{\varrho}{P_0}}$. It implies that, due to the fading channel and the devices' power budget, the transmission probability is upper-bounded.

Algorithm 1 ESOAFL Algorithm

Initialization: Initialize the global model $\mathbf{w}^0$ and set $\mathbf{w}_k^{0,0} = \mathbf{w}^0, \forall k \in \mathcal{K}$; Set the learning rate γ and η, local computing iterations H, and the channel gain threshold g_{th}

Initialize the communication index $r = 0$ and the local computing iteration count $h = 0$

1: **while** $r < R$ **do**
2: **for** $h = 0, ..., H-1$ **do**
3: Each device k computes the unbiased stochastic gradients $\nabla F_k(\mathbf{w}_k^{r,h})$ of $f_k(\mathbf{w}_k^r)$ with one batch size of data from the dataset $\mathcal{D}_k$
4: Each device k in parallel updates its local model: $\mathbf{w}_k^{r,h+1} = \mathbf{w}_k^{r,h} - \eta \nabla F_k(\mathbf{w}_k^{r,h}), \quad \forall k$
5: **end for**
6: Each device k calculates the accumulated gradients with gradient quantization as $Q\left(\eta \sum_{h=0}^{H-1} \nabla F_k(\mathbf{w}_k^{r,h})\right)$.
7: Each device k transmits the quantized accumulated gradients if the observed channel gain larger than the pre-selected threshold g_{th} , i.e., $|h_k|^2 \geq g_{\text{th}}$; otherwise, no transmission.
8: All the local gradients are aggregated over the air to update the global model via Eq. (7).
9: Update $r \leftarrow r + 1$.
10: Each device k updates its local model $\mathbf{w}_k^{r,0} = \mathbf{w}^r$.
11: **end while**

B. M-AirComp-based FL Design

Based on M-AirComp, this subsection presents an **E**nergy and **S**pectrum Efficient **O**ver the **A**ir **F**ederated **L**earning (**ESOAFL**) algorithm integrating gradient quantization, where the overview is illustrated in Fig. 2. The pseudocode of our ESOAFL is given in Alg. 1, and the details are described in the following.

Following the ESOAFL, all the participating users start the training procedure with the initialized model parameters. Here, we assume a synchronized FL setting where every user periodically performs the same number of local iterations, i.e., H, with mini-batch size data drawn from its own dataset for model aggregation. After the local training, a uniform gradient quantization operator $Q(\cdot)$ is utilized to quantize the updated

TABLE I
SUMMARY OF NOTATIONS.

$k(K)$	Index (number) of FL clients	$r(R)$	Index (number) of FL global round
$\mathbf{w}_k^{r,h}$	Local model of client k	$\mathbf{w}^r$	Global model at the r-th round
p_k	Transmitting power of client k	h_k	Channel coefficient of client k
x_k	Transmit signal of client k	$\widehat{y}$	Global aggregation at the server with AirComp
H	The number of local iterations	g_{th}	Power-cutoff threshold for transmitting
ρ	Transmission probability	ϵ	Target training accuracy
P^{comm}	Transmission power per global round	T^{comm}	Transmission time per global round
P^{comp}	Transmission power per global round	T^{comp}	Transmission time per global round

gradients into low bits, i.e., 4-bit or 8-bit. Taking b-bit quantization for example, the local updates of all the participants are quantized to 2^b levels with a specific maximum/minimum value, catering to the digital wireless transmission scheme. Next, every 2 gradient element is modulated into one digital symbol for transmission. We assume the symbol-level synchronization among all the mobile devices that ensures coherent and concurrent transmission. This assumption can be realized by dedicating the bandwidth for mobile device synchronization, e.g., 1.08 MHz primary synchronization channel (PSCH) and secondary synchronization channel (SSCH) in LTE system [8], or the AirShare [9] for distributed MIMO synchronization. Then we employ the M-AirComp operator $\text{Air}(\cdot)$, along with the proposed energy efficient power control policy. In specific, the threshold g_{th} is determined firstly, which is one-to-one mapped with transmission probability ρ by $\rho = e^{-\lambda g_{\text{th}}}$, and then the FL user whose channel gain larger than g_{th} can be allowed to transmit its gradient updates. Because M-AirComp integrates wireless transmission and model aggregation over the air, the server receives only the aggregated gradients, based on which the global model is updated by:

$$\mathbf{w}^{r+1} = \mathbf{w}^r - \text{Air}\left(\left\{Q\left(\eta \sum_{h=0}^{H-1} \nabla F_k(\mathbf{w}_k^{r,h})\right)\right\}_{\mathcal{K}}\right). \quad (7)$$

After that, the server will broadcast the global model to all devices for the next-round federated training. We repeat the above procedure for R rounds until the model converges to a stationary point. Particularly, the convergence requirement can be represented as $\frac{1}{R}\sum_{r=0}^{R-1} \|\nabla f^r\|_2^2 \leq \epsilon$, where ϵ denotes the target training loss and ∇f^r is the global function gradient at the r-th communication round.

IV. Spectrum and Energy Efficient FL: Formulation and Solution

In this section, we formulate an overall energy minimization problem and establish the communication and computation energy models of the proposed ESOAFL algorithm. Based on the derived convergence analysis, we then optimize the control policy in terms of the transmission probability ρ and local computing iterations H to minimize the overall energy consumption. In Table I, we summarize the important notations we use throughout the paper.

A. Energy Minimization Problem Formulation

It is challenging to deploy energy-hungry FL tasks on mobile devices due to their limited battery capacity. Hence, in this work, we aim to minimize the total energy consumption of FL training via joint control of local computing iterations H and transmission probability ρ. The average energy consumption per communication round of mobile device is cast as $E = E^{comm}(\rho) + E^{comp}H$. Here, $E^{comm}(\rho)$ is the communication energy consumed to transmit the updated gradients, which is related to the transmission probability ρ, and E^{comp} is the computing energy of performing one local iteration. Our goal is to minimize the overall energy consumption during the federated training while guaranteeing the model convergence, which is formulated as:

$$\begin{aligned} \min \quad & \mathbb{E}\left[E_{tot}\right] \triangleq \mathbb{E}\left[RE^{comm}(\rho)\right] + \mathbb{E}\left[RE^{comp}H\right] \\ \text{s.t.,} \quad & \frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}\left[\|\nabla f^r\|_2^2\right] \leq \epsilon. \end{aligned} \quad (8)$$

Here, ϵ is the target FL accuracy, and R indicates the number of global communication rounds required for convergence. Note that the value of R is related to the model update behaviors and the target training accuracy, which is difficult to determine before completing the training. Thus, in the following, we first give the energy models for edge devices, and then quantify the number of global communication rounds required for achieving a ϵ-global model convergence, i.e., satisfying $\frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}\left[\|\nabla f^r\|_2^2\right] \leq \epsilon$, via rigorous convergence analysis.

B. Energy Model

1) Communication energy model: If we consider the M-AirComp power control policy with transmission probability ρ whose value is smaller than $p_b^{\max}$, the threshold channel gain is mapped as $g_{\text{th}} := -\frac{1}{\lambda}\ln\rho$. In this way, the average power consumption among all the users and time slots will be:

$$\begin{aligned} P^{comm} =& \rho\varrho \int_{g_{\text{th}}}^{\infty} \lambda \frac{1}{x} e^{-\lambda x} \mathrm{d}x \\ =& -\rho\varrho\lambda \text{Ei}\left(-\lambda g_{\text{th}}\right) = -\rho\varrho\lambda \text{Ei}\left(\ln\rho\right), \end{aligned} \quad (9)$$

where $\text{Ei}(x)$ is the exponential integral function denoted as $\text{Ei}(x) = \int_{-\infty}^{x} \frac{e^t}{t}\mathrm{d}x$. Due to the fact that $-\ln\rho$ is positive, we have $\text{Ei}\left(\ln\rho\right) = -\text{E}_1\left(-\ln\rho\right)$ where $\text{E}_1(x) = \int_x^{\infty} \frac{e^{-t}}{t}\mathrm{d}x$. Then we have $P^{comm} = -\rho\varrho\lambda\text{Ei}\left(\ln\rho\right) = \rho\varrho\lambda\text{E}_1\left(-\ln\rho\right)$. For any x with positive real value, $\text{E}_1(x)$ can be tightly upper bounded by an elementary function as follows:

$$\text{E}_1(x) < e^{-x}\ln\left(1 + \frac{1}{x}\right). \quad (10)$$

We note that the gap between the original $\mathrm{E}_1(x)$ function and its bound is negligible, but the calculation of the function $\mathrm{E}_1(x)$ is much more complex than that of its bound due to the integral operator. Recall that we wish to reduce the transmission energy consumption that is proportional to the value of $\mathrm{E}_1(-\ln p_b)$. For ease of solution, we replace $\mathrm{E}_1(-\ln p_b)$ ($-\ln \rho \geq 0$ always holds) with its upper bound $e^{\ln p_b} \ln\left(1+\frac{1}{-\ln p_b}\right)$, implying that we minimize the transmission energy consumption for the worst case. Then we have

$$P^{comm} \approx \rho\varrho\lambda e^{\ln \rho} \ln\left(1+\frac{1}{-\ln \rho}\right) = \varrho\lambda\rho^2 \ln(1-\frac{1}{\ln \rho}). \tag{11}$$

