Received 10 June 2022; revised 13 July 2022; accepted 23 July 2022. Date of publication 28 July 2022; date of current version 10 August 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3194821

Multiple Parallel Federated Learning via Over-the-Air Computation

GAOXIN SHI[1,2] (Student Member, IEEE), SHUAISHUAI GUO[1,2] (Member, IEEE), JIA YE[3] (Student Member, IEEE), NASIR SAEED[4] (Senior Member, IEEE), AND SHUPING DANG[5] (Member, IEEE)

[1] School of Control Science and Engineering, Shandong University, Jinan 250061, China

[2] Shandong Provincial Key Laboratory of Wireless Communication Technologies, Shandong University, Jinan 250061, China

[3] Electrical Engineering, Computer Electrical and Mathematical Sciences and Engineering Division,
King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia

[4] Department of Electrical Engineering, Northern Border University, Arar 9280, Saudi Arabia

[5] Department of Electrical and Electronic Engineering, University of Bristol, Bristol BS8 1UB, U.K.

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 This paper investigates multiple parallel federated learning in cellular networks, where a base station schedules several FL tasks in parallel and each task has a group of devices involved. To reduce the communication overhead, over-the-air computation is introduced by utilizing the superposition property of multiple access channels (MAC) to accomplish the aggregation step. Since all devices use the same radio resource to transfer their local updates to the BS, in order to separate the received signals of different tasks, we use the zero-forcing receiver combiner to mitigate the mutual interference across different groups. Besides, we analyze the impact of receiver combiner and device selection on the convergence of our multiple parallel FL framework. Also, we formulate an optimization problem that jointly considers receiver combiner vector design and device selection for improving FL performance. We address the problem by decoupling it into two sub-problems and solve them alternatively, adopting successive convex approximation (SCA) to derive the receiver combiner vector, and then solve the device scheduling problem with a greedy algorithm. Simulation results demonstrate that the proposed framework can effectively solve the straggler issue in FL and achieve a near-optimal performance on all tasks.

INDEX TERMS Device selection, federated learning, multiple access channel, over-the-air computation, receiver combiner.

I. INTRODUCTION

THE RAPID growth of technologies, such as the Internet of Things (IoT) and social networking, lead to an exponential explosion of data at edge devices [1]. These trends promote the implementation of smart services based on machine learning, e.g., computer vision [2], natural language processing [3], and speech recognition [4]. In general, conventional machine learning trains a model in a centralized server to collect data from edge devices, which might be impractical soon due to rising privacy concerns and communication burdens. Meanwhile, the improvement of smart devices' computing and storage capacity makes it possible for devices to process data locally [5]. To overcome the challenges of centralized machine learning, federated learning (FL) has been proposed to train a shared learning model collaboratively at edge devices under the schedule of the central server, e.g., a base station (BS), which avoids uploading private and sensitive user data from local clients [6].

Nevertheless, communication bandwidth is a key bottleneck affecting the performance of FL, while the straggler issue caused by the system heterogeneity of computation capability and wireless channel condition makes it even worse [7], [8], [9], [10]. A common way to overcome this difficulty is to reduce the number of participating devices via some scheduling policies [11], [12], [13], [14]. Another way is to reduce the amount of parameters required to upload from clients to the central server via quantization [15], [16], [17] or sparsification [18], [19]. Although the above methods successfully reduce the communication costs, FL performance is still constrained by the communication capability of the network, especially when a large number of devices participate in the FL process with limited communication bandwidth. This is because all these methods suppose that uplink communications between BS and devices use conventional orthogonal-access schemes, e.g., orthogonal frequency division multiple access (OFDMA) or time division multiple access (TDMA), such that the spectral resource allocated to each device will drop sharply as the number of devices increases.

In order to reduce the communication costs, over-the-air computation was introduced to aggregate data in sensor networks [20]. In this setting, all users transmit their data simultaneously via the same radio resources over multiple access channels (MAC); then, without decoding the information of each device, the computation is done utilizing the superposition property of the wireless channel. Compared with conventional orthogonal-access schemes, over-the-air computation can significantly improve communication efficiency, especially when there are a large number of devices in wireless networks because the required communication resource will not increase with the number of devices. Moreover, it is worth noting that over-the-air computation is only applicable when the BS wants to obtain the uniform summation or its variant results from all devices [21]. Due to its energy efficiency, over-the-air computation is more suitable for IoT network. In [22], the authors proposed a framework that was robust against synchronization errors. The work [23] considered a generalized IoT network where multiple different clusters of sensors independently compute different target function. A sensor selection algorithm to improve the computation performance was proposed in [24]. In [25], the authors built a experiment platform to verify the validity of over-the-air computation.

In a typical FL process, the BS receives distributed updates (model parameters or gradients) uploaded from edge devices via MAC and then averages them to update the global model, which is a classic adaptation scenario of over-the-air computation.

A. RELATED WORKS

The over-the-air computation-based FL aggregation was firstly introduced in [26], where the authors derived two trade-offs between communication and learning to quantify the selected device population. At the same time, the parallel work [27] considered the same trade-offs and maximized the number of devices with respect to the mean squared error (MSE) of gradient error. Then the author in [26] extended their work to one-bit over-the-air computation FL in [28] and [29], where a new scheme featuring one-bit quantization followed by modulation at edge devices and majority-voting based decoding at the edge server was proposed. The works [30], [31] supposed that the model update vector is sparse and projected the resultant sparse vector into a low-dimensional vector for reducing the bandwidth in the over-the-air FL. Moreover, the power control of the over-the-air computation FL was studied in [32] and [33]. The goal of [32] is to minimize the MSE of gradients by optimizing the transmit power at each device subject to average power constraints. Further, the authors in [32] analyzed the convergence of the over-the-air computation FL under any given power control policy to optimize the transmit power. The tractable FL convergence analysis of full gradient descent optimization was done in [34], [35]. In [36], reconfigurable intelligent surfaces (RIS) were leveraged to improve the performance of the over-the-air FL. Specifically, the authors developed a convergence analysis framework of the RIS-aided over-the-air computation FL and tried to solve the straggler issue by device scheduling.

B. MOTIVATIONS

The aforementioned works have well optimized the performance of FL. However, they all consider a single FL task over wireless networks. When FL become a service or a popular application in the network [37], the central server (e.g., the base station) may need to schedule multiple FL processes simultaneously.In this situation, the mutual interference across different FL processes need to be considered, as it may bring a huge reduction to learning performance. Different from the latest paper [38] which focuses on multiple FL tasks over multi-cell wireless networks, in this paper we study multiple parallel FL via over-the-air computation over wireless networks where the central server schedule multiple FL processes simultaneously. We jointly consider receiver combiner vector design and device selection policy and effectively solve the high communication costs and straggler issue in FL under the premise of sacrificing only slight FL performance.

C. CONTRIBUTIONS

The main contribution of this paper is to propose a novel framework for the implementation of multiple over-the-air FL in wireless networks by jointly taking the receiver combiner design and device selection into account. To our best knowledge, this is the first work that considers multiple FL process via over-the-air computation. The contributions of this paper are summarized as follows:

- We propose a novel over-the-air computation FL framework, in which one BS services multiple groups of devices to train different FL models. In the uploading stage, all devices from different groups use the same radio resources to transmit their local updates to the BS, and then, by utilizing the superposition property of MAC, the BS receives the sum of signals from all groups of devices. To perform FL machine learning models accurately for every group, we propose a zero-forcing receiver combiner design to separate the received signals of different tasks.

- We analyze the convergence of FL within our framework. Specifically, we derive an upper bound on the gap between the realistic and ideal optima value of global loss function with respect to the aggregation error caused by transmission distortion and device selection, and find how combiner design and device selection policy affect FL performance, e.g., convergence and FL loss function. Based on this analysis, we formulate a mixed-integer non-convex programming problem that jointly optimizes the combiner vector and device set.
- To solve the formulated mixed-integer non-convex programming problem, we first decouple it into two sub-problems, namely, receiver combiner, and device scheduling. Then, we solve them separately in an alternate way. Specifically, given the device selection policy, we use successive convex approximation (SCA) proposed in [39] to derive the receiver combiner vector. Further, based on the derived receiver combiner vector, we solve the device scheduling problem with the greedy algorithm.
- Simulation results show that our proposed framework can efficiently transfer gradient information of all tasks. Besides, we can see that the straggler issue severely and adversely affects FL performance of conventional FL systems. Our proposed framework effectively solves the straggler issue and achieves near-optimal (noise-less aggregation) performance on all processed FL tasks.

D. ORGANIZATION

The remainder of this paper is organized as follows. Section II introduces the FL model, the MAC communication model, and the multiple FL aggregation framework via over-the-air computation. In Section III, we analyze the FL expected convergence rate and formulate the optimization problem to minimize the FL training loss. The optimal receiver combiner design and user selection policy are determined in Section IV. Simulation results are analyzed in Section V. Conclusions are drawn in Section VI.

E. NOTATIONS

In this paper, scalars, vetors, and matrices are denoted by regular letters, boldface lowercase letters, and boldface uppercase letters, respectively. $\mathbb{R}$ and $\mathbb{C}$ are used to denote the real and complex number set, respectively. $(\cdot)^T$ and $(\cdot)^H$ denote the transpose operator and complex conjugate transpose operator, respectively. $\mathcal{CN}(\mu, \sigma^2)$ represents the circularly symmetric complex Gaussian random distribution with mean μ and variance σ. The l_2-norm of a vector is denoted by $||\cdot||$ and the size of set $\mathcal{S}$ is denoted by $|\mathcal{S}|$. diag$(\cdot)$ stands for a diagonal matrix of vector whose diagonal entities are specified by the vector enclosed, and $\mathbb{E}[\,\cdot\,]$ means expectation.

