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

Dian Shi, *Student Member*, *IEEE*, Chenpei Huang, *Student Member*, *IEEE*, Liang Li, *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 on local devices. 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. energy-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 develop an energy-efficient FL over 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 method outperforms existing ones in terms of spectrum utilization, energy efficiency, and learning accuracy.**

***Index Terms*—Federated Learning, Multi-Bit Over-the-Air Computation, Gradient Quantization, Energy Minimization.**

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 and generated 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 good 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 transmitted 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 updates 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, hampering the scalability of FL, i.e., FL over a large number of mobile devices, if there are limited spectrum resources. On the other hand, transmitting a large volume of model updates periodically and executing heavy local on-device computing tasks can quickly drain out the energy of batter-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 transmissions simultaneously. By applying AirComp to FL, all the participating devices can transmit their model updates on the same channel. Due to the fact that 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 works in the literature 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 one iteration transmission performance of the Aircomp FL [5], [6], and the effects of AirComp on overall FL training performance, especially the FL convergence, are rarely discussed.

Therefore, in this paper, we design a novel multi-bit Aircomp (M-AirComp) FL scheme, named ESOAFL, which is

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

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).

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).

compatible with the most common Quadrature Amplitude Modulation (QAM) to transmit the model updates, so that we do not need to modify the modulation protocols manufactured within commercial off-the-shelf mobile devices. Specifically, gradient quantization is incorporated into the ESOAFL scheme to facilitate the digital modulation, and only part of the gradients are selected to transmit to cope with the channel fading. In addition to handling the spectrum issue in FL over wireless networks, our scheme is battery-friendly to the participating mobile devices. Here, the energy consumption is considered from the long-term learning perspective where local computing (i.e., "working") and wireless communication (i.e., "talking") are two main focuses. Our M-AirComp FL scheme only requires updated gradients with good channel conditions to transmit, which further saves the communication energy compared with other AirComp schemes. Moreover, we theoretically analyze the convergence property of our ESOAFL approach, based on which we quantify the number of communication rounds needed for achieving the convergence, and the overall long-term energy consumption is further modeled. Finally, we develop a joint transmission probability and local computing control approach to balance "working" and "talking," thus minimizing overall energy consumption. Our salient contributions are summarized as follows.

- We propose an energy and spectrum efficient AirComp FL approach (ESOAFL) with high precision M-AirComp design, where updated gradients are quantized into multi-bit adapting to the digital modulation settings. Additionally, we adopt an energy efficient power control policy to facilitate the M-AirComp, where only updated gradients with good channel conditions are selected to participate in the FL training.
- To help minimize the overall energy consumption of the proposed ESOAFL approach, the corresponding convergence analysis is derived, which quantitatively indicates the impacts of the M-AirComp on FL. Besides, the gradients transmission probability and local computing iterations are further optimized to achieve the overall energy efficiency.
- We conduct extensive simulations and verify the effectiveness of our proposed ESOAFL scheme and the corresponding control approach under various learning models, datasets, and multiple wireless environmental settings. Compared with other schemes, our proposed method shows significant spectrum utilization and energy efficiency superiority. The ESOAFL approach has the potential to improve spectral efficiency dozens of times and save at least half of the energy consumption.

The rest of this paper is organized as follows. Section II provides some preliminaries of AirComp and FL. Section III describes our M-AirComp design and the corresponding ESOAFL approach. The convergence analysis of the proposed ESOAFL approach is derived in Section IV, and the formulation and solution of the energy efficient control scheme are also presented. Numerical simulations are provided in Section V, and VI reviews related works of the AirComp FL. 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 in each FL round, where each user $k \in \{1, 2, \ldots, K\}$ has its own data set, 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 $r \in \{1, 2, \ldots, R\}$ denote the FL global round index between the server and local devices, and H be the number of local computing iterations executed between two consecutive global communication rounds. Moreover, We define $\mathbf{w}^r$ as the global model at the r-th communication round and $\mathbf{w}_k^{r,h}$ as the local model of user k at the h-th local iteration under the r-th communication round. Therefore, the local updating process of user k under the r-th communication round is denoted as:

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

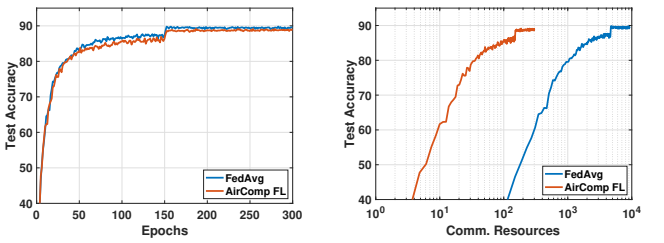
where $\nabla F_k(\mathbf{w}_k^{r,h})$ is a stochastic gradient of f with a random batch-size data, and η is the local learning rate. Here, $\nabla F_k(\mathbf{w}_k^{r,h})$ is the unbiased estimation of $\nabla f_k(\mathbf{w}_k^{r,h})$, i.e., $\mathbb{E}_{\xi\sim\mathcal{D}_k}\left[\nabla F_k(\mathbf{w}) \mid \boldsymbol{\xi}\right] = \nabla f_k(\mathbf{w})$, where ξ represents the randomness like the batch-size index. After finishing the local training, every participate upload 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 training process, all users need to transmit their local updates to the server for aggregation, which causes severe transmission congestion and requires a lot of communication resources, especially with massive participating users. As one of the advanced wireless techniques, over-the-air computation (AirComp) enables all users to simultaneously transmit the local gradients over the same wireless medium without interference and aggregates gradients together in the transmission process, significantly improving the spectrum utilization and saving communication resources.

