Learning Rate Optimization for Federated Learning Exploiting Over-the-Air Computation

Chunmei Xu, *Student Member, IEEE*, Shengheng Liu, *Member, IEEE*, Zhaohui Yang, *Member, IEEE*, Yongming Huang, *Senior Member, IEEE*, and Kai-Kit Wong, *Fellow, IEEE*

***Abstract*—Federated learning (FL) as a promising edge-learning framework can effectively address the latency and privacy issues by featuring distributed learning at the devices and model aggregation in the central server. In order to enable efficient wireless data aggregation, over-the-air computation (AirComp) has recently attracted great attention. However, fading of wireless channels can produce aggregate distortions in an AirComp-based FL scheme. In this paper, we propose a modified federated averaging (FedAvg) algorithm by introducing the local learning rates and present the convergence analysis. To combat the distortion, the local learning rate is optimized to adapt the fading channel, which is termed as dynamic learning rate (DLR). We begin our discussion by considering multiple-input-single-output (MISO) scenario, since the underlying optimization problem is convex and has a closed-form solution. Our studies are extended to a more general multiple-input-multiple-output (MIMO) case and an iterative method is derived. We also present the asymptotic analysis and give a near-optimal and closed-form receive beamforming solution when the number of antennas approaches infinity. Extensive simulation results demonstrate the effectiveness of the proposed DLR scheme in reducing the aggregate distortion and guaranteeing the testing accuracy on the MNIST and CIFAR10 datasets. In addition, the asymptotic analysis and the close-form solution are verified through numerical simulations.**

***Index Terms*—Distributed algorithm, federated learning, over-the-air computation, learning rate, beamforming.**

I. Introduction

FUTURE sixth-generation (6G) communication networks are envisioned to undergo a profound transformation, from connected things to connected intelligence. This transformation comes with stringent requirements such as dense networking, strict security, high energy efficiency, and high intelligence [2], [3]. Artificial intelligence (AI) technologies, which enable automatic analysis of a large mass of data generated in wireless networks and the subsequent optimization of highly dynamic and complex networks [4]–[6], will shape the landscape of 6G. Conversely, 6G will give renewed impetus to the AI-empowered mobile applications by supplying advanced wireless communication and mobile computing technologies [7].

AI tasks entail intensive computation workloads. Hence, they are generally trained on the server center with the availability of data collected from the edge devices/sensors [8], [9]. The data volume can be considerably large and, thereby, imposing heavy transmission traffic burden and increasing the latency. Another critical problem comes from the serious concern of privacy leakage, since the generated data, e.g., photos and social-networking records, at the devices are often privacy sensitive. An intuitive way to counteract the above issues would be to conduct the training and inference processes directly at the network edge using locally generated real-time data [10], [11].

Federated learning (FL) tackles the aforementioned concerns by collaboratively training a shared global model with locally stored data [12]–[15]. As a typical FL algorithm, the federated averaging (FedAvg) proposed in [12] iteratively performs two phases: (I) The devices receive the global model from the aggregator and update local models for multiple steps. (II) These local models are sent to the aggregator to obtain the global model. Note that the size of the local models are much smaller than the size of the raw data. Nonetheless, uploading local models via wireless links is resource-demanding, which becomes the main bottleneck to implement FL in practice. In this regard, developing communication-efficient methods are of paramount importance. Some recent works have considered asynchronous mechanism [16], quantization [17], [18], and sparsification [19], [20] to reduce the transmission overhead, which, however, ignore the aspects of physical and network layers.

In the second phase of FL, the aggregator averages the local model parameters from the distributed devices, which is essentially wireless data aggregation. Conventional multiple-access schemes, such as orthogonal frequency-division multiple access (OFDM), are based on the separated-communication-and-computing principle. In [21], a time-division multiple access (TDMA) system was considered, where a joint batch-size selection and communication resource allocation scheme was developed aiming at accelerating the training process

Manuscript received February 24, 2021; revised July 25, 2021; accepted August 28, 2021. Date of publication October 6, 2021; date of current version November 22, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1806600, in part by the National Natural Science Foundation of China under Grant 62001103 and Grant 61720106003, and in part by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/T015985/1. This article was presented at the IEEE International Conference on Communications (ICC) Workshop on Edge Learning for 5G Mobile Networks and Beyond, Montreal, QC, Canada, June 2021 [DOI:10.1109/ICCWorkshops50388.2021.9473663]. *(Corresponding authors: Yongming Huang; Shengheng Liu.)*

Chunmei Xu, Shengheng Liu, and Yongming Huang are with the National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University, Nanjing 210096, China, and also with the Purple Mountain Laboratories, Nanjing 211111, China (e-mail: xuchunmei@seu.edu.cn; s.liu@seu.edu.cn; huangym@seu.edu.cn).

Zhaohui Yang and Kai-Kit Wong are with the Department of Electronic and Electrical Engineering, University College London, London WC1E 6BT, U.K. (e-mail: zhaohui.yang@ucl.ac.uk; kaikit.wong@ucl.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2021.3118402.

Digital Object Identifier 10.1109/JSAC.2021.3118402

and improving the learning efficiency. The impact of three different scheduling policies on the FL performance were analytically studied in the large-scale wireless networks [22]. Such communication-and-computing approaches could result in a sharp rise in the consumption of wireless resources [10], which is communication inefficient in the implementation of FL. As an alternative, over-the-air computation (AirComp) scheme leverages the waveform superposition property of multiple-access channels, which is fundamentally different from the traditional separated-communication-and-computing principle [23]. By aggregating the data simultaneously received from distributed devices in an analog manner, the AirComp technique can further improve the communication efficiency [24]–[26].

The AirComp technique has recently been integrated into the implementation of FL. A gradient sparsification and random linear projection based approach was proposed, and the reduced data was transmitted via AirComp, which was shown to outperform its digital counterpart [24], [25]. As a matter of fact, the distortions caused by fading and noisy channels are critical for learning tasks as a large aggregation error may lead to the degradation of inference performance. Aiming at the reduction of the aggregate error, the authors in [26] designed the transmission power and proposed two scheduling schemes to align the received signals from the distributed devices in a single-input-single-output (SISO) system. Meanwhile in [27], the joint device selection and receive beamforming design was investigated in a single-input-multiple-output (SIMO) configuration, where a novel unified difference-of-convex (DC) function was proposed. Considering intelligent reflection surface (IRS), the authors proposed a joint optimization scheme of the passive beamforming and linear detection vector in a cloud radio access network (C-RAN) system to minimize the distortion [28]. A novel resource optimization and device selection method was developed to minimize the aggregate error while maximizing the selected devices with the aid of multi-IRS [29]. It is observed that these existing works utilized wireless resources, such as power control, device selection and beamforming design, as well as channel configuration, to align the received signals from distributed devices to mitigate the aggregate error. Nevertheless, they did not unleash the potential of the learning rate from the perspective of machine learning (ML) in reducing the distortion.

Learning rate is one of the key hyper-parameters that can greatly influence the convergence and the convergence rate of the learning tasks. A large learning rate may cause the loss value to bounce or even result in divergence, while a small learning rate may lead to slow convergence [30]. Thereby, selecting a proper learning rate is always a critical and tricky issue for learning algorithms. One feasible approach is to adopt the learning rate scheduler, which is able to adjust the learning rate. However, it has to be designed in advance and is unable to adapt to the characteristics of the dataset [31]. Adaptive learning rates, such as Adagrad, can adapt to the data and change with the gradients, which are widely used in the deep learning community [32]. Later, cyclical learning rate (CLR) was proposed to allow the learning rate cyclically vary between reasonable boundary values, which incurs less computational cost and enhance the learning performance [33]. The essence behind CLR originates from the observation that increasing the learning rate allows more rapid traversal of saddle point plateaus, and thus achieves a longer term beneficial effect. Inspired by this study, we propose dynamic learning rate (DLR) within a reasonable range to adapt to the fading channels, in order to further reduce the aggregate error.

In this paper, we consider FL for AI-empowered mobile applications, such as e-health services, which will be supported by 6G networks. Instead of adopting conventional separated-communication-and-computation pattern, we incorporate AirComp to aggregate local models from distributed devices so as to improve the communication efficiency. In AirComp-based schemes, minimizing the resultant aggregate distortion is important to guarantee the performance of AI tasks. To mitigate the wireless distortion measured by mean square error (MSE), we introduce local learning rate to adapt the time-varying channels, which is referred to as the DLR scheme. The receive beamforming optimization is also considered. The technical contributions of this work are summarized below.

- We propose a modified FedAvg by introducing the local learning rate and theoretically prove its convergence. The proof is also applicable for the case where the local learning rates adapt to the channels.
- To our best knowledge, this is the first work to investigate the use of DLR for FL over wireless communications to reduce the aggregate error. This is fundamentally different from existing works that only consider the optimization of wireless resources. We theoretically prove that the MSE can be further decreased by using the DLR scheme.
- Both MISO and MIMO scenarios are considered, and the respective closed-form solution and iterative algorithm are developed. Extensive simulation results demonstrate the effectiveness of the proposed scheme in further reducing the MSE as well as guaranteeing the learning performance on MNIST and CIFAR10 datasets.
- In addition, we present the asymptotic beamforming solution in closed form when the number of transmit/receive antennas tends to infinity. Simulation results verify the theoretical analysis as well as the receive beamforming design.

The outline of this paper is organized as follows: In Section II, we first present the modified FedAvg, and give the mathematical descriptions of AirComp. The DLR for channel adaption is elaborated in Section III. In Section IV and Section V, the DLR optimization problems in MISO and MIMO scenarios are respectively formulated and solved. Next, we present the theoretical asymptotic analysis and, on this basis, propose a near-optimal and closed-form receive beamforming solution in Section VI. Then, in Section VII, numerical simulations are provided to showcase the advantages of the proposed DLR scheme, and verify the asymptotic analysis. Finally, the paper is concluded in Section VIII.

II. Preliminary

In this work, we consider the problem of FL over wireless networks, which consist of K devices with N_d antennas each and an aggregator with N_t antennas. The set of devices are denoted as $\mathcal{K}$. Each device k updates its model based on locally stored data $\mathcal{D}_k$, which cannot be shared with other entities out of latency and privacy concerns. This section first presents the proposed modified FedAvg and then introduces its incorporation with AirComp technique.

A. Modified FedAvg

FL has recently emerged as an effective distributed approach to enable edge devices to collaboratively build a shared learning model with training taken place locally. Assume that the devices have data of equal size. The objective of FL is to find the optimal model $\boldsymbol{w}^*$ that minimizes the loss, i.e.,

$$\boldsymbol{w}^* \triangleq \underset{\boldsymbol{w}}{\operatorname{argmin}} \frac{1}{K}\sum\nolimits_{k=1}^{K} F_k\left(\boldsymbol{w}\right), \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^D$ is a vector containing the model parameters, D is the dimension of the FL model, $F_k\left(\boldsymbol{w}\right)$ is the local loss value based on $\mathcal{D}_k$ at device k, given by

$$F_k\left(\boldsymbol{w}\right) \triangleq \frac{1}{|\mathcal{D}_k|}\sum\nolimits_{n=1}^{|\mathcal{D}_k|} Q_k\left(\boldsymbol{w};\boldsymbol{x}_n,y_n\right), \tag{2}$$

where Q_k is the loss function on the sample $(\boldsymbol{x}_n, y_n)$ with $\boldsymbol{x}_n, y_n$ the input and the label.

Denote $\boldsymbol{w}_t^k$ as the local model of device k at the t-th step. The global model aggregation takes place every E steps, and the set of the communication rounds is denoted by $\mathcal{I}_E = \{nE | n = 1, 2, \ldots\}$. We define a local learning rate $\mu_t^k = \mu_t r_k$ used for local model update, where μ_t and r_k are respectively the unified learning rate for all devices at the t-th step and the learning rate ratio for device k. The learning rate ratio r_k is restricted within the interval of $[r_{\min}, r_{\max}]$, indicating that the learning rate can be neither too large nor too small. Then, we modify the FedAvg algorithm in [12] as

$$\boldsymbol{v}_{t+1}^k = \boldsymbol{w}_t^k - \mu_t r_k \nabla F_k\left(\boldsymbol{w}_t^k, \mathcal{B}_t^k\right), \tag{3}$$

and

$$\boldsymbol{w}_{t+1}^k = \begin{cases} \boldsymbol{v}_{t+1}^k & \text{if} \;\; t+1 \notin \mathcal{I}_E, \\ \dfrac{1}{K}\sum_{k=1}^{K} \dfrac{1}{r_k}\boldsymbol{v}_{t+1}^k & \text{if} \;\; t+1 \in \mathcal{I}_E, \end{cases} \tag{4}$$

where $\boldsymbol{v}_{t+1}^k$ is an additional variable representing the immediate result updated from $\boldsymbol{w}_t^k$. $\nabla F_k\left(\boldsymbol{w}_t^k, \mathcal{B}_t^k\right)$ is the gradient of F_k with respect to $\boldsymbol{w}_t^k$ on the mini-batch of $\mathcal{B}_t^k$. Note that the global model is the weighted average of local models with the weight $\frac{1}{r_k}, \forall k \in \mathcal{K}$.

