Federated Learning via Intelligent Reflecting Surface

Zhibin Wang, *Graduate Student Member, IEEE*, Jiahang Qiu, Yong Zhou, *Member, IEEE*, Yuanming Shi, *Senior Member, IEEE*, Liqun Fu, *Senior Member, IEEE*, Wei Chen, *Senior Member, IEEE*, and Khaled B. Letaief, *Fellow, IEEE*

***Abstract*— Over-the-air computation (AirComp) based federated learning (FL) is capable of achieving fast model aggregation by exploiting the waveform superposition property of multiple-access channels. However, the model aggregation performance is severely limited by the unfavorable wireless propagation channels. In this paper, we propose to leverage intelligent reflecting surface (IRS) to achieve fast yet reliable model aggregation for AirComp-based FL. To optimize the learning performance, we present the convergence analysis of our proposed IRS-assisted AirComp-based FL system, based on which we propose to maximize the number of scheduled devices of each communication round under certain mean-squared error (MSE) requirements. To tackle the formulated highly-intractable problem, we propose a two-step optimization framework. Specifically, we induce the sparsity of device selection in the first step, followed by solving a series of MSE minimization problems to find the maximum feasible device set in the second step. We then propose an alternating optimization framework, supported by the difference-of-convex programming for low-rank optimization, to efficiently design the aggregation beamformers at the BS and phase shifts at the IRS. Simulation results demonstrate that our proposed algorithm and the deployment of an IRS can achieve a higher FL prediction accuracy than the baseline schemes.**

***Index Terms*— Federated learning, intelligent reflecting surface, over-the-air computation, sparse optimization.**

Manuscript received November 13, 2020; revised May 8, 2021; accepted July 14, 2021. Date of publication July 30, 2021; date of current version February 14, 2022. The work of Yong Zhou was supported by the National Natural Science Foundation of China (NSFC) under Grant 62001294. The work of Liqun Fu was supported by the National Natural Science Foundation of China (NSFC) under Grant 61771017. The associate editor coordinating the review of this article and approving it for publication was D. Gunduz. *(Corresponding authors: Yong Zhou; Liqun Fu.)*

Zhibin Wang is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China, also with Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: wangzhb@shanghaitech.edu.cn).

Jiahang Qiu and Liqun Fu are with the School of Informatics and the Key Laboratory of Underwater Acoustic Communication and Marine Information Technology Ministry of Education, Xiamen University, Xiamen 361005, China (e-mail: jiahang@stu.xmu.edu.cn; liqun@xmu.edu.cn).

Yong Zhou and Yuanming Shi are with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (e-mail: zhouyong@shanghaitech.edu.cn; shiym@shanghaitech.edu.cn).

Wei Chen is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: wchen@tsinghua.edu.cn).

Khaled B. Letaief is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, and also with Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: eekhaled@ust.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2021.3099505.

Digital Object Identifier 10.1109/TWC.2021.3099505

I. Introduction

RECENT years have witnessed a bloom of artificial intelligence (AI) applications, such as chess play, natural language generation, and image classification. By adopting advanced machine learning techniques, particularly reinforcement learning and deep learning, computers are able to mimic human behaviours by exploiting tremendous computing power and large amounts of data. With the further rise of edge computing and Internet of Things (IoT), there emerges a new AI paradigm, named *edge AI* [1]–[3], which pushes the AI frontier from the cloud center to the network edge. As the data collection and processing are mostly performed at the network edge, the service latency and energy consumption of edge devices can be significantly reduced by edge AI. As a promising framework for edge AI, *federated learning* (FL) [4], [5] has recently been proposed to coordinate multiple edge devices to collaboratively train a global AI model. Specifically, FL iteratively performs the following two processes [4]: 1) *model aggregation*: the edge server receives the local model updates from the edge devices over multiple-access channels, and then updates the global model by averaging over the received local model updates; and 2) *model dissemination*: the edge server broadcasts its updated global model to the edge devices, each of which updates the local model based on its own local dataset. As only model parameters rather than the real raw data are transmitted to the edge server in the model aggregation process, FL is capable of achieving privacy protection.

As the edge devices are usually connected to the edge server over wireless channels, the model parameters received by the edge server are inevitably distorted by channel fading and additive noise. To tackle this issue, several *digital FL* schemes have been proposed to achieve reliable model aggregation [6]–[10]. Specifically, each edge device is allocated an orthogonal resource block to upload its local model parameters, while the edge server is assumed to correctly decode the received local models by adopting the adaptive modulation and coding scheme [6]. Due to limited communication resources, only a subset of edge devices can be scheduled to participate in FL to ensure the accuracy of model transmission. To improve the learning efficiency, the scheduling policy design was considered according to the instantaneous and time-average signal-to-noise ratio (SNR) [7], diversities in multiuser channels and edge devices' local gradients [8], delay and energy consumption requirements [9], and latency constraints [10]. However, the aforementioned studies adopted orthogonal multiple access (OMA) based resource allocation schemes, such as time division multiple access (TDMA) and orthogonal frequency division multiple access (OFDMA), where the required radio resources were linearly scaling with the number of edge devices that participate in FL. When the number of edge devices is large, a substantial communication latency is introduced in the model aggregation process and in turn becomes the performance-limiting factor of FL.

To address above challenges, *over-the-air computation* (AirComp) empowered *analog FL* emerged to enhance the learning performance under limited communication bandwidth and stringent latency requirements. AirComp merges the concurrent data transmission from multiple devices and the function computation via exploiting the waveform superposition property of multiple-access channels [11]–[15]. Meanwhile, as the edge server in FL is merely interested in the aggregated model rather than individual local models, AirComp, as a non-orthogonal multiple access (NOMA) scheme, is recognized as a promising solution for achieving spectral-efficient and low-latency FL [16]–[21]. Specifically, the authors in [16] proposed a fast model aggregation approach by jointly optimizing the device selection and the receive beamforming to improve FL performance under certain aggregation error requirements. The authors in [17] developed a broadband analog aggregation scheme for low-latency FL by considering the communication-and-learning trade-off, where the devices within a certain communication range were scheduled. Model sparsification and compression methods were proposed in [18], [19] to enable efficient use of the available limited channel bandwidth. In [20], the authors developed a gradient-based algorithm to directly deal with noise distorted gradients for FL over wireless channels. In addition, the authors in [21] studied the optimal power control problem for AirComp-based FL by considering gradient statistics. It is worth noting that to achieve an average behaviour of local updates during model aggregation, magnitude alignment should be achieved at the edge server to reduce the aggregation error of AirComp [15]. However, unfavorable propagation environment inevitably leads to magnitude reduction and misalignment, which in turn degrade the model aggregation accuracy of AirComp-based FL.

To overcome the detrimental effect of channel fading in wireless networks, *intelligent reflecting surface* (IRS) is a cost-effective technology for improving the spectral and energy efficiency via reconfiguring the wireless propagation environment [22]–[28]. In particular, a large number of low-cost passive reflecting elements contained in an IRS are capable of adjusting the phase shift of the incident signal, and thus altering the propagation of the reflected signal. The signal reflected by IRS can be constructively superposed with the signal over the direct link to boost the received signal power [22]. Due to the passive nature, the power consumption of the IRS is negligible compared with that of the traditional full-duplex amplify-and-forward relay [23]. In [23], an IRS was deployed to minimize the transmit power of the multi-antenna access point (AP) by jointly optimizing active and passive beamforming, while satisfying the signal-to-interference-plus-noise ratio (SINR) constraints. A joint design of the downlink transmit power and the phase shifts of IRS was developed in [24] to maximize the energy efficiency. The authors in [25] utilized the IRS to enhance the physical layer security by jointly optimizing the beamformers at the base station (BS) and the reflecting coefficients at the IRS. Moreover, the authors in [26] and [27] leveraged the IRS to minimize the distortion of AirComp in wireless networks. The aforementioned studies demonstrated the potential gains of deploying an IRS in harsh wireless environment, which motivates us to leverage IRS to compensate for magnitude reduction and misalignment of AirComp in FL systems, so as to achieve a higher prediction accuracy in fewer communication rounds.

A. Contributions

In this paper, we exploit the advantages of IRS to design a communication-efficient model aggregation scheme for AirComp-based FL systems. Developing such a scheme to facilitate fast yet reliable model aggregation is challenging. On one hand, selecting more devices to participate in FL at each communication round is able to simultaneously collect more local model updates, which has a positive impact on the convergence rate of the training process. On the other hand, selecting more devices in each communication round enlarges the model aggregation error due to the inevitable magnitude misalignment at the edge server, which is detrimental to the convergence rate of the training process. As a result, the edge devices should be appropriately selected to speed up the overall convergence rate of FL. The main contributions of this paper are summarized as follows.

- We propose an IRS-assisted AirComp-based FL system that is able to achieve fast yet reliable model aggregation, where an IRS is deployed to mitigate the magnitude misalignment at the edge server during model aggregation. We then derive the convergence analysis of the proposed IRS-assisted AirComp-based FL system, which will demonstrate the impact of the number of selected devices and the aggregation error on the convergence performance.
- We propose to jointly optimize the device selection, the aggregation beamformer at the BS, and the phase shifts at the IRS, which, however, is highly intractable due to the sparse objective function as well as the biquadratic constraints due to the coupling between the aggregation beamformer at the BS and the phase shifts at the IRS.
- We first propose a two-step optimization framework to tackle the sparse objective function. Specifically, we induce the sparsity of the device selection by adopting ℓ_1-relaxation for the ℓ_0-norm objective function in the first step, followed by solving a series of mean-squared error (MSE) minimization problems in the second step to find the maximum feasible device selection set. Then, we propose an alternating optimization algorithm based on difference-of-convex (DC) programming to obtain high-quality solutions.

Simulation results demonstrate that the proposed IRS-assisted AirComp-based FL system is able to schedule more devices in each communication round under certain MSE requirements. The proposed two-step alternating DC algorithm achieves a more accurate feasible set detection than the semidefinite relaxation (SDR) [29] technique, which simply drops the low-rank constraints. Moreover, our proposed algorithm enables FL to converge faster and achieve a more accurate prediction in the experiment of training image classifier models on the MNIST [30] and CIFAR-10 [31] datasets than the baseline schemes, including alternating SDR with IRS, random phase shifts with IRS, and the scheme without IRS.

B. Organization and Notations

The rest of this paper is organized as follows. Section II describes the system model in IRS-assisted AirComp-based FL system. Section III presents the convergence analysis and the problem formulation. In Section IV, we propose a two-step framework to tackle the problem. Section V presents a two-step alternating DC algorithm for solving the problem. The simulation results are provided in Section VI. Finally, Section VII concludes this work.

Italic, boldface lower-case, and boldface upper-case letters denote scalar, vector, and matrix, respectively. $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$ denote the real and complex domain with the space of $m\times n$, respectively. The operators $(\cdot)^{\mathsf{T}}$, $(\cdot)^{\mathsf{H}}$, $\mathrm{tr}(\cdot)$, and $\mathrm{diag}(\cdot)$ denote the transpose, Hermitian transpose, trace, and diagonal matrix, respectively. $\mathbb{E}[\cdot]$ denotes the statistical expectation. The operator $|\cdot|$ denotes the cardinality of a set or the absolute value of a scalar number, and $\|\cdot\|$ denotes the Euclidean norm.

II. System Model

In this section, we develop a computation and communication co-design for fast and reliable model aggregation in AirComp-based FL systems, where an IRS is deployed to compensate for the magnitude reduction and misalignment of AirComp.

A. FL Model

The IRS-assisted AirComp-based FL system under consideration consists of an M-antenna BS serving as an edge server, K single-antenna edge devices, and an IRS with N passive reflecting elements, as shown in Fig. 1. Edge device $k \in \mathcal{K} = \{1, 2, \ldots, K\}$ has its own local dataset $\mathcal{D}_k$ with $D_k = |\mathcal{D}_k|$ labeled data samples $\{(\boldsymbol{u}_i, v_i)\}_{i=1}^{D_k} \in \mathcal{D}_k$, where $(\boldsymbol{u}_i, v_i)$ denotes the input-output data pair consisting of feature $\boldsymbol{u}_i$ and its ground-truth label v_i. For a given d-dimensional model parameter $\boldsymbol{z} \in \mathbb{R}^d$, the local loss function for device k is defined as

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

where $f(\boldsymbol{z}; \boldsymbol{u}_i, v_i)$ denotes the sample-wise loss function. Without loss of generality, we assume that all local datasets have a uniform size, i.e., $D_k = D, \forall\, k \in \mathcal{K}$, as in [17].

