One Bit Aggregation for Federated Edge Learning With Reconfigurable Intelligent Surface: Analysis and Optimization

Heju Li, *Student Member, IEEE*, Rui Wang, *Senior Member, IEEE*,
Wei Zhang, *Fellow, IEEE*, and Jun Wu, *Senior Member, IEEE*

***Abstract*— As one of the most popular and attractive frameworks for model training, federated edge learning (FEEL) presents a new paradigm, which avoids direct data transmission by collaboratively training a global learning model across multiple distributed edge devices, thus overcoming the disadvantage of centralized machine learning in resource limitations, delay constraints, and privacy issues. However, due to the heavy cost of communicating gradient among edge devices, sharing the parameters of a large-scale neural network can still be time-intensive. To alleviate this bottleneck, an efficient scheme, called SignSGD has been recently proposed, where the one-bit gradient quantization with majority vote is featured at edge devices. Nevertheless, the performance of one-bit aggregation will inevitably deteriorate due to the undesirable propagation error introduced by wireless channels. To address this issue, we propose in this work a novel reconfigurable intelligent surface (RIS)-aided one-bit communication optimization scheme under orthogonal frequency division multiple access (OFDMA) to relieve the negative influence of communication error on the SignSGD-based FEEL. Specifically, a learning convergence analysis is firstly presented to quantitatively characterize the impact of wireless communication error measured by the union bound on pairwise bit error rate (BER) on the performance of SignSGD-based FEEL. Immediately, a unified communication-learning optimization problem is further formulated to jointly optimize the sub-band assignment strategy, the power allocation vector, and the RIS configuration matrix. Numerical experiments show that the proposed design achieves substantial performance improvement compared with the state-of-the-art approaches.**

***Index Terms*— Federated edge learning, SignSGD, reconfigurable intelligent surface, bit error rate, orthogonal frequency division multiple access.**

I. Introduction

THE evolvement of next-generation of wireless networks will enable numerous machine learning (ML) applications and advanced tools to efficiently analyze sundry types of data collected by edge devices for inference, autonomy, and decision [2]. However, due to the challenges in resource limitations, delay constraints, and privacy issues, it is impractical for edge devices to upload their entire collected datasets to a cloud server for centrally model training or inference purposes. To address these challenges, federated edge learning (FEEL), as one of the most attractive paradigms of edge ML, has been developed, where geo-distributed devices are able to collaboratively train a global model while keeping the raw data processed locally [3]. Compared with centralized learning paradigms, FEEL can effectively preserve user privacy and data security [4] by avoiding the transmission of privacy-sensitive data over wireless channels. Moreover, the ML frontier is pushed from the cloud center to the network edge, and thus edge devices only need to communicate with the base station (BS) on the up-to-date model parameters [5]. By doing so, the communication cost can be significantly reduced in a distributed manner, thus overcoming the drawback of excessive propagation delay caused by the potential network congestion. Despite the above advantages of FEEL, communication overhead for transmitting millions of locally trained parameters over wireless channels from each edge device to the BS during the iterative model update process is still a significant bottleneck. To alleviate this bottleneck, an efficient scheme called SignSGD [6] has been recently proposed, where every edge device sends the sign of local gradient up to the BS, which aggregates the quantified signs and sends back only the majority decision. Since all communication to and from the BS is compressed to one bit, the need for communication efficiency is further accommodated. Nevertheless, as the edge devices are usually connected to the BS over the wireless channel, the model parameters received by the BS are inevitably distorted by channel fading and additive noise. As the analysis in [7], the performance of one-bit aggregation will inevitably deteriorate due to the undesirable propagation error introduced by wireless channels. Thence, designing uplink

Manuscript received 30 December 2021; revised 16 May 2022 and 7 August 2022; accepted 9 August 2022. Date of publication 22 August 2022; date of current version 13 February 2023. This work was supported in part by the Shanghai Science and Technology Innovation Action Plan With Project under Grant 21220713100, in part by the Natural Science Foundation of Shanghai under Grant 22ZR1465100, in part by the National Key Research and Development Program of China under Grant 2020YFA0711400, in part by the National Science Foundation China under Grant 61831018, and in part by the CCF-Baidu Open Fund under Grant OF2021021. An earlier version of this paper was presented in part at IEEE GLOBECOM 2022 [1]. The associate editor coordinating the review of this article and approving it for publication was J. Xu. *(Corresponding author: Rui Wang.)*

Heju Li is with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: 2010456@tongji.edu.cn).

Rui Wang is with the College of Electronics and Information Engineering and the Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 201804, China (e-mail: ruiwang@tongji.edu.cn).

Wei Zhang is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: w.zhang@unsw.edu.au).

Jun Wu is with the School of Computer Science, Fudan University, Shanghai 200433, China (e-mail: wujun@fudan.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2022.3198881.

Digital Object Identifier 10.1109/TWC.2022.3198881

communication to achieve a more reliable transmission during model update is still critical for FEEL training.

A. State-of-the-Art

As discussed earlier, overcoming the uplink communication bottleneck over wireless channels is already acknowledged as a significant challenge confronting the implementation of FEEL. To handle this, several strategies have been proposed by taking into account wireless channel hostilities and the scarcity of radio resources. The first scheme that studied FEEL by optimizing the physical layer resource constraints focused on a broadband over-the-air computation (AirComp) aggregation system [8], called broadband analog aggregation (BAA), where the transmitting gradient/model of each edge device are averaged over frequency sub-channels, thus leading to substantial latency reduction and narrower bandwidth requirement compared with the orthogonal multiple access schemes. The extended version of [8] was later proposed in [7], namely one-bit over-the-air computation, which adopts in this work a truncated-channel-inversion power control scheme to implement a novel digital version of broadband AirComp aggregation. A compressed analog communication scheme (CA-DSGD) was proposed in [9] by introducing error accumulation and gradient sparsification in addition to AirComp. Although AirComp aggregation can effectively mitigate the communication bottleneck in wireless networks, the stringent requirements in synchronization and accurate power alignment are necessary, or using massive antennas [10], alternatively. [11] presented a hierarchical FEEL framework and provided a rigorous end-to-end latency analysis for the communication latency in heterogeneous cellular networks, which is used to design a resource allocation strategy by minimizing the end-to-end latency over orthogonal frequency division multiplexing (OFDM) wireless channel. Reference [12] derived a closed-form expression of the expected convergence rate of FL algorithm with respect to wireless factors, specifically, wireless resource allocation and user selection. Based on this expression, an optimization problem minimizing an FL loss function was formulated. Interestingly, by considering the local updates in FEEL are not equally important for learning convergence, the performance improvement was further accomplished in [13], where a novel probabilistic scheduling framework was proposed to yield unbiased gradient aggregation, thus achieving the optimal trade-off between channel quality and update importance. In [14], an importance-aware joint data selection and resource allocation algorithm was presented, where the critical data is selected according to the gradient norm to speed up the learning process, and a learning efficiency maximization problem was further formulated by jointly considering the communication resource allocation and data selection.

Recently, reconfigurable intelligent surface (RIS) has been regarded as a vital enabler of the next-generation wireless networks [15], [16] by installing massive low-cost passive reflecting elements on the programmable surfaces and smartly reconfiguring the propagation environment of wireless signals. For instance, in inter-cell interference coordination (ICIC), cell edge devices often suffer from poor communication quality or unfavorable wireless propagation conditions, which dominate the overall model aggregation error and training delay in the FEEL system since the server must wait for its gradients to aggregate. In this case, the deployment of the RIS is able to proactively manipulate the wireless channels between the BS and edge devices by judiciously inducing independent phase shift of each reflecting element in real time [17], thereby hugely improving the channel conditions of cell edge users, reducing EFFL aggregation errors and training delay. [18] considered the critical energy efficiency issue in the RIS-aided wireless communication network where an energy consumption minimization problem in an federated learning system was formulated subject to the completion training time constraint. A fast yet reliable model aggregation for the FEEL aided by the RIS was proposed in [19], where an optimization problem that jointly optimizes the device selection, the RIS phase shifts, and the BS beamformer is further formulated, thus maximizing the number of participating devices in each communication round of FEEL under certain mean-squared-error (MSE) requirements. Reference [20] established an AirComp FEEL framework by developing a convergence analysis framework with respect to device selection and model aggregation error, thus formulating a unified communication-learning optimization problem. To integrate over-the-air and non-orthogonal multiple access into an universal FL framework, [21] proposed to maximize the achievable hybrid rate by jointly optimizing the transmit power, the receive scalar, and the RIS reflection coefficients. The aforementioned works demonstrated the potential benefits of deploying a RIS to promote wireless channel quality, thereby achieving better performance of FEEL in fewer communication rounds.

B. Motivations and Contributions

Although the effectiveness of the RIS in improving the model aggregation quality has been demonstrated, most of the state-of-the-art work overwhelmingly focuses on the improvement of communication level, and thus cannot fully unleash the gains of deploying the RIS in FEEL systems. In addition, there is currently no mature work where the RIS is deployed to relieve the negative influence of communication error for the SignSGD-based FEEL.

Thence, we explore in this work the benefits of the RIS in enhancing the SignSGD-based FEEL under the orthogonal frequency division multiple access (OFDMA) system by developing a unified communication-learning convergence analysis framework, where the impact of communication error on the FEEL training loss is ingeniously tracked. Moreover, an optimization problem is further formulated to jointly optimize the sub-band assignment strategy, the power allocation vector, and the RIS configuration matrix. The key contributions of this paper can be summarized as

- To the best of our knowledge, we are the first to integrate the RIS technology into SignSGD-based FEEL under the OFDMA system to relieve the negative influence of communication error, which is measured by the union bound on pairwise bit error rate (BER) during transmission.
- We innovatively derive a closed-form expression of the convergence rate under the RIS-aided SignSGD-based FEEL system from the perspective of communication

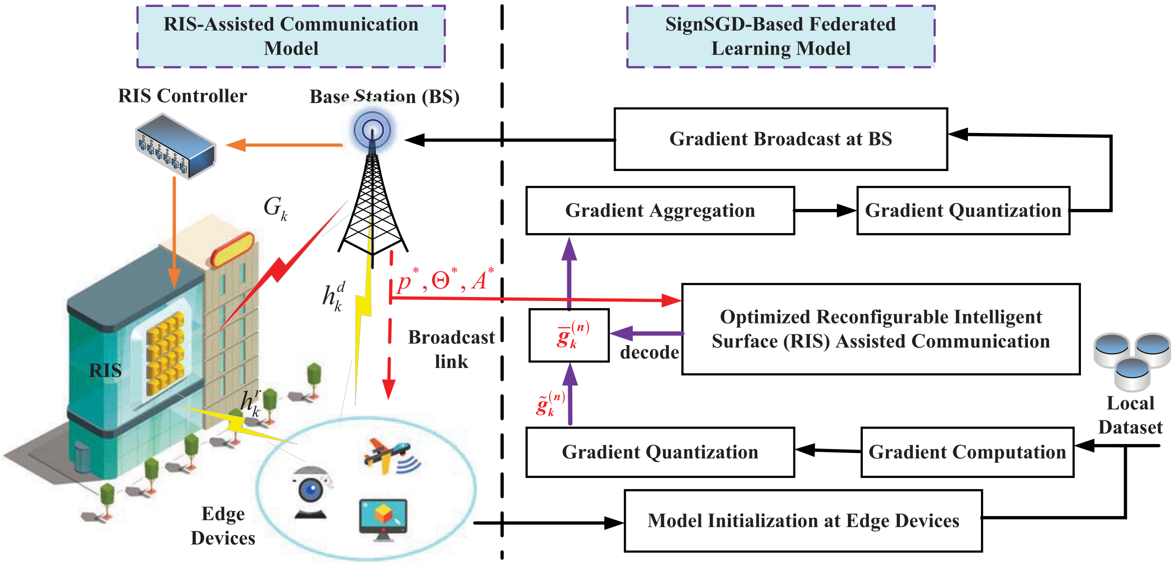


Fig. 1. The RIS-assisted OFDMA-based FL system.

BER. Based on the convergence analysis, we show that the communication error, i.e., BER, slows down the convergence rate of the considered system. However, this reduction can be significantly relieved by jointly designing the sub-band assignment strategy, the power allocation vector, and the RIS configuration matrix. By comparison with the convergence rate over the error-free channel, a unified communication-learning optimization problem with respect to the sub-band assignment strategy, the power allocation vector, and the RIS configuration matrix is further formulated.

- To tackle the non-convex communication-learning optimization problem, we propose an effective algorithm to decouple the tricky problem into several tractable subproblems by a alternating optimization method until the algorithm converges, where the sub-band assignment strategy is solved by a penalty-based successive convex approximation (SCA), and the power allocation vector together with the RIS configuration matrix are optimized by quasi-convex optimization with relaxed ℓ_0 norm approximation and the difference-of-convex (DC) programming, respectively.
- We implement the RIS-assisted Sign-SGD based FEEL prototype and evaluate its performance with several benchmarks, open real-world datasets for data categorization. Simulation results confirm that a substantial performance improvement is realized using our proposed approach compared with the existing solutions.

The rest of this paper is organized as follows. We present the underlying learning and communication models in Section II, followed by the analysis of the convergence performance, and accordingly formulate the learning optimization problem that minimizes the training loss in Section III. An efficient solution to jointly optimize the sub-band assignment strategy, the power allocation vector, and the RIS configuration matrix is described in detail in Section IV. Then illustrative numerical results of our proposed scheme are presented in Section V. Finally, conclusions are drawn in Section VI.

Notations: $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$ separately denote the real and complex number sets with the space of $m \times n$. Regular letters, bold small letters, and bold capital letters represent scalars, vectors, and matrices, respectively. $(\bullet)^T$ and $(\bullet)^H$ denote the transpose operator and the conjugate transpose operator. x_i is used to denote the i-th entry of vector $\mathbf{x}$, and $[\mathbf{X}]_{(i,j)}$ denotes the (i, j)-th entry of matrix $\mathbf{X}$. The circularly-symmetric complex normal distribution with mean μ and covariance δ^2 is represented by $\mathcal{CN}(\mu, \sigma^2)$. diag$\{\mathbf{x}\}$ constructs a diagonal matrix with the diagonal entries specified by $\mathbf{x}$, and $\mathbb{E}[\bullet]$ is the expectation operator.

II. Learning and Communication Models

In this section, the learning and communication models are respectively presented to achieve a fast, low-cost and reliable model aggregation over the wireless channel in SignSGD-based FEEL under the OFDMA system, as shown in Fig. 1, where a RIS is deployed to recompense the reduction of signal magnitude and the dislocation of wireless communication.

A. SignSGD-FL-Based Learning Model

We consider a signSGD-FL system consisting of a J-antenna BS coordinating the cooperative modeling across K single-antenna edge devices, where a RIS is deployed to assist the communication. Each device $k \in \mathcal{K} \triangleq \{1, 2, \ldots, K\}$, holds its local dataset D_k, which consists of labeled data samples $\{(x_i, y_i)\} \in D_k$ with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ denoting the input feature and the associated label respectively. The parameter vector $\mathbf{w} \in \mathbb{R}^q$ is collaboratively trained across the edge devices, and legitimately orchestrated through the BS.

Formally, the local learning objective is to minimize an empirical loss function $F_k(\mathbf{w})$ on D_k

$$F_k(\mathbf{w}) = \frac{1}{|D_k|} \sum\nolimits_{(x_i, y_i)\in D_k} f(\mathbf{w}, x_i, y_i), \tag{1}$$

where $f(\mathbf{w}, x_i, y_i)$ denotes a sample-wise loss function quantifying the prediction deviation of model $\mathbf{w}$ on the training sample x_i with respect to its authentic label y_i, and it abbreviated as $f_i(\mathbf{w})$ for convenience. $|D_k|$ is the size of the local dataset in device k. We assume $|D_i| = |D_j| = D$, for any $i \neq j$, which means the dataset size is uniform across edge devices.