II. SYSTEM MODEL

In this paper, we consider a cellular network in which a set $\mathcal{M}$ with M groups of devices perform M different FL tasks with different training models via the same BS, as shown in Fig. 1.

A. FEDERATED LEARNING MODEL

In the system, each group m $(1 \leq m \leq M)$ trains a machine learning model represented by the parameter vector $\boldsymbol{w_m} \in \mathbb{R}^{D_m \times 1}$ with D_m denoting the model size. The learning objective of group m is done in a way to solve the following optimization problem:

$$\min_{\boldsymbol{w}_m \in \mathbb{R}^{D_m \times 1}} F_m(\boldsymbol{w}_m) = \frac{1}{K_m} \sum_{k=1}^{K_m} f_m\left(\boldsymbol{w}_m; \mathbf{x}_m^k, y_m^k\right), \tag{1}$$

where K_m is the total number of training samples of group m; $(\mathbf{x}_m^k, y_m^k)$ is the kth training sample with $\mathbf{x}_m^k$ and y_m^k denoting the input feature and output label respectively; $f_m(\boldsymbol{w}_m; \mathbf{x}_m^k, y_m^k)$ denotes the loss function with respect to $(\mathbf{x}_m^k, y_m^k)$. Suppose that there is a set $\mathcal{I}_m$ with I_m devices in group m, and the ith device has $K_{m,i}$ training sampels with $\sum_{i=1}^{I_m} K_{m,i} = K_m$. The training dataset at the ith device in group m is represented by $\mathcal{D}_{m,i} = \{(\mathbf{x}_{m,i}^k, y_{m,i}^k) : 1 \leq k \leq K_{m,i}\}$ with $|\mathcal{D}_{m,i}| = K_{m,i}$, and then the objective in (1) turns into:

$$F_m(\boldsymbol{w}_m) = \frac{1}{K_m} \sum_{i \in \mathcal{I}_m} K_{m,i} F_{m,i}\left(\boldsymbol{w}_m; \mathcal{D}_{m,i}\right), \tag{2}$$

with

$$F_{m,i}(\boldsymbol{w}_m; \mathcal{D}_{m,i}) \triangleq \frac{1}{K_{m,i}} \sum_{\left(\mathbf{x}_{m,i}^k, y_{m,i}^k\right) \in \mathcal{D}_{m,i}} f_m\left(\boldsymbol{w}_m; \mathbf{x}_{m,i}^k, y_{m,i}^k\right). \tag{3}$$

To overcome the bottleneck of limited network bandwidth, federated averaging (FedAvg) is developed to reduce communication rounds between devices and the BS [6]. Specifically, at the t-th round in group m, the following processing will be conducted in sequence by FedAvg:

- The BS selects a subset of devices $\mathcal{I}_m^t \subseteq \mathcal{I}_m$ to participate in the current round;
- The BS sends the current global model $\boldsymbol{w}_m^t$ to the selected devices via multicast;
- Each device adopts a standard gradient descent method to compute their local gradients respect to the local dataset as specified in [40]. Specifically, the gradient of device i in group m is given by

$$\boldsymbol{g}_{m,i}^t \triangleq \nabla F_{m,i}\left(\boldsymbol{w}_m^t; \mathcal{D}_{m,i}\right) \in \mathbb{R}^{D_{m,i} \times 1}, \tag{4}$$

where $i \in \mathcal{I}_m^t$, $\nabla F_{m,i}(\boldsymbol{w}_m^t; \mathcal{D}_{m,i})$ is the gradient of $F_{m,i}(\cdot)$ at $\boldsymbol{w} = \boldsymbol{w}_m^t$.

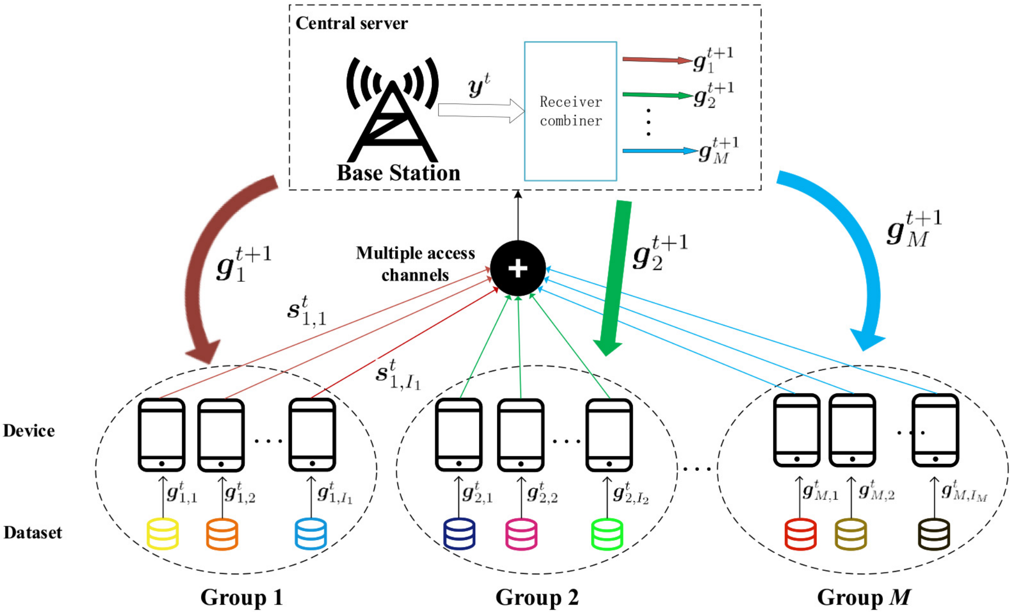

FIGURE 1. System model of multiple federated learning via over-the-air computation.

- The devices upload $\boldsymbol{g}^t_{m,i}$ to the BS, and then the BS performs FedAvg to update the global model. In this case, we can estimate $\boldsymbol{r}^t_m \triangleq \sum_{i\in\mathcal{I}^t_m} K_{m,i}\boldsymbol{g}^t_{m,i}$ at the BS from the received signals. We denote $\hat{\boldsymbol{r}}^t_m$ as the estimate of $\boldsymbol{r}^t_m$ (true value); accordingly the global model of group m is updated by

$$\boldsymbol{w}^{t+1}_m = \boldsymbol{w}^t_m - \frac{\lambda}{\sum_{i\in\mathcal{I}^t_m} K_{m,i}}\hat{\boldsymbol{r}}^t_m, \tag{5}$$

where λ denotes the learning rate.

B. COMMUNICATION MODEL

In this paper, we focus on uplink transmissions between single-antenna devices and an N-antenna BS over a MAC based on the fact that the uploading process dominates the convergence of FL systems, and we consider over-the-air computation for fast update aggregation by exploiting the superposition property of MAC. We assume a block fading channel where channel coefficients remain constant within a communication round, but may change over different communication rounds. Besides, we assume that the channel state information (CSI) is available at all participating entities.

At the tth communication round, we denote the channel coefficient vector between the BS and the ith device in the mth group by $\mathbf{h}_{m,i} \in \mathbb{C}^{N\times 1}$, $i \in \mathcal{I}_m$, $m \in \mathcal{M}$. Letting $\mathbf{s}^t_{m,i}[d]$ denote the transmit signal from device i, the received signal at the BS, denoted by $\mathbf{y}^t[d]$, is given by

$$\mathbf{y}^t[d] = \sum_{m\in\mathcal{M}}\sum_{i\in\mathcal{I}_m} \mathbf{h}_{m,i}\mathbf{s}^t_{m,i}[d] + \mathbf{n}^t[d], \tag{6}$$

where $\mathbf{n}^t[d]$ is an additive white Gaussian noise (AWGN) vector with the entries following the distribution of $\mathcal{CN}(0, \sigma_n^2)$.

To simplify the notation, we omit the time index t. Denote the dth elements of $\boldsymbol{g}_{m,i}$ by $\boldsymbol{g}_{m,i}[d]$. In order to exploiting the superposition property of MAC to accomplish FedAvg, $\{\boldsymbol{g}_{m,i}[d] : 1 \leq d \leq D_{m,i}, i \in \mathcal{I}_m, m \in \mathcal{M}\}$ are first processed to $D_{m,i}$ slot transmit signal $\{\mathbf{s}_{m,i}[d]{:}1 \leq d \leq D_{m,i}, i \in \mathcal{I}_m, m \in \mathcal{M}\}$. First, each device compute the local gradient statistics by

$$\bar{\boldsymbol{g}}_{m,i} = \frac{1}{D_{m,i}}\sum_{d=1}^{D_{m,i}} \boldsymbol{g}_{m,i}[d],$$
$$\nu^2_{m,i} = \frac{1}{D_{m,i}}\sum_{d=1}^{D_{m,i}} \left(\boldsymbol{g}_{m,i}[d] - \bar{\boldsymbol{g}}_{m,i}\right)^2. \tag{7}$$

Then, each device transfers $\boldsymbol{g}_{m,i}[d]$ to $\mathbf{s}_{m,i}[d]$ by

$$\mathbf{s}_{m,i}[d] = p_{m,i}\mathbf{x}_{m,i}[d] \text{ with } \mathbf{x}_{m,i}[d] \triangleq \frac{\boldsymbol{g}_{m,i}[d] - \bar{\boldsymbol{g}}_{m,i}}{\nu_{m,i}}, \quad \forall d, \tag{8}$$

where $p_{m,i} \in \mathbb{C}$ is the transmitter scalar used to combat channel fading and accomplish the weighting process of over-the-air FedAvg. The normalization step in (8) ensures that $\mathbb{E}[|\mathbf{x}_{m,i}[d]|] = 0$ and $\mathbb{E}[|\mathbf{x}_{m,i}[d]|^2] = 1$, such that the transmit power constraint at device i is constrained by

$$\mathbb{E}\left[|\mathbf{s}_{m,i}[d]|^2\right] = |p_{m,i}|^2 \leq P_0, \quad \forall i, m, \tag{9}$$

with $P_0 > 0$ as the maximum transmit power.

