Robust Design for RIS-Assisted Over-the-Air Federated Learning With Imperfect Cascaded CSI

Qiaochu An, Hongbin Zhu, *Member, IEEE*, Qiang Ye, *Senior Member, IEEE*, Ning Zhang, *Senior Member, IEEE*, Yuanming Shi, *Senior Member, IEEE*, and Yong Zhou, *Senior Member, IEEE*

***Abstract*—Over-the-air computation (AirComp) enables spectral-efficient global model aggregation for federated learning (FL) by supporting concurrent transmission and harnessing co-channel interference. However, unfavorable channel conditions and inaccurate channel estimation are two performance-limiting factors of AirComp-assisted FL. In this paper, we leverage reconfigurable intelligent surface (RIS) to assist AirComp for gradient aggregation with imperfect cascaded channel state information (CSI), taking into account both the expectation-based and worst-case error models (i.e., stochastic and deterministic models). Guided by the convergence analysis, we minimize the gradient aggregation distortion measured by the mean-squared-error (MSE), taking into account the unit modulus constraints of RIS phase-shifts. To alleviate the performance degradation due to imperfect channel estimation, we propose two robust algorithms to optimize the receive beamforming at the edge server, RIS phase-shifts, and transmit power at the edge devices. Both algorithms are designed under an alternating optimization framework, where Schur's complement and penalty convex–concave procedure are adopted for the worst-case error model, and Lagrange duality and difference-of-convex programming are used for the expectation-based error model. Simulations are conducted to validate the learning performance superiority of the proposed algorithms over baseline schemes, inducing the robustness against CSI errors.**

***Index Terms*—Federated learning, imperfect channel state information, reconfigurable intelligent surface, over-the-air computation.**

Received 30 April 2025; revised 26 November 2025; accepted 28 January 2026. Date of publication 6 February 2026; date of current version 13 February 2026. The work of Yong Zhou was supported in part by the Natural Science Foundation of Shanghai under Grant 23ZR1442800 and in part by the Science and Technology Commission Foundation of Shanghai under Grant 25DP1501900. The work of Yuanming Shi was supported in part by the National Natural Science Foundation of China under Grant 62522117 and Grant 62271318, Yangtze River Delta Science and Technology Innovation Community Joint Research (Basic Research) Project under Grant BK 2024CSJZN00303, and the Science and Technology Commission Foundation of Shanghai under Grant 25DP1500100. The associate editor coordinating the review of this article and approving it for publication was M. H. Rehmani. *(Corresponding author: Yong Zhou.)*

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

Hongbin Zhu is with the Fintech Institute, Fudan University, Shanghai 200438, China (e-mail: rgzhuhb@gmail.com).

Qiang Ye is with the Department of Electrical and Software Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada (e-mail: qiang.ye@ucalgary.ca).

Ning Zhang is with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada (e-mail: ning.zhang@uwindsor.ca).

Digital Object Identifier 10.1109/TCCN.2026.3661518

I. Introduction

IN THE forthcoming 6G era, intelligent devices with sensing, communication, and computation capabilities are envisioned to support various intelligent applications. However, the conventional centralized machine learning framework, which often involves frequent and extensive raw data exchange, could result in substantial communication overhead and severe privacy leakage [1]. Fortunately, federated learning (FL) is an effective distributed machine learning solution to these issues by only transferring models/gradients of edge devices rather than their private raw data [2], [3], [4].

A typical FL system always comprises multiple edge devices and a central edge server. Each device trains a local model using its own dataset and transmits only the updated model parameters to the server. The server then aggregates these local parameters to generate a global model, which is subsequently broadcast back to all edge devices for further local training. This cycle repeats iteratively until a predefined number of rounds is completed or the model converges. Throughout the process, all raw data remain stored locally on the devices, thereby effectively mitigating the risk of raw data leakage. Even with the bright future, it is challenging for an edge server to aggregate high-dimensional models/gradients from many geographically dispersed edge devices within a short period of time due to scarce radio spectrum. Recently, researchers have initiated the utilization of over-the-air computation (AirComp) in FL. The core is to combine the signal transmission of edge devices with the distributed computing process, utilizing the wave-form superposition characteristics in wireless channels, and directly implementing parallel aggregation of multi edge devices' model parameters. With proper transceiver design, the interference due to concurrent transmission can be fully exploited to facilitate efficient model aggregation. To alleviate the signal distortion of aggregated models, the authors in [5] design a threshold-based transmit power control scheme for AirComp-assisted FL. Specifically, full power transmission is the optimal strategy when the quality indicator is less than the threshold, while the channel-inversion scheme is preferred otherwise. Considering the non-identical data distribution among edge

devices, the authors in [6] design a gradient-aware power control strategy to minimize the aggregation error based on the estimated gradient statistics. In addition to the transmission design at edge devices, the authors in [7] develop a low-complexity phase-shift scheme of receive beamforming at edge server under an unit-modulus constraint. The authors in [8] manage the inter-cell interference in multi-cell AirComp-assisted FL systems by designing the transceiver beamforming in both model uploading and downloading. Besides, to take full advantage of limited spectrum, it is critically important to schedule edge devices that can promote the FL performance. The authors in [9] propose a joint communication-and-learning design that selects the appropriate participants considering the mean-squared-error (MSE) threshold of AirComp. The authors in [10] select high-quality edge devices according to their data importance which is jointly evaluated by the newly defined data state information and channel condition. As FL is a long-term training paradigm involving multiple communication rounds, it is challenging to predict the accurate future system states (i.e., channel conditions, energy consumption, and gradient information) in advance. The authors in [11] propose an dynamic device selection method that allows for online scheduling only relying on the current system state. The authors in [12] and [13] propose Lyapunov optimization based online algorithms to minimize the long-term transmission distortion in an AirComp-assisted federated graph learning network, and a reinforcement learning based algorithm is developed in [14]. The synchronization problem, as an important issue of AirComp-assisted FL, is studied in [15] and [16]. The authors in [15] tackle the synchronization issues by adopting a shared reference clock, while maximum likelihood and aligned-sample estimators are developed in [16] to estimate the arithmetic mean of the mis-aligned transmitted symbols. Despite the above benefits, the model aggregation via AirComp is generally limited by devices with poor channel conditions, which may be in the regime of deep fade due to random channel fading, thereby deteriorating the FL performance.

Reconfigurable intelligent surface (RIS) has the potential to enhance the achievable performance of AirComp by dynamically reconfiguring the propagation environment based on instantaneous channel conditions. RIS consists of a considerable number of passive reflecting elements, which can be independently software-controlled to adjust the reflecting coefficients (i.e., amplitude and phase) and provide a cost-effective and hardware-efficient solution for reshaping wireless environments [17], [18], [19]. Several representative works considered the implementation of RIS in practical systems. The authors in [20] introduced stacked intelligent RIS in the uplink of cell-free MIMO systems to enhance both spectral and energy efficiency. Specifically, the weight vector is designed based on the minimum mean square error criterion to detect the aggregate multi-user signal by taking the effects of hardware impairments into consideration. The authors in [21] proposed an energy efficient beamforming design in the realistic hardware impairment scenarios. The authors in [22] designed an RIS-empowered low earth orbit satellite system, where holographic metasurfaces are deployed to assist terrestrial users to maximize the downlink channel gain. The authors in [23] analyzed the theoretically asymptotic capacity bound of the near-field scenario in cell-free networks with non-ideal hardware. Since RIS is capable of combating the unfavorable channel conditions, it can be utilized to facilitate the multi-round FL training in both model uploading and downloading. In particular, the authors in [24] first reveal that RIS can effectively enhance the quality of over-the-air model aggregation. The authors in [25] introduced a computationally efficient joint optimization algorithm for the active and passive beamforming of transceivers and RIS by using a graph neural network. A similar low-complexity Riemannian optimization method is proposed in [26]. The estimated channel state information (CSI) may be distorted, which further degrades the AirComp accuracy for subsequent model aggregation. Thus, the authors in [27] develop a transmitter CSI-free method that only designs the phase-shift matrix on RIS and receive beamforming at edge server to avoid the erroneous transmission due to CSI distortion. However, these works only concentrate on the optimization from communication perspective. The authors in [28] and [29] further propose the uniform optimization frameworks which incorporate device scheduling to maximize the potential benefits of RIS and AirComp in FL. In addition, the dynamic channel reconfigurability of RIS is leveraged in [30] to jointly optimize the test accuracy and privacy of AirComp-assisted FL. Similarly, the model integrity is also guaranteed in [31]. Although the advantages of RIS have been demonstrated, all the aforementioned works [24], [26], [27], [28], [29], [30], [31] are studied by assuming that the CSI is perfectly known.

However, the channel estimation for RIS-aided link in wireless networks is challenging, due to its passive signal processing characteristics. Particularly, a straightforward approach is to divide the cascaded channel into multiple sub-channels and sequentially estimate each reflecting element based on the receiver's channel measurements. This can be achieved by switching reflecting elements ON/OFF at the RIS in conjunction with transmitter pilot symbols by following the method studied in [32], [33], and [34], by developing special matrices pilots (e.g., the discrete Fourier transform (DFT) matrix, Hadamard matrix, and circulant matrix generated by Zadoff-Chu sequence) in full-ON RIS training reflection pattern [35], [36], [37], and by jointly designing the RIS training reflection pattern and the transmitter's pilot sequence with the aim of minimizing channel estimation error [36], [38], [39]. To estimate the channels with an affordable training overhead, the authors in [40] develop an efficient posterior mean method. In addition, the authors in [41] estimated the RIS-involved channel by deploying several active reflecting elements with two operation modes on the RIS. Taking into account the quantization error and feedback delay, the estimated CSI in RIS-assisted wireless networks is usually inaccurate. Assuming estimated channel state information to be error-free will unavoidably leads to system performance deterioration. Under this circumstance, developing robust communication strategies is essential to mitigate channel estimation errors. There are a few works investigating the robust beamforming design in RIS-assisted wireless networks [42], [43]. Particularly, the

authors in [42] propose a robust design under the worst-case error model but merely considered the presence of imperfect CSI of the links from the RIS to devices. Subsequently, a more comprehensive framework of robust design is presented in [43], where both partial and full channel uncertainties are studied. However, the aforementioned studies mainly focus on the beamforming design in the downlink transmission but cannot be directly applied in the uplink model transmission scenarios. Besides, the authors in [31] and [44] propose the joint robust beamforming design in RIS-assisted FL network under the assumption that the statistics of channel error is available. However, comprehensive beamforming design for AirComp-assisted FL in different scenarios with imperfect CSI remains unexplored.

In this paper, we exploit the superiority of RIS and AirComp to enable reliable and fast model aggregation over a single-cell FL network, where a multi-antennas edge server aggregates distributed local gradients that are uploaded in parallel by all edge devices. We carry out the convergence analysis to evaluate the impact of over-the-air model aggregation, and consider the scenario of imperfect estimation on cascaded server-RIS-device channels. Under this circumstance, we aim at promoting FL performance by jointly optimizing the robust beamforming design of transceiver and RIS phase-shift, taking into account both the expectation-based and worst-case error models. We summarize the main contributions of this paper as follows.

- We first perform the convergence analysis of the proposed FL framework and demonstrate that its performance can be promoted by alleviating the aggregation error. To this end, we jointly optimize the transmit scalars at the edge devices, the RIS phase-shift vector, and the receive beamforming vector at the edge server, while considering the transmit power budgets of devices as well as the unit modulus constraints of RIS reflecting elements. In addition, we study the robust transmission design under two channel uncertainty models, i.e., expectation-based and worst-case error models, which are utilized to model errors due to inaccurate estimation and quantization, respectively.
- For the expectation-based error model, the original problem is reformulated as a non-convex quadratically constrained quadratic programming (QCQP) problem. We further optimize the AirComp transceivers based on closed-form solution and the Karush-Kuhn-Tucker (KKT) conditions, while the RIS phase-shift vector is optimized by proposing a difference-of-convex (DC) algorithm.
- For the worst-case error model, the original problem is reformulated as a min-max problem with constraints involving random errors, which are approximated by Schur's complement. Then, we decouple the optimization variables based on an alternating optimization method. The AirComp transceivers are optimized via solving convex subproblems. In addition, the penalty convex–concave procedure (PCCP) is adopted to optimize the RIS phase-shifts.
- We conduct extensive simulations to demonstrate the potential of exploiting RIS and AirComp for promoting FL. Compared with the baseline schemes, the proposed algorithms achieve significant performance gains, validating the superiority of RIS for enhancing gradients aggregation accuracy. Simulations also show that the learning performance under both the expectation-based and worst-case error models is better than that under the non-robust counterpart, validating the superiority of the proposed algorithm in overcoming CSI uncertainties.

TABLE I
IMPORTANT NOTATIONS AND THEIR DEFINITIONS

Notation	Definition
$\mathcal{K}$	Set of all edge devices.
$\mathcal{D}_k$	Local dataset at edge device k.
$\boldsymbol{z}$	Global model parameter of FL task.
$l(\boldsymbol{z};\boldsymbol{x}_i,y_i)$	Sample-wise loss functions of FL task.
$L_k(\boldsymbol{z})$	Local loss function of FL task at device k.
$L(\boldsymbol{z})$	Global local loss function of FL task.
$\boldsymbol{g}_k^{[t]}$	Local gradient of FL task at device k in the t-th training round.
$\boldsymbol{E}_k$	The actual channel response between the edge server and the k-th edge device.
$\boldsymbol{s}_k^{[t]}$	Normalized local gradient of at device k in the t-th training round.
$\boldsymbol{\omega}^{[t]}$	Receive beamforming vector in the t-th training round.
$\boldsymbol{G}^{[t]} \in \mathbb{C}^{N\times M}$	Channel responses between the edge server and the RIS in the t-th training round.
$\boldsymbol{h}_k^{[t]} \in \mathbb{C}^{N}$	Channel responses between the RIS and device k in the t-th training round.
$\boldsymbol{\Theta}^{[t]}$	RIS phase-shift matrix in the t-th training round.

We organize the rest of this work as follows. In Section II, we describe the FL model and the process of gradient aggregation via AirComp, and present the problem formulation under both the expectation-based and worst-case error models. Section III describes the joint design for the expectation-based error model, while that for the worst-case error model is elaborated in Section IV. We discuss the computation complexity of proposed algorithms in Section V. Simulation results are shown in Section VI and we conclude this paper in Section VII.