After executing a fixed number of local iterations, each device is required to quantize the updated gradients into low-bit precision for digital transmission. Here, we adapt MASK to modulate the gradients, which means the magnitude of each symbol is sufficient to decode the transmission gradient. Let T_s denote the symbol duration that is in inverse proportion to channel bandwidth. To transmit the gradients with the size of d, $d/2$ symbol is required according to the M-AirComp design. Thus, the transmission time can be represented as $T^{comm} = \frac{d}{2M_s}T_s$, where M_s symbols are transmitted in parallel. Accordingly, the communication energy consumption for each device in each communication round is computed as

$$E^{comm} = P^{comm} \times T^{comm}. \tag{12}$$

2) Computational energy model: With massive data stored and processed by edge devices, on-device training can naturally be treated as computation-hungry tasks. Luckily, most modern smart devices are equipped with high-performance GPUs and can handle such heavy training tasks. This work considers the GPU computational energy model. We model the energy consumed to process a mini-batch of data in one iteration as

$$E^{comp} = P^{comp} \times T^{comp}, \tag{13}$$

where P^{comp} and T^{comp} are runtime power and execution time of the edge device, respectively. Both of them are related to the GPU core frequency/voltage and the memory frequency in the forms of [10]

$$P^{comp} = P^0 + af^{mem} + b(v^{core})^2 f^{core}, \tag{14}$$

$$T^{comp} = T^0 + \frac{u}{f^{mem}} + \frac{v}{f^{core}}. \tag{15}$$

Here, P_0 and T_0 are the static power and static time consumption. f^{core}/v^{core} and f^{mem} represent the core frequency/voltage and memory frequency, respectively. a, b, u, and v are constants reflecting the sensitivity of the task execution to GPU memory and core frequency/voltage scaling [10], [11]. Given a specific FL task, i.e., a neural network model with a dataset, these constants can be well estimated based on experiments by measuring the average runtime energy consumption. Since every user performs H local iterations between two consecutive communication rounds, the energy consumption of local computing in one communication round can be calculated as the product of the energy consumption of one iteration and the number of local iterations, i.e., $E^{comp} \cdot H$.

C. Impacts of Control Variables on ESOAFL Convergence

In this subsection, we theoretically analyze the impacts of control variables ρ and H on the convergence rate of ESOAFL. We consider the following three standard assumptions.

Assumption 1 (Smoothness)**.** *The objective function f_k is differentiable and L-smooth :*

$$\|\nabla f_k(\mathbf{x}) - \nabla f_k(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|, \forall k. \tag{16}$$

Assumption 2 (Bounded variances and second moments)**.** *The variance and the second moments of stochastic gradients evaluated with a mini-batch can be bounded as*

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_i} \left\|\nabla F_i\left(\mathbf{w};\xi_i\right) - \nabla f(\mathbf{w})\right\|^2 \leq \sigma^2, \forall \mathbf{w}, \forall i, \tag{17}$$

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_i} \left\|\nabla F_i\left(\mathbf{w};\xi_i\right)\right\|^2 \leq \delta^2, \forall \mathbf{w}, \forall i, \tag{18}$$

where σ and δ are positive constants.

Assumption 3 (Quantization bounded variances)**.** *The output of a q-quantization operator $Q(x)$ is an unbiased estimator of its input x, and its variance grows with the squared of L_2-norm of its argument, i.e., $\mathbb{E}[Q(x)] = x$ and $\mathbb{E}[||Q(x) - x||^2] \leq q||x||^2$, where the expectation $\mathbb{E}[\cdot]$ is taken over the randomness of Q. Here, q could be a function reflecting compression distortion w.r.t the dimension of the input and the number of quantization levels.*

Basically, Assumption 3 is customary in the analysis of distributed learning methods with compression [12], [13], and there are some quantization operators subjecting to the conditions in the assumption, such as QSGD [14], Stochastic Quantization [15], [16], etc. Based on the above assumptions, we have the following lemma on the bounded variances of M-AirComp, where the power control policy with a transmission probability ρ is applied for gradient uploading.

Lemma 1 (M-AirComp bounded variances)**.** *The output of the M-AirComp operator Air($\mathcal{X}$) with the proposed power control scheme is an unbiased estimator of its input set $\mathcal{X}$,i.e., $\mathbb{E}$[Air($\mathcal{X}$)] $= y$, and the transmission probability ρ affects the variance of M-AirComp by*

$$\mathit{Var}(\mathit{Air}(\mathcal{X})) = \frac{1}{K^2}(\frac{1}{\rho}-1)\sum_{x_k \in \mathcal{X}} x_k^2 + \frac{\sigma_z^2}{K^2\rho^2}. \tag{19}$$

Proof. Let $\mathcal{X}$ be the input set of the M-AirComp operator. Given the transmission probability ρ of the FL users, the expected output of M-AirComp is

$$\begin{aligned}\mathbb{E}[\mathrm{Air}(\mathcal{X})] =& \mathbb{E}\left[\frac{1}{\rho K}\left[\sum_{x_k \in \mathcal{X}} x_k + n\right]\right] \\ =& \frac{1}{\rho K}\left[\sum_{x_k \in \mathcal{X}} (x_k\rho + 0\cdot(1-\rho)) + \mathbb{E}\left[n\right]\right] = y,\end{aligned} \tag{20}$$

and the mean of the square of $\text{Air}(\mathcal{X})$ is given by

$$\begin{aligned}
&\mathbb{E}[(\text{Air}(\mathcal{X}))^2] \qquad (21)\\
=&\mathbb{E}\left[\frac{1}{\rho^2K^2}\left(\sum_{x_k\in\mathcal{X}}x_k+n\right)^2\right]\\
=&\mathbb{E}\left[\frac{1}{\rho^2K^2}\left(\sum_{x_k,x_{k'}\in\mathcal{X}}x_kx_{k'}+2\sum_{x_k\in\mathcal{X}}x_kn+n^2\right)\right]\\
=&\frac{1}{\rho^2K^2}\left[\sum_{x_k,x_{k'}\in\mathcal{X},k\neq k'}x_k\rho x_{k'}\rho+\sum_{x_k\in\mathcal{X}}x_k^2\rho+\sigma_z^2\right]\\
=&\frac{1}{\rho^2K^2}\left[\rho^2\left((\sum_{x_k\in\mathcal{X}}x_k)^2-\sum_{x_k\in\mathcal{X}}x_k^2\right)+\rho\sum_{x_k\in\mathcal{X}}x_k^2+\sigma_z^2\right]\\
=&\frac{1}{K^2}\left[(\sum_{x_k\in\mathcal{X}}x_k)^2+(\frac{1}{\rho}-1)\sum_{x_k\in\mathcal{X}}x_k^2\right]+\frac{\sigma_z^2}{K^2\rho^2} \qquad (22)
\end{aligned}$$

Thus, the variance is calculated as:

$$\begin{aligned}
\text{Var}(\text{Air}(\mathcal{X})) =&\mathbb{E}[(\text{Air}(\mathcal{X}))^2]-\mathbb{E}[\text{Air}^2(\mathcal{X})]\\
=&y^2+\frac{1}{K^2}(\frac{1}{\rho}-1)\sum_{x_k\in\mathcal{X}}x_k^2+\frac{\sigma_z^2}{K^2\rho^2}-y^2\\
=&\frac{1}{K^2}(\frac{1}{\rho}-1)\sum_{x_k\in\mathcal{X}}x_k^2+\frac{\sigma_z^2}{K^2\rho^2}.
\end{aligned} \qquad (23)$$

□

Theorem 1. *For the proposed ESOAFL approach, under the above assumptions, if learning rates θ and η satisfy*

$$1\geq L^2\eta^2H^2+HL\theta\eta\frac{q(2-\rho)+K\rho}{K\rho}, \qquad (24)$$

the convergence rate after R communication rounds can be bounded as:

$$\begin{aligned}
\frac{1}{R}\sum_{r=0}^{R-1}\|\nabla f^r\|_2^2\leq&\frac{2(f(\mathbf{w}^0)-f(\mathbf{w}^*)}{\eta\theta HR}+\frac{\eta\theta L}{K}\frac{(\rho+q)}{\rho}\sigma^2\\
&+\eta^2L^2H\sigma^2+\frac{\theta\eta L}{HK^2\rho^2}\sigma_z^2,
\end{aligned} \qquad (25)$$

where q is the gradient quantization precision, ρ is the M-AirComp transmission probability, H is the local computing iterations, and $f(\mathbf{w}^*)$ is the minimum value of the loss.