By substituting (8) into (6), the received signal at the BS in time slot d is given by

$$\begin{aligned}\mathbf{y}[d] &= \sum_{m\in\mathcal{M}}\sum_{i\in\mathcal{I}_m} \mathbf{h}_{m,i}\mathbf{x}_{m,i}[d] + \mathbf{n}[d]\\ &= \sum_{m\in\mathcal{M}}\sum_{i\in\mathcal{I}_m} \mathbf{h}_{m,i}\frac{p_{m,i}}{\nu_{m,i}}(\boldsymbol{g}_{m,i}[d] - \bar{\boldsymbol{g}}_{m,i}) + \mathbf{n}[d]. \end{aligned}\tag{10}$$

TABLE 1. List of notations.

Notation	Description	Notation	Description
M	Number of tasks	$\mathbf{h}_{m,i} \in \mathbb{C}^{N\times 1}$	Channel coefficient vector of device i in task m
$\mathcal{M}$	Group set	$\mathbf{y}$	Received signal at the BS
N	Number of BS antennas i	$\bar{\boldsymbol{g}}_{m,i}$	Mean of all gradient elements of device i in task m
$\mathcal{I}_m$	User set of task m	$\nu_{m,i}$	Variance of all gradient elements of device i in task m
$\boldsymbol{w_m} \in \mathbb{R}^{D_m\times 1}$	Global model of task m	$\mathbf{s}_{m,i}$	Gradient signal of device i in task m
D_m	Model size of task m	$p_{m,i}$	Transmit scalar of device i in task m
$\mathcal{D}_{m,i}$	Dataset collected by user i of task m	P_0	Maximum transmit power of each device
K_m	Total number of data samples of task	η_m	Normalization scalar of task m
$K_{m,i}$	Number of data samples of device i in task m	$\boldsymbol{\Omega} \in \mathbb{C}^{N\times M}$	Receiver combiner matrix
$F_m(\boldsymbol{w}_m)$	Global FL loss of task m	$\mathbf{f}_m \in \mathbb{C}^{N\times 1}$	Receiver combiner vector of task m
$F_{m,i}(\boldsymbol{w}_m; \mathcal{D}_{m,i})$	Local FL loss of device i in task m	$\mathbf{n}$	Additive white Gaussian noise
$f_m(\boldsymbol{w}_m; \mathbf{x}_m^k, y_m^k)$	Machine learning loss of task m	σ_n^2	Variance of noise
$\boldsymbol{g}_{m,i} \in \mathbb{R}^{D_m\times 1}$	Gradient of device i in task m	ξ	Rate of selected devices
$\boldsymbol{r}_m \in \mathbb{R}^{D_m\times 1}$	Weighted sun of gradient of task m i	ε	Threshold of SCA

To perform over-the-air model aggregation, the BS computes the estimate of $\boldsymbol{r}_m[d] = \sum_{i\in\mathcal{I}_m} K_{m,i}\boldsymbol{g}_{m,i}[d]$ from $\mathbf{y}[d]$ as[1]

$$\begin{aligned}\boldsymbol{r} &= \frac{1}{\sqrt{\eta_m}}\Omega^H\mathbf{y}[d] + \bar{\boldsymbol{g}}\\ &= \sum_{m\in\mathcal{M}}\left(\frac{1}{\sqrt{\eta_m}}\sum_{i\in\mathcal{I}_m}\Omega^H\mathbf{h}_{m,i}\mathbf{s}_{m,i}[d]\right)\\ &= \sum_{m\in\mathcal{M}}\left(\frac{1}{\sqrt{\eta_m}}\sum_{i\in\mathcal{I}_m}\Omega^H\mathbf{h}_{m,i}\frac{p_{m,i}}{v_{m,i}}(\boldsymbol{g}_{m,i}[d]-\bar{\boldsymbol{g}}_{m,i})\right)\\ &\quad +\Omega^H\mathbf{n}[d] + \bar{\boldsymbol{g}}, \end{aligned} \tag{11}$$

where $\boldsymbol{r} = [\hat{\boldsymbol{r}}_1[d], \hat{\boldsymbol{r}}_2[d], \ldots, \hat{\boldsymbol{r}}_M[d]]^T$; $\Omega \in \mathbb{C}^{N\times M}$ is the receiver matrix; $\bar{\boldsymbol{g}} = (\bar{\boldsymbol{g}}_1, \bar{\boldsymbol{g}}_2, \ldots, \bar{\boldsymbol{g}}_M)$ with $\bar{\boldsymbol{g}}_m \triangleq \sum_{i\in\mathcal{I}_m} K_{m,i}\bar{\boldsymbol{g}}_{m,i}, \forall m \in \mathcal{M}$ used to restore the subtracted mean value of transmit signal in the regularization according to step (7); and $\eta_m > 0$ is a normalization scalar.

Taking the mth($m \in \mathcal{M}$) FL model as an example, the first term after the second equal sign of the above formula can be rewritten as

$$\begin{aligned}&\sum_{m\in\mathcal{M}}\left(\frac{1}{\sqrt{\eta_m}}\sum_{i\in\mathcal{I}_m}\Omega^H\mathbf{h}_{m,i}\mathbf{s}_{m,i}[d]\right)\\ &= \left(\frac{1}{\sqrt{\eta_m}}\sum_{i\in\mathcal{I}_m}\Omega^H\mathbf{h}_{m,i}\mathbf{s}_{m,i}[d]\right.\\ &\quad \left. + \sum_{n\in\mathcal{M}\backslash\{m\}}\frac{1}{\sqrt{\eta_n}}\sum_{j\in\mathcal{I}_n}\Omega^H\mathbf{h}_{n,j}\mathbf{s}_{n,j}[d]\right). \end{aligned} \tag{12}$$

As we can observe from (12), the received signal at the base station after receiving combining is the sum of gradient information from all participating devices. However, when the BS targets the mth FL process, only gradient information from mth group is needed, while the gradient information from other groups will be treated as interference. Therefore, it is necessary to design a combining scheme to decode the received signals of different groups in a separate manner.

C. ZERO-FORCING RECEIVER COMBINER DESIGN

The signal received by the base station is the weighted sum of the model signals of all tasks. However, machine learning based on gradient descent method is sensitive to interference. For the purpose of training models of the group m accurately, the receiver combiner matrix Ω needs to separate the received signals of different tasks, which reminds us of the zero-forcing receiver. Specifically, the BS should treat the gradients of different tasks as mutual interference and force the interference to zero, such that $\Omega = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M]$ should be designed to meet the following criterion:

$$\mathbf{f}_m^H\mathbf{h}_{n,i} = 0, \forall n \in \mathcal{M}\backslash m, \forall i \in \mathcal{I}_n, \tag{13}$$

where $\mathbf{f}_m \in \mathbb{C}^{N\times 1}$, and $m = 1, 2, \ldots, M$.

Note that the proposed scheme can not only reduce the communication cost, but also minimize the training delay, because it schedule multiple FL tasks in a parallel way. For clear presentation, we list all notations used in this paper in Table 1.

III. PERFORMANCE ANALYSIS AND PROBLEM FORMULATION

In this section, we analyze how the device selection and communication noise affects the performance of the federated learning under the over-the-air model aggregation framework.

A. LEARNING PERFORMANCE ANALYSIS

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

Assumption 1: $F(\cdot)$ is rigorously convex with positive parameter μ, such that for any $\boldsymbol{w}$ and $\boldsymbol{w}'$:

$$F(\boldsymbol{w}) \geq F(\boldsymbol{w}') + (\boldsymbol{w} - \boldsymbol{w}')^T\nabla F(\boldsymbol{w}') + \frac{\mu}{2}\|\boldsymbol{w} - \boldsymbol{w}'\|^2. \tag{14}$$

[1] Since $p_{m,i}$ and $v_{m,i}$ are scalar quantities and a typical machine learning model consists of a large number of parameters, we suppose that these two scalars can be uploaded with no error and communication costs for modeling and analytical simplify.

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

$$\|\nabla F(\boldsymbol{w}) - \nabla F\left(\boldsymbol{w}'\right)\| \leq L\|\boldsymbol{w} - \boldsymbol{w}'\|. \tag{15}$$

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

Assumption 4: The gradient computed by each sample is bounded as a function of the true gradient as follow:

$$\left\|\nabla f\left(\boldsymbol{w}; \mathbf{x}_i^k, y_i^k\right)\right\|^2 \leq \beta_1 + \beta_2\|\nabla F(\boldsymbol{w})\|^2, \quad \forall i, \forall k, \tag{16}$$

where $\beta_1, \beta_2 \geq 0$.

