Over-the-Air Federated Multi-Task Learning Over MIMO Multiple Access Channels

Chenxi Zhong\textsuperscript{\textdagger}, Graduate Student Member, IEEE, Huiyuan Yang\textsuperscript{\textdagger}, and Xiaojun Yuan\textsuperscript{\textdagger}, Senior Member, IEEE

Abstract—With the explosive growth of data and wireless devices, federated learning (FL) over wireless medium has emerged as a promising technology for large-scale distributed intelligent systems. Yet, the urgent demand for ubiquitous intelligence will generate a large number of concurrent FL tasks, which may seriously aggravate the scarcity of communication resources. By exploiting the analog superposition of electromagnetic waves, over-the-air computation (AirComp) is an appealing solution to alleviate the burden of communication required by FL. However, sharing frequency-time resources in over-the-air computation inevitably brings about the problem of inter-task interference, which poses a new challenge that needs to be appropriately addressed. In this paper, we study over-the-air federated multi-task learning (OA-FMTL) over the multiple-input multiple-output (MIMO) multiple access (MAC) channel. We propose a novel model aggregation method for the alignment of local gradients of different devices, which alleviates the straggler problem in over-the-air computation due to the channel heterogeneity. We establish a communication-learning analysis framework for the proposed OA-FMTL scheme by considering the spatial correlation between devices, and formulate an optimization problem for the design of transceiver beamforming and device selection. To solve this problem, we develop an algorithm by using alternating optimization (AO) and fractional programming (FP), which effectively mitigates the impact of inter-task interference on the FL learning performance. We show that due to the use of the new model aggregation method, device selection is no longer essential, thereby avoiding the heavy computational burden involved in selecting active devices. Numerical results demonstrate the validity of the analysis and the superb performance of the proposed scheme.

Index Terms—Federated multi-task learning, over-the-air model aggregation, multiple-input multiple-output multiple access channel, alternating optimization, fractional programming.

I. INTRODUCTION

In recent years, the explosive increase of wireless data has promoted widespread applications of artificial intelligence in many fields, such as computer vision [2] and natural language processing [3]. To exploit the diversity of wireless data, the conventional centralized machine learning (ML) paradigm requires edge devices to upload their local data to a central parameter server (PS) for joint training. However, data uploading brings about huge costs in communication resources and potentially threatens user privacy. To tackle these challenges, federated learning (FL) has been proposed as a promising substitute [4]. In FL, the PS first broadcasts the global model parameters to the selected edge devices. Subsequently, each selected device calculates a local gradient based on its local dataset and uploads the gradient to the PS. The global model is then updated by the PS based on the received local gradients. Clearly, by replacing data sharing with gradient sharing, FL reduces communication costs and protects user privacy.

Despite the appealing advantages, the huge uplink communication cost is still a critical bottleneck for FL, especially when uploading local updates through wireless channels. In the literature, there are two approaches for the aggregation of local updates in the uplink, i.e., the aggregation in the digital domain [5], [6], [7], [8], [9], [10], [11], and the aggregation in the analog domain, also known as (a.k.a.) over-the-air computation (AirComp) [12], [13], [14], [15], [16], [17], [18]. In digital aggregation, the devices encode their local updates in the digital domain and send their encoded signals to the PS; the PS decodes the individual local update of each device and then aggregates these local updates to acquire the global update. On the other hand, in over-the-air FL (OA-FL), edge devices transmit their local updates by using shared radio resources, and the local updates are aggregated over-the-air due to the superposition property of electromagnetic waves. Compared with the digital approach, the communication resource required by AirComp does not increase with the number of devices [13], which largely relieves the communication bottleneck in FL. This appealing property motivates us to explore the OA-FL scheme. Pioneering work has confirmed that OA-FL has strong noise tolerance [14] and reduces latency substantially compared with the schemes based on conventional orthogonal multiple access (OMA) protocols [15], [16], [17], [18].

In spite of the abovementioned benefits, introducing AirComp into FL uplink brings about the so-called straggler\textsuperscript{1} problem [16], [17]. To aggregate the local gradients at the PS,

\textsuperscript{1}In computer science literature, the word “straggler” usually refers to a device with low computation capacity. But here a straggler refers to a device with poor channel condition.
the devices with better channel conditions have to lower their transmitting powers to align themselves with the stragglers (the devices with the worst channel conditions); see e.g. the state-of-the-art OA-FL literature [17], [19], [20]. Clearly, since this strict-alignment approach potentially reduces the aggregation signal-to-noise ratio (SNR), it conceivably leads to a degradation of the FL performance. Refs. [16] and [17] propose to exclude the stragglers from model aggregation to alleviate this problem. However, this exclusion of devices reduces the available dataset for FL training, thereby deteriorating the FL performance. Thus, developing more efficient approaches to deal with the straggler problem is highly desirable.

Another problem brought by the over-the-air FL uplink appears in the situation of multi-task. Specifically, when multiple tasks are performed simultaneously at the wireless edge,\(^2\) the communication cost of this federated multi-task learning system will be further exacerbated, which to a greater extent necessitates the use of AirComp. However, introducing AirComp in this multi-task situation implies that the FL tasks share time-frequency resources for gradients uploading, which inevitably introduces inter-task interference. In this paper, we attempt to suppress the inter-task interference by using the multi-antenna technique.

Moreover, the local gradients of FL exhibit strong correlations temporally (over communication rounds) and spatially (among devices) [22], [23]. The temporal correlation has been exploited to improve the uplink efficiency of OA-FL in [24], whereas an efficient use of the spatial correlation has not been well explored. The existing approaches in OA-FL [19], [25] ignore this correlation and assume spatially independent local gradients in the system optimization, potentially leading to a substantial performance loss. Developing a spatial-correlation-aware OA-FL system optimization scheme will be a non-trivial step to exploit the spatial correlation of the local gradients.

In this paper, we address the above three challenges by developing a novel over-the-air federated multi-task learning (OA-FMTL) scheme where multiple FL tasks are trained simultaneously over a multiple-input multiple-output (MIMO) multiple access (MAC) channel. The MIMO MAC channel consists of a multi-antenna central PS and a number of multi-antenna edge devices. These FL tasks share time-frequency resources and thus generally interfere with each other in model uploading. We establish a communication-learning analysis framework for the considered OA-FMTL scheme. Based on the framework, we analyze the convergence of the OA-FMTL scheme and formulate a transceiver beamforming problem for learning performance enhancement and inter-task interference suppression. We propose a low-complexity algorithm based on alternating optimization (AO) and fractional programming (FP) to optimize the transmit and receive beamforming vectors. The main novelties of our approach are listed as follows:

- **The OA-FMTL analysis framework:** We establish a communication-learning analysis framework for the considered OA-FMTL scheme. Based on the analysis results, we formulate the transceiver beamforming problem and develop a low-complexity solution to the problem.

- **Inter-task interference suppression:** To the best of our knowledge, this is the first attempt to solve the problem of inter-task interference suppression in OA-FMTL utilizing the MIMO technique.

- **Spatial-correlation-aware design:** We establish a probability model to capture the gradient correlation among the devices, allowing us to design the OA-FMTL scheme with the awareness of the spatial correlation. We show that the awareness of the spatial correlation brings substantial improvement in learning performance.

- **Soft straggler alignment:** We propose a misalignment-tolerant aggregation approach for the gradient aggregation at the PS. This approach corrects the stereotype that gradients need to be strictly aligned without going to the other extreme, i.e., to drop out the stragglers, thereby significantly relieving the performance degradation due to the straggler problem.

Extensive numerical results demonstrate that our proposed OA-FMTL scheme achieves significant performance improvements than the existing schemes. Furthermore, the test accuracies of the proposed scheme are very close to those of the ideal error-free benchmarks even in the presence of severe inter-task interference, demonstrating the effectiveness of our proposed scheme.

The remainder of this paper is organized as follows. In Section II, the FL model, the MIMO MAC channel model, the communication system, and the over-the-air model aggregation method are described. In Section III, we analyze the learning performance of the proposed OA-FMTL scheme. In Section IV, we formulate the optimization problem that minimizes the total training loss of the FL tasks and propose algorithms to jointly optimize transmit and receive beamforming vectors, as well as device selection. In Section V, numerical results are presented to evaluate the proposed scheme. Finally, we draw the conclusions in Section VI.

**Notation:** We use \(\mathbb{R}\) and \(\mathbb{C}\) to denote the real and complex number sets, respectively. We denote scalars in italic type, vectors in straight bold small letters, and matrices in straight bold capital letters. \((\cdot)^{\dagger}\), \((\cdot)^{T}\), and \((\cdot)^{\mathsf{H}}\) are used to denote the conjugate, the transpose, and the conjugate transpose, respectively. We use \(s[d]\) to denote the \(d\)-th entry of vector \(s\), \(s(1:D)\) to denote a sub-vector of \(s\) with entries indexed from 1 to \(D\), \(\mathcal{CN}(\mu, \sigma^2)\) to denote the circularly-symmetric complex normal distribution with mean \(\mu\) and covariance \(\sigma^2\), and \(|S|\) to denote the cardinality of set \(S\). We use \(E[\cdot]\) to denote the expectation operator w.r.t. the gradients and the channel noise, and \(E[\cdot]\) to denote the expectation operator w.r.t. the channel matrices, the gradients, and the channel noise. \(I_N\), \(I_{N \times M}\), and \(0_{N \times M}\) are used to denote the \(N \times N\) identity matrix, the \(N \times M\) all-one matrix, and the \(N \times M\) all-zero matrix, respectively. We use \(\|\cdot\|\) to denote the \(l_2\)-norm. \([K]\) denotes the set \(\{k|1 \leq k \leq K\}\).
II. SYSTEM MODEL

In this paper, we consider a $K$-task OA-FMTL system with one central PS and $M$ wireless devices, where $M_k$ wireless devices are collaboratively training an identical model for task $k$, $k \in [K]$, and $M = \sum_{k=1}^{K} M_k$. A two-task OA-FMTL system is illustrated in Fig. 1. In the following, we introduce the FL system, the underlying communication system, the over-the-air model aggregation, and OA-FMTL framework, sequentially.