To perform FL, each edge device locally updates model $\mathbf{w}_k$ by minimizing $F_k(\mathbf{w})$ using gradient descent, and the BS aggregates the local updates to produce the global model. Then, the training optimization problem of FL algorithm is established with the form

$$\min_{\mathbf{w}} \quad F(\mathbf{w}) = \tfrac{1}{K}\textstyle\sum_{k=1}^{K} F_k(\mathbf{w}). \tag{2}$$

In this paper, we adopt the popular idea of one-bit gradient quantization with majority vote to get available $\mathbf{w}$, inspired by signSGD [6]. Suppose the model $\mathbf{w}$ is computed iteratively with N training iterations. The following procedures are performed during the n-th iteration, $1 \leq n \leq N$:

- **Device selection**: A subset of edge devices $\lambdabar \subset \mathcal{K}$ with size K_c is selected by BS to participate in the training process.
- **Local gradient estimation**: Each selected device $k \in \lambdabar$ computes a local estimate of the gradient approximately with respect to the loss function in Equation (1), denoted by $\mathbf{g}_k^{(n)} \in \mathbb{R}^q$, using the current parameter-vector $\mathbf{w}^{(n)}$ broadcasted from BS, and the chosen subset $\tilde{D}_k$ with batchsize n_b from local dataset D_k. Then we have
$$\mathbf{g}_k^{(n)} = \frac{1}{n_b}\sum\nolimits_{i\in\tilde{D}_k} \nabla f_i(\mathbf{w}^{(n)}), \tag{3}$$
with ∇ achieving the gradient operator. Note that $n_b = |D_k| = D$, means all the samples in local dataset are selected to estimate gradient.
- **Quantization**: Each active device $k \in \lambdabar$ takes the signs of the local gradient parameters element-wise as $\tilde{\mathbf{g}}_k^{(n)} = \text{sign}(\mathbf{g}_k^{(n)})$, where sign($\mathbf{x}$) returns the sign of each coordinate value of the input vector $\mathbf{x}$ if it is non-zero and a sign chosen uniformly at random otherwise.
- **One-bit gradient aggregation**: If the quantized local gradient can be reliably recovered in the BS, i.e., the communication is error-free, the BS sums the sign vectors $\tilde{\mathbf{g}}_k^{(n)}$ as
$$\tilde{\mathbf{g}}^{(n)} = \sum\nolimits_{k\in\lambdabar} \tilde{\mathbf{g}}_k^{(n)}, \tag{4}$$
to generate a global gradient estimate $\mathbf{v}^{(n)}$ by simply taking the element-wise sign of $\tilde{\mathbf{g}}^{(n)}$, i.e., $\mathbf{v}^{(n)} = \text{sign}(\tilde{\mathbf{g}}^{(n)})$. Essentially, the operation computes the global gradient by taking the median of all selected devices' signs at every position of the update vectors, i.e., voting on each coordinate point of the sign vector. Then, the BS broadcasts $\mathbf{v}^{(n)}$ to all the devices to initiate the next training iteration via
$$\mathbf{w}^{(n+1)} = \mathbf{w}^{(n)} - \eta\mathbf{v}^{(n)}, \tag{5}$$
or, terminate the training process once the convergence criterion is satisfied, for instance, the maximum number of communications is reached.

Referring to the analysis of learning process, the quantized gradient information needs to be exchanged between the BS and devices. Intuitively, once the bit error caused by channel distortion occurs, the subsequent one-bit aggregation will have an adverse effect inevitably. This motivates the deployment of the RIS to improve the channel conditions in the communication process.

B. RIS-Assisted Communication Model

As illustrated in Fig. 1, we consider a RIS-assisted FEEL system, where a RIS is embedded in a surrounding building acting as a passive relay. Specifically, to handle the challenge in frequency selective fading and inter-symbol interference, the available bandwidth B is divided into M sub-bands by utilizing OFDM modulation technique denoted by the set $\mathcal{M} \triangleq \{1, 2, \ldots, M\}$ with $M \gg K$. Moreover, each sub-band consists of S orthogonal sub-channels (or sub-carriers) indexed by $\mathcal{S} \triangleq \{1, 2, \ldots, S\}$.[1] The channels within each sub-band are presumed to be definite, e.g., frequency-flat, but supposed to vary among different sub-bands [22]. To avoid inter-device interference, each sub-band is assigned to at most one device. Additionally, we assume that each device transmits quantized gradient using only one sub-band without loss of generality. For simplicity, the wireless channel across all sub-channels is invariant during the whole learning process, which is consistent with the settings in [19] and [20].[2] The study on time-varying channels is exhibited in Section V.

The deployed RIS consists of L passive reflecting elements, referred as $\mathcal{L} \triangleq \{1, 2, \ldots, L\}$, and associates with a controller, which is used for controlling signal reflection by adjusting the reflection coefficients of the RIS elements. A separate control link is deployed between the RIS controller and the BS to acquire the information required for designing the reflection coefficient. Within each sub-band m, there are two links where the signals sent from the edge device k arrive at the BS, respectively called, direct link $\mathbf{h}_{k,m}^{d} \in \mathbb{C}^J$ and the k-th device-RIS-server cascade link $\mathbf{h}_{k,m}^{c} \in \mathbb{C}^J$ [23], with $\mathbf{h}_{k,m}^{c} = \mathbf{G}_{k,m}\boldsymbol{\Theta}\mathbf{h}_{k,m}^{r}$. $\mathbf{G}_{k,m} \in \mathbb{C}^{J\times L}$ denotes the channel between the RIS and the BS on the frequency of m-th sub-band where device k transmits the quantified gradient. $\boldsymbol{\Theta} = \text{diag}\{\boldsymbol{\phi}\} \in \mathbb{C}^{L\times L}$ represents the diagonal matrix with the phase shift vector $\boldsymbol{\phi} = [\phi_1, \phi_2, \ldots, \phi_L]^T$ with $|\phi_l| = 1$. Actually, the reflection operation on the RIS element resembles multiplying the incident signal with ϕ_l, and then forwarding this composite signal as if from a point source, which is the main difference from the active reflection surface [24]. $\mathbf{h}_{k,m}^{r} \in \mathbb{C}^L$ denotes the channel between the edge device k and the RIS. For

[1] It is noted that when $S = 1$, the concept of sub-band degenerates into sub-channel. Introducing sub-band is to obtain a better resource utilization during the transmission of gradient symbols. As in SigSGD scheme, each quantized gradient has only 1 bit, which does not require a large bandwidth for transmission. Thence, the quantization gradient can be transmitted more efficiently once each sub-band is divided into multiple sub-channels, due to the frequency diversity.

[2] This assumption is reasonable since in most federated learning scenarios, where the locations of devices, the BS together with the RIS are fixed basically, and thus the channels are generally considered to be constant under the case of small fading with sufficiently small fluctuation (e.g., the mmWave communication).

each involved channel, the channel state information (CSI) is able to estimate perfectly at both the BS and the RIS,[3] which is the same as [20] and [27]. In addition, we show by numerical results in Section V the impact of imperfect CSI on the learning performance as well. Note that, a common reflection coefficients of the RIS is configured to cater all the involved channels [22].

To facilitate the analysis in the sequel, a fixed digital constellation is adopted across all edge devices [20]. For simplicity, once generating quantized local gradient, devices need to map each quantized element to one digital symbol, i.e., $\mathcal{Q} \triangleq \{-1, 1\}$ by adopting binary phase shift keying (BPSK) modulation without loss of generality.[4] Let $\tilde{\mathbf{g}}_k^{(n)} = [\tilde{g}_{k,1}, \tilde{g}_{k,2}, \ldots, \tilde{g}_{k,q}]^T$ stores the channel input vector consisting of quantized gradient with size q of device k within the n-th iteration, where $\tilde{g}_{k,q} \in \mathcal{Q}$. All the devices transmit simultaneously over the sub-channels belonging to the allocated sub-band during the gradient-uploading phase. Assume the m_k-th sub-band is assigned to the k-th device. As a result, gradient-uploading duration will consist of $T_s = \frac{q}{S}$ OFDM symbols for transmitting quantized gradient in each learning iteration [7]. Also, it is assumed that the power is identically distributed over the sub-channels belonging to each sub-band. To facilitate the problem formulation, we vectorize the power allocation factor as $\boldsymbol{p} = [p_1, p_2, \ldots, p_K]^T$, where p_k indicates the transmit power at each sub-channel of device k, which will be optimized as presented in the sequel. Thence, at the t-th OFDM symbol and s-th sub-channel within m_k-th sub-band, the BS receives the i-th quantized gradient $\tilde{g}_{k,i}$, with $i = (t-1)S + s$, as

$$\hat{\mathbf{y}}_{k,i} = \left(\mathbf{G}_{k,m_k}\boldsymbol{\Theta}\mathbf{h}_{k,m_k}^r + \mathbf{h}_{k,m_k}^d\right)\sqrt{p_k}\tilde{g}_{k,i} + \mathbf{n}_{k,m_k}, \tag{6}$$

with $\mathbf{n}_{k,m_k} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_J)$ modeling the additive white Gaussian noise (AWGN) vector, whose entries obey a zero-mean variance σ^2 complex Gaussian distribution. Note that $p_k = 0$ means the k-th device is not selected, i.e., $k \notin \lambda$. The associated power control policy p_k satisfies a one-shot transmission power constraint, i.e., $S\sum_{k=1}^{K} p_k \le P_0$, with P_0 denoting the maximum total power.

By employing a maximum likelihood estimator (MLE), the BS computes the estimate of $\tilde{g}_{k,i}$ as $\max\limits_{\bar{g}_{k,i} \in \mathcal{Q}} \mathbb{P}(\hat{\mathbf{y}}_{k,i} | \bar{g}_{k,i})$, where $\bar{g}_{k,i}$ denotes the quantized gradient after decoding. Note that, due to the existence of wireless noise and fading, certain error inevitably occurs in this estimation, where the union bound on pairwise bit error rate (BER) of $\tilde{g}_{k,i}$ being erroneously detected as $\bar{g}_{k,i}$, i.e., $\mathbb{P}(\tilde{g}_{k,i} \to \bar{g}_{k,i})$ can be written as [23] $\mathbb{P}(\tilde{g}_{k,i} \to \bar{g}_{k,i}) = Q(\sqrt{\frac{\Omega_k}{2\sigma^2}})$, with $\Omega_k = \left\|(\mathbf{G}_{k,m_k}\boldsymbol{\Theta}\mathbf{h}_{k,m_k}^r + \mathbf{h}_{k,m_k}^d)(\tilde{g}_{k,i} - \bar{g}_{k,i})\right\|^2 p_k$, $\forall \tilde{g}_{k,i}, \bar{g}_{k,i} \in \mathcal{Q}$. Let $\mathbf{h}_{k,m_k}^{\Delta} = \mathbf{G}_{k,m_k}\boldsymbol{\Theta}\mathbf{h}_{k,m_k}^r + \mathbf{h}_{k,m_k}^d$. By substituting Ω_k, we have

$$\mathbb{P}_k^{trans} = \mathbb{P}(\tilde{g}_{k,i} \to \bar{g}_{k,i}) = Q\left(\sqrt{\frac{2\left\|\mathbf{h}_{k,m_k}^{\Delta}\right\|^2 p_k}{\sigma^2}}\right), \tag{7}$$

with $Q(x)$ denoting the tail function of the standard normal distribution, given as $Q(x) = \int_x^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$.

After estimating all quantized gradients in the OFDM symbol t, a decoded gradient vector can be constructed as $\bar{\mathbf{g}}_k^t = \mathrm{vec}\left(\{\bar{g}_{k,i} | (t-1)S + 1 \le i \le tS\}\right)$. Next, by cascading all T_s OFDM symbols, we can recover the full-dimension quantized gradient as $\bar{\mathbf{g}}_k = \left[\left[\bar{\mathbf{g}}_k^1\right]^T, \left[\bar{\mathbf{g}}_k^2\right]^T, \ldots, \left[\bar{\mathbf{g}}_k^{T_s}\right]^T\right]^T$. Once $\bar{\mathbf{g}}_k$, for $\forall k$, is obtained, the one-bit gradient aggregation can be executed as Equation (4), i.e., generating the sum of sign vectors $\bar{\mathbf{g}}$, e.g., for i-th entry of $\bar{\mathbf{g}}^{(n)}$, it is produced by $\bar{g}_i^{(n)} = \sum_{k \in \lambda} \bar{g}_{k,i}^{(n)}$, and the global gradient estimate $\bar{\mathbf{v}}$ can be obtained by simply taking the element-wise sign of $\bar{\mathbf{g}}$, i.e., $\bar{\mathbf{v}} = \mathrm{sign}(\bar{\mathbf{g}})$.

Thanks to the enough transmit power available at the BS and the whole downlink bandwidth can be used for broadcasting, we suppose that the global gradient parameters $\bar{\mathbf{v}}$ can be perfectly transmitted to the devices, as assumed in [7] and [20]. Finally, by utilizing Equation (5), devices can update the global model and then start the next learning iteration. It is worth emphasizing that, the BER caused by fading and communication noise over wireless channel may potentially bring about estimation error in $\bar{\mathbf{g}}_k$, thus inevitably affecting $\bar{\mathbf{v}}$. As a result, the global model update in Equation (5) may produce an inaccurate global model update, so as to delay the convergence of FEEL. To alleviate this concern, we will quantitatively characterize this impact in the next section.

III. Convergence Rate Analysis and Problem Formulation

In this section, we formally analyze the learning performance of the RIS-aided SignSGD-based FEEL system. Specifically, several standard assumptions of the stochastic optimization involving the loss function and gradient are introduced in Section III-A. According to these assumptions, we derive an easy-to-understand intuition of how the characteristics of wireless networks affect the learning performance by analyzing the level of the gradient noise introduced by the data-stochasticity and the wireless channel, as described in Section III-B. Immediately, in Section III-C, we establish the system design foundation as a unified gradient noise minimization optimization task over the sub-band assignment strategy, the transmit power vector $\boldsymbol{p}$, and the RIS configuration matrix $\boldsymbol{\Theta}$.

A. Assumptions and Preliminaries

Similar to [7], we define a non-convex loss function in our convergence analysis, thus allowing the derived theories

[3] To obtain relatively accurate CSI, all RIS elements are assumed to equip with receiving radio frequency (RF) chains, and thus that conventional channel estimation methods can be effectively applied at both RIS and BS. Before training procedure, all devices send orthogonal pilots to the BS, which subsequently performs channel estimation using collected signals to obtain a relatively accurate CSI. Some early attempts in channel estimation can be found in [25] and [26]. For instance, [25] proposed a brute-force method, in which the CSI with respect to each RIS element is estimated sequentially by the BS while turning off other elements, and the full CSI with low training overhead is obtained by a channel construct approach based on compressive sensing tools in [26].

[4] Even though the BPSK modulation is adopted for simplicity, we emphasize that the extension of our system to higher-order modulation configuration is straightforward by simply combining multiple quantized gradients to construct a high-order modulation symbol. Simultaneously, our convergence analysis framework also applies well once the bit error rate expression of the corresponding high-order modulation strategy is clearly known.

to be applicable to the popular neural networks in the same way.

Assumption 1 (Lower bound): The associated loss function of arbitrary parameter model $\mathbf{w}$ has a lower bound F^*, i.e., $F(\mathbf{w}) \geq F^*$, $\forall \mathbf{w}$, with F^* being a constant.