For any communication round $t = nE$, the global model can be reformulated as

$$\boldsymbol{w}_{nE} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{r_k}\boldsymbol{v}_{nE}^k = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{r_k}\left(\boldsymbol{w}_{(n-1)E} - r_k \Delta\boldsymbol{G}_{nE}^k\right), \tag{5}$$

where $\Delta\boldsymbol{G}_{nE}^k = \sum_{t=(n-1)E}^{nE-1} \mu_t \nabla F_k\left(\boldsymbol{w}_t^k, \mathcal{B}_t^k\right)$ is the cumulative gradients from $\boldsymbol{w}_{(n-1)E}$. To ensure that the coefficient of $\boldsymbol{w}_{(n-1)E}$ in (5) is unit, we let

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{r_k} = 1. \tag{6}$$

Note that the modified FedAvg is equivalent to FedAvg with $E = 1$. For multiple local updates, i.e., $E > 1$, the convergence analysis of the proposed modified FedAvg is provided in the Appendix. For notational brevity, we omit the subscript t of μ_t^k and μ_t hereafter.

B. AirComp

As introduced earlier, AirComp integrates computation and communication by exploiting the waveform superposition property, which harnesses interference to help functional computation [23]. The AirComp comprises three stages: (i) Pre-processing at the transmitters; (ii) Superposition over the air; and (iii) Post-processing at the receiver [34]. In this work, the pre-processing is assumed to be identity mapping. At the n-th communication round, each parameter in a local model is modulated as a symbol, then the symbol vector $\boldsymbol{s}_n^k \triangleq \boldsymbol{v}_{nE}^k \in \mathbb{C}^D$ is obtained. The symbol vector is assumed to be normalized by the unit variance, which is given by $\mathbb{E}\left[\boldsymbol{s}_n^k\left(\boldsymbol{s}_n^k\right)^{\mathrm{H}}\right] = \mathbf{I}$. The configuration of the system under investigation is depicted in Fig. 1.

Considering the multiple access channel property of wireless communication, the received signal is a linear sum of the transmitted signal plus uncertainty. For notation simplicity, the d-th element of $\boldsymbol{s}_n^k$, $\boldsymbol{w}_{nE}$ and $\Delta\boldsymbol{G}_{nE}^k$, i.e., $\boldsymbol{s}_n^k[d]$, $\boldsymbol{w}_{nE}\left[d\right]$ and $\Delta\boldsymbol{G}_{nE}^k[d]$, are written as s_n^k, w_{nE} and ΔG_{nE}^k. Then, at the aggregator the desired signal based on (5) and the received signal after post-processing are respectively written by

$$y_{\mathrm{des}} = w_{nE} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{r_k}\left(w_{(n-1)E} - r_k\Delta G_{nE}^k\right), \tag{7}$$

and

$$\begin{aligned} y &= \sqrt{\eta}\left(\sum\nolimits_{k=1}^{K} A_k s_n^k + B\right) = \sqrt{\eta}\left(\sum\nolimits_{k=1}^{K} A_k v_{nE}^k + B\right) \\ &= \sqrt{\eta}\left(\sum\nolimits_{k=1}^{K} A_k\left(w_{(n-1)E} - r_k\Delta G_{nE}^k\right) + B\right), \end{aligned} \tag{8}$$

where η is a scaling factor. Variables A_k and B depend on the specific scenario settings. In particular, we have $A_k = h_k b_k$ and $B = n$ for SISO; $A_k = \boldsymbol{h}_k^{\mathrm{T}}\boldsymbol{b}_k$ and $B = n$ for MISO. For SIMO and MIMO scenarios, we have $A_k = \boldsymbol{m}^{\mathrm{H}}\boldsymbol{h}_k b_k$, $B = \boldsymbol{m}^{\mathrm{H}}\boldsymbol{n}$, and $A_k = \boldsymbol{m}^{\mathrm{H}}\boldsymbol{H}_k\boldsymbol{b}_k$, $B = \boldsymbol{m}^{\mathrm{H}}\boldsymbol{n}$, respectively. Assume that $h_k\left(\boldsymbol{h}_k, \boldsymbol{H}_k\right)$ is the independent Rayleigh fading channel (vector/matrix) from device k to the aggregator, which is assumed to be block fading and remains constant during the process of model transmission; $b_k\left(\boldsymbol{b}_k\right)$ is the transmit coefficient (vector) at device k; $\boldsymbol{m}$ is the receive beamforming vector at the aggregator; $n\left(\boldsymbol{n}\right)$ is the noise (vector) with power of σ^2.

Instead of obtaining the individual transmit symbol first and then computing at the aggregator, we employ AirComp technique to estimate the desired signal directly. The induced error is derived as

$$\begin{aligned} e \triangleq y_{\mathrm{des}} - y = &\sum_{k=1}^{K}\left(\frac{1}{Kr_k} - \sqrt{\eta}A_k\right)w_{(n-1)E} \\ &- \sum_{k=1}^{K}\left[\left(\frac{1}{K} - \sqrt{\eta}r_k A_k\right)\Delta G_{nE}^k\right] - \sqrt{\eta}B. \end{aligned} \tag{9}$$

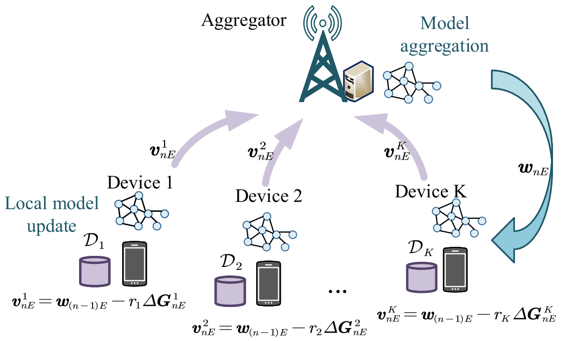

Aggregator
Model aggregation
$\boldsymbol{v}_{nE}^1$
$\boldsymbol{v}_{nE}^2$
$\boldsymbol{v}_{nE}^K$
$\boldsymbol{w}_{nE}$
Device 1
Device 2
Device K
Local model update
$\mathcal{D}_1$
$\mathcal{D}_2$
$\mathcal{D}_K$
...
$\boldsymbol{v}_{nE}^1 = \boldsymbol{w}_{(n-1)E} - r_1 \Delta\boldsymbol{G}_{nE}^1$
$\boldsymbol{v}_{nE}^2 = \boldsymbol{w}_{(n-1)E} - r_2 \Delta\boldsymbol{G}_{nE}^2$
$\boldsymbol{v}_{nE}^K = \boldsymbol{w}_{(n-1)E} - r_K \Delta\boldsymbol{G}_{nE}^K$

Fig. 1. An FL system over the wireless communication system.

According to (9), the aggregate distortion originates from both fading and noise, which are reflected in A_k and B, respectively.

III. DLR for Channel Adaption

Existing works [26]–[29] minimize the aggregate error by means of wireless resource optimization, IRS-based channel reconfiguration, and device selection. These approaches correspond to optimize the transmit coefficient (vector) b_k ($\boldsymbol{b}_k$), the receive beamforming vector $\boldsymbol{m}$, the scaling factor η, and the passive beamforming vector Θ, or select a subset of the devices. Though appears different at first glance, these approaches share a common aim, which is to align the received signals and minimize the induced error. Whereas in this work, we take a different perspective and propose to mitigate the distortion error by optimizing the local learning rate. A strategy is designed to let the local learning rate μ^k, to be exact, the learning rate ratio r_k, adapt to the time-varying wireless environment. Thereby, both μ^k and r_k are dynamic, which are termed as DLR and DLR ratio respectively. It is worthy noting here that the modified FedAvg with the DLR ratios still converges, since the convergence rate is only related to the $r_{\max}$ in Theorem 4.

AirComp results in two sources of error, i.e., the misalignment error due to fading and the additive error due to the noise, when aggregating the model. It has been shown in [35] that, though forcing the fading-related error to zero yielded a suboptimal solution, the obtained performance was close to the optimal MSE in high signal-to-noise ratio region. By further considering multiple antennas, the gap to the optimal performance can be narrowed. As such, we aim at eliminating the fading-related error while mitigating the noise-related error in this work. Based on the aforementioned features of the modified FedAve and AirComp technique, we arrive at the following theorem.

Theorem 1: To eliminate the distortion caused by the fading, we let

$$\sqrt{\eta}\sum\nolimits_{k=1}^{K} A_k = 1, \quad r_k = \frac{1}{K\sqrt{\eta}A_k}, \tag{10}$$

which respectively guarantee that the estimated $\boldsymbol{w}_{nE}$ via AirComp is updated starting from $\boldsymbol{w}_{(n-1)E}$, and that the cumulative gradient $\Delta\boldsymbol{G}_{nE}^k, \forall k \in \mathcal{K}$ for all devices have equal contributions on the update of global model.

Proof: According to (9), the fading-related error is eliminated if $\sum_{k=1}^{K}\left(\frac{1}{Kr_k} - \sqrt{\eta}A_k\right) = 0$ and $\frac{1}{K} - \sqrt{\eta}r_k A_k = 0, \forall k \in \mathcal{K}$. Together with (6), we complete the proof. □

It is observed from Theorem 1 that $\sqrt{\eta}$ and r_k are respectively inversely linear with $\Sigma_{k=1}^{K} A_k$ and $\sqrt{\eta}A_k$. Once the fading-related error is eliminated, the residual aggregate error is the noise-related term $\sqrt{\eta}B$, which can be measured by

$$\mathrm{MSE} = \eta\mathbb{E}\left(\|B\|^2\right). \tag{11}$$

For simplicity, we consider the retransmission mechanism to model the effect of the error on the learning performance. If there exists aggregate error, the probability of retransmission is

$$P = 1 - \exp\left(-\frac{a\mathrm{MSE}}{p_{\mathrm{des}}}\right), \tag{12}$$

where a is the modulation-related parameter [36], p_{des} denotes the power of the desired signal. Intuitively, a larger aggregate error leads to a larger retransmission probability, which results in lower communication efficiency and longer training time.

IV. Problem Formulation

The objective is to minimize the MSE metric given in (11), subject to equality constraints (10), and the boundary constraint of $r_k \in [r_{\min}, r_{\max}]$. In this section, we establish the problem formulations for both MISO and MIMO cases, which are shown in the following to be respectively convex and nonconvex. Conventionally, MISO and MIMO are used to achieve reliable and high-data-rate communication by reaping the diversity and multiplexing gains. In this work, we exploit the spatial degrees-of-freedom provided by MISO/MIMO in the AirComp framework to spatially multiplex functional streams and to reduce the aggregate errors. It is important to note that SISO and SIMO can be regarded as the special cases of MISO and MIMO scenarios, respectively.

A. MISO

In the MISO scenario, the devices equipped with N_d antennas each transmit their local models to the single-antenna aggregator. Under this scenario, we have $A_k = \boldsymbol{h}_k^{\mathrm{T}}\boldsymbol{b}_k$ and $B = n$, where $\boldsymbol{h}_k \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$ and $n \sim \mathcal{CN}\left(0, \sigma^2\right)$. The residual aggregate error by forcing the fading-related error to be zero is measured by MSE, given by

$$\mathrm{MSE} = \eta\mathbb{E}\left(\|B\|^2\right) = \eta\mathbb{E}\left(\|n\|^2\right) = \eta\sigma^2. \tag{13}$$

Since the noise power σ^2 is independent from the optimized variables, the problem can be formulated as

$$\min_{\eta, b_k, r_k} \quad \eta \tag{14a}$$

$$\text{s.t.} \quad \sqrt{\eta}\sum\nolimits_{k=1}^{K} \boldsymbol{h}_k^{\mathrm{T}}\boldsymbol{b}_k = 1, \tag{14b}$$

$$r_k = \frac{1}{K\sqrt{\eta}\boldsymbol{h}_k^{\mathrm{T}}\boldsymbol{b}_k}, \quad \forall k, \tag{14c}$$

$$r_k \in [r_{\min}, r_{\max}], \quad \forall k, \tag{14d}$$

$$\|\boldsymbol{b}_k\|^2 \le P_k, \quad \forall k, \tag{14e}$$

where P_k is the maximum transmit power of device k, and $\boldsymbol{h}_k \in \mathbb{C}^{N_d \times 1}$ is the channel vector between device k and the aggregator. Both equality constraints (14b) and (14c) conspire to guarantee the elimination of the fading-related error as per Theorem 1. Motivated by the uniform-forcing transceiver design in [37], the transmit vector $\boldsymbol{b}_k$ in this work is designed as

$$\boldsymbol{b}_k = \frac{\left(\boldsymbol{h}_k^{\mathrm{T}}\right)^{\mathrm{H}}}{K\sqrt{\eta}\left\|\boldsymbol{h}_k\right\|^2 r_k}. \tag{15}$$

Power constraint (14e) further suggests that $\frac{1}{K^2 P_k r_k^2 \|\boldsymbol{h}_k\|^2} \leq \eta$, and thus we have

$$\eta = \max_k \frac{1}{K^2 P_k r_k^2 \left\|\boldsymbol{h}_k\right\|^2}. \tag{16}$$

We learn from (14c) that $\boldsymbol{h}_k^{\mathrm{T}} \boldsymbol{b}_k = \frac{1}{K\sqrt{\eta} r_k}$. By substituting it back to (14b), we have $\sum_{k=1}^{K} \frac{1}{K r_k} = 1$. Thereby, problem (14) is equivalent to

$$\min_{r_k} \ \max_k \frac{1}{K^2 P_k r_k^2 \left\|\boldsymbol{h}_k\right\|^2} \tag{17a}$$

$$\text{s.t.} \ \sum_{k=1}^{K} \frac{1}{K r_k} = 1, \tag{17b}$$

$$r_k \in \left[r_{\min}, r_{\max}\right], \quad \forall k. \tag{17c}$$

Remark 1: As a special case of the MISO scenario where $A_k = h_k b_k$ and $B = n$, the problem formulated under the SISO case is similar to (17). The only difference lies in the channel and transmit coefficients, which are both complex scalars instead of vectors.