A. Federated Learning System

We start with the description of the $K$-task FL system. Let $A_k$ denote the training dataset of the $k$-th FL task (which is stored on the $M_k$ devices). Let $Q_k$ be the cardinality of $A_k$. Each FL task $k$ has an empirical loss function based on $A_k$, given by

$$F_k(w_k) \triangleq \frac{1}{Q_k} \sum_{n=1}^{Q_k} f_k(w_k; \xi_{k,n}), \quad \forall k \in [K],$$  

(1)

where $w_k \in \mathbb{R}^{D_k}$ denotes the parameter vector of task $k$, $\xi_{k,n} \in A_k$, $f_k(w_k; \xi_{k,n})$ denotes the sample-wise loss function with respect to (w.r.t.) $\xi_{k,n}$. Let subscript $(k,i)$ denote the index of the $i$-th device in the $k$-th group, $k \in [K]$, $i \in [M_k]$.

We assume that the local training datasets of each device $(k,i)$ in group $k$ are independently and randomly drawn from $A_k$, $k \in [K]$. Let $Q_{k,i}$ denote the sample size of the $(k,i)$-th device. Naturally, we have $Q_k = \sum_{i=1}^{M_k} Q_{k,i}$. Then $F_k(w_k)$ can be rewritten as

$$F_k(w_k) = \frac{1}{Q_k} \sum_{i=1}^{M_k} Q_{k,i} F_{k,i}(w_k), \quad \forall k \in [K],$$  

(2)

where $F_{k,i}(w_k) = \frac{1}{Q_{k,i}} \sum_{n=1}^{Q_{k,i}} f_k(w_k; \xi_{k,i,n}) / Q_{k,i}$ denotes the local loss function of the $(k,i)$-th device with $\xi_{k,i,n}$ being the $n$-th training sample on the $(k,i)$-th device. The overall loss function of the considered OA-FMTL is given by

$$F(w) \triangleq \sum_{k=1}^{K} F_k(w_k),$$  

(3)

where $w \triangleq [w_T^1, \ldots, w_T^K]^T$.

Fig. 1. An illustration of the OA-FMTL with two tasks.

OA-FMTL aims to minimize $F(w)$ by separately minimize each $F_k(w_k)$ using gradient descent (GD) [4]. We now focus on the training of the $k$-th task. To be specific, at the $t$-th communication round, the training of each FL task $k$ consists of the following five steps:

- **Device selection**: The parameter server (PS) selects a subgroup $M_k$ of group $k$ to participate in the updating.
- **Global model broadcasting**: The PS broadcasts global model $w_k(t)$ to the selected devices through error-free links.
- **Local gradients computing**: Each selected device computes its gradient based on the local training dataset, i.e., to compute the local gradient $g_{k,i}^{(t)}$ by

$$g_{k,i}^{(t)} = \nabla F_{k,i}(w_k^{(t)}) = \frac{1}{Q_{k,i}} \sum_{n=1}^{Q_{k,i}} \nabla f_k(w_k^{(t)}; \xi_{k,i,n}), \quad \forall i \in M_k.$$  

(4)

- **Local gradients uploading**: Selected devices upload their local gradients to the PS.
- **Global model updating**: The PS updates the model $w_k(t)$ by

$$w_k(t+1) = w_k(t) - \eta_k g_k^{(t)},$$  

(5)

where $\eta_k \in \mathbb{R}$ is the learning rate of task $k$.

B. MIMO MAC Channel

We now introduce the wireless channel model of OA-FMTL. Following the common practice in OA-FL [13], [26], [27], we assume the global model is broadcast to the devices via error-free links and the local gradient uploading is synchronized among all the edge devices. Recall that OA-FMTL is composed of a PS (i.e., a base station) and $M$ edge devices. We assume the PS and each edge device are respectively equipped with $N_R$ and $N_T$ antennas, which leads to a multiple-input multiple-output (MIMO) multiple access (MAC) channel. We further assume a block-fading channel model, where the channel state keeps invariant during the step of local gradients uploading. In addition, we assume that the channel statistics keep invariant during the entire training process, and the channel state varies independently from round to round.

- In practice, there exist communication errors due to inter-task interference and channel noise. In this case, the schemes proposed by [5], [6] can be used for downlink communication. This is, however, not the focus of this paper, and thus is omitted here for brevity.
- The symbol-level synchronization between the devices can be realized by using the existing techniques, e.g., the timing advance (TA) mechanism for uplink synchronization in 4G Long Term Evolution (LTE) [28].
- We assume that perfect CSI is available at the PS. Studies on CSI acquisition over MIMO MAC channels can be found, e.g., in [5], [6], [29], [30].
Fig. 2. An illustration of the over-the-air gradient aggregation.

We next focus on the local gradient uploading process. Let \( C \) denote the total number of channel uses in each communication round. Then, at the \( t \)-th communication round, the received signal matrix at the PS is denoted by

\[
Y(t) = \sum_{k=1}^{K} \sum_{i \in M_k} H_{(k,i)}^{(t)} X_{(k,i)}^{(t)} + N(t), \tag{7}
\]

where \( H_{(k,i)}^{(t)} \in \mathbb{C}^{N_R \times N_T} \) denotes the channel matrix between the \( (k,i) \)-th device and the PS, \( X_{(k,i)}^{(t)} \) is the transmitted signal from the \( (k,i) \)-th device, \( N(t) \) is an additive white Gaussian noise (AWGN) matrix whose entries are independently drawn from \( \mathcal{CN}(0,\sigma^2) \), and \( y(t) \) denotes the received signal at the PS at the \( t \)-th channel use. Note that the two sums in (7) imply that all the edge devices are allowed to participate in transmit training using shared radio resources. The superposition characteristic of electromagnetic waves can be utilized for signal aggregation, as suggested by AirComp, while this non-orthogonal transmission introduces inter-task interference. In the next subsection, we show how to aggregate local gradients over the air using the underlying communication channel in (7).

C. Over-the-Air Model Aggregation

In this subsection, we introduce the approach of over-the-air gradient aggregation based on the underlying communication system in Section II-B. As illustrated in Fig. 2, the local gradients \( \{g_{(k,i)}^{(t)}\}_{k \in [K], i \in M_k} \) are first pre-processed to yield signals \( \{X_{(k,i)}^{(t)}\}_{k \in [K], i \in M_k} \), which are then transmitted to the PS via the underlying communication channel in (7). After the PS receives \( Y(t) \), it separately performs post-processing to obtain \( \{g_{(k,i)}^{(t)}\}_{k=1}^{K} \), which are estimates of the desired aggregated gradients \( \{g_{(k,i)}^{(t)}\}_{k=1}^{K} \). In the following, we introduce the pre-processing and post-processing steps.

1) Pre-Processing: This step performs separately on each participating device, and aims to generate appropriate transmitting signals. Specifically, we first normalize \( g_{(k,i)}^{(t)} \) to \( \tilde{g}_{(k,i)}^{(t)} \in \mathbb{R}^D \) by

\[
\tilde{g}_{(k,i)}^{(t)}[d] = (g_{(k,i)}^{(t)} - \bar{g}_{(k,i)}^{(t)}) / \sqrt{v_{(k,i)}^{(t)}}, \tag{8}
\]

where \( \bar{g}_{(k,i)}^{(t)} = \frac{1}{D} \sum_{d=1}^{D} g_{(k,i)}^{(t)}[d] \) and \( v_{(k,i)}^{(t)} = \frac{1}{D} \sum_{d=1}^{D} \left| g_{(k,i)}^{(t)}[d] - \bar{g}_{(k,i)}^{(t)} \right|^2 \). Following the common practice \cite{17, 31}, we assume \( \{g_{(k,i)}^{(t)}, v_{(k,i)}^{(t)}\}_{k \in [K], t \in [T]} \) are transmitted to the PS via error-free links.

To match the complex communication system in (7), we then modulate (analog modulation) the normalized gradient \( \tilde{g}_{(k,i)}^{(t)} \) to a complex data vector \( r_{(k,i)}^{(t)} \), i.e.,

\[
r_{(k,i)}^{(t)} \triangleq \tilde{g}_{(k,i)}^{(t)} \left( 1 : \frac{D}{2} \right) + j\tilde{g}_{(k,i)}^{(t)} \left( \frac{D}{2} : D \right) \in \mathbb{C}^C, \tag{9}
\]

where \( C = D/2 \). We transmit \( r_{(k,i)}^{(t)} \) with \( C \) times of channel use (one element of \( r_{(k,i)}^{(t)} \) for one time of channel use). Specifically, the transmitted signal is given by

\[
X_{(k,i)}^{(t)} \triangleq u_{(k,i)}^{(t)} r_{(k,i)}^{(t) \top} \in \mathbb{C}^{N_T \times C}, \tag{10}
\]

where \( u_{(k,i)}^{(t)} \in \mathbb{C}^{N_T} \) is the transmit beamforming vector of the \( (k,i) \)-th device. Recall that \( X_{(k,i)}^{(t)} = [x_{(k,i)}^{(t)}[1], \ldots, x_{(k,i)}^{(t)}[C]] \). Thus we have \( x_{(k,i)}^{(t)}[c] = r_{(k,i)}^{(t)\top} u_{(k,i)} \) denoting the transmitted signal of the \( (k,i) \)-th device at the \( t \)-th channel use, satisfying the following power constraint:

\[
\mathbb{E} \left[ \|x_{(k,i)}^{(t)}[c]\|^2 \right] = \|u_{(k,i)}^{(t)}\|^2 \leq P_0, \tag{11}
\]

where the equality follows the normalization in (8) (implying \( \mathbb{E} \left[ \|r_{(k,i)}^{(t)\top}[c]\|^2 \right] = 2 \).

