Analog Gradient Aggregation for Federated Learning Over Wireless Networks: Customized Design and Convergence Analysis

Huayan Guo, *Member, IEEE*, An Liu, *Senior Member, IEEE*, and Vincent K. N. Lau, *Fellow, IEEE*

***Abstract*—This article investigates the analog gradient aggregation (AGA) solution to overcome the communication bottleneck for wireless federated learning applications by exploiting the idea of analog over-the-air transmission. Despite the various advantages, this special transmission solution also brings new challenges to both transceiver design and learning algorithm design due to the nonstationary local gradients and the time-varying wireless channels in different communication rounds. To address these issues, we propose a novel design of both the transceiver and learning algorithm for the AGA solution. In particular, the parameters in the transceiver are optimized with the consideration of the nonstationarity in the local gradients based on a simple feedback variable. Moreover, a novel learning rate design is proposed for the stochastic gradient descent algorithm, which is adaptive to the quality of the gradient estimation. Theoretical analyses are provided on the convergence rate of the proposed AGA solution. Finally, the effectiveness of the proposed solution is confirmed by two separate experiments based on linear regression and the shallow neural network. The simulation results verify that the proposed solution outperforms various state-of-the-art baseline schemes with a much faster convergence speed.**

***Index Terms*—Distributed data aggregation, distributed machine learning, federated learning (FL), Internet of Things (IoT), over-the-air transmission.**

I. Introduction

INTERNET of Things (IoT) is one important feature of future 5G wireless networks. In future machine learning applications, raw data are usually available at the IoT sensors. To train the neural network at the cloud, sensors need to upload raw data over the radio interface. However, there are privacy issues as well as the spectral efficiency issue for such brute-force uploading of raw data. Federated learning (FL) is a recently proposed approach to address these issues [1]–[4]. Specifically, in FL, the IoT sensors (also known as clients) collaboratively learn a shared prediction model while keeping all the raw data locally, as illustrated in Fig. 1 [5]–[7]. In each FL iteration, each client generates a local update according to the current global model broadcast from the cloud server. There are two variations of local updates in FL, namely, the *model aggregation solution* and the *gradient aggregation solution*. In both cases, only the local updates (instead of the raw data) are uploaded to the server and this resolves the privacy issue.

To alleviate the radio resource requirements, some works have focused on improving the communication efficiency by designing new classes of gradient descent (GD) methods to reduce the updating frequency [7]–[9] or to compress the updated messages [10]–[13]. In these schemes, the clients are assumed to be allocated with dedicated radio resource and the updates are assumed to be error-free through the help of a powerful error correction code. However, the required number of radio resource blocks needs to scale with the number of clients, which is highly undesirable because a large number of clients will be involved. To address the scalability issues, the analog transmission scheme is proposed to utilize the free over-the-air aggregation of signals in the wireless multiple-access channel (MAC) [14]–[21]. This idea is inspired by the over-the-air fusion of sensor measurements in consensus applications [22]–[25] and modulation-free remote state estimations [26], [27]. Specifically, all the clients share a common resource block in the wireless interface to upload the updates to the server. When several active clients transmit simultaneously, the received signal at the server side will be a noisy and weighted aggregation of the transmitted symbols. Therefore, one widely investigated challenge of analog over-the-air aggregation is to extract an unbiased estimator from the received signal. One approach is the uniform-forcing transmitter based on the channel inversion policy [22]–[24]. Analog over-the-air aggregation has many advantages, such as having a highly scalable radio resource requirement to support a large number of clients. It also substantially simplifies the MAC protocol because collisions between clients are no longer harmful. This can result in virtually zero access latency for the IoT network [28].

The analog model aggregation solution for FL is investigated in [14]–[16]. These works all employ the uniform-forcing transmitter, and due to the channel inversion policy, it is found that the average receive signal-to-noise ratio (SNR)

Manuscript received March 3, 2020; revised May 29, 2020; accepted June 7, 2020. Date of publication June 16, 2020; date of current version December 21, 2020. The work of Huayan Guo and Vincent K. N. Lau was supported in part by the Hong Kong Innovation and Technology Fund Project under Grant GHP-016-18GD. The work of An Liu was supported in part by the Huawei Technologies under Grant YBN2020045196, and in part by the China Recruitment Program of Global Young Experts. *(Corresponding author: An Liu.)*

Huayan Guo and Vincent K. N. Lau are with the Department of Electronics and Communication Engineering, Hong Kong University of Science and Technology, Hong Kong (e-mail: eeguohuayan@ust.hk; eeknlau@ust.hk).

An Liu is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: anliu@zju.edu.cn).

Digital Object Identifier 10.1109/JIOT.2020.3002925

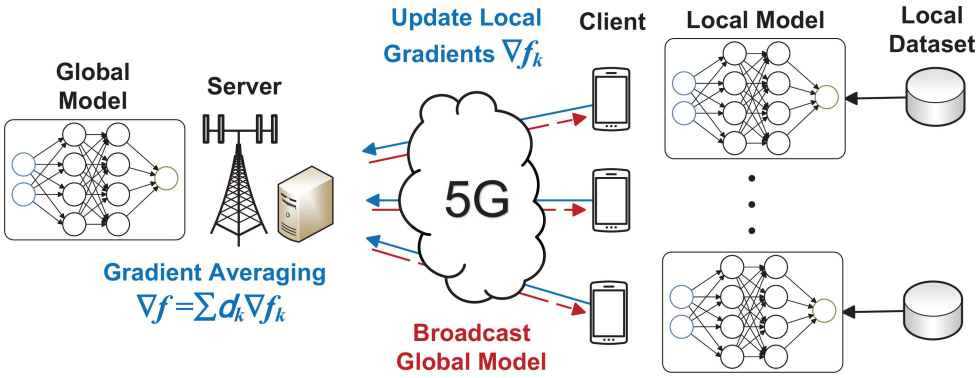


Fig. 1. Architecture of FL using gradient aggregation.

decreases as the number of clients increases. The tradeoff between the number of active clients and the SNR is investigated in [14] and [15]. Then, in [16], the unbalanced size of the local data sets is jointly considered in the tradeoff. The analog gradient aggregation (AGA) solution for FL is investigated in [17]–[21]. In [17], the convergence of the AGA solution is analyzed, while employing the GD algorithm with a constant learning rate. It is pointed out that the AGA suffers from a nonzero optimality gap due to the bias on the gradient estimator and the use of a constant learning rate. In [18]–[20], the dimension of the local updates is further compressed by exploiting gradient sparsity, and the effectiveness of the proposed algorithms is verified by simulation. In [21], the aggregated local gradients are quantized to one-bit digital symbols for compatibility with the modern digital communication infrastructure.

Despite the above works on analog aggregation solutions for FL applications, there are still various technical challenges yet to be addressed. In this article, we focus on the gradient aggregation solution using an analog over-the-air transmission scheme. In contrast to the existing works, we propose a novel AGA solution to achieve a zero convergence gap with significantly faster convergence speed, corresponding to higher learning accuracy and a lower communication overhead, respectively. The following summarizes the critical issues to be addressed and the contributions of the article.

1) *Nonstationarity of Gradient Updates:* In previous works [14]–[20], the uniform-forcing transceiver is employed and the transceiver parameters are optimized, which only exploits the channel state information (CSI). However, the effectiveness of this design relies on the assumption that the transmitted local gradients are stationary, which does not hold in the AGA solution for FL applications. In FL applications, the L_2 norm of the local gradient decreases gradually as the model training converges. As a result, the traditional CSI-only transceiver-adaptation is suboptimal. In this article, we optimize the parameters in the uniform-forcing transceiver through a novel data-and-CSI-aware design with the help of a low overhead feedback mechanism. This is critical to achieve superior convergence speed in FL iterations.
2) *Efficient Noise Mitigation of Stochastic GD:* Due to the noisy transmission, analog aggregation solutions suffer from aggregation errors in the updates. In contrast to the traditional stochastic GD (SGD), the noise variance of the estimated gradient varies in different communication rounds. Exploiting the above properties, we propose a novel learning rate design for the SGD algorithm, which is adaptive to the quality of the gradient estimation in every communication round.
3) *Fast Convergence Rate With Zero Convergence Gap:* We provide a theoretical analysis of the convergence rate of the proposed AGA solution. We prove that, under special conditions, the proposed solution can achieve a fast linear convergence rate, which is an orderwise improvement compared to the traditional designs. The simulation results verify that the proposed solution outperforms various state-of-the-art baseline schemes with a much faster convergence speed, even in general cases.

The remainder of the article is organized as follows. Section II briefly introduces the FL system. Then, in Section III, the AGA solution is presented, and the remaining design tasks on the transceiver-parameter optimization and the learning rate design are separately pointed out. In Section IV, we propose the data-and-CSI-aware transceiver-parameter optimization for the AGA solution. In Section V, we propose a novel learning rate design for the SGD in the AGA solution. The effectiveness of the proposed solution is verified in Section VI by two separate examples on linear regression and a shallow neural network. Finally, Section VII concludes the article. The notations used in this article are listed in Table I.

II. Learning Model and Basic Assumptions

A. Gradient-Aggregation Federated Learning

We investigate the gradient-aggregation FL system consisting of one edge server and K distributed clients, as illustrated in Fig. 1. A global machine learning model is shared by all the clients and the server. In particular, the architecture of the model, such as the number of layers and the structure of each layer, is fixed. Meanwhile, the parameters, such as the bias and weight values, need to be collaboratively trained, which is denoted by vector $\mathbf{w} \in \mathbb{R}^B$, where B is the dimension of $\mathbf{w}$. In the training process, in contrast to traditional learning schemes, the local training data sets are not sent to the server.

TABLE I
LIST OF SYMBOLS AND NOTATIONS

Symbols and Notations	Description
$\mathbb{E}[\cdot]$	Statistical expectation
$\mathcal{CN}(\mu, \sigma^2)$	Complex Gaussian distribution with mean μ and variance σ^2
$\nabla f(\mathbf{w})$	Gradient of function $f(\cdot)$ at $\mathbf{w}$
$\|\mathbf{w}\|_2$	Euclidean norm of vector $\mathbf{w}$
$\mathcal{R}[x]$	Real part of a complex number x
$\max\{a, b\}$	The largest value given real numbers a and b
x^+	The positive part of the real number x
$\mathbf{G}^{\mathrm{T}}$	The transpose of matrix $\mathbf{G}$
$\mathbf{G}^{\mathrm{H}}$	The conjugate transpose of matrix $\mathbf{G}$

Instead, in every communication round, the server first broadcasts $\mathbf{w}$ to synchronize all the models in the clients over an error-free shared link. Then, the local gradients with respect to the current model are computed by the clients based on their own data sets. Finally, all the local gradients are aggregated over the wireless channel to update $\mathbf{w}$ at the server.

