Over-the-Air Federated Edge Learning with Hierarchical Clustering

Ozan Aygün, *Student Member, IEEE,* Mohammad Kazemi, *Member, IEEE,* Deniz Gündüz, *Fellow, IEEE,* and Tolga M. Duman, *Fellow, IEEE*

***Abstract*—We examine federated learning (FL) with over-the-air (OTA) aggregation, where mobile users (MUs) aim to reach a consensus on a global model with the help of a parameter server (PS) that aggregates the local gradients. In OTA FL, MUs train their models using local data at every training round and transmit their gradients simultaneously using the same frequency band in an uncoded fashion. Based on the received signal of the superposed gradients, the PS performs a global model update. While the OTA FL has a significantly decreased communication cost, it is susceptible to adverse channel effects and noise. Employing multiple antennas at the receiver side can reduce these effects, yet the path-loss is still a limiting factor for users located far away from the PS. To ameliorate this issue, in this paper, we propose a wireless-based hierarchical FL scheme that uses intermediate servers (ISs) to form clusters in the areas where the MUs are more densely located. Our scheme utilizes OTA cluster aggregations for the communication of the MUs with their corresponding IS, and OTA global aggregations from the ISs to the PS. We present a convergence analysis for the proposed algorithm, and show through numerical evaluations of the derived analytical expressions and experimental results that utilizing ISs results in a faster convergence and a better performance than the OTA FL alone while using less transmit power. We also validate the results on the performance using different numbers of cluster iterations with different datasets and data distributions. We conclude that the best choice of cluster aggregations depends on the data distribution among the MUs and the clusters.**

***Index Terms*—Machine learning, over-the-air communications, federated learning, wireless communications, over-the-air aggregation, hierarchical clustering**

I. Introduction

We are surrounded by devices that continuously gather all kinds of information from images, videos, and sound to various sensor measurements. The abundance of generated data has been essential for the rapid advancements in machine learning (ML) in different domains. Traditionally, ML relies on accumulating all the data on a server to train a powerful model with many parameters. However, such centralized training and data accumulation leads to concerns regarding data privacy, communication cost, and latency. Firstly, users are concerned about sharing their personal datasets as it may leak information about the owner beyond the intended use [2]. Secondly, offloading collected data samples to a remote sensor, typically for high-rate data such as images and videos, requires significant communication resources. Thirdly, applications that need to operate in real-time might be affected by the increased latency since their performance depends on the model response of the simultaneously collected data [3]. *Federated learning* (FL) offers an attractive alternative to centralized training, where the training is distributed across user devices, and does not require collecting data at a centralized server [4].

In FL, a parameter server (PS), which keeps track of the global model orchestrates training across a set of mobile users (MUs). At each iteration, the current global model parameters are shared with a subset of the MUs, selected depending on their battery states, computing capabilities, data qualities, or distance to the PS [5]. These devices are asked to perform stochastic gradient descent (SGD) on the current model using their local datasets. After completing several local training iterations, each MU sends its model update to the PS. The PS performs model aggregation using these local updates to update the global model, and sends the new model back to a potentially different subset of MUs for the next iteration. Lately, the literature on FL has focused on topics such as data heterogeneity [6], convergence analysis for non-convex data distributions [7], privacy [8], energy efficiency and latency analysis [9]. *Federated Averaging* is the simplest and the most popular model aggregation method, which employs simple averaging operation at the PS [10]. Since the performance mostly depends on the data distribution at the MUs, recent studies focus on heterogeneous datasets across MUs [6], [11]–[14]. Even though FL aims to protect local data privacy, some studies show that it is possible to infer user data even from the gradient information, and present approaches to enhance data privacy [8], [14], [15]. Numerous works have been reported on FL's power consumption and latency, and developed power-efficient FL schemes for MUs with limited power [9], [16], [17]. User heterogeneity is another direction being investigated, where a subset of devices are selected efficiently based on their available power, computing capabilities, and distance [11], [18]–[21].

While FL among edge devices has a great premise, its

Part of this work was presented at the 2022 IEEE International Conference on Communications (ICC) [1].

Ozan Aygün is with the Department of Electrical and Computer Engineering, New York University Tandon School of Engineering, Brooklyn, NY 11201 USA (e-mail: ozan@nyu.edu)

Mohammad Kazemi and Tolga M. Duman are with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara, 06800, Turkey (email: {kazemi, duman}@ee.bilkent.edu.tr)

Deniz Gündüz is with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2BT, U.K. (e-mail: d.gunduz@imperial.ac.uk)

The authors acknowledge support from TUBITAK through CHIST-ERA project SONATA (CHIST-ERA-20-SICT-004, funded by TUBITAK, Turkey Grant 221N366 and EPSRC, UK Grant EP/W035960/1).

Ozan Aygün's research was also supported by Turkcell A.S. within the framework of 5G and Beyond Joint Graduate Support Programme coordinated by Information and Communication Technologies Authority.

implementation in practical wireless scenarios requires combatting adverse channel effects and optimization under limited channel resources. Since the cost of communication in FL depends on the size of the underlying model, the periodic transmission of a large model increases the communication costs and the required bandwidth even though no raw data is being sent. In order to use the bandwidth more efficiently, over-the-air (OTA) aggregation has become a widely used method, where the local updates are sent using the same frequency band, thereby performing computation and transmission simultaneously [22]. However, accurate OTA aggregation requires mitigating the adverse effects of wireless channels so that the transmitted signals arrive at the PS at similar power levels. While this is originally achieved by applying channel inversion when accurate channel state information (CSI) is available at the MUs, it is shown in [23] that the effects of wireless fading can also be alleviated by increasing the number of receive antennas at the PS even when the transmitter side has no CSI. Recently, the focus in OTA FL has been on user scheduling [18]–[21] and analysis of different wireless channel models [8], [23]–[28]. Moreover, OTA communication in multiple-input multiple-output (MIMO) channels is also investigated in [29], where the angle of arrival is used to cluster users. Other approaches on wireless FL include model compression for the local updates and location-based user scheduling [9], [22], [28], [30].

As a possible solution for the increased communication costs, hierarchical federated learning (HFL) has been proposed. In HFL, intermediate servers (ISs) are employed in areas where the number of MUs is high to form cluster-like structures [31]. In this model, the MUs carry out multiple SGD iterations before transmitting their model differences to their corresponding IS. After several cluster aggregations between the ISs and their corresponding MUs, global aggregation is carried out at the PS, using the IS cluster updates. Studies on HFL focus on the system performance when the users have non-independent and identically distributed (non-i.i.d.) data distributions [13], power, latency, and convergence analysis [31]–[34], and optimal resource allocation schemes [17].

In this paper, we study an HFL scenario, similar to the one in [31], but with OTA aggregations both at the ISs and the PS. Our specific contributions in this framework are as follows:

- We provide the system model and channel specifications for the considered wireless hierarchical federated learning (W-HFL) system including the intra-cluster and inter-cluster interference effects. We present the details of the proposed OTA aggregation algorithm.
- We conduct a detailed convergence analysis for the proposed model, where the effects of interference and noise terms can be clearly identified. We also provide an upper bound on the convergence rate and numerically compare the convergence rate with that of the conventional FL, where all the MUs communicate with PS directly without the need for an IS. We show through numerical evaluations of the analytical results that the proposed algorithm has a higher convergence rate than conventional FL, and has a competitive performance compared to the baseline scheme with error-free links.

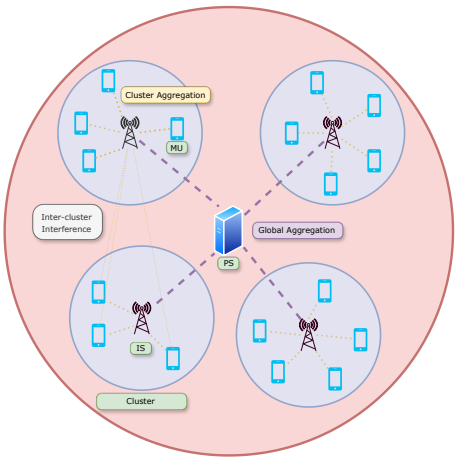


Fig. 1: Illustration of the W-HFL system model

- We demonstrate via experimental results on MNIST and CIFAR-10 datasets with different data distributions that the proposed scheme exhibits a faster convergence behavior and converges to a more reliable model compared to that of the conventional FL while also using less power at the edge.

The rest of the paper is organized as follows. In Section II, we introduce the learning objective as well as the structure of W-HFL. In Section III, we provide the communication model of the proposed algorithm. In Section IV, the convergence analysis of W-HFL is presented, and it is upper-bounded under some convexity assumptions. In Section V, we give experimental and numerical results to compare our algorithm with the conventional FL as well as the baseline approaches, and we conclude the paper in Section VI.

II. System Model

The objective of W-HFL is to minimize a loss function $F(\boldsymbol{\theta})$ with respect to the model weight vector $\boldsymbol{\theta} \in \mathbb{R}^{2N}$, where $2N$ is the model dimension. Our system consists of C clusters each containing one IS and M MUs, and a PS, as depicted in Fig. 1. The dataset of the m-th MU in the c-th cluster is denoted as $\mathcal{B}_{c,m}$, and we define $B \triangleq \sum_{c=1}^{C}\sum_{m=1}^{M}|\mathcal{B}_{c,m}|$. We have

$$F(\boldsymbol{\theta}) = \sum_{c=1}^{C}\sum_{m=1}^{M}\frac{|\mathcal{B}_{c,m}|}{B}F_{c,m}(\boldsymbol{\theta}), \tag{1}$$

where $F_{c,m}(\boldsymbol{\theta}) \triangleq \frac{1}{|\mathcal{B}_{c,m}|}\sum_{u\in\mathcal{B}_{c,m}} f(\boldsymbol{\theta},u)$, with $f(\boldsymbol{\theta},u)$ denoting the loss function corresponding to parameter vector $\boldsymbol{\theta}$ and data sample u.

We consider a hierarchical and iterative approach consisting of global, cluster, and user iterations to minimize (1). In every cluster iteration, the MUs carry out τ user iterations using their local datasets, then send their model updates to their corresponding ISs for cluster aggregation. I cluster iterations are performed at each IS before all the updated models are forwarded to the PS for global aggregation. Consider the j-th user iteration of the i-th cluster iteration of the t-th global iteration by the m-th user in the c-th cluster. The weight update is performed employing SGD as follows:

$$\boldsymbol{\theta}_{c,m}^{i,j+1}(t) = \boldsymbol{\theta}_{c,m}^{i,j}(t) - \eta_{c,m}^{i,j}(t)\nabla F_{c,m}(\boldsymbol{\theta}_{c,m}^{i,j}(t), \boldsymbol{\xi}_{c,m}^{i,j}(t)), \tag{2}$$

where $\eta_{c,m}^{i,j}(t)$ is the learning rate, $\nabla F_{c,m}(\boldsymbol{\theta}_{c,m}^{i,j}(t), \boldsymbol{\xi}_{c,m}^{i,j}(t))$ denotes the stochastic gradient estimate for the weight vector $\boldsymbol{\theta}_{c,m}^{i,j}(t)$ and a randomly sampled batch of data samples $\boldsymbol{\xi}_{c,m}^{i,j}(t)$ sampled from the dataset $\mathcal{B}_{c,m}$. Initially, $\boldsymbol{\theta}_{c,m}^{1,1}(t) = \boldsymbol{\theta}_{IS,c}^{i}(t), \forall i \in [I]$, where $[I] \triangleq \{1, 2, \ldots, I\}$, and $\boldsymbol{\theta}_{IS,c}^{1}(t) = \boldsymbol{\theta}_{PS}(t)$, where $\boldsymbol{\theta}_{PS}(t)$ is the global model at the PS at the t-th global iteration and $\boldsymbol{\theta}_{IS,c}^{i}(t)$ denotes the local model of the IS in the c-th cluster at the i-th cluster iteration. The purpose of employing ISs is to accumulate the local model differences within each cluster more frequently over smaller areas before obtaining the global model $\boldsymbol{\theta}_{PS}(t)$ for the next global iteration. Also, note that $\nabla F_{c,m}(\boldsymbol{\theta}_{c,m}^{i,j}(t), \boldsymbol{\xi}_{c,m}^{i,j}(t))$ is an unbiased estimator of $\nabla F_{c,m}(\boldsymbol{\theta}_{c,m}^{i,j}(t))$, i.e., $\mathbb{E}_{\xi}\left[\nabla F_{c,m}(\boldsymbol{\theta}_{c,m}^{i,j}(t), \boldsymbol{\xi}_{c,m}^{i,j}(t))\right] = \nabla F_{c,m}(\boldsymbol{\theta}_{c,m}^{i,j}(t))$, where the expectation is over the randomness due to SGD.

III. Wireless Hierarchical Federated Learning (W-HFL)

A. Ideal Communication

We first consider the case in which all the communications among all the units are error-free as a benchmark. In this case, after τ user iterations, each MU calculates its model difference to be sent to its corresponding IS as

$$\Delta\boldsymbol{\theta}_{c,m}^{i}(t) = \boldsymbol{\theta}_{c,m}^{i,\tau+1}(t) - \boldsymbol{\theta}_{IS,c}^{i}(t). \tag{3}$$

Then, the cluster aggregation at the c-th cluster is performed as

$$\boldsymbol{\theta}_{IS,c}^{i+1}(t) = \boldsymbol{\theta}_{IS,c}^{i}(t) + \frac{1}{M}\sum_{m=1}^{M}\Delta\boldsymbol{\theta}_{c,m}^{i}(t). \tag{4}$$

After completing I cluster iterations in each cluster, ISs send their model differences to the PS, which can be written as

$$\Delta\boldsymbol{\theta}_{PS,c}(t) = \boldsymbol{\theta}_{IS,c}^{I+1}(t) - \boldsymbol{\theta}_{PS}(t). \tag{5}$$

The global update rule is $\Delta\boldsymbol{\theta}_{PS}(t) = \frac{1}{C}\sum_{c=1}^{C}\Delta\boldsymbol{\theta}_{PS,c}(t)$. Using recursion, we can conclude that

$$\Delta\boldsymbol{\theta}_{PS}(t) = \frac{1}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\Delta\boldsymbol{\theta}_{c,m}^{i}(t). \tag{6}$$

After the global aggregation, the model at the PS is updated as $\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\boldsymbol{\theta}_{PS}(t)$.

