ON FEDERATED LEARNING OVER WIRELESS CHANNELS WITH OVER-THE-AIR AGGREGATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

By
Ozan Aygün
July 2022

ON FEDERATED LEARNING OVER WIRELESS CHANNELS WITH OVER-THE-AIR AGGREGATION

By Ozan Aygün

July 2022

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Tolga Mete Duman (Advisor)

Orhan Arıkan

Elif Tuğçe Ceran Arslan

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

ON FEDERATED LEARNING OVER WIRELESS CHANNELS WITH OVER-THE-AIR AGGREGATION

Ozan Aygün
M.S. in Electrical and Electronics Engineering
Advisor: Tolga Mete Duman
July 2022

A decentralized machine learning (ML) approach called federated learning (FL) has recently been at the center of attention since it secures edge users' data and decreases communication costs. In FL, a parameter server (PS), which keeps track of the global model orchestrates local training and global model aggregation across a set of mobile users (MUs). While there exist studies on FL over wireless channels, its performance on practical wireless communication scenarios has not been investigated very well. With this motivation, this thesis considers wireless FL schemes that use realistic channel models, and analyze the impact of different wireless channel effects.

In the first part of the thesis, we study hierarchical federated learning (HFL) where intermediate servers (ISs) are utilized to make the server-side closer to the MUs. Clustering approach is used where MUs are assigned to ISs to perform multiple cluster aggregations before the global aggregation. We first analyze the performance of a partially wireless approach where the MUs send their gradients through a channel with path-loss and fading using over-the-air (OTA) aggregation. We assume that there is no inter-cluster interference and the gradients from the ISs to the PS are sent error-free. We show through numerical and experimental analysis that our proposed algorithm offers a faster convergence and lower power consumption compared to the standard FL with OTA aggregation. As an extension, we also examine a fully-wireless HFL setup where both the MUs and ISs send their gradients through OTA aggregation, taking into account the effect of inter-cluster interference. Our numerical and experimental results reveal that utilizing ISs results in a faster convergence and a better performance than the OTA FL without any IS while using less transmit power. It is also shown that the best choice of cluster aggregations depends on the data distribution among

the MUs and the clusters.

In the second part of the thesis, we study FL with energy harvesting MUs with stochastic energy arrivals. In every global iteration, the MUs with enough energy in their batteries perform local SGD iterations, and transmit their gradients using OTA aggregation. Before sending the gradients to the PS, the gradients are scaled with respect to the idle time and data cardinality of each MU, through a cooldown multiplier, to amplify the importance of the MUs that send less frequent local updates. We provide a convergence analysis of the proposed setup, and validate our results with numerical and neural network simulations under different energy arrival profiles. The results show that the OTA FL with energy harvesting devices performs slightly worse than the OTA FL without any energy restrictions, and that utilizing the excess energy for more local SGD iterations gives a better convergence rate than simply increasing the transmit power.

Keywords: Distributed machine learning, federated learning, wireless channels, path-loss, fading channels, energy harvesting communications.

ÖZET

KABLOSUZ KANALLAR ÜZERİNDEN HAVADA BİRLEŞTİRME İLE FEDERE ÖĞRENME

Ozan Aygün
Elektrik ve Elektronik Mühendisliği, Yüksek Lisans
Tez Danışmanı: Tolga Mete Duman
Temmuz 2022

Federe öğrenme (FL) adı verilen merkezi olmayan makine öğrenimi (ML) yaklaşımı, kullanıcıların verilerini güvence altına aldığı ve iletişim maliyetlerini azalttığı için son zamanlarda ilgi odağı olmuştur. FL'de, global modelin kaydını tutan bir parametre sunucusu (PS), bir dizi mobil kullanıcı (MU) arasında yerel eğitimi ve global model toplamasını düzenler. Kablosuz kanallar üzerinden FL ile ilgili çalışmalar mevcut olmakla birlikte, pratik kablosuz iletişim senaryolarındaki performansı çok iyi araştırılmamıştır. Bu motivasyonla, bu tez, gerçekçi kanal modelleri kullanan ve farklı kablosuz kanal etkilerinin etkisini analiz eden kablosuz FL şemalarını ele almaktadır.

Tezin ilk bölümünde, sunucu tarafını MU'lara daha yakın hale getirmek için ara sunucuların (IS'ler) kullanıldığı hiyerarşik federe öğrenmeyi (HFL) inceliyoruz. Kümeleme yaklaşımı, küresel toplamadan önce çoklu küme toplamaları gerçekleştirmek için MU'ların IS'lere atandığı durumlarda kullanılır. İlk olarak, MU'ların gradyanlarını havadan (OTA) toplama kullanarak yol kaybı ve sönümlemeli bir kanal üzerinden gönderdiği kısmen kablosuz bir yaklaşımın performansını analiz ediyoruz. Kümeler arası girişim olmadığını ve IS'lerden PS'ye olan gradyanların hatasız gönderildiğini varsayıyoruz. Önerilen algoritmamızın OTA toplamalı standart FL'ye kıyasla daha hızlı yakınsama ve daha düşük güç tüketimi sunduğunu sayısal ve deneysel analizlerle gösteriyoruz. Bir uzantı olarak, hem MU'ların hem de IS'lerin gradyanlarını kümeler arası girişimin etkisini dikkate alarak OTA toplama yoluyla gönderdiği tamamen kablosuz bir HFL kurulumunu da inceliyoruz. Sayısal ve deneysel sonuçlarımız, IS'lerin kullanılmasının, daha az iletim gücü kullanırken herhangi bir IS'siz OTA FL'den daha hızlı bir yakınsama ve daha iyi bir performans ile sonuçlandığını ortaya koymaktadır. Ayrıca, küme kümelemelerinin en iyi seçiminin, MU'lar ve kümeler

arasındaki veri dağılımına bağlı olduğu da gösterilmiştir.

Tezin ikinci bölümünde, stokastik enerji gelişleri ile enerji hasadı MU'ları ile FL'yi inceliyoruz. Her global yinelemede, pillerinde yeterli enerjiye sahip MU'lar yerel SGD yinelemelerini gerçekleştirir ve gradyanlarını OTA toplamasını kullanarak iletir. Gradyanları PS'ye göndermeden önce, daha az sıklıkta yerel güncellemeler gönderen MU'ların önemini artırmak için gradyanlar, bir bekleme süresi çarpanı aracılığıyla her bir MU'nun boşta kalma süresi ve veri kardinalitesine göre ölçeklendirilir. Önerilen kurulumun yakınsama analizini sağlarız ve sonuçlarımızı farklı enerji varış profilleri altında sayısal ve sinir ağı simülasyonları ile doğrularız. Sonuçlar, enerji toplama cihazlarına sahip OTA FL'nin herhangi bir enerji kısıtlaması olmaksızın OTA FL'den biraz daha kötü performans gösterdiğini ve fazla enerjiyi daha fazla yerel SGD yinelemeleri için kullanmanın, yalnızca iletim gücünü artırmaktan daha iyi bir yakınsama oranı sağladığını göstermektedir.

Anahtar sözcükler: Dağıtılmış makine öğrenmesi, federe öğrenme, kablosuz kanallar, yol kaybı, sönümleme kanalları, enerji hasatı ile haberleşme.

Acknowledgement

First and foremost, I would like to express my gratitude to my advisor Prof. Tolga M. Duman for his invaluable support, motivation, and patience throughout my M.S. study. I feel very lucky to work with an extremely knowledgeable mentor, and it would not be possible to conduct research without him.

I would like to thank Prof. Orhan Arıkan and Asst. Prof. Elif Tuğçe Ceran for their time and comments.

This work was supported by Turkcell A.Ş. through BTK 5G VATS graduate research fellowship program and partially supported by TUBITAK through CHIST-ERA project SONATA (CHIST-ERA-20-SICT-004) funded by TUBITAK, Turkey Grant 221N366, and I sincerely acknowledge the financial support.

I feel lucky having İlker Demirel, Emre Mumcuoğlu, and Mohammad Kazemi around me. It has been a pleasure brainstorming ideas and having conversations with them.

I would also like to thank the members of the Bilkent Communication Theory and Application Research (CTAR) Lab, Javad Haghighat, Mohammad Kazemi, Mücahit Gümüş, Mert Özateş, Büşra Tegin, Mohammad Javad Ahmadi, Berke Eren, Sadra Charandabi, Muhammad Atif Ali, and Uras Kargı.

Last but not least, I would like to express my heartfelt gratitude to my family for their unconditional support and encouragement.

Contents

1 Introduction **1**

1.1 Overview 1

1.2 Contributions of the Thesis 2

1.3 Thesis Outline 4

2 Preliminaries and Literature Review **5**

2.1 Federated Learning 6

2.1.1 System Model 7

2.1.2 Literature Review on Federated Learning 9

2.2 Wireless FL 11

2.2.1 System Model 12

2.2.2 Literature Review on Wireless FL 15

2.3 Hierarchical FL 18

2.3.1 System Model 18

2.3.2 Literature Review on HFL 20

2.4 Federated Learning with Energy Harvesting Devices 22

2.5 Chapter Summary 24

3 Hierarchical Over-the-Air Federated Edge Learning 25

3.1 System Model 26

3.1.1 Ideal Communication Scenario 26

3.1.2 OTA Communication Scenario 26

3.2 Convergence Analysis 29

3.3 Numerical Examples 34

3.4 Chapter Summary 37

4 Over-the-Air Federated Edge Learning with Hierarchical Clustering 38

4.1 System Model 39

4.1.1 Cluster Aggregation 39

4.1.2 Global Aggregation 41

4.2 Convergence Analysis 44

4.3 Numerical Examples 47

4.4 Chapter Summary 52

5 Over-the-Air Federated Learning with Energy Harvesting Devices 54

5.1 System Model 55

5.1.1 System Model for Energy Harvesting Devices with Unit Battery 55

5.1.2 System Model for Energy Harvesting Devices with Discrete Battery 58

5.2 Convergence Analysis 61

5.2.1 Convergence Analysis for OTA FL with Energy Harvesting Devices with Unit Battery 61

5.2.2 Convergence Analysis for OTA FL with Energy Harvesting Devices with Discrete Battery 66

5.3 Numerical Examples 70

5.3.1 Unit Battery Case 70

5.3.2 The Case With with Discrete Battery 72

5.4 Chapter Summary 74

6 Conclusions and Future Directions 76

A Proof of Lemma 5 & 8 87

B Proof of Lemma 7 92

List of Figures

2.1 Illustration of a traditional ML system model. 6

2.2 FL system model. 7

2.3 HFL system model. 19

3.2 HOTAFL Test accuracy for non-i.i.d. MNIST data with $\tau = 3$. . . 35

3.1 HOTAFL Test accuracy for i.i.d. MNIST data with $\tau = 1$. 35

3.3 HOTAFL Test accuracy for i.i.d. CIFAR-10 data with $\tau = 5$. . . . 36

3.4 Convergence rate for i.i.d. MNIST data with $\tau = 1$. 37

4.1 W-HFL Test accuracy for i.i.d. MNIST data with $\tau = 1$. 49

4.2 W-HFL Test accuracy for user non-i.i.d. MNIST data with $\tau = 3$. 50

4.3 W-HFL Test accuracy for cluster non-i.i.d. with $\tau = 1$. 50

4.4 W-HFL Test accuracy for i.i.d. CIFAR-10 data with $\tau = 5$. 51

4.5 Convergence rate for Figure 4.1. 52

5.1 Energy harvesting OTA FL test accuracy for $\tau = 1$ 71

5.2 Energy harvesting OTA FL test accuracy for $\tau = 3$ 72

5.3 Upper bound on $\mathbb{E}\left[F(\boldsymbol{\theta}_{PS}(t)) - F^*\right]$ 73

5.4 Upper bound on $\mathbb{E}\left[F(\boldsymbol{\theta}_{PS}(t)) - F^*\right]$ 74

List of Tables

4.1 CNN Architecture for CIFAR-10 Training 48

List of Abbreviations

AWGN Additive white gaussian noise

CSI Channel state information

FL Federated learning

HFL Hierarchical federated learning

HOTAFL Hierarchical Over-the-Air Federated Edge Learning

i.i.d. Independent and identically distributed

IS Intermediate server

ML Machine learning

MU Mobile user

OTA Over-the-air

PS Parameter server

SGD Stochastic gradient descent

W-HFL Wireless hierarchical federated learning

Chapter 1

Introduction

1.1 Overview

Extensive amounts of collected data from various devices such as mobile phones and Internet-of-things (IoT) sensors have enabled the accelerating rise of machine learning (ML) algorithms. Traditionally, ML algorithms require all the data to be collected at a cloud server for model training, which raises concerns regarding privacy, cost, and latency. Firstly, data owners may be sensitive about sharing their personal data; secondly, the increasing quality and volume of collected data results in higher communication costs; and finally, solutions that work in real-time are faced with latency issues. To overcome these problems, a decentralized approach called *federated learning* (FL) has been introduced, where models are trained locally instead of using a centralized server.

In FL, several data owners, called mobile users (MUs), are selected at each iteration based on some criteria such as their computing capability, available power, and location. The parameter server (PS) sends the current global model to the selected MUs. Each of these MUs trains a local model by carrying out multiple stochastic gradient descent (SGD) iterations using its own data and computing power. Then, each MU sends only the weight updates to the PS,

which performs model aggregation to update the global model. These steps are repeated until a convergence criterion is met.

Despite its superiority over traditional ML, adverse channel effects in wireless setups and increased communication costs pose challenges for the feasibility of conventional FL in practical scenarios. To address the communication cost concerns, over-the-air (OTA) aggregation has become a popular method thanks to its efficient strategy that allocates all the MUs to the same bandwidth, thereby handling the transmission and aggregation of the gradient updates simultaneously (over the air). For this framework, one approach to deal with the channel effects (particularly when there is no transmit side channel state information) is to increase the number of receive antennas at the PS. Nevertheless, the disparity among the channel gains is still a critical concern, e.g., when some MUs are far away from the PS, as this would introduce bias across the updates.

1.2 Contributions of the Thesis

In this thesis, we investigate the performance of OTA FL while taking wireless channel effects into account. Specifically, we study on the convergence rate and the performance of different OTA FL schemes.

We first propose a hierarchical federated learning (HFL) with OTA aggregation, where intermediate servers (ISs) are employed in areas where the number of MUs is high to form cluster-like structures. In the proposed approach, MUs carry out multiple SGD iterations before transmitting their model differences to their corresponding IS using OTA aggregation. After several cluster aggregation steps between ISs and their corresponding MUs, global aggregation is made at the PS, using the IS cluster updates. We provide a theoretical analysis on the convergence rate of the proposed algorithm with different number of cluster aggregations to show that a hierarchical structure gives a faster convergence than the conventional FL with OTA aggregation which does not employ any intermediate servers.

