# SegOTA: Accelerating Over-the-Air Federated Learning with Segmented Transmission

Chong Zhang*, Min Dong†, Ben Liang*, Ali Afana‡, Yahia Ahmed‡
*Department of Electrical and Computer Engineering, University of Toronto, Canada
†Department of Electrical, Computer and Software Engineering, Ontario Tech University, Canada, ‡Ericsson Canada

***Abstract*—Federated learning (FL) with over-the-air computation efficiently utilizes the communication resources, but it can still experience significant latency when each device transmits a large number of model parameters to the server. This paper proposes the Segmented Over-The-Air (SegOTA) method for FL, which reduces latency by partitioning devices into groups and letting each group transmit only one segment of the model parameters in each communication round. Considering a multi-antenna server, we model the SegOTA transmission and reception process to establish an upper bound on the expected model learning optimality gap. We minimize this upper bound, by formulating the per-round online optimization of device grouping and joint transmit-receive beamforming, for which we derive efficient closed-form solutions. Simulation results show that our proposed SegOTA substantially outperforms the conventional full-model OTA approach and other common alternatives.**

## I. Introduction

Federated learning (FL) [1] enables multiple worker devices to collaboratively train a machine learning model using their local datasets, with a parameter server (PS) aggregating their local updates into a global model. In wireless FL, the PS often is hosted by a base station (BS) [2]. However, limited wireless resources and signal distortion in wireless links degrade the performance of wireless FL, making efficient communication design a necessity.

Most prior wireless FL works focused on improving the communication efficiency in uplink aggregation of local model parameters from the devices to the BS [3]–[13]. Early works [3]–[5] studied digital transmission-then-aggregation schemes using orthogonal channels, which can consume large bandwidth and cause high latency with many devices. Later, analog transmission-and-aggregation schemes were proposed [6]–[8], which adopt analog modulation and superposition for over-the-air computation of local parameters via the multiple access channel. Analog schemes result in significant communication savings and lower latency compared with digital approaches. However, these analog schemes are designed for single-antenna BSs, making their solutions and convergence analysis unsuitable for the multi-antenna BSs commonly used in practical wireless systems.

In the multi-antenna communication, beamforming plays a critical role in improving communication quality in wireless networks. Receive beamforming was considered in [14], [15] to boost performance of the uplink analog over-the-air computation. Various uplink beamforming designs have since been proposed to improve the training performance of wireless FL [9]–[13]. These works demonstrate that well-designed beamforming schemes can significantly enhance the over-the-air computation for wireless FL.

The existing uplink analog over-the-air works [6]–[13] typically adopt the traditional full-model transmission approach illustrated in Fig. 1(left). In each channel use, all devices simultaneously send parameters from the same location in their model parameter vectors to the BS, which aggregates these parameters via over-the-air computation. However, this full-model transmission approach can lead to substantial latency when the model parameter vector is long, which degrades the overall performance of wireless FL.

To address this issue, we propose the Segmented Over-The-Air (SegOTA) method for wireless FL, as illustrated in Fig. 1(right). SegOTA divides the locations of a model parameter vector into equal-sized segments and assigns the transmission task of each segment to a group of devices. By allowing simultaneous transmission of parameters in different segments, SegOTA can substantially reduce the communication latency, while maintaining satisfactory learning performance by allowing the BS to aggregate and update global parameters for each segment. However, it also introduces wireless interference among the different segments, which requires careful transmission design to balance the tradeoff between communication efficiency and OTA computation accuracy, in order to optimize the overall FL performance.

The main contribution of this paper is summarized as follows:

- We propose a novel SegOTA method to allow simultaneous transmission of parameters from multiple segments in wireless FL. We formulate an optimization problem to carefully manage the inter-segment interference through device grouping, BS receive beamforming, and device transmit power control, in order to minimize the expected model optimality gap after a given number of FL communication rounds. As far as we are aware, this is the first study on segmented OTA transmission in FL.
- We analyze the SegOTA transmission and reception processes and derive an upper bound on the expected model optimality gap. We show that minimization of this bound can be decomposed into per-round online opti-

This work was supported in part by Ericsson, the Natural Sciences and Engineering Research Council of Canada, and Mitacs.

ISBN 978-3-903176-73-7 © 2025 IFIP

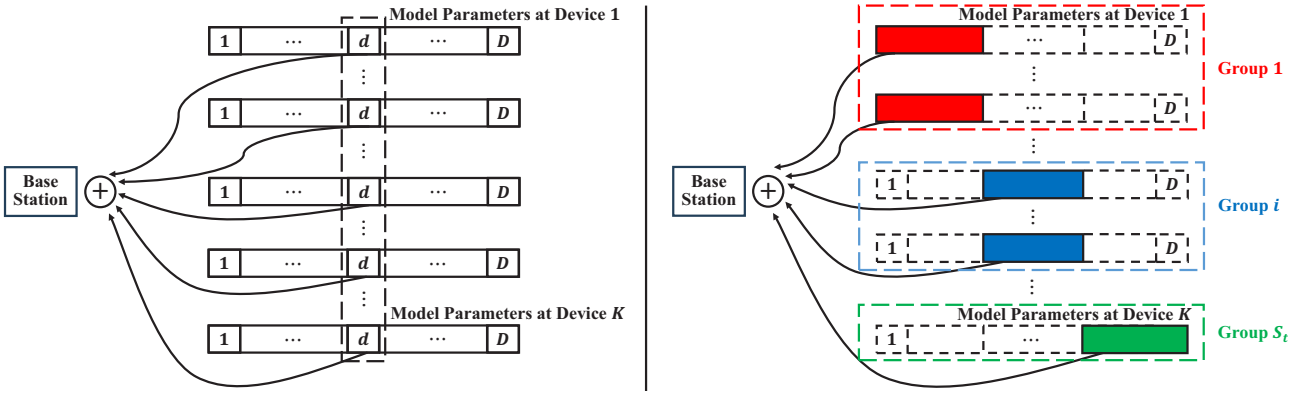


Fig. 1. Uplink analog OTA aggregation for wireless FL. Left: traditional full-model OTA approach; Right: proposed SegOTA (Each colored box represents a model segment that consists of $I_t$ parameters; different colors indicate different segments).

mization problems of device grouping and uplink joint transmit-receive beamforming. We apply a spherical $k$-means method for device grouping and then optimize joint transmit-receive beamforming, obtaining efficient closed-form solutions at each round. The proposed solution is guaranteed to converge to a stationary point.

- Simulation under typical wireless network settings shows that SegOTA substantially outperforms the conventional approach of full-model OTA aggregation, as well as other alternatives such as segmented OTA with the popular zero-forcing beamforming.

## II. Wireless FL System Model

We consider an FL system consisting of a server and $K$ worker devices that collaboratively train a machine learning model. Let $\mathcal{K}_{\text{tot}} = \{1, \ldots, K\}$ denote the total set of devices. Each device $k \in \mathcal{K}_{\text{tot}}$ holds a local training dataset of size $A_k$, denoted by $\mathcal{A}_k = \{(\mathbf{a}_{k,i}, y_{k,i}) : 1 \le i \le A_k\}$, where $\mathbf{a}_{k,i} \in \mathbb{R}^b$ is the $i$-th data feature vector and $y_{k,i}$ is the label for this data sample. Let $\boldsymbol{\theta} \in \mathbb{R}^D$ be the parameter vector of the machine learning model, which has $D$ parameters. The local training loss function that represents the training error at device $k$ is given by

$$F_k(\boldsymbol{\theta}) = \frac{1}{A_k} \sum_{i=1}^{A_k} L(\boldsymbol{\theta}; \mathbf{a}_{k,i}, y_{k,i})$$

where $L(\cdot)$ is the sample-wise training loss corresponding to each data sample. The global training loss function is the weighted sum of the local loss function $F_k(\boldsymbol{\theta})$ of each device $k$, expressed as

$$F(\boldsymbol{\theta}) = \frac{1}{\sum_{k=1}^{K} A_k} \sum_{k=1}^{K} A_k F_k(\boldsymbol{\theta}). \tag{1}$$

The learning objective is to find the optimal global model $\boldsymbol{\theta}^\star$ that minimizes $F(\boldsymbol{\theta})$.

The $K$ devices communicate with the server via separate downlink and uplink channels to exchange the model training information iteratively. The FL training procedure in each communication round $t = 0, 1, \ldots$ is given as follows:

- *Downlink broadcast*: The server sends the parameter vector of the current global model $\boldsymbol{\theta}_t$ to all $K$ devices. We make the common assumption that the downlink channel is error-free.
- *Local model update*: Device $k$ divides $\mathcal{A}_k$ into smaller mini-batches, and applies the standard mini-batch stochastic gradient descent (SGD) algorithm with $J$ iterations to generate an updated local model based on $\boldsymbol{\theta}_t$. Let $\boldsymbol{\theta}_{k,t}^\tau$ be the local model update by device $k$ at iteration $\tau \in \{0, \ldots, J-1\}$, with $\boldsymbol{\theta}_{k,t}^0 = \boldsymbol{\theta}_t$, and let $\mathcal{B}_{k,t}^\tau$ denote the mini-batch at iteration $\tau$. Then, the local model update is given by

$$\begin{aligned} \boldsymbol{\theta}_{k,t}^{\tau+1} &= \boldsymbol{\theta}_{k,t}^\tau - \eta_t \nabla F_k(\boldsymbol{\theta}_{k,t}^\tau; \mathcal{B}_{k,t}^\tau) \\ &= \boldsymbol{\theta}_{k,t}^\tau - \frac{\eta_t}{|\mathcal{B}_{k,t}^\tau|} \sum_{(\mathbf{a},y)\in\mathcal{B}_{k,t}^\tau} \nabla L(\boldsymbol{\theta}_{k,t}^\tau; \mathbf{a}, y) \end{aligned} \tag{2}$$

  where $\eta_t$ is the learning rate in communication round $t$, and $\nabla F_k$ and $\nabla L$ are the gradient functions with respect to (w.r.t.) $\boldsymbol{\theta}_{k,t}^\tau$. After $J$ iterations, device $k$ obtains the updated local model $\boldsymbol{\theta}_{k,t}^J$.
- *Uplink aggregation*: The devices send their updated local models $\{\boldsymbol{\theta}_{k,t}^J\}_{k\in\mathcal{K}_{\text{tot}}}$ to the server through the uplink channels. The server aggregates $\boldsymbol{\theta}_{k,t}^J$'s to generate an updated global model $\boldsymbol{\theta}_{t+1}$ for the next communication round $t+1$. In the existing full-model OTA approach [6]–[13], this consists of analog transmission of each of the parameters in the $\boldsymbol{\theta}_{k,t}^J$ vector by device $k$, and OTA aggregation by the BS, as shown in Fig. 1(left). In this paper, we will propose a segmented OTA approach to improve the communication efficiency of this step.

For the model exchange between the server and devices through a wireless system, we assume the server is hosted by a BS equipped with $N$ antennas, and each device has a single antenna. In this paper, we propose a segmented transmission approach for uplink OTA aggregation and optimize the corresponding uplink beamforming to accelerate the FL training convergence.

## III. Uplink Segmented Over-the-Air Aggregation

We propose an efficient uplink aggregation approach, named SegOTA. Under SegOTA, each device only sends one segment

of its $D$ local parameters to the BS in each communication round, shown in Fig. 1(right) as a colored segment.

In particular, at the beginning of communication round $t$, the BS partitions the $K$ devices into $S_t$ groups, which remain unchanged during this round. Let $\mathcal{K}_{i,t}$ denote the set of devices of group $i \in \{1,\ldots,S_t\}$ in round $t$, with $K_{i,t} \triangleq |\mathcal{K}_{i,t}|$. The BS also divides the model parameter vector into $S_t$ equal-sized segments, with each segment having a length of $I_t \triangleq \lceil \frac{D}{S_t} \rceil$. If $D$ is not a multiple of $S_t$, the last segment will be padded with zero. Let $\mathcal{S}_t \triangleq \{1,\ldots,S_t\}$ be the index set of model segments in round $t$. Devices in group $\mathcal{K}_{i,t}$ are assigned to send segment $\hat{m}(i,t) \in \mathcal{S}_t$ to the BS. Each group is assigned a unique segment, which can be either at random or in a round-robin fashion. The BS then aggregates the received local segments from the devices in group $\mathcal{K}_{i,t}$ to update segment $\hat{m}(i,t)$ of the global model $\boldsymbol{\theta}_{t+1}$ for the next communication round $t+1$. Below, we detail the formulation of uplink aggregation under the proposed SegOTA.

Let $\mathbf{s}_{m,t}^{k,J} \in \mathbb{R}^{I_t}$ denote the segment $m$ of the local model update $\boldsymbol{\theta}_{k,t}^{J}$ at device $k$ in communication round $t$. For efficient transmission, we represent $\mathbf{s}_{m,t}^{k,J} \in \mathbb{R}^{I_t}$ using an equivalent complex vector $\tilde{\mathbf{s}}_{m,t}^{k,J}$, whose real and imaginary parts contain the first and second halves of the elements in $\mathbf{s}_{m,t}^{k,J}$, respectively. That is, $\tilde{\mathbf{s}}_{m,t}^{k,J} = \tilde{\mathbf{s}}_{m,t}^{k,J\text{re}} + j\tilde{\mathbf{s}}_{m,t}^{k,J\text{im}} \in \mathbb{C}^{\frac{I_t}{2}}$, where $\tilde{\mathbf{s}}_{m,t}^{k,J\text{re}}$ contains the first $\frac{I_t}{2}$ elements in $\mathbf{s}_{m,t}^{k,J}$ and $\tilde{\mathbf{s}}_{m,t}^{k,J\text{im}}$ contains the other $\frac{I_t}{2}$ elements.

We denote the channel from device $k$ to the BS in communication round $t$ by $\mathbf{h}_{k,t}$ and assume it is known at the BS. We denote the transmit beamforming weight at device $k$ in round $t$ by $a_{k,t} \in \mathbb{C}$. Device $k$ in group $i$ applies $a_{k,t}$ to the normalized complex model segment $\frac{\tilde{\mathbf{s}}_{\hat{m}(i,t),t}^{k,J}}{\|\tilde{\mathbf{s}}_{\hat{m}(i,t),t}^{k,J}\|}$, and all $K$ devices send their respective segments simultaneously to the BS with $\frac{I_t}{2}$ channel uses. Let $\frac{\tilde{s}_{\hat{m}(i,t)l,t}^{k,J}}{\|\tilde{\mathbf{s}}_{\hat{m}(i,t),t}^{k,J}\|}$ be the $l$-th element in segment $\frac{\tilde{\mathbf{s}}_{\hat{m}(i,t),t}^{k,J}}{\|\tilde{\mathbf{s}}_{\hat{m}(i,t),t}^{k,J}\|}$ sent in the $l$-th channel use. The received signal vector at the BS in the $l$-th channel use is given by

$$\mathbf{v}_{l,t} = \sum_{i=1}^{S_t} \sum_{k\in\mathcal{K}_{i,t}} \mathbf{h}_{k,t} a_{k,t} \frac{\tilde{s}_{\hat{m}(i,t)l,t}^{k,J}}{\|\tilde{\mathbf{s}}_{\hat{m}(i,t),t}^{k,J}\|} + \mathbf{u}_{l,t}$$

where $\mathbf{u}_{l,t} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$ is the receiver additive white Gaussian noise vector with variance $\sigma^2$.

The BS applies receive beamforming on $\mathbf{v}_{l,t}$ to aggregate the local segments $\{\tilde{\mathbf{s}}_{\hat{m}(i,t),t}^{k,J}\}_{k\in\mathcal{K}_{i,t}}$ from each group $i$. Let $\mathbf{w}_{i,t} \in \mathbb{C}^N$ denote the receive beamforming vector at the BS for group $i$ in communication round $t$, which is normalized as $\|\mathbf{w}_{i,t}\|^2 = 1$. The effective channel from device $k \in \mathcal{K}_{i,t}$ to the BS after applying $\mathbf{w}_{i,t}$ is given by $\alpha_{k,t} \triangleq \frac{\mathbf{w}_{i,t}^{\mathsf{H}}\mathbf{h}_{k,t}a_{k,t}}{\|\tilde{\mathbf{s}}_{\hat{m}(i,t),t}^{k,J}\|}$. Then, the post-processed received signal vector for the aggregated local segments $\{\tilde{\mathbf{s}}_{\hat{m}(i,t),t}^{k,J}\}_{k\in\mathcal{K}_{i,t}}$ over the $\frac{I_t}{2}$ channel uses is

$$\mathbf{z}_{\hat{m}(i,t),t} = \sum_{k\in\mathcal{K}_{i,t}} \alpha_{k,t}\tilde{\mathbf{s}}_{\hat{m}(i,t),t}^{k,J} + \sum_{j\neq i}\sum_{q\in\mathcal{K}_{j,t}} \mathbf{w}_{i,t}^{\mathsf{H}}\mathbf{h}_{q,t}a_{q,t}\frac{\tilde{\mathbf{s}}_{\hat{m}(j,t),t}^{q,J}}{\|\tilde{\mathbf{s}}_{\hat{m}(j,t),t}^{q,J}\|} + \mathbf{n}_{\hat{m}(i,t),t} \quad (3)$$

where $\mathbf{n}_{\hat{m}(i,t),t} \in \mathbb{C}^{\frac{I_t}{2}}$ is the post-processed receiver noise vector with the $l$-th element being $\mathbf{w}_{i,t}^{\mathsf{H}}\mathbf{u}_{l,t}$, and the last term is the inter-segment interference.

After receive beamforming, the BS receiver finally performs scaling to obtain the global model update. Let $\mathbf{s}_{m,t} \in \mathbb{R}^{I_t}$ denote segment $m$ of the global model update $\boldsymbol{\theta}_t$ in communication round $t$, and let $\tilde{\mathbf{s}}_{m,t}$ be the equivalent complex representation of $\mathbf{s}_{m,t}$. Let $\hat{i}(m,t)$ represent the device group that transmits segment $m$ in round $t$. The BS scales $\mathbf{z}_{m,t}$ by $\alpha_{m,t}^{\text{s}} \triangleq \sum_{k\in\mathcal{K}_{\hat{i}(m,t),t}} \alpha_{k,t}$ to obtain segment $m$ of the global model update $\boldsymbol{\theta}_{t+1}$ for the next round $t+1$:

$$\tilde{\mathbf{s}}_{m,t+1} = \frac{\mathbf{z}_{m,t}}{\alpha_{m,t}^{\text{s}}}. \quad (4)$$

Based on (3)(4), we obtain the following updating equation for segment $\tilde{\mathbf{s}}_{m,t}$ in each round:

$$\tilde{\mathbf{s}}_{m,t+1} = \tilde{\mathbf{s}}_{m,t} + \sum_{k\in\mathcal{K}_{\hat{i}(m,t),t}} \rho_{k,t}\Delta\tilde{\mathbf{s}}_{m,t}^{k} + \tilde{\mathbf{n}}_{m,t} + \frac{1}{\alpha_{m,t}^{\text{s}}}\sum_{j\neq\hat{i}(m,t)}\sum_{q\in\mathcal{K}_{j,t}} \mathbf{w}_{\hat{i}(m,t),t}^{\mathsf{H}}\mathbf{h}_{q,t}a_{q,t}\frac{\tilde{\mathbf{s}}_{\hat{m}(j,t),t}^{q,J}}{\|\tilde{\mathbf{s}}_{\hat{m}(j,t),t}^{q,J}\|} \quad (5)$$

where $\Delta\tilde{\mathbf{s}}_{m,t}^{k} \triangleq \tilde{\mathbf{s}}_{m,t}^{k,J} - \tilde{\mathbf{s}}_{m,t}^{k,0}$ is the difference in the local segment update after the local training, $\rho_{k,t} \triangleq \frac{\alpha_{k,t}}{\alpha_{m,t}^{\text{s}}}$ represents the weight of each device $k$ in group $\hat{i}(m,t)$, and $\tilde{\mathbf{n}}_{m,t} \triangleq \frac{\mathbf{n}_{m,t}}{\alpha_{m,t}^{\text{s}}}$ is post-processed receiver noise vector.

Finally, segment $m$ of the global model update $\boldsymbol{\theta}_{t+1}$, *i.e.,* $\mathbf{s}_{m,t+1}$, can be recovered from its complex version as $\mathbf{s}_{m,t+1} = [\Re\mathfrak{e}\{\tilde{\mathbf{s}}_{m,t+1}\}^{\mathsf{T}}, \Im\mathfrak{m}\{\tilde{\mathbf{s}}_{m,t+1}\}^{\mathsf{T}}]^{\mathsf{T}}$.

## IV. SegOTA Design Optimization

### A. Joint Optimization Formulation

We focus on the uplink communication design involved in SegOTA, aiming to maximize the FL training convergence rate. For effective SegOTA, we consider joint transmit-receive beamforming in the uplink, where the device transmit beamforming weights $\{a_{k,t}\}_{k\in\mathcal{K}_{i,t}}$ and the BS receive beamforming vector $\mathbf{w}_{i,t}$ are designed jointly for each group $i$ in communication round $t$. In communication round $t$, the model segments $\{\tilde{\mathbf{s}}_{\hat{m}(i,t),t}^{k,J}\}_{k\in\mathcal{K}_{i,t}}$ from devices within the same group $i$ need to be combined coherently, as indicated in the first term of (3). To achieve this, we phase-align the effective channels $\{a_{k,t}\}_{k\in\mathcal{K}_{i,t}}$ among the devices in group $i$ via the joint transmit-receive beamforming. In particular, we set the transmit beamforming weight at device $k$ in group $i$ as $a_{k,t} = \sqrt{p_{k,t}}\frac{\mathbf{h}_{k,t}^{\mathsf{H}}\mathbf{w}_{i,t}}{|\mathbf{h}_{k,t}^{\mathsf{H}}\mathbf{w}_{i,t}|}$, where $p_{k,t}$ is the transmit power used for sending the entire segment at device $k$. Then, the effective channel of this device $k$ is given by

$$\alpha_{k,t} = \frac{\mathbf{w}_{i,t}^{\mathsf{H}}\mathbf{h}_{k,t}a_{k,t}}{\|\tilde{\mathbf{s}}_{\hat{m}(i,t),t}^{k,J}\|} = \frac{\sqrt{p_{k,t}}|\mathbf{h}_{k,t}^{\mathsf{H}}\mathbf{w}_{i,t}|}{\|\tilde{\mathbf{s}}_{\hat{m}(i,t),t}^{k,J}\|}, \ k \in \mathcal{K}_{i,t}. \quad (6)$$

Given a total number of device groups $S_t$, our goal is to optimize the device grouping and joint transmit-receive beamforming to minimize the expected optimality gap $\mathbb{E}[\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^\star\|^2]$ after $T$ communication rounds. Let $\mathbf{p}_t \triangleq [\mathbf{p}_{1,t}^\mathsf{T}, \ldots, \mathbf{p}_{S_t,t}^\mathsf{T}]^\mathsf{T}$, where $\mathbf{p}_{i,t} \in \mathbb{R}^{K_{i,t}}$ contains the transmit power of each device in group $i$ of round $t$, $p_{k,t}$, $k \in \mathcal{K}_{i,t}$. Also, $\mathbf{w}_t \triangleq [\mathbf{w}_{1,t}^\mathsf{H}, \ldots, \mathbf{w}_{S_t,t}^\mathsf{H}]^\mathsf{H} \in \mathbb{C}^{S_t N}$ contains the BS receive beamforming vectors for all $S_t$ groups in round $t$. Let $\mathcal{T} \triangleq \{0, \ldots, T-1\}$. Then, the joint optimization problem is formulated as

$$\mathcal{P}_o: \min_{\{\{\mathcal{K}_{i,t}\}_{i=1}^{S_t}, \mathbf{w}_t, \mathbf{p}_t\}_{t=0}^{T-1}} \mathbb{E}[\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^\star\|^2]$$

$$\text{s.t. } \mathcal{K}_{i,t} \bigcap \mathcal{K}_{j,t} = \varnothing, \quad i \neq j,\ i,j \in \mathcal{S}_t,\ t \in \mathcal{T}, \tag{7}$$

$$\bigcup_{i \in \mathcal{S}_t} \mathcal{K}_{i,t} = \mathcal{K}_{\text{tot}}, \quad t \in \mathcal{T}, \tag{8}$$

$$p_{k,t} \leq I_t P_k, \quad k \in \mathcal{K}_{\text{tot}},\ t \in \mathcal{T}, \tag{9}$$

$$\|\mathbf{w}_{i,t}\|^2 = 1, \quad i \in \mathcal{S}_t,\ t \in \mathcal{T} \tag{10}$$

where $\mathbb{E}[\cdot]$ is taken w.r.t. receiver noise and mini-batch local data samples at each device, (7) and (8) are the device grouping constraints, $P_k$ is the average transmit power budget at device $k$ for sending a signal and constraint (9) specifies the per-device average transmit power budget for sending the entire model segment in each communication round.

Problem $\mathcal{P}_o$ is a $T$-horizon stochastic optimization problem. To tackle this challenging problem, we first analyze the training convergence rate and develop a more tractable upper bound on $\mathbb{E}[\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^\star\|^2]$. Then, we propose an efficient scheme for device grouping and joint transmit-receive beamforming to minimize this upper bound.

### *B. SegOTA Training Convergence Analysis*

To analyze the FL training convergence speed, we make the following assumptions. They are commonly used in the existing literature for conventional full-model OTA aggregation, and here we have extended them to segmented OTA.

**Assumption 1.** The local loss function $F_k(\cdot)$ is $L$-smooth and $\lambda$-strongly convex, $k \in \mathcal{K}_{\text{tot}}$.

Moreover, let $\nabla F$ denote the gradient of the global loss function given in (1) w.r.t. $\boldsymbol{\theta}_t$. Denote segments $m$ of gradients $\nabla F(\boldsymbol{\theta}_t)$ and $\nabla F_k(\boldsymbol{\theta}_{k,t}^\tau)$ respectively by $\nabla F^m(\boldsymbol{\theta}_t)$ and $\nabla F_k^m(\boldsymbol{\theta}_{k,t}^\tau)$. We then make Assumption 2 below.

**Assumption 2.** Bounded gradient divergence: for all $t$, $\tau$, and $m$, $\mathbb{E}[\|\nabla F^m(\boldsymbol{\theta}_t) - \sum_k c_k \nabla F_k^m(\boldsymbol{\theta}_{k,t}^\tau)\|^2] \leq \phi$, for some $\phi \geq 0$ and $0 \leq c_k \leq 1$ such that $\sum_k c_k = 1$. Furthermore, for all $t$, $\tau$, $k$, and $m$, $\mathbb{E}[\|\nabla F_k^m(\boldsymbol{\theta}_{k,t}^\tau) - \nabla F_k^m(\boldsymbol{\theta}_{k,t}^\tau; \mathcal{B}_{k,t}^\tau)\|^2] \leq \mu$, for some $\mu \geq 0$.

We apply $a_{k,t}$ and $\alpha_{k,t}$ given at the beginning of Section IV-A into the segment updating equation (5) and further define $\Delta\tilde{\mathbf{s}}_{m,t} \triangleq \sum_{k \in \mathcal{K}_{\hat{i}(m,t),t}} \rho_{k,t} \Delta\tilde{\mathbf{s}}_{m,t}^k$. Moreover, let $\tilde{\mathbf{e}}_{m,t}$ be the fourth term for the inter-segment interference on the right-hand side of (5). Stacking all $S_t$ segments, $\tilde{\mathbf{s}}_{m,t+1}$, $m \in \mathcal{S}_t$, together, following (5), we express the entire global model update $\tilde{\boldsymbol{\theta}}_{t+1}$ from $\tilde{\boldsymbol{\theta}}_t$ as

$$\tilde{\boldsymbol{\theta}}_{t+1} = \tilde{\boldsymbol{\theta}}_t + \Delta\tilde{\boldsymbol{\theta}}_t + \tilde{\mathbf{n}}_t + \tilde{\mathbf{e}}_t \tag{11}$$

where $\Delta\tilde{\boldsymbol{\theta}}_t$ is the vector that stacks $\Delta\tilde{\mathbf{s}}_{m,t}$, $m \in \mathcal{S}_t$, and $\tilde{\mathbf{n}}_t$ and $\tilde{\mathbf{e}}_t$ are similarly defined.

We analyze the expected optimality gap, $\mathbb{E}[\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star\|^2]$ at round $t+1$. Based on (11), we can show that its upper bound is a function of $\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star\|^2]$. In particular, we first consider an ideal centralized model training procedure using the full gradient descent training algorithm and all the device datasets. We assume the BS has the device datasets and implements this centralized training procedure without exchanging any model updating information with the devices. Let $\mathbf{v}_t^\tau$ be the model update at iteration $\tau \in \{0, \ldots, J-1\}$, with $\mathbf{v}_t^0 = \boldsymbol{\theta}_t$. The ideal centralized model update is given by

$$\mathbf{v}_t^{\tau+1} = \mathbf{v}_t^\tau - \eta_t \nabla F(\mathbf{v}_t^\tau). \tag{12}$$

Let $\tilde{\boldsymbol{\theta}}^\star$, $\nabla\tilde{F}_k^m(\boldsymbol{\theta}_{k,t}^\tau)$, $\tilde{\mathbf{v}}_t^\tau$, and $\nabla\tilde{F}(\mathbf{v}_t^\tau)$ respectively denote the equivalent complex representations of $\boldsymbol{\theta}^\star$, $\nabla F_k^m(\boldsymbol{\theta}_{k,t}^\tau)$, $\mathbf{v}_t^\tau$, and $\nabla F(\mathbf{v}_t^\tau)$. Based on (11)(12), we have

$$\begin{aligned}
&\tilde{\boldsymbol{\theta}}_{t+1} - \tilde{\boldsymbol{\theta}}^\star \\
&= \tilde{\boldsymbol{\theta}}_t - \eta_t \sum_{\tau=0}^{J-1} \nabla\tilde{F}(\mathbf{v}_t^\tau) - \tilde{\boldsymbol{\theta}}^\star + \eta_t \sum_{\tau=0}^{J-1} \nabla\tilde{F}(\mathbf{v}_t^\tau) + \Delta\tilde{\boldsymbol{\theta}}_t + \tilde{\mathbf{n}}_t + \tilde{\mathbf{e}}_t \\
&= \tilde{\mathbf{v}}_t^J - \tilde{\boldsymbol{\theta}}^\star + \eta_t \sum_{\tau=0}^{J-1} \nabla\tilde{F}(\mathbf{v}_t^\tau) + \Delta\tilde{\boldsymbol{\theta}}_t + \tilde{\mathbf{n}}_t + \tilde{\mathbf{e}}_t \\
&= \tilde{\mathbf{v}}_t^J - \tilde{\boldsymbol{\theta}}^\star + \underbrace{\eta_t \sum_{\tau=0}^{J-1} \nabla\tilde{F}(\mathbf{v}_t^\tau) - \Delta\bar{\boldsymbol{\theta}}_t}_{\triangleq \tilde{\boldsymbol{\alpha}}_t} + \underbrace{\Delta\bar{\boldsymbol{\theta}}_t + \Delta\tilde{\boldsymbol{\theta}}_t}_{\triangleq \tilde{\boldsymbol{\beta}}_t} + \underbrace{\tilde{\mathbf{n}}_t + \tilde{\mathbf{e}}_t}_{\triangleq \tilde{\boldsymbol{\delta}}_t}
\end{aligned} \tag{13}$$

where $\Delta\bar{\boldsymbol{\theta}}_t$ stacks $\eta_t \sum_{k \in \mathcal{K}_{\hat{i}(m,t),t}} \rho_{k,t} \sum_{\tau=0}^{J-1} \nabla\tilde{F}_k^m(\boldsymbol{\theta}_{k,t}^\tau)$, $m \in \mathcal{S}_t$. Following the above, we have

$$\begin{aligned}
&\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_{t+1} - \tilde{\boldsymbol{\theta}}^\star\|^2] = \mathbb{E}[\|\tilde{\mathbf{v}}_t^J - \tilde{\boldsymbol{\theta}}^\star + \tilde{\boldsymbol{\alpha}}_t + \tilde{\boldsymbol{\beta}}_t + \tilde{\boldsymbol{\delta}}_t\|^2] \\
&\leq \mathbb{E}[(\|\tilde{\mathbf{v}}_t^J - \tilde{\boldsymbol{\theta}}^\star\| + \|\tilde{\boldsymbol{\alpha}}_t\| + \|\tilde{\boldsymbol{\beta}}_t\| + \|\tilde{\boldsymbol{\delta}}_t\|)^2] \\
&\overset{(a)}{\leq} 4\big(\mathbb{E}[\|\tilde{\mathbf{v}}_t^J - \tilde{\boldsymbol{\theta}}^\star\|^2] + \mathbb{E}[\|\tilde{\boldsymbol{\alpha}}_t\|^2] + \mathbb{E}[\|\tilde{\boldsymbol{\beta}}_t\|^2] + \mathbb{E}[\|\tilde{\boldsymbol{\delta}}_t\|^2]\big)
\end{aligned} \tag{14}$$

where $(a)$ is based on a specific case of the Cauchy–Schwarz inequality $(\sum_{i=1}^G x_i)^2 \leq G \sum_{i=1}^G x_i^2, \forall x_i \in \mathbb{R}$, for some $G \in \mathbb{N}^+$. We upper bound each term in (14) below.

We first obtain an upper bound for $\mathbb{E}[\|\tilde{\mathbf{v}}_t^J - \tilde{\boldsymbol{\theta}}^\star\|^2]$ in Lemma 1. The proof uses the same technique as in [16, Lemma 2] and thus is omitted.

**Lemma 1** (Bounding $\mathbb{E}[\|\tilde{\mathbf{v}}_t^J - \tilde{\boldsymbol{\theta}}^\star\|^2]$)**.** Consider SegOTA described in Section III and the ideal centralized training described in Section IV-B. For $\eta_t < \frac{1}{L}$, $\forall t \in \mathcal{T}$, under Assumption 1, $\mathbb{E}[\|\tilde{\mathbf{v}}_t^J - \tilde{\boldsymbol{\theta}}^\star\|^2]$ is upper bounded as

$$\mathbb{E}[\|\tilde{\mathbf{v}}_t^J - \tilde{\boldsymbol{\theta}}^\star\|^2] \leq (1 - \eta_t\lambda)^{2J} \mathbb{E}[\|\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}^\star\|^2],\ t \in \mathcal{T}. \tag{15}$$

Next, we bound the terms $\mathbb{E}[\|\tilde{\boldsymbol{\alpha}}_t\|^2]$, $\mathbb{E}[\|\tilde{\boldsymbol{\beta}}_t\|^2]$, and $\mathbb{E}[\|\tilde{\boldsymbol{\delta}}_t\|^2]$ respectively in the following lemma.

**Lemma 2** (Bounding $\mathbb{E}[\|\tilde{\boldsymbol{\alpha}}_t\|^2]$, $\mathbb{E}[\|\tilde{\boldsymbol{\beta}}_t\|^2]$, and $\mathbb{E}[\|\tilde{\boldsymbol{\delta}}_t\|^2]$)**.** Consider SegOTA described in Section III and joint transmit-receive beamforming described at the beginning of Section IV-A. Let $\nu \triangleq \max_{k\in\mathcal{K}_{\text{tot}},m\in\mathcal{S}_t,t\in\mathcal{T}} \|\tilde{\mathbf{s}}_{m,t}^{k,J}\|^2$ and $\tilde{\sigma}_t^2 \triangleq \sigma^2 I_t/2$. Under Assumption 2, $\mathbb{E}[\|\tilde{\boldsymbol{\alpha}}_t\|^2]$, $\mathbb{E}[\|\tilde{\boldsymbol{\beta}}_t\|^2]$, and $\mathbb{E}[\|\tilde{\boldsymbol{\delta}}_t\|^2]$ are respectively upper bounded as

$$\mathbb{E}[\|\tilde{\boldsymbol{\alpha}}_t\|^2] \le \eta_t^2 J^2 S_t \phi,\ t\in\mathcal{T}, \tag{16}$$

$$\mathbb{E}[\|\tilde{\boldsymbol{\beta}}_t\|^2] \le \eta_t^2 J^2 S_t K^2 \mu,\ t\in\mathcal{T}, \tag{17}$$

$$\mathbb{E}[\|\tilde{\boldsymbol{\delta}}_t\|^2] \le \nu \sum_{i=1}^{S_t} \frac{\tilde{\sigma}_t^2}{(\sum_{k\in\mathcal{K}_{i,t}} \sqrt{p_{k,t}}|\mathbf{h}_{k,t}^{\mathsf{H}}\mathbf{w}_{i,t}|)^2} + \nu \sum_{i=1}^{S_t}\sum_{j\neq i} K_{j,t} \frac{\sum_{j\neq i}\sum_{q\in\mathcal{K}_{j,t}} p_{q,t}|\mathbf{h}_{q,t}^{\mathsf{H}}\mathbf{w}_{i,t}|^2}{(\sum_{k\in\mathcal{K}_{i,t}} \sqrt{p_{k,t}}|\mathbf{h}_{k,t}^{\mathsf{H}}\mathbf{w}_{i,t}|)^2},\ t\in\mathcal{T}. \tag{18}$$

*Proof:* See Appendix A. ∎

Using the above, we obtain an upper bound on $\mathbb{E}[\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^\star\|^2]$ in the following proposition.

**Proposition 1.** For SegOTA described in Section III, under Assumptions 1–2 and for $\eta_t < \frac{1}{L}$, $\forall t\in\mathcal{T}$, the expected model optimality gap after $T$ communication rounds is bounded by

$$\mathbb{E}[\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^\star\|^2] \le \sum_{t=0}^{T-1} \overline{G}_t\big(H_t(\{\mathcal{K}_{i,t}\},\mathbf{w}_t,\mathbf{p}_t) + C_t\big) + \Gamma \prod_{t=0}^{T-1} G_t \tag{19}$$

where $\Gamma \triangleq \mathbb{E}[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\|^2]$, $G_t \triangleq 4(1-\eta_t\lambda)^{2J}$, $C_t \triangleq 4\eta_t^2 J^2 S_t(\phi + K^2\mu)$, $\overline{G}_t \triangleq \prod_{s=t+1}^{T-1} G_s$ with $\overline{G}_{T-1}=1$, and

$$H_t(\{\mathcal{K}_{i,t}\},\mathbf{w}_t,\mathbf{p}_t) \triangleq 4\nu \sum_{i=1}^{S_t} \frac{\tilde{\sigma}_t^2}{(\sum_{k\in\mathcal{K}_{i,t}} \sqrt{p_{k,t}}|\mathbf{h}_{k,t}^{\mathsf{H}}\mathbf{w}_{i,t}|)^2} + 4\nu \sum_{i=1}^{S_t}\sum_{j\neq i} K_{j,t} \frac{\sum_{j\neq i}\sum_{q\in\mathcal{K}_{j,t}} p_{q,t}|\mathbf{h}_{q,t}^{\mathsf{H}}\mathbf{w}_{i,t}|^2}{(\sum_{k\in\mathcal{K}_{i,t}} \sqrt{p_{k,t}}|\mathbf{h}_{k,t}^{\mathsf{H}}\mathbf{w}_{i,t}|)^2}. \tag{20}$$

*Proof:* Combining (14)–(18), we have

$$\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_{t+1} - \tilde{\boldsymbol{\theta}}^\star\|^2] \le G_t \mathbb{E}[\|\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}^\star\|^2] + H_t(\{\mathcal{K}_{i,t}\},\mathbf{w}_t,\mathbf{p}_t) + C_t. \tag{21}$$

Summing up both sides of (21) over $t\in\mathcal{T}$ and rearranging the terms, we have (19). ∎

### C. Beamforming Optimization for SegOTA

We replace the objective function in $\mathcal{P}_o$ with the upper bound in (19). Omitting the constant terms in (19) that do not depend on the optimization variables, we arrive at an equivalent optimization problem with the objective function $\sum_{t=0}^{T-1} \overline{G}_t H_t(\{\mathcal{K}_{i,t}\},\mathbf{w}_t,\mathbf{p}_t)$. By Proposition 1 and Assumption 1, we have $\eta_t \le \frac{1}{L} \le \frac{1}{\lambda}$, $\forall t\in\mathcal{T}$, which leads to $G_t > 0$ and further $\overline{G}_t > 0$. Hence, we separate this optimization problem into $T$ per-round problems, each minimizing $H_t(\{\mathcal{K}_{i,t}\},\mathbf{w}_t,\mathbf{p}_t)$ in communication round $t$.

Next, we apply $\sum_{k\in\mathcal{K}_{i,t}} p_{k,t}|\mathbf{h}_{k,t}^{\mathsf{H}}\mathbf{w}_{i,t}|^2 \le (\sum_{k\in\mathcal{K}_{i,t}} \sqrt{p_{k,t}}|\mathbf{h}_{k,t}^{\mathsf{H}}\mathbf{w}_{i,t}|)^2$ to the denominators in (20) to further upper bound $H_t(\{\mathcal{K}_{i,t}\},\mathbf{w}_t,\mathbf{p}_t)$. Then, we arrive at the following per-round online optimization problem:

$$\begin{aligned}\mathcal{P}_{1,t}: \min_{\{\mathcal{K}_{i,t}\}_{i=1}^{S_t},\mathbf{w}_t,\mathbf{p}_t}\ & \sum_{i=1}^{S_t} \frac{Z_{i,t}\sum_{j\neq i}\sum_{q\in\mathcal{K}_{j,t}} p_{q,t}|\mathbf{f}_{q,t}^{\mathsf{H}}\mathbf{w}_{i,t}|^2 + 1}{\sum_{k\in\mathcal{K}_{i,t}} p_{k,t}|\mathbf{f}_{k,t}^{\mathsf{H}}\mathbf{w}_{i,t}|^2} \\ \text{s.t.}\ & \mathcal{K}_{i,t}\bigcap\mathcal{K}_{j,t} = \varnothing,\quad i\neq j,\ i,j\in\mathcal{S}_t, \\ & \bigcup_{i\in\mathcal{S}_t}\mathcal{K}_{i,t} = \mathcal{K}_{\text{tot}}, \\ & p_{k,t} \le I_t P_k,\quad k\in\mathcal{K}_{\text{tot}}, \\ & \|\mathbf{w}_{i,t}\|^2 = 1,\quad i\in\mathcal{S}_t\end{aligned}$$

where $Z_{i,t} \triangleq \sum_{j\neq i} K_{j,t}$ and $\mathbf{f}_{k,t} \triangleq \mathbf{h}_{k,t}/\tilde{\sigma}_t$. Note that the objective function represents a sum of the inverse of received signal-to-interference-and-noise ratio (SINR) corresponding to received aggregated segment from each device group.

The above is a mixed-integer programming problem. Furthermore, the objective function is nonconvex w.r.t. the power vector $\mathbf{p}_t$ and the beamforming vector $\mathbf{w}_t$, which is challenging to solve. We propose a device grouping scheme based on spherical *k*-means to first obtain $\{\mathcal{K}_{i,t}\}$. Based on this, we then optimize the joint transmit-receive beamforming $(\mathbf{w}_t,\mathbf{p}_t)$ in $\mathcal{P}_{1,t}$.

*1) Device grouping via spherical k-means:* Since receive beamforming $\mathbf{w}_{i,t}$ is applied to the devices of the same group $i$, more spatially correlated device channels can lead to higher received beamforming gain for this group [17]. For this purpose, we propose a device grouping scheme that uses the clustering idea to find the spatially correlated device groups. The scheme is based on the spherical *k*-means framework [18], which is a variant of the standard *k*-means that captures the cosine similarity among data points to form clusters. In particular, we first define the feature space for the purpose of device grouping. Let $\mathcal{X}_t$ denote the feature space spanned by the device uplink channels in round $t$, given by

$$\mathcal{X}_t = \left\{\mathbf{x}_{k,t} : \mathbf{x}_{k,t} \triangleq \frac{\mathbf{h}_{k,t}}{\|\mathbf{h}_{k,t}\|} e^{-j\angle h_{1k,t}},\ \forall k\in\mathcal{K}_{\text{tot}}\right\}$$

where $\angle h_{1k,t}$ denotes the phase of the first element in $\mathbf{h}_{k,t}$. Each data point $\mathbf{x}_{k,t}$ in $\mathcal{X}_t$ is phase-adjusted such that its first element is phase-aligned to 0 degree. This is to ensure that during the iterative updating process, all $\mathbf{x}_{k,t}$'s are phase-aligned to sum up properly in computing the centroid. Let $\mathbf{c}_{r,t}$ denote the centroid of cluster $r=1,\ldots,S_t$ in $\mathcal{X}_t$ with $\|\mathbf{c}_{r,t}\| = 1$. We consider the following metric to measure distance from each data point $\mathbf{x}_{k,t}$ in $\mathcal{X}_t$ to a centroid $\mathbf{c}_{r,t}$:

$$\delta(\mathbf{x},\mathbf{c}) = |\mathbf{x}_{k,t}^{\mathsf{H}}\mathbf{c}_{r,t}|,\ \forall \mathbf{x}_{k,t}\in\mathcal{X}_t \tag{22}$$

where $\delta(\mathbf{x},\mathbf{c}) \in [0,1]$. This distance metric measures the correlation level between the channel vector and the centroid, with 1 being fully correlated and 0 being orthogonal. A data point $\mathbf{x}_{k,t}$ will be included in cluster $\mathbf{c}_{r,t}$ where it has the largest $\delta(\mathbf{x},\mathbf{c})$ among all clusters. Specifically, denote the set of $\mathbf{x}_{k,t}$'s in cluster $\mathbf{c}_{r,t}$ by

$$\mathcal{Y}_{r,t} = \{\mathbf{x}_{k,t}\in\mathcal{X}_t : \delta(\mathbf{x}_{k,t},\mathbf{c}_{r,t}) > \delta(\mathbf{x}_{k,t},\mathbf{c}_{r't}),\ r'\neq r\}.$$

Given $\mathcal{X}_t$ and $\delta(\mathbf{x},\mathbf{c})$, we then apply the spherical $k$-means method [18] to form $S_t$ centroid points and clusters in $\mathcal{X}_t$ and iteratively update the centroid points $\mathbf{c}_{r,t}$'s. The centroid update $\mathbf{c}_{r,t}^{(l+1)}$ at iteration $l$ is then given by

$$\mathbf{c}_{r,t}^{(l+1)} = \frac{\sum_{\mathbf{x}_{k,t}\in\mathcal{Y}_{r,t}} \mathbf{x}_{k,t}}{|\mathcal{Y}_{r,t}|};\ \ \mathbf{c}_{r,t}^{(l+1)} \leftarrow \frac{\mathbf{c}_{r,t}^{(l+1)}}{\|\mathbf{c}_{r,t}^{(l+1)}\|}. \tag{23}$$

The above procedure is repeated until convergence to obtain a device grouping solution $\{\mathcal{K}_{i,t}\}_{i=1}^{S_t}$.

*2) Joint transmit-receive beamforming:* Given $\{\mathcal{K}_{i,t}\}_{i=1}^{S_t}$, we now optimize uplink beamforming $(\mathbf{w}_t,\mathbf{p}_t)$ in $\mathcal{P}_{1,t}$. Since the problem is nonconvex, we propose to alternatingly optimize receive beamforming $\mathbf{w}_t$ and the device transmit powers in $\mathbf{p}_t$ via the block coordinate descent (BCD) method [19]. The two subproblems are given below:

*i) Updating* $\mathbf{w}_t$*:* Given $\mathbf{p}_t$, $\mathcal{P}_{1,t}$ can be equivalently decomposed into $S_t$ subproblems, one for each beamformer $\mathbf{w}_{i,t}$ for each device group $i$ as

$$\mathcal{P}_{1,t,i}^{\text{wsub1}}:\ \min_{\mathbf{w}_{i,t}} \frac{Z_{i,t}\sum_{j\neq i}\sum_{q\in\mathcal{K}_{j,t}} p_{q,t}|\mathbf{f}_{q,t}^{\mathsf{H}}\mathbf{w}_{i,t}|^2+1}{\sum_{k\in\mathcal{K}_{i,t}} p_{k,t}|\mathbf{f}_{k,t}^{\mathsf{H}}\mathbf{w}_{i,t}|^2} \quad \text{s.t.}\ \ \|\mathbf{w}_{i,t}\|^2=1.$$

After expanding the quadratic terms in $\mathcal{P}_{1,t,i}^{\text{wsub1}}$, $\mathbf{w}_{i,t}$ can be moved outside of the summation at both the numerator and denominator, given by

$$\mathcal{P}_{1,t,i}^{\text{wsub2}}:\ \min_{\mathbf{w}_{i,t}} \frac{\mathbf{w}_{i,t}^{\mathsf{H}}\left(Z_{i,t}\sum_{j\neq i}\sum_{q\in\mathcal{K}_{j,t}} p_{q,t}\mathbf{f}_{q,t}\mathbf{f}_{q,t}^{\mathsf{H}}+\mathbf{I}\right)\mathbf{w}_{i,t}}{\mathbf{w}_{i,t}^{\mathsf{H}}\left(\sum_{k\in\mathcal{K}_{i,t}} p_{k,t}\mathbf{f}_{k,t}\mathbf{f}_{k,t}^{\mathsf{H}}\right)\mathbf{w}_{i,t}} \quad \text{s.t.}\ \ \|\mathbf{w}_{i,t}\|^2=1,$$

which is a generalized eigenvalue problem. Specifically, let $\mathbf{A}_{i,t} \triangleq Z_{i,t}\sum_{j\neq i}\sum_{q\in\mathcal{K}_{j,t}} p_{q,t}\mathbf{f}_{q,t}\mathbf{f}_{q,t}^{\mathsf{H}}+\mathbf{I}$ and $\mathbf{B}_{i,t} \triangleq \sum_{k\in\mathcal{K}_{i,t}} p_{k,t}\mathbf{f}_{k,t}\mathbf{f}_{k,t}^{\mathsf{H}}$. We re-express problem $\mathcal{P}_{1,t,i}^{\text{wsub2}}$ as

$$\mathcal{P}_{1,t,i}^{\text{wsub3}}:\ \min_{\mathbf{w}_{i,t}} \frac{\mathbf{w}_{i,t}^{\mathsf{H}}\mathbf{A}_{i,t}\mathbf{w}_{i,t}}{\mathbf{w}_{i,t}^{\mathsf{H}}\mathbf{B}_{i,t}\mathbf{w}_{i,t}} \quad \text{s.t.}\ \ \|\mathbf{w}_{i,t}\|^2=1.$$

The optimal solution $\mathbf{w}_{i,t}$ of $\mathcal{P}_{1,t,i}^{\text{wsub3}}$ is the generalized eigenvector corresponding to the smallest generalized eigenvalue in the generalized eigenvalue problem of $\mathbf{A}_{i,t}\mathbf{\Phi}_{i,t} = \mathbf{B}_{i,t}\mathbf{\Phi}_{i,t}\mathbf{\Lambda}_{i,t}$, where $\mathbf{\Phi}_{i,t}$ is the eigenvector matrix and $\mathbf{\Lambda}_{i,t}$ a diagonal matrix with its diagonal elements being the eigenvalues. Let $\mathbf{\Phi}_{i,t}^{\text{A}}$ and $\mathbf{\Phi}_{i,t}^{\text{B}}$ denote the eigenvector matrices of $\mathbf{A}_{i,t}$ and $\mathbf{B}_{i,t}$, respectively. Let $\mathbf{\Lambda}_{i,t}^{\text{B}}$ be a diagonal matrix with its diagonal elements being the eigenvalues of $\mathbf{B}_{i,t}$. Then, the generalized eigenvector matrix is given by

$$\mathbf{\Phi}_{i,t} = \mathbf{\Phi}_{i,t}^{\text{B}}(\mathbf{\Lambda}_{i,t}^{\text{B}})^{-1/2}\mathbf{\Phi}_{i,t}^{\text{A}}. \tag{24}$$

The optimal solution $\mathbf{w}_{i,t}$ is a column vector in $\mathbf{\Phi}_{i,t}$ corresponding to the smallest generalized eigenvalue.

*ii) Updating* $\mathbf{p}_t$*:* Let $\mathbf{g}_{ij,t}$ be the vector containing $\{g_{iq,t} \triangleq |\mathbf{f}_{q,t}^{\mathsf{H}}\mathbf{w}_{i,t}|^2,\ q\in\mathcal{K}_{j,t}\}$ from group $j$ after applying receive beamformer $\mathbf{w}_{i,t}$. Given $\mathbf{w}_t$, we can rewrite $\mathcal{P}_{1,t}$ as

$$\mathcal{P}_{1,t}^{\text{psub1}}:\ \min_{\mathbf{p}_t} \sum_{i=1}^{S_t} \frac{Z_{i,t}\sum_{j\neq i}\mathbf{g}_{ij,t}^{\mathsf{T}}\mathbf{p}_{j,t}+1}{\mathbf{g}_{ii,t}^{\mathsf{T}}\mathbf{p}_{i,t}} \quad \text{s.t.}\ \ p_{k,t}\le I_t P_k,\quad k\in\mathcal{K}_{\text{tot}}.$$

To efficiently compute $\mathbf{p}_t$, we adopt BCD to update $\mathbf{p}_{1,t},\ldots,\mathbf{p}_{S_t,t}$ alternatingly, one for each group $i$. Specifically, given $\mathbf{p}_{j,t}$, $\forall j\in\mathcal{S}_t, j\neq i$, the optimization of $\mathbf{p}_{i,t}$ for group $i$ is given by

$$\mathcal{P}_{1,t,i}^{\text{psub2}}:\ \min_{\mathbf{p}_{i,t}} \frac{Z_{i,t}\sum_{j\neq i}\mathbf{g}_{ij,t}^{\mathsf{T}}\mathbf{p}_{j,t}+1}{\mathbf{g}_{ii,t}^{\mathsf{T}}\mathbf{p}_{i,t}} + \sum_{j\neq i}\frac{Z_{j,t}\mathbf{g}_{ji,t}^{\mathsf{T}}}{\mathbf{g}_{jj,t}^{\mathsf{T}}\mathbf{p}_{j,t}}\mathbf{p}_{i,t} \quad \text{s.t.}\ \ p_{k,t}\le I_t P_k,\quad k\in\mathcal{K}_{i,t}.$$

Problem $\mathcal{P}_{1,t,i}^{\text{psub2}}$ is convex w.r.t. $\mathbf{p}_{i,t}$, for which the optimal $\mathbf{p}_{i,t}$ can be obtained in closed-form. Specifically, let

$$\beta_{i,t}^{\min} \triangleq \min_{k\in\mathcal{K}_{i,t}} \left(\frac{Z_{i,t}\sum_{j\neq i}\mathbf{g}_{ij,t}^{\mathsf{T}}\mathbf{p}_{j,t}+1}{\sum_{j\neq i}\frac{Z_{j,t}g_{jk,t}}{\mathbf{g}_{jj,t}^{\mathsf{T}}\mathbf{p}_{j,t}}} g_{ik,t}\right)^{1/2},$$

and let $k'\in\mathcal{K}_{i,t}$ be the corresponding device index that achieves $\beta_{i,t}^{\min}$. Let $\overline{\mathbf{P}}_i$ be the vector containing the maximum power of devices in group $i$ $\{P_k, k\in\mathcal{K}_{i,t}\}$. Then, each device transmit power $p_{k,t}, k\in\mathcal{K}_{i,t}$ in the optimal $\mathbf{p}_{i,t}$ is given by

$$p_{k,t} = \begin{cases} I_t P_k & \text{for } k\in\mathcal{K}_{i,t}, k\neq k' \\ I_t P_{k'} - \dfrac{\left[I_t\mathbf{g}_{ii,t}^{\mathsf{T}}\overline{\mathbf{P}}_i - \beta_{i,t}^{\min}\right]^+}{g_{ik',t}} & \text{for } k=k' \end{cases}$$

where $[\beta]^+=\max\{\beta,0\}$. All $\mathbf{p}_{i,t}$'s are updated alternatingly using the above solution.

Since subproblems $\mathcal{P}_{1,t,i}^{\text{wsub3}}$ and $\mathcal{P}_{1,t,i}^{\text{psub2}}$ for each $i$ are solved optimally, and our optimization objective is lower bounded by zero, by the monotone convergence theorem of BCD [19], our proposed algorithm for computing $(\mathbf{w}_t,\mathbf{p}_t)$ is guaranteed to converge to a stationary point.

## V. Simulation Results

*1) Simulation Setup:* We evaluate our proposed SegOTA for FL on image classification over a simulated wireless network under typical wireless specifications with system bandwidth 1 MHz, carrier frequency 2 GHz, and per-device maximum transmit power at each device $P_k = 23$ dBm. We generate channels as $\mathbf{h}_{k,t} = \sqrt{G_k}\bar{\mathbf{h}}_{k,t}$, where $\bar{\mathbf{h}}_{k,t}\sim\mathcal{CN}(\mathbf{0},\mathbf{I})$ and the path gain $G_k[\text{dB}] = -136.3 - 35\log_{10} d_k - \psi_k$, with BS-device distance $d_k$ being uniformly distributed within $(0.02, 0.5)$ km and shadowing variable $\psi_k\sim\mathcal{N}(0,\sigma_\psi^2)$ with $\sigma_\psi = 8$ dB. We set the receiver noise power $\sigma^2 = -79$ dBm, which accounts for both thermal noise and inter-cell interference.

We use the MNIST dataset for model training and testing. It consists of $6\times10^4$ training samples and $1\times10^4$ test samples. We trains a convolutional neural network with an $8\times3\times3$ ReLU convolutional layer, a $2\times2$ max pooling layer, a ReLU fully-connected layer with 20 units, and a softmax output layer, with $D = 2.735\times10^4$ model parameters. We use the $10^4$ test samples to measure the test accuracy of the global model update $\boldsymbol{\theta}_t$ at each round $t$. The training samples are randomly and evenly distributed across $K$ devices, with local dataset size $A_k = \frac{6\times10^4}{K}$ at device $k$. For local training using SGD, we set $J = 100$, mini-batch size $|\mathcal{B}_{k,t}^{\tau}| = 600/K, \forall t,\tau,k,$

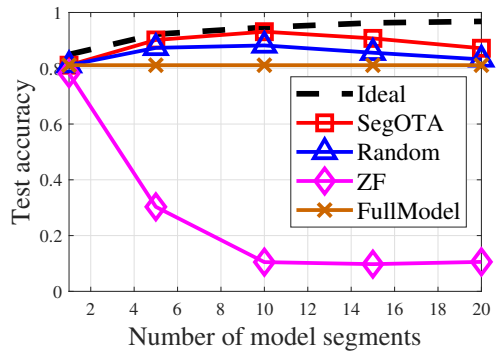


Fig. 2. Test accuracy vs. number of model segments $S_t$ ($(N, K) = (32, 50)$).

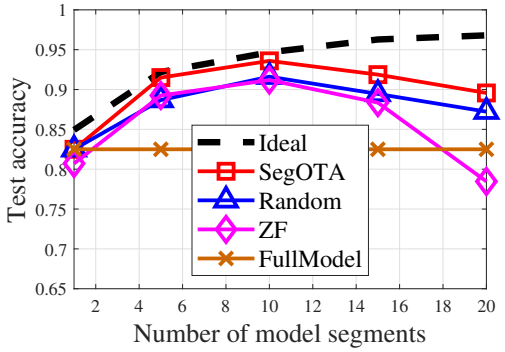


Fig. 3. Test accuracy vs. number of model segments $S_t$ ($(N, K) = (64, 50)$).

and learning rate $\eta_t = 0.1$, $\forall t$. All results are obtained by averaging over 20 channel realizations.

*2) Performance Comparison:* For comparison, we consider the following five schemes: i) **Ideal**: SegOTA via (5) with noise-interference-free uplink and perfect recovery of model parameters, which serves as a performance upper bound for all schemes. ii) **SegOTA**: our proposed method. iii) **Random**: the same as the proposed SegOTA, except that devices are randomly partitioned into equal-sized groups instead of our grouping method proposed in Section IV-C1. iv) **ZF**: the same as the proposed SegOTA and device grouping in Section IV-C1, except that we apply zero-forcing receive beamforming under maximum device transmit powers [20]. v) **FullModel**: traditional full-model OTA approach, which is equivalent to SegOTA with $S_t = 1, \forall t$.

We study the tradeoff between communication efficiency in the uplink model transmission and test accuracy of the global model updates. Figs. 2–4 compare the test accuracy performance of the five methods using a total of $2.736 \times 10^4$ channel uses per device for uplink model transmission. Each device group is randomly assigned a unique segment, as mentioned at the beginning of Section III. Fig. 2 and 3 show the test accuracy vs. $S_t$ model segments for the overloaded setup $(N, K) = (32, 50)$ and the underloaded setup $(N, K) = (64, 50)$, respectively. SegOTA outperforms both Random and ZF for all $S_t > 1$. Furthermore, SegOTA nearly attains the performance of Ideal for $S_t \leq 10$. Fig. 4 shows the test

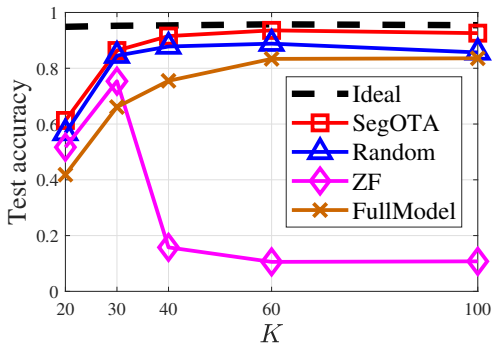


Fig. 4. Test accuracy vs. number of devices $K$ for $N = 32$ and $S_t = 10$.

accuracy vs. $K$ devices for $N = 32$ and $S_t = 10$. Again, we see that SegOTA nearly attains the optimal performance under Ideal and outperforms the other alternatives. We also observe that ZF can perform worse than FullModel, especially when $S_t$ or $K$ is large. This shows the effectiveness of our proposed beamforming algorithm for SegOTA targeting at FL training performance, while conventional ZF for interference cancellation may not be effective.

## VI. Conclusion

This paper proposes a segmented transmission approach SegOTA to reduce the latency in uplink OTA aggregation for wireless FL. Under SegOTA, devices are divided into groups, where each group is assigned a segment of the model for OTA aggregation, and all segments are sent via uplink simultaneously. Based on the segment global updating equation, we derive an upper bound on the model training optimality gap to formulate a joint transmit-receive beamforming problem along with device grouping. We propose a device grouping scheme based on their spatial channel correlations via spherical $k$-means and an iterative uplink beamforming algorithm with fast closed-form updates. Simulation results show the proposed SegOTA with proposed device grouping and uplink beamforming outperforms the traditional full-model OTA approach and other alternatives.

## Appendix A
## Proof of Lemma 2

*Proof:* For bounding $\mathbb{E}[\|\tilde{\boldsymbol{\alpha}}_t\|^2]$, we have

$$\mathbb{E}[\|\tilde{\boldsymbol{\alpha}}_t\|^2] \overset{(a)}{=} \eta_t^2 \sum_{m=1}^{S_t} \mathbb{E}\left[\left\| \sum_{\tau=0}^{J-1} \nabla \tilde{F}^m(\mathbf{v}_t^\tau) - \sum_{k \in \mathcal{K}_{\hat{i}(m,t),t}} \rho_{k,t} \sum_{\tau=0}^{J-1} \nabla \tilde{F}_k^m(\boldsymbol{\theta}_{k,t}^\tau) \right\|^2\right]$$

$$\leq \eta_t^2 \sum_{m=1}^{S_t} \mathbb{E}\left[\left( \sum_{\tau=0}^{J-1} \left\| \nabla \tilde{F}^m(\mathbf{v}_t^\tau) - \sum_{k \in \mathcal{K}_{\hat{i}(m,t),t}} \rho_{k,t} \nabla \tilde{F}_k^m(\boldsymbol{\theta}_{k,t}^\tau) \right\| \right)^2\right]$$

$$\overset{(b)}{\leq} \eta_t^2 J \sum_{m=1}^{S_t} \sum_{\tau=0}^{J-1} \mathbb{E}\Bigg[\Bigg\|\nabla \tilde{F}^m(\mathbf{v}_t^\tau) - \sum_{k\in\mathcal{K}_{\hat{i}(m,t),t}} \rho_{k,t} \nabla \tilde{F}_k^m(\boldsymbol{\theta}_{k,t}^\tau)\Bigg\|^2\Bigg]$$
$$\overset{(c)}{\leq} \eta_t^2 J^2 S_t \phi.$$

where $(a)$ uses the expression of $\tilde{\boldsymbol{\alpha}}_t$ in (13), $(b)$ is based on $(\sum_{i=1}^G x_i)^2 \leq G\sum_{i=1}^G x_i^2, \forall x_i \in \mathbb{R}$, for some $G \in \mathbb{N}^+$, and $(c)$ follows Assumption 2. Thus, we have (16).

For bounding $\mathbb{E}[\|\tilde{\boldsymbol{\beta}}_t\|^2]$, we have

$$\mathbb{E}[\|\tilde{\boldsymbol{\beta}}_t\|^2] \overset{(a)}{=} \eta_t^2 \sum_{m=1}^{S_t} \mathbb{E}\Bigg[\Bigg\| \sum_{k\in\mathcal{K}_{\hat{i}(m,t),t}} \rho_{k,t} \sum_{\tau=0}^{J-1} \Big(\nabla \tilde{F}_k^m(\boldsymbol{\theta}_{k,t}^\tau) - \nabla \tilde{F}_k^m(\boldsymbol{\theta}_{k,t}^\tau; \mathcal{B}_{k,t}^\tau)\Big)\Bigg\|^2\Bigg]$$
$$\leq \eta_t^2 \sum_{m=1}^{S_t} \mathbb{E}\Bigg[\Bigg(\sum_{k\in\mathcal{K}_{\hat{i}(m,t),t}} \sum_{\tau=0}^{J-1} |\rho_{k,t}| \Big\|\nabla \tilde{F}_k^m(\boldsymbol{\theta}_{k,t}^\tau) - \nabla \tilde{F}_k^m(\boldsymbol{\theta}_{k,t}^\tau; \mathcal{B}_{k,t}^\tau)\Big\|\Bigg)^2\Bigg]$$
$$\overset{(b)}{\leq} \eta_t^2 \sum_{m=1}^{S_t} \sum_{k\in\mathcal{K}_{\hat{i}(m,t),t}} \sum_{\tau=0}^{J-1} J K_{\hat{i}(m,t),t} \mathbb{E}\Big[\Big\|\nabla \tilde{F}_k^m(\boldsymbol{\theta}_{k,t}^\tau) - \nabla \tilde{F}_k^m(\boldsymbol{\theta}_{k,t}^\tau; \mathcal{B}_{k,t}^\tau)\Big\|^2\Big]$$
$$\overset{(c)}{\leq} \eta_t^2 J^2 S_t K^2 \mu$$

where $(a)$ uses the expression of $\tilde{\boldsymbol{\delta}}_t$ in (13), $(b)$ is based on $(\sum_{i=1}^G x_i)^2 \leq G\sum_{i=1}^G x_i^2, \forall x_i \in \mathbb{R}$, for some $G \in \mathbb{N}^+$, and $(c)$ follows Assumption 2. Thus, we have (17).

For bounding $\mathbb{E}[\|\tilde{\boldsymbol{\delta}}_t\|^2]$, we have

$$\mathbb{E}[\|\tilde{\boldsymbol{\delta}}_t\|^2] \overset{(a)}{=} \sum_{m=1}^{S_t} \mathbb{E}\Bigg[\Bigg\|\tilde{\mathbf{n}}_{m,t} + \frac{1}{\alpha_{m,t}^{\mathrm{s}}} \cdot \sum_{j\neq\hat{i}(m,t)} \sum_{q\in\mathcal{K}_{j,t}} \frac{\mathbf{h}_{q,t}^{\mathsf{H}} \mathbf{w}_{j,t} \mathbf{w}_{\hat{i}(m,t),t}^{\mathsf{H}} \mathbf{h}_{q,t}}{|\mathbf{h}_{q,t}^{\mathsf{H}} \mathbf{w}_{j,t}|} \cdot \frac{\sqrt{p_{q,t}} \tilde{\mathbf{s}}_{\hat{m}(j,t),t}^{q,J}}{\|\tilde{\mathbf{s}}_{\hat{m}(j,t),t}^{q,J}\|}\Bigg\|^2\Bigg]$$
$$\leq \sum_{m=1}^{S_t} \mathbb{E}\Bigg[\Bigg(\|\tilde{\mathbf{n}}_{m,t}\| + \frac{1}{\alpha_{m,t}^{\mathrm{s}}} \cdot \sum_{j\neq\hat{i}(m,t)} \sum_{q\in\mathcal{K}_{j,t}} \Bigg\|\frac{\mathbf{h}_{q,t}^{\mathsf{H}} \mathbf{w}_{j,t} \mathbf{w}_{\hat{i}(m,t),t}^{\mathsf{H}} \mathbf{h}_{q,t}}{|\mathbf{h}_{q,t}^{\mathsf{H}} \mathbf{w}_{j,t}|} \cdot \frac{\sqrt{p_{q,t}} \tilde{\mathbf{s}}_{\hat{m}(j,t),t}^{q,J}}{\|\tilde{\mathbf{s}}_{\hat{m}(j,t),t}^{q,J}\|}\Bigg\|\Bigg)^2\Bigg]$$
$$\overset{(b)}{\leq} \nu \sum_{i=1}^{S_t} \frac{\tilde{\sigma}_t^2}{(\sum_{k\in\mathcal{K}_{i,t}} \sqrt{p_{k,t}} |\mathbf{h}_{k,t}^{\mathsf{H}} \mathbf{w}_{i,t}|)^2} + \nu \sum_{i=1}^{S_t} \sum_{j\neq i} K_{j,t} \frac{\sum_{j\neq i} \sum_{q\in\mathcal{K}_{j,t}} p_{q,t} |\mathbf{h}_{q,t}^{\mathsf{H}} \mathbf{w}_{i,t}|^2}{(\sum_{k\in\mathcal{K}_{i,t}} \sqrt{p_{k,t}} |\mathbf{h}_{k,t}^{\mathsf{H}} \mathbf{w}_{i,t}|)^2}$$

where $(a)$ uses the expression of $\tilde{\boldsymbol{\delta}}_t$ in (13), and $(b)$ is based on $(\sum_{i=1}^G x_i)^2 \leq G\sum_{i=1}^G x_i^2, \forall x_i \in \mathbb{R}$, for some $G \in \mathbb{N}^+$. Thus, we have (18). ■

## REFERENCES


[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, Apr. 2017, pp. 1273–1282.

[2] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, Jan. 2020.

[3] Y. Du, S. Yang, and K. Huang, "High-dimensional stochastic gradient quantization for communication-efficient edge learning," *IEEE Trans. Signal Process.*, vol. 68, pp. 2128–2142, Mar. 2020.

[4] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Convergence of update aware device scheduling for federated learning at the wireless edge," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3643–3658, Jun. 2021.

[5] Y. Wang, Y. Xu, Q. Shi, and T.-H. Chang, "Quantized federated learning under transmission delay and outage constraints," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 323–341, Jan. 2022.

[6] 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.

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

[8] J. Wang, B. Liang, M. Dong, G. Boudreau, and H. Abou-Zeid, "Joint online optimization of model training and analog aggregation for wireless edge learning," *IEEE/ACM Trans. Netw.*, vol. 32, no. 2, pp. 1212–1228, Apr. 2024.

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

[10] 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.

[11] M. Kim, A. L. Swindlehurst, and D. Park, "Beamforming vector design and device selection in over-the-air federated learning," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 7464–7477, Nov. 2023.

[12] F. M. Kalarde, M. Dong, B. Liang, Y. A. E. Ahmed, and H. T. Cheng, "Beamforming and device selection design in federated learning with over-the-air aggregation," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 1710–1723, Mar. 2024.

[13] C. Zhang, M. Dong, B. Liang, A. Afana, and Y. Ahmed, "Uplink over-the-air aggregation for multi-model wireless federated learning," in *Proc. IEEE SPAWC*, Sept. 2024, pp. 36–40.

[14] L. Chen, N. Zhao, Y. Chen, F. R. Yu, and G. Wei, "Over-the-air computation for IoT networks: Computing multiple functions with antenna arrays," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 5296–5306, Dec. 2018.

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

[16] N. Bhuyan, S. Moharir, and G. Joshi, "Multi-model federated learning with provable guarantees," in *Proc. VALUETOOLS*, Nov. 2022, pp. 207–222.

[17] S. Kiani, M. Dong, S. ShahbazPanahi, G. Boudreau, and M. Bavand, "Learning-based user clustering in NOMA-aided MIMO networks with spatially correlated channels," *IEEE Trans. Commun.*, vol. 70, no. 7, pp. 4807–4821, Jul. 2022.

[18] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Mach. Learn.*, vol. 42, pp. 143–175, Jan. 2001.

[19] D. Bertsekas, *Nonlinear Programming*, Belmont, MA, USA.: Athena Scientific, 2016.

[20] M. Sadeghi, E. Björnson, E. G. Larsson, C. Yuen, and T. L. Marzetta, "Max–min fair transmit precoding for multi-group multicasting in massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1358–1373, Feb. 2018.