Let $\mathcal{X} := \{x_1, x_2, ..., x_K\}$ be the input set and $\tilde{y}$ be the output objective of the system. In other words, x_k denotes gradients to be transmitted for user k in FL, and $\tilde{y}$ denotes the aggregation result received at the server with AirComp. Specifically, with AirComp, a wireless communication system usually adopts precoding and amplification at transmitters, while receivers often have equalization blocks for signal detection. Therefore, AirComp computes the aggregated objective at each time slot as

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



(a) Test accuracy vs. epochs (b) Test accuracy vs. comm. resources

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

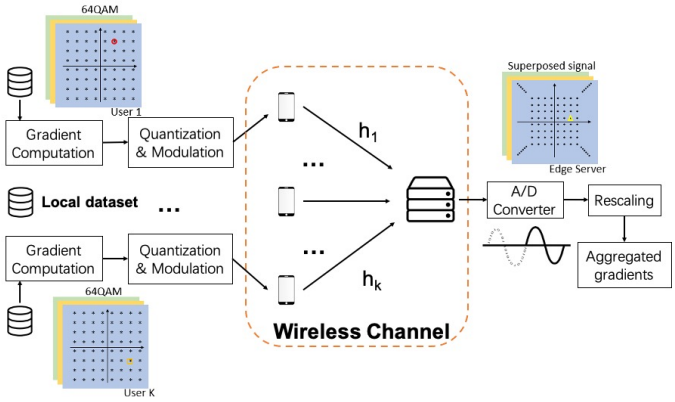


Fig. 2. Multi-bit Over-the-Air computation design.

where $h_k \in \mathbb{C}$ is the channel coefficient between user k and the aggregator, h_k satisfies the Rayleigh distribution, i.e., $h_k \sim CN(0, \sqrt{\lambda})$. $n \sim N(0, \sigma_z^2)$ 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 jointly amplifies the transmitted data. The goal of the Tx-scaling is to ensure that each distributed user contributes equally at the receiver antenna and the superposed signal is proportional to the ideal summation. Note that the ideal summation is defined as the average operation over the input set without the AirComp, i.e., $y := \frac{1}{K}\sum_{k=1}^{K} x_k$. Accordingly, the Rx-scaling factor $a \in \mathbb{R}$ acts as a equalizer, and rescales the sampled analog result to its expected value.

C. Preliminary Experiments on AirComp FL

To better show the benefits of AirComp FL, we show the performance of FL schemes with AirComp (AirComp FL) and without AirComp (FedAvg) in this subsection. We consider an FL task involving 10 participants to collaboratively train a ResNet-18 model with CIFAR-10 datasets. We allocate the same communication bandwidth for these two schemes and train both models with the same number of data epochs. The simulation results are depicted in Fig. 1. Fig. 1(a) shows that, compared with FedAvg, AirComp FL only has a minor loss of convergence in terms of data epochs. In other words, AirComp FL requires a little more or even the same number of data epochs to achieve the target test accuracy. For the communication spectrum, all users cannot take the concurrent transmission with the same bandwidth in FedAvg. Therefore, communication resources consumption of the AirComp FL is much less than that of FedAvg. The overall communication cost during the training process of these two approaches is shown in Fig. 1(b), which demonstrates that AirComp FL can significantly improve spectrum efficiency compared with FedAvg. Here, we assume that AirComp FL consumes a unit communication resource in each communication round.

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

A. The Design of Multi-Bit Over-the-Air Computation

Different from the most existing AirComp approaches with an analogy modulation scheme, we implement the digital modulation scheme for the AirComp to pair with the commercial transmit devices and design a Multi-bit AirComp scheme (M-AirComp). In this situation, the Rx-scaling factor a acts 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, each gradient is clipped and quantized as the Multiple Amplitude Shift Keying (MASK) symbol, which is compatible with modern digital devices. In the following, two MASK modulated gradients can be transmitted orthogonally using in-phase (I) and quadrature (Q) channel simultaneously. We notice that it is equivalent to map two separate gradients onto a symbol from the square M^2 QAM constellation. 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 shown in Fig. 2. In this way, altering the value M at the transmitter allows full digital data transmission while preserving b-bit resolution, according to the estimated channel gain.

We assume the edge server will equip 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). However, 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.

During the transmission process, each device is constrained by an average transmitting power budget P^0. Thus, assume all devices have the same power budget, and the average

transmission power constraint is:

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

Due to the transmit power constrain, some users facing the poor fading channels cannot completely align their amplitude, which means the Tx-scaling factor $p_k \in \mathbb{C}$ cannot be infinitely enlarged to meet the amplitude alignment requirement. Therefore, we adopt an energy efficient power control policy that the gradients of users with poor channel conditions are not transmitted. In other words, we set the transmit power of these users as 0, thus saving the transmission energy. Let g_{th} denote as a channel threshold, and the power control policy p_k can be represented as:

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

For the above equation, ϱ is a scaling factor to guarantee the desired SNR. Under the above power control policy, only gradient elements facing channel gain larger than the threshold g_{th} can be allowed to transmit. Note that the threshold g_{th} can be adjusted to control the gradient transmission. Due to the power constraint in Eq. 4, the threshold g_{th} can be set as an arbitrary value larger than a minimum value $g_{\text{th}}^{\min} := h^2 = \frac{\varrho}{p^2} = \frac{\varrho}{P^0}$. Specifically, under a certain communication environment, the greater the threshold g_{th} we set, the larger the number of allowable transmitting gradients. By changing the threshold g_{th}, our M-AirComp design has the potential to only involve gradients with good channel conditions, which just require low transmit power in an energy efficient manner. Thus, we define a long-term average transmission probability p_b to indicate the degree of gradient participation. To be specific, each specific threshold corresponds to a transmission probability. Since the channel coefficient is Rayleigh distributed $h_k \sim CN(0, \sqrt{\lambda})$, the channel gain is an exponential distribution. Therefore, the transmission probability p_b corresponding to the threshold g_{th} can be calculated as:

$$p_b = \int_{g_{\text{th}}}^{\infty} \lambda e^{-\lambda x} \mathrm{d}x = e^{-\lambda g_{\text{th}}} \,. \tag{6}$$

If the probability of keeping these gradient elements to transmit is p_b, the Rx-scaling factor a will be set as $\frac{1}{\sqrt{\varrho} p_b}$ to rescale the received signal. Due to the property of the Rayleigh fading channel and the power constrain of the local user devices, the achievable highest transmission probability $p_b^{\max}$ is calculated as $p_b^{\max} = e^{-\lambda g_{\text{th}}^{\min}} = e^{-\lambda \frac{\varrho}{P_0}}$.

B. M-AirComp Based FL

To improve spectrum efficiency and reduce energy consumption, we design an **E**nergy and **S**pectrum Efficient **O**ver the **A**ir **F**ederated **L**earning (**ESOAFL**) approach with gradient quantization technique, which is shown in Fig. 2 and described in detail in Alg. 1. All mobile devices start the training procedure with the initialized model parameters. Specifically, each user executes H local computing SGD steps with mini-batch size data of its own dataset. After the local training, we adopt the uniform gradient quantization operator $Q(\cdot)$ to quantize the updated gradients with low bits, i.e., 4-bit or 8-bit. Taking b-bit quantization as an example, the gradients of all participants are quantized to 2^b levels with a specific maximum/minimum value, catering to the digital wireless transmission scheme. Next, for the transmission process, every 2 gradient element is modulated into one digital symbol over the sub-channel according to our M-AirComp design. 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 $\text{Air}(\cdot)$ and apply the proposed energy efficient power control scheme. The threshold g_{th} is determined firstly, and then gradient element whose corresponding channel gain larger than this threshold can be allowed to transmit. In this way, the long-term transmission probability p_b can be calculated as $e^{-\lambda g_{\text{th}}}$. Because M-AirComp integrates wireless transmissions and aggregation over the air, the server receives only the aggregated updated gradients. Finally, the server updates the global model with the aggregated updated gradients, which can be represented as:

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

Algorithm 1 **E**nergy and **S**pectrum Efficient **O**ver the **A**ir **F**ederated **L**earning Algorithm (ESOAFL)

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 $h = 0$

while $r < R$ **do**
  for $h = 0, ..., H-1$ **do**
    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 $\mathcal{D}_k$
    Each device k in parallel updates its local model: $\mathbf{w}_k^{r,h+1} = \mathbf{w}_k^{r,h} + \eta \nabla F_k(\mathbf{w}_k^{r,h}), \quad \forall k$
  end for
  Each device k calculates the accumulated gradients with gradient quantization as $Q\left(\eta \sum_{h=0}^{H-1} \nabla F_k(\mathbf{w}_k^{r,h})\right)$
  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
  All transmitted gradients are aggregated over the air and the global model is updated as in Eq. (7)
  Update $r \leftarrow r + 1$
  Each device k updates its local model $\mathbf{w}_k^{r,0} = \mathbf{w}^r$
end while

After updating the global model, the server will broadcast the global model to all devices for continuing 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 round r.

IV. Spectrum and Energy Efficient FL: Formulation and Solutions

In this section, we first formulate an overall energy minimization problem and establish the communication and computation energy models of the proposed ESOAFL approach. Based on the derived convergence analysis, we then find the optimal control policy in terms of the transmission probability p_b and local computing iterations H to minimize the overall energy consumption.

A. Energy Minimization Problem Formulation

Deploying energy-hungry FL training on mobile devices is challenging due to the limited battery capacity of mobile devices. Hence, in this work, we aim to minimize the total energy consumption of FL training via local computing iterations H and transmission probability p_b control. The average energy consumption per communication round of mobile device casts as $E = E^{comm}(p_b) + E^{comp}H$. Here, $E^{comm}(p_b)$ is the communication energy to transmit the updated gradients, which is related to the transmission probability p_b, and E^{comp} is the computing energy of performing one local iteration. The goal is to minimize the overall energy consumption in FL training while guaranteeing the model convergence, denoted as

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

B. Communication and Computation Energy Models

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

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

where $\mathrm{Ei}(x)$ is the exponential integral function denoted as $\mathrm{Ei}(x) = \int_{-\infty}^{x}\frac{e^t}{t}\mathrm{d}x$. Due to the fact that $-\ln p_b$ is positive, we have $\mathrm{Ei}\left(\ln p_b\right) = -\mathrm{E}_1\left(-\ln p_b\right)$, where $\mathrm{E}_1(x) = \int_x^{\infty}\frac{e^{-t}}{t}\mathrm{d}x$. Therefore, we get $P^{comm} = -p_b\varrho\lambda\mathrm{Ei}\left(\ln p_b\right) = p_b\varrho\lambda\mathrm{E}_1\left(-\ln p_b\right)$. For positive real values of the x, $\mathrm{E}_1(x)$ can be bracketed by elementary functions as follows:

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

Due to $-\ln p_b > 0$, we have

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

After receiving the full precision gradients, each device is required to quantize the gradient into low-bit precision for digital transmission. Then, 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, which is in inverse proportion to sub-channel bandwidth B. Therefore, for transmitting the model with the number of d gradients, $d/2$ 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 the product of the average transmission power and the transmission time, as:

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

2) Computational model: With massive data are generated or collected on mobile devices, local on-device computing 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 efficiently. Therefore, we consider the GPU computational energy model here. We model the energy consumption for processing a mini-batch of data in one iteration as a product of the runtime power and the execution time, i.e.,

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

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

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

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

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 constant coefficients that reflect 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 and the corresponding dataset, such coefficients can be accurately estimated according to the hardware experiments. Since there are H local computing iterations between two sequential 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 local iteration number, i.e., $E^{comp} \times H$.

C. Impacts of Control Variables on ESOAFL Convergence

In this subsection, we derive the convergence analysis of the ESOAFL approach, where we theoretically analyze the impacts of control variables p_b and H on training convergence. Firstly, we have the following model assumptions.

Assumption 1 (Smoothness)**.** *The objective function f_i 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 the quantization operator $Q(x)$ is an unbiased estimator of its input x, and its variance grows with the squared of L2-norm of its argument, i.e., $\mathbb{E}[Q(x)] = x$ and $\mathbb{E}[||Q(x)-x||^2] = q||x||^2$.*

In our work, we consider the Rayleigh channel and employ the power control policy with a transmission probability p_b. Thus, the following assumption is obtained.

Assumption 4 (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 $\mathcal{X}$, and its variance decreases with the increasing of the transmission probability and grows with the squared of its argument, i.e., $\mathbb{E}[Air(\mathcal{X})] = y$ and $\mathbb{E}[||Air(\mathcal{X}) - y||^2] = \frac{1}{K^2}(\frac{1}{p_b}-1)\sum_{x_k\in\mathcal{X}} x_k^2 + \frac{\sigma_z^2}{K^2{p_b}^2}$.*

Proof. Let $\mathcal{X}$ be the input set of the M-AirComp operator, and we further define $\bar{\mathcal{X}}$ as the successful transmit set to help the proof. Accordingly, the mean and the mean of the square values can be expressed as:

$$\begin{aligned}
&\mathbb{E}[\text{Air}(\mathcal{X})] \qquad (19)\\
=&\mathbb{E}\left[\frac{1}{p_b K}\left[\sum_{x_k\in\bar{\mathcal{X}}} x_k + \sum_{x_k\notin\bar{\mathcal{X}}} x_k + n\right]\right]\\
=&\frac{1}{p_b K}\left[\sum_{x_k\in\bar{\mathcal{X}}} x_k\cdot p_b + \sum_{x_k\notin\bar{\mathcal{X}}} 0\cdot(1-p_b) + \mathbb{E}\left[n\right]\right] = y,\\
&\mathbb{E}[(\text{Air}(\mathcal{X}))^2] \qquad (20)\\
=&\mathbb{E}\left[\frac{1}{{p_b}^2K^2}\left(\sum_{x_k\in\mathcal{X}} x_k + n\right)^2\right]\\
=&\mathbb{E}\left[\frac{1}{{p_b}^2K^2}\left(\sum_{x_i\in\mathcal{X}}\sum_{x_j\in\mathcal{X}} x_i x_j + 2\sum_{x_k\in\mathcal{X}} x_k n + n^2\right)\right]\\
=&\frac{1}{{p_b}^2K^2}\left[\sum_{x_i\in\mathcal{X}}\sum_{x_j\in\mathcal{X},i\neq j} x_i p_b x_j p_b + \sum_{x_k\in\mathcal{X}} x_k^2 p_b\right] + \frac{\sigma_z^2}{K^2{p_b}^2}\\
=&\frac{1}{{p_b}^2K^2}\left[{p_b}^2\left((\sum_{x_k\in\mathcal{X}} x_k)^2 - \sum_{x_k\in\mathcal{X}} x_k^2\right) + p_b\sum_{x_k\in\mathcal{X}} x_k^2 + \sigma_z^2\right]\\
=&\frac{1}{K^2}\left((\sum_{x_k\in\mathcal{X}} x_k)^2 + (\frac{1}{p_b}-1)\sum_{x_k\in\mathcal{X}} x_k^2\right) + \frac{\sigma_z^2}{K^2{p_b}^2}
\end{aligned}$$

Thus, the variance is equal to the mean of the square value minus the square of the mean value, which is represented as:

$$\begin{aligned}
&\text{Var}(\text{Air}(\mathcal{X})) = \mathbb{E}[(\text{Air}(\mathcal{X}))^2] - \mathbb{E}[\text{Air}^2(\mathcal{X})]\\
=&y^2 + \frac{1}{K^2}(\frac{1}{p_b}-1)\sum_{x_k\in\mathcal{X}} x_k^2 + \frac{\sigma_z^2}{K^2{p_b}^2} - y^2\\
=&\frac{1}{K^2}(\frac{1}{p_b}-1)\sum_{x_k\in\mathcal{X}} x_k^2 + \frac{\sigma_z^2}{K^2{p_b}^2}.
\end{aligned} \tag{21}$$

□

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

$$1 \geq L^2\eta^2H^2 + HL\theta\eta\frac{q(2-p_b)+Kp_b}{Kp_b}, \tag{22}$$

and with considering the gradient quantization q, the M-AirComp transmission probability p_b, and the local computing iterations H, the convergence rate after R communication rounds can be bounded as:

$$\begin{aligned}
\frac{1}{R}\sum_{r=0}^{R-1}\|\nabla f^r\|_2^2 \leq& \frac{2(f(\mathbf{w}^0)-f(\mathbf{w}^*)}{\eta\theta HR}\\
+&\frac{\eta\theta L}{K}\frac{(p_b+q)}{p_b}\sigma^2 + \eta^2L^2H\sigma^2 + \frac{\theta\eta L}{HK^2{p_b}^2}\sigma_z^2,
\end{aligned} \tag{23}$$

where $f(\mathbf{w}^*)$ is the minimum value of the loss.