Notations: Scalar, vector, and matrix are represented by italic, boldface lowercase and uppercase, respectively. Superscripts $(\cdot)^*$, $(\cdot)^{\mathsf{T}}$, $(\cdot)^{\mathsf{H}}$, and $(\cdot)^{-1}$ refer to conjugate, transpose, Hermitian transpose and inverse. Operators $\mathrm{diag}(\cdot)$, $\mathrm{Tr}(\cdot)$ and $\mathbb{E}(\cdot)$ stand for diagonal matrix, trace and expectation, respectively. Operators $|\cdot|$, $\|\cdot\|_2$, $\|\cdot\|_*$ and $\|\cdot\|_F$ denote the absolute value of scalar numbers, L_2-norm of vectors or spectral norm of matrices, nuclear norm and Frobenius norm, respectively. $\mathbb{C}^{m\times n}$ represents the space of $m\times n$ complex matrices. $\mathrm{vec}(\cdot)$ refers to the vectorization of a matrix and the real part of complex numbers is denoted by $\Re\{\cdot\}$.

II. RIS-ASSISTED OVER-THE-AIR FL

A. FL Model

Consider a single-cell FL network, in which an edge server co-located with a base station cooperatively trains the global model parameter $\boldsymbol{z} = [z_1,\ldots,z_d]^{\mathsf{T}} \in \mathbb{R}$ with a set $\mathcal{K} = \{1,2,\ldots,K\}$ of K edge devices, as depicted in Fig. 1. The edge server has M antennas and each single-antenna edge

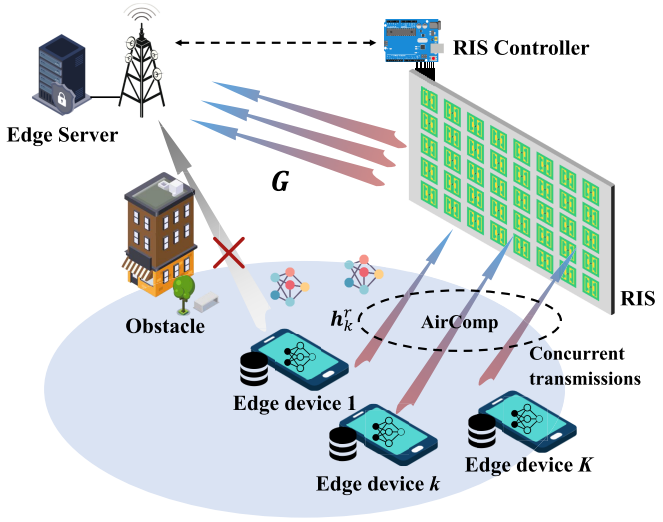


Fig. 1. Illustration of gradient aggregation in an RIS-assisted over-the-air FL network.

device $k \in \mathcal{K}$ holds a non-overlapping and equal-sized local dataset $\mathcal{D}_k = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{D_k}$ with $D_k = |\mathcal{D}_k|$ data pairs satisfying $\mathcal{D}_k \cap \mathcal{D}_j = \emptyset$ and $D_k = D_j$, $\forall\, k \neq j$, where $(\boldsymbol{x}_i, y_i)$ stands for the data pair consisting of sample $\boldsymbol{x}_i$ and its corresponding ground-truth label y_i.

For any FL task, we strive for an optimal model parameter $\boldsymbol{z}^\star$ minimizing the average prediction error measured by loss function over the global dataset $\mathcal{D} = \cup_{k\in\mathcal{K}}\mathcal{D}_k$ with size $D = |\mathcal{D}|$, i.e.

$$\boldsymbol{z}^\star = \arg\min_{\boldsymbol{z}\in\mathbb{R}^d} L(\boldsymbol{z}) = \arg\min_{\boldsymbol{z}\in\mathbb{R}^d} \frac{1}{D}\sum_{(\boldsymbol{x}_i,y_i)\in\mathcal{D}} l(\boldsymbol{z};\boldsymbol{x}_i,y_i), \quad (1)$$

where $L(\boldsymbol{z})$ and $l(\boldsymbol{z};\boldsymbol{x}_i,y_i) \in \mathbb{R}$ denote the global and sample-wise loss functions, respectively. Problem (1) can be addressed by performing local training on edge devices. Accordingly, we have

$$L(\boldsymbol{z}) = \frac{1}{D}\sum_{k\in\mathcal{K}}\sum_{(\boldsymbol{x}_i,y_i)\in\mathcal{D}_k} l(\boldsymbol{z};\boldsymbol{x}_i,y_i) = \sum_{k\in\mathcal{K}}\frac{D_k}{D}L_k(\boldsymbol{z}), \quad (2)$$

where $L_k(\boldsymbol{z}) = \frac{1}{D_k}\sum_{(\boldsymbol{x}_i,y_i)\in\mathcal{D}_k} l(\boldsymbol{z};\boldsymbol{x}_i,y_i)$ denotes the local loss function at device k. In a specific training round indexed as t, the following steps are carried out:

- *Model broadcast:* The global model $\boldsymbol{z}$ is broadcast to edge devices by the edge server. We assume that the downlink transmission from the edge server is error-free.[1]
- *Local training:* Once the global model is decoded, the local model parameter at each device is set as $\boldsymbol{z}_k^{[t-1]} = \boldsymbol{z}^{[t-1]}$ and it then computes the local gradient as

$$\begin{aligned}\boldsymbol{g}_k^{[t]} &= \nabla L_k(\boldsymbol{z}^{[t-1]}) \\ &= \frac{1}{D_k}\sum_{(\boldsymbol{x}_i,y_i)\in\mathcal{D}_k} \nabla l(\boldsymbol{z}^{[t-1]};\boldsymbol{x}_i,y_i), \ \forall\, k\in\mathcal{K}. \quad (3)\end{aligned}$$

- *Gradient aggregation:* After computing their gradients $\{\boldsymbol{g}_k^{[t]}\}$, all local devices subsequently upload them to the edge server. It is inevitable that the received gradients at edge server would be deteriorated due to the limited spectrum and power budgets. The edge server then proceeds to update the global model by taking into account the distorted gradients $\{\hat{\boldsymbol{g}}_k^{[t]}\}$ from all participating devices as follows

$$\boldsymbol{z}^{[t]} = \boldsymbol{z}^{[t-1]} - r\hat{\boldsymbol{g}}^{[t]}, \quad (4)$$

where $\hat{\boldsymbol{g}}^{[t]} = \sum_{k\in\mathcal{K}} D_k\hat{\boldsymbol{g}}_k^{[t]} / \sum_{k\in\mathcal{K}} D_k$ is the aggregated global gradient and r denotes the learning rate.

The operations elaborated above are performed iteratively until the optimal model parameter $\boldsymbol{z}^\star$ is obtained.

B. Gradient Aggregation via AirComp

The performance of wireless FL is predominantly governed by the signal transmission and aggregation accuracy of model gradients. Traditional orthogonal multiple access method requires successful demodulation of all devices' gradients before aggregation, which can result in significant communication overhead and long training time. To address this challenge, we resort to leveraging AirComp to enable efficient gradient aggregation over the wireless channel.

Specifically, each device concurrently sends a normalized version $\boldsymbol{s}_k^{[t]}$ of its gradient $\boldsymbol{g}_k^{[t]}$ to the edge server, where the normalization defined as

$$\boldsymbol{s}_k^{[t]} = \frac{\boldsymbol{g}_k^{[t]} - \bar{g}_k^{[t]}\mathbf{1}}{v_k^{[t]}}, \forall\, k, \quad (5)$$

where $\bar{g}_k^{[t]} = \frac{1}{d}\sum_{j=1}^d g_{k,j}^{[t]}$ denotes the mean value of d elements in gradient $\boldsymbol{g}_k^{[t]}$ and $v_k^{[t]} = \left(\frac{1}{d}\sum_{j=1}^d \left(g_{k,j}^{[t]} - \bar{g}_k^{[t]}\right)^2\right)^{0.5}$ denotes the corresponding standard deviation. We consider the scenario that device-server links are blocked by obstacles [46] and an N-elements adjustable RIS is installed on the surrounding wall to assist the gradient aggregation. Under this circumstance, the edge server can only receive the signals reflected by the RIS. We denote the diagonal RIS phase-shift matrix as $\boldsymbol{\Theta}^{[t]} = \text{diag}\left(\boldsymbol{\theta}^{[t]}\right)$, where $\boldsymbol{\theta}^{[t]} = [\theta_1^{[t]}, \theta_2^{[t]}, \ldots, \theta_N^{[t]}]^\mathsf{T}$ and $\theta_n^{[t]} = \beta_n^{[t]}(e^{j\kappa_n})^{[t]}$ with $\beta_n^{[t]}$ being the reflection amplitude and $\kappa_n^{[t]} \in [0, 2\pi)$ being the phase-shift of element n. We assume, without loss of generality, that the amplitude reflection coefficient is set as $\beta_n^{[t]} = 1, \forall\, n$ and $\theta_n^{[t]}$ can be independently adjusted [47]. Let $\boldsymbol{G}^{[t]} \in \mathbb{C}^{N\times M}$ and $\boldsymbol{h}_k^{[t]} \in \mathbb{C}^N$ denote the channel responses between the edge server and the RIS, and between the RIS and device k in the t-th training round, respectively. With AirComp, all edge devices concurrently transmit $\{s_k^{[t]}\}$ to the edge server. All edge devices are assumed to be synchronized.[2] After being superposed by the

[1] As in [45], the edge server with sufficient transmit power is capable of combating the channel fading and noise.

[2] The synchronization of received signals at the edge server can be guaranteed by either referring to a common clock across wireless devices [15], or applying the timing advance technique widely adopted in 4G and 5G [48].

wireless channels, the signal observed at the edge server is given by

$$\begin{aligned}\boldsymbol{y}^{[t]} &= \sum_{k\in\mathcal{K}} (\boldsymbol{G}^{[t]})^{\mathsf{H}} \boldsymbol{\Theta}^{[t]} \boldsymbol{h}_k^{[t]} \alpha_k^{[t]} s_{k,j}^{[t]} + \boldsymbol{n}^{[t]} \\ &= \sum_{k\in\mathcal{K}} \boldsymbol{E}_k^{[t]} \boldsymbol{\theta}^{[t]} \alpha_k^{[t]} s_{k,j}^{[t]} + \boldsymbol{n}^{[t]}, \end{aligned} \tag{6}$$

where $s_{k,j}^{[t]}$ denotes the j-th entry in $\boldsymbol{s}_k^{[t]}$ and $\alpha_k^{[t]} \in \mathbb{C}$ denotes the transmit scalar of the k-th device, $\boldsymbol{E}_k^{[t]} = (\boldsymbol{G}^{[t]})^{\mathsf{H}} \mathrm{diag}(\boldsymbol{h}_k^{[t]})$ denotes the cascaded channel coefficient between device k and the edge server via RIS, and $\boldsymbol{n}^{[t]} \sim \mathcal{CN}(\boldsymbol{0}, \sigma_n^2 \boldsymbol{I})$ denotes additive white Gaussian noise (AWGN). Since we only consider the single-cell setting, the inter-cell interference can be ignored. Note that when synchronization errors are inevitable, the method proposed in [49] can be adopted. Specifically, the received signal (6) is first oversampled using a rectangular pulse to collect sufficient statistics. Then, a sum-product maximum likelihood estimator and an aligned-sample estimator are devised to estimate the arithmetic sum of the transmitted symbols from the edge devices. We assume that there are only single-reflection RIS signals in the network, due to the severe path loss [47]. After performing receive beamforming, the normalized gradient information transmitted via AirComp can be expressed as

$$\hat{f}_j^{[t]} = (\boldsymbol{\omega}^{[t]})^{\mathsf{H}} \boldsymbol{y}^{[t]} = (\boldsymbol{\omega}^{[t]})^{\mathsf{H}} \sum_{k\in\mathcal{K}} \boldsymbol{E}_k^{[t]} \boldsymbol{\theta}^{[t]} \alpha_k^{[t]} s_{k,j}^{[t]} + (\boldsymbol{\omega}^{[t]})^{\mathsf{H}} \boldsymbol{n}^{[t]}, \tag{7}$$

where $\boldsymbol{\omega}^{[t]} \in \mathbb{C}^M$ denotes the receive beamforming vector. The edge server aims to acquire the following target value via AirComp is

$$f_j^{[t]} = \sum_{k\in\mathcal{K}} v_k^{[t]} s_{k,j}^{[t]}. \tag{8}$$

After successfully receiving all gradient entries $\{\boldsymbol{g}_k^{[t]}\}$, the edge server carries out the following post-processing to reconstruct the global gradient,

$$\hat{\boldsymbol{g}}^{[t]} = \frac{1}{K} \left(\hat{\boldsymbol{f}}^{[t]} + \sum_{k\in\mathcal{K}} \bar{g}_k^{[t]} \boldsymbol{1} \right), \tag{9}$$

where $\hat{\boldsymbol{f}}^{[t]} = [\hat{f}_1^{[t]}, \ldots, \hat{f}_d^{[t]}]$.

C. FL Convergence Analysis

To evaluate how transmission quality and AirComp accuracy influence FL performance, we provide the analysis of proposed framework in Proposition 1 with the assistance of the following key assumptions.