Remark 1: Assumptions 1-4 are satisfied for most machine learning loss functions, such as squared support vector machine (SVM) and linear regression [41], and are widely used in the literature of performance analysis for FL [10], [42]. Although some machine learning models, such as neural network, might not satisfy Assumption 1, our experimental results presented later will clearly show that the proposed receiver combiner and device selection policy based on these four assumptions work well.

Assumptions 1-4 leads to an upper bound on $\|\nabla F(\boldsymbol{w}^t)\|^2$ with a proper learning rate λ. According the analysis in [43], we have

$$\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|^2 \leq 2L\left[F\left(\boldsymbol{w}^t\right) - F\left(\boldsymbol{w}^\star\right)\right] \tag{17}$$

where the learning rate is given as $\lambda = \frac{1}{L}$ and $F(\boldsymbol{w}^\star)$ denotes the global optima.

Based on (5), the global model at iteration t is updated by the relation given infra:

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \frac{\lambda}{\sum_{i\in\mathcal{I}} K_i}\hat{\boldsymbol{r}}^t = \boldsymbol{w}^t - \lambda\left(\nabla F\left(\boldsymbol{w}^t\right) - \boldsymbol{e}^t\right), \tag{18}$$

where $\boldsymbol{e}^t = \nabla F(\boldsymbol{w}^t) - \hat{\boldsymbol{r}}^t/\sum_{i\in\mathcal{I}^t} K_i$, which denotes the gradient error caused by device selection and communication noise. According the analysis in [43], the upper bound on $\mathbb{E}[F(\boldsymbol{w}^{t+1}) - F(\boldsymbol{w}^\star)]$ can be given by:

$$\begin{aligned}\mathbb{E}\left[F\left(\boldsymbol{w}^{t+1}\right) - F(\boldsymbol{w}^\star)\right] \leq \left(1 - \frac{\mu}{L}\right)\mathbb{E}\left[F\left(\boldsymbol{w}^t\right) - F\left(\boldsymbol{w}^\star\right)\right]\\ + \frac{1}{2L}\mathbb{E}\left[\|\boldsymbol{e}^t\|^2\right],\end{aligned} \tag{19}$$

where $\mathbb{E}(\cdot)$ returns the expected value of the random variable/quantity enclosed.

In order to lessen the gap between the realistic and ideal optima value of global loss function, we need reduce the value of $\boldsymbol{e}^t$. Since the gradient error $\boldsymbol{e}^t$ is determined by device selection and communication noise, given the device selection policy, the transmitter scalar is determined in the following proposition.

Proposition 1: Given the channel coefficient, receiver combiner vector and device selection policy, the optimal transmitter scalar that minimizes the gradient error is designed by

$$p_i^* = \frac{K_i\sqrt{\eta}\nu_i\left(\mathbf{f}^H\mathbf{h}_i\right)^H}{\left|\mathbf{f}^H\mathbf{h}_i\right|^2}, \quad \forall i. \tag{20}$$

Proof: See Appendix A. ■

Considering the transmit power constraint formulated in (9) and Proposition 1, the optimal η that minimizes the gradient error can be computed as

$$\eta^* = \min_{i\in\mathcal{I}^t} \frac{P_0\left|\mathbf{f}^H\mathbf{h}_i\right|^2}{K_i^2\nu_i^2}. \tag{21}$$

From Proposition 1, we can get the tractable expression of the gradient error $\boldsymbol{e}^t$, based on which we can derive an upper bound on $\mathbb{E}[F(\boldsymbol{w}^{t+1}) - F(\boldsymbol{w}^\star)]$ in the following theorem with respect to any given device selection policy $\mathcal{I}^t$ and receiver combiner vector $\mathbf{f}$.

Theorem 1: Supposing that the assumptions 1-4 holds, with p_i and η given in (20) and (21), for arbitrary $\{\mathcal{I}^t, \mathbf{f}\}$, we have

$$\begin{aligned}&\mathbb{E}\left[F\left(\boldsymbol{w}^{t+1}\right) - F\left(\boldsymbol{w}^\star\right)\right]\\ &\quad\leq [\psi]^{t+1}\mathbb{E}\left[F\left(\boldsymbol{w}^t\right) - F\left(\boldsymbol{w}^\star\right)\right]\\ &\qquad+ \frac{\beta_1}{L}d\left(\mathcal{I}^t, \mathbf{f}\right)\frac{1 - [\psi]^t}{1 - \psi},\end{aligned} \tag{22}$$

where $\psi = 1 - \frac{\mu}{L} + d(\mathcal{I}^t, \mathbf{f})$ and $[\psi]^t$ denotes exponentiation with base ψ and power t; $d(\mathcal{I}^t, \mathbf{f}) = \frac{\sigma_n^2}{(\sum_{i\in\mathcal{I}^t} K_i)^2} \max_{i\in\mathcal{I}^t} \frac{K_i^2}{P_0|\mathbf{f}^H\mathbf{h}_i|^2} + 4(\frac{\sum_{i\in\mathcal{I}} K_i - \sum_{i\in\mathcal{I}^t} K_i}{\sum_{i\in\mathcal{I}} K_i})^2$.

Proof: See Appendix B. ■

From Theorem 1, we can see a gap, $\frac{\beta_1}{L}d(\mathcal{I}^t, \mathbf{f})\frac{1-[\psi]^t}{1-\psi}$, exists between $\mathbb{E}[F(\boldsymbol{w}^{t+1})]$ and $\mathbb{E}[F(\boldsymbol{w}^\star)]$. This gap is caused by communication noise and device selection policy. Specifically, as the communication noise decreases, the gap between $\mathbb{E}[F(\boldsymbol{w}^{t+1})]$ and $\mathbb{E}[F(\boldsymbol{w}^\star)]$ decreases. Meanwhile, when the number of training samples used to perform FL algorithm increases, the gap also decreases.

B. PROBLEM FORMULATION

From Theorem 1, we can observe that ψ controls the convergence rate of the FL algorithm. A smaller ψ means faster convergence rate and the FL algorithm will not converge when $\psi \geq 1$. Therefore, in this paper we only consider the case where $\psi < 1$. As a result, as $t \to \infty$, we have $[\psi]^t = 0$, the gap between $\mathbb{E}[F(\boldsymbol{w}^{t+1})]$ and $\mathbb{E}[F(\boldsymbol{w}^\star)]$ can be rewritten as $\frac{\beta_1}{L}d(\mathcal{I}^t, \mathbf{f})$. Moreover, from the expression of ψ and the gap, we can see that the convergence rate ψ and the gap are both monotonic functions of $d(\cdot)$. As a result, a smaller $d(\cdot)$ leads to faster convergence and a smaller gap. Besides, we see that the selected device set $\mathcal{I}_m^t$ and receiver combiner vector $\mathbf{f}_m$ determine the value of $d(\cdot)$. Based on the above observations, we formulate the following minimization problem for task m:

$$\begin{aligned}\min_{\mathcal{I}_m^t, \mathbf{f}_m} \ & 4\left(\frac{K - \sum_{i\in\mathcal{I}_m^t} K_i}{K}\right)^2\\ &+ \frac{\sigma_n^2}{\left(\sum_{i\in\mathcal{I}_m^t} K_i\right)^2} \max_{i\in\mathcal{I}_m^t} \frac{K_i^2\|\mathbf{f}_m\|^2}{P_0|\mathbf{f}_m^H\mathbf{h}_{m,i}|^2},\end{aligned}$$

$$\begin{aligned} \text{s.t.}\ \ \text{C1} \quad & \mathcal{I}_m^t \subseteq \mathcal{I}_m \\ \text{C2} \quad & \mathbf{f}_m^H \mathbf{h}_{n,i} = 0, \forall n \in \mathcal{M}\backslash m, \ \ \forall i \in \mathcal{I}_n^t \end{aligned} \tag{23}$$

where C2 is zero-forcing constraint. Obviously, the objective function in (23) is non-convex, and the optimization variable $\mathcal{I}_m^t$ is a set. Hence, the objective problem (23) is a mixed-integer non-convex optimization problem. Due to the heterogeneity of the system, different users have different amounts of data and channel state information. From the optimization problem we can see that, The first item of the objective function requires selecting a device with a large amount of data, but this may cause the effect of noise in the second item to be amplified. In addition, the second term requires the selection of devices with good channel conditions and the design of the receiver vector to minimize the effects of channel fading.

IV. JOINT OPTIMIZATION OF RECEIVER COMBINER AND DEVICE SELECTION

In this section, our goal is to solve the minimization problem fomulated in (23). However, (23) is a mixed-integer non-convex optimization problem with non-convex objective and constraints. To facilitate solving this type of optimization problems, certain tactics are necessary. First, we decouple it into two sub-problems, namely, receiver combiner and device scheduling, and solve them alternately. Specifically, given the device selection policy, we use SCA to derive the receiver combiner vector. Further, based on the derived receiver combiner vector, we solve the device scheduling problem with the greedy algorithm.

A. RECEIVER COMBINER DESIGN

Given the device scheduling policy of task m at the tth round $\mathcal{I}_m^t$, the minimization problem (23) can be written as follows:

$$\begin{aligned} \min_{\mathbf{f}_m} \max_{i\in\mathcal{I}_m^t} \ & \frac{\|\mathbf{f}_m\|^2}{|\mathbf{f}_m^H \mathbf{h}_{m,i}|^2} \\ \text{s.t.}\ \ & \mathbf{f}_m^H \mathbf{h}_{n,i} = 0, \forall n \in \mathcal{M}\backslash m, \forall i \in \mathcal{I}_n^t \end{aligned} \tag{24}$$

Problem (24) is a min-max optimization problem with non-convex objective, we first transform it through the following proposition.