Proof. Please refer to the detailed proof[1] at Github. □

The proof of Theorem 1 can be 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. Next, we can bound the updated gradients with M-AirComp and quantization operators. Finally, by integrating the derived results above, we can obtain the convergence analysis of the ESOAFL approach.

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 $\theta\eta = O\left(\frac{1}{L}\sqrt{\frac{Kp_b}{RH(p_b+q)}}\right)$, the convergence rate can be represented as:*

$$\begin{aligned}
\frac{1}{R}\sum_{r=0}^{R-1}\|\nabla f^r\|_2^2 \leq& \frac{2L(f(\mathbf{w}^0)-f(\mathbf{w}^*)\sqrt{(p_b+q)}}{\sqrt{KRHp_b}} +\\
&\frac{\sqrt{p_b+q}}{\sqrt{KRHp_b}}\sigma^2 + \frac{K}{R\theta^2}\sigma^2 + \sqrt{\frac{1}{K^3RH^3(p_b+q){p_b}^3}}\sigma_z^2\\
\overset{(a)}{=}& O\left(\frac{\sqrt{p_b+q}}{\sqrt{KRHp_b}}(2L(f(\mathbf{w}^0)-f(\mathbf{w}^*)+\sigma^2)) + \frac{K}{R\theta^2}\sigma^2\right)\\
\overset{(b)}{=}& O\left(\frac{\chi}{\sqrt{KRH}}\right) + O\left(\frac{K}{R}\right),
\end{aligned} \tag{24}$$

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

[1] https://github.com/shidian117/ESOAFL/blob/main/proof.pdf

Algorithm 2 JCP Control Algorithm

Initialization: $\epsilon, \xi, \iota = 10^{-5}$; $\gamma^0 \in (0, 1]$; $\kappa = 0$

1: **repeat**
2: Solve (30) and set the optimal value as $\boldsymbol{\phi}^*(\boldsymbol{\phi}^\kappa)$
3: Set $\boldsymbol{\phi}^{\kappa+1} = \boldsymbol{\phi}^\kappa + \gamma^0(\boldsymbol{\phi}^*(\boldsymbol{\phi}^\kappa) - \boldsymbol{\phi}^\kappa)$
4: Set $\kappa = \kappa + 1$
5: Set $\gamma^\kappa = \gamma^{\kappa-1}(1 - \xi\gamma^{\kappa-1})$
6: **until** $||\boldsymbol{\phi}^\kappa - \boldsymbol{\phi}^{\kappa-1}||_2^2 \leq \iota$
7: Round the current H to the nearest integer in $\mathcal{H}$
8: **return** The current solutions of p_b and H.

After establishing the communication and computing energy models, another key component to formulate the overall energy consumption problem is to obtain the required communication rounds. Accordingly, we can obtain it from the derived convergence analysis.

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*

$$\begin{aligned} R &= O\left(\frac{2\epsilon\sigma^2 HK^2 + \chi^2(\delta+\sigma^2)^2\theta^2}{2\epsilon^2\theta^2 HK}\right) \\ &+ O\left(\frac{+\chi(\delta+\sigma^2)\theta\sqrt{4\epsilon\sigma^2 HK^2 + \chi^2(\delta+\sigma^2)^2\theta^2}}{2\epsilon^2\theta^2 HK}\right) \\ &= O\left(K\right) + O\left(\frac{\chi^2}{HK}\right) + O\left(\frac{\chi}{\sqrt{H}}\right), \end{aligned} \tag{25}$$

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

D. Overall Energy Minimization Reformulation and Solution

Aiming at minimizing the energy consumption during the entire training process, we reformulate the Joint local Computing and transmission Probability (JCP) control problem as:

$$\begin{aligned} \min_{p_b, H} \quad & R \times (E^{comm} + HE^{comp}) && \text{(26a)} \\ & = \left(\frac{A_0(p_b+q)}{p_b H} + \frac{B_0\sqrt{p_b+q}}{\sqrt{p_b H}} + C_0\right) \\ & \times \left(\varrho\lambda {p_b}^2 \ln(1 - \frac{1}{\ln p_b})T^{comm} + HE^{comp}\right) \\ s.t. \quad & 0 < p_b \leq p_b^{max}, && \text{(26b)} \\ & H \in \mathcal{H}, && \text{(26c)} \end{aligned}$$

where A_0, B_0, and C_0 are constants used to approximate the big-O notion in Eq. 25. From the above formula, we observe that increasing the local computing iterations H reduces the needed communication rounds R ("talking"), but increases the computing energy consumption per round ("working"). Similarly, adjusting p_b also affects the required communication rounds and the energy consumption of each round. Thus, it is necessary to optimize H and p_b to balance the "working" and "talking", thus minimizing the overall energy consumption.