Proof. Please refer to the Appendix. A for the proof. □

The above Theorem 1 is derived based on the L-smoothness gradient assumption on global objective [12]. After expanding the inequality of the global objective, we first bound the inner product between the stochastic gradient and full batch gradient, while we can also bound the distance between the global model and the local model. Further, we bound the updated gradients with M-AirComp and quantization operators. Finally, by integrating the derived results above, we finish the convergence analysis of the ESOAFL algorithm.

Corollary 1. *To achieve the linear speedup, we need to have $\theta\eta=O\left(\frac{\sqrt{K}}{\sqrt{RH}}\right)$. If we further choose $\theta\eta=O\left(\frac{1}{L}\sqrt{\frac{K\rho}{RH(\rho+q)}}\right)$, the convergence rate can be represented as:*

$$\begin{aligned}
\frac{1}{R}\sum_{r=0}^{R-1}\|\nabla f^r\|_2^2\leq&\frac{2L(f(\mathbf{w}^0)-f(\mathbf{w}^*)\sqrt{(\rho+q)}}{\sqrt{KRH\rho}}+\\
&\frac{\sqrt{\rho+q}}{\sqrt{KRH\rho}}\sigma^2+\frac{K}{R\theta^2}\sigma^2+\sqrt{\frac{1}{K^3RH^3(\rho+q)\rho^3}}\sigma_z^2\\
\overset{(a)}{=}&O\left(\frac{\sqrt{\rho+q}}{\sqrt{KRH\rho}}(2L(f(\mathbf{w}^0)-f(\mathbf{w}^*)+\sigma^2))+\frac{K}{R\theta^2}\sigma^2\right)\\
\overset{(b)}{=}&O\left(\frac{\chi}{\sqrt{KRH}}\right)+O\left(\frac{K}{R}\right),
\end{aligned} \qquad (26)$$

where (a) is due to the fact that $O(\sqrt{\frac{1}{K^3R}})$ decays faster than $O(\sqrt{\frac{1}{KR}})$, and we replace $\sqrt{\frac{\rho+q}{\rho}}$ by χ in (b).

We note that, for the ESOAFL without probabilistic transmission, i.e., $\rho=1$, the bound in Eq. (26) matches the best-known rate given by [12] with a tight convergence analysis. This implies that our ESOAFL will retain the same linear speedup property as its counterpart without probabilistic transmission and M-AirComp operation. Based on the convergence analysis, we further give the following corollary on the communication complexity, i.e., the number of communication rounds required for achieving convergence, of our ESOAFL algorithm.

Corollary 2. *From the Corollary 1, the required maximum number of communications for achieving the ϵ target training loss, i.e., satisfying $\epsilon=\frac{1}{R}\sum_{r=0}^{R-1}\|\nabla f^r\|_2^2$, is given by*

$$\begin{aligned}
R=&O\left(\frac{2\epsilon\sigma^2HK^2+\chi^2(\delta+\sigma^2)^2\theta^2}{2\epsilon^2\theta^2HK}\right) \qquad (27)\\
&+O\left(\frac{+\chi(\delta+\sigma^2)\theta\sqrt{4\epsilon\sigma^2HK^2+\chi^2(\delta+\sigma^2)^2\theta^2}}{2\epsilon^2\theta^2HK}\right)\\
=&O(K)+O\left(\frac{\chi^2}{HK}\right)+O\left(\frac{\chi}{\sqrt{H}}\right), \qquad (28)
\end{aligned}$$

where $\chi=\sqrt{\frac{\rho+q}{\rho}}$ and $\delta=2L(f(\mathbf{w}^0)-f(\mathbf{w}^*))$.

D. Overall Energy Minimization Reformulation and Solution

With the above models, we calculate the total energy consumed by the participating mobile devices during the entire training process as:

$$\begin{aligned}
\Theta(\rho,H)=&R\times(E^{comm}+HE^{comp})\\
=&\left(\frac{A_0(\rho+q)}{\rho H}+\frac{B_0\sqrt{\rho+q}}{\sqrt{\rho H}}+C_0\right)\\
&\cdot\left(\varrho\lambda\rho^2\ln(1-\frac{1}{\ln\rho})T^{comm}+HE^{comp}\right),
\end{aligned} \qquad (29)$$

where A_0, B_0, and C_0 are constants used to approximate the big-O notion in Eq. (27). From the above formula, we observe that a larger H lead to the reduced number of communication rounds R ("talking"), but increases the computational energy consumption per round ("working"). Also, adjusting ρ affects

Algorithm 2 JCP Control Algorithm

Initialization: $\xi = 10^{-5}$, $\iota = 10^{-5}$, $\gamma^0 \in (0, 1]$, $\kappa = 0$.
Input: Parameters p_b^{max}, H_{min} and H_{max}; Value access to function $\Theta(\cdot)$.

repeat
  Solve (34) and set the optimal value as $\boldsymbol{\phi}^*(\boldsymbol{\phi}^\kappa)$
  Set $\boldsymbol{\phi}^{\kappa+1} = \boldsymbol{\phi}^\kappa + \gamma^0(\boldsymbol{\phi}^*(\boldsymbol{\phi}^\kappa) - \boldsymbol{\phi}^\kappa)$
  Set $\kappa = \kappa + 1$ and $\gamma^\kappa = \gamma^{\kappa-1}(1 - \xi\gamma^{\kappa-1})$
until $||\boldsymbol{\phi}^\kappa - \boldsymbol{\phi}^{\kappa-1}||_2^2 \leq \iota$
Round the current H to the nearest integer in $\mathcal{H}$
return The current solutions of ρ and H.

the required communication rounds and the communication energy consumption in each round. Thus, it is necessary to optimize H and ρ to balance "working" and "talking" for minimizing the overall energy consumption. To this end, we formulate the Joint local Computing and transmission Probability (JCP) control problem as:

$$\min_{\rho,H} \quad \left(\frac{A_0(\rho+q)}{\rho H} + \frac{B_0\sqrt{\rho+q}}{\sqrt{\rho H}} + C_0\right) \tag{30a}$$

$$\cdot \left(\varrho\lambda\rho^2 \ln(1 - \frac{1}{\ln\rho})T^{comm} + HE^{comp}\right) \tag{30b}$$

$$s.t. \quad 0 < \rho \leq p_b^{max}, \tag{30c}$$

$$H \in \mathcal{H}. \tag{30d}$$

For notational brevity, we define $\boldsymbol{\phi} = \{\rho, H\}$ and represent the objective function as $\Theta(\boldsymbol{\phi}) = \Theta_1(\boldsymbol{\phi}) \times \Theta_2(\boldsymbol{\phi})$, where

$$\Theta_1(\boldsymbol{\phi}) = \frac{A_0(\rho+q)}{\rho H} + \frac{B_0\sqrt{\rho+q}}{\sqrt{\rho H}} + C_0, \tag{31}$$

$$\Theta_2(\boldsymbol{\phi}) = \varrho\lambda\rho^2 \ln(1 - \frac{1}{\ln\rho})T^{comm} + HE^{comp}. \tag{32}$$

Noticing the decoupled constraints in (30c-30d), we relax the constraint in (30d) as $H_{min} \leq H \leq H_{max}$, where H_{min} and H_{max} are the minimum and the maximum integer in $\mathcal{H}$, respectively. Moreover, we can identify that both function $\Theta_1(\boldsymbol{\phi})$ and $\Theta_2(\boldsymbol{\phi})$ are positive and convex after calculating the first and second-order partial derivative of these two functions. (Please refer to Appendix. B for the detailed derivation.)