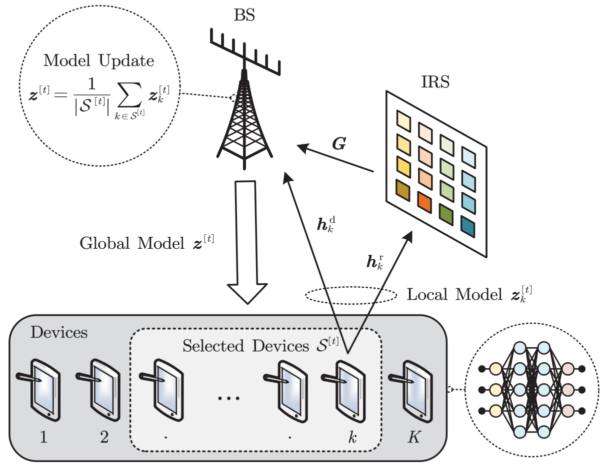


Fig. 1. Illustration of an IRS-assisted AirComp-based FL system.

Then, the global loss function with model parameter $\boldsymbol{z}$ can be represented as

$$F(\boldsymbol{z}) = \frac{1}{\sum_{k\in\mathcal{K}} D_k} \sum_{k\in\mathcal{K}} D_k F_k(\boldsymbol{z}) = \frac{1}{K}\sum_{k\in\mathcal{K}} F_k(\boldsymbol{z}). \tag{2}$$

The learning process aims to optimize the model parameter $\boldsymbol{z}$ that minimizes the global loss function, i.e.,

$$\boldsymbol{z}^\star = \arg\min_{\boldsymbol{z}\in\mathbb{R}^d} F(\boldsymbol{z}). \tag{3}$$

To achieve this purpose, we leverage the Federated Averaging (FedAvg) [4], also referred to as model averaging, to train a global model. Specifically, at the t-th communication round, the BS and the edge devices perform the following procedures:

- The BS broadcasts the current global model $\boldsymbol{z}^{[t-1]}$ to the edge devices belonging to a selected set, denoted as $\mathcal{S}^{[t]} \subseteq \mathcal{K}$.
- Based on the received global model $\boldsymbol{z}^{[t-1]}$, each edge device $k \in \mathcal{S}^{[t]}$ performs a local model update via the gradient descent algorithm with its local dataset $\mathcal{D}_k$ to obtain an updated local model, which is given by

$$\begin{aligned}\boldsymbol{z}_k^{[t]} &= \boldsymbol{z}^{[t-1]} - \varsigma^{[t]}\nabla F_k(\boldsymbol{z}^{[t-1]})\\ &= \boldsymbol{z}^{[t-1]} - \frac{\varsigma^{[t]}}{D_k}\sum_{(\boldsymbol{u}_i, v_i)\in\mathcal{D}_k} \nabla f(\boldsymbol{z}^{[t-1]}; \boldsymbol{u}_i, v_i),\end{aligned} \tag{4}$$

  where $\varsigma^{[t]}$ denotes the learning rate.
- All the local model updates are aggregated at the BS by taking an average to obtain an updated global model $\boldsymbol{z}^{[t]}$, which is given by

$$\boldsymbol{z}^{[t]} = \frac{1}{|\mathcal{S}^{[t]}|}\sum_{k\in\mathcal{S}^{[t]}} \boldsymbol{z}_k^{[t]}. \tag{5}$$

Note that our proposed algorithm can be directly extended to consider gradient aggregation by updating (4) and (5) accordingly.

In the following, we train image classifier models on the MNIST and CIFAR-10 datasets by using the FedAvg algorithm to show the impact of the number of selected devices on the test accuracy under different model aggregation errors. The

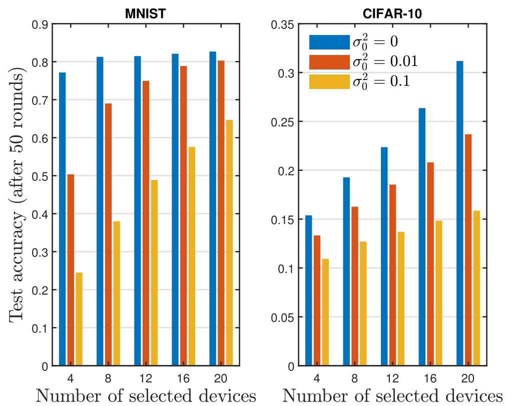


Fig. 2. Test accuracy versus the number of selected devices under different model aggregation errors.

aggregated global model at the t-th communication round is given by

$$\hat{\boldsymbol{z}}^{[t]} = \frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}} \boldsymbol{z}_k^{[t]} + \boldsymbol{e}_0^{[t]}, \tag{6}$$

where $\boldsymbol{e}_0^{[t]} \sim \mathcal{N}(0, \sigma_0^2 \boldsymbol{I}_d)$. As shown in Fig. 2, selecting more devices to participate in the training process is able to obtain a model that provides a higher test accuracy. Besides, the test accuracy decreases as the model aggregation error increases under the same number of selected devices. Therefore, it is critical to schedule more devices and reduce the aggregation error in each communication round for training a high quality model in wireless FL systems.

B. Communication Model for IRS-Assisted AirComp

Since the average sum in (5) for model aggregation falls into the category of nomographic functions [13], AirComp as a promising technique can be utilized to enhance the efficiency of model aggregation from distributed edge devices. For ease of presentation, we omit the time indices in this subsection. Let $\phi_k(\cdot)$ denote the normalization function at device k and $\psi(\cdot)$ denote the de-normalization function at the BS. The target function for aggregating the local model updates at the BS can then be expressed as

$$\boldsymbol{z} = \psi\left(\sum_{k\in\mathcal{S}} \phi_k(\boldsymbol{z}_k)\right), \tag{7}$$

where $\mathcal{S}$ is the device selection set. According to [17, Appendix A], we first normalize the local model via

$$\boldsymbol{s}_k = \phi_k(\boldsymbol{z}_k) = \frac{\boldsymbol{z}_k - \bar{z}}{\iota} = \frac{\boldsymbol{z}_k - \frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}}\bar{z}_k}{\frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}}\iota_k}, \ \forall k \in \mathcal{S}, \tag{8}$$

where $\bar{z}_k$ and ι_k denote the mean and standard deviation of d entries of the local model $\boldsymbol{z}_k$, respectively. They are defined as

$$\bar{z}_k = \frac{1}{d}\sum_{j=1}^{d} z_{k,j}, \ \iota_k^2 = \frac{1}{d}\sum_{j=1}^{d}\left(z_{k,j} - \bar{z}_k\right)^2, \ \forall k \in \mathcal{S}. \tag{9}$$

Therefore, the transmit symbols have zero mean and unit variance, i.e., $\mathbb{E}[\boldsymbol{s}_k \boldsymbol{s}_k^{\mathsf{T}}] = \boldsymbol{I}_d$, $\forall k \in \mathcal{S}$, and they are assumed to be independent of each other.

For edge devices with limited data storage and computing capacity, the dimension of the model parameters is often assumed to be tens of thousands of entries, which can be transmitted within one transmission interval in 5G networks [32]–[35], which is assumed in this paper. For high-dimensional model parameters, we can leverage specification and compressed sensing methods such as those proposed in [18], [19] to further reduce the model dimension, which is however out of the scope of this paper. Let $\tilde{s}_k$ denote a typical entry of $\boldsymbol{s}_k$ within one communication interval. Based on (8), the target function to be estimated at the BS is given by

$$r = \psi(g) = \frac{1}{|\mathcal{S}|}\left(\iota g + |\mathcal{S}|\bar{z}\right) \text{ with } g = \sum_{k\in\mathcal{S}} \tilde{s}_k. \tag{10}$$

However, the transmitted signals may encounter detrimental channel conditions during the model aggregation process through AirComp in the uplink, which leads to magnitude reduction and misalignment, thereby enlarging the aggregation error at the BS. To tackle this issue, we propose to deploy an IRS to alleviate the distortion of AirComp.

Let $\boldsymbol{h}_k^{\mathrm{d}} \in \mathbb{C}^M$, $\boldsymbol{h}_k^{\mathrm{r}} \in \mathbb{C}^N$, and $\boldsymbol{G} \in \mathbb{C}^{M\times N}$ denote the channel responses from device k to the BS, from device k to the IRS, and from the IRS to the BS, respectively. The channel gain of each link is assumed to be invariant within one transmission interval. In addition, many promising channel estimation methods have recently been proposed for IRS-assisted wireless networks [36]–[41]. Due to the high dimensionality of the model parameters that are allowed to be transmitted in one coherence block [35], the overhead introduced by channel estimation [40] is quite small by comparison. Therefore, we assume that perfect CSI is available at the BS and the overhead for channel estimation is negligible in this paper, as in [22]–[27]. The diagonal phase-shift matrix of the IRS is denoted by $\boldsymbol{\Theta} = \mathrm{diag}(\beta e^{j\theta_1}, \ldots, \beta e^{j\theta_N}) \in \mathbb{C}^{N\times N}$, where $\theta_n \in [0, 2\pi)$ denotes the phase shift of element n and $\beta \in [0,1]$ is the amplitude reflection coefficient on the incident signals. Without loss of generality, we assume $\beta = 1$ in this paper [23]. Compounded with reflected signals, the received signal at the BS is given by[1]

$$\boldsymbol{y} = \sum_{k\in\mathcal{S}} (\boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_k^{\mathrm{r}} + \boldsymbol{h}_k^{\mathrm{d}}) w_k \tilde{s}_k + \boldsymbol{n}, \tag{11}$$

where $w_k \in \mathbb{C}$ is the transmit scalar of device k for transmit power control and channel fading compensation, and $\boldsymbol{n} \sim \mathcal{CN}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right)$ is the additive white Gaussian noise (AWGN).

[1]In AirComp-based systems, synchronization is required among the selected devices [15], [42]. Since the IRS is passive and generally deployed close to BS/devices [43], the reflecting link introduced by IRS can be regarded as an additional propagation path similar to the reflecting and scattering propagation paths in wireless networks without IRS. Therefore, the synchronization methods used in wireless networks without IRS can be applied in our proposed system, such as sharing a reference-clock across distributed devices [44].

By denoting the aggregation beamforming vector at the BS as $\boldsymbol{m} \in \mathbb{C}^M$, we have

$$\begin{aligned}\hat{g} &= \frac{1}{\sqrt{\eta}}\boldsymbol{m}^{\mathsf{H}}\boldsymbol{y} \\ &= \frac{1}{\sqrt{\eta}}\boldsymbol{m}^{\mathsf{H}}\sum_{k\in\mathcal{S}}(\boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_k^{\mathrm{r}} + \boldsymbol{h}_k^{\mathrm{d}})w_k\tilde{s}_k + \frac{1}{\sqrt{\eta}}\boldsymbol{m}^{\mathsf{H}}\boldsymbol{n}, \end{aligned} \quad (12)$$

where η denotes the denoising factor for signal amplitude alignment and noise interference reduction at the BS. Thus, a typical entry of the global model at the BS can be updated as

$$\hat{r} = \frac{1}{|\mathcal{S}|}\left(\iota\hat{g} + |\mathcal{S}|\bar{z}\right). \quad (13)$$

After consecutively collecting d received signals $\hat{r}$ within one transmission interval, the global model can be generated by rearranging them into a d-dimensional vector.

III. Convergence Analysis and Problem Formulation

In this section, we provide the convergence analysis of our proposed IRS-aided AirComp-based FL system, which will guide us to formulate a joint learning-and-communication optimization problem.

A. Convergence Analysis

We make several standard assumptions on the loss functions and gradients as follows [45], [46].

Assumption 1 (Smoothness): The global loss function $F(\cdot)$ is L-smooth that satisfies

$$F(\tilde{\boldsymbol{z}}) \leq F(\boldsymbol{z}) + \nabla F(\boldsymbol{z})^{\mathsf{T}}(\tilde{\boldsymbol{z}} - \boldsymbol{z}) + \frac{L}{2}\|\tilde{\boldsymbol{z}} - \boldsymbol{z}\|^2, \quad \forall\, \boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathbb{R}^d. \quad (14)$$

Assumption 2 (Bounded Gradient): For any labeled data sample $(\boldsymbol{u}, v) \in \mathbb{R}^d \times \mathbb{R}$ and model parameter $\boldsymbol{z} \in \mathbb{R}^d$, we have

$$\|\nabla f(\boldsymbol{z}; \boldsymbol{u}, v)\|^2 \leq \kappa \quad (15)$$

for some constants $\kappa \geq 0$.