Assumption 1: [Smoothness] For any two different model parameter vectors $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^d$, there exists a Lipschitz constant S making the global loss function $L(\cdot)$ such that

$$L(\boldsymbol{z}') \le L(\boldsymbol{z}) + \nabla L(\boldsymbol{z})^{\mathsf{T}} (\boldsymbol{z}' - \boldsymbol{z}) + \frac{S}{2} \|\boldsymbol{z}' - \boldsymbol{z}\|^2. \tag{10}$$

Assumption 2: [μ-Polyak-Łojasiewicz (PL) Condition [50]] For any differentiable global loss function $L(\cdot) : \mathbb{R}^d \to \mathbb{R}$ and model parameter vectors $\boldsymbol{z}, \boldsymbol{z}^\star \in \mathbb{R}^d$, there exists a constant $\mu > 0$ such that

$$\|\nabla L(\boldsymbol{z})\|^2 \ge 2\mu \big(L(\boldsymbol{z}) - L(\boldsymbol{z}^\star) \big), \forall\, \boldsymbol{z} \in \mathbb{R}^d. \tag{11}$$

Proposition 1: Following the above assumptions, the deviation of the global loss from its optima after T training rounds is upper-bounded by

$$\begin{aligned} &\mathbb{E}\left[L(\boldsymbol{z}^{[T]})\right] - \mathbb{E}\left[L(\boldsymbol{z}^\star)\right] \\ &\le \underbrace{\Delta^T \left(\mathbb{E}\left[L(\boldsymbol{z}^{[0]})\right] - \mathbb{E}\left[L(\boldsymbol{z}^\star)\right] \right)}_{\text{Initial error}} \\ &\quad + \underbrace{\sum_{t=0}^{T-1} \frac{rd}{2K^2} \Delta^{T-1-t} \mathsf{MSE}(\hat{f}_j^{[t]}, f_j^{[t]})}_{\text{Aggregation error}}. \end{aligned} \tag{12}$$

where $\mathsf{MSE}(\hat{f}_j^{[t]}, f_j^{[t]})$ stands for the MSE in the t-th training round between the distorted normalized gradient $\hat{f}_j^{[t]}$ and the ground-truth $f_j^{[t]}$, which is computed as

$$\begin{aligned} &\mathsf{MSE}(\hat{f}_j^{[t]}, f_j^{[t]}) \\ &= \mathbb{E}\left[|\hat{f}_j^{[t]} - f_j^{[t]}|^2 \right] \\ &= \sum_{k\in\mathcal{K}} \left| (\boldsymbol{\omega}^{[t]})^{\mathsf{H}} \boldsymbol{E}_k^{[t]} \boldsymbol{\theta}^{[t]} \alpha_k^{[t]} - v_k^{[t]} \right|^2 + \sigma_n^2 \|\boldsymbol{\omega}^{[t]}\|^2. \end{aligned} \tag{13}$$

Proof: See Appendix A. □

Observed from (12), we find that the performance of proposed FL framework after T training rounds is primarily affected by the combined error due to initialization and non-ideal gradient aggregation. The initial error can be mitigated through performing additional training rounds, while the gradient aggregation error is caused by the imperfect AirComp design owing to noise, unfavorable phase-shift on RIS and imperfect channel estimation. Accordingly, we can accelerate the FL convergence by alleviating the adverse effect of gradient aggregation error in each training round. Unless otherwise specified, the time index is omitted for convenience in the following sections.

D. Problem Formulation

Based on the aforementioned analysis, we aim to design an AirComp scheme to facilitate FL by jointly optimizing the transceiver beamforming (i.e., $\boldsymbol{\omega}$ and $\{\alpha_k\}$) and the RIS phase-shift vector $\boldsymbol{\theta}$ with the goal minimizing the MSE between $\hat{f}_j$ and f_j in each training round. Mathematically, the optimization problem can be expressed as

$$\begin{aligned} \mathscr{P}_1 : \underset{\{\alpha_k\},\boldsymbol{\omega},\boldsymbol{\theta}}{\text{minimize}} \quad & \mathsf{MSE}(\hat{f}_j, f_j) && \text{(14a)} \\ \text{subject to} \quad & |\alpha_k|^2 \le \bar{P}_k,\ \forall\, k, && \text{(14b)} \\ & |\theta_n| = 1,\ \forall\, n, && \text{(14c)} \end{aligned}$$

where (14b) refers to the maximum transmit power budget $\bar{P}_k$ of device $k \in \mathcal{K}$ and (14c) is the unit-modulus constraint of each reflecting element at the RIS. Note that the optimal transmit scalar $\{\alpha_k\}$ can be obtained in a closed form under the

assumption of perfect CSI [51]. However, for problems with imperfect CSI, which will be further discussed in the following subsection, this zero-forcing approach is not applicable.

E. Transceiver Design Under Imperfect CSI

It is crucial for the edge server to obtain the perfect CSI so that it can take full advantage of RIS. However, because of the passive nature of RIS, obtaining perfect CSI of the cascaded channel is challenging [46]. Considering the inaccurate estimation of the cascaded channel, we characterize the CSI uncertainty by adopting the additive error model as in [43]. In particular, the actual channel response between the edge server and the k-th edge device is expressed as

$$\boldsymbol{E}_k = \hat{\boldsymbol{E}}_k + \Delta \boldsymbol{E}_k, \ \forall k, \tag{15}$$

where $\hat{\boldsymbol{E}}_k$ denotes the nominal estimated cascaded channel vectors and $\Delta \boldsymbol{E}_k$ is the corresponding random error that takes different forms under different error models. We consider the robust transceiver design under the following two different CSI error models.

1) Expectation-Based Error Model: The expectation-based error model is a stochastic method that finds extensive application in the cases that channel statistics can be easily acquired. Such a model is applicable when the errors are mainly caused by inaccurate channel estimation [52].

Definition 1 (Expectation-Based Error Model [53]): In the expectation-based error model, the CSI error is modeled as a random variable that obeys the Gaussian distribution with zero mean and variance $\sigma_{e,k}^2$, i.e.,

$$\mathbb{E}\{\mathrm{vec}(\Delta \boldsymbol{E}_k)\} = \boldsymbol{0} \text{ and } \mathbb{E}\{\mathrm{vec}(\Delta \boldsymbol{E}_k)\mathrm{vec}^{\mathsf{H}}(\Delta \boldsymbol{E}_k)\} = \sigma_{e,k}^2 \boldsymbol{I}. \tag{16}$$

In light of the expectation-based channel uncertainty model, the objective function (14a) can thus be reformulated as

$$\begin{aligned} & f(\{\alpha_k\}, \boldsymbol{\omega}, \boldsymbol{\theta}) \\ &= \sum_{k\in\mathcal{K}} |\alpha_k|^2 \left(\boldsymbol{\omega}^{\mathsf{H}} \boldsymbol{A}_k \boldsymbol{\omega}\right) \\ &\quad - 2\sum_{k\in\mathcal{K}} \Re\left\{\boldsymbol{\omega}^{\mathsf{H}} \boldsymbol{a}_k \alpha_k \upsilon_k\right\} + \sigma_n^2 \|\boldsymbol{\omega}\|^2 + \sum_{k\in\mathcal{K}} \upsilon_k^2, \end{aligned} \tag{17}$$

where $\boldsymbol{a}_k = \boldsymbol{E}_k \boldsymbol{\theta}$ and $\boldsymbol{A}_k = \boldsymbol{a}_k \boldsymbol{a}_k^{\mathsf{H}} + \sigma_{e,k}^2 ||\theta||^2 \boldsymbol{I}_M$ as $\mathbb{E}\{\Delta \boldsymbol{G}^{\mathsf{H}} \mathbf{H} \Delta \boldsymbol{G}\} = \sigma_g^2 \mathrm{Tr}\{\mathbf{H}\} \boldsymbol{I}_M$ for any $\mathbf{H} \in \mathbb{C}^{N\times N}$. Therefore, we transform the original problem (14) as

$$\begin{aligned} \mathscr{P}_2 : \underset{\{\alpha_k\}, \boldsymbol{\omega}, \boldsymbol{\theta}}{\text{minimize}} \quad & f(\{\alpha_k\}, \boldsymbol{\omega}, \boldsymbol{\theta}) \\ \text{subject to} \quad & \text{(14b)}, \text{(14c)}. \end{aligned}$$

Problem $\mathscr{P}_2$ is highly intractable due to the coupled phase-shift vector (i.e., $\boldsymbol{\theta}$) and the beamforming variables (i.e., $\{\alpha_k\}$, $\boldsymbol{\omega}$) in (17). Besides, the unit-modulus constraint (14c) is also non-convex. To solve problem $\mathscr{P}_2$, we develop an alternating optimization algorithm together with the Lagrangian dual method and DC programming in Section III.

2) Worst-Case Error Model: The worst-case error model is a deterministic model, where the CSI uncertainty is specified by a bounded uncertainty set. Such a model is commonly adopted for characterizing the instantaneous estimated CSI, where errors are predominantly introduced by quantization [54], or in scenarios that the CSI statistic is unavailable.[3]

Definition 2 (Worst-Case Error Model [55]): Under the worst-case error model, the Frobenius norm of random CSI error is upper-bounded by a constant, i.e.,

$$\|\Delta \boldsymbol{E}_k\|^2 \le \zeta_{e,k}^2, \forall\, k$$

where $\zeta_{e,k}$ is the radius of the bounded channel uncertainty region.

Considering the worst-case CSI error, problem (14) is rewritten as

$$\mathscr{P}_3 : \underset{\{\alpha_k\}, \boldsymbol{\theta}, \boldsymbol{\omega}}{\text{minimize}} \ \underset{\Delta \boldsymbol{E}_k}{\text{maximize}} \sum_{k\in\mathcal{K}} \left\| \boldsymbol{\omega}^{\mathsf{H}} \boldsymbol{E}_k \boldsymbol{\theta} \alpha_k - \upsilon_k \right\|^2 + \sigma_n^2 \|\boldsymbol{\omega}\|^2 \tag{18a}$$

$$\text{subject to } \|\Delta \boldsymbol{E}_k\|^2 \le \zeta_{e,k}^2, \ \forall\, k, \quad \text{(14b)}, \text{(14c)}, \tag{18b}$$

where constraint (18b) warrants the robustness for any channel estimation realization in the uncertainty region.

Solving the formulated min-max optimization problem $\mathscr{P}_3$ is challenging. One challenge is that both the objective function (18a) with coupled optimization variables and the unit-modulus constraint (14c) are non-convex. Another challenge is that it is not clear which $\Delta \boldsymbol{E}_k$ maximizes the inner problem. To address these challenges, we shall develop an alternating optimization algorithm using Schur's complement to decouple problem $\mathscr{P}_3$ into several convex subproblems in Section IV.

III. Robust Design Under Expectation-Based Error Model

In this section, the expectation-based robust design problem formulated in Section II-E is tackled by developing an alternating minimization algorithm.

A. Alternating Minimization for Expectation-Based Robust Design

Given $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$, the transmit scalar $\{\alpha_k, \forall\, k\}$ is optimized by solving the following convex problem

$$\begin{aligned} \mathscr{P}_{2.1} : \underset{\{\alpha_k\}}{\text{minimize}} \quad & f(\{\alpha_k\}) \\ \text{subject to} \quad & \text{(14b)}. \end{aligned} \tag{19a}$$

By introducing Lagrange multipliers $\{\mu_k, \forall\, k\}$ to be associated with constraint (14b), the corresponding Lagrangian function is constructed as

$$L(\alpha_k, \mu_k) = f(\{\alpha_k\}) + \sum_{k\in\mathcal{K}} \mu_k (|\alpha_k|^2 - \bar{P}_k), \tag{20}$$

[3]The expectation-based error model demonstrates superior robustness and is often regarded as the solution with higher priority. However, when the statistics or historical samples of channel error is unavailable, the worst-case error model, applicable to more general scenarios, can be utilized as an alternative.

and we can optimize transmit scalar $\{\alpha_k\}$ by solving its dual problem. The optimal solutions $\{\alpha_k^\star\}$ should satisfy the following KKT conditions

$$\alpha_k = \frac{\upsilon_k \boldsymbol{a}_k^{\mathsf{H}} \boldsymbol{\omega}}{\boldsymbol{\omega}^{\mathsf{H}} \boldsymbol{A}_k \boldsymbol{\omega} + \mu_k}, \tag{21a}$$

$$\mu_k \geq 0, \tag{21b}$$

$$|\alpha_k|^2 - \bar{P}_k \leq 0, \tag{21c}$$

$$\mu_k(|\alpha_k|^2 - \bar{P}_k) = 0. \tag{21d}$$

Based on (21a), we can calculate the optimal transmit scalar $\alpha_k^\star$ after obtaining the value of multiplier μ_k. Although the closed-form expression of μ_k is not available, it can be obtained by utilizing the bisection search method. Specifically, by iteratively bisecting the search range of μ_k, we can update the value of α_k in (21a) until the complementary slackness condition (21d) is satisfied [56].

Now we optimize the receive beamforming vector $\boldsymbol{\omega}$ with given transmit scalar $\{\alpha_k\}$ and phase-shift vector $\boldsymbol{\theta}$. The original problem $\mathscr{P}_3$ can be rewritten as

$$\mathscr{P}_{2.2}: \underset{\boldsymbol{\omega}}{\text{minimize}} \;\; f(\boldsymbol{\omega}). \tag{22}$$

This is an unconstrained convex optimization problem. By taking the first derivative of $f(\boldsymbol{\omega})$ with respect to $\boldsymbol{\omega}$, we obtain the optimal $\boldsymbol{\omega}^\star$, as

$$\boldsymbol{\omega}^\star = \left(\sum_{k\in\mathcal{K}} |\alpha_k|^2 \boldsymbol{A}_k + \sigma_n^2 \boldsymbol{I}_M\right)^{-1} \sum_{k\in\mathcal{K}} \boldsymbol{a_k} \alpha_k \upsilon_k. \tag{23}$$

When the beamforming variables $\{\alpha_k\}$ and $\boldsymbol{\omega}$ are fixed, the subproblem of $\mathscr{P}_2$ with respect to $\boldsymbol{\theta}$ is simplified as

$$\mathscr{P}_{2.3}: \underset{\boldsymbol{\theta}}{\text{minimize}} \;\; f(\boldsymbol{\theta}) \quad \text{subject to } |\boldsymbol{\theta}_n| = 1, \; \forall n. \tag{24a}$$

By omitting the constant terms in (17) and introducing auxiliary variables as $\boldsymbol{\Phi}_k = \sigma_{e,k}\|\boldsymbol{\omega}\alpha_k^*\|\boldsymbol{I}_N$ and $\boldsymbol{\gamma}_k = \hat{\boldsymbol{E}}_k^{\mathsf{H}}\boldsymbol{\omega}\alpha_k^*$, the objective function can be equivalently rewritten as

$$\sum_{k\in\mathcal{K}} \boldsymbol{\theta}^{\mathsf{H}}\left(\boldsymbol{\Phi_k}\boldsymbol{\Phi_k}^{\mathsf{H}} + \boldsymbol{\gamma}_k\boldsymbol{\gamma}_k^{\mathsf{H}}\right)\boldsymbol{\theta} - \sum_{k\in\mathcal{K}} 2\Re\left\{\boldsymbol{\theta}^{\mathsf{H}}\boldsymbol{\gamma}_k\upsilon_k\right\}. \tag{25}$$

To further simplify problem $\mathscr{P}_{2.3}$, we denote $\boldsymbol{b} = [\boldsymbol{\theta}, 1]^{\mathsf{T}}$ and matrices $\{\boldsymbol{Q}_k\}$ as

$$\boldsymbol{Q}_k = \begin{bmatrix} \boldsymbol{\gamma}_k\boldsymbol{\gamma}_k^{\mathsf{H}} + \boldsymbol{\Phi_k}\boldsymbol{\Phi_k}^{\mathsf{H}} & -\upsilon_k\boldsymbol{\gamma}_k \\ -(\upsilon_k\boldsymbol{\gamma}_k^{\mathsf{H}}) & 0 \end{bmatrix}. \tag{26}$$

Subsequently, problem (24) can be reformulated as follows

$$\underset{\boldsymbol{b}}{\text{minimize}} \quad \sum_{k\in\mathcal{K}} \boldsymbol{b}^{\mathsf{H}}\boldsymbol{Q}_k\boldsymbol{b} \tag{27a}$$

$$\text{subject to} \quad |v_n|^2 = 1, \forall n, \tag{27b}$$

which is a QCQP problem. To address the non-convexity of unit modulus constraint (27b), we utilize the matrix lifting technique and further transform problem (27) into a semi-definite programming (SDP) problem, where the variables are semi-positive definite matrices [57]. Specifically, by denoting $\boldsymbol{B} = \boldsymbol{b}\boldsymbol{b}^{\mathsf{H}}$, problem (27) is recast as the following low-rank matrix SDP problem

$$\underset{\boldsymbol{B}}{\text{minimize}} \quad \sum_{k\in\mathcal{K}} \text{Tr}(\boldsymbol{Q}_k\boldsymbol{B}) \tag{28a}$$

$$\text{subject to} \quad \boldsymbol{B}_{n,n} = 1, \forall n, \tag{28b}$$

$$\boldsymbol{B} \succeq 0, \tag{28c}$$

$$\text{rank}(\boldsymbol{B}) = 1. \tag{28d}$$

Nevertheless, the rank-one constraint (28d) still renders problem (28) non-convex. To solve this problem, particularly when the optimization variables are high dimensional, we develop an effective DC algorithm to solve problem (28).

