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

Chenxi Zhong, Huiyuan Yang, and Xiaojun Yuan, *Senior Member, IEEE*

arXiv:2112.13603v2 [eess.SP] 6 May 2022

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

Multi-task federated learning, over-the-air model aggregation, multiple-input multiple-output multiple access channel, alternating optimization, fractional programming.

C. Zhong, H. Yang and X. Yuan are with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu, China (e-mail: cxzhong@std.uestc.edu.cn; hyyang@std.uestc.edu.cn; xjyuan@uestc.edu.cn). The corresponding author is Xiaojun Yuan.

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 [1] and natural language processing [2]. 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 [3]. 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 upload 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 through wireless channels. Recently, over-the-air computation (AirComp) has been adopted to improve the communication efficiency of the FL uplink. [4] In over-the-air FL (OA-FL), edge devices transmit their local gradients by using shared radio resources, and aggregate them over the air by utilizing the superposition property of electromagnetic waves. Pioneering work has confirmed that over-the-air FL has strong noise tolerance [5] and reduces latency substantially compared with the schemes based on conventional orthogonal multiple access (OMA) protocols [6]–[9].

However, given the above mentioned benefits, introducing AirComp into FL uplink also brings some tricky problems, such as the so-called straggler[1] problem [6], [7]. To aggregate the local gradients at the PS, 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 [7], [10], [11]. 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. [6] and [7] propose to exclude the stragglers from model aggregation to alleviate this problem. However, this exclusion of devices reduces

---

[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 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 multi-tasking situation. Specifically, when multiple tasks are performed simultaneously at the wireless edge[2], the communication cost of this multi-task FL 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) [13], [14]. The temporal correlation has been exploited to improve the uplink efficiency of OA-FL in [15], whereas an efficient use of the spatial correlation has not been well explored. The existing approaches in OA-FL [10], [16] 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 analyse 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

---

[2]The fast development of intelligent systems spawns a large number of FL tasks to meet various demands of intelligence [12].

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 analyse 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. $(\cdot)^{\dagger}$, $(\cdot)^{\mathrm{T}}$, and $(\cdot)^{\mathrm{H}}$ 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 $\mathbf{s}$, $\mathbf{s}(1:D)$ to denote a sub-vector of $\mathbf{s}$ with entries indexed from 1 to D, $\mathcal{CN}(\mu, \sigma^2)$ to denote the circularly-symmetric complex normal distribution with mean μ and covariance σ^2, $\mathbb{E}[\cdot]$ to denote the expectation operator, and $|\mathcal{S}|$ to denote the cardinality of set $\mathcal{S}$. $\mathbf{I}_N$, $\mathbf{1}_{N \times M}$, and $\mathbf{0}_{N \times M}$ 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[3], $k \in [K]$, 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.

Fig. 1. FL interference network with two tasks.

A. Federated Learning System

We start with the description of the K-task FL system. Let $\mathcal{A}_k$ 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 $\mathcal{A}_k$. Each FL task k has an empirical loss function based on $\mathcal{A}_k$, given by

$$F_k(\mathbf{w}_k) \triangleq \frac{1}{Q_k} \sum_{n=1}^{Q_k} f_k\left(\mathbf{w}_k; \boldsymbol{\xi}_{k,n}\right), \ \forall k \in [K], \tag{1}$$

where $\mathbf{w}_k \in \mathbb{R}^D$ denotes the parameter vector of task k, [4], $\boldsymbol{\xi}_{k,n} \in \mathcal{A}_k$ denotes the n-th training sample in $\mathcal{A}_k$, and $f_k\left(\mathbf{w}_k; \boldsymbol{\xi}_{k,n}\right)$ denotes the sample-wise loss function with respect to (w.r.t.) $\boldsymbol{\xi}_{k,n}$. Let subscript $\langle k, i \rangle$ denote the index of the i-th device in the k-th group, $\forall k \in [K]$, $i \in [M_k]$. We assume that the local training datasets of each device $\langle k, i \rangle$ in group k are independently

---

[3]For simplicity, we assume that each device is assigned to a single task. The extension to the case of one device serving multiple tasks is straightforward. We omit detailed discussions due to space limitation.

[4]If the lengths of the parameter vectors vary among the tasks, zero padding is employed to ensure that $\{\mathbf{w}_k\}_{k=1}^{K}$ have identical length D.

and randomly drawn from $\mathcal{A}_k$, $k \in [K]$. Let $Q_{\langle k,i\rangle}$ denote the sample size of the $\langle k, i\rangle$-th device. Naturally, we have $Q_k = \sum_{i=1}^{M_k} Q_{\langle k,i\rangle}$. Then $F_k(\mathbf{w}_k)$ can be rewritten as

$$F_k(\mathbf{w}_k) = \frac{1}{Q_k} \sum_{i=1}^{M_k} Q_{\langle k,i\rangle} F_{\langle k,i\rangle}(\mathbf{w}_k), \ \forall k \in [K], \tag{2}$$

where $F_{\langle k,i\rangle}(\mathbf{w}_k) \triangleq \sum_{n=1}^{Q_{\langle k,i\rangle}} f_k\left(\mathbf{w}_k; \boldsymbol{\xi}_{\langle k,i\rangle,n}\right)/Q_{\langle k,i\rangle}$ denotes the local loss function of the $\langle k, i\rangle$-th device with $\boldsymbol{\xi}_{\langle k,i\rangle,n}$ being the n-th training sample on the $\langle k, i\rangle$-th device. The overall loss function of the considered OA-FMTL is given by

$$\mathcal{F}(\mathbf{w}) \triangleq \sum_{k=1}^{K} F_k(\mathbf{w}_k), \tag{3}$$

where $\mathbf{w} \triangleq [\mathbf{w}_1^{\mathrm{T}}, \ldots, \mathbf{w}_K^{\mathrm{T}}]^{\mathrm{T}}$.

OA-FMTL aims to minimize $\mathcal{F}(\mathbf{w})$ by separately minimize each $F_k(\mathbf{w}_k)$ using gradient descent (GD) [3]. 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 sub-group $\mathcal{M}_k$ of group k to participate in the updating.
- *Global model broadcasting*: The PS broadcasts *global model* $\mathbf{w}_k^{(t)}$ 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 $\mathbf{g}_{\langle k,i\rangle}^{(t)}$ by

$$\mathbf{g}_{\langle k,i\rangle}^{(t)} \triangleq \nabla F_{\langle k,i\rangle}(\mathbf{w}_k^{(t)}) = \frac{1}{Q_{\langle k,i\rangle}} \sum_{n=1}^{Q_{\langle k,i\rangle}} \nabla f_k\left(\mathbf{w}_k^{(t)}; \boldsymbol{\xi}_{\langle k,i\rangle,n}\right), \forall i \in \mathcal{M}_k. \tag{4}$$

- *Local gradients uploading*: Selected devices upload their local gradients to the PS.
- *Global model updating*: The PS computes the aggregated gradient $\mathbf{g}_k^{(t)}$ by

$$\mathbf{g}_k^{(t)} \triangleq \frac{\sum_{i \in \mathcal{M}_k} Q_{\langle k,i\rangle} \mathbf{g}_{\langle k,i\rangle}^{(t)}}{\sum_{i \in \mathcal{M}_k} Q_{\langle k,i\rangle}}. \tag{5}$$

The PS then updates the model $\mathbf{w}_k^{(t)}$ by

$$\mathbf{w}_k^{(t+1)} = \mathbf{w}_k^{(t)} - \eta_k \mathbf{g}_k^{(t)}, \tag{6}$$

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 [17]–[19], we assume the global model is broadcast to the devices via error-free links and the local gradient uploading is synchronized among the devices [5]. 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[6].

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

$$\mathbf{Y}^{(t)} = \sum\nolimits_{k=1}^{K} \sum\nolimits_{i \in \mathcal{M}_k} \mathbf{H}_{\langle k,i \rangle}^{(t)} \mathbf{X}_{\langle k,i \rangle}^{(t)} + \mathbf{N}^{(t)}, \tag{7}$$

where $\mathbf{H}_{\langle k,i \rangle}^{(t)} \in \mathbb{C}^{N_R \times N_T}$ denotes the channel matrix between the $\langle k, i \rangle$-th device and the PS, $\mathbf{X}_{\langle k,i \rangle}^{(t)} \triangleq \left[\mathbf{x}_{\langle k,i \rangle}^{(t)}[1], \cdots, \mathbf{x}_{\langle k,i \rangle}^{(t)}[C] \right] \in \mathbb{C}^{N_T \times C}$ with $\mathbf{x}_{\langle k,i \rangle}^{(t)}[c]$ denotes the signal transmitted by the $\langle k, i \rangle$-th device at the c-th channel use, $\mathbf{N}^{(t)} \triangleq [\mathbf{n}^{(t)}[1], \cdots, \mathbf{n}^{(t)}[C]] \in \mathbb{C}^{N_R \times C}$ is an additive white Gaussian noise (AWGN) matrix whose entries are independently drawn from $\mathcal{CN}(0, \sigma^2)$, and $\mathbf{Y}^{(t)} \triangleq \left[\mathbf{y}^{(t)}[1], \cdots, \mathbf{y}^{(t)}[C] \right] \in \mathbb{C}^{N_R \times C}$ with $\mathbf{y}^{(t)}[c]$ denotes the received signal at the PS at the c-th channel use.

Note that the two sums in (7) imply that all the edge devices are allowed to participate in training transmit 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*

---

[5]The synchronization can be realized by using the existing techniques, e.g., the timing-advance mechanism for uplink synchronization in 4G Long Term Evolution (LTE) [20].

[6]We assume that perfect CSI is available at the PS. Studies on CSI acquisition over MIMO MAC channels can be found, e.g., in [21], [22].