Capturing such the "product-of-convexity" property of the objective function $\Theta(\boldsymbol{\phi})$, we use the inner convex approximation method [17] to solve the relaxed JCP control problem by optimizing a sequence of strongly convex inner approximations of $\Theta(\boldsymbol{\phi})$ in the form: given $\boldsymbol{\phi}^\kappa$

$$\Theta(\boldsymbol{\phi}; \boldsymbol{\phi}^\kappa) = \Theta_1(\boldsymbol{\phi})\Theta_2(\boldsymbol{\phi}^\kappa) + \Theta_1(\boldsymbol{\phi}^\kappa)\Theta_2(\boldsymbol{\phi}), \tag{33}$$

where $\boldsymbol{\phi}^\kappa = \{\rho^\kappa, H^\kappa\}$ refers to the intermediate $\boldsymbol{\phi}$ obtained in the κ-th iteration. Obviously, the approximated objective function in (33) is strongly convex with the fixed $\boldsymbol{\phi}^\kappa$. With the surrogate function above, we are essentially required to compute the optimal solutions of the following convex optimization problem in each iteration, while preserving the feasibility of the iterates to the original problem in (30).

$$\min_{\rho,H} \quad \Theta(\boldsymbol{\phi}; \boldsymbol{\phi}^\kappa) \tag{34a}$$

$$s.t. \quad 0 < \rho \leq p_b^{max}, \tag{34b}$$

$$H_{min} \leq H \leq H_{max}. \tag{34c}$$

Notice that the problem (34) can be solved by various commercial solvers, e.g., IBM CPLEX optimizer [18]. The formal description of the Joint Power and Aggregation Control Algorithm is presented in Alg. 2. Starting from a feasible point $\boldsymbol{\phi}^0$, the method consists in iteratively computing the solution $\boldsymbol{\phi}^*(\boldsymbol{\phi}^\kappa)$ to the surrogate problem (34), and then taking a step from $\boldsymbol{\phi}^\kappa$ towards $\boldsymbol{\phi}^*(\boldsymbol{\phi}^\kappa)$. Here, instead of using a constant step-size, we use a diminishing step-size rule, i.e, $\gamma^\kappa = \gamma^{\kappa-1}(1 - \xi\gamma^{\kappa-1})$, as it is more efficient to control the iteration complexity and the convergence speed in practice [17]. The process is repeated until it meets the termination criterion, and the value of H is rounded afterward to ensure its feasibility, i.e., $H \in \mathcal{H}$.

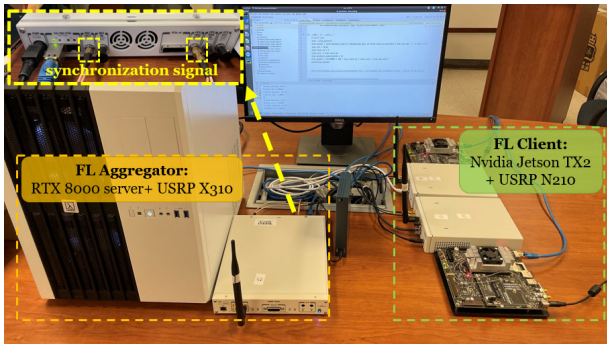


Fig. 3. M-AirComp based FL testbed.

V. Performance Evaluation

A. Implementation of M-AirComp

As shown in Fig. 3, we first set up experiments to elaborate on the usage of M-AirComp for an FL testbed. The system consists of one edge server and two edge devices. We let one RTX-8000 server with one USRP X310 play the role of the over-the-air FL aggregator. Each FL client consists of the NVIDIA Jetson TX2 as the computing unit and USRP N210 as the wireless transmitter. We also use WBX 50-2200 MHz Rx/Tx USRP daughterboards, with up to 200mW output power. The synchronization is provided by USRP X310 REF and PPS output ports through cable connection. In the end, all the USRPs are connected to an internet switch. We run MATLAB codes from the Communication Toolbox Support Package for USRP Radio to control the transmitting and receiving in different sessions on the RTX-8000 server.

We first verify the feasibility of M-AirComp by the in-lab experiments, where two edge devices transmit QAM symbols with quantization, e.g., 16 QAM for 4-bit quantization. From the constellation in Fig. 4, the receiving symbol set is expanded into a constellation for higher-order modulations,

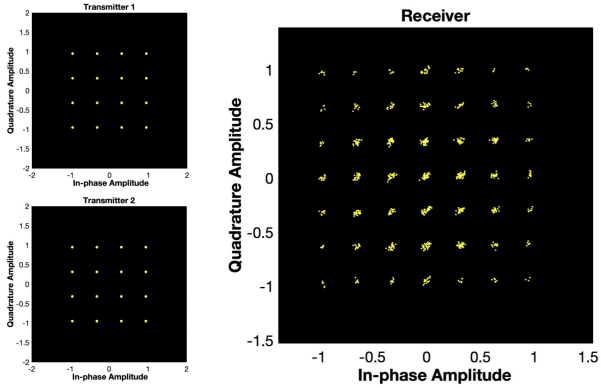


Fig. 4. Constellation diagram of M-AirComp demo (left: transmitter; right: receiver).

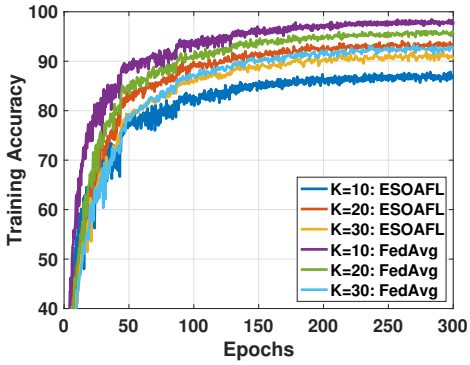


Fig. 5. Training performance under poor channel conditions.

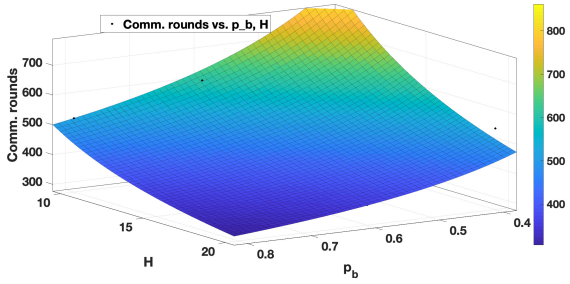


Fig. 6. Communication rounds with varying p_b and H.

which explains the addition carried by the over-the-air computation from the communication point of view. The aggregated symbol will be further decoded as a quantized model update, with a certain probability of bit error with regards to the signal-to-noise ratio (SNR).

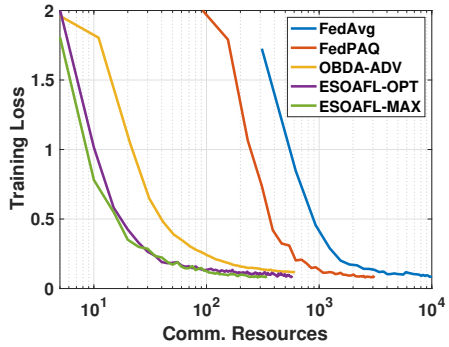
B. Some Observations of the ESOAFL

As we have discussed in Sec. II-C, AirComp can dramatically improve the spectrum efficiency in the FL training process. Particularly, our ESOAFL scheme has great potential to retain the training performance in the case of many participating devices, even if the communication environment (i.e., channel condition) is extremely poor. In Fig. 5, we consider a severe communication environment with $SNR = 5dB$ over different number of FL participants, i.e., $K = 10$, 20, and 30, and train the ResNet-20 model with the CIFAR-10 dataset. Here, we partition the dataset into several sub-datasets, where one sub-dataset is for one FL participant. As a result, the size of the local dataset decreases as the number of participants K grows. This results in the degrading performance of FedAvg with increasing K. Unlike such the monotone impact in FedAvg, the impacts of K on the performance of ESOAFL could be more complicated. With the increasing number of participants, the variance of AirComp is decreasing, as indicated in Eq. (23), which helps improve the performance of our ESOAFL. This can be validated by the shrinking gap between the ESOAFL scheme and its ideal case (i.e., FedAvg without channel noise) as the number of devices grows in Fig. 5. Especially with a large set of participants (e.g., $K = 30$), the performance of ESOAFL is very close to that of FedAvg, which also implies that our ESOAFL scheme has strong ability to resist on the poor channel condition.

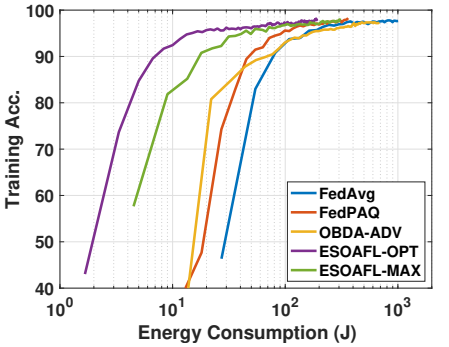
Recall that the parameters A_0, B_0, and C_0 exist in the JCP control problem, which are related to the specific learning model and dataset. Here, we use a sampling-based method to estimate the values of the constants A_0, B_0, and C_0, where we empirically sample different combinations (ρ,H) and use the derived convergence bound in Eq. (27) to infer their values. In specific, we repeatedly train a ResNet-20 model on the CIFAR-10 dataset using varying local computing iterations H and transmission probabilities ρ, where we set a fixed target training loss and record the number of communication rounds correspondingly. With these experimental results, we use the non-linear least squares curve fitting algorithm [19] to estimate the values of A_0, B_0, and C_0 in the JCP control problem. Fig. 6 shows the fitting results. We observe that, with the increase of local computing iterations H and transmission probability ρ, the number of required communication rounds is decreasing, but this effect is gradually weakened. At the same time, the computing energy consumption of each round increases linearly with the incremental of local computing iterations H. Thus, the trade-off between local computing and wireless communications has to be considered to reduce the overall energy consumption, where local iterations H and transmission probability ρ are necessary to carefully determine.