We then extend our study on OTA HFL to a fully-wireless case where both the cluster and global aggregations are performed using OTA computing. Moreover, we remove the assumption that the interference in a cluster aggregation comes only from the MUs inside the cluster, i.e., we include the inter-cluster interference due to all the MUs. We present a detailed analysis and study the performance of the proposed approach with different data distributions and datasets.

In the last part of the thesis, we focus on OTA FL with energy harvesting MUs. In this case, the MUs collect the required energy for both local computations and transmissions from their ambient environment. We start our analysis with energy harvesting MUs equipped with unit batteries where the energy arrivals are based on a Bernoulli process or a uniform process with a pre-defined length. We conduct a theoretical analysis of energy harvesting OTA FL with Bernoulli energy arrivals and compare it with the performance of OTA FL with full participation. We also extend the energy harvesting OTA FL approach to a case where the MUs can have finite battery levels, and the energy is used both for local computations and gradient transmissions. We propose two different power consumption policies where the excess energy is either used to increase the number of local SGDs, or to increase the transmit power for the OTA aggregation. We show through a theoretical analysis that utilizing the excess energy for additional local computations gives a faster convergence rate than utilizing it to increase the transmission power.

Our results in this thesis are reported in two conference papers (one published, one accepted), and an additional manuscript submitted for publication:

- O. Aygün, M. Kazemi, D. Gündüz, T. M. Duman, "Hierarchical over-the-air federated edge learning" in *IEEE Int. Conf. Commun. (ICC)*, Seoul, South Korea, May 2022. [1]
- O. Aygün, M. Kazemi, D. Gündüz, T. M. Duman "Over-the-air federated learning with energy harvesting devices" *IEEE Glob. Telecommun. Conf. (GLOBECOM)*, Rio de Janeiro, Brazil, Dec. 2022. [2]
- O. Aygün, M. Kazemi, D. Gündüz, T. M. Duman "Over-the-Air Federated

Edge Learning with Hierarchical Clustering" *IEEE Trans. Wireless Commun.*, 2022 (under review). Available: https://arxiv.org/pdf/2207.09232.pdf [3]

1.3 Thesis Outline

In Chapter 2, we provide the motivation behind OTA FL by firstly explaining the fundamentals regarding FL with its literature review, then introducing the more practical approaches such as the communication model of OTA FL and its hierarchical version. In Chapter 3, we explain the design and the system model of our basic HFL setup with OTA communications in cluster aggregation, and error-free transmission for the global aggregation. We also provide its convergence analysis and experimental results with different datasets and data distributions. In Chapter 4, we propose an extended version of the previous algorithm to a more realistic model where OTA aggregation is used in both cluster and global aggregation steps with inter-cluster interference included in the cluster aggregation stage. Similar to before, we also provide the convergence analysis and experimental results to verify our work and compare with the other schemes. In Chapter 5, we work on OTA FL using energy harvesting devices, and provide an algorithm for MUs who have unit-sized battery with either deterministic or stochastic energy arrivals. Moreover, we provide a convergence analysis for the MUs with stochastic energy arrivals and report its experimental results with different setups. We also work on an extension of energy harvesting OTA FL where the MUs have discrete energy arrivals between two consecutive global iterations based on a Poisson process. We take into account the energy costs of local SGD computations and gradient transmissions separately, and propose different energy consumption profiles where the excess energy is used either for increasing the number of local computations or increasing the transmit power, and provide theoretical results to compare different scenarios. Finally, in Chapter 6, we conclude our work and provide possible future research directions.

Chapter 2

Preliminaries and Literature Review

In this chapter, we provide the basic preliminaries and a literature review required for the material in this thesis. Firstly, federated learning is introduced in detail. Secondly, wireless FL is explained to set the framework for the subsequent chapters. Then, hierarchical FL schemes are discussed as background on Chapters 3 and 4. Finally, energy harvesting devices with different capabilities are introduced as preliminaries for the proposed solutions in Chapter 5.

The chapter is organized as follows. In Section 2.1, federated learning scheme with error-free transmission is presented. In Section 2.2, wireless FL schemes including OTA aggregation is explained. In Section 2.3, hierarchical FL approach is described. In Section 2.4, preliminaries regarding energy harvesting devices with limited battery are given. The chapter is concluded in Section 2.5.

2.1 Federated Learning

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 at a server to train the input data and give a representative model. An example of the traditional ML model is given in Figure 2.1 where the participating users send their data to a cloud server to obtain a global model that represents all the users' data. The server shares the resulting model with the users after the training.

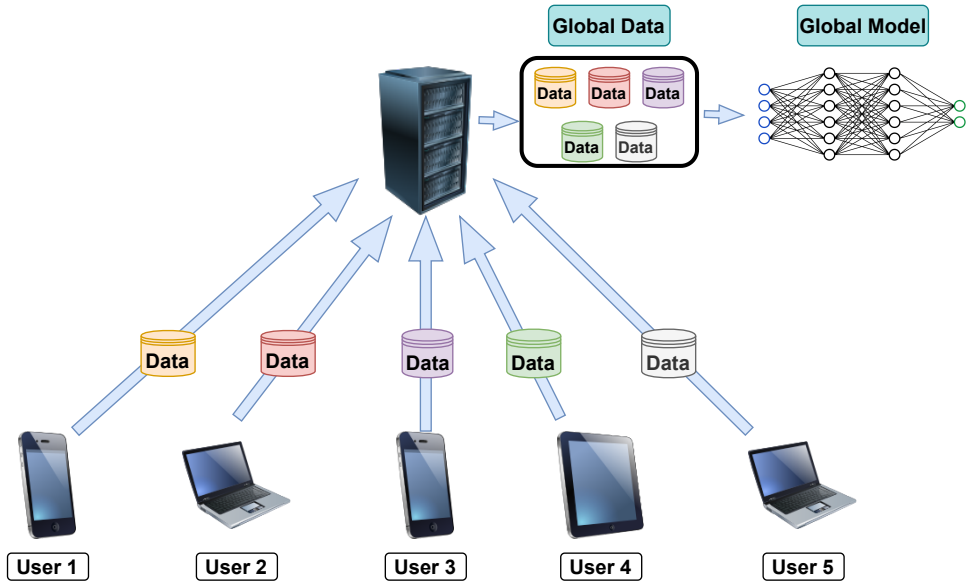

Figure 2.1: Illustration of a traditional ML system model.

Recent work on ML has shown that there may be concerns about data privacy, communication costs, and latency. Firstly, users may be reluctant to share their data since they want to ensure that their sensitive information is safe. Secondly, the sensor devices produce data with a much higher quality than ever before, which requires more time or higher rates for transmission. Thirdly, applications that need to operate in real-time might be affected by the latency since their performance depends on the model response of the simultaneously collected data. To overcome these problems, a decentralized approach called *federated learning* (FL) has been introduced, where models are trained locally instead of using a centralized server.

In FL, several data owners, called mobile users (MUs), are selected at each iteration based on some criteria such as their computing capability, available power, and location. The parameter server (PS) sends the current global model to the selected MUs. Each of these MUs trains a local model by carrying out multiple stochastic gradient descent (SGD) iterations using its own data and computing power. Then, each MU sends only the weight updates to the PS, which performs model aggregation to update the global model. These steps are repeated until a convergence criterion is met.

2.1.1 System Model

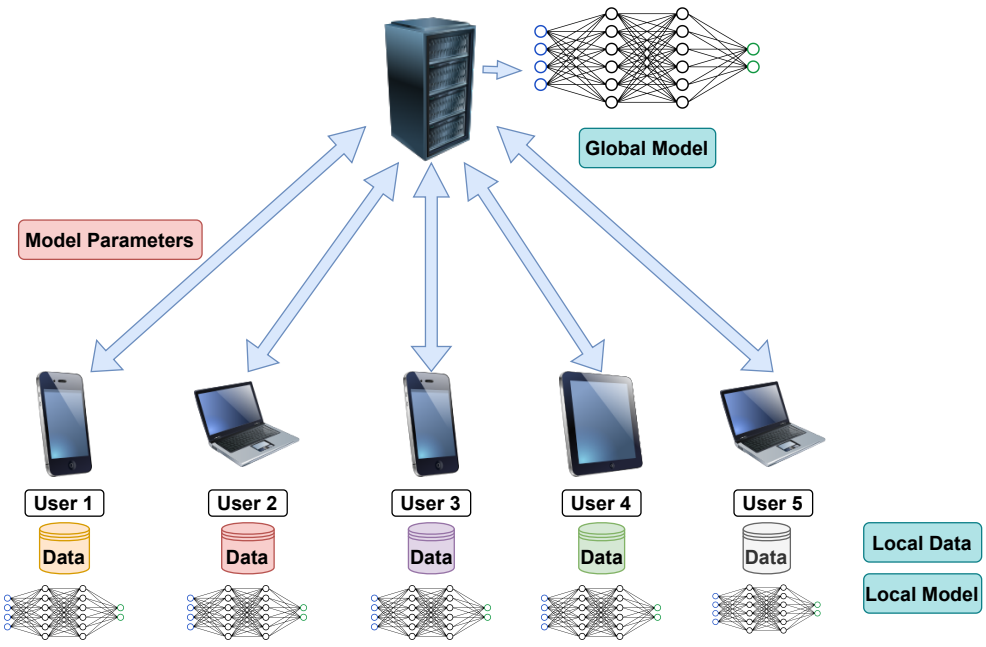

Figure 2.2: FL system model.

In [4], McMahan et. al. introduced federated learning (FL) where model training with the user data can be done without collecting them at a centralized server. The objective of FL 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. The system consists of M MUs and a PS as depicted in Fig. 2.2. The dataset of the m-th MU is

denoted as $\mathcal{B}_m$, and we define $B \triangleq \sum_{m=1}^{M} |\mathcal{B}_m|$. We have

$$F(\boldsymbol{\theta}) = \sum_{m=1}^{M} \frac{|\mathcal{B}_m|}{B} F_m(\boldsymbol{\theta}), \tag{2.1}$$

where $F_m(\boldsymbol{\theta}) \triangleq \frac{1}{|\mathcal{B}_m|} \sum_{u \in \mathcal{B}_m} f(\boldsymbol{\theta}, u)$, with $f(\boldsymbol{\theta}, u)$ denoting the corresponding loss of u-th data sample. In every global iteration, the MUs carry out τ user iterations on their own, and then send the model updates to the corresponding PSs for the global aggregation. At the j-th user iteration, the weight update is performed employing stochastic gradient descent (SGD) for the m-th MU is as follows

$$\boldsymbol{\theta}_m^{j+1}(t) = \boldsymbol{\theta}_m^{j}(t) - \eta_m^j(t) \nabla F_m(\boldsymbol{\theta}_m^j(t), \boldsymbol{\xi}_m^j(t)), \tag{2.2}$$

where $\eta_m^j(t)$ is the learning rate, $\nabla F_m(\boldsymbol{\theta}_m^j(t), \boldsymbol{\xi}_m^j(t))$ denotes the stochastic gradient estimate for the weight vector $\boldsymbol{\theta}_m^j(t)$ and a randomly sampled batch of data samples $\xi_m^j(t)$ from the dataset of the m-th MU at the t-th global and j-th user iteration. The PS performs model aggregation using the local updates to obtain an updated global model and sends the new model back to the MUs for the next iteration. The first proposed aggregation method is called *FedAvg*, proposed in [5], which finds the global model for the next global iteration as

$$\boldsymbol{\theta}_{PS}(t+1) = \sum_{m=1}^{M} \frac{|\mathcal{B}_m|}{B} \boldsymbol{\theta}_m^{\tau}(t). \tag{2.3}$$

The steps given in (2.2) and (2.3) are repeated until a desirable accuracy is obtained.

Alternatively, one can also send local gradients to the parameter server for convenience. That way, the global model is not lost, instead, it is updated in every global iteration. In this case, after performing τ local SGD iterations, each MU calculates its model difference to be sent to the PS as

$$\Delta\boldsymbol{\theta}_m(t) = \boldsymbol{\theta}_m^{\tau+1}(t) - \boldsymbol{\theta}_m^{1}(t). \tag{2.4}$$

The global update rule is

$$\Delta\boldsymbol{\theta}_{PS}(t) = \sum_{m=1}^{M} \frac{|\mathcal{B}_m|}{B} \Delta\boldsymbol{\theta}_{PS,c}(t). \tag{2.5}$$

After the global aggregation, the model at the PS is updated as

$$\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\boldsymbol{\theta}_{PS}(t). \tag{2.6}$$

Throughout the thesis, we will consider the FL notion for which the gradients are sent to PS.

2.1.2 Literature Review on Federated Learning

FL has gained a significant attention because of its privacy preserving nature, and utilizing computation at the edge to address the latency problem while offloading the data to the server [4]. One of the first FL algorithms that is proposed is called *FedAvg*, where the global aggregation is performed via a simple averaging of the local weights received from the MUs [5]. Since not all the MUs can have equal amount of data samples, a weighted averaging operation with respect to the data cardinality is used in order not to bias the system towards the MUs with more data samples. Therefore, the authors in [5] have added a weight depending on the number of data samples each MU has, and normalizing those weights with respect to the total number of data samples contributing to the aggregation.

In practical applications, not all the MUs have the same amount of data diversity, i.e., some MUs might only have data samples with only a small portion of all the available labels, which creates a non-independent and identically distributed data distribution scheme across the MUs. To overcome the data distribution issue and reduce the bias toward some labels, the authors in [6] have proposed to use the data distribution of MUs to weigh them based on how independent and identically distributed (i.i.d.) their data is when compared to the other MUs using the earth mover's distance [7], and achieved a performance increase when the MUs have non-i.i.d. data distribution. If an MU's data distribution is too skewed when compared to the others, a small percentage of the data sharing is requested from the MU to the PS to support the global aggregation process. They show that with 5% data sharing, a performance increase can be observed with both i.i.d. and non-i.i.d. data distributions. Similar to this scheme, [8] introduces another method to compare the variance of the local models among different MUs using

maximum mean discrepancy (MMD) to decrease the bias towards the MUs that have skewed data distributions [9]. Using the global model as a reference during their local training process, they aim to minimize the MMD loss between the local and global models. There are also studies supporting that more local SGD iterations at the edge before the global aggregation can give higher accuracies and achieve faster convergence when compared to the case where only a single SGD iteration is performed, but at the risk of being stuck at the local minima values in non-i.i.d. settings, see, for instance, [6].