B. MIMO

In the MIMO scenario, each device and the aggregator are equipped with N_d and N_t antennas, respectively. The terms A_k and B in (8) are $A_k = \boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_k \boldsymbol{b}_k$, $B = \boldsymbol{m}^{\mathrm{H}} \boldsymbol{n}$ with $\boldsymbol{m}$ the receive beamforming vector, where each element of $\boldsymbol{H}_k$ follows $\mathcal{CN}(0,1)$ and $\boldsymbol{n} \sim \mathcal{CN}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right)$. Accordingly, the aggregate error measured by MSE is expressed as

$$\mathrm{MSE} = \eta \mathbb{E}\left(\|B\|^2\right) = \eta \mathbb{E}\left(\left\|\boldsymbol{m}^{\mathrm{H}} \boldsymbol{n}\right\|^2\right) = \sigma^2 \|\boldsymbol{m}\|^2 \eta. \tag{18}$$

Based on Theorem 1, the MSE minimization problem can be formulated as

$$\min_{\boldsymbol{m}, \eta, r_k, \boldsymbol{b}_k} \ \|\boldsymbol{m}\|^2 \eta \tag{19a}$$

$$\text{s.t.} \ \sqrt{\eta} \sum_{k=1}^{K} \boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_k \boldsymbol{b}_k = 1, \tag{19b}$$

$$r_k = \frac{1}{K\sqrt{\eta} \boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_k \boldsymbol{b}_k}, \quad \forall k, \tag{19c}$$

$$r_k \in \left[r_{\min}, r_{\max}\right], \quad \forall k, \tag{19d}$$

$$\left\|\boldsymbol{b}_k\right\|^2 \leq P_k, \quad \forall k, \tag{19e}$$

where $\boldsymbol{b}_k \in \mathbb{C}^{N_d}$, $\boldsymbol{H}_k \in \mathbb{C}^{N_t \times N_d}$, and $\boldsymbol{m} \in \mathbb{C}^{N_t}$ are the transmit beamforming vector at device k, the channel matrix between the aggregator and device k, and the receive beamforming vector at the aggregator, respectively. Similarly, the transmit beamforming vector is designed as

$$\boldsymbol{b}_k = \frac{\boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{m}}{K\sqrt{\eta} r_k \left\|\boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_k\right\|^2}. \tag{20}$$

Power constraint (19e) further indicates:

$$\eta = \max_k \frac{1}{K^2 P_k r_k^2 \left\|\boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_k\right\|^2}. \tag{21}$$

Similar to the MISO scenario, the ratio r_k satisfies $\sum_{k=1}^{K} \frac{1}{K r_k} = 1$, which can be easily derived from the equality constraints (14b) and (14c). Thus, problem (19) can then be rewritten as

$$\min_{\boldsymbol{m}, r_k} \ \max_k \frac{\|\boldsymbol{m}\|^2}{K^2 P_k r_k^2 \left\|\boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_k\right\|^2} \tag{22a}$$

$$\text{s.t.} \ \sum_{k=1}^{K} \frac{1}{K r_k} = 1, \tag{22b}$$

$$r_k \in \left[r_{\min}, r_{\max}\right], \quad \forall k. \tag{22c}$$

Proposition 1: Problem (22) is equivalent to

$$\min_{\boldsymbol{m}, r_k} \ \max_k \frac{\|\boldsymbol{m}\|^2}{K^2 P_k r_k^2 \left\|\boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_k\right\|^2} \tag{23a}$$

$$(22\text{b}), \ (22\text{c}),$$

$$\|\boldsymbol{m}\| = 1. \tag{23b}$$

Proof: For any vector $\boldsymbol{m}$, it is the multiplication of its ℓ_2-norm and the unit direction vector, i.e., $\boldsymbol{m} = \|\boldsymbol{m}\| \frac{\boldsymbol{m}}{\|\boldsymbol{m}\|}$. Denoting $\tilde{\boldsymbol{m}} = \frac{\boldsymbol{m}}{\|\boldsymbol{m}\|}$, the objective function of problem (22) is equivalent to $\max\limits_k \frac{\|\tilde{\boldsymbol{m}}\|^2}{K^2 P_k r_k^2 \left\|\tilde{\boldsymbol{m}}^{\mathrm{H}} \boldsymbol{H}_k\right\|^2}$, where $\|\tilde{\boldsymbol{m}}\| = 1$. This completes the proof. □

Remark 2: The SIMO scenario is a special case of the MIMO case, where $A_k = \boldsymbol{m}^{\mathrm{H}} \boldsymbol{h}_k b_k$, $B = \boldsymbol{m}^{\mathrm{H}} \boldsymbol{n}$. The problem formulated under SIMO case is the same as problem (23) except that the channel and transmit coefficients are respectively a vector and a scalar.

V. DLR Optimization

In this section, we develop two algorithms to solve problems (17) and (23), respectively. For the MISO scenario, problem (17) is convex and a closed-form solution is derived. For the nonconvex problem (23), we decompose it into two sub-problems and propose an iterative method. Note that the channel gain and the full channel information of the edge devices are respectively needed at the aggregator in MISO and MIMO scenarios.

A. MISO

Obviously, problem (17) is convex. For notational simplicity, we let $l_k = \frac{1}{r_k}$ and $c_k = \frac{1}{K\sqrt{P_k}\|\boldsymbol{h}_k\|}$. As such, problem (17) can be written as

$$\min_{l_k} \ \max_k \left(c_k l_k\right)^2 \tag{24a}$$

$$\text{s.t.} \ \sum_{k=1}^{K} l_k = K, \tag{24b}$$

$$\frac{1}{r_{\max}} \leq l_k \leq \frac{1}{r_{\min}}, \quad \forall k. \tag{24c}$$

The solution that minimizes $\max_k (c_k l_k)^2$ also minimizes $\max_k c_k l_k$. Consequently, problem (24) is further equivalent to the following problem

$$\min_{l_k} \max_{k} c_k l_k \quad \text{s.t.} \, (24\text{b}), \, (24\text{c}), \tag{25}$$

which is a typical linear programming problem.

Assuming $k = \arg\max_i c_i l_i$, we then have $c_i l_i \le c_k l_k, \forall i \in \mathcal{K}$. Following the equation constraint (24b), we have

$$K = \sum\nolimits_{i=1}^{K} l_i \le \sum\nolimits_{i=1}^{K} \frac{c_k l_k}{c_i} = c_k l_k \sum\nolimits_{i=1}^{K} \frac{1}{c_i}. \tag{26}$$

Thus

$$c_k l_k \ge \frac{K}{\sum_{i=1}^{K} \frac{1}{c_i}}. \tag{27}$$

Denoting the MSE obtained with and without considering DLR as $\mathrm{MSE}^{\mathrm{d}}$ and $\mathrm{MSE}^{\mathrm{n}}$, we have the following Theorem.

Theorem 2: In a MISO system, the MSE is lower bounded by $\mathrm{MSE}^{\mathrm{lb}} = \frac{\sigma^2}{\left(\sum_{i=1}^{K} \sqrt{P_i} \|\boldsymbol{h}_i\|\right)^2}$*, and we always have*

$$\mathrm{MSE}^{\mathrm{n}} \overset{(a)}{\ge} \mathrm{MSE}^{\mathrm{d}} \overset{(b)}{\ge} \mathrm{MSE}^{\mathrm{lb}}, \tag{28}$$

where equality (a) *holds if and only if* $\sqrt{P_i} \left\|\boldsymbol{h}_i\right\| = \sqrt{P_j} \left\|\boldsymbol{h}_j\right\|, \forall i, j \in \mathcal{K}$*, and equality* (b) *holds if and only if* $r_i \sqrt{P_i} \left\|\boldsymbol{h}_i\right\| = r_j \sqrt{P_j} \left\|\boldsymbol{h}_j\right\|, \forall i, j \in \mathcal{K}$.

Proof: According to (27), the lower bound of $c_k l_k$ is $\frac{K}{\sum_{i=1}^{K} \frac{1}{c_i}}$. Hence, the MSE is lower bounded by

$$\mathrm{MSE}^{\mathrm{lb}} = \left(\frac{K}{\sum_{i=1}^{K} \frac{1}{c_i}}\right)^2 \sigma^2 = \frac{\sigma^2}{\left(\sum_{i=1}^{K} \sqrt{P_i} \left\|\boldsymbol{h}_i\right\|\right)^2}. \tag{29}$$

The lower bound is achieved if and only if $c_i l_i = c_j l_j, \forall i, j \in \mathcal{K}$, which means that $r_i \sqrt{P_i} \left\|\boldsymbol{h}_i\right\| = r_j \sqrt{P_j} \left\|\boldsymbol{h}_j\right\|, \forall i, j \in \mathcal{K}$.

With loss of generality, we assume that c_k is in a descending order, i.e., $c_i \ge c_j, \forall i > j$. Thus, the MSE without considering DLR can be readily obtained as

$$\mathrm{MSE}^{\mathrm{n}} = c_1^2 \sigma^2 = \frac{\sigma^2}{K^2 P_1 \left\|\boldsymbol{h}_1\right\|^2}. \tag{30}$$

When taking the DLR into consideration, we can always find a feasible set of coefficients $[l_1, l_2, \ldots, l_K]$, which guarantees $c_1 \ge c_1 l_1$ and $c_1 l_1 = \max_i c_i l_i$. Consequently, we have

$$\mathrm{MSE}^{\mathrm{d}} = c_1^2 l_1^2 \sigma^2 \le c_1^2 \sigma^2 = \mathrm{MSE}^{\mathrm{n}}, \tag{31}$$

the equality of which holds if and only if when $c_i = c_j, \forall i, j \in \mathcal{K}$, i.e., $\sqrt{P_i} \left\|\boldsymbol{h}_i\right\| = \sqrt{P_j} \left\|\boldsymbol{h}_j\right\|, \forall i, j \in \mathcal{K}$. Finally, we complete the proof. □

Theorem 2 shows that optimizing DLR can further reduce the MSE, since the DLR ratio offers an additional dimension to narrow the gap among $r_i \sqrt{P_i} \left\|\boldsymbol{h}_i\right\|, \forall i \in \mathcal{K}$. The theorem also provides the condition of achieving the lower bound of MSE in the MISO case. Based on the above analysis, the optimal solution on l_i of problem (24) under constraint (24c) is given by

$$l_i = \mathrm{clip}\left(\frac{c_k l_k}{c_i}, \left[\frac{1}{r_{\max}}, \frac{1}{r_{\min}}\right]\right), \tag{32}$$

where $\sum_{i=1}^{K} l_i = K$. Operation $\mathrm{clip}\,(x, [a, b])$ truncates x to the specified interval of $[a, b]$.

Algorithm 1 Optimal Learning Rate for MISO

```
Input: $c_k$, $\mathrm{obj} = \frac{\max_k (c_k)}{r_{\min}}$, $\delta$
Output: $l_i^{\mathrm{opt}}, \mathrm{obj}$
Initialize: $l_k^{\min} = \frac{1}{r_{\max}}$, $l_k^{\max} = \frac{1}{r_{\min}}$
for $k = 1 : K$
    while $|\sum_{i=1}^{K} l_i - K| \le \delta$
        $l_k = (l_k^{\max} + l_k^{\min})/2$
        $l_i = \mathrm{clip}\left(\frac{c_k l_k}{c_i}, [\frac{1}{r_{\max}}, \frac{1}{r_{\min}}]\right), \forall i \in \mathcal{K}$
        if $(\sum_{i=1}^{K} l_i \ge K)$
            $l_k^{\max} \leftarrow l_k$
        else
            $l_k^{\min} \leftarrow l_k$
    if $\mathrm{obj} \ge \max_i c_i l_i$
        $\mathrm{obj} = \max_i c_i l_i$
        $l_i^{\mathrm{opt}} = l_i, \forall i \in \mathcal{K}$
```

The overall procedure to solve problem (24) is shown in Algorithm 1. According to (32), the proposed scheme needs to know the user index k with the maximum value $c_k l_k$. To find the device index, we exhaustively search all devices, which shows that the number of iterations in the outer layer is K. For a given device index k, the bisection technique is applied to find a solution l_k, the complexity of which is $\mathcal{O}(\log_2(1/\delta_1))$ with accuracy δ_1.

B. MIMO

Denoting $c_k = \frac{\|\boldsymbol{m}\|}{K \sqrt{P_k} \|\boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_k\|}$, $l_k = \frac{1}{r_k}$, and $c_k l_k = \max_i c_i l_i$, we arrive at the following theorem.