Denote the local data set at the kth client by $\mathcal{D}_k$, and further assume $\mathcal{D}_k \cap \mathcal{D}_{k'} = \emptyset$, for $k \neq k'$. Then, the learning problem is expressed as follows:

$$\mathcal{P}(A) \quad \min_{\mathbf{w}} \ f(\mathbf{w}) \triangleq \sum_{k=1}^{K} d_k f_k(\mathbf{w})$$

where $f(\mathbf{w})$ is the global objective function, $\mathbf{w}$ is the shared model parameter, $d_k = (|\mathcal{D}_k|/|\mathcal{D}|)$ is the weight of the kth client, and $\mathcal{D}$ denotes the global data set with $|\mathcal{D}| = \sum_{k=1}^{K} |\mathcal{D}_k|$. $f_k(\mathbf{w})$ is the local objective function at the kth client

$$\text{(Local objective)} \quad f_k(\mathbf{w}) = \frac{1}{|\mathcal{D}_k|} \sum_{d=1}^{|\mathcal{D}_k|} \ell\big(\mathbf{w}; \mathbf{s}_{k,d}, \mathbf{q}_{k,d}\big) \quad (1)$$

where $\mathbf{s}_{k,d}$ and $\mathbf{q}_{k,d}$ is the dth data and dth label in client k, respectively, and $\ell\big(\mathbf{w}; \mathbf{s}_{k,d}, \mathbf{q}_{k,d}\big)$ is the loss function.

We assume a synchronous update scheme, in which all the clients calculate their local gradients $\nabla f_k(\mathbf{w}_t)$ based on the same current shared model $\mathbf{w}_t$ in the tth round. In addition, we assume that the full gradient is utilized for updating $\mathbf{w}_t$.[1] In an ideal case, the server aggregates the global gradient via the error-free channel, and obtains the global gradient

$$\text{(Global gradient)} \quad \nabla f(\mathbf{w}_t) = \sum_{k=1}^{K} d_k \nabla f_k(\mathbf{w}_t). \quad (2)$$

For ease of presentation, denote

$$\mathbf{g}_t = \nabla f(\mathbf{w}_t)$$
$$\mathbf{g}_{k,t} = \nabla f_k(\mathbf{w}_t)$$

[1]The reasons for adopting the full gradient are two-fold. First, the GD algorithm with full gradient does not suffer sampling noise in every communication round, and thus we can concentrate on the distortion in the gradient aggregation due to the channel noise. Second, in the FL application, the number of clients is massive, while the size of the local data sets is relatively small. The GD algorithm with the full gradient may achieve better convergence speed with a slightly higher computation cost, resulting in higher communication efficiency.

where $\mathbf{g}_t = [g_{t,1}, \ldots, g_{t,B}]^T$ and $\mathbf{g}_{k,t} = [g_{k,t,1}, \ldots, g_{k,t,B}]^T$. Then, (2) becomes

$$\mathbf{g}_t = \sum_{k=1}^{K} d_k \mathbf{g}_{k,t}. \quad (3)$$

Finally, the standard GD algorithm is adopted for the global model update as follows, if all the local gradients $\mathbf{g}_{k,t}$ are sent through the error-free channel:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_t \quad (4)$$

where η_t is the learning rate. However, in practice, the communication channels between the clients and sever always contain noise, and the server can only obtain an estimated global gradient. In this case, it is challenging to design a good communication (gradient aggregation) scheme and the associated learning rate η_t to achieve good learning performance with limited radio resources. In Sections III–V, we will introduce a novel AGA solution together with its parameter optimization method and associated learning rate design to address this challenge.

B. Basic Assumptions for the Learning Algorithm

In this section, we present two standard basic assumptions for the learning problem.

Assumption 1 (L-Smooth): The local objectives $f_k(\mathbf{w})$ are L-smooth for all k; that is, for any $\mathbf{w}$ and $\bar{\mathbf{w}}$, we have

$$f_k(\mathbf{w}) \leq f_k(\bar{\mathbf{w}}) + \nabla f_k(\bar{\mathbf{w}})^T(\mathbf{w} - \bar{\mathbf{w}}) + \frac{L}{2}\|\mathbf{w} - \bar{\mathbf{w}}\|_2^2.$$

Assumption 2 (μ-Strong Convexity): The local objectives $f_k(\mathbf{w})$ are μ-strongly convex for all k; that is, for all $\mathbf{w}$ and $\bar{\mathbf{w}}$, we have

$$f_k(\mathbf{w}) \geq f_k(\bar{\mathbf{w}}) + \nabla f_k(\bar{\mathbf{w}})^T(\mathbf{w} - \bar{\mathbf{w}}) + \frac{\mu}{2}\|\mathbf{w} - \bar{\mathbf{w}}\|_2^2.$$

Combining the above two assumptions, we have $\mu \leq L$. Furthermore, one may obtain the following corollary.

Corollary 1 (Properties of the Global Objective Function): Given Assumptions 1 and 2, $f(\mathbf{w})$ is L-smooth and μ-strongly convex.

Assumptions 1 and 2 are standard, and have been widely used in the convergence analysis of FL algorithms, e.g., [8]–[13].

C. Illustration With Two Application Examples

In this section, we illustrate the learning model using two important application examples, i.e., the linear regression and the shallow neural network. Then we will show that the local objection functions $f_k(\mathbf{w})$ in both examples satisfy the above two key assumptions.

1) Linear Regression: The output of the linear regression can be denoted by $\hat{q}_{k,d} = \mathbf{w}^T \mathbf{s}_{k,d}$ for the dth data sample in client k (i.e., $\mathbf{s}_{k,d}$). Then, the loss function is given by

$$\ell\big(\mathbf{w}; \mathbf{s}_{k,d}, q_{k,d}\big) = \frac{1}{2}|q_{k,d} - \mathbf{w}^T \mathbf{s}_{k,d}|^2$$

and the local objective function in client k is given by

$$f_k(\mathbf{w}) = \frac{1}{2|\mathcal{D}_k|}\sum_{d=1}^{|\mathcal{D}_k|}|q_{k,d} - \mathbf{w}^T\mathbf{s}_{k,d}|^2.$$

One can verify that $f_k(\mathbf{w})$ is smooth and strongly convex, and thus satisfies Assumptions 1 and 2.

2) Shallow Neural Network: The shallow neural network consists of one input layer, one fully connected layer, and one softmax layer, and the output can be expressed concisely by

$$\hat{\mathbf{q}}\big(\mathbf{w}; \mathbf{s}_{k,d}\big) = \text{Softmax}\big(\mathbf{W}\mathbf{s}_{k,d} + \mathbf{b}\big)$$

where the model parameters are denoted by $\mathbf{w} = (\mathbf{W}, \mathbf{b})$. Moreover, we adopt "CrossEntropy" as the loss function [29]

$$\ell\big(\mathbf{w}; \mathbf{s}_{k,d}, \mathbf{q}_{k,d}\big) = \text{CrossEntropy}\big(\hat{\mathbf{q}}\big(\mathbf{w}; \mathbf{s}_{k,d}\big), \mathbf{q}_{k,d}\big) + \lambda\|\mathbf{w}\|_2^2$$

where $\lambda = 10^{-3}$ is the regularization parameter. Finally, the local objective function in client k is given by

$$f_k(\mathbf{w}) = \frac{1}{|\mathcal{D}_k|}\sum_{d=1}^{|\mathcal{D}_k|}\text{CrossEntropy}\big(\hat{\mathbf{q}}\big(\mathbf{w}; \mathbf{s}_{k,d}\big), \mathbf{q}_{k,d}\big) + \lambda\|\mathbf{w}\|_2^2$$

which is smooth and strongly convex, and thus satisfies Assumptions 1 and 2 [29].

III. Analog Gradient Aggregation Solution

In a conventional wireless FL scheme, the local gradient vector $\mathbf{g}_{k,t}$ is sent to the central server with dedicated radio resource, so that the transmission error can be suppressed to an arbitrarily low level through the help of a powerful error correction code [7]–[13]. However, the radio resource blocks required will scale linearly with the number of clients, and thus such a naive transmission scheme is highly undesirable because a large number of clients will be involved.

In this section, we present the AGA solution to reduce the radio resource requirement and the multiple access latency by exploiting the free-aggregation of signals over the wireless channel. In particular, the AGA solution contains two main components. One is the uniform-forcing transceiver, which obtains an unbiased estimation of the global gradient $\mathbf{g}_t$, denoted by $\hat{\mathbf{g}}_t$, and the other is the corresponding SGD algorithm for the model update based on $\hat{\mathbf{g}}_t$. Finally, we point out two remaining tasks to further improve the performance of the AGA solutions which will be investigated in detail in Sections IV and V, separately.

A. Uniform-Forcing Transmitter for Unbiased Gradient Estimation

We assume that the base station connecting the cloud server has M antennas, while each client device has a single antenna. Since all the client devices share a common radio resource block, the received signal at the base station over the wireless fading channel in the tth communication round is given by

$$\mathbf{y}_{t,b} = \sum_{k=1}^{K}\mathbf{h}_{k,t}a_{k,t}x_{k,t,b} + \mathbf{z}_{t,b} \tag{5}$$

where $b = 1, \ldots, B$ is the index of $x_{k,t,b}$, $\mathbf{h}_{k,t} \in \mathbb{C}^{M\times 1}$ is the channel fading coefficient from client k to the base station in the tth round, $\mathbf{z}_{t,b}\sim\mathcal{CN}(0, \sigma_0^2)$ is the additive Gaussian noise, $a_{k,t}$ is the power gain, and the information-bearing symbol is given by

$$x_{k,t,b} = d_k g_{k,t,b}. \tag{6}$$

Define $\mathbf{x}_{k,t} = [x_{k,t,1}, \ldots, x_{k,t,B}]^T$, we have

$$\mathbf{x}_{k,t} = d_k\mathbf{g}_{k,t}. \tag{7}$$

Hence, $\mathbf{g}_t = \sum_{k=1}^{K}\mathbf{x}_{k,t}$. The base station implements a simple linear receiver, and the estimation of $g_{t,b}$ is expressed as follows[2]:

$$\begin{aligned}\hat{g}_{t,b} &= \frac{1}{\sqrt{\alpha_t}}\mathcal{R}\big[\mathbf{v}_t^H\mathbf{y}_{t,b}\big] \\ &= \frac{1}{\sqrt{\alpha_t}}\mathcal{R}\left[\mathbf{v}_t^H\sum_{k=1}^{K}\mathbf{h}_{k,t}a_{k,t}x_{k,t,b}\right] + \frac{1}{\sqrt{\alpha_t}}u_{t,b}\end{aligned} \tag{8}$$

where $\mathbf{v}_t \in \mathbb{C}^{M\times 1}$ is the receive beamforming vector with unit power $\|\mathbf{v}_t\|_2 = 1$, α_t is the scaling factor, and $u_{t,b} = \mathcal{R}[\mathbf{v}_t^H\mathbf{z}_{t,b}]\sim\mathcal{N}(0, [1/2]\sigma_0^2)$.

