Received 6 October 2023; accepted 26 October 2023. Date of publication 31 October 2023; date of current version 20 November 2023.

Digital Object Identifier 10.1109/OJCOMS.2023.3328931

Joint Pre-Equalization and Receiver Combining Design for Federated Learning With Misaligned Over-the-Air Computation

JIANDA WANG [1,2] (Graduate Student Member, IEEE), AND SHUAISHUAI GUO [1,2] (Senior Member, IEEE)

[1] School of Control Science and Engineering, Shandong University, Jinan 250061, China
[2] Shandong Key Laboratory of Wireless Communication Technologies, Shandong University, Jinan 250061, China

CORRESPONDING AUTHOR: S. GUO (e-mail: shuaishuai_guo@sdu.edu.cn)

The work was supported in part by the National Natural Science Foundation of China under Grant 62171262; in part by the Shandong Provincial Natural Science Foundation under Grant ZR2021YQ47; in part by the Major Scientific and Technological Innovation Project of Shandong Province under Grant 2020CXGC010109; and in part by the Tashan Young Scholar under Grant tsqn201909043.

ABSTRACT With the growth of terminal devices and data traffic, privacy concerns have inspired an innovative edge learning framework, called federated learning (FL). Over-the-air computation (OAC) has been introduced to reduce communication overhead for FL, however, requires stringent time alignment. Misaligned OAC has been proposed by recent research where the symbol-timing misaligned superimposed signal can be recovered via whitening matched filtering and sampling (WMFS), followed by maximum likelihood (ML) estimation. Similarly to aligned OAC, misaligned OAC also suffers from the straggler issue, leading to FL's poor performance under low EsN0. To solve this issue, we propose a novel framework of misaligned OAC FL for accurate model aggregation on wireless networks. First, we analyze the effect of aggregation error on the convergence of FL. Then, we formulate an optimization problem to minimize the distortion of the aggregation measured by mean square error (MSE) w.r.t. the transmitter equalization and receiver combining. Finally, a successive convex approximation (SCA)-based optimization algorithm is further developed to solve the resulting quadratic constrained quadratic programming. Comprehensive experiments show that the proposed algorithm achieves substantial learning performance improvement compared to existing baseline schemes and achieves the near-optimal performance of the ideal benchmark with aligned and noiseless aggregation.

INDEX TERMS Federated learning, multiple access channels, over-the-air computation, asynchronous, pre-equalization, receiver combining, successive convex approximation.

I. INTRODUCTION

OVER the past few decades, we have witnessed phenomenal growth in modern science and technology, especially artificial intelligence (AI), which has stimulated the dramatic increase in the use of mobile edge devices such as smartphones, tablets, and Internet of Things (IoT) sensors leading to a tremendous growth of global data traffic [1]. The enormous critical data can be utilized for real-time decision-making, predictive health care [2], etc., which can fuel the performance of machine learning techniques, especially deep learning, while being exploited to improve the user experience of AI model [3].

Traditional machine learning algorithms work in a centralized fashion, where all raw data is collected into a parameter server (PS) or cloud to train AI models [4]. To avoid the high data accumulation cost and privacy issues of centralized training, the recent trend is to deploy AI algorithms to the edge of distributed networks, with the substantial increase in computing power and storage capacity of modern smart terminals [5], [6]. This has inspired an innovative edge learning framework, called federated learning (FL) which enables edge devices to collaboratively learn a shared AI model while retaining all training data on individual devices to avoid compromising privacy [7], [8].

A typical FL algorithm such as the Federated Average (FedAvg) algorithm alternates among four phases until the global model converges. To be specific, at each communication round (outer iteration), the PS first broadcasts the current global model to active edge devices; next, the edge devices implement local mini-batch stochastic gradient descent several times, using their own data based on received current global model; and then, the edge devices updated local models or gradients back to the PS; finally, the PS aggregates these updates and obtains a new average global model [2], [7]. However, due to the repeated transmission of a large number of model parameters through wireless channels between the PS and edge devices during the iterative training process, the scarcity of communication resources (such as bandwidth, energy, and power) becomes a bottleneck for FL, mainly in the uplink communications [2].

In addition to general solutions such as speeding up local updates [9], [10], [11], discarding updates from slow-response edge device (stragglers) [12], [13], and parameters compression via quantization [14], [15] or sparsification [16], [17], recent research focuses on the implementation of uplink communication. Compared with traditional orthogonal multiple access schemes [18], [19], [20], analog over-the-air computation (OAC) is a promising solution, which allows edge devices to simultaneously transmit their local updates using complete radio resources in an uncoded manner. It exploits the superposition property of multiple access channels (MAC) to directly compute the desired function (i.e., sum function) of every single update without decoding each individual message [21], [22]. This way of combining the communication and computation reduces latency and bandwidth requirements, which alleviates the uplink communication bottleneck of FL to a great extent.

This has sparked a growing body of comprehensive research on OAC in the context of distributed machine learning, FL, and IoT networks, among others. In [23], the authors considered a wireless fading MAC for distributed machine learning and scheduled each entry of the gradient vector according to the channel conditions. In [24], the authors exploited the sparsity of gradients to compress the dimension of parameter vectors and proposed digital and analog (via OAC) distributed stochastic gradient descent (D-DSGD and A-DSGD) algorithms in Gaussian MAC and extended them to fading MAC in [6] and [25]. In [26], [27], the authors proposed a convergent over-the-air federated learning (COTAF) algorithm in Gaussian MAC and fading MAC with a time-varying precoding factor. In [28], the authors analyze the convergence behavior of FedAvg with aggregation errors, and then jointly optimize the transmission power and the PS scaling factor to improve performance. In [29], the authors proposed a Byzantine resilient OAC model aggregation scheme. In [30], the authors propose an online model updating with analog aggregation (OMUAA) algorithm in fading MAC.

The implementation of the above scheme often relies on the perfect channel state information (CSI) of the transmitter, while [31], [32], [33] shows that enough PS antennas and a proper receiver combining design can reduce the demand for perfect CSI. Other studies in CSI-free scenarios focus on signal processing [34], [35], [36] and receiver combining strategy for multi-antenna systems [37].

A. RELATED WORKS AND MOTIVATIONS

Another vein of research on OAC implementation focuses on synchronization over edge devices [22], [38], [39], [40]. Over-the-air FL requires multiple devices to synchronously transmit gradients to PS. In practice, achieving this strict synchronization is very expensive. In [39], a solution called AirShare has been developed for synchronizing sensors by broadcasting shared clocks. However, there is an additional cost involved in broadcasting and maintaining a shared clock, and for non-ideal hardware, there can be residual asynchronies among the signals at the PS. In [38], the sensing value of each sensor is modulated as the power of the transmission signal to relax synchronization requirements. This design converts the function calculation at the receiver into power detection, while synchronization error manifests as random noise which degrades computing performance. Reference [36] considered both the lack of perfect CSI at the transmitter and the asynchronous transmission timing of each edge device, i.e., *misaligned OAC*, and proposed a *whitening matched filtering and sampling* (WMFS) scheme. By oversampling the superimposed but symbol-timing misaligned signal, a series of independent samples were obtained, from which a *maximum likelihood* (ML) estimator was designed to recover the arithmetic sum of transmitted symbols from different edge devices. This allows for transmitted symbol-timing asynchrony between different edge devices with a maximum transmission delay of less than a symbol period, which relaxes the synchronization requirement. However, for the single-antenna scenario considered in this paper, the accuracy of ML estimation is highly susceptible to communication noise, which is caused by the stragglers mentioned later.

It is well known that OAC is vulnerable to noise, due to the transmission of uncoded analog signals. From the perspective of the heterogeneity of communication capabilities among edge devices, the devices with weak channels (*stragglers in communications*) [41] affect the overall model aggregation error. Specifically, aligned OAC is achieved by linear superposition of signals. As component signals from different devices, they need to be aligned in magnitude for accurate aggregation at the receiver [23]. While, for devices with weaker channels, the transmitter typically requires additional power to counteract channel fading, which can account for most of the transmit power. Accordingly, a scaling factor is needed to scale the transmission power of each device as a whole, so that those devices experiencing deep channel fading can satisfy the maximum power constraint. The existence of stragglers determines the size of the scaling factor, which affects the aggregation error [42].

In this case, selecting an appropriate set of devices is a natural solution, the so-called device scheduling/selection. For instance, in [23], the channels experiencing deep fading do not transmit, and in [42], the author's goal is to maximize the number of participating devices while meeting the requirements of mean square error (MSE). Moreover, in [41], the author jointly optimizes the reconfigurable intelligent surface configuration and device selection. In [43], the author proposes a uniform-forcing transceiver design for OAC. In [44], the author proposes a dynamic learning rate (DLR) scheme for OAC based FL by defining the local learning rate and presents its convergence analysis, in order to mitigate the wireless distortion measured by MSE, taking into account both MISO and MIMO scenarios. In [45], the author considers receiver combining design and device selection in multiple parallel federated learning.

Despite the effectiveness of these works in aligned OAC, how to analyze and address the straggler problem in misaligned OAC and how to characterize the impact of misaligned OAC model aggregation errors on FL learning performance are still open issues, which motivate the current work in this paper.

For misaligned OAC, the pre-equalization at the transmitter is considered to be inaccurate, i.e., there are residual channel fading coefficients, which can be counteracted by the ML estimation at the receiver. At this point, stragglers account for the overall model aggregation error, since deep fading causes a large channel equalization factor, which will amplify the effect of noise at the receiver. Meanwhile, due to the increase in sampling resolution, the ML estimation can recover a separate transmission symbol sequence for each edge device, breaking the aligned amplitude requirement, compared to the aligned OAC. That means the design of the transmitter equalization factor can be flexible and can be further optimized appropriately. Due to the decoupling among device transmits powers, selecting a set of devices that do not contain stragglers is not so necessary. In general, compared with the single-antenna scenario considered in [36], multiple antennas with reasonable transceiver design can overcome unfavorable channel conditions and non-uniform channel fading. To take advantage of the decoupling of transmit power, we consider a transmitter pre-equalization and receiver combining design in multi-antenna scenarios to address this issue and propose a novel framework for misaligned OAC FL. After convergence analysis, a joint pre-equalization and receiver combining design scheme is proposed.

B. CONTRIBUTIONS

We study a misaligned FL system consisting of multiple edge devices and one PS. In each training round, among randomly selected devices, the stragglers account for the majority of aggregation errors. To relieve this issue, and improve the learning performance of FL, we propose a novel framework of misaligned OAC FL for accurate model aggregation on wireless networks. This is achieved by jointly pre-equalization and receiver combining design. To the best of our knowledge, this is the first work to consider the straggler problem in misaligned OAC and conduct a comprehensive study of the learning performance of FL algorithms with misaligned OAC. The contributions of this paper are summarized as follows:

- We study and propose a novel framework for misaligned OAC FL that contains multiple edge devices coordinated by a PS. In the model aggregation stage, the PS recovers the superimposed but symbol-timing misaligned transmitted symbol sequence from the edge devices via ML estimation. We derive a closed-form expression of the model aggregation error for the misaligned OAC, then derive an upper bound on the training loss between the training model and the global optimal value, and give a sufficient condition for the convergence of FL.
- Based on the observation of the problem in misaligned OAC model aggregation and obtained theoretical results, we formulate an optimization problem to minimize the distortion of the aggregation which is measured by MSE, w.r.t. the transmitter equalization factor and receiver combining vector. However, this is a non-convex quadratic constrained quadratic programming (QCQP) problem.
- To address this problem, we propose an approach based on the successive convex approximation (SCA) to deal with the non-convexity of quadratic constraints. Specifically, we first perform a linear approximation of the non-convex constraints, and then transform the prime problem into a series of convex problems, thus proposing the SCA-based optimization algorithm.

We conducted extensive experimental research. Experiment results show that the proposed algorithm achieves substantial learning performance improvement compared to existing baseline schemes and achieves the near-optimal performance of the ideal benchmark with aligned and noiseless aggregation.

C. ORGANIZATION

The remainder of this paper is organized as follows. In Section II, we introduce the FL model, the misaligned OAC communication model, WMFS Scheme, and ML estimation. In Section III, we analyze how aggregation error affects the FL learning performance with misaligned OAC. In Section IV we formulate the pre-equalization and receiver combining design problem that minimizes the MSE and proposes an SCA-based optimization algorithm to deal with the non-convexity of quadratic constraints. In Section V we conduct extensive experiments to evaluate the proposed algorithm. Conclusions are drawn in Section VI.