Local gradients calculated at the MUs can have different mean and variance values, depending on the neural network architecture employed. Using the observation that most parameter values are sparsely distributed and close to zero, another approach called edge stochastic gradient descent (eSGD) is proposed to reduce the amount redundant gradient transmissions [10]. Only a portion of the gradients are sent, and the algorithm determines the importance of current training gradients based on improvements in loss value. Positive hidden weights are given to important parameters. Small gradient values are accumulated as residual values and once their residuals reach a threshold, they are chosen to replace the least important hidden weights. To reduce the communication costs, [11] proposes to assign relevancy scores to the updates from each MU. The local update is performed only if it is above a predefined relevancy score threshold. They show that while guaranteeing its convergence, the conventional FL accuracies can be acheved with significantly less number of communication rounds.

One of the main reasons that FL is chosen over traditional ML is that it protects the user data, i.e., the local data do not need to be offloaded to a cloud server. Despite its advantages in terms of privacy, recent studies show that there can be concerns regarding the privacy and security of the MUs, especially when there are malicious participants among the MUs whose intentions are to infer the user data from the shared local models [12], or to provide the PS with false information [13].

One of the approaches that can jeopardize the privacy of the system rely on the fact that local models represent the local data, and an adversary user

can regenerate a victim user's data by obtaining the representative model and utilizing generative models such as generative adversarial networks (GANs) [14]. By training two different networks for both generating the data and discriminating between fake and real data, it is possible to mimic the data of target users by using their neural network model. Authors in [15] explore this idea to introduce malicious participants into the system in order to infer the unshared data from the local model. Results show that it is possible to generate the local data with a high accuracy. A possible approach to increase the user privacy is to use differential privacy techniques [16]. In [17], the authors propose to add some "noise" to the local updates before sending them to the PS in order to increase the privacy of the system. As a more established approach, encryption can also be used to increase the data privacy [18].

Security is another important issue that needs to be studied at in order to preserve the reliability of the FL system. Even though a malicious participant does not give any privacy issues, it can still disturb the system by sending false information, or no information at all. One approach that needs to be dealt with is the "data poisoning attacks", where a malicious participant sends dirty-label data in order to confuse the PS [19]. It is possible to almost completely falsify the PS even with a small number of samples when compared to the total amount of data, so one needs to be aware of the possibility that the incoming labels might not be correct. Another way of exploiting this issue is to perform "model poisoning attack" where the sent model is entirely false, and its only purpose is to misguide the PS into a wrong direction [20]. It is shown that model poisoning attacks are much more effective than the data poisoning attacks since they directly modify the global model [21].

2.2 Wireless FL

Since the edge users are usually far away from the PS, the gradients need to be sent through wireless channels. However, one needs to be aware that the amount of required bandwidth increases linearly with the number of participating

users. Therefore, the required bandwidth becomes extremely high in settings that include a large number of MUs. To address the communication cost concerns, over-the-air (OTA) aggregation [22] has become a popular method thanks to its efficient strategy that allocates all the MUs to the same bandwidth, thereby handling the transmission and aggregation of the gradient updates simultaneously (over the air). Even though we lose the values of individual gradient values coming from different MUs, the PS does not need the individual values since they will be added up for the global aggregation.

Despite its benefits over traditional ML, adverse channel effects in wireless setups pose challenges for the feasibility of conventional FL in practical scenarios. Because of the interference effects when OTA aggregation is used, the performance of wireless FL with a single antenna becomes poor. Adverse effects of the wireless channel can be reduced by increasing the number of receive antennas at the PS. In [23], authors show that the interference and noise terms can be alleviated as the number of receive antennas increases. They show that even when the tranmitter side has no channel state information (CSI), and the receiver has imperfect CSI, a good performance can still be achieved.

2.2.1 System Model

We now consider the scheme referred as wireless FL that uses OTA aggregation, for which the links between the MUs and the PS are modeled as wireless channels. Since a common wireless medium is used in local aggregations, noisy versions of the model updates $\Delta\boldsymbol{\theta}_{PS}(t)$ are received at the PS. In our setup, the PS is equipped with K antennas, and we assume perfect CSI at the receivers and no CSI at the MUs.

In wireless FL, in order to increase the spectral efficiency, the model differences are grouped to form a complex vector $\Delta\boldsymbol{\theta}_m^{cx}(t) \in \mathbb{C}^N$ with the following real and

imaginary parts

$$\Delta\boldsymbol{\theta}_m^{re}(t) \triangleq \left[\Delta\theta_m^1(t), \Delta\theta_m^2(t), \ldots, \Delta\theta_m^N(t)\right]^T, \tag{2.7a}$$

$$\Delta\boldsymbol{\theta}_m^{im}(t) \triangleq \left[\Delta\theta_m^{N+1}(t), \Delta\theta_m^{N+2}(t), \ldots, \Delta\theta_m^{2N}(t)\right]^T. \tag{2.7b}$$

The received signal at the PS for the k-th antenna at the t-th global iteration can be represented as

$$\boldsymbol{y}_{PS,k}(t) = P_t \sum_{m=1}^{M} \boldsymbol{h}_{m,k}(t) \circ \Delta\boldsymbol{\theta}_m^{cx}(t) + \boldsymbol{z}_{PS,k}(t), \tag{2.8}$$

where P_t is a multiplier at the t-th global iteration used to adjust the average transmitted power, "$\circ$" denotes the element-wise product, $\boldsymbol{z}_{PS,k}(t) \in \mathbb{C}^N$ with i.i.d. entries $z_{PS,k}^n(t) \sim \mathcal{CN}(0, \sigma_z^2)$. The channel coefficients are modelled as

$$\boldsymbol{h}_{m,k}(t) = \sqrt{\beta_m}\, \boldsymbol{g}_{m,k}(t), \tag{2.9}$$

where $\boldsymbol{g}_{m,k}(t) \in \mathbb{C}^N$ with entries $g_{m,k}^n(t) \sim \mathcal{CN}(0, \sigma_h^2)$ (i.e., Rayleigh fading), β_m is the large-scale fading coefficient modeled as $\beta_m = (d_m)^{-p}$, where p represents the path loss exponent, and d_m denotes the distance between the m-th MU and the PS.

Knowing the CSI perfectly, the PS combines the received signals as

$$\boldsymbol{y}_{PS}(t) = \frac{1}{K} \sum_{k=1}^{K} \Big(\sum_{m=1}^{M} \boldsymbol{h}_{m,k}(t) \Big)^* \circ \boldsymbol{y}_{PS,k}(t). \tag{2.10}$$

For the n-th symbol, the combined signal can be written as

$$y_{PS}^n(t) = \frac{1}{K}\sum_{k=1}^{K}\Big(\sum_{m=1}^{M} h_{m,k}^n(t)\Big)^* y_{PS,k}^n(t) \tag{2.11}$$

$$\begin{aligned} &= \underbrace{P_t \sum_{m=1}^{M}\Big(\frac{1}{K}\sum_{k=1}^{K}|h_{m,k}^n(t)|^2\Big)\Delta\theta_m^{n,cx}(t)}_{y_{PS}^{n,sig}(t)\text{ (signal term)}} \\ &\quad + \underbrace{\frac{P_t}{K}\sum_{m=1}^{M}\sum_{\substack{m'=1\\ m'\neq m}}^{M}\sum_{k=1}^{K}(h_{m,k}^n(t))^* h_{m',k}^n(t)\Delta\theta_{m'}^{n,cx}(t)}_{y_{PS}^{n,itf}(t)\text{ (interference term)}} \\ &\quad + \underbrace{\frac{1}{K}\sum_{m=1}^{M}\sum_{k=1}^{K}(h_{m,k}^n(t))^* z_{PS,k}^n(t)}_{y_{PS}^{n,no}(t)\text{ (noise term)}}. \end{aligned} \tag{2.12}$$

Aggregated model differences can be recovered by

$$\Delta\hat{\theta}_{PS}^n(t) = \frac{1}{P_t M\sigma_h^2\bar{\beta}}\,\mathrm{Re}\{y_{PS}^n(t)\}, \tag{2.13a}$$

$$\Delta\hat{\theta}_{PS}^{n+N}(t) = \frac{1}{P_t M\sigma_h^2\bar{\beta}}\,\mathrm{Im}\{y_{PS}^n(t)\}, \tag{2.13b}$$

where $\bar{\beta} = \sum_{m=1}^{M}\beta_m$. After estimating the model difference values, the model update is written as

$$\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\hat{\boldsymbol{\theta}}_{PS}(t), \tag{2.14}$$

where $\Delta\hat{\boldsymbol{\theta}}_{PS}(t) = \big[\Delta\hat{\theta}_{PS}^1(t)\ \Delta\hat{\theta}_{PS}^2(t)\ \cdots\ \Delta\hat{\theta}_{PS}^{2N}(t)\big]^T$.

The OTA FL framework above can be implemented using orthogonal frequency-division multiplexing (OFDM) in a practical and efficient manner since an OFDM based implementation makes the user synchronization and over-the-air data aggregation very simple. We also note that even though the ML model dimensions may be large, the amount of time needed for the transmission is practical. As an example, consider a scenario where the available bandwidth is 100 MHz with a subcarrier spacing of 120 kHz. Therefore, there are 8192 subcarriers for each OFDM word. Looking at the previous works on OTA FL, some of

the most demanding neural network architectures have model dimensions around 300000 [23], corresponding to around 150000 complex symbol transmissions. Using this setup, we need around 18 OFDM words to transmit one gradient vector from the MU to the PS. Each OFDM word is of duration as $\frac{1}{120\text{kHz}} \approx 0.01$ ms. In practice, a cyclic prefix will also be needed, but it will not increase the OFDM word length drastically. For 18 OFDM words, we will only need around 0.2 ms, hence we conclude that the OTA FL is applicable to a real setup where OFDM is employed, and the symbol durations are not that high, even in scenarios when the neural network model is complex.

2.2.2 Literature Review on Wireless FL

Bandwidth is a serious issue in wireless communication schemes since it is one of the most scarce resources that needs to be used carefully. In multi-user scenarios, the amount of required bandwidth increases linearly with the number of participants. With this motivation, wireless FL has drawn a significant attention, thanks to the OTA aggregation process where the gradients can be sent through the same wireless medium since only their sum is needed and they do not need to be restored separately at the receiver side [24, 25]. By utilizing OTA aggregation, the gradient transmission and aggregation can be handled simultaneously over-the-air, while using the channel bandwidth efficiently. In [26], the authors have introduced this concept as broadband analog aggregation (BAA), compared its performance with FL using orthogonal frequency division multiplexing (OFDM) as well as a device scheduling algorithm based on the distance to the PS to show that similar performances can be obtained while using less bandwidth.

Authors in [22] combine the ideas of OTA aggregation, gradient sparsification, and low rank approximation with digital and analog approaches. In the digital scheme called digital-distributed stochastic gradient descent (D-DSGD), gradients are clipped, where a pre-defined number of largest and smallest gradient values in local models are set to zero, then the means of the positive and negative gradients are calculated, and all the gradient values are equated to the mean of

the positive-valued gradients if it is greater than the mean of the negative-valued gradients, and vice versa. The system also keeps track of the difference between the actual gradients and the sparsified version of the gradients, which will be sent to the PS. Before calculating the sparsified gradients, this accumulated error vector is added on top of the local gradients in order to keep track of the sparsification error. Finally, the gradients are sent digitally to the PS for the global aggregation. In order to demonstrate the performance of the OTA aggregation, authors also propose the analog version of this method, called analog-distributed stochastic gradient descent (A-DSGD), in which similar sparsification and error accumulation approaches are used, but the gradients are sent in an uncoded manner. Before transmitting the gradients, the gradient vector is multiplied with a pseudo-random matrix whose seed is known by both the MUs and the PS, in order to get a low-rank approximation of the local gradients. Then, the final form is transmitted using OTA aggregation and an estimate of the global update is obtained by using an approximate message passing (AMP) algorithm [27]. They derive the convergence rate of the algorithm and show that the proposed A-DSGD algorithm performs better than its digital version.

Inspired from [22], recent works in wireless FL also apply model compression in order to use wireless resources more effectively. In [28], the authors aim to decrease the convergence time while lowering the global loss. First, they apply a probabilistic device selection approach where the heterogeneous devices have different probabilities to participate in the federation. Then, they solve an optimization problem which aims to have the highest number of participating devices possible, while keeping the transmission delay to the PS low. Finally, they apply random lattice quantization in order to reduce the length of the transmitted vector to the PS [29]. They show through simulations that they are able to obtain a faster convergence while keeping the loss similar to the uncompressed cases. Other works on model compression in FL models can also be found in [30–34].

For setups with lots of heterogeneous devices, one needs to be aware that the interference can be high when OTA aggregation is employed, and the required

channel bandwidth becomes extremely high in setups where the orthogonal channel resources are used. Therefore, device scheduling and selection are also important aspects of wireless FL where the subset of devices are adjusted based on their available computing power, energy, distance, data and gradient quality [35, 36]. In [37], a scheduling algorithm is considered where they choose the participating users based on the quality of the local update and wireless channel reliability. They also show the relationship between the scheduling thresholds and their effects on the performance through simulations. Similar to [37], authors in [38] compare the importance of channel awareness and the quality of gradient updates in scheduling, and show the situational uses of different scheduling approaches. There are also other approaches such as those taking the age of the update into account while adjusting the scheduling policy [39], or those choosing the devices in such a way that convergence takes the least amount of time [40].

Another direction in wireless FL is to analyze practical schemes taking into account more realistic wireless channels, similar to [41, 42]. For example, the authors in [37] remove the assumption that the transmission of the global model from the PS to the MUs are ideal, and analyze the convergence rate of the FL system with a noisy downlink. Similarly, in [43], the uplink is modeled as the multiple access fading channel, and the downlink is modeled as a fading broadcast channel, where the transmitted gradients are calculated based on federated distillation [44]. OTA FL with heterogeneous data distribution at MUs is another challenge that is studied in [45] where the authors provide a convergence rate analysis and develop performance guarantees.

In OTA aggregation setups, since the MUs use a shared wireless medium, interference becomes a serious issue that needs to be dealt with [46]. The authors in [23] consider a wireless FL setup where the PS has multiple receive antennas. By using a single-input multiple-output (SIMO) communication setup, and employing combining techniques at the receiver side with no CSI at the transmitters, the authors show that increasing the number of receive antennas at the PS improves the performance, and ideally, the noise and interference terms can be mitigated fully as the number of antennas go infinity. The analyses are also extended to the imperfect CSI scenario at the receiver.

2.3 Hierarchical FL

In FL, not all the users are located in a dense environment, and they can be far away from the PS. The distance between the MUs and the PS increases the communication costs for the wireless setups, and the overall latency. Therefore, a hierarchical FL setup can been used in order to reduce the latency and the communication costs, where intermediate servers (ISs) are employed in areas where the number of MUs is high to form cluster-like structures [47]. The MUs carry out multiple SGD iterations before transmitting their model differences to their corresponding ISs. After several cluster aggregation steps between ISs and their corresponding MUs, global aggregation is made at the PS, using the IS cluster updates.