C. Spectrum and Energy Efficiency of the ESOAFL

After the parameter estimation, we implement the proposed JCP control scheme to find the optimal local computing iterations H and transmission probability ρ. Here, we use two different image classification datasets, i.e., MNIST and CIFAR-10, to verify the effectiveness of our proposed approach, both of which consist of 50000 training images and 10000 test images in 10 classes. In particular, the MNIST dataset contains 28x28 black and white images of handwritten digits, while the CIFAR-10 dataset is rather complicated that contains 32x32 color images of animals and vehicles. A LeNet model and a ResNet-20 model are trained on the two datasets respectively, where the former is light and the latter has a more complex structure to fit the dataset. We set the batch

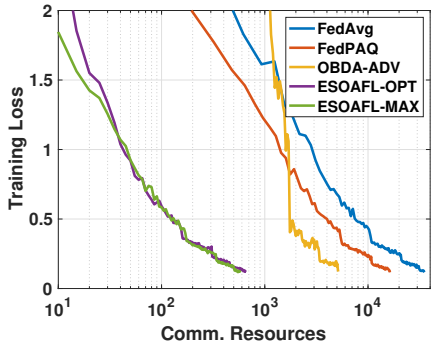


(a) Training loss vs. comm. resources

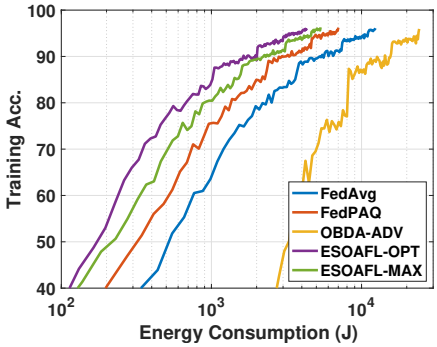


(b) Training acc. vs. energy cons.

Fig. 7. Training performance of LeNet on MNIST dataset.



(a) Training loss vs. comm. resources



(b) Training acc. vs. energy cons.

Fig. 8. Training performance of ResNet-20 on CIFAR-10 dataset.

size as 128 for ResNet-20 and 32 for LeNet. In each round of FL, we set $K = 10$ participating devices to execute H iterations of stochastic gradient descent (SGD) in parallel, and the maximum transmission probability $\rho^{\max}$ is set to 0.77 according to the simulated communication environment and the power constraint. The initial learning rate is $\eta = 0.2$ with a fixed decay rate. Particularly, we compare our ESOAFL-OPT (i.e., ESOAFL with optimal JCP control) with the following schemes:

- FedAvg [1]: FL without AirComp and gradient quantization, where ideal noise-free transmission is supposed.
- FedPAQ [20]: FL with gradient quantization, where the users transmit the quantized model updates in every communication round.
- OBDA-ADV [21]: A modified version of the OBDA (one-bit digital AirComp), where we improve the original scheme by ignoring the quantization at the receiver to preserve the learning precision.
- ESOAFL-MAX: the proposed ESOAFL scheme without the transmission control, where we adopt the maximum transmission probability $\rho^{\max}$ to transmit the model updates.

We assume the same communication bandwidth for all the schemes. We consider the Nvidia TX2 as FL device and utilize the Jtop [22] tool to measure the computing energy. It measures that the LeNet model consumes 0.03J and the ResNet model consumes 0.5J for one training iteration. For example, training the ResNet model for one iteration takes 130ms, and the GPU power is nearly 4W. We assume the AirComp is deployed in the commercial LTE system for wireless transmissions. For all the schemes, we set the maximum transmit power as 0.2W and set the average SNR as 15dB for the FL participants, whose channel quality can be reflected by the CQI (Channel Quality Indicator) category 11. In this case, the modulation scheme, code rate, bits per resource element are 64QAM, 0.8525, 5.115, respectively.

Fig. 7(a) and Fig. 7(b) show the performance of training a LeNet model on the MNIST dataset. Here, we set the target training loss ϵ as 0.07 and assume the data samples are independent and identically distributed (IID). The local computing iteration and transmission probability used in ESOAFL-OPT are with the values of $H = 3$ and $\rho = 0.29$ respectively, which are obtained by performing the JCP control algorithm in Alg. 2. Here, we integrate the local SGD method (i.e., taking several training steps among the sequential communication rounds) into OBDA-ADV scheme for a fair comparison. Let the spectrum resource consumed in each round of ESOAFL be a unit communication resource. We set the gradient quantization level as 4-bit in ESOAFL and FedPAQ. Fig. 7(a) illustrates the communication resources consumption during the training procedure, and we can obviously find that the proposed ESOAFL significantly improves the spectrum efficiency compared with FedAvg and FedPAQ. This is because that the FL devices in FedAvg and FedPAQ cannot take the concurrent transmission with the same bandwidth as ESOAFL does. Besides, ESOAFL allows each pair of gradients to be transmitted orthogonally using in-phase (I) and quadrature (Q) channels simultaneously, while FedAvg and FedPAQ only

TABLE II
PERFORMANCE COMPARISON UNDER DIFFERENT LEARNING SETTINGS (RESNET20 ON CIFAR-10)

		IID			**Non-IID:** $\varsigma = 0.3$			**Non-IID:** $\varsigma = 0.5$			**Non-IID:** $\varsigma = 0.8$		
		Comm.	**Energy**	**Acc.**	**Comm.**	**Energy**	**Acc.**	**Comm.**	**Energy**	**Acc.**	**Comm.**	**Energy**	**Acc.**
K=10	**FedAvg**	35030	12379	88.1%	37820	13365	87.9%	42470	15008	85.1%	53010	18951	68.4%
B=128	**FedPAQ**	16320	7011	88.1%	17632	7550	87.4%	22080	9471	85.1%	27040	11600	68.6%
H=10	**ESOAFL**	**656**	**4323**	87.4%	**674**	**4590**	87.1%	**693**	**4692**	84.8%	**861**	**5848**	68.1%
K=100	**FedAvg**	350300	9977	87.7%	418500	11920	86.7%	461900	13156	81.0%	492900	14039	63.1%
B=32	**FedPAQ**	187200	5544	87.3%	214400	6349	86.6%	238400	7060	81.0%	254400	7535	58.9%
H=5	**ESOAFL**	**600**	**1530**	87.1%	**675**	**1721**	86.4%	**740**	**1887**	81.0%	**785**	**785**	53.2%

allow each resource element to carry several bits of a gradient for fitting in with the LTE protocol. Since OBDA-ADV applies one-bit digital AirComp, the precision of the model updates can be seriously scarified in every communication round. Due to such information distortion, it is required to take more communication rounds to achieve a specific accuracy, and thereby consumes more communication resources than our high-precision AirComp FL scheme during training, as shown in Fig. 7(a). Fig. 7(b) further illustrates the behaviours of energy consumption during FL. The results show that our ESOAFL scheme consumes the least energy among all the schemes. Specifically, when achieving the same target training loss, the energy consumption of FL devices in ESOAFL-OPT is twice and three times lower than that of FedPAQ and OBDA-ADV, respectively. This is because the energy efficient power control policy and the digital modulation scheme in the M-AirComp design save both the transmit power and time. Moreover, since the optimized transmission probability is much lower than the maximum value, our ESOAFL-OPT approach only consumes nearly half of the ESOAFL-MAX approach's energy, which demonstrates the necessity of the JCP control. Note that the low-precision OBDA-ADV approach cannot reach the target training loss we set, and thus we consider the target loss $\epsilon = 0.12$ especially for the OBDA-ADV approach to present the results.

Fig. 8(a) and Fig. 8(b) demonstrate the performance comparison of all the schemes using ResNet-20 model on the CIFAR-10 dataset. We set the target training loss ϵ as 0.12, and obtain the optimal control strategies $H = 11$ and $\rho = 0.51$ for ESOAFL-OPT. As expected, the proposed ESOAFL approach dramatically improves the spectrum efficiency and reduces the energy consumption of devices. Particularly, ESOAFL-OPT saves hundreds of times of communication resources compared with FedAvg and FedPAQ in this case. It also saves more than $8\times$ of communication resources compared with the OBDA method. Besides, our proposed ESOAFL-OPT scheme saves nearly one-third and two-thirds of energy consumption than FedPAQ and FedAvg schemes. Notice that the OBDA-ADV scheme has relatively poor convergence performance compared with other approaches due to the high precision requirement of the complex ResNet-20 model.