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 $\{\mathbf{g}^{(t)}_{\langle k,i\rangle} \mid k \in [K], i \in \mathcal{M}_k\}$ are first pre-processed to yield signals $\{\mathbf{X}^{(t)}_{\langle k,i\rangle} \mid k \in [K], i \in \mathcal{M}_k\}$, which are then transmitted to the PS via the underlying communication channel in (7). After the PS receives $\mathbf{Y}^{(t)}$, it separately performs post-processing to obtain $\{\hat{\mathbf{g}}^{(t)}_k\}^K_{k=1}$, which are estimates of the desired aggregated gradients $\{\mathbf{g}^{(t)}_k\}^K_{k=1}$. In the following, we introduce the pre-processing and post-processing steps.

Fig. 2. Over-the-air gradient aggregation.

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

$$\tilde{\mathbf{g}}^{(t)}_{\langle k,i\rangle}[d] = \left(\mathbf{g}^{(t)}_{\langle k,i\rangle} - \bar{g}^{(t)}_{\langle k,i\rangle}\mathbf{1}_{D\times1}\right)\Big/\sqrt{v^{(t)}_{\langle k,i\rangle}}, \tag{8}$$

where $\bar{g}^{(t)}_{\langle k,i\rangle} = \frac{1}{D}\sum_{d=1}^{D} g^{(t)}_{\langle k,i\rangle}[d]$ and $v^{(t)}_{\langle k,i\rangle} = \frac{1}{D}\sum_{d=1}^{D}\left|g^{(t)}_{\langle k,i\rangle}[d] - \bar{g}^{(t)}_{\langle k,i\rangle}\right|^2$. Following the common practice [7], [23], we assume $\{\bar{g}^{(t)}_{\langle k,i\rangle}, v^{(t)}_{\langle k,i\rangle}|\forall k,i,t\}$ 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 $\tilde{\mathbf{g}}^{(t)}_{\langle k,i\rangle}$ to a complex data vector $\mathbf{r}^{(t)}_{\langle k,i\rangle}$, i.e.,

$$\mathbf{r}^{(t)}_{\langle k,i\rangle} \triangleq \tilde{\mathbf{g}}^{(t)}_{\langle k,i\rangle}\left(1:\frac{D}{2}\right) + j\tilde{\mathbf{g}}^{(t)}_{\langle k,i\rangle}\left(\frac{D+2}{2}:D\right) \in \mathbb{C}^C, \tag{9}$$

where $C = D/2$[7]. We transmit $\mathbf{r}^{(t)}_{\langle k,i\rangle}$ with C times of channel use (one element of $\mathbf{r}^{(t)}_{\langle k,i\rangle}$ for one time of channel use). Specifically, the transmitted signal is given by

$$\mathbf{X}^{(t)}_{\langle k,i\rangle} \triangleq \mathbf{u}^{(t)}_{\langle k,i\rangle}\mathbf{r}^{(t)}_{\langle k,i\rangle}{}^{\mathrm{T}} \in \mathbb{C}^{N_{\mathrm{T}}\times C}, \tag{10}$$

---

[7]For simplicity, we assume that D is even.

where $\mathbf{u}_{\langle k,i \rangle} \in \mathbb{C}^{N_\mathrm{T}}$ is the transmit beamforming vector of the $\langle k,i \rangle$-th device. Recall that $\mathbf{X}_{\langle k,i \rangle}^{(t)} = \left[\mathbf{x}_{\langle k,i \rangle}^{(t)}[1], \cdots, \mathbf{x}_{\langle k,i \rangle}^{(t)}[C] \right]$. Thus we have $\mathbf{x}_{\langle k,i \rangle}^{(t)}[c] = r_{\langle k,i \rangle}^{(t)}[c] \mathbf{u}_{\langle k,i \rangle}$ denoting the transmitted signal of the $\langle k,i \rangle$-th device at the c-th channel use, satisfying the following power constraint:

$$\mathbb{E}\left[\|\mathbf{x}_{\langle k,i \rangle}^{(t)}[c]\|^2 \right] = 2\|\mathbf{u}_{\langle k,i \rangle}\|^2 \le P_0, \tag{11}$$

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

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

$$\hat{\mathbf{r}}_k^{(t)} = \zeta_k \left(\mathbf{f}_k^\mathrm{H} \mathbf{Y}_k^{(t)} \right)^\mathrm{T} = \zeta_k \Big(\sum_{i \in \mathcal{M}_k} \mathbf{r}_{\langle k,i \rangle} (\mathbf{H}_{\langle k,i \rangle}^{(t)} \mathbf{u}_{\langle k,i \rangle})^\mathrm{T} + \sum_{l \ne k} \sum_{i \in \mathcal{M}_l} \mathbf{r}_{\langle l,i \rangle}^{(t)} (\mathbf{H}_{\langle l,i \rangle}^{(t)} \mathbf{u}_{\langle l,i \rangle})^\mathrm{T} + \mathbf{N}^{(t)\mathrm{T}} \Big) \mathbf{f}_k^\dagger, k \in [K], \tag{12}$$

where $\zeta_k \in \mathbb{R}$ is a weighting factor and $\mathbf{f}_k \in \mathbb{C}^{N_\mathrm{R}}$ is the normalized receive beamforming vector of task k with $\|\mathbf{f}_k\| = 1$. Estimates of the desired aggregated gradients $\{\hat{\mathbf{g}}_k^{(t)}\}_{k=1}^K$ is reconstructed from $\{\hat{\mathbf{r}}_k^{(t)}\}_{k=1}^K$ by

$$\hat{\mathbf{g}}_k^{(t)} = \frac{1}{\sum_{i \in \mathcal{M}_k} Q_{\langle k,i \rangle}} \left[\mathrm{Re}\{\hat{\mathbf{r}}_k^{(t)}\}^\mathrm{T}, \, \mathrm{Im}\{\hat{\mathbf{r}}_k^{(t)}\}^\mathrm{T} \right]^\mathrm{T} + \bar{g}_k^{(t)} \mathbf{1}_{D \times 1} \in \mathbb{R}^D, \ k \in [K], \tag{13}$$

where $\bar{g}_k^{(t)} \triangleq \left(\sum_{i \in \mathcal{M}_k} Q_{\langle k,i \rangle} \bar{g}_{\langle k,i \rangle}^{(t)} \right) / \left(\sum_{i \in \mathcal{M}_k} Q_{\langle k,i \rangle} \right)$.

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 $\{\hat{\mathbf{g}}_k^{(t)}\}_{k=1}^K$ inevitably suffer from distortions caused by the inter-task interference and the channel noise. The global model $\mathbf{w}_k^{(t)}$ of task k is updated with the gradient aggregation $\hat{\mathbf{g}}_k^{(t)}$:

$$\mathbf{w}_k^{(t+1)} = \mathbf{w}_k^{(t)} - \eta_k \hat{\mathbf{g}}_k^{(t)} = \mathbf{w}_k^{(t)} - \eta_k \left(\nabla F_k(\mathbf{w}_k^{(t)}) - \mathbf{e}_k^{(t)} \right), \tag{14}$$

where $\nabla F_k(\mathbf{w}_k^{(t)}) \triangleq \frac{1}{Q_k} \sum_{n=1}^{Q_k} \nabla f_k \left(\mathbf{w}_k; \boldsymbol{\xi}_{k,n} \right)$ is the gradient of the loss function of the k-th task $F_k(\mathbf{w}_k)$ at $\mathbf{w}_k = \mathbf{w}_k^{(t)}$, and $\mathbf{e}_k^{(t)}$ is the error caused by gradient uploading, which can be divided into two parts:

$$\mathbf{e}_k^{(t)} = \underbrace{\nabla F_k(\mathbf{w}_k^{(t)}) - \mathbf{g}_k^{(t)}}_{\mathbf{e}_{\mathrm{ds},k}^{(t)}} + \underbrace{\mathbf{g}_k^{(t)} - \hat{\mathbf{g}}_k^{(t)}}_{\mathbf{e}_{\mathrm{com},k}^{(t)}}, \tag{15}$$

where $\mathbf{e}_{\mathrm{ds},k}^{(t)}$ denotes the error caused by device selection, and $\mathbf{e}_{\mathrm{com},k}^{(t)}$ denotes the communication error due to the inter-task interference and the noise. To alleviate the impact of $\mathbf{e}_k^{(t)}$ 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: $\mathcal{M} \triangleq \{\mathcal{M}_k\}_{k=1}^K$, $\mathbf{f} \triangleq \{\mathbf{f}_k\}_{k=1}^K$, $\mathbf{u}_k \triangleq \{\mathbf{u}_{\langle k,i \rangle}\}_{i \in \mathcal{M}_k}$, and $\mathbf{u} \triangleq \{\mathbf{u}_k\}_{k=1}^K$. In the following section, we establish the connection between the model aggregation errors in (15) and the learning performance of the OA-FMTL.

Algorithm 1 OA-FMTL framework

Input: T, $\{Q_{\langle k,i \rangle}\}$.

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

III. PERFORMANCE ANALYSIS

In this section, we analyse 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 [24], [25], 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 ω_k, i.e.,*

$$\|\nabla F_k(\mathbf{w}) - \nabla F_k(\mathbf{w}')\| \leq \omega_k \|\mathbf{w} - \mathbf{w}'\|, \forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^D, k \in [K]. \tag{16}$$

Assumption 2. *For each task k, loss function F_k is strongly convex with positive parameter μ_k:*

$$F_k(\mathbf{w}) \geq F_k(\mathbf{w}') + (\mathbf{w} - \mathbf{w}')^{\mathrm{T}} \nabla F_k(\mathbf{w}') + \frac{\mu_k}{2} \|\mathbf{w} - \mathbf{w}'\|^2, \forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^D, k \in [K]. \tag{17}$$

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

Assumption 4. *The gradient vector is upper bounded by*

$$\|\nabla f_k\left(\mathbf{w}_k; \boldsymbol{\xi}_{k,n}\right)\|^2 \leq \beta_1 + \beta_2 \|\nabla F_k(\mathbf{w}_k^{(t)})\|^2, \forall k \in [K], \tag{18}$$

for some constants $\beta_1 \geq 0$ and $\beta_2 \geq 0$. Both β_1 and β_2 are constants shared by $\{F_k(\cdot)\}_{k=1}^K$.