Proposition 2: The problem formulated in (24) is equivalent to the following problem:

$$\begin{aligned} \min_{\mathbf{f}_m} \ \ & \|\mathbf{f}_m\|^2 \\ \text{s.t.}\ \ \text{C1:} \quad & \left|\mathbf{f}_m^H \mathbf{h}_{m,i}\right|^2 \geq 1, \forall i \in \mathcal{I}_m^t \\ \text{C2:} \quad & \mathbf{f}_m^H \mathbf{h}_{n,i} = 0, \forall n \in \mathcal{M}\backslash m, \forall i \in \mathcal{I}_n^t \end{aligned} \tag{25}$$

Proof: See Appendix C. ■

The problem written in (25) is a quadratically constrained quadratic programming (QCQP) problem with non-convex constraints. Thus, we are able to solve it iteratively through SCA. Specifically, at the lth iteration, we derive the optimal $\mathbf{f}_m$ by solving the following problem:

$$\begin{aligned} \min_{\mathbf{f}_m} \ \ & \|\mathbf{f}_m\|^2 \\ \text{s.t.}\ \ \text{C1:} \quad & \|\mathbf{c}_i^{(l)}\|^2 + 2\left(\mathbf{c}_i^{(l)}\right)^T \left(\mathbf{c}_i - \mathbf{c}_i^{(l)}\right) \geq 1, \forall i \in \mathcal{I}_m^t \\ \text{C2:} \quad & \mathbf{c}_i = \left[\text{Re}\left(\mathbf{f}_m^H \mathbf{h}_{m,i}\right), \text{Im}\left(\mathbf{f}_m^H \mathbf{h}_{m,i}\right)\ \right] \\ \text{C3:} \quad & \mathbf{f}_m^H \mathbf{h}_{n,i} = 0, \forall n \in \mathcal{M}\backslash m, \ \ \forall i \in \mathcal{I}_n^t \end{aligned} \tag{26}$$

where the first constraint is obtained by performing the second-order Taylor expansion.

Algorithm 1: SCA Based Receiver Combiner Vector Design

```
Input: $l = 0$, $\varepsilon$
   Randomly initialize $\mathbf{f}_m^\star$
   Set $\mathbf{c}_i^{(0)} = \left[\text{Re}\left(\mathbf{f}_m^{\star H}\mathbf{h}_{m,i}\right), \text{Im}\left(\mathbf{f}_m^{\star H}\mathbf{h}_{m,i}\right)\ \right]$
   repeat
      Solve the convex optimization problem (26)
   until $\sum_{i\in\mathcal{I}_m^t} \|\mathbf{c}_i^{(l+1)} - \mathbf{c}_i^{(l)}\| \leq \varepsilon$
Output: $\mathbf{f}_m$
```

The problem given in (26) is convex, and we solve it through a standard convex optimization solver, e.g., CVX. Besides, we initialize $\mathbf{c}_i^{(0)}$ in a random way, and the iteration stops when the difference of $\mathbf{c}_i$ between two consecutive iterations is less than a preset threshold ε. The algorithm for optimizing $\mathbf{f}_m$ is summarized in Algorithm 1.

B. DEVICE SELECTION

In this subsection, we adopt a greedy device selection algorithm based on (23). Specifically, at the kth iteration with $k = 1, 2, \ldots, \xi I_m$, given the device scheduling policy $\mathcal{I}_m^{(k-1)}$, we first remove each device of $\mathcal{I}_m^{(k-1)}$ and perform Algorithm 1 to derive the corresponding optimal receiver combiner vector, based on which we compute the target value of (23) corresponding to the each removed device. Finally, we find the removed device with the minimum target value, and this device is the one needed to be deleted in this step. We suppose that in step 1, all devices are selected to participate in the FL algorithm. In this algorithm, ξ is the rate of device selection and is treated as a tunable hyperparameter. The algorithm for optimizing $\mathcal{I}_m$ is summarized in Algorithm 2.

Remark 2: From the analysis in Section III-B, we obtain that as the proportion of device selection increases, the amount of data involved in training will increase, leading to the data distribution being closer to the true distribution. This is conducive to the convergence of FL. However, choosing more devices may cause the effects of noise to be amplified, adversely affecting FL performance. On the contrary, when too few devices are selected, the distribution of data involved in training may deviate significantly from the true distribution, which will also bring serious disadvantages to the convergence of FL. Therefore, for scenarios with poor network conditions, the proportion of selected devices should be appropriately reduced, while for scenarios with less data held by devices, it is better to get more devices involved.

Algorithm 2: Greedy Device Selection Algorithm

```
Input: $\mathcal{I}_m^0 = \mathcal{I}_m$, $I_m$, $\mathbf{f}_m^H$, selection rate $\xi$, $k = 1$
  While $k < \xi I_m$ do
    While $i < I_m^{(k)}$ do
      Remove device $i$ from $\mathcal{I}_m^{(k)}$
      Design receiver combiner vector via Algorithm 1
      Compute $obj = d(\mathcal{I}_m^{(k)}, \mathbf{f})$
      Find device with the minimum $obj$ and delete it from $\mathcal{I}_m^{(k)}$
    $k \Leftarrow k + 1$
Output: $\mathcal{I}_m^t$
```

C. COMPUTATION COMPLEXITY

The problem developed in (26) is a second-order cone programming problem, and, therefore, the worst-case complexity during each iteration of the algorithm is $\mathcal{O}(N^3)$. The computational complexity of device selection is thus $\mathcal{O}((2-\xi)\xi I^2)$, where I is the total number of devices.

V. SIMULATION RESULTS

In this section, we evaluate the performance of the proposed joint receiver combiner and device selection algorithm. The simulation set up is introduced in section V-A, and in section V-B, we numerically demonstrate the performance of proposed algorithm for two image classification tasks with different settings and benchmarks described in section V-A.

A. SIMULATION SETUP

For our simulations, we consider a square network area with one BS placed at its center servicing $M \times I$ uniformly distributed devices. We simulate the channel to experience small-scale fading multiplied by large-scale fading, where the small-scale fading follows the standard independent and identically distributed (i.i.d.) Gaussian distribution and the large-scale fading follows the free-space path loss as $G_{BS} G_D (\frac{3*10^8 m/s}{4\pi f_c d_{BD}})^{PL}$, where G_{BS} and G_D are the antenna gains of BS and each device; PL is the free-space path loss coefficient; f_c is the carrier frequency, and d_{BD} is the distance between BS and device. We consider the following two settings on data and device location distribution:

(1) One cluster device with equal data: the $M \times I$ devices are uniformly distributed in a square network area $\{(x, y)$: $-10 < x < 10, -10 < y < 10\}$ and each device have 1000 training samples.

(2) Two cluster device with unequal data: the $M \times I$ devices are uniformly distributed in two square network areas: half devices are in $\{(x, y) : -10 < x < 10, -10 < y < 10\}$, and the other half are in $\{(x, y) : 40 < x < 60, -10 < y < 10\}$. Besides, the number of training samples for each device is unequal. We randomly set half devices with [1500, 2000] training samples and the other half with [300, 500] training samples.

The muti-FL algorithm is simulated by using PyTorch for two image classification tasks on MNIST [44] and FMNIST [45] datasets. Since the sample size and the sample space of the two datasets are the same, we can use the same neural network structure to conduct the classification task. Specifically, each device trains a CNN consisting of two convolution layers with 5×5 kernel size, and each convolution layer is followed by a 2×2 max pool layer, a batch normalization layer, a fully connected layer, a ReLu activation layer, and a softmax output layer. The total number of neurons is 21921. The loss function is the cross-entropy loss.

TABLE 2. Simulation parameters.

Simulation parameters	Value
Number of antennas N	32
Number of FL task M	2
Number of device I_m	40
Rate of device selection ξ	0.5
Device power constant P_0	-10dB
Variance of noise σ_n^2	-100dB
Path loss coefficient PL	3.76
Antenna gain at the BS G_{BS}	5 dBi
Antenna gain at devices G_D	0 dBi
Carrier frequency f_c	915 MHz
Threshold of SCA ε	0.01

For the purpose of comparison, we use the three benchmarks as follows:

(a) Noiseless aggregation: Suppose that the gradient information uploaded by the device to BS is undistorted, which means the BS directly uses the gradient calculated by the device to perform the FedAvg algorithm. Meanwhile, all devices are selected to participate in the FL process.

(b) Optimizes receiver combiner with random device selection: Suppose that devices are randomly selected, and the receiver combiner vector $\mathbf{f}$ is optimized by the SCA algorithm.

(c) OFDMA scheme: Orthogonal frequency division multiple access communication scheme with the same user selection sets as proposed algorithm.

(d) Proposed algorithm: A wireless optimization algorithm that optimizes the receiver combiner vector $\mathbf{f}$ via SCA and the device selection policy via greedy algorithm.

In our simulations, we stipulate that the BS has 64 antennas and 40 devices in each task, and we select half of all devices to participate in the FL process. We perform 1000 FL rounds in each task, and the learning rate of each device is set to be 0.01. The values of parameters used in simulations are listed in Table 2.