B. OTA Communication

We now introduce the scheme referred to as OTA communications to be used for all the links from the users to the ISs, and from the ISs to the PS. Since model differences are transmitted via a common wireless medium in both cluster and global updates, estimated versions of $\Delta\boldsymbol{\theta}_{IS,c}(t)$ and $\Delta\boldsymbol{\theta}_{PS}(t)$ are received at the ISs and the PS, where the system noise and inter/intra cluster interference are present. In our setup, ISs and PS have K and K' receive antennas, respectively, while both ISs and the MUs are equipped with a single transmit antenna [1]. Also, we assume perfect channel state information (CSI) at the receiver ends.

[1]For the case of multiple transmit antennas at the ISs, the same setup is applicable, as long as each IS transmits the weighted and phase shifted versions of the same stream, i.e., employs beamforming.

1) Cluster Aggregation: In OTA communication, the local updates $\Delta\boldsymbol{\theta}_{c,m}^{i}(t) \in \mathbb{R}^{2N}$ are sent without any coding. In order to increase the spectral efficiency, the model differences are grouped to form a complex vector $\Delta\boldsymbol{\theta}_{c,m}^{i,cx}(t) \in \mathbb{C}^{N}$ with entries $\Delta\theta_{c,m}^{i,n,cx}(t)$ for $m \in [M], c \in [C], i \in [I]$, with the real and imaginary parts $\Delta\boldsymbol{\theta}_{c,m}^{i,re}(t) \triangleq \left[\Delta\theta_{c,m}^{i,1}(t), \ldots, \Delta\theta_{c,m}^{i,N}(t)\right]^T$ and $\Delta\boldsymbol{\theta}_{c,m}^{i,im}(t) \triangleq \left[\Delta\theta_{c,m}^{i,N+1}(t), \ldots, \Delta\theta_{c,m}^{i,2N}(t)\right]^T$, respectively, where $\Delta\theta_{c,m}^{i,n}(t)$ denotes the n-th entry of $\Delta\boldsymbol{\theta}_{c,m}^{i}(t)$ for $n \in [2N]$. The resulting complex vector is transmitted through the wireless medium. The received signal at the k-th antenna of the c-th IS in the i-th cluster iteration can be represented as

$$\boldsymbol{y}_{IS,c,k}^{i}(t) = P_t\sum_{c'=1}^{C}\sum_{m=1}^{M}\boldsymbol{h}_{c',m,c,k}^{i}(t)\circ\Delta\boldsymbol{\theta}_{c',m}^{i,cx}(t) + \boldsymbol{z}_{IS,c,k}^{i}(t), \tag{7}$$

where P_t is the power multiplier at the t-th global iteration, $\circ$ denotes the element-wise (Hadamard) product, $\boldsymbol{z}_{IS,c,k}^{i}(t) \in \mathbb{C}^{N}$ is the circularly symmetric additive white Gaussian noise (AWGN) vector with i.i.d. entries with zero mean and variance of σ_z^2; i.e., $z_{IS,c,k}^{i,n}(t) \sim \mathcal{CN}(0, \sigma_z^2), n \in [N]$. $\boldsymbol{h}_{c',m,c,k}^{i}(t) \in [N]$ is the channel coefficient vector between the m-th MU in the c'-th cluster and the c-th IS, whose n-th entry is modelled as $h_{c',m,c,k}^{i,n}(t) = \sqrt{\beta_{c',m,c}}\,g_{c',m,c,k}^{i,n}(t)$, where $g_{c',m,c,k}^{i,n}(t) \sim \mathcal{CN}(0, \sigma_h^2)$ is the small-scale fading coefficient (i.e., Rayleigh fading), and $\beta_{c',m,c}$ is the large-scale fading coefficient modeled as $\beta_{c',m,c} = \left(d_{c',m,c}\right)^{-p}$, where p represents the path-loss exponent and $d_{c',m,c}$ is the distance between the m-th user in the c'-th cluster and the c-th IS.

Knowing the CSI perfectly, the c-th IS combines the received signals as

$$\boldsymbol{y}_{IS,c}^{i}(t) = \frac{1}{K}\sum_{k=1}^{K}\Big(\sum_{m=1}^{M}\boldsymbol{h}_{c,m,c,k}^{i}(t)\Big)^{*}\circ\boldsymbol{y}_{IS,c,k}^{i}(t), \tag{8}$$

whose n-th entry can be written as

$$y_{IS,c}^{i,n}(t) = \frac{1}{K}\sum_{k=1}^{K}\Big(\sum_{m=1}^{M}h_{c,m,c,k}^{i,n}(t)\Big)^{*}y_{IS,c,k}^{i,n}(t), \tag{9}$$

where $y_{IS,c,k}^{i,n}(t)$ denotes the n-th entry of $\boldsymbol{y}_{IS,c,k}^{i}(t), n \in [N]$. Substituting (7) into (8), and using (9), we get

$$\begin{aligned}
y_{IS,c}^{i,n}(t) = {} & \underbrace{\frac{P_t}{K}\sum_{m,k}|h_{c,m,c,k}^{i,n}(t)|^2\Delta\theta_{c,m}^{i,n,cx}(t)}_{y_{IS,c}^{i,n,sig}(t)\text{ (signal term)}} \\
& + \underbrace{\frac{P_t}{K}\sum_{\substack{m,k,\\ m'\neq m}}h_{\substack{c,m,c,k\\ c,m',c,k}}^{i,n}(t)\Delta\theta_{c,m'}^{i,n,cx}(t)}_{y_{IS,c}^{i,n,int1}(t)\text{ (Intra-cluster interference)}} \\
& + \underbrace{\frac{P_t}{K}\sum_{\substack{m,m',k,\\ c'\neq c}}h_{\substack{c,m,c,k\\ c',m',c,k}}^{i,n}(t)\Delta\theta_{c',m'}^{i,n,cx}(t)}_{y_{IS,c}^{i,n,int2}(t)\text{ (Inter-cluster interference)}} \\
& + \underbrace{\frac{1}{K}\sum_{m,k}(h_{c,m,c,k}^{i,n}(t))^{*}z_{IS,c,k}^{i,n}(t)}_{y_{IS,c}^{i,n,no}(t)\text{ (noise term)}} \\
= {} & y_{IS,c}^{i,n,sig}(t) + y_{IS,c}^{i,n,int1}(t) + y_{IS,c}^{i,n,int2}(t) + y_{IS,c}^{i,n,no}(t),
\end{aligned} \tag{10}$$

where we define $h_{\substack{i,j,q,w\\ p,r,s,v}}^{i,n}(t) = \left(h_{i,j,q,w}^{i,n}(t)\right)^{*}h_{p,r,s,v}^{i,n}(t)$.

Aggregated model differences for $n \in [N]$, are recovered as

$$\Delta\hat{\theta}^{i,n}_{IS,c}(t) = \frac{1}{P_t M \sigma_h^2 \bar{\beta}_c} \mathrm{Re}\{y^{i,n}_{IS,c}(t)\},$$
$$\Delta\hat{\theta}^{i,n+N}_{IS,c}(t) = \frac{1}{P_t M \sigma_h^2 \bar{\beta}_c} \mathrm{Im}\{y^{i,n}_{IS,c}(t)\}, \quad (11)$$

where $\bar{\beta}_c = \sum_{m=1}^{M} \beta_{c,m,c}$, and $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ denote the real and imaginary parts, respectively. Finally, the cluster model update can be written as $\boldsymbol{\theta}^{i+1}_{IS,c}(t) = \boldsymbol{\theta}^{i}_{IS,c}(t) + \Delta\hat{\boldsymbol{\theta}}^{i}_{IS,c}(t)$, where $\Delta\hat{\boldsymbol{\theta}}^{i}_{IS,c}(t) \triangleq \left[\Delta\hat{\theta}^{i,1}_{IS,c}(t) \Delta\hat{\theta}^{i,2}_{IS,c}(t) \dots \Delta\hat{\theta}^{i,2N}_{IS,c}(t)\right]^T$.

2) Global Aggregation: Global aggregation is very similar to cluster aggregation, where each IS has a single transmit antenna and the PS has K' receive antennas. After I cluster iterations, to obtain the signal to be transmitted from the c-th IS, model differences are grouped to form a complex vector $\Delta\boldsymbol{\theta}^{cx}_{PS,c} \in \mathbb{C}^N$, with the real and imaginary parts $\Delta\boldsymbol{\theta}^{re}_{PS,c}(t) \triangleq \left[\Delta\theta^1_{PS,c}(t), \theta^2_{PS,c}(t), \dots, \Delta\theta^N_{PS,c}(t)\right]^T$ and $\Delta\boldsymbol{\theta}^{im}_{PS,c}(t) \triangleq \left[\Delta\theta^{N+1}_{PS,c}(t), \theta^{N+2}_{PS,c}(t), \dots, \Delta\theta^{2N}_{PS,c}(t)\right]^T$, respectively, where $\Delta\theta^n_{PS,c}(t)$ denotes the n-th gradient value at the c-th IS. The received signal at the k'-th antenna of the PS can be written as

$$\boldsymbol{y}_{PS,k'}(t) = P_{IS,t} \sum_{c=1}^{C} \boldsymbol{h}_{PS,c,k'}(t) \circ \Delta\boldsymbol{\theta}^{cx}_{PS,c}(t) + \boldsymbol{z}_{PS,k'}(t), \quad (12)$$

where $P_{IS,t}$ is the power multiplier of the c-th IS at the t-th global iteration, $\boldsymbol{z}_{PS,k'}(t) \in \mathbb{C}^N$ is the circularly symmetric AWGN noise with i.i.d. entries with zero mean and variance σ_z^2; i.e., $z^n_{PS,k'}(t) \sim \mathcal{CN}(0, \sigma_z^2)$. The channel coefficient between the c-th IS and the PS is modelled as $\boldsymbol{h}_{PS,c,k'}(t) = \sqrt{\beta_{IS,c}}\ \boldsymbol{g}_{PS,c,k'}(t)$, where $\boldsymbol{g}_{PS,c,k'}(t) \in \mathbb{C}^N$ is the small-scale fading coefficient vector with entries $g^n_{PS,c,k'}(t) \sim \mathcal{CN}(0, \sigma_h^2)$, $\beta_{IS,c}$ is the large-scale fading coefficient modeled as $\beta_{IS,c} = \left(d_{IS,c}\right)^{-p}$, where $d_{IS,c}$ denotes the distance between the c-th IS and the PS.

Knowing the CSI perfectly, the received signal at the PS is combined as

$$\boldsymbol{y}_{PS}(t) \triangleq \frac{1}{K'} \sum_{k'=1}^{K'} \Big(\sum_{c=1}^{C} \boldsymbol{h}_{PS,c,k'}(t)\Big)^* \circ \boldsymbol{y}_{PS,k'}(t). \quad (13)$$

Estimated global model differences at the PS can be recovered as

$$\Delta\hat{\theta}^n_{PS}(t) = \frac{1}{P_{IS,t} C \sigma_h^2 \bar{\beta}} \mathrm{Re}\{y^n_{PS}(t)\}, \quad (14a)$$
$$\Delta\hat{\theta}^{n+N}_{PS}(t) = \frac{1}{P_{IS,t} C \sigma_h^2 \bar{\beta}} \mathrm{Im}\{y^n_{PS}(t)\}, \quad (14b)$$

where $\bar{\beta} = \sum_{c=1}^{C} \beta_{IS,c}$. Finally, the global aggregation is performed using $\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\hat{\boldsymbol{\theta}}_{PS}(t)$, where $\Delta\hat{\boldsymbol{\theta}}_{PS}(t) = \left[\Delta\hat{\theta}^1_{PS}(t) \Delta\hat{\theta}^2_{PS}(t) \dots \Delta\hat{\theta}^{2N}_{PS}(t)\right]^T$.

The n-th symbol can be written as

$$\begin{aligned} y^n_{PS}(t) =& \frac{1}{K'} \sum_{k'=1}^{K'} \Big(\sum_{c=1}^{C} h_{PS,c,k'}(t)\Big)^* y^n_{PS,k'}(t) \\ =& \underbrace{P_{IS,t} \sum_{c=1}^{C} \Big(\frac{1}{K'} \sum_{k'=1}^{K'} |h^n_{PS,c,k'}(t)|^2\Big) \Delta\theta^{n,cx}_{PS,c}(t)}_{\text{Signal Term}} \\ &+ \underbrace{\frac{P_{IS,t}}{K'} \sum_{c=1}^{C} \sum_{\substack{c'=1 \\ c' \neq c}}^{C} \sum_{k'=1}^{K'} h^{PS,n}_{\substack{c,k' \\ c',k'}}(t) \Delta\theta^{n,cx}_{PS,c'}(t)}_{\text{Interference Term}} \\ &+ \underbrace{\frac{1}{K'} \sum_{c=1}^{C} \sum_{k'=1}^{K'} \big(h^n_{PS,c,k'}(t)\big)^* z^n_{PS,k'}(t)}_{\text{Noise Term}} \\ =& y^{n,sig}_{PS}(t) + y^{n,int}_{PS}(t) + y^{n,noise}_{PS}(t), \end{aligned} \quad (15)$$

where $h^{PS,n}_{\substack{i,j \\ p,r}}(t) = \big(h^n_{PS,i,j}(t)\big)^* h^n_{PS,p,r}(t)$.