Assumption 3 (Bounded Variance): For any model parameter $\boldsymbol{z} \in \mathbb{R}^d$, the variance of its d entries is upper bounded, i.e., $\iota^2 \leq \Gamma$ for some constant $\Gamma \geq 0$, where

$$\iota^2 = \frac{1}{d}\sum_{j=1}^{d}\left(z_j - \frac{1}{d}\sum_{p=1}^{d} z_p\right)^2. \quad (16)$$

This assumption is reasonable because ι^2 is the variance of specific model parameters whose entries have finite values.

Based on above assumptions, we have the following lemma.

Lemma 1: Suppose that the loss functions satisfy the above assumptions. By setting $0 < \varsigma^{[t]} \equiv \varsigma \leq 1/L$, after T communication rounds, then the average norm of the global gradients is upper bounded by

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{z}^{[t]})\right\|^2\right] &\leq \frac{2}{\varsigma T}\left(F(\boldsymbol{z}^{[0]}) - F(\boldsymbol{z}^{\star})\right) \\ &+ \frac{2}{T}\sum_{t=0}^{T-1}\left(4\kappa\left(1 - \frac{|\mathcal{S}^{[t]}|}{K}\right)^2 + \frac{\Gamma d}{\varsigma^2|\mathcal{S}^{[t]}|^2}\mathsf{MSE}^{[t]}\right), \end{aligned} \quad (17)$$

where the expectation $\mathbb{E}[\cdot]$ is taken over the transmit symbols and AWGN variables, and

$$\begin{aligned}&\mathsf{MSE}^{[t]} \\ &= \mathsf{MSE}_j^{[t]} = \mathbb{E}\left[|\hat{g}_j^{[t]} - g_j^{[t]}|^2\right] \\ &= \sum_{k\in\mathcal{S}^{[t]}}\left|\frac{1}{\sqrt{\eta^{[t]}}}(\boldsymbol{m}^{[t]})^{\mathsf{H}}(\boldsymbol{G}^{[t]}\boldsymbol{\Theta}^{[t]}(\boldsymbol{h}_k^{\mathrm{r}})^{[t]} + (\boldsymbol{h}_k^{\mathrm{d}})^{[t]})w_k^{[t]} - 1\right|^2 \\ &\quad + \frac{\sigma^2\|\boldsymbol{m}^{[t]}\|^2}{\eta^{[t]}}, \ \forall j \in \{1, 2, \ldots, d\}. \end{aligned} \quad (18)$$

Proof: Please refer to Appendix A. □

B. Problem Formulation

It is observed that as $T \to \infty$, the reduction rate of the global gradient is mainly determined by the second term in (17). To mitigate the adverse effect of the second term on the convergence rate, we need to optimize the device selection set $\mathcal{S}^{[t]}$ and the model aggregation error quantified by $\mathsf{MSE}^{[t]}$ in each communication round, which is also indicated by the simulation results shown in Fig. 2. Therefore, we propose to maximize the number of selected devices while satisfying the given MSE requirement of the model aggregation in each communication round, with an objective of speeding up the convergence of the training process and avoiding notable reduction of the prediction accuracy. To simplify the notations, we omit the time indices in the following. Given the MSE requirement $\gamma > 0$ for model aggregation, the optimization problem can be formulated as[2]

$$\begin{aligned}&\underset{\substack{\mathcal{S}\subseteq\mathcal{K},\boldsymbol{m},\boldsymbol{\Theta},\\ \{w_k\},\eta}}{\text{maximize}} \ |\mathcal{S}| \\ &\text{subject to } \mathsf{MSE} \leq \gamma, \\ &\qquad |w_k|^2 \leq P_0, \quad \forall\, k \in \mathcal{S}, \\ &\qquad |\boldsymbol{\Theta}_{n,n}| = 1, \quad \forall\, n \in \{1, \ldots, N\}. \end{aligned} \quad (19)$$

The following proposition presents the optimal transmit scalars at the edge devices to minimize the MSE.

Proposition 1: Given the aggregation beamforming vector $\boldsymbol{m}$ and the phase-shift matrix $\boldsymbol{\Theta}$, the minimum MSE is obtained by using the following optimal transmit scalar

$$w_k^{\star} = \sqrt{\eta}\frac{(\boldsymbol{m}^{\mathsf{H}}(\boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_k^{\mathrm{r}} + \boldsymbol{h}_k^{\mathrm{d}}))^{\mathsf{H}}}{|\boldsymbol{m}^{\mathsf{H}}(\boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_k^{\mathrm{r}} + \boldsymbol{h}_k^{\mathrm{d}})|^2}, \ \forall\, k \in \mathcal{S}. \quad (20)$$

Proof: Please refer to Appendix B. □

The transmit power of device k is constrained by a given maximum transmit power $P_0 > 0$, i.e., $|w_k|^2 \leq P_0$. With the optimal transmit scalar $w_k^{\star}$ given in (20), we have

$$\eta^{\star} = P_0 \min_{k\in\mathcal{S}}|\boldsymbol{m}^{\mathsf{H}}(\boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_k^{\mathrm{r}} + \boldsymbol{h}_k^{\mathrm{d}})|^2. \quad (21)$$

[2]By exploiting the channel reciprocity, the BS with full CSI and strong computation capabilities can efficiently solve problem (19) with the methods described in the following, and then inform each scheduled device of its transmit scalar. Thus, the devices do not need to gather CSI for transmit scalar optimization. Besides, the phase shift design of the IRS can also be transmitted simultaneously with the transmit scalars from the BS to the smart controller attached to the IRS via a separate control link [23].

Therefore, the minimum MSE is given by

$$\mathsf{MSE}=\frac{\sigma^2}{P_0}\max_{k\in\mathcal{S}}\frac{\|\boldsymbol{m}\|^2}{\left|\boldsymbol{m}^{\mathsf{H}}(\boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_k^{\mathrm{r}}+\boldsymbol{h}_k^{\mathrm{d}})\right|^2}. \tag{22}$$

As a result, problem (19) can be rewritten as

$$\begin{aligned}&\underset{\mathcal{S}\subseteq\mathcal{K},\boldsymbol{m},\boldsymbol{\Theta}}{\text{maximize}}\ |\mathcal{S}| && \text{(23a)}\\ &\text{subject to}\ \max_{k\in\mathcal{S}}\frac{\|\boldsymbol{m}\|^2}{\left|\boldsymbol{m}^{\mathsf{H}}(\boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_k^{\mathrm{r}}+\boldsymbol{h}_k^{\mathrm{d}})\right|^2}\le\tilde{\gamma}, && \text{(23b)}\\ &\qquad\qquad |\boldsymbol{\Theta}_{n,n}|=1,\ \forall\, n\in\{1,\ldots,N\}, && \text{(23c)}\end{aligned}$$

where $\tilde{\gamma}=\gamma P_0/\sigma^2$. To facilitate the algorithm design, the MSE constraint (23b) can be rewritten as nonconvex constraints with quadratic and biquadratic terms, as presented in Proposition 2.

Proposition 2: The constraint (23b) *can be equivalently rewritten as the following constraints*:

$$\|\boldsymbol{m}\|^2-\tilde{\gamma}\,|\boldsymbol{m}^{\mathsf{H}}(\boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_k^{\mathrm{r}}+\boldsymbol{h}_k^{\mathrm{d}})|^2\le 0,\quad \forall\, k\in\mathcal{S}, \tag{24}$$

where $\|\boldsymbol{m}\|^2\ge 1$.

Proof: Please refer to Appendix C. □

According to Proposition 2, the objective function $|\mathcal{S}|$ represents the number of feasible MSE constraints (24), which should be maximized under the regularity condition $\|\boldsymbol{m}\|^2\ge 1$. By adding an auxiliary variable $\boldsymbol{x}$ [47], we equivalently transform the problem of maximizing the number of feasible MSE constraints into the problem of minimizing the number of nonzero x_k's. Hence, we turn to solve the following sparse optimization problem

$$\begin{aligned}&\underset{\boldsymbol{x}\in\mathbb{R}_+^K,\boldsymbol{m},\boldsymbol{\Theta}}{\text{minimize}}\ \ \|\boldsymbol{x}\|_0 && \text{(25a)}\\ &\text{subject to}\ \ \|\boldsymbol{m}\|^2-\tilde{\gamma}\,|\boldsymbol{m}^{\mathsf{H}}(\boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_k^{\mathrm{r}}+\boldsymbol{h}_k^{\mathrm{d}})|^2\le x_k,\ \forall\, k\in\mathcal{K}, && \text{(25b)}\\ &\qquad\qquad \|\boldsymbol{m}\|^2\ge 1, && \text{(25c)}\\ &\qquad\qquad |\boldsymbol{\Theta}_{n,n}|=1,\quad \forall\, n\in\{1,\ldots,N\}. && \text{(25d)}\end{aligned}$$

Note that the selection of each edge device is indicated by the sparse structure of $\boldsymbol{x}$, i.e., $x_k=0$ indicates that device k can be selected while satisfying the MSE requirement. Due to the sparse objective function and nonconvex constraints with biquadratic (25b) and quadratic (25c) terms, problem (25) is computationally difficult. To tackle this issue, we shall propose a two-step alternating low-rank optimization framework in the following section.

IV. Alternating Low-Rank Optimization Framework for Model Aggregation

In this section, we propose a two-step framework to solve problem (25) for IRS-assisted AirComp-based FL with device selection, followed by proposing to use the alternating optimization approach to solve the problem in each step.

A. Proposed Two-Step Framework

The main idea of our proposed two-step framework is to induce the sparsity of $\boldsymbol{x}$ in the first step, so as to determine the priority for each device to be selected. With the obtained priority vector, we then solve a series of MSE minimization problems to find the maximum feasible device set while satisfying the MSE requirement in the second step.

1) Sparsity Inducing: For the nonconvex sparse objective function being in the form of ℓ_0-norm, we adopt the well-recognized ℓ_1-norm as a convex surrogate [48]. To solve problem $\mathscr{P}$, we shall solve the following problem in the first step

$$\begin{aligned}\mathscr{P}_1:\ &\underset{\boldsymbol{x}\in\mathbb{R}_+^K,\boldsymbol{m},\boldsymbol{\Theta}}{\text{minimize}}\ \|\boldsymbol{x}\|_1\\ &\text{subject to constraints (25b), (25c), (25d).}\end{aligned} \tag{26}$$

After solving problem $\mathscr{P}_1$, we proceed to the second step to check the feasibility of the selected devices and find the maximum number of edge devices under the MSE constraint.

2) Feasibility Detection: The value of x_k obtained from the first step characterizes the disparity between the MSE requirement and the achievable MSE for device k. Therefore, the smaller the value of x_k, the higher priority device k being selected in the second step. We sort $\{x_k\}_{k=1}^K$ in an ascending order $x_{\pi(1)}\le x_{\pi(2)}\le\cdots\le x_{\pi(K)}$ to determine the priority of edge devices, where $x_{\pi(k)}$ denotes the k-th smallest element in $\{x_k\}_{k=1}^K$. We adopt the bisection method to find the maximum value of $\tilde{k}$ that enables all devices in the set $\mathcal{S}_{\tilde{k}}=\{\pi(1),\pi(2),\ldots,\pi(\tilde{k})\}$ to be feasibly selected. Specifically, for a given device selection set $\mathcal{S}_{\tilde{k}}$, we check the feasibility via comparing the MSE requirement γ with the MSE of selected devices in $\mathcal{S}_{\tilde{k}}$ obtained from the following problem

$$\begin{aligned}\mathscr{P}_2:\ &\underset{\boldsymbol{m},\boldsymbol{\Theta}}{\text{minimize}}\ \max_{k\in\mathcal{S}_{\tilde{k}}}\frac{\|\boldsymbol{m}\|^2}{|\boldsymbol{m}^{\mathsf{H}}(\boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_k^{\mathrm{r}}+\boldsymbol{h}_k^{\mathrm{d}})|^2} && \text{(27a)}\\ &\text{subject to}\ |\boldsymbol{\Theta}_{n,n}|=1,\quad \forall\, n\in\{1,\ldots,N\}. && \text{(27b)}\end{aligned}$$

If the optimal objective value of problem (27) is less than the required MSE, then set $\mathcal{S}_{\tilde{k}}$ is considered as a feasible set.

B. Alternating Low-Rank Optimization

It can be observed that constraint (25b) and objective function (27a) are both nonconvex due to the coupled optimization variables. To address this issue, we propose to apply alternating optimization.