B. DC Algorithm for Rank-One Constrained Problem

For rank-one matrix $\boldsymbol{X} \in \mathbb{C}^{L\times L}$, there is only one non-zero singular value, which implies that

$$\left\|\{\sigma_i(\boldsymbol{X})\}_{i=1}^L\right\|_0 = 1 \Leftrightarrow \|\boldsymbol{X}\|_* = \sum_{i=1}^L \sigma_i(\boldsymbol{X}) = \sigma_1(\boldsymbol{X}), \tag{29}$$

where $\sigma_i(\boldsymbol{X})$ represents the i-th largest singular value of matrix $\boldsymbol{X}$. Thus, the rank-one constraint (28d) can be substituted by an alternative expression as shown in Proposition 2.

Proposition 2: [DC representation] If positive semi-definite (PSD) matrix $\boldsymbol{X} \in \mathbb{C}^{L\times L}$ satisfies $\text{Tr}(X) \geq 0$, then $\text{rank}(\boldsymbol{X}) = 1$ can be equivalently expressed as the zero difference between its nuclear norm and spectral norm [58], i.e.,

$$\text{rank}(\boldsymbol{X}) = 1 \Leftrightarrow \|\boldsymbol{X}\|_* - \|\boldsymbol{X}\|_2 = 0 \tag{30}$$

where $\|\boldsymbol{X}\|_* = \text{Tr}(\boldsymbol{X}) = \sum_{i=1}^L \sigma_l(\boldsymbol{X})$ and $\|\boldsymbol{X}\|_2 = \sigma_1(\boldsymbol{X})$.

Based on the above proposition, the rank-one constraint (28d) can be rewritten as

$$\text{Tr}(\boldsymbol{B}) - \|\boldsymbol{B}\|_2 = 0. \tag{31}$$

By taking the left-hand-side of (31) as a penalty component, problem (28) is reformulated as

$$\underset{\boldsymbol{B}}{\text{minimize}} \quad \sum_{k\in\mathcal{K}} \text{Tr}(\boldsymbol{Q}_k\boldsymbol{B}) + \rho_e(\text{Tr}(\boldsymbol{B}) - \|\boldsymbol{B}\|_2) \tag{32a}$$

$$\text{subject to} \quad \boldsymbol{B}_{n,n} = 1, \forall n \tag{32b}$$

$$\boldsymbol{B} \succeq 0, \tag{32c}$$

where the penalty parameter $\rho_e > 0$. When the penalty term vanishes, a rank-one solution can be obtained.

Because of the concave term $-\rho_e\|\boldsymbol{B}\|_2$ in the penalty component, problem (32) is non-convex. With successive convex approximation, this challenge can be solved by linearizing this concave term in a series of convex subproblems [59]. Specifically, the j_1-th subproblem is given by

$$\underset{\boldsymbol{B}}{\text{minimize}} \quad \sum_{k\in\mathcal{K}} \text{Tr}(\boldsymbol{Q}_k\boldsymbol{B}) + \rho_e\langle\boldsymbol{B}, \boldsymbol{I} - \partial_{\boldsymbol{B}^{j_1-1}}\|\boldsymbol{B}\|_2\rangle \tag{33a}$$

$$\text{subject to} \quad \boldsymbol{B}_{n,n} = 1, \forall n, \tag{33b}$$

$$\boldsymbol{B} \succeq 0, \tag{33c}$$

where $\partial_{\boldsymbol{B}^{j_1-1}}\|\boldsymbol{B}\|_2$ is the subgradient of $\|\boldsymbol{B}\|_2$ with respect to $\boldsymbol{B}$ at the optimal solution of the (j_1-1)-th subproblem $\boldsymbol{B}^{j_1-1}$. By denoting $\boldsymbol{u}_1$ as the eigenvector that corresponds to the largest eigenvalue $\sigma_1(\boldsymbol{B})$, the subgradient $\partial_{\boldsymbol{B}^{j_1-1}}\|\boldsymbol{B}\|_2$ can be computed as $\boldsymbol{u}_1\boldsymbol{u}_1^{\mathsf{H}}$. Obviously, subproblem (33) is convex and can thus be solved by adopting the CVX [60] solver.

In summary, the MSE minimization problem $\mathscr{P}_3$, under the expectation-based error model, is solved by Algorithm 1 with random initial $\boldsymbol{\theta}^0$ and $\boldsymbol{\omega}^0$. Transmit scalar $\{\alpha_k^\star\}$ is optimized by solving the Lagrangian dual problem of (19) and then updated based on the KKT conditions (21) in each iteration. Besides, the optimal receive beamforming vector $\boldsymbol{\omega}^\star$ is derived based on (23). An exact rank-one solution $\boldsymbol{B}^\star$ can be obtained by iteratively solving subproblems (33)until the penalty term goes to zero. According to Cholesky decomposition $\boldsymbol{B}^\star = \boldsymbol{b}^\star(\boldsymbol{b}^\star)^{\mathsf{H}}$, the feasible solution of problem (24) can be computed by $\boldsymbol{\theta}^\star = [\boldsymbol{b}^\star/b^\star_{N+1}]_{(1:N)}$, where $[\boldsymbol{y}]_{(1:N)}$ represents the first N entries of vector $\boldsymbol{y}$. The variables $\{\alpha_k\}$, $\boldsymbol{\omega}$, and $\boldsymbol{\theta}$ are updated in an iterative manner until the decrease of the objective function of problem $\mathscr{P}_2$ below a preset threshold ϵ, and uppercase $[j]$ and $[j_1]$ denote the j-th iteration of the proposed alternating algorithm and j_1-th sub-problem in DC algorithm.

IV. Robust Design Under Worst-Case Error Model

In this section, we first develop a robust algorithm in the presence of the CSI error that obeys the worst-case error model, and subsequently design a Schur's complement and PCCP based alternating optimization algorithm.

A. Alternating Minimization for Worst-Case Robust Design

By introducing auxiliary variables $\{u_k, \forall k\}$, we can reformulate problem $\mathscr{P}_3$ as

$$\begin{aligned}&\underset{\{\alpha_k\},\boldsymbol{\theta},\boldsymbol{\omega},\{u_k\}}{\text{minimize}} \sum_{k\in\mathcal{K}} u_k + \sigma_n^2\|\boldsymbol{\omega}\|^2\\ &\text{subject to} \quad \left\|\boldsymbol{\omega}^{\mathsf{H}}\left(\hat{\boldsymbol{E}}_k + \Delta\boldsymbol{E}_k\right)\boldsymbol{\theta}\alpha_k - \upsilon_k\right\|^2 \le u_k, \quad (34a)\\ &\qquad u_k \ge 0,\\ &\qquad (14b), (14c), (18b). \quad (34b)\end{aligned}$$

Due to the channel uncertainty in constraints (18b) and (34a), problem (34) is still intractable. As the value of the determinant of any PSD matrix is greater than or equal to zero, constraint (34a) can be equivalently expressed as (35), as shown at the bottom of the page. Then, two important lemmas widely applied in the robust optimization to handle the random error are introduced as follows.

Lemma 1: [Schur's complement [56]] For any Hermitian matrix $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}^{\mathsf{H}}\\ \boldsymbol{B} & \boldsymbol{C}\end{bmatrix}$, $\boldsymbol{X} \succeq \boldsymbol{0}$ holds if and only if $\boldsymbol{A} - \boldsymbol{B}^{\mathsf{H}}\boldsymbol{C}^{-1}\boldsymbol{B} \succeq \boldsymbol{0}$ with $\boldsymbol{C}$ being invertible, or $\boldsymbol{C} - \boldsymbol{B}^{\mathsf{H}}\boldsymbol{A}^{-1}\boldsymbol{B} \succeq \boldsymbol{0}$ with $\boldsymbol{A}$ being invertible.

Lemma 2: [S-procedure [56]] Let $h_m(\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{H}}\boldsymbol{F}_m\boldsymbol{x} + 2\Re\{\boldsymbol{g}_m^{\mathsf{H}}\boldsymbol{x}\} + h_m, m = 1, 2$ be two quadratic functions with respect to $\boldsymbol{x}$, where $\boldsymbol{F}_m \in \mathbf{S}^L$, $\boldsymbol{g}_m \in \mathbb{C}^{L\times 1}$, and $h_m \in \mathbb{R}$. If there exists $\hat{\boldsymbol{x}}$ satisfying $h_1(\hat{\boldsymbol{x}}) > 0$, then $h_1(\boldsymbol{x}) \ge 0 \Rightarrow h_2(\boldsymbol{x}) \ge 0$ holds if and only if there exists $\lambda \ge 0$ such that

$$\begin{bmatrix}\boldsymbol{F}_2 & \boldsymbol{g}_2\\ \boldsymbol{g}_2 & h_2\end{bmatrix} \succeq \lambda \begin{bmatrix}\boldsymbol{F}_1 & \boldsymbol{g}_1\\ \boldsymbol{g}_1 & h_1\end{bmatrix}. \quad (36)$$

According to Lemma 1, constraint (35) can be reformulated as

$$\boldsymbol{X}_k - \left[\boldsymbol{y}_k^{\mathsf{H}}\mathrm{vec}(\Delta\boldsymbol{E}_k)\boldsymbol{c} + \boldsymbol{c}^{\mathsf{H}}\mathrm{vec}^{\mathsf{H}}(\Delta\boldsymbol{E}_k)\boldsymbol{y}_k\right] \succeq \boldsymbol{0}, \forall\, k, \quad (37)$$

where

$$\begin{aligned}\boldsymbol{X}_k &= \begin{bmatrix} u_k & (\boldsymbol{\omega}^{\mathsf{H}}\hat{\boldsymbol{E}}_k\boldsymbol{\theta}\alpha_k - \upsilon_k)^*\\ (\boldsymbol{\omega}^{\mathsf{H}}\hat{\boldsymbol{E}}_k\boldsymbol{\theta}\alpha_k - \upsilon_k) & 1\end{bmatrix},\\ \boldsymbol{y}_k &= \begin{bmatrix}\boldsymbol{0} & -\mathrm{vec}(\boldsymbol{\omega}\alpha_k^*\boldsymbol{\theta}^{\mathsf{H}})\end{bmatrix} \text{ and } \boldsymbol{c} = \begin{bmatrix}1 & \boldsymbol{0}\end{bmatrix}. \quad (38)\end{aligned}$$

To further solve the challenge of the uncertainty vector in (18b) and (35), another lemma is given the following.