2) Post-Processing: The received signals from the \( N_T \) receive antennas are combined separately using \( K \) receive beamforming vectors to obtain \( f_k^{(t)} \), i.e.,

\[
f_k^{(t)} = \zeta_k \left( f_k^{(t)\top} Y(t) \right) \in \mathbb{C}^{K \times N} = \zeta_k \left( \sum_{i \in M_k} r_{(k,i)}^{(t)\top} (H_{(k,i)}^{(t)})^\top u_{(k,i)} \right) + \sum_{l \neq k} \sum_{i \in M_l} r_{(l,i)}^{(t)\top} (H_{(l,i)}^{(t)})^\top u_{(l,i)} + N^{(t)\top} f_k^{(t)}, \forall k \in [K], \tag{12}
\]

where \( \zeta_k \in \mathbb{R} \) is a weighting factor and \( f_k \in \mathbb{C}^{N_N} \) is the normalized receive beamforming vector of task \( k \) with

\[\text{D} \text{D} \text{D}\]
\[ \| \mathbf{e}_k \|_2 = 1. \] Estimates of the desired aggregated gradients \( \{ \hat{g}_k(t)^{(t)} \}_{k=1}^K \) is reconstructed from \( \{ \hat{r}_k(t)^{(t)} \}_{k=1}^K \) by
\[
\hat{g}_k(t) = \frac{1}{\sum_{i \in M_k} Q(k,i)} \left[ \sum_{i \in M_k} \hat{r}_k(t)^{(t)} \right] + \hat{g}_k(t) 1_{D \times 1} \in \mathbb{R}^D, \forall k \in [K],
\]
where \( \hat{g}_k(t) = (\sum_{i \in M_k} Q(k,i) \hat{g}_{k(i),i})/(\sum_{i \in M_k} Q(k,i)) \).

### D. OA-FMTL Framework

We first describe the model aggregation error of each task \( k \), and then summarize the OA-FMTL framework in this subsection. From the channel model in (7), the model aggregations \( \{ g_k(t)^{(t)} \}_{k=1}^K \) inevitably suffer from distortions caused by the inter-task interference and the channel noise. The global model \( w_k(t) \) of task \( k \) is updated with the gradient aggregation \( \hat{g}_k(t) \):
\[
w_k(t+1) = w_k(t) - \eta_k \hat{g}_k(t) = w_k(t) - \eta_k \left( \nabla F_k(w_k(t)) - e_k(t) \right),
\]
where \( \nabla F_k(w_k(t)) \) is the gradient of the loss function of the \( k \)-th task \( F_k(w_k) \) at \( w_k = w_k(t) \), and \( e_k(t) \) is the error caused by gradient uploading, which can be divided into two parts:
\[
e_k(t) = e_{da,k}^{(t)} + e_{com,k}^{(t)},
\]
where \( e_{da,k}^{(t)} \) denotes the error caused by device selection, and \( e_{com,k}^{(t)} \) denotes the communication error due to the inter-task interference and the noise. To alleviate the impact of \( e_k(t) \) on the learning performance, the PS optimizes the device selection, transmit and receive beamforming.

We summarize the OA-FMTL framework described above in Algorithm 1. Besides, we define the following variables for notational brevity: \( M \triangleq \{ M_k \}_{k=1}^K, T \triangleq \{ t_k \}_{k=1}^K, \mathbf{u}_k \triangleq \{ u_{k(i),i} \}_{i \in M_k} \), and \( \mathbf{u} \triangleq \{ u_{k(i),i} \}_{k=1}^K \). In the following section, we establish the connection between the model aggregation errors in (15) and the learning performance of the OA-FMTL.

### III. PERFORMANCE ANALYSIS

In this section, we analyze the learning performance of the OA-FMTL. We start with some standard assumptions on the loss functions \( \{ F_k(\cdot) \}_{k=1}^K \).

#### A. Preliminaries

To conduct convergence analysis, following the stochastic optimization literature [32], [33], we make the following four assumptions on \( \{ F_k(\cdot) \}_{k=1}^K \):

**Assumption 1:** For each task \( k \), the loss function \( F_k \) is continuously differentiable, and the gradient \( \nabla F_k(\cdot) \) is uniformly Lipschitz continuous with parameter \( \omega_k \), i.e.,
\[
\| \nabla F_k(\mathbf{w}) - \nabla F_k(\mathbf{w}') \| \leq \omega_k \| \mathbf{w} - \mathbf{w}' \|,
\]
\[
\forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^D, \ k \in [K].
\]

**Algorithm 1 OA-FMTL Framework**

**Require:** \( T, \{ Q(k,i) \} \).

1. **Initialization:** \( t = 0 \), the global models \( \{ w_k^{(0)} \} \) on the PS.
2. **for** \( t \in [T] \) **do**
   3. The PS estimates the CSI and optimize \( (M, f, u) \);
   4. The PS sends the global models \( \{ w_k^{(t)} \} \) to the devices through orthogonal transmission;
   5. **for** \( k \in [K], i \in M_k \) **in parallel**
      6. Device \( (k,i) \) computes its local gradient \( g_{k(i),i}^{(t)} \) based on the local dataset via (4);
      7. Device \( (k,i) \) uploads its gradient \( g_{k(i),i}^{(t)} \) to the PS via (8)-(11);
   8. **end for**
   9. The PS recovers the aggregated gradient \( \hat{g}_k^{(t)} \) of each task \( k \) based on (12) and (13);
10. The PS updates the global models \( \{ w_k^{(t+1)} \} \) based on (14);
11. **end for**

**Assumption 2:** For each task \( k \), loss function \( F_k \) is strongly convex with positive parameter \( \mu_k \):
\[
F_k(\mathbf{w}) \geq F_k(\mathbf{w}') + (\mathbf{w} - \mathbf{w}')^T \nabla F_k(\mathbf{w}') + \frac{\mu_k}{2} \| \mathbf{w} - \mathbf{w}' \|^2, \forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^D, k \in [K].
\]

**Assumption 3:** \( F_k(\cdot), k \in [K] \) are twice-continuously differentiable.

**Assumption 4:** The gradient vector is upper bounded by
\[
\| \nabla F_k(\mathbf{w}_k; \xi_{k,n}) \|^2 \leq \beta_1 + \beta_2 \| \nabla F_k(\mathbf{w}_k(t)) \|^2, \forall k \in [K],
\]
for some constants \( \beta_1 \geq 0 \) and \( \beta_2 \geq 0 \). Both \( \beta_1 \) and \( \beta_2 \) are constants shared by \( \{ F_k(\cdot) \}_{k=1}^K \).

Furthermore, unlike the approaches in [19], [25] where the local gradients from a common task are treated as independent, we find that these local gradients are highly correlated in a typical learning task. As such, it is of critical importance to understand the impact of the spatial correlation between gradients on the learning performance. To this end, we introduce a probability model for the gradients as follows. Let \( \widetilde{G}_k(t) \triangleq \{ G_k(t)^{(t)} \}_{k=1}^K \) be the local gradients from the \( M_k \) devices of task \( k \) at the \( t \)-th round. Let \( p_k^{(t)}(\{ G_k(t)^{(t)} \}_{k=1}^K) \in \mathbb{R}^{D \times M_k} \) be the local gradient matrices on the \( M_k \) devices of task \( k \), and \( G_{k,d}^{(t)} \) is the \( d \)-th row of \( G_k(t)^{(t)} \), as shown in Fig. 3. To track the correlation of the gradients, we make the following assumption on the distribution of the gradients elements.

**Assumption 5:** For the \( t \)-th communication round, the gradient matrices \( \{ G_k(t)^{(t)} \}_{k=1}^K \) are independent and non-identically distributed. For the \( k \)-th task, the gradients \( \{ G_{k,d}^{(t)} \}_{d \in [D]} \) are independent and identically distributed. That is,
\[
p_k^{(t)}(\{ G_k(t)^{(t)} \}_{k=1}^K) = \prod_{k=1}^K \prod_{d=1}^D p_k^{(t)}(G_{k,d}^{(t)}), \ \forall d.
\]
the gradients are updated ideally without any transmission error. Intuitively, the darker the color of a pixel, the stronger the spatial correlation between the gradients from the corresponding two devices. From Fig. 4, we see that the correlation grows substantially at the beginning of training, say, for communication round \( t \leq 50 \). When the training approaches convergence, most of the elements of \( \rho_k^{(t)} \) reduce to less than 0.3, which indicates that the local gradients have weaker cross-correlation when the models are close to convergence.

### B. Convergence Analysis of OA-FMTL

We start with the analysis of the loss function \( F_k(\cdot) \) of task \( k \) at each communication round. From [32, Lemma 2.1], Assumptions 1-4 lead to an upper bound of the loss function \( F_k(\cdot) \) at the \( t \)-th round of the recursive updates in (14). We therefore have the following theorem.

**Theorem 1:** Under Assumptions 1-5, the expected loss function of each task \( k \) at the \( t \)-th communication round is bounded by

\[
\mathbb{E}[F_k(\mathbf{w}_k^{(t+1)})] \leq \mathbb{E}[F_k(\mathbf{w}_k^{(t)})] - \frac{1}{2\omega_k} \left( \mathbb{E}||\nabla F_k(\mathbf{w}_k^{(t)})||^2 \right) - 2\mathbb{E}||\mathbf{e}_d^{(t)}||^2 - 2\mathbb{E}||\mathbf{e}_{\text{com},k}^{(t)}||^2, \tag{21}
\]

where the device selection MSE is bounded by

\[
\mathbb{E}||\mathbf{e}_d^{(t)}||^2 \leq \frac{4}{Q_k^2} \left( Q_k - \sum_{i \in M_k} Q_{(k,i)} \right)^2 \tag{22}
\]

and the communication MSE is given by (23), shown at the bottom of the next page.