One can seen from (8) that $\mathbf{x}_{k,t} = d_k\mathbf{g}_{k,t}$ of all the K clients are updated in one common resource block in the wireless interface so that the multiple access latency is significantly reduced. However, $\hat{\mathbf{g}}_t = [\hat{g}_{t,1}, \ldots, \hat{g}_{t,B}]^T$ is contaminated by the channel noise. In order to extract an unbiased estimation of $\mathbf{g}_{k,t}$ from $\hat{\mathbf{g}}_t$, the beamforming vector $\mathbf{v}_t$ and the power gain $a_{k,t}$ in (8) should be carefully designed. To this end, we apply the uniform-forcing transmitter in [22]–[25], in which the power gain $a_{k,t}$ is given by

$$a_{k,t} = \sqrt{\alpha_t}\frac{\big(\mathbf{v}_t^H\mathbf{h}_{k,t}\big)^H}{\big|\mathbf{v}_t^H\mathbf{h}_{k,t}\big|^2}. \tag{9}$$

Then the gradient estimation in (8) becomes

$$\hat{g}_{t,b} = \sum_{k=1}^{K}x_{k,t,b} + \hat{u}_{t,b} \tag{10}$$

where $\hat{u}_{t,b} = [1/(\sqrt{\alpha_t})]u_{t,b}$, and thus $\hat{u}_{t,b}\sim\mathcal{N}(0, [1/(2\alpha_t)]\sigma_0^2)$. Let $\hat{\mathbf{u}}_t = [\hat{u}_{t,1}, \ldots, \hat{u}_{t,B}]^T$. Combining $\mathbf{g}_t = \sum_{k=1}^{K}\mathbf{x}_{k,t}$, we have

$$\begin{aligned}\text{(Gradient estimation)}\quad \hat{\mathbf{g}}_t &= \sum_{k=1}^{K}\mathbf{x}_{k,t} + \hat{\mathbf{u}}_t \\ &= \mathbf{g}_t + \hat{\mathbf{u}}_t.\end{aligned} \tag{11}$$

Therefore, we have $\mathbb{E}_{\hat{\mathbf{u}}_t}[\hat{\mathbf{g}}_t] = \mathbf{g}_t$ since the noise $\hat{\mathbf{u}}_t$ is zero-mean and independent to $\mathbf{g}_t$, and thus $\hat{\mathbf{g}}_t$ is an unbiased estimation of $\mathbf{g}_t$.

[2]We choose the real part of the received signal here because the model parameter $\mathbf{w}$ is defined as the real vector in this article, which is also true in practical machine learning applications. The derivations in this article can be easily extended to the learning model using complex variables by simply removing the $\mathcal{R}[\cdot]$ operation in (8).

B. Model Update for Analog Gradient Aggregation

The existing works on the AGA solution suggest updating the model by the SGD algorithm with full gradient and a constant learning rate [16], [17] or the ADAM optimizer [18]–[20]. However, these algorithms are mainly applied in the centralized learning system, and it has been verified by experiments that these algorithms cannot converge to the optimal solution due to the estimation noise in $\hat{\mathbf{g}}_t$ [16]–[20]. In this article, we employ the following SGD algorithm at the server to update the global model:

$$\text{(Model update)} \quad \mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \hat{\mathbf{g}}_t \tag{12}$$

where $\hat{\mathbf{g}}_t$ is an unbiased estimation of the full gradient $\mathbf{g}_t$. In contrast to [16] and [17], the learning rate η_t will be carefully designed instead of naively assigning it a constant value.

C. Performance Improvement for the AGA Solution

The channel noise in $\hat{\mathbf{g}}_t$ is harmful for both the convergence rate and the learning accuracy of the AGA solution. Therefore, we can improve the performance of the AGA solution by the following designs.

1) *Noise Variance Minimization Through Transceiver-Parameter Optimization:* Since the noise variance in (10) is $(1/2\alpha_t)\sigma_0^2$, minimizing the noise variance is equivalent to maximizing α_t. We assume that the transmit power of each client in every communication round is constrained by

$$\left\| a_{k,t}\mathbf{x}_{k,t} \right\|_2^2 \le P \quad \forall k. \tag{13}$$

Then, the transceiver-parameter optimization for the noise variance reduction can be formulated by

$$\begin{aligned} \mathcal{P}(B_0) \quad \max_{\mathbf{v}_t, \alpha_t} \ & \alpha_t \\ \textbf{s.t.} \ & \left\| a_{k,t}\mathbf{x}_{k,t} \right\|_2^2 \le P \quad \forall k \\ & \|\mathbf{v}_t\|_2^2 = 1 \end{aligned}$$

where $\mathbf{v}_t$ and α_t are the transceiver parameters in the uniform-forcing transmitter, which needs to be optimized. The details on solving $\mathcal{P}(B_0)$ will be investigated in Section IV.

2) *Convergence Speed Acceleration by Learning Rate Design:* It is known that if $\hat{\mathbf{g}}_t = \mathbf{g}_t$, the learning algorithm with the constant learning rate $\eta_t = (1/L)$ will converge to the optimal solution with a linear convergence rate. However, due to the noise in $\hat{\mathbf{g}}_t$, directly applying $\eta_t = (1/L)$ will result in a nonzero convergence gap from the optimal solution [30]. Therefore, an efficient learning rate η_t is desirable to speed up the convergence speed while achieving a zero convergence gap. The detailed design will be investigated in Section V.

IV. Data-and-CSI-Aware Transceiver-Parameter Optimization

In the traditional over-the-air fusion applications, e.g., [22]–[25], parameters $\mathbf{v}_t$ and α_t are optimized only based on the CSI. However, this CSI-only transceiver-adaptation is suboptimal for the AGA solution since the updated gradients are nonstationary. In this section, we propose the data-and-CSI-aware transceiver-parameter optimization to address the nonstationarity of the local gradients. In addition, a low-complexity scheme is proposed to further reduce the computation complexity with little performance degradation.

A. Traditional CSI-Only Transceiver-Parameter Optimization With Stationary Source Symbol Assumption

In traditional over-the-air fusion in consensus applications in sensor networks [22]–[25], the source symbols $\{x_{k,t,1}, \ldots, x_{k,t,B}\}$ at client k are assumed to be stationary over different rounds with $\mathbb{E}_{k,t,b}[|x_{k,t,b}|^2] = \chi_k$. One may approximately have $\|\mathbf{x}_{k,t}\|_2^2 = B\chi_k$ since the dimension of $\mathbf{x}_{k,t}$ (i.e., B) is usually large. In this case, it suffices to consider CSI-only parameter adaptation and the transceiver-parameter optimization becomes

$$\begin{aligned} \mathcal{P}(C) \quad \max_{\mathbf{v}_t, \alpha_t} \ & \alpha_t \\ \textbf{s.t.} \ & \alpha_t \le \frac{P}{B\chi_k} \left| \mathbf{v}_t^H \mathbf{h}_{k,t} \right|^2 \quad \forall k \qquad \text{(15a)} \\ & \|\mathbf{v}_t\|_2^2 = 1. \qquad \text{(15b)} \end{aligned}$$

In most existing works on the analog gradient (or model) aggregation solution for FL, such as [14]–[20], the transceiver-parameter optimization naively adopts the CSI-only optimization in $\mathcal{P}(C)$. However, the stationary source symbol assumption does not hold in these applications. Specifically, the transmit symbols in AGA schemes are given by $x_{k,t,b} = d_k g_{k,t,b}$. The stochastic gradient $\mathbb{E}_{k,t,b}[|g_{k,t,b}|^2]$ is not constant but decreases gradually as the model training converges. As a result, CSI-only transceiver-adaptation is suboptimal.

B. Proposed Data-and-CSI-Aware Transceiver-Parameter Optimization

In this section, we first introduce the concept of data-and-CSI-aware transceiver-parameter optimization to address the nonstationarity of the local gradients. Then, we elaborate the method to solve the problem.

1) Addressing the Nonstationarity With the Data State Information: The local gradient $\mathbf{g}_{k,t}$ can be decomposed into two components: the direction vector $\bar{\mathbf{g}}_{k,t}$ and the magnitude $E_{k,t}$, which are defined as follows:

$$E_{k,t} = \left\| \mathbf{g}_{k,t} \right\|_2 \tag{16}$$

$$\bar{\mathbf{g}}_{k,t} = \frac{\mathbf{g}_{k,t}}{E_{k,t}}. \tag{17}$$

Since we always have $\|\bar{\mathbf{g}}_{k,t}\|_2^2 = 1$ for all k and t, the nonstationarity in $\mathbf{g}_{k,t}$ is contributed by $E_{k,t}$. As such, we refer to $E_{k,t}$ as the data state information (DSI).

To address the problem caused by the nonstationarity, we propose a data-and-CSI-aware design, in which the server needs to collect both the CSI and DSI from client k. These CSI and DSI can be combined into a new state variable called

the effective channel coefficient, as follows:

$$\bar{\mathbf{h}}_{k,t} = \frac{\mathbf{h}_{k,t}}{E_{k,t}}. \tag{18}$$

Substituting $x_{k,t,b} = d_k g_{k,t,b}$ and $a_{k,t} = \sqrt{\alpha_t}([(\mathbf{v}_t^H \mathbf{h}_{k,t})^H]/[|\mathbf{v}_t^H \mathbf{h}_{k,t}|^2])$ into the power constraint (13), it becomes

$$\alpha_t \leq \frac{P\left|\mathbf{v}_t^H \mathbf{h}_{k,t}\right|^2}{d_k^2 \left\|\mathbf{g}_{k,t}\right\|_2^2} \quad \forall k. \tag{19}$$

Additionally, with the knowledge of both CSI and DSI, power constraint (19) can be converted to

$$\alpha_t \leq \frac{P}{d_k^2}\left|\mathbf{v}_t^H \bar{\mathbf{h}}_{k,t}\right|^2 \quad \forall k. \tag{20}$$

Finally, $\mathcal{P}(B_0)$ is equivalently represented by

$$\begin{aligned} \mathcal{P}(B) \ \max_{\mathbf{v}_t, \alpha_t} \ & \alpha_t \\ \textbf{s.t.} \ & \alpha_t \leq \frac{P}{d_k^2}|\mathbf{v}_t^H \bar{\mathbf{h}}_{k,t}|^2 \quad \forall k \\ & \|\mathbf{v}_t\|_2^2 = 1 \end{aligned}$$

where $d_k = (|\mathcal{D}_k|/|\mathcal{D}|)$ can be known *a priori* at the server since it is independent of t. Since the DSI (i.e., $E_{k,t}$) is a scalar, the additional overhead for collecting $E_{k,t}$ is quite small.[3]

Remark: The transceiver parameters $\mathbf{v}_t$ and α_t in $\mathcal{P}(B)$ will be adaptive to both the DSI and CSI in every communication round. As a result, a client with a large DSI will also be given higher priority compared to the CSI-only design.

2) Solving the Transceiver-Parameter Optimization: We solve $\mathcal{P}(B)$ in two steps. In step one, we derive the optimal α_t given $\mathbf{v}_t$ by using the primal decomposition. Specifically, according to the power constraint in (20), the optimal α_t given the received beamforming vector $\mathbf{v}_t$ can be derived in a closed-form expression as follows:

$$\alpha_t(\mathbf{v}_t) = P \min_k \frac{1}{d_k^2}\left|\mathbf{v}_t^H \bar{\mathbf{h}}_{k,t}\right|^2. \tag{22}$$

One can see that with the knowledge of both CSI and DSI, the power constraint (13) is converted to (19), and the resulting solution can then fully exploit the transmit power to reduce the noise variance.

In step two, we plug $\alpha_t(\mathbf{v}_t)$ into $\mathcal{P}(B)$, and then obtain an equivalent optimization problem with $\mathbf{v}_t$ as the only variable:

$$\begin{aligned} \mathcal{P}(B_1) \ \max_{\mathbf{v}_t} \ & \min_k \ \frac{P}{d_k^2}\left|\mathbf{v}_t^H \bar{\mathbf{h}}_{k,t}\right|^2 \\ \textbf{s.t.} \ & \|\mathbf{v}_t\|_2^2 = 1. \end{aligned}$$

Problem $\mathcal{P}(B_1)$ has the same form as the max–min fair beamforming problem for physical-layer multicasting [31, Sec. V], which has been extensively investigated and can be solved by the semi-definite relaxation (SDR) technique [31] or the successive convex approximation technique [32], [33]. We outline an efficient algorithm [33] for solving $\mathcal{P}(B_1)$ in Appendix A.