Assumption 2 (Smooth): Let $\mathbf{g}$ represent the gradient of the associated loss function $F(\mathbf{w})$ estimated at point model vector $\mathbf{w} = [w_1, w_2, \ldots, w_q]$ where q is the total number of model $\mathbf{w}$. For $\forall \mathbf{w}$, $\mathbf{w}'$, $\exists \mathbf{L} = [L_1, L_2, \ldots, L_q]$, where L_q, $\forall q$ is non-negative constant, satisfy

$$|F(\mathbf{w}') - [F(\mathbf{w}) + \mathbf{g}^T(\mathbf{w}' - \mathbf{w})]| \leq \frac{1}{2}\sum\nolimits_{i=1}^{q} L_i(w_i' - w_i)^2. \quad (8)$$

Assumption 3 (Variance bound): Upon any $\mathbf{w} \in \mathbb{R}^q$, the stochastic gradient estimate $\{\mathbf{g}_j\}$ in Equation (3) provides independent and unbiased estimates of the batch gradient $\mathbf{g} = \nabla F(\mathbf{w})$ that has coordinate bounded variance $\mathbb{E}[\mathbf{g}_j] = \mathbf{g}$, $\forall j$, and $\mathbb{E}[(g_{j,i} - g_i)^2] \leq \sigma_i^2$, $\forall j$, i, for a vector of non-negative constants $\boldsymbol{\sigma} = [\sigma_1, \sigma_1, \ldots, \sigma_q]$, with $g_{j,i}$ and g_i denoting the i-th entry of $\mathbf{g}_j(\mathbf{w})$ and $\mathbf{g}(\mathbf{w})$ respectively.

Assumption 4 (Unimodal, Symmetric Gradient Noise): For arbitrary $\mathbf{w}$, entries of stochastic gradient $\mathbf{g}_j(\mathbf{w})$, $\forall j$, obey a unimodal distribution and are symmetrical around its mean.[5]

Assumption 1 formally states an essential lower bound to guarantee convergence to a stationary point [28], i.e., it ensures that a global optimum $\mathbf{w}$ exists for the loss function F; Assumptions 2–3 are standard assumptions among stochastic optimization literature, e.g., [28]; Assumption 4 reveals the gradient noise caused by data-stochasticity [7], which is generated by the gradient calculation strategy of devices, i.e., using randomly sampled batch instead of the ground-truth full-batch samples, thus directly bringing about the discrepancy between $\mathbf{g}_j$ and $\mathbf{g}$, as verified in [6].

Referring to the above assumptions, once learning rate η^* is found, the following upper bound on the loss function $F(\mathbf{w}^{(n+1)})$ with respect to the recursion in Equation (5) can be proved.

Lemma 1: Suppose that the loss function F satisfies Assumptions 1–4 and the parameter vector at the n-th learning iteration is described as $\mathbf{w}^{(n)}$. If we set learning rate $\eta = \eta^*$, we have

$$\mathbb{E}[F^{(n+1)} - F^{(n)}|\mathbf{w}^{(n)}] \leq -\eta\|\mathbf{g}^{(n)}\|_1 + \frac{\eta^2}{2}\|\mathbf{L}\|_1 + 2\eta\sum\nolimits_{i=1}^{q}\left|g_i^{(n)}\right|\mathbb{P}[\text{sign}(\bar{g}_i^{(n)}) \neq \text{sign}(g_i^{(n)})], \quad (9)$$

with $\mathbb{P}[\text{sign}(\bar{g}_i^{(n)}) \neq \text{sign}(g_i^{(n)})]$ denoting the error probability of the sign of each entry of the stochastic gradient $\bar{g}_i^{(n)}$ compared with the true gradient $g_i^{(n)} = \nabla F_i^{(n)}$. The expectation and the probability are over the dynamics of the wireless channels.

Proof: See Appendix A.

B. Learning Convergence Analysis

With the above assumptions, tractable convergence analysis can be presented as follows. Note that the results establish the convergence behavior of our system. Throughout the paper, the learning rate is set as $\eta = \frac{1}{\sqrt{\|L\|_1 n_b}}$ with the batchsize n_b being set to $\frac{1}{\gamma}N$, where γ denotes an arbitrary non-negative constant, and N indicates the number of communication rounds.

Utilizing Lemma 1 and plugging $\bar{g}_i^{(n)}$ into Equation (9), we have

$$\mathbb{E}[F^{(n+1)} - F^{(n)}|\mathbf{w}^{(n)}] \leq -\eta\|\mathbf{g}^{(n)}\|_1 + \frac{\|\mathbf{L}\|_1\eta^2}{2} + 2\eta\sum\nolimits_{i=1}^{q}\left|g_i^{(n)}\right|\mathbb{P}[\text{sign}\left(\sum\nolimits_{k\in\bar{\lambda}}\bar{g}_{k,i}^{(n)}\right) \neq \text{sign}(g_i^{(n)})]. \quad (10)$$

Thence, to analyze the upper bound of $\mathbb{E}[F^{(n+1)} - F^{(n)}|\mathbf{w}^{(n)}]$, we need derive a tractable expressions (upper bound) for $\mathbb{P}[\text{sign}(\sum_{k\in\bar{\lambda}}\bar{g}_{k,i}^{(n)}) \neq \text{sign}(g_i^{(n)})]$, which mainly depends on the degree of similarity between the gradient symbol after majority vote and the true gradient intuitively.

To begin with, we first derive the convergence rate over error-free channel of our system, for comparison in the sequel. To this end, a tractable expressions (upper bound) for $\mathbb{P}[\text{sign}(\sum_{k\in\bar{\lambda}}\bar{g}_{k,i}^{(n)}) \neq \text{sign}(g_i^{(n)})]$ can be established, which is given in the following lemma.

Lemma 2: With Assumptions 1–4, we can derive a probability upper bound over error-free channel as $\mathbb{P}[\text{sign}(\sum_{k\in\bar{\lambda}}\bar{g}_{k,i}^{(n)}) \neq \text{sign}(g_i^{(n)})] \leq \frac{1}{\Sigma_i\sqrt{K_c}}$, with $\Sigma_i = \sqrt{n_b}\frac{|g_i^{(n)}|}{\sigma_i}$ denoting the gradient-signal-to-data-noise ratio. $\bar{\lambda}$ indicates the index of participating devices with size K_c.

Proof: See Appendix B.

Next, by plugging Lemma 2 into Equation (10), we can stablish a non-convex convergence rate over error-free channel.

Theorem 1 (Convergence rate over error-free channel): Run one-bit gradient quantization with majority vote for N iterations over error-free channel under Assumptions 1–4, and set the learning rate η and mini-batch size n_b for each edge device as $\eta = \frac{1}{\sqrt{\|L\|_1 n_b}}$ and $n_b = \frac{1}{\gamma}N$. We have

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1}\|\mathbf{g}^{(n)}\|_1\right] \leq \frac{a_{\text{ERF}}}{\sqrt{N}}\left(\sqrt{\frac{\|\mathbf{L}\|_1}{\gamma}}\left(F^{(0)} - F^* + \frac{\gamma}{2}\right) + 2b_{\text{ERF}}\sqrt{\gamma}\|\boldsymbol{\sigma}\|_1\right), \quad (11)$$

where the scaling factor a_{ERF} and b_{ERF} are respectively expressed as $a_{\text{ERF}} = 1$, $b_{\text{ERF}} = \frac{1}{\sqrt{K_c}}$.[6]

Proof: See Theorem 2 derived in [6]. The complete proof is skipped due to space limitation.

Remark 1: Theorem 1 relates the norm of the gradient to the expected improvement made in a single algorithmic step, which is used to compare with the total possible improvement under Assumption 1, thus providing an upper bound on the

[5] Clearly, Gaussian noise is a special case. Note that even for a moderate mini-batch size, we expect the central limit theorem to kick in rendering typical gradient noise distributions close to Gaussian, whicn is the same as [7].

[6] The definition of a_{ERF} and b_{ERF} is to facilitate the comparison with the convergence rate over RIS-assisted fading channel, which reflects the effect of data-stochasticity on the convergence rate.

average gradient norm under non-convex loss function.[7] The effect of data-stochasticity on the non-convex convergence rate of signSGD-FL can be represented by the bias term $2b_{\text{ERF}}\sqrt{\gamma}\|\boldsymbol{\sigma}\|_1$, with $\boldsymbol{\sigma}$ being the standard deviation vector of the gradient noise. This is generated by the gradient calculation strategy of devices, i.e., using randomly sampled batch instead of the ground-truth full-batch samples. It can be seen that as the number of participating edge devices $K_c \to \infty$, $b_{\text{ERF}} \to 0$ and the bias term $2b_{\text{ERF}}\sqrt{\gamma}\|\boldsymbol{\sigma}\|_1 \to 0$, thus achieving a better convergence rate.

Similar to the way to obtain the convergence rate over error-free channel, we start our derivation by modeling a tractable upper bound over RIS-assisted fading channel, which is jointly determined by communication noise and data-stochasticity, unlike the error-free channel.

Lemma 3: With Assumptions 1–4, we obtain a probability upper bound over RIS-assisted fading channel described in Section II-B as

$$\mathbb{P}[\text{sign}(\textstyle\sum_{k\in\bar{\lambda}} \bar{g}_{k,i}^{(n)}) \neq \text{sign}(g_i^{(n)})] \tag{12}$$

$$\leq \frac{\sqrt{K_c}}{\Sigma_i \sum_{k\in\bar{\lambda}} \Xi_k} + \frac{\sqrt{K_c}}{2\sum_{k\in\bar{\lambda}} \Xi_k}\sqrt{1-\left(\frac{1}{\sqrt{K}\sqrt{K_c}}\sum_{k\in\bar{\lambda}} \Xi_k\right)^2}, \tag{13}$$

where $\Xi_k = 1 - 2\mathbb{P}_k^{trans}$ can be seen as the wireless channel perfection, and $\mathbb{P}_k^{trans}$ indicates the BER defined in Equation (7), which reveals the effect of the communication noise on the correct decoding of quantization gradient during data transmission.

Proof: See Appendix C.

Following the analogous convergence rate analysis with the error-free channel and utilizing Lemma 3, the non-convex convergence rate over RIS-assisted fading channel can be established.

Theorem 2 (Convergence Rate Over RIS-Assisted Fading Channel): Run one-bit gradient quantization with majority vote for N iterations over RIS-assisted fading channel under Assumptions 1–4, and set the learning rate η and mini-batch size n_b for each edge device as $\eta = \frac{1}{\sqrt{\|L\|_1 n_b}}$ and $n_b = \frac{1}{\gamma}N$. We have the following improved convergence rate

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1}\|\mathbf{g}^{(n)}\|_1\right] \leq \frac{a_{\text{RIS}}}{\sqrt{N}}\left(\sqrt{\frac{\|\mathbf{L}\|_1}{\gamma}}\left(F^{(0)} - F^* + \frac{\gamma}{2}\right) + 2b_{\text{RIS}}\sqrt{\gamma}\|\boldsymbol{\sigma}\|_1\right), \tag{14}$$

with the scaling factor $a_{\text{RIS}} = \frac{1}{1-\frac{\sqrt{K}_c\sqrt{1-\left(\frac{1}{\sqrt{K}\sqrt{K_c}}\sum_{k\in\bar{\lambda}}\Xi_k\right)^2}}{\sum_{k\in\bar{\lambda}}\Xi_k}}$ and $b_{\text{RIS}} = \frac{\sqrt{K}_c}{\sum_{k\in\bar{\lambda}}\Xi_k}$, respectively.[8]

Proof: See Appendix D.

Remark 2: The existence of BER slows down the convergence rate by increasing the value of two positive scaling factors a_{ERF} and b_{ERF} defined in Theorem 1, and thus two larger scaling items, a_{RIS} and b_{RIS} are generated with $a_{\text{RIS}} \geq a_{\text{ERF}}$ and $b_{\text{RIS}} \geq b_{\text{ERF}}$. In this case, the upper bound on the time-averaged gradient norm inevitably increases compared with the error-free scenario.[9] By analyzing the forms of a_{RIS} and b_{RIS}, we can clearly see that as the number of participating edge devices is fixed, the smaller the sum of BERs of participating devices, the faster the FL converges.

C. Problem Formulation

Theorem 2 reveals that the effect of channel hostilities and data-stochasticity on the non-convex convergence rate of signSGD-FL can be translated into two scaling factors $a_{\text{RIS}} \in (1, \infty)$ and $b_{\text{RIS}} \in (0, \infty)$, which slow down the convergence rate potentially. Nevertheless, as shown in a_{RIS} and b_{RIS}, we can conclude that a_{RIS} is a monotonically increasing function with respect to b_{RIS}. Thence, by minimizing b_{RIS}, a_{RIS} will decrease simultaneously, thereby reducing the upper bound on the time-averaged gradient norm and achieving faster convergence. This inspires us to treat b_{RIS} as the metric of the communication-learning co-design over RIS-aided fading channel by optimizing the sub-band assignment strategy, the transmit power vector $\boldsymbol{p}$, and the RIS configuration matrix $\boldsymbol{\Theta}$.

Note that $\sqrt{K_c}$ can be equivalently rewritten as $\sqrt{\|\boldsymbol{p}\|_0}$ where $\|\boldsymbol{p}\|_0$ refers to the number of nonzero elements in the vector $\boldsymbol{p}$, i.e., $\|\boldsymbol{p}\|_0 \triangleq |\{i : p_i \neq 0\}|$. Simultaneously, let an auxiliary variable $\boldsymbol{A}$ denote the sub-band assignment matrix, where $[\boldsymbol{A}]_{(k,m)} \in \{0, 1\}$ indicates whether the m-th sub-band is allocated to device k, i.e., $[\boldsymbol{A}]_{(k,m)} = 1$ if sub-band m is assigned to device k, and $[\boldsymbol{A}]_{(k,m)} = 0$ otherwise. Let $\mathbf{h}_{k,m_k}^{\Delta} = \mathbf{G}_{k,m_k}\boldsymbol{\Theta}\mathbf{h}_{k,m_k}^{r} + \mathbf{h}_{k,m_k}^{d}$. Define

$$\Pi_k = 1 - 2Q\left(\sqrt{\frac{2\sum_{m=1}^{M}[\boldsymbol{A}]_{(k,m)}\left\|\mathbf{h}_{k,m_k}^{\Delta}\right\|^2 p_k}{\sigma^2}}\right). \tag{15}$$

We can easily verify that $\Pi_k = \Xi_k, \forall k$, under the constraints of $\sum_{k=1}^{K}[\boldsymbol{A}]_{(k,m)} \leq 1, \forall m$ and $\sum_{m=1}^{M}[\boldsymbol{A}]_{(k,m)} \leq 1, \forall k$ corresponding to the sub-band assignment rules. Actually, once $k \notin \bar{\lambda}$, i.e., $p_k = 0$, we always have $\Pi_k = 0$ due to the fact that $Q(0) = \frac{1}{2}$. Then, for an arbitrary $\bar{\lambda}$, we always have $\sum_{k\in\bar{\lambda}}\Pi_k = \sum_{k\in\bar{\lambda}}\Pi_k + \sum_{k\notin\bar{\lambda}}\Pi_k = \sum_{k=1}^{K}\Pi_k$. Thence, the communication-learning design problem, with aims to minimize b_{RIS} over the feasible set of $\{\boldsymbol{A}, \boldsymbol{\Theta}, \boldsymbol{p}\}$, can be formulated as

$$(\mathcal{P})\ \min_{\boldsymbol{A},\boldsymbol{\Theta},\boldsymbol{p}} \frac{\sqrt{\|\boldsymbol{p}\|_0}}{\sum\limits_{k=1}^{K}\Pi_k} \tag{16a}$$

[7] The upper bound decays like $O\left(\frac{1}{\sqrt{N}}\right)$, until the gradient vanishes, thus establishing convergence. It is the same as non-convex convergence rate of SGD.

[8] The convergence analysis in Theorem 2 can be easily extended to the scenario with time-varying channels among different training iterations. By adding the iteration identification to the probability upper bound in Lemma 3 and subsequently substituting the results into Equation (10), the convergence rate under time-varying channels can be derived analogously.