Since we can write $\Delta\theta^{n,cx}_{PS,c}(t) = \Delta\theta^n_{PS,c}(t) + j\Delta\theta^{n+N}_{PS,c}(t)$, using (5) and recursively adding previous cluster iterations, we obtain

$$\begin{aligned} \Delta\theta^{n,cx}_{PS,c}(t) =& \big(\Delta\theta^{I+1,n}_{IS,c}(t) - \Delta\theta^{1,n}_{IS,c}(t)\big) \\ &+ j\big(\Delta\theta^{I+1,n+N}_{IS,c}(t) - \Delta\theta^{1,n+N}_{IS,c}(t)\big) \\ =& \sum_{i=1}^{I} \Delta\hat{\theta}^{i,n}_{IS,c}(t) + j\Delta\hat{\theta}^{i,n+N}_{IS,c}(t) \\ =& \frac{1}{P_t M \sigma_h^2 \bar{\beta}_c} \sum_{i=1}^{I} y^{i,n}_{IS,c}(t). \end{aligned} \quad (16)$$

Substituting (16) into (15), we have

$$\begin{aligned} y^n_{PS}(t) =& \frac{P_{IS,t}}{P_t K' M \sigma_h^2 \bar{\beta}_c} \sum_{c,k',i} |h^n_{PS,c,k'}(t)|^2 y^{i,n}_{IS,c}(t) \\ &+ \frac{1}{K'} \sum_{c,k'} \big(h^n_{PS,c,k'}(t)\big)^* z^n_{PS,k'}(t) \\ &+ \frac{P_{IS,t}}{P_t K' M \sigma_h^2 \bar{\beta}_{c'}} \sum_{\substack{c,k',i, \\ c' \neq c}}^{C'} h^{PS,n}_{\substack{c,k' \\ c',k'}}(t) y^{i,n}_{IS,c'}(t). \end{aligned} \quad (17)$$

Substituting (11) into (17), we can write y^n_{PS} as $y^n_{PS}(t) = \sum_{l=1}^{9} y^n_{PS,l}$, with $\lambda_{t,c} = \frac{P_{IS,t}}{K K' M \sigma_h^2}$, each term can be written as

$$y^n_{PS,1}(t) = \sum_{\substack{c,m,i, \\ k,k'}} \frac{\lambda_{t,c}}{\bar{\beta}_c} |h^n_{PS,c,k'}(t)|^2 |h^{i,n}_{c,m,c,k}(t)|^2 \Delta\theta^{i,n,cx}_{c,m}(t),$$

$$y^n_{PS,2}(t) = \sum_{\substack{c,m, \\ m' \neq m, \\ i,k,k'}} \frac{\lambda_{t,c}}{\bar{\beta}_c} |h^n_{PS,c,k'}(t)|^2 h^{i,n}_{\substack{c,m,c,k \\ c,m',c,k}}(t) \Delta\theta^{i,n,cx}_{c,m'}(t),$$

$$y^n_{PS,3}(t) = \sum_{\substack{c,c' \neq c,m, \\ m',i,k,k'}} \frac{\lambda_{t,c}}{\bar{\beta}_c} |h^n_{PS,c,k'}(t)|^2 h^{i,n}_{\substack{c,m,c,k \\ c,m',c',k}}(t) \Delta\theta^{i,n,cx}_{c',m'}(t),$$

$$y^n_{PS,4}(t) = \sum_{\substack{c,m, \\ i,k, \\ k'}} \frac{\lambda_{t,c}}{P_{IS,t} \bar{\beta}_c} |h^n_{PS,c,k'}(t)|^2 \big(h^{i,n}_{c,m,c,k}(t)\big)^* z^{i,n}_{IS,c,k}(t),$$

$$y^n_{PS,5}(t) = \sum_{\substack{c,c' \neq c, \\ m,i,k,k'}} \frac{\lambda_{t,c}}{\bar{\beta}_{c'}} h^{PS,n}_{\substack{c,k' \\ c',k'}}(t) |h^{i,n}_{c',m,c',k}(t)|^2 \Delta\theta^{i,n,cx}_{c',m}(t),$$

$$y^n_{PS,6}(t) = \sum_{\substack{c,c' \neq c,m, \\ m' \neq m,i,k,k'}} \frac{\lambda_{t,c}}{\bar{\beta}_{c'}} h^{PS,n}_{\substack{c,k' \\ c',k'}}(t) h^{i,n}_{\substack{c',m,c',k \\ c',m',c',k}}(t) \Delta\theta^{i,n,cx}_{c',m'}(t),$$

$$y^n_{PS,7}(t) = \sum_{\substack{c,c' \neq c, \\ c'' \neq c',m, \\ m',i,k,k'}} \frac{\lambda_{t,c}}{\bar{\beta}_{c'}} h^{PS,n}_{\substack{c,k' \\ c',k'}}(t) h^{i,n}_{\substack{c',m,c',k \\ c',m',c'',k}}(t) \Delta\theta^{i,n,cx}_{c'',m'}(t),$$

$$y^n_{PS,8}(t) = \sum_{\substack{c,c' \neq c, \\ m,i, \\ k,k'}} \frac{\lambda_{t,c}}{P_{IS,t} \bar{\beta}_{c'}} h^{PS,n}_{\substack{c,k' \\ c',k'}}(t) \big(h^{i,n}_{c',m,c',k}(t)\big)^* z^{i,n}_{IS,c',k}(t),$$

$$y^n_{PS,9}(t) = \sum_{c,k'} \frac{1}{K'} \big(h^n_{PS,c,k'}(t)\big)^* z^n_{PS,k}(t). \quad (18)$$

IV. Convergence Analysis

In this section, we present an upper bound on the global loss function at the PS, which represents the difference between the expected loss at the PS and the optimal loss after a certain number of iterations, which helps us to understand and quantify the relationship between the system parameters and performance. We define the solution that minimizes the loss $F(\boldsymbol{\theta})$ as

$$\boldsymbol{\theta}^* \triangleq \arg\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}). \tag{19}$$

Also, the minimum value of the loss function is denoted as $F^* = F(\boldsymbol{\theta}^*)$, the minimum value of the local loss function $F_{c,m}$ is given as $F^*_{c,m}$, and the bias in the dataset is defined as

$$\Gamma \triangleq F^* - \sum_{c=1}^{C}\sum_{m=1}^{M}\frac{B_{c,m}}{B}F^*_{c,m} \geq 0. \tag{20}$$

In addition, we assume that the learning rate of the overall system does not change in user and cluster iterations, i.e., $\eta^{i,j}_{c,m}(t) = \eta(t)$. Therefore, we can write the global update rule as $\boldsymbol{\theta}^{i,j+1}_{c,m}(t) = \boldsymbol{\theta}^{i,j}_{c,m}(t) - \eta(t)\nabla F_{c,m}(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t))$, which can also be written as

$$\boldsymbol{\theta}^{i,j+1}_{c,m}(t) - \boldsymbol{\theta}^{i,1}_{c,m}(t) = -\eta(t)\sum_{l=1}^{j}\nabla F_{c,m}(\boldsymbol{\theta}^{i,l}_{c,m}, \boldsymbol{\xi}^{i,l}_{c,m}(t)). \tag{21}$$

We make the following two assumptions as in [23].

Assumption 1. *All the loss functions $F_{1,1}, \ldots, F_{C,M}$ for all the clusters and users are L-smooth and μ-strongly convex, which are, respectively $\forall \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^{2N}$, $\forall m \in [M], \forall c \in [C]$,*

$$F_{c,m}(\boldsymbol{v}) - F_{c,m}(\boldsymbol{w}) \leq \langle \boldsymbol{v} - \boldsymbol{w}, \nabla F_{c,m}(\boldsymbol{w})\rangle + \frac{L}{2}\|\boldsymbol{v} - \boldsymbol{w}\|_2^2, \tag{22}$$

$$F_{c,m}(\boldsymbol{v}) - F_{c,m}(\boldsymbol{w}) \geq \langle \boldsymbol{v} - \boldsymbol{w}, \nabla F_{c,m}(\boldsymbol{w})\rangle + \frac{\mu}{2}\|\boldsymbol{v} - \boldsymbol{w}\|_2^2. \tag{23}$$

Assumption 2. *The expected value of the squared l_2 norm of the stochastic gradients are bounded, which is, $\forall j \in [\tau], i \in [I], \forall m \in [M], \forall c \in [C]$, $\mathbb{E}_\xi\left[\left\|\nabla F_{c,m}(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t))\right\|_2^2\right] \leq G^2$, which in turn translates into $\mathbb{E}_\xi\left[\nabla F_{c,m}(\theta^{i,j,n}_{c,m}(t), \xi^{i,j,n}_{c,m}(t))\right] \leq G$, $\forall n \in [2N]$.*

Theorem 1. *In W-HFL, for $0 \leq \eta(t) \leq min\left\{1, \frac{1}{\mu\tau I}\right\}$, the global loss function can be upper bounded as*

$$\mathbb{E}\left[\|\boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^*\|_2^2\right] \leq \left(\prod_{a=0}^{t-1}X(a)\right)\|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 + \sum_{b=0}^{t-1}Y(b)\prod_{a=b+1}^{t-1}X(a), \tag{24}$$

where $X(t) = (1 - \mu\eta(t)I\left(\tau - \eta(t)(\tau - 1)\right))$, and

$$\begin{aligned}
Y(t) = {} & \frac{\eta^2(t)G^2I^2\tau^2}{M^2C^2}\sum_{c_1,c_2,m_1,m_2}A(m_1,m_2,c_1,c_2) \\
& + \frac{\big(2 + (M-1)(C-2)(K-1)(I-1)\big)\eta^2(t)IG^2\tau^2}{K(K')M^3C^2(C-1)\bar{\beta}^2} \\
& \times \sum_{\substack{m_1,m_2,\\ c,c'\neq c}}\frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m_1,c'}\beta_{c',m_2,c'}}{\bar{\beta}^2_{c'}} \\
& + \frac{\eta^2(t)G^2I\tau^2}{KK'M^2C^2\bar{\beta}^2}\sum_{c,m}\Bigg(\frac{(K'+1)\beta^2_{IS,c}\beta_{c,m,c}}{\bar{\beta}^2_c} \\
& \times \Big(\sum_{m'\neq m}\beta_{c,m',c} + \sum_{c'\neq c,m'}\beta_{c,m',c'}\Big)\Bigg) \\
& + \frac{\sigma_z^2N}{K'C^2\sigma_h^2\bar{\beta}^2}\sum_c\beta_{IS,c}\Bigg(\frac{1}{P^2_{IS,t}} + \frac{I}{KM^2} \\
& \times \sum_m\Big(\frac{(K'+1)\beta_{IS,c}\beta_{c,m,c}}{P_t^2\bar{\beta}^2_c} + \sum_{c'\neq c}\frac{\beta_{IS,c'}\beta_{c',m,c'}}{P^2_{IS,t}\bar{\beta}^2_{c'}}\Big)\Bigg) \\
& + (1 + \mu(1-\eta(t))\,\eta^2(t)IG^2\frac{\tau(\tau-1)(2\tau-1)}{6} \\
& + \eta^2(t)I(\tau^2 + \tau - 1)G^2 + 2\eta(t)I(\tau-1)\Gamma,
\end{aligned} \tag{25}$$

with $A(m_1,m_2,c_1,c_2) = 1 - \frac{\beta_{c_1,m_1,c_1}\beta_{IS,c_1}}{\bar{\beta}\bar{\beta}_{c_1}} - \frac{\beta_{c_2,m_2,c_2}\beta_{IS,c_2}}{\bar{\beta}\bar{\beta}_{c_2}} + \frac{\beta_{c_1,m_1,c_1}\beta_{c_2,m_2,c_2}\beta_{IS,c_1}\beta_{IS,c_2}}{MCKK'I\bar{\beta}^2\bar{\beta}_{c_1}\bar{\beta}_{c_2}}$
$\times\big(4 + 2(K'-1) + (M-1)(K-1)(I-1)\big(2 + (K'-1)(C-1)\big)\big)$.

Proof: See Appendix A. ∎

Corollary 1. *Assuming L-smoothness, after T global iterations, the loss function can be upper-bounded as*

$$\begin{aligned}
\mathbb{E}\left[F(\boldsymbol{\theta}_{PS}(T))\right] - F^* \leq & \frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}_{PS}(T) - \boldsymbol{\theta}^*\|_2^2\right], \\
\leq & \frac{L}{2}\left(\prod_{a=0}^{T-1}X(a)\right)\|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 \\
& + \frac{L}{2}\sum_{b=0}^{T-1}Y(b)\prod_{a=b+1}^{T-1}X(a).
\end{aligned} \tag{26}$$

Remark 1. Since the fourth term in $Y(a)$ is independent of $\eta(a)$, even for $\lim_{t\to\infty}\eta(t) = 0$, we have $\lim_{t\to\infty}\mathbb{E}[F(\boldsymbol{\theta}_{PS}(t))] - F^* \neq 0$. $Y(a)$ is also proportional to I and τ, meaning that more user iterations and cluster aggregations do not always provide faster convergence. However, since the MUs experience lower path-loss in W-HFL than in the conventional FL, W-HFL can reach a higher accuracy. Moreover, increasing the number of clusters C leads to a faster convergence, at the cost of employing more ISs.

Corollary 2. *For a simplified setting with $I = \tau = 1, P_{IS,t} \gg P_t, P_t = P, \forall t, \beta_{c,m,c} = \beta, \beta_{IS,c} = \beta_{IS}, \forall m \in [M], \forall c \in [C]$, we have $X(t) = \big(1 - \mu\eta(t)\big)$ and*

$$\begin{aligned}
Y(t) & \approx \frac{\eta^2(t)G^2}{KK'M^3C^3}MCK' + 2\eta^2G^2\Big(1 - \frac{1}{MC}\Big) \\
& \quad + \frac{\sigma_z^2N}{KC^3\sigma_h^2}\Big(\frac{1}{P_{IS,t}} + \frac{(K'+1)}{KM^3P^2} + \frac{(C-1)}{P^2_{IS,t}M^2}\Big) \\
& \approx \frac{\eta^2(t)G^2}{KM^2C^2} + 2\eta^2G^2 + \frac{\sigma_z^2N}{KM^3C^3\sigma_h^2P^2} \\
& \approx 2\eta^2(t)G^2 + \frac{\sigma_z^2N}{KM^3C^3\sigma_h^2P^2},
\end{aligned} \tag{27}$$

which, when $\eta(t) = \eta, \forall t$, simplify the upper bound on the loss function as

$$\begin{aligned}
\mathbb{E}\left[F(\boldsymbol{\theta}_{PS}(T))\right] - F^* \leq & \frac{L}{2}\big(1 - \mu\eta\big)^T\|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 \\
& + \frac{L}{2\mu\eta}\Big(2\eta^2G^2 + \frac{\sigma_z^2N}{KM^3C^3\sigma_h^2P^2}\Big)\Big(1 - \big(1 - \mu\eta\big)^T\Big).
\end{aligned} \tag{28}$$

Remark 2. As expected, it is observed that the numbers of receive antennas, MUs and ISs have positive effects on the convergence, while the model dimension has an adverse effect.