Fig. 3. An illustration of the gradients in the OA-FMTL

Furthermore, unlike the approaches in [10], [16] 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 $\tilde{\mathbf{G}}_k^{(t)} \triangleq [\tilde{\mathbf{g}}_{\langle k,1 \rangle}^{(t)}, \cdots, \tilde{\mathbf{g}}_{\langle k, M_k \rangle}^{(t)}] \in \mathbb{R}^{D \times M_k}$ be the local gradients from the M_k devices of task k at the t-th round. Let $\mathbf{z}_{k,d}^{(t)} \in \mathbb{R}^{M_k}$ be the d-th dimension of the local gradients from a common task, and $\mathbf{z}_{k,d}^{(t) \, \mathrm{T}}$ is the d-th row of $\tilde{\mathbf{G}}_k^{(t)}$, 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 $\{\tilde{\mathbf{G}}_k^{(t)} | k \in [K]\}$ are independent and non-identically distributed. For the k-th task, the gradients $\{\mathbf{z}_{k,d}^{(t)} | d \in [D]\}$, are*

independent and identically distributed. That is,

$$p^{(t)}\left(\{\tilde{\mathbf{G}}_k^{(t)}|k \in [K]\}\right) = \prod_{k=1}^{K}\prod_{d=1}^{D} p_k^{(t)}\left(\mathbf{z}_{k,d}^{(t)}\right), \forall d, \tag{19}$$

where $p^{(t)}(\cdot)$ denotes the distribution of the elements of $\{\tilde{\mathbf{G}}_k^{(t)}|k \in [K]\}$, and $p_k^{(t)}(\cdot)$ denotes the distribution of the elements of $\mathbf{z}_{k,d}^{(t)}$. Furthermore, for each task k, the local gradients of the M_k devices have the same degree of variation, i.e., $v_{\langle k,1\rangle}^{(t)} = \cdots = v_{\langle k,M_k\rangle}^{(t)} = v_k^{(t)}$.

We now focus on the spatial correlation between the local gradients in a common task, i.e., between the entries of $\mathbf{z}_{k,d}^{(t)}$. The auto-correlation matrix for task k is then defined by

$$\boldsymbol{\rho}_k^{(t)} \triangleq \mathbb{E}\left[\mathbf{z}_{k,d}^{(t)}(\mathbf{z}_{k,d}^{(t)})^{\mathrm{T}}\right] \in \mathbb{R}^{M_k \times M_k}, \forall d \tag{20}$$

where the (i,j)-th entry is denoted as $\rho_{\langle k,i\rangle,\langle k,j\rangle}^{(t)}$, measuring the spatial correlation between device i and device j of task k. Note that each element of $\mathbf{z}_{k,d}^{(t)}$, $\tilde{g}_{\langle k,i\rangle}^{(t)}[d]$, has zero mean and unit variance. Thus, $\rho_{\langle k,i\rangle,\langle k,i\rangle}^{(t)} = 1$, and $\rho_{\langle k,i\rangle,\langle k,j\rangle}^{(t)} \in [-1,1]$, for $\forall k,i,j$.

Fig. 4. Experimental results of $\boldsymbol{\rho}_k^{(t)}$ versus communication rounds t. We train three LeNet [26]-based FL models, with $M_1 = M_2 = M_3 = 10$ and 20 Monte Carlo trials. The learning rate is set to $\eta_1 = \eta_2 = \eta_3 = 0.002$, and the momentum is set to 0.9. The loss function is the cross-entropy loss. Each training dataset is assigned 60000 samples and the local training data is assigned 6000 samples i.i.d. drawn from the dataset. The local updates have 5 times of stochastic gradient descents (SGD), and the mini-batch size is set to be 1200. The gradients uploading is error-free and all devices are selected. We approximate $\rho_{\langle k,i\rangle,\langle k,j\rangle}^{(t)}$ by $\rho_{\langle k,i\rangle,\langle k,j\rangle}^{(t)} \approx \frac{1}{D}\sum_{d=1}^{D} \tilde{g}_{\langle k,i\rangle}^{(t)}[d]\tilde{g}_{\langle k,j\rangle}^{(t)}[d]$.

The heatmaps in Fig. 4 illustrate the experimental results of $\boldsymbol{\rho}_k^{(t)}$ versus the communication round t of three datasets, MNIST [27], Fashion-MNIST [28], and KMNIST [29], where 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 $\boldsymbol{\rho}_k^{(t)}$ 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 [24, 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*

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

where the device selection MSE is bounded by

$$\mathbb{E}[\|\mathbf{e}_{\mathrm{ds},k}^{(t)}\|^2] \leq \frac{4}{Q_k^2}\left(Q_k - \sum_{i \in \mathcal{M}_k} Q_{\langle k,i\rangle}\right)^2 \left(\beta_1 + \beta_2 \mathbb{E}\left[\|\nabla F_k(\mathbf{w}_k^{(t)})\|^2\right]\right), \quad (22)$$

and the communication MSE is given by

$$\mathbb{E}[\|\mathbf{e}_{\mathrm{com},k}^{(t)}\|^2] = \frac{1}{(\sum_{i \in \mathcal{M}_k} Q_{\langle k,i\rangle})^2} \sum_{c=1}^{C}\left(\underbrace{\mathbb{E}\left[\left|\sum_{i \in \mathcal{M}_k}\left(Q_{\langle k,i\rangle}\sqrt{v_k^{(t)}} - \zeta_k \mathbf{f}_k^{\mathrm{H}}\mathbf{H}_{\langle k,i\rangle}^{(t)}\mathbf{u}_{\langle k,i\rangle}\right)r_{\langle k,i\rangle}^{(t)}[c]\right|^2\right]}_{\text{the first term: the misalignment error}}\right.$$

$$\left. + \underbrace{\zeta_k^2 \sum_{l \neq k} \mathbb{E}\left[\left|\sum_{i \in \mathcal{M}_l} \mathbf{f}_k^{\mathrm{H}}\mathbf{H}_{\langle l,i\rangle}^{(t)}\mathbf{u}_{\langle l,i\rangle}r_{\langle l,i\rangle}^{(t)}[c]\right|^2\right]}_{\text{the second term: the interference}} + \underbrace{\zeta_k^2 \mathbb{E}\left[\left|\mathbf{f}_k^{\mathrm{H}}\mathbf{n}^{(t)}[c]\right|^2\right]}_{\text{the third term: the noise}}\right). \quad (23)$$

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 $\mathcal{M}_k$ and beamforming $\mathbf{f}_k$ and $\{\mathbf{u}_{\langle k,i\rangle}\}_{i \in \mathcal{M}_k}$. Thus, we have the following corollary.

Corollary 1. *The optimal ζ_k is given by*

$$\zeta_k^* = \frac{\sqrt{v_k^{(t)}} \sum_{i,j \in \mathcal{M}_k} \rho_{\langle k,i \rangle, \langle k,j \rangle}^{(t)} \left(Q_{\langle k,i \rangle} (\mathbf{f}_k^{\mathrm{H}} \mathbf{H}_{\langle k,j \rangle}^{(t)} \mathbf{u}_{\langle k,j \rangle})^{\mathrm{H}} + Q_{\langle k,j \rangle} \mathbf{f}_k^{\mathrm{H}} \mathbf{H}_{\langle k,i \rangle}^{(t)} \mathbf{u}_{\langle k,i \rangle} \right)}{2 \left(\sum_{l=1}^K \sum_{i,j \in \mathcal{M}_l} \rho_{\langle l,i \rangle, \langle l,j \rangle}^{(t)} (\mathbf{f}_k^{\mathrm{H}} \mathbf{H}_{\langle l,i \rangle}^{(t)} \mathbf{u}_{\langle l,i \rangle})^{\mathrm{H}} \mathbf{f}_k^{\mathrm{H}} \mathbf{H}_{\langle l,j \rangle}^{(t)} \mathbf{u}_{\langle l,j \rangle} + \sigma^2 \|\mathbf{f}_k\|^2 / 2 \right)}. \qquad (24)$$

Proof. Please refer to Appendix B. □

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. [7], [10], [11] force this component to zero, which leads to the constraints of $Q_{\langle k,i \rangle} \sqrt{v_k^{(t)}} - \zeta_k \mathbf{f}_k^{\mathrm{H}} \mathbf{H}_{\langle k,i \rangle}^{(t)} \mathbf{u}_{\langle k,i \rangle} = 0, \forall i \in \mathcal{M}_k$, and then minimize the rest of the communication MSE to determine ζ_k. For all devices in the k-th task, to satisfy the constraints and the transmit power budgets in (11), the choice of ζ_k is therefore given by $\zeta_k^2 = \max_{i \in \mathcal{M}_k} Q_{\langle k,i \rangle}^2 v_k^{(t)} / (P_0 \|\mathbf{f}_k^{\mathrm{H}} \mathbf{H}_{\langle k,i \rangle}^{(t)}\|^2)$. Thus, ζ_k is dominated by the device with the worst channel condition (in terms of $\|\mathbf{f}_k^{\mathrm{H}} \mathbf{H}_{\langle k,i \rangle}^{(t)}\|^2$). 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 tolerances the misalignment error but requires the design of ζ_k to directly minimize the overall communication MSE for task k. In this way, ζ_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 [16] since we aware 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 $\mathcal{F}\left(\mathbf{w}^{(t+1)}\right)$ and the optimal $\mathcal{F}\left(\mathbf{w}^*\right)$, i.e., $\mathbb{E}\left[\mathcal{F}\left(\mathbf{w}^{(t+1)}\right) - \mathcal{F}\left(\mathbf{w}^*\right)\right]$. 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 1-4, the overall loss function at the t-th round $\mathcal{F}(\mathbf{w}^{(t+1)})$*

satisfies the following inequality:

$$\mathbb{E}[\mathcal{F}(\mathbf{w}^{(t+1)}) - \mathcal{F}(\mathbf{w}^*)] \leq \mathbb{E}[\mathcal{F}(\mathbf{w}^{(t)}) - \mathcal{F}(\mathbf{w}^*)]\left(1 - \frac{\mu}{\omega}\right)$$
$$+ \left(\frac{2\mu\beta_2}{\omega}\mathbb{E}[\mathcal{F}(\mathbf{w}^{(t)}) - \mathcal{F}(\mathbf{w}^*)] + \frac{\beta_1}{\omega}\right)\mathcal{E}^{(t)}(\mathcal{M}, \mathbf{f}, \mathbf{u}), \qquad (25)$$

where $\omega \triangleq \max_k \omega_k$, $\mu \triangleq \min_k \mu_k$, $\mathcal{E}^{(t)}(\mathcal{M}, \mathbf{f}, \mathbf{u}) \triangleq \sum_{k=1}^K d_k^{(t)}(\mathcal{M}_k, \mathbf{f}_k, \mathbf{u}_k)$, *and* $d_k^{(t)}(\mathcal{M}_k, \mathbf{f}_k, \mathbf{u}_k)$ *is denoted by*

$$d_k^{(t)}(\mathcal{M}_k, \mathbf{f}_k, \mathbf{u}_k) \triangleq \frac{4}{Q_k^2}\left(Q_k - \sum_{i \in \mathcal{M}_k} Q_{\langle k,i\rangle}\right)^2 + \frac{1}{\left(\sum_{i \in \mathcal{M}_k} Q_{\langle k,i\rangle}\right)^2}\left(\sum_{i,j \in \mathcal{M}_k} \rho_{\langle k,i\rangle, \langle k,j\rangle}^{(t)} Q_{\langle k,i\rangle} Q_{\langle k,j\rangle}\right.$$
$$\left.- \frac{\left(\sum_{i,j \in \mathcal{M}_k} \rho_{\langle k,i\rangle, \langle k,j\rangle}^{(t)}\left(Q_{\langle k,i\rangle}(\mathbf{f}_k^{\mathrm{H}}\mathbf{H}_{\langle k,j\rangle}^{(t)}\mathbf{u}_{\langle k,j\rangle})^{\mathrm{H}} + Q_{\langle k,j\rangle}\mathbf{f}_k^{\mathrm{H}}\mathbf{H}_{\langle k,i\rangle}^{(t)}\mathbf{u}_{\langle k,i\rangle}\right)\right)^2}{4\left(\sum_{l=1}^K \sum_{i,j \in \mathcal{M}_l} \rho_{\langle l,i\rangle, \langle l,j\rangle}^{(t)}(\mathbf{f}_k^{\mathrm{H}}\mathbf{H}_{\langle l,i\rangle}^{(t)}\mathbf{u}_{\langle l,i\rangle})^{\mathrm{H}}\mathbf{f}_k^{\mathrm{H}}\mathbf{H}_{\langle l,j\rangle}^{(t)}\mathbf{u}_{\langle l,j\rangle} + \sigma^2\|\mathbf{f}_k\|^2/2\right)}\right). \qquad (26)$$

Proof. Please refer to Appendix C. □

Theorem 2 provides a metric for evaluating the learning performance of the OA-FMTL model. In the next section, we formulate the optimization problem based on this metric, and propose an efficient algorithm to solve this problem.

IV. SYSTEM OPTIMIZATION

To achieve better learning performance, we aim to minimize the upper bound of $\mathbb{E}[\mathcal{F}(\mathbf{w}^{(t+1)}) - \mathcal{F}(\mathbf{w}^*)]$ in (25), or equivalently, to minimize $\mathcal{E}^{(t)}(\mathcal{M}, \mathbf{f}, \mathbf{u})$ over device selection set $\mathcal{M}$, receive beamforming $\mathbf{f}$ and transmit beamforming $\mathbf{u}$, as detailed below.

A. Problem Formulation

From Theorem 2, the gap $\mathbb{E}[\mathcal{F}(\mathbf{w}^{(t+1)}) - \mathcal{F}(\mathbf{w}^*)]$ has an upper bound in (25). The upper bound is monotonically increasing w.r.t. $\mathcal{E}^{(t)}(\mathcal{M}, \mathbf{f}, \mathbf{u})$ since $\frac{2\mu\beta_2}{\omega}\mathbb{E}[\mathcal{F}(\mathbf{w}^{(t)}) - \mathcal{F}(\mathbf{w}^*)] + \frac{\beta_1}{\omega} > 0$. With the target to minimize the gap $\mathbb{E}[\mathcal{F}(\mathbf{w}^{(t+1)}) - \mathcal{F}(\mathbf{w}^*)]$, we minimize $\mathcal{E}^{(t)}(\mathcal{M}, \mathbf{f}, \mathbf{u})$ at round t, and

we formulate the optimization problem P1 as:

$$(\text{P1}): \min_{\mathcal{M},\mathbf{f},\mathbf{u}} \quad \mathcal{E}^{(t)}(\mathcal{M},\mathbf{f},\mathbf{u}) = \sum_{k=1}^{K} d_k^{(t)}(\mathcal{M}_k,\mathbf{f}_k,\mathbf{u}_k) \tag{27a}$$

$$\text{s.t.} \quad \mathcal{M}_k \subset [M_k], k \in [K], \tag{27b}$$

$$\|\mathbf{f}_k\| = 1, k \in [K], \tag{27c}$$

$$\left\|\mathbf{u}_{\langle k,i\rangle}\right\|^2 \le P_0/2, k \in [K], i \in [M_k], \tag{27d}$$

where

$$d_k^{(t)}(\mathcal{M}_k,\mathbf{f}_k,\mathbf{u}_k) = \frac{4}{Q_k^2}\left(Q_k - \sum_{i\in\mathcal{M}_k} Q_{\langle k,i\rangle}\right)^2 + \frac{\sum_{i,j\in\mathcal{M}_k}\rho_{\langle k,i\rangle,\langle k,j\rangle}^{(t)}Q_{\langle k,i\rangle}Q_{\langle k,j\rangle} - \frac{a_k^{(t)}(\mathcal{M}_k,\mathbf{f}_k,\mathbf{u}_k)^2}{4b_k^{(t)}(\mathcal{M}_k,\mathbf{f}_k,\mathbf{u}_k)}}{\left(\sum_{i\in\mathcal{M}_k} Q_{\langle k,i\rangle}\right)^2}, \tag{28}$$

with

$$a_k^{(t)}(\mathcal{M}_k,\mathbf{f}_k,\mathbf{u}_k) \triangleq \sum_{i,j\in\mathcal{M}_k}\rho_{\langle k,i\rangle,\langle k,j\rangle}^{(t)}\left(Q_{\langle k,i\rangle}(\mathbf{f}_k^{\mathrm{H}}\mathbf{H}_{\langle k,j\rangle}^{(t)}\mathbf{u}_{\langle k,j\rangle})^{\mathrm{H}} + Q_{\langle k,j\rangle}\mathbf{f}_k^{\mathrm{H}}\mathbf{H}_{\langle k,i\rangle}^{(t)}\mathbf{u}_{\langle k,i\rangle}\right), \tag{29a}$$

$$b_k^{(t)}(\mathcal{M}_k,\mathbf{f}_k,\mathbf{u}_k) \triangleq \sum_{l=1}^{K}\sum_{i,j\in\mathcal{M}_l}\rho_{\langle l,i\rangle,\langle l,j\rangle}^{(t)}(\mathbf{f}_k^{\mathrm{H}}\mathbf{H}_{\langle l,i\rangle}^{(t)}\mathbf{u}_{\langle l,i\rangle})^{\mathrm{H}}\mathbf{f}_k^{\mathrm{H}}\mathbf{H}_{\langle l,j\rangle}^{(t)}\mathbf{u}_{\langle l,j\rangle} + \sigma^2\|\mathbf{f}_k\|^2/2. \tag{29b}$$

P1 is an optimization problem w.r.t. device selection set $\mathcal{M}$, receive beamforming vectors $\mathbf{f}$ and transmit beamforming vectors $\mathbf{u}$, 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 $\mathcal{M}$, $\mathbf{f}$ and $\mathbf{u}$. 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 $\mathcal{M}$. Second, with fixed beamforming vectors, we optimize device selection set $\mathcal{M}$ with Gibbs sampling [7]. The two steps iterate until convergence. The details are discussed in the following subsections.

B. *Optimization of* $\mathbf{f}$ *and* $\mathbf{u}$ *with fixed* $\mathcal{M}$

We first optimize beamforming vectors $\mathbf{f}$ and $\mathbf{u}$ with fixed device selection set $\mathcal{M}$. By inspection of (P1), $d_k^{(t)}(\mathcal{M}_k,\mathbf{f}_k,\mathbf{u}_k)$ in (28) is invariant to the value of $\|\mathbf{f}_k\|$. Therefore, the unit-length constraint of $\mathbf{f}_k$ in (27c) can be ignored without changing the minimum of (P1).

Further, we drop the constant terms in the objective function $\mathcal{E}^{(t)}(\mathcal{M}, \mathbf{f}, \mathbf{u})$ to obtain

$$\min_{\{\mathbf{f}_k\}, \mathbf{u}} \quad -\sum_{k=1}^{K} \frac{a_k^{(t)}(\mathcal{M}_k, \mathbf{f}_k, \mathbf{u}_k)^2}{4 \left(\sum_{i \in \mathcal{M}_k} Q_{\langle k,i \rangle}\right)^2 b_k^{(t)}(\mathcal{M}_k, \mathbf{f}_k, \mathbf{u}_k)}, \text{ s.t. } (27\text{d}). \tag{30}$$

(30) is still non-convex because both the numerator and the denominator in the k-th summand contain the optimization variables $\mathbf{f}_k$ and $\mathbf{u}_k$. We adopt the quadratic transform in fractional programming (FP) [30] to decouple the numerator and the denominator as

$$\min_{\{\mathbf{f}_k\}, \mathbf{u}, \mathbf{y}} \quad -\sum_{k=1}^{K} \left(\frac{y_k a_k^{(t)}(\mathcal{M}_k, \mathbf{f}_k, \mathbf{u}_k)}{\sum_{i \in \mathcal{M}_k} Q_{\langle k,i \rangle}} - y_k^2 b_k^{(t)}(\mathcal{M}_k, \mathbf{f}_k, \mathbf{u}_k) \right), \text{ s.t. } (27\text{d}). \tag{31}$$

where $\mathbf{y} = [y_1, \cdots, y_k]^{\mathrm{T}} \in \mathbb{R}^K$ is an auxiliary vector introduced by FP. Note that (31) reduces to (30) by letting each y_k take its optimal form as

$$y_k = \frac{a_k^{(t)}(\mathcal{M}_k, \mathbf{f}_k, \mathbf{u}_k)}{2 \left(\sum_{i \in \mathcal{M}_k} Q_{\langle k,i \rangle}\right) b_k^{(t)}(\mathcal{M}_k, \mathbf{f}_k, \mathbf{u}_k)}. \tag{32}$$

1) Optimizing $\mathbf{u}_{\langle k,i \rangle}$ with fixed $\{\mathbf{u}_{\langle k,i \rangle}\}_{j \neq i}$ and $\{\mathbf{f}_k\}$: With fixed $\{\mathbf{u}_{\langle k,i \rangle}\}_{j \neq i}$ and $\{\mathbf{f}_k\}$, when $i \notin \mathcal{M}_k$, i.e., the device $\langle k, i \rangle$ is not selected, we obtain $\mathbf{u}_{\langle k,i \rangle} = \mathbf{0}_{N_{\mathrm{T}} \times 1}$ directly. When $i \in \mathcal{M}_k$, i.e., the device $\langle k, i \rangle$ is selected, (31) reduces to

$$(\text{P2}): \quad \min_{\mathbf{u}_{\langle k,i \rangle}} \quad \mathbf{u}_{\langle k,i \rangle}^{\mathrm{H}} \mathbf{A}_{\langle k,i \rangle}^{(t)} \mathbf{u}_{\langle k,i \rangle} - 2 \operatorname{Re} \left\{ \mathbf{b}_{\langle k,i \rangle}^{(t)\mathrm{H}} \mathbf{u}_{\langle k,i \rangle} \right\}, \text{ s.t. } (27\text{d}).$$

where $\mathbf{A}_{\langle k,i \rangle}^{(t)} \in \mathbb{C}^{N_{\mathrm{T}} \times N_{\mathrm{T}}}$ and $\mathbf{b}_{\langle k,i \rangle}^{(t)} \in \mathbb{C}^{N_{\mathrm{T}}}$ are defined by

$$\mathbf{A}_{\langle k,i \rangle}^{(t)} \triangleq \sum_{l=1}^{K} y_l^2 (\mathbf{H}_{\langle k,i \rangle}^{(t)})^{\mathrm{H}} \mathbf{f}_l \mathbf{f}_l^{\mathrm{H}} \mathbf{H}_{\langle k,i \rangle}^{(t)}, \tag{33a}$$

$$\mathbf{b}_{\langle k,i \rangle}^{(t)} \triangleq \frac{y_k \sum_{j \in \mathcal{M}_k} \rho_{\langle k,i \rangle, \langle k,j \rangle}^{(t)} Q_{\langle k,j \rangle} (\mathbf{f}_k^{\mathrm{H}} \mathbf{H}_{\langle k,i \rangle}^{(t)})^{\mathrm{H}}}{\sum_{i \in \mathcal{M}_k} Q_{\langle k,i \rangle}} - \sum_{l=1}^{K} y_l^2 \sum_{j \in \mathcal{M}_k, j \neq i} \rho_{\langle k,i \rangle, \langle k,j \rangle}^{(t)} (\mathbf{f}_l^{\mathrm{H}} \mathbf{H}_{\langle k,i \rangle}^{(t)})^{\mathrm{H}} \mathbf{f}_l^{\mathrm{H}} \mathbf{H}_{\langle k,j \rangle}^{(t)} \mathbf{u}_{\langle k,j \rangle}. \tag{33b}$$

Note that $\mathbf{A}_{\langle k,i \rangle}^{(t)}$ is a positive semidefinite matrix, and that the constraint (27d) is convex w.r.t. $\mathbf{u}_{\langle k,i \rangle}$. Thus, P2 is a convex quadratically constrained quadratic programming (QCQP) problem w.r.t. $\mathbf{u}_{\langle k,i \rangle}$, which can be solved by standard convex optimization tools.

2) Optimizing $\mathbf{f}_k$ with fixed $\mathbf{u}$: Similarly to IV-B1, with fixed $\mathbf{u}$, (31) reduces to

$$(\text{P3}): \quad \min_{\mathbf{f}_k} \quad \mathbf{f}_k^{\mathrm{H}} \mathbf{A}_k^{(t)} \mathbf{f}_k - 2 \operatorname{Re}\{\mathbf{b}_k^{(t)\mathrm{H}} \mathbf{f}_k\}$$

where $\mathbf{A}_k^{(t)} \in \mathbb{C}^{N_R \times N_R}$ and $\mathbf{b}_k^{(t)} \in \mathbb{C}^{N_R}$ are denoted by

$$\mathbf{A}_k^{(t)} \triangleq y_k^2 \sum_{l=1}^{K} \sum_{i,j \in \mathcal{M}_l} \rho_{\langle l,i \rangle, \langle l,j \rangle}^{(t)} \mathbf{H}_{\langle l,i \rangle}^{(t)} \mathbf{u}_{\langle l,i \rangle} (\mathbf{H}_{\langle l,j \rangle}^{(t)} \mathbf{u}_{\langle l,j \rangle})^{H} + (y_k^2 \sigma^2/2) \mathbf{I}_{N_R} \tag{34a}$$

$$\mathbf{b}_k^{(t)} \triangleq \frac{y_k}{\sum_{i \in \mathcal{M}_k} Q_{\langle k,i \rangle}} \sum_{i,j \in \mathcal{M}_k} \rho_{\langle k,i \rangle, \langle k,j \rangle}^{(t)} Q_{\langle k,j \rangle} \mathbf{H}_{\langle k,i \rangle}^{(t)} \mathbf{u}_{\langle k,i \rangle}. \tag{34b}$$

P3 is convex w.r.t. $\mathbf{f}_k$ by noting the positive semidefinite matrix $\mathbf{A}_k^{(t)}$. Note that $\mathbf{f}_k$ here is obtained by dropping the unit-length constraint in (27c). Thus, we finally obtain the optimal $\mathbf{f}_k$ by scaling the obtained $\mathbf{f}_k$ to a unit vector. We summarize the optimization of $\mathbf{f}$ and $\mathbf{u}$ in Algorithm 2.

Algorithm 2 AO Algorithm to Optimize $\mathbf{f}$ and $\mathbf{u}$

Input: $\mathcal{M}, \{\boldsymbol{\rho}_k^{(t)}, \mathbf{H}_{k, \langle l,i \rangle}^{(t)}, Q_{\langle k,i \rangle} | k, l \in [K], i \in [M_k]\}$, and $I_{\max}$.
1: **Initialization:** $\mathbf{y}$, $\mathbf{f}$ and $\mathbf{u}$.
2: **for** $\tau \in [I_{\max}]$ **do**
3:     **for** $k \in [K]$ **do**
4:         **for** $i \in \mathcal{M}_k$ **do**
5:             Compute $\mathbf{A}_{\langle k,i \rangle}^{(t)}, \mathbf{b}_{\langle k,i \rangle}^{(t)}$ based on (33) ;
6:             Optimize $\mathbf{u}_{\langle k,i \rangle}$ by solving (P2);
7:         **end for**
8:         Compute $\mathbf{A}_k^{(t)}, \mathbf{b}_k^{(t)}$ based on (33);
9:         Optimize $\mathbf{f}_k$ by solving (P3), update $\mathbf{f}_k$ (by scaling to a unit vector)
10:        Updates $\mathbf{y}$ based on (32);
11:     **end for**
12: **end for**
Output: $(\mathbf{f}, \mathbf{u})$.

C. *Optimizing $\mathcal{M}$ with fixed $\mathbf{f}$ and $\mathbf{u}$*

In this subsection, we optimize device selection set $\mathcal{M}$ with fixed beamforming $\mathbf{f}$ and $\mathbf{u}$. For notational convenience, we use a binary indication vector $\mathbf{s}$ to represent the device selection set $\mathcal{M}$, i.e., $\mathbf{s} \triangleq [\mathbf{s}_1^T, \cdots, \mathbf{s}_k^T]^T \in \{0, 1\}^M$, where $\mathbf{s}_k \in \{0, 1\}^{M_k}$ with the i-th entry $s_{\langle k,i \rangle} = 1$ meaning that the $\langle k, i \rangle$-th device is selected and $s_{\langle k,i \rangle} = 0$ otherwise. Note that $\mathcal{M}$ and $\mathbf{s}$ can be interchanged. Since $\mathbf{s}$ is discrete, we introduce the Gibbs sampling to optimize $\mathbf{s}$ by following the approach in [7]. We denote $\mathbf{s}^{\text{old}}$ as the sampling device selection solution obtained from the proceeding sampling round. At the current sampling round, the sampling set $\mathcal{S}$ is generated from $\mathbf{s}^{\text{old}}$, given by $\mathcal{S} \triangleq \{\mathbf{s}^{\text{old}}\} \cup \{\mathbf{s}_{\langle k,i \rangle}^{\text{old}} | k \in [K], i \in [M_k]\}$, where $\mathbf{s}_{\langle k,i \rangle}^{\text{old}}$ denotes the indication

vector that differs from $\mathbf{s}^{\text{old}}$ only at the $\langle k, i \rangle$-th element, corresponding to the $\langle k, i \rangle$-th device. We sample $\mathbf{s}^{\text{new}}$ according to the following distribution $\pi(\mathbf{s}^{\text{new}})$:

$$\pi(\mathbf{s}^{\text{new}}) \triangleq \frac{\exp\left(-\phi(\mathbf{s}^{\text{new}})/\beta\right)}{\sum_{k=1}^{K} \sum_{i=1}^{M_k} \exp\left(-\phi(\mathbf{s}^{\text{old}}_{\langle k,i \rangle})/\beta\right) + \exp\left(-\phi(\mathbf{s}^{\text{old}})/\beta\right)}, \tag{35}$$

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

Algorithm 3 AO Algorithm plus Gibbs Sampling

Input: $\{\boldsymbol{\rho}_k^{(t)}, \mathbf{H}_{\langle l,i \rangle}^{(t)}, Q_{\langle k,i \rangle} | k, l \in [K], i \in [M_k]\}, J_{max}, \beta, \gamma.$
 1: **Initialization:** $\mathbf{s}^{\text{old}} = \mathbf{1}_{M \times 1}$.
 2: **for** $j \in [J_{max}]$ **do**
 3: $\mathbf{s}^{\text{old}} = \mathbf{s}^{\text{new}}$;
 4: Generate $\mathcal{S}$;
 5: **for** every $\mathbf{s}^{\text{old}}_{\langle k,i \rangle} \in \mathcal{S}$ **do**
 6:   Optimize $(\mathbf{f}, \mathbf{u})$ by solving P2 with given $\mathbf{s}^{\text{old}}_{\langle k,i \rangle}$, with Algorithm 2;
 7: **end for**
 8: Sample $\mathbf{s}^{\text{new}}$ according to (35);
 9: Refresh $\beta = \gamma\beta$ for a certain $\gamma \in (0, 1)$;
10: **end for**
Output: $\mathbf{s}^{\text{new}}$ with corresponding $(\mathbf{f}, \mathbf{u})$.

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 $\mathcal{O}(I_{\max}(K + M)N^{3.5})$, where $N \triangleq \max\{N_{\text{T}}, N_{\text{R}}\}$ denotes the maximum number of the transmit or receive antennas, $I_{\max}$ 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 $J_{max}M$ times to optimize device selection set $\mathcal{M}$. Thus, the complexity of Algorithm 3 is $\mathcal{O}(J_{max}I_{\max}(KM + M^2)N^{3.5})$.

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 optimises the weighting factor ζ_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.

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 (δ, θ, χ), where δ, θ and χ 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 Δ. We set the location of the $\langle k, i \rangle$-th device to $(\delta_{\langle k,i \rangle}, \theta_{\langle k,i \rangle}, 0)$, where $\delta_{\langle k,i \rangle}^2$ is uniform in $[0, \Delta]$, and $\theta_{\langle k,i \rangle}$ is uniform in $[0, 2\pi)$. The PS is placed at the center of the circle, i.e.,$(0, 0, 10)$. We adopt the channel model in [31], given by $\mathbf{H}_{\langle l,i \rangle} = \sqrt{G_{\mathrm{S}} G_{\mathrm{D}} \kappa \tilde{\delta}_{\langle l,i \rangle}^{-\alpha}} \tilde{\mathbf{H}}_{\langle l,i \rangle}$, where the entries of $\tilde{\mathbf{H}}_{\langle l,i \rangle}$ 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, κ is the path loss at the reference distance $\delta_0 = 1\,\mathrm{m}$ [32], α is the path loss exponent, and $\tilde{\delta}_{\langle l,i \rangle} \triangleq \sqrt{\delta_{\langle l,i \rangle}^2 + 10^2}$ is the distance between the $\langle l, i \rangle$-th device and the PS. The simulation settings are given in Table I.

TABLE I
SYSTEM PARAMETERS

Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
N_{T}	2	N_{R}	8	$I_{\max}$	50	J_{max}	50
α	3.8	κ	$-60\,\mathrm{dB}$	G_{S}	$5\,\mathrm{dBi}$	G_{D}	$0\,\mathrm{dBi}$
P_0	$1\,\mathrm{W}$	β	1	γ	0.9	Δ	$100\,\mathrm{m}$
σ^2	$-80\,\mathrm{dBm}$	Q_k	60000				

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 5×5 convolution layers (the first with 16

Fig. 5. An illustration of locations for the devices and the PS in the OA-FMTL on the vertical view.

channels, the second with 32, each followed by 2×2 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 $\mathcal{A}_k$ 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 $\frac{Q_k}{5M_k}$ 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 $\{\boldsymbol{\rho}_k^{(t)}\}$. Specifically, we consider the following three approximations of the correlation matrix:

- **Approximation 1**: $\boldsymbol{\rho}_k^{(t)} = \frac{1}{D}\sum_{d=1}^{D}\mathbf{z}_{\langle k,d\rangle}^{(t)}\mathbf{z}_{\langle k,d\rangle}^{(t)}{}^{\mathrm{T}}$.
- **Approximation 2**: $\boldsymbol{\rho}_k^{(t)}$ is approximated by $\boldsymbol{\rho}_k^{(t)} = \epsilon\mathbf{1}_{M_k}+(1\!-\!\epsilon)\mathbf{I}_{M_k}$, where ϵ 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 $\boldsymbol{\rho}_k^{(t)}$ requires the knowledge of the local gradients uploaded by the devices. In contrast, Approximation 2 is more practical in implementation, where ϵ 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.

Fig. 6. FL test accuracy of the proposed Algorithm 2 with different approximations of $\boldsymbol{\rho}_k^{(t)}$ 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.

We plot the FL test accuracy curve of the proposed Algorithm 2 with the above approximations of the correlation matrix $\boldsymbol{\rho}_k^{(t)}$ in Fig. 6. We train two FL tasks, namely, tasks 1 and 2, both on i.i.d. data and on non-i.i.d. data. We set the numbers of devices $M_1 = M_2 = 20$, the learning rates $\eta_1 = \eta_2 = 0.05$, and the momentum $= 0.5$. The local updates consist of 5 times of SGD. The results are averaged over 20 Monte Carlo trials. In Approximation 2, we set $\epsilon = 0$, $\epsilon = 0.5$, and $\epsilon = 1$. From Fig. 6, we see that in the case of i.i.d. data, both Approximation 1 and Approximation 2 with $\epsilon = 1$ achieve test accuracies close to the error-free bound on both two tasks, and Approximation 2 with $\epsilon = 0$ has the worst learning performance. This is because Approximation 2 with $\epsilon = 0$ suffers from a serious aggregation error for ignoring the correlation between the local gradients. On the other hand, in the case of non-i.i.d. data, the accuracies achieved by Approximation 1 and Approximation 2 with $\epsilon = 0.5$ are close to the error-free bound, since $\epsilon = 0.5$ approximates the spatial correlation more precisely in the case of non-i.i.d. data. Thus, Approximation 2 with $\epsilon = 1$ for i.i.d. data and Approximation 2 with $\epsilon = 0.5$ for non-i.i.d. data are preferred in the system design.

We next study the necessity of device selection by comparing the proposed Algorithms 2 and 3. We simulate Algorithms 2 and 3 on tasks 1 and 2, with $\epsilon = 1$ on i.i.d. data and $\epsilon = 0.5$ on

non-i.i.d. data. The numbers of devices are set to $M_1 = M_2 = 10$. The learning rates are set to $\eta_1 = \eta_2 = 0.05$, and the momentum is set to 0.5. The local updates have 10 times of SGD. The results are averaged over 10 Monte Carlo trials [8]. In Fig. 7, we present the test accuracy of Algorithms 2 and 3 versus communication rounds on i.i.d. and non-i.i.d. data. From Fig. 7, we see that Algorithms 2 and 3 always perform closely. This implies that device selection is no longer necessary to our proposed scheme, which avoids the high computational complexity involved in the implementation of device selection. Thus, we henceforce always employ Algorithm 2 performance comparison when we refer to the proposed AO algorithm.

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

C. Comparisons With Existing Schemes

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

- SOCP-based cooperative power control [16]: This method assumes that all the devices selected, and that the local gradients from devices in the same task are independent with each other. The phases of transmit beamforming $\mathbf{u}_{\langle k,i \rangle}, \forall k, i$ are given by the zero-forcing.

---

[8]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.

The design of transmit power $\|\mathbf{u}_{\langle k,i \rangle}\|^2, \forall k,i$ are formulated as an second-order cone programming (SOCP) problem solved by the bisection method.

- SCA-based optimization and device selection [7]: $\mathcal{M}_k$, $\mathbf{f}_k$, and $\{\mathbf{u}_{\langle k,i \rangle}\}$ 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 ζ_k determined by zero-forcing. With given optimized $\mathbf{f}_k$, $\mathbf{u}_k$ and ζ_k, device selection set $\mathcal{M}_k$ is optimized via Gibbs sampling.

- Receive beamforming by differential geometry programming [8]: 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 [10]: 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 $\sigma^2 = -60\,\mathrm{dBm}$. We set $\epsilon = 0.5$. Besides, we introduce the normalized mean square error (NMSE) of each task k at round t, defined as $\mathrm{NMSE}_k^{(t)} \triangleq 10 \log_{10}\left(\mathbb{E}\left[\|\hat{\mathbf{g}}_k^{(t)} - \mathbf{g}_k^{(t)}\|_2^2 / \|\mathbf{g}_k^{(t)}\|_2^2 \right] \right)$. 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 [16] achieve much better NMSEs on all the three FL tasks than the other methods. We also see that our algorithm significantly outperforms the method in [16]. This is attributed to the careful optimization of the transceiver beamforming based on the proposed analytical framework.