[3]A similar feedback mechanism can be found in [16] for the analog model aggregation solution with almost the same signaling overhead.

C. Low-Complexity Data-and-CSI-Aware Transceiver-Parameter Optimization Based on Selection Combining

In $\mathcal{P}(B)$, the number of constraints increases with the number of users K. The worst case complexity to solve $\mathcal{P}(B)$ with SDR is $\mathcal{O}([M+K]^{3.5})$, which increases significantly with K. On the other hand, the beamforming gain might decrease as K increases since we usually have $M < K$. Therefore, a low-complexity optimization scheme is desirable for large K cases.

In this section, we propose a low-complexity optimization scheme based on the selection combining rule. To this end, we add one constraint that $\mathbf{v}_t$ is the one-hot vector. The set of M-dimensional one-hot vectors is defined by

$$\mathbb{H} = \left\{ \mathbf{v} \middle| \mathbf{v} = [v_1, \ldots, v_M]^T, v_m \in \{0, 1\}, \sum_{m=1}^{M} v_m = 1 \right\}.$$

The transceiver-parameter optimization problem is given by

$$\mathcal{P}(D) \ \max_{\mathbf{v}_t \in \mathbb{H}} \ \min_k \ \frac{P}{d_k^2 E_{k,t}^2}\left|\mathbf{v}_t^H \mathbf{h}_{k,t}\right|^2.$$

Since $\mathbf{v}_t$ is a one-hot vector, we define m^* by

$$m^* = \arg\max_m \ \min_k \ \frac{1}{d_k^2 E_{k,t}^2}\left|h_{k,t,m}\right|^2$$

where $\mathbf{h}_{k,t} = [h_{k,t,1}, \ldots, h_{k,t,M}]^T$. The optimal $\mathbf{v}_t = [v_{t,1}, \ldots, v_{t,M}]^T$ of $\mathcal{P}$(D) is given by $v_{t,m^*} = 1$ and $v_{t,m} = 0$ for all $m \neq m^*$. One can see that by adding the constraint $\mathbf{v}_t \in \mathbb{H}$, the complexity for solving $\mathcal{P}$(D) is $\mathcal{O}(KM)$, which is much smaller than solving the original $\mathcal{P}(B)$. Meanwhile, this low-complexity scheme can still enjoy the diversity gain of multiple receive antennas, which will be verified by simulations.

V. Convergence Speed Acceleration by Learning Rate Design

In this section, we further design an efficient η_t to accelerate the convergence speed. In particular, we first present two learning rate designs in Sections V-A and V-B. Both designs can achieve the optimal solution for the FL problem $\mathcal{P}(A)$, yet they both have shortcomings. In the end, the overall learning rate design is proposed in Section V-C, which combines the advantages of the two previous designs.

A. Diminishing Learning Rate Scheme

In the traditional training task, the SGD algorithm usually calculates the gradient with only a part of the whole training data set (which is known as a batch) in order to reduce the complexity.[4] As such, the variance of the gradient estimation noise is a constant determined by the batch size, and the diminishing learning rate has been widely used to tackle that noisy estimated gradient [30, Th. 4.7]. However, the main difference

[4]Note that since the full gradient are adopted in this article, we do not consider this "sampling noise" issue and focus only on the channel noise. The aggregated gradient is still "stochastic" since the channel noise is a random variable.

here is that when using the proposed transceiver-parameter optimization in Section IV, the noise in $\hat{\mathbf{g}}_t$ comes from the wireless channel whose variance varies over the communication rounds due to the nonstationary local gradients and the time-varying wireless channels. Fortunately, we will show that the SGD with a diminishing learning rate can still guarantee the convergence.

Lemma 1 (Diminishing Learning Rate): Under Assumptions 1 and 2, suppose that the learning rate η_t for all t is designed as follows:

$$\eta_t = \frac{\beta}{\tau + t} \tag{26}$$

where $\tau > 0$, $\beta > (1/\mu)$ and $\eta_1 \leq (1/L)$. Then, the upper bound of $\Delta_{t+1} = \mathbb{E}[f(\mathbf{w}_{t+1})] - f^*$ for the SGD algorithm is given by

$$\Delta_{t+1} \leq \max\left\{\frac{\beta^2 LB\sigma_0^2}{4\bar{\alpha}_t(\beta\mu - 1)(\tau + t + 1)}, \frac{\Delta_t(\tau + t)}{\tau + t + 1}\right\} \tag{27}$$

where $f^* = f(\mathbf{w}^*)$, $\mathbf{w}^*$ is the optimal solution of $\mathcal{P}(A)$, and $\bar{\alpha}_t = \mathbb{E}[\alpha_t]$ in which the expectation is taken over all the received noise $\hat{\mathbf{u}}_i$ for all $i < t$.

Proof: See Appendix B. ∎

One can observe from (27) that the expected optimality gap Δ_t may not monotonically decrease with the increase of t, due to the change of $\bar{\alpha}_t$ in different communication rounds. In the following, we will prove that the convergence can still be established under some mild conditions.

Corollary 2 (Sublinear Convergence Rate): Under Assumptions 1 and 2, Δ_{t+1} in (27) approaches zero with the convergence rate $\mathcal{O}(1/t)$ if the channel power gain is bounded away from zero.

Proof: Define the following function:

$$J(t, \bar{\alpha}_t) = \frac{\beta^2 LB\sigma_0^2}{4\bar{\alpha}_t(\beta\mu - 1)(\tau + t + 1)}. \tag{28}$$

We will prove that $\lim_{t\to\infty} J(t, \bar{\alpha}_t) = 0$. The convergence is established based on this. See the detailed proof in Appendix C. ∎

As shown above, the SGD with η_t in (26) achieves sublinear convergence. Then, the remaining task is the fine-tuning of τ and β.

1) Fine-Tuning τ: We first discuss the fine-tuning of τ, while assuming β is fixed. As seen from (27), the second term in (27) does not contain any variables related to the wireless transmission. Therefore, when the $\bar{\alpha}_t$ are sufficiently large for all t, the first term in (27) is very small, and we will obtain

$$\Delta_{t+1} \leq \frac{\Delta_1(\tau + 1)}{\tau + t + 1}$$

which is the best convergence performance we may realize by using η_t in (26). In this case, we prefer a small τ to reduce the above gap. However, for small $\bar{\alpha}_t$, the first term in (27) dominates the gap

$$\Delta_{t+1} \leq J(t, \bar{\alpha}_t)$$

and we need a large τ to suppress $J(t, \bar{\alpha}_t)$. In practice, it is better to set τ slightly large to guarantee the robust convergence, especially when the channel condition is not good or the power budget is limited.

2) Fine-Tuning β: In (27), β is only contained in the first term. Therefore, if τ is given, fine-tuning β to minimize

$$\max\left\{\frac{\beta^2 LB\sigma_0^2}{4\bar{\alpha}_t(\beta\mu - 1)(\tau + t + 1)}, \frac{\Delta_t(\tau + t)}{\tau + t + 1}\right\}$$

for all t is equivalent to minimizing the first term in the max operation. The optimal β to minimize $([\beta^2 LB\sigma_0^2]/[4\bar{\alpha}_t(\beta\mu - 1)(\tau + t + 1)])$ in (27) for all t is the solution for minimizing function $[\beta^2/(\beta\mu - 1)]$

$$\beta = \frac{2}{\mu}. \tag{29}$$

Recall that the fine-tuning of τ and β should also satisfy the constraint $\eta_1 \leq (1/L)$. Hence, one practical method is to first set $\beta = (2/\mu)$, and then fine tune τ in $[(2L/\mu), +\infty)$.

B. Transmission Quality Awareness Learning Rate

The SGD with η_t in (26) achieves sublinear convergence. However, η_t is chosen by fine-tuning τ and β according to the trend of Δ_t, which is determined by the overall quality of all the analog over-the-air gradient aggregation rounds. When the transmission quality varies drastically, τ and β have to be fine-tuned to cater to the worst case performance, resulting in a slow convergence speed.

In this section, we propose a new learning rate design, which may vary adaptively to the wireless transmission quality. To characterize the transmission quality, we define one new variable as follows:

$$\zeta_t = \frac{\mathbb{E}\left[\|\mathbf{g}_t\|_2^2\right]}{\mathbb{E}\left[\|\mathbf{g}_t\|_2^2 + \frac{B}{2\alpha_t}\sigma_0^2\right]} \tag{30}$$

which is the ratio of the expected power of the useful information to the expected total received power at time index t, in which the two expectations are both taken over all the received noise $\hat{\mathbf{u}}_i$ for all $i < t$. Based on ζ_t in (30), we will propose a novel learning step η_t for the tth round to track the instantaneous transmission quality.

Lemma 2 (Transmission Quality Awareness Learning Rate): Under Assumptions 1 and 2, suppose that the learning rate η_t for all t is designed as follows:

$$\eta_t = \frac{1}{L}\zeta_t. \tag{31}$$

Then, the upper bound of Δ_{t+1} is given by

$$\Delta_{t+1} \leq \prod_{i=1}^{t}\left(1 - \frac{\mu}{L}\zeta_i\right)\Delta_1. \tag{32}$$

In addition, the improvement of the expected optimality gap at the tth round is lower bounded by

$$\Delta_t - \Delta_{t+1} \geq \frac{\zeta_t}{2L}\mathbb{E}\left[\|\mathbf{g}_t\|_2^2\right] \tag{33}$$

where the expectation is taken over all $\hat{\mathbf{u}}_i$ for $i < t$.

Proof: See Appendix D. ■

As shown in (33), the proposed learning rate in (31) guarantees the decrease of Δ_t in every round. In the following, we analyze the convergence of the SGD using the proposed learning rate in (31). Recall that, in Section IV, the parameters of the transceiver in the AGA solution are optimized by exploiting the DSI $E_{k,t} = \|\mathbf{g}_{k,t}\|_2$ from all clients. In order to further analyze the convergence behavior of the AGA solution, we derive a lower bound of ζ_t based on a new variable defined by $C_t = \sum_{k=1}^{K} d_k^2 E_{k,t}^2$.

Lemma 3 (Lower Bound of ζ_t): Under Assumptions 1 and 2, suppose the learning rate in (31) is used, given any training accuracy $\delta_0 > 0$, for any t that satisfies $\Delta_t \geq \delta_0$, we have $\zeta_t \geq \bar{\zeta}_t$ where

$$\bar{\zeta}_t = \frac{2\mu}{2\mu + \frac{C_t B \sigma_0^2}{2\delta_0 P \min_k |h_{k,t}|^2}}. \tag{34}$$

Proof: See Appendix E. ■

Using Lemma 3, we present the convergence analysis for a given training accuracy δ_0 in the following theorem.

Theorem 1 (Linear Convergence Rate): Under Assumptions 1 and 2, suppose the local gradients are uniformly bounded by a constant as $C_0 \geq C_t$ for all t, and the learning rate in (31) is used; given any training accuracy $\delta_0 > 0$, for any t that satisfies $\Delta_t \geq \delta_0$, we have

$$\Delta_t \leq \left(1 - \frac{\mu}{L}\zeta_0\right)^{t-1} \Delta_1 \tag{35}$$

where ζ_0 is given by

$$\zeta_0 = \frac{2\mu}{2\mu + \frac{C_0 B \sigma_0^2}{2\delta_0 P \min_k |h_{k,t}|^2}}.$$

In other words, Δ_t converges linearly with a rate faster than $1 - (\mu/L)\zeta_0$ until it reaches δ_0.