1) Sparsity Inducing: In the first step, variables $(\boldsymbol{x},\boldsymbol{m})$ and $\boldsymbol{\Theta}$ of problem $\mathscr{P}_1$ can be optimized alternately. Specifically, when the phase-shift matrix $\boldsymbol{\Theta}$ is fixed (i.e., the combined channel vector $\boldsymbol{h}_k=\boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_k^{\mathrm{r}}+\boldsymbol{h}_k^{\mathrm{d}}$ between device k and the BS is fixed), the problem can be expressed as

$$\begin{aligned}&\underset{\boldsymbol{x}\in\mathbb{R}_+^K,\boldsymbol{m}}{\text{minimize}}\ \|\boldsymbol{x}\|_1\\ &\text{subject to constraints (25b), (25c).}\end{aligned} \tag{28}$$

To address the nonconvexity of biquadratic and quadratic constraints (25b) and (25c), we further transform problem (28) into a semidefinite programming (SDP) problem via the matrix lifting technique [29]. By denoting $\boldsymbol{M}=\boldsymbol{m}\boldsymbol{m}^{\mathsf{H}}$ and $\boldsymbol{H}_k=\boldsymbol{h}_k\boldsymbol{h}_k^{\mathsf{H}}$, problem (28) can be rewritten as a low-rank optimization problem

$$\begin{aligned}\mathscr{P}_{1,1}:\ &\underset{\boldsymbol{x}\in\mathbb{R}_+^K,\boldsymbol{M}}{\text{minimize}}\ \|\boldsymbol{x}\|_1\\ &\text{subject to}\ \operatorname{tr}(\boldsymbol{M})-\tilde{\gamma}\operatorname{tr}(\boldsymbol{M}\boldsymbol{H}_k)\le x_k,\ \forall\, k\in\mathcal{K},\\ &\qquad\operatorname{tr}(\boldsymbol{M})\ge 1,\\ &\qquad\boldsymbol{M}\succeq\boldsymbol{0},\ \operatorname{rank}(\boldsymbol{M})=1.\end{aligned}\tag{29}$$

On the other hand, when the auxiliary vector $\boldsymbol{x}$ and the aggregation beamforming vector $\boldsymbol{m}$ are fixed, problem $\mathscr{P}_1$ is reduced to be a feasibility detection problem of phase-shift matrix $\boldsymbol{\Theta}$. By denoting $\boldsymbol{v}=[e^{j\theta_1},\ldots,e^{j\theta_N}]^{\mathsf{T}}$, $\boldsymbol{a}_k^{\mathsf{H}}=\boldsymbol{m}^{\mathsf{H}}\boldsymbol{G}\operatorname{diag}(\boldsymbol{h}_k^{\mathrm{r}})$, and $c_k=\boldsymbol{m}^{\mathsf{H}}\boldsymbol{h}_k^{\mathrm{d}}$, the problem can be expressed as

$$\text{find}\ \boldsymbol{v}\tag{30a}$$
$$\text{subject to}\ \|\boldsymbol{m}\|^2-\tilde{\gamma}\,|\boldsymbol{a}_k^{\mathsf{H}}\boldsymbol{v}+c_k|^2\le x_k,\ \forall\, k\in\mathcal{K},\tag{30b}$$
$$|v_n|=1,\ \forall\, n\in\{1,\ldots,N\}.\tag{30c}$$

We denote $\tilde{\boldsymbol{v}}=[\boldsymbol{v},t]^{\mathsf{T}}$ by introducing an auxiliary variable t, and thus constraints (30b) can be rewritten as

$$\|\boldsymbol{m}\|^2-\tilde{\gamma}\,(\tilde{\boldsymbol{v}}^{\mathsf{H}}\boldsymbol{R}_k\tilde{\boldsymbol{v}}+|c_k|^2)\le x_k,\ \forall\, k\in\mathcal{K},\quad\text{with}\ \boldsymbol{R}_k=\begin{bmatrix}\boldsymbol{a}_k\boldsymbol{a}_k^{\mathsf{H}} & c_k\boldsymbol{a}_k\\ c_k^*\boldsymbol{a}_k^{\mathsf{H}} & 0\end{bmatrix}.\tag{31}$$

Since $\tilde{\boldsymbol{v}}^{\mathsf{H}}\boldsymbol{R}_k\tilde{\boldsymbol{v}}=\operatorname{tr}(\boldsymbol{R}_k\tilde{\boldsymbol{v}}\tilde{\boldsymbol{v}}^{\mathsf{H}})$, we lift $\tilde{\boldsymbol{v}}$ as a positive semidefinite (PSD) matrix $\boldsymbol{V}=\tilde{\boldsymbol{v}}\tilde{\boldsymbol{v}}^{\mathsf{H}}$ with $\operatorname{rank}(\boldsymbol{V})=1$. Hence, problem (30) can be equivalently reformulated as the following low-rank matrix optimization problem

$$\begin{aligned}\mathscr{P}_{1,2}:\ &\text{find}\ \boldsymbol{V}\\ &\text{subject to}\ \|\boldsymbol{m}\|^2-\tilde{\gamma}\,(\operatorname{tr}(\boldsymbol{R}_k\boldsymbol{V})+|c_k|^2)\le x_k,\\ &\qquad\forall\, k\in\mathcal{K},\\ &\qquad\boldsymbol{V}_{n,n}=1,\ \forall\, n\in\{1,\ldots,N+1\},\\ &\qquad\boldsymbol{V}\succeq\boldsymbol{0},\ \operatorname{rank}(\boldsymbol{V})=1.\end{aligned}\tag{32}$$

2) Feasibility Detection: In the second step, we first reformulate problem $\mathscr{P}_2$ as the following problem [26]

$$\begin{aligned}&\underset{\boldsymbol{m},\boldsymbol{\Theta}}{\text{minimize}}\ \|\boldsymbol{m}\|^2\\ &\text{subject to}\ |\boldsymbol{m}^{\mathsf{H}}(\boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_k^{\mathrm{r}}+\boldsymbol{h}_k^{\mathrm{d}})|^2\ge 1,\ \forall\, k\in\mathcal{S}_{\tilde{k}},\\ &\qquad|\boldsymbol{\Theta}_{n,n}|=1,\ \forall\, n\in\{1,\ldots,N\}.\end{aligned}\tag{33}$$

To decouple the optimization variables, we optimize the aggregation beamforming vector $\boldsymbol{m}$ and the phase-shift matrix $\boldsymbol{\Theta}$ alternately. Specifically, given the phase-shift matrix $\boldsymbol{\Theta}$, we have

$$\begin{aligned}&\underset{\boldsymbol{m}}{\text{minimize}}\ \|\boldsymbol{m}\|^2\\ &\text{subject to}\ |\boldsymbol{m}^{\mathsf{H}}\boldsymbol{h}_k|^2\ge 1,\ \forall\, k\in\mathcal{S}_{\tilde{k}}.\end{aligned}\tag{34}$$

This problem can be further represented as a low-rank matrix optimization problem:

$$\begin{aligned}\mathscr{P}_{2,1}:\ &\underset{\boldsymbol{M}}{\text{minimize}}\ \operatorname{tr}(\boldsymbol{M})\\ &\text{subject to}\ \operatorname{tr}(\boldsymbol{M}\boldsymbol{H}_k)\ge 1,\ \forall\, k\in\mathcal{S}_{\tilde{k}},\\ &\qquad\boldsymbol{M}\succeq\boldsymbol{0},\ \operatorname{rank}(\boldsymbol{M})=1.\end{aligned}\tag{35}$$

On the other hand, given the aggregation beamforming vector $\boldsymbol{m}$, we have

$$\begin{aligned}&\text{find}\ \boldsymbol{v}\\ &\text{subject to}\ |\boldsymbol{a}_k^{\mathsf{H}}\boldsymbol{v}+c_k|^2\ge 1,\ \forall\, k\in\mathcal{S}_{\tilde{k}},\\ &\qquad|v_n|=1,\ \forall\, n\in\{1,\ldots,N\},\end{aligned}\tag{36}$$

and its corresponding low-rank matrix optimization problem is given by

$$\begin{aligned}\mathscr{P}_{2,2}:\ &\text{find}\ \boldsymbol{V}\\ &\text{subject to}\ \operatorname{tr}(\boldsymbol{R}_k\boldsymbol{V})+|c_k|^2\ge 1,\ \forall\, k\in\mathcal{S}_{\tilde{k}},\\ &\qquad\boldsymbol{V}_{n,n}=1,\ \forall\, n\in\{1,\ldots,N+1\},\\ &\qquad\boldsymbol{V}\succeq\boldsymbol{0},\ \operatorname{rank}(\boldsymbol{V})=1.\end{aligned}\tag{37}$$

The resulting problems $\mathscr{P}_{1,1}$, $\mathscr{P}_{1,2}$, $\mathscr{P}_{2,1}$, and $\mathscr{P}_{2,2}$ in the alternating low-rank optimization are still nonconvex because of the fixed rank-one constraints. This nonconvexity issue can be tackled by simply dropping the rank-one constraints via the SDR technique, leading to standard SDP problems [29]. If the obtained solution fails to be rank-one, the Gaussian randomization method can be adopted to obtain a suboptimal solution [29]. However, if the number of antennas and the number of reflecting elements are large, the performance of the SDR technique degenerates in the resulting high-dimensional optimization problems due to the low probability of returning rank-one solutions [49]. To address the limitations of the SDR technique, we present a DC programming approach for inducing rank-one solutions in the next section.

V. Alternating DC Approach for Low-Rank Optimization

In this section, we present a DC formulation for the rank-one constrained SDP problems in the alternating procedure, followed by proposing a two-step alternating DC algorithm to solve problem (19) in IRS-assisted AirComp-based FL systems.

A. DC Formulation for Rank-One Constrained Problems

The accurate detection of the rank-one constraint plays a critical role in precisely detecting the feasibility of nonconvex quadratic constraints, which is important in our two-step framework for device selection. Therefore, we provide a DC representation for the rank-one constraints in the aforementioned problems in Section IV-B.

The rank-one constraint of PSD matrix $\boldsymbol{M}\in\mathbb{C}^{M\times M}$ can be equivalently rewritten as

$$\|\{\sigma_i(\boldsymbol{M})\}_{i=1}^M\|_0=1,\tag{38}$$

where $\sigma_i(\boldsymbol{M})$ denotes the i-th largest singular value of matrix $\boldsymbol{M}$. Furthermore, since the trace norm and the spectral norm are represented by

$$\operatorname{tr}(\boldsymbol{M})=\sum_{i=1}^{M}\sigma_i(\boldsymbol{M})\ \text{and}\ \|\boldsymbol{M}\|_2=\sigma_1(\boldsymbol{M}),\tag{39}$$

respectively, we have [16]

$$\operatorname{rank}(\boldsymbol{M})=1\Leftrightarrow\operatorname{tr}(\boldsymbol{M})-\|\boldsymbol{M}\|_2=0\tag{40}$$

with $\mathrm{tr}(\boldsymbol{M}) > 0$. Therefore, we can use a DC penalty to induce rank-one solutions. The corresponding DC formulation for problem $\mathscr{P}_{1,1}$ is given by

$$\begin{aligned}\underset{\boldsymbol{x}\in\mathbb{R}_+^K,\boldsymbol{M}}{\text{minimize}}\ & \|\boldsymbol{x}\|_1 + \rho\left(\mathrm{tr}(\boldsymbol{M}) - \|\boldsymbol{M}\|_2\right)\\ \text{subject to}\ & \mathrm{tr}(\boldsymbol{M}) - \tilde{\gamma}\,\mathrm{tr}(\boldsymbol{M}\boldsymbol{H}_k) \le x_k,\ \forall\, k\in\mathcal{K},\\ & \mathrm{tr}(\boldsymbol{M}) \ge 1,\ \boldsymbol{M}\succeq \boldsymbol{0},\end{aligned} \tag{41}$$

where $\rho > 0$ denotes the penalty parameter. Hence, we are able to obtain a rank-one matrix when the DC penalty term is enforced to be zero. Then, the feasible aggregation beamforming vector $\boldsymbol{m}$ of problem $\mathscr{P}_1$ can be recovered by utilizing Cholesky decomposition for $\boldsymbol{M}^\star = \boldsymbol{m}\boldsymbol{m}^{\mathsf{H}}$. Similarly, we detect the feasibility of problem $\mathscr{P}_{1,2}$ by minimizing the DC representation term that is given by

$$\begin{aligned}\underset{\boldsymbol{V}}{\text{minimize}}\ & \mathrm{tr}(\boldsymbol{V}) - \|\boldsymbol{V}\|_2\\ \text{subject to}\ & \|\boldsymbol{m}\|^2 - \tilde{\gamma}\left(\mathrm{tr}(\boldsymbol{R}_k\boldsymbol{V}) + |c_k|^2\right) \le x_k,\ \forall\, k\in\mathcal{K},\\ & \boldsymbol{V}\succeq\boldsymbol{0},\ \boldsymbol{V}_{n,n} = 1,\ \forall\, n\in\{1,\ldots,N+1\}.\end{aligned} \tag{42}$$

Once the objective value becomes zero, we can obtain an exact rank-one feasible solution and extract $\tilde{\boldsymbol{v}} = [\boldsymbol{v}_0, t_0]^{\mathsf{T}}$ by utilizing Cholesky decomposition for $\boldsymbol{V}^\star = \tilde{\boldsymbol{v}}\tilde{\boldsymbol{v}}^{\mathsf{H}}$. Then, by computing $\boldsymbol{v} = \boldsymbol{v}_0/t_0$, the phase-shift matrix can be recovered according to $\boldsymbol{\Theta} = \mathrm{diag}(\boldsymbol{v})$.