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

$$\Theta_1(\boldsymbol{\phi}) = \frac{A_0(p_b+q)}{p_b H} + \frac{B_0\sqrt{p_b+q}}{\sqrt{p_b H}} + C_0, \tag{27}$$

$$\Theta_2(\boldsymbol{\phi}) = \varrho\lambda {p_b}^2 \ln(1 - \frac{1}{\ln p_b})T^{comm} + HE^{comp}. \tag{28}$$

Noticing the simple and decoupled constraints in (26b-26c), we relax the constraint in (26c) 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 easily observe 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.

Capturing such the "product-of-convexity" property of the objective function $\Theta(\boldsymbol{\phi})$, we can use the inner convex approximation method [13] to solve the relaxed JCP control problem by optimizing a sequence of strongly convex inner approximations of $\Theta(\boldsymbol{\phi})$ in the form: given $\boldsymbol{\phi}^\kappa \in \Phi$

$$\Theta(\boldsymbol{\phi}, \boldsymbol{\phi}^\kappa) = \Theta_1(\boldsymbol{\phi})\Theta_2(\boldsymbol{\phi}^\kappa) + \Theta_1(\boldsymbol{\phi}^\kappa)\Theta_2(\boldsymbol{\phi}), \tag{29}$$

where $\boldsymbol{\phi}^\kappa = \{H^\kappa, {p_b}^\kappa\}$ refers to the intermediate $\boldsymbol{\phi}$ obtained in the κ-th iteration. Obviously, the approximated objective function in (29) is strongly convex with the fixed $\boldsymbol{\phi}^\kappa$. With the surrogate function above, we are actually required to efficiently compute the optimal solutions of the following convex optimization problem in each iteration, while preserving the feasibility of the iterates to the original problem (26).

$$\begin{aligned} \min_{p_b, H} \quad & \Theta(\boldsymbol{\phi}, \boldsymbol{\phi}^\kappa) && \text{(30a)} \\ s.t. \quad & 0 < p_b \leq p_b^{max}, && \text{(30b)} \\ & H_{min} \leq H \leq H_{max}. && \text{(30c)} \end{aligned}$$

Notice that the problem (30) can be solved by various commercial solvers, e.g., IBM CPLEX optimizer [14]. 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 (30), and then taking a step from $\boldsymbol{\phi}^\kappa$ towards $\boldsymbol{\phi}^*(\boldsymbol{\phi}^\kappa)$. The process is repeated until it meets the termination criterion, and the value of H is rounded afterward to ensure its feasibility.

V. Performance Evaluation

A. Implementation of M-AirComp

As shown in Fig. 3, we first set up experiments to elaborate the usage of M-AirComp for FL testbed in the lab. The system comprises one edge server as well as two edge devices. We let one RTX-8000 server with one USRP X310 play the role of the over-the-air FL aggregator. Besides, each FL client consists of the NVIDIA Jetson TX2 as the computation unit and USRP N210 as the wireless transmitter. We also employ WBX 50-2200 MHz Rx/Tx USRP daughterboards, with up to 200 mW output power. The synchronization is provided by USRP X310 REF and PPS output ports through cable connection. In the end, all USRPs are connected to an internet switch.

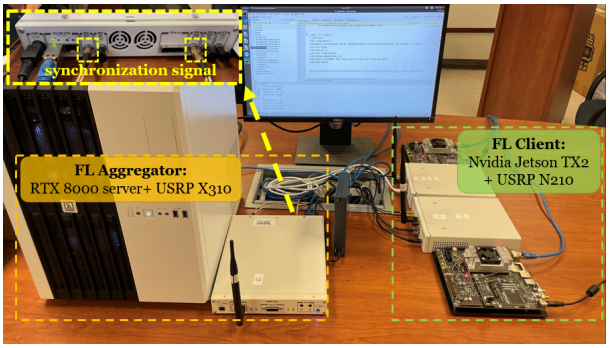


Fig. 3. Federated learning via M-AirComp testbed in the lab.

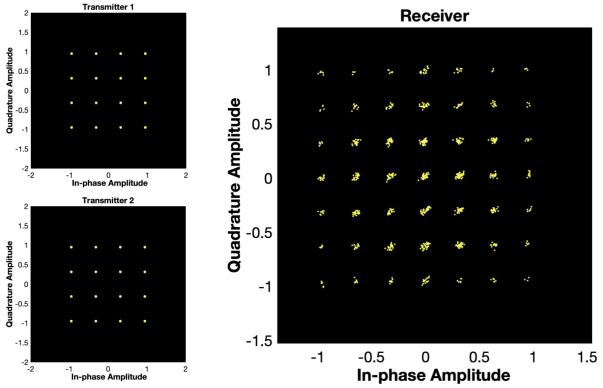


Fig. 4. Constellation diagram of M-AirComp demo (left: transmitter; right: receiver).

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 show the feasibility of M-AirComp by the in-lab experiments. In our M-AirComp demo, incorporated with quantization, two clients are transmitting QAM symbols, for example, 16 QAM for 4-bit quantization given in Fig. 4. From the constellation, the receiving symbol set is expanded into a constellation for higher-order modulations, which explains the addition carried in 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. In addition, if the communication environment (i.e., channel condition) is extremely poor, our proposed ESOAFL approach can still retain the performance in the case of many participating devices. We consider a severe communication environment with a SNR = 5dB over different numbers of participants(e.g., K = 10, 20, and 30). Here, we train the ResNet-18 model with the CIFAR-10 dataset. As shown in Fig. 5, with the increasing number of participating devices, the convergence gap between the ESOAFL approach and its ideal case (i.e., FedAvg without channel noise) gradually decreases. This verifies that the AirComp variance is decreasing with the number of participating devices K, which is also shown in Eq. 21. Moreover, especially with a large number of participants (e.g., K=30), the training curve of the ESOAFL approach is similar to its ideal case (i.e., FedAvg) in terms of training epochs, which also exhibits the strong anti-interference ability of the proposed ESOAFL approach.