TABLE II
COMMUNICATION NMSE

Optimization method	NMSE (dB)					
	MNIST ($k=1$)		Fashion-MNIST ($k=2$)		KMNIST ($k=3$)	
	$t=40$	$t=90$	$t=40$	$t=90$	$t=40$	$t=90$
Proposed AO algorithm	-1.43	-1.37	-1.81	-1.40	-1.98	-1.75
SOCP-based cooperative power control [16]	-0.84	-1.01	-1.10	-1.01	-1.11	-0.98
SCA & Gibbs [7]	-0.40	-0.47	-0.50	-0.52	-0.58	-0.52
Differential geometry [8]	-0.16	-0.20	-0.21	-0.19	-0.22	-0.21
DC and device selection [10]	-0.17	-0.20	-0.15	-0.21	-0.06	-0.13

Fig. 8. Proportion of devices versus range of allocated power with tasks 1, 2 and 3. $t = 40$ for (a)-(c); $t = 90$ for (d)-(f).

In Fig. 8, we plot the histogram of allocated transmission powers for various optimization methods. The methods based on zero-forcing [8], [10], [16] 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 [7] 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. 9, we present the test accuracies of various optimization algorithms versus communication rounds. As shown in Fig. 9, the proposed algorithm achieves an accuracy close to the error-free bound in all the three FL tasks and significantly outperforms the other baselines, which clearly demonstrates the superiority of our proposed scheme.

VI. CONCLUSION

In this paper, we studied a problem of designing an OA-FMTL system over 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 an optimization problem with respect to device selection,

Fig. 9. 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.

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 to demonstrate the learning accuracy outstanding improvements of the proposed algorithm by comparison with the state-of-the-art methods.

APPENDIX A

PROOF OF THEOREM 1

From [24, 3.2], the device selection MSE $\mathbb{E}[\|\mathbf{e}_{\mathrm{ds},k}^{(t)}\|^2]$ is bounded by (22). We next consider the communication MSE of the k-th task given by

$$\mathbb{E}\left[\|\mathbf{e}_{\mathrm{com},k}^{(t)}\|^2\right] = \sum_{d=1}^{D} \mathbb{E}\left[\left|g_k^{(t)}[d] - \hat{g}_k^{(t)}[d]\right|^2\right]. \tag{36}$$

By plugging (5), (8), (9) and (13) into (36), we obtain

$$\mathbb{E}[\|\mathbf{e}_{\mathrm{com},k}^{(t)}\|^2] = \frac{1}{\left(\sum_{i\in\mathcal{M}_k} Q_{\langle k,i\rangle}\right)^2} \sum_{c=1}^{C} \mathbb{E}\left[\left|\sum_{i\in\mathcal{M}_k} Q_{\langle k,i\rangle}\sqrt{v_k^{(t)}}r_{\langle k,i\rangle}^{(t)}[c] - \sum_{l=1}^{K}\sum_{i\in\mathcal{M}_l} \zeta_k \mathbf{f}_k^{\mathrm{H}}\mathbf{H}_{\langle l,i\rangle}^{(t)}\mathbf{u}_{\langle l,i\rangle}r_{\langle l,i\rangle}^{(t)}[c] - \zeta_k \mathbf{f}_k^{\mathrm{H}}\mathbf{n}^{(t)}[c]\right|^2\right]. \tag{37}$$

Based on Assumption 5, for $\forall k \neq l$, we have $\mathbb{E}[r_{\langle k,i\rangle}[c]r_{\langle l,j\rangle}[c]^\dagger] = 0$. Thus, we obtain (23) by expanding (37). What remains is to prove (21). From [24, Lemma 2.1], Assumptions 1-4 lead to an upper bound of $F_k(\cdot)$ at the t-th round. Thus, we have the following lemma.

Lemma 1. *Assume that $F_k(\cdot)$ satisfies Assumptions 1-4, at the t-th communication round with the learning rate η_k is set to $1/\omega_k$. Then*

$$\mathbb{E}[F_k(\mathbf{w}_k^{(t+1)})] \leq \mathbb{E}[F_k(\mathbf{w}_k^{(t)})] - \frac{1}{2\omega_k}\mathbb{E}[\|\nabla F_k(\mathbf{w}_k^{(t)})\|^2] + \frac{1}{2\omega_k}\mathbb{E}[\|\mathbf{e}_k^{(t)}\|^2], \tag{38}$$

where ω_k is the Lipschitz continuity parameter defined in (16), and $\mathbb{E}[\cdot]$ is the expectation w.r.t. $\{n_{k,z}^{(t)}[c], g_{\langle k,i\rangle}^{(t)}[d] | k \in [K], z \in [N_R], c \in [C], i \in [M_k], d \in [D], \tau \in [t+1]\}$.

Proof. See [24, Lemma 2.1]. □

In addition, we obtain

$$\mathbb{E}[\|\mathbf{e}_k^{(t)}\|^2] \overset{(a)}{=} \mathbb{E}[\|\mathbf{e}_{\text{ds},k}^{(t)} + \mathbf{e}_{\text{com},k}^{(t)}\|^2] \overset{(b)}{\leq} 2\left(\mathbb{E}[\|\mathbf{e}_{\text{ds},k}^{(t)}\|^2] + \mathbb{E}[\|\mathbf{e}_{\text{com},k}^{(t)}\|^2]\right), \tag{39}$$

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 (39) into (38), we obtain (21).

<center>APPENDIX B</center>

<center>PROOF OF COROLLARY 1</center>

From (23), we obtain

$$\mathbb{E}\left[\|\mathbf{e}_{\text{com},k}^{(t)}\|^2\right] \overset{(a)}{=} \frac{C}{\left(\sum_{i\in\mathcal{M}_k}Q_{\langle k,i\rangle}\right)^2}\Bigg(\sum_{i,j\in\mathcal{M}_k}2\rho_{\langle k,i\rangle,\langle k,j\rangle}^{(t)}Q_{\langle k,i\rangle}Q_{\langle k,j\rangle}v_k^{(t)}$$

$$-\zeta_k\sqrt{v_k^{(t)}}\sum_{i,j\in\mathcal{M}_k}2\rho_{\langle k,i\rangle,\langle k,j\rangle}^{(t)}\left(Q_{\langle k,i\rangle}(\mathbf{f}_k^{\text{H}}\mathbf{H}_{\langle k,j\rangle}^{(t)}\mathbf{u}_{\langle k,j\rangle})^{\text{H}} + Q_{\langle k,j\rangle}\mathbf{f}_k^{\text{H}}\mathbf{H}_{\langle k,i\rangle}^{(t)}\mathbf{u}_{\langle k,i\rangle}\right)$$

$$+\zeta_k^2\left(\sum_{l=1}^K\sum_{i,j\in\mathcal{M}_l}2\rho_{\langle l,i\rangle,\langle l,j\rangle}^{(t)}(\mathbf{f}_k^{\text{H}}\mathbf{H}_{\langle l,i\rangle}^{(t)}\mathbf{u}_{l,i})^{\text{H}}\mathbf{f}_k^{\text{H}}\mathbf{H}_{\langle l,j\rangle}^{(t)}\mathbf{u}_{\langle l,j\rangle} + \sigma^2\|\mathbf{f}_k\|^2)\right)\Bigg) \tag{40a}$$

$$\overset{(b)}{\geq} \frac{2Cv_k^{(t)}}{\left(\sum_{i\in\mathcal{M}_k}Q_{\langle k,i\rangle}\right)^2}\Bigg(\sum_{i,j\in\mathcal{M}_k}\rho_{\langle k,i\rangle,\langle k,j\rangle}^{(t)}Q_{\langle k,i\rangle}Q_{\langle k,j\rangle}$$

$$-\frac{\left(\sum_{i,j\in\mathcal{M}_k}\rho_{\langle k,i\rangle,\langle k,j\rangle}^{(t)}\left(Q_{\langle k,i\rangle}(\mathbf{f}_k^{\text{H}}\mathbf{H}_{\langle k,j\rangle}^{(t)}\mathbf{u}_{\langle k,j\rangle})^{\text{H}} + Q_{\langle k,j\rangle}\mathbf{f}_k^{\text{H}}\mathbf{H}_{\langle k,i\rangle}^{(t)}\mathbf{u}_{\langle k,i\rangle}\right)\right)^2}{4\left(\sum_{l=1}^K\sum_{i,j\in\mathcal{M}_l}\rho_{\langle l,i\rangle,\langle l,j\rangle}^{(t)}(\mathbf{f}_k^{\text{H}}\mathbf{H}_{\langle l,i\rangle}^{(t)}\mathbf{u}_{\langle l,i\rangle})^{\text{H}}\mathbf{f}_k^{\text{H}}\mathbf{H}_{\langle l,j\rangle}^{(t)}\mathbf{u}_{\langle l,j\rangle} + \sigma^2\|\mathbf{f}_k\|^2/2\right)}\Bigg), \tag{40b}$$

where step (a) is from $\mathbb{E}[r_{\langle k,i\rangle}[c]r_{\langle k,j\rangle}[c]^\dagger] = 2\rho_{\langle k,i\rangle,\langle k,j\rangle}$ 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. ζ_k and the minimizer ζ_k^* is given by (24).