D. NOTATIONS

Throughout this paper, vectors, and matrices are denoted by boldface lowercase letters (e.g., $\boldsymbol{s}$) and boldface uppercase letters (e.g., $\boldsymbol{A}$), respectively. let $\boldsymbol{A}^{-1}$ denote inverse of a matrix $\boldsymbol{A}$. $\mathbb{R}$ and $\mathbb{C}$ represent the sets of real and complex values, respectively. The operator $\mathfrak{R}(\cdot)$, $\mathfrak{I}(\cdot)$, $\|\cdot\|_p$, $(\cdot)^T$, $(\cdot)^H$ stand for the real part and imaginary part of a complex number, the ℓ_p norm, the transpose, and complex conjugate transpose, respectively. $\mathcal{CN}(\mu, \sigma^2)$ represents the circularly symmetric complex Gaussian random distribution with mean

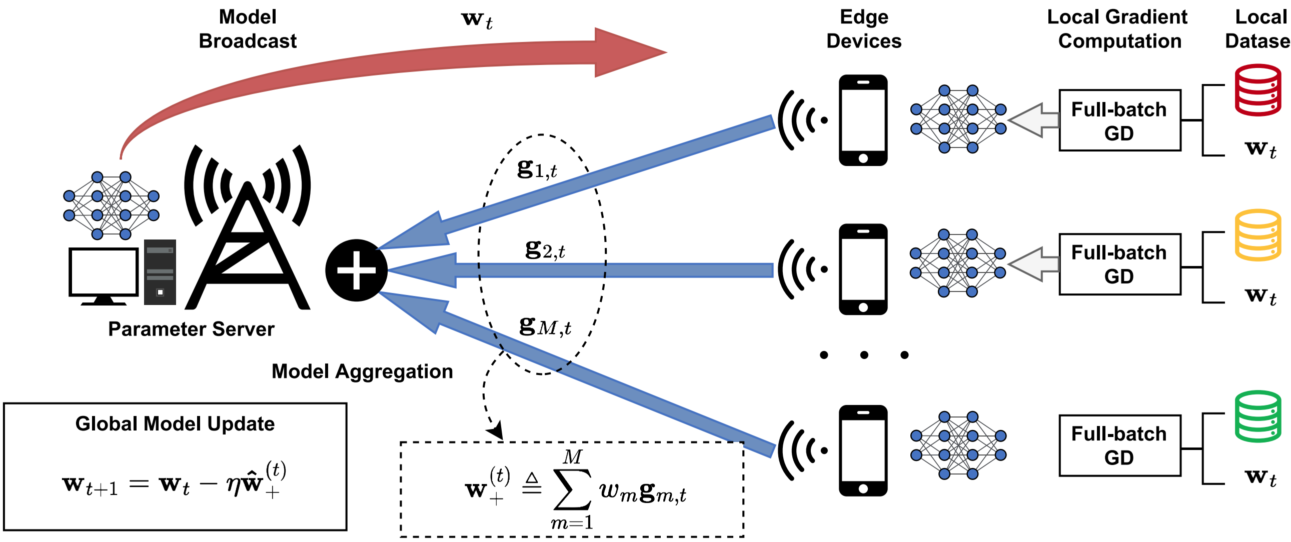


FIGURE 1. System model of FL.

μ and variance σ^2. diag$(\boldsymbol{x})$ denotes a diagonal matrix with the diagonal entries specified by $\boldsymbol{x}$, $\mathbb{E}(\cdot)$ denotes the expectation operator, and sgn$(\cdot)$ denotes the sign function.

II. SYSTEM MODEL

A. FEDERATED LEARNING MODEL

We consider a general FL system comprising of M' edge devices coordinated by a PS, equipped with N antennas, as shown in Fig. 1. The training process of FL aims to collaboratively learn a shared model, and the local data of each device is not exchanged. We denote the $\mathcal{D}_m$ as the local dataset collected at the m-th edge device. For each device, the learning objective is to solve the following optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^{D\times 1}} F_m(\mathbf{w}) = \frac{1}{|\mathcal{D}_m|}\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_m} f(\mathbf{w};\mathbf{x}_i,y_i), \tag{1}$$

where $\mathbf{w}$ is the D-dimensional model parameter vector; $|\mathcal{D}_m|$ denotes the cardinality of $|\mathcal{D}_m|$, and $\mathcal{D} = \bigcup_{m=1}^{M'}\{\mathcal{D}_m\}$ is the global dataset with $|\mathcal{D}| = \sum_{m=1}^{M'}|\mathcal{D}_m|$; $F_m(\mathbf{w})$ is the *local loss function* of m-th device on $\mathcal{D}_m$; $f(\mathbf{w};\mathbf{x}_i,y_i)$ is the sample-wise loss function indicating the prediction error on example $(\mathbf{x}_i,y_i)$ with model parameters $\mathbf{w}$. Then, we calculate the weighted average of all local loss functions in (1) according to the data volume of each device. Hence, the *global loss function* can be represented as

$$F(\mathbf{w}) = w_m\sum_{m=1}^{M'} F_m(\mathbf{w}), \tag{2}$$

where $w_m = \frac{|\mathcal{D}_m|}{|\mathcal{D}|}$ is the weight of m-th edge device, with $\sum_{m=1}^{M'} w_m = 1$. FL aims at finding an optimal model parameter $\mathbf{w}^*$ to minimize $F(\mathbf{w})$, on distributed devices, i.e.,

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

FL is usually performed on wireless channels with repeated training and communication processes. Specifically, at the t-th round:

- *Model broadcast*: The PS first randomly selects M active edge devices from M' edge devices, i.e., the subset of all devices, denoted as $\mathcal{M}$, to participate in the learning process, and broadcasts the current global model $\mathbf{w}_t$ to them.
- *Local gradient computation*: Each active edge device $m \in \mathcal{M}$ computes its local gradient based on its local dataset and received global model $\mathbf{w}_t$. Specifically, the m-th device's local gradient is given by

$$\mathbf{g}_{m,t} \triangleq \nabla F_m(\mathbf{w}_t), \tag{4}$$

  where $\nabla F_m(\mathbf{w}_t)$ is the gradient of $F_m(\cdot)$ at $\mathbf{w} = \mathbf{w}_t$.
- *Model aggregation*: Each active edge device m uploads its local gradient to the PS through wireless channels. Based on the received signals, the PS intends to compute the weighted average arithmetic sum of the local gradients $\mathbf{w}_+^{(t)} \triangleq \sum_{m=1}^{M} w_m\mathbf{g}_{m,t}$. Due to the existence of estimation error caused by channel fading and communication noise, the received signal is actually the estimation of $\mathbf{w}_+^{(t)}$, denoted by $\hat{\mathbf{w}}_+^{(t)}$, and $\hat{\mathbf{w}}_+^{(t)} \neq \hat{\mathbf{w}}_+^{(t)}$.
- *Global model update*: The PS aggregates these transmitted gradients to obtain a new global model $\mathbf{w}_{t+1}$ by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\hat{\mathbf{w}}_+^{(t)}. \tag{5}$$

OAC has emerged as a promising method for computing a nomographic function (e.g., arithmetic sum) which can improve communication efficiency, and reduce the required communication bandwidth [23].

In this paper, we focus on misaligned OAC to realize accurate analog aggregation of uplink parameters with low aggregation error.

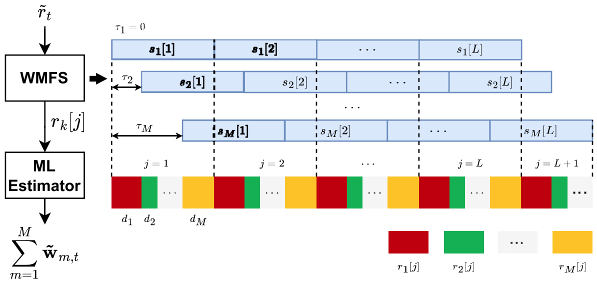


FIGURE 2. In the WMFS scheme, the received signal after matched filtered is oversampled to obtain the symbol sequence $r_k[j]$.

B. MISALIGNED OVER-THE-AIR AGGREGATION

We consider implementing the FedAvg algorithm in a misaligned OAC scenario. The target for aggregating local updates in FedAvg is to estimate $\mathbf{w}_+$. To simplify the notation, we omit the index t and let $\tilde{\mathbf{w}}_m = w_m\mathbf{g}_{m,t}$.

The transmission symbol sequences are zero-centered and transmitted in packets for a total of N_s. Specifically, for i-th packet we have $\boldsymbol{s}_m^{(i)} = [s_m^{(i)}[1], s_m^{(i)}[2], \dots, s_m^{(i)}[L]]^T \in \mathbb{C}^{L\times 1}$. In addition, the m-th active edge device set the transmit sequence $\{x_m[\ell] : 1 \le \ell \le L\}$ as

$$x_m[\ell] = b_m s_m[\ell], \quad \mathbb{E}\Big[|x_m[\ell]|^2\Big] \le P_0, \quad \forall m, \tag{6}$$

where $b_m \in \mathbb{C}$ is the transmit equalization factors; $P_0 > 0$ is the maximum transmit power.

In misaligned OAC scenarios, the transmission of each edge device has a certain transmission timing delay less than a symbol duration T denoted by $\tau_m, \forall m \in \{1, 2, \dots, M\}$, with $\tau_1 = 0 < \tau_2 < \dots < \tau_M < T$, without loss of generality. The equivalent received signal in the time domain after combining computed at the PS is given as

$$\tilde{r}(t) = \sum_{m=1}^{M} \boldsymbol{f}^H\boldsymbol{h}_m b_m \sum_{\ell=1}^{L} s_m[\ell]p(t-\tau_m-\ell T) + \boldsymbol{f}^H\tilde{\boldsymbol{n}}(t), \tag{7}$$

where $\boldsymbol{h}_m \in \mathbb{C}^{N\times 1}$ is the channel vector between active edge device m and the PS; $p(t) = 1/2[\operatorname{sgn}(t+T) - \operatorname{sgn}(t)]$ is a rectangular pulse of duration T; $\tilde{\boldsymbol{n}}(t)$ is AWGN whose one-sided power spectral density of each entry is N_0, and the noise gains are assumed to be independent and identically distributed (i.i.d.) across different PS antennas; $\boldsymbol{f} \in \mathbb{C}^{N\times 1}$ is the normalized receiver combining vector with $\|\boldsymbol{f}\|_2^2 = 1$.

To recover the gradient $\mathbf{w}_+$ from the superimposed but symbol-timing misaligned transmitted signal, we adopt the *whitening matched filtering and sampling* (WMFS) scheme and the *maximum likelihood* (ML) estimation introduced in [36].

C. WMFS SCHEME AND ML ESTIMATION

Based on the WMFS which is depicted in Fig. 2, we obtain a series of independent sample sequences $r_k[j]$ of length $(M(L+1)-1)$ via filtering by a series of matched filters and oversampling, given by

$$r_k[j] \triangleq \sum_{m=1}^{M} \boldsymbol{f}^H\boldsymbol{h}_m b_m s_m\big[j - \mathbf{1}_{m>k}\big] + \boldsymbol{f}^H\tilde{\boldsymbol{n}}_k[j], \tag{8}$$

with $\forall k \in \{1, 2, \dots, M\}$, where $\mathbf{1}$ is the indicator function; the length of the k-th matched filter is $d_k = \tau_{k+1} - \tau_k$, and $d_M = T - \tau_M$; $\tilde{\boldsymbol{n}}_k[j] \in \mathbb{C}^{N\times 1}$ is the sample noise vector introduced by multiple antennas, with each entry an i.i.d. random variable following the distribution of $\mathcal{CN}(0, N_0/d_k)$, and $\tilde{n}_k^l[j]$ is independent for different k and i [36].

Furthermore, we can vectorize the above process as

$$\boldsymbol{r} = \boldsymbol{K}\boldsymbol{s} + \boldsymbol{n}, \tag{9}$$

where $\boldsymbol{r}$ is $(M(L+1)-1)$-dimensional sample vector, and $\boldsymbol{s}$ is ML-dimensional signal vector. the coefficient matrix $\boldsymbol{K}$ is $M(L+1)-1$ by ML, giving

$$\boldsymbol{K} = \begin{bmatrix}
K_1 & & & & & & & & \\
K_1 & K_2 & & & & & & & \\
\cdots & K_2 & \dots & & & & & & \\
K_1 & \dots & \dots & K_M & & & & & \\
 & K_2 & \dots & K_M & K_1 & & & & \\
 & & \dots & \dots & K_1 & K_2 & & & \\
 & & & K_M & \dots & K_2 & \dots & & \\
 & & & & K_1 & \dots & \dots & K_M & \\
 & & & & & K_2 & \dots & K_M & \dots \\
 & & & & & & \dots & \dots & \dots \\
 & & & & & & & K_M & \dots \\
 & & & & & & & & \dots
\end{bmatrix},$$

where $K_m = \boldsymbol{f}^H\boldsymbol{h}_m b_m, \forall m$; The vector $\boldsymbol{n} = (\boldsymbol{f}^H\boldsymbol{N})^H$ is $M(L+1)-1$ by 1 dimension and from Appendix B, we can see that the covariance matrix of $\boldsymbol{n}$ is a diagonal matrix denoted by $\boldsymbol{\Sigma}$.