One difficulty in the problem-solving process is to estimate the values of A_0, B_0, and C_0, which are related to the specific learning model and dataset. Here, we conduct the sampling-based methods to estimate these parameters, where we empirically sample different combinations (H, p_b) and employ the derived bound in (25) to approximate these values. Note that the estimation overhead is marginal. Take the ResNet-18 model with the CIFAR-10 dataset as an example. We first implement various local computing iterations H and different transmission probabilities p_b for the training task. Then we set the target training loss and record the corresponding number of communication rounds. After receiving these records, we utilize the Non-linear least squares curve fitting algorithm [15] to estimate values of A_0, B_0, and C_0, and the estimation results are shown in Fig. 6. From Fig. 6, obviously, with the increase of local computing iterations H and transmission probability p_b, the number of required communication rounds required 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 energy trade-off between local computing and wireless communications has to be considered to minimize the overall energy consumption, where H and transmission probability p_b are required to be carefully selected.

C. Spectrum and Energy Efficiency of the ESOAFL

After finishing the estimation process and perceiving the above observations, we implement the proposed JCP control scheme to find the optimal local computing iterations H and transmission probability p_b. Here, we consider two different image classification models and datasets to verify the effectiveness of our proposed approach, where the LeNet model on the MNIST dataset is relatively light, and ResNet-18 on CIFAR-10 is relatively complex. Both datasets consist of 50000 training images and 10000 test images in 10 classes, and we set batch size as 128 and 32 for ResNet and LeNet, respectively. In each round of FL, we set $K = 10$ participating mobile devices executing H steps of SGD in parallel, and the maximum transmission probability $p_b^{\max}$ 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. We consider several popular FL schemes as baseline approaches compared with our proposed ESOAFL-OPT approach (i.e., ESOAFL with optimal JCP control).

- FedAvg [1]: the FL approach without AirComp, where the ideal transmission is taken without channel noise.
- FedPAQ [16]: all participants transmit the quantized version of model updates to the edge server.

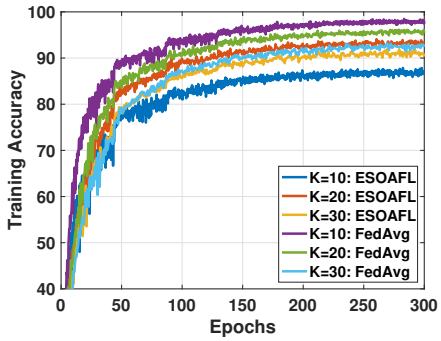


Fig. 5. Training performance under poor channel conditions.

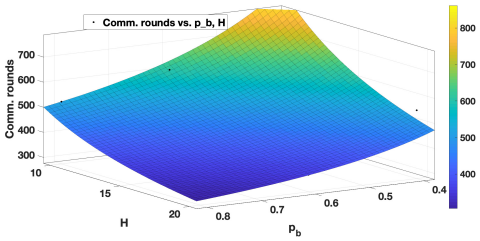


Fig. 6. Comm. rounds vs. p_b and H.

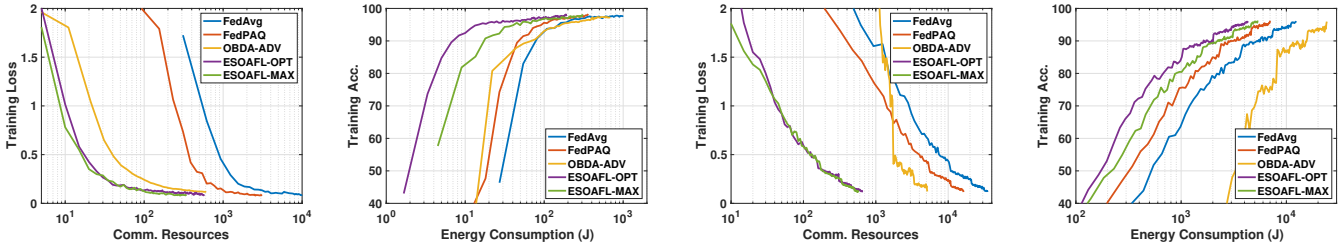
- OBDA-ADV [17]: a modified version of the OBDA (one-bit digital AirComp), where we improve the original scheme without considering the quantization at the receiver to preserve the learning precision.
- ESOAFL-MAX: the proposed ESOAFL scheme without the transmission control, where we adopt the maximum transmission probability $p_b^{\max}$ to transmit gradients.

We assume all schemes can utilize the same amount of communication bandwidth. Furthermore, We utilize the Nvidia TX2 as the mobile device and deploy Jtop [18] tool to measure the computing energy, where 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 consumes 130ms, and the GPU power is nearly 4W. We assume the AirComp can be deployed in the commercial LTE system for wireless transmissions. The resource block is 180 kHz, and we can obtain the transmission time of local updates with the specific model size accordingly. Moreover, we assume the average maximum transmit power is 0.2W. Thus, the transmission energy consumption can be calculated as the product of transmit power and transmission time. In all schemes, We set the average SNR=15dB for 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, in FedAvg and FedPAQ for a fair comparison.