**Proof:** Please refer to Appendix A.

From Theorem 1, we obtain an upper bound of the loss function \( F_k(\cdot) \) w.r.t. the device selection MSE and the communication MSE. By inspection, the communication MSE in (23) consists of three terms, where the first term represents the misalignment error of the aggregation gradients from the devices in task \( k \), the second term represents the error caused by the interference from the devices associated with other tasks, and the third term represents the error caused by the channel noise. Note that the expression in (23) is convex w.r.t.
ζ_k for any fixed device selection set M_k and beamforming f_k and \{u_{k,i}\}_{i \in M_k}. Thus, we have the following corollary.

Corollary 1: The optimal \( \zeta_k \) is given by (24), which is shown at the bottom of the page.

Proof: Please refer to Appendix B. \( \square \)

We emphasize that the design strategy of the weighting factors \( \{\zeta_k\}_{k=1}^K \) in (24) is different from that of the existing over-the-air FL approaches. For each task \( k \), the first component in (23) represents the misalignment error of the gradient aggregation. In the existing scheme, Refs. [17], [19], [20] force this component to zero, which leads to the constraints of \( Q_{(k,i)}v_{k}^{(t)} - \zeta_k f_k^H H_{(k,i)}^t u_{(k,i)} = 0, \forall i \in M_k \), and then minimize the rest of the communication MSE to determine \( \zeta_k \). For all devices in the \( k \)-th task, to satisfy the constraints and the transmit power budgets in (11), the choice of \( \zeta_k \) is therefore given by

\[
\zeta_k^2 = \max_{i \in M_k} Q_{(k,i)}^2 / ||f_k^H H_{(k,i)}^t||^2.
\]

Thus, \( \zeta_k \) is dominated by the device with the worst channel condition (in terms of \( ||f_k^H H_{(k,i)}^t||^2 \)). That is, the device with the worst channel becomes the bottleneck of the overall scheme, also known as the straggler problem. However, instead of zero-forcing, our approach tolerates the misalignment error by requiring the design of \( \zeta_k \) to directly minimize the overall communication MSE for task \( k \). In this way, \( \zeta_k \) is no longer solely determined by the worst channel condition, which significantly relieves the straggler problem. Note that our approach is also different from the approach in [25] since we are aware of the spatial correlation between the local gradients, which brings substantial improvement in learning performance. Numerical results will be presented later in Section V for verification.

In the former of this subsection, we obtain an upper bound of the loss function \( F_k(\cdot) \) of task \( k \) at the \( t \)-th communication round in Theorem 1. In the following, we analyse the convergence performance of the entire OA-FMTL model at the \( t \)-th round, and obtain an upper bound of the average difference between the overall loss function at the \((t + 1)\)-th round \( F(w^{(t+1)}) \) and the optimal \( F(w^*) \), i.e.,

\[
\mathbb{E}[F(w^{(t+1)}) - F(w^*)].
\]

Based on Theorem 1 and Corollary 1, we bring the device selection MSE and the communication MSE into an analysis framework, and obtain the following theorem.

Theorem 2: Based on Assumptions 1-5, the overall loss function at the \( t \)-th round \( F(w^{(t+1)}) \) satisfies the following inequality:

\[
\mathbb{E}[F(w^{(t+1)}) - F(w^*)] \leq \mathbb{E}[F(w^{(t)}) - F(w^*)] \left( 1 - \frac{\mu}{\omega} \right) + \left( \frac{2\mu^2}{\omega} \right) \mathbb{E}[F(w^{(t)}) - F(w^*)] + \left( \frac{\beta}{\omega} \right) \mathbb{E}(t)(M, f, u),
\]

where \( \omega \triangleq \max_k \omega_k, \mu \triangleq \min_k \mu_k, \mathbb{E}(t)(M, f, u) \triangleq \sum_{k=1}^K d_k^t(M_k, f_k, u_k), \) and \( d_k^t(M_k, f_k, u_k) \) is denoted by (25), shown at the bottom of the page.

Proof: Please refer to Appendix C. \( \square \)

From Theorem 2, we see that \( \mathbb{E}[F(w^{(t+1)}) - F(w^*)] \) is upper bounded by a quantity monotonic in \( \mathbb{E}(t)(M, f, u) \). Since the CSI matrices are time-varying, \( \mathbb{E}(t)(M, f, u) \) also varies from round to round. Thus, we give a further convergence analysis in the sense of averaging over channel realizations. Let \( \mathbb{E}[\cdot] \) denote the expectation taken over the
CSI matrices, the gradients, and the channel noise. Note that \( \{ M, f, u \} \) are generally functions of the CSI matrices, since the CSI matrices are known by the transceivers. With the expectation taken over the CSI matrices, \( \mathbb{E} \left[ \mathcal{E}^{(t)}(M, f, u) \right] \) is no longer a function of \( \{ M, f, u \} \). Thus, we obtain \( \mathbb{E} \left[ \mathcal{E}^{(t)}(M, f, u) \right] = \bar{E}, \forall t \in [T] \), where \( \bar{E} \) is a constant. Taking expectation \( \mathbb{E} [ \cdot ] \) on both sides of (26), we have

\[
\mathbb{E} [\mathcal{F}(w^{(t)}(t)) - \mathcal{F}(w^*)] \leq \mathbb{E} [\mathcal{F}(w^{(t)}(t)) - \mathcal{F}(w^*)] \frac{\beta_1 \bar{E}}{\omega} \quad (27)
\]

where \( \bar{E} = \mathbb{E} [\psi(t)(M, f, u)] = 1 - \frac{\mu}{2} + \frac{2 \mu \beta_2}{\omega}, \forall t \in [T] \), with \( \psi(t)(M, f, u) = 1 - \frac{\mu}{2} + \frac{2 \mu \beta_2}{\omega} \psi^{(t)}(M, f, u) \). Recursively applying (27) for \( (t + 1) \) times, we obtain

\[
\mathbb{E} [\mathcal{F}(w^{(t+1)}(t+1)) - \mathcal{F}(w^*)] \leq \left( \mathcal{F}(w^{(t)}(t)) - \mathcal{F}(w^*) \right) \frac{\beta_1 \bar{E}}{\omega} + \frac{\beta_1}{\omega - 1} \bar{E}. \quad (28)
\]

The key insights from (27) and (28) are summarized as below:

- When \( \bar{E} < 1 \), the upper bound of \( \mathbb{E}[\mathcal{F}(w^{(t+1)}(t+1)) - \mathcal{F}(w^*)] \) in (28) converges as \( T \to \infty \) with speed \( \bar{E} \). Clearly, minimizing \( \bar{E} \) can speed up the convergence of the upper bound of \( \mathbb{E}[\mathcal{F}(w^{(t+1)}(t+1)) - \mathcal{F}(w^*)] \).
- Minimizing \( \bar{E} \) also reduces the upper bound limit of \( \mathbb{E}[\mathcal{F}(w^{(t+1)}(t+1)) - \mathcal{F}(w^*)] \), i.e.,

\[
\lim_{T \to \infty} \mathbb{E}[\mathcal{F}(w^{(t)}(T)) - \mathcal{F}(w^*)] \leq \frac{\beta_1 \bar{E}}{(\omega - 1) \bar{E}} \leq \frac{\beta_1}{\omega - 2 \mu \beta_2}. \quad (29)
\]

Therefore, minimizing \( \bar{E} \) is desirable since it both speeds up the convergence speed and minimizes the upper bound in (29). This implies that we need to minimize \( \mathcal{E}^{(t)}(M, f, u) \) for each channel realization, as detailed in the next section.

### IV. System Optimization

To achieve better learning performance, we aim to minimize the upper bound of \( \mathbb{E}[\mathcal{F}(w^{(t+1)}(t+1)) - \mathcal{F}(w^*)] \) in (26), or equivalently, to minimize \( \mathcal{E}^{(t)}(M, f, u) \) over device selection set \( M \), receive beamforming \( f \) and transmit beamforming \( u \), as detailed below.

#### A. Problem Formulation

From Theorem 2, the gap \( \mathbb{E}[\mathcal{F}(w^{(t+1)}(t+1)) - \mathcal{F}(w^*)] \) has an upper bound in (26). The upper bound is monotonically increasing w.r.t. \( \mathcal{E}^{(t)}(M, f, u) \) since \( \frac{2 \mu \beta_2}{\omega} E[F(w^{(t)}(t)) - F(w^*)] + \frac{\beta_1}{\omega} > 0 \). With the target to minimize the gap \( \mathbb{E}[\mathcal{F}(w^{(t+1)}(t+1)) - \mathcal{F}(w^*)] \), we minimize \( \mathcal{E}^{(t)}(M, f, u) \) at round \( t \), and we formulate the optimization problem P1 as:

\[
(P1) \quad \min_{M, f, u} \mathcal{E}^{(t)}(M, f, u) = \sum_{k=1}^{K} d_k^{(t)}(M_k, f_k, u_k)
\]

\[
\text{s.t.} \quad M_k \subseteq [M_k], k \in [K]
\]

\[
\|f_k\|_2 = 1, k \in [K]
\]

\[
\|u_k\|_2 \leq P_0/2, k \in [K], i \in [M_k]
\]

\[
\text{where}
\]

\[
d_k^{(t)}(M_k, f_k, u_k) = \frac{4}{Q_k} Q_k - \sum_{i \in M_k} Q_{(k,i)}^2 + \sum_{l,j \in M_k} \rho^{(t)}_{k,l,j} Q_{(k,i)} Q_{(l,j)} - \frac{a_k^{(t)}(M_k, f_k, u_k)^2}{(\sum_{i \in M_k} Q_{(k,i)}^2)^2}.
\]

\[
(31)
\]

with

\[
a_k^{(t)}(M_k, f_k, u_k) \triangleq \sum_{i,j \in M_k} \rho^{(t)}_{k,i,j} \times \left( Q_{(k,i)} f_k^H H_{(k,j)}^i u_{(k,j)}^H \right) + \sum_{i,j \in M_k} \left( \rho^{(t)}_{k,l,j} \times f_k^H H_{(k,j)}^i u_{(l,j)}^H \right) + \sigma^2 \|f_k\|_2^2/2.
\]

(P1) is an optimization problem w.r.t. device selection set \( M \), receive beamforming vectors \( f \) and transmit beamforming vectors \( u \), respectively.