In one packet, we focus on estimating sum symbol vector $\boldsymbol{s}_+ \in \mathbb{C}^{L\times 1}$ which is given by $\boldsymbol{s}_+ = \boldsymbol{V}\boldsymbol{s} + \bar{s}$, where $\bar{s} \triangleq \sum_{m=1}^{M} \bar{s}_m$; $\boldsymbol{V}$ is a known matrix used to superpose the recovered signals $\boldsymbol{s}$. Accordingly, we can estimate $\boldsymbol{s}_+$ from $\boldsymbol{r}$ via pre-multiplying both sides of the Eq. (9) by $\boldsymbol{U} = \boldsymbol{V}(\boldsymbol{K}^H\boldsymbol{\Sigma}^{-1}\boldsymbol{K})^{-1}\boldsymbol{K}^H\boldsymbol{\Sigma}^{-1}$. Then, we obtain the ML estimation of $\boldsymbol{s}_+$,[1] defined as

$$\hat{\boldsymbol{s}}_+ \triangleq \boldsymbol{s}_+ + \boldsymbol{V}\Big(\boldsymbol{K}^H\boldsymbol{\Sigma}^{-1}\boldsymbol{K}\Big)^{-1}\boldsymbol{K}^H\boldsymbol{\Sigma}^{-1}\boldsymbol{n} + \bar{s}. \tag{10}$$

After collecting $\hat{\boldsymbol{s}}_+$ in each packet, for N_s times, during one communication round, we first restore $\hat{\mathbf{w}}_+$ with is the estimate of $\mathbf{w}_+$, and then update the global model by (5). Note that, as Eq. (10), the presence of communication noise leads to inevitable estimation errors in $\hat{\boldsymbol{s}}_+$. As a result, the weighted average arithmetic sum of the local gradients in (5) becomes inaccurate, thereby affecting the convergence of FL. In the next section, we quantitatively describe the impact of communication errors on the convergence of FL.

[1] Note that we can get the estimated value of the signal vector $\boldsymbol{s}$ by multiplying the matrix $(\boldsymbol{K}^H\boldsymbol{\Sigma}^{-1}\boldsymbol{K})^{-1}\boldsymbol{K}^H\boldsymbol{\Sigma}^{-1}$ at both sides of (9) to recover a separate transmission symbol sequence for each edge device.

III. LEARNING PERFORMANCE ANALYSIS

In this section, we analyze how aggregation error affects the FL learning performance with misaligned OAC, and derive a closed-form expression of the model aggregation error. Finally, we obtain a sufficient condition for FL convergence.

Due to the estimation error of gradients, the global model update recursion is given by

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \eta \hat{\mathbf{w}}_+^{(t)} = \mathbf{w}_t + \eta(\nabla F(\mathbf{w}_t) - \mathbf{e}_t), \tag{11}$$

where $\nabla F(\mathbf{w}_t) \triangleq \frac{1}{|\mathcal{D}|}\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}} \nabla f(\mathbf{w}_t;\mathbf{x}_i,y_i)$ is the gradient of $F(\mathbf{w})$ at $\mathbf{w} = \mathbf{w}_t$; and $\mathbf{e}_t$ denotes the gradient error vector caused by device selection and model aggregation at t-th communication round, given by

$$\mathbf{e}_t = \nabla F(\mathbf{w}_t) - \hat{\mathbf{w}}_+^{(t)}. \tag{12}$$

In addition, $\mathbf{e}_t$ can be further divided into: device selection error $\mathbf{e}_{1,t} = \nabla F(\mathbf{w}_t) - \mathbf{w}_+^{(t)}$ and model aggregation error $\mathbf{e}_{2,t} = \mathbf{w}_+^{(t)} - \hat{\mathbf{w}}_+^{(t)}$.

To facilitate the analysis, we first make the following assumptions on the loss function $F(\cdot)$:

Assumption 1: F is strongly convex with positive parameter μ, such that for all $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^{D\times 1}$

$$F(\mathbf{w}) \geq F(\mathbf{w}') + (\mathbf{w} - \mathbf{w}')^T \nabla F(\mathbf{w}') + \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2. \tag{13}$$

Assumption 2: The gradient $\nabla F(\cdot)$ of $F(\cdot)$ is Lipschitz continuous with parameter ω, i.e., we have:

$$\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\|_2 \leq \omega\|\mathbf{w} - \mathbf{w}'\|_2. \tag{14}$$

Assumption 3: $F(\cdot)$ is twice-continuously differentiable.

Assumption 4: For any training sample, $\nabla f(\cdot)$ has an upper bound.

$$\|\nabla f(\mathbf{w};\mathbf{x}_i,y_i)\|_2^2 \leq \alpha_1 + \alpha_2\|\nabla F(\mathbf{w})\|_2^2, \quad \forall i. \tag{15}$$

where α_1 and α_2 are non-negative constants.

The above assumptions are common in the random optimization literature, e.g., [18], [20], [41], [46], and can be satisfied by several widely used loss functions such as mean squared error, logistic regression, etc. Although the objective function of some models (such as neural networks) may not be strongly convex, our subsequent experimental results will show that the model is still convergent.

To begin with, we will introduce the following lemma.

Lemma 1: At each communication round t of FL, with $\eta = 1/\omega$, we have

$$\mathbb{E}[F(\mathbf{w}_{t+1})] \leq \mathbb{E}[F(\mathbf{w}_t)] - \frac{1}{2\omega}\|\nabla F(\mathbf{w}_t)\|_2^2 + \frac{1}{2\omega}\mathbb{E}[\|\mathbf{e}_t\|_2^2], \tag{16}$$

where ω is defined in Assumption 2.

Proof: See Appendix A.

From lemma 1, we can further obtain the upper limit of the expectation of the difference between the training loss and the optimal loss. For $\mathbf{e}_t$, one of the tractable expressions of $\mathbb{E}[\|\mathbf{e}_{2,t}\|_2^2]$ is given in the following lemma.

Lemma 2: For any given $\{\boldsymbol{f}, b_m, \forall m\}$, we have

$$\mathbb{E}[\|\mathbf{e}_{2,t}\|_2^2] = N_0\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x}, \tag{17}$$

where N_0 is one-sided power spectral density of additive noise; $\boldsymbol{A}$ is a known $M\times M$ hermitian matrix defined in (37), and $\boldsymbol{x} \in \mathbb{C}^{M\times 1}$ is given by

$$\boldsymbol{x} = \left[(\boldsymbol{f}^H\boldsymbol{h}_1 b_1)^{-1}, (\boldsymbol{f}^H\boldsymbol{h}_2 b_2)^{-1}, \ldots, (\boldsymbol{f}^H\boldsymbol{h}_m b_m)^{-1}\right]^T, \tag{18}$$

Proof: See Appendix C.

Moreover, $\mathbb{E}[\|\mathbf{e}_{2,t}\|_2^2]$ also denotes the MSE between the weighted average arithmetic sum of the local gradients $\mathbf{w}_+^{(t)}$ and its estimation $\hat{\mathbf{w}}_+^{(t)}$ multiplied by the model dimension D, which is defined as

$$\mathrm{MSE}(\mathbf{w}_+^{(t)}, \hat{\mathbf{w}}_+^{(t)}) = \frac{1}{D}\mathbb{E}(\|\mathbf{w}_+^{(t)} - \hat{\mathbf{w}}_+^{(t)}\|_2^2), \tag{19}$$

Lemma 2 gives the tractable expression of $\mathbb{E}[\|\mathbf{e}_{2,t}\|_2^2]$, thus we can derive an upper bound on $\mathbb{E}[F(\mathbf{w}_{t+1}) - F(\mathbf{w}^\star)]$ in the following theorem.

Theorem 1: With above assumptions as well as suitable conditions given by $\frac{1}{|b_m|^2} \leq \max_m v_m^2$ and $\eta = \frac{1}{\omega}$, for any given $\{\boldsymbol{f}, b_m, \forall m\}$ and optimal global FL model $\mathbf{w}^\star$, at t-th round, we have

$$\mathbb{E}[F(\mathbf{w}_{t+1}) - F(\mathbf{w}^\star)] \leq \Psi^t\mathbb{E}(F(\mathbf{w}_0) - F(\mathbf{w}^\star)) + \frac{\alpha_1}{\omega}\Phi\frac{1-\Psi^t}{1-\Psi} \tag{20}$$

where v_m^2 is the variance of the gradient $\tilde{\mathbf{w}}_{m,t}$ produced by the m-th active edge device, defined in (42); $\mathbf{w}_0$ is the initial global model; and the functions Φ and Ψ are defined as

$$\Phi = \frac{N_0\lambda_m M}{L}\max_m \frac{1}{|\boldsymbol{f}^H\boldsymbol{h}_m|^2} + 4\left(1 - \frac{\sum_{m=1}^M|\mathcal{D}_m|}{|\mathcal{D}|}\right)^2,$$

$$\Psi = 1 - \frac{\mu}{\omega} + \frac{2\mu\alpha_2}{\omega}\Phi, \tag{21}$$

where λ_m is the largest eigenvalue of matrix $\boldsymbol{A}$.

Proof: See Appendix D.

In addition, $\frac{1}{|b_m|^2} \leq \max_m v_m^2$ and $\eta = \frac{1}{\omega}$ are preconditions for Theorem 1. The former means that P_0 must be large enough to make the $|b_m|^2, \forall m$ large enough that its reciprocal is smaller than the maximum value of v_m^2, under the constraints of power constraint $v_m^2|b_m|^2 \leq P_0$. The latter shows that a proper learning rate η is important.

On the basis of Theorem 1, the following corollary is derived to guarantee the convergence of the FL algorithm.

Corollary 1: Satisfying the assumptions and conditions in Theorem 1 and $\Phi < \frac{1}{2\alpha_2}$, as $t \to \infty$, we have

$$\lim_{t\to\infty}\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^\star)] \leq \frac{\alpha_1\Phi}{\mu(1-2\alpha_2\Phi)}. \tag{22}$$

Proof: When $\Phi < \frac{1}{2\alpha_2}$, we have $\Psi < 1$. From Theorem 1, we see that when $\Psi < 1$, $\lim_{t\to\infty}\Psi^t = 0$. Hence, we take the limit on both ends of (20) and complete the proof.

From the above corollary, we can see that under certain conditions, there is an upper bound for the expectation of the gap between $F(\mathbf{w}_\infty)$ and $F(\mathbf{w}^\star)$ when the training rounds tend to infinity. In addition, Φ is an enlarged version of $\mathbb{E}[\|\mathbf{e}_t\|_2^2]$ from Eq. (39) and (43). Combining condition $\Phi < \frac{1}{2\alpha_2}$, we can obtain a sufficient condition for FL convergence which is given by

$$\boldsymbol{x}^H \boldsymbol{A}\boldsymbol{x} < \frac{1}{N_0}\left(\frac{\alpha_1}{\alpha_2} + \|\nabla F(\mathbf{w}_t)\|_2^2\right) \times \left[\frac{1}{2} - 4\alpha_2\left(1 - \frac{\sum_{m=1}^{M}|\mathcal{D}_m|}{|\mathcal{D}|}\right)^2\right], \tag{23}$$

and other conditions in Theorem 1. See Appendix E.

Therefore, given an arbitrary N_0 and device selection method, we can make $\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x}$ as small as possible to make FL more prone to convergence. The MSE that drops in tandem with it also improves the learning performance of FL.

IV. JOINT OPTIMIZATION OF PRE-EQUALIZATION AND RECEIVER COMBINER

In this section, we formulate the pre-equalization and receiver combining design problem as an aggregation distortion minimization task and propose an SCA-based algorithm to deal with the non-convexity of quadratic constraints.

A. PROBLEM FORMULATION

We previously proved the convergence of FL with misaligned OAC under sufficient conditions. Based on the above theoretical analysis, we formulate the pre-equalization and receiver combining design problem as the following minimization problem:

$$\begin{aligned} \underset{b_m\in\mathbb{C},\boldsymbol{f}\in\mathbb{C}^{N\times 1}}{\text{minimize}} \quad & \boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x} \\ \text{subject to} \quad & v_m^2|b_m|^2 \le P_0, \quad \forall m, && \text{(24a)} \\ & \|\boldsymbol{f}\|_2^2 = 1, && \text{(24b)} \\ & \frac{1}{|b_m|^2} \le \max_m v_m^2, && \text{(24c)} \\ & x_m = (\boldsymbol{f}^H h_m b_m)^{-1}, && \text{(24d)} \end{aligned}$$

where x_m is the m-th entry of $\boldsymbol{x}$ and $\boldsymbol{A}$ is a known $M \times M$ hermitian matrix defined in (37); v_m^2 is the variance of m-th local gradients; constraint (24c) comes from Theorem 1.

From the definition of $\boldsymbol{x}$ in (18), we can see that $|b_m|^2 = 1/|\boldsymbol{x_m}|^2|\boldsymbol{f}^H\boldsymbol{h}_m|$. Substituting $|b_m|^2$ into the constraints (24a), (24c) and treating $\boldsymbol{x}$ as the optimization variable instead of b_m, we have

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{C}^{M\times 1},\boldsymbol{f}\in\mathbb{C}^{N\times 1}}{\text{minimize}} \quad & \boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x} \\ \text{subject to} \quad & \left|x_m\boldsymbol{f}^H\boldsymbol{h}_m\right|^2 \ge \frac{1}{P_{0,m}}, \quad \forall m, && \text{(25a)} \\ & \left|x_m\boldsymbol{f}^H\boldsymbol{h}_m\right|^2 \le \max_m v_m^2, \quad \forall m, && \text{(25b)} \\ & \|\boldsymbol{f}\|_2^2 = 1, && \text{(25c)} \end{aligned}$$

where $P_{0,m} = \frac{P_0}{v_m^2}$; $\max_m v_m^2$ is a constant. Unfortunately, the problem written in (25) is a QCQP problem with non-convex constraints (25a), (25b) and (25c), which is highly intractable.