B. SIMULATION RESULTS

In this section, we simulate the performance of the proposed algorithm for two image classification tasks with different settings and benchmarks described in Section V-A.

Fig. 2 and Fig. 3 show the test accuracy of MNIST and FMNIST classification tasks with setting 1. From the two figures, we see that the random device selection benchmark can approximate the OFDMA scheme and proposed algorithm, and all of them nearly achieve optimal performance (noiseless aggregation). This is due to the fact that devices

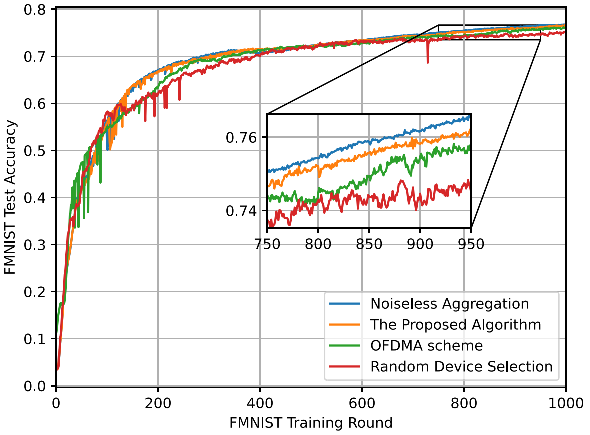

FIGURE 2. Test accuracy of FMNIST classification task under setting 1.

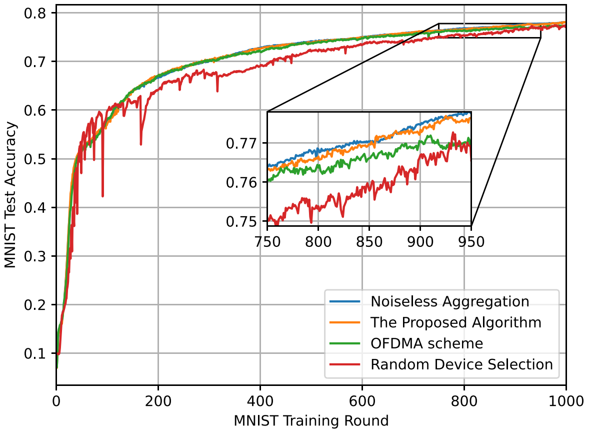

FIGURE 3. Test accuracy of MNIST classification task under setting 1.

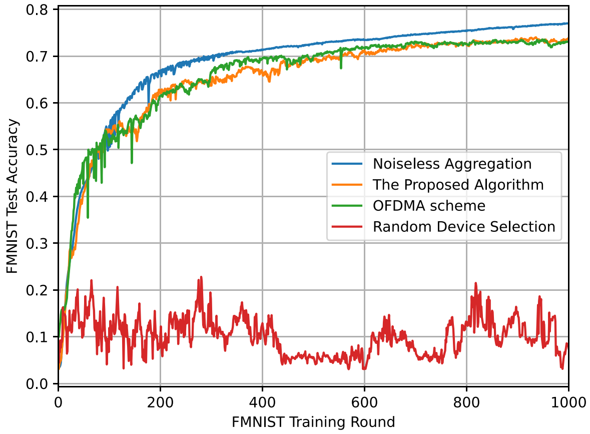

FIGURE 4. Test accuracy of FMNIST classification task under setting 2.

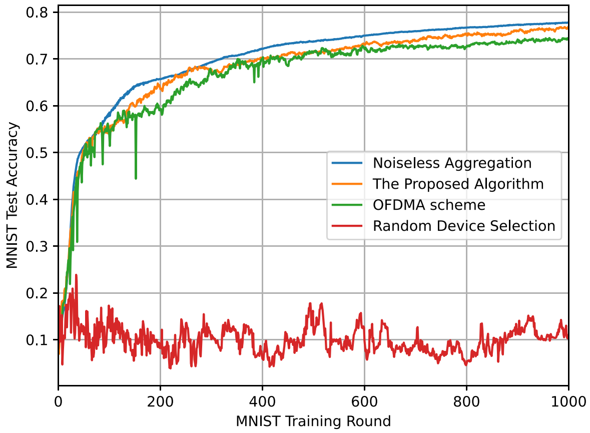

FIGURE 5. Test accuracy of MNIST classification task under setting 2.

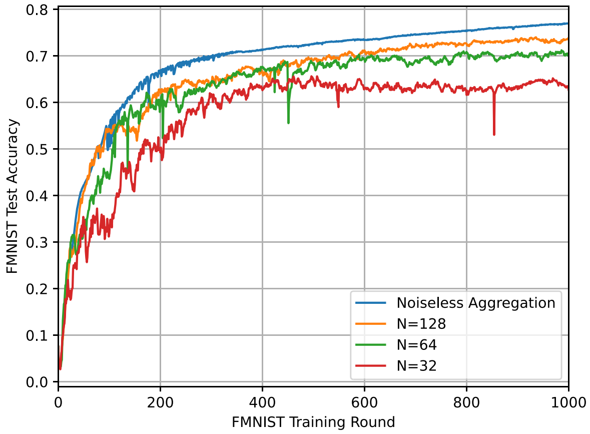

FIGURE 6. Test accuracy of FMNIST classification task versus different numbers of BS receive antennas under setting 2.

are all close to the BS under setting 1, such that there are no significant stragglers. Besides, Fig. 2 and Fig. 3 also show that the proposed algorithm has a better performance than the OFDMA scheme and nearly reach the same performance as the optimal FL. The improvement stems from the fact that the proposed algorithm optimizes receiver combiner vector based on FL convergence speed and error. Above results verifies that the proposed algorithm can not only improve the performance of multiple parallel FL but also effectively reduce the communication costs since the spectrum resources for the OFDMA scheme increase proportionally with the number of devices.

Fig. 4 and Fig. 5 show the test accuracy of MNIST and FMNIST classification tasks under setting 2. From the two figures, we observe that the FL performance is significantly affected when randomly selecting devices because of stragglers. However, due to our greedy device selection algorithm, the performance of the proposed algorithm is still near-optimal. Besides, with the same device selection policy, the OFDMA scheme can also achieve a relatively satisfactory performance.

Fig. 6 and Fig. 7 show the test accuracy of FMNIST and MNIST classification tasks versus different numbers of BS receive antennas under setting 2. From the two figures, we can see that, as the number of BS receive antennas increases, the values of the FL test accuracy increase on both tasks. This is because, as the number of antennas increases, the dimension of the vector $\mathbf{f}$ increases, which helps the SCA algorithm find a better solution. In addition, as the dimension

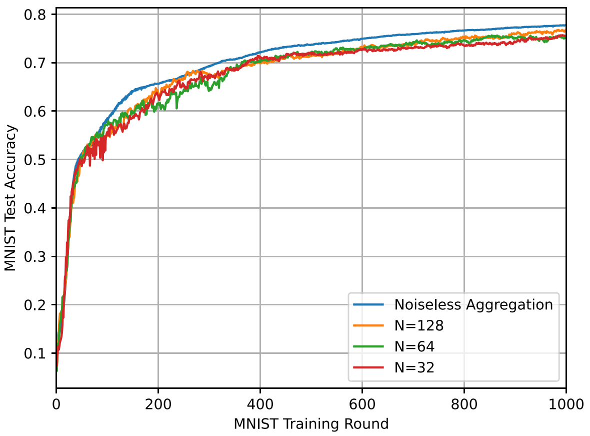

FIGURE 7. Test accuracy of MNIST classification task versus different numbers of BS receive antennas under setting 2.

of the vector $\mathbf{f}$ increases, the dimension of the solution space of the zero-forcing constraint given in (13) increases, which is beneficial for the SCA algorithm to find a better solution as well.

VI. CONCLUSION

In this paper, we developed a multiple FL framework via over-the-air computation in wireless networks. We proposed the zero-forcing receiver combiner to separate the received signals of different computing tasks. Also, we analyzed the convergence of FL under our framework and derived an upper bound on the difference between the loss function and its optimal value, which reveals how the receiver combiner vector and device selection policy affect FL performance. Based on this discovery, we formulated an optimization problem that jointly considers receiver combiner vector design and device selection for improving FL performance. We addressed the problem by alternately optimizing the receiver combiner vector and device selection policy. In particular, we adopted SCA to derive the receiver combiner vector and solve the device scheduling problem with a greedy algorithm. Simulation results show that our proposed framework effectively solves the straggler issue and achieves near-optimal performance for all processed learning tasks.

APPENDIX A
PROOF OF PROPOSITION 1

Let $\mathcal{N}^t$ denote the complement of $\mathcal{I}^t$, so that $\mathcal{I}^t \cup \mathcal{N}^t = \mathcal{I}$, and then the gradient residual in (18) is bounded by the expression derived in (28)^2 at the top of the next page, where $\mathbf{N} = (\mathbf{n}[1], \mathbf{n}[2], \ldots, \mathbf{n}[D]) \in \mathbb{C}^{N \times D}$, and the inequality is achieved by the inequality of arithmetic and geometric means. To minimize the gradient residual, transmitter scalar p_i should satisfy $(K_i - \frac{\mathbf{f}^H \mathbf{h}_i p_i}{\sqrt{\eta} v_i^t}) = 0$ for $i \in \mathcal{I}^t$, thus we get the p_i in (20).