We next design an AO-based algorithm to solve the optimization problem P1. P1 contains \( 2M + K \) optimization variables, i.e., \( M \) device selection indices, \( K \) receive beamforming vectors at the PS and \( M \) transmit beamforming vectors at the devices. P1 is non-convex due to the coupling of \( M, f \) and \( u \). Thus, we adopt the AO framework to solve the problem in a suboptimal fashion. First, we optimize the beamforming vectors with fixed device selection set \( M \). Second, with fixed beamforming vectors, we optimize device selection set \( M \) with Gibbs sampling [17]. The two steps iterate until convergence. The details are discussed in the following subsections.

#### B. Optimization of \( f \) and \( u \) With Fixed \( M \)

We first optimize beamforming vectors \( f \) and \( u \) with fixed device selection set \( M \). By inspection of (P1), \( d_k^{(t)}(M_k, f_k, u_k) \) in (31) is invariant to the value of \( \|f_k\|_2 \). Therefore, the unit-length constraint of \( f_k \) in (30c) can be ignored without changing the minimum of (P1). Further, we drop the constant terms in the objective function \( \mathcal{E}^{(t)}(M, f, u) \) to obtain

\[
\min_{\{f_k\}, u} \sum_{k=1}^{K} d_k^{(t)}(M_k, f_k, u_k) - \sum_{k=1}^{K} 4 \left( \sum_{i \in M_k} Q_{(k,i)}^2 \right)^2 b_k^{(t)}(M_k, f_k, u_k),
\]

\[
\text{s.t.} \quad (30d).
\]

(33) is still non-convex because both the numerator and the denominator in the \( k \)-th summand contain the optimization variables \( f_k \) and \( u_k \). We adopt the quadratic transform in fractional programming (FP) [38] to decouple the numerator.
and the denominator as

\[
\min_{\{f_k\}, u, y} \sum_{k=1}^{K} \left( y_k L_k^f (M_k, f_k, u_k) - y_k L_k^b (M_k, f_k, u_k) \right), \tag{34}
\]

s.t. \hspace{1em} (30d).

where \( y = [y_1, \cdots, y_K]^T \in \mathbb{R}^K \) is an auxiliary vector introduced by FP. Note that (34) reduces to (33) by letting each \( y_k \) take its optimal form as

\[
y_k = \frac{a_k^f (M_k, f_k, u_k)}{2 \left( \sum_{i \in M_k} Q(k,i) b_k^b (M_k, f_k, u_k) \right)}. \tag{35}
\]

1) Optimizing \( u_{(k,i)} \) With Fixed \( \{u_{(k,j)}\}_{j \neq i} \) and \( \{f_k\} \): With fixed \( \{u_{(k,j)}\}_{j \neq i} \) and \( \{f_k\} \), when \( i \notin M_k \), i.e., the device \( \langle k, i \rangle \) is not selected, we obtain \( u_{(k,i)} = \mathbf{0}_{N_T \times 1} \) directly. When \( i \in M_k \), i.e., the device \( \langle k, i \rangle \) is selected, (34) reduces to

\[
(P2) : \min_{u_{(k,i)}} A_{(k,i)}^H u_{(k,i)} - 2 \Re \{ \mathbf{b}_{(k,i)}^H u_{(k,i)} \},
\]

s.t. \hspace{1em} (30d).

where \( A_{(k,i)}^H \in \mathbb{C}^{N_T \times N_T} \) and \( b_{(k,i)}^H \in \mathbb{C}^{N_T} \) are defined by

\[
A_{(k,i)}^H = \sum_{l=1}^{K} y_k^2 \sigma^4 \mathbf{H}_{(k,i)}^H \mathbf{H}_{(k,i)}, \tag{36a}
\]

\[
b_{(k,i)}^H = y_k \sum_{j \in M_k, j \neq i} \rho_{(k,i), (k,j)} \mathbf{Q}_{(k,i)}^H \mathbf{Q}_{(k,j)}^H \mathbf{H}_{(k,i)}^H \mathbf{H}_{(k,j)}^H u_{(k,j)}.
\]

Note that \( A_{(k,i)}^H \) is a positive semidefinite matrix, and that the constraint (30d) is convex w.r.t. \( u_{(k,i)} \). Thus, P2 is a convex quadratically constrained quadratic programming (QCQP) problem w.r.t. \( u_{(k,i)} \), which can be solved by standard convex optimization tools.

2) Optimizing \( f_k \) With Fixed \( u \): Similarly to IV-B.1, with fixed \( u \), (34) reduces to

\[
(P3) : \min_{f_k} f_k^H A_k^H f_k - 2 \Re \{ \mathbf{b}_{k}^H \mathbf{f}_k \}
\]

where \( A_k^H \in \mathbb{C}^{N_K \times N_K} \) and \( b_k^H \in \mathbb{C}^{N_K} \) are denoted by

\[
A_k^H = \sum_{l=1}^{K} \sum_{i, j \in M_l} \rho_{(l,i), (l,j)} \mathbf{H}_{(l,i)}^H \mathbf{u}_{(l,i)} (\mathbf{H}_{(l,j)}^H \mathbf{u}_{(l,j)})^H + \left( y_k^2 \sigma^2 / 2 \right) \mathbf{I}_{N_K}, \tag{37a}
\]

\[
b_k^H = \sum_{i, j \in M_k} Q_{(k,i), (k,j)} \rho_{(k,i), (k,j)} \mathbf{H}_{(k,i)}^H \mathbf{u}_{(k,i)} \mathbf{u}_{(k,j)}^H.
\]

P3 is convex w.r.t. \( f_k \) by noting the positive semidefinite matrix \( A_k^H \). Note that \( f_k \) here is obtained by dropping the unit-length constraint in (30c). Thus, we finally obtain the optimal \( f_k \) by scaling the obtained \( f_k \) to a unit vector. We summarize the optimization of \( f \) and \( u \) in Algorithm 2.

\begin{algorithm}
\caption{AO Algorithm to Optimize \( f \) and \( u \)}
\label{alg:ao}
\begin{algorithmic}
\Require: \( M, \{\rho_k^f, \mathbf{H}_{k, (l,i), (l,j)}^H, Q_{(k,i), (k,j)}\}_{k, l, i, j} \), and \( I_{\max} \).
\Ensure: \( (f, u) \).
\State Initialization: \( y, f \) and \( u \).
\For {\( \tau \in [I_{\max}] \)}
\For {\( k \in [K] \)}
\State Compute \( A_{k, (l,i)}^f, b_{k, (l,i)}^f \) based on (36);
\State Optimize \( u_{(k,i)} \) by solving (P2);
\EndFor
\State Compute \( A_{k, (l,i)}^b, b_{k, (l,i)}^b \) based on (36);
\State Optimize \( f_k \) by solving (P3), update \( f_k \) (by scaling to a unit vector)
\EndFor
\State Updates \( y \) based on (35);
\EndFor
\Ensure: \( (f, u) \).
\end{algorithmic}
\end{algorithm}

C. Optimizing \( M \) With Fixed \( f \) and \( u \)

In this subsection, we optimize device selection set \( M \) with fixed beamforming \( f \) and \( u \). For notational convenience, we use a binary indication vector \( s \) to represent the device selection set \( M \), i.e., \( s \triangleq \left[ s_1^T, \cdots, s_K^T \right]^T \in \{0, 1\}^M \), where \( s_k \in \{0, 1\}^M \) with the \( i \)-th entry \( s_{(k,i)} = 1 \) meaning that the \( \langle k, i \rangle \)-th device is selected and \( s_{(k,i)} = 0 \) otherwise. Note that \( M \) and \( s \) can be interchanged. Since \( s \) is discrete, we introduce the Gibbs sampling to optimize \( s \) by following the approach in [17]. We denote \( s^{\text{old}} \) as the sampling device selection solution obtained from the proceeding sampling round. At the current sampling round, the sampling set \( S \) is generated from \( s^{\text{old}} \), given by \( S \triangleq \{s^{\text{old}}\} \cup \{s^{\text{old}}(k,i) \in [K], i \in [M_k]\} \), where \( s^{\text{old}}(k,i) \) denotes the indication vector that differs from \( s^{\text{old}} \) only at the \( \langle k, i \rangle \)-th element, corresponding to the \( \langle k, i \rangle \)-th device. We sample \( s^{\text{new}} \) according to the following distribution \( \pi(s^{\text{new}}) \):

\[
\pi(s^{\text{new}}) \triangleq \frac{\exp \left(-\phi(s^{\text{old}})/\beta\right)}{\sum_{k=1}^{K} \sum_{i=1}^{M_k} \exp \left(-\phi(s^{\text{old}})/\beta\right)},
\tag{38}
\]

where \( \phi(s^{\text{old}}) \) denotes the objective \( \mathcal{E}^{(s^{\text{old}})}(M, f, u) \) with \( f \) and \( u \) obtained by Algorithm 2 with the device selection set \( M \) corresponding to \( s^{\text{old}}(k,i) \), and \( \beta > 0 \) denotes the “temperature parameter” to accelerate convergence. We summarize the overall algorithm for the optimization of \( M \), \( f \) and \( u \) in Algorithm 3.

D. Complexity Analysis

We now briefly discuss the computational complexity involved in Algorithms 2 and 3. For Algorithm 2, both P2 and P3 are QCQP problems that can be solved by existing optimization solvers based on the interior point method. Thus, the worst-case complexity of Algorithm 2 is given by \( O(I_{\max} (K + M)N^3) \), where \( N \triangleq \max\{N_T, N_R\} \) denotes.
the maximum number of the transmit or receive antennas, $I_{\text{max}}$ is the max iteration times of optimization, and $M$ is the total number of the devices in the OA-FMTL. Algorithm 3 invokes Algorithm 2 for $J_{\text{max}}M$ times to optimize device selection set $\mathcal{M}$. Thus, the complexity of Algorithm 3 is $O(J_{\text{max}}I_{\text{max}}(KM + M^2)N^{3.5})$.