Remark 1: Note that $\boldsymbol{A}$ is a given matrix, which is related to the preset parameters M, N, τ_m and the initial receiver combining vector $\boldsymbol{f}_0$. Without loss of generality, we randomly initialize normalized $\boldsymbol{f}_0$ to ensure that $\|\boldsymbol{f}_0\|_2^2 = 1$ thus initializing $\boldsymbol{A}$. However, unlike other parameters, the receiver combining vector is the variable to optimize, not a fixed value, and it would be pointless to optimize this variable if $\boldsymbol{A}$ changes. In fact, we just need to guarantee that the optimal $\|\boldsymbol{f}^\star\|_2^2$ is equal to $\|\boldsymbol{f}_0\|_2^2$, because the effect of $\boldsymbol{f}_0$ on $\boldsymbol{A}$ is only that $\boldsymbol{A}$ contains the scaling factor $\|\boldsymbol{f}_0\|_2^2$.

Remark 2: Note that the vector $\boldsymbol{x}$ contains $1/\boldsymbol{f}$, and its effect is reflected as a scaling factor $1/\|\boldsymbol{f}\|_2^2$ in $\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x}$ if we ignore the weight between the entries inside $\boldsymbol{f}$. Moreover, when the optimal value $\boldsymbol{f}^\star$ is obtained, the actual scaling factor is $\|\boldsymbol{f}_0\|_2^2/\|\boldsymbol{f}^\star\|_2^2$. Consequently, the role $\|\boldsymbol{f}\|_2^2$ of Problem (25), is just to scale the magnitude of $\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x}$, and a bigger $\|\boldsymbol{f}\|_2^2$ leads to smaller objective value.[2]

Based on the above observations, we can easily extend the non-convex feasible region of $\boldsymbol{f}$ to a convex set $\|\boldsymbol{f}\|_2^2 \le 1$.[3] In the following part, we will provide one method to solve the non-convexity of another constraint (25a), (25b).

B. SCA-BASED OPTIMIZATION METHOD

SCA is a treatment method for solving non-convex optimization problems. Its basic idea is to transform problem (25) into a series of convex subproblems by linearizing the non-convex term in constraints. Motivated by [43], we introduce two auxiliary variables $\mathbf{c}_m = [\Re(x_m), \Im(x_m)]^T$ and $\mathbf{d}_m = [\Re(\boldsymbol{f}^H\boldsymbol{h}_m), \Im(\boldsymbol{f}^H\boldsymbol{h}_m)]^T$ to solve the nonconvex constraint. The problem (25) can be rewritten as

$$\begin{aligned} \underset{\mathbf{c}_m,\mathbf{d}_m,\forall m}{\text{minimize}} \quad & \boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x} \\ \text{subject to} \quad & \|\mathbf{c}_m\|^2\|\mathbf{d}_m\|^2 \ge \frac{1}{P_{0,m}}, \forall m, && \text{(26a)} \\ & \|\mathbf{c}_m\|^2\|\mathbf{d}_m\|^2 \le \max_m v_m^2, \forall m, && \text{(26b)} \\ & \mathbf{c}_m = [\Re(x_m), \Im(x_m)]^T, \forall m, && \text{(26c)} \\ & \mathbf{d}_m = [\Re(\boldsymbol{f}^H\boldsymbol{h}_m), \Im(\boldsymbol{f}^H\boldsymbol{h}_m)]^T, \forall m, && \text{(26d)} \\ & \|\boldsymbol{f}\|_2^2 \le 1. && \text{(26e)} \end{aligned}$$

To handle the non-convex constraints (26a) and (26b), we approximate them with iterative linear constraints, resulting in a series of convex problems. Afterward, the Problem (26) is converted into a second-order cone-programming (SOCP) problem, which has a lower computational complexity and

[2] Outside of Problem (25), $\|\boldsymbol{f}\|_2^2$ is independent of $\mathbb{E}[\|\mathbf{e}_t\|_2^2]$, because in this case the scaling factor in $\boldsymbol{A}$ always equals to $\|\boldsymbol{f}\|_2^2$, which can counteract the effect of $1/\boldsymbol{f}$ in $\boldsymbol{x}$.

[3] Actually, the weight between each entry in $\boldsymbol{f}$ is the true optimization variable, and a larger $\|\boldsymbol{f}\|_2^2$ is more likely to be obtained. As a result, $\|\boldsymbol{f}\|_2^2$ can always reach its upper bound 1.

Algorithm 1: SCA-Based Optimization Algorithm

```
Input: $\boldsymbol{A}$, $l = 0$, $\varepsilon$, $I_{\max}$;
  Randomly initialize $\boldsymbol{x}^{(0)}, \boldsymbol{f}_0$;
  Set $\mathbf{c}_m^{(0)} = \left[\Re(x_m^{(0)}), \Im(x_m^{(0)})\right]^T, \forall m$;
  Set $\mathbf{d}_m^{(0)} = [\Re(\boldsymbol{f}_0^H \boldsymbol{h}_m), \Im(\boldsymbol{f}_0^H \boldsymbol{h}_m)]^T, \forall m$;
  for $l = 1, 2, \ldots, I_{\max}$
    Solve Problem (27) to obtain $\boldsymbol{x}$ and $\boldsymbol{f}$;
    Update $\mathbf{c}_m^{(l)}$ and $\mathbf{d}_m^{(l)}$;
    if $\sum_m^M \left(\left\|\mathbf{c}_m^{(l+1)} - \mathbf{c}_m^{(l)}\right\| + \left\|\mathbf{d}_m^{(l+1)} - \mathbf{d}_m^{(l)}\right\|\right) \le \varepsilon$,
    early stop;
  end for
Output: $\boldsymbol{x}, \boldsymbol{f}$;
```

can be efficiently solved. Specifically, at l-th iteration, we solve the following convex problem:

$$\begin{aligned}
&\underset{\mathbf{c}_m, \mathbf{d}_m, \forall m}{\text{minimize}} \quad \boldsymbol{x}^H \boldsymbol{A} \boldsymbol{x} \\
&\text{subject to} \\
&\quad \left\|\mathbf{c}_m^{(l)}\right\|^2 \left\|\mathbf{d}_m^{(l)}\right\|^2 + 2\left\|\mathbf{d}_m^{(l)}\right\|^2 \left(\mathbf{c}_m^{(l)}\right)^T \left(\mathbf{c}_m - \mathbf{c}_m^{(l)}\right) \\
&\quad + 2\left\|\mathbf{c}_m^{(l)}\right\|^2 \left(\mathbf{d}_m^{(l)}\right)^T \left(\mathbf{d}_m - \mathbf{d}_m^{(l)}\right) \ge \frac{1}{P_{0,m}}, \forall m, && \text{(27a)} \\
&\quad \left\|\mathbf{c}_m^{(l)}\right\|^2 \left\|\mathbf{d}_m^{(l)}\right\|^2 + 2\left\|\mathbf{d}_m^{(l)}\right\|^2 \left(\mathbf{c}_m^{(l)}\right)^T \left(\mathbf{c}_m - \mathbf{c}_m^{(l)}\right) \\
&\quad + 2\left\|\mathbf{c}_m^{(l)}\right\|^2 \left(\mathbf{d}_m^{(l)}\right)^T \left(\mathbf{d}_m - \mathbf{d}_m^{(l)}\right) \le \max_m v_m^2, \forall m, && \text{(27b)} \\
&\quad \mathbf{c}_m = [\Re(x_m), \Im(x_m)]^T, \forall m, && \text{(27c)} \\
&\quad \mathbf{d}_m = [\Re(\boldsymbol{f}^H \boldsymbol{h}_m), \Im(\boldsymbol{f}^H \boldsymbol{h}_m)]^T, \forall m, && \text{(27d)} \\
&\quad \|\boldsymbol{f}\|_2^2 \le 1, && \text{(27e)}
\end{aligned}$$

where $\mathbf{c}_m^{(l)}$ and $\mathbf{d}_m^{(l)}$ are the optimized solutions at l-th iteration.

We initialize $\mathbf{c}_m^{(0)}$ and $\mathbf{d}_m^{(0)}$ in a random fashion and stop the iteration when the sum of distance between two consecutive iterations of $\mathbf{c}_m$ and $\mathbf{d}_m$ is less than a preset threshold ε, or when the maximum number of iterations $I_{\max}$ is reached. The corresponding algorithm is summarized in Algorithm 1.

Note that the proposed joint pre-equalization and receiver combining design with misaligned over-the-air computation relies on the perfect CSI and the variance of local gradients of m-th device. The proposed algorithm runs in a centralized manner at the PS. The variance of the gradient of each device is uploaded to the PS by a conventional method, such as orthogonal frequency-division multiple access after the channel estimation is completed. After optimization, the PS transmits the obtained transmitter pre-equalization factor b_m to each device respectively. Channel training for estimating CSI at the PS can be accomplished by transmitting pilot sequences from each device [47].

C. COMPUTATIONAL COMPLEXITY

Problem (27) is a SOCP problem. To solve it, the existing solvers usually apply the interior point method, and the worst-case computational complexity of each iteration is $\mathcal{O}((N+M)^3)$. Consequently, the complexity of Algorithm 1 is upper bounded by $\mathcal{O}(I_{\max}(N+M)^3)$. Considering the computational complexity, in practical systems, it is often necessary to limit the number of devices selected in each round of communication and determine a reasonable maximum number of iterations.

V. EXPERIMENTS AND DISCUSSIONS

In this section, we conduct comprehensive experiments to compare the proposed SCA-based algorithm with baseline schemes for federated learning with misaligned OAC to examine its effectiveness. Simulation codes are available at https://github.com/Forgethson/JointDesignMisAlignedFL.

A. SIMULATION SETUP

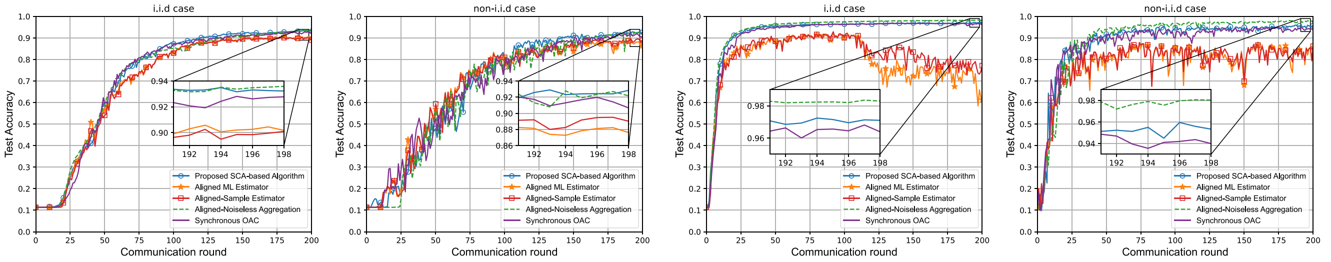
For our simulations, we consider a typical FL network. We randomly select M active devices in each iteration of FL training. We consider channel variation transmit timing variation under different packets and assume that the channel coefficients and transmission timing delays over a packet are fixed. The wireless channels from the edge devices to the PS follow i.i.d. Rayleigh channel model, i.e., $\boldsymbol{h}_m \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{I})$. For the transmission timing delays across M active devices, we set $\tau_1 = 0$, fix a maximum time offset τ_M and the rest follow the uniform distribution of $\mathcal{U}(0, \tau_M)$. The effect of noise is measured by a given EsN0, i.e., the expectation of average energy per symbol for all transmitters to noise power spectral density ratio, defined as

$$\text{EsN0} = \frac{\mathbb{E}_m\left[\mathbb{E}_i\left[|b_m s_m[i]|^2\right]\right]}{N_0}. \tag{28}$$

To verify the effectiveness of the proposed SCA-based Algorithm 1, we simulate the image classification task on MNIST and CIFAR-10 data sets via PyTorch. FL performance is evaluated by test accuracy, i.e., the number of correctly classified test images to the size of the test set ratio. For all active devices, at each model aggregation stage, the model update sequences are transmitted in multiple packets with length L. The PS recovers the weighted average arithmetic sum of the local gradients $\mathbf{w}_+$ from the superimposed but symbol-timing misaligned transmitted signal via WMFS and ML estimation. Unless otherwise specified, the values of simulation parameters are given in Table 1.