Problems $\mathscr{P}_{2,1}$ and $\mathscr{P}_{2,2}$ in the second step can be reformulated in the similar DC formulation to guarantee the feasibility of the rank-one constraint, which are rewritten as

$$\begin{aligned}\underset{\boldsymbol{M}}{\text{minimize}}\ & \mathrm{tr}(\boldsymbol{M}) + \rho\left(\mathrm{tr}(\boldsymbol{M}) - \|\boldsymbol{M}\|_2\right)\\ \text{subject to}\ & \mathrm{tr}(\boldsymbol{M}\boldsymbol{H}_k) \ge 1,\ \forall\, k\in\mathcal{S}_{\tilde{k}},\\ & \boldsymbol{M}\succeq\boldsymbol{0},\end{aligned} \tag{43}$$

and

$$\begin{aligned}\underset{\boldsymbol{V}}{\text{minimize}}\ & \mathrm{tr}(\boldsymbol{V}) - \|\boldsymbol{V}\|_2\\ \text{subject to}\ & \mathrm{tr}(\boldsymbol{R}_k\boldsymbol{V}) + |c_k|^2 \ge 1,\ \forall\, k\in\mathcal{S}_{\tilde{k}},\\ & \boldsymbol{V}\succeq\boldsymbol{0},\ \boldsymbol{V}_{n,n} = 1,\ \forall\, n\in\{1,\ldots,N+1\},\end{aligned} \tag{44}$$

respectively.

B. DC Algorithm

Although the DC programs are still nonconvex, their problem structures of minimizing the difference of two convex functions can be exploited to develop an efficient DC algorithm [50] by successively linearizing the concave part. Specifically, the penalty terms can be approximated by

$$\mathrm{tr}(\boldsymbol{X}) - \|\boldsymbol{X}\|_2 \le \mathrm{tr}(\boldsymbol{X}) - \left(\|\boldsymbol{X}^{q-1}\|_2 + \langle\partial_{\boldsymbol{X}^{q-1}}\|\boldsymbol{X}\|_2, \boldsymbol{X}\rangle\right), \tag{45}$$

where $\langle\boldsymbol{X},\boldsymbol{Y}\rangle = \Re[\mathrm{tr}(\boldsymbol{X}^{\mathsf{H}}\boldsymbol{Y})]$ denotes the inner product of two matrices, and $\partial_{\boldsymbol{X}^{q-1}}\|\boldsymbol{X}\|_2$ denotes the subgradient of $\|\boldsymbol{X}\|_2$ with respect to the solution obtained at iteration $q-1$. Here, $\partial_{\boldsymbol{X}^{q-1}}\|\boldsymbol{X}\|_2$ can be efficiently computed by $\boldsymbol{\varpi}_1\boldsymbol{\varpi}_1^{\mathsf{H}}$ [16], where $\boldsymbol{\varpi}_1$ is the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{X}^{q-1}$. Consequently, in the first step, the resulting subproblems of (41) and (42) at the q-th iteration are, respectively, given by

$$\begin{aligned}\underset{\boldsymbol{x}\in\mathbb{R}_+^K,\boldsymbol{M}}{\text{minimize}}\ & \|\boldsymbol{x}\|_1 + \rho\left(\mathrm{tr}(\boldsymbol{M}) - \langle\partial_{\boldsymbol{M}^{q-1}}\|\boldsymbol{M}\|_2, \boldsymbol{M}\rangle\right)\\ \text{subject to}\ & \mathrm{tr}(\boldsymbol{M}) - \tilde{\gamma}\,\mathrm{tr}(\boldsymbol{M}\boldsymbol{H}_k) \le x_k,\ \forall\, k\in\mathcal{K},\\ & \mathrm{tr}(\boldsymbol{M}) \ge 1,\ \boldsymbol{M}\succeq\boldsymbol{0},\end{aligned} \tag{46}$$

and

$$\begin{aligned}\underset{\boldsymbol{V}}{\text{minimize}}\ & \mathrm{tr}(\boldsymbol{V}) - \langle\partial_{\boldsymbol{V}^{q-1}}\|\boldsymbol{V}\|_2, \boldsymbol{V}\rangle\\ \text{subject to}\ & \|\boldsymbol{m}\|^2 - \tilde{\gamma}\left(\mathrm{tr}(\boldsymbol{R}_k\boldsymbol{V}) + |c_k|^2\right) \le x_k,\ \forall\, k\in\mathcal{K},\\ & \boldsymbol{V}\succeq\boldsymbol{0},\ \boldsymbol{V}_{n,n} = 1,\ \forall\, n\in\{1,\ldots,N+1\}.\end{aligned} \tag{47}$$

Likewise, in the second step, we can transform the DC programs (43) and (44) into a series of subproblems as

$$\begin{aligned}\underset{\boldsymbol{M}}{\text{minimize}}\ & (1+\rho)\,\mathrm{tr}(\boldsymbol{M}) - \rho\,\langle\partial_{\boldsymbol{M}^{q-1}}\|\boldsymbol{M}\|_2, \boldsymbol{M}\rangle\\ \text{subject to}\ & \mathrm{tr}(\boldsymbol{M}\boldsymbol{H}_k) \ge 1,\ \forall\, k\in\mathcal{S}_{\tilde{k}},\\ & \boldsymbol{M}\succeq\boldsymbol{0},\end{aligned} \tag{48}$$

and

$$\begin{aligned}\underset{\boldsymbol{V}}{\text{minimize}}\ & \mathrm{tr}(\boldsymbol{V}) - \langle\partial_{\boldsymbol{V}^{q-1}}\|\boldsymbol{V}\|_2, \boldsymbol{V}\rangle\\ \text{subject to}\ & \mathrm{tr}(\boldsymbol{R}_k\boldsymbol{V}) + |c_k|^2 \ge 1,\ \forall\, k\in\mathcal{S}_{\tilde{k}},\\ & \boldsymbol{V}\succeq\boldsymbol{0},\ \boldsymbol{V}_{n,n} = 1,\ \forall\, n\in\{1,\ldots,N+1\},\end{aligned} \tag{49}$$

respectively.

It can be verified that the above subproblems are convex and thus can be efficiently solved by using CVX [51]. Furthermore, it has been shown in [50] that the solving procedure with the DC algorithm always converges to the critical points of the DC programs from any feasible initial points. In summary, the entire proposed two-step alternating DC algorithm for solving the sparse and low-rank optimization problem (19) is presented in Algorithm 1.

C. Computation Complexity Analysis

In our proposed Algorithm 1, we need to solve a sequence of SDP problems (46), (47) in the first step, and (48), (49) in the second step. To solve each SDP problem, the worst-case computational complexity by using the interior-point method [29] is $\mathcal{O}(\max\{M,K\}^4\ M^{1/2}\log(1/\epsilon_s))$ in problems (46) and (48), and is $\mathcal{O}(\max\{N,K\}^4\ N^{1/2}\log(1/\epsilon_s))$ in problems (47) and (49), where $\epsilon_s > 0$ denotes the solution precision. Supposing those problems converging to critical points of the DC programs with $Q > 1$ iterations, the computational cost of solving a DC program, i.e., one of problems (41), (42), (43), and (44), is $\mathcal{O}(Q\max\{M,K\}^4\ M^{1/2}\log(1/\epsilon_s))$ or $\mathcal{O}(Q\max\{N,K\}^4\ N^{1/2}\log(1/\epsilon_s))$. Note that we merely need to solve the SDP problem once for problems $\mathscr{P}_{1,1}$, $\mathscr{P}_{1,2}$, $\mathscr{P}_{2,1}$, and $\mathscr{P}_{2,2}$ via SDR technique with simply dropping the rank-one constraints, i.e., $Q = 1$ in this case. The proposed DC algorithm has a higher computational complexity than the SDR method. Nevertheless, the sacrifice of the computational complexity results in significant improvement on the system performance, which will be demonstrated in the following section.

Algorithm 1 Two-Step Alternating DC Algorithm for Solving Problem (19) in FL With Device Selection

Step 1: Sparsity Inducing
Input: Initial point $\mathbf{\Theta}^0$ and predefined threshold $\epsilon > 0$.
for $t_o \leftarrow 1, 2, \ldots$ **do**
  Given $\mathbf{\Theta}^{t_o - 1}$, obtain solution $(\boldsymbol{x}^{t_o}, \boldsymbol{m}^{t_o})$ by solving problem (41).
  Given $(\boldsymbol{x}^{t_o}, \boldsymbol{m}^{t_o})$, obtain solution $\mathbf{\Theta}^{t_o}$ by solving problem (42).
  if *Decrease of the objective value of problem* $\mathscr{P}_1$ *is below* ϵ **then**
    break.
  end
end
Output: $\boldsymbol{x}^\star \leftarrow \boldsymbol{x}^{t_o}$.

Step 2: Feasibility Detection
Input: Set $\mathcal{S}_K = \{\pi(1), \pi(2), \ldots, \pi(K)\}$ obtained by ordering $\boldsymbol{x}^\star$ in an ascending order as $x_{\pi(1)} \leq \cdots \leq x_{\pi(K)}$, $K_{\text{low}} \leftarrow 0$, $K_{\text{up}} \leftarrow K$, and predefined threshold $\epsilon > 0$.
$\tilde{k} \leftarrow K$.
while $K_{\text{up}} - K_{\text{low}} > 1$ **do**
  Initialize $\mathbf{\Theta}^0$ and $\mathcal{S}_{\tilde{k}} \leftarrow \left\{\pi(1), \pi(2), \ldots, \pi(\tilde{k})\right\}$.
  for $t_o \leftarrow 1, 2, \ldots$ **do**
    Given $\mathbf{\Theta}^{t_o - 1}$, obtain solution $\boldsymbol{m}^{t_o}$ by solving problem (43).
    if $\mathsf{MSE} \leq \gamma$ **then**
      $K_{\text{low}} \leftarrow \tilde{k}$.
      $\tilde{\boldsymbol{m}} \leftarrow \boldsymbol{m}^{t_o}$.
      $\tilde{k} \leftarrow \lfloor \frac{K_{\text{low}} + K_{\text{up}}}{2} \rfloor$.
      Break.
    else if *Decrease of the objective value of problem* $\mathscr{P}_2$ *is below* ϵ **then**
      $K_{\text{up}} \leftarrow \tilde{k}$.
      $\tilde{k} \leftarrow \lfloor \frac{K_{\text{low}} + K_{\text{up}}}{2} \rfloor$.
      Break.
    end
    Given $\boldsymbol{m}^{t_o}$, obtain solution $\mathbf{\Theta}^{t_o}$ by solving problem (44).
  end
end
Output: Aggregation beamforming vector $\boldsymbol{m}^\star \leftarrow \tilde{\boldsymbol{m}}$ and the set of selected devices $\mathcal{S}^\star \leftarrow \{\pi(1), \pi(2), \ldots, \pi(k^\star)\}$ with $k^\star \leftarrow K_{\text{low}}$.