Lemma 3: For function $\mathbf{F}(\boldsymbol{X}) = \boldsymbol{A} - (\boldsymbol{B}^{\mathsf{H}}\boldsymbol{X}\boldsymbol{C} + \boldsymbol{C}^{\mathsf{H}}\boldsymbol{X}^{\mathsf{H}}\boldsymbol{B})$, where Hermitian matrix $\boldsymbol{A} \in \mathbb{C}^{L\times L}, \boldsymbol{B} \in \mathbb{C}^{M\times L}, \boldsymbol{C} \in \mathbb{C}^{N\times L}$, and variable $\boldsymbol{X} \in \mathbb{C}^{M\times N}$, we have

$$\mathbf{F}(\boldsymbol{X}) \succeq \boldsymbol{0},\ \forall\, \boldsymbol{X} : \|\boldsymbol{X}\|_F < \epsilon$$

when there exists $\lambda \ge 0$, such that

$$\begin{bmatrix}\boldsymbol{A} - \lambda\boldsymbol{C}^{\mathsf{H}}\boldsymbol{C} & -\epsilon\boldsymbol{B}^{\mathsf{H}}\\ -\epsilon\boldsymbol{B} & \lambda\boldsymbol{I}\end{bmatrix} \succeq \boldsymbol{0}.$$

Proof: According to the definition of PSD matrix, $\mathbf{F}(\boldsymbol{X}) \succeq \boldsymbol{0}, \forall\, \boldsymbol{X} : \|\boldsymbol{X}\|_F \le \epsilon$ holds if and only if for $\forall\, \boldsymbol{z} \ne \boldsymbol{0}$,

$$\boldsymbol{z}^H\boldsymbol{A}\boldsymbol{z} - \boldsymbol{z}^H(\boldsymbol{B}^{\mathsf{H}}\boldsymbol{X}\boldsymbol{C} + \boldsymbol{C}^{\mathsf{H}}\boldsymbol{X}^{\mathsf{H}}\boldsymbol{B})\boldsymbol{z} \ge 0. \quad (39)$$

Based on Lemma 1, inequality (39) is equivalent to

$$\begin{aligned}&\boldsymbol{z}^H\boldsymbol{A}\boldsymbol{z}\\ &\ge \max \boldsymbol{z}^H(\boldsymbol{B}^{\mathsf{H}}\boldsymbol{X}\boldsymbol{C} + \boldsymbol{C}^{\mathsf{H}}\boldsymbol{X}^{\mathsf{H}}\boldsymbol{B})\boldsymbol{z} = 2\epsilon\|\boldsymbol{C}^{\mathsf{H}}\boldsymbol{z}\|\|\boldsymbol{B}\boldsymbol{z}\|. \quad (40)\end{aligned}$$

Based on the Cauchy-Schwarz inequality, inequality (40) can be further transformed as

$$\boldsymbol{z}^H\boldsymbol{A}\boldsymbol{z} - 2\epsilon\eta\boldsymbol{C}^{\mathsf{H}}\boldsymbol{z} \ge 0, \forall\, \eta : |\eta| \le \|\boldsymbol{B}\boldsymbol{z}\|. \quad (41)$$

By applying the S-procedure in Lemma 2, inequality (41) holds when there exists a $\lambda \ge 0$ such that

$$\begin{bmatrix}\boldsymbol{A} - \lambda\boldsymbol{C}^{\mathsf{H}}\boldsymbol{C} & -\epsilon\boldsymbol{B}^{\mathsf{H}}\\ -\epsilon\boldsymbol{B} & \lambda\boldsymbol{I}\end{bmatrix} \succeq \boldsymbol{0}.$$

□

$$\begin{bmatrix} u_k & \left[(\boldsymbol{\omega}^{\mathsf{H}}\hat{\boldsymbol{E}}_k\boldsymbol{\theta}\alpha_k - \upsilon_k) + \mathrm{vec}^{\mathsf{H}}(\boldsymbol{\omega}\alpha_k^*\boldsymbol{\theta}^{\mathsf{H}})\mathrm{vec}(\Delta\boldsymbol{E}_k)\right]^*\\ (\boldsymbol{\omega}^{\mathsf{H}}\hat{\boldsymbol{E}}_k\boldsymbol{\theta}\alpha_k - \upsilon_k) + \mathrm{vec}^{\mathsf{H}}(\boldsymbol{\omega}\alpha_k^*\boldsymbol{\theta}^{\mathsf{H}})\mathrm{vec}(\Delta\boldsymbol{E}_k) & 1\end{bmatrix} \succeq \boldsymbol{0}, \forall\, k, \quad (35)$$

Algorithm 1 Alternating Optimization Algorithm Under the Expectation-Based Error Model

Input: Initial arbitrary $\boldsymbol{\theta}^0$, $\boldsymbol{\omega}^0$, and $\epsilon > 0$
$j = 1$
repeat
  Optimize $\{\alpha_k^{[j]}\}$, $\forall\, k \in \mathcal{K}$, by solving (21) with fixed $\boldsymbol{\omega}^{[j-1]}$ and $\boldsymbol{\theta}^{[j-1]}$
  Optimize $\boldsymbol{\omega}^{[j]}$ according to (23) with fixed $\boldsymbol{\theta}^{[j-1]}$ and $\{\alpha_k^{[j]}\}$,
  repeat
    Compute a subgradient $\partial_{\boldsymbol{B}^{[j_1-1]}} \|\boldsymbol{B}\|_2$
    Optimize $\boldsymbol{B}^{[j_1]}$ according to subproblem (28) with given $\{\alpha_k^{[j]}\}$ and $\boldsymbol{\omega}^{[j]}$
    $j_1 = j_1 + 1$
  until *penalty component in (32a) is zero*;
  Recover $\boldsymbol{b}^{[j]}$ based on Cholesky decomposition $\boldsymbol{B}^{j_1} = \boldsymbol{b}^{[j]}(\boldsymbol{b}^{[j]})^{\mathsf{H}}$
  Update $\boldsymbol{\theta}^{[j]} = [\boldsymbol{b}^{[j]}/\boldsymbol{b}_{N+1}^{[j]}]_{(1:N)}$
  $j = j + 1$
until *Problem (28) is infeasible or decrease of the objective function of problem $\mathscr{P}_2$ is below ϵ*;
Output: $\{\alpha_k^{[j]}\}$, $\boldsymbol{\omega}^{[j]}$, $\boldsymbol{\theta}^{[j]}$

Based on Lemma 3, constraints (18b) and (37) can be combined into a convex PSD constraint as follows

$$\begin{bmatrix} u_k - \lambda_k & (\boldsymbol{\omega}^{\mathsf{H}}\hat{\boldsymbol{E}}_k\boldsymbol{\theta}\alpha_k - \upsilon_k)^* & \mathbf{0} \\ (\boldsymbol{\omega}^{\mathsf{H}}\hat{\boldsymbol{E}}_k\boldsymbol{\theta}\alpha_k - \upsilon_k) & 1 & \zeta_{e,k}\mathrm{vec}^{\mathsf{H}}(\boldsymbol{\omega}\alpha_k^*\boldsymbol{\theta}^{\mathsf{H}}) \\ \mathbf{0} & \zeta_{e,k}\mathrm{vec}(\boldsymbol{\omega}\alpha_k^*\boldsymbol{\theta}^{\mathsf{H}}) & \lambda_k\boldsymbol{I}_{MN} \end{bmatrix} \succeq \mathbf{0}, \exists\, \lambda_k \geq 0. \tag{42}$$

If such a λ_k exists, constraint (42) ensures that the MSE constraint (34a) can be satisfied for all random CSI errors in uncertainty set (18b). Thus, problem (34) can be rewritten as

$$\begin{aligned} \underset{\{\alpha_k\},\boldsymbol{\omega},\boldsymbol{\theta},\{u_k\},\{\lambda_k\}}{\text{minimize}} \quad & \sum_{k\in\mathcal{K}} u_k + \sigma_n^2\|\boldsymbol{\omega}\|^2 \\ \text{subject to} \quad & (14b), (14c), (34b), (42) \\ & \lambda_k \geq 0. \end{aligned} \tag{43a}$$

However, this problem is not jointly convex because of the coupled optimization variables (i.e., $\{\alpha_k\}, \boldsymbol{\theta}, \boldsymbol{\omega}$). To handle this problem, we propose an alternating optimization method to update $\{\alpha_k\}, \boldsymbol{\omega}$ and $\boldsymbol{\theta}$ sequentially in an iterative manner.

We first optimize $\{\alpha_k\}$ of edge devices when $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ are given. Thus, we recast problem (43) as

$$\mathscr{P}_{3.1}: \underset{\{\alpha_k\},\{u_k\},\{\lambda_k\}}{\text{minimize}} \quad \sum_{k\in\mathcal{K}} u_k \tag{44a}$$

$$\text{subject to} \quad (14b), (42), (34b), (43a). \tag{44b}$$

Similarly, by fixing transmit power $\{\alpha_k\}$ and phase-shift vector $\boldsymbol{\theta}$, the corresponding subproblem with respect to $\boldsymbol{\omega}$ is given by

$$\mathscr{P}_{3.2}: \underset{\boldsymbol{\omega},\{u_k\},\{\lambda_k\}}{\text{minimize}} \quad \sum_{k\in\mathcal{K}} u_k + \sigma_n^2\|\boldsymbol{\omega}\|^2 \tag{45a}$$

$$\text{subject to} \quad (42), (34b), (43a). \tag{45b}$$

We note that both subproblems $\mathscr{P}_{3.1}$ and $\mathscr{P}_{3.2}$ can be solved with the existing solvers (e.g., CVX).

On the other hand, we optimize the phase-shift vector $\boldsymbol{\theta}$ by fixing transmit power $\{\alpha_k\}$ and receive beamforming vector $\boldsymbol{\omega}$. Accordingly, problem (43) is reformulated as

$$\mathscr{P}_{3.3}: \underset{\boldsymbol{\theta},\{u_k\},\{\lambda_k\}}{\text{minimize}} \quad \sum_{k\in\mathcal{K}} u_k$$
$$\text{subject to} \quad (14c), (42), (34b), (43a).$$

However, problem $\mathscr{P}_{3.3}$ still remains non-convex due to the unit-modulus constraints (14c). To address this issue, we utilize the PCCP algorithm in the following.

B. PCCP for Unit-Modulus Constrained Problem

In particular, we reformulate the initial unit-modulus constraint (14c) as a couple of convex ones, i.e.,

$$|\theta_n|^2 \leq 1, \ \forall\, n \tag{46}$$

$$1 \leq |\theta_n|^2, \ \forall\, n. \tag{47}$$

For the non-convex constraint in (47), it is further linearly approximated as

$$|\theta_n^{[j_1]}|^2 - 2\Re(\theta_n^*\theta_n^{[j_1]}) \leq -1, \ \forall\, n, \tag{48}$$

at fixed $\theta_n^{[j_1]}$. Thus, based on PCCP, problem $\mathscr{P}_{3.3}$ can be tackled by solving a series of subproblems in a successive manner. The specific subproblem in the t-th iteration is given as

$$\underset{\boldsymbol{\theta},\{u_k\},\{\lambda_k\},\boldsymbol{q}}{\text{minimize}} \quad \sum_{k\in\mathcal{K}} u_k + \rho_w^{[j_1]}\|\boldsymbol{q}\|_1$$

$$\text{subject to} \quad (42), (34b), (43a), \tag{49a}$$

$$|\theta_n^{[j_1-1]}|^2 - 2\Re\{\theta_n^*\theta_n^{[j_1-1]}\} \leq q_n - 1, \ \forall\, n, \tag{49b}$$

$$|\theta_n|^2 \leq 1 + q_{N+n}, \ \forall\, n, \tag{49c}$$

$$\boldsymbol{q} \geq \mathbf{0}, \tag{49d}$$

where $\rho_w^{[j_1]}$ is the penalty parameter in the j_1-th subproblem and $\boldsymbol{q} = [q_1, \ldots, q_{2M}]^{\mathsf{T}}$ contains the slack variables introduced to ensure feasible solutions in the phase-shift optimization [43]. With the iteration of the subproblems, the penalty parameter $\rho_w^{[j_1]}$ not only diminishes the slack variables $\|\boldsymbol{q}\|_1$ but also forces the solution to converge toward satisfying the original constraints. The above problem is convex and can be solved by using CVX.

The proposed algorithm for achieving the robust transceiver design of an RIS-assisted AirComp system in the worst-case error model is summarized in Algorithm 2 with random initial $\boldsymbol{\theta}^0$ and $\boldsymbol{\omega}^0$. In j-th iteration of the proposed alternating algorithm, transmit scalar $\{\alpha_k^\star\}$ and receive beamforming vector $\boldsymbol{\omega}^\star$ are updated by solving the SPD problem $\mathscr{P}_{3.1}$ an $\mathscr{P}_{3.2}$, respectively. The optimal phase-shift vector $\boldsymbol{\theta}$ is updated based on the PCCP algorithm by solving a series of subproblems (49) until the value of penalty term $\|\boldsymbol{q}\|_1$ below threshold ϵ_1 and the phase-shift vector $\boldsymbol{\theta}$ converges. The maximum number of iterations $T_{\max}$ warrants that the solution derived from problem (49) is also feasible for

problem $\mathscr{P}_{3.3}$. The variables $\{\alpha_k\}$, $\boldsymbol{\omega}$, and $\boldsymbol{\theta}$ are updated in an iterative manner until the decrease of the objective function of problem $\mathscr{P}_3$ below a preset threshold ϵ.

V. Computational Complexity Analysis

The computational complexity of the proposed two robust designs are given as follows. The proposed Algorithm 1 consists of three sub-problems. The computational complexity of sub-problems $\mathscr{P}_{2.1}$ and $\mathscr{P}_{2.2}$ are $\mathcal{O}(KMN + KM^2)$ and $\mathcal{O}(M^3 + MN + KM)$, respectively. The sub-problem $\mathscr{P}_{2.3}$ involves a series of SDP problems. By adopting the primal-dual path-following method, the worst-case computation complexity for solving each SDP problem is $\mathcal{O}(M^{4.5}\log(1/\epsilon))$, where $\epsilon > 0$ is the given solution precision. Thus the computational complexity of $\mathscr{P}_{2.3}$ is $\mathcal{O}(QM^{4.5}\log(1/\epsilon))$, where Q is the number of DC iterations. The total computational complexity of Algorithm 1 is $C_{\alpha_k}^{\exp} + C_{\boldsymbol{\omega}}^{\exp} + C_{\boldsymbol{\theta}}^{\exp}$. When the number of edge devices (e.g, K) scales up, the computational complexities of solving transmit scalar $\{\alpha_k\}$ and receive beamforming vector $\boldsymbol{\omega}$ grows linearly. Thus, the overall computational complexity increases accordingly. When more reflecting elements are deployed at RIS (e.g, N), the computational complexities of solving transmit scalar $\{\alpha_k\}$, receive beamforming vector $\boldsymbol{\omega}$, and RIS phase-shift matrix $\boldsymbol{\Theta}$ increase, leading to a growth in the overall computational complexity.

For the proposed Algorithm 2, we can adopt standard interior-point method to solve each sub-problem [61]. Thus the computational complexity of $\mathscr{P}_{3.1}$ is $C_{\alpha_k}^{\text{worst}} = \mathcal{O}([K(MN+2)]^{0.5}(K^3 + K^2(MN+2)^3 + K^3(MN+2)^2))$, of $\mathscr{P}_{3.2}$ is $C_{\boldsymbol{\omega}}^{\text{worst}} = \mathcal{O}([K(MN+2)]^{0.5}(M^3 + MK(MN+2)^3 + M^2K(MN+2)^2))$, and of $\mathscr{P}_{3.3}$ is $C_{\boldsymbol{\theta}}^{\text{worst}} = \mathcal{O}([2N + K(MN+2)]^{0.5}(2N^3 + NK(MN+2)^3 + N^2K(MN+2)^2))$. The total computational complexity of Algorithm 2 is $C_{\alpha_k}^{\text{worst}} + C_{\boldsymbol{\omega}}^{\text{worst}} + C_{\boldsymbol{\theta}}^{\text{worst}}$. Similar to Algorithm 1, the overall computational complexity of Algorithm 2 is also monotonically increasing with K and N.