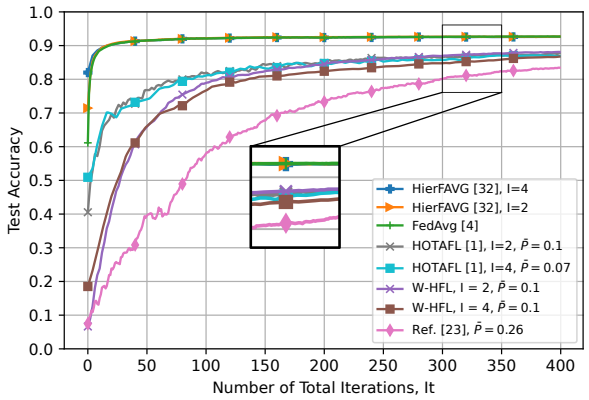


Fig. 2: I.i.d. with $\tau = 1$.

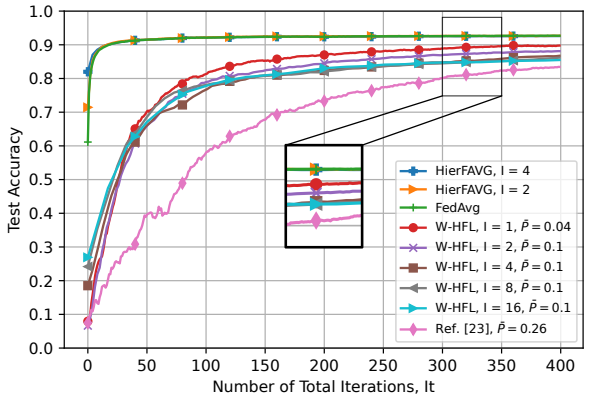


Fig. 3: I.i.d. with $\tau = 1$ for different cluster aggregation values.

V. Simulation Results

In this section, we evaluate and compare the performance of W-HFL with that of the conventional FL under different scenarios. Via numerical experiments, we observe the power consumption, as well as the convergence speed of the learning algorithm with different numbers of cluster aggregations, I. In our experiments, we use two different image classification datasets, MNIST [35] and CIFAR-10 [36]. For the MNIST dataset, we train a single-layer neural network with 784 input neurons and 10 output neurons with $2N = 7850$; and, for CIFAR-10, we employ a convolutional neural network (CNN) architecture which has two $3 \times 3 \times 32$, two $3 \times 3 \times 64$, and two $3 \times 3 \times 128$ convolutional layers, each of them with the same padding, batch normalization, and ReLU function. After every two convolutional layers, it has 2×2 max pooling and dropout with $0, 2, 0.3,$ and 0.4, respectively. In the end, we have a fully-connected layer with 10 output neurons and softmax activation, which corresponds to $2N = 307498$. We employ Adam optimizer [37] for training both networks.

We consider a hierarchical structure with $D = 20$ MUs, $C = 4$ circular clusters each with a single IS in the middle and $M = 5$ MUs, and a single PS. MUs are randomly distributed at a normalized distance between 0.5 and 1 units from their corresponding IS. Also, these clusters are randomly placed at a normalized distance between 0.5 and 3 units from the PS.

The experiments are performed with two different data distributions. In the i.i.d. experiments, all the training data is randomly and equally distributed across MUs. In the non-i.i.d. case, we split the training data into $3MC$ groups each consisting of samples with the same label, and randomly assign 3 groups to each MU randomly. As a second non-i.i.d. case, we distribute the labels to different clusters in such a way that each cluster pair has 6 shared labels, and assigned labels are distributed randomly across MUs in each cluster. In order to make the comparison fair, we use a normalized time IT in the accuracy plots where T denotes the number of global iterations. In the experiments, the total time IT is set to 400, the mini-batch size is $|\boldsymbol{\xi}^i_{c,m}(t)| = 500$ for MNIST training and $|\boldsymbol{\xi}^i_{c,m}(t)| = 128$ for CIFAR-10 training, the path loss exponent p is set to 4, $\sigma_h^2 = 1$, $\sigma_z^2 = 10$ for the MNIST, and $\sigma_z^2 = 1$ for the CIFAR-10 training. Each IS and the PS has $5MC = 100$ receive antennas. Also, the power multipliers are set to $P_t = 1+10^{-2}t$, and $P_{IS,t} = 20P_t$, $t \in [T]$. In order to make the average transmit power levels consistent among different simulations, $P_{t,low} = 0.5P_t$ is used for the cases with $I = 1$.

In Figs. 2-5, we present the performance of W-HFL with different numbers of cluster aggregations I using the MNIST dataset. We also report the average transmit power per total number of iterations at the edge for each case. We consider W-HFL with $I \in [1, 2, 4, 8]$, as well as the conventional FL scheme with no IS in between. To assess their performance, we compare the results with wireless and error-free conventional FL schemes with no IS in between [4], [23] as well as recent HFL schemes, including our previous work with error-free links between ISs and PS [1], and HierFAVG scheme with error-free links both in between MUs and ISs, and ISs and PS [32]. We can observe in Fig. 3 that W-HFL outperforms [23] while using less power at the edge. This is mainly because in W-HFL, MUs have a closer server (IS) to transmit their signals to, thereby being less affected by the path-loss effects. Also, it can be seen that the performance slightly deteriorates as I increases, while consuming less transmit power at the edge, where we simulated W-HFL including $I = 16$ to show the trend. The system performs better in i.i.d. distribution when the ISs perform fewer cluster aggregations, and the best performance is observed with $I = 1$, where the ISs just relay the received cluster updates. Moreover, we can observe that HOTAFL achieves faster convergence than W-HFL due to the reduced number of wireless layers.

In Fig. 4, we consider MNIST with non-i.i.d. distribution across MUs, and evaluate the system performances for $\tau = 3$. We can see the change in the order of performance when the distribution changes and τ increases since having more cluster updates before the global aggregation provides a more powerful update for the model than having a frequent global model update with less trained non-i.i.d. datasets. Moreover, we evaluate the performance when clusters are non-i.i.d. in Fig. 5, where we observe a slight decrease in accuracies compared to the i.i.d. data distribution.

In Fig. 6, we also depict the performance of the proposed algorithm on the CIFAR-10 dataset with i.i.d. data distribution across MUs. We can see a similar trend with the i.i.d. MNIST results. However, the average transmit power values have increased when compared to MNIST results since the used

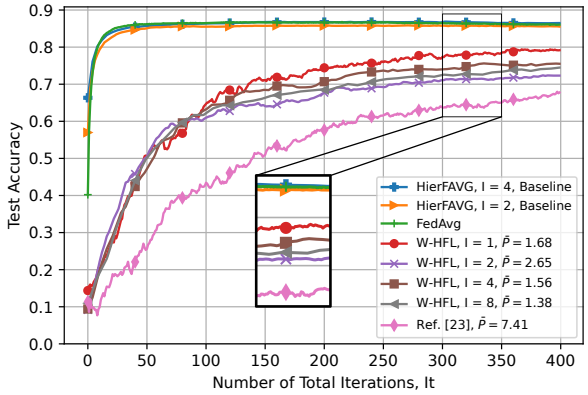


Fig. 4: Non-i.i.d. clusters with $\tau = 3$.

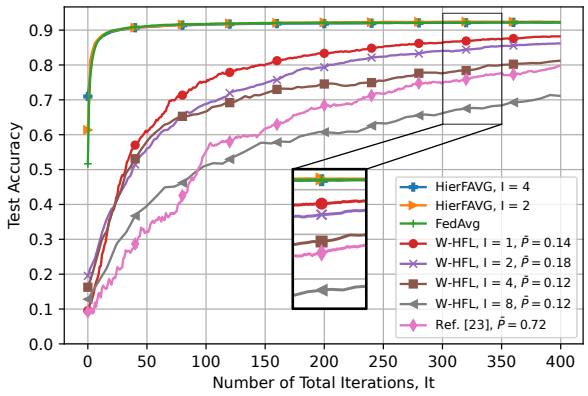


Fig. 5: Non-i.i.d. clusters with $\tau = 1$.

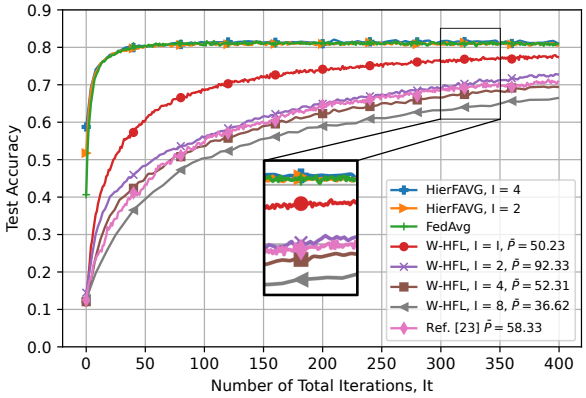


Fig. 6: CIFAR-10 training with $\tau = 5$.

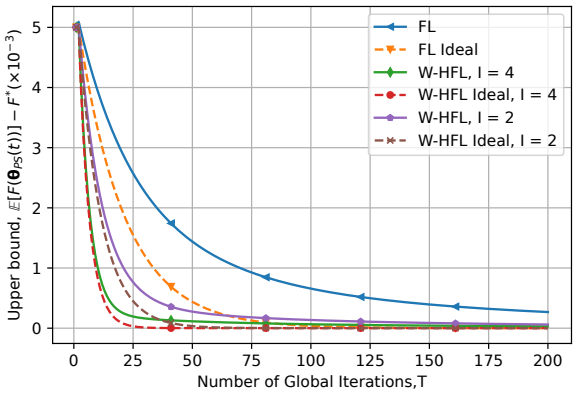


Fig. 7: Convergence rate for Fig. 3.

model contains more parameters in CIFAR-10 simulations to tackle the more challenging dataset. It can be observed that using the ISs as relays gives the best performance while using less transmit power. W-HFL with $I = 2$ uses more transmit power than $I = 4$ since it performs more global iteration rounds with an increased P_t. We can also see that W-HFL starts to perform worse than [23] when $I = 8$ since the model updates become more dependent on cluster updates, thereby making less use of overall existing data in the system.

In Fig. 7, we numerically analyze the convergence rate of W-HFL, with the results presented in Corollary 1. The setting from MNIST i.i.d. training is used with $2N = 7850, L = 10, \mu = 1, G^2 = 1, \Gamma = 1, \eta(t) = 5 \cdot 10^{-2} - 2 \cdot 10^{-5}t, P_t = 1 + 10^{-2}t, P_{IS,t} = 10P_t, \|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 = 10^3$. We can observe that W-HFL converges faster than the conventional FL, and performs similarly to the baseline.

VI. Conclusions

We have proposed a W-HFL scheme, where edge devices exploit nearby local servers called ISs for model aggregation. After several OTA cluster aggregations, ISs transmit their model differences to the PS to update the global model for the next iteration. We have considered the effects of inter-cluster interference during local cluster aggregations, and adopted OTA aggregation also at the PS. We provided a detailed system model as well as a convergence analysis for the proposed algorithm that gives an upper bound on the global loss function. We showed through numerical evaluations and experimental results with different data distributions and datasets that bringing the server-side closer to the more densely located MUs can improve the final model accuracy and result in faster convergence compared to the conventional FL. We also observed that using fewer cluster aggregations in W-HFL can lead to higher accuracies, but with an increased average transmit power at the edge. In a nutshell, W-HFL provides a bandwidth-efficient realistic framework using OTA aggregation. Even though poor channel states and large scale use cases are potential limitations to W-HFL, we left their analyses as future work.

Appendix A: Proof of Theorem 1

Let us define the auxiliary variable $\boldsymbol{v}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\boldsymbol{\theta}_{PS}(t)$. Then, we have

$$\begin{aligned}\left\|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{\theta}^*\right\|_2^2 =& \left\|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1) + \boldsymbol{v}(t+1) - \boldsymbol{\theta}^*\right\|_2^2 \\ =& \left\|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1)\right\|_2^2 + \left\|\boldsymbol{v}(t+1) - \boldsymbol{\theta}^*\right\|_2^2 \\ &+ 2\langle \boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1), \boldsymbol{v}(t+1) - \boldsymbol{\theta}^*\rangle. \end{aligned} \tag{29}$$

Next, we provide upper bounds on the three terms of (29).

Lemma 1. $\mathbb{E}\left[\left\|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1)\right\|_2^2\right]$

$$\leq \frac{\eta^2(t) G^2 I^2 \tau^2}{M^2 C^2} \sum_{\substack{c_1, c_2, \\ m_1, m_2}} A(m_1, m_2, c_1, c_2)$$

$$+\frac{\big(2+(M-1)(C-2)(K-1)(I-1)\big)\eta^2(t)IG^2\tau^2}{K(K')M^3C^2(C-1)\bar{\beta}^2}\times\sum_{\substack{c,m_1,m_2\\c'\neq c}}\frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m_1,c'}\beta_{c',m_2,c'}}{\bar{\beta}^2_{c'}}+\frac{\eta^2(t)G^2I\tau^2}{KK'M^2C^2\bar{\beta}^2}\sum_{c,m}\Bigg(\frac{(K'+1)\beta^2_{IS,c}\beta_{c,m,c}}{\bar{\beta}^2_c}\times\Big(\sum_{m'\neq m}\beta_{c,m',c}+\sum_{c'\neq c,m'}\beta_{c,m',c'}\Big)\Bigg)+\frac{\eta^2(t)G^2I\tau^2}{KK'M^2C^2\bar{\beta}^2}\sum_{\substack{c,m\\c'\neq c}}\Bigg(\frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}}{\bar{\beta}^2_{c'}}\times\Big(\sum_{m'\neq m}\beta_{c',m',c'}+\sum_{c''\neq c'}\beta_{c',m',c''}\Big)\Bigg)+\frac{\sigma_z^2N}{K'C^2\sigma_h^2\bar{\beta}^2}\sum_c\beta_{IS,c}\Bigg(\frac{1}{P^2_{IS,t}}+\frac{I}{KM^2}\times\sum_m\Big(\frac{(K'+1)\beta_{IS,c}\beta_{c,m,c}}{P_t^2\bar{\beta}^2_c}+\sum_{c'\neq c}\frac{\beta_{IS,c'}\beta_{c',m,c'}}{P^2_{IS,t}\bar{\beta}^2_{c'}}\Big)\Bigg),\quad(30)$$

where $A(m_1,m_2,c_1,c_2)$ is given in Theorem 1.