VI. Simulation Results

In this section, we present the simulation results to demonstrate the advantages of the proposed two-step alternating DC algorithm for FL with device selection. The effectiveness of deploying an IRS for the AirComp-based FL system will also be illustrated. We consider a three-dimensional coordinate system, where the antennas at the BS and the reflecting elements at the IRS are placed as a uniform linear array and a uniform rectangular array, respectively. The locations of the BS and the IRS are, respectively, set as $(3, 0, 6)$ meters and $(0, 100, 6)$ meters, while the edge devices are distributed in the region of $([0, 6], [100, 106], 0)$ meters surrounding the IRS. The path loss model is given by

$$B(d) = C_0 (l/l_0)^{-\alpha}, \tag{50}$$

where C_0 denotes the path loss at the reference distance $l_0 = 1$ meter, l is the link distance, and α is the path loss exponent. All channels are assumed to suffer from Rician fading [23], where the channel coefficient can be expressed as

$$\boldsymbol{\varrho} = \sqrt{\frac{\chi}{1+\chi}} \boldsymbol{\varrho}_{\text{LoS}} + \sqrt{\frac{1}{1+\chi}} \boldsymbol{\varrho}_{\text{NLoS}}, \tag{51}$$

where χ is the Rician factor, $\boldsymbol{\varrho}_{\text{LoS}}$ denotes the line-of-sight (LoS) component, and $\boldsymbol{\varrho}_{\text{NLoS}}$ denotes the non-line-of-sight (NLoS) component. In our simulations, the channel coefficients are given by $\boldsymbol{G} = \sqrt{B(d_{\text{BI}})} \boldsymbol{\varrho}_{\text{BI}}$, $\boldsymbol{h}_k^{\text{r}} = \sqrt{B(d_{\text{ID},k})} \boldsymbol{\varrho}_{\text{ID},k}$, and $\boldsymbol{h}_k^{\text{d}} = \sqrt{B(d_{\text{BD},k})} \boldsymbol{\varrho}_{\text{BD},k}$, where d_{BI}, $d_{\text{ID},k}$ and $d_{\text{BD},k}$ denote the distance between BS and IRS, the distance between IRS and device k, and the distance between BS and device k, respectively. As in [52], the Rician factors of $\boldsymbol{\varrho}_{\text{IB}}$, $\boldsymbol{\varrho}_{\text{DI},k}$, and $\boldsymbol{\varrho}_{\text{DB},k}$ are set to be 3 dB, 0, and 0, respectively, and the path loss exponents for the BS-device channel, the BS-IRS channel, and the IRS-device channel are set to be 3.6, 2.2, and 2.8, respectively. Unless stated otherwise, other parameters are set as follows: $C_0 = -30$ dB, $P_0 = 20$ dBm, $\sigma^2 = -60$ dBm, $\epsilon = 10^{-3}$, $K = 20$, $M = 20$, and $N = 64$.

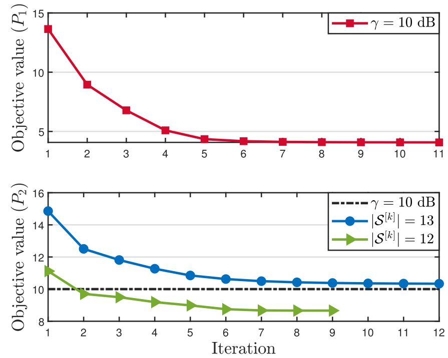
A. Convergence Behaviour and Optimality

To demonstrate the effectiveness of our proposed two-step alternating DC algorithm for device selection, we first show the convergence behaviour of the sparsity inducing step and the feasibility detection step in Fig. 3(a). It is observed that the objective values of problems $\mathscr{P}_1$ and $\mathscr{P}_2$ are both able to converge to the stationary points by accurately finding rank-one solutions with DC programming. Moreover, we present the simulation results for the proposed algorithm and the brute-force search method when $K = 10$, $M = 10$, and $N = 25$, as shown in Fig. 3. It is observed that our proposed algorithm, which has a much lower computational complexity, achieves a very close performance to the brute-force search method in terms of the average number of selected devices.

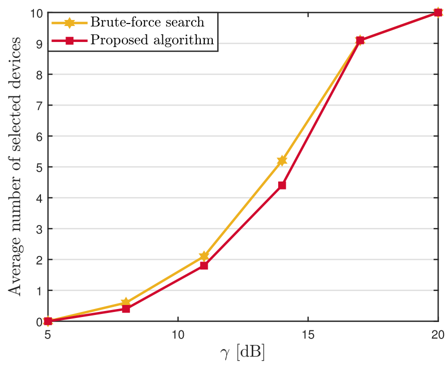
B. Device Selection

Under the two-step framework, we compare the proposed alternating DC-based device selection algorithm (i.e., Algorithm 1) with the following baseline schemes:

- **Alternating SDR with IRS**: In this scheme, the SDR technique is applied to solve problems $\mathscr{P}_{1,1}$, $\mathscr{P}_{1,2}$, $\mathscr{P}_{2,1}$, and $\mathscr{P}_{2,2}$.
- **Random phase shifts**: In this scheme, the phase shift of each reflecting element at the IRS is uniformly and independently generated from $[0, 2\pi)$. We merely solve problem $\mathscr{P}_{1,1}$ in the first step and problem $\mathscr{P}_{2,1}$ in the second step with the proposed DC algorithm.



(a) Convergence behaviour.



(b) Performance comparison between the proposed algorithm and the brute-force search method.

Fig. 3. Convergence behaviour and optimality of the proposed algorithm.

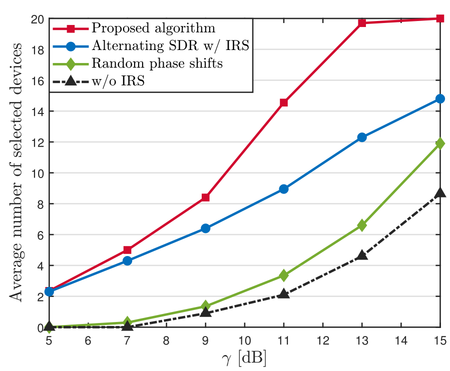


Fig. 4. Average number of selected devices versus the MSE threshold.

- **Without IRS**: In the circumstance without IRS, only problem $\mathscr{P}_{1,1}$ in the first step and problem $\mathscr{P}_{2,1}$ in the second step need to be solved with the proposed DC algorithm by setting $\mathbf{\Theta} = \mathbf{0}$.

Fig. 4 shows the average number of selected devices under different schemes versus the MSE threshold γ for FL systems with and without IRS. As γ increases, the average number of selected devices becomes larger. This is because reducing the requirement for the aggregation error is capable of inducing more edge devices to participate in the training process of FL. In contrast to the scenario without IRS, deploying an IRS in the FL system can support much more devices for concurrent

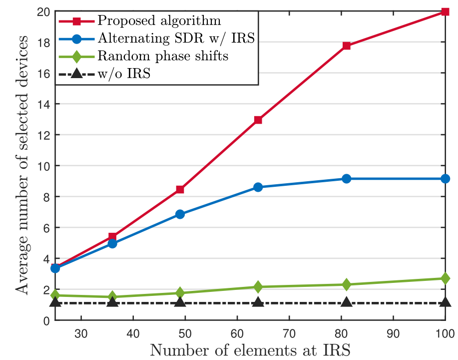


Fig. 5. Average number of selected devices versus the number of reflecting elements at IRS.

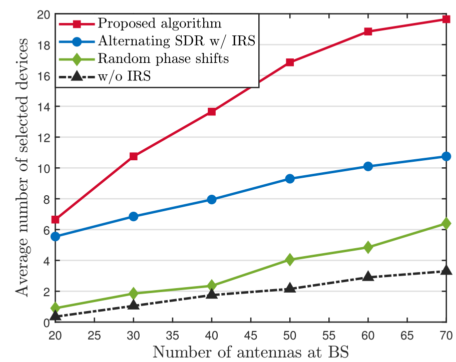
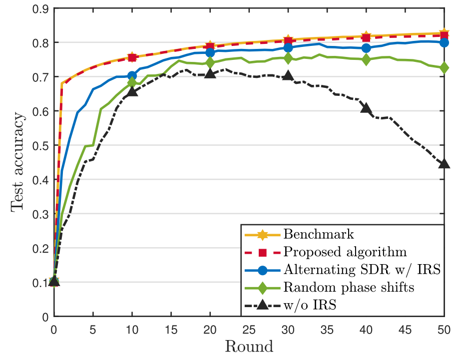


Fig. 6. Average number of selected devices versus the number of antennas at BS.

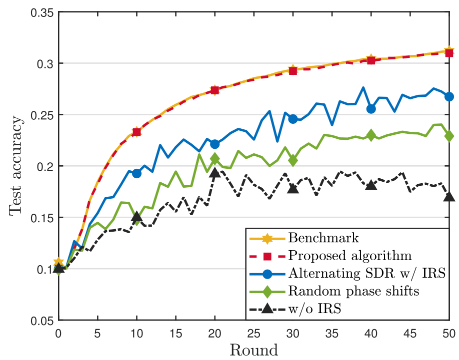
model aggregation under a certain MSE requirement. Besides, the scheme with random phase shifts performs worse than both our proposed algorithm and the alternating SDR algorithm, which demonstrates the importance of jointly optimizing the device selection, the aggregation beamformer at the BS, and the phase shifts at the IRS. Moreover, due to the effectiveness of obtaining the rank-one solutions with the DC algorithm, our proposed DC-based method significantly outperforms the SDR method.

Fig. 5 illustrates the impact of the number of reflecting elements at the IRS on the average number of selected devices when $\gamma = 10$ dB. As the number of reflecting elements increases, the IRS generates more accurate passive reflective beamforming for the incident signals, thereby effectively reducing the aggregation error at the BS. Therefore, the system is capable of selecting more edge devices to participate in FL, while satisfying the MSE requirement. In addition, since the SDR method has a high probability of failing to return rank-one solutions for high-dimensional optimization problems, it is observed that the gap between our proposed algorithm and alternating SDR algorithm increases as the number of reflecting elements at the IRS becomes larger.

Fig. 6 shows the impact of the number of antennas at the BS on the average number of selected devices when $\gamma = 8$ dB. As the number of antennas at the BS increases, the channel gain between the BS and each edge device is enhanced by gathering signals from more antennas. Therefore, the adverse



(a) MNIST dataset



(b) CIFAR-10 dataset

Fig. 7. Test accuracy versus the number of communication rounds.

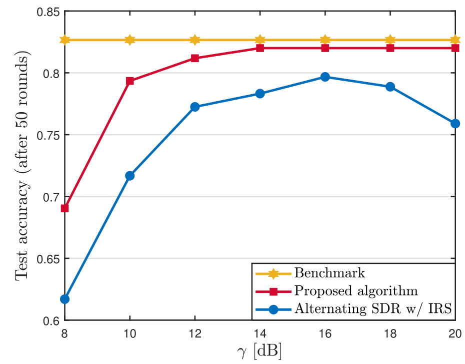


Fig. 8. Test accuracy versus the MSE threshold with MNIST dataset.

impact of additive noise at the BS can be alleviated and in turn the aggregation error is reduced, thereby being able to schedule more edge devices to participate in FL under a certain MSE requirement. In addition, it is observed that even when the number of antennas at the BS is doubled, it is still difficult for the system without IRS to achieve a similar performance to the scenario with an IRS by jointly optimizing the aggregation beamformer at the BS and the phase shifts at the IRS. This observation implies that deploying an IRS not only enhances the system performance but also reduces the hardware complexity at the BS. Therefore, it is an efficient way to achieve fast and reliable model aggregation from the edge devices under a certain MSE requirement by deploying an IRS in AirComp-based FL system.

C. Performance Comparison for FL

To directly show the performance of our proposed two-step alternating DC algorithm for dealing with FL tasks, we train image classifier models on the widely-used MNIST and CIFAR-10 datasets. In the simulations, we first sort the dataset by their labels and then divide it into 40 shards of size 1500 with MNIST or 1250 with CIFAR-10. After that, we assign each device 2 shards without replacement. The case that all devices are selected at each communication round and without any aggregation error serves as the **Benchmark**. All results are averaged over 20 experiments.

Fig. 7 shows the test accuracy versus the number of communication rounds when $\gamma = 16$ dB. It is observed that in the beginning, the performance of all schemes shows an increasing trend, and our proposed algorithm achieves a very close performance as the benchmark. This demonstrates that the learning performance is mainly affected by the number of selected devices at early phases of FL. As the number of communication rounds increases, the accumulated model aggregation error becomes the main factor affecting the learning performance, thereby increasing the gap between the benchmark and other baseline schemes.

Fig. 8 shows the test accuracy versus the MSE threshold with MNIST dataset. It is observed, for the alternating SDR algorithm, the learning performance increases when $\gamma \leq 16$ dB and then decreases when $\gamma > 16$ dB, which demonstrates that scheduling more devices with larger aggregation errors may lead to a worse learning performance than scheduling fewer devices with smaller aggregation errors. In other words, there is a trade-off between the number of selected devices and the aggregation error, which is also verified by the convergence analysis (17). In addition, since our proposed algorithm can schedule more devices with low aggregation errors, its learning performance is able to monotonically increase under the current settings.