VI. Simulation Results

Through simulation experiments in this section, we validate the superiority of the proposed algorithm against imperfect CSI in a wireless FL network aided by AirComp and RIS.

A. Simulation Setup

In the simulations, we consider a three-dimensional coordinate system where the edge devices and server collaboratively implement an image classification task with multinomial logistic regression on MNIST [62] and CIFAR-10 [63] datasets. The edge server and the RIS are deployed at $(0, 0, 15)$ and $(0, 30, 10)$ meters, respectively. All devices are uniformly located within a half-circle region that is centered at the RIS with radius $R = 25$ meters. We model the path loss as

$$L(d) = L_0(d/d_0)^{-\iota}, \tag{50}$$

where L_0 denotes the path loss at the reference distance $d_0 = 1$ meter, d denotes the link distance, and ι refers to the path loss exponent. We consider that each segment of the cascaded channel suffers from Rician fading as follows

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

where χ denotes the energy ratio factor, $\boldsymbol{H}^{\text{LoS}}$ is the line-of-sight component, and $\boldsymbol{H}^{\text{NLoS}}$ is the non-line-of-sight components. Thus, the channel coefficients are calculated by $\boldsymbol{G} = \sqrt{L(d_{\text{FI}})}\boldsymbol{H}_{\text{FI}}$ and $\boldsymbol{h}_k = \sqrt{L(d_{\text{I,k}})}\boldsymbol{H}_{\text{I,k}}$, where d_{FI} and $d_{\text{I,k}}$ denote the distances from edge server to RIS and from RIS to device $k, \forall k$, respectively. In addition, the variance of $\{\Delta\boldsymbol{E}_k, \forall k,\}$ in the expectation-based error model is defined as $\sigma_{e,k}^2 = \delta_e^2\|\hat{\boldsymbol{E}}_k\|_F^2, \forall k$, where uncertainty ratio $\delta_e \in [0, 1)$.

For fair comparison, the radius of the bounded uncertainty region under the worst-case error model is set as [64]

$$\zeta_{e,k} = \sqrt{\sigma_{e,k}^2 F_{2MN}^{-1}(P_{\text{in}})/2}, \tag{52}$$

where $F_X^{-1}(\cdot)$ denotes the inverse cumulative distribution function (CDF) of the Chi-square distribution with X degrees of

Algorithm 2 Alternating Optimization Algorithm Under the Worst-Case Error Model

Input: Initial arbitrary $\boldsymbol{\theta}^0$, $\boldsymbol{\omega}^0$, $\epsilon > 0$ and $\gamma > 1$
$j = 1$
repeat
  Update $\{\alpha_k^{[j]}\}, \forall k \in \mathcal{K}$, by solving problem $\mathscr{P}_{3.1}$ with fixed $\boldsymbol{\omega}^{[j-1]}$ and $\boldsymbol{\theta}^{[j-1]}$
  Update $\boldsymbol{\omega}^{[j]}$ by solving problem $\mathscr{P}_{3.2}$ with fixed $\boldsymbol{\theta}^{[j-1]}$ and $\{\alpha_k^{[j-1]}\}$
  $j_1 \leftarrow 1$, $\boldsymbol{\theta}^{[j_1-1]} \leftarrow \boldsymbol{\theta}^{[j-1]}$
  repeat
    if $j_1 \le T_{\max}$ **then**
      Update $\boldsymbol{\theta}^{[j_1]}$ by solving problem (49)
      $\rho_w^{[j_1+1]} = \min\{\gamma\rho_w^{[j_1]}, \rho_{max}\}$
      $j_1 \leftarrow j_1 + 1$
    else
      $j_1 \leftarrow 1$ and initialize a new random $\boldsymbol{\theta}^{[j_1-1]}$
    end
  until $\|\boldsymbol{q}\|_1 \le \epsilon_1 \text{and} \|\boldsymbol{\theta}^{[j_1]} - \boldsymbol{\theta}^{[j_1-1]}\| \le \epsilon_2$;
  Update $\boldsymbol{\theta}^{[j]} \leftarrow \boldsymbol{\theta}^{[j_1]}$
  $j = j + 1$
until *The decrease of objective function in problem $\mathscr{P}_3$ is below ϵ*;
Output: $\{\alpha_k^{[j]}\}$, $\boldsymbol{\omega}^{[j]}$, $\boldsymbol{\theta}^{[j]}$

TABLE II
Simulation Parameters

Parameters	Values
Maximum power budget P_k	200 mW
Path loss at the reference distance L_0	-30 dB
Path loss exponent of edge server-RIS link ι_g	2.6
Path loss exponent of RIS-device link ι_k	2.4
Energy ratio factor in Rician fading χ	5
Penalty parameter ρ_e	10
Noise power σ^2	-110 dBm
Uncertainty ratio in the expectation-based error model δ_e	1%
Probability of error within uncertainty region P_{in}	95%

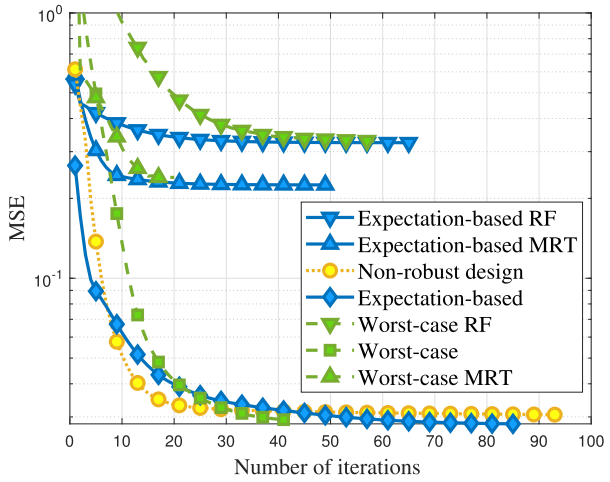


Fig. 2. Convergence of proposed robust algorithms with different benchmarks.

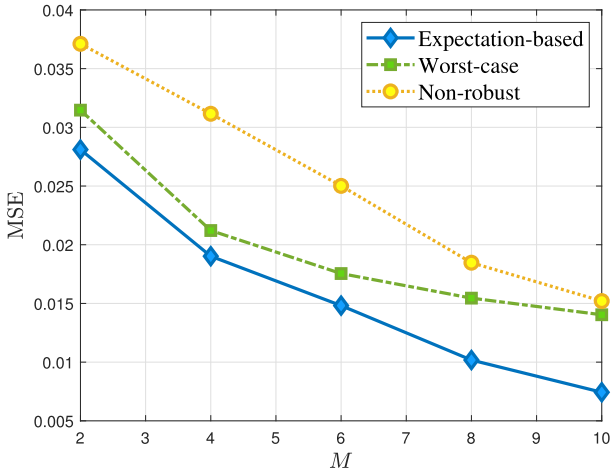


Fig. 3. MSE versus the number of antennas at edge server with different benchmarks.

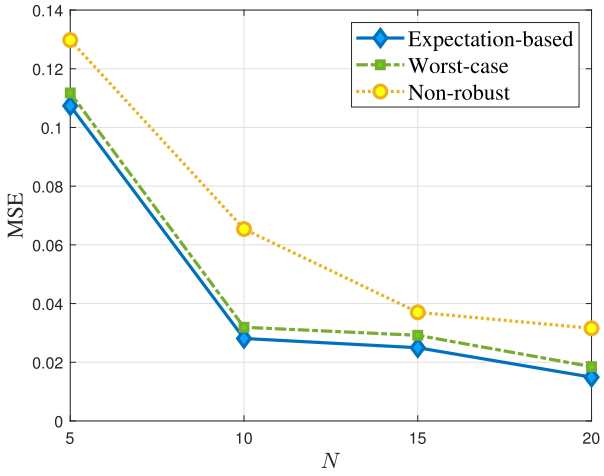


Fig. 4. MSE versus the number of RIS elements with different benchmarks.

freedom and P_{in} is the probability that the CSI error is within the uncertainty region, i.e., $\text{Prob}(\|\Delta \boldsymbol{E}_k\|^2 \leq \zeta_{e,k}^2) = P_{\text{in}}$. Unless specified otherwise, other simulation parameters are listed in Table II. All simulations are performed using the CVX solver on a computer equipped with a 3.8 GHz AMD Ryzen 9 3900X processor and 32 GB of RAM.

B. Performance Comparison

We consider four benchmarks to validate the superiority of proposed robust algorithms, and the descriptions are specified as follows:

- **Perfect CSI** [43], [65]: In this scheme, edge server, RIS and devices have perfect CSI when optimizing variables.
- **Non-robust design** [66]: In this scheme, the errors caused by channel estimation and quantization are ignored throughout the optimization.
- **Random phase-shifts (RF)** [47]: We set each reflecting element at the RIS with a fixed and randomly chosen phase-shift from $[0, 2\pi)$ when optimizing other variables.
- **Maximum-ratio transmission (MRT)** [47]: The transmit scalar of each device is set as $\alpha_k = \sqrt{\bar{P}_k} \frac{(\boldsymbol{\omega}^{\text{H}} \boldsymbol{E}_k \boldsymbol{\theta})^*}{\|\boldsymbol{\omega}^{\text{H}} \boldsymbol{E}_k \boldsymbol{\theta}\|_2}$.

We first show the convergence of our proposed robust algorithms when $N = 10, M = 2$, and $K = 5$ with other benchmarks in Fig. 2. As can be observed, the proposed expectation-based and worst-case robust algorithms outperform their non-robust counterpart, and achieve lower MSE of received gradients, indicating their effectiveness in overcoming the imperfect CSI. This robustness stems from incorporating inherent channel errors into the framework design. By jointly optimizing the AirComp transceivers and RIS phase shifts under channel uncertainty, our proposed algorithms mitigate gradient distortion at the edge server and lead to accurate gradient aggregation. Besides, benefiting from the well-optimized phase-shift at each reflecting element, the proposed robust algorithms also obtain more accurate model parameter aggregation and achieve lower MSE than the random phase-shift approach, which proves that optimizing the phase-shift of RIS is effective in tackling the unfavorable propagation environments. Compared with the MRT scheme, the proposed algorithms also show better gradient aggregation performance. By optimizing the transmit scalar for each device, the proposed algorithms effectively mitigate performance degradation due to the power loss when beam deviates from the optimum.

Fig. 3 shows the average MSE for different number of antennas under different benchmarks when $N = 10$ and $K = 5$. Generally, the average MSE exhibits a decreasing trend with growing antenna scale, owing to improved power gain by receiving more copies of the signals. Hence, deploying a larger number of antennas at the edge server leads to a lower distortion, i.e., a better AirComp performance. In addition, the integration of inherent channel errors into the design framework enables the two proposed robust algorithms to achieve superior performance over the non-robust algorithm, regardless of the number of antennas at the edge server. Compared to the worst-case robust design, the expectation-based algorithm yields lower MSE and more accurate gradient aggregation by leveraging the available channel error statistics.

Fig. 4 shows the average MSE for different number of RIS reflecting elements under different benchmarks when $M = 2$ and $K = 5$. As can be observed, there is a negative correlation between the MSE performance and the number of reflecting

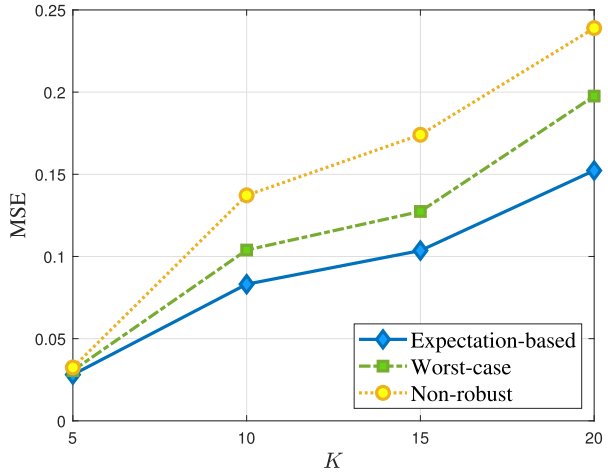


Fig. 5. MSE versus the number of edge devices with different benchmarks.

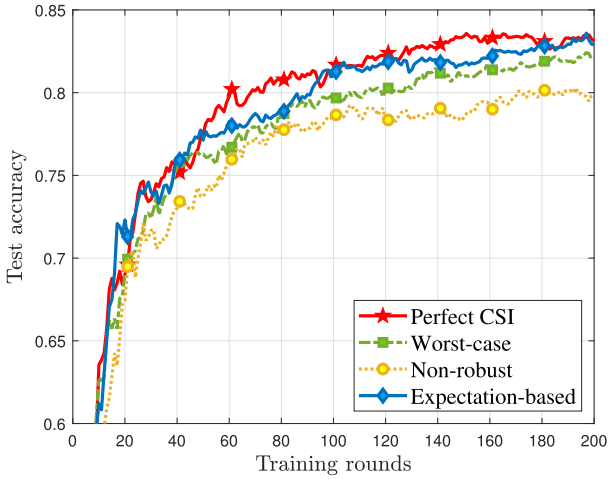


Fig. 6. Test accuracy versus the number of training rounds with different benchmarks on MNIST dataset.

elements. By leveraging more reflecting elements, the RIS provides more accurate passive beamforming, which enhances the incident signal power, thereby reducing the aggregation error at the edge server and improving the overall AirComp performance.

Figs. 5 shows the MSE performance of the proposed robust transceiver design for different number of edge devices under different error models, when $N = 10$ and $M = 2$. It is observed that the MSE of all schemes under consideration increases with the growth of the number of devices. This agrees with our intuition that more edge devices increase the difficulty of designing a common receive beamformer to equalize all channels. Besides, it is observed that the robust design in each error model achieves a lower MSE than the non-robust scheme and the expectation-based algorithm outperforms its worst-case counterpart, as the same reasons discussed in Fig. 2 and Fig. 3. Besides, the performance gain becomes more significant with the growth in the number of edge devices. This is because the challenge of accurately aggregating model parameters grows with the number of edge devices, which in turn heightens the importance of effective beamforming design.