2.3.1 System Model

The objective HFL is similar to FL where the aim is to minimize a loss function $F(\boldsymbol{\theta})$ with respect to the model weight vector $\boldsymbol{\theta} \in \mathbb{R}^{2N}$. Our system consists of C clusters each containing an IS and M MUs as depicted in Fig. 2.3.

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

$$F(\boldsymbol{\theta}) = \sum_{c=1}^{C} \sum_{m=1}^{M} \frac{|\mathcal{B}_{c,m}|}{B} F_{c,m}(\boldsymbol{\theta}), \tag{2.15}$$

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.

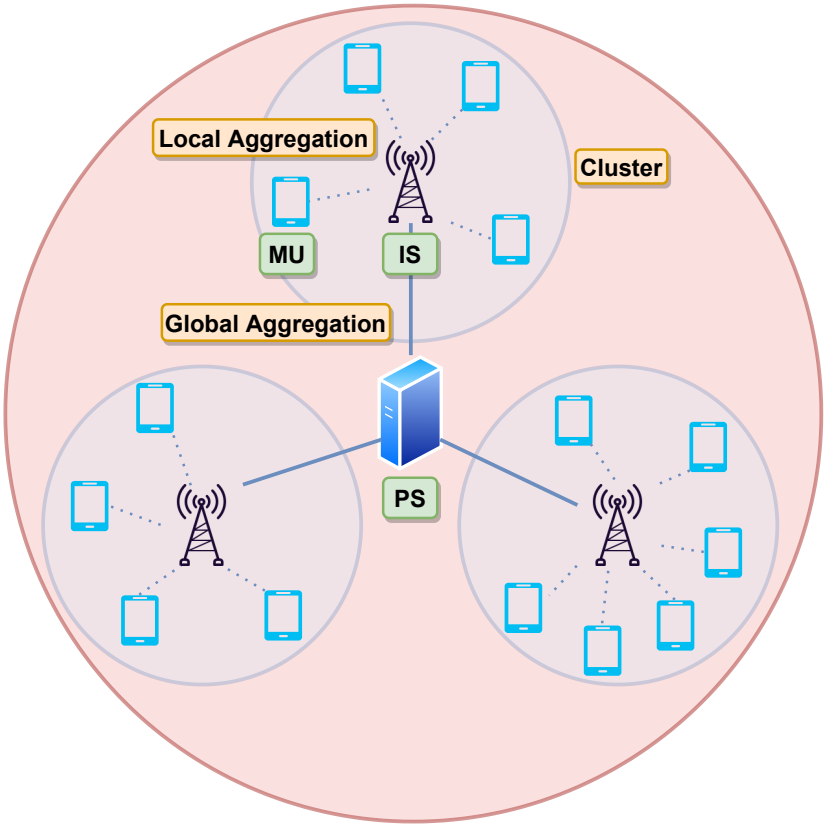

Figure 2.3: HFL system model.

We consider a hierarchical and iterative approach consisting of global, cluster, and user iterations to minimize (2.15). 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:

$$\boldsymbol{\theta}_{c,m}^{i,j+1}(t) = \boldsymbol{\theta}_{c,m}^{i,j}(t) - \eta_{c,m}^{i,j}(t)\nabla F_{c,m}(\boldsymbol{\theta}_{c,m}^{i,j}(t), \boldsymbol{\xi}_{c,m}^{i,j}(t)), \tag{2.16}$$

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 $\boldsymbol{\theta}^i_{IS,c}(t)$ 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 $\nabla F_{c,m}(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t))$ is an unbiased estimator of $\nabla F_{c,m}(\boldsymbol{\theta}^{i,j}_{c,m}(t))$, i.e., $\mathbb{E}_\xi \left[\nabla F_{c,m}(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t))\right] = \nabla F_{c,m}(\boldsymbol{\theta}^{i,j}_{c,m}(t))$, where the expectation is over the randomness due to the SGD.

After τ user iterations, each MU calculates its model difference to be sent to its corresponding IS as

$$\Delta\boldsymbol{\theta}^i_{c,m}(t) = \boldsymbol{\theta}^{i,\tau+1}_{c,m}(t) - \boldsymbol{\theta}^i_{IS,c}(t). \tag{2.17}$$

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

$$\boldsymbol{\theta}^{i+1}_{IS,c}(t) = \boldsymbol{\theta}^i_{IS,c}(t) + \frac{1}{M}\sum_{m=1}^{M} \Delta\boldsymbol{\theta}^i_{c,m}(t). \tag{2.18}$$

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

$$\Delta\boldsymbol{\theta}_{PS,c}(t) = \boldsymbol{\theta}^{I+1}_{IS,c}(t) - \boldsymbol{\theta}_{PS}(t). \tag{2.19}$$

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 conclude that

$$\Delta\boldsymbol{\theta}_{PS}(t) = \frac{1}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I} \Delta\boldsymbol{\theta}^i_{c,m}(t). \tag{2.20}$$

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

2.3.2 Literature Review on HFL

HFL has recently been investigated due to its latency reducing, energy efficient, and more reliable nature. For the ideal communication model, [48] have proposed the clustering of local updates to obtain convergence in less number of global iterations. Their results with different hyperparameter setups show that the

proposed hierarchical scheme outperforms the conventional FL with only MUs and PS. As a more specific approach, [49] focuses on HFL with non-i.i.d. data distribution among the MUs. Since each cluster may have a different set of data samples, and some of them might not be able to have access to some of the labels, the authors investigate which scenarios of HFL with non-i.i.d. data distribution works the best.

The authors in [47] analyze the relationship between uplink/downlink latency and data rates, report the latencies with different cluster aggregation numbers and path loss exponents, and give experimental results with their proposed hierarchical clustering approach. Their results show that HFL has a lower latency thanks to ISs in between the MUs and the PS, thereby requiring less time to converge than conventional FL. Similar to [47], [50] also investigates the system latency of the HFL, but the authors expand the analysis to the optimization of MU scheduling and the energy use at the edge. They propose an algorithm where the MUs constantly check which cluster to join based on the approximate latency and the available energy. Then, the PS determines which MUs should be scheduled at which iteration and which cluster they will be assigned. Their results demonstrate that a better performance with less amount of energy consumption is possible when compared to different conventional FL scenarios with different scheduling solutions and energy settings. Moreover, the authors in [51] present a more detailed scenario where the transmitted gradients are also quantized before sending them. They also conduct an analysis on the latency of the proposed HFL system. All of the above works show the superiority of HFL in terms of latency, energy, and the convergence rate, at the cost of employing multiple ISs between the MUs and the PS.

Most of the HFL works in the literature are based on an IS between MUs and the PS, making it a two-layered system. Different from them, [52] and [53] extend this idea into a multi-level HFL where there are multiple (or adjusted number of) intermediate nodes in the network. They give an analysis based on the optimal number of level of hierarchy in different setups and demonstrate the trade-off between the number of clusters, levels, and MUs in each clusters with different data distributions.

2.4 Federated Learning with Energy Harvesting Devices

Despite the success of FL in practical scenarios, the energy consumption and carbon footprint of MUs for training and sharing their local models create serious concerns about the sustainability of future smart systems [54]. As a more sustainable approach, energy harvesting devices, which can acquire energy from their surroundings [55], have been widely considered for mobile networks. These devices are typically equipped with a rechargeable battery to store the harvested energy, and perform the required computations and communications if they have sufficient energy in their battery.

Energy harvesting devices, also known as energy scavenging devices [56], have been studied extensively in the last decade to build more sustainable communications systems [57]. With the emerging green communications area, the recent works focus on various ways to harvest, for example, solar, electromagnetic, or thermal energy [55]. The limiting factor for energy harvesting devices is that the energy arrivals are intermittent, i.e., the harvested energy usually has an underlying stochastic process that depends on the intensity of the energy source, which may or may not be known by the device itself. In this scenario, devices try to make sure that they send the required message as efficiently as possible [58]. Therefore, one of the most common issues that needs to be deal with in energy harvesting is how to optimally allocate the available energy in the nodes so that the throughput is at a desirable level [59]. Another direction to investigate is to send the message with a maximum rate without running out of available energy [60].

Authors in [61] present a comprehensive analysis on both throughput maximization and transmission completion time with given battery levels and the energy arrival processes for the energy harvesting devices. They present both offline and online optimal policies. The offline policy for the energy usage can be described using a directional water-filling algorithm, while the online policy is given by a dynamic programming algorithm, which assumes the knowledge on

the underlying stochastic processes for the causal energy arrival information and fading variations. The results show that their approach gives a higher throughput with a given average recharge rate of the nodes. Moreover, the authors in [62] consider the similar problem of throughput maximization with energy harvesting devices with a more practical scenario where there are possible energy leakages or other losses during charging/discharging of the battery. They show that it is more beneficial to use the energy without storing it in the battery when the energy leakage rate increases. Packet scheduling among energy harvesting nodes is also a recent topic of study [63], where it is shown that, in an M-user broadcast AWGN channel, the optimal packet transmission strategy can also be found using a directional water-filling algorithm.

Federated learning with energy harvesting devices is a new research area that utilizes energy harvesting nodes as edge users. The users employ the harvested energy for local SGD computations and model transmissions [64]. Some of the works assume an ideal transmission model in which the gradients are sent without any wireless channel in between, and the harvested energy is spent to send the gradient vector to the PS without any loss [65, 66]. Both of these study FL with energy harvesting nodes, and show the performance of the system with different scheduling policies and energy arrival profiles. However, [66] only covers the energy harvesting devices with a unit-capacity battery. The most recent topic in energy harvesting FL is the analysis of how much the wireless channel can affect the performance and the convergence rate of the algorithm as studied in [2, 67].

In order to investigate energy harvesting FL in a more practical scenario, we study energy harvesting FL using OTA aggregation in Chapter 5. We start our analysis with energy harvesting MUs with a unit battery, and then, we consider a case where MUs can have discrete battery levels.

2.5 Chapter Summary

In this chapter, we have discussed federated learning approaches that will be helpful in the subsequent chapters. We first introduced conventional FL with error-free transmissions and provided a literature review. Secondly, we discussed wireless FL and OTA aggregation along with its transmission and combining approaches. We then presented the idea of hierarchical FL that introduces intermediary servers to bring the server-side closer to the MUs. Finally, we motivated the FL model with energy harvesting MUs, and summarized the relevant literature. The rest of the thesis presents our novel studies and findings on FL over wireless channels.

Chapter 3

Hierarchical Over-the-Air Federated Edge Learning

In this chapter, in order to make distant MUs more resilient to the channel effects, we propose *hierarchical over-the-air federated learning* (HOTAFL), where MUs communicate with their corresponding ISs through wireless links. In this setup, each MU shares its local training result with its corresponding IS through OTA (cluster) aggregation. After several local iterations with the MUs in their clusters, the ISs send the results to the PS to complete the global aggregation, which constitutes one global iteration. We examine the performance of HOTAFL and compare the results with those of the conventional FL and error-free HFL both through analytical results and numerical experiments. The results show that the proposed framework outperforms conventional OTA FL and leads to a better model accuracy and faster convergence.

The chapter is organized as follows. In Sections 3.1 we introduce the specific communication model and the HOTAFL framework. In Section 3.2, we provide a convergence analysis of HOTAFL under certain convexity assumptions on the loss function. We present our numerical results in Section 3.3, and conclude the chapter in Section 3.4.

3.1 System Model

3.1.1 Ideal Communication Scenario

We refer to the case in which all the communication among all the units is error-free as the ideal communication scenario. In this case, after performing SGD, each MU calculates its model difference to be sent to its corresponding IS as

$$\Delta\boldsymbol{\theta}^{i}_{m,c}(t) = \boldsymbol{\theta}^{i,\tau+1}_{m,c}(t) - \boldsymbol{\theta}^{i}_{IS,c}(t). \tag{3.1}$$

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

$$\boldsymbol{\theta}^{i+1}_{IS,c}(t) = \boldsymbol{\theta}^{i}_{IS,c}(t) + \frac{1}{M}\sum_{m=1}^{M}\Delta\boldsymbol{\theta}^{i}_{m,c}(t). \tag{3.2}$$

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

$$\Delta\boldsymbol{\theta}_{PS,c}(t) = \boldsymbol{\theta}^{I+1}_{IS,c}(t) - \boldsymbol{\theta}_{PS}(t). \tag{3.3}$$

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

$$\Delta\boldsymbol{\theta}_{PS}(t) = \frac{1}{MC}\sum_{c=1}^{C}\sum_{i=1}^{I}\sum_{m=1}^{M}\Delta\boldsymbol{\theta}^{i}_{m,c}(t). \tag{3.4}$$

After the global aggregation, the model at the PS is updated as

$$\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\boldsymbol{\theta}_{PS}(t). \tag{3.5}$$

3.1.2 OTA Communication Scenario

We now consider the scheme referred as OTA communications, for which the links between the MUs and the ISs are wireless with OTA aggregation, however, the links between ISs and the PS is assumed to be error-free. Since a common wireless medium is used in local aggregations, noisy versions of the model updates $\Delta\boldsymbol{\theta}_{IS,c}(t)$ are received at the ISs. In our setup, the ISs are equipped with K

antennas, and we assume perfect channel state information (CSI) at the receivers and no CSI at the MUs. For the k-th antenna, the received signal at the c-th IS can be written as[1]

$$\boldsymbol{y}^i_{IS,c,k}(t) = \sum_{m=1}^{M} \boldsymbol{h}^i_{m,c,k}(t) \circ \boldsymbol{x}^i_{m,c,k}(t) + \boldsymbol{z}^i_{IS,c,k}(t), \tag{3.6}$$

where $\circ$ denotes the element-wise product, $\boldsymbol{x}^i_{m,c,k}(t) \in \mathbb{C}^N$, $\boldsymbol{z}^i_{IS,c,k}(t) \in \mathbb{C}^N$ with i.i.d. entries $z^{i,n}_{IS,c,k}(t) \sim \mathcal{CN}(0, \sigma_z^2)$. The channel coefficients are modelled as $\boldsymbol{h}^i_{m,c,k}(t) = \sqrt{\beta_{m,c}}\ \boldsymbol{g}^i_{m,c,k}(t)$, where $\boldsymbol{g}_{m,c,k}(t) \in \mathbb{C}^N$ with entries $g^{i,n}_{m,c,k}(t) \sim \mathcal{CN}(0, \sigma_h^2)$ (i.e., Rayleigh fading), $\beta_{m,c}$ is the large-scale fading coefficient modeled as $\beta_{m,c} = (d_{m,c})^{-p}$, where p represents the path loss exponent, and $d_{m,c}$ denotes the distance between the m-th MU in the c-th cluster and the IS in that cluster.