We further show the scalability of the ESOAFL scheme with more learning settings. Here, we consider different data distributions in the content of different levels of non-IID data. Let $\varsigma \in [0, 1]$ denotes the non-IID level [23]. For example, $\varsigma = 0.3$ indicates that 30% of the data belong to one label and the remaining 70% data belong to others. Following this setting, we generate the local dataset for each user by drawing the data from the whole dataset with specific labels, instead of evenly partition the dataset with all the labels. We ignore the OBDA-ADV scheme since its performance is not good in non-IID data settings. From Table. II, we can observe that training with non-IID data incurs a larger energy and communication resources consumption to converge. Despite all this, our ESOAFL, compared with FedAvg and FedPAQ, achieves the indistinguishable testing accuracy at all non-IID levels while saving communication resources and overall energy consumption. We also conducted the simulations with $K = 100$ participants, obtaining the similar observations. Here, compared with $K = 10$ participants settings, we put less computing loads ($B = 32$, $H = 5$) in each communication round of the $K = 100$ setting, thus causing more communication loads. Therefore, the communication resources consumption of FedAvg and FedPAQ at $K = 100$ increases significantly compared with the scenario of $K = 10$. Benefiting from concurrent transmission, ESOAFL does not introduce extra communication resource consumption as K increases, revealing its significant potentials for involving massive FL participants.

VI. RELATED WORKS

Much attention has recently been paid to improve the energy efficiency of wireless FL over mobile devices via integrating various advanced techniques [24]. For saving the energy consumed for communication, gradient sparsification [25]–[27] and gradient quantization [14], [15] techniques are used to compress the model updates and thereby reduce the transmission load in every FL round. In [28], [29], momentum GD/SGD methods are adopted to accelerate the convergence where the involved communications and energy consumption during training can be reduced accordingly. For saving the energy consumed for local training, some researchers propose to quantize the model parameters into low bit-width at edge devices to facilitate computationally-efficient on-device training [30], [31]. Despite their benefits in improve the energy efficiency of FL, these methods are mainly considered from the perspective of learning algorithms and widely ignore the wireless communication environments, especially with the physical-layer aspects of communication [32].

By exploiting the waveform superposition property of the wireless medium [33], some pioneering works propose the AirComp FL to enable a large number of simultaneous local model uploading for improving the spectrum efficiency during FL [34]. Cao et al. in [35], and Amiri and Gündüz in [36] apply

AirComp to mitigate the communication bottleneck when a large number of participants aggregate the data together, where power allocation schemes are derived to satisfy the mean square error requirements. The authors in [5], [6] propose a joint device selection and communication scheme to improve the learning performance for AirComp FL. Some works further utilize the reconfigurable intelligent surface (RIS) technology to mitigate the communication bottleneck and relieve the straggler issue in FL by reconfiguring the wireless propagation environment [37]–[39]. All these works take analogy modulation schemes for wireless transmission, which are difficult to be implemented on commercial devices. In addition, the convergence analysis for the whole FL training procedure is rarely discussed in existing works. Noticing the limitations above, Zhu et al. [21] applies the 1-bit digital modulation and derives the convergence analysis accordingly. However, the 1-bit based scheme could seriously scarify the precision, and the energy consumption issue is overlooked in designing the scheme. Different from the existing approaches, our design targets at facilitating the general multi-bit digital modulation scheme, where a convergence-guaranteed FL scheme integrating both the AirComp and the gradient quantization is proposed to improve the energy and spectrum efficiency simultaneously.

VII. Conclusion and Future Work

In this paper, we proposed the ESOAFL scheme for energy and spectrum efficient FL over mobile devices, where M-AirComp was applied for model updates transmission in a joint compute-and-communicate manner. A high-precision digital modulation scheme with multi-bit gradient quantization was designed for the participating devices to upload their model updates during FL. With the theoretical convergence analysis of the modified FL algorithm, we further developed a joint local computing and transmission probability control approach aiming to minimize the overall energy consumed by all devices. Extensive simulations were conducted to verify our theoretical analysis, and the results showed that the ESOAFL scheme effectively improves the spectrum efficiency with the learning precision guarantee. Besides, it also saved at least half of energy consumption compared with other FL schemes. We hope our analysis will promote future endeavors in improving the energy and spectrum efficiency of FL. For example, non-orthogonal multiple access and reconfigurable intelligent surface techniques can be integrated into an M-AirComp FL framework, which may improve the FL performance when facing massive connectivity and unfavorable propagation channel.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, April 2017.

[2] S. T. Apple, "Hey siri: An on-device dnn-powered voice trigger for apple's personal assistant," https://machinelearning.apple.com/research/hey-siri, accessed May, 2021.

[3] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi, "Federated learning of predictive models from federated electronic health records," *International journal of medical informatics*, vol. 112, pp. 59–67, April 2018.

[4] M. Pan, C. Zhang, P. Li, and Y. Fang, "Joint routing and link scheduling for cognitive radio networks under uncertain spectrum supply," in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, Shanghai, China, April 2011.

[5] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2022–2035, 2020.

[6] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Joint optimization of communications and federated learning over the air," *arXiv:2104.03490*, April 2021.

[7] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3gpp device-to-device proximity services," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 40–48, 2014.

[8] M. Sriharsha, S. Dama, and K. Kuchi, "A complete cell search and synchronization in lte," *EURASIP Journal on Wireless Communications and Networking*, vol. 111, no. 1, pp. 1–14, March 2017.

[9] O. Abari, H. Rahul, D. Katabi, and M. Pant, "Airshare: Distributed coherent transmission made seamless," in *IEEE Conference on Computer Communications (INFOCOM)*, Hong Kong, April 2015.

[10] X. Mei, X. Chu, H. Liu, Y.-W. Leung, and Z. Li, "Energy efficient real-time task scheduling on cpu-gpu hybrid clusters," in *Proc. of IEEE Conference on Computer Communications (INFOCOM)*, Atlanta, GA, May 2017.

[11] Y. Abe, H. Sasaki, S. Kato, K. Inoue, M. Edahiro, and M. Peres, "Power and performance characterization and modeling of gpu-accelerated systems," in *2014 IEEE 28th international parallel and distributed processing symposium*, Phoenix, AZ, August 2014.

[12] F. Haddadpour, M. M. Kamani, A. Mokhtari, and M. Mahdavi, "Federated learning with compression: Unified analysis and sharp guarantees," in *Proc. International Conference on Artificial Intelligence and Statistics*. Virtual Conference: PMLR, April 2021, pp. 2350–2358.

[13] D. Basu, D. Data, C. Karakus, and S. Diggavi, "Qsparse-local-sgd: Distributed sgd with quantization, sparsification and local computations," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2019.

[14] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient SGD via gradient quantization and encoding," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, Long Beach, CA, December 2017.

[15] H. Tang, S. Gan, C. Zhang, T. Zhang, and J. Liu, "Communication compression for decentralized training," in *Proc. of Advances in Neural Information Processing Systems*, Montréal, Canada, December 2018.

[16] A. T. Suresh, X. Y. Felix, S. Kumar, and H. B. McMahan, "Distributed mean estimation with limited communication," in *Proc. of International conference on machine learning (ICML)*, Sydney, Australia, August 2017.

[17] G. Scutari, F. Facchinei, and L. Lampariello, "Parallel and distributed methods for constrained nonconvex optimization—part i: Theory," *IEEE Transactions on Signal Processing*, vol. 65, no. 8, pp. 1929–1944, 2016.

[18] IBM, "Ibm cplex optimizer," https://www.ibm.com/analytics/cplex-optimizer, accessed April 4, 2021.

[19] C. L. Lawson and R. J. Hanson, *Solving least squares problems*. SIAM, 1995.

[20] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," in *Proc. International Conference on Artificial Intelligence and Statistics*, Virtual Conference, August 2020.

[21] G. Zhu, Y. Du, D. Gündüz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 2120–2135, 2020.

[22] R. Bonghi, "Jetson stats," https://github.com/rbonghi/jetson_stats, accessed March, 2021.

[23] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing federated learning on non-iid data with reinforcement learning," in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, Toronto, Canada, July 2020.

[24] D. Shi, L. Li, R. Chen, P. Prakash, M. Pan, and Y. Fang, "Toward energy-efficient federated learning over 5g+ mobile devices," *IEEE Wireless Communications*, vol. 29, no. 5, pp. 44–51, 2022.