For performance comparison, we consider the following baselines:

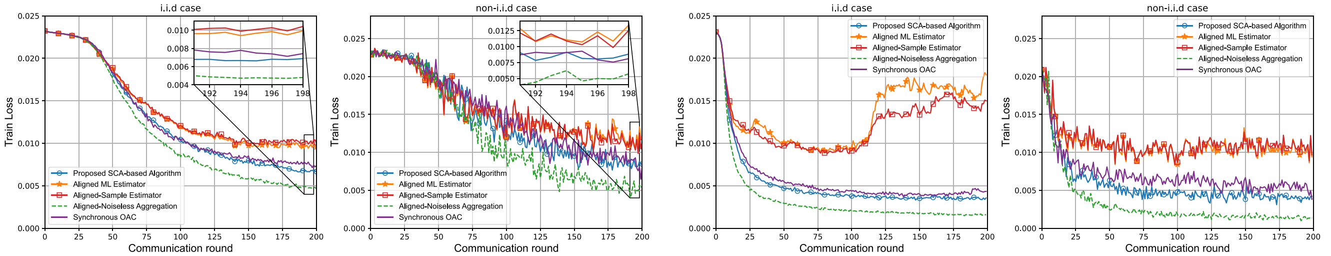
- *Aligned ML estimator:* The ML estimator is one of the baseline estimators proposed in [36]. Note that in the aligned ML estimator, the residual channel-fading coefficient $h_m' = b_m h_m$ at the receiver is set to 1 (both phase-aligned and amplitude-aligned case), which is equivalent to $b_m = 1/h_m$.



(a) full-batch gradient descent case

(b) mini-batch gradient descent case

FIGURE 3. Test accuracy versus communication round of proposed SCA-based scheme and baseline schemes.



(a) full-batch gradient descent case

(b) mini-batch gradient descent case

FIGURE 4. Train loss versus communication round of proposed SCA-based scheme and baseline schemes.

TABLE 1. Simulation parameters.

Parameter	Description	Value
M'	Number of edge devices	40
M	Number of active edge devices	4
N	Number of PS antennas	5
η	Learning rate	0.01
P_0	Maximum transmit power	10 dB
L	Length of packet	260
B	Local batch-size	128
$I_{\max}$	Maximum SCA loops	100
ε	Threshold of SCA early stopping criteria	0.01
T	Sample period	1
τ_M	Maximum time offset	0.9
EsN0	The expectation of average energy per symbol to N_0 ratio	-20 dB

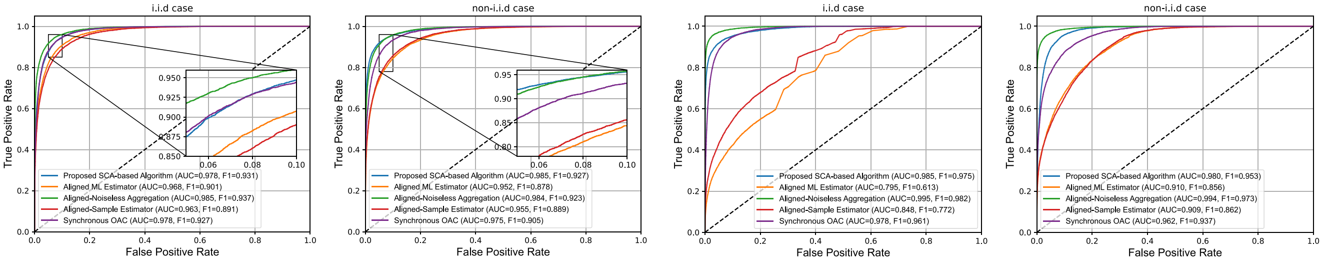
- *Aligned-sample estimator:* As another baseline estimator, the aligned-sample (AS) estimator directly uses the m-th matched filter to generate the result sequence, i.e., the sequence $\{r_M[1], r_M[2], \ldots, r_M[L]\}$. Due to the lack of ML estimation at the receiver, it cannot resist the influence of channel fading on the original signal. In order to achieve the AS estimator's ideal performance [36], we fix the channel constant with $h_m = 1 + 0_J, \forall m$. It can also be considered that the transmitter pre-equalization is accurate (for channel fading), that is $K_m = 1, \forall m$.
- *Aligned-noiseless aggregation:* Suppose that the channels are noiseless and transmission timing is synchronous, i.e., $\hat{\mathbf{w}}_+ = \mathbf{w}_+$, $\tau_M = 0$. In other words, as the ideal aggregation benchmark, the PS directly utilizes an undistorted gradient to implement FL.
- *Synchronous OAC aggregation:* This traditional OAC method based on channel-inversion only uses transmitter pre-equalization to handle channel fading, assuming that the transmission timing is synchronized.

B. SIMULATIONS ON MNIST DATASET

In this subsection, we examine the performance of the proposed algorithm on a handwritten digit identification task. The training and test sets of the MNIST dataset contain 60,000 examples and 10,000 examples respectively with a total of 10 classes, and each example is a 28×28 single-channel image. For dataset partition, we consider both the i.i.d. setting where the data is shuffled and uniformly distributed across all edge devices, and the non-i.i.d. setting where samples sorted by digit label are divided into 200 shards with a size of 300, and assign 5 shards to each of the 40 edge devices [7]. Subsequently, we train a convolutional neural network, which has 21,840 parameters and the loss function is cross-entropy loss. For the MNIST dataset, our simulations consist of the results of the full-batch gradient descent method and the results of the mini-batch gradient descent method. In terms of the mini-batch gradient descent method, the local dataset $\mathcal{D}$ is randomly partitioned into $\lceil|\mathcal{D}|/B\rceil$ batches $\{\mathcal{D}_m^{(1)}, D_m^{(2)} \ldots, \mathcal{D}_m^{(\lceil|\mathcal{D}|/B\rceil)}\}$ where B is batch size; $\lceil\cdot\rceil$ is the ceiling function. Each device needs $\lceil|\mathcal{D}|/B\rceil$ iterations for each batch dataset, thus the gradient of m-th device is $\mathbf{g}_{m,t} = (\mathbf{w}_{m,t}^{(\lceil|\mathcal{D}|/B\rceil)} - \mathbf{w}_t)/\eta$.

In Fig. 3 and Fig. 4, we plot the test accuracy and training loss versus the communication round with the full-batch gradient descent method and mini-batch gradient descent method with batch size $B = 128$ under two dataset partition settings on MNIST to compare the performance of all four schemes. In addition, the ROC curve is shown in Fig. 5, and the Area Under Curve (AUC) and F1-score of different schemes are given in the legend.

For the full-batch gradient descent case, we can see from Fig. 3(a), the proposed SCA-based scheme achieves



(a) full-batch gradient descent case

(b) mini-batch gradient descent case

FIGURE 5. **ROC curves of the proposed SCA-based scheme and the baseline scheme.**

an average test accuracy of 94.91% over the past 20 communication rounds, 82.59% in the aligned ML estimator scheme, 95.33% in the aligned noiseless aggregation scheme, 74.72% in the AS estimator scheme, and 93.91% in synchronous OAC scheme. A similar situation is shown from the loss curve in Fig. 4(a) and Fig. 5(a). The average training losses of the last 20 rounds of the above scheme are 0.56 (SCA), 1.52 (aligned ML), 0.36 (aligned noiseless), 1.87 (AS), and 0.66 (synchronous OAC). The AUC and F1 scores of the different schemes are given in the legend, and the results are similar to the loss curves. The above F1 scores are 0.931 (SCA), 0.901 (aligned ML), 0.937 (aligned noiseless), 0.891 (AS), and 0.927 (synchronous OAC). In the non-i.i.d. case, due to the heterogeneity of datasets across different devices, the model obtained by distributed training and averaged in a linear fashion differs greatly from the model obtained by centralized training, which degrades the learning performance of FL. The results for non-i.i.d. cases are similar to those for i.i.d. cases.

For the mini-batch gradient descent case, Fig. 3(b) demonstrates a similar result to the full-batch gradient case. The proposed SCA-based scheme achieves an average test accuracy of 96.95% over the past 20 communication rounds, 70.64% in the aligned ML estimator scheme, 98.22% in the aligned noiseless aggregation scheme, 76.57% in the AS estimator scheme, and 96.45% in synchronous OAC scheme. A similar situation is shown from the loss curve in Fig. 4(b) and Fig. 5(b). The average training losses of the last 20 rounds of the above scheme are 0.35 (SCA), 1.63 (aligned ML), 0.16 (aligned noiseless), 1.46 (AS), and 0.43 (synchronous OAC). The AUC and F1 scores of the different schemes are given in the legend, and the results are similar to the loss curves. The above F1 scores are 0.975 (SCA), 0.613 (aligned ML), 0.982 (aligned noiseless), 0.772 (AS), and 0.961 (synchronous OAC).

The above results demonstrate that the proposed SCA-based scheme outperforms two baseline schemes, is comparable to or even better than the synchronous OAC scheme, and achieves the near-optimal performance of the ideal aggregation benchmark. In addition, the original ML estimator scheme was unable to converge the FL training in the above experiments.

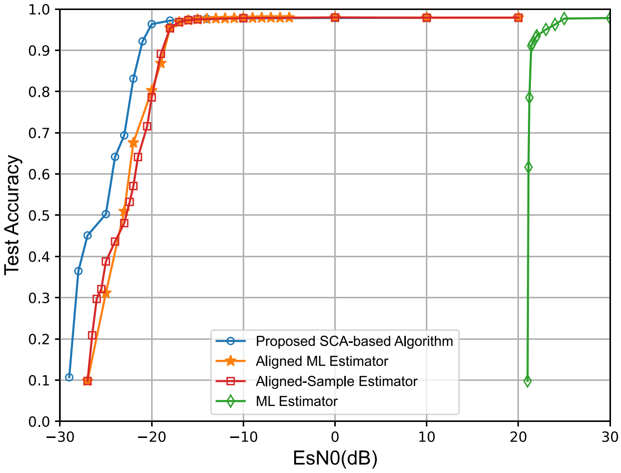


FIGURE 6. **Test accuracy versus EsN0 of different schemes under i.i.d setting.**

Fig. 6 shows that compared to the other baseline schemes, the proposed SCA-based scheme can achieve the same test accuracy at lower EsN0. ML estimator does not work at low EsN0 (less than 21dB). Furthermore, we calculate the MSE in one packet of the three schemes on 500 random channel realizations respectively with $M = 4$, $N = 10$, $N_0 = 1$ and let τ_M a random value within the range 0.5 to 0.9. For proposed scheme and ML scheme, the MSE equals $\frac{N_0}{L}\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x}$. Note that since there is no ML estimation in the AS scheme, its MSE equals N_0/d_M, where $d_M = T - \tau_M$ is the duration of the M-th matched filter. The proposed scheme has an average MES of 0.0038, which corresponds to 0.1154 for the ML estimator scheme and 0.0156 for the AS estimator scheme.

For different PS antenna quantities, Fig. 7 demonstrates the test accuracy discrepancy of FL under the above three different values of N with $M = 8$. It shows that a larger number of antennas can improve the performance of FL. Under different N, the results over 500 channel realizations with $M = 8$ are illustrated in Fig. 8. We first fixed N_0 and optimized 500 random channel samples under different $N \in \{1, 2, \ldots, 20\}$. Then we show the MES of the optimization results for $N = 6$, $N = 12$, and $N = 20$ in the form of a scatterplot and the average MSE curve for 500 samples under different N.

As can be seen from the above figures, increasing the number of PS antennas can lead to a better solution, i.e., a smaller MSE which is the reason for the higher test accuracy of FL. This is because, as the number of antennas increases,

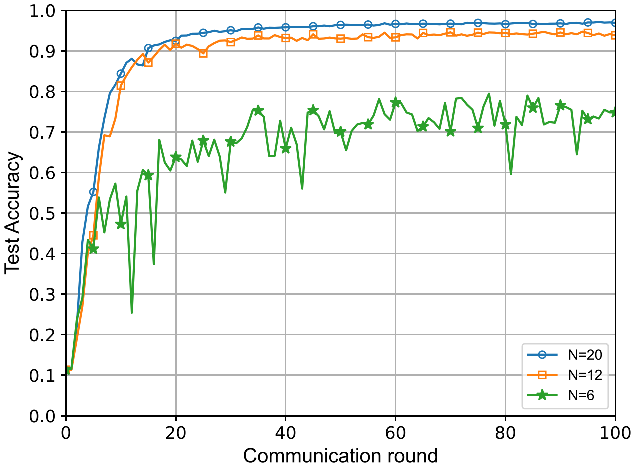


FIGURE 7. Test accuracy versus communication round of proposed SCA-based scheme over the number of PS antennas.