Proof: The proof is straightforward by setting $\zeta_0 = \min_t \bar{\zeta}_t$. ■

In conclusion, given one training accuracy requirement δ_0, the expectation of the optimal gap Δ_{t+1} will converge with a rate faster than $1 - (\mu/L)\zeta_0$. Then, given a smaller δ_0', the bound of the convergence rate is switched to $1 - (\mu/L)\zeta_0'$. This way, the SGD with the learning rate in (31) can achieve arbitrary precision, and the algorithm will finally converge to the optimal solution.

C. Overall Learning Rate Design

1) Discussions on Performance and Implementation: From Lemma 3, the convergence speed of the transmission quality awareness learning rate at round t is highly related to (C_t/δ_0). Specifically, if (C_t/δ_0) is large, we have a large $\bar{\zeta}_t$, as shown in (34), and thus a high convergence speed. In a special case where all the data samples stored in the clients are i.i.d., we have $\mathbf{g}_t = \mathbf{g}_{k,t}$ for all k, and then C_t and δ_0 decrease to zero at the same order of speed. As a result, $\bar{\zeta}_t$ in (34) remains a constant order for all t so that the expected optimality gap Δ_t approaches to zero at a very fast speed.

However, in practice, the data samples are not i.i.d., and they are also not entirely unrelated. It is pointed out in [9] that when δ_0 is higher than some threshold, one can observe that C_t and δ_0 decrease at the same order of speed. However, beyond this threshold, C_t may float up and down around some value. As a result, the convergence speed by using the learning rate in (31) decreases as δ_0 decreases. Generally speaking, this threshold for δ_0 is determined by how the training samples are distributed in the local data sets, such as the relation among the local data sets, as well as the size of the local data sets. Fortunately, it is also verified in [9] that, the threshold for δ_0 is small enough for several real federated data examples (see [9, Appendix C.3.2]).

Hence, the following two conditions are necessary for the transmission-quality-awareness learning rate in (31) to outperform the diminishing learning rate in (26).

1) *ζ_t Is Not Too Small:* According to (32) and (33), the improvement of the convergence speed by using the learning rate in (31) is significant only when the transmission quality is high.
2) *t Is Not Too Large:* As discussed above, the fast convergence speed of the learning rate in (31) can be maintained when δ_0 is higher than some threshold in practice. When t is large, Δ_t and $\|\mathbf{g}_t\|_2^2$ have been trained very small, and thus the factor $1 - (\mu/L)\zeta_0$ in (35) is close to 1, resulting in a slow convergence speed.

Another important issue is that due to the channel noise, the exact value of ζ_t in (30) is hard to be obtained, and instead, ζ_t should be replaced with its estimated value $\hat{\zeta}_t$ during practical implementation. In this article, we use the following variable as the estimation of ζ_t:

$$\hat{\zeta}_t = \frac{\left\|\hat{\mathbf{g}}_t\right\|_2^2 - \frac{B}{2\alpha_t}\sigma_0^2}{\left\|\hat{\mathbf{g}}_t\right\|_2^2} \tag{36}$$

since $\mathbb{E}_{\hat{\mathbf{u}}_t}[\|\hat{\mathbf{g}}_t\|_2^2] = \|\mathbf{g}_t\|_2^2 + (B/2\alpha_t)\sigma_0^2$. Our simulations will show that this choice of $\hat{\zeta}_t$ in (36) works well in practice.

2) Overall Proposed Learning Rate: In the end, we propose the following learning rate for the SGD algorithm in the AGA solution by combining the two learning rate designs in (26) and (31):

$$\eta_t = \begin{cases} \frac{1}{L}\hat{\zeta}_t^+, & \text{if } t < \tilde{t} \text{ or } \hat{\zeta}_t > \tilde{\zeta} \\ \frac{\beta}{\tau + t}, & \text{otherwise} \end{cases} \tag{37}$$

where $\tilde{t}$ and $\tilde{\zeta}$ are the predesigned thresholds. One can see that the proposed learning rate may accelerate the convergence speed when the transmission quality ζ_t is high, and the final convergence is guaranteed by the diminishing learning rate for a large t.

VI. Simulation Results and Discussions

This section evaluates the performance of the proposed AGA solution on the linear regression and shallow neural network. In particular, in the simulation with linear regression, the data samples in different clients are highly related. However, in the simulation with a shallow neural network, the samples in different clients are entirely unrelated. For simplicity, we assume all the clients have the same path loss of 106 dBm, and noise variance $\sigma_0^2 = -117$ dBm. Then, all the

fading channel coefficients are randomly generated according to the standard complex Gaussian distribution. The randomness of the channel coefficients and the channel noise are smoothed by 100 training processes.

A. Introduction of the Curves in the Figures

In the simulation, we consider four baseline schemes to benchmark the proposed scheme:

1) *Baseline 1 [Centralized]:* This approach benchmarks the theoretically optimal performance. The learning model in this approach is updated by the noise-free gradients and the constant learning rate $\eta_t = (1/L)$.
2) *Baseline 2 [Traditional]:* This approach naively uses the traditional CSI-only transceiver-parameter optimization for the over-the-air fusion introduced in Section IV-A. The transceiver parameters are optimized by assuming the local gradients follow a uniform distribution, and the power constraint is guaranteed by setting a large χ so that $\|\mathbf{x}_{k,t}\|_2^2 \leq B\chi$ for all k and t.
3) *Baseline 3 [One Bit]:* In this approach, the one-bit gradient quantization proposed in [21] is used so that the transmit symbols become stationary.
4) *Baseline 4 [Open Loop]:* This approach employs the transceiver proposed in [17], which does not address the small-scale fading. In particular, in round t, the base station broadcasts one random $\mathbf{v}_t$ first. Then, each client compensates the phase of the channel and sends its local gradient with the maximum power P according to its local CSI and DSI.

Note that baselines 2 and 3 belong to the CSI-only designs since the transceiver parameters are optimized without the DSI feedback, and then the diminishing learning rate in (26) is used for model update.

Two curves are drawn to illustrate the performance of the proposed AGA solution, both of which adopt the learning rate proposed in (37). However, the adopted data-and-CSI-aware transceiver-parameter optimization are slightly different.

1) *Proposed BF:* The transceiver parameters are optimized by $\mathcal{P}(B)$ with standard receive beamforming $\mathbf{v}_t \in \mathbb{C}^{M\times 1}$.
2) *Proposed SC:* The transceiver parameters are optimized with selection combining.

Finally, one auxiliary curve is drawn to assist the performance evaluation.

1) *Auxiliary:* The transceiver-parameter optimization uses the selection combining solution. However, the diminishing learning rate in (26) is used for the SGD.

By comparing the auxiliary curve with baselines 2–4, one can verify the performance gain of the proposed data-and-CSI-aware transceiver-parameter optimization. By comparing the auxiliary curve with "Proposed SC," one can verify the performance gain of the proposed learning rate design.

B. Linear Regression

We first examine the proposed AGA solution with one simple linear regression task. The optimality gap $f(\mathbf{w}_t) - f^*$ for the training process is used as the performance metric.

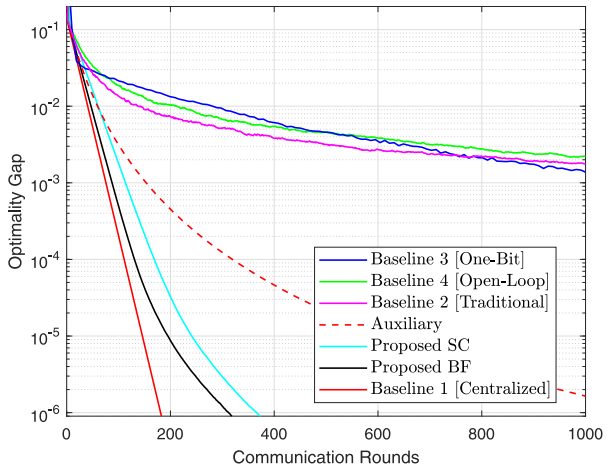


Fig. 2. Convergence of the optimality gap for linear regression, when $K = 5$, $M = 4$, and $P = -10$ dBm.

1) Objective Function: As introduced in Section II-B, the global objective function of the linear regression is

$$f(\mathbf{w}) = \frac{1}{2|\mathcal{D}|}\sum_{k=1}^{K}\sum_{d=1}^{|\mathcal{D}_k|}\left|q_{k,d} - \mathbf{w}^T\mathbf{s}_{k,d}\right|^2. \tag{38}$$

The gradient and Hessian of $f(\mathbf{w})$ is

$$\nabla f(\mathbf{w}) = \frac{1}{|\mathcal{D}|}\left(\sum_{k,d}\mathbf{s}_{k,d}\mathbf{s}_{k,d}^T\right)\mathbf{w} - \frac{1}{|\mathcal{D}|}\sum_{k,d}q_{k,d}\mathbf{s}_{k,d}$$

$$\nabla^2 f(\mathbf{w}) = \frac{1}{|\mathcal{D}|}\sum_{k,d}\mathbf{s}_{k,d}\mathbf{s}_{k,d}^T$$

respectively. Then, one may obtain μ and L by calculating the minimum and maximum eigenvalues of $\nabla^2 f(\mathbf{w})$. According to the discussions in Section V-A, we set $\beta = (2/\mu)$ for the diminishing learning rate, and then τ is chosen from the set $\{(2L/\mu), (20L/\mu), (200L/\mu)\}$.

2) Data Sets: In the simulation, we assume all clients store a total of $|\mathcal{D}| = 6 \times 10^5$ data samples, and the size of the local data set for every client is the same, i.e., $|\mathcal{D}|/K$. The data samples are generated according to the function

$$q_{k,d} = \mathbf{c}^T\mathbf{s}_{k,d} + \mathbf{u}_{k,d}$$

where $B = 10$ is the dimension of $\mathbf{s}_{k,d}$, $\mathbf{c} = [1, 2, \ldots, 10]^T$, and $\mathbf{u}_{d,k}\sim\mathcal{N}(0, \sigma_k^2\mathbf{I}_{10})$. To imitate the non-i.i.d. data distributions, we randomly generate σ_k for each client k, and then normalize all the variances by $(1/K)\sum_k \sigma_k^2 = 1$.

3) Simulation Results: Fig. 2 illustrates the convergence of the optimality gap for different schemes when $K = 5$, $M = 4$, and $P = -10$ dBm. We set $\chi = 0.5$ as the approximated second moment for the local gradient elements for baselines 2 and 4, and the thresholds for the proposed learning rate are $\tilde{t} = 400$ and $\tilde{\zeta} = 0.1$.

Since the samples in the local data sets are highly correlated and we have a large $\bar{\zeta}_t$ when the optimality gap is not small, as expected, we observe that the curve of each of the proposed schemes achieves a fast linear convergence rate when the optimality gap is larger that 10^{-4}. However, when t becomes

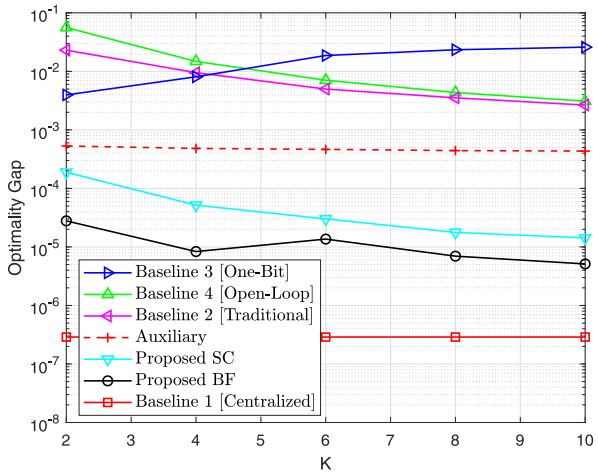


Fig. 3. Optimality gap for K increasing from 2 to 10, when $M = 4$, $P = -10$ dBm, and $t = 200$.