VII. Conclusion

In this paper, we proposed a novel IRS-assisted AirComp approach for fast model aggregation in a FL system. To accelerate the convergence and enhance the learning performance of FL, we developed a two-step alternating low-rank optimization framework to maximize the number of selected devices under the MSE requirement for model aggregation according to the derived convergence analysis. We presented a DC formulation for rank-one constrained problems in the alternating procedure, followed by proposing the DC algorithm for solving the resulting DC programs. Simulation results demonstrated that our proposed algorithm can achieve a higher test accuracy by selecting more devices under certain MSE requirements compared with the baseline scheme without IRS.

It is worth noting that there is a trade-off between the overhead introduced by channel estimation and the training performance of FL in each communication round. Specifically, by deploying an IRS with more reflecting elements, longer training pilots are required for accurate channel estimation and more overhead is needed for the feedback on IRS configuration. On the other hand, deploying an IRS with more reflecting elements yields a higher beamforming gain, which can be exploited to schedule more devices and/or achieve a smaller MSE. Characterizing such a trade-off requires the joint design on channel estimation and resource allocation in IRS-assisted AirComp-based FL systems, which is an interesting topic for future work.

Appendix A
Proof of Lemma 1

Based on (4) and (13), the update of the global model can be expressed as

$$\begin{aligned}
&\boldsymbol{z}^{[t+1]}\\
&= \hat{\boldsymbol{r}}^{[t]}\\
&= \frac{1}{K}\sum_{k\in\mathcal{K}} \boldsymbol{z}_k^{[t+1]} + \left(\hat{\boldsymbol{r}}^{[t]} - \boldsymbol{r}^{[t]}\right) + \left(\boldsymbol{r}^{[t]} - \frac{1}{K}\sum_{k\in\mathcal{K}} \boldsymbol{z}_k^{[t+1]}\right)\\
&= \boldsymbol{z}^{[t]} - \varsigma^{[t]}\frac{1}{K}\sum_{k\in\mathcal{K}} \nabla F_k(\boldsymbol{z}^{[t]}) + \boldsymbol{e}_{\mathrm{cmm}}^{[t]} + \boldsymbol{e}_{\mathrm{sel}}^{[t]}\\
&= \boldsymbol{z}^{[t]} - \varsigma^{[t]}\left(\nabla F(\boldsymbol{z}^{[t]}) - \boldsymbol{e}^{[t]}\right),
\end{aligned} \tag{52}$$

where $\boldsymbol{e}^{[t]} = (\boldsymbol{e}_{\mathrm{cmm}}^{[t]} + \boldsymbol{e}_{\mathrm{sel}}^{[t]})/\varsigma^{[t]}$, and $\boldsymbol{e}_{\mathrm{cmm}}^{[t]}$ and $\boldsymbol{e}_{\mathrm{sel}}^{[t]}$ denote the error caused by wireless communication and device selection, respectively, and they are computed by

$$\boldsymbol{e}_{\mathrm{cmm}}^{[t]} = \hat{\boldsymbol{r}}^{[t]} - \boldsymbol{r}^{[t]} = \frac{\iota^{[t]}}{|\mathcal{S}^{[t]}|}\left(\hat{\boldsymbol{g}}^{[t]} - \boldsymbol{g}^{[t]}\right), \tag{53}$$

$$\begin{aligned}
\boldsymbol{e}_{\mathrm{sel}}^{[t]} &= \boldsymbol{r}^{[t]} - \frac{1}{K}\sum_{k\in\mathcal{K}} \boldsymbol{z}_k^{[t+1]}\\
&= \varsigma^{[t]}\left(\nabla F(\boldsymbol{z}^{[t]}) - \frac{1}{|\mathcal{S}^{[t]}|}\sum_{k\in\mathcal{S}^{[t]}} \nabla F_k(\boldsymbol{z}^{[t]})\right).
\end{aligned} \tag{54}$$

Since $\boldsymbol{z}^{[t]}$ is determined by the realizations of normalized transmit symbols

$$\mathcal{A} = \left\{\{\boldsymbol{s}_k^{[0]}\}_{k\in\mathcal{S}^{[0]}}, \{\boldsymbol{s}_k^{[1]}\}_{k\in\mathcal{S}^{[1]}}, \ldots, \{\boldsymbol{s}_k^{[t-1]}\}_{k\in\mathcal{S}^{[t-1]}}\right\} \tag{55}$$

and AWGN variables

$$\mathcal{B} = \left\{\{\boldsymbol{n}_j^{[0]}\}_{j=1}^d, \{\boldsymbol{n}_j^{[1]}\}_{j=1}^d, \ldots, \{\boldsymbol{n}_j^{[t-1]}\}_{j=1}^d\right\}, \tag{56}$$

the total expectation of $F(\boldsymbol{z}^{[t]})$ for any $t \in \mathbb{N}_+$ can be taken as $\mathbb{E}\left[F(\boldsymbol{z}^{[t]})\right] = \mathbb{E}_{\mathcal{A}}\mathbb{E}_{\mathcal{B}}\left[F(\boldsymbol{z}^{[t]})\right]$ [45].

Based on (52) and Assumption 1 with $0 < \varsigma^{[t]} \equiv \varsigma \le 1/L$, we have

$$\begin{aligned}
&F(\boldsymbol{z}^{[t+1]}) - F(\boldsymbol{z}^{[t]})\\
&\le \varsigma\left(\frac{L\varsigma}{2} - 1\right)\|\nabla F(\boldsymbol{z}^{[t]})\|^2 + (1 - L\varsigma)\varsigma\left\langle \nabla F(\boldsymbol{z}^{[t]}), \boldsymbol{e}^{[t]}\right\rangle\\
&\quad + \frac{L\varsigma^2}{2}\|\boldsymbol{e}^{[t]}\|^2\\
&\overset{\diamond_1}{\le} -\frac{\varsigma}{2}\|\nabla F(\boldsymbol{z}^{[t]})\|^2 + \frac{\varsigma}{2}\|\boldsymbol{e}^{[t]}\|^2\\
&\overset{\diamond_2}{\le} -\frac{\varsigma}{2}\|\nabla F(\boldsymbol{z}^{[t]})\|^2 + \frac{1}{\varsigma}\left(\|\boldsymbol{e}_{\mathrm{cmm}}^{[t]}\|^2 + \|\boldsymbol{e}_{\mathrm{sel}}^{[t]}\|^2\right),
\end{aligned} \tag{57}$$

where $\diamond_1$ is according to $\left\langle \nabla F(\boldsymbol{z}^{[t]}), \boldsymbol{e}^{[t]}\right\rangle \le \frac{1}{2}\left(\|\nabla F(\boldsymbol{z}^{[t]})\|^2 + \|\boldsymbol{e}^{[t]}\|^2\right)$, and $\diamond_2$ is based on $\|\boldsymbol{e}^{[t]}\|^2 = \|\boldsymbol{e}_{\mathrm{cmm}}^{[t]} + \boldsymbol{e}_{\mathrm{sel}}^{[t]}\|^2/\varsigma^2 \le 2(\|\boldsymbol{e}_{\mathrm{cmm}}^{[t]}\|^2 + \|\boldsymbol{e}_{\mathrm{sel}}^{[t]}\|^2)/\varsigma^2$.

Then, according to [45, Section 3.1] and Assumption 2, the norm of device selection error is upper bounded by

$$\|\boldsymbol{e}_{\mathrm{sel}}^{[t]}\|^2 \le 4\varsigma^2\kappa\left(1 - \frac{|\mathcal{S}^{[t]}|}{K}\right)^2. \tag{58}$$

By substituting (53) and (58) into (57), we have

$$\begin{aligned}
&\mathbb{E}\left[F(\boldsymbol{z}^{[t+1]})\right] - \mathbb{E}\left[F(\boldsymbol{z}^{[t]})\right]\\
&\le -\frac{\varsigma}{2}\mathbb{E}\left[\|\nabla F(\boldsymbol{z}^{[t]})\|^2\right] + 4\varsigma\kappa\left(1 - \frac{|\mathcal{S}^{[t]}|}{K}\right)^2\\
&\quad + \frac{(\iota^{[t]})^2}{\varsigma|\mathcal{S}^{[t]}|^2}\mathbb{E}\left[\|\hat{\boldsymbol{g}}^{[t]} - \boldsymbol{g}^{[t]}\|^2\right].
\end{aligned} \tag{59}$$

Then, summing both sides of (59) for $t \in \{0, 1, \ldots, T-1\}$, and recalling Assumptions 3 gives

$$\begin{aligned}
&F(\boldsymbol{z}^\star) - F(\boldsymbol{z}^{[0]}) \le \mathbb{E}\left[F(\boldsymbol{z}^{[T]})\right] - F(\boldsymbol{z}^{[0]})\\
&\le -\frac{\varsigma}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(\boldsymbol{z}^{[t]})\|^2\right] + 4\varsigma\kappa\sum_{t=0}^{T-1}\left(1 - \frac{|\mathcal{S}^{[t]}|}{K}\right)^2\\
&\quad + \frac{\Gamma}{\varsigma}\sum_{t=0}^{T-1}\frac{1}{|\mathcal{S}^{[t]}|^2}\sum_{j=1}^{d}\mathbb{E}\left[|\hat{g}_j^{[t]} - g_j^{[t]}|^2\right].
\end{aligned} \tag{60}$$

Finally, dividing both sides of (60) by T communication rounds and rearranging it further yields (17).

Appendix B
Proof of Proposition 1

The transmitter scalar $\{w_k\}$ in (18) has the zero-forcing structure to enforce

$$\sum_{k\in\mathcal{S}}\left|\frac{1}{\sqrt{\eta}}\boldsymbol{m}^{\mathsf{H}}(\boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_k^{\mathrm{r}} + \boldsymbol{h}_k^{\mathrm{d}})w_k - 1\right|^2 = 0. \tag{61}$$

Moreover, we have $\mathsf{MSE} \ge \sigma^2\|\boldsymbol{m}\|^2/\eta$ from (18). We thus obtain the form of zero-forcing transmitter scalar given in Proposition 1 which minimizes the MSE.

Appendix C
Proof of Proposition 2

The constraint (23b) can be reformulated as $E_k(\boldsymbol{m}) = \|\boldsymbol{m}\|^2 - \tilde{\gamma}\,|\boldsymbol{m}^{\mathsf{H}}(\boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_k^{\mathrm{r}} + \boldsymbol{h}_k^{\mathrm{d}})|^2 \le 0, \forall\, k \in \mathcal{S}$, where $\boldsymbol{m} \ne \boldsymbol{0}$. We can further rewrite it as $E_k(\boldsymbol{m}/\sqrt{\tau}) = E_k(\boldsymbol{m})/\tau \le 0, \forall\, k \in \mathcal{S}$, where $\|\boldsymbol{m}\|^2 \ge \tau$ and $\tau > 0$. By introducing optimization variable $\tilde{\boldsymbol{m}} = \boldsymbol{m}/\sqrt{\tau}$, the constraint (23b) can be equivalently reformulated as $\|\tilde{\boldsymbol{m}}\|^2 - \tilde{\gamma}\,|\tilde{\boldsymbol{m}}^{\mathsf{H}}(\boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_k^{\mathrm{r}} + \boldsymbol{h}_k^{\mathrm{d}})|^2 \le 0, \forall\, k \in \mathcal{S}$, where $\|\tilde{\boldsymbol{m}}\|^2 \ge 1$. Thus, we obtain the equivalent form of constraint (23b) given in (24).

REFERENCES

[1] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning," *IEEE Netw.*, vol. 33, no. 5, pp. 156–165, Sep. 2019.

[2] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J.-A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[3] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge AI: Algorithms and systems," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2167–2191, 4th Quart., 2020.

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

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

[6] A. Elgabli, J. Park, C. B. Issaid, and M. Bennis, "Harnessing wireless channels for scalable and privacy-preserving federated learning," *IEEE Trans. Commun.*, early access, May 10, 2021, doi: 10.1109/TCOMM.2021.3078783.

[7] H. H. Yang, Z. Liu, T. Q. S. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 317–333, Jan. 2020.

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

[9] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, Jan. 2021.

[10] W. Shi, S. Zhou, Z. Niu, M. Jiang, and L. Geng, "Joint device scheduling and resource allocation for latency constrained wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 453–467, Jan. 2021.

[11] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3498–3516, Oct. 2007.

[12] L. Chen, X. Qin, and G. Wei, "A uniform-forcing transceiver design for over-the-air function computation," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 942–945, Dec. 2018.

[13] G. Zhu and K. Huang, "MIMO over-the-air computation for high-mobility multimodal sensing," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6089–6103, Aug. 2019.