APPENDIX B
PROOF OF THEOREM 1

Since $\boldsymbol{g}_i^t \triangleq \nabla F_i(\boldsymbol{w}^t; \mathcal{D}_i)$ and $F_i(\boldsymbol{w}; \mathcal{D}_i) \triangleq \frac{1}{K_i} \sum_{(\mathbf{x}_i^k, y_i^k) \in \mathcal{D}_i} f(\boldsymbol{w}; \mathbf{x}_i^k, y_i^k)$, the first term at the right side of (28) is bounded as follows

$$\begin{aligned}
&2\left\| \frac{\sum_{i\in\mathcal{N}^t} K_i \boldsymbol{g}_i^t}{\sum_{i\in\mathcal{I}} K_i} - \frac{\sum_{i\in\mathcal{N}^t} K_i \sum_{i\in\mathcal{I}^t} K_i \boldsymbol{g}_i^t}{\sum_{i\in\mathcal{I}} K_i \sum_{i\in\mathcal{I}^t} K_i} \right\|^2 \\
&\leq 2\left[\left\| \frac{\sum_{i\in\mathcal{N}^t} K_i \boldsymbol{g}_i^t}{\sum_{i\in\mathcal{I}} K_i} \right\| + \left\| \frac{\sum_{i\in\mathcal{N}^t} K_i \sum_{i\in\mathcal{I}^t} K_i \boldsymbol{g}_i^t}{\sum_{i\in\mathcal{I}} K_i \sum_{i\in\mathcal{I}^t} K_i} \right\| \right]^2 \\
&\leq 2\left[\frac{\sum_{i\in\mathcal{N}^t} \| K_i \boldsymbol{g}_i^t \|}{\sum_{i\in\mathcal{I}} K_i} + \frac{\sum_{i\in\mathcal{N}^t} K_i \sum_{i\in\mathcal{I}^t} \| K_i \boldsymbol{g}_i^t \|}{\sum_{i\in\mathcal{I}} K_i \sum_{i\in\mathcal{I}^t} K_i} \right]^2 \\
&\leq 8\left(\frac{\sum_{i\in\mathcal{I}} K_i - \sum_{i\in\mathcal{I}^t} K_i}{\sum_{i\in\mathcal{I}} K_i} \right)^2 (\beta_1 + \beta_2 \|\nabla F(\boldsymbol{w}^t)\|^2), \quad (27)
\end{aligned}$$

where the first two inequalities are achieved by the triangle-inequality, and the last one is achieved based on Assumption 4.

Substituting (20) into the last term at the right side of (28) yields

$$\begin{aligned}
&2\left\| \frac{\sum_{i\in\mathcal{I}^t} \left((K_i - \frac{\mathbf{f}^H \mathbf{h}_i p_i}{\sqrt{\eta} v_i^t})(\boldsymbol{g}_i^t - \bar{g}_i^t) \right) + \frac{\mathbf{f}^H \mathbf{N}}{\sqrt{\eta}}}{\sum_{i\in\mathcal{I}^t} K_i} \right\|^2 \\
&= \frac{2}{\eta \left(\sum_{i\in\mathcal{I}^t} K_i \right)^2} \left\| \mathbf{f}^H \mathbf{N} \right\|^2 = \frac{2D \|\mathbf{f}\|^2 \sigma_n^2}{\eta \left(\sum_{i\in\mathcal{I}^t} K_i \right)^2}. \quad (29)
\end{aligned}$$

From (21), we have

$$\frac{2D\sigma_n^2}{\eta \left(\sum_{i\in\mathcal{I}^t} K_i \right)^2} = \frac{2D\sigma_n^2}{\left(\sum_{i\in\mathcal{I}^t} K_i \right)^2} \max_{i\in\mathcal{I}^t} \frac{K_i^2 v_i^2 \|\mathbf{f}\|^2}{P_0 |\mathbf{f}^H \mathbf{h}_i|^2}. \quad (30)$$

Based on (7), we have

$$\begin{aligned}
v_i^2 &= \frac{1}{D_i} \sum_{d=1}^{D_i} (\boldsymbol{g}_i[d] - \bar{g}_i)^2 \\
&= \frac{1}{D_i} \left(\sum_{d=1}^{D_i} \boldsymbol{g}_i^2[d] - \frac{1}{D_i} \left(\sum_{d=1}^{D_i} \boldsymbol{g}_i[d] \right)^2 \right) \\
&\leq \frac{1}{D_i} \sum_{d=1}^{D_i} \boldsymbol{g}_i^2[d] \\
&= \frac{1}{D_i} \|\boldsymbol{g}_i^t\|^2 \\
&= \frac{1}{D_i} \left\| \frac{1}{K_i} \sum_{(\mathbf{x}_i^k, y_i^k) \in \mathcal{D}_i} \nabla f\left(\boldsymbol{w}^t; \mathbf{x}_i^k, y_i^k \right) \right\|^2 \\
&\leq \frac{1}{D_i} \left(\beta_1 + \beta_2 \|\nabla F\left(\boldsymbol{w}^t\right)\|^2 \right), \quad (31)
\end{aligned}$$

where the last inequality is derived based on Assumption 4.

[2] In this expression, vector $\boldsymbol{g}$ subtracting scalar $\bar{g}$ means each entitiy of the vector $\boldsymbol{g}$ subtracting scalar $\bar{g}$. Besides, addition and subtraction operations between vectors and scalars involved in other formulas in this paper also obey the above principle.

$$\begin{aligned}
\|\boldsymbol{e}^t\|^2 &= \left\| \nabla F\left(\boldsymbol{w}^t\right) - \frac{\lambda}{\sum_{i\in\mathcal{I}} K_i}\hat{\boldsymbol{r}}^t \right\|^2 \\
&= \left\| \frac{\sum_{i\in\mathcal{I}} K_i \boldsymbol{g}_i^t}{\sum_{i\in\mathcal{I}} K_i} - \frac{\frac{1}{\sqrt{\eta}}\mathbf{f}^H\mathbf{y}^t + \bar{g}^t}{\sum_{i\in\mathcal{I}^t} K_i} \right\|^2 \\
&= \left\| \frac{\sum_{i\in\mathcal{I}^t} K_i \boldsymbol{g}_i^t + \sum_{i\in\mathcal{N}^t} K_i \boldsymbol{g}_i^t}{\sum_{i\in\mathcal{I}} K_i} - \frac{\frac{1}{\sqrt{\eta}}\mathbf{f}^H \sum_{i\in\mathcal{I}^t} \mathbf{h}_i p_i\left(\frac{\boldsymbol{g}_i^t-\bar{g}_i^t}{v_i^t}\right) + \frac{\mathbf{f}^H\mathbf{N}}{\sqrt{\eta}} + \sum_{i\in\mathcal{I}^t} K_i\bar{g}_i^t}{\sum_{i\in\mathcal{I}^t} K_i} \right\|^2 \\
&= \left\| \frac{\sum_{i\in\mathcal{N}^t} K_i \boldsymbol{g}_i^t}{\sum_{i\in\mathcal{I}} K_i} - \frac{\sum_{i\in\mathcal{I}^t}\left(\frac{(\sum_{i\in\mathcal{I}} K_i)\mathbf{f}^H\mathbf{h}_i p_i}{\sqrt{\eta}v_i^t} - \left(\sum_{i\in\mathcal{I}^t} K_i\right)K_i\right)\boldsymbol{g}_i^t}{\sum_{i\in\mathcal{I}} K_i \sum_{i\in\mathcal{I}^t} K_i} + \frac{\sum_{i\in\mathcal{I}^t}\left(\frac{\mathbf{f}^H\mathbf{h}_i p_i}{\sqrt{\eta}v_i^t} - K_i\right)\bar{g}_i^t + \frac{\mathbf{f}^H\mathbf{N}}{\sqrt{\eta}}}{\sum_{i\in\mathcal{I}^t} K_i} \right\|^2 \\
&= \left\| \frac{\sum_{i\in\mathcal{N}^t} K_i \boldsymbol{g}_i^t}{\sum_{i\in\mathcal{I}} K_i} - \frac{\sum_{i\in\mathcal{N}^t} K_i \sum_{i\in\mathcal{I}^t} K_i \boldsymbol{g}_i^t}{\sum_{i\in\mathcal{I}} K_i \sum_{i\in\mathcal{I}^t} K_i} + \frac{\sum_{i\in\mathcal{I}^t}\left(\left(K_i - \frac{\mathbf{f}^H\mathbf{h}_i p_i}{\sqrt{\eta}v_i^t}\right)\left(\boldsymbol{g}_i^t - \bar{g}_i^t\right)\right) + \frac{\mathbf{f}^H\mathbf{N}}{\sqrt{\eta}}}{\sum_{i\in\mathcal{I}^t} K_i} \right\|^2 \\
&\le \left[\left\| \frac{\sum_{i\in\mathcal{N}^t} K_i \boldsymbol{g}_i^t}{\sum_{i\in\mathcal{I}} K_i} - \frac{\sum_{i\in\mathcal{N}^t} K_i \sum_{i\in\mathcal{I}^t} K_i \boldsymbol{g}_i^t}{\sum_{i\in\mathcal{I}} K_i \sum_{i\in\mathcal{I}^t} K_i} \right\| + \left\| \frac{\sum_{i\in\mathcal{I}^t}\left(\left(K_i - \frac{\mathbf{f}^H\mathbf{h}_i p_i}{\sqrt{\eta}v_i^t}\right)\left(\boldsymbol{g}_i^t - \bar{g}_i^t\right)\right) + \frac{\mathbf{f}^H\mathbf{N}}{\sqrt{\eta}}}{\sum_{i\in\mathcal{I}^t} K_i} \right\| \right]^2 \\
&\le 2\left\| \frac{\sum_{i\in\mathcal{N}^t} K_i \boldsymbol{g}_i^t}{\sum_{i\in\mathcal{I}} K_i} - \frac{\sum_{i\in\mathcal{N}^t} K_i \sum_{i\in\mathcal{I}^t} K_i \boldsymbol{g}_i^t}{\sum_{i\in\mathcal{I}} K_i \sum_{i\in\mathcal{I}^t} K_i} \right\|^2 + 2\left\| \frac{\sum_{i\in\mathcal{I}^t}\left(\left(K_i - \frac{\mathbf{f}^H\mathbf{h}_i p_i}{\sqrt{\eta}v_i^t}\right)\left(\boldsymbol{g}_i^t - \bar{g}_i^t\right)\right) + \frac{\mathbf{f}^H\mathbf{N}}{\sqrt{\eta}}}{\sum_{i\in\mathcal{I}^t} K_i} \right\|^2 \qquad (28)
\end{aligned}$$

Combining (29), (30) and (31), we obtain

$$\begin{aligned}
&2\left\| \frac{\sum_{i\in\mathcal{I}^t}\left(\left(K_i - \frac{\mathbf{f}^H\mathbf{h}_i p_i}{\sqrt{\eta}v_i^t}\right)\left(\boldsymbol{g}_i^t - \bar{g}_i^t\right)\right) + \frac{\mathbf{f}^H\mathbf{N}}{\sqrt{\eta}}}{\sum_{i\in\mathcal{I}^t} K_i} \right\|^2 \\
&\le \frac{2\sigma_n^2}{\left(\sum_{i\in\mathcal{I}^t} K_i\right)^2}\left(\beta_1 + \beta_2\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|^2\right)\max_{i\in\mathcal{I}^t}\frac{K_i^2\|\mathbf{f}\|^2}{P_0|\mathbf{f}^H\mathbf{h}_i|^2}. \qquad (32)
\end{aligned}$$

Given (19), (27) and (32), we have

$$\begin{aligned}
&\mathbb{E}\left[F\left(\boldsymbol{w}^{t+1}\right) - F\left(\boldsymbol{w}^\star\right)\right] \\
&\le \psi\mathbb{E}\left[F\left(\boldsymbol{w}^t\right) - F\left(\boldsymbol{w}^\star\right)\right] + \frac{\beta_1\psi}{L}, \qquad (33)
\end{aligned}$$

where ψ is defined in Theorem 1.

Applying (33) recursively, we complete the proof.

APPENDIX C
PROOF OF PROPOSITION 2

By introducing an auxiliary variable $\tau = \min_{i\in\mathcal{I}_m^t}|\mathbf{f}_m^H\mathbf{h}_{m,i}|^2$, the problem developed in (24) can be rewritten as

$$\begin{aligned}
\min_{\mathbf{f}_m} \quad & \|\mathbf{f}_m\|^2/\tau \\
\text{s.t.} \quad & |\mathbf{f}_m^H\mathbf{h}_{m,i}|^2 \ge \tau, \forall i \in \mathcal{I}_m^t \qquad (34)
\end{aligned}$$

Then introducing a new optimization variable $\tilde{\mathbf{f}}_m = \mathbf{f}_m/\sqrt{\tau}$, the above problem is equivalently transferred to

$$\begin{aligned}
\min_{\tilde{\mathbf{f}}_m} \quad & \|\tilde{\mathbf{f}}_m\|^2 \\
\text{s.t.} \quad & |\tilde{\mathbf{f}}_m^H\mathbf{h}_{m,i}|^2 \ge 1, \forall i \in \mathcal{I}_m^t \qquad (35)
\end{aligned}$$

which completes the proof.

REFERENCES

[1] M. Chiang and T. Zhang, “Fog and IoT: An overview of research opportunities,” *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. NeurIPS*, vol. 25. Lake Tahoe, NV, USA, 2012, pp. 84–90.

[3] R. Collobert and J. Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” in *Proc. 25th Int. Conf. Mach. Learn.*, New York, NY, USA, 2008, pp. 160–167.

[4] G. Hinton *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[5] 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.

[6] 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. Ft. Lauderdale, FL, USA, Apr. 2017, pp. 1273–1282.

[7] K. Bonawitz *et al.*, “Towards federated learning at scale: System design,” 2019, *arXiv:1902.01046*.

[8] P. Kairouz *et al.*, “Advances and open problems in federated learning,” 2019, *arXiv:1912.04977*.

[9] 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.

[10] 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.

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

[12] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. Vincent Poor, "Update aware device scheduling for federated learning at the wireless edge," in *Proc. IEEE Int. Symp. Inf. Theory*, Los Angeles, CA, USA, Jun. 2020, pp. 2598–2603.

[13] J. Chen, X. Pan, R. Monga, S. Bengio, and R. Jozefowicz, "Revisiting distributed synchronous SGD," 2016, *arXiv:1604.00981*.

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

[15] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2016, *arXiv:1610.05492*.

[16] 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. Long Beach, CA, USA, Dec. 2017, pp. 1707–1718.

[17] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "UVeQFed: Universal vector quantization for federated learning," *IEEE Trans. Signal Process.*, vol. 69, pp. 500–514, Dec. 2021.

[18] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," 2017, *arXiv:1704.05021*.

[19] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, "The convergence of sparsified gradient methods," in *Proc. NeurIPS*, vol. 31. Montreal, QC, Canada, 2018, pp. 5977–5987.

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

[21] 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.

[22] 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.

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

[24] S.-W. Jeon and B. C. Jung, "Opportunistic function computation for wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4045–4059, Jun. 2016.

[25] S. Sigg, P. Jakimovski, and M. Beigl, "Calculation of functions on the RF-channel for IoT," in *Proc. 3rd IEEE Int. Conf. Internet Things*, Wuxi, China, Oct. 2012, pp. 107–113.

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

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

[28] G. Zhu, Y. Du, D. Gündüz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning," in *Proc. IEEE Global Commun. Conf.*, Taipei, Taiwan, 2020, pp. 1–6.

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

[30] 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.

[31] 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.

[32] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5115–5128, Aug. 2021.

[33] X. Cao, G. Zhu, J. Xu, and S. Cui, "Optimized power control for over-the-air federated edge learning," in *Proc. IEEE Int. Conf. Commun.*, Montreal, QC, Canada, 2021, pp. 1–6.

[34] 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.

[35] H. Guo, A. Liu, and V. K. N. Lau, "Analog gradient aggregation for federated learning over wireless networks: Customized design and convergence analysis," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 197–210, Jan. 2021.

[36] 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.

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

[38] Z. Wang, Y. Zhou, Y. Shi, and W. Zhuang, "Interference management for over-the-air federated learning in multi-cell wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, p. 1, Aug. 2022.

[39] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang, "Decomposition by partial linearization: Parallel optimization of multi-agent systems," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 641–656, Feb. 2014.

[40] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," 2016, *arXiv:1610.02527*.

[41] S. Wang *et al.*, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, Jun. 2019.

[42] 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.

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

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

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

GAOXIN SHI (Student Member, IEEE) was born in in Shandong, China. He received the B.Sc. degree in automation from the School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, China, in 2020. He is currently pursuing the M.S. degree with Shandong University, Jinan, China. His main research interests include the communication efficiency and robustness of federated leaning systems.

SHUAISHUAI GUO (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.

JIA YE (Student Member, IEEE) was born in Chongqing, China. She received the B.Sc. degree in communication engineering from Southwest University, Chongqing, in 2018, and the M.S. degree from the King Abdullah University of Science and Technology, Saudi Arabia, in 2020, where she is currently pursuing the Ph.D. degree. Her main research interests include the performance analysis and modeling of wireless/wireless communication systems.

NASIR SAEED (Senior Member, IEEE) received the B.Sc. degree in telecommunication from the University of Engineering and Technology, Peshawar, Pakistan, in 2009, the M.Sc. degree in satellite navigation from the Polito di Torino, Italy, in 2012, and the Ph.D. degree in electronics and communication engineering from Hanyang University, Seoul, South Korea, in 2015. He was an Assistant Professor with the Department of Electrical Engineering, Gandhara Institute of Science and IT, Peshawar, from August 2015 to September 2016. He has worked as an Assistant Professor with IQRA National University, Peshawar, from October 2016 to July 2017. From July 2017 to December 2020, he was a Postdoctoral Research Fellow with the Communication Theory Laboratory, King Abdullah University of Science and Technology. He is currently an Associate Professor with the Department of Electrical Engineering, Northern Border University, Arar, Saudi Arabia. His current research interests include cognitive radio networks, underwater wireless communications, aerial networks, dimensionality reduction, and localization.

SHUPING DANG (Member, IEEE) received the first B.Eng. degree (with First Class Hons.) in electrical and electronic engineering from the University of Manchester and the second B.Eng. degree in electrical engineering and automation from Beijing Jiaotong University in 2014 via a joint ‘2+2’ dual-degree program, and the D.Phil. degree in engineering science from the University of Oxford in 2018. He joined with the Research and Development Center, Huanan Communication Company Ltd., after graduating with the University of Oxford and worked as a Postdoctoral Fellow with the Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology. He is currently a Lecturer with the Department of Electrical and Electronic Engineering, University of Bristol. His research interests include 6G communications, wireless communications, wireless security, and machine learning for communications.