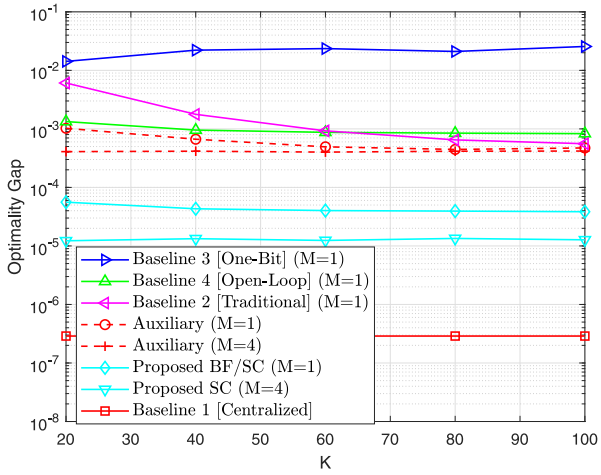


Fig. 4. Optimality gap for K increasing from 20 to 100, when $P = -10$ dBm and $t = 200$.

larger, the convergence speed decreases due to the limited size of the local data sets. In addition, the selection combining scheme only suffers a small performance loss compared to the beamforming optimization scheme while costing much lower complexity. On the other hand, baselines 2–4 and the auxiliary curve all achieve a slow sublinear convergence. The auxiliary curve shows significant performance gain compared to these baselines, thanks to the tracking of the nonstationarity of the local gradients during the transceiver-parameter optimization.

Fig. 3 shows the optimality gap for different K when $t = 200$ to verify that the performance of the proposed AGA solution will not decrease when the user number K increases. All the other simulation parameters are the same as those in Fig. 2. We observe that the optimality gap in most curves even decreases as K increases. This is because the weighted local gradients $\mathbf{x}_{k,t} = d_k \mathbf{g}_{k,t}$ become smaller with the decease of $d_k = (1/K)$. However, in the "Proposed BF" curve, the performance of $K = 4$ is better than that of $K = 6$. This is because when $K > M$, the multi-antenna gain decreases significantly. For the same reason, the performance gap between "Proposed BF" and "Proposed SC" becomes smaller when $K > M$. Moreover, the auxiliary curve decreases very slowly, which implies that under this simulation setup, the optimality gap is dominated by the second term in (27), which is independent of the SNR. Furthermore, the optimality gap of baseline 3 even becomes larger as K increases since the gradient estimation error increases due to the one-bit quantization on the local gradients.

Next, in Fig. 4, we continue investigating the optimality gap vs K, in which K is increased by ten times. We set $M = 1$ since the complexity of beamforming optimization is very high. Meanwhile, we still draw the $M = 4$ curves for the low-complexity selection combining scheme. The proposed scheme still has superior performance, and we observe that the performance varies only a little with the increase of K. This is because we have fixed the total size of the data samples $|\mathcal{D}|$, and when K is increased beyond some threshold, the size of the local data sets becomes so small that $\bar{\zeta}_t$ decreases drastically [9]. As a result, in the tth round, although d_k deceases, $\|\mathbf{g}_{k,t}\|_2^2$ becomes large due to large K, and $d_k^2 \|\mathbf{g}_{k,t}\|_2^2$ may not necessarily decrease. Fortunately, although we do not observe the same performance gain in Fig. 3, the performance also does not degrade for large K.

C. Shallow Neural Network

In this section, we examine the proposed AGA solution on a shallow neural network, which is exploited to identify the handwritten digits from zero to nine. Besides the optimality gap, we will also provide the corresponding test accuracy for the performance evaluation.

1) Objective Function: As introduced in Section II-B, the global objective function for $\mathcal{P}(A)$ is given by

$$f(\mathbf{w}) = \frac{1}{|\mathcal{D}|} \sum_{k=1}^{K} \sum_{d=1}^{|\mathcal{D}_k|} \text{CrossEntropy}\big(\hat{\mathbf{q}}\big(\mathbf{w}; \mathbf{s}_{k,d}\big), \mathbf{q}_{k,d}\big) + \lambda \|\mathbf{w}\|_2^2.$$

We denote the dimensions of $\mathbf{s}_{k,d}$ and $\mathbf{q}_{k,d}$ by B_s and B_q, respectively, and then the dimension of $\mathbf{w}$ is $B_q(B_s + 1)$.

2) Data Sets: In this section, we use the sample data sets provided by the Deep Learning Toolbox of MATLAB with "digitTrain4DArrayData" for training and digitTest4DArrayData for testing. The data samples $\mathbf{s}_d$ are images consisting of 28×28 pixels, so the dimension of $\mathbf{s}_d$ is $B_s = 784$. Each data set contains 500 images for every digital number from zero to nine, so the dimension of the label $\mathbf{q}_d$ is $B_q = 10$. Therefore, the dimension of $\mathbf{w}$ to be learned is $B = 7850$. We find that the L of the objective function on this training set is about 1. So we set $L = 1$ in the simulation. Additionally, for the diminishing learning rate, we set $\beta = 10$, and then τ is chosen from the set $\{10, 100, 1000, 10\,000\}$.

We fix the number of clients to $K = 5$. In addition, we imitate a very special case in which the local data sets are entirely unrelated (also non-i.i.d.) to verify the effectiveness of the proposed learning rate in extreme conditions. In particular, in the system, different clients store images labeled by different digital numbers, i.e., every client stores 1000 images, which are labeled by only two digital numbers. In this case, C_t decreases slowly even for small t [9], so that the

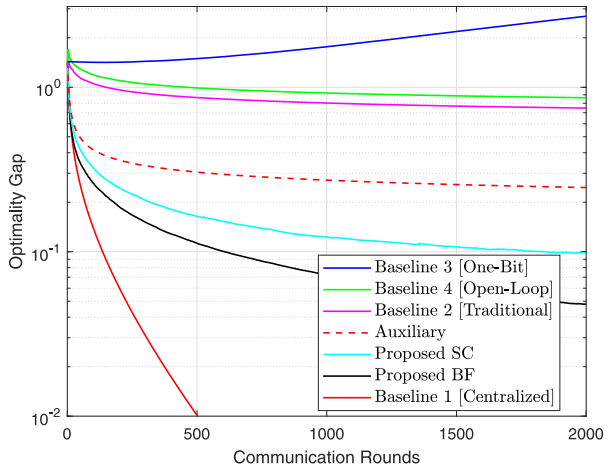


Fig. 5. Convergence of the optimality gap for handwritten-digit recognition.

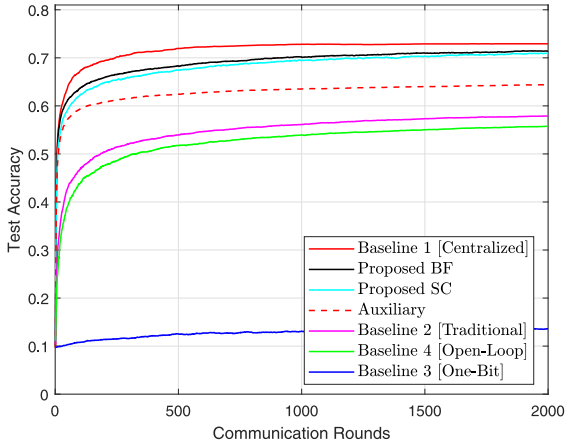


Fig. 6. Test accuracy of different communication rounds for handwritten-digit recognition.

proposed learning rate cannot keep the same fast convergence speed as that shown in the simulation for linear regression in Section VI-B. However, we will show that, even in this special case, the convergence of the proposed scheme is still much better than baselines by using the proposed learning rate in (37).

3) Simulation Results: Figs. 5 and 6 show the changes of the optimality gap and the test accuracy, respectively, as the communication round t increases. Since there are much more elements to be learned in $\mathbf{w}$, we increase the power constraint to $P = 20$ dBm. The number of antennas at the base station is still $M = 4$, and we set $\chi = 0.2$, $\tilde{t} = 1800$, and $\tilde{\zeta} = 0.05$. Moreover, we note that part of the gradient elements are in idle state with a very small value. Thus, we only consider the gradient elements that satisfy $\hat{g}_{t,b} \geq 0.1 \max_b \hat{g}_{t,b}$ for the estimation of ζ_t in (36) to improve the accuracy.

One can see that the auxiliary curve outperforms baselines 2–4, as expected. In addition, Proposed SC still achieves a superior convergence speed and higher learning accuracy than the auxiliary curve. We also observe that Proposed BF and Proposed SC achieve almost the same test accuracy, although their training losses are much different. Moreover, the performance gap between the proposed schemes and baseline 1 is relatively small, although the performance loss in

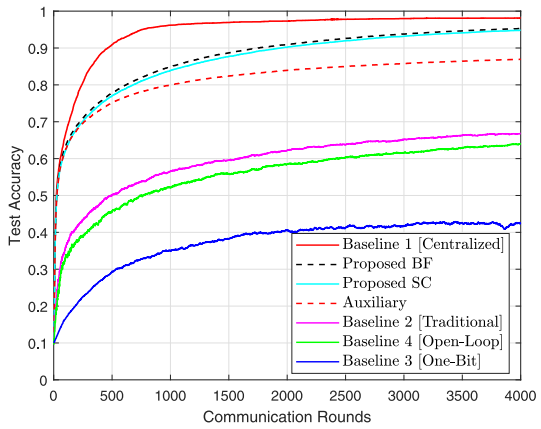


Fig. 7. Test accuracy of different communication rounds for handwritten-digit recognition when using a larger neural network.

Fig. 5 is slightly large. This implies that for the recognition task on the handwritten digits images, we do not need very high training accuracy in practice.

Finally, we enhance the test accuracy by adding a middle layer consisting of 100 "Relu" nodes in the neural network, while all the other simulation setups do not change. The test accuracy of this larger neural network is shown in Fig. 7. However, the optimality gap is difficult to verify since the training problem becomes nonconvex. It can be seen that the test accuracy is greatly improved thanks to the increase of the neural network size. In addition, our proposed solution still achieves superior performance compared to the baselines, even for this nonconvex training problem.

VII. Conclusion

In this article, we investigated the AGA solution to overcome the scarcity of radio resource on wireless FL applications. A novel design of both the communication system and learning algorithm has been proposed to improve the gradient aggregation quality as well as to accelerate the convergence speed by addressing the challenges caused by the nonstationary local gradients and the time-varying wireless channels. Specifically, the transceiver-parameter optimization is adaptive to both the data and the CSI, and the proposed learning rate is adaptive to the quality of the gradient estimation. The effectiveness of the proposed solution has been confirmed by both theoretical analyses and simulation results, and it has been verified that the proposed solution outperforms various state-of-the-art baseline schemes with a faster convergence speed and higher learning accuracy.