Proof: See Appendix B. ∎

Lemma 2. $\mathbb{E}\Big[\big\|\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\big\|_2^2\Big]$
$\leq(1-\mu\eta(t)I\left(\tau-\eta(t)(\tau-1)\right))\,\mathbb{E}\Big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\Big]+2\eta(t)I(\tau-1)\Gamma$
$+(1+\mu(1-\eta(t))\,\eta^2(t)IG^2\frac{\tau(\tau-1)(2\tau-1)}{6}+\eta^2(t)I(\tau^2+\tau-1)G^2.$

Proof: See Appendix C. ∎

Lemma 3. $\mathbb{E}\left[\langle\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1),\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\rangle\right]=0.$

Proof: $\mathbb{E}\left[\langle\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1),\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\rangle\right]=\mathbb{E}\Big[\langle\Delta\hat{\boldsymbol{\theta}}_{PS}(t)-\Delta\boldsymbol{\theta}_{PS}(t),\boldsymbol{\theta}_{PS}(t)+\Delta\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\rangle\Big]$. Then, knowing that channel realizations are independent at different user and cluster updates of the same global iteration t, we have $\mathbb{E}\Big[\langle\Delta\hat{\boldsymbol{\theta}}_{PS}(t)-\Delta\boldsymbol{\theta}_{PS}(t),\boldsymbol{\theta}_{PS}(t)+\Delta\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\rangle\Big]=0.$ ∎

Recursively iterating through the results of Lemmas 1, 2, and 3 concludes the theorem.

Appendix B: Proof of Lemma 1

Using (29), we have

$$\mathbb{E}\big[||\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)||_2^2\big]=\mathbb{E}\left[||\Delta\hat{\boldsymbol{\theta}}_{PS}(t)-\Delta\boldsymbol{\theta}_{PS}(t)||_2^2\right],=\sum_{n=1}^{2N}\mathbb{E}\Big[(\Delta\hat{\theta}^n_{PS}(t)-\Delta\theta^n_{PS}(t))^2\Big].\quad(31)$$

Note that $\Delta\hat{\theta}^n_{PS}(t)=\sum_{l=1}^{9}\Delta\hat{\theta}^n_{PS,l}(t)$. Using the independence of different channel realizations over different users, clusters, and noise, we can write

$$\mathbb{E}\left[||\Delta\hat{\theta}^n_{PS}(t)-\Delta\theta^n_{PS}(t)||_2^2\right]=\sum_{l=2}^{9}\mathbb{E}\left[\Big(\Delta\hat{\theta}^n_{PS,l}(t)\Big)^2\right]+\mathbb{E}\left[\Big(\Delta\hat{\theta}^n_{PS,1}(t)-\Delta\theta^n_{PS}(t)\Big)^2\right].\quad(32)$$

Lemma 4. $\mathbb{E}\Big[\Delta\theta^{i_1,n}_{c_1,m_1}(t)\Delta\theta^{i_2,n}_{c_2,m_2}(t)\Big]\leq\eta^2(t)G^2\tau^2.$

Proof: $\mathbb{E}\Big[\Delta\theta^{i_1,n}_{c_1,m_1}(t)\Delta\theta^{i_2,n}_{c_2,m_2}(t)\Big]$
$=\eta^2(t)\sum_{j_1=1}^{\tau}\sum_{j_2=1}^{\tau}\mathbb{E}\Big[\nabla F_{c_1,m_1}(\theta^{i_1,j_1,n}_{c_1,m_1}(t),\boldsymbol{\xi}^{i_1,j_1,n}_{c_1,m_1}(t))\times\nabla F_{c_2,m_2}(\theta^{i_2,j_2,n}_{c_2,m_2}(t),\boldsymbol{\xi}^{i_2,j_2,n}_{c_2,m_2}(t))\Big]$
$\overset{(a)}{\leq}\eta^2(t)G^2\tau^2,$

where (a) holds due to Assumption 2. ∎

Lemma 5. $\mathbb{E}\Big[\big\|\Delta\boldsymbol{\theta}^i_{c,m}(t)\big\|_2^2\Big]\leq\eta^2(t)G^2\tau^2.$

Proof:

$$\mathbb{E}\Big[\big\|\Delta\boldsymbol{\theta}^i_{c,m}(t)\big\|_2^2\Big]=\eta^2(t)\mathbb{E}\Big[\Big\|\sum_{j=1}^{\tau}\nabla F_{c,m}(\boldsymbol{\theta}^{i,j}_{c,m}(t),\boldsymbol{\xi}^{i,j}_{c,m}(t))\Big\|_2^2\Big]\overset{(a)}{\leq}\eta^2(t)\tau\sum_{j=1}^{\tau}\mathbb{E}\Big[\Big\|\nabla F_{c,m}(\boldsymbol{\theta}^{i,j}_{c,m}(t),\boldsymbol{\xi}^{i,j}_{c,m}(t))\Big\|_2^2\Big]\overset{(b)}{\leq}\eta^2(t)G^2\tau^2,$$

where (a) is obtained using the convexity of $\|\|_2^2$ and (b) holds because of Assumption 2. ∎

Lemma 6. $\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,1}(t)-\Delta\theta^n_{PS}(t)\big)^2\Big]$
$\leq\frac{\eta^2(t)G^2I^2\tau^2}{M^2C^2}\sum_{c_1=1}^{C}\sum_{c_2=1}^{C}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}A(m_1,m_2,c_1,c_2).$

Proof: Using (6) and (18), we have

$$\begin{aligned}
&\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,1}(t)-\Delta\theta^n_{PS}(t)\big)^2\Big]\\
&=\mathbb{E}\Bigg[\Big(\frac{1}{MC}\sum_{c,m,i}\Delta\theta^{i,n}_{c,m}(t)\Big(\Big(\frac{1}{KK'\sigma_h^4\bar{\beta}\bar{\beta}_c}\times\sum_{k,k'}|h^n_{IS,c,k'}(t)|^2|h^{i,n}_{c,m,c,k}(t)|^2\Big)-1\Big)\Big)^2\Bigg],\\
&=\frac{1}{M^2C^2}\mathbb{E}\Bigg[\sum_{\substack{c_1,c_2,m_1,\\m_2,i_1,i_2}}\Delta\theta^{i_1,n}_{c_1,m_1}(t)\Delta\theta^{i_2,n}_{c_2,m_2}(t)\times\Big(1-\frac{2}{KK'\sigma_h^4\bar{\beta}\bar{\beta}_{c_1}}\sum_{k_1,k_1'}|h^n_{PS,c_1,k_1'}(t)|^2|h^{i_1,n}_{c_1,m_1,c_1,k_1}(t)|^2\\
&\quad+\frac{1}{K^2(K')^2\sigma_h^8\bar{\beta}^2\bar{\beta}_{c_1}\bar{\beta}_{c_2}}\sum_{\substack{k_1,k_1',\\k_2,k_2'}}|h^n_{PS,c_1,k_1'}(t)|^2|h^n_{IS,c_2,k_2'}(t)|^2\times|h^{i_1,n}_{c_1,m_1,c_1,k_1}(t)|^2|h^{i_2,n}_{c_2,m_2,c_2,k_2}(t)|^2\Big)\Bigg],\\
&=\frac{1}{M^2C^2}\mathbb{E}\Bigg[\sum_{\substack{c_1,c_2,m_1,\\m_2,i_1,i_2}}\Delta\theta^{i_1,n}_{c_1,m_1}(t)\Delta\theta^{i_2,n}_{c_2,m_2}(t)\times\Big(1-\frac{\beta_{c_1,m_1,c_1}\beta_{IS,c_1}}{\bar{\beta}\bar{\beta}_{c_1}}-\frac{\beta_{c_2,m_2,c_2}\beta_{IS,c_2}}{\bar{\beta}\bar{\beta}_{c_2}}\\
&\quad+\frac{\beta_{c_1,m_1,c_1}\beta_{c_2,m_2,c_2}\beta_{IS,c_1}\beta_{IS,c_2}}{MCKK'I\bar{\beta}^2\bar{\beta}_{c_1}\bar{\beta}_{c_2}}\times\big(4+2(K'-1)+(M-1)\\
&\quad\times(K-1)(I-1)(2+(K'-1)(C-1))\big)\Big)\Bigg],\\
&=\mathbb{E}\Big[\frac{1}{M^2C^2}\sum_{\substack{c_1,c_2,m_1,\\m_2,i_1,i_2}}\Delta\theta^{i_1,n}_{c_1,m_1}(t)\Delta\theta^{i_2,n}_{c_2,m_2}(t)A(m_1,m_2,c_1,c_2)\Big],
\end{aligned}\quad(33)$$

where $A(m_1, m_2, c_1, c_2)$ is given in Theorem 1. Combining for all symbols, we have

$$\sum_{n=1}^{2N} \mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,1}(t) - \Delta\theta^n_{PS}(t)\big)^2\Big]$$
$$= \frac{1}{M^2C^2}\sum_{n=1}^{2N} \sum_{\substack{c_1,c_2,m_1,\\ m_2,i_1,i_2}} A(m_1,m_2,c_1,c_2)\mathbb{E}\Big[\Delta\theta^{i_1,n}_{c_1,m_1}(t)\Delta\theta^{i_2,n}_{c_2,m_2}(t)\Big]$$
$$\overset{(a)}{\leq} \frac{\eta^2(t)G^2I^2\tau^2}{M^2C^2} \sum_{\substack{c_1,c_2,\\ m_1,m_2}} A(m_1,m_2,c_1,c_2), \tag{34}$$

where (a) is obtained using Lemma 4. ∎

Lemma 7. $\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\Big]$
$\leq \frac{(K'+1)\eta^2(t)G^2I\tau^2}{KK'M^2C^2\bar{\beta}^2} \sum_{\substack{c,m,\\ m'\neq m}} \frac{\beta^2_{IS,c}\beta_{c,m,c}\beta_{c,m',c}}{\bar{\beta}^2_c}.$

Proof: For $1 \leq n \leq N$, we have

$$\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\Big] = \frac{1}{K^2(K')^2M^2C^2\sigma_h^8\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{\substack{c_1,c_2,m_1,m_2,\\ m_1'\neq m_1,m_2'\neq m_2,\\ i_1,i_2,k_1,k_2,k_1',k_2'}}$$
$$\times \frac{1}{\bar{\beta}_{c_1}\bar{\beta}_{c_2}}|h^n_{PS,c_1,k_1'}(t)|^2 \operatorname{Re}\Big\{h^{i_1,n}_{\substack{c_1,m_1,c_1,k_1\\ c_1,m_1',c_1,k_1}}(t)\Delta\theta^{i_1,n,cx}_{c_1,m_1'}(t)\Big\}$$
$$\times |h^n_{PS,c_2,k_2'}(t)|^2 \operatorname{Re}\Big\{h^{i_2,n}_{\substack{c_2,m_2,c_2,k_2\\ c_2,m_2',c_2,k_2}}(t)\Delta\theta^{i_2,n,cx}_{c_2,m_2'}(t)\Big\}\Bigg]. \tag{35}$$

In order for the expectation not to be zero, we need to have $c_1 = c_2, i_1 = i_2$ and $k_1 = k_2$ because of the independence of different channel realizations. Then, using $\mathbb{E}\left[|h^n_{IS,c,k}(t)|^4\right] = 2\beta^2_{IS,c}\sigma_h^4$, we have

$$= \frac{(K'+1)}{K^2K'M^2C^2\sigma_h^4\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{\substack{c,m,i,\\ m'\neq m,k}} \frac{\beta^2_{IS,c}}{\bar{\beta}^2_c}\operatorname{Re}\Big\{h^{i,n}_{\substack{c,m,c,k\\ c,m',c,k}}(t)\Delta\theta^{i,n,cx}_{c,m'}(t)\Big\}^2$$
$$+ \operatorname{Re}\Big\{h^{i,n}_{\substack{c,m,c,k\\ c,m',c,k}}(t)\Delta\theta^{i,n,cx}_{c,m'}(t)\Big\}\operatorname{Re}\Big\{h^{i,n}_{\substack{c,m',c,k,\\ c,m,c,k}}(t)\Delta\theta^{i,n,cx}_{c,m}(t)\Big\}\Bigg],$$
$$= \frac{(K'+1)}{2KK'M^2C^2\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{\substack{c,m,i,\\ m'\neq m}} \frac{\beta^2_{IS,c}\beta_{c,m,c}\beta_{c,m',c}}{\bar{\beta}^2_c}$$
$$\times\Big(\Big(\Delta\theta^{i,n}_{c,m'}(t)\Big)^2 + \Big(\Delta\theta^{i,n+N}_{c,m'}(t)\Big)^2$$
$$+ \Delta\theta^{i,n}_{c,m}(t)\Delta\theta^{i,n}_{c,m'}(t) - \Delta\theta^{i,n+N}_{c,m}(t)\Delta\theta^{i,n+N}_{c,m'}(t)\Big)\Bigg]. \tag{36}$$

For $N+1 \leq n \leq 2N$, we can similarly obtain

$$= \frac{(K'+1)}{2KK'M^2C^2\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{\substack{c,m,i\\ m'\neq m}} \frac{\beta^2_{IS,c}\beta_{c,m,c}\beta_{c,m',c}}{\bar{\beta}^2_c}$$
$$\times\Big(\Big(\Delta\theta^{i,n}_{c,m'}(t)\Big)^2 + \Big(\Delta\theta^{i,n-N}_{c,m'}(t)\Big)^2$$
$$+ \Delta\theta^{i,n}_{c,m}(t)\Delta\theta^{i,n}_{c,m'}(t) - \Delta\theta^{i,n-N}_{c,m}(t)\Delta\theta^{i,n-N}_{c,m'}(t)\Big)\Bigg]. \tag{37}$$