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

[15] G. Zhu, J. Xu, K. Huang, and S. Cui, "Over-the-air computing for wireless data aggregation in massive IoT," 2020, *arXiv:2009.02181*. [Online]. Available: http://arxiv.org/abs/2009.02181

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

[17] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, Jan. 2020.

[18] M. M. Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, vol. 68, pp. 2155–2169, 2020.

[19] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, May 2020.

[20] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Signal Process.*, vol. 68, pp. 2897–2911, 2020.

[21] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning," *IEEE Trans. Wireless Commun.*, early access, Mar. 19, 2021, doi: 10.1109/TWC.2021.3065748.

[22] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2020.

[23] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[24] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.

[25] J. Chen, Y.-C. Liang, Y. Pei, and H. Guo, "Intelligent reflecting surface: A programmable wireless environment for physical layer security," *IEEE Access*, vol. 7, pp. 82599–82612, 2019.

[26] T. Jiang and Y. Shi, "Over-the-air computation via intelligent reflecting surfaces," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[27] Z. Wang, Y. Shi, Y. Zhou, H. Zhou, and N. Zhang, "Wireless-powered over-the-air computation in intelligent reflecting surface-aided IoT networks," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1585–1598, Feb. 2021.

[28] K. Yang, Y. Shi, Y. Zhou, Z. Yang, L. Fu, and W. Chen, "Federated machine learning for intelligent IoT via reconfigurable intelligent surface," *IEEE Netw.*, vol. 34, no. 5, pp. 16–22, Sep. 2020.

[29] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[31] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. 4, Apr. 2009.

[32] S. Ravi, "Efficient on-device models using neural projections," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 5370–5379.

[33] W. Debaenst, A. Feys, I. Cuiñas, M. G. Sánchez, and J. Verhaevert, "RMS delay spread vs. coherence bandwidth from 5G indoor radio channel measurements at 3.5 GHz band," *Sensors*, vol. 20, no. 3, p. 750, Jan. 2020.

[34] S. Wang *et al.*, "Doppler shift and coherence time of 5G vehicular channels at 3.5 GHz," in *Proc. IEEE Int. Symp. Antennas Propag. USNC/URSI Nat. Radio Sci. Meeting*, Jul. 2018, pp. 2005–2006.

[35] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 170–185, Jan. 2021.

[36] Y. Yang, B. Zheng, S. Zhang, and R. Zhang, "Intelligent reflecting surface meets OFDM: Protocol design and rate maximization," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4522–4535, Jul. 2020.

[37] C. You, B. Zheng, and R. Zhang, "Channel estimation and passive beamforming for intelligent reflecting surface: Discrete phase shift and progressive refinement," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2604–2620, Nov. 2020.

[38] Z. Wang, L. Liu, and S. Cui, "Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6607–6620, Oct. 2020.

[39] H. Liu, X. Yuan, and Y.-J.-A. Zhang, "Matrix-calibration-based cascaded channel estimation for reconfigurable intelligent surface assisted multiuser MIMO," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2621–2636, Nov. 2020.

[40] X. Guan, Q. Wu, and R. Zhang, "Anchor-assisted intelligent reflecting surface channel estimation for multiuser communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[41] M.-M. Zhao, Q. Wu, M.-J. Zhao, and R. Zhang, "Intelligent reflecting surface enhanced wireless networks: Two-timescale beamforming optimization," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 2–17, Jan. 2021.

[42] Y. Shao, D. Gündüz, and S. C. Liew, "Federated edge learning with misaligned over-the-air computation," 2021, *arXiv:2102.13604*. [Online]. Available: http://arxiv.org/abs/2102.13604

[43] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: A tutorial," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3313–3351, May 2021.

[44] O. Abari, H. Rahul, D. Katabi, and M. Pant, "AirShare: Distributed coherent transmission made seamless," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2015, pp. 1742–1750.

[45] M. P. Friedlander and M. Schmidt, "Hybrid deterministic-stochastic methods for data fitting," *SIAM J. Sci. Comput.*, vol. 34, no. 3, pp. A1380–A1405, Jan. 2012.

[46] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, 2018.

[47] Y. Shi, J. Cheng, J. Zhang, B. Bai, W. Chen, and K. B. Letaief, "Smoothed L_p-minimization for green cloud-RAN with user admission control," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1022–1036, Apr. 2016.

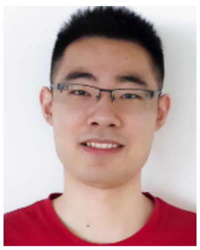
[48] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[49] Z.-Q. Luo, N. D. Sidiropoulos, P. Tseng, and S. Zhang, "Approximation bounds for quadratic optimization with homogeneous quadratic constraints," *SIAM J. Optim.*, vol. 18, no. 1, pp. 1–28, 2007.

[50] P. D. Tao and L. T. H. An, "Convex analysis approach to DC programming: Theory, algorithms and applications," *Acta Math. Vietnamica*, vol. 22, no. 1, pp. 289–355, 1997.

[51] M. Grant and S. Boyd. (2014). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1*. [Online]. Available: http://cvxr.com/cvx

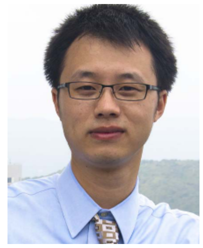
[52] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1838–1851, May 2020.



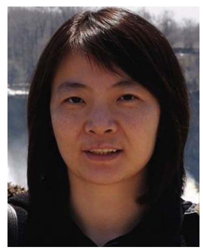
Zhibin Wang (Graduate Student Member, IEEE) received the B.S. degree from the School of Telecommunications Engineering, Xidian University, Xi'an, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. His research interests include the Internet of Things, intelligent reflecting surface, and federated learning.

Jiahang Qiu received the B.S. degree in communication engineering from Zhejiang Gongshang University, Hangzhou, China, in 2017, and the M.S. degree in electronic and communication engineering from Xiamen University, Xiamen, China, in 2021. His research interests include wireless network optimization and machine learning.

Yong Zhou (Member, IEEE) received the B.Sc. and M.Eng. degrees from Shandong University, Jinan, China, in 2008 and 2011, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2015. From November 2015 to January 2018, he worked as a Post-Doctoral Research Fellow with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada. He is currently an Assistant Professor with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. His research interests include the Internet of Things, edge computing, and reconfigurable intelligent surface.



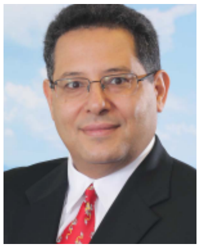
Yuanming Shi (Senior Member, IEEE) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2011, and the Ph.D. degree in electronic and computer engineering from The Hong Kong University of Science and Technology (HKUST) in 2015. Since September 2015, he has been with the School of Information Science and Technology, ShanghaiTech University, where he is currently a tenured Associate Professor. He visited the University of California, Berkeley, CA, USA, from October 2016 to February 2017. His research areas include optimization, statistics, machine learning, signal processing, and their applications to 6G, the IoT, and AI. He was a recipient of the 2016 IEEE Marconi Prize Paper Award in Wireless Communications and the 2016 Young Author Best Paper Award by the IEEE Signal Processing Society. He is an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS.



Liqun Fu (Senior Member, IEEE) received the Ph.D. degree in information engineering from The Chinese University of Hong Kong in 2010.

From 2011 to 2013, she was a Post-Doctoral Research Fellow with the Institute of Network Coding, The Chinese University of Hong Kong. From 2013 to 2015, she was a Post-Doctoral Research Fellow with the ACCESS Linnaeus Centre, KTH Royal Institute of Technology. From 2015 to 2016, she was an Assistant Professor with ShanghaiTech University. She is currently a Full Professor with the School of Informatics, Xiamen University, China. Her research interests are mainly in communication theory, optimization theory, game theory, and learning theory, with applications in wireless networks. She serves as a TPC member for many leading conferences in communications and networking, such as IEEE INFOCOM, ICC, and GLOBECOM. She served as the Technical Program Co-Chair for the GCCCN Workshop of the IEEE INFOCOM 2014, the Publicity Co-Chair for the GSNC Workshop of the IEEE INFOCOM 2016, and the Web Chair for the IEEE WiOpt 2018. She is on the Editorial Board of IEEE ACCESS and the *Journal of Communications and Information Networks* (JCIN).

Wei Chen (Senior Member, IEEE) received the B.S. and Ph.D. degrees (Hons.) from Tsinghua University in 2002 and 2007, respectively. From 2005 to 2007, he was a Visiting Ph.D. Student with The Hong Kong University of Science and Technology. Since 2007, he has been a Faculty Member with Tsinghua University, where he is currently a tenured Full Professor and a University Council Member. He visited the University of Southampton in 2010, Telecom Paris Tech in 2014, and Princeton University, Princeton, NJ, USA, in 2016. From 2014 to 2016, he was the Deputy Head of the Department of Electronic Engineering, Tsinghua University. From 2016 to 2021, he was the Director of the Degree Office, Tsinghua University. His research interest is in the real-time communication theory. He is a Cheung Kong Young Scholar and a member of the National Program for Special Support of Eminent Professionals, also known as 10 000 talent program. He has also been supported by the National 973 Youth Project, the NSFC Excellent Young Investigator Project, the New Century Talent Program of the Ministry of Education, and the Beijing Nova Program. He is a Standing Committee Member of All-China Youth Federation and the Secretary-General of its Education Board. He received the IEEE Marconi Prize Paper Award in 2009 and the IEEE Comsoc Asia-Pacific Board Best Young Researcher Award in 2011. He was a recipient of the National May 1st Labor Medal and the China Youth May 4th Medal. He has served as a TPC Co-Chair for IEEE VTC-Spring in 2011 and a Symposium Co-Chair for IEEE ICC and GLOBECOM. He served as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS.



Khaled B. Letaief (Fellow, IEEE) received the B.S. (Hons.), M.S., and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in December 1984, August 1986, and May 1990, respectively.

From 1990 to 1993, he was a Faculty Member at the University of Melbourne, Australia. Since 1993, he has been with The Hong Kong University of Science and Technology (HKUST). While at HKUST, he has held many administrative positions, including the Dean of engineering, the Head of the Electronic and Computer Engineering Department, the Director of the Wireless IC Design Center and Hong Kong Telecom Institute of Information Technology, and the Founding Director of Huawei Innovation Laboratory. He served as a Consultant for different organizations, including Huawei, ASTRI, ZTE, Nortel, PricewaterhouseCoopers, and Motorola. He is an internationally recognized leader in wireless communications and networks with research interest in artificial intelligence, big data analytics systems, mobile cloud and edge computing, tactile Internet, and 5G systems and beyond. In these areas, he has over 630 papers with over 38 600 citations and an H-index of 87 along with 15 patents, including 11 U.S. inventions. He is also recognized by Thomson Reuters as an ISI Highly Cited Researcher and was listed among the 2020 top 30 of AI 2000 Internet of Things Most Influential Scholars.

Dr. Letaief is a member of United States National Academy of Engineering and Hong Kong Academy of Engineering Sciences. He is a fellow of Hong Kong Institution of Engineers. He has been involved in organizing many flagship international conferences. From 2018 to 2019, he served as the President of the IEEE Communications Society, the world's leading organization for communications professionals with headquarter in New York City and members in 162 countries. He is well recognized for his dedicated service to professional societies and IEEE where he has served in many leadership positions. These include a Treasurer of IEEE Communications Society, the IEEE Communications Society Vice-President for conferences, the Chair of IEEE Committee on Wireless Communications, an Elected Member of IEEE Product Services and Publications Board, and the IEEE Communications Society Vice-President for technical activities. He was a recipient of many distinguished awards and honors including the 2019 Distinguished Research Excellence Award by HKUST School of Engineering (Highest research award and only one recipient/three years is honored for his/her contributions), the 2019 IEEE Communications Society and Information Theory Society Joint Paper Award, the 2018 IEEE Signal Processing Society Young Author Best Paper Award, the 2017 IEEE Cognitive Networks Technical Committee Publication Award, the 2016 IEEE Signal Processing Society Young Author Best Paper Award, the 2016 IEEE Marconi Prize Paper Award in Wireless Communications, the 2011 IEEE Wireless Communications Technical Committee Recognition Award, the 2011 IEEE Communications Society Harold Sobol Award, the 2010 Purdue University Outstanding Electrical and Computer Engineer Award, the 2009 IEEE Marconi Prize Award in Wireless Communications, the 2007 IEEE Communications Society Joseph LoCicero Publications Exemplary Award, and over 16 IEEE best paper awards. He is the Founding Editor-in-Chief of the prestigious IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and has served as the Editor-in-Chief on the Editorial Board for other premier journals including the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS in wireless series.