We note that the complexity of Algorithm 3 is quadratic in $M$, due to the use of Gibbs sampling in the optimization of device selection. When the number of devices $M$ in the OA-FMTL is large, Gibbs sampling causes a tremendous computation burden. Recall from Section III-B that the device selection is adopted by the existing works to reduce the impact of stragglers. As pointed out in Section III-B, our proposed scheme optimizes the weighting factor $\zeta_k$ to minimize the communication MSE, which relieves the straggler problem significantly. We observe from experiments that the improvement of device selection in Algorithm 3 is negligible, as compared to the performance of Algorithm 2. Therefore, we prefer to use Algorithm 2 in the system optimization, since Algorithm 2 has a much lower complexity.

We note that Algorithm 2 needs to be done within each transmission block. In this regard, the complexity of Algorithm 2 may be still too high, especially in relatively fast-changing channels with short channel coherence time. It is possible to extend the work in this paper to perform system optimization in a longer timescale, e.g., by following the approach in [5] and [6]. However, the details are out of the scope of this paper.

V. NUMERICAL RESULTS

A. Simulation Setups

We consider a three-dimensional (3-D) simulation scenario as shown in Fig. 5. The point locations are represented by cylindrical coordinate triples $(\delta, \theta, \chi)$, where $\delta$, $\theta$ and $\chi$ denote the radial distance, the azimuth, and the height, respectively. The locations of the devices are distributed as follows. All the devices in the OA-FMTL are located in a circle with center $O = (0, 0, 0)$ and radius $\Delta$. We set the location of the $(k, i)$-th device to $(\delta_{(k,i)}, \theta_{(k,i)}, 0)$, where $\delta_{(k,i)}$ is uniform in $[0, \Delta]$, and $\theta_{(k,i)}$ is uniform in $[0, 2\pi)$. The PS is placed at the center of the circle, i.e., $(0, 0, 0)$. We adopt the channel model in [39], given by $H_{(l,i)} = \sqrt{G_S G_D \delta_{(l,i)}^{-\alpha_z}} \tilde{H}_{(l,i)}$, where the entries of $\tilde{H}_{(l,i)}$ are modeled as i.i.d. circularly symmetric complex Gaussian (CSCG) random variables with zero-mean and unit-variance, $G_S$ and $G_D$ are the antenna gains at the PS and the devices, respectively, $\kappa$ is the path loss at the reference distance $\delta_0 = 1$ m [40], $\alpha$ is the path loss exponent, and $\tilde{\delta}_{(l,i)} \triangleq \sqrt{\tilde{\delta}_{(l,i)}^2 + \Delta^2}$ is the distance between the $(l, i)$-th device and the PS. The simulation settings are given in Table I.

We set three image classification tasks as FL tasks in the OA-FMTL, with each FL task being trained on an individual dataset, i.e., MNIST for task 1, Fashion-MNIST for task 2, and KMNIST for task 3. For each FL task, we train a CNN with two $5 \times 5$ convolution layers (the first with 16 channels, the second with 32, each followed by $2 \times 2$ max pooling), a fully connected layer with 50 units and ReLu activation, and a final softmax output layer ($D = 39408$ total parameters). The loss function is the cross-entropy loss. We study two ways of partitioning the dataset $A_k$ over devices: 1) i.i.d., where the data are shuffled, and then assigned evenly to the $M_k$ devices; 2) non-i.i.d., where each device randomly selects 5 classes, and then randomly draws $Q_{SM_k}$ samples from each selected class.

B. Comparisons of the Proposed Algorithms Under Various Settings

In this subsection, we study the impact of various approximations of the correlation matrices $\{\rho_{(l,i)}^{(t)}\}$. Specifically, we consider the following three approximations of the correlation matrix:

- **Approximation 1**: $\rho_{(l,i)}^{(t)} = \frac{1}{D} \sum_{d=1}^{D} z_{(k,d)}^{(t)} z_{(k,d)}^{(t)\top}$.
- **Approximation 2**: $\rho_{(l,i)}^{(t)}$ is approximated by $\rho_{k}^{(t)} = \epsilon \mathbf{1}_{M_k} + (1 - \epsilon) \mathbf{1}_{M_k}$, where $\epsilon$ is an empirical parameter to represent the degree of correlation.

Approximation 1 estimates the spatial correlation between the gradients of the devices from the same task $k$ at each round. However, Approximation 1 is impractical since the calculation of $\rho_{k}^{(t)}$ requires the knowledge of the local gradients uploaded by the devices. In contrast, Approximation 2 is more practical in implementation, where $\epsilon$ can be set empirically. Therefore, the proposed schemes with Approximation 1 are only used as a performance baseline, and the schemes with Approximation 2 are used in performance comparison with other counterpart schemes. In addition, we adopt the error-free case as a performance upper bound of each FL task in the OA-FMTL:
• Error-free bound: Each FL task is trained independently with all the devices being selected, and the model aggregation is error-free at each communication round.

We plot the FL test accuracy curve of the proposed Algorithm 2 with different approximations of \( \rho_k^{(t)} \) versus the communication rounds, with \( K = 2 \), \( M_1 = M_2 = 20 \). Left: MNIST; right: Fashion-MNIST; top: i.i.d. data; bottom: non-i.i.d. data.

**Algorithm 3 AO Algorithm Plus Gibbs Sampling**

Require: \( \{ \rho_k^{(t)}, H_{l,i}^{(t)} \}, Q_{[k,i]}[k,l \in [K], i \in [M_k]], J_{\max}, \beta, \gamma \).
1. Initialization: \( s^{\text{old}} = 1_{M \times 1} \).
2. for \( j \in [J_{\max}] \) do
3. \( s^{\text{old}} = s^{\text{new}} \).
4. Generate \( S \);
5. for every \( s^{\text{old}}_{[k,i]} \in S \) do
6. Optimize \((f,u)\) by solving P2 with given \( s^{\text{old}}_{[k,i]} \), with Algorithm 2;
7. end for
8. Sample \( s^{\text{new}} \) according to (38);
9. Refresh \( \beta = \gamma \beta \) for a certain \( \gamma \in (0,1) \);
10. end for
Ensure: \( s^{\text{new}} \) with corresponding \((f,u)\).

![Image](312x555 to 563x555)

![Image](312x720 to 563x720)

![Image](312x733 to 563x734)

Fig. 6. FL test accuracy of the proposed Algorithm 2 with different approximations of \( \rho_k^{(t)} \) versus the communication rounds, with \( K = 2 \), \( M_1 = M_2 = 20 \). Left: MNIST; right: Fashion-MNIST; top: i.i.d. data; bottom: non-i.i.d. data.

![Image](312x301)

![Image](312x309)

Fig. 7. Test accuracy versus communication round, with \( K = 2 \), \( M_1 = M_2 = 10 \). Left: MNIST; right: Fashion-MNIST; top: i.i.d. data; bottom: non-i.i.d. data.

Note that we choose a relatively small number of trials due to the high computational complexity of device selection in Algorithm 3. In simulations, we use a personal computer with an Intel(R) Core(TM) i7-10700 CPU and a GTX 1050Ti GPU. One Monte Carlo trial of Algorithm 3 takes about 10 hours.
ψ

the tasks, i.e., each task is assigned with an orthogonal time slot to avoid inter-task interference, and we have $t_0(\psi) = \sum_{k=1}^{K} t_k(\psi)$. Fig. 8 depicts the total required rounds versus the relative target accuracy on both i.i.d. data and non-i.i.d. data, with $K = 3$. We see that $t_0(\psi)$ is nearly $1/K$ of $t_k(\psi)$ at the same $\psi$ on both i.i.d. data and non-i.i.d. data. $t_0(\psi)$ does not increase with the growth of the number of FL tasks, which shows the saving of communication resources by using the proposed non-orthogonal scheme.

C. Comparisons With Existing Schemes

In this subsection, we present the performance of system optimization obtained by the proposed AO algorithm on both i.i.d. data and non-i.i.d. data. We consider the following baselines for comparison:

- **SOCP-based cooperative power control** [25]: This method assumes that all the devices are selected and that the local gradients from devices in the same task are independent with each other. The phases of transmit beamforming $\mathbf{u}_{(k,i)}, \forall k, i$ are given by the zero-forcing. The design of transmit power $\|\mathbf{u}_{(k,i)}\|^2, \forall k, i$ are formulated as a second-order cone programming (SOCP) problem solved by the bisection method.

- **SCA-based optimization and device selection** [17]: $\mathcal{M}_k, \mathbf{f}_k$, and $\{\mathbf{u}_{(k,i)}\}$ for each FL task are optimized separately. For task $k$, the receive beamforming $\mathbf{f}_k$ is optimized by the successive convex approximation (SCA)-based optimization algorithm, with given transmit beamforming $\mathbf{u}_k$ and the weighting factor $\zeta_k$ determined by zero-forcing. With given optimized $\mathbf{f}_k$, $\mathbf{u}_k$ and $\zeta_k$, device selection set $\mathcal{M}_k$ is optimized via Gibbs sampling.

- **Receive beamforming by differential geometry programming** [15]: This method optimizes each task separately, with all the devices selected. $\mathbf{f}_k$ is optimized on the Grassmann manifold via differential geometry programming, and $\mathbf{u}_k$ is given by the zero-forcing.

- **Difference-of-convex (DC) programming and device selection** [19]: This method optimizes each task separately. For each task, the method maximizes the number of selected devices by a two-step framework based on DC programming, with a given threshold of the communication MSE.