[9] It is worth noting that [7] establishes a relationship between the transmitter signal to noise ratio and the non-convex convergence rate under over-the-air aggregation and the truncation-based power allocation strategy. However, the influence of the wireless fading channel on the FEEL convergence is not completely reflected. By contrast, Theorem 2 indicates that the wireless fading potentially affects the communication BER thereby causing the inevitable errors during one-bit aggregation, and decelerating the FEEL convergence. Moreover, this performance attenuation can be greatly alleviated by the deployment of the RIS.

$$\text{s.t.}\quad \sum\nolimits_{k=1}^{K} [\boldsymbol{A}]_{(k,m)} \le 1, \forall m, \tag{16b}$$

$$\sum\nolimits_{m=1}^{M} [\boldsymbol{A}]_{(k,m)} \le 1, \forall k, \tag{16c}$$

$$[\boldsymbol{A}]_{(k,m)} \in \{0,1\}, \forall m, k, \tag{16d}$$

$$\left|[\boldsymbol{\Theta}]_{(i,i)}\right| = 1, \forall i = 1, \cdots, L, \tag{16e}$$

$$S\sum\nolimits_{k=1}^{K} p_k = P_0, p_k \ge 0, \forall k = 1, \cdots, K \quad , \tag{16f}$$

with constraints (16b) and (16c) guaranteeing each sub-band is assigned to at most one device, and each device transmits gradient using only one sub-band. Constraints (16d) ensures the assignment variable to be binary, and constraint (16e) denotes the unit-modulus requirements of the RIS elements, separately; constraint (16f) is derived from the maximum power constraint and the following key intuition.

Proposition 1: Given the sub-band assignment strategy $\boldsymbol{A}$ and the RIS configuration matrix $\boldsymbol{\Theta}$, the optimal power control strategy obeys an allocation rule where the maximum power is fully consumed by participating devices.

Proof: Given the sub-band assignment strategy $\boldsymbol{A}$ and the RIS configuration matrix $\boldsymbol{\Theta}$, $Q\left(\sqrt{\alpha p_k}\right)$ with a positive constant α is a decreasing function of $p_k \ge 0$ as the first derivative of $Q\left(\sqrt{\alpha p_k}\right)$ satisfying $\frac{\partial Q\left(\sqrt{\alpha p_k}\right)}{p_k} = -\sqrt{\frac{\alpha}{8\pi}} e^{-\alpha p_k/2} p_k^{-1/2} < 0$. Once all power is allocated completely to the participating devices, the denominator in the objective function of problem $(\mathcal{P})$ reaches the maximum level, thus achieving optimal solution. ■

Despite the conciseness of problem $(\mathcal{P})$, it turns out to be a mixed-integer nonlinear programming problem, which is highly intractable due to the non-convex objective function with the ℓ_0 norm term and the coupling of $\boldsymbol{A}$, $\boldsymbol{\Theta}$, and $\boldsymbol{p}$, together with the non-convex unit-modulus constraint and binary assignment variables. To tackle this issue, an alternating optimization (AO) framework is invoked as an intuitive approach to solve the formulated joint optimization problem $(\mathcal{P})$ in an efficient manner, as described in the following section.

IV. Alternating Optimization for Communication-Learning Co-Design

In this section, we propose to decouple problem $(\mathcal{P})$ into several tractable sub-problems by introducing AO method, where the sub-band assignment matrix $\boldsymbol{A}$, the power allocation vector $\boldsymbol{p}$, and the RIS configuration matrix $\boldsymbol{\Theta}$ are optimized in an alternative manner until the algorithm converges.

A. Optimizing Sub-Band Assignment Strategy Given Transmit Power and RIS Phase Shifts

For given RIS configuration matrix $\boldsymbol{\Theta}$ and the power allocation vector $\boldsymbol{p}$, the optimal sub-band assignment strategy can be established by solving the following optimization problem

$$(\mathcal{P}_1)\quad \max_{\boldsymbol{A}} \quad \sum_{k=1}^{K} \Pi_k \quad \text{s.t.}\quad (16\text{b}), (16\text{c}), (16\text{d}). \tag{17a}$$

We note that the non-convexity of problem $(\mathcal{P}_1)$ is presented by a binary assignment variable which makes the strategy design very challenging. Fortunately, by adopting the difference-of-convex penalty-based method [29], the binary variables $[\boldsymbol{A}]_{(k,m)} \in \{0,1\}$ can be relaxed to continuous variables $0 \le [\boldsymbol{A}]_{(k,m)} \le 1$ without loss of optimality under the constraint of

$$\Gamma(\boldsymbol{A}) = \sum\nolimits_{k=1}^{K} \sum\nolimits_{m=1}^{M} \left([\boldsymbol{A}]_{(k,m)} - \left([\boldsymbol{A}]_{(k,m)}\right)^2\right) \le 0, \tag{18}$$

which involves a difference of convex functions. According to the Proposition 1 within [29], with the penalty term $\mu\Gamma(\boldsymbol{A})$, there exists $\bar{\mu} \ge 0$ such that, for any $\mu \in [0, \bar{\mu}]$, the original problem $(\mathcal{P}_1)$ has the same optimal solutions with the following penalty-based problem

$$(\mathcal{P}_{1.1})\ \max_{\boldsymbol{A}} \sum\nolimits_{k=1}^{K} \Pi_k - \mu\Gamma(\boldsymbol{A}) \tag{19a}$$

$$\text{s.t.}\ (16\text{b}), (16\text{c}), \tag{19b}$$

$$0 \ \le\ [\boldsymbol{A}]_{(k,m)} \le 1, \forall m, k. \tag{19c}$$

Note that the penalty will be plenty large, such that, the non-integer solutions to $\boldsymbol{A}$ are penalized.[10]

The challenge of solving the penalty problem $(\mathcal{P}_{1.1})$ arises from the concavity of $-\left([\boldsymbol{A}]_{(k,m)}\right)^2$, which can be effectively handled by leveraging the SCA approximation to sequentially convexify the concave function. In particular, for a given feasible point $[\boldsymbol{A}]^{\nu}_{(k,m)}$ obtained from the ν-th iteration, a global underestimator of $\left([\boldsymbol{A}]_{(k,m)}\right)^2$ can be constructed as $\left([\boldsymbol{A}]_{(k,m)}\right)^2 \ge 2[\boldsymbol{A}]_{(k,m)}[\boldsymbol{A}]^{\nu}_{(k,m)} - \left([\boldsymbol{A}]^{\nu}_{(k,m)}\right)^2$ based on the first-order Taylor expansion, and thus the optimization problem solved in the $(\nu+1)$-th iteration is given by

$$(\mathcal{P}_{1.2})\ \max_{\boldsymbol{A}} \sum\nolimits_{k=1}^{K} \Pi_k - \mu \sum\nolimits_{k=1}^{K} \sum\nolimits_{m=1}^{M} [\boldsymbol{R}]^{\nu}_{(k,m)} \tag{20a}$$

$$\text{s.t.}\ (16\text{b}), (16\text{c}), (19\text{c}). \tag{20b}$$

with $[\boldsymbol{R}]^{\nu}_{(k,m)} = [\boldsymbol{A}]_{(k,m)} - 2[\boldsymbol{A}]_{(k,m)}[\boldsymbol{A}]^{\nu}_{(k,m)} + \left([\boldsymbol{A}]^{\nu}_{(k,m)}\right)^2$.

To proceed further, a tight and simple upper bound of the Q-function, defined as $Q(x) \le \vartheta(x) = \frac{1}{6}e^{-2x^2} + \frac{1}{12}e^{-x^2} + \frac{1}{4}e^{-x^2/2}$ [30], [31] can be adopted. By replacing the Q-function in Π_k with ϑ-function, problem $(\mathcal{P}_{1.2})$ can be efficiently solved by a standard convex optimization solver such as CVXPY [32].

B. Optimizing Transmit Power Given Sub-Band Assignment Strategy and RIS Phase Shifts

For given RIS configuration matrix $\boldsymbol{\Theta}$ and sub-band assignment strategy $\boldsymbol{A}$, problem $(\mathcal{P})$ can be simplified into

$$(\mathcal{P}_2)\quad \min_{\boldsymbol{p}}\quad \frac{\sqrt{\|\boldsymbol{p}\|_0}}{\sum\limits_{k=1}^{K} \Pi_k} \quad \text{s.t.}\quad (16\text{f}). \tag{21a}$$

The objective function of problem $(\mathcal{P}_2)$ is intractable due to the non-convex ℓ_0 norm and tricky Q-function, which need to be simplified into an easy-to-handle form. Similarly, the upper

[10] The weight of penalty μ must be chosen as the same magnitude of the original objective, thus sufficiently approximating the original binary problem.

Algorithm 1 Sloving Problem ($\mathcal{P}_2$) by Quasi-Convex Optimization With Relaxed ℓ_0 Norm Approximation.

```
Input: Reweighting function $\boldsymbol{\varsigma}^{(1)} = \mathbf{1}$.
for each $i = 1, 2, \cdots, i^{\max}$ do
    Initialize $\boldsymbol{p}^{(1)}$.
    for each $j = 1, 2, \cdots, j^{\max}$ do
        Given bisection interval $\cup = [l, u]$, tolerance $\epsilon$.
        while $u - l \leq \epsilon$ do
            $\chi := \frac{l+u}{2}$.
            Solve the convex feasibility problem ($\mathcal{P}_{2.3}$).
            if Problem ($\mathcal{P}_{2.3}$) is feasible then
                $u := \chi$;
            else
                $l := \chi$.
        Output $\boldsymbol{p}^{(j+1)}$.
    Output $\boldsymbol{p}^{(i)}$, and update $\boldsymbol{\varsigma}^{(i+1)}$ using Equation (23).
```

bound of the Q-function defined as $\vartheta(x)$ can be introduced. However, problem ($\mathcal{P}_2$) is still a non-convex problem due to the ℓ_0 norm optimization. Thence, we will creatively convert problem ($\mathcal{P}_2$) into another solvable form. Let $f_1(\boldsymbol{p})$ refer to the denominator term approximated by $\vartheta(x)$ in the objective function of ($\mathcal{P}_2$). Referring to the ℓ_0 norm relax method in [33], we can further approximate the non-convex ℓ_0 norm by introducing a reweighting function to enforce the sparsity in a democratic way, as

$$(\mathcal{P}_{2.1}) \quad \min_{\boldsymbol{p}} \quad \frac{\sqrt{\sum_{k=1}^{K} \varsigma_k^{(i)} p_k^2}}{f_1(\boldsymbol{p})} \quad \textbf{s.t.} \quad (16\text{f}), \tag{22a}$$

with superscript (i) denoting the i-th iteration. The k-th entry of the reweighting function $\boldsymbol{\varsigma}$ in each iteration can be established as

$$\varsigma_k = \frac{\varrho}{2}\left[\left(p_k^{(i)}\right)^2 + \lambda^2\right]^{\frac{\varrho}{2}-1}, \tag{23}$$

where $0 \leq \varrho \leq 1$, and the regularizer parameter $\lambda > 0$ is added to avoid yielding infinite values when some $\varsigma_k^{(i)}$ become zeros in the iterations. Notice that the optimization ($\mathcal{P}_{2.1}$) is executed in an iterative manner such that the coefficients $\boldsymbol{\varsigma}^{(i+1)}$ are updated using the $\boldsymbol{\varpi}$ after the i-th iteration.

Next, we begin to analyze the problem ($\mathcal{P}_{2.1}$) that needs to be solved in each iteration. It can be found that both the denominator and molecular term in the relaxed objective function are concave due to the fact that the second derivatives of $\frac{\partial Q\left(\sqrt{\alpha x}\right)}{x^2} = \frac{1}{2}\sqrt{\frac{\alpha}{8\pi}} e^{-\alpha x/2} x^{-1/2}\left(\frac{1}{x} + \alpha\right) > 0$ with $\alpha \geq 0$ and $x \geq 0$. Thence, problem ($\mathcal{P}_{2.1}$) is a tricky non-convex optimization problem, which cannot be efficiently solved by a standard convex optimization solver. To address this issue, we propose to linearize the molecular term by adopting the continuous first-order Taylor expansion, thus converting the objective function into a form of quasi-convex function. Specifically, in the j-th approximation iteration, each subproblem of ($\mathcal{P}_{2.1}$) can be rewritten as

$$(\mathcal{P}_{2.2}) \min_{\boldsymbol{p}_j} \frac{\sqrt{\sum_{k=1}^{K} \varsigma_k^{(i)} \tilde{p}_k^2} + \frac{1}{2}\frac{1}{\sqrt{\sum_{k=1}^{K} \varsigma_k^{(i)} \tilde{p}_k^2}} \boldsymbol{\upsilon}^T \left(\boldsymbol{p}_j - \tilde{\boldsymbol{p}}\right)}{f_1(\boldsymbol{p}_j)} \tag{24a}$$

$$\textbf{s.t.} \ (16\text{f}), \tag{24b}$$

with $\tilde{\boldsymbol{p}}$ denoting the value of $\boldsymbol{p}$ in the $(j-1)$-th iteration. $\boldsymbol{\upsilon}$ is a column vector with size $K \times 1$, each element of which can be expressed as $\upsilon_k = 2\varsigma_k^{(i)} \tilde{p}_k$.

Now, it is easy to verify that problem ($\mathcal{P}_{2.2}$) is a quasi-convex optimization problem, which can be solved efficiently via bisection method. Specifically, let $f_2(\boldsymbol{p}_j) = \sqrt{\sum_{k=1}^{K} \varsigma_k^{(i)} \tilde{p}_k^2} + \frac{1}{2}\frac{1}{\sqrt{\sum_{k=1}^{K} \varsigma_k^{(i)} \tilde{p}_k^2}} \boldsymbol{\upsilon}^T \left(\boldsymbol{p}_j - \tilde{\boldsymbol{p}}\right)$, then problem ($\mathcal{P}_{2.2}$) can be turned into a feasibility check problem, as

$$(\mathcal{P}_{2.3})\text{find} \quad \boldsymbol{p}_j \quad \textbf{s.t.} \ f_2(\boldsymbol{p}_j) - \chi f_1(\boldsymbol{p}_j) \leq 0, (16\text{f}), \tag{25a}$$

which can be efficiently solved by a standard convex optimization solver. Let $\boldsymbol{p}_j^*$ be the optimal value of problem ($\mathcal{P}_{2.3}$). We can check whether the optimal $\frac{f_2(\boldsymbol{p}_j^*)}{f_1(\boldsymbol{p}_j^*)}$ is less than or more than a given value χ by solving the convex problem ($\mathcal{P}_{2.3}$). If ($\mathcal{P}_{2.3}$) is feasible, we have $\frac{f_2(\boldsymbol{p}_j^*)}{f_1(\boldsymbol{p}_j^*)} \leq \chi$, thus decreasing χ accordingly. Conversely, we have $\frac{f_2(\boldsymbol{p}_j^*)}{f_1(\boldsymbol{p}_j^*)} \geq \chi$, and increase χ.

The quasi-convex optimization with relaxed ℓ_0 norm approximation for solving problem ($\mathcal{P}_2$) is summarized as Algorithm 1. Once the objective value realizes convergence, an exactly feasible solution $\boldsymbol{p}^*$ can be obtained.

C. Optimizing RIS Phase Shifts Given Sub-Band Assignment Strategy and Transmit Power

For given power allocation vector $\boldsymbol{p}$ and sub-band assignment strategy $\boldsymbol{A}$, problem ($\mathcal{P}$) can be rewritten as

$$(\mathcal{P}_3) \min_{\boldsymbol{\Theta}} \sum_{k=1}^{K} Q\left(\sqrt{\frac{2p_k\left\|\mathbf{G}_{k,m_k^*}\boldsymbol{\Theta}\mathbf{h}_{k,m_k^*}^r + \mathbf{h}_{k,m_k^*}^d\right\|^2}{\sigma^2}}\right) \tag{26a}$$

$$\textbf{s.t.} \ (16\text{e}), \tag{26b}$$

with $m_k^* \triangleq \left\{m \middle| [\boldsymbol{A}]_{(k,m)} = 1, k \in \mathcal{K}\right\}$ denoting the device mapping for each sub-band.

However, the objective function in problem ($\mathcal{P}_3$) is non-convex with respect to $\boldsymbol{\Theta}$, and the unit-modulus constraint (16e) is also intrinsically non-convex. Thence, we will convert ($\mathcal{P}_3$) into another solvable form inspired by [34]. Let $\mathbf{D}_k = \text{diag}\{\mathbf{h}_k^r\}$,[11] then $\mathbf{G}_k\boldsymbol{\Theta}\mathbf{h}_k^r = \mathbf{G}_k\mathbf{D}_k\boldsymbol{\phi} = \mathbf{A}_k\boldsymbol{\phi}$ with $\boldsymbol{\phi} = [\phi_1, \phi_2, \ldots, \phi_L]^T$. Define $Z = \left\|\mathbf{G}_k\boldsymbol{\Theta}\mathbf{h}_k^r + \mathbf{h}_k^d\right\|^2$. By replacing $\mathbf{G}_k\boldsymbol{\Theta}\mathbf{h}_k^r = \mathbf{A}_k\boldsymbol{\phi}$, we have

$$Z = \boldsymbol{\phi}^H\mathbf{A}_k^H\mathbf{A}_k\boldsymbol{\phi} + \boldsymbol{\phi}^H\mathbf{A}_k^H\mathbf{h}_k^d + \left(\mathbf{h}_k^d\right)^H\mathbf{A}_k\boldsymbol{\phi} + \left\|\mathbf{h}_k^d\right\|^2. \tag{27}$$

[11] Given sub-band assignment strategy $\boldsymbol{A}$, m_k^* is distinctly known, and thus $\mathbf{h}_{k,m_k^*}^r$ can be rewritten as $\mathbf{h}_k^r$ for ease of notation, whenever no confusion is incurred, and $\mathbf{h}_k^d$ and $\mathbf{G}_k$ are also derived from this predigestion.

Let $\hat{\boldsymbol{\phi}} = \left[\boldsymbol{\phi}^H, 1\right]$. Now we can rewrite Z as $Z = \hat{\boldsymbol{\phi}}^H \boldsymbol{\Lambda}_k \hat{\boldsymbol{\phi}} + \left\|\mathbf{h}_k^d\right\|^2 = \mathrm{tr}\left(\boldsymbol{\Lambda}_k \mathbf{X}\right) + \left\|\mathbf{h}_k^d\right\|^2$, with $\boldsymbol{\Lambda}_k = \begin{bmatrix} \mathbf{A}_k^H \mathbf{A}_k & \mathbf{A}_k^H \mathbf{h}_k^d \\ \left(\mathbf{h}_k^d\right)^H \mathbf{A}_k & 0 \end{bmatrix}$, and $\mathbf{X} \in \mathbb{C}^{(L+1)\times(L+1)}$, is defined as

$$\mathbf{X} = \hat{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}^H = \begin{bmatrix} \boldsymbol{\phi} \\ 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\phi}^H & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\phi}\boldsymbol{\phi}^H & \boldsymbol{\phi} \\ \boldsymbol{\phi}^H & 1 \end{bmatrix}. \tag{28}$$

Thus, problem $(\mathcal{P}_3)$ can be re-formulated as

$$(\mathcal{P}_{3.1}) \ \min_{\boldsymbol{\phi}} \sum_{k=1}^{K} Q\left(\sqrt{\frac{2p_k\left(\mathrm{tr}\left(\boldsymbol{\Lambda}_k \mathbf{X}\right) + \left\|\mathbf{h}_k^d\right\|^2\right)}{\sigma^2}}\right) \tag{29a}$$

$$\text{s.t. } (16\text{e}). \tag{29b}$$

Similar to Section IV-B, an upper bound of the Q-function can be introduced to replace the objective function in $(\mathcal{P}_{3.1})$ with $f_{2,k}(\mathbf{X}) = \vartheta\left(\sqrt{\frac{2p_k\left(\mathrm{tr}\left(\boldsymbol{\Lambda}_k \mathbf{X}\right) + \left\|\mathbf{h}_k^d\right\|^2\right)}{\sigma^2}}\right)$. However, the problem is still non-convex due to the modulus constraint (16e).

From $\boldsymbol{\phi}\boldsymbol{\phi}^H$ in Equation (28), we can realize that the diagonal entries in $\mathbf{X}$ embody the modulus of the elements in $\boldsymbol{\phi}$. This motivates us to define a simple matrix $\mathbf{E}$ with (i,j)-th entry being given by $[\mathbf{E}]_{(i,j)} = 1, \forall i = j$, and $[\mathbf{E}]_{(i,j)} = 0, \forall i \neq j$. As a result, problem $(\mathcal{P}_{3.1})$ is converted into

$$(\mathcal{P}_{3.2}) \ \min_{\mathbf{X}} \sum\nolimits_{k=1}^{K} f_{2,k}(\mathbf{X}) \tag{30a}$$

$$\text{s.t. } \mathbf{X} \geq 0, \mathrm{tr}\left(\mathbf{E}\mathbf{X}\right) = 1, \tag{30b}$$

$$\mathrm{rank}\left(\mathbf{X}\right) = 1, \tag{30c}$$

where the constraint in (30c) is responsible for strictly guaranteeing that 1) the resolved $\mathbf{X}$ can be decomposed into $\mathbf{X} = \hat{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}^H$, and 2) the solution of the phase shift in $\boldsymbol{\phi}$ in the resolved $\mathbf{X}$ is equivalent to the solution of the phase shift in $\boldsymbol{\Theta}$ in $(\mathcal{P}_3)$.

To further address the nonconvexity due to the constraint (30c) in problem $(\mathcal{P}_{3.2})$, the semidefinite relaxation (SDR) technique by simply dropping $\mathrm{rank}\left(\mathbf{X}\right) = 1$ is applicable to obtain a feasible solution $\mathbf{X}^*$. By this means, if $\mathbf{X}^*$ is rank-one, the optimal solution to the original problem $\boldsymbol{\phi}^*$ can be recovered by rank-one decomposition, and thus $\boldsymbol{\Theta}^* = \mathrm{diag}\{\boldsymbol{\phi}^*\}$ can be determined; otherwise, if $\mathbf{X}^*$ fails to be rank-one, additional step, e.g., Gaussian randomization [35] needs to be applied, thus extracting a suboptimal solution for the original problem. However, for the high-dimensional optimization problems, e.g., L is large enough, the probability of returning a rank-one solution $\mathbf{X}^*$ becomes low, which yields significant performance deterioration [36], [37], [38]. Thence, we present a DC programming approach [19], [37] to induce a rank-one solutions $\mathbf{X}^*$ for $(\mathcal{P}_{3.2})$, thus addressing the limitations of the SDR technique.

To begin with, a key intuition on the rank-one constraint needs to be revealed.

Lemma 4: For arbitrary positive semidefinite matrix $\mathbf{X} \in \mathbb{C}^{L\times L}$ with $\mathrm{tr}(\mathbf{X}) \geq 1$, we have $\mathrm{rank}\left(\mathbf{X}\right) = 1 \Leftrightarrow \mathrm{tr}\left(\mathbf{X}\right) - \left\|\mathbf{X}\right\|_2 = 0$ where trace norm $\mathrm{tr}\left(\mathbf{X}\right) = \sum_{i=1}^{L} \sigma_i\left(X\right)$ and spectral norm $\left\|\mathbf{X}\right\|_2 = \sigma_1\left(\mathbf{X}\right)$ with $\sigma_i\left(\mathbf{X}\right)$ denoting the i-th largest singular value of matrix $\mathbf{X}$.

Proof: See Proposition 3 in [39]. The complete proof is skipped due to space limitation. ■

By utilizing Lemma 4, we treat the DC function as a penalty component, which can be added to the objective function, instead of simply dropping the non-convex rank-one constraint via the SDR technique, thus enhancing a low-rank solution for problem $(\mathcal{P}_{3.2})$, yielding

$$(\mathcal{P}_{3.3}) \ \min_{\mathbf{X}} \sum\nolimits_{k=1}^{K} f_{2,k}(\mathbf{X}) + \rho\left(\mathrm{tr}\left(\mathbf{X}\right) - \left\|\mathbf{X}\right\|_2\right) \tag{31a}$$

$$\text{s.t. } (30\text{b}), \tag{31b}$$

with ρ denoting the penalty parameter. As the non-negative component $\mathrm{tr}\left(\mathbf{X}\right) - \left\|\mathbf{X}\right\|_2 \rightarrow 0$, an exact rank-one solution $\mathbf{X}^*$ can be obtained.

However, problem $(\mathcal{P}_{3.3})$ is still non-convex, due to the concave term $-\rho\left\|\mathbf{X}\right\|_2$. Fortunately, this term can be linearized by leveraging majorization-minimization techniques, yielding a DC algorithm [40], where problem $(\mathcal{P}_{3.3})$ can be transformed into a series of subproblems. At iteration υ, the subproblem can be established as

$$(\mathcal{P}_{3.4}) \ \min_{\mathbf{X}} \sum_{k=1}^{K} f_{2,k}(\mathbf{X}) + \rho\left\langle \mathbf{X}, \mathbf{I} - \partial\left\|\mathbf{X}^{\upsilon-1}\right\|_2 \right\rangle \tag{32a}$$

$$\text{s.t. } (30\text{b}), \tag{32b}$$

with $\mathbf{X}^{\upsilon-1}$ denoting the optimal solution of the subproblem at iteration $\upsilon - 1$, and $\langle \mathbf{X}, \mathbf{Y} \rangle = \mathcal{R}\{\mathbf{tr}\left(\mathbf{X}^H\mathbf{Y}\right)\}$ outputing the inner product of matrices $\mathbf{X}$ and $\mathbf{Y}$. The subgradient $\partial\left\|\mathbf{X}^{\upsilon-1}\right\|_2$ can be calculated efficiently as $\mathbf{v}\mathbf{v}^H$, with $\mathbf{v}$ representing the leading eigenvector of matrix $\mathbf{X}^{\upsilon-1}$ [39]. Obviously, problem $(\mathcal{P}_{3.4})$ is convex and can be solved efficiently by existing solvers. According to [40], the convergent critical points can be easily guaranteed by the DC algorithm from any feasible initial points, and thus a feasible solution $\mathbf{X}^*$ can be obtained. Then, the related $\boldsymbol{\phi}^*$ is computed by adopting the Cholesky decomposition, thus recovering the RIS configuration matrix by $\boldsymbol{\Theta}^* = \mathrm{diag}\{\boldsymbol{\phi}^*\}$.

D. Implementation and Complexity

Based on the above analysis, the overall AO algorithm for maximizing the learning performance in Theorem 2 by alternately optimizing the sub-band assignment strategy $\boldsymbol{A}$, the power allocation vector $\boldsymbol{p}$ and the RIS configuration matrix $\boldsymbol{\Theta}$ is summarized as Algorithm 2.[12]

The computational cost of the proposed AO algorithm is mainly derived from the step for solving problem $(\mathcal{P}_{1.2})$ by fixing RIS phase shifts $\boldsymbol{\Theta}$ and power allocation vector $\boldsymbol{p}$, solving problem $(\mathcal{P}_{2.3})$ by fixing RIS phase shifts $\boldsymbol{\Theta}$ and sub-band assignment strategy $\boldsymbol{A}$, plus the step for solving a sequence of the problems $(\mathcal{P}_{3.4})$ with fixed transmit power $\boldsymbol{p}$ and sub-band assignment strategy $\boldsymbol{A}$.

To solve $(\mathcal{P}_{1.2})$, according to [41], the worst-case computational complexity is $\mathcal{O}\left(\left(KM\right)^{3.5}\right)$ by adopting the interior

[12] Once the optimization strategy is executed entirely and converges to a critical point, a dedicated control channel will be adopted at the BS to feed back the optimal power allocation scheme and sub-band assignment strategy to the device in the form of control signaling. Thanks to the enough transmit power available at the BS, we suppose that the control signaling can be sent to the devices without error.

Algorithm 2 Alternating Optimization Algorithm for Solving Problem ($\mathcal{P}$).

```
Input: Initial the RIS configuration matrix $\boldsymbol{\Theta}^1$ and $\boldsymbol{p}^1$, halting criterion $\epsilon > 0$.
for each $t = 1, 2, \cdots, a^{\max}$ do
    Given $\boldsymbol{\Theta}^t$ and $\boldsymbol{p}^t$, solve problem ($\mathcal{P}_{1.2}$) to obtain the solution $\boldsymbol{A}^{t+1}$.
    Given $\boldsymbol{A}^{t+1}$ and $\boldsymbol{\Theta}^t$, solve problem ($\mathcal{P}_2$) to obtain the solution $\boldsymbol{p}^{t+1}$ using Algorithm 1.
    Given $\boldsymbol{A}^{t+1}$ and $\boldsymbol{p}^{t+1}$, solve problem ($\mathcal{P}_{3.4}$) to obtain the solution $\mathbf{X}^{t+1}$ by DC algorithm.
    Update $\boldsymbol{\phi}^{t+1}$ by Cholesky decomposition, and obtain $\boldsymbol{\Theta}^{t+1}$ using $\boldsymbol{\Theta}^{t+1} = \text{diag}\left\{\boldsymbol{\phi}^{t+1}\right\}$.
    if The decrease of the objective function of problem ($\mathcal{P}$) is below $\epsilon$ then
        break
```

point method. Assume the convergence requires $d^{\max}$ rounds for the SCA approximation, which is a finite number and not very large in practice. Thence, the total computational cost of optimizing the sub-band assignment strategy in each iteration is given by $\mathcal{O}\left(d^{\max}(KM)^{3.5}\right)$. To solve ($\mathcal{P}_{2.3}$), the worst-case computational complexity by using the interior point method is $\mathcal{O}\left(K^{3.5}\right)$ [36]. In bisection method, the length of the interval after τ iterations is $2^{-\tau}\left(u-l\right)$, where $u-l$ denotes the length of the initial interval defined in Algorithm 1. It follows $\lfloor\log_2\left((u-l)/\varepsilon\right)\rfloor$ iterations are required before the bisection method terminates. As a result, the computational cost of optimizing the power allocation vector in each alternating iteration can be expressed as $P_c = \mathcal{O}\left(i^{\max} j^{\max}\lfloor\log_2\left((u-l)/\varepsilon\right)\rfloor K^{3.5}\right)$. Simultaneously, when each ($\mathcal{P}_{3.4}$) is solved by the second-order interior point method [35], the worst-case complexity is bounded by $\mathcal{O}\left(\left(L^2\right)^{3.5}\right)$. Supposing those problems converging to critical points with $e^{\max} \geq 1$ iterations, during each alternate iteration, the computational cost of optimizing the RIS phase shifts in the worst case is $\mathcal{O}\left(e^{\max}\left(L^2\right)^{3.5}\right)$. As a result, the complexity of Algorithm 2 is upper bounded by $\mathcal{O}\left(a^{\max}\left(P_c + e^{\max}\left(L^2\right)^{3.5} + d^{\max}(KM)^{3.5}\right)\right)$.

V. Numerical Results and Analysis

In this section, numerical results are provided to examine the effectiveness of the proposed optimization algorithms.

A. Simulation Setup

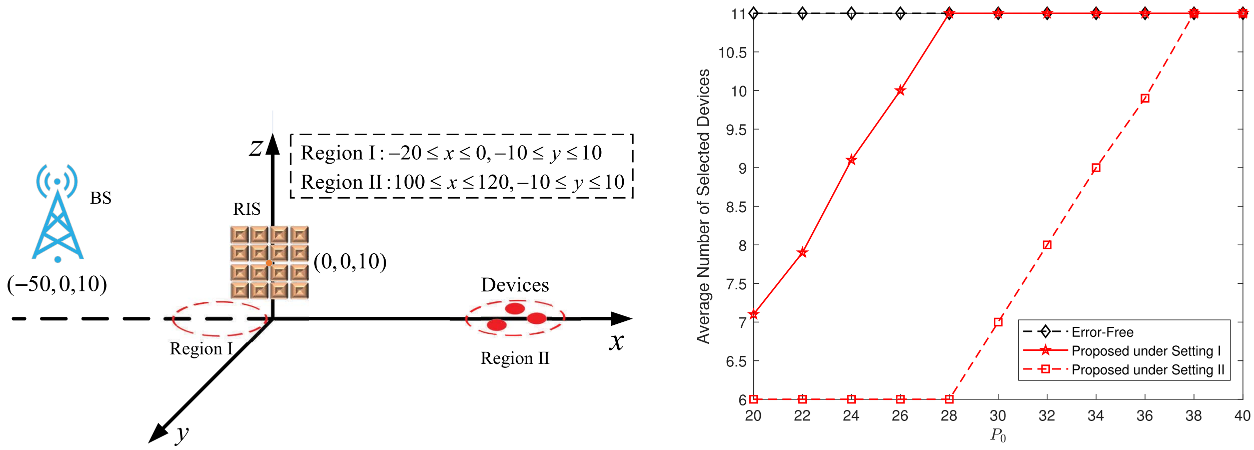
In our simulations, a three-dimensional Cartesian coordinate system as illustrated in Fig. 2(a) is conducted, where the BS and the RIS equipped with a uniform linear array are placed at $(50, 0, 10)$ and $(0, 0, 10)$, respectively. The edge devices are randomly and uniformly distributed in two circle regions with the radius of 10m, i.e., Region I $\propto$ $\{(x, y, 0) : -20 \leq x \leq 0, -10 \leq y \leq 10\}$ and Region II $\propto \{(x, y, 0) : 100 \leq x \leq 120, -10 \leq y \leq 10\}$. For simplicity, we assume that half of the K devices are randomly distributed in Region I, and the other half are randomly distributed in Region II.

The distances for the k-th direct device-BS link, the RIS-BS link, and the device-RIS link are denoted by d_{UB}^k d_{IB}^k and d_{UI}^k, respectively. Within each sub-band, the distance-dependent path loss for all channels is given by $L\left(d\right) = \zeta_0(d/d_0)^{-\alpha}$, with ζ_0 representing the path loss with respect to reference distance $d_0 = 1$ meter. The link distance is denoted by d, and α indicates the path loss exponent. All channels suffer from Rician fading [16] where the channel coefficient is expressed as $\boldsymbol{\iota} = \sqrt{\frac{\varsigma}{1+\varsigma}}\boldsymbol{\iota}_{Los} + \sqrt{\frac{1}{1+\varsigma}}\boldsymbol{\iota}_{NLos}$, with ς denoting the Rician factor. $\boldsymbol{\iota}_{Los}$ and $\boldsymbol{\iota}_{NLos}$ respectively represent the line-of-sight component and the non-line-of-sight component. Then, the corresponding channel coefficients involved can be given by $\mathbf{G}_k = \sqrt{L\left(d_{IB}^k\right)}\boldsymbol{\iota}_{IB}^k$, $\mathbf{h}_k^r = \sqrt{L\left(d_{UI}^k\right)}\boldsymbol{\iota}_{UI}^k$, $\mathbf{h}_k^d = \sqrt{L\left(d_{UB}^k\right)}\boldsymbol{\iota}_{UB}^k$, respectively. Following [42], we set the Rician factor of ι_{IB}^k, ι_{UI}^k and ι_{UB}^k to be 3 dB, 0 dB, and 0 dB. The path loss exponents for the direct device-BS channel, the RIS-BS channel, and the device-RIS channel are set to 4.8, 2.2, and 2.2, respectively. Unless stated otherwise, we set $P_0 = 30$dBm, $\sigma^2 = -50$dBm, $\zeta_0 = -30$dB, $J = 5$, $K = 11$, $S = 1$, $L = 40$, and $\epsilon = 1e-5$.

We consider the image classification task on the well-known FEMNIST dataset, which consists of 10 classes of apparel [43]. Specifically, we construct a 6-layer convolutional neural network (CNN), which consists of two 5×5 convolution layers, a 2×2 max pooling layer, followed by a batch normalization layer, a fully connected layer with 50 units, a ReLu activation layer, and a softmax output layer ($q = 21921$). The cross-entropy loss is defined, and the local training data of each edge device are drawn independently and identically from the training set of FEMNIST.

We compare the performance of our proposed algorithm with the following baseline schemes, thus verifying the effectiveness of our proposed system:

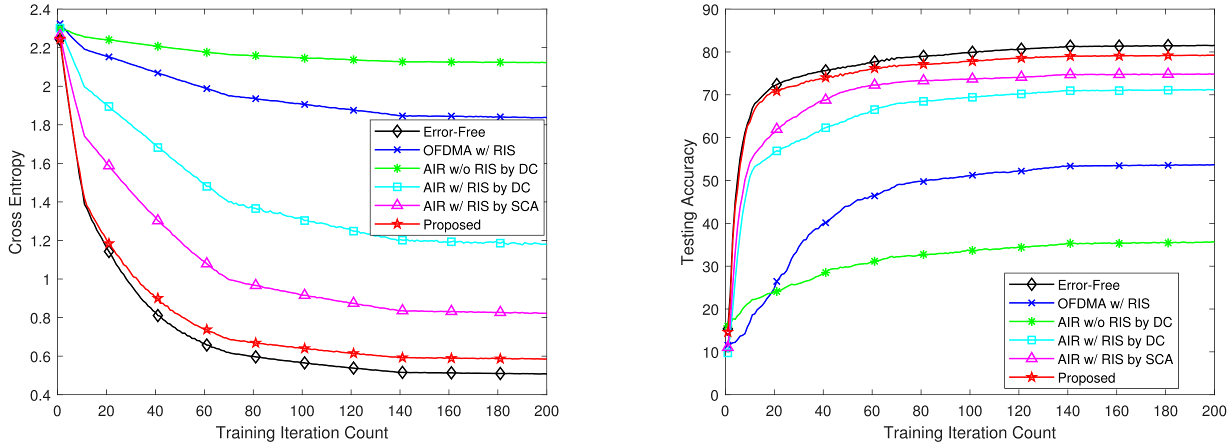
- **Joint optimization for RIS-enhanced OFDMA with RIS (OFDMA w/ RIS)** [22]: A RIS is employed to aid a multiuser OFDMA communication system, where the RIS configuration matrix, the OFDMA sub-band assignment, and the power allocation are jointly optimized for maximizing the common rate among all devices.
- **DC-based optimization without RIS (AIR w/o RIS by DC)** [39]: The RIS is not considered under Aircomp system, i.e., $\boldsymbol{\phi} = \mathbf{0}$. The device selection and the receiver beamforming are jointly optimized by the DC programming to maximize the number of active devices under certain mean-squared error (MSE) requirements φ.
- **DC-based alternating optimization with RIS (AIR w/ RIS by DC)** [19]: A RIS is deployed to assist AirComp-based FL system. The device selection, the receiver beamforming, and the configuration matrix at the RIS are jointly optimized using DC programming such that the number of active devices is maximized while the communication MSE is guaranteed.
- **SCA-based optimization with RIS (AIR w/ RIS by SCA)** [20]: A RIS is employed to assist the over-the-air model aggregation, where the active devices, the receiver beamforming, and the RIS configuration matrix are jointly optimized based on SCA approximation and Gibbs sampling.



(a) Simulation setup of our system.

(b) Average number of selected devices versus P_0.

Fig. 2. Simulation setup and the average number of selected devices versus P_0.



(a) Cross entropy as the number of iteration varies.

(b) Testing accuracy as the number of iteration varies.

Fig. 3. Performance of the proposed algorithm.

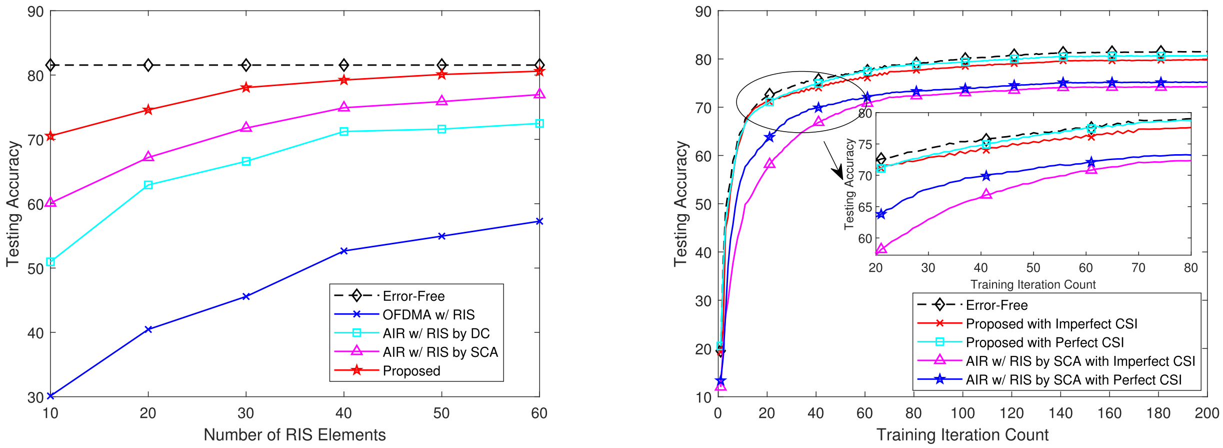
Additionally, the case where the channels are noiseless and all devices are selected at each learning iteration (**Error-Free**), serves for the comparison.

B. Performance on Device Selection

As illustrated in Fig. 2(b), we exhibit how the average number of selected devices varies with the maximum total power P_0. We consider the following two settings on the locations of the devices: 1) Setting I: the K devices are randomly and uniformly allocated in Region I; 2) Setting II: half of the K devices are randomly and uniformly distributed in Region I, and the other half are randomly distributed in Region II. It is observed that when P_0 is of lower magnitude, our proposed scheme tends to discard the partial devices with weak channels. A potential reason is that the available power resources are insufficient to support reliable transmission for all the devices. Thence, the limited power resources will be allocated to the devices with better channel conditions as much as possible to reduce the influence of the aggregation error caused by wireless channels. With the increase of P_0, the average number of selected devices of our method on Setting I becomes larger. This demonstrates that, with sufficient power resources, more edge devices will be encouraged to participate in the training process to eliminate the noise induced by data-stochasticity. Note that, under Setting II, when the increase of P_0 is not sufficient, the average number of selected devices may stay invariant. This is because devices under Setting II suffer from more wireless fading, and thus supporting the reliable transmission of additional devices requires a more pronounced power boost.

C. Performance on Training Convergence

In this section, we examine the convergence of our proposed algorithm in the image classification task described in Section V-A. Fig. 3(a) and Fig. 3(b) illustrate the results on cross entropy and testing accuracy of the proposed algorithm and above benchmarks versus the number of training iterations respectively. We can see that, the proposed method achieves a more excellent convergence performance than other benchmarks. Comparing the proposed method with AirComp-based system, e.g., (**AIR w/ RIS by SCA**) and (**AIR w/ RIS by DC**), the performance gap may be mainly



(a) Testing accuracy versus L.

(b) Testing accuracy of under imperfect CSI.

Fig. 4. Testing accuracy of the proposed algorithm under different number of RIS elements and imperfect CSI.

determined by the fact that the overall model aggregation error in Aircomp system is dominated by the devices with weak channels since the devices with better channel qualities have to lower their transmit power to align the local models at the BS. However, alignment operation in Aircomp will reduce the transmission power, thus resulting in less energy consumption. Besides, from the aspect of the spectral efficiency, the Aircomp-based algorithm can provide K_c times the spectral efficiency of our system due to the fact that the edge devices share the same sub-band. It is worth mentioning that there exists a tiny but unignorable gap compared to the error-free case due to the fact that the convergence rate depends on the level of wireless hostility and device selection loss, thus resulting in a deviated gradient for model updating. These observations are aligned with our analysis presented in Theorem 2.

D. Performance Versus The Number of RIS Elements

The effect of the number of reflecting elements at the RIS on the testing accuracy is illustrated in Fig. 4(a). Clearly, both methods can realize sizable performance promotion with a large L. This is because more precise passive reflective beamforming for the incident signals can be produced as the number of reflecting elements increases, thus suppressing the aggregation error effectively at the BS. However, the loss gap between the proposed method and other benchmarks still exists.

E. Performance on Imperfect Wireless Channels

As illustrated in Fig. 4(b), we consider the impact of imperfect CSI on the performance of the proposed method and (**AIR w/ RIS by SCA**). Both the direct channel and the cascade link from the device to BS via the RIS $\mathbf{C}_{k,m} = \mathbf{G}_{k,m}\mathrm{diag}\{\mathbf{h}^r_{k,m}\}$ are assumed to be imperfect [44]. Specifically, the direct channel is given by $\hat{\mathbf{h}}^d_{k,m} = \mathbf{h}^d_{k,m} + \Delta\mathbf{h}^d_{k,m}$, where $\mathbf{h}^d_{k,m}$ is the estimated channel known at the BS and $\Delta\mathbf{h}^d_{k,m}$ is the unknown channel error. Similarly, the cascade link can be represented as $\hat{\mathbf{C}}_{k,m} = \mathbf{C}_{k,m} + \Delta\mathbf{C}_{k,m}$, with $\mathbf{C}_{k,m}$ and $\Delta\mathbf{C}_{k,m}$ denoting the corresponding estimated channel and the estimation error, respectively. The CSI error vector $\Delta\mathbf{h}^d_{k,m}$ and $\mathrm{vec}(\Delta\mathbf{C}_{k,m})$ follow the circularly symmetric complex Gaussian distribution, i.e., $\Delta\mathbf{h}^d_{k,m} \sim \mathcal{CN}(\mathbf{0}, \varsigma^2_{k,m}\mathbf{I})$ and $\mathrm{vec}(\Delta\mathbf{C}_{k,m}) \sim \mathcal{CN}(\mathbf{0}, \varepsilon^2_{k,m}\mathbf{I})$, with $\varsigma^2_{k,m} = \sigma_h^2 \left\|\mathbf{h}^d_{k,m}\right\|_2^2$ and $\varepsilon^2_{k,m} = \sigma_c^2 \left\|\mathbf{C}_{k,m}\right\|_2^2$. $\sigma_h^2 \in [0,1)$ and $\sigma_c^2 \in [0,1)$ measure the amount of CSI uncertainties, which are set to 0.1. As it can be seen, the proposed method achieves stronger robustness than (**AIR w/ RIS by SCA**), although both approaches are robust against the imperfect CSI.

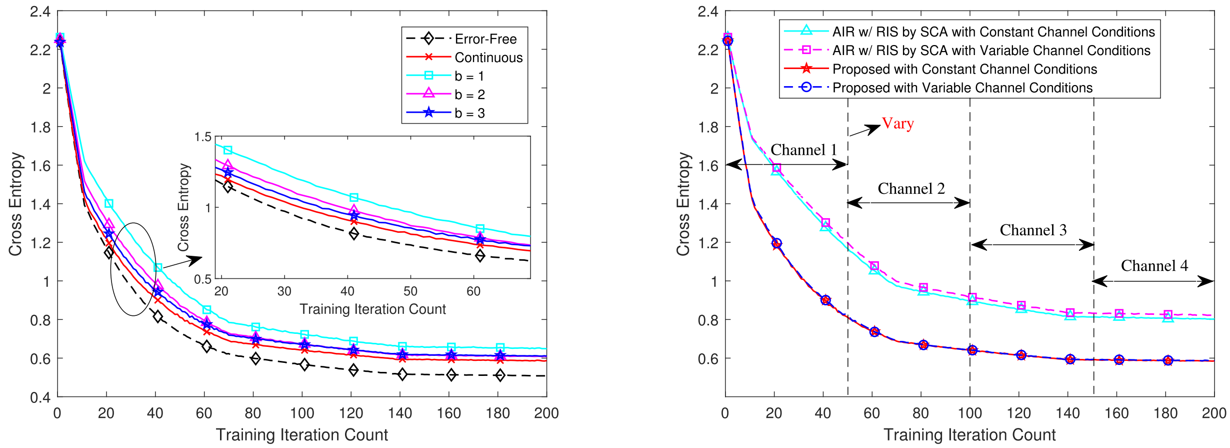
F. Performance on Discrete RIS Phase Shifts

As disclosed in [20] and [42], the continuous phase shift model $|\phi_l| = 1$ of RIS potentially incurs high implementation cost. Thence, we further suppose that only a finite number of discrete values can be taken for each phase shift element following [20] and [42], where $\phi_l \in \mathcal{F} = \left\{\exp\left(\frac{j2\pi m}{2^b}\right)\right\}_{m=0}^{2^b-1}$ with b phase resolution in number of bits [45]. $\mathcal{F}$ is the discrete feasible set of the reflection coefficient. When $b = \infty$, $\mathcal{F}$ becomes a continuous set, i.e., $\mathcal{F}_d = \{\phi : |\phi_l| = 1, \forall l\}$. As described in [20], the case with discrete RIS phase shifts can be accommodated by projecting the solution under the proposed algorithm to $\mathcal{F}_d$. The cross entropy of the proposed method under various choices of $b \in \{1, 2, 3\}$ is presented in Fig. 5(a). Under the case with $b = 1$, there exists 0.06 loss ascension gap compared with the continuous case, although this gap gradually becomes smaller when $b \in \{2, 3\}$. This is because low-bit phase shifts introduce an additional mismatch, thus reducing the training performance.

G. Performance on Time-Varying Channels

We further simulate the proposed method under the case where the small-scale fading channel coefficients of all the channels vary independently every 50 training iterations. In this circumstance, the optimal solution of the sub-band assignment matrix $\boldsymbol{A}$, the power allocation vector $\mathbf{p}$, and the RIS configuration matrix $\boldsymbol{\Theta}$ need to be updated when the channels change.[13] As illustrated in Fig. 5(b), we plot the cross entropy

[13] The device will send pilot symbols sequence to the BS at regular intervals, and the BS will estimate the channel according to the pilot symbols information [25], so as to check whether the channel of devices has changed.



(a) Cross entropy under discrete phase shift.

(b) Cross entropy under time-varying channels.

Fig. 5. Cross entropy of the proposed algorithm under discrete phase shift and time-varying Rician fading channels.

of the proposed method and (**AIR w/ RIS by SCA**) with time-varying channels. Obviously, both algorithms still achieve a more excellent convergence performance when the channels change, demonstrating the robustness of both algorithms under time-varying fading channels.

VI. Conclusion

In this paper, we proposed a RIS-aided SignSGD-based learning system to achieve an effective FEEL across wireless devices. We derived a closed-form expression of the convergence rate by characterizing the performance loss due to the communication noise, which is measured by the union bound on BER caused by the wireless channel. Based on the convergence analysis, a unified communication-learning optimization problem with respect to the sub-band assignment strategy, the power allocation vector, and the RIS configuration matrix was further formulated. To tackle the non-convex communication-learning optimization problem, we proposed an effective algorithm to decouple the tricky problem into several tractable subproblems by an alternating optimization method, where the sub-band assignment strategy is established by a penalty-based SCA method, and the power allocation vector together with the RIS configuration matrix are optimized by quasi-convex optimization with relaxed ℓ_0 approximation and the DC programming, respectively. Finally, some extensive numerical results demonstrate the convergence improvements of our proposed algorithm compared with the existing methods. This work represents a communication-learning tradeoff in the SignSGD-based FEEL over wireless channel, which is an initial attempt to theoretically prove that RIS can help the convergence of one-bit aggregation performance. For future work, more generalized settings such as heterogeneous data distribution will be taken into account, where the more complex convergence theory needs to be established, due to the uneven data-stochasticity of each edge device.

Appendix A
Proof of Lemma 1

To begin with, in every single step, the improvement of the objective based on Assumption 2 for one instantiation of the noise induced by data-stochasticity and communication error can be bounded. $g_i^{(n)} = \nabla F_i^{(n)}$ and $\bar{g}_i^{(n)}$ respectively denote the i-th component of the true gradient $\mathbf{g}^{(n)}$ and $\bar{\mathbf{g}}^{(n)}$. By substituting Equation (5) to Equation (8) and decomposing the improvement to reveal the error induced by stochasticity and communication, we can get

$$\begin{aligned} &F^{(n+1)} - F^{(n)} \\ &\quad \leq (\mathbf{g}^{(n)})^T(\mathbf{w}^{(n+1)} - \mathbf{w}^{(n)}) \\ &\qquad + \frac{1}{2}\sum\nolimits_{i=1}^{q} L_i \left(w_i^{(n+1)} - w_i^{(n)}\right)^2 \\ &\quad = -\eta(\mathbf{g}^{(n)})^T \text{sign}(\bar{\mathbf{g}}^{(n)}) + \eta^2 \sum\nolimits_{i=1}^{q} \frac{L_i}{2} \\ &\quad = -\eta\|\mathbf{g}^{(n)}\|_1 + \frac{\eta^2}{2}\|\mathbf{L}\|_1 + \\ &\qquad 2\eta \sum\nolimits_{i=1}^{q} |g_i^{(n)}| \mathbb{I}[\text{sign}(\bar{g}_i^{(n)}) \neq \text{sign}(g_i^{(n)})]. \end{aligned} \tag{33}$$

To find the expected improvement at iteration $n+1$ conditioned on the previous iterate, we have

$$\begin{aligned} \mathbb{E}[F^{(n+1)} - F^{(n)}|\mathbf{w}^{(n)}] &\leq -\eta\|\mathbf{g}^{(n)}\|_1 + \frac{\eta^2}{2}\|\mathbf{L}\|_1 \\ &+ 2\eta \sum_{i=1}^{q} |g_i^{(n)}| \mathbb{P}[\text{sign}(\bar{g}_i^{(n)}) \neq \text{sign}(g_i^{(n)})] \end{aligned} \tag{34}$$

with $\mathbf{g}^{(n)}$ keeping constant under the conditioning, which completes the proof. ■

Appendix B
Proof of Lemma 2

The main idea is to formulate an equivalent mathematical event, which can be described as a well-defined random variable with known distributions [6], [7], for $\mathbb{P}[\text{sign}(\sum_{k\in\mathcal{X}} \bar{g}_{k,i}^{(n)}) \neq \text{sign}(g_i^{(n)})]$. To this end, we define a random event for each device, as

$$X_k = \begin{cases} 1, & \text{with probability } \; p_i, \\ 0, & \text{with probability } \; q_i. \end{cases}, \tag{35}$$

with $p_i = \mathbb{P}[\text{sign}(\bar{g}_{k,i}^{(n)}) \neq \text{sign}(g_i^{(n)})]$ and $q_i = \mathbb{P}[\text{sign}(\bar{g}_{k,i}^{(n)}) = \text{sign}(g_i^{(n)})]$ respectively denoting the success probability and failure probability of Bernoulli trial X_k, and keeping constant among devices under the independent and identical setting. Let $Z = \sum_{k \in \lambdabar} X_k$, which represents the number of edge devices whose sign bit at the i-th entry is not equal to the sign of true gradient. Thence, Z is the sum of K_c independent Bernoulli trials, i.e., binomial with success probability p_i. Next, by computing the mean and variance of Z, we have $\mathbb{E}[Z] = K_c p_i = K_c\left(\frac{1}{2} - \epsilon_i\right)$, and $\mathbb{V}[Z] = K_c p_i q_i = K_c\left(\frac{1}{4} - \epsilon_i^2\right)$, with $\epsilon_i = \frac{1}{2} - p_i = q_i - \frac{1}{2}$ denoting the distance from p_i to $\frac{1}{2}$.

To determine $\mathbb{P}[\text{sign}(\sum_{k \in \lambdabar} \bar{g}_{k,i}^{(n)}) \neq \text{sign}(g_i^{(n)})]$, we must derive the probability that at least half of the K_c devices make a wrong approximation to the true gradient, i.e., $\mathbb{P}[Z \geq \frac{K_c}{2}]$. We have

$$\begin{aligned}
\mathbb{P}[Z \geq \frac{K_c}{2}] &= \mathbb{P}[Z - \mathbb{E}\left[Z\right] \geq \frac{K_c}{2} - \mathbb{E}\left[Z\right]] \\
&\overset{(a)}{\leq} \frac{\mathbb{V}\left[Z\right]}{\mathbb{V}\left[Z\right] + \left(\frac{K_c}{2} - \mathbb{E}\left[Z\right]\right)^2} \\
&= \frac{1}{1 + \frac{\left(\frac{K_c}{2} - \mathbb{E}[Z]\right)^2}{\mathbb{V}[Z]}} \overset{(b)}{\leq} \frac{1}{2}\frac{\sqrt{\mathbb{V}\left[Z\right]}}{\frac{K_c}{2} - \mathbb{E}\left[Z\right]} \\
&\overset{(c)}{=} \frac{1}{2}\sqrt{\frac{\frac{1}{4} - \varepsilon_i^2}{K_c \varepsilon_i^2}} = \frac{1}{2}\sqrt{\frac{1}{K_c}\left(\frac{1}{4\varepsilon_i^2} - 1\right)},
\end{aligned} \tag{36}$$

where (a) is from the Cantellis' inequality, i.e., $\mathbb{P}[X - \mathbb{E}[X] \geq \lambda] \leq \frac{\mathbb{V}[Z]}{\mathbb{V}[Z] + \lambda^2}$, $\lambda > 0$; (b) is from the fact that $1 + a^2 \geq 2a$; and (c) is derived by variable substitution using $\mathbb{E}[Z]$ and $\mathbb{V}[Z]$ calculated above. Now, it can be expected that next we need to derive an upper bound for $\frac{1}{4\varepsilon_i^2} - 1$.

Lemma 5: Suppose that the unimodal symmetric gradient noise is defined in Assumption 4, then we can derive $\frac{1}{4\varepsilon_i^2} - 1 \leq \frac{4}{\Sigma_i^2}$, with $\Sigma_i = \sqrt{n_b}\frac{|g_i^{(n)}|}{\sigma_i}$ denoting the gradient-signal-to-data-noise ratio.

Proof: Thanks to the unimodal symmetric gradient noise assumption stated in Assumption 4, the upper bound of probability p_i for the sign bit of a single device can be established. Specifically, recall the known Gauss' inequality [46] for a unimodal symmetric random variable X, with mean μ and expected squared deviation from the mean σ^2, then we have

$$\mathbb{P}[|X - \mu| > x] \leq \begin{cases} \dfrac{4}{9}\dfrac{\sigma^2}{x^2}, & \text{if } \frac{x}{\sigma} > \frac{2}{\sqrt{3}}, \\ 1 - \dfrac{x}{\sqrt{3}\sigma}, & \text{otherwise.} \end{cases} \tag{37}$$

Let g_i hold negative without loss of generality. By adopting Gauss's inequality, we get

$$\begin{aligned}
p_i &= \mathbb{P}\left[\text{sign}(\bar{g}_{k,i}^{(n)}) \neq \text{sign}(g_i^{(n)})\right] = \mathbb{P}\left[\bar{g}_{k,i}^{(n)} - g_i^{(n)} \geq |g_i^{(n)}|\right] \\
&= \frac{1}{2}\mathbb{P}\left[|\bar{g}_{k,i}^{(n)} - g_i^{(n)}| \geq |g_i^{(n)}|\right], \\
&\leq \begin{cases} \dfrac{2}{9}\dfrac{1}{\Sigma_i^2}, & \text{if } \Sigma_i > \frac{2}{\sqrt{3}}, \\ \dfrac{1}{2} - \dfrac{\Sigma_i}{2\sqrt{3}}, & \text{otherwise.} \end{cases} \quad \forall i,
\end{aligned} \tag{38}$$

with the term $\sqrt{n_b}$ being due to the fact that $\bar{g}_{k,i}^{(n)}$ is computed over a mini-batch of size n_b.

Well plugging the equation $\epsilon_i = \frac{1}{2} - p_i$ into Equation (38), we have

$$\epsilon_i \geq \begin{cases} \dfrac{1}{2} - \dfrac{2}{9}\dfrac{1}{\Sigma_i^2}, & \text{if } \Sigma_i > \frac{2}{\sqrt{3}}, \\ \dfrac{\Sigma_i}{2\sqrt{3}}, & \text{otherwise.} \end{cases} \quad \forall i. \tag{39}$$

For $\Sigma_i \leq \frac{2}{\sqrt{3}}$, $\frac{1}{4\varepsilon_i^2} - 1 \leq \frac{3}{\Sigma_i^2} - 1 < \frac{4}{\Sigma_i^2}$; For $\Sigma_i \geq \frac{2}{\sqrt{3}}$, $\frac{1}{4\varepsilon_i^2} - 1 \leq \frac{3}{\Sigma_i^2} - 1 \leq \frac{1}{\Sigma_i^2}\frac{\frac{8}{9} - \frac{16}{81}\frac{1}{\Sigma_i^2}}{1 - \frac{8}{9}\frac{1}{\Sigma_i^2} + \frac{16}{81}\frac{1}{\Sigma_i^4}} < \frac{1}{\Sigma_i^2}\frac{\frac{8}{9}}{1 - \frac{8}{9}\frac{1}{\Sigma_i^2}} < \frac{4}{\Sigma_i^2}$.

By plugging the upper bound in Lemma 5 into Equation (36), we can derive a tractable upper bound of $\mathbb{P}[\text{sign}(\sum_{k \in \lambdabar} \bar{g}_{k,i}^{(n)}) \neq \text{sign}(g_i^{(n)})]$, as $\mathbb{P}[Z \geq \frac{K_c}{2}] \leq \frac{1}{\Sigma_i \sqrt{K_c}}$, which completes the proof. ■

APPENDIX C
PROOF OF LEMMA 3

Following the proof of Lemma 2, transforming $\mathbb{P}[\text{sign}(\sum_{k \in \lambdabar} \bar{g}_{k,i}^{(n)}) \neq \text{sign}(g_i^{(n)})]$ into an equivalent mathematical event described by well-defined random variables with known distributions is an essential step. However, we cannot directly model the estimation of each device to the i-th entry of true gradient as an equal-probability Bernoulli trial like the proof of Lemma 2, due to the fact that the estimation is determined by the data-stochasticity and the heterogeneous communication conditions, which leads to a different probability distribution of each bernoulli trial intuitively.

Recall the form of bit error probability of each device $k \in \lambdabar$, i.e., $p_{k,i} = \mathbb{P}[\text{sign}(\bar{g}_{k,i}^{(n)}) \neq \text{sign}(g_i^{(n)})]$, which intuitively depends on the level of the noise introduced by the data-stochasticity and communication noise. To formalize the intuition, we further decouple the decoding bit error probability into a joint expression, as $p_{k,i} = \mathbb{P}_i^{sam}(1 - \mathbb{P}_k^{trans}) + \mathbb{P}_k^{trans}(1 - \mathbb{P}_i^{sam})$. $\mathbb{P}_i^{sam} = \mathbb{P}\left[\text{sign}(\tilde{g}_{k,i}^{(n)}) \neq \text{sign}(g_i^{(n)})\right] \leq \frac{1}{2}$ denotes the probability of the sign of a component of the stochastic gradient $\tilde{g}_{k,i}^{(n)}$ being incorrect compared with the true gradient $g_i^{(n)}$, which is equal across devices under the independent and identically case as suggested in [7], and can be effectively expressed by Equation (38). $\mathbb{P}_k^{trans} = \mathbb{P}\left[\text{sign}(\bar{g}_{k,i}^{(n)}) \neq \text{sign}(\tilde{g}_{k,i}^{(n)})\right]$ reveals the effect of the communication noise on the correct decoding of quantization gradient during data transmission, i.e., the BER, which can be written as Equation (7). Note that we rewrite $\mathbb{P}_{k,i}^{trans}$ as $\mathbb{P}_k^{trans}$ since the channel state remains constant when sending each bit.

By rearranging the expression, we have $p_{k,i} = \mathbb{P}_i^{sam}(1 - 2\mathbb{P}_k^{trans}) + \mathbb{P}_k^{trans}$. Define

$$X_k = \begin{cases} 1, & \text{with probability} \quad p_{k,i} \ \forall\, \boldsymbol{\Theta}, \boldsymbol{p}, \mathbf{A}, \\ 0, & \text{with probability} \quad q_{k,i} \ \forall\, \boldsymbol{\Theta}, \boldsymbol{p}, \mathbf{A}, \end{cases} \tag{40}$$

with $q_{k,i} = 1 - p_{k,i}$, and let $Z = \sum_{k \in \lambdabar} X_k$, which represents the number of edge devices with incorrect sign bit at the i-th entry of the true gradient. Thence, Z is the sum of K_c independent Bernoulli random variables, with respective success probabilities $p_{k,i}$, known commonly as the Poisson binomial distribution (PBD), which is well approximated by the binomial distribution, with a known bound [47] on the total variation distance from the PBD.

To find $\mathbb{P}[\text{sign}(\sum_{k\in\bar{\lambda}} \bar{g}_{k,i}^{(n)}) \neq \text{sign}(g_i^{(n)})]$, the probability upper bound for $\mathbb{P}[Z \geq \frac{K_c}{2}]$ should be derived. Note that $\mathbb{P}[Z \geq \frac{K_c}{2}] = 1-\mathbb{P}[Z \leq \frac{K_c}{2}]$, which means deriving the upper bound of $\mathbb{P}[Z \geq \frac{K_c}{2}]$ naturally equivalents to calculating the lower bound of $\mathbb{P}[Z \leq \frac{K_c}{2}]$. This encourages us to cleverly introduce Lemma 6 as follows to deal with the tricky PDB of event Z.

Lemma 6: Let $X \rightarrow \text{PB}(p_1,\ldots,p_n)$, $\bar{X} \rightarrow \text{Bin}(n,\bar{p})$, with PB and Bin denoting the PBD and the binomial distribution respectively, then we have $\mathbb{P}[X \leq k] \geq \mathbb{P}[\bar{X} \leq k]$ with $n\bar{p} \leq k \leq n$.

Proof: The original proof was brute-force by using the idea of majorization and Schur convexity, as detailed in [47]. The complete proof is skipped due to space limitation.

Next, let $X = Z$, $k = \frac{K_c}{2}$, and $\bar{p} = \frac{1}{K_c}\sum_{k\in\bar{\lambda}} p_{k,i}$. We have

$$\begin{aligned}\mathbb{P}\left(Z \geq \frac{K_c}{2}\right) &= 1-\mathbb{P}\left(Z \leq \frac{K_c}{2}\right)\\ &\leq 1-\mathbb{P}\left(\bar{Z} \leq \frac{K_c}{2}\right) = \mathbb{P}\left(\bar{Z} \geq \frac{K_c}{2}\right),\end{aligned} \tag{41}$$

which inspires us to derive the upper bound of $\mathbb{P}[\bar{Z} \geq \frac{K_c}{2}]$.

Define

$$\varepsilon_i = \frac{1}{2} - \underbrace{\frac{1}{K_c}\sum\nolimits_{k\in\bar{\lambda}} p_{k,i}}_{p_i=\bar{p}} = \underbrace{1-\frac{1}{K_c}\sum\nolimits_{k\in\bar{\lambda}} p_{k,i}}_{q_i} - \frac{1}{2}. \tag{42}$$

Plugging $p_{k,i}$ into ε_i, we get

$$\begin{aligned}\varepsilon_i &= \frac{1}{2} - \frac{1}{K_c}\sum\nolimits_{k\in\bar{\lambda}}\left(\mathbb{P}_i^{sam}\left(1-2\mathbb{P}_k^{trans}\right)+\mathbb{P}_k^{trans}\right)\\ &= \frac{1}{K_c}\sum\nolimits_{k\in\bar{\lambda}}\left(\frac{1}{2}\left(1-2\mathbb{P}_k^{trans}\right)-\mathbb{P}_i^{sam}\left(1-2\mathbb{P}_k^{trans}\right)\right)\\ &= \left(\frac{1}{2}-\mathbb{P}_i^{sam}\right)\frac{1}{K_c}\sum\nolimits_{k\in\bar{\lambda}}\Xi_k,\end{aligned} \tag{43}$$

with $\Xi_k = 1-2\mathbb{P}_k^{trans}$ in the last equation, and thus $\varepsilon_i^2 = (\frac{1}{2}-\mathbb{P}_i^{sam})^2(\frac{1}{K_c}\sum_{k\in\bar{\lambda}}\Xi_k)^2$. Now we begin to give a lower bound of $(\frac{1}{2}-\mathbb{P}_i^{sam})^2$. From Lemma 5, we have $\frac{1}{4(\frac{1}{2}-\mathbb{P}_i^{sam})^2} - 1 \leq \frac{4}{\Sigma_i^2}$, thus $(\frac{1}{2}-\mathbb{P}_i^{sam})^2 \geq \frac{\Sigma_i^2}{16+4\Sigma_i^2}$. Combining ε_i^2, Equation (36) and Equation (41), we have

$$\begin{aligned}&\mathbb{P}[Z \geq \frac{K_c}{2}]\\ &\leq \frac{\sqrt{K_c}}{\Sigma_i\sum_{k\in\bar{\lambda}}\Xi_k} + \frac{\sqrt{K_c}}{2\sum_{k\in\bar{\lambda}}\Xi_k}\sqrt{1-\left(\frac{1}{K_c}\sum\nolimits_{k\in\bar{\lambda}}\Xi_k\right)^2}\\ &\leq \frac{\sqrt{K_c}}{\Sigma_i\sum_{k\in\bar{\lambda}}\Xi_k} + \frac{\sqrt{K_c}}{2\sum_{k\in\bar{\lambda}}\Xi_k}\sqrt{1-\left(\frac{1}{\sqrt{K}\sqrt{K_c}}\sum\nolimits_{k\in\bar{\lambda}}\Xi_k\right)^2}\end{aligned} \tag{44}$$

which completes the proof. ■

Appendix D
Proof of Theorem 2

By plugging Lemma 3 into Equation (10), we have

$$\begin{aligned}&\mathbb{E}[F^{(n+1)}-F^{(n)}|\mathbf{w}^{(n)}] \leq -\eta\|\mathbf{g}^{(n)}\|_1 + \frac{\eta^2}{2}\|\mathbf{L}\|_1\\ &\quad + 2\eta\sum\nolimits_{i=1}^{q}\left|g_i^{(n)}\right|\mathbb{P}[\text{sign}(\sum\nolimits_{k\in\bar{\lambda}}\bar{g}_{k,i}^{(n)}) \neq \text{sign}(g_i^{(n)})]\\ &= \eta\left(\frac{\sqrt{K_c}}{\sum_{k\in\bar{\lambda}}\Xi_k}\sqrt{1-\left(\frac{1}{\sqrt{K}\sqrt{K_c}}\sum\nolimits_{k\in\bar{\lambda}}\Xi_k\right)^2}-1\right)\left\|g^{(n)}\right\|_1\\ &\quad + \frac{\eta^2}{2}\|\mathbf{L}\|_1 + \eta\frac{2\sqrt{K_c}\|\sigma\|_1}{\sqrt{n_b}\sum_{k\in\bar{\lambda}}\Xi_k},\end{aligned} \tag{45}$$

with Σ_i denoting the gradient-signal-to-data-noise ratio. Substituting the learning rate $\eta = \frac{1}{\sqrt{\|L\|_1 n_b}}$ and batch schedule $n_b = \frac{1}{\gamma}N$, we get

$$\begin{aligned}&\mathbb{E}[F^{(n+1)}-F^{(n)}|\mathbf{w}^{(n)}]\\ &\leq \frac{\sqrt{\gamma}\left\|g^{(n)}\right\|_1}{\sqrt{\|\mathbf{L}\|_1 N}}\left(\underbrace{\frac{\sqrt{K_c}}{\sum_{k\in\bar{\lambda}}\Xi_k}\sqrt{1-\left(\frac{1}{\sqrt{K}\sqrt{K_c}}\sum\nolimits_{k\in\bar{\lambda}}\Xi_k\right)^2}-1}_{G_a}\right)\\ &\quad + \frac{\gamma}{2N} + 2\frac{\gamma\sqrt{K_c}\|\sigma\|_1}{N\sqrt{\|\mathbf{L}\|_1}\sum_{k\in\bar{\lambda}}\Xi_k}.\end{aligned} \tag{46}$$

Further, by extending the expectation over the randomness in the optimization trajectory, and telescoping sum over the iterations, we have

$$\begin{aligned}&F^{(0)}-F^* \geq F^{(0)}-\mathbb{E}[F^{(n)}] = \mathbb{E}\left[\sum_{n=0}^{N-1}F^{(n)}-F^{(n+1)}\right]\\ &\geq \mathbb{E}\left[\sum_{n=0}^{N-1}\left[\frac{-\sqrt{\gamma}G_a}{\sqrt{\|\mathbf{L}\|_1 N}}\left\|g^{(n)}\right\|_1 - \frac{\gamma}{2N} - \frac{2\gamma\frac{\sqrt{K_c}}{\sum_{k\in\bar{\lambda}}\Xi_k}}{N\sqrt{\|\mathbf{L}\|_1}}\|\sigma\|_1\right]\right]\\ &= \sqrt{\frac{\gamma N}{\|\mathbf{L}\|_1}}\left(1-\frac{\sqrt{K_c}}{\sum_{k\in\bar{\lambda}}\Xi_k}\sqrt{1-\left(\frac{1}{\sqrt{K}\sqrt{K_c}}\sum\nolimits_{k\in\bar{\lambda}}\Xi_k\right)^2}\right) \qquad (47)\\ &\quad *E\left[\frac{1}{N}\sum_{n=0}^{N-1}\left\|g^{(n)}\right\|_1\right] - \frac{2\gamma}{\sqrt{\|\mathbf{L}\|_1}}\frac{\sqrt{K_c}}{\sum_{k\in\bar{\lambda}}\Xi_k}\|\sigma\|_1 - \frac{\gamma}{2}.\end{aligned} \tag{48}$$

Now rearranging the terms in Equation (48), yields the bound given by

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1}\|\mathbf{g}^{(n)}\|_1\right]\\ &\leq \frac{a_{\text{RIS}}}{\sqrt{N}}\left(\sqrt{\frac{\|\mathbf{L}\|_1}{\gamma}}\left(F^{(0)}-F^*+\frac{\gamma}{2}\right)\right. \qquad (49)\\ &\quad \left.+2b_{\text{RIS}}\sqrt{\gamma}\|\boldsymbol{\sigma}\|_1\right),\end{aligned} \tag{50}$$

with the scaling factor $a_{\text{RIS}} = \frac{1}{1-\frac{\sqrt{K}_c\sqrt{1-\left(\frac{1}{\sqrt{K}\sqrt{K_c}}\sum_{k\in\bar{\lambda}}\Xi_k\right)^2}}{\sum_{k\in\bar{\lambda}}\Xi_k}}$ and $b_{\text{RIS}} = \frac{\sqrt{K}_c}{\sum_{k\in\bar{\lambda}}\Xi_k}$, which completes the proof. ■

REFERENCES

[1] H. Li, R. Wang, W. Zhang, and J. Wu, "Federated edge learning via reconfigurable intelligent surface with one-bit quantization," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2022, pp. 1–6.

[2] M. Chen *et al.*, "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3579–3605, Dec. 2021.

[3] X. Lyu, C. Ren, W. Ni, H. Tian, R. P. Liu, and E. Dutkiewicz, "Optimal online data partitioning for geo-distributed machine learning in edge of wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2393–2406, Oct. 2019.

[4] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[5] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang, and M. Guizani, "Reliable federated learning for mobile networks," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 72–80, Feb. 2020.

[6] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 560–569.

[7] G. Zhu, Y. Du, D. Gündüz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2120–2135, Mar. 2021.

[8] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, Jan. 2019.

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

[10] M. M. Amiri, T. M. Duman, and D. Gündüz, "Collaborative machine learning at the wireless edge with blind transmitters," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2019, pp. 1–5.

[11] M. S. H. Abad, E. Ozfatura, D. Gündüz, and O. Ercetin, "Hierarchical federated learning ACROSS heterogeneous cellular networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 8866–8870.

[12] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, Oct. 2020.

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

[14] Y. He, J. Ren, G. Yu, and J. Yuan, "Importance-aware data selection and resource allocation in federated edge learning system," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13593–13605, Nov. 2020.

[15] Z. Yang, W. Xu, C. Huang, J. Shi, and M. Shikh-Bahaei, "Beamforming design for multiuser transmission through reconfigurable intelligent surface," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 589–601, Jan. 2021.

[16] 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.

[17] Z. Yang *et al.*, "Energy-efficient wireless communications with distributed reconfigurable intelligent surfaces," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 665–679, Jan. 2021.

[18] T. Zhang and S. Mao, "Energy-efficient federated learning with intelligent reflecting surface," *IEEE Trans. Green Commun. Netw.*, vol. 6, no. 2, pp. 845–858, Jun. 2022.

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

[20] H. Liu, X. Yuan, and Y.-J.-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.

[21] W. Ni, Y. Liu, Z. Yang, H. Tian, and X. Shen, "Integrating over-the-air federated learning and non-orthogonal multiple access: What role can RIS play?" *IEEE Trans. Wireless Commun.*, early access, Jun. 16, 2022, doi: 10.1109/TWC.2022.3181214.

[22] Y. Yang, S. Zhang, and R. Zhang, "IRS-enhanced OFDMA: Joint resource allocation and passive beamforming optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 760–764, Jun. 2020.

[23] J. Ye, S. Guo, and M.-S. Alouini, "Joint reflecting and precoding designs for SER minimization in reconfigurable intelligent surfaces assisted MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5561–5574, Aug. 2020.

[24] S. Hu, F. Rusek, and O. Edfors, "Beyond massive MIMO: The potential of data transmission with large intelligent surfaces," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2746–2758, May 2018.

[25] Q.-U.-U. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Intelligent reflecting surface assisted wireless communication: Modeling and channel estimation," 2019, *arXiv:1906.02360*.

[26] C. Liaskos *et al.*, "Joint compressed sensing and manipulation of wireless emissions with intelligent surfaces," in *Proc. 15th Int. Conf. Distrib. Comput. Sensor Syst. (DCOSS)*, May 2019, pp. 318–325.

[27] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, "Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3064–3076, May 2020.

[28] Z. Allen-Zhu, "Natasha 2: Faster non-convex optimization than SGD," 2017, *arXiv:1708.08694*.

[29] S. Lucidi and F. Rinaldi, "Exact penalty functions for nonlinear integer programming problems," *J. Optim. Theory Appl.*, vol. 145, no. 3, pp. 479–488, Jun. 2010.

[30] M. Chiani, D. Dardari, and M. K. Simon, "New exponential bounds and approximations for the computation of error probability in fading channels," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 840–845, Jul. 2003.

[31] T. Ma, Y. Xiao, X. Lei, P. Yang, X. Lei, and O. A. Dobre, "Large intelligent surface assisted wireless communications with spatial modulation and antenna selection," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2562–2574, Nov. 2020.

[32] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2909–2913, Jan. 2016.

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

[34] Z. Xing, R. Wang, J. Wu, and E. Liu, "Achievable rate analysis and phase shift optimization on intelligent reflecting surface with hardware impairments," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5514–5530, Sep. 2021.

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

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

[37] 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.

[38] E. Chen and M. Tao, "ADMM-based fast algorithm for multi-group multicast beamforming in large-scale wireless systems," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2685–2698, Jun. 2017.

[39] 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.

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

[41] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. Philadelphia, PA, USA: SIAM, 2001.

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

[43] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.

[44] G. Zhou, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "Outage constrained transmission design for IRS-aided communications with imperfect cascaded channels," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[45] C. Huang, G. C. Alexandropoulos, A. Zappone, M. Debbah, and C. Yuen, "Energy efficient multi-user MISO communication using low resolution large intelligent surfaces," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2018, pp. 1–6.

[46] C. F. Gauss, *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*, vol. 2, H. Dieterich, Ed. Charleston, SC, USA: Nabu Press, 1823.

[47] W. Hoeffding, "On the distribution of the number of successes in independent trials," *Ann. Math. Statist.*, vol. 27, no. 3, pp. 713–721, Sep. 1956.