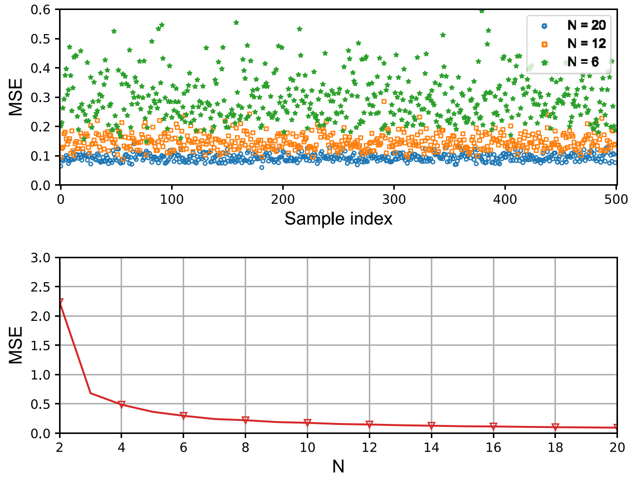
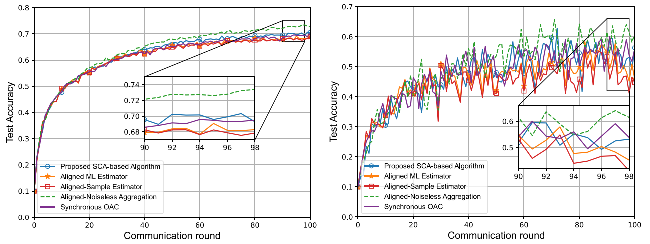


FIGURE 8. Effect of the number of PS antennas on optimization objective value with $M = 8$, $N_0 = 1$.



(a) i.i.d. case

(b) non-i.i.d. case

FIGURE 9. Test accuracy versus communication round of proposed SCA-based scheme and baseline schemes under two dataset partition settings on the CIFAR-10 dataset.

the dimensionality of the receiver combining vector $\boldsymbol{f}$ also increases, which helps the proposed SCA-based algorithm find better $\boldsymbol{f}$. In addition, a higher dimension $\boldsymbol{f}$ is also conducive to satisfying constraint (25a) and (25b), which in disguise expands the solution space of $\boldsymbol{x}$, ultimately helping the algorithm to find better solutions. Moreover, Fig. 8 also indicates that when the number of antennas exceeds 14, further increasing the number of antennas will not significantly decrease the MSE, but will further increase the deployment cost and power consumption. This phenomenon shows that there is a trade-off between the deployment cost and power consumption brought by the high antenna count, and the FL performance improvement brought by the low MSE.

C. SIMULATIONS ON CIFAR-10 DATASET

In this subsection, we examine the performance of the proposed algorithm on a more challenging task. For the CIFAR-10 dataset, we only consider the mini-batch gradient case with $B = 128$. The training and test sets of the CIFAR-10 dataset contain 50,000 examples and 10,000 examples respectively with a total of 10 classes, and each example is a 32×32 three-channel image. For dataset partition, we also consider the i.i.d. case (the same as the setting in Section V-B) and the non-i.i.d. case. For the non-i.i.d. in this case, the samples are first sorted by label, then divided into 400 shards of size 125, and 10 shards are assigned to each of the 40 edge devices. The learning rate η is set to 0.05 and the EsN0 is set to -8 dB.

Fig. 9 shows that the proposed algorithm still outperforms other baseline schemes, and achieves the near-optimal performance of ideal aggregation benchmark. For the i.i.d. case, in Fig. 9(a), the proposed SCA-based scheme achieves an average test accuracy of 69.68% over the past 20 communication rounds, 68.06% in the aligned ML estimator scheme, 72.35% in the aligned noiseless aggregation scheme, 67.82% in the AS estimator scheme, and 68.85% in synchronous OAC scheme. For the non-i.i.d. case, in Fig. 9(b), the proposed SCA-based scheme achieves an average test accuracy of 53.75% over the past 20 communication rounds, 51.97% in the aligned ML estimator scheme, 59.20% in the aligned noiseless aggregation scheme, 48.82% in the AS estimator scheme, and 54.88% in synchronous OAC scheme. For more difficult classification tasks, the effectiveness of the algorithm is confirmed.

D. FURTHER DISCUSSION

From the preceding theoretical analysis, we obtain a sufficient condition (23) for FL convergence, and we know that, for any transmission power, when EsN0 is large enough that the MSE cannot meet (23), the convergence of FL cannot be guaranteed. Fig. 6 shows that with the decrease of EsN0, the test accuracy of FL remains unchanged at first, and then decreases until the convergence fails. To analyze the theoretical results, we can see that when EsN0 is greater than a certain threshold (different for different schemes), the performance of FL reaches the equivalent performance under ideal conditions, i.e., aligned-noiseless aggregation scheme. In this case, since N_0 is small enough, FL can converge completely, and the influence of noise is negligible. As EsN0 decreases, for the same average transmission power, the N_0 increases accordingly. At this point, a relatively small $\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x}$ is required to ensure convergence condition, otherwise, the

MSE between the true value of the gradient $\mathbf{w}_+$ and the estimated value $\hat{\mathbf{w}}_+$ is so large that the FL training fails to converge.

Moreover, different schemes have different sensitivity to noise. Due to the heterogeneity of communication capabilities among edge devices, the stragglers account for the overall model aggregation error, since deep fading causes a significant channel equalization factor, which will amplify the effect of noise at the receiver. For the ML estimator scheme, without accurate transmitter pre-equalization, the entry of coefficient matrix in (9) is related to residual channel fading, i.e., $K_m = h'_m$. Thus, the deep channel fading will amplify the effect of noise at the receiver. This explains why the misaligned FL based on the original ML estimator cannot converge at low EsN0. The computational complexity of the ML estimator is $\mathcal{O}(L^2M^2\log(LM))$ caused by matrix inversion. As a special case of the ML estimator with the same computational complexity and $h'_m = 1$, the aligned ML estimator performs better than ML estimators at low EsN0. However, it does not take advantage of the flexible design potential of the transmitter coefficients to further improve noise mitigation. For the AS estimator scheme, the transmitter pre-equalization is assumed to be accurate, i.e., $K_m = 1$. Although the AS estimator has shown good performance and low computational complexity which can be regarded as $\mathcal{O}(L)$, it is highly dependent on the accurate pre-equalization of the transmitter. No further processing at the receiver makes it unable to cope with inaccurate pre-equalization and larger noise effects at lower EsN0. For the proposed SCA-based scheme, there is additional calculation complexity $\mathcal{O}(I_{\max}(N{+}M)^3)$ from Algorithm 1, but the entry of coefficient matrix $K_m = \boldsymbol{f}^H\boldsymbol{h}_m b_m$ can be flexibly designed to find the appropriate b_m and $\boldsymbol{f}$, to resist non-uniform fading and noise effects to a large extent, which shows that completely accurate pre-equalization on the transmitter is not necessarily optimal.

To sum up, compared with the other baseline schemes which suffer from significant aggregation errors caused by stragglers, the proposed SCA-based scheme overcomes the unfavorable channel conditions and non-uniform fading via channel pre-equalization and receiver combining design, while largely reducing the effect of noise on model aggregation. Finally improved the performance of FL.

VI. CONCLUSION

In this paper, we studied the misaligned OAC FL system design problem and implemented a novel framework of FL with misaligned OAC for accurate model aggregation on wireless networks. We first derive an upper bound on the training loss between the training model and the global optimal value and give a sufficient condition for the convergence of FL. Based on the observation of the problem in misaligned OAC model aggregation and obtained theoretical results, we formulate an optimization problem to minimize the distortion of the aggregation which is measured by MSE w.r.t. the transmitter equalization factor and receiver combining vector. Then we proposed an SCA-based optimization algorithm to solve the resulting quadratic constrained quadratic program. Finally, our extensive experiments show that the proposed algorithm has improved test accuracy compared to existing baseline schemes and is close to the ideal benchmark, which proves its effectiveness.

APPENDIX A
PROOF OF LEMMA 1

Assumption 2 is equivalent to

$$F(\mathbf{w}) \leq F(\mathbf{w}') + (\mathbf{w}-\mathbf{w}')^T\nabla F(\mathbf{w}') + \frac{\omega}{2}\|\mathbf{w}-\mathbf{w}'\|_2^2. \quad (29)$$

Replacing $\mathbf{w}$ and $\mathbf{w}'$ with $\mathbf{w}_{t+1}$ and $\mathbf{w}_t$, we have:

$$\begin{aligned} F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) &+ (\mathbf{w}_{t+1}-\mathbf{w}_t)^T\nabla F(\mathbf{w}_t) \\ &+ \frac{\omega}{2}\|\mathbf{w}_{t+1}-\mathbf{w_t}\|_2^2 \end{aligned} \quad (30)$$

Combining (30), (12), (5), and $\eta = 1/\omega$, we have

$$\begin{aligned} F(\mathbf{w}_{t+1}) &\leq F(\mathbf{w}_t) - \frac{1}{\omega}(\nabla F(\mathbf{w}_t)-\mathbf{e}_t)^T\nabla F(\mathbf{w}_t) \\ &\quad + \frac{1}{2\omega}\|\nabla F(\mathbf{w}_t)-\mathbf{e}_t\|_2^2 \\ &= F(\mathbf{w}_t) - \frac{1}{\omega}\|\nabla F(\mathbf{w}_t)\|_2^2 + \frac{1}{\omega}\nabla F(\mathbf{w}_t)^T\mathbf{e}_t \\ &\quad + \frac{1}{2\omega}\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{1}{\omega}\nabla F(\mathbf{w}_t)^T\mathbf{e}_t + \frac{1}{2\omega}\|\mathbf{e}_t\|_2^2 \\ &= F(\mathbf{w}_t) - \frac{1}{2\omega}\|\nabla F(\mathbf{w}_t)\|_2^2 + \frac{1}{2\omega}\|\mathbf{e}_t\|_2^2. \end{aligned} \quad (31)$$

By taking the expectation with respect to communication noise, we complete the proof.

APPENDIX B

The $\boldsymbol{N}$ is a $N \times M(L+1)-1$ matrix given by

$$\begin{aligned} \boldsymbol{N} = [&\tilde{\boldsymbol{n}}_1[1], \tilde{\boldsymbol{n}}_2[1], \ldots, \tilde{\boldsymbol{n}}_M[1], \tilde{\boldsymbol{n}}_1[2], \tilde{\boldsymbol{n}}_2[2], \ldots, \tilde{\boldsymbol{n}}_M[2], \ldots, \\ &\tilde{\boldsymbol{n}}_1[L], \tilde{\boldsymbol{n}}_2[L], \ldots, \tilde{\boldsymbol{n}}_M[L], \tilde{\boldsymbol{n}}_1[L+1], \tilde{\boldsymbol{n}}_2[L+1], \ldots, \\ &\tilde{\boldsymbol{n}}_{M-1}[L+1]]^T. \end{aligned}$$

The detailed form of $\boldsymbol{N}^H$ is given by

$$\left\{\begin{matrix} \tilde{n}_1^1[1] & \tilde{n}_1^2[1] & \ldots & \tilde{n}_1^N[1] \\ \tilde{n}_2^1[1] & \tilde{n}_2^2[1] & \ldots & \tilde{n}_2^N[1] \\ \ldots & \ldots & \ldots & \\ \tilde{n}_M^1[1] & \tilde{n}_M^2[1] & \ldots & \tilde{n}_M^N[1] \\ \tilde{n}_1^1[2] & \tilde{n}_1^2[2] & \ldots & \tilde{n}_1^N[2] \\ \tilde{n}_2^1[2] & \tilde{n}_2^2[2] & \ldots & \tilde{n}_2^N[2] \\ \ldots & \ldots & \ldots & \\ \tilde{n}_M^1[2] & \tilde{n}_M^2[2] & \ldots & \tilde{n}_M^N[2] \\ \ldots & \ldots & \ldots & \\ \tilde{n}_1^1[L] & n_1^2[L] & \ldots & n_1^N[L] \\ \tilde{n}_2^1[L] & \tilde{n}_2^2[L] & \ldots & \tilde{n}_2^N[L] \\ \ldots & \ldots & \ldots & \\ \tilde{n}_M^1[L] & \tilde{n}_M^2[L] & \ldots & \tilde{n}_M^N[L] \\ \tilde{n}_1^1[L+1] & \tilde{n}_1^2[L+1] & \ldots & \tilde{n}_1^N[L+1] \\ \tilde{n}_2^1[L+1] & \tilde{n}_2^2[L+1] & \ldots & \tilde{n}_2^N[L+1] \\ \ldots & \ldots & \ldots & \\ \tilde{n}_{M-1}^1[L+1] & \tilde{n}_{M-1}^2[L+1] & \ldots & \tilde{n}_{M-1}^N[L+1] \end{matrix}\right\}$$

For $\boldsymbol{N}^H$, each column is the noise sequence for the l-th antenna at the receiver denoted by $\hat{\boldsymbol{n}}_l \in \mathbb{C}^{M(L+1)-1}$ and each row is denoted by $\tilde{\boldsymbol{n}}_k^H[j]$.

We denote diagonal matrix $\bar{\boldsymbol{\Sigma}}^l$ is the covariance matrix of $\hat{\boldsymbol{n}}_l$, and for any given k and j, $\tilde{n}_k^l[j]$ is i.i.d., $\forall l \in \{1, 2, \ldots, N\}$ [36]. Moreover, $\bar{\boldsymbol{\Sigma}}^l = \hat{\boldsymbol{n}}_l \hat{\boldsymbol{n}}_l^H$ is a diagonal matrix, with $\bar{\boldsymbol{\Sigma}}^l \equiv \bar{\boldsymbol{\Sigma}}, \forall l$ across different PS antennas. And the $(M(L+1)-1) \times (M(L+1)-1)$ diagonal matrix $\bar{\boldsymbol{\Sigma}}$ can be constructed by

$$\bar{\boldsymbol{\Sigma}} = \operatorname{diag}\left(\left[\frac{N_0}{d_1}, \frac{N_0}{d_2}, \ldots, \frac{N_0}{d_M}, \frac{N_0}{d_1}, \frac{N_0}{d_2}, \ldots, \frac{N_0}{d_M} \ldots, \right.\right.$$
$$\left.\left. \frac{N_0}{d_1}, \frac{N_0}{d_2}, \ldots \frac{N_0}{d_{M-1}}\right]\right),$$

where N_0 is the one-sided power spectral density of additive noise.