Combining the two cases, it becomes

$$\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\Big] = \frac{(K'+1)}{KK'M^2C^2\bar{\beta}^2}$$
$$\times \sum_{\substack{c,m,i,\\ m'\neq m}} \frac{\beta^2_{IS,c}\beta_{c,m,c}\beta_{c,m',c}}{\bar{\beta}^2_c}\mathbb{E}\Big[\big\|\Delta\boldsymbol{\theta}^i_{c,m'}(t)\big\|_2^2\Big],$$
$$\overset{(a)}{\leq} \frac{(K'+1)\eta^2(t)G^2I\tau^2}{KK'M^2C^2\bar{\beta}^2} \sum_{\substack{c,m,\\ m'\neq m}} \frac{\beta^2_{IS,c}\beta_{c,m,c}\beta_{c,m',c}}{\bar{\beta}^2_c}, \tag{38}$$

where (a) is obtained using Lemma 5. ∎

Lemma 8. $\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,3}(t)\big)^2\Big]$
$\leq \frac{(K'+1)\eta^2(t)G^2I\tau^2}{KK'M^2C^2\bar{\beta}^2} \sum_{\substack{c,m,m',\\ c'\neq c}} \frac{\beta^2_{IS,c}\beta_{c,m,c}\beta_{c,m',c'}}{\bar{\beta}^2_c}$

Proof: For $1 \leq n \leq N$, we have

$$\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,3}(t)\big)^2\Big]$$
$$= \frac{1}{K^2(K')^2M^2C^2\sigma_h^8\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{\substack{c_1,c_2,m_1,m_2,m_1',\\ m_2',c_1'\neq c_1,c_2'\neq c_2,\\ i_1,i_2,k_1,k_2,k_1',k_2'}} \frac{1}{\bar{\beta}_{c_1}\bar{\beta}_{c_2}}$$
$$\times |h^n_{PS,c_1,k_1'}(t)|^2 \operatorname{Re}\Big\{h^{i_1,n}_{\substack{c_1,m_1,c_1,k_1\\ c_1,m_1',c_1',k_1}}(t)\Delta\theta^{i_1,n,cx}_{c_1',m_1'}(t)\Big\}$$
$$\times |h^n_{PS,c_2,k_2'}(t)|^2 \operatorname{Re}\Big\{h^{i_2,n}_{\substack{c_2,m_2,c_2,k_2\\ c_2,m_2',c_2',k_2}}(t)\Delta\theta^{i_2,n,cx}_{c_2',m_2'}(t)\Big\}\Bigg]. \tag{39}$$

In order for the expectation not to be zero, we need to have $c_1 = c_2, c_1' = c_2', m_1 = m_2, m_1' = m_2', i_1 = i_2$ and $k_1 = k_2$ because of the independence of different channel realizations. We get

$$= \frac{(K'+1)}{K^2K'M^2C^2\sigma_h^4\bar{\beta}^2}$$
$$\times \mathbb{E}\Bigg[\sum_{\substack{c,m,m',\\ c'\neq c,i,k}} \frac{\beta^2_{IS,c}}{\bar{\beta}^2_c}\Big(\operatorname{Re}\Big\{h^{i,n}_{\substack{c,m,c,k,\\ c,m',c',k}}(t)\Delta\theta^{i,n,cx}_{c',m'}(t)\Big\}\Big)^2\Bigg],$$
$$= \frac{(K'+1)}{2KK'M^2C^2\bar{\beta}^2}$$
$$\times \mathbb{E}\Bigg[\sum_{\substack{c,m,m',\\ c'\neq c,i}} \frac{\beta^2_{IS,c}\beta_{c,m,c}\beta_{c,m',c'}}{\bar{\beta}^2_c}\Big(\big(\Delta\theta^{i,n}_{c',m'}(t)\big)^2 + \big(\Delta\theta^{i,n+N}_{c',m'}(t)\big)^2\Big)\Bigg]. \tag{40}$$

A similar expression can be obtained for $N+1 \leq n \leq 2N$. Combining two cases, it becomes

$$\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,3}(t)\big)^2\Big] = \sum_{n=1}^{N}\frac{(K'+1)}{KK'M^2C^2\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{\substack{c,m,m',\\ c'\neq c,i}}$$
$$\times \frac{\beta^2_{IS,c}\beta_{c,m,c}\beta_{c,m',c'}}{\bar{\beta}^2_c}\Big(\big(\Delta\theta^{i,n}_{c',m'}(t)\big)^2 + \big(\Delta\theta^{i,n+N}_{c',m'}(t)\big)^2\Big)\Bigg]$$
$$\overset{(a)}{\leq} \frac{(K'+1)\eta^2(t)G^2I\tau^2}{KK'M^2C^2\bar{\beta}^2} \sum_{\substack{c,m,m',\\ c'\neq c}} \frac{\beta^2_{IS,c}\beta_{c,m,c}\beta_{c,m',c'}}{\bar{\beta}^2_c}, \tag{41}$$

where (a) is obtained using Lemma 5. ∎

Lemma 9. $\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,4}(t)\big)^2\Big]$
$= \frac{(K'+1)I\sigma_z^2N}{KK'M^2C^2P_t^2\sigma_h^2\bar{\beta}^2} \sum_{c,m} \frac{\beta^2_{IS,c}\beta_{m,c,c}}{\bar{\beta}^2_c}$

Proof: For $1 \le n \le N$, we have

$$\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,4}(t)\big)^2\Big] = \frac{1}{P^2_{IS,t}K^2\big(K'\big)^2M^2C^2\sigma_h^8\bar{\beta}^2} \times \mathbb{E}\Big[\sum_{\substack{c_1,c_2,m_1,m_2,i_1,\\ i_2,k_1,k_2,k_1',k_2'}} \frac{1}{\bar{\beta}_{c_1}\bar{\beta}_{c_2}} \times |h^n_{PS,c_1,k_1'}(t)|^2 \operatorname{Re}\Big\{\big(h^{i_1,n}_{c_1,m_1,c_1,k_1}(t)\big)^* z^{i_1,n}_{c_1,k_1}(t)\Big\} \times |h^n_{PS,c_2,k_2'}(t)|^2 \operatorname{Re}\Big\{\big(h^{i_2,n}_{c_2,m_2,c_2,k_2}(t)\big)^* z^{i_2,n}_{c_2,k_2}(t)\Big\}\Big]. \quad (42)$$

For a non-zero result, we need to have $c_1 = c_2, i_1 = i_2, m_1 = m_2$ and $k_1 = k_2$. Then, we get

$$= \frac{(K'+1)}{P^2_{IS,t}K^2K'M^2C^2\sigma_h^4\bar{\beta}^2}\mathbb{E}\Big[\sum_{c,m,i,k}\frac{\beta^2_{IS,c}}{\bar{\beta}^2_c} \times \Big(\operatorname{Re}\Big\{\big(h^{i,n}_{c,m,c,k}(t)\big)^* z^{i,n}_{c,k}(t)\Big\}\Big)^2\Big],$$
$$= \frac{(K'+1)I\sigma_z^2}{2P^2_{IS,t}KK'M^2C^2\sigma_h^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{\beta^2_{IS,c}\beta_{c,m,c}}{\bar{\beta}^2_c}. \quad (43)$$

The derivation is similar for $N+1 \le n \le 2N$. Combining the two cases, we get

$$\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,4}(t)\big)^2\Big] = \frac{(K'+1)I\sigma_z^2N}{P^2_{IS,t}KK'M^2C^2\sigma_h^2\bar{\beta}^2}\sum_{c,m}\frac{\beta^2_{IS,c}\beta_{c,m,c}}{\bar{\beta}^2_c}. \quad (44)$$

■

Lemma 10. $\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,5}(t)\big)^2\Big]$
$\le \frac{\big(2+(M-1)(C-2)(K-1)(I-1)\big)\eta^2(t)IG^2\tau^2}{K(K')M^3C^2(C-1)\bar{\beta}^2} \times \sum_{\substack{c,m_1,m_2,\\ c'\neq c}} \frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m_1,c'}\beta_{c',m_2,c'}}{\bar{\beta}^2_{c'}}$

Proof: For $1 \le n \le N$, the equation becomes

$$\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,5}(t)\big)^2\Big] = \frac{1}{K^2(K')^2M^2C^2\sigma_h^8\bar{\beta}^2}\mathbb{E}\Big[\sum_{\substack{c_1,c_2,m_1,m_2,i_1,\\ c_1'\neq c_1,c_2'\neq c_2,\\ i_2,k_1,k_2,k_1',k_2'}} \times \frac{1}{\bar{\beta}_{c_1'}\bar{\beta}_{c_2'}}|h^{i_1,n}_{c_1',m_1,c_1',k_1}(t)|^2 \operatorname{Re}\Big\{h^{PS,n}_{\substack{c_1,k_1'\\ c_1',k_1'}}(t)\Delta\theta^{i_1,n,cx}_{c_1',m_1}(t)\Big\} \times |h^{i_2,n}_{c_2',m_2,c_2',k_2}(t)|^2 \operatorname{Re}\Big\{h^{PS,n}_{\substack{c_2,k_2'\\ c_2',k_2'}}(t)\Delta\theta^{i_2,n,cx}_{c_2',m_2}(t)\Big\}\Big]. \quad (45)$$

For a non-zero answer, we need to have $k_1' = k_2'$. The expression becomes

$$= \frac{\big(2+(M-1)(C-2)(K-1)(I-1)\big)}{4(K')M^3C^2(C-1)KI\bar{\beta}^2} \mathbb{E}\Big[\sum_{\substack{c,c'\neq c,m_1,\\ m_2,i_1,i_2}} \beta_{IS,c}\beta_{IS,c'}\beta_{c',m_1,c'}\Big(\frac{\beta_{c',m_2,c'}}{\bar{\beta}^2_{c'}} \times \Big(\Delta\theta^{i_1,n}_{c',m_1}(t)\Delta\theta^{i_2,n}_{c',m_2}(t) + \Delta\theta^{i_1,n+N}_{c',m_1}(t)\Delta\theta^{i_2,n+N}_{c',m_2}(t)\Big) + \frac{\beta_{c,m_2,c}}{\bar{\beta}_c\bar{\beta}_{c'}}\Big(\Delta\theta^{i_1,n}_{c',m_1}(t)\Delta\theta^{i_2,n}_{c,m_2}(t) - \Delta\theta^{i_1,n+N}_{c',m_1}(t)\Delta\theta^{i_2,n+N}_{c,m_2}(t)\Big)\Big)\Big]. \quad (46)$$

The result is similar for $N+1 \le n \le 2N$. Overall, it becomes

$$\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,5}(t)\big)^2\Big] = \frac{\big(2+(M-1)(C-2)(K-1)(I-1)\big)}{2K(K')M^3C^2(C-1)I\bar{\beta}^2} \times \sum_{n=1}^{N}\sum_{\substack{c,c'\neq c,\\ m_1,m_2,i_1,i_2}} \frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m_1,c'}\beta_{c',m_2,c'}}{\bar{\beta}^2_{c'}} \times \mathbb{E}\Big[\Big(\Delta\theta^{i_1,n}_{c',m_1}(t)\Delta\theta^{i_2,n}_{c',m_2}(t) + \Delta\theta^{i_1,n+N}_{c',m_1}(t)\Delta\theta^{i_2,n+N}_{c',m_2}(t)\Big)\Big]$$
$$\overset{(a)}{\le} \frac{\big(2+(M-1)(C-2)(K-1)(I-1)\big)\eta^2(t)IG^2\tau^2}{K(K')M^3C^2(C-1)\bar{\beta}^2} \times \sum_{\substack{c,c'\neq c,\\ m_1,m_2}} \frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m_1,c'}\beta_{c',m_2,c'}}{\bar{\beta}^2_{c'}}, \quad (47)$$

where (a) is obtained using Lemma 4. ■

Lemma 11. $\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,6}(t)\big)^2\Big]$
$\le \frac{\eta^2(t)IG^2\tau^2}{KK'M^2C^2\bar{\beta}^2}\sum_{\substack{c,c'\neq c,\\ m,m'\neq m}} \frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m',c'}}{\bar{\beta}^2_{c'}}$

Proof: For $1 \le n \le N$, we have

$$\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,6}(t)\big)^2\Big] = \frac{1}{K^2(K')^2M^2C^2\sigma_h^8\bar{\beta}^2}\mathbb{E}\Big[\sum_{\substack{c_1,c_2,m_1,m_2,i_1,\\ i_2,k_1,k_2,k_1',k_2',\\ c_1'\neq c_1,c_2'\neq c_2,\\ m_1'\neq m_1,m_2'\neq m_2}} \times \frac{1}{\bar{\beta}_{c_1'}\bar{\beta}_{c_2'}} \operatorname{Re}\Big\{h^{PS,n}_{\substack{c_1,k_1'\\ c_1',k_1'}}(t)h^{i_1,n}_{\substack{c_1',m_1,c_1',k_1\\ c_1',m_1',c_1',k_1}}(t)\Delta\theta^{i_1,n,cx}_{c_1',m_1'}(t)\Big\} \times \operatorname{Re}\Big\{h^{PS,n}_{\substack{c_2,k_2'\\ c_2',k_2'}}(t)h^{i_2,n}_{\substack{c_2',m_2,c_2',k_2\\ c_2',m_2',c_2',k_2}}(t)\Delta\theta^{i_2,n,cx}_{c_2',m_2'}(t)\Big\}\Big]. \quad (48)$$

For a non-zero result, we need to have $k_1 = k_2$, $k_1' = k_2'$, $i_1 = i_2$, and $c_1 = c_2$, which leads to $c_1' = c_2'$.