Here we simulate the case of 3 FL tasks on i.i.d. data. The numbers of devices for the three tasks are set to $M_1 = M_2 = M_3 = 20$. The learning rates are set to $\eta_1 = \eta_2 = \eta_3 = 0.05$. The local updates contain 5 mini-batches of SGD. The noise power is set to $\sigma^2 = 51 - 60$ dBm, and $\epsilon$ is set to $\epsilon = 1.0$. Besides, we introduce the normalized mean square error (NMSE) of each task $k$ at round $t$, defined as $\text{NMSE}^{(t)}_k = \frac{10 \log_{10} \left( \mathbb{E}[\|\mathbf{g}^{(t)}_k - \hat{\mathbf{g}}^{(t)}_k\|_2^2/\|\mathbf{g}^{(t)}_k\|_2^2] \right)}{2}$. In Table II, we list the average NMSE over 20 Monte Carlo trials at $t = 40$ and $t = 90$. Benefiting from interference awareness, the proposed AO algorithm and the method in [25] achieve much better NMSEs on all these three FL tasks than the other methods. We also see that our algorithm significantly outperforms the method in [25]. This is attributed to the careful optimization of the transceiver beamforming based on the proposed analytical framework.

In Fig. 9, we plot the histogram of allocated transmission powers for various optimization methods. The methods based on zero-forcing [15, 19, 25] only allocate full power to less than 20% of all the devices, due to the stragglers with the worst channel conditions. The SCA & Gibbs algorithm in [17] excludes several stragglers through Gibbs sampling, improving the number of full-power-allocated devices, but the percentage is still below 50%. We see that the transmission powers of most devices are allocated fully in the proposed AO algorithm. This is because our proposed scheme relaxes the hard requirement for all the devices to align their gradients with the stragglers, which gives freedom to the devices to fully exploit the power budgets.

In Fig. 10, we present the test accuracies of various optimization algorithms versus communication rounds on i.i.d. data. As shown in Fig. 10, the proposed algorithm achieves an accuracy close to the error-free bound in all these three FL tasks and significantly outperforms the other baselines, which clearly demonstrates the superiority of our proposed scheme.
TABLE II
COMMUNICATION NMSE

| Optimization method                      | MNIST (k = 1) | Fashion-MNIST (k = 2) | KMNIST (k = 3) |
|------------------------------------------|--------------|-----------------------|---------------|
| Proposed AO algorithm                    | −1.43        | −1.37                 | −1.81         |
| SODCP-based cooperative power control    | −0.84        | −1.01                 | −1.10         |
| SCA & Gibbs [17]                         | −0.40        | −0.47                 | −0.50         |
| Differential geometry [15]               | −0.16        | −0.20                 | −0.21         |
| DC and device selection [19]             | −0.17        | −0.20                 | −0.15         |

Fig. 10. Test accuracy versus communication rounds on (a) MNIST, (b) Fashion-MNIST, and (c) KMNIST, with $K = 3$, $M_1 = M_2 = M_3 = 20$, i.i.d. data.

Fig. 11. Test accuracy versus communication rounds on (a) MNIST, (b) Fashion-MNIST, and (c) KMNIST, with $K = 3$, $M_1 = M_2 = M_3 = 20$, non-i.i.d. data.

Fig. 11 depicts the test accuracies of various optimization algorithms versus communication rounds on non-i.i.d. data. The noise power is set to $\sigma^2 = −70$ dBm, and $\epsilon$ is set to $\epsilon = 0.5$. As shown in Fig. 11, the proposed algorithm outperforms other baselines on all these three FL tasks, which demonstrates the performance of our proposed algorithm.

VI. CONCLUSION

In this paper, we studied a problem of designing an OA-FMTL system over the MIMO MAC channel. We proposed a misalignment-tolerant strategy to align the local gradients at model aggregation at the PS side to relieve the straggler problem. We further derived a communication-learning framework to analyze the OA-FMTL performance by characterizing the performance loss due to device selection, inter-task interference, and communication noise. Based on the analytical framework, we formulated the optimization problem with respect to device selection, transmit beamforming, and receive beamforming. We captured the spatial correlation between the local gradients to enhance the optimization and proposed a low-complexity algorithm to solve the communication-learning problem based on AO framework. Finally, we performed extensive numerical experiments by comparison with the state-of-the-art methods, demonstrating the outstanding improvements of the proposed algorithm on the learning accuracy.

APPENDIX A
PROOF OF THEOREM 1

From [32, 3.2], the device selection MSE $E[\|e_{(1)}^{(i)}\|^2]$ is bounded by (22). We next consider the communication MSE

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
of the $k$-th task given by
\[
\mathbb{E} \left[ ||\mathbf{e}_{\text{com},k}^{(t)}||^2 \right] = \sum_{d=1}^{D} \mathbb{E} \left[ ||g_k^{(t)}(d) - \hat{g}_k^{(t)}(d)||^2 \right].
\] (39)

By plugging (5), (8), (9) and (13) into (39), we obtain
\[
\mathbb{E}[||\mathbf{e}_{\text{com},k}^{(t)}||^2] = \sum_{i \in M_k} \frac{C}{(\sum Q_{(k,i)})^2} \sum_{j \in M_k} 2\rho_{(k,i),(k,j)}^{(t)} Q(k,j) Q(k,j) \hat{r}_{(k,j)}^{(t)}
\]
\[
- \zeta k \sqrt{v_k^{(t)}} \sum_{i,j \in M_k} 2\rho_{(k,i),(j,k)}^{(t)} (Q(k,i) (f_k^H H_{(k,j)}^H u_{(k,j)})^H + Q(k,j) f_k^H H_{(k,j)}^H u_{(k,i)})
\]
\[
+ \zeta k \sum_{i,j \in M_k} 2\rho_{(t,i),(l,j)} (f_k^H H_{(l,j)}^H u_{(l,j)})^H f_k^H H_{(l,j)}^H u_{(l,j)} + \sigma^2 ||f_k||^2)
\]
\[
\geq \frac{2Cv_k^{(t)}}{(\sum_{i \in M_k} Q_{(k,i)})^2} \left( \sum_{i,j \in M_k} \rho_{(k,i),(j,k)}^{(t)} Q(k,j) Q(k,j) \right)
\]
\[
- \left( \sum_{i,j \in M_k} \rho_{(k,i),(j,k)}^{(t)} (Q(k,i) (f_k^H H_{(k,j)}^H u_{(k,j)})^H + Q(k,j) f_k^H H_{(k,j)}^H u_{(k,i)}) \right)^2
\]
\[
\geq \frac{\beta_1 + \beta_2 \mathbb{E}[||\nabla F_k(\mathbf{w}_k^{(t)})||^2]}{(\sum_{i \in M_k} Q_{(k,i)})^2} \left( \sum_{i,j \in M_k} \rho_{(k,i),(j,k)}^{(t)} Q(k,j) Q(k,j) \right)
\]
\[
- \left( \sum_{i,j \in M_k} \rho_{(k,i),(j,k)}^{(t)} (Q(k,i) (f_k^H H_{(k,j)}^H u_{(k,j)})^H + Q(k,j) f_k^H H_{(k,j)}^H u_{(k,i)}) \right)^2
\],
\] (43)

\[
\mathbb{E}[||\mathbf{e}_{\text{com},k}^{(t)}||^2] \leq \frac{\beta_1 + \beta_2 \mathbb{E}[||\nabla F_k(\mathbf{w}_k^{(t)})||^2]}{(\sum_{i \in M_k} Q_{(k,i)})^2} \left( \sum_{i,j \in M_k} \rho_{(k,i),(j,k)}^{(t)} Q(k,j) Q(k,j) \right)
\]
\[
- \left( \sum_{i,j \in M_k} \rho_{(k,i),(j,k)}^{(t)} (Q(k,i) (f_k^H H_{(k,j)}^H u_{(k,j)})^H + Q(k,j) f_k^H H_{(k,j)}^H u_{(k,i)}) \right)^2
\],
\] (44)

In addition, we obtain
\[
\mathbb{E}[||\mathbf{e}_k^{(t)}||^2] \leq 2 \mathbb{E}[||\mathbf{e}_{\text{com},k}^{(t)}||^2] + \mathbb{E}[||\mathbf{e}_{\text{ds},k}^{(t)} + \mathbf{e}_{\text{com},k}^{(t)}||^2]
\]
\[
\leq 2 \left( \mathbb{E}[||\mathbf{e}_{\text{ds},k}^{(t)}||^2] + \mathbb{E}[||\mathbf{e}_{\text{com},k}^{(t)}||^2] \right),
\] (42)

where step (a) is from the expression of $\mathbf{e}_k^{(t)}$ in (15), and step (b) is from the inequality of arithmetic and geometric means. By plugging (42) into (41), we obtain (21).

**Appendix B**

**Proof of Corollary 1**

From (23), we obtain (43), shown at the bottom of the page, where step (a) is from $\mathbb{E}[r_{(k,i)} c r_{(k,j)} c] = 2\rho_{(k,i),(k,j)}$ based on Assumption 5, and step (b) is because $\mathbb{E}[||\mathbf{e}_{\text{com},k}^{(t)}||^2]$ is a convex quadratic function w.r.t. $\zeta_k$ and the minimizer $\zeta_k^*$ is given by (24).

**Appendix C**

**Proof of Theorem 2**

We first derive an upper bound w.r.t. the communication MSE in (44), shown at the bottom of the page. For device $(k,i), \forall k,i$, we have
\[
2Cv_k^{(t)} = 2Cv_k^{(t)} = \sum_{d=1}^{D} \mathbb{E}[||g_k^{(t)}(d) - |d||^2] - \mathbb{E}[|g_k^{(t)}(d)|^2]
\]
\[
\leq \mathbb{E}[||\mathbf{e}_k^{(t)}||^2]
\]
\[
\leq \mathbb{E}
\]
\[
\leq \mathbb{E}
\]
\[
\leq \mathbb{E}
\]
\[
\leq \mathbb{E}
\]
where step (a) is from \( \mathbb{E}[t_{k,i}(d)] \geq 0 \); (b) is from the fact that \( \mathbb{E}[t_{k,i}(d)] = \mathbb{E}[\nabla F_k(w_k^{(t)})] \) and with definition of \( F_{k,i}(w_k^{(t)}) \) given below (2); (c) is from the triangle inequality; and (d) is from (18). Note that (43) is obtained \( d \) where step (a) is from \( \mathbb{E}[t_{k,i}(d)] \geq 0 \). By plugging (45) into (43), we obtain an upper bound w.r.t. \( \mathbb{E}[\|e_{ds,k}^{(t)}\|^2] \) in (44).