Theorem 3: In a MIMO system given any feasible $\boldsymbol{m}$*, the MSE is lower bounded by* $\mathrm{MSE}^{\mathrm{lbm}} = \frac{\sigma^2}{\left(\sum_{i=1}^{K} \sqrt{P_i} \|\boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_i\|\right)^2}$*, and we always have*

$$\mathrm{MSE}^{\mathrm{n}} \overset{(a)}{\ge} \mathrm{MSE}^{\mathrm{d}} \overset{(b)}{\ge} \mathrm{MSE}^{\mathrm{lbm}}, \tag{33}$$

where the equalities of (a) *and* (b) *hold when* $\sqrt{P_i} \left\|\boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_i\right\| = \sqrt{P_j} \left\|\boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_j\right\|$*, and* $r_i \sqrt{P_i} \left\|\boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_i\right\| = r_j \sqrt{P_j} \left\|\boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_j\right\|, \forall i, j \in \mathcal{K}$*, respectively.*

Proof: Denote the equivalent channel vector $\boldsymbol{h}_i' = \boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_i \in \mathbb{C}^{N_d}$. The MIMO scenario given $\boldsymbol{m}$ is equivalent to the MISO scenario with the channel of $\boldsymbol{h}_i', \forall i \in \mathcal{K}$. Based on Theorem 2, we complete the proof. □

It is shown in Theorem 3 that the MSE in MIMO scenario given the feasible $\boldsymbol{m}$ is lower bounded, and can be further reduced in consideration of DLR. In the MIMO scenario, problem (23) is difficult due to the nonconvex objective (23a) and constraint (23b). By introducing an auxiliary variable τ, problem (23) is further equivalent to the following problem:

$$\min_{\boldsymbol{m}, r_k, \tau} \quad \tau \tag{34a}$$

$$\text{s.t.} \quad \|\boldsymbol{m}\|^2 \le \tau K^2 P_k \left\|\boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_k\right\|^2 r_k^2, \, \forall k, \tag{34b}$$

$$(22\text{b}), (22\text{c}), (23\text{b}).$$

To solve problem (34), we decompose it into two sub-problems by alternately fixing the DLR ratio r_k and the receive beamforming vector $\boldsymbol{m}$.

Given the DLR ratio r_k, the original problem (34) is reduced into the following sub-problem:

$$\min_{\boldsymbol{m},\tau} \ \tau \quad \text{s.t.} \ \ (23\text{b}), (34\text{b}), \tag{35}$$

which is nonconvex due to constraints (23b) and (34b). To solve problem (35), we have the following lemma.

Lemma 1: Define $\boldsymbol{A}_k = \boldsymbol{H}_k \boldsymbol{H}_k^{\mathrm{H}}$, *which is semidefinite. If* $\det \boldsymbol{A}_k > 0$, *the range of* $\frac{\|\boldsymbol{m}\|^2}{K^2 P_k \|\boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_k\|^2 r_k^2}$ *is* $\left[\frac{1}{K^2 P_k r_k^2 \lambda_{\max}}, \frac{1}{K^2 P_k r_k^2 \lambda_{k,\min}}\right]$; *otherwise* $\left[\frac{1}{K^2 P_k r_k^2 \lambda_{k,\max}}, \infty\right]$, *where* $\lambda_{k,\max}$ *and* $\lambda_{k,\min}$ *are the maximum and minimum eigenvalues of* $\boldsymbol{A}_k$, *respectively.*

Proof: By using the Rayleigh-Ritz property of matrix $\boldsymbol{A}_k = \boldsymbol{H}_k \boldsymbol{H}_k^{\mathrm{H}}$, we have $\frac{\|\boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_k\|^2}{\|\boldsymbol{m}\|^2} \in [\lambda_{k,\min}, \lambda_{\mathrm{k,max}}]$. Hence, we complete the proof. □

Remark 3: Let $\tau_k^{\mathrm{low}} = \frac{1}{K^2 P_k r_k^2 \lambda_{k,\max}}$, $\tau_k^{\mathrm{up}} = \frac{1}{K^2 P_k r_k^2 \lambda_{k,\min}}$ if $\lambda_{\min} > 0$; otherwise $\tau_k^{\mathrm{up}} = \infty$, and then $\frac{\|\boldsymbol{m}\|^2}{K^2 P_k \|\boldsymbol{m}^{\mathrm{H}} \boldsymbol{H}_k\|^2 r_k^2} \in \left[\tau_k^{\mathrm{low}}, \tau_k^{\mathrm{up}}\right]$. The necessary condition of τ that guarantees the feasibility of problem (35) is $\tau \in \left[\tau^{\mathrm{low}}, \tau^{\mathrm{up}}\right]$, where

$$\tau^{\mathrm{low}} = \min_k \left(\tau_k^{\mathrm{low}}\right), \ \tau^{\mathrm{up}} = \max_k \left(\tau_k^{\mathrm{up}}\right). \tag{36}$$

For any given $\tau \in \left[\tau^{\mathrm{low}}, \tau^{\mathrm{up}}\right]$, sub-problem (35) is interpreted as finding a receive beamforming vector $\boldsymbol{m}$ that makes the sub-problem feasible, which is a check problem. By introducing $\boldsymbol{M} = \boldsymbol{m}\boldsymbol{m}^{\mathrm{H}}$, the sub-problem given τ is converted to

$$\min_{\boldsymbol{M}} \ 0 \tag{37a}$$

$$\text{s.t.} \ \operatorname{Tr}(\boldsymbol{M}) \leq \tau K^2 P_k \operatorname{Tr}\left(\boldsymbol{M}\boldsymbol{H}_k \boldsymbol{H}_k^{\mathrm{H}}\right) r_k^2, \ \forall k, \tag{37b}$$

$$\boldsymbol{M} \succcurlyeq 0, \tag{37c}$$

$$\operatorname{Tr}(\boldsymbol{M}) = 1, \tag{37d}$$

$$\operatorname{rank}(\boldsymbol{M}) = 1. \tag{37e}$$

The only difficulty of the above problem lies in the rank one constraint (37e). The problem can be solved by first dropping the constraint (37e) to obtain solution $\boldsymbol{M}^*$, and then obtaining the $\boldsymbol{m}^*$ using the eigenvector approximation method or the randomization technique [38]. Such approach that replaces $\boldsymbol{m}\boldsymbol{m}^{\mathrm{H}}$ by $\boldsymbol{M}$ is termed as the semidefinite relaxation (SDR) method. However, this approach is sub-optimal especially when $\boldsymbol{M}$ is large. To guarantee the rank one constraint, we utilize a DC representation [27], [39], and convert problem (37) to

$$\min_{\boldsymbol{M}} \ \operatorname{Tr}(\boldsymbol{M}) - \|\boldsymbol{M}\|_2 \quad \text{s.t.} \ (37\text{b}), (37\text{c}), (37\text{d}), \tag{38}$$

where $\|\boldsymbol{M}\|_2$ is the maximum eigenvalue of matrix $\boldsymbol{M}$. Rank one constraint is satisfied if $\operatorname{Tr}(\boldsymbol{M}) - \|\boldsymbol{M}\|_2 = 0$. The problem can be efficiently solved by DC programming with complexity of $\mathcal{O}\left(N_t^6\right)$.

To find the solution τ, we utilize the bisection technique. Specifically, the interval $\left[\tau^{\mathrm{low}}, \tau^{\mathrm{up}}\right]$ is divided into two sub-intervals of $\left[\tau^{\mathrm{low}}, \tau\right]$ and $[\tau, \tau^{\mathrm{up}}]$ with $\tau = \left(\tau^{\mathrm{low}} + \tau^{\mathrm{up}}\right)/2$. If problem (38) is solved given τ, then the solution is within $\left[\tau^{\mathrm{low}}, \tau\right]$; otherwise within $[\tau, \tau^{\mathrm{up}}]$. Repeatedly check the problem until $\tau^{\mathrm{up}} - \tau^{\mathrm{low}} < \delta_2$, where δ_2 is the accuracy of bisection search. Such bisection technique involves $\log_2 \frac{1}{\delta_2}$ repetitive operations and, hence, solving the problem (35) requires the computational complexity of $\mathcal{O}\left(N_t^6 \log_2 \frac{1}{\delta_2}\right)$. The algorithm is summarized in Algorithm 2.

Algorithm 2 Receive Beamforming Design for MIMO

```
Input: $\boldsymbol{H}_k^{\mathrm{H}}$, $r_k$, $\delta$
Output: $\boldsymbol{m}$, $\tau$
calculate $\tau^{\mathrm{low}}$, $\tau^{\mathrm{up}}$ based on (36)
while $\left(\tau^{\mathrm{up}} - \tau^{\mathrm{low}}\right) > \delta$
    $\tau = \left(\tau^{\mathrm{up}} + \tau^{\mathrm{low}}\right)/2$
    if problem (38) is infeasible
        $\tau^{\mathrm{up}} \leftarrow \tau$
    else
        $\tau^{\mathrm{low}} \leftarrow \tau$
solve problem (38) to obtain $\boldsymbol{m}$
```

Algorithm 3 Iterative Learning Rate and Receive Beamforming

```
Input: $\boldsymbol{H}_k$
Output: $\boldsymbol{m}$, $r_k$, $\tau$
Initialize: $r_k = 1$
do loop
    given $r_k$, solve problem (35) using Algorithm 2
    given $\boldsymbol{m}$, solve problem (39) using Algorithm 1
until $\tau$ converges
```

Given the obtained receive beamforming vector $\boldsymbol{m}$, the original problem (34) is reduced into the following sub-problem:

$$\min_{r_k,\tau} \ \tau \quad \text{s.t.} \ \ (22\text{b}), (22\text{c}), (34\text{b}). \tag{39}$$

Denoting the equivalent channel vector $\boldsymbol{h}_k'$ as $\boldsymbol{m}^{\mathrm{H}}\boldsymbol{H}_k$, the sub-problem (39) can be transformed into the problem under the MISO case with channel vector $\boldsymbol{h}_k'$. Note that the sub-problem under the SIMO case is equivalent to that of the SISO case by defining equivalent channel coefficient h_k' as $\boldsymbol{m}^{\mathrm{H}}\boldsymbol{h}_k$. Therefore, letting $l_k = \frac{1}{r_k}$ and $c_k = \frac{1}{K\sqrt{P_k}\left\|\boldsymbol{h}_k'\right\|}$, we can readily derive the closed-form solution of $l_i = \operatorname{clip}\left(\frac{c_k l_k}{c_i}, [\frac{1}{r_{\max}}, \frac{1}{r_{\min}}]\right)$, where $\sum_{i=1}^K l_i = K$.

Thus, the sub-optimal solution of problem (34) can be obtained by alternatively solving sub-problems (38) and (39). For each iteration, the computational complexity is $\mathcal{O}\left(K \log_2 \frac{1}{\delta_1} + N_t^6 \log_2 \frac{1}{\delta_2}\right)$, where δ_1 and δ_2 are the accuracy of solving DLR ratio r_k and receive beamforming vector $\boldsymbol{m}$, respectively. The proposed iterative method is summarized in Algorithm 3.

VI. Asymptotic Analysis and Receive Beamforming Design

This section presents the theoretical analysis of the MSE and the DLR ratio in the MISO and MIMO scenarios when N_d

and N_t increase to infinity. Based on the asymptotic analysis, we give a near-optimal and closed-form receive beamforming solution. In the following, we assume that the maximum power ratio of any two devices, i.e., $P_i/P_j, \forall i \neq j$, is within the interval of $\left[\frac{r_{\min}^2}{r_{\max}^2}, \frac{r_{\max}^2}{r_{\min}^2}\right]$.

A. MISO

In the MISO case, we present asymptotic analysis when the number of antennas N_d at the devices goes to infinity. Since the channels follow $\boldsymbol{h}_k \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$, we have

$$\|\boldsymbol{h}_k\| \to \sqrt{N_d}, \tag{40}$$

$$c_k = \frac{1}{K\sqrt{P_k}\,\|\boldsymbol{h}_k\|} \to \frac{1}{K\sqrt{P_k}\sqrt{N_d}}, \tag{41}$$

which suggest that $\frac{c_i}{c_j} \in \left[\frac{r_{\min}}{r_{\max}}, \frac{r_{\max}}{r_{\min}}\right], \forall i, j \in \mathcal{K}$. As a consequence, the equality in (b) of (28) is guaranteed, and the lower bound $\mathrm{MSE}^{\mathrm{lb}}$ can be achieved according to Theorem 2. Accordingly, the MSE and r_k become

$$\mathrm{MSE} = \frac{\sigma^2}{\left(\sum_{i=1}^K \sqrt{P_i}\,\|\boldsymbol{h}_i\|\right)^2} \to \frac{\sigma^2}{\left(\sum_{i=1}^K \sqrt{P_i}\right)^2 N_d}, \tag{42}$$

$$r_k = \frac{1}{l_k} = \frac{\sum_{i=1}^K \sqrt{P_i}\,\|\boldsymbol{h}_i\|}{K\sqrt{P_k}\,\|\boldsymbol{h}_k\|} \to \frac{\sum_{i=1}^K \sqrt{P_i}}{K\sqrt{P_k}}, \tag{43}$$

where the achieved MSE is inversely proportional to $\left(\sum_{i=1}^K \sqrt{P_i}\right)^2$ and N_d. In particular, if the devices have equal maximum transmit power, i.e., $P_i = P_j = P, \forall i, j \in \mathcal{K}$, we have $\mathrm{MSE} \to \frac{\sigma^2}{PK^2N_d}$ and $r_k \to 1$.

B. MIMO

In the MIMO case, both devices and the aggregator have multiple antennas. This sub-section presents the analyses when N_d and N_t go to infinity, respectively. The equivalent channel vector between device i and the aggregator is denoted as $\boldsymbol{h}_i' = \boldsymbol{m}^{\mathrm{H}}\boldsymbol{H}_i$, where $\boldsymbol{H}_i \in \mathbb{C}^{N_t \times N_d}$. Since $\boldsymbol{H}_i\boldsymbol{H}_i^{\mathrm{H}}$ is semidefinite and $\|\boldsymbol{m}\| = 1$, the norm of $\boldsymbol{h}_i'$ satisfies

$$\left\|\boldsymbol{h}_i'\right\| = \sqrt{\boldsymbol{m}^{\mathrm{H}}\boldsymbol{H}_i\boldsymbol{H}_i^{\mathrm{H}}\boldsymbol{m}} = \sqrt{\frac{\boldsymbol{m}^{\mathrm{H}}\boldsymbol{H}_i\boldsymbol{H}_i^{\mathrm{H}}\boldsymbol{m}}{\boldsymbol{m}^{\mathrm{H}}\boldsymbol{m}}}. \tag{44}$$

According to the Rayleigh-Ritz property, $\left\|\boldsymbol{h}_i'\right\|$ is within the range of $\left[\sqrt{\lambda_{i,\min}}, \sqrt{\lambda_{i,\max}}\right]$, where $\lambda_{i,\min}$ and $\lambda_{i,\max}$ are respectively the minimum and maximum eigenvalues of $\boldsymbol{H}_i\boldsymbol{H}_i^{\mathrm{H}}$.