$$= \frac{1}{K^2(K')^2M^2C^2\sigma_h^8\bar{\beta}^2} \mathbb{E}\Big[\sum_{\substack{c,c'\neq c,m,\\ m'\neq m,i,k,k'}} \frac{1}{\bar{\beta}^2_{c'}}\Big(\Big(\operatorname{Re}\Big\{h^{PS,n}_{\substack{c,k'\\ c',k'}}(t)h^{i,n}_{\substack{c',m,c',k,\\ c',m',c',k}}(t)\Delta\theta^{i,n,cx}_{c',m'}(t)\Big\}\Big)^2 + \operatorname{Re}\Big\{h^{PS,n}_{\substack{c,k'\\ c',k'}}(t)h^{i,n}_{\substack{c',m,c',k,\\ c',m',c',k}}(t)\Delta\theta^{i,n,cx}_{c',m'}(t)\Big\} \times \operatorname{Re}\Big\{h^{PS,n}_{\substack{c,k'\\ c',k'}}(t)h^{i,n}_{\substack{c',m',c',k,\\ c',m,c',k}}(t)\Delta\theta^{i,n,cx}_{c',m}(t)\Big\}\Big)\Big],$$
$$= \mathbb{E}\Big[\sum_{\substack{c,c'\neq c,i,\\ m,m'\neq m}} \frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m',c'}}{2K(K')M^2C^2\bar{\beta}^2\bar{\beta}^2_{c'}} \times \Big((\Delta\theta^{i,n}_{c',m'}(t))^2 + (\Delta\theta^{i,n+N}_{c',m'}(t))^2\Big)\Big]. \quad (49)$$

The result is similar for $N+1 \le n \le 2N$. Combining the two parts, we have

$$\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,6}(t)\big)^2\Big] = \sum_{\substack{c,c'\neq c,i,\\ m,m'\neq m}} \frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m',c'}}{K(K')M^2C^2\bar{\beta}^2\bar{\beta}^2_{c'}}\mathbb{E}\Big[\big\|\Delta\boldsymbol{\theta}^i_{c',m'}(t)\big\|_2^2\Big],$$

$$\overset{(a)}{\leq} \frac{\eta^2(t) I G^2 \tau^2}{K K' M^2 C^2 \bar{\beta}^2} \sum_{\substack{c,c' \neq c, \\ m,m' \neq m}} \frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m',c'}}{\bar{\beta}_{c'}^2}, \quad (50)$$

where (a) is obtained using Lemma 5. ∎

Lemma 12. $\sum_{n=1}^{2N} \mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,7}(t)\big)^2\Big]$
$\leq \frac{\eta^2(t) I G^2 \tau^2}{K K' M^2 C^2 \bar{\beta}^2} \sum_{\substack{c,c' \neq c, \\ c'' \neq c', m}} \frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m,c''}}{\bar{\beta}_{c'}^2}$

Proof: For $1 \leq n \leq N$, we have

$$\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,7}(t)\big)^2\Big] = \frac{1}{K^2(K')^2M^2C^2\sigma_h^8\bar{\beta}^2} \sum_{\substack{c_1,c_2,m_1,m_2,m_1', \\ m_2',i_1,i_2,k_1,k_2, \\ k_1',c_1'\neq c_1,c_2'\neq c_2, \\ k_2',c_1''\neq c_1',c_2''\neq c_2'}}$$

$$\times \mathbb{E}\Bigg[\frac{1}{\bar{\beta}_{c_1'}\bar{\beta}_{c_2'}} \mathrm{Re}\Big\{h^{PS,n}_{\substack{c_1,k_1' \\ c_1',k_1'}}(t) h^{i_1,n}_{\substack{c_1',m_1,c_1',k_1 \\ c_1',m_1',c_1'',k_1}}(t)\Delta\theta^{i_1,n,cx}_{c_1'',m_1'}(t)\Big\}$$

$$\times \mathrm{Re}\Big\{h^{PS,n}_{\substack{c_2,k_2' \\ c_2',k_2'}}(t) h^{i_2,n}_{\substack{c_2',m_2,c_2',k_2 \\ c_2',m_2',c_2'',k_2}}(t)\Delta\theta^{i_2,n,cx}_{c_2'',m_2'}(t)\Big\}\Bigg]. \quad (51)$$

For a non-zero answer, we need to have $m_1 = m_2$, $m_1' = m_2'$ $k_1 = k_2$, $k_1' = k_2'$, $i_1 = i_2$, and $c_1' = c_2'$ and $c_1'' = c_2''$, which leads to $c_1 = c_2$. Then, we have

$$= \frac{1}{K^2(K')^2M^2C^2\sigma_h^8\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{\substack{c,c'\neq c,c''\neq c', \\ m,m',i,k,k'}} \frac{1}{\bar{\beta}_{c'}^2}$$

$$\times \Big(\mathrm{Re}\Big\{h^{PS,n}_{\substack{c,k' \\ c',k'}}(t) h^{i,n}_{\substack{c',m,c',k, \\ c',m',c'',k}}(t)\Delta\theta^{i,n,cx}_{c'',m}(t)\Big\}\Big)^2\Bigg],$$

$$= \mathbb{E}\Bigg[\sum_{\substack{c,c'\neq c,c''\neq c', \\ m,m',i}} \frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m',c''}}{2KK'M^2C^2\bar{\beta}^2\bar{\beta}_{c'}^2}$$

$$\times \Big((\Delta\theta^{i,n}_{c'',m'}(t))^2 + (\Delta\theta^{i,n+N}_{c'',m'}(t))^2\Big)\Bigg]. \quad (52)$$

The derivation is similar for $N+1 \leq n \leq 2N$. Combining the two parts, we have

$$\sum_{n=1}^{2N} \mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,7}(t)\big)^2\Big]$$
$$= \sum_{n=1}^{N} \sum_{\substack{c,c'\neq c,c''\neq c', \\ m,m',i,k,k'}} \frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m',c''}}{K(K')M^2C^2\bar{\beta}^2\bar{\beta}_{c'}^2}$$
$$\times \mathbb{E}\Big[(\Delta\theta^{i,n}_{c'',m'}(t))^2 + (\Delta\theta^{i,n+N}_{c'',m'}(t))^2\Big]$$
$$\overset{(a)}{\leq} \frac{\eta^2(t) I G^2 \tau^2}{K K' M^2 C^2 \bar{\beta}^2} \sum_{\substack{c,c' \neq c, \\ c'' \neq c', m}} \frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m,c''}}{\bar{\beta}_{c'}^2}. \quad (53)$$

where (a) is due to Lemma 5. ∎

Lemma 13. $\sum_{n=1}^{2N} \mathbb{E}\Big[\Big(\Delta\hat{\theta}^n_{PS,6}(t)\Big)^2\Big]$
$= \frac{\sigma_z^2 I N}{P_{IS,t}^2 K(K')M^2C^2\sigma_h^2\bar{\beta}^2} \sum_{\substack{c,m, \\ c'\neq c}} \frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}}{\bar{\beta}_{c'}^2}$

Proof: For $1 \leq n \leq N$, we have

$$\mathbb{E}\Big[\Big(\Delta\hat{\theta}^n_{PS,6}(t)\Big)^2\Big]$$
$$= \frac{1}{P_{IS,t}^2K^2(K')^2M^2C^2\sigma_h^8\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{\substack{c_1,c_2,m_1,m_2,i_1, \\ i_2,k_1,k_2,k_1',k_2', \\ c_1'\neq c_1,c_2'\neq c_2}} \frac{1}{\bar{\beta}_{c_1'}\bar{\beta}_{c_2'}}$$

$$\times \mathrm{Re}\Big\{\big(h^{PS,n}_{\substack{c_1,k_1' \\ c_1',k_1'}}(t)\big(h^{i_1,n}_{c_1',m_1,c_1',k_1}(t)\big)^* z^{i_1,n}_{IS,c_1',k_1}(t)\Big\}$$

$$\times \mathrm{Re}\Big\{\big(h^{PS,n}_{\substack{c_2,k_2' \\ c_2',k_2'}}(t)\big(h^{i_2,n}_{c_2',m_2,c_2',k_2}(t)\big)^* z^{i_2,n}_{IS,c_2',k_2}(t)\Big\}\Bigg]. \quad (54)$$

For a non-zero answer, we have $m_1 = m_2, c_1 = c_2, c_1' = c_2', k_1 = k_2, k_1' = k_2', i_1 = i_2$. Then, it becomes

$$\mathbb{E}\Big[\Big(\Delta\hat{\theta}^n_{PS,6}(t)\Big)^2\Big] = \frac{\sigma_z^2 I \sum_{\substack{c,m, \\ c'\neq c}} \frac{1}{\bar{\beta}_{c'}^2}\big(\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\big)}{2P_{IS,t}^2K(K')M^2C^2\sigma_h^2\bar{\beta}^2}. \quad (55)$$

The solution is the same for $N+1 \leq n \leq 2N$. Adding all the terms concludes the lemma. ∎

Lemma 14. $\sum_{n=1}^{2N} \mathbb{E}\Big[\Big(\Delta\hat{\theta}^n_{PS,7}(t)\Big)^2\Big] = \frac{\sigma_z^2 N}{P_{IS,t}^2(K')C^2\sigma_h^2\bar{\beta}^2} \sum_{c=1}^{C} \beta_{IS,c}.$

Proof: For $1 \leq n \leq N$, we have

$$\mathbb{E}\Big[\Big(\Delta\hat{\theta}^n_{PS,7}(t)\Big)^2\Big]$$
$$= \frac{1}{P_{IS,t}^2(K')^2C^2\sigma_h^4\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{\substack{c_1,c_2, \\ k_1',k_2'}} \mathrm{Re}\Big\{\big(h^n_{PS,c_1,k_1'}(t)\big)^* z^n_{PS,k_1'}(t)\Big\}$$

$$\times \mathrm{Re}\Big\{\big(h^n_{PS,c_2,k_2'}(t)\big)^* z^n_{PS,k_2'}(t)\Big\}\Bigg]. \quad (56)$$

For a non-zero answer, we have $c_1 = c_2$ and $k_1' = k_2'$. Then, it becomes

$$\mathbb{E}\Big[\Big(\Delta\hat{\theta}^n_{PS,7}(t)\Big)^2\Big] = \frac{\mathbb{E}\Big[\sum_{c,k'}\Big(\mathrm{Re}\Big\{(h^n_{PS,c,k'}(t))^* z^n_{PS,k'}(t)\Big\}\Big)^2\Big]}{P_{IS,t}^2(K')^2C^2\sigma_h^4\bar{\beta}^2}$$
$$= \frac{\sigma_z^2}{2P_{IS,t}^2(K')C^2\sigma_h^2\bar{\beta}^2} \sum_{c=1}^{C} \beta_{IS,c}. \quad (57)$$

The solution is similar for $N+1 \leq n \leq 2N$. Summing over all the symbols concludes the lemma. ∎

Combining Lemmas 6-14 completes the proof of Lemma 1.

Appendix C: Proof of Lemma 2

We have

$$\mathbb{E}\big[\|v(t+1) - \boldsymbol{\theta}^*\|_2^2\big] = \mathbb{E}\big[\|\boldsymbol{\theta}_{PS}(t) + \Delta\boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^*\|_2^2\big],$$
$$= \mathbb{E}\big[\|\boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^*\|_2^2\big] + \mathbb{E}\big[\|\Delta\boldsymbol{\theta}_{PS}(t)\|_2^2\big]$$
$$+ 2\mathbb{E}\big[\langle\boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^*, \Delta\boldsymbol{\theta}_{PS}(t)\rangle\big]. \quad (58)$$

Where the second term can be bounded as

$$\mathbb{E}\big[\|\Delta\boldsymbol{\theta}_{PS}(t)\|_2^2\big] = \mathbb{E}\Big[\Big\|\frac{1}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\Delta\boldsymbol{\theta}^i_{c,m}(t)\Big\|_2^2\Big]$$
$$\overset{(a)}{\leq} \frac{1}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\mathbb{E}\big[\|\Delta\boldsymbol{\theta}^i_{c,m}(t)\|_2^2\big]$$
$$\overset{(b)}{=} \frac{\eta^2(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\mathbb{E}\Big[\Big\|\sum_{j=1}^{\tau}\nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)\big)\Big\|_2^2\Big]$$
$$\leq \frac{\eta^2(t)\tau}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=1}^{\tau}\mathbb{E}\big[\|\nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)\big)\|_2^2\big]$$

$$\overset{(c)}{\leq} \eta^2(t) I \tau^2 G^2, \tag{59}$$

where (a) is due to the convexity of $\|\|_2^2$, (b) comes from utilizing (21), and (c) is obtained using Assumption 2. Plugging in the result to (58), we have

$$\begin{aligned}\mathbb{E}\big[\big\|v(t+1)-\boldsymbol{\theta}^*\big\|_2^2\big] \leq & \mathbb{E}\big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\big] + \eta^2(t) I \tau^2 G^2 \\ & + 2\mathbb{E}\big[\langle \boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*, \Delta\boldsymbol{\theta}_{PS}(t)\rangle\big].\end{aligned} \tag{60}$$

The last term of (60), we have

$$\begin{aligned}&2\mathbb{E}\big[\langle \boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*, \Delta\boldsymbol{\theta}_{PS}(t)\rangle\big] \\ &= \frac{2}{MC}\sum_{c,m,i}\mathbb{E}\big[\langle \boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*, \Delta\boldsymbol{\theta}^i_{c,m}(t)\rangle\big], \\ &= \frac{2\eta(t)}{MC}\sum_{c,m,i}\mathbb{E}\Big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}_{PS}(t), \sum_{j=1}^{\tau}\nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)\big)\rangle\Big], \\ &= \frac{2\eta(t)}{MC}\sum_{c,m,i}\mathbb{E}\big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}_{PS}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}_{PS}(t), \boldsymbol{\xi}^{1,1}_{c,m}(t)\big)\rangle\big] \\ &\quad + \frac{2\eta(t)}{MC}\sum_{c,m,i}\mathbb{E}\Big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}_{PS}(t), \sum_{j=2}^{\tau}\nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)\big)\rangle\Big].\end{aligned} \tag{61}$$

For the first term of (61), we have

$$\begin{aligned}&\frac{2\eta(t)}{MC}\sum_{c,m,i}\mathbb{E}\big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}_{PS}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}_{PS}(t), \boldsymbol{\xi}^{1,1}_{c,m}(t)\big)\rangle\big] \\ &\overset{(a)}{=} \frac{2\eta(t)}{MC}\sum_{c,m,i}\mathbb{E}\big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}_{PS}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}_{PS}(t)\big)\rangle\big] \\ &\overset{(b)}{\leq} \frac{2\eta(t)}{MC}\sum_{c,m,i}\mathbb{E}\big[F_{c,m}(\boldsymbol{\theta}^*) - F_{c,m}(\boldsymbol{\theta}_{PS}(t)) - \frac{\mu}{2}\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\big] \\ &= 2\eta(t) I\Big(F^* - \mathbb{E}\big[F(\boldsymbol{\theta}_{PS}(t))\big] - \frac{\mu}{2}\mathbb{E}\big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\big]\Big) \\ &\overset{(c)}{\leq} -\eta(t) I \mu \mathbb{E}\big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\big],\end{aligned} \tag{62}$$

where (a) comes from $\mathbb{E}_{\xi}\Big[\nabla F_{c,m}\Big(\boldsymbol{\theta}_{PS}(t), \boldsymbol{\xi}^{1,1}_{c,m}(t)\Big)\Big] = \nabla F_{c,m}(\boldsymbol{\theta}_{PS}(t))$, (b) holds due to the μ-strong convexity of $F_{c,m}$, and (c) follows since $F^* \leq F(\boldsymbol{\theta}(t))$. For the second term of (61), we have

$$\begin{aligned}&\frac{2\eta(t)}{MC}\sum_{c,m,i}\mathbb{E}\Big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}_{PS}(t), \sum_{j=2}^{\tau}\nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)\big)\rangle\Big] \\ &= \frac{2\eta(t)}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}_{PS}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)\big)\rangle\big], \\ &= \frac{2\eta(t)}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\big[\langle \boldsymbol{\theta}^{i,j}_{c,m}(t)-\boldsymbol{\theta}_{PS}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)\big)\rangle\big] \\ &\quad + \frac{2\eta(t)}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}^{i,j}_{c,m}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)\big)\rangle\big].\end{aligned} \tag{63}$$

Using Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}&\frac{2\eta(t)}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\big[\langle \boldsymbol{\theta}^{i,j}_{c,m}(t)-\boldsymbol{\theta}_{PS}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)\big)\rangle\big] \\ &\leq \frac{\eta(t)}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\Big[\frac{1}{\eta(t)}\big\|\boldsymbol{\theta}^{i,j}_{c,m}(t)-\boldsymbol{\theta}_{PS}(t)\big\|_2^2 \\ &\qquad\qquad + \eta(t)\big\|\nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)\big)\big\|_2^2\Big] \\ &\overset{(a)}{\leq} \frac{1}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\Big[\frac{1}{\eta(t)}\big\|\boldsymbol{\theta}^{i,j}_{c,m}(t)-\boldsymbol{\theta}_{PS}(t)\big\|_2^2\Big] + \eta^2(t) I(\tau-1) G^2,\end{aligned} \tag{64}$$

where (a) is obtained using Assumption 2. The following lemma will give an upper bound on the second term of (63).

Lemma 15. $\frac{2\eta(t)}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}^{i,j}_{c,m}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)\big)\rangle\big]$

$$\begin{aligned}&= -\mu\eta(t)(1-\eta(t)) I(\tau-1)\mathbb{E}\big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\big] \\ &+ \frac{\mu(1-\eta(t))}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\big[\big\|\boldsymbol{\theta}^{i,j}_{c,m}(t)-\boldsymbol{\theta}_{PS}(t)\big\|_2^2\big] + 2\eta(t) I(\tau-1)\Gamma.\end{aligned}$$

Proof: See Appendix D. ∎

Using the results in (64) and (70), we can write (63) as

$$\begin{aligned}&\frac{2\eta(t)}{MC}\sum_{c,m,i}\mathbb{E}\Big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}_{PS}(t), \sum_{j=2}^{\tau}\nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)\big)\rangle\Big] \\ &= -\mu\eta(t)(1-\eta(t)) I(\tau-1)\mathbb{E}\big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\big] \\ &\quad + \frac{(1+\mu(1-\eta(t)))}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\big[\big\|\boldsymbol{\theta}^{i,j}_{c,m}(t)-\boldsymbol{\theta}_{PS}(t)\big\|_2^2\big] \\ &\quad + \eta^2(t) I(\tau-1) G^2 + 2\eta(t) I(\tau-1)\Gamma.\end{aligned} \tag{65}$$

Also, we have

$$\begin{aligned}&\frac{1}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\big[\big\|\boldsymbol{\theta}^{i,j}_{c,m}(t)-\boldsymbol{\theta}_{PS}(t)\big\|_2^2\big] \\ &\quad = \frac{\eta^2}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\big[\big\|\nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)\big)\big\|_2^2\big] \\ &\quad \overset{(a)}{\leq} \eta^2 I G^2 \frac{\tau(\tau-1)(2\tau-1)}{6},\end{aligned} \tag{66}$$

where (a) is due to the convexity of L_2 norm and Assumption 2. For $\eta(t) \leq 1$, we have

$$\begin{aligned}&\frac{2\eta(t)}{MC}\sum_{c,m,i}\mathbb{E}\Big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}_{PS}(t), \sum_{j=2}^{\tau}\nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)\big)\rangle\Big] \\ &\quad \leq \mu\eta(t)(1-\eta(t)) I(\tau-1)\mathbb{E}\big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\big] \\ &\quad + \big(1+\mu(1-\eta(t))\big)\eta^2(t) I G^2 \frac{\tau(\tau-1)(2\tau-1)}{6} \\ &\quad + \eta^2(t) I(\tau-1) G^2 + 2\eta(t) I(\tau-1)\Gamma.\end{aligned} \tag{67}$$

Substituting the results in (62) and (67) into (61), we get

$$\begin{aligned}&2\mathbb{E}\big[\langle \boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*, \Delta\boldsymbol{\theta}_{PS}(t)\rangle\big] \\ &\quad \leq \mu\eta(t) I(\tau-\eta(t)(\tau-1))\mathbb{E}\big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\big] \\ &\quad + \big(1+\mu(1-\eta(t))\big)\eta^2(t) I G^2 \frac{\tau(\tau-1)(2\tau-1)}{6} \\ &\quad + \eta^2(t) I(\tau-1) G^2 + 2\eta(t) I(\tau-1)\Gamma.\end{aligned} \tag{68}$$

Lemma 2 is concluded by plugging (68) into (60).

Appendix D: Proof of Lemma 15

We have

$$\begin{aligned}&\frac{2\eta(t)}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}^{i,j}_{c,m}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)\big)\rangle\big] \\ &\overset{(a)}{\leq} \frac{2\eta(t)}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}^{i,j}_{c,m}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}^{i,j}_{c,m}(t)\big)\rangle\big] \\ &\overset{(b)}{\leq} \frac{2\eta(t)}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\Big[F_{c,m}(\boldsymbol{\theta}^*) - F_{c,m}(\boldsymbol{\theta}^{i,j}_{c,m}(t)) \\ &\quad - \frac{\mu}{2}\big\|\boldsymbol{\theta}^{i,j}_{c,m}(t)-\boldsymbol{\theta}^*\big\|_2^2\Big] \\ &= \frac{2\eta(t)}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\Big[F_{c,m}(\boldsymbol{\theta}^*) - F^*_{c,m} + F^*_{c,m}\end{aligned}$$

$$
\begin{aligned}
&- F_{c,m}(\boldsymbol{\theta}_{c,m}^{i,j}(t)) - \frac{\mu}{2}\left\|\boldsymbol{\theta}_{c,m}^{i,j}(t) - \boldsymbol{\theta}^*\right\|_2^2\Big] \\
&= 2\eta(t)I(\tau-1)\Big(F^* - \frac{1}{MC}\sum_{c,m} F_{c,m}^*\Big) \\
&\quad + \frac{2\eta}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\Big(F_{c,m}^* - \mathbb{E}\left[F_{c,m}(\boldsymbol{\theta}_{c,m}^{i,j})\right]\Big) \\
&\quad - \frac{\mu\eta(t)}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\big[\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t) - \boldsymbol{\theta}^*\big\|_2^2\big] \\
&\overset{(c)}{\leq} 2\eta(t)I(\tau-1)\Gamma - \frac{\mu\eta}{MC}\sum_{c,m,i}\sum_{j=2}^{\tau}\mathbb{E}\big[\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t) - \boldsymbol{\theta}^*\big\|_2^2\big], \quad (69)
\end{aligned}
$$

where (a) is obtained using $\mathbb{E}_\xi\left[\nabla F_{c,m}\left(\boldsymbol{\theta}_{c,m}^{i,j}(t), \boldsymbol{\xi}_{c,m}^{i,j}(t)\right)\right] = F_{c,m}(\boldsymbol{\theta}_{c,m}^{i,j}(t)), \forall i,j,m,c,t$, (b) is using the μ-strong convexity of $F_{c,m}$, (c) follows since $F_{c,m}^* \leq F_{c,m}(\boldsymbol{\theta}_{c,m}^{i,j}(t))$. Also

$$
\begin{aligned}
&- \left\|\boldsymbol{\theta}_{c,m}^{i,j}(t) - \boldsymbol{\theta}^*\right\|_2^2 \\
&= - \left\|\boldsymbol{\theta}_{c,m}^{i,j}(t) - \boldsymbol{\theta}_{PS}(t)\right\|_2^2 - \left\|\boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^*\right\|_2^2 \\
&\quad - 2\langle \boldsymbol{\theta}_{c,m}^{i,j}(t) - \boldsymbol{\theta}_{PS}(t), \boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^*\rangle \\
&\overset{(a)}{\leq} - \left\|\boldsymbol{\theta}_{c,m}^{i,j}(t) - \boldsymbol{\theta}_{PS}(t)\right\|_2^2 - \left\|\boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^*\right\|_2^2 \\
&\quad + \frac{1}{\eta(t)}\left\|\boldsymbol{\theta}_{c,m}^{i,j}(t) - \boldsymbol{\theta}_{PS}(t)\right\|_2^2 + \eta(t)\left\|\boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^*\right\|_2^2 \\
&= - (1-\eta(t))\left\|\boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^*\right\|_2^2 + \left(\frac{1}{\eta(t)} - 1\right)\left\|\boldsymbol{\theta}_{c,m}^{i,j}(t) - \boldsymbol{\theta}_{PS}(t)\right\|_2^2, \quad (70)
\end{aligned}
$$

where (a) is due to Cauchy-Schwarz inequality. Plugging (69) and (70) concludes Lemma 15.

REFERENCES

[1] O. Aygün, M. Kazemi, D. Gündüz, and T. M. Duman, "Hierarchical over-the-air federated edge learning," in *2022 IEEE International Conference on Communications (ICC)*, Seoul, South Korea, May 2022.

[2] K. Wei *et al.*, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[3] W. Y. B. Lim *et al.*, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, 2020.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[5] D. Gündüz *et al.*, "Communicate to learn at the edge," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 14–19, 2020.

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

[7] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2021–2031.

[8] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 170–185, 2020.

[9] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Joint resource management and model compression for wireless federated learning," in *2021 IEEE International Conference on Communications (ICC)*, Montreal, Canada, Jun. 2021, pp. 1–6.

[10] J. Konečnỳ *et al.*, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[11] W. Zhang, X. Wang, P. Zhou, W. Wu, and X. Zhang, "Client selection for federated learning with non-iid data in mobile edge computing," *IEEE Access*, vol. 9, pp. 24 462–24 474, 2021.

[12] Y. Zhao *et al.*, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[13] C. Briggs, Z. Fan, and P. Andras, "Federated learning with hierarchical clustering of local updates to improve training on non-iid data," in *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, Sep. 2020, pp. 1–9.

[14] D. Data and S. Diggavi, "Byzantine-resilient sgd in high dimensions on heterogeneous data," in *2021 IEEE Int'l Symp. on Inform. Theory (ISIT)*, Melbourne, Australia, Jul. 2021, pp. 2310–2315.

[15] J. So, B. Güler, and A. S. Avestimehr, "Byzantine-resilient secure federated learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2168–2181, 2020.

[16] C. T. Dinh *et al.*, "Federated learning over wireless networks: Convergence analysis and resource allocation," *IEEE/ACM Trans. on Networking*, vol. 29, no. 1, pp. 398–409, 2020.

[17] S. Luo, X. Chen, Q. Wu, Z. Zhou, and S. Yu, "Hfel: Joint edge association and resource allocation for cost-efficient hierarchical federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6535–6548, 2020.

[18] 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, 2021.

[19] Y. Sun, S. Zhou, Z. Niu, and D. Gündüz, "Dynamic scheduling for over-the-air federated edge learning with energy constraints," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 227–242, 2021.

[20] M. M. Amiri, S. R. Kulkarni, and H. V. Poor, "Federated learning with downlink device selection," in *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Lucca, Italy, Sept. 2021, pp. 306–310.

[21] J. Ren *et al.*, "Scheduling for cellular federated edge learning with importance and channel awareness," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7690–7703, 2020.

[22] M. M. Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, vol. 68, pp. 2155–2169, 2020.

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

[24] 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, 2019.

[25] Y. Shao, D. Gündüz, and S. C. Liew, "Federated edge learning with misaligned over-the-air computation," *IEEE Trans. Wireless Commun.*, 2021.

[26] X. Wei and C. Shen, "Federated learning over noisy channels: Convergence analysis and design examples," *IEEE Trans. Cogn. Commun. Netw.*, 2022.

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

[28] G. Zhu, Y. Du, D. Gündüz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2120–2135, 2020.

[29] D. Wen, G. Zhu, and K. Huang, "Reduced-dimension design of mimo over-the-air computing for data aggregation in clustered iot networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5255–5268, 2019.

[30] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication-efficient federated learning," *Proceedings of the National Academy of Sciences*, vol. 118, no. 17, 2021.

[31] M. S. H. Abad, E. Ozfatura, D. Gunduz, and O. Ercetin, "Hierarchical federated learning across heterogeneous cellular networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 8866–8870.

[32] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, Dublin, Ireland, Jun. 2020, pp. 1–6.

[33] ——, "Hierarchical quantized federated learning: Convergence analysis and system design," *arXiv preprint arXiv:2103.14272*, 2021.

[34] J. Wang, S. Wang, R.-R. Chen, and M. Ji, "Local averaging helps: Hierarchical federated learning and convergence analysis," *arXiv preprint arXiv:2010.12998*, 2020.

[35] Y. LeCun, "The MNIST database of handwritten digits," *http://yann.lecun.com/exdb/mnist/*, 1998.

[36] A. Krizhevsky *et al.*, "Learning multiple layers of features from tiny images," 2009.
[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.