The detailed form of $\boldsymbol{n} = (\boldsymbol{f}^H \boldsymbol{N})^H \in \mathbb{C}^{M(L+1)-1}$ is given by

$$\left[\sum_{l=1}^{N} \tilde{n}_1^l[1] f_l, \sum_{l=1}^{N} \tilde{n}_2^l[1] f_l, \ldots, \sum_{l=1}^{N} \tilde{n}_M^l[1] f_l, \sum_{l=1}^{N} \tilde{n}_1^l[2] f_l, \sum_{l=1}^{N} \tilde{n}_2^l[2] f_l, \right.$$
$$\ldots, \sum_{l=1}^{N} \tilde{n}_M^l[2] f_l, \ldots, \sum_{l=1}^{N} \tilde{n}_1^l[L] f_l, \sum_{l=1}^{N} \tilde{n}_2^l[L] f_l, \ldots, \sum_{l=1}^{N} \tilde{n}_M^l[L] f_l,$$
$$\left. \sum_{l=1}^{N} \tilde{n}_1^l[L+1] f_l, \sum_{l=1}^{N} \tilde{n}_2^l[L+1] f_l, \ldots, \sum_{l=1}^{N} \tilde{n}_{M-1}^l[L+1] f_l \right]^T,$$

where f_l is l-th entry of $\boldsymbol{f}$. Since the noise gains are assumed to be i.i.d. across different PS antennas, we have $\boldsymbol{\Sigma} = \boldsymbol{n}\boldsymbol{n}^H$ is the covariance matrix of $\boldsymbol{n}$ which can be represented by

$$\boldsymbol{\Sigma} = \|\boldsymbol{f}\|_2^2 \bar{\boldsymbol{\Sigma}} = N_0 \tilde{\boldsymbol{\Sigma}}, \tag{32}$$

where $\tilde{\boldsymbol{\Sigma}}$ is the auxiliary matrix without factor N_0 given by

$$\tilde{\boldsymbol{\Sigma}} = \operatorname{diag}\left(\|\boldsymbol{f}\|_2^2 \left[\frac{1}{d_1}, \frac{1}{d_2}, \ldots, \frac{1}{d_M}, \frac{1}{d_1}, \frac{1}{d_2}, \ldots, \frac{1}{d_M} \ldots, \right.\right.$$
$$\left.\left. \frac{1}{d_1}, \frac{1}{d_2}, \ldots \frac{1}{d_{M-1}}\right]\right).$$

APPENDIX C
PROOF OF LEMMA 2

Recall that we construct the transmitted symbols from local gradients and send these symbols in packets. Consequently, at one communication round, we first obtain N_s combined received signal vector $\hat{\boldsymbol{s}}_+ \in \mathbb{C}^{L\times 1}$ at PS. Next, we stack these vectors and obtain $\hat{\tilde{\boldsymbol{s}}}_+ \in \mathbb{C}^{\frac{D}{2}\times 1}$. Finally, we restore the true gradient estimate vector $\hat{\mathbf{w}}_+ \in \mathbb{R}^D$ from $\hat{\tilde{\boldsymbol{s}}}_+$.

From (12), we have $\mathbb{E}[\|\mathbf{e}_t\|_2^2] = \|\mathbf{w}_+^{(t)} - \hat{\mathbf{w}}_+^{(t)}\|_2^2$, which is equals to $\|\tilde{\boldsymbol{s}}_+^{(t)} - \hat{\tilde{\boldsymbol{s}}}_+^{(t)}\|_2^2$, as well as, $\sum_{i=1}^{N_s} \|\boldsymbol{s}_+^{(i)} - \hat{\boldsymbol{s}}_+^{(i)}\|_2^2$ at t-th round. Note that the value of $\mathbb{E}[\|\mathbf{e}_t\|_2^2]$ has nothing to do with whether the symbol is packed or not. For convenience and without loss of generality, we let $L = D$, thus Eq. (12) can be rewritten as

$$\begin{aligned} \mathbf{e}_{2,t} &= \boldsymbol{s}_+ - \hat{\boldsymbol{s}}_+ \\ &= \boldsymbol{V}\left(\boldsymbol{K}^H \boldsymbol{\Sigma}^{-1} \boldsymbol{K}\right)^{-1} \boldsymbol{K}^H \boldsymbol{\Sigma}^{-1} \boldsymbol{n}. \end{aligned} \tag{33}$$

Then we have

$$\begin{aligned} \mathbb{E}[\|\mathbf{e}_{2,t}\|_2^2] &= \mathbb{E}\left[\|\boldsymbol{V}\left(\boldsymbol{K}^H \boldsymbol{\Sigma}^{-1} \boldsymbol{K}\right)^{-1} \boldsymbol{K}^H \boldsymbol{\Sigma}^{-1} \boldsymbol{n}\|_2^2\right] \\ &= \mathbb{E}\left[\|\boldsymbol{U}\boldsymbol{n}\|_2^2\right] \\ &= \mathbb{E}\left[\operatorname{Tr}\left[(\boldsymbol{U}\boldsymbol{n})(\boldsymbol{U}\boldsymbol{n})^H\right]\right] \\ &= \mathbb{E}\left[\operatorname{Tr}\left(\boldsymbol{U}\boldsymbol{n}\boldsymbol{n}^H\boldsymbol{U}^H\right)\right] \\ &= \operatorname{Tr}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^H) \\ &= \operatorname{Tr}(\boldsymbol{V}^H\boldsymbol{V}(\boldsymbol{K}^H\boldsymbol{\Sigma}^{-1}\boldsymbol{K})^{-1}), \end{aligned} \tag{34}$$

where $U = \boldsymbol{V}(\boldsymbol{K}^H\boldsymbol{\Sigma}^{-1}\boldsymbol{K})^{-1}\boldsymbol{K}^H\boldsymbol{\Sigma}^{-1}$, is a auxiliary matrix with L by $M(L+1)-1$ dimension; $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{n}$ defined in Appendix B; $\operatorname{Tr}(\cdot)$ returns the trace of the matrix.

For further simplification, we let $\boldsymbol{V}^H\boldsymbol{V} = \boldsymbol{Q}$, and $\boldsymbol{K} = \boldsymbol{P}\boldsymbol{K_h}$, where $\boldsymbol{P}$ can be obtain by changing the non-zero entry in $\boldsymbol{K}$ to 1 and $ML \times ML$ diagonal matrix $\boldsymbol{K_h}$ can be given as

$$\boldsymbol{K_h} = \operatorname{diag}(\boldsymbol{f}\boldsymbol{h}_1 b_1, \boldsymbol{f}\boldsymbol{h}_2 b_2, \ldots \boldsymbol{f}\boldsymbol{h}_m b_m, \boldsymbol{f}\boldsymbol{h}_1 b_1, \boldsymbol{f}\boldsymbol{h}_2 b_2, \ldots).$$

Substituting the above into (34), we have

$$\begin{aligned} \mathbb{E}[\|\mathbf{e}_{2,t}\|_2^2] &= \operatorname{Tr}\left(\boldsymbol{Q}[(\boldsymbol{P}\boldsymbol{K_h})^H\boldsymbol{\Sigma}^{-1}\boldsymbol{P}\boldsymbol{K_h}]^{-1}\right) \\ &= \operatorname{Tr}\left(\boldsymbol{Q}[\boldsymbol{K_h}^H\boldsymbol{P}^H\boldsymbol{\Sigma}^{-1}\boldsymbol{P}\boldsymbol{K_h}]^{-1}\right) \\ &= \operatorname{Tr}\left(\boldsymbol{Q}(\boldsymbol{K_h}^H\boldsymbol{W}\boldsymbol{K_h})^{-1}\right) \\ &= \operatorname{Tr}\left(\boldsymbol{K_h}^{-1}\boldsymbol{W}^{-1}\boldsymbol{K_h}^{-H}\boldsymbol{Q}\right) \\ &\overset{(a)}{=} \tilde{\boldsymbol{x}}^T(\boldsymbol{W}^{-1} \odot \boldsymbol{Q}^T)\bar{\tilde{\boldsymbol{x}}} \\ &= \tilde{\boldsymbol{x}}^H\bar{\boldsymbol{A}}\tilde{\boldsymbol{x}}, \end{aligned} \tag{35}$$

where $\boldsymbol{W} = \boldsymbol{P}^H\boldsymbol{\Sigma}^{-1}\boldsymbol{P}$ is a $ML \times ML$ hermitian matrix; (a) stems from the equality $\operatorname{Tr}(\mathbf{D_x A D_y B}^T) = \mathbf{x}^T(\mathbf{A} \odot \mathbf{B})\mathbf{y}$, [48] with $\mathbf{D_x} = \operatorname{diag}(\mathbf{x})$ and $\mathbf{D_y} = \operatorname{diag}(\mathbf{y})$, and $\odot$ denotes Hadamard product. $\bar{\boldsymbol{A}} = \boldsymbol{W}^{-1} \odot \boldsymbol{Q}^T$ is a $ML \times ML$ hermitian matrix, and, $\tilde{\boldsymbol{x}} = [(\boldsymbol{f}\boldsymbol{h}_1 b_1)^{-1}, (\boldsymbol{f}\boldsymbol{h}_2 b_2)^{-1}, \ldots (\boldsymbol{f}\boldsymbol{h}_m b_m)^{-1}, (\boldsymbol{f}\boldsymbol{h}_1 b_1)^{-1}, (\boldsymbol{f}\boldsymbol{h}_2 b_2)^{-1}, \ldots]^T \in \mathbf{C}^{ML\times 1}$.

To characterize the influence of noise scale on $\mathbf{e}_{2,t}$, we can further rewrite Eq. (35) as

$$\mathbb{E}[\|\mathbf{e}_{2,t}\|_2^2] = N_0\tilde{\boldsymbol{x}}^H\tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}}, \tag{36}$$

where $\tilde{\boldsymbol{A}} = (\boldsymbol{P}^H\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{P})^{-1} \odot \boldsymbol{Q}^T$, and the auxiliary matrix $\tilde{\boldsymbol{\Sigma}} = \frac{1}{N_0}\boldsymbol{\Sigma}$ is defined in (32).

Note that there are L groups of repeated elements in the vector $\tilde{\boldsymbol{x}}$, we can further simplify (35) by matrix blocking. Specifically, $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{x}}$ can be written as

$$\tilde{\boldsymbol{A}} = \begin{Bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1L} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{1L} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{A}_{L1} & \boldsymbol{A}_{L2} & \cdots & \boldsymbol{A}_{LL} \end{Bmatrix}, \tilde{\boldsymbol{x}} = \begin{Bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_L \end{Bmatrix}^T,$$

respectively, where $\boldsymbol{A}_{ij}$ is $M \times M$ matrix, and the vector $\boldsymbol{x}$ has been defined in (18), with $\boldsymbol{x}_i \equiv \boldsymbol{x} \in \mathbb{C}^{M\times 1}, \forall i \in \{1, \ldots, L\}$.

Consequently, Eq. (36) can be rewritten as

$$\mathbb{E}[\|\mathbf{e}_{2,t}\|_2^2] = N_0 \boldsymbol{x}^H \boldsymbol{A}\boldsymbol{x}, \tag{37}$$

where $\boldsymbol{A} = \sum_{i=1}^{L}\sum_{j=1}^{L} \boldsymbol{A}_{ij}$.

APPENDIX D PROOF OF THEOREM 1

From (12), we know that $\mathbf{e}_t = \mathbf{e}_{1,t} + \mathbf{e}_{2,t}$. Accordingly, we have

$$\begin{aligned}\mathbb{E}\Big[\|\mathbf{e}_t\|_2^2\Big] &= \mathbb{E}\Big[\|\mathbf{e}_{1,t} + \mathbf{e}_{2,t}\|_2^2\Big]\\ &\overset{(b)}{\leq} \mathbb{E}\Big[\big(\|\mathbf{e}_{1,t}\|_2 + \|\mathbf{e}_{2,t}\|_2\big)^2\Big]\\ &\overset{(c)}{\leq} 2\Big(\|\mathbf{e}_{1,t}\|_2^2 + \mathbb{E}\Big[\|\mathbf{e}_{2,t}\|_2^2\Big]\Big),\end{aligned} \tag{38}$$

where (b) is from the triangle inequality, and (c) is from the inequality $\|\mathbf{x}_1 + \mathbf{x}_2\|_2^2 \leq 2\|\mathbf{x}_1\|_2^2 + 2\|\mathbf{x}_2\|_2^2$. Due to $\mathbf{e}_{1,t}$ being from device selection and independent of the communication noise, the expectation on $\|\mathbf{e}_{1,t}\|_2^2$ can be removed.