First, we give the analysis when $N_d \to \infty$ for $N_d > N_t$. In this case, the channel between device i and each antenna at the aggregator is asymptotically orthogonal. Consequently, we have

$$\boldsymbol{H}_i\boldsymbol{H}_i^{\mathrm{H}} \to N_d\boldsymbol{I}_{N_t \times N_t}, \tag{45}$$

whose eigenvalues share the same value of N_d. Thus, for any beamforming vector $\boldsymbol{m}$ with $\|\boldsymbol{m}\| = 1$, we have

$$\left\|\boldsymbol{h}_i'\right\| \to \sqrt{N_d},\ \forall i \in \mathcal{K}, \tag{46}$$

which suggests that $\frac{c_i}{c_j} \in \left[\frac{r_{\min}}{r_{\max}}, \frac{r_{\max}}{r_{\min}}\right], \forall i, j \in \mathcal{K}$. The equality condition of (b) in (33) is also guaranteed. Based on Theorem 3, the MSE and r_k are obtained such that

$$\mathrm{MSE} = \frac{\sigma^2}{\left(\sum_{i=1}^K \sqrt{P_i}\,\left\|\boldsymbol{h}_i'\right\|\right)^2} \to \frac{\sigma^2}{\left(\sum_{i=1}^K \sqrt{P_i}\right)^2 N_d}, \tag{47}$$

$$r_k = \frac{1}{l_k} = \frac{\sum_{i=1}^K \sqrt{P_i}\,\left\|\boldsymbol{h}_i'\right\|}{K\sqrt{P_k}\,\left\|\boldsymbol{h}_k'\right\|} \to \frac{\sum_{i=1}^K \sqrt{P_i}}{K\sqrt{P_k}}. \tag{48}$$

Note that the MSE achieved is inversely proportional to N_d, irrespective of N_t when $N_d \to \infty$ for $N_d > N_t$. The reason is that $\boldsymbol{H}_i$ has full row rank with equal singular value of $\sqrt{N_d}$, which indicates that the equivalent channel $\boldsymbol{h}_i'$ has the same power of N_d regardless of N_t. When the maximum power of all devices is equal, we have $\mathrm{MSE} \to \frac{\sigma^2}{PK^2N_d}$ and $r_k \to 1$.

Next, the analysis of the MSE when $N_t \to \infty$ for $N_t > N_d$ is provided. The channels between each antenna of each device i and the aggregator are asymptotically orthogonal with power of N_t. Consequently, the column spaces spanned by $\boldsymbol{H}_i, \forall i \in \mathcal{K}$ are orthogonal, i.e.,

$$\mathrm{span}\left(\boldsymbol{H}_i\right) \perp \mathrm{span}\left(\boldsymbol{H}_j\right),\ \forall i \neq j, \tag{49}$$

which suggests that $\boldsymbol{h}_i \perp \boldsymbol{h}_j$ for any $\boldsymbol{h}_i \in \mathrm{span}\left(\boldsymbol{H}_i\right)$ and $\boldsymbol{h}_j \in \mathrm{span}\left(\boldsymbol{H}_j\right)$.

The rank of $\boldsymbol{H}_i$ is $r = \mathrm{rank}\left(\boldsymbol{H}_i\right) = N_d$. By using singular value decomposition (SVD), we readily have

$$\boldsymbol{H}_i\boldsymbol{H}_i^{\mathrm{H}} = \boldsymbol{U}_i\boldsymbol{\Sigma}^2\boldsymbol{U}_i^{\mathrm{H}} \to \boldsymbol{U}_i\begin{bmatrix} N_t\boldsymbol{I}_{N_d \times N_d} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{U}_i^{\mathrm{H}}, \tag{50}$$

where the first N_d columns of $\boldsymbol{U}_i$ could be $\boldsymbol{U}_{i,r} = \left[\frac{\boldsymbol{H}_i[:,1]}{\|\boldsymbol{H}_i[:,1]\|}, \frac{\boldsymbol{H}_i[:,2]}{\|\boldsymbol{H}_i[:,2]\|}, \ldots, \frac{\boldsymbol{H}_i[:,N_d]}{\|\boldsymbol{H}_i[:,N_d]\|}\right]$, where $\boldsymbol{H}_i\left[:, n\right]$ is the n-th column of $\boldsymbol{H}_i$. Thereby, the minimum and maximum eigenvalues of $\boldsymbol{H}_i\boldsymbol{H}_i^{\mathrm{H}}$ are respectively $\lambda_{i,\min} = 0$, $\lambda_{i,\max} = N_t$. Based on the above observations, the receive beamforming vector can be designed as

$$\boldsymbol{m} = \frac{\widetilde{\boldsymbol{h}}_1 + \widetilde{\boldsymbol{h}}_2 + \ldots + \widetilde{\boldsymbol{h}}_K}{\left\|\widetilde{\boldsymbol{h}}_1 + \widetilde{\boldsymbol{h}}_2 + \ldots + \widetilde{\boldsymbol{h}}_K\right\|}, \tag{51}$$

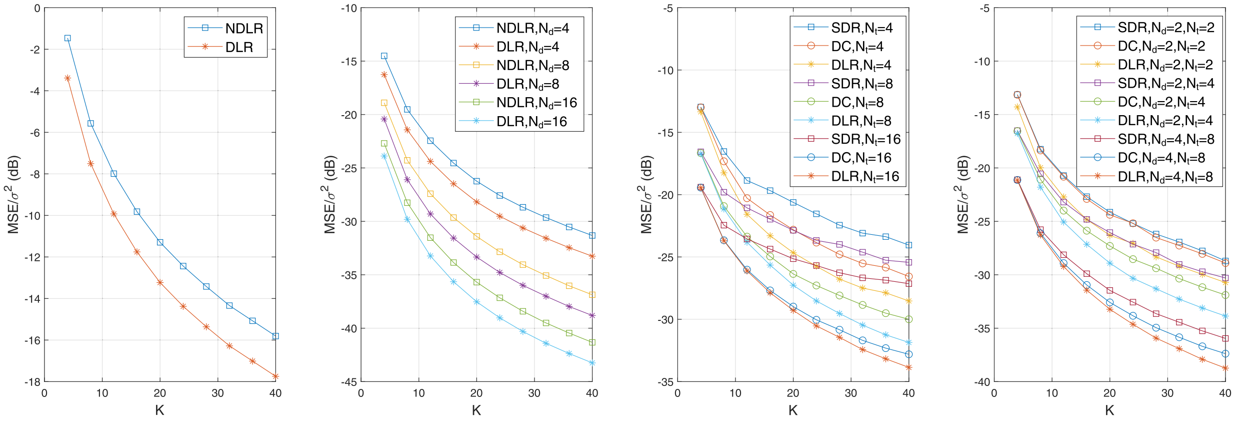
where $\widetilde{\boldsymbol{h}}_i$ is one of the column vectors of $\boldsymbol{U}_{i,r}$. Then, the equivalent channel can be expressed as

$$\boldsymbol{h}_i' = \boldsymbol{m}^{\mathrm{H}}\boldsymbol{H}_i \to \frac{\widetilde{\boldsymbol{h}}_i^{\mathrm{H}}\boldsymbol{H}_i}{\sqrt{K}} = \frac{\sqrt{N_t}}{\sqrt{K}}\boldsymbol{I}_e, \tag{52}$$

where $\boldsymbol{I}_e = [\underbrace{0, \ldots,}_{e-1} 1, \underbrace{\ldots, 0}_{N_d - e}]$ if $\widetilde{\boldsymbol{h}}_i$ is the e-th column of $\boldsymbol{U}_{i,r}$. Hence, we have

$$\left\|\boldsymbol{h}_i'\right\| = \left\|\frac{\sqrt{N_t}}{\sqrt{K}}\boldsymbol{I}_e\right\| \to \sqrt{\frac{N_t}{K}}. \tag{53}$$

Note that in the SIMO scenario where $N_d = 1$, the equivalent channel has $h_i' = \boldsymbol{m}^{\mathrm{H}}\boldsymbol{h}_i = \sqrt{\frac{N_t}{K}}$. According to Theorem 3, the equality (b) in (33) is satisfied, and the MSE



(a) (b) (c) (d)

Fig. 2. The aggregate error versus the number of devices K. (a) SISO scenario. (b) MISO scenario. (c) SIMO scenario. (d) MIMO scenario.

and r_k can be obtained as

$$\text{MSE} = \frac{\sigma^2}{\left(\sum_{i=1}^{K}\sqrt{P_i}\left\|\boldsymbol{h}_i'\right\|\right)^2} \to \frac{K\sigma^2}{\left(\sum_{i=1}^{K}\sqrt{P_i}\right)^2 N_t}, \tag{54}$$

$$r_k = \frac{1}{l_k} = \frac{\sum_{i=1}^{K}\sqrt{P_i}\left\|\boldsymbol{h}_i'\right\|}{K\sqrt{P_k}\left\|\boldsymbol{h}_k'\right\|} \to \frac{\sum_{i=1}^{K}\sqrt{P_i}}{K\sqrt{P_k}}. \tag{55}$$

The achieved MSE is inversely proportional to N_t, irrespective of N_d when $N_t \to \infty$ for $N_t > N_d$, since $\boldsymbol{H}_i$ is a column full rank matrix with equal singular value of $\sqrt{N_t}$. We have MSE $\to \frac{\sigma^2}{PKN_t}$ and $r_k \to 1$ when $P_i = P_j = P, \forall i, j \in \mathcal{K}$.

C. Observations

We observe the following facts from the asymptotic analysis under the consumption of $P_i/P_j \in \left[\frac{r_{\min}^2}{r_{\max}^2}, \frac{r_{\max}^2}{r_{\min}^2}\right], \forall i \neq j$:

- The deployment of more antennas evidently provides more degree of freedom to align the signals from the distributed devices. This in turn brings down the performance gain obtained by the proposed DLR. The DLR ratio r_k is pushed closer to $\frac{\sum_{i=1}^{K}\sqrt{P_i}}{K\sqrt{P_k}}$, which is approaching 1 if $P_i = P_j, \forall i \neq j$.
- In the MISO scenario, the low bound MSE^{lb} is optimal, which can be achieved by designing DLR ratios as $r_k = \frac{\sum_{i=1}^{K}\sqrt{P_i}}{K\sqrt{P_k}}, \forall k \in \mathcal{K}$. In the MIMO scenario, the designed closed-from receive beamforming $\boldsymbol{m}$ can achieve the lower bound MSE^{lbm}.
- In the MIMO scenario, two cases, i.e. $N_d \to \infty$ for $N_d > N_t$, and $N_t \to \infty$ for $N_t > N_d$, are considered. The attained MSE is $\frac{\sigma^2}{\left(\sum_{i=1}^{K}\sqrt{P_i}\right)^2 N_d}$ in the former case, which is proportional to $\frac{1}{N_d}$ regardless of N_t. Whereas the obtained MSE in the latter case is linear with $\frac{K}{N_t}$ independent of N_d, i.e., $\frac{K\sigma^2}{\left(\sum_{i=1}^{K}\sqrt{P_i}\right)^2 N_t}$. Such results can be explained by the facts that the equivalent channel gains given the closed-form beamforming vector are respectively $\sqrt{N_d}$ and $\sqrt{\frac{N_t}{K}}$.

VII. Simulation Results

Simulation results are given in this section to demonstrate the effectiveness of the proposed DLR design in terms of the MSE performance as well as the learning performance. We also present the MSE performance of the proposed near-optimal and closed-form receive beamforming solution when a massive number of antennas are deployed. The proposed method is compared with the existing approaches without optimizing the DLR ratios. The performance of the compared method is labelled as ‘NDLR’ in the SISO and MISO scenarios, which does not optimize the receive beamforming but only optimizes the transmit coefficients (vector) b_k $(\boldsymbol{b}_k)$ using the method proposed in [37]. In the SIMO and MIMO scenarios, the first reference method, labelled as ‘SDR’, obtains the receive beamforming vector by SDR technique [38], which drops rank one constraint. Another reference method, labelled as ‘DC’, utilizes DC technique to represent the rank one constraint to solve the problem. We set the maximum transmit power of each device k as $P_k = 0$ dB. To show the impact of DLR on the learning performance, we implement the modified FedAvg to learn the classification tasks on MNIST and CIFAR10 datasets.

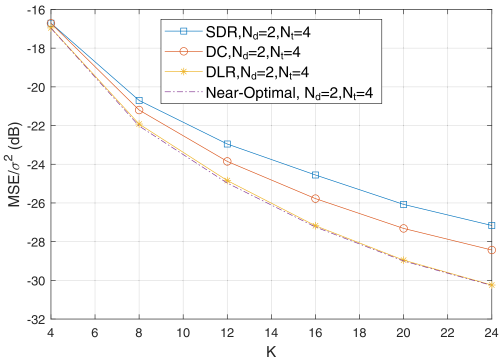


Fig. 3. The optimality of the proposed iterative learning rate and receive beamforming.