APPENDIX C

PROOF OF THEOREM 2

We first derive an upper bound w.r.t. the communication MSE in (40b). For device $\langle k, i \rangle, \forall k, i$, we have

$$
\begin{aligned}
2C v_k^{(t)} = 2C v_{\langle k,i \rangle}^{(t)} &= \sum_{d=1}^{D} \left(\mathbb{E}[|g_{\langle k,i \rangle}^{(t)}[d]|^2] - |\mathbb{E}[g_{\langle k,i \rangle}^{(t)}[d]]|^2 \right) \stackrel{(a)}{\leq} \mathbb{E}[\|\mathbf{g}_{\langle k,i \rangle}^{(t)}\|^2] \\
&\stackrel{(b)}{=} \mathbb{E}\left[\|\frac{1}{Q_{\langle k,i \rangle}} \sum_{n=1}^{Q_{\langle k,i \rangle}} \nabla f_k(\mathbf{w}_k^{(t)}; \boldsymbol{\xi}_{\langle k,i \rangle, n}) \|^2 \right] \stackrel{(c)}{\leq} \mathbb{E}\left[(\frac{1}{Q_{\langle k,i \rangle}} \sum_{n=1}^{Q_{\langle k,i \rangle}} \|\nabla f_k(\mathbf{w}_k^{(t)}; \boldsymbol{\xi}_{\langle k,i \rangle, n}) \|)^2 \right] \\
&\stackrel{(d)}{\leq} \mathbb{E}\left[\left(\frac{1}{Q_{\langle k,i \rangle}} \sum_{n=1}^{Q_{\langle k,i \rangle}} \sqrt{\beta_1 + \beta_2 \|\nabla F_k(\mathbf{w}_k^{(t)})\|^2} \right)^2 \right] = \beta_1 + \beta_2 \mathbb{E}[\|\nabla F_k(\mathbf{w}_k^{(t)})\|^2] \quad (41)
\end{aligned}
$$

where step (a) is from $\left| \mathbb{E}[g_{\langle k,i \rangle}^{(t)}[d]] \right|^2 \geq 0$; (b) is from the fact that $\mathbb{E}[\mathbf{g}_{\langle k,i \rangle}^{(t)}] = \mathbb{E}[\nabla F_{\langle k,i \rangle}(\mathbf{w}_k^{(t)})]$ and with definition of $F_{\langle k,i \rangle}(\mathbf{w}_k^{(t)})$ given below (2); (c) is from the triangle inequality; and (d) is from (18). Note that (40b) is obtained by the expression of $\mathbb{E}[\|\mathbf{e}_{\mathrm{com},k}^{(t)}\|^2]$ in (23) as $\zeta_k = \zeta_k^*$. By plugging (41) into (40b), we obtain an upper bound w.r.t. $\mathbb{E}[\|\mathbf{e}_{\mathrm{com},k}^{(t)}\|^2]$:

$$
\begin{aligned}
\mathbb{E}[\|\mathbf{e}_{\mathrm{com},k}^{(t)}\|^2] \leq &\frac{\beta_1 + \beta_2 \mathbb{E}[\|\nabla F_k(\mathbf{w}_k^{(t)})\|^2]}{\left(\sum_{i \in \mathcal{M}_k} Q_{\langle k,i \rangle} \right)^2} \left(\sum_{i,j \in \mathcal{M}_k} \rho_{\langle k,i \rangle, \langle k,j \rangle}^{(t)} Q_{\langle k,i \rangle} Q_{\langle k,j \rangle} \right. \\
&\left. - \frac{\left(\sum_{i,j \in \mathcal{M}_k} \rho_{\langle k,i \rangle, \langle k,j \rangle}^{(t)} \left(Q_{\langle k,i \rangle} (\mathbf{f}_k^{\mathrm{H}} \mathbf{H}_{\langle k,j \rangle}^{(t)} \mathbf{u}_{\langle k,j \rangle})^{\mathrm{H}} + Q_{\langle k,j \rangle} \mathbf{f}_k^{\mathrm{H}} \mathbf{H}_{\langle k,i \rangle}^{(t)} \mathbf{u}_{\langle k,i \rangle} \right) \right)^2}{4 \left(\sum_{l=1}^{K} \sum_{i,j \in \mathcal{M}_l} \rho_{\langle l,i \rangle, \langle l,j \rangle}^{(t)} (\mathbf{f}_k^{\mathrm{H}} \mathbf{H}_{\langle l,i \rangle}^{(t)} \mathbf{u}_{\langle l,i \rangle})^{\mathrm{H}} \mathbf{f}_k^{\mathrm{H}} \mathbf{H}_{\langle l,j \rangle}^{(t)} \mathbf{u}_{\langle l,j \rangle} + \sigma^2 \|\mathbf{f}_k\|^2/2 \right)} \right), \quad (42)
\end{aligned}
$$

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

$$
\mathbb{E}[\mathcal{F}(\mathbf{w}^{(t+1)})] \leq \mathbb{E}[\mathcal{F}(\mathbf{w}^{(t)})] - \frac{1}{2\omega} \sum_{k=1}^{K} \left(\mathbb{E}[\|\nabla F_k(\mathbf{w}_k^{(t)})\|^2] - 2(\mathbb{E}[\|\mathbf{e}_{\mathrm{ds},k}^{(t)}\|^2] + \mathbb{E}[\|\mathbf{e}_{\mathrm{com},k}^{(t)}\|^2]) \right). \quad (43)
$$

where $\omega = \max_k \omega_k$, and $\mu = \min_k \mu_k$. Substituting $\mathbb{E}[\|\mathbf{e}_{\mathrm{ds},k}^{(t)}\|^2]$ and $\mathbb{E}[\|\mathbf{e}_{\mathrm{com},k}^{(t)}\|^2]$ respectively with (22) and (42), we have an upper bound of the overall loss function $\mathcal{F}_k(\cdot)$ at the round $(t+1)$ as

$$
\mathbb{E}[\mathcal{F}(\mathbf{w}^{(t+1)})] \leq \mathbb{E}[\mathcal{F}(\mathbf{w}^{(t)})] - \frac{1}{2\omega} \sum_{k=1}^{K} \left(\mathbb{E}[\|\nabla F_k(\mathbf{w}_k^{(t)})\|^2](1 - 2\beta_2 d_k^{(t)}(\mathcal{M}_k, \mathbf{f}_k, \mathbf{u}_k)) - 2\beta_1 d_k^{(t)}(\mathcal{M}_k, \mathbf{f}_k, \mathbf{u}_k) \right), \quad (44)
$$

where $d_k^{(t)}(\mathcal{M}_k, \mathbf{f}_k, \mathbf{u}_k)$ is defined by (26). Furthermore, based on Assumption 2, an upper bound of $\mathbb{E}[\|\nabla F_k(\mathbf{w}_k^{(t)})\|^2]$ is obtained from [24, eq. (2.4)] as

$$
\mathbb{E}[\|\nabla F_k(\mathbf{w}_k^{(t)})\|^2] \geq 2\mu \mathbb{E}\left[F_k(\mathbf{w}_k^{(t)}) - F_k(\mathbf{w}_k^*) \right]. \quad (45)
$$

Plugging (45) into (44) and subtracting $\mathcal{F}(\mathbf{w}^*)$ on the both sides of (44), we obtain (25).

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[2] 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. 3, pp. 55–75, Aug. 2018.

[3] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," Oct. 2016, [Online] Available: https://arxiv.org/abs/1610.02527.

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

[5] 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. Wirel. Commun.*, vol. 20, no. 3, pp. 2120–2135, Mar. 2020.

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

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

[8] G. Zhu, L. Chen, and K. Huang, "MIMO over-the-air computation: Beamforming optimization on the Grassmann manifold," in *IEEE Global Communications Conference (GLOBECOM)*, Dec. 2018, pp. 1–6.

[9] 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.

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

[11] Y. Shi, Y. Zhou, and Y. Shi, "Over-the-air decentralized federated learning," Jun. 2021, [Online] Available: https://arxiv.org/abs/2106.08011.

[12] 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, Sept. 2020.

[13] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," Oct. 2015, [Online] Available: https://arxiv.org/abs/1506.02626.

[14] C.-Y. Chen, J. Ni, S. Lu, X. Cui, P.-Y. Chen, X. Sun, N. Wang, S. Venkataramani, V. Srinivasan, and W. Zhang, "Scalecom: Scalable sparsified gradient compression for communication-efficient distributed training," Apr. 2021, [Online] Available: https://arxiv.org/abs/2104.11125.

[15] 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.

[16] X. Cao, G. Zhu, J. Xu, and K. Huang, "Cooperative interference management for over-the-air computation networks," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 4, pp. 2634–2651, Apr. 2021.

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

[18] 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.*, 2022.

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

[20] G. Zhu, J. Xu, K. Huang, and S. Cui, "Over-the-air computing for wireless data aggregation in massive iot," *IEEE Wirel. Commun.*, vol. 28, no. 4, pp. 57–65, Sep. 2021.

[21] S. L. H. Nguyen and A. Ghrayeb, "Compressive sensing-based channel estimation for massive multiuser MIMO systems," in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, Oct. 2013, pp. 2890–2895.

[22] 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. Wirel. Commun.*, vol. 14, no. 3, pp. 1356–1368, Oct. 2014.

[23] Z. Lin, X. Li, V. K. Lau, Y. Gong, and K. Huang, "Deploying federated learning in large-scale cellular networks: Spatial convergence analysis," Mar. 2021, [Online] Available: https://arxiv.org/abs/2103.06056.

[24] 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.

[25] D. P. Bertsekas and J. N. Tsitsiklis, "Neuro-dynamic programming: An overview," in *Proc. 1995 34th IEEE conf. Decis. and Control*, vol. 1, 1995, pp. 560–564.

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

[27] Y. LeCun, C. Cortes, and C. Burges, "The MNIST database of handwritten digits," 1998, [Online] Available: http://yann.lecun.com/exdb/mnist.

[28] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," Sep. 2017, [Online] Available: https://arxiv.org/abs/1708.07747.

[29] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, "Deep learning for classical Japanese literature," Dec. 2018, [Online] Available: https://arxiv.org/abs/1812.01718.

[30] 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.

[31] A. Goldsmith, *Wireless Communication*. Cambridge Univ. Press, Aug. 2005.

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