To bound $\|\mathbf{e}_{1,t}\|_2^2$, we have the following equation in [46, Sec. 3.1],

$$\|\mathbf{e}_{1,t}\|_2^2 \leq 4\left(1 - \frac{\sum_{m=1}^{M}|\mathcal{D}_m|}{|\mathcal{D}|}\right)^2\Big(\alpha_1 + \alpha_2\|\nabla F(\mathbf{w}_t)\|_2^2\Big). \tag{39}$$

In addition, We scale through inequality to get an enlarged version of $\mathbb{E}[\|\mathbf{e}_{2,t}\|_2^2]$. Specifically,

$$\begin{aligned}\mathbb{E}\Big[\|\mathbf{e}_{2,t}\|_2^2\Big] &= N_0\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x}\\ &\overset{(d)}{\leq} N_0\lambda_m\|\boldsymbol{x}\|_2^2\\ &\leq N_0\lambda_m M \max_m |x_m|^2\\ &\leq N_0\lambda_m M \max_m \frac{1}{|\boldsymbol{f}^H\boldsymbol{h}_m|^2|b_m|^2}\\ &\leq N_0\lambda_m M \max_m \frac{1}{|\boldsymbol{f}^H\boldsymbol{h}_m|^2}\max_m\frac{1}{|b_m|^2}\\ &\leq N_0\lambda_m M \max_m \frac{1}{|\boldsymbol{f}^H\boldsymbol{h}_m|^2}\max_m v_m^2,\end{aligned} \tag{40}$$

where (d) is from the proof below.

Proof of (d): We let $\boldsymbol{x} = \boldsymbol{R}\boldsymbol{y}$, where $\boldsymbol{y}$ is a vector $\in \mathbb{C}^{m\times 1}$, and $\boldsymbol{R}$ is a unitary matrix, satisfying $\boldsymbol{R}^H\boldsymbol{A}\boldsymbol{R} = \boldsymbol{\Lambda}$, where Λ is a diagonal matrix. Substituting $\boldsymbol{x}$ with $\boldsymbol{R}\boldsymbol{y}$, we obtain

$$\begin{aligned}\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}^H\boldsymbol{\Lambda}\boldsymbol{y} = \sum_{i=1}^{m}\lambda_i|y_i|^2 \leq \sum_{i=1}^{m}\lambda_m|y_i|^2 = \lambda_m\boldsymbol{y}^H\boldsymbol{y}\\ = \lambda_m\boldsymbol{x}^H\boldsymbol{R}\boldsymbol{R}^H\boldsymbol{x} = \lambda_m\|\boldsymbol{x}\|_2^2.\end{aligned} \tag{41}$$

Moreover, for all m, we have

$$\begin{aligned}v_m^2 &= \frac{1}{L}\sum_{\ell=1}^{L}\left(\tilde{\mathbf{w}}_{m,t}[\ell] - \frac{1}{L}\sum_{\ell'=1}^{L}\tilde{\mathbf{w}}_{m,t}[\ell']\right)^2\\ &= \frac{1}{L}\sum_{\ell=1}^{L}\tilde{\mathbf{w}}_{m,t}^2[\ell] - \left(\frac{1}{L}\sum_{\ell'=1}^{L}\tilde{\mathbf{w}}_{m,t}[\ell']\right)^2\\ &\leq \frac{1}{L}\sum_{\ell=1}^{L}\tilde{\mathbf{w}}_{m,t}^2[\ell]\\ &= \frac{1}{L}\|\nabla F_m(\mathbf{w}_t)\|_2^2\\ &= \frac{1}{|\mathcal{D}_m|^2L}\|\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_m}\nabla f(\mathbf{w}_t;\mathbf{x}_i,y_i)\|_2^2\\ &\overset{(e)}{\leq} \frac{1}{|\mathcal{D}_m|^2L}\left(\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_m}\|\nabla f(\mathbf{w}_t;\mathbf{x}_i,y_i)\|_2\right)^2\\ &\overset{(f)}{\leq} \frac{1}{|\mathcal{D}_m|^2L}\left(\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_m}\sqrt{\alpha_1 + \alpha_2\|\nabla F(\mathbf{w}_t)\|_2^2}\right)^2\\ &= \frac{1}{L}\Big(\alpha_1 + \alpha_2\|\nabla F(\mathbf{w}_t)\|_2^2\Big),\end{aligned} \tag{42}$$

where (e) is from the triangle inequality and (f) is from Assumption 4, i.e., Eq. (15). Substituting (42) into (40), we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{e}_{2,t}\|_2^2] \leq \frac{N_0\lambda_m M}{L}\max_m\frac{1}{|\boldsymbol{f}^H\boldsymbol{h}_m|^2}\\ \times\Big(\alpha_1 + \alpha_2\|\nabla F(\mathbf{w}_t)\|_2^2\Big).\end{aligned} \tag{43}$$

Combining (16), (38), (39) and (43), we obtain

$$\begin{aligned}\mathbb{E}\big[F(\mathbf{w}_{t+1})\big] \leq \mathbb{E}[F\ \ (\mathbf{w}_t)] + \frac{\alpha_1}{\omega}\Phi\\ - \frac{\|\nabla F(\mathbf{w}_t)\|_2^2}{2\omega}(1 - 2\alpha_2\Phi),\end{aligned} \tag{44}$$

where Φ has been defined in (21).

Subtracting $F(\mathbf{w}^\star)$ on both sides of (44), and using the equality $\|\nabla F(\mathbf{w}_t)\|_2^2 \geq 2\mu(F(\mathbf{w}_t) - F(\mathbf{w}^\star))$ [46], we have

$$\mathbb{E}\big[F(\mathbf{w}_{t+1}) - F\big(\mathbf{w}^\star\big)\ \big] \leq \frac{\alpha_1}{\omega}\Phi + \mathbb{E}\big[F(\mathbf{w}_t) - F\big(\mathbf{w}^\star\big)\big]\Psi, \tag{45}$$

where Ψ has been defined in (21). Applying (45) recursively, we complete the proof.

APPENDIX E PROOF OF SUFFICIENT CONDITIONS FOR FL CONVERGENCE

Combining (39), (43), we have

$$\|\mathbf{e}_{1,t}\|_2^2 + \mathbb{E}[\|\mathbf{e}_{2,t}\|_2^2] \leq \Phi\Big(\alpha_1 + \alpha_2\|\nabla F(\mathbf{w}_t)\|_2^2\Big). \tag{46}$$

Substituting Lemma2 and the condition $\Phi < \frac{1}{2\alpha_2}$ into (46), we have

$$N_0\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x} < \frac{1}{2}\left(\frac{\alpha_1}{\alpha_2}\|\nabla F(\mathbf{w}_t)\|_2^2\right) - \|\mathbf{e}_{1,t}\|_2^2. \tag{47}$$

From (39), it can be observed that $\|\mathbf{e}_{1,t}\|_2^2$ has an upper bound. Therefore, as long as the above inequality holds when $\|\mathbf{e}_{1,t}\|_2^2$ takes this bound, then for any other $\|\mathbf{e}_{1,t}\|_2^2$, the $N_0\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x}$ is already sufficiently small. Substituting the upper bound $4(1 - \frac{\sum_{m=1}^{M}|\mathcal{D}_m|}{|\mathcal{D}|})^2(\alpha_1 + \alpha_2\|\nabla F(\mathbf{w}_t)\|_2^2)$, we have a sufficient condition for convergence.

REFERENCES

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

[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] 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.

[4] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[5] 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.

[6] 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.

[7] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Stat.*, vol. 54, Apr. 2017, pp. 1273–1282.

[8] A. Zhang, S. Guo, and S. Liu, "Private federated learning with dynamic power control via non-coherent over-the-air computation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2023, pp. 1–8.

[9] T. Lin, S. U. Stich, K. K. Patel, and M. Jaggi. "Don't use large mini-batches, use local SGD." 2018. [Online]. Available: https://arxiv.org/abs/1808.07217

[10] S. U. Stich. "Local SGD converges fast and communicates little." 2018. [Online]. Available: https://arxiv.org/abs/1805.09767

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

[12] X. Pan, J. Chen, R. Monga, S. Bengio, and R. Józefowicz. "Revisiting distributed synchronous SGD." 2017. [Online]. Available: https://arxiv.org/abs/1702.05800

[13] R. Tandon, Q. Lei, A. G. Dimakis, and N. Karampatziakis, "Gradient coding: Avoiding stragglers in distributed learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 70, Aug. 2017, pp. 3368–3376.

[14] D. Alistarh, D. Grubic, J. Z. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. NeurIPS*, vol. 30, Dec. 2017, pp. 1707–1718.

[15] S. Lv, S. Guo, and H. Zhang, "Distribution-aware weight compression for federated averaging learning over wireless edge networks," in *Proc. Int. Conf. Commun. China (ICCC)*, Jul. 2021, pp. 1107–1112.

[16] F. Sattler, S. Wiedemann, K.-R. Muller, and W. Samek, "Sparse binary compression: Towards distributed deep learning with minimal communication," in *Proc. Int. J. Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[17] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Copenhagen, Denmark, Sep. 2017, pp. 440–445. [Online]. Available: https://aclanthology.org/D17-1045

[18] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, Jan. 2021.

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

[20] W. Shi, S. Zhou, Z. Niu, M. Jiang, and L. Geng, "Joint device scheduling and resource allocation for latency constrained wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 453–467, Jan. 2021.

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

[22] M. Goldenbaum, H. Boche, and S. Stanczak, "Harnessing interference for analog function computation in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 61, no. 20, pp. 4893–4906, Oct. 2013.

[23] 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.

[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, "Over-the-air machine learning at the wireless edge," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2019, pp. 1–5.

[26] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "COTAF: Convergent over-the-air federated learning," in *Proc. IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–6.

[27] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Over-the-air federated learning from heterogeneous data," *IEEE Trans. Signal Process.*, vol. 69, pp. 3796–3811, 2021.

[28] X. Cao, G. Zhu, J. Xu, and S. Cui, "Transmission power control for over-the-air federated averaging at network edge," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1571–1586, May 2022.

[29] S. Huang, Y. Zhou, T. Wang, and Y. Shi, "Byzantine-resilient federated machine learning via over-the-air computation," in *Proc. IEEE Int. Conf. Commun. Workshops*, Jun. 2021, pp. 1–6.

[30] J. Wang, M. Dong, B. Liang, G. Boudreau, and H. Abou-Zeid, "Online model updating with analog aggregation in wireless edge learning," in *Proc. IEEE INFOCOM*, May 2022, pp. 1229–1238.

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

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

[33] M. Goldenbaum and S. Stanczak, "On the channel estimation effort for analog computation over wireless multiple-access channels," *IEEE Wireless Commun. Lett.*, vol. 3, no. 3, pp. 261–264, Jun. 2014.

[34] J. Dong, Y. Shi, and Z. Ding, "Blind over-the-air computation and data fusion via provable Wirtinger flow," *IEEE Trans. Signal Process.*, vol. 68, pp. 1136–1151, 2020.

[35] B. Tegin and T. M. Duman, "Blind federated learning at the wireless edge with low-resolution ADC and DAC," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 7786–7798, Dec. 2021.

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

[37] E. Becirovic, Z. Chen, and E. G. Larsson, "Optimal MIMO combining for blind federated edge learning with gradient sparsification," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2022, pp. 1–5.

[38] M. Goldenbaum and S. Stanczak, "Robust analog function computation via wireless multiple-access channels," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 3863–3877, Sep. 2013.

[39] O. Abari, H. Rahul, D. Katabi, and M. Pant, "AirShare: Distributed coherent transmission made seamless," in *Proc. IEEE INFOCOM*, May 2015, pp. 1742–1750.

[40] M. H. Adeli and A. Şahin, "Multi-cell non-coherent over-the-air computation for federated edge learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2022, pp. 4944–4949.

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

[42] 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.

[43] 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.

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

[45] G. Shi, S. Guo, J. Ye, N. Saeed, and S. Dang, "Multiple parallel federated learning via over-the-air computation," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 1252–1264, 2022.

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

[47] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[48] X.-D. Zhang, *Matrix Analysis and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

JIANDA WANG (Graduate Student Member, IEEE) received the B.Sc. degree in mechanical and electronic engineering from the School of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou, China, in 2021. He is currently pursuing the M.S. degree with Shandong University, Jinan, China. His research interests lie in the areas of communication efficiency and robustness of federated learning systems, and artificial intelligence techniques (machine learning, deep learning methods, and evolutionary algorithms).

SHUAISHUAI GUO (Senior Member, IEEE) received the B.E. and Ph.D. degrees in communication and information systems from the School of Information Science and Engineering, Shandong University, Jinan, China, in 2011 and 2017, respectively. He visited the University of Tennessee at Chattanooga, USA, from 2016 to 2017. He worked as a Postdoctoral Research Fellow with the King Abdullah University of Science and Technology, Saudi Arabia, from 2017 to 2019. He is currently working as a Full Professor with Shandong University. His research interests include 6G communications and machine learning.