Fig. 7(a) and 7(b) show the simulation results for LeNet on MNIST. Here, we set the target training loss ϵ as 0.07 and assume the data samples are independent and identically distributed (IID). For the OBDA-ADV scheme, we bring the local SGD method (i.e., taking several training steps among the sequential communication rounds) into the original scheme. Let the spectrum resource consumed in each communication round of the ESOAFL approach as the 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, FedPAQ. One reason is that FedAvg and FedPAQ consume more communication resources in each round because all devices cannot take the concurrent transmission with the same bandwidth. Another reason is that each pair of gradients can be transmitted orthogonally using in-phase (I) and quadrature (Q) channels simultaneously in the proposed ESOAFL scheme. However, according to the LTE protocol, each resource element can only carry several bits of a gradient in FedAvg and FedPAQ. In the meantime, Fig. 7(b) presents the energy consumption during FL training, where $H = 3$ and $p_b = 0.29$ is obtained for optimal controlling of ESOAFL. The results show that our ESOAFL scheme consumes the least energy among all schemes. Specifically, when achieving the same target training loss, the energy efficiency of ESOAFL-OPT is twice and three times higher 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 scheme. Noted that the low-precision OBDA-ADV approach cannot reach the target training loss we set, and thus we consider the training loss $\epsilon = 0.12$ for the OBDA-ADV approach.

Fig. 7(c) and 7(d) demonstrate the performance comparison of all schemes with ResNet-18 model on CIFAR-10 dataset. We set the target training loss ϵ as 0.12, and obtain the optimal control strategies $H = 11$ and $p_b = 0.51$. Like the conclusions described above, the proposed ESOAFL approach dramatically improves the spectrum efficiency and reduces the required energy. In this situation, our proposed ESOAFL-OPT saves hundreds of times of communication resources compared with FedAvg and FedPAQ. It also saves more than $8\times$ of communication resources compared with the OBDA method. Accordingly, our proposed ESOAFL-OPT scheme saves nearly one-third and two-thirds of energy consumption than FedPAQ and FedAvg schemes. Furthermore, the OBDA-ADV approach has relatively poor convergence performance compared with other approaches due to the high precision requirement of the complex ResNet-18 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 [19]. For example, $\varsigma = 0.3$ indicates that 30% of the data belong to one label and the remaining 70% data belong to others. We ignore the OBDA-ADV scheme since its performance is not good



(a) Training loss vs. comm. resources (b) Training acc. vs. energy cons. (c) Training loss vs. comm. resources (d) Training acc. vs. energy cons.

Fig. 7. Simulation results on various architectures and datasets. ((a-b): LeNet on MNIST; (c-d): ResNet-18 on CIFAR-10.)

TABLE I
PERFORMANCE COMPARISON UNDER DIFFERENT LEARNING SETTINGS (RESNET18 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%

in non-IID data settings. From Table. I, we can observe that training with non-IID data incurs a larger energy and communication resources consumption to converge. Moreover, compared with FedAvg and FedPAQ, the proposed ESOAFL achieves the indistinguishable final 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 similar observations. Note that 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 $K = 10$. However, because of the concurrent property of the AirComp, communication resources consumption at $K = 100$ in the ESOAFL scheme remains constant compared with the scenario of $K = 10$, which indicates the potential of involving large mounts of participants in the ESOAFL scheme.

VI. RELATED WORKS

Recently, much attention has been paid to the energy-efficient FL over mobile devices, where several advanced techniques are utilized to save energy during the FL training [20]. On the one hand, gradient sparsification [21], [22] and gradient quantization [23], [24] techniques can compress model updates in the transmission process, significantly reducing the communication burdens [25]. On the other hand, some researchers consider applying a weight quantization scheme to reduce the required computing energy [26]. Although these methods can effectively reduce the energy cost, they are mainly considered from the perspective of learning algorithms and widely ignore the communication components, especially with the physical layer aspects of communication. Realizing the above problem, some pioneering works exploit the waveform superposition property of the wireless medium and propose the AirComp FL [27]. Cao et al. in [28], and Amiri and Gündüz in [29] apply AirComp to solve 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 (MSE) requirements. Additionally, the works in [5], [6] propose joint device selection and communication scheme design methods to improve the learning performance for AirComp FL. 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 not discussed in these works. Noticing the limits above, Zhu et al. [30] applies the 1-bit digital modulation and derives the convergence analysis accordingly. However, 1-bit based scheme tremendously scarifies precision without considering the energy consumption during the training process. Different from the existing approaches, our design targets the general digital modulation scheme with multiple bits, where the convergence-guaranteed FL approach integrates both the AirComp and the gradient quantization techniques. Moreover, the energy consumption issue, including both computing and communication, is well studied accordingly.

VII. CONCLUSION

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 guaranteed. Besides, it also saved at least half of energy consumption compared with other FL methods.

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] 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, December 2016.

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

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

[16] 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, August 2020.

[17] 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*, November 2020, doi:10.1109/TWC.2020.3039309.

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

[19] 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.

[20] D. Shi, L. Li, R. Chen, P. Prakash, M. Pan, and Y. Fang, "Towards energy efficient federated learning over 5g+ mobile devices," *accepted by IEEE Wireless Communications*, 2021.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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: PMLR, July 2020, pp. 3304–3314.

[27] 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.

[28] 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.

[29] 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.

[30] 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, 2021.