In Fig. 6, an image classification task is implemented to compare the proposed robust algorithms with other benchmarks on the MNIST dataset when $N = 10$, $M = 2$, and $K = 5$. It is observed that the proposed robust algorithms achieve higher test accuracy than the non-robust design. After 200 training rounds, the expectation-based and worst-case algorithms yield 4% and 3% performance gain compared with the non-robust scheme, respectively. This is attributed to our consideration of inherent channel errors during framework design. By accounting for the channel uncertainties, the gradient distortion at the edge server is mitigated by the joint robust design of AirComp transceivers and phase-shift on RIS. Therefore, more accurate gradient aggregation is achieved at the edge server, which improves the robustness of the proposed FL network against imperfect CSI and enhances the FL performance. Besides, the proposed expectation-based algorithm achieves higher test accuracy compared with the worst-case one, and attains a performance close to the perfect

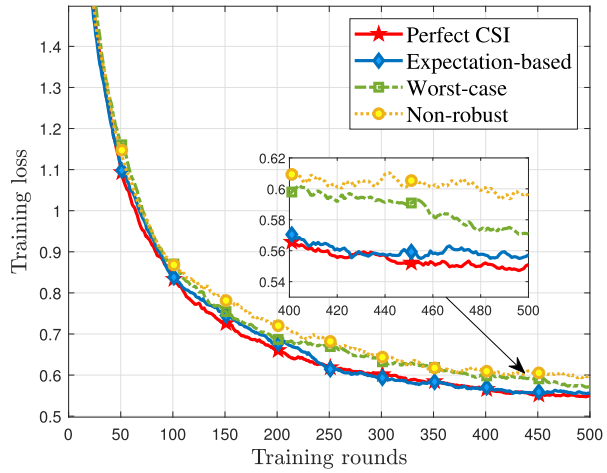


Fig. 7. Training loss versus the number of training rounds with different benchmarks on MNIST dataset.

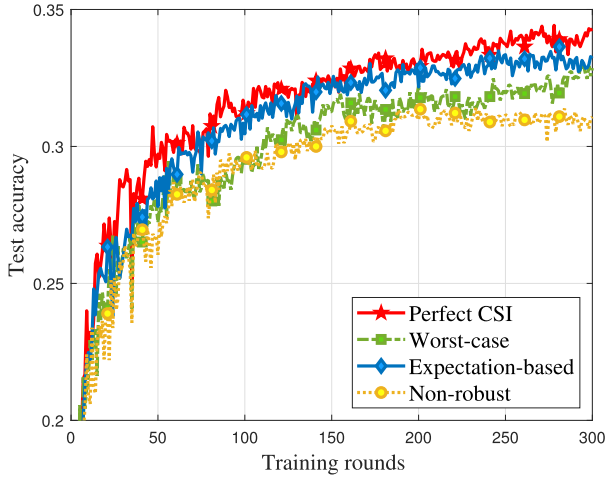


Fig. 8. Test accuracy versus the number of training rounds with different benchmarks on CIFAR-10 dataset.

CSI scheme. This results from utilizing a more accurate channel error model in the expectation-based algorithm, achieving gradient aggregation accuracy approaching the perfect CSI

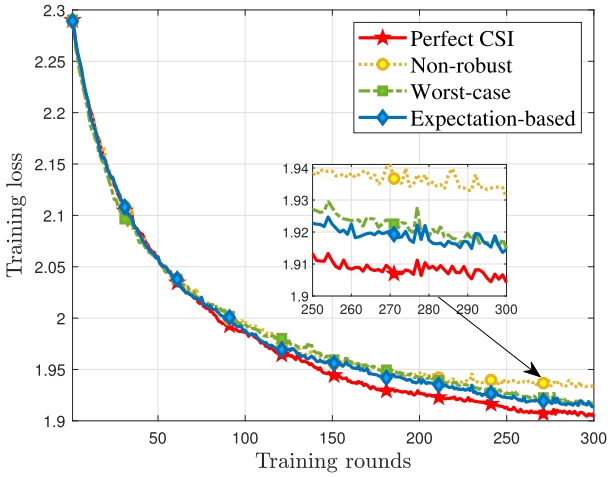


Fig. 9. Training loss versus the number of training rounds with different benchmarks on CIFAR-10 dataset.

scenario which only suffers from channel fading and noise, thereby demonstrating its superiority. Moreover, Fig. 7 further presents training loss of the proposed robust algorithms with other benchmarks on the MNIST dataset and shows similar performance due to the same reason discussed above.

In Figs. 8 and 9, an image classification task is implemented to compare the proposed robust algorithms with other benchmarks on the CIFAR-10 dataset when $N = 10$, $M = 2$, and $K = 5$. By virtue of the available statistics of channel error, the expectation-based error model attains a near-perfect CIS learning performance while outperforming the worst-case one. Compared with the non-robust design, the worst-case error model achieves a better learning performance, demonstrating its robustness.

VII. Conclusion

In this paper, we investigated a joint transceiver design problem in an RIS-assisted over-the-air FL network with imperfect cascaded CSI. After analyzing the convergence of the proposed framework, we formulated a joint optimization problem with an objective to minimize the MSE of the concurrently received model parameters at the edge server. We took into account both the expectation-based and worst-case error models. Both reformulated problems were solved by developing an alternating optimization framework, where a DC algorithm is introduced for dealing with the rank-one constraints in the expectation-based error model, and a PCCP algorithm is proposed to deal with the non-convex unit-modulus constraints in the worst-case error model. Simulation results demonstrated that RIS contributes a significant performance gain for FL model parameters aggregation via AirComp. Besides, we compared the achievable test accuracy under the worst-case and expectation-based error models with a fair setting of parameters. In comparison with the non-robust baseline schemes, the proposed robust algorithms can efficiently overcome the CSI errors. For future work, we plan to extend the proposed framework to multi-cell settings, where server coordination and interference management should be considered.

Appendix

A. Proof of Proposition 1

Based on the local gradients in (3), the AirComp framework aggregates global models through

$$\begin{aligned}
&\boldsymbol{w}^{[t+1]}\\
&=\boldsymbol{w}^{[t]}-r^{[t]}\hat{\boldsymbol{g}}^{[t]}\\
&=\boldsymbol{w}^{[t]}-r^{[t]}\left[\left(\hat{\boldsymbol{g}}^{[t]}-\frac{\sum_{k\in\mathcal{K}}D_k\boldsymbol{g}_k^{[t]}}{\sum_{k\in\mathcal{K}}D_k}\right)+\frac{\sum_{k\in\mathcal{K}}D_k\boldsymbol{g}_k^{[t]}}{\sum_{k\in\mathcal{K}}D_k}\right]\\
&=\boldsymbol{w}^{[t]}-r^{[t]}\left(\frac{\sum_{k\in\mathcal{K}}D_k\boldsymbol{g}_k^{[t]}}{\sum_{k\in\mathcal{K}}D_k}+\boldsymbol{e}^{[t]}\right)\\
&=\boldsymbol{w}^{[t]}-r^{[t]}\Big(\nabla F(\boldsymbol{w}^{[t]})+\boldsymbol{e}^{[t]}\Big),
\end{aligned}\tag{53}$$

where $\boldsymbol{e}^{[t]}$ stands for the gradient distortion error in t-th training round due to unfavorable wireless transmission and it can be computed as

$$\begin{aligned}
&\boldsymbol{e}^{[t]}\\
&=\hat{\boldsymbol{g}}^{[t]}-\frac{\sum_{k\in\mathcal{K}}D_k\boldsymbol{g}_k^{[t]}}{\sum_{k\in\mathcal{K}}D_k}\\
&=\frac{1}{K}\sum_{k\in\mathcal{K}}\boldsymbol{g}_k^{[t]}+\frac{1}{K}\Bigg[\sum_{k\in\mathcal{K}}(\boldsymbol{\omega}^{[t]})^{\mathsf{H}}\boldsymbol{E}_k^{[t]}\boldsymbol{\theta}^{[t]}\alpha_k^{[t]}\boldsymbol{s}_k^{[t]}\\
&\quad+(\boldsymbol{\omega}^{[t]})^{\mathsf{H}}\boldsymbol{n}^{[t]}\Bigg]-\frac{1}{K}\sum_{k\in\mathcal{K}}\boldsymbol{g}_k^{[t]}+\frac{1}{K}\sum_{k\in\mathcal{K}}\bar{g}_k^{[t]}\mathbf{1}-\frac{1}{K}\sum_{k\in\mathcal{K}}\boldsymbol{g}_k^{[t]}\\
&=\frac{1}{K}\sum_{k\in\mathcal{K}}\Big[(\boldsymbol{\omega}^{[t]})^{\mathsf{H}}\boldsymbol{E}_k^{[t]}\alpha_k^{[t]}-v_k^{[t]}\Big]\boldsymbol{s}_k^{[t]}+\frac{1}{K}(\boldsymbol{\omega}^{[t]})^{\mathsf{H}}\boldsymbol{n}^{[t]}.
\end{aligned}\tag{54}$$

Combining the smoothness assumption of the global loss function and (53), we can derive

$$\begin{aligned}
&L\left(\boldsymbol{z}^{[t+1]}\right)-L\left(\boldsymbol{z}^{[t]}\right)\\
&\le\left(\frac{Sr^2}{2}-r\right)\left\|\nabla L\left(\boldsymbol{z}^{[t]}\right)\right\|^2\\
&\quad+(Sr^2-r)\left\langle\nabla L\left(\boldsymbol{z}^{[t]}\right),\boldsymbol{e}^{[t]}\right\rangle+\frac{Sr^2}{2}\left\|\boldsymbol{e}^{[t]}\right\|^2\\
&\le-\frac{r}{2}\left\|\nabla L\left(\boldsymbol{z}^{[t]}\right)\right\|^2+\frac{r}{2}\left\|\boldsymbol{e}^{[t]}\right\|^2,
\end{aligned}\tag{55}$$

where the last inequality holds because $-\left\langle\nabla L\left(\boldsymbol{z}^{[t]}\right),\boldsymbol{e}^{[t]}\right\rangle\le(\left\|\nabla L\left(\boldsymbol{z}^{[t]}\right)\right\|^2+\left\|\boldsymbol{e}^{[t]}\right\|^2)/2$. Thus, we have

$$\begin{aligned}
&\mathbb{E}\left[L\left(\boldsymbol{z}^{[t+1]}\right)\right]-\mathbb{E}\left[L\left(\boldsymbol{z}^{[t]}\right)\right]\\
&\le-\frac{r}{2}\mathbb{E}\left[\left\|\nabla L\left(\boldsymbol{z}^{[t]}\right)\right\|^2\right]+\frac{r}{2K^2}\mathbb{E}\left[\left\|\hat{\boldsymbol{f}}^{[t]}-\boldsymbol{f}^{[t]}\right\|^2\right]\\
&=-\frac{r}{2}\mathbb{E}\left[\left\|\nabla L\left(\boldsymbol{z}^{[t]}\right)\right\|^2\right]+\frac{r}{2K^2}\sum_{j=1}^{d}\mathsf{MSE}^{[t]}(\hat{f}_j,f_j).
\end{aligned}\tag{56}$$

By utilizing Assumption 2 (μ-PL condition), we can further derive

$$\begin{aligned}
&\mathbb{E}\left[L\left(\boldsymbol{z}^{[t+1]}\right)\right]-\mathbb{E}\left[L\left(\boldsymbol{z}^{[t]}\right)\right]\\
&\le-r\mu\Big(\mathbb{E}\left[L\left(\boldsymbol{z}^{[t]}\right)\right]\\
&\quad-\mathbb{E}\left[L\left(\boldsymbol{z}^{\star}\right)\right]\Big)+\frac{rd}{2K^2}\mathsf{MSE}^{[t]}(\hat{f}_j,f_j).
\end{aligned}\tag{57}$$

Subtracting $\mathbb{E}\left[L\left(\boldsymbol{z}^{\star}\right)\right]$ from both sides of the inequality above, we have

$$\begin{aligned}&\mathbb{E}\left[L\left(\boldsymbol{z}^{[t+1]}\right)\right]-\mathbb{E}\left[L\left(\boldsymbol{z}^{\star}\right)\right]\\&\leq(1-r\mu)\Big(\mathbb{E}\left[L\left(\boldsymbol{z}^{[t]}\right)\right]\\&\quad-\mathbb{E}\left[L\left(\boldsymbol{z}^{\star}\right)\right]\Big)+\frac{rd}{2K^2}\mathsf{MSE}^{[t]}(\hat{f}_j,f_j).\end{aligned}\tag{58}$$

Denoting $\Delta=(1-r\mu)$ and applying (58) recursively for $t\in\{0,1,\ldots,T-1\}$, we have

$$\begin{aligned}&\mathbb{E}\left[L(\boldsymbol{z}^{[T]})\right]-\mathbb{E}\left[L(\boldsymbol{z}^{\star})\right]\\&\leq\Delta^{T}\Big(\mathbb{E}\left[L(\boldsymbol{z}^{[0]})\right]-\mathbb{E}\left[L(\boldsymbol{z}^{\star})\right]\Big)\\&\quad+\sum_{t=0}^{T-1}\frac{rd}{2K^2}\Delta^{T-1-t}\mathsf{MSE}^{[t]}(\hat{f}_j,f_j).\end{aligned}\tag{59}$$

REFERENCES

[1] Y. Shi, Y. Zhou, D. Wen, Y. Wu, C. Jiang, and K. B. Letaief, "Task-oriented communications for 6G: Vision, principles, and technologies," *IEEE Wireless Commun.*, vol. 30, no. 3, pp. 78–85, Jun. 2023.

[2] 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.

[3] M. Tao et al., "Federated edge learning for 6G: Foundations, methodologies, and applications," *Proc. IEEE*, vol. 113, no. 9, pp. 1075–1113, Sep. 2025.

[4] Y. Zhou, Y. Shi, H. Zhou, J. Wang, L. Fu, and Y. Yang, "Toward scalable wireless federated learning: Challenges and solutions," *IEEE Internet Things Mag.*, vol. 6, no. 4, pp. 10–16, Dec. 2023.

[5] X. Cao, G. Zhu, J. Xu, Z. Wang, and S. Cui, "Optimized power control design for over-the-air federated edge learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 342–358, Jan. 2022.

[6] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5115–5128, Aug. 2021.

[7] S. Wang, Y. Hong, R. Wang, Q. Hao, Y.-C. Wu, and D. W. K. Ng, "Edge federated learning via unit-modulus over-the-air computation," *IEEE Trans. Commun.*, vol. 70, no. 5, pp. 3141–3156, May 2022.