Next, we derive the upper bound of \( \mathbb{E}[\mathcal{F}(w^{(t+1)}) - \mathcal{F}(w^*)] \). We take summation on both sides of (21) to obtain the following inequality:

\[
\mathbb{E}[\mathcal{F}(w^{(t+1)})] \leq \mathbb{E}[\mathcal{F}(w^{(t)})] - \frac{1}{2\omega} \sum_{k=1}^{K} \left( \mathbb{E}[\|\nabla F_k(w_k^{(t)})\|^2] \right) - 2\mathbb{E}[\|e_{ds,k}^{(t)}\|^2 + \mathbb{E}[\|e_{com,k}^{(t)}\|^2]),
\]

where \( \omega = \max_{k} \lambda_k \omega_k \) and \( \mu = \min_{k} \mu_k \). Substituting \( \mathbb{E}[\|e_{ds,k}^{(t)}\|^2] \) and \( \mathbb{E}[\|e_{com,k}^{(t)}\|^2] \) respectively with (22) and (44), we have an upper bound of the overall loss function \( F_k(\cdot) \) at the round \( (t+1) \) as

\[
\mathbb{E}[\mathcal{F}(w^{(t)})] \leq \mathbb{E}[\mathcal{F}(w^{(t)})] - \frac{1}{2\omega} \sum_{k=1}^{K} \left( \mathbb{E}[\|\nabla F_k(w_k^{(t)})\|^2] \right) - 2\mathbb{E}[1\mathbb{E}[\|e_{ds,k}^{(t)}\|^2 + \mathbb{E}[\|e_{com,k}^{(t)}\|^2])].
\]

Plugging (48) into (47) and subtracting \( \mathcal{F}(w^*) \) on the both sides of (47), we obtain (26).

REFERENCES

[1] C. Zhong, H. Yang, and X. Yuan, “Multi-task federated learning with over-the-air computation for MIMO interference channels,” in Proc. Int. Symp. Netw. Comput. Commun. (ISNCC), Jul. 2022, pp. 1–6.

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[3] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent trends in deep learning based natural language processing,” IEEE Comput. Intell. Mag., vol. 13, no. 5, pp. 55–75, Aug. 2018.

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

[5] T. Tu, D. T. Ngo, N. H. Tran, H. Q. Ngo, M. N. Dao, and R. H. Middleton, “Cell-free massive MIMO for wireless federated learning,” IEEE Trans. Wireless Commun., vol. 19, no. 10, pp. 6377–6392, Oct. 2020.

[6] T. Y. Shi, D. T. Ngo, H. Q. Ngo, M. N. Dao, N. H. Tran, and R. H. Middleton, “Joint resource allocation to minimize execution time of federated learning in cell-free massive MIMO,” IEEE Internet Things J., vol. 9, no. 21, pp. 21736–21750, Nov. 2022.

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

[8] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, “Energy-efficient resource management for federated edge learning with CPU-GPU heterogeneous computing,” IEEE Trans. Wireless Commun., vol. 20, no. 12, pp. 7947–7962, Dec. 2021.

[9] Q.-V. Pham, M. Le, T. Huynh-The, Z. Han, and W.-J. Hwang, “Energy-efficient federated learning over UAV-enabled wireless powered communications,” IEEE Trans. Veh. Technol., vol. 71, no. 5, pp. 4977–4990, May 2022.

[10] Y. Wu, Y. Song, T. Wang, L. Qian, and T. Q. S. Quek, “Non-orthogonal multiple access assisted federated learning via wireless power transfer: A cost-efficient approach,” IEEE Trans. Commun., vol. 70, no. 4, pp. 2853–2869, Apr. 2022.

[11] P. S. Bouzouinas, P. D. Diamantoulakis, and G. K. Karagiannidis, “Wireless federated learning (WFL) for 6G networks—II: The compute-then-transmit NOMA paradigm,” IEEE Commun. Lett., vol. 26, no. 1, pp. 8–12, Jan. 2022.

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

[13] M. M. Amiri and D. Gunduz, “Federated learning over wireless fading channels,” IEEE Trans. Wireless Commun., vol. 19, no. 5, pp. 3546–3557, May 2020.

[14] G. Zhu, Y. Du, D. Gunduz, 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.

[15] G. Zhu, L. Chen, and K. Huang, “MIMO over-the-air computation: Beamforming optimization on the Grassmann manifold,” in Proc. IEEE Global Commun. Conf. (GLOBECOM), Dec. 2018, pp. 1–6.

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

[17] 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.

[18] H. Liu, X. Yuan, and Y.-J.-A. Zhang, “CSIT-free model aggregation for federated edge learning via reconfigurable intelligent surface,” IEEE Wireless Commun. Lett., vol. 10, no. 11, pp. 2440–2444, Nov. 2021.

[19] X. Yang, T. Jiang, Y. Fang, and X. Ding, “Federated learning via over-the-air computation,” IEEE Trans. Wireless Commun., vol. 19, no. 3, pp. 2022–2035, Mar. 2020.

[20] Y. Shi, Y. Zhou, and Y. Shi, “Over-the-air decentralized federated learning,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jul. 2021, pp. 455–460.

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

[22] S. Han, J. Pool, J. Tran, and W. Dally, “Learning both weights and connections for efficient neural network,” in Proc. Adv. Neural Inf. Process. Syst., vol. 28, 2015, pp. 1–9.

[23] C.-Y. Chen et al., “ScaleCom: Scalable sparsified gradient compression for communication-efficient distributed training,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 13551–13563.

[24] D. Fan, X. Yuan, and Y.-J.-A. Zhang, “Temporal-structure-assisted gradient aggregation for over-the-air federated edge learning,” IEEE J. Sel. Areas Commun., vol. 39, no. 12, pp. 3757–3771, Dec. 2021.

[25] X. Cao, G. Zhu, J. Xu, and K. Huang, “Cooperative interference management for over-the-air computation networks,” IEEE Wireless Commun., vol. 20, no. 4, pp. 2634–2651, Apr. 2021.

[26] 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.

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

[28] LTE: Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures, 3GPP, document TS 36.213, Sep. 2017.

[29] S. L. H. Nguyen and A. Ghrayeb, “Compressive sensing-based channel estimation for massive multiuser MIMO systems,” in Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), Apr. 2013, pp. 2890–2895.
[30] C.-K. Wen, S. Jin, K.-K. Wong, J.-C. Chen, and P. Ting, “Channel estimation for massive MIMO using Gaussian-mixture Bayesian learning,” IEEE Trans. Wireless Commun., vol. 14, no. 3, pp. 1356–1368, Mar. 2015.

[31] Z. Lin, X. Li, V. K. N. Lau, Y. Gong, and K. Huang, “Deploying federated learning in large-scale cellular networks: Spatial convergence analysis,” IEEE Trans. Wireless Commun., vol. 21, no. 3, pp. 1542–1556, Mar. 2022.

[32] 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.

[33] D. P. Bertsekas and J. N. Tsitsiklis, “Neuro-dynamic programming: An overview,” in Proc. 34th IEEE Conf. Decis. Control, vol. 1, Dec. 1995, pp. 560–564.

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

[35] Y. LeCun, C. Cortes, and C. Burges. (1998). The MNIST Database of Handwritten Digits. [Online]. Available: http://yann.lecun.com/exdb/mnist

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

[37] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, “Deep learning for classical Japanese literature,” 2018, arXiv:1812.01718.

[38] K. Shen and W. Yu, “Fractional programming for communication systems—Part I: Power control and beamforming,” IEEE Trans. Signal Process., vol. 66, no. 10, pp. 2616–2630, May 2018.

[39] A. Goldsmith, Wireless Communication. Cambridge, U.K.: Cambridge Univ. Press, Aug. 2005.

[40] Q. Wu and R. Zhang, “Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming,” IEEE Trans. Wireless Commun., vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

Chenxi Zhong (Graduate Student Member, IEEE) received the B.S. degree in communication engineering from the School of Information and Communication Engineering, University of Electronic Science and Technology of China, in 2021. He is currently pursuing the M.S. degree in electrical and information engineering with the National Key Laboratory of Science and Technology on Communications. His current research interests include MIMO assisted wireless communications and distributed learning.

Huiyuan Yang received the B.S. degree in communication engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2019, where he is currently pursuing the Ph.D. degree. His current research interests include reconfigurable intelligent surfaces assisted wireless communications and distributed learning.

Xiaojun Yuan (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the City University of Hong Kong in 2009. From 2009 to 2011, he was a Research Fellow at the Department of Electronic Engineering, City University of Hong Kong. He was also a Visiting Scholar at the Department of Electrical Engineering, University of Hawaii at Manoa, in spring and summer 2009, as well as in the same period of 2010. From 2011 to 2014, he was a Research Assistant Professor with the Institute of Network Coding, The Chinese University of Hong Kong. From 2014 to 2017, he was an Assistant Professor with the School of Information Science and Technology, Shanghai Tech University. He is currently a State-Specially Recruited Professor with the University of Electronic Science and Technology of China. His research interests include signal processing, machine learning, and wireless communications, including but not limited to intelligent communications, structured signal reconstruction, Bayesian approximate inference, and distributed learning. He has published over 220 peer-reviewed research papers in the leading international journals and conferences in the related areas. He has served on several technical programs for international conferences. He was a co-recipient of the Best Paper Award of IEEE International Conference on Communications (ICC) 2014, the Best Journal Paper Award of IEEE Technical Committee on Green Communications and Computing (TCGCC) 2017, and the IEEE Heinrich Hertz Award for Best Communication Letter 2022. He was an Editor of IEEE leading journals, including IEEE Transactions on Wireless Communications and IEEE Transactions on Communications.