[25] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified sgd with memory," in *Proc. of Advances in Neural Information Processing Systems*, Montréal, Canada, December 2018.

[26] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, "The convergence of sparsified gradient methods," in

Proc. of Advances in Neural Information Processing Systems, Montréal, Canada, December 2018.

[27] L. Li, D. Shi, R. Hou, H. Li, M. Pan, and Z. Han, "To talk or to work: Flexible communication compression for energy efficient federated learning over heterogeneous mobile edge devices," in *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, Virtual Conference, May 2021.

[28] S. Zheng, Z. Huang, and J. Kwok, "Communication-efficient distributed blockwise momentum SGD with error-feedback," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2019.

[29] W. Liu, L. Chen, Y. Chen, and W. Zhang, "Accelerating federated learning via momentum gradient descent," *IEEE Transactions on Parallel & Distributed Systems*, vol. 31, no. 08, pp. 1754–1766, 2020.

[30] R. Chen, L. Li, K. Xue, C. Zhang, M. Pan, and Y. Fang, "Energy efficient federated learning over heterogeneous mobile devices via joint design of weight quantization and wireless transmission," *IEEE Transactions on Mobile Computing*, 2022.

[31] F. Fu, Y. Hu, Y. He, J. Jiang, Y. Shao, C. Zhang, and B. Cui, "Don't waste your bits! squeeze activations and gradients for deep neural networks via tinyscript," in *Proc. International Conference on Machine Learning*. Virtual Conference: PMLR, July 2020, pp. 3304–3314.

[32] F. Ang, L. Chen, N. Zhao, Y. Chen, W. Wang, and F. R. Yu, "Robust federated learning with noisy communication," *IEEE Transactions on Communications*, vol. 68, no. 6, pp. 3452–3464, 2020.

[33] L. Chen, N. Zhao, Y. Chen, F. R. Yu, and G. Wei, "Over-the-air computation for iot networks: Computing multiple functions with antenna arrays," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 5296–5306, 2018.

[34] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 491–506, October 2020.

[35] X. Cao, G. Zhu, J. Xu, and K. Huang, "Optimized power control for over-the-air computation in fading channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7498–7513, August 2020.

[36] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3546–3557, February 2020.

[37] H. Liu, X. Yuan, and Y.-J. A. Zhang, "Reconfigurable intelligent surface enabled federated learning: A unified communication-learning design approach," *IEEE Transactions on Wireless Communications*, vol. 20, no. 11, pp. 7595–7609, 2021.

[38] W. Ni, Y. Liu, Z. Yang, and H. Tian, "Over-the-air federated learning and non-orthogonal multiple access unified by reconfigurable intelligent surface," in *Proc. of IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Virtual Conference, May 2021.

[39] J. Zheng, H. Tian, W. Ni, W. Ni, and P. Zhang, "Balancing accuracy and integrity for reconfigurable intelligent surface-aided over-the-air federated learning," *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 10 964–10 980, 2022.

APPENDIX

A. Proof of Theorem 1

We consider a non-convex FL model setting. From the L-smoothness gradient assumption on global objective f, we have

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}_Q\left[\mathbb{E}_{\text{Air}}\left[f(\mathbf{w}^{r+1})-f(\mathbf{w}^r)\right]\right]\right]\\ \leq&-\theta\eta\mathbb{E}\left[\mathbb{E}_Q\left[\mathbb{E}_{\text{Air}}\left[\left\langle\nabla f^r,\nabla F_Q^r\right\rangle\right]\right]\right]\\ &+\frac{\theta^2\eta^2L}{2}\mathbb{E}\left[\mathbb{E}_Q\left[\mathbb{E}_{\text{Air}}\left[\|\nabla F_Q^r\|^2\right]\right]\right],\end{aligned}\tag{35}$$

where we take the expectation over the sampling and operations.

Next, we give three important lemmas where the first two are borrowed from [12] and the last one is proved in the following.

Lemma 2. *The inner product between the stochastic gradient ∇F_Q^r and full batch gradient ∇f^r can be bounded by*

$$\begin{aligned}&\mathbb{E}_{\xi^{(r)}}\mathbb{E}_Q\mathbb{E}_{Air}\left[\left\langle\nabla f^r,\nabla F_Q^r\right\rangle\right]\\ =&\mathbb{E}_{\xi^{(r)}}\left[\left\langle\nabla f^r,\frac{1}{K}\sum_{k=1}^{K}\sum_{h=0}^{H-1}\nabla F_k^{r,h}\right\rangle\right]\\ \leq&\frac{1}{2K}\sum_{k=1}^{K}\sum_{h=0}^{H}\left[-\|\nabla f^r\|_2^2-\|\nabla f_k^{r,h}\|_2^2+L^2\|\mathbf{w}^r-\mathbf{w}_k^{r,h}\|_2^2\right].\end{aligned}\tag{36}$$

Here, we set $\nabla F_k^r = \sum_{h=0}^{H-1}\nabla F_k^{r,h}$ and $\nabla F_{k,Q}^r = Q\left(\sum_{h=0}^{H-1}\nabla F_k^{(h,r)}\right)$. We further define $\nabla F_Q^r = Air_{\mathcal{K}}\left(Q\left(\sum_{h=0}^{H-1}\nabla F_k^{r,h}\right)\right)$.

Lemma 3. *Under Assumption 2, the distance between the global model and the local model at r-th communication round can be bounded by*

$$\mathbb{E}\left[\|\mathbf{w}^r-\mathbf{w}_k^{r,h}\|_2^2\right]\leq\eta^2H\sigma^2+\eta^2\sum_{h=0}^{H-1}H\|\nabla f_k^{r,h}\|_2^2\tag{37}$$

Lemma 4. *The last term in (35) can be calculated as*

$$\begin{aligned}&\mathbb{E}_{\xi^{(r)}}\mathbb{E}_Q\mathbb{E}_{Air}\left[\|Air_{\mathcal{K}}\left(Q\left(\sum_{h=0}^{H-1}\nabla F_k^{r,h}\right)\right)\|^2\right]\leq\frac{\sigma_z^2}{K^2\rho^2}\\ &+\sum_{k=1}^{K}\frac{q+\rho}{K^2\rho}\operatorname{Var}(\nabla F_k^r)+\sum_{k=1}^{K}\frac{q(2-\rho)+K\rho}{K^2\rho}\|\nabla f_k^r\|^2\end{aligned}\tag{38}$$

Proof. Applying Lemma 1 into the left-hand-side of(38), we get (39). Then we complete the proof of Lemma 4

□

From Assumption 2, we have $\operatorname{Var}(\nabla F_k^r)\leq H\sigma^2$. Further, we have $\|\nabla f_k^r\|^2=\|\sum_{h=0}^{H-1}\nabla f_k^{r,h}\|^2\leq H\sum_{h=0}^{H-1}\|\nabla f_k^{r,h}\|^2$ and

$$\begin{aligned}&\mathbb{E}_{\xi^{(r)}}\mathbb{E}_Q\mathbb{E}_{\text{Air}}\left[\|\text{Air}_{\mathcal{K}}\left(Q\left(\sum_{h=0}^{H-1}\nabla F_k^{r,h}\right)\right)\|^2\right]\\ \leq&\frac{q+\rho}{K\rho}H\sigma^2+H\frac{q(2-\rho)+K\rho}{K^2\rho}\sum_{k=1}^{K}\sum_{h=0}^{H}\|\nabla f_k^{r,h}\|^2+\frac{\sigma_z^2}{K^2\rho^2}\end{aligned}\tag{40}$$

Applying Lemma 2, 3, and 4 together into (35), we get (41). By taking $1-L^2\eta^2H^2-HL\theta\eta\frac{q(2-\rho)+K\rho}{K\rho}\geq 0$, we have

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}_Q\left[\mathbb{E}_{\text{Air}}\left[f(\mathbf{w}^{r+1})-f(\mathbf{w}^r)\right]\right]\right]\leq-\frac{\eta\theta H}{2}\|\nabla f^r\|_2^2\\ &+\frac{\theta\eta^2LH}{2K}(\eta LHK+\frac{(\rho+q)\theta}{\rho})\sigma^2+\frac{\theta^2\eta^2\sigma_z^2L}{2K^2\rho^2}\end{aligned}\tag{42}$$

Recursively applying the above inequality from $r=0$ to $r=R-1$ yields (43).