[8] Z. Wang, Y. Zhou, Y. Shi, and W. Zhuang, "Interference management for over-the-air federated learning in multi-cell wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2361–2377, Aug. 2022.

[9] 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.

[10] J. Du, B. Jiang, C. Jiang, Y. Shi, and Z. Han, "Gradient and channel aware dynamic scheduling for over-the-air computation in federated edge learning systems," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 1035–1050, Apr. 2023.

[11] Y. Sun, S. Zhou, Z. Niu, and D. Gündüz, "Dynamic scheduling for over-the-air federated edge learning with energy constraints," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 227–242, Jan. 2022.

[12] G. Gao, Q. An, Z. Wang, Z. Wang, Y. Shi, and Y. Zhou, "Over-the-air computation assisted federated learning with progressive training," in *Proc. ICC-IEEE Int. Conf. Commun.*, Jun. 2024, pp. 5664–5669.

[13] Q. An, Y. Zhou, Z. Wang, H. Shan, Y. Shi, and M. Bennis, "Online optimization for over-the-air federated learning with energy harvesting," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7291–7306, Jul. 2024.

[14] Z. Wang, Y. Zhou, and Y. Shi, "Over-the-air federated graph learning," *IEEE Trans. Wireless Commun.*, vol. 23, no. 12, pp. 18669–18683, Dec. 2024.

[15] 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.

[16] Y. Shao, D. Gündüz, and S. C. Liew, "Federated edge learning with misaligned over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 3951–3964, Jun. 2022.

[17] Y.-C. Liang, R. Long, Q. Zhang, J. Chen, H. V. Cheng, and H. Guo, "Large intelligent surface/antennas (LISA): Making reflective radios smart," *J. Commun. Inf. Netw.*, vol. 4, no. 2, pp. 40–50, Jun. 2019.

[18] S. Gong et al., "Toward smart wireless communications via intelligent reflecting surfaces: A contemporary survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2283–2314, 4th Quart., 2020.

[19] 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.

[20] Q. Li, M. El-Hajjar, C. Xu, J. An, C. Yuen, and L. Hanzo, "Stacked intelligent metasurfaces for holographic MIMO-aided cell-free networks," *IEEE Trans. Commun.*, vol. 72, no. 11, pp. 7139–7151, Nov. 2024.

[21] Q. Li, M. El-Hajjar, Y. Sun, I. Hemadeh, A. Shojaeifard, and L. Hanzo, "Energy-efficient reconfigurable holographic surfaces operating in the presence of realistic hardware impairments," *IEEE Trans. Commun.*, vol. 72, no. 8, pp. 5226–5238, Aug. 2024.

[22] Q. Li, M. El-Hajjar, K. Cao, C. Xu, H. Haas, and L. Hanzo, "Holographic metasurface-based beamforming for multi-altitude LEO satellite networks," *IEEE Trans. Wireless Commun.*, vol. 24, no. 4, pp. 3103–3116, Apr. 2025.

[23] Q. Li, M. El-Hajjar, Y. Sun, and L. Hanzo, "Performance analysis of reconfigurable holographic surfaces in the near-field scenario of cell-free networks under hardware impairments," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 11972–11984, Sep. 2024.

[24] 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.

[25] Z. Wang, Y. Zhou, Y. Zou, Q. An, Y. Shi, and M. Bennis, "A graph neural network learning approach to optimize RIS-assisted federated learning," *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6092–6106, Sep. 2023.

[26] L. Hu, Z. Wang, H. Zhu, and Y. Zhou, "RIS-assisted over-the-air federated learning in millimeter wave MIMO networks," *J. Commun. Inf. Netw.*, vol. 7, no. 2, pp. 145–156, Jun. 2022.

[27] H. Liu, X. Yuan, and Y.-J.-A. Zhang, "CSIT-free model aggregation for federated edge learning via reconfigurable intelligent surface," *IEEE Wireless Commun. Lett.*, vol. 10, no. 11, pp. 2440–2444, Nov. 2021.

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

[29] Z. Wang et al., "Federated learning via intelligent reflecting surface," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 808–822, Feb. 2022.

[30] Y. Yang, Y. Zhou, Y. Wu, and Y. Shi, "Differentially private federated learning via reconfigurable intelligent surface," *IEEE Internet Things J.*, vol. 9, no. 20, pp. 19728–19743, Oct. 2022.

[31] J. Zheng, H. Tian, W. Ni, W. Ni, and P. Zhang, "Balancing accuracy and integrity for reconfigurable intelligent surface-aided over-the-air federated learning," *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 10964–10980, Dec. 2022.

[32] Z. Zhou, N. Ge, Z. Wang, and L. Hanzo, "Joint transmit precoding and reconfigurable intelligent surface phase adjustment: A decomposition-aided channel estimation approach," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1228–1243, Feb. 2021.

[33] 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.

[34] D. Mishra and H. Johansson, "Channel estimation and low-complexity beamforming design for passive intelligent surface assisted MISO wireless energy transfer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 4659–4663.

[35] B. Zheng and R. Zhang, "Intelligent reflecting surface-enhanced OFDM: Channel estimation and reflection optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 518–522, Apr. 2020.

[36] B. Zheng, C. You, and R. Zhang, "Fast channel estimation for IRS-assisted OFDM," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 580–584, Mar. 2021.

[37] T. L. Jensen and E. De Carvalho, "An optimal channel estimation scheme for intelligent reflecting surfaces based on a minimum variance unbiased estimator," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 5000–5004.

[38] B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface assisted multi-user OFDMA: Channel estimation and training design," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8315–8329, Dec. 2020.

[39] J.-M. Kang, "Intelligent reflecting surface: Joint optimal training sequence and refection pattern," *IEEE Commun. Lett.*, vol. 24, no. 8, pp. 1784–1788, Aug. 2020.

[40] 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.
[41] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Deep learning for large intelligent surfaces in millimeter wave and massive MIMO systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.
[42] G. Zhou, C. Pan, H. Ren, K. Wang, M. D. Renzo, and A. Nallanathan, "Robust beamforming design for intelligent reflecting surface aided MISO communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1658–1662, Oct. 2020.
[43] G. Zhou, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "A framework of robust transmission design for IRS-aided MISO communications with imperfect cascaded channels," *IEEE Trans. Signal Process.*, vol. 68, pp. 5092–5106, 2020.
[44] X. Zhang, H. Tian, W. Ni, and Z. Yang, "Joint beamforming design for multifunctional RIS-aided over-the-air federated learning," *IEEE Internet Things J.*, vol. 12, no. 12, pp. 21720–21739, Jun. 2025.
[45] 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.
[46] S. Hong, C. Pan, H. Ren, K. Wang, K. K. Chai, and A. Nallanathan, "Robust transmission design for intelligent reflecting surface-aided secure communication systems with imperfect cascaded CSI," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2487–2501, Apr. 2021.
[47] 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.
[48] A. Mahmood, M. I. Ashraf, M. Gidlund, J. Torsner, and J. Sachs, "Time synchronization in 5G wireless edge: Requirements and solutions for critical-MTC," *IEEE Commun. Mag.*, vol. 57, no. 12, pp. 45–51, Dec. 2019.
[49] Y. Shao, D. Gündüz, and S. C. Liew, "Federated edge learning with misaligned over-the-air computation," in *Proc. IEEE 22nd Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, May 2021, pp. 236–240.
[50] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2016, pp. 795–811.
[51] 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.
[52] Q. Zhang, C. He, and L. Jiang, "Per-stream MSE based linear transceiver design for MIMO interference channels with CSI error," *IEEE Trans. Commun.*, vol. 63, no. 5, pp. 1676–1689, May 2015.
[53] A. Gershman, N. Sidiropoulos, S. Shahbazpanahi, M. Bengtsson, and B. Ottersten, "Convex optimization-based beamforming," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 62–75, May 2010.
[54] P. Ubaidulla and A. Chockalingam, "Relay precoder optimization in MIMO-relay networks with imperfect CSI," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5473–5484, Nov. 2011.
[55] S. A. Vorobyov, A. B. Gershman, and Z.-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 313–324, Feb. 2003.
[56] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[57] Z.-Q. Luo, W.-K. Ma, A. M. 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.
[58] J.-Y. Gotoh, A. Takeda, and K. Tono, "DC formulations and algorithms for sparse optimization problems," *Math. Program.*, vol. 169, no. 1, pp. 141–176, May 2018.
[59] Q. An, Y. Shi, and Y. Zhou, "Reconfigurable intelligent surface assisted non-orthogonal unicast and broadcast transmission," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, May 2020, pp. 1–5.
[60] M. Grant and S. Boyd. (Mar. 2014). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1*. [Online]. Available: http://cvxr.com/cvx
[61] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. Philadelphia, PA, USA: SIAM, 2001.
[62] Y. LeCun. (1998). *The Mnist Database of Handwritten Digits*. [Online]. Available: http://yann.lecun.com/exdb/mnist/
[63] A. Krizhevsky, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf
[64] A. Pascual-Iserte, D. P. Palomar, A. I. Perez-Neira, and M. A. Lagunas, "A robust maximin approach for MIMO communications with imperfect channel state information based on convex optimization," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 346–360, Jan. 2006.
[65] F. Ang, L. Chen, N. Zhao, Y. Chen, and F. R. Yu, "Robust design for massive CSI acquisition in analog function computation networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2361–2373, Mar. 2019.
[66] W. Zhang, J. Xu, W. Xu, X. You, and W. Fu, "Worst-case design for RIS-aided over-the-air computation with imperfect CSI," *IEEE Commun. Lett.*, vol. 26, no. 9, pp. 2136–2140, Sep. 2022.

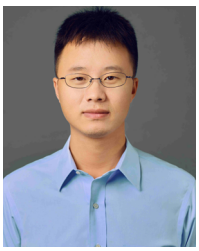
Qiaochu An received the B.S. degree from the School of Information and Communications Engineering, Zhengzhou University, Zhengzhou, China, in 2019, and the Ph.D. degree from the School of Information Science and Technology, ShanghaiTech University, Shanghai, China, in 2025. His research interests include the Internet of Things, intelligent reflecting surface, and federated learning.

Hongbin Zhu (Member, IEEE) received the B.S. degree in electronic engineering from the Ocean University of China, Qingdao, China, in 2014, and the joint Ph.D. degree in electronic engineering with co-supervision from ShanghaiTech University, Shanghai, China, and Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, in 2019. He is currently a Pre-Tenure Associate Professor with the Institute of Fintech, Fudan University. His research interests include FinTech, graph learning, and generative AI.

Qiang (John) Ye (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2016. Before joining UCalgary, he was as a Faculty Member with the Memorial University of Newfoundland, NL, Canada, from 2021 to 2023, and with the Minnesota State University, Mankato, USA, from 2019 to 2021. He was with the Department of Electrical and Computer Engineering, University of Waterloo, as a Post-Doctoral Fellow and then a Research Associate from 2016 to 2019. Since 2023, he has been a Faculty Member with the Department of Electrical and Software Engineering, Schulich School of Engineering, University of Calgary, Calgary, AB, Canada. He has published more than 90 research articles on top-ranked journals and conference proceedings. He received the Best Paper Award in the IEEE/CIC ICCC in 2024, the IEEE OJ-COMS Exemplary Editor Award in 2025, the IEEE TCCN Exemplary Editor Award in 2023, the Early Career Research Excellence Award, Schulich School of Engineering, and the University of Calgary, in 2024. He has been named among the World's Top 2% Scientists from 2023 to 2025 (by Stanford/Elsevier). He is/was the General, Publication, and Program Co-Chairs for different reputable international conferences and workshops (e.g., IEEE INFOCOM, GLOBECOM, VTC, ICCC, ICCT, WISEE, and SWC). He also serves/served for the IEEE Vehicular Technology Society (VTS) Region 7 Chapter Coordinator in 2024, the IEEE Communications Society (ComSoc) Southern Alberta Chapter Vice Chair in 2024, and the VTS Regions 1-7 Chapters Coordinator from 2022 to 2023. He is the Leading SIG Co-Chair of the IEEE ComSoc-IoT-AHSN Technical Committee. He serves as an Associate Editor for prestigious IEEE journals, such as IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, and IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY. He has been selected as an IEEE Communication Society Distinguished Lecturer for the class from 2025 to 2026.

Ning Zhang (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Canada, in January 2015.

After that, he was a Post-Doctoral Research Fellow with the University of Waterloo and the University of Toronto. He is currently a Professor and Canada Research Chair with the Department of Electrical and Computer Engineering, University of Windsor. His research interests include connected vehicles, mobile edge computing, wireless networking, and security. He is a member of the Royal Society of Canada College, a Distinguished Lecturer of IEEE, and a Highly Cited Researcher (Web of Science). He serves as the Chair for the IEEE Technical Committee on Cognitive Networks and the Vice Chair for the IEEE Technical Committee on Big Data. He is/was the Editor-in-Chief (EIC) of the IEEE ComSoc Technical Committees Newsletter (TCN) and the Founding EIC of Journal of Pervasive Intelligence and Internet of Things. He also serves/served as an Associate Editor for IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, IEEE INTERNET OF THINGS JOURNAL, and IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING; and the TPC/General Chair for numerous international conferences.



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. He visited the University of California, Berkeley, CA, USA, from October 2016 to February 2017. Since September 2015, he has been with the School of Information Science and Technology, ShanghaiTech University, where he is a Full Professor. His research interests include edge artificial intelligence and large-scale optimization. He is an IET Fellow. He was a recipient of the IEEE Marconi Prize Paper Award in Wireless Communications in 2016, the Young Author Best Paper Award by the IEEE Signal Processing Society in 2016, the IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2021, Chinese Institute of Electronics First Prize in Natural Science in 2022, and China Institute of Communications First Prize in Natural Science in 2024. He is an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, *Journal of Communications and Information Networks*, and *Space Habitation*.

Yong Zhou (Senior 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 was a Post-Doctoral Researcher Fellow with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada. Since March 2018, he has been with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China, where he is currently a Tenured Associate Professor. His research interests include 6G communications, edge intelligence, and the Internet of Things. He was the Track Co-Chair of IEEE VTC 2020 Fall and IEEE VTC 2023 Spring, and the Co-Chair of IEEE ICC 2022 workshop on edge artificial intelligence for 6G and IEEE Globecom 2024 workshop on space computing power networks. He serves as an Associate Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY.