Performance-Oriented Design for Intelligent Reflecting Surface-Assisted Federated Learning

Yapeng Zhao, *Graduate Student Member, IEEE*, Qingqing Wu, *Senior Member, IEEE*, Wen Chen, *Senior Member, IEEE*, Celimuge Wu, *Senior Member, IEEE*, and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*— To efficiently exploit the massive amounts of raw data that are increasingly being generated in mobile edge networks, federated learning (FL) has emerged as a promising distributed learning technique by collaboratively training a shared learning model on edge devices. The number of resource blocks when using traditional orthogonal transmission strategies for FL linearly scales with the number of participating devices, which conflicts with the scarcity of communication resources. To tackle this issue, over-the-air computation (AirComp) has emerged recently which leverages the inherent superposition property of wireless channels to perform *one-shot* model aggregation. However, the aggregation accuracy in AirComp suffers from the unfavorable wireless propagation environment. In this paper, we consider the use of intelligent reflecting surfaces (IRSs) to mitigate this problem and improve FL performance with AirComp. Specifically, a novel performance-oriented long-term design scheme that integrated design multiple communication rounds to minimize the optimality gap of the loss function is proposed. We first analyze the convergence behavior of the FL procedure with the absence of channel fading and noise. Based on the obtained optimality gap which characterizes the impact of channel fading and noise in different communication rounds on the ultimate performance of FL, we propose both online and offline schemes to tackle the resulting design problem. Simulation results demonstrate that such a long-term design strategy can achieve higher test accuracy than the conventional isolated design approach in FL. Both the theoretical analysis and numerical results exhibit a "later-is-better" principle, which demonstrates the later rounds in the FL procedure are more sensitive to aggregation error, and hence more resources are required over time.**

***Index Terms*— Intelligent reflecting surface, over-the-air computation, federated learning, Lyapunov framework, transceiver design, passive beamforming.**

Manuscript received 27 October 2022; revised 1 April 2023 and 22 May 2023; accepted 31 May 2023. Date of publication 7 June 2023; date of current version 18 September 2023. The work of Qingqing Wu was supported by Sichuan Science and Technology Program (Grant 2023YFH0092) and Guangdong Science and Technology Program (Grant 2022A0505050011). The work of Wen Chen was supported by National Key Project 2020YFB1807700, NSFC 62071296, Shanghai 22JC1404000, 20JC1416502, and PKX2021-D02. The work of Celimuge Wu was supported in part by the ROIS NII Open Collaborative Research 23S0601, and in part by JSPS KAKENHI grant number 21H03424. The work of H. Vincent Poor was supported by U.S. National Science Foundation under Grant CNS-2128448. The associate editor coordinating the review of this article and approving it for publication was J. Zhang. *(Corresponding author: Qingqing Wu.)*

Yapeng Zhao is with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau, China, and also with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: yc17435@connect.um.edu.mo).

Qingqing Wu and Wen Chen are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: qingqingwu@sjtu.edu.cn; wenchen@sjtu.edu.cn).

Celimuge Wu is with the Meta-Networking Research Center, The University of Electro-Communications, Chofu-shi, Tokyo 182-8585, Japan (e-mail: celimuge@uec.ac.jp).

H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2023.3283799.

Digital Object Identifier 10.1109/TCOMM.2023.3283799

I. Introduction

RECENT years have witnessed a significant increase in artificial intelligence (AI) applications, such as image recognition and natural language processing. Moreover, some delay-sensitive applications of AI like autonomous driving, require timely processing of real-time data. Conventional approaches require a data center to collect all the raw data for centralized model training. However, collecting massive amounts of data from distributed devices incurs high latency and very high energy and bandwidth cost [1], [2], [3]. Furthermore, the computation resources of edge devices are wasted in such a centralized model, and potential privacy violations exist since local data is collected and processed at the central server. This has thus motivated the migration of AI applications from the center of networks to the network edge, where implicit knowledge can be locally distilled to provide timely and economical intelligence as well as improved privacy. Chief among techniques for edge AI is federated learning (FL), in which local model parameters or gradients are exchanged instead of raw data [4], [5].

Since the training procedure in a state-of-the-art deep neural network (DNN) may involve millions of parameters [4], uplink communication overhead incurred during the iterative model update process becomes a critical bottleneck for FL given the limited communication bandwidth in practice. To address this issue, there has been considerable interest in communication-efficient uploading strategies for FL [6]. Conventionally, edge devices are orthogonally scheduled to upload their local model parameters/gradients, and the edge server sequentially decodes the received signals. However, the radio resources required in such orthogonal multiple access (OMA) protocols scale significantly with the number of participating devices and the model dimension. Therefore, different strategies have been proposed to alleviate this communication overhead, e.g., device sampling based on the network topology [7], model sparsification through the time correlations [8], and gradient quantization via lossy

0090-6778 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.

compression [9], [10]. However, all of these schemes sacrifice performance in order to deal with the limited communication resources. As an alternative, the recently aroused over-the-air computation (AirComp) technique has shown its efficiency in the *one-shot* aggregation of simultaneously transmitted local model parameters/gradients as compared to OMA protocols [11]. AirComp essentially turns the air into a computer by leveraging the inherent waveform superposition property of wireless channels. Particularly, it allows all the participating devices in FL to access all the radio resources simultaneously instead of only a fraction of them as in conventional OMA schemes. Moreover, its capability for *one-shot* function computation has been validated in a series of works, such as the foundational study from an information theoretic perspective [12], [13], performance analysis for various systems from the signal processing perspective [14], [15], and prototype validation in practical implementation [16].

AirComp empowered FL systems have been comprehensively investigated recently, including how to accelerate the convergence of FL under tight resources constraints (e.g., communication and computation resources) [17] or reduce resources consumption while maintaining satisfactory learning performance [18]. Specifically, there have been two main lines of research in existing works: one is system optimization from the communication perspective including device scheduling [19], transceiver design [20], and parameter/gradient compressing [21], the other is the inherent design of hyperparameters in the learning procedure, such as the learning rate [22] and batch size [23]. For example, the authors in [19] proposed a two-step framework for joint device scheduling and receive beamformer design to improve the test accuracy. The local learning rate at edge devices was optimized in [22] to combat the distortion induced by fading channels. Although AirComp has been envisioned and further validated to be a promising scalable model aggregation solution in FL, a crucial defect obstructs its implementation in practice. Specifically, the devices whose channels are deeply faded dominate the aggregation error in AirComp-based FL [14], [15]. This is because the *one-shot* aggregation demands that the signals transmitted by different devices be equally superimposed at the receiver, and thus the remaining devices with better channel conditions have to reduce their transmit power in order to perform signal alignment at the receiver. To this end, the strength of the obtained signal is diminished, and the aggregation accuracy is more sensitive to the inherent noise. Therefore, the unfavorable propagation environment inevitably limits the performance of AirComp-based FL. Fortunately, intelligent reflecting surfaces (IRSs), a promising technology for the beyond fifth-generation (B5G) and the future sixth-generation (6G) network, has shown its potential to overcome this detrimental effect [24], [25], [26]. By intelligently tuning signal reflections via a large number of low-cost passive reflecting elements, IRSs are capable of dynamically altering wireless channels to enable precise model aggregation, thus enhancing the FL performance. Recent works have demonstrated the effectiveness of IRSs in improving the performance of AirComp-based FL, e.g., weak channels were enhanced by IRSs to involve more devices in collaborative model training while sustaining precise aggregation thus accelerating the model convergence in [19], [27], and [28]. Besides, IRS was utilized to flexibly adjust the decoding order of heterogeneous data to serve both the FL and non-orthogonal multiple access (NOMA) users on the same time-frequency resource in [29] and minimize the energy consumption of engaged devices in [30].

However, the communication system design in the existing literature has not been tailored to the inherent characteristics of FL. The existing system design schemes in AirComp-based FL mainly focused on the isolated system design in each communication round [19], [28]. In fact, FL is a long-term process consisting of many progressive learning rounds that collaboratively determine the ultimate learning performance [31], [32]. Different learning rounds may have varying significance toward the convergence rate and the final model accuracy due to this intrinsic nature. Hence, resources need to be balanced among different iterations in FL by analyzing the collective impact of successive communication rounds on the ultimate performance. The isolated resource allocation strategy in the existing works equally treats each learning round, which inevitably results in performance loss.

In this paper, we consider an AirComp-based FL system, in which an IRS is employed to configure a favorable wireless channel to provide precise model aggregation in each iteration. For the first time, we propose a performance-oriented long-term design approach to fully unleash the available communication resources, thus obtaining higher test accuracy. The first step toward resource allocation across different learning rounds is to evaluate the impact of resources on the overall learning accuracy and convergence. To this end, we characterize the optimality gap of the loss function in arbitrary communication rounds, which unveils the relationship between communication accuracy in each communication round and ultimate learning performance. Accordingly, a long-term system design scheme is proposed to minimize the optimality gap via integrated design over multiple communication rounds. Different from existing works that focused on the isolated design in each communication round, it is interesting to find that such an integrated approach exhibits a *later-is-better* principle, which confirms that the later rounds in the FL procedure are more sensitive to the aggregation error, and hence more resources are required over time. The main contributions of this paper can be summarized as follows.

- We derive the general optimality gap of the loss function, which characterizes the impact of gradient aggregation errors in different communication rounds on the convergence performance of AirComp-based FL. From the obtained optimality gap, it is observed that within a finite number of communication rounds, the aggregation errors in later rounds have a greater impact on the optimality gap than those in earlier rounds. Hence, the later rounds are more essential to the learning performance of FL, and more resources need to be allocated to them. This observation provides important guidance for practical system design. However, the conventional isolated system design approach neglects

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.

the significant property and treats all the communication rounds equally.

- To enhance the ultimate performance of FL, we integrated design multiple communication rounds to minimize the optimality gap and establish the long-term system design schemes. We first propose an offline design approach in which the entire FL procedure is sequentially decomposed based on the available lookahead information, and then leverage the Lyapunov framework [33] to construct an online design problem over each communication round without foreseeing the future. The transceiver design and the IRS phase shift tuning are decoupled through the block coordinate descent (BCD) method. Furthermore, we propose an element-wise successive refinement algorithm with low complexity to guide the practical implementation of an IRS with discrete phase shift constraint.
- We conduct extensive simulations to evaluate the performance of the proposed performance-oriented design framework on the MNIST dataset [34] and the CIFAR-10 dataset [35] with convolution neural networks (CNNs). Simulation results show that the proposed long-term design schemes can achieve higher test accuracy than the conventional isolated design approach eventually, although it may lag behind in the beginning. Furthermore, it is observed that the online system design approach can acquire satisfactory performance compared to the offline solution without foreseeing the future, which further validates the *later-is-better* principle. Simulation results demonstrate that the proposed online scheme is a practical and promising approach for FL system design.

The remainder of this paper is organized as follows. In Section II, we describe the FL model and the IRS-assisted AirComp framework. In Section III, we analyze the FL performance and accordingly formulate the performance-oriented optimization problem. In Section IV, we harness an efficient BCD approach to alternatively optimize transceiver and IRS phase shifts. In Section V, we present extensive numerical results to evaluate the proposed algorithm. Finally, this paper concludes in Section VI.

Notation: Scalars are denoted by italic letters, and vectors and matrices are denoted by bold-face lower-case and uppercase letters, respectively. $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$ denote the space of $m\times n$ real-valued and complex-valued matrices, respectively. For a complex-valued vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|$ represents the Euclidean norm of $\boldsymbol{x}$, $\arg(\boldsymbol{x})$ denotes the phase of $\boldsymbol{x}$, and $\mathrm{diag}(\boldsymbol{x})$ denotes a diagonal matrix whose main diagonal elements are extracted from the vector $\boldsymbol{x}$. For a square matrix $\boldsymbol{S}$, $\mathrm{tr}(\boldsymbol{S})$ and $\boldsymbol{S}^{-1}$ denote its trace and inverse, respectively, while $\boldsymbol{S}\succeq\boldsymbol{0}$ means that $\boldsymbol{S}$ is positive semi-definite, where $\boldsymbol{0}$ is a zero matrix of proper size. For any general matrix $\boldsymbol{A}$, $\boldsymbol{A}^{\mathrm{H}}$, $\mathrm{rank}(\boldsymbol{A})$, and $\boldsymbol{A}_{i,j}$ denote its conjugate transpose, rank, and (i,j)th entry, respectively. $\boldsymbol{I}_M$ denotes an identity matrix of size $M\times M$. $\jmath$ denotes the imaginary unit, i.e., $\jmath^2=-1$. $\mathbb{E}[\cdot]$ denotes the statistical expectation. $|\cdot|$ denotes the cardinality of a given set. $\mathcal{O}(\cdot)$ is the big-O computational complexity notation.

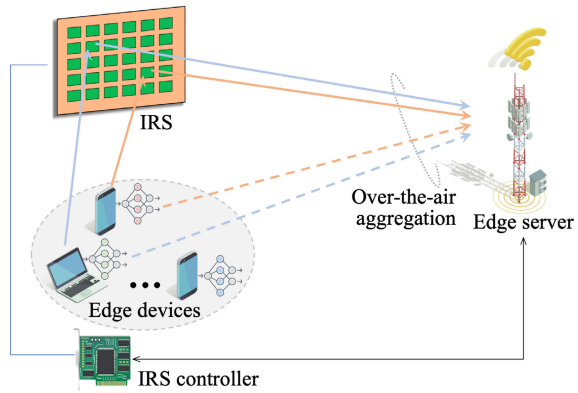


Fig. 1. An IRS-assisted AirComp-based FL system.

II. System Model

As depicted in Fig. 1, we consider an IRS-assisted FL system comprising of K single antenna devices, an IRS with N passive elements, and a base station (BS) with M antennas serving as an edge server. Devices and the IRS are coordinated by the BS to realize the collaborative model training through periodical communication and computation. The details of the FL procedure and the IRS-assisted over-the-air aggregation are introduced as follows.

A. Federated Learning Procedure

The goal of the FL system is to collectively acquire the desired parameter $\mathbf{w}^*\in\mathbb{R}^d$ that minimizes the global loss function with respect to the entire dataset. Instead of uploading all the raw data to the server, each device processes its local data in parallel and uploads the correspondingly generated local parameters or gradients to the server. We adopt the gradient sharing strategy in the FL procedure, i.e., the edge devices compute their local gradients and upload them to the edge server, and then the edge server broadcasts the global gradient which is computed based on the aggregated gradients to the edge devices for synchronizing local models' update.

Specifically, we assume that each edge device holds a subset of training data with $|\mathcal{D}_k|$ samples that are sampled i.i.d. from a source distribution $\mathcal{D}=\bigcup_{k\in\mathcal{K}}\mathcal{D}_k$, which is denoted by $\mathcal{D}_k=\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_{|\mathcal{D}_k|}\}$. Notation $\mathbf{u}_i$ represents a data sample, which can be represented as a set of feature vectors and its label in supervised learning. The regularized local loss function of device k is given by

$$F_k(\mathbf{w})=\frac{1}{|\mathcal{D}_k|}\sum_{\mathbf{u}_i\in\mathcal{D}_k}f(\mathbf{w};\mathbf{u}_i)+\lambda' R(\mathbf{w}), \tag{1}$$

where $R(\mathbf{w})$ is a strongly convex regularization function, $\lambda'\geq 0$ is a scaling hyperparameter. The global loss function evaluated at model parameter $\mathbf{w}$ is

$$F(\mathbf{w})=\frac{1}{D_{\mathrm{tot}}}\sum_{k=1}^{K}|\mathcal{D}_k|F_k(\mathbf{w}), \tag{2}$$

where $D_{\mathrm{tot}}=\sum_k|\mathcal{D}_k|$. This amounts to the regularized empirical average of the sample-wise loss functions on the global data set $\mathcal{D}$. The FL procedure aims to obtain the optimal

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.

$\mathbf{w}^*$ that minimizes the global loss function:

$$\mathbf{w}^* = \arg\min \ F(\mathbf{w}). \tag{3}$$

To perform cooperative training among edge devices, the edge devices compute their local gradients by minimizing $F_k(\mathbf{w})$ in parallel, and then the BS obtains the global gradient based on gathered local gradients. Specifically, based on a randomly sampled mini-batch $\hat{\mathcal{D}}_k$ from the local dataset, each device k computes a gradient vector $\mathbf{g}_k^{(t)}$ at communication round t as

$$\begin{aligned}\mathbf{g}_k^{(t)} &= \nabla F_k(\mathbf{w}^{(t)}) \\ &= \frac{1}{|\hat{\mathcal{D}}_k|}\sum_{\mathbf{u},v\in\hat{\mathcal{D}}_k} \nabla f(\mathbf{w};\boldsymbol{u},v) + \lambda\nabla R(\mathbf{w}^{(t)}),\end{aligned} \tag{4}$$

where $\mathbf{w}^{(t)}$ is the model parameter before updating. In this paper, we assume that all the mini-batch sizes are equal among devices.[1] Hence, the aggregated global gradient is

$$\mathbf{g}^{(t)} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{g}_k^{(t)}. \tag{5}$$

Then the BS broadcasts the global gradient to edge devices for synchronously local model updating. We assume that downlink communication is ideal, hence each device can obtain precise global gradient $\mathbf{g}^{(t)}$ for local model updating as

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \alpha^{(t)}\cdot\overline{\mathbf{g}}^{(t)}, \tag{6}$$

in which $\alpha^{(t)}$ denotes the learning rate in the communication round t.

B. Over-the-Air Aggregation

We adopt the AirComp technique to perform the communication-efficient gradient aggregation, and the IRS is deployed to assist the signal transmission from devices to the BS. Generally, the target function for aggregating the local gradient at the BS can be expressed as

$$\mathbf{g}^{(t)} = \chi^{(t)}(\varrho_k^{(t)}(\mathbf{g}_k^{(t)})), \tag{7}$$

where $\varrho_k^{(t)}(\cdot)$ denotes the pre-process function at device k for normalization in round t, and $\chi^{(t)}(\cdot)$ denotes the post-process function for de-normalization at the BS. Hence, the transmit sequence is $\mathbf{s}_k^{(t)} = \varrho_k^{(t)}(\mathbf{g}_k^{(t)})$. The baseband equivalent channels in communication round t from the IRS to BS, from the device k to IRS, and from the device k to BS are denoted by $\boldsymbol{G}^{(t)}\in\mathbb{C}^{M\times N}$, $\boldsymbol{h}_{r,k}^{(t)}\in\mathbb{C}^{N\times 1}$ and $\boldsymbol{h}_{d,k}^{(t)}\in\mathbb{C}^{M\times 1}$, respectively. The CSI in current communication round is assumed to be perfectly known based on the various channel acquisition methods discussed in [24]. Due to the severe path loss, the power of the signals that are reflected by the IRS two or more times is assumed to be negligible [36]. With the assistance of IRS, the signal received at the BS is given by

$$\boldsymbol{Y}^{(t)} = \sum_{k\in\mathcal{K}^{(t)}}\left(\boldsymbol{h}_{d,k}^{(t)} + \boldsymbol{G}^{(t)}\boldsymbol{\Theta}^{(t)}\boldsymbol{h}_{r,k}^{(t)}\right)b_k^{(t)}(\mathbf{s}_k^{(t)})^{\mathrm{T}} + \boldsymbol{Z}^{(t)}, \tag{8}$$

where $b_k^{(t)}\in\mathbb{C}$ is the transmit factor of device k that controls the power consumption in round t, and $\boldsymbol{Z}^{(t)}\in\mathbb{C}^{M\times d}$ with each entry $z_i \sim \mathcal{CN}(0,\sigma_z^2)$ denotes the additional white Gaussian noise at the receiver. Besides, $\boldsymbol{\Theta}^{(t)} = \mathrm{diag}(\beta_1^{(t)}e^{\jmath\theta_1^{(t)}},\ldots,\beta_N^{(t)}e^{\jmath\theta_N^{(t)}})$ denotes the diagonal phase shift matrix of the IRS, where $\theta_n^{(t)}\in[0,2\pi), \beta_n^{(t)}\in[0,1], \forall n\in\mathcal{N}, \forall t\in\mathcal{T}$ denote the phase shift and the amplitude reflection coefficient on the incident signal of element n. We assume $\beta_n^{(t)} = 1, \forall n\in\mathcal{N}, \forall t\in\mathcal{T}$ without loss of generality [36]. By applying receive beamformer $\boldsymbol{m}^{(t)}$, the obtained signal at BS is

$$\begin{aligned}&(\hat{\mathbf{s}}^{(t)})^{\mathrm{T}} \\ &= (\boldsymbol{m}^{(t)})^{\mathrm{H}}\left(\sum_{k\in\mathcal{K}}\tilde{\boldsymbol{h}}_k^{(t)}b_k^{(t)}(\mathbf{s}_k^{(t)})^{\mathrm{T}} + \boldsymbol{Z}^{(t)}\right) \\ &= \underbrace{\sum_{k\in\mathcal{K}}(\mathbf{s}_k^{(t)})^{\mathrm{T}}}_{\text{desired signal: }(\overline{\mathbf{s}}^{(t)})^{\mathrm{T}}} \\ &\quad + \underbrace{\sum_{k\in\mathcal{K}}\left((\boldsymbol{m}^{(t)})^{\mathrm{H}}\tilde{\boldsymbol{h}}_k^{(t)}b_k^{(t)} - 1\right)(\mathbf{s}_k^{(t)})^{\mathrm{T}} + (\boldsymbol{m}^{(t)})^{\mathrm{H}}\boldsymbol{Z}^{(t)}}_{\text{aggregation error: }(\boldsymbol{\varepsilon}_s^{(t)})^{\mathrm{T}}}.\end{aligned} \tag{9}$$

where $\tilde{\boldsymbol{h}}_k^{(t)} = \boldsymbol{h}_{d,k}^{(t)} + \boldsymbol{G}^{(t)}\boldsymbol{\Theta}^{(t)}\boldsymbol{h}_{r,k}^{(t)}$ denotes the superimposed channel from devices to BS, and $\boldsymbol{\varepsilon}_s^{(t)}$ is the aggregation error caused by the uplink transmission via AirComp. Hence, the corresponding estimated global gradient $\hat{\boldsymbol{g}}^{(t)}$ and the gradient distortion $\boldsymbol{\varepsilon}_g^{(t)}$ can be presented as

$$\begin{aligned}\hat{\mathbf{g}}^{(t)} &= \chi^{(t)}(\hat{\mathbf{s}}^{(t)}), \\ \boldsymbol{\varepsilon}_g^{(t)} &= \chi^{(t)}(\hat{\mathbf{s}}^{(t)}) - \mathbf{g}^{(t)} = \chi^{(t)}(\mathbf{s}^{(t)} + \boldsymbol{\varepsilon}_s^{(t)}) - \chi^{(t)}(\mathbf{s}^{(t)}).\end{aligned} \tag{10}$$

Hence, the local model updating at each device becomes

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \alpha^{(t)}\cdot\hat{\mathbf{g}}^{(t)} = \mathbf{w}^{(t)} - \alpha^{(t)}\cdot\left(\overline{\mathbf{g}}^{(t)} + \boldsymbol{\varepsilon}_g^{(t)}\right). \tag{11}$$

Notice that the aggregated gradient at BS in (10) is biased due to perturbation caused by the channel fading and noise. In the next section, we analyze the convergence behavior of the AirComp-based FL with perturbed gradients.

III. Convergence Analysis and Problem Formulation

In this section, we analyze the convergence behavior of FL. First, we introduce several assumptions on the loss function $F(\mathbf{w})$ and stochastic gradients to facilitate the convergence analysis, that are commonly made in the existing literature. Then, the generalized optimality gap, which is suitable for generic wireless networks, is derived to characterize the learning efficiency between two arbitrary communication rounds. The obtained optimality gap sheds light on how the imperfect gradient updates affect the convergence of FL. Next, we focus on the IRS-assisted AirComp to provide precise gradients aggregation for FL, and formulate the corresponding performance-oriented design problems.

[1]The utilized FL framework can be readily extended to the case that devices are equipped with different data sizes. In this case, (5) can be revised to a weighted-average form.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.

A. Convergence Analysis

In this subsection, we present several assumptions which have been widely used in the convergence analysis of FL, see e.g., [37], [38] and [39].

Assumption 1 (Smoothness): The global loss function $F(\mathbf{w})$ is smooth at any point $\mathbf{w} \in \mathbb{R}^d$ with constant $L > 0$, that is, it is continuously differentiable and the gradient $\nabla F(\mathbf{w})$ is Lipschitz continuous with constant L, i.e.,

$$\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\| \leq L\|\mathbf{w} - \mathbf{w}'\|, \quad \forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d. \tag{12}$$

Assumption 2 (Polyak-Łojasiewicz (PL) Condition [40]): The global loss function $F(\mathbf{w})$ satisfies the PL condition, i.e., for constant $\mu > 0$,

$$\left\|\nabla F(\mathbf{w}^{(t)})\right\|^2 \geq 2\mu(F(\mathbf{w}^{(t)}) - F(\mathbf{w}^*)), \tag{13}$$

where $\mathbf{w}^*$ is the global minimizer of the loss function.

Assumption 3 (Unbiased Estimation): The stochastic gradient evaluated on a mini-batch $\hat{\mathcal{D}}_k \subset \mathcal{D}$ and at any point $\mathbf{w}$ is an unbiased estimator of the partial full gradient, i.e. $\mathbb{E}[\mathbf{g}_k^{(t)}] = \nabla F(\mathbf{w}^{(t)})$.

As shown in the following theorem, the convergence behavior for generic wireless networks with perturbed gradients is proposed based on the above assumptions.

Theorem 1 (Generalized Optimality Gap): Suppose that an FL procedure satisfies Assumptions 1-3, and the learning rate $\alpha^{(t)} \equiv \alpha$ with $\alpha \leq \frac{1}{\mu}$ and $\alpha \leq \frac{1}{L}$,[2] for arbitrary T_2, T_1 satisfy $T_2 > T_1$, the optimality gap at the end of T_2-th round compared to the one of T_1-th is bounded by

$$\begin{aligned}
&\mathbb{E}\left[F\left(\mathbf{w}^{(T_2+1)}\right)\right] - F(\mathbf{w}^*) \\
&\quad \leq (1-\mu\alpha)^{(T_2-T_1)}\left(\mathbb{E}\left[F\left(\mathbf{w}^{(T_1+1)}\right)\right] - F(\mathbf{w}^*)\right) \\
&\qquad + \sum_{t=T_1+1}^{T_2}(1-\mu\alpha)^{(T_2-t)}\left\{\frac{\alpha(1-L\alpha)}{2}\left\|\mathbb{E}\left[\boldsymbol{\varepsilon}_g^{(t)}\right]\right\|^2\right. \\
&\qquad \left. + \frac{L\alpha^2}{2}\mathbb{E}\left[\left\|\boldsymbol{\varepsilon}_g^{(t)}\right\|^2\right] + \frac{L\alpha^2}{2}\mathbb{E}\left[\left\|\overline{\mathbf{g}}^{(t)}\right\|^2\right]\right\},
\end{aligned} \tag{14}$$

where $\boldsymbol{\varepsilon}_g^{(t)} = \chi^{(t)}(\mathbf{s}^{(t)} + \boldsymbol{\varepsilon}_s^{(t)}) - \chi^{(t)}(\mathbf{s}^{(t)})$ is the gradient distortion.

Proof: See Appendix A. ■

The presented optimality gap in Theorem 1 is shown as a weighted sum of distortions $\boldsymbol{\varepsilon}_g^{(t)}$ in each communication round. For arbitrary communication round t_1, t_2 satisfy $T_1 < t_1 < t_2 < T_2$, the time-related weights $(1-\mu\alpha)^{(T_2-t_1)} < (1-\mu\alpha)^{(T_2-t_2)}$ due to that the base $(1-\mu\alpha)$ belongs to $(0,1)$ and the exponents $T_2 - t_1 > T_2 - t_2 \geq 1$. Hence, more communication resources are required to diminish the aggregation error in the later rounds thus minimizing the optimality gap which exhibits the "*latter-is-better*" principle. Next, we formulate the offline and online problems based on Theorem 1.

[2]Note that the optimality gap with time-varying learning rate is also given in Appendix A, and the following problem formulation and proposed solution can be readily implemented on the one with varying $\alpha^{(t)}$.

B. Problem Formulation

The distortion caused by the channel fading and noise have a strong impact on the FL convergence, therefore, we utilize the IRS to configure favorable channels thereby reducing the aggregation error. As we focus on the ultimate performance of FL, the goal is to directly minimize the optimality gap in the final round with the maximum power constraint and long-term energy budget constraint of each device. Note that distortion $\boldsymbol{\varepsilon}_g^{(t)}$ in (14) is highly related to two aspects, one is the wireless channel, the other is normalization and de-normalization processes. Typically, the transmit symbols in AirComp are normalized into zero-mean unit-variance symbols based on current local gradient statistics [19], [27]. Although (14) can be readily transformed into the case by setting $T_2 = T$, and $T_1 = 0$, one critical issue appears that such a long-term system optimization requires foreseeing the future information, i.e., complete offline channel state information (CSI) and the normalization and de-normalization functions shown in (7) over the entire FL period (i.e., all the T learning rounds) which is impractical. In this part, we propose two schemes, one is the offline design with lookahead information served as the benchmark, and the other is the online design conducted on each communication round.

1) Offline Design With Lookahead Information: We divide the entire FL period into $R \geq 1$ periods, each comprising of $\rho \geq 1$ learning rounds such that $T = R\rho$. The offline framework is based on the ρ-rounds lookahead CSI and estimated ℓ_2-norm bound of local gradients in each round. We assume the CSI in the next ρ rounds is known in advance which is reasonable when the CSI during all the ρ-rounds remains static in a slow-fading channel, or the CSI can be precisely predicted via deep learning or other techniques [41]. In addition, the ℓ_2-norm bound of local gradients in the next ρ rounds are estimated by the heuristic methods provided in [42]. In that case, the normalization process at each device is conducted as

$$\mathbf{s}_k^{(t)} = \frac{\mathbf{g}_k^{(t)}}{\sqrt{\gamma^{(t)}}}, \quad \forall k, \tag{15}$$

where $\sqrt{\gamma^{(t)}}$ is the estimated ℓ_2-norm bound according to [42] and [43]. The de-normalization process at the BS is

$$\hat{\mathbf{g}}^{(t)} = \frac{\sqrt{\gamma^{(t)}}}{K}\hat{\mathbf{s}}^{(t)} = \overline{\mathbf{g}}^{(t)} + \frac{\sqrt{\gamma^{(t)}}}{K}\boldsymbol{\varepsilon}_s^{(t)}. \tag{16}$$

Hence, the corresponding gradient aggregation error is given by

$$\boldsymbol{\varepsilon}_g^{(t)} = \frac{\sqrt{\gamma^{(t)}}}{K}\boldsymbol{\varepsilon}_s^{(t)}, \tag{17}$$

and we have

$$\begin{aligned}
&\left\|\mathbb{E}\left[\boldsymbol{\varepsilon}_g^{(t)}\right]\right\|^2 \\
&\quad = \left|\frac{1}{K}\sum_{k\in\mathcal{K}}\left((\boldsymbol{m}^{(t)})^{\mathrm{H}}\tilde{\boldsymbol{h}}_k^{(t)}b_{f,k}^{(t)} - 1\right)\right|^2\|\nabla F(\mathbf{w}^{(t)})\|^2 \\
&\quad \leq \left|\frac{1}{K}\sum_{k\in\mathcal{K}}\left((\boldsymbol{m}^{(t)})^{\mathrm{H}}\tilde{\boldsymbol{h}}_k^{(t)}b_{f,k}^{(t)} - 1\right)\right|^2\gamma^{(t)},
\end{aligned} \tag{18}$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.

$$\mathbb{E}\left[\left\|\boldsymbol{\varepsilon}_g^{(t)}\right\|^2\right] \leq \left(\frac{1}{K^2}\sum_{k\in\mathcal{K}}\left|(\boldsymbol{m}^{(t)})^{\mathrm{H}}\tilde{\boldsymbol{h}}_k^{(t)}b_{f,k}^{(t)}-1\right|^2+\frac{1}{K^2}\|\boldsymbol{m}^{(t)}\|^2 d\sigma_z^2\right) \times \gamma^{(t)}, \tag{19}$$

where $b_{f,k}^{(t)}$ denotes the transmit factor of the offline scheme.

According to (14), the optimality gap at the end of period r compared to that of period $r-1$ is bounded as (20), shown at the bottom of the next page, where $\omega_1^{(t)} = (1-\mu\alpha)^{(T_2-t)}\frac{\alpha(1-L\alpha)}{2}$, and $\omega_2^{(t)} = (1-\mu\alpha)^{(T_2-t)}\frac{L\alpha^2}{2}$. Hence, the ultimate gap at round T (i.e., the end of period R) is bounded as

$$\mathbb{E}\left[F\left(\mathbf{w}^{(T+1)}\right)\right]-F(\mathbf{w}^*) \leq (1-\mu\alpha)^T\left(\mathbb{E}\left[F\left(\mathbf{w}^{(1)}\right)\right]-F(\mathbf{w}^*)\right) + \sum_{r=1}^{R}(1-\mu\alpha)^{(R-r)\rho}\Lambda^{(r)}(\{b_k^{(t)}\},\{\boldsymbol{\Theta}^{(t)}\},\{\boldsymbol{m}^{(t)}\}). \tag{21}$$

We consider the maximum transmit power on an individual symbol to be $P_k^{\max}$, thus each edge device is subject to a maximum power constraint $dP_k^{\max}$ on each communication round, i.e.,

$$\mathbb{E}\left(\left\||b_{f,k}^{(t)}|\mathbf{s}_k^{(t)}\right\|^2\right) \leq |b_{f,k}^{(t)}|^2 \leq dP_k^{\max}, \quad \forall k\in\mathcal{K},\ \forall t\in\mathcal{T}, \tag{22}$$

Besides, to reveal the communication resources allocation between different communication rounds, we further consider the average power constraint of each period which equivalently substitutes the long-term energy budget constraint:

$$\sum_{t=(r-1)\rho+1}^{r\rho}\mathbb{E}\left(\left\||b_{f,k}^{(t)}|\mathbf{s}_k^{(t)}\right\|^2\right) \leq \sum_{t=(r-1)\rho+1}^{r\rho}|b_{f,k}^{(t)}|^2 \leq \rho dP_k^{\mathrm{avg}}, \quad \forall k\in\mathcal{K}. \tag{23}$$

By discarding the constant term in (21), i.e., the initial optimality gap, the corresponding optimization problem[3] is given by

$$(\mathrm{P1}): \min_{\substack{\{b_{f,k}^{(t)}\},\\ \{\boldsymbol{\Theta}^{(t)}\},\{\boldsymbol{m}^{(t)}\}}} \sum_{r=1}^{R}(1-\mu\alpha)^{(R-r)\rho}\Lambda^{(r)}(\{b_{f,k}^{(t)}\}, \{\boldsymbol{\Theta}^{(t)}\},\{\boldsymbol{m}^{(t)}\}) \tag{24a}$$

$$\text{s.t. } |b_{f,k}^{(t)}|^2 \leq dP_k^{\max}, \quad \forall k\in\mathcal{K},\ \forall t\in\mathcal{T}, \tag{24b}$$

$$\sum_{t=(r-1)\rho+1}^{r\rho}|b_{f,k}^{(t)}|^2 \leq \rho dP_k^{\mathrm{avg}}, \quad \forall k\in\mathcal{K},\ \forall r\in[1,R], \tag{24c}$$

$$0 \leq \theta_n^{(t)} \leq 2\pi, \quad \forall n\in\{1,\ldots,N\},\ \forall t\in\mathcal{T}. \tag{24d}$$

[3]The formulated problem can be readily extended to the one including device scheduling, and two kinds of extensions are provided in Appendix B.

where the objective function is a weighted sum of the optimality gap in each communication round, and the latter rounds have larger weights. Furthermore, (P1) can be divided into R subproblems since both the objective function and the constraints are independent among different periods. Each subproblem designs the resource allocation in one period in which the communication resources can be flexibly allocated. In practice, the CSI and the estimated gradient bounds may have errors, thereby the energy consumption needs to be corrected at the end of each period r. Note that the presented offline design scheme is served as a benchmark for the online scheme shown below.

2) Online Design via Lyapunov Technique: (P1) needs the ρ-rounds lookahead CSI as well as the estimated gradient norm bound to allocate the communication resources. However, the CSI varies rapidly within a period when facing a fast-varying channel, and the channel prediction techniques may induce additional errors. In addition, the estimated ℓ_2-norm bound of local gradients may induce additional perturbation to the obtained optimality gap. Hence, we utilize the Lyapunov optimization framework [33] to construct a virtual energy deficit queue $e_k(t)$ for each device k to guide the power allocation over sequential communication rounds without foreseeing the future. According to [19] and [27], the normalization process at each device is conducted as

$$\mathbf{s}_k^{(t)} = \frac{\mathbf{g}_k^{(t)} - \frac{1}{K}\sum_{k\in\mathcal{K}}\xi_k^{(t)}}{\frac{1}{K}\sqrt{\sum_{k\in\mathcal{K}}(\iota_k^{(t)})^2}}, \quad \forall k, \tag{25}$$

where $\xi_k^{(t)}$, $(\iota_k^{(t)})^2$ is the local gradient statistics (i.e., the means and variances) in current round t that are computed as

$$\xi_k^{(t)} = \frac{1}{D}\sum_{d=1}^{D}g_{k,d}^{(t)}, \quad (\iota_k^{(t)})^2 = \frac{1}{D}\sum_{d=1}^{D}(g_{k,d}^{(t)}-\xi_k^{(t)})^2, \quad \forall k. \tag{26}$$

The de-normalization process at the BS is

$$\hat{\mathbf{g}}^{(t)} = \frac{1}{K}\left(\frac{\sqrt{\sum_{k\in\mathcal{K}}(\iota_k^{(t)})^2}}{K}\hat{\mathbf{s}}^{(t)} + \sum_{k\in\mathcal{K}}\xi_k^{(t)}\right), \tag{27}$$

thereby, the corresponding gradient aggregation error is

$$\boldsymbol{\varepsilon}_g^{(t)} = \hat{\mathbf{g}}^{(t)} - \overline{\mathbf{g}}^{(t)} = \frac{\sqrt{\sum_{k\in\mathcal{K}}(\iota_k^{(t)})^2}}{K^2}\boldsymbol{\varepsilon}_s^{(t)}, \tag{28}$$

and we have

$$\left\|\mathbb{E}\left[\boldsymbol{\varepsilon}_g^{(t)}\right]\right\|^2 = 0,$$
$$\mathbb{E}\left[\left\|\boldsymbol{\varepsilon}_g^{(t)}\right\|^2\right] \leq \frac{d\sum_{k\in\mathcal{K}}(\iota_k^{(t)})^2}{K^4}\left(\sum_{k\in\mathcal{K}}\left|(\boldsymbol{m}^{(t)})^{\mathrm{H}}\tilde{\boldsymbol{h}}_k^{(t)}b_k^{(t)}-1\right|^2 + \|\boldsymbol{m}^{(t)}\|^2\sigma_z^2\right). \tag{29}$$

Consider the maximum transmit power on an individual symbol to be $P_k^{\max}$ as the offline scheme, thereby the maximum power constraint of the online scheme is given by

$$|b_k^{(t)}|^2 \leq P_k^{\max}, \quad \forall k\in\mathcal{K},\ \forall t\in\mathcal{T}. \tag{30}$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.

The virtual energy queue of device k starts with $e_k(t) \geq 0, \forall k$, and is updated at the end of each round t as

$$e_k(t+1) = \max\{e_k(t) + d|b_k^{(t)}|^2 - dP_k^{\text{avg}}, 0\}, \tag{31}$$

where $d|b_k^{(t)}|^2$ is the power consumption on communication round t. Hence, $e_k(t)$ indicates the deviation of the current energy consumption of device k from its long-term energy constraint $\rho d P_k^{\text{avg}}$. Let $\boldsymbol{e}(t) = \{e_1(t), e_2(t), \ldots, e_K(t)\}$ collect the energy deficit queues for all devices. The constructed optimization problem of communication round t is shown as a weighted sum of the per-round optimality gap and the energy consumption at the devices, i.e.,

$$(\text{P2}):$$

$$\min_{\substack{\{b_k^{(t)}\},\{\boldsymbol{\Theta}^{(t)}\},\\ \{\boldsymbol{m}^{(t)}\}}} V_r\,(1-\mu\alpha)^{(R-r)\rho}\,\omega_2^{(t)}\mathbb{E}\left[\left\|\boldsymbol{\varepsilon}_g^{(t)}\right\|^2\right] + \sum_{k\in\mathcal{K}} d e_k(t)|b_k^{(t)}|^2 \tag{32a}$$

$$\text{s.t.}\quad |b_k^{(t)}|^2 \leq P_k^{\max}, \quad \forall k \in \mathcal{K},\ \forall t \in \mathcal{T}, \tag{32b}$$

$$0 \leq \theta_n^{(t)} \leq 2\pi, \quad \forall n \in \{1,\ldots,N\},\ \forall t \in \mathcal{T}, \tag{32c}$$

where parameter $V_r \geq 0$ is represented as an importance weight to adjust the emphasis on the objective function (i.e., optimality gap minimization) in different communication rounds. Notice that if $e_k(t)$ increases in round t, then minimizing energy consumption is more critical in round $t+1$. Intuitively, when the energy queue is stable, the constraint in (24c) is satisfied. Although problems (P2) and (P1) are not equivalent due to the different objective functions, constraints, and even different normalization/de-normalization processes, the following theorem unveils that the results of (P2) are comparable to (P1) and are within a bounded deviation from the optimal results of the omniscient (P1) with precise future information.

Theorem 2: Assume that the omniscient (P1) is an offline scheme that obtains precise CSI and local gradient statistics at the beginning of each period and conducts the normalization/de-normalization as the one in the online scheme. (P2) can achieve a performance-backlog tradeoff of $[\mathcal{O}(1/V), \mathcal{O}(\sqrt{V})]$ with compared to omniscient (P1). We use G^* to represent the obtained optimality gap via the optimal offline solution and $G^\dagger$ to represent obtained the optimality gap via (P2). Specifically, the FL performance is $\mathcal{O}(1/V)$-optimal, which is bounded by

$$\sum_{t=1}^{T} G_t^\dagger \leq \sum_{t=1}^{T} G_t^* + \sum_{r=1}^{R-1} \frac{C_r}{V_r}, \tag{33}$$

where $C_r = \rho C_e + \sum_{t=r\rho+1}^{(r+1)\rho}\sum_{k\in\mathcal{K}} E_{\max}\big((t-1)E_{\max} + e_k(r\rho+1)\big)$, and $E_{\max} = \max_{k,t}(d|b_k^{(t)}|^2 - dP_k^{\text{avg}})$. Besides, the energy consumption of each device is $\mathcal{O}(\sqrt{V})$-bounded as

$$\sum_{t=1}^{T}\left(d|b_k^{(t)}|^2 - dP_k^{\text{avg}}\right) \leq \sum_{r=1}^{R-1}\left(\sqrt{2\left(C_r + V_r\sum_{t=r\rho+1}^{(r+1)\rho} G_t^*\right)} - e_k(r\rho+1)\right). \tag{34}$$

Proof: See Appendix C. ■

IV. Proposed Algorithms for Offline and Online Design

In the previous section, we present an offline optimization problem based on the available lookahead information and an online optimization problem via the Lyapunov technique without foreseeing the future. It is observed that (P1) can be divided into R subproblems, each of which corresponds to one period, and (P2) is an online design problem within each communication round. Note that although variables are highly coupled in the objective function, all the constraints corresponding to each variable are uncoupled with others in both problems. This thus motivates us to apply the BCD method to solve them efficiently by properly partitioning the optimization variables into different blocks. One can observe that, with any given transmit factor, the optimization procedures for receiver $\boldsymbol{m}^{(t)}$ and IRS phase shift $\boldsymbol{\Theta}^{(t)}$ of (P2) is a simplification of (P1) with drop the term related to $\left\|\mathbb{E}\left[\boldsymbol{\varepsilon}_g^{(t)}\right]\right\|^2$. Besides, there exist similarities in the optimization procedures for power allocation of (P1) and (P2). Therefore, we propose the optimization algorithms for the two problems from the perspective of different variable blocks.

A. Power Allocation

For any given $\{\boldsymbol{\Theta}^{(t)}\}$ and $\{\boldsymbol{m}^{(t)}\}$, the transmit factors need to be precisely allocated among different communication rounds to minimize the ultimate optimality gap. Recalling (P1) and (P2), we can obtain the following remark.

Remark 1: To minimize the objective functions of (P1) and (P2), each device k in communication round t always adjusts

$$\mathbb{E}\left[F\left(\mathbf{w}^{(r\rho+1)}\right)\right] - F(\mathbf{w}^*) \leq (1-\mu\alpha)^{\rho}\underbrace{\left(\mathbb{E}\left[F(\mathbf{w}^{((r-1)*\rho+1)})\right] - F(\mathbf{w}^*)\right)}_{\text{Optimality gap in the end of period } r-1} + \underbrace{\sum_{t=(r-1)\rho+1}^{r\rho}\omega_1^{(t)}\left\|\mathbb{E}\left[\boldsymbol{\varepsilon}_g^{(t)}\right]\right\|^2 + \sum_{t=(r-1)\rho+1}^{r\rho}\omega_2^{(t)}\left(\mathbb{E}\left[\left\|\boldsymbol{\varepsilon}_g^{(t)}\right\|^2\right] + \mathbb{E}\left[\left\|\overline{\mathbf{g}}^{(t)}\right\|^2\right]\right)}_{\Lambda^{(r)}(\{b_k^{(t)}\},\{\boldsymbol{\Theta}^{(t)}\},\{\boldsymbol{m}^{(t)}\})}, \tag{20}$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.

the phase of its transmit factor for phase compensation with respect to its equivalent fading channel $\bar{h}_k^{(t)} = (\boldsymbol{m}^{(t)})^{\mathrm{H}}\tilde{\boldsymbol{h}}_k^{(t)}$, i.e., $\arg(b_{f,k}^{(t)}) = \arg(b_k^{(t)}) = \arg\big((\bar{h}_k^{(t)})^{-1}\big)$.

In that case, only the magnitudes of $\{b_{f,k}^{(t)}\}$ have effects on the value of the objective function. By fixing $\arg(b_{f,k}^{(t)}) = \arg\big((\bar{h}_k^{(t)})^{-1}\big)$, and letting $\bar{b}_{f,k}^{(t)} = |b_{f,k}^{(t)}|, \forall k \in \mathcal{K}$, the power allocation problem of (P1) over one dedicated period can be expressed in the real form as

$$(\mathrm{P1.1}): \min_{\{\bar{b}_{f,k}^{(t)}\}} \sum_{t=(r-1)\rho+1}^{r\rho} \frac{\gamma^{(t)}\omega_1^{(t)}}{K^2}\left(\sum_{k\in\mathcal{K}} |\bar{h}_k^{(t)}|\bar{b}_{f,k}^{(t)} - K\right)^2 + \sum_{t=(r-1)\rho}^{r\rho} \frac{\gamma^{(t)}\omega_2^{(t)}}{K^2}\sum_{k\in\mathcal{K}}\left(|\bar{h}_k^{(t)}|\bar{b}_{f,k}^{(t)} - 1\right)^2 \tag{35a}$$

$$\text{s.t. } (\bar{b}_{f,k}^{(t)})^2 \le dP_k^{\max}, \quad \forall t \in [(r-1)\rho+1, r\rho], \quad \forall k \in \mathcal{K}, \tag{35b}$$

$$\sum_{t=(r-1)\rho+1}^{r\rho} (\bar{b}_{f,k}^{(t)})^2 \le \rho d P_k^{\mathrm{avg}}, \quad \forall k \in \mathcal{K}. \tag{35c}$$

Note that (P1.1) is a convex quadratic optimization problem which can be optimally solved by standard convex optimization techniques such as the interior point method or alternating direction method of multipliers (ADMM). Instead, we next resort to the Lagrange duality method to derive the structured optimal solution for problem (P1.1) to gain engineering insights. Let λ_k^* denote the optimal dual variable associated with the k-th constraint in (35c). By adopting the first-order optimality condition, the optimal solution of $\bar{b}_k^{(t)}$ is given by

$$(\bar{b}_{f,k}^{(t)})^* = \min\left\{\frac{\gamma^{(t)}\omega_1^{(t)}(K - \sum_{i\neq k}|\bar{h}_i^{(t)}|\bar{b}_i^{(t)}) + \gamma^{(t)}\omega_2^{(t)}}{\big(\gamma^{(t)}\omega_1^{(t)} + \gamma^{(t)}\omega_2^{(t)}\big)|\bar{h}_k^{(t)}| + \frac{K^2}{|\bar{h}_k^{(t)}|}\lambda_k^*}, \sqrt{dP_k^{\max}}\right\}, \tag{36}$$

where the optimal dual variables $\{\lambda_k^*\}$ can be obtained through the subgradient-based methods as shown in [20].

Consider the special case that the average power budget is not less than the maximum power at all devices, i.e., $P_k^{\mathrm{avg}} \ge P_k^{\max}, \forall k \in \mathcal{K}$. Then, the average power constraints are not activated, and the dual variables are equal to zero (i.e., $\lambda_k^* = 0, \forall k$). The proposed optimal power allocation strategy is reduced to the isolated optimization across different communication rounds, and exhibits a threshold-based structure: all the devices $k \in \mathcal{K}'$ that cannot compensate their channel fading, i.e., satisfying $|\bar{h}_i^{(t)}|\sqrt{dP_k^{\max}} - 1 < 0$, should transmit signals with the largest power; otherwise, the devices $k \notin \mathcal{K}'$ that satisfying $|\bar{h}_i^{(t)}|\sqrt{dP_k^{\max}} - 1 \ge 0$, should enlarge their transmit power moderately (overcompensate their channel) to minimize the optimality gap exacerbated by the deep faded channel of the devices $k \in \mathcal{K}'$. As for the general case that the average power budgets are less than the maximum power at all devices, i.e., $P_k^{\mathrm{avg}} < P_k^{\max}, \forall k \in \mathcal{K}$, the dual variables in (36) are larger than zero (i.e., $\lambda_k^* > 0$). From (36), one can observe that the power allocation is closely related to the chronological order and the current channel condition. Larger power will be allocated to the devices in the later communication rounds and/or with better channel conditions.

The power control of (P2) in one dedicated communication round is a convex optimization problem presented as

$$\min_{\{b_k^{(t)}\}} V_r(1-\mu\alpha)^{(R-r)\rho}\omega_2^{(t)}\frac{\sum(\iota_k^{(t)})^2}{K^4}\sum_{k\in\mathcal{K}}\left||\bar{h}_k^{(t)}|\bar{b}_k^{(t)} - 1\right|^2 + \sum_{k\in\mathcal{K}} e_k(t)|b_k^{(t)}|^2 \tag{37a}$$

$$\text{s.t. } |b_k^{(t)}|^2 \le P_k^{\max}, \quad \forall k \in \mathcal{K}, \ \forall t \in \mathcal{T}, \tag{37b}$$

and the optimal solution is given by

$$(\bar{b}_k^{(t)})^\dagger = \min\left\{\frac{1}{|\bar{h}_k^{(t)}| + \frac{e_k(t)K^4}{V_r(1-\mu\alpha)^{(R-r)\rho}\omega_2^{(t)}\sum(\iota_k^{(t)})^2|\bar{h}_k^{(t)}|}}, \sqrt{P_k^{\max}}\right\}. \tag{38}$$

It can be observed that the power allocation is closely related to the current energy queue $e_k(t)$ and the parameter V_r which acted as an importance weight to balance the emphasis on the optimality gap and energy consumption minimization, in addition to the chronological order and the current channel condition. We can obtain a similar result as (P1) that larger power will be allocated to the devices in the later communication rounds and/or with better channel conditions. Besides, the allocated power will be diminished when facing a higher energy queue $e_k(t)$ and/or less importance weight V_r on the optimality gap in order to satisfy the long-term energy constraint.

B. Receiver Design

For any given power allocation and phase shift of IRS, the optimization problems with respect to the receive beamformers $\{\boldsymbol{m}^{(t)}\}$ of (P1) can be separated among different rounds. The corresponding subproblem of (P1) with omitting the time index can be expressed as

$$(\mathrm{P1.2}): \min_{\boldsymbol{m}} \ \omega_1\left|\boldsymbol{m}^{\mathrm{H}}\sum_{k\in\mathcal{K}}\tilde{\boldsymbol{h}}_k b_{f,k} - K\right|^2 + \omega_2\Big(\sum_{k\in\mathcal{K}}\left|\boldsymbol{m}^{\mathrm{H}}\tilde{\boldsymbol{h}}_k b_{f,k} - 1\right|^2 + d\sigma_z^2\|\boldsymbol{m}\|^2\Big). \tag{39}$$

Since problem (P1.2) is an unconstrained convex problem, by exploiting the first-order optimality condition, the closed-form solution for $\boldsymbol{m}$ is given by

$$\boldsymbol{m}^* = \left(\omega_1\Big(\sum_{k\in\mathcal{K}}\tilde{\boldsymbol{h}}_k b_{f,k}\Big)^2 + \omega_2\sum_{k\in\mathcal{K}}|b_{f,k}|^2\tilde{\boldsymbol{h}}_k\tilde{\boldsymbol{h}}_k^{\mathrm{H}} + \omega_2 d\sigma_z^2\boldsymbol{I}\right)^{-1} \times\Big((K\omega_1 + \omega_2)\sum_{k\in\mathcal{K}}\tilde{\boldsymbol{h}}_k b_{f,k}\Big). \tag{40}$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.

The optimal receiver (40) of the proposed offline design scheme is shown as a weighted summed structure, with the corresponding weights being the ones of the bias term and the MSE term in (14). Similarly, it can be obtained that the optimal receiver of the online design scheme is given by $\left(\sum_{k\in\mathcal{K}}|b_k|^2\tilde{\boldsymbol{h}}_k\tilde{\boldsymbol{h}}_k^{\mathrm{H}}+\sigma_z^2\mathbf{I}\right)^{-1}\left(\sum_{k\in\mathcal{K}}\tilde{\boldsymbol{h}}_k b_k\right)$ that targets at computing the summation of local gradients.

C. Phase Shift Optimization of the IRS

For any given power allocation and the receiver beamformer, the optimization problem with respect to the phase shift $\{\boldsymbol{\Theta}^{(t)}\}$ of (P1) can be separated among different communication rounds. Note that optimization procedures for $\{\boldsymbol{\Theta}^{(t)}\}$ of (P2) is part of the subproblem of (P1) that is only related to the MSE term $\mathbb{E}\left[\left\|\boldsymbol{\varepsilon}_g^{(t)}\right\|^2\right]$. Hence, we analyze the corresponding subproblem of (P1) and obtain the solutions of the two problems. By multiplying by a constant term K^2, the subproblem with respect to the phase shift $\{\boldsymbol{\Theta}\}$ of (P1) is given by

$$(\mathrm{P1.3}):\ \min_{\boldsymbol{\Theta}}\ \omega_1\left|\boldsymbol{m}^{\mathrm{H}}\sum_{k\in\mathcal{K}}\tilde{\boldsymbol{h}}_k b_{f,k}-K\right|^2 + \omega_2\sum_{k\in\mathcal{K}}\left|\boldsymbol{m}^{\mathrm{H}}\tilde{\boldsymbol{h}}_k b_{f,k}-1\right|^2 \tag{41a}$$

$$\text{s.t. } 0\le\theta_n\le 2\pi,\quad \forall n\in\{1,\dots,N\}. \tag{41b}$$

Since (P1.3) is a non-convex optimization problem due to the per-phase constraints in (41b). We first apply the semidefinite relaxation (SDR) strategy to relax (P1.3) into a semidefinite programming (SDP) problem. To reduce the computational complexity, we further propose an element-vise optimization approach to achieve near-optimal performance. Besides, it can be readily extended to the practical scenario that with discrete phase shifts constraint.

1) SDR Approach: Let $\boldsymbol{v}=[e^{\jmath\theta_1},\dots,e^{\jmath\theta_N}]^{\mathrm{H}}$, (P1.3) can be transformed to a convex optimation problem with norm-one constraint. Furthermore, the well-known SDR technique can be applied to convert it to a convex SDP problem [36], [44], [45] which can be solved efficiently in polynomial time by existing convex optimization solvers such as CVX. Besides, additional steps such as Gaussian randomization [45] need to be applied to extract a suboptimal solution when the optimal solution to the SDP problem is obtained with a higher rank.

2) Low-Complexity Suboptimal Solution: Although a satisfactory solution to (P1.3) can be obtained through the SDR approach, the lifted optimization variable induces huge computation complexity. Solving a series of high-dimensional SDP problems drastically increases the computational burden. Besides, the additional steps to extract a suboptimal solution are often with intolerable computation consumption, especially when facing an IRS with large-scale elements that are common in practice. Furthermore, the obtained continuous phase shifts are practically difficult to implement due to hardware limitations. Generally, the discrete phase shifts optimization problem is shown as an integer linear program problem that can be optimally solved via the branch-and-bound with the worst-case exponential complexity [46]. To reduce the computational complexity and provide the near-optimal solution for the IRS with discrete phase shifts constraint, we successively refine the phase shift of each element until converges. It is noted that the phase shifts of all elements are fully separable in the constraint, and only coupled in the objective function. Hence, we can successively optimize the phase shift of each element until converges. For a given $n\in\mathcal{N}$, by fixing the others, the objective function in (P1.3) is linear with respect to $e^{\jmath\theta_n}$, which is written as

$$\begin{aligned}
&\omega_1\Big(\boldsymbol{v}^{\mathrm{H}}\underbrace{\boldsymbol{\psi}\boldsymbol{\psi}^{\mathrm{H}}}_{\boldsymbol{\Psi}'}\boldsymbol{v}+2\Re\{\boldsymbol{v}^{\mathrm{H}}\underbrace{\boldsymbol{\psi}\zeta^{\mathrm{H}}}_{\zeta'}\}+|\zeta|^2\Big)\\
&\quad+\omega_2\sum_{k\in\mathcal{K}}\Big(\boldsymbol{v}^{\mathrm{H}}\underbrace{\boldsymbol{\phi}_k\boldsymbol{\phi}_k^{\mathrm{H}}}_{\boldsymbol{\Phi}_k'}\boldsymbol{v}+2\Re\{\boldsymbol{v}^{\mathrm{H}}\underbrace{\boldsymbol{\phi}_k\varphi_k^{\mathrm{H}}}_{\varphi_k'}\}+|\varphi_k|^2\Big)\\
&=\omega_1\Big(\sum_{l\neq n}^{N}\sum_{i\neq n}^{N}\boldsymbol{\Psi}'(l,i)e^{\jmath(\theta_l-\theta_i)}+2\Re\{e^{\jmath\theta_n}q_n^{(1)}\}+C'\Big)\\
&\quad+\omega_2\Big(\sum_{k=1}^{K}\sum_{l\neq n}^{N}\sum_{i\neq n}^{N}\boldsymbol{\Phi}_k'(l,i)e^{\jmath(\theta_l-\theta_i)}+2\Re\Big\{e^{\jmath\theta_n}\sum_{k=1}^{K}q_{k,n}^{(2)}\Big\}\\
&\qquad+\sum_{k=1}^{K}C_k\Big),
\end{aligned}\tag{42}$$

where $\boldsymbol{\psi}=\mathrm{diag}(\boldsymbol{m}^{\mathrm{H}}\boldsymbol{G})\boldsymbol{H}_r\boldsymbol{b}$, $\zeta=\boldsymbol{m}^{\mathrm{H}}\boldsymbol{H}_d\boldsymbol{b}-K$, $\boldsymbol{\phi}_k=\mathrm{diag}(\boldsymbol{m}^H\boldsymbol{G})\boldsymbol{h}_{r,k}b_{f,k}$, $\varphi_k=\boldsymbol{m}^H\boldsymbol{h}_{d,k}b_{f,k}-1$, and $\boldsymbol{H}_r,\boldsymbol{H}_d,\boldsymbol{b}$ collect the channel and the power allocation of devices, respectively. Besides, the constant terms in (42) are given by

$$q_n^{(1)}=\sum_{l\neq n}^{N}\boldsymbol{\Psi}'(n,l)e^{-\jmath\theta_l}+\zeta'(n)=|q_n^{(1)}|e^{\jmath\nu_n},\tag{43}$$

$$C'=\boldsymbol{\Psi}'(n,n)+2\Re\{\sum_{l\neq n}^{N}e^{\jmath\theta_l}\zeta'(l)\}+|\zeta'|^2,\tag{44}$$

$$\sum_{k=1}^{K}q_{k,n}^{(2)}=\sum_{k=1}^{K}\left(\sum_{l\neq n}^{N}\boldsymbol{\Phi}'(n,l)e^{-\jmath\theta_l}+\varphi_k'(n)\right)=|q_n^{(2)}|e^{\jmath\varsigma_n},\tag{45}$$

$$C_k=\boldsymbol{\Phi}'(n,n)+2\Re\left\{\sum_{l\neq n}^{N}e^{\jmath\theta_\ell}\varphi_k'(l)\right\}+|\varphi_k'|^2.\tag{46}$$

By leveraging the trigonometric identities, the part related to element n can be equivalently transformed to

$$\begin{aligned}
&\omega_1\Re\{e^{\jmath\theta_n}q_n^{(1)}\}+\omega_2\Re\left\{e^{\jmath\theta_n}\sum_{k=1}^{K}q_{k,n}^{(2)}\right\}\\
&=\omega_1|q_n^{(1)}|\sin(\theta_n+\nu_n+\frac{\pi}{2})+\omega_2|q_n^{(2)}|\sin(\theta_n+\varsigma_n+\frac{\pi}{2})\\
&=\hat{c}_n\sin(\theta_n+\hat{\theta}_n),
\end{aligned}\tag{47}$$

where $\hat{\theta}_n=\mathrm{atan}\frac{\omega_2|q_n^{(2)}|\sin(\varsigma_n-\nu_n)}{\omega_1|q_n^{(1)}|+\omega_2|q_n^{(2)}|\cos(\varsigma_n-\nu_n)}\in[-\frac{\pi}{2},\frac{\pi}{2}]$ if $\omega_1|q_n^{(1)}|+\omega_2|q_n^{(2)}|\cos(\varsigma_n-\nu_n)>0$, otherwise $\hat{\theta}_n=\pi+\mathrm{atan}\frac{\omega_2|q_n^{(2)}|\sin(\varsigma_n-\nu_n)}{\omega_1|q_n^{(1)}|+\omega_2|q_n^{(2)}|\cos(\varsigma_n-\nu_n)}\in[\frac{\pi}{2},\frac{3\pi}{2}]$.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.

Hence, the optimal phase shift of (P1.3) for the element n is

$$\theta_n^* = \frac{3}{2}\pi - \hat{\theta}_n. \tag{48}$$

Besides, the optimal phase shift of the element n to minimize the objective function of (P2) is $\theta_n^* = \arg\min_{\theta_n} \sin(\theta_n + \varsigma_n + \frac{\pi}{2})$. For ease of practical implementation, we consider that the phase shift at each element of the IRS can take only a finite number of discrete values. Let δ denote the number of bits used to indicate the number of phase shift levels. For simplicity, we assume that such discrete phase shift values are obtained by uniformly quantizing the interval $[0, 2\pi)$. Thus, the set of discrete phase shift values at each element is given by $\mathcal{S} = \{0, \Delta\theta, \ldots, (2^\delta - 1)\Delta\theta\}$ with $\Delta\theta = 2\pi/2^\delta$. Hence, the optimal configuration with finite quantization bit is $\bar{\theta}_n^* = \arg\min_{\theta\in\mathcal{S}} |\theta - \theta_n^*|$. With successively setting the phase shifts of all elements based on the above, the objective value will be non-increasing over the iterations until converges.

Overall Algorithm and Computational Complexity Analysis: Based on the provided solutions to the above three subproblems, an efficient BCD algorithm is proposed, where the IRS beamforming vector and transceiver are alternately optimized until convergence is achieved. Note that the objective value of problem (P1) and (P2) is bounded and non-decreasing by alternately optimizing $\{\boldsymbol{m}^{(t)}\}$, $\{b_k^{(t)}\}$ and $\{\boldsymbol{\Theta}^{(t)}\}$. Besides, the proposed BCD algorithm is guaranteed to converge to the stationary points of the two problems since the variables are only coupled in the objective functions. The mainly computational complexity of the proposed BCD algorithm lies in solving the sub-problems (P1.1), (P1.2), and (P1.3). Specifically, the corresponding computational complexity is given by $\mathcal{O}((\rho K)^{3.5})$, $\mathcal{O}(\rho M)$ and $\mathcal{O}(\rho N I'_{iter})$, where I'_{iter} denotes the number of iterations required in the element-vise optimization of discrete phase shifts. Therefore, the total complexity of the BCD algorithm is $\mathcal{O}\Big(\big((\rho K)^{3.5} + \rho(NI'_{iter} + M)\big)I_{iter}\Big)$, where I_{iter} denotes the number of iterations required to reach convergence of the objective function.

(P2) is an online system design problem that is operated in each communication round. The corresponding complexity of the BCD algorithm for (P2) is $\mathcal{O}\Big(\big(K + NI'_{iter} + M\big)I_{iter}\Big)$. From Theorem 2, we can obtain (P2) is only a bounded substitution of omniscient (P1). V_r is utilized to control the trade-off between the sizes of the queue backlogs and the objective function value. It is noted that an increasing V_r will strengthen the emphasis on the optimality gap minimization rather than power minimization in the later rounds, therefore the ultimate performance can be promoted. Besides, if the $e_k(r\rho + 1)$ is initialized to 0 as in the existing literature [31], (P2) is reduced to an optimality gap minimization problem in the initial rounds without virtual energy constraint. In that case, the obtained power allocation in the initial is most like to exceed the average power constraint, which obviously deviates from the optimal allocation. Therefore, the value of V_r and $e_k(r\rho + 1)$ are highly related to the FL performance, and we will show this phenomenon in the next section.

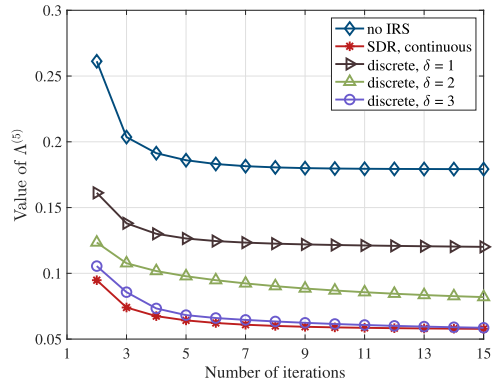


Fig. 2. Convergence behavior of the proposed algorithm.

V. Simulation Results

In this section, we conduct extensive simulations to validate the performance of the proposed performance-oriented long-term design approach for IRS-assisted FL. Simulations are operated on the MNIST dataset [34] and the CIFAR-10 dataset [35]. First, the convergence behavior of the proposed optimality gap minimization algorithm was presented. Then, we analyze the test accuracy under different parameters, e.g., ρ which determines the available future information which is highly related to the power allocation among different communication rounds, and the number of IRS elements N which determines the capability to configure the wireless channel. Besides, we compare the performance of offline and online design approaches under different settings, and provide the power allocation over the entire FL procedure to depict the *later-is-better* principle.

A. Simulation Setup

We consider a three-dimensional coordinate system, where the BS and the IRS are respectively located at $(0, 0, 30)$ and $(0, 50, 20)$ meters. In addition, the edge devices are randomly distributed in the circle region centered at $(50, 40, 0)$ with the radius equals to 20 meters. The wireless channels from the devices to the BS over different communication rounds follow i.i.d. Rayleigh fading, and the channels from IRS to BS and devices follow i.i.d. Rician fading. The path loss model under consideration is $L(d) = T_0(d/d_0)^\alpha$, where $T_0 = -30$ dB is the path loss at reference distance $d_0 = 1$ meter, d is the signal distance, and α is the path loss exponent. The path loss exponents for the BS-device link, the BS-IRS link, and the IRS-device link are set to 3.5, 2.2, and 2.5, respectively. Other parameters are set as follows: $P_k^{\max} = 20$ dBm and $P_k^{\text{avg}} = 17$ dBm, $\sigma_z^2 = -80$ dBm.

B. Performance Evaluation

First, the convergence behavior of the proposed offline scheme on the MNIST dataset [34] is shown in Fig. 2. We set $T = 100$, $R = 10$, and $\rho = 10$, the convergence behavior of the proposed optimality gap minimization algorithm conducted on the period 5 (i.e., 40 to 50 communication rounds) is presented as an example. Besides, parameters in the

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.

assumptions are set as $\mu = 0.2, L = 10$, respectively [47]. The number of devices is fixed to $K = 20$. For the case without IRS, the phase shifts matrix is fixed to $\mathbf{\Theta} = \mathbf{0}$. For the case with IRS, the number of the IRS elements is fixed to $N = 40$, and the number of antennas equipped at BS is set to $M = 5$. It is shown that the IRS can significantly diminish the optimality gap even with a finite configuration range such as the quantization bit equals to 1. Besides, the performance of IRS with a quantization level of 3 can achieve nearly the same performance as the one with continuous phase shifts.

Then, the IRS-assisted FL was deployed for handwritten recognition on the MNIST dataset [34] and image classification on the CIFAR-10 dataset [35]. We implement a 3-layer CNN as the recognition model for the MNIST dataset, which consists of an input layer, a final softmax output layer, and a midterm convolution layer with max pooling. The local batch size at each edge device is set to $B = 64$, and the learning rate α_t is fixed to 0.005. Besides, three convolution layers followed by max-pooling layers are developed for classification on the CIFAR-10 dataset, and other parameters are set as $\mu = 1, L = 5, B = 256, \alpha_t = 0.01$ [48]. We compare the learning performance with the following two benchmark schemes:

- **Optimal aggregation**: The wireless channels are ideal, thus the server can aggregate precise local gradients without the aggregation error presented in (9).
- **Isolated Design**: Variables are isolated optimized in each communication round as the existing works to minimize the aggregation error, hence the maximum transmit power constraint becomes P_{avg}.

Fig. 3a and 4a present the test accuracy of the offline scheme on the two datasets with different ρ. It is observed that the offline scheme with $\rho = 1$ has the same performance as the conventional isolated design approach since the offline scheme degenerates to the isolated design with no future information available. Besides, the performance of $\rho = 10$ and $\rho = 20$ are inferior to the isolated design approach initially, but it speeds up and eventually outperforms in the later rounds. This is due to the fact that the contributions of aggregation errors to the optimality gap are distinct at different communication rounds, which cannot be captured by the conventional isolated design scheme. Note that the system optimization is operated based on each ρ rounds lookahead information, therefore, the available information is closely related to the system performance. It is observed that the offline scheme with $\rho = 20$ outperforms the one with $\rho = 10$. Moreover, one can observe that the performance declines in the initial as the ρ goes, but rises and exceeds in the later rounds. These observations are consistent with the convergence analysis in the former sections which further demonstrate the necessity and efficiency of such a performance-oriented design approach. The ultimate performance of the proposed algorithm monotonically increases with a moderate ρ since more communication rounds are integrated optimized. However, it is observed that the offline scheme with $\rho = 50$ has the worst performance compared to $\rho = 10$ and $\rho = 20$ since the lookahead CSI and the estimated gradient bounds in the distant rounds may have errors thus degrading the system performance. Besides, the performance of the FL without the assistance of the IRS deteriorates heavily as the number of communication rounds increases. This is due to the model aggregation error in the latter rounds deteriorating the learning performance greater than the beginning.

TABLE I
PERFORMANCE OF THE ONLINE SCHEME WITH DIFFERENT ρ

	$\rho = 10$	$\rho = 20$	$\rho = 40$	$\rho = 100$	$\rho = 200$
$V = 20$	70.72%	72.56%	74.14%	75.29%	77.30%
$V = 100$	78.24%	80.86%	82.23%	83.17%	84.27%
$V = 300$	70.91%	73.30%	75.43%	77.90%	79.29%

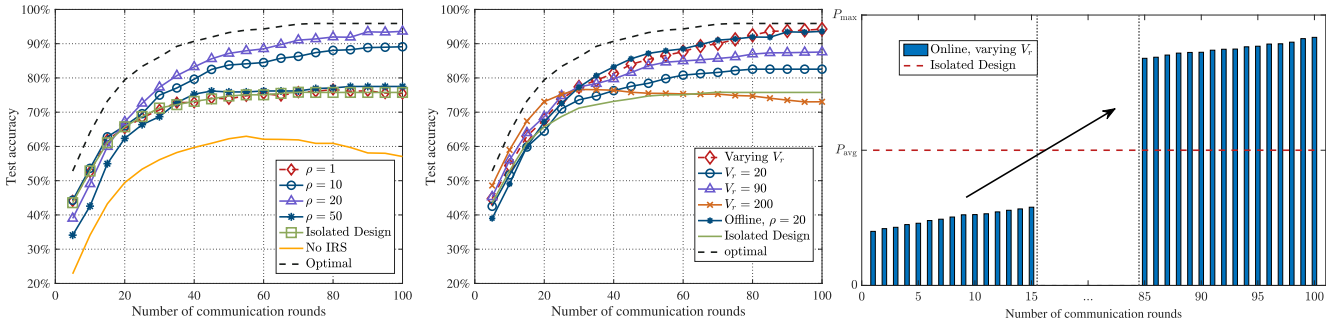
TABLE II
PERFORMANCE OF OFFLINE/ONLINE SCHEME UNDER DIFFERENT POWER SEQUENCE

	Descending power	Equal power	Proposed algorithm
Offline scheme	64.05%	76.87%	95.10%
Online scheme	62.48%	76.87%	95.46%

Fig. 3b and 4b depict the test accuracy of the online learning scheme on the two datasets. The FL procedures contain $T = 100$ rounds and $T = 200$ rounds on the MNIST and the CIFAR-10 dataset, respectively. We choose varying V_r sequences with $V_r = 20\sqrt{(10r + 1)}$ for both the MNIST and CIFAR-10 dataset. Besides, the benchmarks with fixed V_r are set as $V_r = 20$, $V_r = 90$, $V_r = 200$ and $V_r = 20$, $V_r = 100$, $V_r = 300$ for the two datasets, respectively. The energy queue $e_k(r\rho + 1), \forall r$ are initialized as $0.2dP_k^{\text{avg}}$. It can be observed that the online design scheme can achieve satisfying performance compared to the offline without foreseeing the future, which can be regarded as a practical and promising approach for FL system design. Besides, Fig. 3b and 4b validate that the online design approach with varying V_r can obtain higher test accuracy than the fixed V_r. The online design schemes with low V_r are more conservative to meet the long-term energy constraints while at the expense of sacrificing the system performance. On the contrary, the online design scheme with high V_r obtains near-optimal performance in the beginning, but it results in extreme energy consumption which is destructive to the latter rounds. It can be observed from the figures that an increasing V_r queue leads to larger power allocation in the latter rounds in FL, thus better performance can be achieved. Furthermore, we present the simulation results of the online scheme on the CIFAR-10 dataset with different parameters in Table I. It is observed that larger ρ leads to better performance under different V. Compared it with Fig. 4b, we can observed that the online scheme with appropriated V can enhance its performance by enlarge ρ since more communication rounds can be incorporated in the Lyapunov framework, while large ρ deteriorates the performance of the offline scheme as shown in Fig. 4a.

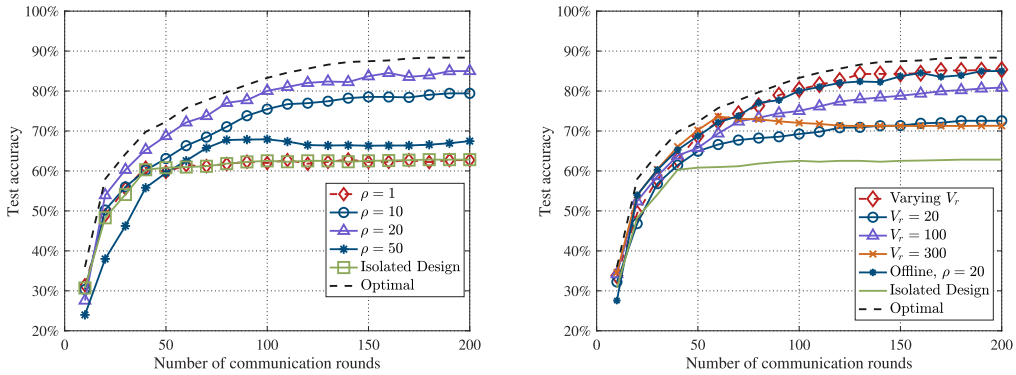
In order to further depict the *later-is-better* theory, we present the power allocation for a specific device of the online scheme on the MNIST dataset in Fig. 3c. The simulation is conducted under a static channel, therefore,

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.



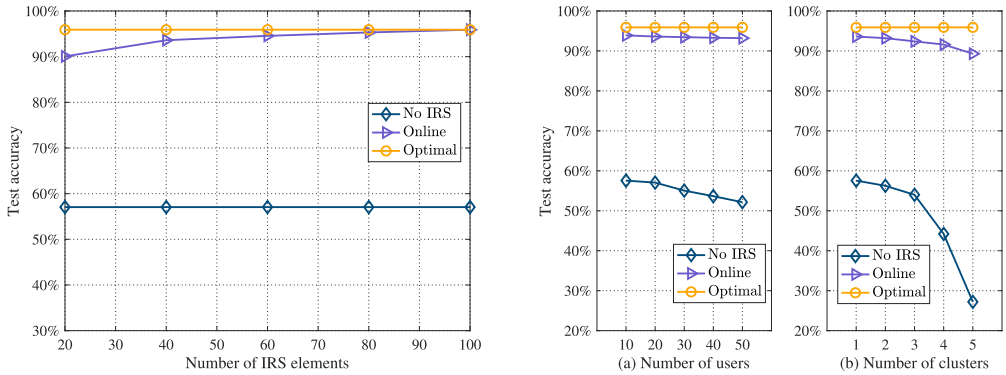
(a) Accuracy of the offline design. (b) Accuracy of the online design. (c) Power allocation of the online design.

Fig. 3. Performance of the offline/online design scheme on MINST dataset.



(a) Test accuracy of the offline design scheme. (b) Test accuracy of the online design scheme.

Fig. 4. Performance of the offline/online design scheme on CIFAR-10 dataset.



(a) Test accuracy vs. N. (b) Test accuracy vs. K.

Fig. 5. Test accuracy on MNIST dataset versus channel parameters.

it can be readily obtained and observed that the transmit power of a fixed device in conventional isolated design schemes remains the same over all the communication rounds. However, the long-term design schemes have a near-monotonic in the transmit power over time. Furthermore, we test the performance of descending and equal power queue on the MNIST dataset with $\rho = 20$ on a stationary channel. We first inverse the obtained power queues of the proposed offline and online schemes to construct the descending power sequences, and then optimize other variables. The equal power queue can be obtained from the conventional isolated design approach since the channel remains static. The simulation results are shown in Table II. It is observed that both the offline and online schemes with descending power sequence have extremely poor performance compared with the proposed algorithm and the equal power queue. Such an observation further validates the "*latter-is-better*" principle. By combing Fig. 3a, 3b, 3c and Table II, we can obtain that the system performance is enhanced by allocating large transmit power to the later communication rounds.

Finally, we investigate the test accuracy versus the number of IRS elements[4] and the number of users on the MNIST

[4]With assuming that the signal reflected by IRS two or more times is negligible and thus ignored, our problem can be readily extended to the multiple IRSs case. In addition, our proposed algorithm is also applicable to the multiple IRSs case without any modifications.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.

dataset. As shown in Fig. 5a, the test accuracy of the proposed algorithm is monotonically increased with N until converges to the optimal. By comparing Fig. 5a to Fig. 3b, we can observe that a large-scale IRS can diminish the performance deterioration due to the lack of lookahead information. Hence, when the CSI is varying or can not be precisely predicted, e.g., undergoes a fast-varying channel, the IRS with more elements can configure a more favorable communication link thus obtaining higher test accuracy in FL. The performance between FL with IRS and the one without IRS is further compared with the varying numbers of devices. Devices are distributed in one cluster in the setting of Fig. 5b(a) and are distributed in various clusters in the setting of Fig. 5b(b). Specifically, in the setting of multiple clusters, devices are randomly distributed in 5 circle regions with each radius equal to 20 meters, and the centers are located from $(50, 40, 0)$ to $(130, 40, 0)$ with the coordinates of the first dimension incremented by 20 in sequence. From Fig. 5b(a), it is observed that the test accuracy decays slightly as the number of devices increases even without the help of IRS, while the one with IRS can still obtain satisfactory test accuracy with tiny performance loss. This result verifies that AirComp is an efficient data aggregation approach for model aggregation in FL which is tolerant of the number of accessed devices. Then redirecting to Fig. 5b(b), one can observe that the performance of the one with or without IRS decays as the number of clusters increases. This is due to that the performance of AirComp is determined by the devices with poor channels. Fortunately, the FL system can still maintain satisfactory performance with the help of IRSs which can configure the deep faded channel to a favorable one.

VI. Conclusion

In this paper, we have proposed a performance-oriented long-term design framework based on the optimality gap for IRS-assisted FL systems. By analyzing the convergence behavior of FL, both offline and online design schemes were established. We adopted the BCD method to tackle the highly-intractable problems. Simulation results have demonstrated that such a long-term design scheme can precisely allocate the resources to different communication rounds, hence achieving a higher test accuracy and faster convergence behavior in FL than the conventional isolated design approach. Besides, we have confirmed that the online design approach based on the Lyapunov framework can achieve satisfactory performance without foreseeing the future which can be regarded as a practical and promising approach for FL system. Further, it has been seen that the employed IRS can assist AirComp in providing precise model aggregation in FL, especially when the lookahead information is limited or devices are widely distributed. The relationship between communication accuracy and learning performance presented in the paper has been rigorously determined via mathematical proof and simulation verification, which have revealed a *later-is-better* principle that can provide significant guidance for the FL system design.

Appendix A
Proof of Theorem 1

According to Assumption 1, we can obtain that

$$\begin{aligned}
&F\left(\mathbf{w}^{(t+1)}\right) - F\left(\mathbf{w}^{(t)}\right) \\
&\leq \left\langle \nabla F\left(\mathbf{w}^{(t)}\right), \left(\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\right)\right\rangle + \frac{L}{2}\left\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\right\|^2 \\
&= -\alpha^{(t)}\left\langle \nabla F\left(\mathbf{w}^{(t)}\right), \left(\overline{\mathbf{g}}^{(t)} + \boldsymbol{\varepsilon}_g^{(t)}\right)\right\rangle \\
&\quad + \frac{L(\alpha^{(t)})^2}{2}\left\|\overline{\mathbf{g}}^{(t)} + \boldsymbol{\varepsilon}_g^{(t)}\right\|^2. \qquad (49)
\end{aligned}$$

With $\alpha^{(t)} \leq \frac{1}{L}$, by taking expectation with respect to the stochastic gradient and the aggregation error on both sides of (49), we have

$$\begin{aligned}
&\mathbb{E}\left[F\left(\mathbf{w}^{(t+1)}\right) - F\left(\mathbf{w}^{(t)}\right)\right] \\
&\leq \frac{L(\alpha^{(t)})^2}{2}\mathbb{E}\left[\left\|\overline{\mathbf{g}}^{(t)} + \boldsymbol{\varepsilon}_g^{(t)}\right\|^2\right] \\
&\quad - \alpha^{(t)}\mathbb{E}\left[\left\langle \nabla F\left(\mathbf{w}^{(t)}\right), \left(\overline{\mathbf{g}}^{(t)} + \boldsymbol{\varepsilon}_g^{(t)}\right)\right\rangle\right] \\
&\overset{a}{=} -\alpha^{(t)}\left\|\nabla F\left(\mathbf{w}^{(t)}\right)\right\|^2 - \alpha_t\left\langle \nabla F\left(\mathbf{w}^{(t)}\right), \mathbb{E}\left[\boldsymbol{\varepsilon}_g^{(t)}\right]\right\rangle \\
&\quad + \frac{L(\alpha^{(t)})^2}{2}\mathbb{E}\left[\left\|\overline{\mathbf{g}}^{(t)}\right\|^2\right] + L(\alpha^{(t)})^2 \\
&\quad \times \left\langle \nabla F\left(\mathbf{w}^{(t)}\right), \mathbb{E}\left[\boldsymbol{\varepsilon}_g^{(t)}\right]\right\rangle + \frac{L(\alpha^{(t)})^2}{2}\mathbb{E}\left[\left\|\boldsymbol{\varepsilon}_g^{(t)}\right\|^2\right] \\
&\overset{b}{\leq} -\alpha^{(t)}\left\|\nabla F(\mathbf{w}^{(t)})\right\|^2 \\
&\quad + \frac{L(\alpha^{(t)})^2}{2}\left(\mathbb{E}\left[\left\|\overline{\mathbf{g}}^{(t)}\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{\varepsilon}_g^{(t)}\right\|^2\right]\right) \\
&\quad + \frac{\alpha^{(t)}(1-L\alpha^{(t)})}{2}\left(\left\|\nabla F\left(\mathbf{w}^{(t)}\right)\right\|^2 + \left\|\mathbb{E}\left[\boldsymbol{\varepsilon}_g^{(t)}\right]\right\|^2\right), \qquad (50)
\end{aligned}$$

where a is due to Assumption 3, b is due to the Cauchy-Schwarz inequality and the inequality of arithmetic and geometric means: $\pm\boldsymbol{x}_1^T\boldsymbol{x}_2 \leq \|\boldsymbol{x}_1\|\,\|\boldsymbol{x}_2\| \leq \frac{\|\boldsymbol{x}_1\|^2}{2} + \frac{\|\boldsymbol{x}_2\|^2}{2}$.

According to Assumption 2, and with $\alpha^{(t)} \leq \frac{1}{\mu}, \alpha^{(t)} \leq \frac{1}{L}$, we have

$$\begin{aligned}
&\mathbb{E}\left[F\left(\mathbf{w}^{(t+1)}\right)\right] - F(\mathbf{w}^*) \\
&\leq (1-\mu\alpha^{(t)})\left(\mathbb{E}\left[F(\mathbf{w}^{(t)})\right] - F(\mathbf{w}^*)\right) \\
&\quad + \frac{L(\alpha^{(t)})^2}{2}\mathbb{E}\left[\left\|\overline{\mathbf{g}}^{(t)}\right\|^2\right] + \frac{\alpha^{(t)}(1-L\alpha^{(t)})}{2}\left\|\mathbb{E}\left[\boldsymbol{\varepsilon}_g^{(t)}\right]\right\|^2 \\
&\quad + \frac{L(\alpha^{(t)})^2}{2}\mathbb{E}\left[\left\|\boldsymbol{\varepsilon}_g^{(t)}\right\|^2\right]. \qquad (51)
\end{aligned}$$

Recursively applying (51), finally we obtain that for $\forall T_2 > T_1$,

$$\begin{aligned}
&\mathbb{E}\left[F\left(\mathbf{w}^{(T_2+1)}\right)\right] - F(\mathbf{w}^*) \\
&\leq \prod_{t=T_1+1}^{T_2}\left(1-\mu\alpha^{(t)}\right)\left(\mathbb{E}\left[F\left(\mathbf{w}^{(T_1+1)}\right)\right] - F(\mathbf{w}^*)\right) \\
&\quad + \sum_{t=T_1+1}^{T_2}\frac{\prod_{t=T_1+1}^{T_2}\left(1-\mu\alpha^{(t)}\right)}{\left(1-\mu\alpha^{(T_2)}\right)}
\end{aligned}$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.

$$\times \left\{ \frac{\alpha^{(t)}(1-L\alpha^{(t)})}{2} \left\| \mathbb{E}\left[\boldsymbol{\varepsilon}_g^{(t)}\right] \right\|^2 + \frac{L(\alpha^{(t)})^2}{2}\mathbb{E}\left[\left\|\boldsymbol{\varepsilon}_g^{(t)}\right\|^2\right] + \frac{L(\alpha^{(t)})^2}{2}\mathbb{E}\left[\left\|\overline{\mathbf{g}}^{(t)}\right\|^2\right] \right\}. \tag{52}$$

By fixing $\alpha^{(t)} \equiv \alpha$, we have

$$\begin{aligned}
&\mathbb{E}\left[F\left(\mathbf{w}^{(T_2+1)}\right)\right] - F(\mathbf{w}^*) \\
&\leq (1-\mu\alpha)^{(T_2-T_1)}\left(\mathbb{E}\left[F\left(\mathbf{w}^{(T_1+1)}\right)\right] - F(\mathbf{w}^*)\right) \\
&\quad + \sum_{t=T_1+1}^{T_2}(1-\mu\alpha)^{(T_2-t)}\left\{\frac{\alpha(1-L\alpha)}{2}\left\|\mathbb{E}\left[\boldsymbol{\varepsilon}_g^{(t)}\right]\right\|^2 \right. \\
&\quad \left. + \frac{L\alpha^2}{2}\mathbb{E}\left[\left\|\boldsymbol{\varepsilon}_g^{(t)}\right\|^2\right] + \frac{L\alpha^2}{2}\mathbb{E}\left[\left\|\overline{\mathbf{g}}^{(t)}\right\|^2\right]\right\}.
\end{aligned} \tag{53}$$

Hence we can obtain the optimality gap in Theorem 1.

Appendix B
Two Kinds of Extensions for Device Scheduling

Two kinds of extensions to include the device scheduling are shown as follows.

Case 1: We can simply construct a bicriterion problem that minimizes the weighted sum of the optimality gap and maximizes the cardinality of the currently selected device set. By adding a term $-w^{(t)}|\mathcal{K}^{(t)}|$ to the objective function, where $w^{(t)} > 0$ is the parameter to achieve a trade-off between optimality gap and device participation, we can obtain a similar problem as [28] that minimize the optimality gap caused by wireless channel while selecting as more devices as possible for convergence accelerating.

Case 2: We can mathematically characterize the gradient residual due to the device selection and then formulate the problem. Consider the general case that the k-th device has $|\mathcal{D}_k|$ training samples with $\sum_k |\mathcal{D}_k| = |\mathcal{D}|$. Hence, the desired global gradient is given by

$$\mathbf{g}^{(t)} = \frac{1}{|\mathcal{D}|}\sum_{k=1}^{K}|\mathcal{D}_k|\mathbf{g}_k^{(t)}, \tag{54}$$

shown as the weighted average of local gradients with the weights proportional to the size of the corresponding local dataset $|\mathcal{D}_k|$. Let $\mathcal{K}^{(t)}$ denotes the selected devices set in round t, and $\widetilde{\mathcal{K}}^{(t)}$ denotes its complement. Then gradient residual due to the device selection is

$$\boldsymbol{\varepsilon}_d^{(t)} = \frac{1}{\sum_{i\in\mathcal{K}^{(t)}}|\mathcal{D}_k|}\sum_{i\in\mathcal{K}^{(t)}}|\mathcal{D}_k|\boldsymbol{g}_i^{(t)} - \frac{1}{|\mathcal{D}|}\sum_{k=1}^{K}|\mathcal{D}_k|\mathbf{g}_k^{(t)}. \tag{55}$$

In that case, the gradient error vector is given by

$$\widetilde{\boldsymbol{\varepsilon}_g}^{(t)} = \boldsymbol{\varepsilon}_g^{(t)} + \boldsymbol{\varepsilon}_d^{(t)}, \tag{56}$$

and we have [37]

$$\begin{aligned}
\mathbb{E}[\|\widetilde{\boldsymbol{\varepsilon}_g}^{(t)}\|^2] &\leq \mathbb{E}[\|\boldsymbol{\varepsilon}_g{}^{(t)}\|^2] + \mathbb{E}[\|\boldsymbol{\varepsilon}_d{}^{(t)}\|^2] \\
&\leq \mathbb{E}[\|\boldsymbol{\varepsilon}_g{}^{(t)}\|^2] + \frac{4(|\mathcal{D}| - \sum_{i\in\mathcal{K}^{(t)}}|\mathcal{D}_k|)}{|\mathcal{D}|^2}\gamma^{(t)},
\end{aligned} \tag{57}$$

where $\gamma^{(t)}$ is the upper bound of sample-wise gradient. In that case, minimizing the optimality gap consists of two aspects, one is decreasing the number of selected devices to reduce the gradient aggregation error $\boldsymbol{\varepsilon}_g^{(t)}$, and the other is increasing the number of selected devices to reduce the gradient residual $\boldsymbol{\varepsilon}_d^{(t)}$ due to the device selection.

Appendix C
Proof of Theorem 2

Without loss of generality, we adopt the quadratic Lyapunov function $L(\boldsymbol{e}(t)) \triangleq \frac{1}{2}\sum_{k\in\mathcal{K}} e_k^2(t)$. Besides the 1-round Lyapunov drift and the R-round Lyapunov drift are represented as $\Delta_1(t) \triangleq L(\boldsymbol{e}(t+1)) - L(\boldsymbol{e}(t))$ and $\Delta_R(t) \triangleq L(\boldsymbol{e}(t+R)) - L(\boldsymbol{e}(t))$, respectively. According to (31), we have

$$\begin{aligned}
&\frac{1}{2}\sum_{k\in\mathcal{K}} e_k^2(t+1) \\
&\leq \frac{1}{2}\sum_{k\in\mathcal{K}}\left(d|b_k^{(t)}|^2 - dP_k^{\text{avg}} + e_k(t)\right)^2 \\
&\leq C_e + \frac{1}{2}\sum_{k\in\mathcal{K}} e_k^2(t) + \sum_{k\in\mathcal{K}} e_k(t)\left(d|b_k^{(t)}|^2 - dP_k^{\text{avg}}\right),
\end{aligned} \tag{58}$$

where $C_e = \frac{1}{2}KE_{\max}^2$ with $E_{\max} = \max_{k,t}(d|b_k^{(t)}|^2 - dP_k^{\text{avg}})$. Thus,

$$\begin{aligned}
\Delta_1(t) &\leq C_e + \sum_{k\in\mathcal{K}} e_k(t)\left(d|b_k^{(t)}|^2 - dP_k^{\text{avg}}\right) \\
&\leq C_e + \sum_{k\in\mathcal{K}} e_k(t) d\left(P_k^{\max} - P_k^{\text{avg}}\right).
\end{aligned} \tag{59}$$

For arbitrary r-th period with V_r, we have

$$\begin{aligned}
&\Delta_R^{\dagger}(r\rho) + V_r\sum_{t=r\rho+1}^{(r+1)\rho} G_t^{\dagger} \\
&\leq \rho C_e + \sum_{t=r\rho+1}^{(r+1)\rho}\sum_{k\in\mathcal{K}} e_k(t)\left(d|b_k^{(t)}|^2 - dP_k^{\text{avg}}\right) \\
&\quad + V_r\sum_{t=r\rho+1}^{(r+1)\rho} G_t^* \\
&\leq \rho C_e + \sum_{t=r\rho+1}^{(r+1)\rho}\sum_{k\in\mathcal{K}} E_{\max}\big((t-1)E_{\max} + e_k(r\rho+1)\big) \\
&\quad + V_r\sum_{t=r\rho+1}^{(r+1)\rho} G_t^*.
\end{aligned} \tag{60}$$

Note that $\Delta_R^{\dagger}(r\rho) \geq 0$, thus

$$\sum_{t=r\rho+1}^{(r+1)\rho} G_t^{\dagger} \leq \sum_{t=r\rho+1}^{(r+1)\rho} G_t^* + \frac{C_r}{V_r}, \tag{61}$$

where $C_r = \rho C_e + \sum_{t=r\rho+1}^{(r+1)\rho}\sum_{k\in\mathcal{K}} E_{\max}\big((t-1)E_{\max} + e_k(r\rho+1)\big)$. Summing over all the R periods, we can obtain the (33) in Theorem 2.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.

Note that

$$\Delta_R(t) \triangleq L(\boldsymbol{e}(t+R)) - L(\boldsymbol{e}(t)) = \sum_k \frac{1}{2} e_k^2(R+1). \quad (62)$$

Hence, we have

$$\begin{aligned}\sum_{t=r\rho+1}^{(r+1)\rho} \left(d|b_k^{(t)}|^2 - dP_k^{\text{avg}}\right) &\leq \sum_{t=r\rho+1}^{(r+1)\rho} \left(e_k(t+1) - e_k(t)\right) \\ &= e_k((r+1)\rho+1) - e_k(r\rho+1) \leq \sqrt{2\Delta_\rho} - e_k(r\rho+1) \\ &\leq \sqrt{2\left(C_r + V_r \sum_{t=r\rho+1}^{(r+1)\rho} G_t^*\right)} - e_k(r\rho+1). \quad (63)\end{aligned}$$

Summing over all the R periods, we can obtain the (34) in Theorem 2.

References

[1] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, Jan. 2020.

[2] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, Jan. 2022.

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

[4] W. Y. B. Lim et al., "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, 3rd Quart., 2020.

[5] 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.

[6] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication-efficient federated learning," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 17, Apr. 2021, Art. no. e2024789118.

[7] S. Wang, M. Lee, S. Hosseinalipour, R. Morabito, M. Chiang, and C. G. Brinton, "Device sampling for heterogeneous federated learning: Theory, algorithms, and implementation," in *Proc. IEEE Conf. Comput. Commun.*, May 2021, pp. 1–10.

[8] E. Ozfatura, K. Ozfatura, and D. Gündüz, "Time-correlated sparsification for communication-efficient federated learning," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 461–466.

[9] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. Adv. NIPS*, Dec. 2017, pp. 1709–1720.

[10] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "SignSGD: Compressed optimisation for non-convex problems," in *Proc. Intl. Conf. Mach. Learn.*, Jul. 2018, pp. 559–568.

[11] Z. Wang, Y. Zhao, Y. Zhou, Y. Shi, C. Jiang, and K. B. Letaief, "Over-the-air computation: Foundations, technologies, and applications," 2022, *arXiv:2210.10524*.

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

[13] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6463–6486, Oct. 2011.

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

[15] M. Fu, Y. Zhou, Y. Shi, W. Chen, and R. Zhang, "UAV aided over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 4909–4924, Jul. 2022.

[16] H. Guo et al., "Over-the-air aggregation for federated learning: Waveform superposition and prototype validation," *J. Commun. Inf. Netw.*, vol. 6, no. 4, pp. 429–442, Dec. 2021.

[17] M. Chen, H. V. Poor, W. Saad, and S. Cui, "Convergence time optimization for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2457–2471, Apr. 2021.

[18] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1935–1949, Mar. 2021.

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

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

[22] C. Xu, S. Liu, Z. Yang, Y. Huang, and K. Wong, "Learning rate optimization for federated learning exploiting over-the-air computation," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3742–3756, Dec. 2021.

[23] J. Ren, G. Yu, and G. Ding, "Accelerating DNN training in wireless federated edge learning systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 219–232, Jan. 2021.

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

[25] N. Rajatheva et al., "Scoring the terabit/s goal: Broadband connectivity in 6G," 2020, *arXiv:2008.07220*.

[26] M. Hua, Q. Wu, D. W. K. Ng, J. Zhao, and L. Yang, "Intelligent reflecting surface-aided joint processing coordinated multipoint transmission," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1650–1665, Mar. 2021.

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

[28] W. Ni, Y. Liu, Z. Yang, H. Tian, and X. Shen, "Federated learning in multi-RIS-aided systems," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9608–9624, Jun. 2022.

[29] 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.*, vol. 21, no. 12, pp. 10083–10099, Dec. 2022.

[30] Y. Hu et al., "Energy minimization for federated learning with IRS-assisted over-the-air computation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 3105–3109.

[31] J. Xu and H. Wang, "Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1188–1200, Feb. 2021.

[32] C. Shen, J. Xu, S. Zheng, and X. Chen, "Resource rationing for wireless federated learning: Concept, benefits, and challenges," *IEEE Commun. Mag.*, vol. 59, no. 5, pp. 82–87, May 2021.

[33] M. J. Neely, *Stochastic Network Optimization With Application to Communication and Queueing Systems*. San Rafael, CA, USA: Morgan & Claypool Publishers, 2010.

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

[35] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2009.

[36] 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.

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

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

[39] J. Wang and G. Joshi, "Cooperative SGD: A unified framework for the design and analysis of local-update SGD algorithms," *J. Mach. Learn. Res.*, vol. 22, no. 213, pp. 1–50, 2021.

[40] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition," in *Proc. ECML PKDD*, 2016, pp. 795–811.

[41] A. Kulkarni, A. Seetharam, A. Ramesh, and J. D. Herath, "DeepChannel: Wireless channel quality prediction using deep learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 443–456, Jan. 2020.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.

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

[43] H. Yu, S. Yang, and S. Zhu, "Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *Proc. AAAI*, Jan. 2019, pp. 5693–5700.

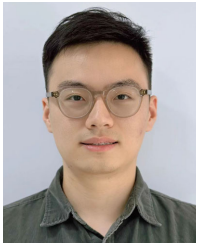
[44] M. Zhao, Q. Wu, M. Zhao, and R. Zhang, "Exploiting amplitude control in intelligent reflecting surface aided wireless communication with imperfect CSI," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 4216–4231, Jun. 2021.

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

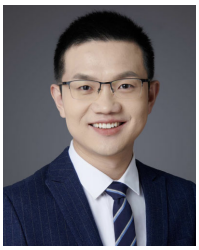
[46] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1838–1851, Mar. 2020.

[47] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Convergence of federated learning over a noisy downlink," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1422–1437, Mar. 2022.

[48] M. M. Amiri, T. M. Duman, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Blind federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5129–5143, Aug. 2021.



Yapeng Zhao (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau, China. His current research interests include intelligent reflecting surface (IRS), over-the-air computation (AirComp), and federated learning (FL).

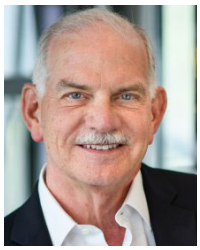


Qingqing Wu (Senior Member, IEEE) received the B.Eng. degree in electronic engineering from the South China University of Technology in 2012 and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University (SJTU) in 2016.

From 2016 to 2020, he was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore. He is currently an Assistant Professor with Shanghai Jiao Tong University. He has coauthored more than 100 IEEE journal articles with 29 ESI highly cited papers and nine ESI hot papers, which have received more than 20,000 Google citations. His current research interests include intelligent reflecting surface (IRS), unmanned aerial vehicle (UAV) communications, and MIMO transceiver design. He was listed as the Clarivate ESI Highly Cited Researcher in 2021 and 2022, and the Most Influential Scholar Award in AI-2000 by Aminer in 2022 and 2023. He was a recipient of the IEEE Communications Society Fred Ellersick Prize, the IEEE Best Tutorial Paper Award in 2023, the IEEE Communications Society Asia Pacific Best Young Researcher Award and Outstanding Paper Award in 2022, the IEEE Communications Society Young Author Best Paper Award in 2021, the Outstanding Ph.D. Thesis Award of China Institute of Communications in 2017, the IEEE ICCC Best Paper Award in 2021, and IEEE WCSP Best Paper Award in 2015. He was the Exemplary Editor of IEEE COMMUNICATIONS LETTERS in 2019 and the exemplary reviewer of several IEEE journals. He serves as an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, and IEEE WIRELESS COMMUNICATIONS LETTERS. He is the Lead Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS on "UAV Communications in 5G and Beyond Networks." He is the Workshop Co-Chair for IEEE ICC (2019–2022) Workshop on "Integrating UAVs into 5G and Beyond" and the Workshop Co-Chair for IEEE GLOBECOM 2020 and ICC 2021 Workshop on "Reconfigurable Intelligent Surfaces for Wireless Communication for Beyond 5G." He serves as the Workshops and Symposia Officer for Reconfigurable Intelligent Surfaces Emerging Technology Initiative and Research Blog Officer of Aerial Communications Emerging Technology Initiative. He is the IEEE Communications Society Young Professional Chair in Asia Pacific Region.

Wen Chen (Senior Member, IEEE) is currently a Tenured Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, China, where he is also the Director of the Broadband Access Network Laboratory. He has published more than 130 articles in IEEE journals and more than 120 papers in IEEE conferences, with citations more than 9000 in Google scholar. His research interests include multiple access, wireless AI, and meta-surface communications. He is a fellow of the Chinese Institute of Electronics and a Distinguished Lecturer of IEEE Communications Society and IEEE Vehicular Technology Society. He is the Shanghai Chapter Chair of IEEE Vehicular Technology Society and an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE ACCESS, and IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY.

Celimuge Wu (Senior Member, IEEE) received the Ph.D. degree from The University of Electro-Communications, Japan. He is currently a Professor and the Director of the Meta-Networking Research Center, The University of Electro-Communications. His research interests include vehicular networks, edge computing, the IoT, and AI for wireless networking and computing. He serves as an Associate Editor for IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, and IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING. He was a recipient of the 2021 IEEE Communications Society Outstanding Paper Award, the 2021 IEEE INTERNET OF THINGS JOURNAL Best Paper Award, the IEEE Computer Society 2020 Best Paper Award, and the IEEE Computer Society 2019 Best Paper Award Runner-Up. He is the Vice Chair (Asia Pacific) of the IEEE Technical Committee on Big Data (TCBD). He is an IEEE Vehicular Technology Society Distinguished Lecturer.



H. Vincent Poor (Life Fellow, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering (EECS), Princeton University, in 1977. From 1977 to 1990, he was a Faculty Member of the University of Illinois at Urbana–Champaign. Since 1990, he has been a Faculty Member at Princeton University, where he is currently the Michael Henry Strater University Professor. From 2006 to 2016, he served as the Dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. Among his publications in these areas is the recent book *Machine Learning and Wireless Communications* (Cambridge University Press, 2022). His research interests include information theory, machine learning and network science, and their applications in wireless networks, and energy systems and related fields. He is a member of the National Academy of Engineering and the National Academy of Sciences and is a Foreign Member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the IEEE Alexander Graham Bell Medal in 2017.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on January 01,2024 at 13:54:43 UTC from IEEE Xplore. Restrictions apply.