A. Performance on MSE Using DLR

To showcase the effectiveness of the proposed DLR, we conduct simulations under the SISO, MISO, SIMO and MIMO scenarios, where the boundaries of DLR ratio are set to $r_{\min} = 1/1.2$ and $r_{\max} = 1/0.8$. Fig. 2 displays MSE/σ^2 with respect to the number of devices K. It shows that the

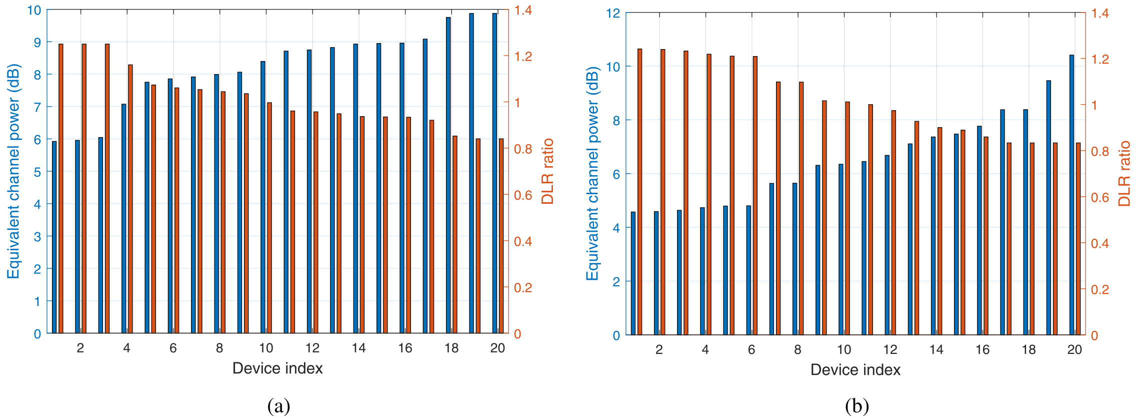


Fig. 4. The equivalent channel power and the corresponding DLR ratio. (a) MISO scenario with $N_d = 8$, (b) MIMO scenario with $N_d = 4, N_t = 4$.

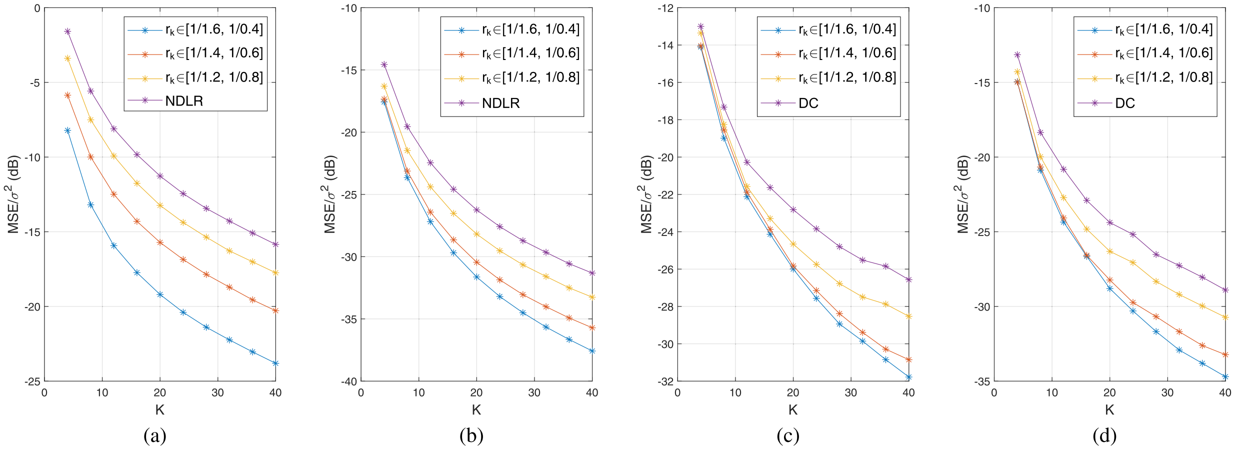


Fig. 5. The impact of $r_{\max}$ and $r_{\min}$. (a) SISO scenario. (b) MISO scenario with $N_d = 4$. (c) SIMO scenario with $N_t = 4$. (d) MIMO scenario with $N_d = 2, N_t = 2$.

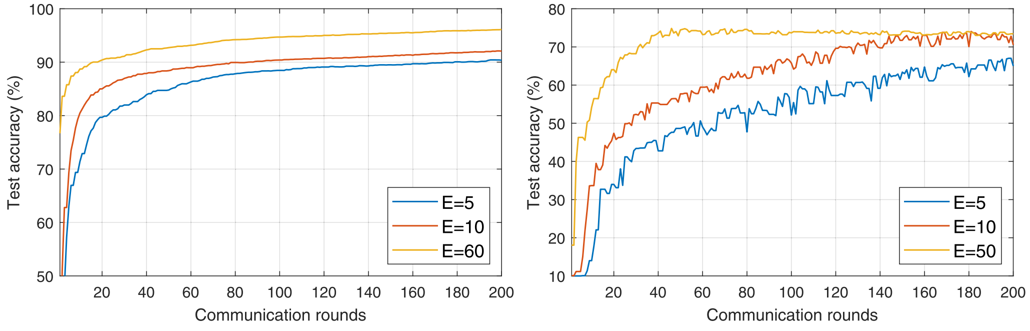
aggregate error decreases with the increase of K in all four cases. The aggregate error is further reduced by additionally considering DLR compared to the reference methods utilizing only wireless resources, which validates Theorem 2 and Theorem 3. In a MISO scenario where there are 20 devices equipped with $N_d = 8$ antennas, the proposed DRL can reduce the MSE up to 35.89% compared to the NDLR scheme. In the MIMO case where $K = 20$, $N_d = 2$, and $N_t = 4$, 48.27% and 30.93% of MSE can be further reduced compared with SDR and DC reference methods. In addition, the performance gap is extended with the increase of K, as more devices lead to a larger difference between the maximum and minimum signal power such that the proposed DLR can provide more potential. Compared with the SISO scenario, multiple antennas at the devices or/and the aggregator offer more degree of freedom to combat fading. Thereby, more antenna deployment results in a smaller aggregate error while the performance gain obtained by DLR is shrinking.

To showcase the optimality of the proposed iterative algorithm in the MIMO scenario, we conduct numerical simulations to obtain the near-optimal solution to problem (23), which is obtained by initializing 10 starting points. Fig. 3 reveals that the proposed iterative learning rate and receive beamforming algorithm can achieve near-optimal performance. In order to examine the impact of DLR ratios, we show the channel/equivalent channel power and the corresponding DLR ratios in Fig. 4. For clear illustration, the devices are indexed by the channel/equivalent channel power in an ascending order. As displayed in Fig. 4, the DLR ratio r_k is smaller (larger) for device k with higher (smaller) channel/equivalent channel power, so does DLR since $\mu^k = r_k\mu$.

Simulations are conducted on different ranges of DLR ratios under four scenarios to illustrate their impact. Note that the reference methods labelled as 'NDLR' and 'DC' are equivalent to the settings of $r_{\min} = r_{\max} = 1$. Fig. 5 shows that a larger range of DLR ratio leads to the decrease of MSE. In Fig. 5(c) and Fig. 5(d), the performance in the cases of $r_k \in [1/1.4, 1/0.6]$ and $r_k \in [1/1.6, 1/0.4]$ is close when $K \leq 10$. This can be explained such that the obtained receive beamforming vector $\boldsymbol{m}$ can well combat the distortion when K is small.

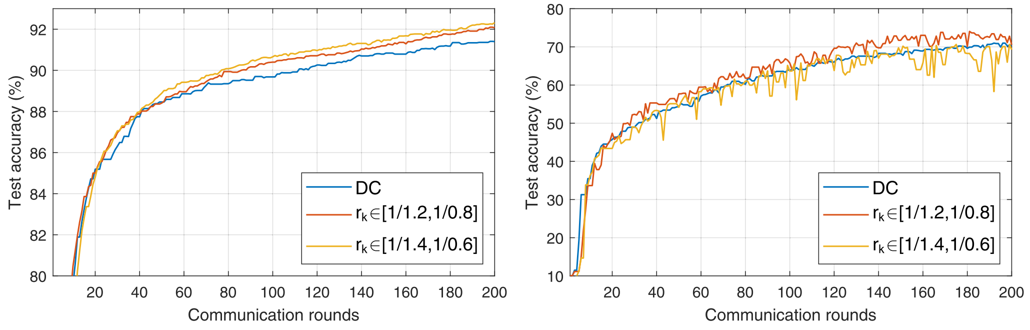
B. Performance of Learning Task Using DLR

To investigate the impact of proposed DLR on the training and testing performance of FL tasks, we utilize the modified FedAvg to perform classification tasks on MNIST and CIFAR10 datasets, which are evenly allocated at 20



(a) (b)

Fig. 6. The test accuracy performance under different steps E with $r_k \in [1/1.2, 1/0.8]$. (a) MNIST dataset. (b) CIFAR10 dataset.



(a) (b)

Fig. 7. The test accuracy performance under different ranges of DLR ratio with $E = 10$. (a) MNIST dataset. (b) CIFAR10 dataset.

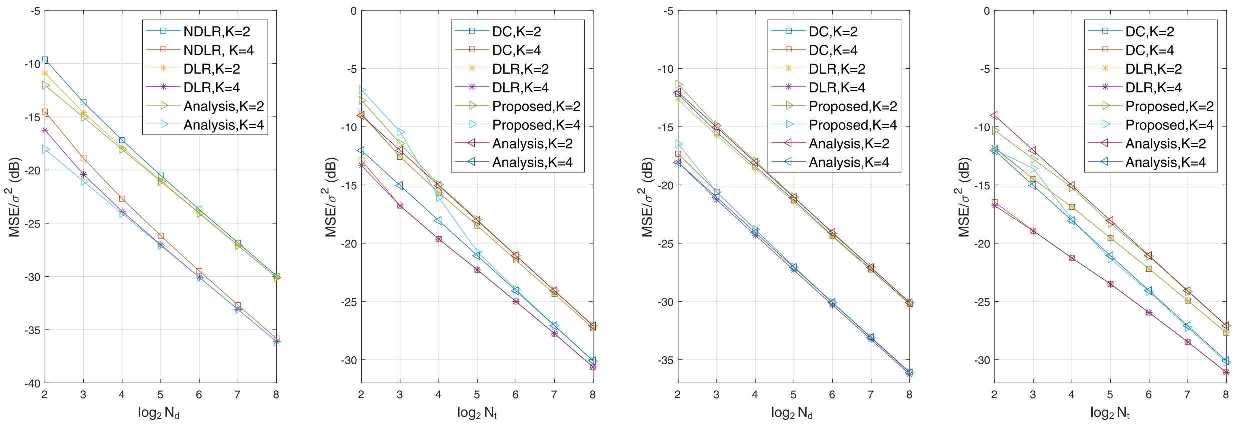
TABLE I
REPORTED ACCURACY ON MNIST AND CIFAR10 AT THE 200-TH COMMUNICATION ROUND

Scenario	Number of devices	$K = 4$		$K = 12$		$K = 20$	
	Dataset	MNIST	CIFAR10	MNIST	CIFAR10	MNIST	CIFAR10
MISO $N_d = 4$	NDLR	85.91%	46.18%	91.58%	65.32%	92.28%	70.44%
	DLR	87.44%	46.71%	92.10%	66.10%	92.51%	70.52%
MIMO $N_d = 2$, $N_t = 4$	DC	85.79%	49.32%	91.78%	65.99%	92.16%	71.90%
	DLR	88.00%	52.59%	91.94%	65.98%	92.54%	71.89%

devices. MLP and ResNet18 neural networks are adopted to train respectively on MNIST and CIFAR10 dataset for 200 communication rounds via mini-batch SGD optimizer. The size of mini-batch is set to 50. Note that the local learning rate μ^k is the product of r_k and μ, i.e., $\mu^k = r_k\mu$, where μ is set to 0.01. The parameter a and the noise power σ^2 in the retransmission model (12) are respectively set to 20 dB and 0 dB.

Simulations are conducted to evaluate the impact of E, i.e. the number of local model updates before aggregation, on the learning performance, which are shown in Fig. 6 with the setting of $r_k \in [1/1.2, 1/0.8]$. It shows that the convergence rate in terms of the communication round accelerates with the increase of E. Besides, the increase of E leads to the reduction of variance in Fig. 6(b). We also give the test accuracy performance under different ranges of DLR ratio in a MISO case with $N_d = 4$ in Fig. 7 to show its impact on the learning performance, where E is set to 10. Since the MSE utilizing DLR is smaller than the MSE without DLR, its re-transmission probability is smaller. Fig. 7 shows that the DLR-based scheme can slightly improve the test accuracy on both datasets compared to 'DC' method without utilizing DLR. In the setting of $r_k \in [1/1.2, 1/0.8]$, the test accuracy is improved approximately by 0.7% and 2% on MNIST and CIFAR10 datasets. This is because an increased learning rate owing to the adaption to the fading channels can help escape the saddle point known as the difficulty in minimizing the loss. Additionally, Fig. 7(b) shows that a larger range of DLR ratio may result in a poorer performance with a bigger variance. The variance can be reduced by increasing E, which is observed from Fig. 6(b). This implies that proper boundaries of DLR ratio should be chosen to guarantee the learning performance, which also validates the necessity of constraining the DLR ratio boundaries.