3.1.2.1 Local Aggregation

In OTA communications, in order to increase the spectral efficiency, the model differences are grouped to form a complex vector $\Delta\boldsymbol{\theta}^{i,cx}_{m,c}(t) \in \mathbb{C}^N$ with the following real and imaginary parts

$$\Delta\boldsymbol{\theta}^{i,re}_{m,c}(t) \triangleq \left[\Delta\theta^{i,1}_{m,c}(t), \Delta\theta^{i,2}_{m,c}(t), \ldots, \Delta\theta^{i,N}_{m,c}(t)\right]^T, \tag{3.7a}$$

$$\Delta\boldsymbol{\theta}^{i,im}_{m,c}(t) \triangleq \left[\Delta\theta^{i,N+1}_{m,c}(t), \Delta\theta^{i,N+2}_{m,c}(t), \ldots, \Delta\theta^{i,2N}_{m,c}(t)\right]^T. \tag{3.7b}$$

Under the assumption that there is no inter-cluster interference, the received signal for the k-th antenna in the c-th cluster at the i-th local iteration can be represented as

$$\boldsymbol{y}^i_{IS,c,k}(t) = P_t \sum_{m=1}^{M} \boldsymbol{h}^i_{m,c,k}(t) \circ \Delta\boldsymbol{\theta}^{i,cx}_{m,c}(t) + \boldsymbol{z}^i_{IS,c,k}(t), \tag{3.8}$$

where P_t is the power constant at the t-th global iteration. Knowing the CSI perfectly, the c-th IS combines the received signals as $\boldsymbol{y}^i_{IS,c}(t) =$

[1]Note that the setup here can be efficiently implemented in practice using orthogonal frequency-division multiplexing (OFDM).

$\frac{1}{K}\sum_{k=1}^{K}\left(\sum_{m=1}^{M}\boldsymbol{h}_{m,c,k}^{i}(t)\right)^{*}\circ\boldsymbol{y}_{IS,c,k}^{i}(t)$. For the n-th symbol, the combined signal can be written as

$$\begin{aligned} y_{IS,c}^{i,n}(t) = & \underbrace{P_t\sum_{m=1}^{M}\left(\frac{1}{K}\sum_{k=1}^{K}|h_{m,c,k}^{i,n}(t)|^2\right)\Delta\theta_{m,c}^{i,n,cx}(t)}_{y_{IS,c}^{i,n,sig}(t)\text{ (signal term)}} \\ & + \underbrace{\frac{P_t}{K}\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{k=1}^{K}(h_{m,c,k}^{i,n}(t))^{*}h_{m',c,k}^{i,n}(t)\Delta\theta_{m',c}^{i,n,cx}(t)}_{y_{IS,c}^{i,n,itf}(t)\text{ (interference term)}} \\ & + \underbrace{\frac{1}{K}\sum_{m=1}^{M}\sum_{k=1}^{K}(h_{m,c,k}^{i,n}(t))^{*}z_{c,k}^{i,n}(t)}_{y_{IS,c}^{i,n,no}(t)\text{ (noise term)}}. \end{aligned} \tag{3.9}$$

Aggregated model differences can be recovered by

$$\Delta\hat{\theta}_{IS,c}^{i,n}(t) = \frac{1}{P_t M\sigma_h^2\bar{\beta}_c}\,\mathrm{Re}\{y_{IS,c}^{i,n}(t)\}, \tag{3.10a}$$

$$\Delta\hat{\theta}_{IS,c}^{i,n+N}(t) = \frac{1}{P_t M\sigma_h^2\bar{\beta}_c}\,\mathrm{Im}\{y_{IS,c}^{i,n}(t)\}, \tag{3.10b}$$

where $\bar{\beta}_c = \sum_{m=1}^{M}\beta_{m,c}$. After estimating the model difference values, the cluster model update is written as

$$\boldsymbol{\theta}_{IS,c}^{i+1}(t) = \boldsymbol{\theta}_{IS,c}^{i}(t) + \Delta\hat{\boldsymbol{\theta}}_{IS,c}^{i}(t), \tag{3.11}$$

where $\Delta\hat{\boldsymbol{\theta}}_{IS,c}^{i}(t) = \left[\Delta\hat{\theta}_{IS,c}^{i,1}(t)\ \Delta\hat{\theta}_{IS,c}^{i,2}(t)\ \cdots\ \Delta\hat{\theta}_{IS,c}^{i,2N}(t)\right]^{T}$.

3.1.2.2 Global Aggregation

This part is similar to ideal communication. The only difference is that the aggregated signals are the estimates of the actual model differences. Letting $\boldsymbol{x}_{PS,c}(t)$ be the transmitted signal from the c-th IS, its n-th symbol can be written as

$$x_{PS,c}^{n}(t) = \Delta\theta_{PS,c}^{n}(t) + j\Delta\theta_{PS,c}^{n+N}(t). \tag{3.12}$$

Then, using (3.3), (3.9), (3.12) and recursion, the received signal for $1 \leq n \leq N$ (similarly for $N+1 \leq n \leq 2N$) can be written as

$$y_{PS}^{n}(t) = \sum_{c=1}^{C} x_{PS,c}^{n}(t) \tag{3.13}$$

$$= \underbrace{\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{\mathrm{Re}\left\{y_{IS,c}^{i,n,sig}(t)\right\}}{P_t M \sigma_h^2}}_{y_{PS,1}^{n}(t)} + \underbrace{\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{\mathrm{Re}\left\{y_{IS,c}^{i,n,itf}(t)\right\}}{P_t M \sigma_h^2}}_{y_{PS,2}^{n}(t)} + \underbrace{\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{\mathrm{Re}\left\{y_{IS,c}^{i,n,no}(t)\right\}}{P_t M \sigma_h^2}}_{y_{PS,3}^{n}(t)}. \tag{3.14}$$

The received signal at the PS is then recovered as $\Delta\hat{\theta}_{PS}^{n}(t) = \frac{1}{C}\,\mathrm{Re}\{y_{PS}^{n}(t)\}$, $\Delta\hat{\theta}_{PS}^{n+N}(t) = \frac{1}{C}\,\mathrm{Im}\{y_{PS}^{n}(t)\}$. Finally, the global aggregation is performed via

$$\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\hat{\boldsymbol{\theta}}_{PS}(t), \tag{3.15}$$

where $\Delta\hat{\boldsymbol{\theta}}_{PS}(t) = \left[\Delta\hat{\theta}_{PS}^{1}(t)\ \Delta\hat{\theta}_{PS}^{2}(t)\ \cdots\ \Delta\hat{\theta}_{PS}^{2N}(t)\right]^{T}$.

3.2 Convergence Analysis

Define the optimal solution as $\boldsymbol{\theta}^{*} \triangleq \arg\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$, the minimum values of the total and the local loss functions as $F^{*} = F(\boldsymbol{\theta}^{*})$ and $F_{m,c}^{*}$, respectively, and the bias in the dataset as $\Gamma \triangleq F^{*} - \sum_{c=1}^{C}\sum_{m=1}^{M}\frac{B_{m,c}}{B}F_{m,c}^{*} \geq 0$. In addition, assume that the learning rate of the overall system does not change in local iterations, i.e., $\eta_{m,c}^{i,j}(t) = \eta(t)$. Therefore, we can write the global update rule as

$$\boldsymbol{\theta}_{m,c}^{i,j+1}(t) = \boldsymbol{\theta}_{m,c}^{i,j}(t) - \eta(t)\nabla F_{m,c}(\boldsymbol{\theta}_{m,c}^{i,j}(t), \boldsymbol{\xi}_{m,c}^{i,j}(t)), \tag{3.16}$$

which can also be written as

$$\boldsymbol{\theta}_{m,c}^{i,j+1}(t) - \boldsymbol{\theta}_{m,c}^{i,1}(t) = -\eta(t)\sum_{l=1}^{j}\nabla F_{m,c}(\boldsymbol{\theta}_{m,c}^{i,l}, \boldsymbol{\xi}_{m,c}^{i,l}(t)). \tag{3.17}$$

Assumption 1. *All the loss functions are L-smooth and μ-strongly convex; i.e., $\forall \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^{2N}$, $\forall m \in [M], \forall c \in [C]$,*

$$F_{m,c}(\boldsymbol{v}) - F_{m,c}(\boldsymbol{w}) \leq \langle \boldsymbol{v}-\boldsymbol{w}, \nabla F_{m,c}(\boldsymbol{w})\rangle + \frac{L}{2}\|\boldsymbol{v}-\boldsymbol{w}\|_2^2, \tag{3.18}$$

$$F_{m,c}(\boldsymbol{v}) - F_{m,c}(\boldsymbol{w}) \geq \langle \boldsymbol{v}-\boldsymbol{w}, \nabla F_{m,c}(\boldsymbol{w})\rangle + \frac{\mu}{2}\|\boldsymbol{v}-\boldsymbol{w}\|_2^2. \tag{3.19}$$

Assumption 2. *The expected value of the squared l_2 norm of the stochastic gradients are bounded; i.e., $\forall j \in [\tau], i \in [I]$,*

$$\mathbb{E}_{\xi}\Big[\big\|\nabla F_{m,c}(\boldsymbol{\theta}_{m,c}^{i,j}(t), \boldsymbol{\xi}_{m,c}^{i,j}(t))\big\|_2^2 \Big] \leq G^2, \tag{3.20}$$

which translates to $\forall n \in [2N]$, $\mathbb{E}_{\xi}\big[\nabla F_{m,c}(\theta_{m,c}^{i,j,n}, \xi_{m,c}^{i,j,n}(t))\big] \leq G$.

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

$$\begin{aligned}
&\mathbb{E}\big[\|\boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^*\|_2^2 \big] \\
&\leq \Bigg(\prod_{a=1}^{t-1} X(a)\Bigg) \|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 + \sum_{b=1}^{t-1} Y(b) \prod_{a=b+1}^{t-1} X(a),
\end{aligned} \tag{3.21}$$

where $X(a) = (1 - \mu\eta(a) I\,(\tau - \eta(a)(\tau - 1)))$ and

$$\begin{aligned}
Y(a) = &\frac{\eta^2(a)\tau^2 G^2 I}{M^2 C^2} \sum_{m_1=1}^{M} \sum_{c_1=1}^{C} \Big(\frac{\beta_{m_1,c_1}^2}{K \bar{\beta}_{c_1}^2} + \Big(\sum_{m_2=1}^{M} \sum_{c_2=1}^{C} A_1 I \Big) \Big) \\
&+ \sum_{m=1}^{M} \sum_{\substack{m'=1 \\ m' \neq m}}^{M} \sum_{c=1}^{C} \frac{\eta^2(a)\tau^2 G^2 I \beta_{m,c} \beta_{m',c}}{M^2 C^2 K \bar{\beta}_c^2} + \frac{\sigma_z^2 I N}{P_a^2 M^2 C^2 K \sigma_h^2} \sum_{m=1}^{M} \sum_{c=1}^{C} \frac{\beta_{m,c}}{\bar{\beta}_c^2} \\
&+ (1 + \mu(1 - \eta(a))\, \eta^2(a) I G^2 \frac{\tau(\tau - 1)(2\tau - 1)}{6} \\
&+ \eta^2(a) I (\tau^2 + \tau - 1) G^2 + 2\eta(a) I (\tau - 1) \Gamma,
\end{aligned} \tag{3.22}$$

with $A_1 = 1 - \frac{\beta_{m_1,c_1}}{\beta_{c_1}} - \frac{\beta_{m_2,c_2}}{\beta_{c_2}} + \frac{\beta_{m_1,c_1}\beta_{m_2,c_2}}{\beta_{c_1}\beta_{c_2}}$.

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

$$\begin{aligned}
\|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{\theta}^*\|_2^2 &= \|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1) + \boldsymbol{v}(t+1) - \boldsymbol{\theta}^*\|_2^2 \\
&= \|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1)\|_2^2 + \|\boldsymbol{v}(t+1) - \boldsymbol{\theta}^*\|_2^2 \\
&\quad + 2\langle \boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1), \boldsymbol{v}(t+1) - \boldsymbol{\theta}^* \rangle.
\end{aligned} \tag{3.23}$$

Next, we provide upper bounds on the three terms of (3.23).

Lemma 1. *We have*

$$\mathbb{E}\Big[\big\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)\big\|_2^2\Big] \le \frac{\eta^2(t)\tau^2G^2I}{M^2C^2}\sum_{m_1=1}^{M}\sum_{c_1=1}^{C}\Big(\frac{\beta_{m_1,c_1}^2}{K\bar{\beta}_{c_1}^2}+\Big(\sum_{m_2=1}^{M}\sum_{c_2=1}^{C}A_1 I\Big)\Big)$$
$$+\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{c=1}^{C}\frac{\eta^2(t)\tau^2G^2I\beta_{m,c}\beta_{m',c}}{M^2C^2K\bar{\beta}_c^2}$$
$$+\frac{\sigma_z^2IN}{P_t^2M^2C^2K\sigma_h^2}\sum_{m=1}^{M}\sum_{c=1}^{C}\frac{\beta_{m,c}}{\bar{\beta}_c^2}. \tag{3.24}$$

Proof. We have $\Delta\hat{\theta}_{PS}^n(t)=\sum_{l=1}^{3}\Delta\hat{\theta}_{PS,l}^n(t)$, for the n-th symbol; using the independence of channel coefficients, we write

$$\mathbb{E}\big[||\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)||_2^2\big]=\mathbb{E}\big[\big\|\Delta\hat{\boldsymbol{\theta}}_{PS}(t)-\Delta\boldsymbol{\theta}_{PS}(t)\big\|_2^2\big]$$
$$=\sum_{n=1}^{2N}\big(\mathbb{E}\big[\big(\Delta\hat{\theta}_{PS,1}^n(t)-\Delta\theta_{PS}^n(t)\big)^2\big]+\sum_{l=2}^{3}\mathbb{E}\big[\big(\Delta\hat{\theta}_{PS,l}^n(t)\big)^2\big]. \tag{3.25}$$

In the following lemmas, we will bound each of these terms.