Appendix A
Method in [33] for Solving $\mathcal{P}(B_1)$

We define an auxiliary variable $\tilde{\mathbf{v}}_t$ that satisfies $\mathbf{v}_t = (\tilde{\mathbf{v}}_t / \|\tilde{\mathbf{v}}_t\|_2)$. Then $\mathcal{P}(B_1)$ is recast to

$$\mathcal{P}(E_1) \quad \min_{\tilde{\mathbf{v}}_t} \quad \left\|\tilde{\mathbf{v}}_t\right\|_2^2$$
$$\text{s.t.} \quad \left|\tilde{\mathbf{v}}_t^H \bar{\mathbf{h}}_{k,t}\right|^2 \geq d_k^2 \quad \forall k = 1, \ldots, K.$$

We employ the feasible point pursuit successive convex approximation (FPP-SCA) algorithm to solve $\mathcal{P}(E_1)$. First, the slack variable $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_K]^T \in \mathbb{R}^K$ is added as a slack penalty

$$\begin{aligned}\mathcal{P}(E_2) \ \min_{\tilde{\mathbf{v}}_t, \boldsymbol{\omega}} \ & \|\tilde{\mathbf{v}}_t\|_2^2 + \lambda\|\boldsymbol{\omega}\|_2^2 \\ \textbf{s.t.} \ & -\left|\tilde{\mathbf{v}}_t^H \bar{\mathbf{h}}_{k,t}\right|^2 \le \omega_k - d_k^2 \quad \forall k = 1, \ldots, K \\ & \omega_k \ge 0 \quad \forall k = 1, \ldots, K.\end{aligned}$$

By introducing $\boldsymbol{\omega}$, $\mathcal{P}(E_2)$ becomes always feasible. Note that $\mathcal{P}(E_2)$ and $\mathcal{P}(E_1)$ are equivalent only when the solution $\{\tilde{\mathbf{v}}_t, \boldsymbol{\omega}\}$ of $\mathcal{P}(E_2)$ has $\boldsymbol{\omega} = \mathbf{0}$. Therefore, we need to set a proper λ to make sure the solution of $\boldsymbol{\omega}$ approaches zero [33].

Next, $\mathcal{P}(E_2)$ is solved by the SCA technique, which is an iterative method. In particular, in the ith iteration, the following convex problem is solved:

$$\begin{aligned}\min_{\tilde{\mathbf{v}}_t, \boldsymbol{\omega}} \ & \left\|\tilde{\mathbf{v}}_t\right\|_2^2 + \lambda\|\boldsymbol{\omega}\|_2^2 \\ \textbf{s.t.} \ & -2\mathcal{R}\left\{\mathbf{z}^H \bar{\mathbf{h}}_{k,t} \bar{\mathbf{h}}_{k,t}^H \tilde{\mathbf{v}}_t\right\} + \left|\mathbf{z}_t^H \bar{\mathbf{h}}_{k,t}\right|^2 \\ & \le \omega_k - d_k^2 \quad \forall k = 1, \ldots, K \\ & \omega_k \ge 0 \quad \forall k = 1, \ldots, K\end{aligned}$$

where $\mathbf{z}$ is the solution of $\tilde{\mathbf{v}}_t$ in the $\{i-1\}$th iteration.

Appendix B
Proof of Lemma 1

Substituting the model update rule in (12) into the L-smooth assumption, we have

$$\begin{aligned}f(\mathbf{w}_{t+1}) - f(\mathbf{w}_t) &\le \mathbf{g}_t^T(\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\ &= -\eta_t \mathbf{g}_t^T \hat{\mathbf{g}}_t + \frac{1}{2}\eta_t^2 L\|\hat{\mathbf{g}}_t\|_2^2.\end{aligned}$$

Taking expectations on both sides with respect to the noise $\hat{\mathbf{u}}_t$, we have

$$\bar{\Delta}_{t+1} \le -\eta_t \mathbf{g}_t^T \mathbb{E}_{\hat{\mathbf{u}}_t}\left[\hat{\mathbf{g}}_t\right] + \frac{1}{2}\eta_t^2 L \mathbb{E}_{\hat{\mathbf{u}}_t}\left[\|\hat{\mathbf{g}}_t\|_2^2\right] \tag{42}$$

where $\bar{\Delta}_{t+1} = \mathbb{E}_{\hat{\mathbf{u}}_t}[f(\mathbf{w}_{t+1})] - f(\mathbf{w}_t)$.

Since the estimation noise is independent of $\mathbf{g}_t$, we have

$$\begin{aligned}\mathbb{E}_{\hat{\mathbf{u}}_t}\left[\left\|\hat{\mathbf{g}}_t\right\|_2^2\right] &= \mathbb{E}_{\hat{\mathbf{u}}_t}\left[\left\|\mathbf{g}_t + \hat{\mathbf{u}}_t\right\|_2^2\right] \\ &= \left\|\mathbf{g}_t\right\|_2^2 + \frac{B\sigma_0^2}{2\alpha_t}.\end{aligned} \tag{43}$$

Substituting (43) into (42), we get

$$\bar{\Delta}_{t+1} \le -\frac{1}{2}(2 - \eta_t L)\eta_t \left\|\mathbf{g}_t\right\|_2^2 + \frac{1}{4\alpha_t}\eta_t^2 LB\sigma_0^2. \tag{44}$$

Combining $\eta_t \le \eta_1$ and $\eta_1 \le (1/L)$, it yields $\eta_t L \le 1$. Then, (44) becomes

$$\bar{\Delta}_{t+1} \le -\frac{1}{2}\eta_t \left\|\mathbf{g}_t\right\|_2^2 + \frac{1}{4\alpha_t}\eta_t^2 LB\sigma_0^2. \tag{45}$$

From Corollary 1, $f(\mathbf{w})$ is μ-strongly convex as well, so it satisfies the Polyak–Łojasiewicz condition

$$2\mu\left(f(\mathbf{w}_t) - f^*\right) \le \left\|\mathbf{g}_t\right\|_2^2 \tag{46}$$

for all $\mathbf{w}$. Substituting (46) into (45), we have

$$\bar{\Delta}_{t+1} \le -\eta_t \mu\left(f(\mathbf{w}_t) - f^*\right) + \frac{1}{4\alpha_t}\eta_t^2 LB\sigma_0^2. \tag{47}$$

Rearranging (47) and taking the total expectation, it yields

$$\Delta_{t+1} \le (1 - \eta_t\mu)\Delta_t + \frac{1}{4\bar{\alpha}_t}\eta_t^2 LB\sigma_0^2. \tag{48}$$

Then, define $\nu_t = \max\{([\beta^2 LB\sigma_0^2]/[4\bar{\alpha}_t(\beta\mu - 1)]), \Delta_t(\tau + t)\}$, it yields

$$\Delta_t \le \frac{\nu_t}{t + \tau}. \tag{49}$$

Substituting $\eta_t = [\beta/(t+\tau)]$ in (26) and the inequality in (49) into (48), it follows that:

$$\begin{aligned}\Delta_{t+1} &\le \left(1 - \frac{\beta\mu}{t+\tau}\right)\frac{\nu_t}{t+\tau} + \frac{\beta^2}{4\bar{\alpha}_t(t+\tau)^2} LB\sigma_0^2 \\ &= \frac{t+\tau-1}{(t+\tau)^2}\nu_t - \frac{\beta\mu - 1}{(t+\tau)^2}\left(\nu_t - \frac{\beta^2 LB\sigma_0^2}{4\bar{\alpha}_t(\beta\mu - 1)}\right).\end{aligned}$$

Since $\beta > (1/\mu)$, we have $([\beta\mu - 1]/(t+\tau)^2)(\nu_t - [(\beta^2 LB\sigma_0^2)/(4\bar{\alpha}_t(\beta\mu - 1))]) \ge 0$. Thus

$$\begin{aligned}\Delta_{t+1} &\le \frac{t+\tau-1}{(t+\tau)^2}\nu_t \\ &\le \frac{\nu_t}{t+1+\tau}.\end{aligned}$$

Appendix C
Proof of Corollary 2

Since the channel power gain is bounded away from zero, we have nonzero α^*, where $\alpha^* \le \bar{\alpha}_t$ for all t. Using $\beta\mu > 1$, we have $J(t, \bar{\alpha}_t) \le J(t, \alpha^*)$, which yields

$$\lim_{t\to\infty} J(t, \alpha_t) \le \lim_{t\to\infty} J\left(t, \alpha^*\right).$$

Since $\lim_{t\to\infty} J(t, \alpha^*) = 0$, $\lim_{t\to\infty} J(t, \alpha_t) = 0$ is proved.

Then, the remaining task is to prove that, if we have

$$J(t, \bar{\alpha}_t) < \frac{\Delta_t(\tau + t)}{\tau + t + 1}$$

for all $t > \bar{t}$, the second term in (27) determines Δ_{t+1}, and in this case the convergence can still be established. Fortunately, using $\Delta_{t+1} \le ([\Delta_t(\tau + t)]/[\tau + t + 1])$, one may obtain

$$\Delta_{t+1} \le \frac{\Delta_{\bar{t}}\left(\tau + \bar{t}\right)}{\tau + t + 1}$$

for all $t > \bar{t}$. Therefore, we have $\Delta_{t+1} \to 0$ when $t \to +\infty$, and thus the corollary is proved.

Appendix D
Proof of Lemma 2

Similar to the proof of Lemma 1, based on the L-smooth assumption, we have the same relation in (42)

$$\bar{\Delta}_{t+1} \le -\eta_t \mathbf{g}_t^T \mathbb{E}_{\hat{\mathbf{u}}_t}\left[\hat{\mathbf{g}}_t\right] + \frac{1}{2}\eta_t^2 L \mathbb{E}_{\hat{\mathbf{u}}_t}\left[\left\|\hat{\mathbf{g}}_t\right\|_2^2\right] \tag{50}$$

where $\bar{\Delta}_{t+1} = \mathbb{E}_{\hat{\mathbf{u}}_t}[f(\mathbf{w}_{t+1})] - f(\mathbf{w}_t)$. Define

$$\bar{\zeta}_t = \frac{\|\mathbf{g}_t\|_2^2}{\|\mathbf{g}_t\|_2^2 + \frac{B}{2\alpha_t}\sigma_0^2}$$

and substitute it together with $\eta_t = (1/L)\zeta_t$ into (50), yielding

$$\begin{aligned}\bar{\Delta}_{t+1} &\leq -\eta_t\|\mathbf{g}_t\|_2^2 + \frac{\eta_t^2 L}{2\bar{\zeta}_t}\|\mathbf{g}_t\|_2^2 \\ &= -\frac{\zeta_t}{L}\left(1 - \frac{\zeta_t}{2\bar{\zeta}_t}\right)\|\mathbf{g}_t\|_2^2. \end{aligned} \tag{51}$$

1) Proof of (33): Substituting $\bar{\Delta}_{t+1} = \mathbb{E}_{\hat{\mathbf{u}}_t}[f(\mathbf{w}_{t+1})] - f(\mathbf{w}_t)$ into (51), we have

$$f(\mathbf{w}_t) - \mathbb{E}_{\hat{\mathbf{u}}_t}\big[f(\mathbf{w}_{t+1})\big] \geq \frac{\zeta_t}{L}\|\mathbf{g}_t\|_2^2 - \frac{\zeta_t^2}{2L\bar{\zeta}_t}\|\mathbf{g}_t\|_2^2. \tag{52}$$