Further numerical simulations are conducted under the MISO and MIMO scenarios where there are $K = 4, 12, 20$ devices. Note that the devices participating in the FL training are fixed. The device selection based on the channel condition is not considered in this paper. The DLR ratio boundaries and the local update steps are set to $r_{\min} = 1/1.2$ and $r_{\max} = 1/0.8$, and $E = 10$, respectively. The reported accuracy performance on both datasets at the 200-th communication



(a) (b) (c) (d)

Fig. 8. The performance of the proposed closed-form receive beamforming design. (a) MISO scenario. (b) SIMO scenario. (c) MIMO scenario with $N_t = 2$. (d) MIMO scenario with $N_d = 2$.

round is given in Table I. The total data under $K = 4$ and $K = 12$ cases is insufficient compared to that of 20 devices, which results in a lower test accuracy. The reported accuracy using DLR may be smaller than that with a fixed learning rate due to the variance.

C. Performance of the Proposed Closed-Form Receive Beamforming Solution

Now we move on to verify the asymptotic analysis and the proposed closed-form receive beamforming design. The maximum power of the devices is set to 0 dB, which suggests that the DLR ratios $r_k \to 1$ according to the analysis. Due to the lack of space, we restrict ourselves to the cases of $K = 2$ and $K = 4$ only. The line labeled as 'Analysis' is the derived theoretical MSE, and 'Proposed' is the performance using the proposed closed-form receive beamforming design.

As shown in Fig. 8, the performance gap shrinks between the proposed DLR and reference methods without considering DLR, as more antennas lead to smaller differences on the channel gain among devices. In Fig. 8(a), both 'DLR' and 'NDLR' approach the optimal 'Analysis' performance eventually, which implies that the equalities of (a) and (b) in (28) are guaranteed, and verifies the asymptotic analysis under the MISO case. In the SIMO and MIMO scenarios, the MSE obtained by the proposed closed-form receive beamforming design is shown to approach the theoretical bound in Fig. 8(b), (c) and (d), which verifies the asymptotic analyses. The 'Analysis' is suboptimal and has a poorer performance than the 'DLR' method in Fig. 8(b), (c) and (d). This is because the lower bound of MSE in the SIMO/MIMO case is derived conditioning on the designed closed-form $\boldsymbol{m}$, wheras the 'DLR' is near-optimal. Besides, the 'Analysis' performance is shown to approach the near-optimal 'DLR' method as the increase of N_t or N_d. This further implies that the proposed closed-form receive beamforming design is near-optimal when the number of antennas becomes large. Therefore, the asymptotic analyses are verified and the effectiveness of the proposed closed-form receive beamforming design in Section VI is confirmed.

VIII. Conclusion

In this paper, we proposed a modified FedAvg by defining the local learning rate and presented its convergence analysis. The AirComp technique was incorporated to improve the communication efficiency, and the resultant aggregate error should be minimized. Different from existing works that mainly optimized the wireless resources to align the received signals, we first studied the potential of the learning rates to adapt to the fading channels, which were termed as the DLR, to further reduce the distortion error. The problems were formulated under the MISO and MIMO scenarios, where a closed-form solution and an iterative method were proposed respectively. Asymptotic analyses were also provided when the number of antennas becomes infinity. On this basis, a near-optimal and closed-form receive beamforming design was proposed by simply summing up the channel vectors. The simulation results have validated the effectiveness of the proposed DLR scheme in terms of both the MSE performance and the testing accuracy on the MNIST and CIFAR10 datasets. Additionally, the asymptotic analyses and the effectiveness of the proposed closed-form receive beamforming design were verified by extensive numerical simulations.

Appendix

To prove the convergence of the proposed modified FedAvg, we first give the additional notations and assumptions, followed by the key Lemmas and the Theorem. Then, the proofs of key lemmas and the theorem are presented.

A. Additional Notations and Assumptions

Define two virtual sequences $\overline{\boldsymbol{v}}_t = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{r_k}\boldsymbol{v}_t^k$ and $\overline{\boldsymbol{w}}_t = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{r_k}\boldsymbol{w}_t^k$. Then, $\overline{\boldsymbol{v}}_{t+1}$ results from a single step of $\overline{\boldsymbol{w}}_t$. Note that these two virtual sequences are inaccessible when $t+1 \notin \mathcal{I}_E$. We can only obtain $\overline{\boldsymbol{w}}_{t+1}$ when $t+1 \in \mathcal{I}_E$. Define $\overline{\boldsymbol{g}}_t = \frac{1}{K}\sum_{k=1}^{K}\nabla F_k\left(\boldsymbol{w}_t^k\right)$, $\boldsymbol{g}_t = \frac{1}{K}\sum_{k=1}^{K}\nabla F_k\left(\boldsymbol{w}_t^k, \mathcal{B}_t^k\right)$, where $\mathcal{B}_t^k$ is a mini-batch of data, $\nabla F_k\left(\boldsymbol{w}_t^k, \mathcal{B}_t^k\right) = \frac{1}{\left|\mathcal{B}_t^k\right|}\sum_{\xi_t^k \in \mathcal{B}_t^k}\nabla F_k\left(\boldsymbol{w}_t^k, \xi_t^k\right)$. Thereby, we have $\overline{\boldsymbol{v}}_{t+1} = \overline{\boldsymbol{w}}_t - \mu_t \boldsymbol{g}_t$ and $\mathbb{E}\boldsymbol{g}_t = \overline{\boldsymbol{g}}_t$.

Assumption 1: $F_1, \ldots, F_K$ are L-smooth: for all $\boldsymbol{v}$ and $\boldsymbol{w}$, we have $F_k(\boldsymbol{v}) \leq F_k(\boldsymbol{w}) + (\boldsymbol{v}-\boldsymbol{w})^T \nabla F_k(\boldsymbol{w}) + \frac{L}{2}\|\boldsymbol{v}-\boldsymbol{w}\|_2^2$.

Assumption 2: $F_1, \ldots, F_K$ are $\frac{1}{r_{\min}}\chi$-strongly convex: for all $\boldsymbol{v}$ and $\boldsymbol{w}$, we have $F_k(\boldsymbol{v}) \geq F_k(\boldsymbol{w}) + (\boldsymbol{v}-\boldsymbol{w})^T \nabla F_k(\boldsymbol{w}) + \frac{\frac{1}{r_{\min}}\chi}{2}\|\boldsymbol{v}-\boldsymbol{w}\|_2^2, \forall k \in \mathcal{K}$.

Assumption 3: The variance of the mini-batch gradient is bounded by $\mathbb{E}\left\|\nabla F_k\left(\boldsymbol{w}_t^k, \mathcal{B}_t^k\right) - \nabla F_k\left(\boldsymbol{w}_t^k\right)\right\| \leq \sigma_k^2, \forall k \in \mathcal{K}$.

Assumption 4: The expectation of the squared norm of the mini-batch gradients is bounded by $\mathbb{E}\left\|\nabla F_k\left(\boldsymbol{w}_t^k, \mathcal{B}_t^k\right)\right\|_2^2 \leq G^2, \forall k \in \mathcal{K}$.

B. Key Lemmas and Theorem

Lemma 2: Assume that Assumptions 1 and 2 hold. If $\mu_t \leq \frac{1}{4L}$, we have

$$\mathbb{E}\left\|\overline{\boldsymbol{v}}_{t+1} - \boldsymbol{w}^*\right\|^2 \leq (1-\chi\mu_t)\,\mathbb{E}\left\|\overline{\boldsymbol{w}}_t - \boldsymbol{w}^*\right\|^2 + 6L\mu_t^2\Gamma + 2\mathbb{E}\sum_{k=1}^{K}\frac{1}{K}\left\|\overline{\boldsymbol{w}}_t - \boldsymbol{w}_t^k\right\|^2 + \mu_t^2\mathbb{E}\left\|\overline{\boldsymbol{g}}_t - \boldsymbol{g}_t\right\|^2,$$

where $\Gamma = \sum_{k=1}^{K}\frac{1}{K}\left(F^* - F_k^*\right)$ with F^* and F_k^* the optimal loss value on the whole dataset and the local dataset at device k, respectively.

Lemma 3: Assume that Assumption 3 holds. We then have

$$\mathbb{E}\left\|\overline{\boldsymbol{g}}_t - \boldsymbol{g}_t\right\|^2 \leq \sum_{k=1}^{K}\frac{1}{K}\sigma_k^2.$$

Lemma 4: The local models are updated for E steps before uploaded to the aggregator. If μ_t is non-increasing and satisfies $\mu_t \leq 2\mu_{t+E}, \forall t \geq 0$, we can obtain:

$$\mathbb{E}\sum_{k=1}^{K}\frac{1}{K}\left\|\overline{\boldsymbol{w}}_t - \boldsymbol{w}_t^k\right\|^2 \leq 4E^2\mu_t^2\left(1 + Kr_{\max}^2\right)G^2.$$

Theorem 4: Assume that Assumptions 1–4 hold. Let $\mu_t = \frac{\beta}{t-\mathrm{mod}(t,E)+\gamma}$, $\beta > \frac{E}{\chi}$, $\gamma > 0$ such that $\mu_1 \leq \min\left\{\frac{1}{\chi}, \frac{1}{4L}\right\}$ and $\mu_t \leq 2\mu_{t+E}$, where $\mathrm{mod}(t,E)$ is the remainder of the t divided by E. The modified FedAvg converges and satisfies

$$\mathbb{E}\left[F\left(\overline{\boldsymbol{w}}_{\boldsymbol{t}}\right) - F^*\right] \leq \frac{L}{2}\Delta_t \leq \frac{L}{2}\frac{v}{\gamma + t - \mathrm{mod}(t,E)},$$

where $v = \max\left(\frac{\beta^2 B}{\beta\chi - E}, \gamma\Delta_1\right)$ with $\Delta_1 = \left\|\overline{\boldsymbol{w}}_1 - \boldsymbol{w}^*\right\|^2$ and $B = 8E^2\left(1 + Kr_{\max}^2\right)G^2 + 6L\Gamma + \sum_{k=1}^{K}\frac{1}{K}\sigma_k^2$ and $\Delta_1 = \left\|\overline{\boldsymbol{w}}_1 - \boldsymbol{w}^*\right\|^2$.

C. Proofs

1) Proof of Theorem 4

Proof: Whether $t+1 \in \mathcal{I}_E$ or $t+1 \notin \mathcal{I}_E$, we always have $\overline{\boldsymbol{w}}_{t+1} = \overline{\boldsymbol{v}}_{t+1}$. Let $\Delta_t = \mathbb{E}\left\|\overline{\boldsymbol{w}}_t - \boldsymbol{w}^*\right\|^2$. Based on Lemmas 2–4, we will prove $\Delta_t \leq \frac{v}{t-\mathrm{mod}(t,E)+\gamma}$:

$$\begin{aligned}\Delta_{t+1} &= \mathbb{E}\left\|\overline{\boldsymbol{w}}_{t+1} - \boldsymbol{w}^*\right\|^2 \leq (1-\chi\mu_t)\,\Delta_t + \mu_t^2 B \\ &\leq \frac{t - \mathrm{mod}(t,E) + \gamma - E}{\left(t - \mathrm{mod}(t,E) + \gamma\right)^2}v + \frac{\beta^2 B}{\left(t - \mathrm{mod}(t,E) + \gamma\right)^2} \\ &\quad - \frac{\chi\beta - E}{\left(t - \mathrm{mod}(t,E) + \gamma\right)^2}v \\ &\leq \frac{v}{t + 1 - \mathrm{mod}(t+1,E) + \gamma},\end{aligned}$$

where the third inequality holds since $\frac{\beta^2 B}{(t-\mathrm{mod}(t,E)+\gamma)^2} - \frac{\chi\beta - E}{(t-\mathrm{mod}(t,E)+\gamma)^2}v \leq 0$ when $v = \max\left(\frac{\beta^2 B}{\chi\beta - E}, \gamma\Delta_1\right)$. According to the L-smoothness of $F(.)$, we have

$$\mathbb{E}\left[F\left(\overline{\boldsymbol{w}}_{\boldsymbol{t}}\right) - F\left(\boldsymbol{w}^*\right)\right] \leq \frac{L}{2}\Delta_t \leq \frac{L}{2}\frac{v}{t - \mathrm{mod}(t,E) + \gamma}.$$

Thus, the modified FedAvg converges. □

2) Proof of Lemma 2:

Due to the page limit, please kindly refer to the proof of Lemma 1 in [40]. Note that we use the additional inequality when proving Lemma 2, i.e.,

$$\begin{aligned}\sum_{k=1}^{K}\frac{1}{K}\frac{1}{r_{\min}}\left\|\boldsymbol{w}_t^k - \boldsymbol{w}^*\right\|^2 &\geq \sum_{k=1}^{K}\frac{1}{K}\frac{1}{r_k}\left\|\boldsymbol{w}_t^k - \boldsymbol{w}^*\right\|^2 \\ &\geq \left\|\sum_{k=1}^{K}\frac{1}{K}\frac{1}{r_k}\left(\boldsymbol{w}_t^k - \boldsymbol{w}^*\right)\right\|^2 = \left\|\overline{\boldsymbol{w}}_t - \boldsymbol{w}^*\right\|^2.\end{aligned}$$

3) Proof of Lemma 3:

Proof: Based on Assumption 3 and Jensen inequality, we have

$$\begin{aligned}\mathbb{E}\left\|\overline{\boldsymbol{g}}_t - \boldsymbol{g}_t\right\|^2 &= \mathbb{E}\left\|\sum_{k=1}^{K}\frac{1}{K}\left(\nabla F_k\left(\boldsymbol{w}_t^k\right) - \nabla F_k\left(\boldsymbol{w}_t^k, \mathcal{B}_t^k\right)\right)\right\|^2 \\ &\leq \mathbb{E}\sum_{k=1}^{K}\frac{1}{K}\left\|\left(\nabla F_k\left(\boldsymbol{w}_t^k\right) - \nabla F_k\left(\boldsymbol{w}_t^k, \mathcal{B}_t^k\right)\right)\right\|^2 \\ &\leq \sum_{k=1}^{K}\frac{1}{K}\sigma_k^2,\end{aligned}$$

which completes the proof. □

4) Proof of Lemma 4:

Proof: For any $t \geq 0$, there exists a $t_0 \leq t$ such that $t - t_0 \leq E - 1$ and $\boldsymbol{w}_{t_0}^k = \overline{\boldsymbol{w}}_{t_0}, \forall k \in \mathcal{K}$. By denoting $\Delta G_{t,t_0}^k = \sum_{t'=t_0}^{t-1}\mu_{t'}\nabla F_k\left(\boldsymbol{w}_{t'}^k, \mathcal{B}_{t'}^k\right)$, $\boldsymbol{w}_t^k$ and $\overline{\boldsymbol{w}}_t$ can be respectively expressed as

$$\begin{aligned}\boldsymbol{w}_t^k &= \overline{\boldsymbol{w}}_{t_0} - \sum_{t'=t_0}^{t-1}\eta_{t'}r_k\nabla F_k\left(\boldsymbol{w}_{t'}^k, \mathcal{B}_{t'}^k\right) = \overline{\boldsymbol{w}}_{t_0} - r_k\Delta G_{t,t_0}^k, \\ \overline{\boldsymbol{w}}_t &= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{r_k}\boldsymbol{w}_t^k = \overline{\boldsymbol{w}}_{t_0} - \frac{1}{K}\sum_{k=1}^{K}\Delta G_{t,t_0}^k.\end{aligned}$$

Thereby, $\left\|\overline{\boldsymbol{w}}_t - \boldsymbol{w}_t^k\right\|^2$ satisfies:

$$\begin{aligned}\left\|\overline{\boldsymbol{w}}_t - \boldsymbol{w}_t^k\right\|^2 &= \left\|-\frac{1}{K}\sum_{k'=1}^{K}\Delta G_{t,t_0}^{k'} + r_k\Delta G_{t,t_0}^k\right\|^2 \\ &= \left\|\frac{1}{K}\Delta G_{t,t_0}^1 + \cdots + \left(\frac{1}{K} - r_k\right)\Delta G_{t,t_0}^k + \cdots + \frac{1}{K}\Delta G_{t,t_0}^K\right\|^2 \\ &\leq K\left(\sum_{k'=1}^{K}\frac{1}{K^2}\left\|\Delta G_{t,t_0}^{k'}\right\|^2 + r_k^2\left\|\Delta G_{t,t_0}^k\right\|^2\right).\end{aligned}$$

Next, we focus on the term $\mathbb{E}\left\|\Delta G_{t,t_0}^k\right\|^2, \forall k \in \mathcal{K}$, which follows

$$\begin{aligned}
\mathbb{E}\left\|\Delta G_{t,t_0}^k\right\|^2 &= \mathbb{E}\left\|\sum_{t'=t_0}^{t-1} \mu_{t'} \nabla F_k\left(\boldsymbol{w}_{t'}^k, \mathcal{B}_{t'}^k\right)\right\|^2 \\
&\leq \mathbb{E}\left((t-t_0)\sum_{t'=t_0}^{t-1} \mu_{t'}^2 \left\|\nabla F_k\left(\boldsymbol{w}_{t'}^k, \mathcal{B}_{t'}^k\right)\right\|^2\right) \\
&\leq \mathbb{E}\left((t-t_0)\sum_{t'=t_0}^{t-1} \mu_{t_0}^2 \left\|\nabla F_k\left(\boldsymbol{w}_{t'}^k, \mathcal{B}_{t'}^k\right)\right\|^2\right) \\
&\leq 4E^2\mu_t^2 G^2.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}\sum_{k=1}^{K} \frac{1}{K}\left\|\overline{\boldsymbol{w}}_t - \boldsymbol{w}_t^k\right\|^2 \\
&\leq \mathbb{E}\sum_{k=1}^{K}\left(\sum_{k'=1}^{K}\frac{1}{K^2}\left\|\Delta G_{t,t_0}^{k'}\right\|^2 + r_k^2\left\|\Delta G_{t,t_0}^k\right\|^2\right) \\
&= \mathbb{E}\left(\sum_{k=1}^{K}\left(\frac{1}{K}+r_k^2\right)\left\|\Delta G_{t,t_0}^k\right\|^2\right) \\
&\leq 4E^2\mu_t^2\left(1+\sum_{k=1}^{K} r_k^2\right)G^2 \\
&\leq 4E^2\mu_t^2\left(1+Kr_{\max}^2\right)G^2.
\end{aligned}$$

It shows that the inequality still holds for the proposed DLR scheme, where the DLR ratios are within the interval of $[r_{\min}, r_{\max}]$. □

REFERENCES

[1] C. Xu, S. Liu, Y. Huang, C. Huang, and Z. Zhang, "Over-the-air learning rate optimization for federated learning," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, Montreal, QC, Canada, Jun. 2021, pp. 1–7.

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

[3] Y. Huang, S. Liu, C. Zhang, X. You, and H. Wu, "True-data testbed for 5G/B5G intelligent network," *Intell. Converged Netw.*, vol. 2, no. 2, pp. 133–149, Jun. 2021.

[4] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.

[5] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2224–2287, 3rd Quart., 2019.

[6] C. Xu, S. Liu, C. Zhang, Y. Huang, Z. Lu, and L. Yang, "Multi-agent reinforcement learning based distributed transmission in collaborative cloud-edge systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 2, pp. 1658–1672, Feb. 2021.

[7] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6G be?" *Nature Electron.*, vol. 3, no. 1, pp. 20–29, Jan. 2020.

[8] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3039–3071, 4th Quart., 2019.

[9] S. Bi, R. Zhang, Z. Ding, and S. Cui, "Wireless communications in the era of big data," *IEEE Commun. Mag.*, vol. 53, no. 10, pp. 190–199, Oct. 2015.

[10] 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.

[11] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, no. 11, pp. 2204–2239, Nov. 2019.

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

[13] Y. Liu, X. Yuan, Z. Xiong, J. Kang, X. Wang, and D. Niyato, "Federated learning for 6G communications: Challenges, methods, and future directions," *China Commun.*, vol. 17, no. 9, pp. 105–118, Sep. 2020.

[14] Z. Yang, M. Chen, K.-K. Wong, H. V. Poor, and S. Cui, "Federated learning for 6G: Applications, challenges, and opportunities," 2021, *arXiv:2101.01338*. [Online]. Available: http://arxiv.org/abs/2101.01338

[15] M. Parimala *et al.*, "Fusion of federated learning and industrial Internet of Things: A survey," 2021, *arXiv:2101.00798*. [Online]. Available: http://arxiv.org/abs/2101.00798

[16] Q. Zhou *et al.*, "Falcon: Addressing stragglers in heterogeneous parameter server via multiple parallelism," *IEEE Trans. Comput.*, vol. 70, no. 1, pp. 139–155, Jan. 2021.

[17] W. Wen *et al.*, "TernGrad: Ternary gradients to reduce communication in distributed deep learning," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 1508–1518.

[18] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 1707–1718.

[19] Y. Lin, S. Han, H. Mao, Y. Wang, and B. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, May 2018, pp. 1–14.

[20] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, 2017, pp. 440–445.

[21] 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.

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

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

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

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

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

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

[28] D. Yu, S. H. Park, O. Simeone, and S. S. Shitz, "Optimizing over-the-air computation in IRS-aided C-RAN systems," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Atlanta, GA, USA, May 2020, pp. 1–5.

[29] W. Ni, Y. Liu, Z. Yang, H. Tian, and X. Shen, "Federated learning in multi-RIS aided systems," 2020, *arXiv:2010.13333*. [Online]. Available: http://arxiv.org/abs/2010.13333

[30] G. B. Orr and K.-R. Müller, *Neural Networks: Tricks of the Trade*. San Francisco, CA, USA: Springer, 2012.

[31] C. Darken, J. Chang, and J. Moody, "Learning rate schedules for faster stochastic gradient search," in *Proc. IEEE Workshop Neural Netw. Signal Process.*, Helsingoer, Denmark, Aug. 1992, pp. 3–12.

[32] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*. [Online]. Available: http://arxiv.org/abs/1609.04747

[33] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Santa Rosa, CA, USA, Mar. 2017, pp. 464–472.

[34] 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.

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

[36] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," 2019, *arXiv:1909.07972*. [Online]. Available: http://arxiv.org/abs/1909.07972

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

[38] 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.

[39] J.-Y. Gotoh, A. Takeda, and K. Tono, "DC formulations and algorithms for sparse optimization problems," *Math. Program.*, vol. 169, no. 1, pp. 141–176, May 2018.

[40] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FeDavg on non-iid data," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020, pp. 1–26. [Online]. Available: https://openreview.net/forum?id=HJxNAnVtDS

Chunmei Xu (Student Member, IEEE) received the B.Eng. degree in information engineering from the School of Information Science and Engineering, Southeast University, Nanjing, China, in 2017, where she is currently pursuing the Ph.D. degree in information and communication engineering with the School of Information Science and Engineering. Her research interests mainly focus on intelligent wireless communications.

Shengheng Liu (Member, IEEE) received the B.Eng. and Ph.D. degrees in electronics engineering from the School of Information and Electronics, Beijing Institute of Technology, Beijing, China, in 2010 and 2017, respectively.

He is currently an Associate Professor with the School of Information Science and Engineering, Southeast University (SEU), Nanjing, China. Prior to joining SEU, he held a post-doctoral position at the Institute for Digital Communications, The University of Edinburgh, Edinburgh, U.K., from 2017 to 2018. From 2015 to 2016, he also worked as a Visiting Research Associate at the Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA, USA, under the support of the China Scholarship Council. He was a recipient of the 2017 National Excellent Doctoral Dissertation Award from the China Institute of Communications. His research interests mainly focus on intelligent sensing and wireless communications.

Zhaohui Yang (Member, IEEE) received the B.S. degree in information science and engineering from Chien-Shiung Wu Honors College, Southeast University, Nanjing, China, in June 2014, and the Ph.D. degree in communication and information system with the National Mobile Communications Research Laboratory, Southeast University, in May 2018. From May 2018 to October 2020, he was a Post-Doctoral Research Associate with the Center for Telecommunications Research, Department of Informatics, King's College London, U.K. He is currently a Visiting Associate Professor with the College of Information Science and Electronic Engineering, Zhejiang Key Laboratory of Information Processing Communication and Networking, Zhejiang University, and also a Research Fellow with the Department of Electronic and Electrical Engineering, University College London, U.K. His research interests include federated learning, reconfigurable intelligent surface, UAV, and NOMA. He is an Associate Editor of the IEEE COMMUNICATIONS LETTERS, *IET Communications*, and *EURASIP Journal on Wireless Communications and Networking*. He has guest edited a feature topic of *IEEE Communications Magazine* on Communication Technologies for Efficient Edge Learning. He was the Co-Chair for workshops on edge learning and wireless communications in several conferences, including the IEEE International Conference on Communications (ICC), the IEEE Global Telecommunication Conference (GLOBECOM), the IEEE Wireless Communications and Networking Conference (WCNC), and the IEEE International Symposium on Personal, Indoor and Mobile Radio Communication (PIMRC). He was a TPC Member of IEEE ICC from 2015 to 2021 and IEEE GLOBECOM from 2017 to 2021. He was an Exemplary Reviewer of IEEE TRANSACTIONS ON COMMUNICATIONS in 2019 and 2020.

Yongming Huang (Senior Member, IEEE) received the B.S. and M.S. degrees from Nanjing University, Nanjing, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from Southeast University, Nanjing, in 2007.

Since March 2007, he has been a faculty with the School of Information Science and Engineering, Southeast University, where he is currently a Full Professor. He has also been the Director of the Pervasive Communication Research Center, Purple Mountain Laboratories, since 2019. From 2008 to 2009, he visited the Signal Processing Lab, KTH Royal Institute of Technology, Stockholm, Sweden. He has published over 200 peer-reviewed articles and hold over 80 invention patents. His current research interests include intelligent 5G/6G mobile communications and millimeter wave wireless communications. He submitted around 20 technical contributions to IEEE standards, and was awarded a certificate of appreciation for outstanding contribution to the development of IEEE standard 802.11aj. He has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, and is serving as the Editor-at-Large for the IEEE Open Journal of the Communications Society and an Associate Editor for the IEEE WIRELESS COMMUNICATIONS LETTERS.

Kai-Kit Wong (Fellow, IEEE) received the B.Eng., M.Phil., and Ph.D. degrees in electrical and electronic engineering from The Hong Kong University of Science and Technology, Hong Kong, in 1996, 1998, and 2001, respectively. After graduation, he took up academic and research positions at The University of Hong Kong, Lucent Technologies, Bell-Labs, Holmdel, the Smart Antennas Research Group, Stanford University, and the University of Hull, U.K. He is the Chair in wireless communications at the Department of Electronic and Electrical Engineering, University College London, U.K. His current research centers around 5G and beyond mobile communications. He was a co-recipient of the 2013 IEEE SIGNAL PROCESSING LETTERS Best Paper Award and the 2000 IEEE VTS Japan Chapter Award at the IEEE Vehicular Technology Conference in Japan in 2000, and a few other international best paper awards. He is a fellow of IET and is also on the editorial board of several international journals. He is the Editor-in-Chief of IEEE WIRELESS COMMUNICATIONS LETTERS since 2020.