Lemma 2.

$$\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}_{PS,1}^n(t)-\Delta\theta_{PS}^n(t)\big)^2\big]=\frac{1}{M^2C^2}\sum_{m_1=1}^{M}\sum_{c_1=1}^{C}\sum_{i_1=1}^{I}\Big(\frac{\beta_{m_1,c_1}^2}{K\bar{\beta}_{c_1}^2}\mathbb{E}\big[\big\|\Delta\boldsymbol{\theta}_{m_1,c_1}^{i_1}(t)\big\|_2^2\big]$$
$$+\Big(\sum_{m_2=1}^{M}\sum_{c_2=1}^{C}\sum_{i_2=1}^{I}\sum_{n=1}^{2N}A_1\mathbb{E}\big[\Delta\theta_{m_1,c_1}^{i_1,n}(t)\Delta\theta_{m_2,c_2}^{i_2,n}(t)\big]\Big)\Big), \tag{3.26}$$

where $A_1=1-\frac{\beta_{m_1,c_1}}{\bar{\beta}_{c_1}}-\frac{\beta_{m_2,c_2}}{\bar{\beta}_{c_2}}+\frac{\beta_{m_1,c_1}\beta_{m_2,c_2}}{\bar{\beta}_{c_1}\bar{\beta}_{c_2}}$.

Proof. Using (3.4) and (3.10), we have

$$\begin{aligned}
&\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,1}(t)-\Delta\theta^n_{PS}(t)\big)^2\big]\\
&=\mathbb{E}\Big[\frac{1}{M^2C^2}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}\sum_{c_1=1}^{C}\sum_{c_2=1}^{C}\sum_{i_1=1}^{I}\sum_{i_2=1}^{I}\Delta\theta^{i_1,n}_{m_1,c_1}(t)\Delta\theta^{i_2,n}_{m_2,c_2}(t)\\
&\Big(1-\frac{1}{K\sigma_h^2\bar{\beta}_{c_1}}\sum_{k_1=1}^{K}|h^{i_1,n}_{m_1,c_1,k_1}(t)|^2-\frac{1}{K\sigma_h^2\bar{\beta}_{c_2}}\sum_{k_2=1}^{K}|h^{i_2,n}_{m_2,c_2,k_2}(t)|^2\\
&\quad+\frac{1}{K^2\sigma_h^4\bar{\beta}^2_{c_1}}\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}|h^{i_1,n}_{m_1,c_1,k_1}(t)|^2|h^{i_2,n}_{m_2,c_2,k_2}(t)|^2\Big)\Big].
\end{aligned}\tag{3.27}$$

Summing over all the symbols and using the independence of channel coefficients result in (3.26). □

Lemma 3.

$$\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\big]=\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{\beta_{m,c}\beta_{m',c}}{M^2C^2K\bar{\beta}_c^2}\mathbb{E}\big[\,\big\|\Delta\boldsymbol{\theta}^i_{m',c}(t)\big\|_2^2\,\big].\tag{3.28}$$

Proof. For $1\leq n\leq N$, using the independence of channel coefficients, we have

$$\begin{aligned}
&\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\big]=\mathbb{E}\Big[\Big(\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{1}{MCK\sigma_h^2\bar{\beta}_c}\\
&\quad\times\sum_{k=1}^{K}\mathrm{Re}\big\{\big(h^{i,n}_{m,c,k}(t)\big)^*h^{i,n}_{m',c,k}(t)\Delta\theta^{i,n}_{m',c}(t)\big\}\Big)^2\Big]\\
&=\mathbb{E}\Big[\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{c=1}^{C}\sum_{i=1}^{I}\frac{\beta_{m,c}\beta_{m',c}}{2M^2C^2K\bar{\beta}_c^2}\\
&\quad\times\big(\big(\Delta\theta^{i,n}_{m',c}(t)\big)^2+\big(\Delta\theta^{i,n+N}_{m',c}(t)\big)^2+\Delta\theta^{i,n}_{m,c}(t)\Delta\theta^{i,n}_{m',c}(t)-\Delta\theta^{i,n+N}_{m,c}(t)\Delta\theta^{i,n+N}_{m',c}(t)\big)\Big]
\end{aligned}\tag{3.29}$$

Obtaining the expressions for $N+1\leq n\leq 2N$ in a similar manner and combining the two, results in (3.28). □

Lemma 4.

$$\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,3}(t)\big)^2\big]=\frac{\sigma_z^2IN}{P_t^2M^2C^2K\sigma_h^2}\sum_{m=1}^{M}\sum_{c=1}^{C}\frac{\beta_{m,c}}{\bar{\beta}_c^2}.\tag{3.30}$$

Proof. Using the independence of channel coefficients, for $1 \leq n \leq N$, we have

$$\begin{aligned}\mathbb{E}\left[\left(\Delta\hat{\theta}^n_{PS,3}(t)\right)^2\right] =& \mathbb{E}\Big[\Big(\sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{i=1}^{I}\sum_{k=1}^{K}\frac{1}{P_t MCK\sigma_h^2\bar{\beta}_c}\,\mathrm{Re}\big\{\big(h^{i,n}_{m,c,k}(t)\big)^* z^{i,n}_{c,k}(t)\big\}\Big)^2\Big] \\ =& \mathbb{E}\Big[\sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{i=1}^{I}\sum_{k=1}^{K}\frac{1}{P_t^2M^2C^2K^2\sigma_h^4\bar{\beta}_c^2}\big(\mathrm{Re}\big\{\big(h^{i,n}_{m,c,k}(t)\big)^* z^{i,n}_{c,k}(t)\big\}\big)^2\Big] \\ =& \frac{\sigma_z^2 I}{2P_t^2M^2C^2K\sigma_h^2}\sum_{m=1}^{M}\sum_{c=1}^{C}\frac{\beta_{m,c}}{\bar{\beta}_c^2}. \end{aligned} \tag{3.31}$$

The same result holds for $N+1 \leq n \leq 2N$. Combining the two results concludes the proof. □

Combining the results in Lemmas 2, 3, and 4 and applying Assumption 2 with (3.17) completes the proof of Lemma 1. □

Lemma 5. *We have*

$$\begin{aligned}\mathbb{E}\Big[\big\|v(t+1)-\boldsymbol{\theta}^*\big\|_2^2\Big] \leq& (1-\mu\eta(t)I(\tau-\eta(t)(\tau-1)))\mathbb{E}\Big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\Big] \\ &+ (1+\mu(1-\eta(t))\,\eta^2(t)IG^2\frac{\tau(\tau-1)(2\tau-1)}{6} \\ &+ \eta^2(t)I(\tau^2+\tau-1)G^2 + 2\eta(t)I(\tau-1)\Gamma. \end{aligned} \tag{3.32}$$

Proof. See Appendix A. □

Lemma 6. $\mathbb{E}\left[\langle \boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1), \boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\rangle\right] = 0.$

Proof. We have

$$\begin{aligned}&\mathbb{E}[\langle \boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1), \boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\rangle] \\ &\quad= \mathbb{E}\Big[\big\langle \Delta\hat{\boldsymbol{\theta}}_{PS}(t)-\Delta\boldsymbol{\theta}_{PS}(t), \boldsymbol{\theta}_{PS}(t)+\Delta\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\rangle\Big]. \end{aligned} \tag{3.33}$$

Then, knowing that channel realizations are independent of the user and cluster updates at the same global iteration t, we have

$$\mathbb{E}\Big[\big\langle \Delta\hat{\boldsymbol{\theta}}_{PS}(t)-\Delta\boldsymbol{\theta}_{PS}(t), \boldsymbol{\theta}_{PS}(t)+\Delta\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\rangle\Big] = 0. \tag{3.34}$$

□

Recursively iterating through the results of Lemmas 1, 5, and 6 concludes the theorem. □

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

$$\mathbb{E}\left[F(\boldsymbol{\theta}_{PS}(T)) - F^*\right] \leq \frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}_{PS}(T) - \boldsymbol{\theta}^*\|_2^2\right]$$
$$\leq \frac{L}{2}\Bigg(\prod_{n=1}^{T-1} X(n)\Bigg)\|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 + \frac{L}{2}\sum_{p=1}^{T-1} Y(p)\prod_{n=p+1}^{T-1} X(n). \tag{3.35}$$

Remark. Since the third 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, meaning that more cluster aggregations do not always provide faster convergence. However, since the MUs face lower path losses in HOTAFL than in the conventional FL, it can reach a higher accuracy. Moreover, increasing the number of clusters C leads to a faster convergence, however, at the cost of employing more ISs.

3.3 Numerical Examples

We consider a hierarchical system with one PS and $C = 4$ non-overlapping clusters, each containing one IS with $K = 5MC$ receive antennas and $M = 5$ MUs. MUs are randomly placed in the clusters in such a way that their distance to the PS is between 0.5 and 3, and between 0.5 and 1 to their corresponding IS. We use the MNIST [68] and CIFAR-10 [69] datasets with Adam optimizer [70], and consider both i.i.d. and non-i.i.d. data distributions. In the i.i.d. case, data samples are randomly distributed among MUs, while in the non-i.i.d. case, the training data is divided into $5MC$ groups each consisting of samples with the same label. Then, 5 groups are assigned to each MU randomly. For CIFAR-10, we use the neural network given in [23] with $2N = 307498$ whereas for MNIST, we employ a one-layer network with $2N = 7850$. Three scenarios are considered: baseline with error-free transmissions, FL with OTA, i.e., ISs are not employed, and all the MUs aggregate parameters at the PS, and HOTAFL. We set the total

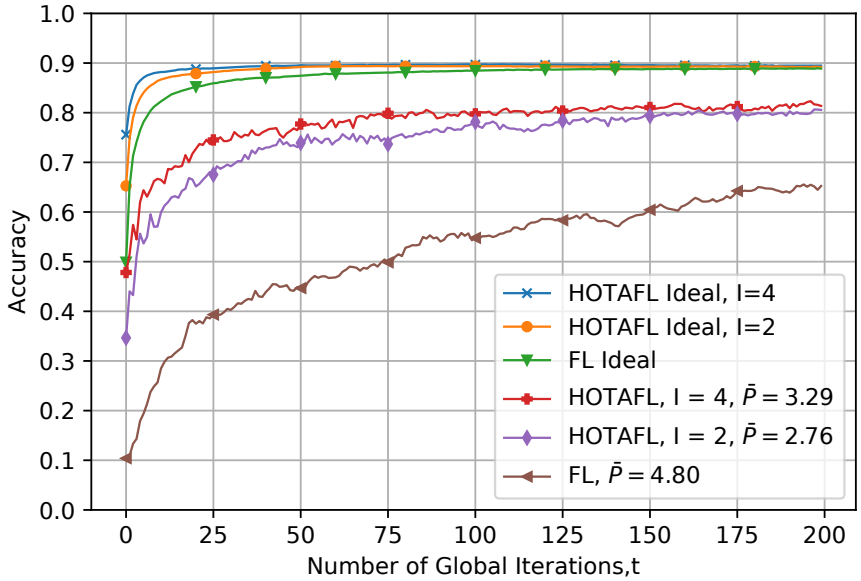

Figure 3.2: HOTAFL Test accuracy for non-i.i.d. MNIST data with $\tau = 3$. number of global iterations T to 200, the mini-batch size to $|\boldsymbol{\xi}^i_{m,c}(t)| = 500$, the path loss exponent p to 4, $\sigma_h^2 = 1$, $\sigma_z^2 = 10$ for the MNIST, and $\sigma_z^2 = 1$ for the CIFAR-10 training. Also, the power multiplier is set to $P_t = 1 + 10^{-2}t$ for HOTAFL, $P_t = 1.5 + 10^{-2}t$ for conventional FL, $t \in [T]$.

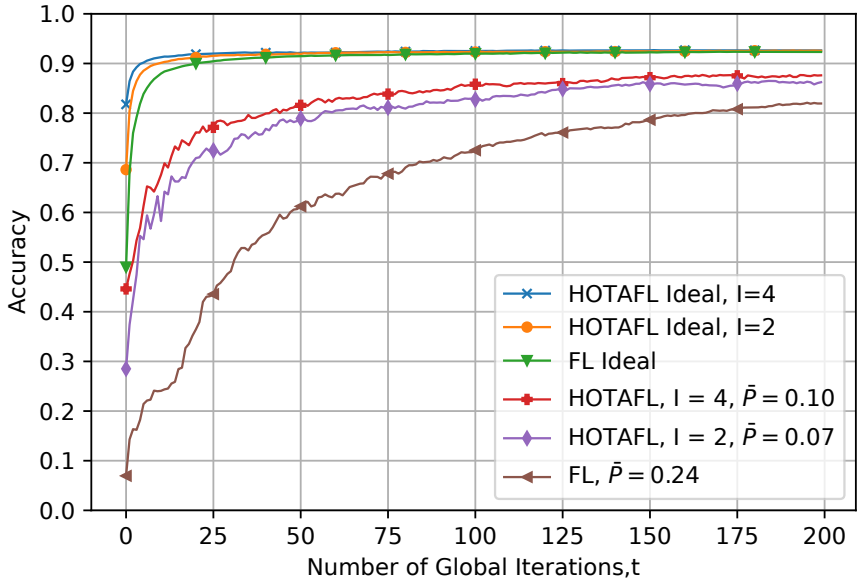

Figure 3.1: HOTAFL Test accuracy for i.i.d. MNIST data with $\tau = 1$.

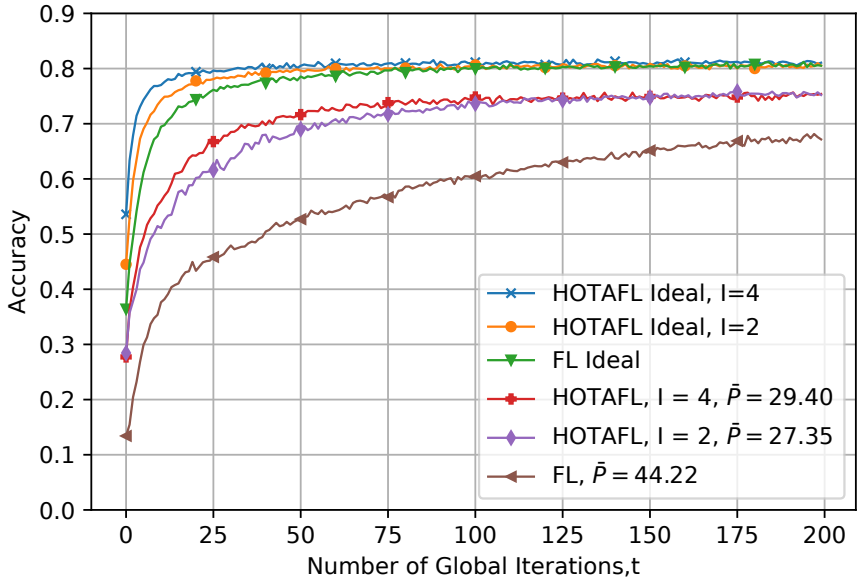

Figure 3.3: HOTAFL Test accuracy for i.i.d. CIFAR-10 data with $\tau = 5$.