It easy to verify that

$$\begin{aligned}\mathbb{E}\left[\frac{\zeta_t}{\bar{\zeta}_t}\|\mathbf{g}_t\|_2^2\right] &= \zeta_t\mathbb{E}\left[\|\mathbf{g}_t\|_2^2 + \frac{B}{2\alpha_t}\sigma_0^2\right] \\ &= \mathbb{E}\left[\|\mathbf{g}_t\|_2^2\right]\end{aligned}$$

where the expectation is taken over all the received noise $\hat{\mathbf{u}}_i$ for all $i < t$, since

$$\zeta_t = \frac{\mathbb{E}\left[\|\mathbf{g}_t\|_2^2\right]}{\mathbb{E}\left[\|\mathbf{g}_t\|_2^2 + \frac{B}{2\alpha_t}\sigma_0^2\right]}.$$

Therefore, taking the total expectation on both sides of (52), yields

$$\mathbb{E}\big[f(\mathbf{w}_t)\big] - \mathbb{E}\big[f(\mathbf{w}_{t+1})\big] \geq \frac{\zeta_t}{2L}\mathbb{E}\left[\|\mathbf{g}_t\|_2^2\right]. \tag{53}$$

Rearranging the above expression, we have

$$\begin{aligned}\frac{\zeta_t}{2L}\mathbb{E}\left[\|\mathbf{g}_t\|_2^2\right] &\leq \mathbb{E}\big[f(\mathbf{w}_t)\big] - f^* - \big(\mathbb{E}\big[f(\mathbf{w}_{t+1})\big] - f^*\big) \\ &= \Delta_t - \Delta_{t+1}.\end{aligned}$$

2) Proof of (32): Substituting the Polyak–Łojasiewicz condition $2\mu(f(\mathbf{w}_t) - f^*) \leq \|\mathbf{g}_t\|_2^2$ in (46) into (53), we have

$$\mathbb{E}\big[f(\mathbf{w}_t)\big] - \mathbb{E}\big[f(\mathbf{w}_{t+1})\big] \geq \frac{\zeta_t\mu}{L}\mathbb{E}\big[f(\mathbf{w}_t) - f^*\big].$$

Rearranging the above expression yields

$$\mathbb{E}\big[f(\mathbf{w}_{t+1})\big] - f^* \leq \left(1 - \frac{\zeta_t\mu}{L}\right)\big(\mathbb{E}\big[f(\mathbf{w}_t)\big] - f^*\big)$$

and thus the lemma is proved.

APPENDIX E
PROOF OF LEMMA 3

The proof will be elaborated in three steps as follows.

1) Lower Bound of $\mathbb{E}[\|\mathbf{g}_t\|_2^2]$*:* Since $f(\mathbf{w})$ is μ-strongly convex, the Polyak–Łojasiewicz condition is satisfied:

$$f(\mathbf{w}_t) - f^* \leq \frac{1}{2\mu}\|\mathbf{g}_t\|_2^2.$$

Taking the expectation on both sides over the received noise, we have

$$\mathbb{E}\left[\|\mathbf{g}_t\|_2^2\right] \geq 2\mu\delta_0. \tag{54}$$

2) Lower Bound of α_t*:* It can be verified that the receive SNR of the proposed transceiver will decrease if M decreases. Therefore, one could derive the lower bound of α_t in the worst case $M = 1$, in which $\alpha_t = \min_k\ [P/(d_k^2\|\mathbf{g}_{k,t}\|_2^2)]|h_{k,t}|^2$. Based on this, we further have

$$\begin{aligned}\alpha_t\sum_k d_k^2\|\mathbf{g}_{k,t}\|_2^2 &= \min_k\ \frac{P\sum_k d_k^2\|\mathbf{g}_{k,t}\|_2^2}{d_k^2\|\mathbf{g}_{k,t}\|_2^2}|h_{k,t}|^2 \\ &\geq \min_k\ P|h_{k,t}|^2.\end{aligned}$$

Therefore, it yields

$$\begin{aligned}\alpha_t &\geq \frac{P\min_k|h_{k,t}|^2}{\sum_k d_k^2\|\mathbf{g}_{k,t}\|_2^2} \\ &= \frac{P\min_k|h_{k,t}|^2}{C_t}.\end{aligned} \tag{55}$$

3) Lower bound of ζ_t*:* Based on (55), $[B/(2\alpha_t)]\sigma_0^2$ in (30) can be upper bounded by

$$\frac{B}{2\alpha_t}\sigma_0^2 \leq \frac{B\sigma_0^2 C_t}{2P\min_k|h_{k,t}|^2}. \tag{56}$$

Using (56), ζ_t in (30) is lower bounded by

$$\zeta_t \geq \frac{\mathbb{E}\left[\|\mathbf{g}_t\|_2^2\right]}{\mathbb{E}\left[\|\mathbf{g}_t\|_2^2\right] + \frac{B\sigma_0^2 C_t}{2P\min_k|h_{k,t}|^2}}. \tag{57}$$

Finally, we further substitute $\mathbb{E}[\|\mathbf{g}_t\|_2^2] \geq 2\mu\delta_0$ in (54) into the above expression, and the lemma is proved:

$$\zeta_t \geq \frac{2\mu\delta_0}{2\mu\delta_0 + \frac{B\sigma_0^2 C_t}{2P\min_k|h_{k,t}|^2}}.$$

REFERENCES

[1] P. Kairouz *et al.* (2019). *Advances and Open Problems in Federated Learning*. [Online]. Available: https://arxiv.org/abs/1912.04977.

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

[3] W. Y. B. Lim *et al.*, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, early access, Apr. 8, 2020, doi: 10.1109/COMST.2020.2986024.

[4] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, p. 12, 2019.

[5] J. Konečný, B. McMahan, and D. Ramage. (2015). *Federated Optimization: Distributed Optimization Beyond the Datacenter*. [Online]. Available: https://arxiv.org/abs/1511.03575

[6] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik. (2016). *Federated Optimization: Distributed Machine Learning for On-Device Intelligence*. [Online]. Available: https://arxiv.org/abs/1610.02527

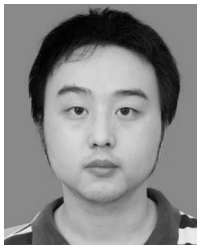
[7] H. B. McMahan, E. Moore, D. Ramage, and S. Hampson. (2016). *Communication-Efficient Learning of Deep Networks From Decentralized Data*. [Online]. Available: https://arxiv.org/abs/1602.05629

[8] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. (2019). *On the Convergence of FedAvg on Non-IID Data*. [Online]. Available: https://arxiv.org/abs/1907.02189

[9] A. K. Sahu, T. Li, M. Sanjabi, M. Zaheer, A. Talwalkar, and V. Smith. (2018). *On the Convergence of Federated Optimization in Heterogeneous Networks*. [Online]. Available: https://arxiv.org/abs/1812.06127
[10] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. (2016). *Federated Learning: Strategies for Improving Communication Efficiency*. [Online]. Available: https://arxiv.org/abs/1610.05492
[11] A. T. Suresh, F. X. Yu, S. Kumar, and H. B. McMahan, "Distributed mean estimation with limited communication," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3329–3337.
[12] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani. (2019). *FedPAQ: A Communication-Efficient Federated Learning Method With Periodic Averaging and Quantization*. [Online]. Available: https://arxiv.org/abs/1909.13014
[13] T. Chen, G. Giannakis, T. Sun, and W. Yin, "LAG: Lazily aggregated gradient for communication-efficient distributed learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5050–5060.
[14] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning based on over-the-air computation," in *Proc. IEEE ICC*, 2019, pp. 1–6.
[15] 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.
[16] 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.
[17] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Signal Process.*, vol. 68, pp. 2897–2911, Apr. 2020.
[18] 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.
[19] 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, Mar. 2020.
[20] M. M. Amiri and D. Gündüz, "Over-the-air machine learning at the wireless edge," in *Proc. IEEE SPAWC*, 2019, pp. 1–5.
[21] G. Zhu, Y. Du, D. Gunduz, and K. Huang. (2020). *One-Bit Over-the-Air Aggregation for Communication-Efficient Federated Edge Learning: Design and Convergence Analysis*. [Online]. Available: https://arxiv.org/abs/2001.05713
[22] 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.
[23] L. Chen, N. Zhao, Y. Chen, F. R. Yu, and G. Wei, "Over-the-air computation for IoT networks: Computing multiple functions with antenna arrays," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 5296–5306, Dec. 2018.
[24] D. Wen, G. Zhu, and K. Huang, "Reduced-dimension design of MIMO over-the-air computing for data aggregation in clustered IoT networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5255–5268, Nov. 2019.
[25] G. Zhu and K. Huang, "MIMO over-the-air computation for high-mobility multi-modal sensing," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6089–6103, Aug. 2019.
[26] S. Cai and V. K. N. Lau, "MSE tail analysis for remote state estimation of linear systems over multi-antenna random access channels," *IEEE Trans. Autom. Control*, vol. 65, no. 5, pp. 2046–2061, May 2020.
[27] M. Gastpar, "Uncoded transmission is exactly optimal for a simple Gaussian 'sensor' network," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 5247–5251, Nov. 2008.
[28] S. Cai and V. K. N. Lau, "Modulation-free M2M communications for mission-critical applications," *IEEE Trans. Signal Process. Netw.*, vol. 4, no. 2, pp. 248–263, Jun. 2018.
[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org
[30] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, 2018.
[31] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.
[32] L.-N. Tran, M. F. Hanif, and M. Juntti, "A conic quadratic programming approach to physical layer multicasting for large-scale antenna arrays," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 114–117, Jan. 2014.
[33] O. Mehanna, K. Huang, B. Gopalakrishnan, A. Konar, and N. D. Sidiropoulos, "Feasible point pursuit and successive approximation of non-convex QCQPs," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 804–808, Jul. 2015.

Huayan Guo (Member, IEEE) received the B.Eng. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2012, and the Ph.D. degree from Peking University, Beijing, in 2017.

He is currently a Postdoctoral Fellow with the Department of ECE, HKUST, Hong Kong. His research interests include wireless communications and machine learning for communications.



An Liu (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Peking University, Beijing, China, in 2011 and 2004, respectively.

From 2008 to 2010, he was a Visiting Scholar with the Department of ECEE, University of Colorado at Boulder, Boulder, CO, USA. He was a Postdoctoral Research Fellow from 2011 to 2013, a Visiting Assistant Professor in 2014, and a Research Assistant Professor from 2015 to 2017, with the Department of ECE, HKUST. He is currently a Distinguished Research Fellow with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. His research interests include wireless communications, stochastic optimization, compressive sensing, and machine/deep learning for communications.

Dr. Liu is serving as an Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and IEEE WIRELESS COMMUNICATIONS LETTERS.

Vincent K. N. Lau (Fellow, IEEE) received the B.Eng. degree (First-Class Hons. with Distinction) from the University of Hong Kong, Hong Kong, in 1992, and the Ph.D. degree from Cambridge University, Cambridge, U.K., in 1997.

He was with Bell Labs, Murray Hill, NJ, USA, from 1997 to 2004, and the Department of ECE, Hong Kong University of Science and Technology, Hong Kong, in 2004, where he is currently a Chair Professor and the Founding Director of Huawei-HKUST Joint Innovation Lab. His current research interests include wireless communications for 5G systems, content-centric wireless networking, wireless networking for mission-critical control, and cloud-assisted autonomous systems.