Until now we complete the proof of Theorem 1.

$$
\begin{aligned}
&\mathbb{E}_{\xi^{(r)}}\mathbb{E}_Q\mathbb{E}_{Air}\left[\|Air_{\mathcal{K}}\left(Q\left(\sum_{h=0}^{H-1}\nabla F_k^{r,h}\right)\right)\|^2\right]\\
=&\mathbb{E}_{\xi^{(r)},Q}\left[\frac{1}{K^2}\left(\|\sum_{k=1}^{K}\nabla F_{k,Q}^r\|^2+(\frac{1}{\rho}-1)\sum_{k=1}^{K}\|\nabla F_{k,Q}^r\|^2\right)+\frac{\sigma_z^2}{K^2\rho^2}\right]\\
=&\mathbb{E}_{\xi^{(r)}}\left[\mathbb{E}_Q\left[\|\frac{1}{K}\sum_{k=1}^{K}\nabla F_{k,Q}^r\|^2+\frac{1}{K^2}(\frac{1}{\rho}-1)\sum_{k=1}^{K}\|\nabla F_{k,Q}^r\|^2\right]\right]+\frac{\sigma_z^2}{K^2\rho^2}\\
=&\mathbb{E}_{\xi^{(r)}}\left[\mathbb{E}_Q\left[\frac{1}{K^2}\sum_{k=1}^{K}\left[\|\nabla F_{k,Q}^r-\nabla F_k^r\|^2\right]\right]+\|\frac{1}{K}\sum_{k=1}^{K}\nabla F_k^r\|^2\right]+\mathbb{E}_{\xi^{(r)}}\left[\frac{1}{K^2}(\frac{1}{\rho}-1)\sum_{k=1}^{K}(\mathbb{E}_Q\left[\|\nabla F_{k,Q}^r-\nabla F_k^r\|^2\right]+\|\nabla F_k^r\|^2\right]+\frac{\sigma_z^2}{K^2\rho^2}\\
=&\mathbb{E}_{\xi^{(r)}}\left[\mathbb{E}_Q\left[\frac{1}{K^2\rho}\sum_{k=1}^{K}\left[\|\nabla F_{k,Q}^r-\nabla F_k^r\|^2\right]\right]+\|\frac{1}{K}\sum_{k=1}^{K}\nabla F_k^r\|^2+\frac{1}{K}(\frac{1}{\rho}-1)\sum_{k=1}^{K}\|\nabla F_k^r\|^2\right]+\frac{\sigma_z^2}{K^2\rho^2}\\
\leq&\mathbb{E}_{\xi^{(r)}}\left[\sum_{k=1}^{K}\frac{q}{K^2p}\|\nabla F_k^r\|^2+\|\frac{1}{K}\sum_{k=1}^{K}\nabla F_k^r\|^2+\frac{1}{K}(\frac{1}{\rho}-1)\sum_{k=1}^{K}\|\nabla F_k^r\|^2\right]+\frac{\sigma_z^2}{K^2\rho^2}\\
=&\sum_{k=1}^{K}\frac{q}{K^2\rho}\left[\mathrm{Var}(\nabla F_k^r)+\|\nabla f_k^r\|^2\right]+\left[\frac{1}{K^2}\sum_{k=1}^{K}\mathrm{Var}(\nabla F_k^r)+\|\frac{1}{K}\sum_{k=1}^{K}\nabla f_k^r\|^2\right]+\frac{1}{K}(\frac{1}{\rho}-1)\sum_{k=1}^{K}\|\nabla f_k^r\|^2+\frac{\sigma_z^2}{K^2\rho^2}\\
\leq&\sum_{k=1}^{K}\frac{q+\rho}{K^2\rho}\mathrm{Var}(\nabla F_k^r)+\sum_{k=1}^{K}\frac{q(2-\rho)+K\rho}{K^2\rho}\|\nabla f_k^r\|^2+\frac{\sigma_z^2}{K^2\rho^2}
\end{aligned}
\tag{39}
$$

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}_Q\left[\mathbb{E}_{\text{Air}}\left[f(\mathbf{w}^{r+1})-f(\mathbf{w}^r)\right]\right]\right]\\
\leq&\frac{\eta\theta}{2K}\sum_{k=1}^{K}\sum_{h=0}^{H}\left[-\|\nabla f^r\|_2^2-\|\nabla f_k^{r,h}\|_2^2+L^2\eta^2H\left[\sigma^2+H\|\nabla f_k^{r,h}\|_2^2\right]\right]+\\
&\frac{(q+\rho)\theta^2\eta^2L}{2K\rho}H\sigma^2+HL\theta^2\eta^2\frac{q(2-\rho)+K\rho}{2K^2\rho}\sum_{k=1}^{K}\sum_{h=0}^{H}\|\nabla f_k^{r,h}\|^2+\frac{\theta^2\eta^2\sigma_z^2L}{2K^2\rho^2}\\
=&-\frac{\eta\theta H}{2}\|\nabla f^r\|_2^2-\frac{\eta\theta}{2K}(1-L^2\eta^2H^2-HL\theta\eta\frac{q(2-\rho)+K\rho}{K\rho})\sum_{k=1}^{K}\sum_{h=0}^{H}\|\nabla f_k^{r,h}\|_2^2+\frac{\theta\eta^2LH}{2K}(\eta LHK+\frac{(\rho+q)\theta}{\rho})\sigma^2+\frac{\theta^2\eta^2\sigma_z^2L}{2K^2\rho^2}
\end{aligned}
\tag{41}
$$

$$
\begin{aligned}
\frac{1}{R}\sum_{r=0}^{R-1}\|\nabla f^r\|_2^2\leq&\frac{2(f(\mathbf{w}^0)-f(\mathbf{w}^*))}{\eta\theta HR}+\frac{\eta L}{K}(\eta LHK+\frac{(\rho+q)\theta}{\rho})\sigma^2+\frac{\theta\eta L}{HK^2\rho^2}\sigma_z^2\\
=&\frac{2(f(\mathbf{w}^0)-f(\mathbf{w}^*))}{\eta\theta HR}+\frac{\eta\theta L}{K}\frac{(\rho+q)}{\rho}\sigma^2+\eta^2L^2H\sigma^2+\frac{\theta\eta L\sigma_z^2}{HK^2\rho^2}
\end{aligned}
\tag{43}
$$

B. Proof of the convexity of Θ_1 and Θ_2

The second-order partial derivative of functions Θ_1 and Θ_2 can be calculated as:

$$
\frac{\partial\Theta_1}{\partial\rho}=-\frac{A_0q}{\rho^2H}-\frac{B_0q}{2\rho^{\frac{3}{2}}H^{\frac{1}{2}}(\rho+q)^{\frac{1}{2}}} \tag{43}
$$

$$
\frac{\partial^2\Theta_1}{\partial\rho^2}=\frac{2q\left(A_0\sqrt{H}\left(q\sqrt{\rho}+\rho^{3/2}\right)\sqrt{\rho+q}+\frac{1}{2}H\rho B_0\left(\rho+\frac{3}{4}q\right)\right)}{H^{3/2}\rho^{7/2}(\rho+q)^{3/2}} \tag{44}
$$

$$
\frac{\partial^2\Theta_1}{\partial\rho\partial H}=\frac{q\left(A_0\sqrt{\rho}\sqrt{H}\sqrt{\rho+q}+\frac{1}{4}B_0\rho H\right)}{H^{5/2}\sqrt{\rho+q}\rho^{5/2}} \tag{45}
$$

$$
\frac{\partial\Theta_1}{\partial H}=-\frac{A_0(\rho+q)}{\rho H^2}-\frac{1}{2}\frac{B_0\sqrt{\rho+q}\rho}{(\rho H)^{3/2}} \tag{46}
$$

$$
\frac{\partial^2\Theta_1}{\partial H^2}=\frac{2\left(A_0(\rho+q)\sqrt{\rho H}+\frac{3}{8}B_0\sqrt{\rho+q}\rho H\right)}{\sqrt{\rho H}\rho H^3} \tag{47}
$$

$$\frac{\partial^2 \Theta_1}{\partial H \partial \rho} = \frac{1}{4} \frac{q(4A_0\sqrt{\rho + q\sqrt{\rho}H + B_0\rho H})}{\rho^2 H^2 \sqrt{\rho + q}\sqrt{\rho H}} \tag{48}$$

$$\frac{\partial^2 \Theta_2}{\partial \rho^2} = \lambda \frac{\ln^2 \rho - 3\ln \rho + 1}{\rho \ln^2 \rho (\ln \rho - 1)^2} T^{comm} \tag{49}$$

$$\frac{\partial^2 \Theta_2}{\partial H^2} = \frac{\partial^2 \Theta_2}{\partial H \partial \rho} = \frac{\partial^2 \Theta_2}{\partial \rho \partial H} = 0 \tag{50}$$

With the equations above, we can easily conclude that the Hessian matrix of both the functions Θ_1 and Θ_2 are positive semi-definite. It implies that Θ_1 and Θ_2 are convex. This completes the proof.