Accuracy plots are presented in Figs. 3.1-3.3, where $\bar{P}$ is the average transmit power. The results show that with the selected geometry, bringing the servers closer to the MUs enhances the learning accuracy significantly. One reason is that the cluster structure enables the MUs share their model differences with a local IS closer than the PS, reducing the adverse effects of the large-scale wireless channel. Another reason is that MUs receive updated models even without communicating with the PS due to local aggregations. We also observe that although more initial power is given to FL, the received signals are distorted more compared to those of HOTAFL due to the more severe wireless channel effects. More local iterations enable faster convergence at the cost of increased transmit power. Increasing τ compensates the accuracy under a more complex model.

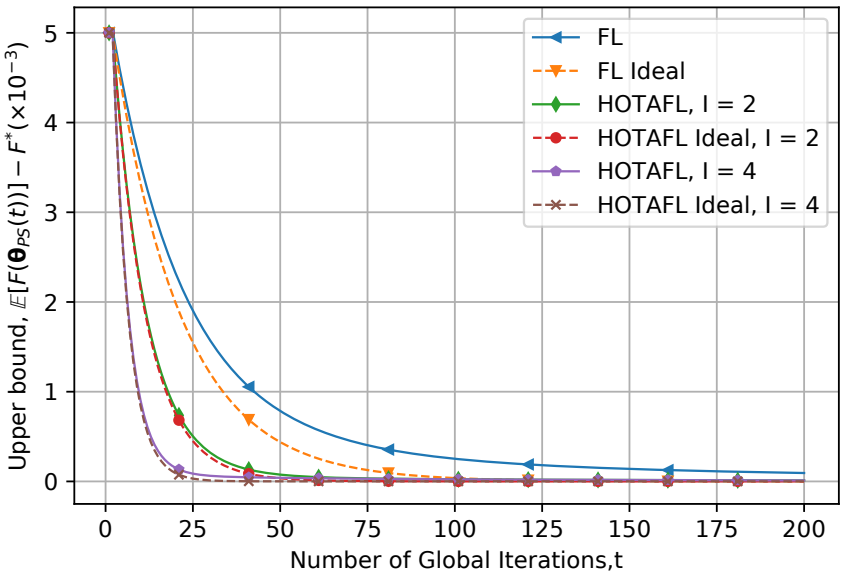

Figure 3.4: Convergence rate for i.i.d. MNIST data with $\tau = 1$.

In Fig. 3.4, we compare the convergence rates of conventional FL and HOTAFL using the upper bound in (3.35), with $2N = 7850, L = 10, \mu = 1, G^2 = 1, \Gamma = 1, \eta(t) = 5 \cdot 10^{-2} - 2 \cdot 10^{-5} t, P_t = 1 + 10^{-2} t, \beta_{m,c} = 1, \forall m \in [M], \forall c \in [C], \|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 = 10^3$. The convergence rate of HOTAFL and the ideal case are very close, and they become almost the same when the number of local iterations is increased.

3.4 Chapter Summary

We have proposed HOTAFL, which enables geographically localized model aggregation by employing ISs located in the areas where the MUs are more densely located. Our framework includes OTA cluster aggregations, which allows the MUs to simultaneously transmit and aggregate their model updates at the ISs over a wireless channel with path-loss and fading. We have analyzed the convergence rate of HOTAFL, and examined its performance with different datasets and data distributions. The results show that HOTAFL outperforms the conventional FL significantly.

Chapter 4

Over-the-Air Federated Edge Learning with Hierarchical Clustering

In this chapter, we propose a wireless-based hierarchical FL scheme (W-HFL) that uses intermediate servers (ISs) to form clusters at 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 number 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.

The chapter is organized as follows. In Section 4.1, introduce the specific communication model and the W-HFL framework. In Section 4.2, the convergence analysis of W-HFL is presented, and it is upper-bounded under some convexity

assumptions. In Section 4.3, 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 4.4.

4.1 System Model

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.

4.1.1 Cluster Aggregation

In OTA communication, the local updates $\Delta\boldsymbol{\theta}^{i}_{c,m}(t) \in \mathbb{R}^{2N}$ are sent without any coding. In order to increase the spectral efficiency, the model differences are grouped to form a complex vector $\Delta\boldsymbol{\theta}^{i,cx}_{c,m}(t) \in \mathbb{C}^{N}$ with entries $\Delta\theta^{i,n,cx}_{c,m}(t)$ for $m \in [M], c \in [C], i \in [I]$, with the following real and imaginary parts

$$\Delta\boldsymbol{\theta}^{i,re}_{c,m}(t) \triangleq \left[\Delta\theta^{i,1}_{c,m}(t), \theta^{i,2}_{c,m}(t), \ldots, \Delta\theta^{i,N}_{c,m}(t)\right]^{T}, \tag{4.1a}$$

$$\Delta\boldsymbol{\theta}^{i,im}_{c,m}(t) \triangleq \left[\Delta\theta^{i,N+1}_{c,m}(t), \theta^{i,N+2}_{c,m}(t), \ldots, \Delta\theta^{i,2N}_{c,m}(t)\right]^{T}, \tag{4.1b}$$

where $\Delta\theta^{i,n}_{c,m}(t)$ denotes the n-th entry of $\Delta\boldsymbol{\theta}^{i}_{c,m}(t)$ 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

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

as

$$\boldsymbol{y}^{i}_{IS,c,k}(t) = P_t \sum_{c'=1}^{C} \sum_{m=1}^{M} \boldsymbol{h}^{i}_{c',m,c,k}(t) \circ \Delta\boldsymbol{\theta}^{i,cx}_{c',m}(t) + \boldsymbol{z}^{i}_{IS,c,k}(t), \tag{4.2}$$

where P_t is the power multiplier at the t-th global iteration, $\circ$ denotes the element-wise (Hadamard) product, $\boldsymbol{z}^{i}_{IS,c,k}(t) \in \mathbb{C}^N$ 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., $z^{i,n}_{IS,c,k}(t) \sim \mathcal{CN}(0, \sigma_z^2), n \in [N]$. $\boldsymbol{h}^{i}_{c',m,c,k}(t) \in [N]$ 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 $h^{i,n}_{c',m,c,k}(t) = \sqrt{\beta_{c',m,c}}\, g^{i,n}_{c',m,c,k}(t)$, where $g^{i,n}_{c',m,c,k}(t) \sim \mathcal{CN}(0, \sigma_h^2)$ is the small-scale fading coefficient (i.e., Rayleigh fading), and $\beta_{c',m,c}$ is the large-scale fading coefficient modeled as $\beta_{c',m,c} = (d_{c',m,c})^{-p}$, 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

$$\boldsymbol{y}^{i}_{IS,c}(t) = \frac{1}{K} \sum_{k=1}^{K} \Big(\sum_{m=1}^{M} \boldsymbol{h}^{i}_{c,m,c,k}(t) \Big)^{*} \circ \boldsymbol{y}^{i}_{IS,c,k}(t), \tag{4.3}$$

whose n-th entry can be written as

$$y^{i,n}_{IS,c}(t) = \frac{1}{K} \sum_{k=1}^{K} \Big(\sum_{m=1}^{M} h^{i,n}_{c,m,c,k}(t) \Big)^{*} y^{i,n}_{IS,c,k}(t), \tag{4.4}$$

where $y^{i,n}_{IS,c,k}(t)$ denotes the n-th entry of $\boldsymbol{y}^{i}_{IS,c,k}(t), n \in [N]$. Substituting (4.2) into (4.3), and using (4.4), we get

$$\begin{aligned} y^{i,n}_{IS,c}(t) = {} & \underbrace{\frac{P_t}{K} \sum_{m=1}^{M} \Big(\sum_{k=1}^{K} |h^{i,n}_{c,m,c,k}(t)|^2 \Big) \Delta\theta^{i,n,cx}_{c,m}(t)}_{y^{i,n,sig}_{IS,c}(t) \text{ (signal term)}} + \underbrace{\frac{1}{K} \sum_{m=1}^{M} \sum_{k=1}^{K} (h^{i,n}_{c,m,c,k}(t))^{*} z^{i,n}_{IS,c,k}(t)}_{y^{i,n,no}_{IS,c}(t) \text{ (noise term)}} \\ & + \frac{P_t}{K} \sum_{m=1}^{M} \sum_{k=1}^{K} (h^{i,n}_{c,m,c,k}(t))^{*} \Bigg(\underbrace{\sum_{\substack{m'=1 \\ m' \neq m}}^{M} h^{i,n}_{c,m',c,k}(t) \Delta\theta^{i,n,cx}_{c,m'}(t)}_{y^{i,n,int1}_{IS,c}(t) \text{ (Intra-cluster interference)}} + \underbrace{\sum_{\substack{c'=1 \\ c' \neq c}}^{C} \sum_{m'=1}^{M} h^{i,n}_{c',m',c,k}(t) \Delta\theta^{i,n,cx}_{c',m'}(t)}_{y^{i,n,int2}_{IS,c}(t) \text{ (Inter-cluster interference)}} \Bigg) \\ = {} & y^{i,n,sig}_{IS,c}(t) + y^{i,n,int1}_{IS,c}(t) + y^{i,n,int2}_{IS,c}(t) + y^{i,n,noise}_{IS,c}(t). \end{aligned} \tag{4.5}$$

Aggregated model differences for $n \in [N]$, can be recovered by

$$\Delta\hat{\theta}^{i,n}_{IS,c}(t) = \frac{1}{P_t M \sigma_h^2 \bar{\beta}_c} \operatorname{Re}\{y^{i,n}_{IS,c}(t)\}, \qquad \Delta\hat{\theta}^{i,n+N}_{IS,c}(t) = \frac{1}{P_t M \sigma_h^2 \bar{\beta}_c} \operatorname{Im}\{y^{i,n}_{IS,c}(t)\}, \tag{4.6}$$

where $\bar{\beta}_c = \sum_{m=1}^{M} \beta_{c,m,c}$, and $\mathrm{Re}\{a\}$ and $\mathrm{Im}\{a\}$ denote the real and imaginary parts of a, respectively. Finally, the cluster model update can be written as

$$\boldsymbol{\theta}_{IS,c}^{i+1}(t) = \boldsymbol{\theta}_{IS,c}^{i}(t) + \Delta\hat{\boldsymbol{\theta}}_{IS,c}^{i}(t), \tag{4.7}$$

where $\Delta\hat{\boldsymbol{\theta}}_{IS,c}^{i}(t) \triangleq \left[\Delta\hat{\theta}_{IS,c}^{i,1}(t)\Delta\hat{\theta}_{IS,c}^{i,2}(t)\ldots\Delta\hat{\theta}_{IS,c}^{i,2N}(t)\right]^T$.

4.1.2 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 are completed to obtain the signal to be transmitted from the c-th IS, model differences are grouped to form a complex vector $\Delta\boldsymbol{\theta}_{PS,c}^{cx} \in \mathbb{C}^N$, with the following real and imaginary parts

$$\Delta\boldsymbol{\theta}_{PS,c}^{re}(t) \triangleq \left[\Delta\theta_{PS,c}^{1}(t), \theta_{PS,c}^{2}(t), \ldots, \Delta\theta_{PS,c}^{N}(t)\right]^T, \tag{4.8a}$$

$$\Delta\boldsymbol{\theta}_{PS,c}^{im}(t) \triangleq \left[\Delta\theta_{PS,c}^{N+1}(t), \theta_{PS,c}^{N+2}(t), \ldots, \Delta\theta_{PS,c}^{2N}(t)\right]^T, \tag{4.8b}$$

where $\Delta\theta_{PS,c}^{n}(t)$ 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

$$\boldsymbol{y}_{PS,k'}(t) = P_{IS,t} \sum_{c=1}^{C} \boldsymbol{h}_{PS,c,k'}(t) \circ \Delta\boldsymbol{\theta}_{PS,c}^{cx}(t) + \boldsymbol{z}_{PS,k'}(t), \tag{4.9}$$

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., $z_{PS,k'}^{n}(t) \sim \mathcal{CN}(0, \sigma_z^2)$. 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 $g_{PS,c,k'}^{n}(t) \sim \mathcal{CN}(0, \sigma_h^2)$, $\beta_{IS,c}$ 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

$$\boldsymbol{y}_{PS}(t) \triangleq \frac{1}{K'} \sum_{k'=1}^{K'} \Big(\sum_{c=1}^{C} \boldsymbol{h}_{PS,c,k'}(t) \Big)^* \circ \boldsymbol{y}_{PS,k'}(t). \tag{4.10}$$

Estimated global model differences at the PS can be recovered as

$$\Delta\hat{\theta}^n_{PS}(t) = \frac{1}{P_{IS,t}C\sigma_h^2\bar{\beta}}\,\mathrm{Re}\{y^n_{PS}(t)\}, \qquad \Delta\hat{\theta}^{n+N}_{PS}(t) = \frac{1}{P_{IS,t}C\sigma_h^2\bar{\beta}}\,\mathrm{Im}\{y^n_{PS}(t)\}, \tag{4.11}$$

where $\bar{\beta} = \sum_{c=1}^{C}\beta_{IS,c}$. Finally, the global aggregation is performed using

$$\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\hat{\boldsymbol{\theta}}_{PS}(t), \tag{4.12}$$

where $\Delta\hat{\boldsymbol{\theta}}_{PS}(t) = \left[\Delta\hat{\theta}^1_{PS}(t)\Delta\hat{\theta}^2_{PS}(t)\ldots\Delta\hat{\theta}^{2N}_{PS}(t)\right]^T$.

The n-th symbol can be written as

$$y^n_{PS}(t) = \frac{1}{K'}\sum_{k'=1}^{K'}\Big(\sum_{c=1}^{C}h_{PS,c,k'}(t)\Big)^* y^n_{PS,k'}(t) \tag{4.13a}$$

$$= \underbrace{P_{IS,t}\sum_{c=1}^{C}\Big(\frac{1}{K'}\sum_{k'=1}^{K'}|h^n_{PS,c,k'}(t)|^2\Big)\Delta\theta^{n,cx}_{PS,c}(t)}_{\text{Signal Term}} \tag{4.13b}$$

$$+ \underbrace{\frac{P_{IS,t}}{K'}\sum_{c=1}^{C}\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{k'=1}^{K'}\big(h^n_{PS,c,k'}(t)\big)^* h^n_{PS,c',k'}(t)\Delta\theta^{n,cx}_{PS,c'}(t)}_{\text{Interference Term}}$$

$$+ \underbrace{\frac{1}{K'}\sum_{c=1}^{C}\sum_{k'=1}^{K'}\big(h^n_{PS,c,k'}(t)\big)^* z^n_{PS,k'}(t)}_{\text{Noise Term}}$$

$$= y^{n,sig}_{PS}(t) + y^{n,int}_{PS}(t) + y^{n,noise}_{PS}(t). \tag{4.13c}$$

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 (2.19) and recursively adding previous cluster iterations, we obtain

$$\Delta\theta^{n,cx}_{PS,c}(t) = \big(\Delta\theta^{I+1,n}_{IS,c}(t) - \Delta\theta^{1,n}_{IS,c}(t)\big) + j\big(\Delta\theta^{I+1,n+N}_{IS,c}(t) - \Delta\theta^{1,n+N}_{IS,c}(t)\big) \tag{4.14}$$

$$= \sum_{i=1}^{I}\Delta\hat{\theta}^{i,n}_{IS,c}(t) + j\Delta\hat{\theta}^{i,n+N}_{IS,c}(t) \tag{4.15}$$

$$= \frac{1}{P_t M\sigma_h^2\bar{\beta}_c}\sum_{i=1}^{I} y^{i,n}_{IS,c}(t). \tag{4.16}$$

Substituting Equation (4.16) into (4.13), we have

$$
\begin{aligned}
y_{PS}^{n}(t)=&P_{IS,t}\sum_{c=1}^{C}\left(\frac{1}{K'}\sum_{k'=1}^{K'}|h_{PS,c,k'}^{n}(t)|^{2}\right)\left(\frac{1}{P_tM\sigma_h^2\bar{\beta}_c}\sum_{i=1}^{I}y_{IS,c}^{i,n}(t)\right)\\
&+\frac{P_{IS,t}}{K'}\sum_{c=1}^{C}\sum_{\substack{c'=1\\c'\neq c}}^{C'}\sum_{k'=1}^{K'}\left(h_{PS,c,k'}^{n}(t)\right)^{*}h_{PS,c',k'}^{n}(t)\left(\frac{1}{P_tM\sigma_h^2\bar{\beta}_{c'}}\sum_{i=1}^{I}y_{IS,c'}^{i,n}(t)\right)\\
&+\frac{1}{K'}\sum_{c=1}^{C}\sum_{k'=1}^{K'}\left(h_{PS,c,k'}^{n}(t)\right)^{*}z_{PS,k'}^{n}(t). \qquad (4.17)
\end{aligned}
$$

Substituting (5.9) into (4.17), we can write y_{PS}^{n} as $y_{PS}^{n}(t)=\sum_{l=1}^{9}y_{PS,l}^{n}$, with $\lambda_{t,c}=\frac{P_{IS,t}}{KK'M\sigma_h^2}$, each term can be written as

$$
\begin{aligned}
y_{PS,1}^{n}(t)=&\sum_{\substack{c,m,i,\\k,k'}}\frac{\lambda_{t,c}}{\bar{\beta}_c}|h_{PS,c,k'}^{n}(t)|^{2}|h_{c,m,c,k}^{i,n}(t)|^{2}\Delta\theta_{c,m}^{i,n,cx}(t),\\
y_{PS,2}^{n}(t)=&\sum_{\substack{c,m,m'\neq m,\\i,k,k'}}\frac{\lambda_{t,c}}{\bar{\beta}_c}|h_{PS,c,k'}^{n}(t)|^{2}\left(h_{c,m,c,k}^{i,n}(t)\right)^{*}h_{c,m',c,k}^{i,n}(t)\Delta\theta_{c,m'}^{i,n,cx}(t),\\
y_{PS,3}^{n}(t)=&\sum_{\substack{c,c'\neq c,m,m',\\i,k,k'}}\frac{\lambda_{t,c}}{\bar{\beta}_c}|h_{PS,c,k'}^{n}(t)|^{2}\left(h_{c,m,c,k}^{i,n}(t)\right)^{*}h_{c,m',c',k}^{i,n}(t)\Delta\theta_{c',m'}^{i,n,cx}(t),\\
y_{PS,4}^{n}(t)=&\sum_{\substack{c,m,i,\\k,k'}}\frac{\lambda_{t,c}}{P_{IS,t}\bar{\beta}_c}|h_{PS,c,k'}^{n}(t)|^{2}\left(h_{c,m,c,k}^{i,n}(t)\right)^{*}z_{IS,c,k}^{i,n}(t),\\
y_{PS,5}^{n}(t)=&\sum_{\substack{c,c'\neq c,m,\\i,k,k'}}\frac{\lambda_{t,c}}{\bar{\beta}_{c'}}\left(h_{PS,c,k'}^{n}(t)\right)^{*}h_{PS,c',k'}^{n}(t)|h_{c',m,c',k}^{i,n}(t)|^{2}\Delta\theta_{c',m}^{i,n,cx}(t),\\
y_{PS,6}^{n}(t)=&\sum_{\substack{c,c'\neq c,m,m'\neq m,\\i,k,k'}}\frac{\lambda_{t,c}}{\bar{\beta}_{c'}}\left(h_{PS,c,k'}^{n}(t)\right)^{*}h_{PS,c',k'}^{n}(t)\left(h_{c',m,c',k}^{i,n}(t)\right)^{*}h_{c',m',c',k}^{i,n}(t)\Delta\theta_{c',m'}^{i,n,cx}(t),\\
y_{PS,7}^{n}(t)=&\sum_{\substack{c,c'\neq c,c''\neq c',\\m,m',i,k,k'}}\frac{\lambda_{t,c}}{\bar{\beta}_{c'}}\left(h_{PS,c,k'}^{n}(t)\right)^{*}h_{PS,c',k'}^{n}(t)\left(h_{c',m,c',k}^{i,n}(t)\right)^{*}h_{c',m',c'',k}^{i,n}(t)\Delta\theta_{c'',m'}^{i,n,cx}(t),\\
y_{PS,8}^{n}(t)=&\sum_{\substack{c,c'\neq c,m,\\i,k,k'}}\frac{\lambda_{t,c}}{P_{IS,t}\bar{\beta}_{c'}}\left(h_{PS,c,k'}^{n}(t)\right)^{*}h_{PS,c',k'}^{n}(t)\left(h_{c',m,c',k}^{i,n}(t)\right)^{*}z_{IS,c',k}^{i,n}(t),\\
y_{PS,9}^{n}(t)=&\sum_{c,k'}\frac{1}{K'}\left(h_{PS,c,k'}^{n}(t)\right)^{*}z_{PS,k}^{n}(t), \qquad (4.18)
\end{aligned}
$$

4.2 Convergence Analysis

In this section, we present an upper bound on the global loss function, which shows how far the global FL model is after a certain number of iterations from the optimal model. Define the optimal solution that minimizes the loss $F(\boldsymbol{\theta})$ as

$$\boldsymbol{\theta}^* \triangleq \arg\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}). \tag{4.19}$$

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

$$\Gamma \triangleq F^* - \sum_{c=1}^{C}\sum_{m=1}^{M} \frac{B_{c,m}}{B} F^*_{c,m} \geq 0. \tag{4.20}$$

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

$$\boldsymbol{\theta}^{i,j+1}_{c,m}(t) = \boldsymbol{\theta}^{i,j}_{c,m}(t) - \eta(t)\nabla F_{c,m}(\boldsymbol{\theta}^{i,j}_{c,m}(t), \boldsymbol{\xi}^{i,j}_{c,m}(t)), \tag{4.21}$$

which can also be written as

$$\boldsymbol{\theta}^{i,j+1}_{c,m}(t) - \boldsymbol{\theta}^{i,1}_{c,m}(t) = -\eta(t)\sum_{l=1}^{j} \nabla F_{c,m}(\boldsymbol{\theta}^{i,l}_{c,m}, \boldsymbol{\xi}^{i,l}_{c,m}(t)). \tag{4.22}$$

Theorem 2. *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*

$$\mathbb{E}\left[\|\boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^*\|_2^2\right] \leq \left(\prod_{a=0}^{t-1} X(a)\right)\|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 + \sum_{b=0}^{t-1} Y(b) \prod_{a=b+1}^{t-1} X(a), \tag{4.23}$$

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

$$
\begin{aligned}
Y(t) &= \frac{\eta^2(t)G^2I^2\tau^2}{M^2C^2}\sum_{c_1=1}^{C}\sum_{c_2=1}^{C}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}A(m_1,m_2,c_1,c_2)\\
&+\frac{\big(2+(M-1)(C-2)(K-1)(I-1)\big)\eta^2(t)IG^2\tau^2}{K(K')M^3C^2(C-1)\bar{\beta}^2}\sum_{c=1}^{C}\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}\frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m_1,c'}\beta_{c',m_2,c'}}{\bar{\beta}_{c'}^2}\\
&+\frac{\eta^2(t)G^2I\tau^2}{KK'M^2C^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{m=1}^{M}\left(\frac{(K'+1)\beta_{IS,c}^2\beta_{c,m,c}}{\bar{\beta}_c^2}\Big(\sum_{\substack{m'=1\\m'\neq m}}^{M}\beta_{c,m',c}+\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{m'=1}^{M}\beta_{c,m',c'}\Big)\right)\\
&+\frac{\eta^2(t)G^2I\tau^2}{KK'M^2C^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{m=1}^{M}\left(\frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}}{\bar{\beta}_{c'}^2}\Big(\sum_{\substack{m'=1\\m'\neq m}}^{M}\beta_{c',m',c'}+\sum_{\substack{c''=1\\c''\neq c'}}^{C}\beta_{c',m',c''}\Big)\right)\\
&+\frac{\sigma_z^2N}{K'C^2\sigma_h^2\bar{\beta}^2}\sum_{c=1}^{C}\beta_{IS,c}\left(\frac{1}{P_{IS,t}^2}+\frac{I}{KM^2}\sum_{m=1}^{M}\left(\frac{(K'+1)\beta_{IS,c}\beta_{c,m,c}}{P_t^2\bar{\beta}_c^2}+\sum_{\substack{c'=1\\c'\neq c}}^{C}\frac{\beta_{IS,c'}\beta_{c',m,c'}}{P_{IS,t}^2\bar{\beta}_{c'}^2}\right)\right)\\
&+(1+\mu(1-\eta(t))\,\eta^2(t)IG^2\frac{\tau(\tau-1)(2\tau-1)}{6}+\eta^2(t)I(\tau^2+\tau-1)G^2+2\eta(t)I(\tau-1)\Gamma, \qquad (4.24)
\end{aligned}
$$

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

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

$$
\begin{aligned}
&\big\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{\theta}^*\big\|_2^2=\big\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)+\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\big\|_2^2\\
&=\big\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)\big\|_2^2+\big\|\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\big\|_2^2+2\langle\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1),\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\rangle. \qquad (4.25)
\end{aligned}
$$

Next, we provide upper bounds on the three terms of (4.25).

Lemma 7. $\mathbb{E}\Big[\big\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)\big\|_2^2\Big] \le$ $\dfrac{\eta^2(t)G^2I^2\tau^2}{M^2C^2}\sum_{c_1=1}^{C}\sum_{c_2=1}^{C}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}A(m_1,m_2,c_1,c_2)$
$+\dfrac{\big(2+(M-1)(C-2)(K-1)(I-1)\big)\eta^2(t)IG^2\tau^2}{K(K')M^3C^2(C-1)\bar{\beta}^2}\sum_{c=1}^{C}\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}\dfrac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m_1,c'}\beta_{c',m_2,c'}}{\bar{\beta}_{c'}^2}$
$+\dfrac{\eta^2(t)G^2I\tau^2}{KK'M^2C^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{m=1}^{M}\left(\dfrac{(K'+1)\beta_{IS,c}^2\beta_{c,m,c}}{\bar{\beta}_c^2}\Big(\sum_{\substack{m'=1\\m'\neq m}}^{M}\beta_{c,m',c}+\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{m'=1}^{M}\beta_{c,m',c'}\Big)\right)$
$+\dfrac{\eta^2(t)G^2I\tau^2}{KK'M^2C^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{m=1}^{M}\left(\dfrac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}}{\bar{\beta}_{c'}^2}\Big(\sum_{\substack{m'=1\\m'\neq m}}^{M}\beta_{c',m',c'}+\sum_{\substack{c''=1\\c''\neq c'}}^{C}\beta_{c',m',c''}\Big)\right)$

$+\frac{\sigma_z^2 N}{K'C^2\sigma_h^2\bar{\beta}^2}\sum_{c=1}^{C}\beta_{IS,c}\left(\frac{1}{P_{IS,t}^2}+\frac{I}{KM^2}\sum_{m=1}^{M}\left(\frac{(K'+1)\beta_{IS,c}\beta_{c,m,c}}{P_t^2\bar{\beta}_c^2}+\sum_{\substack{c'=1\\c'\neq c}}^{C}\frac{\beta_{IS,c'}\beta_{c',m,c'}}{P_{IS,t}^2\bar{\beta}_{c'}^2}\right)\right)$, *where* $A(m_1,m_2,c_1,c_2)$ *is given in Theorem 2.*

Proof. See Appendix B. □

Lemma 8. $\mathbb{E}\left[\left\|\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\right\|_2^2\right] \leq (1-\mu\eta(t)I\left(\tau-\eta(t)(\tau-1)\right))\,\mathbb{E}\left[\left\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\right\|_2^2\right]$
$+\left(1+\mu(1-\eta(t))\,\eta^2(t)IG^2\frac{\tau(\tau-1)(2\tau-1)}{6}+\eta^2(t)I(\tau^2+\tau-1)G^2+2\eta(t)I(\tau-1)\Gamma.\right.$

Proof. See Appendix A. □

Lemma 9. $\mathbb{E}\left[\langle\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1),\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\rangle\right]=0.$

Proof. $\mathbb{E}[\langle\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1),\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\rangle]=\mathbb{E}\Big[\langle\Delta\hat{\boldsymbol{\theta}}_{PS}(t)-\Delta\boldsymbol{\theta}_{PS}(t),\boldsymbol{\theta}_{PS}(t)+\Delta\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\rangle\Big]$. 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 7, 8, and 9 concludes the theorem. □

Corollary 2. *Assuming L-smoothness, after T global iterations, the loss function can be upper-bounded as*

$$\begin{aligned}\mathbb{E}\left[F\big(\boldsymbol{\theta}_{PS}(T)\big)\right]-F^* &\leq \frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}_{PS}(T)-\boldsymbol{\theta}^*\|_2^2\right],\\ &\leq \frac{L}{2}\left(\prod_{a=0}^{T-1}X(a)\right)\|\boldsymbol{\theta}_{PS}(0)-\boldsymbol{\theta}^*\|_2^2+\frac{L}{2}\sum_{b=0}^{T-1}Y(b)\prod_{a=b+1}^{T-1}X(a).\end{aligned} \tag{4.26}$$

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, it can reach a higher accuracy. Moreover, increasing the number of clusters C leads to a faster convergence, however, at the cost of employing more ISs.

Corollary 3. *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*

$$\begin{aligned} Y(t) &\approx \frac{\eta^2(t)G^2}{KK'M^3C^3}MCK' + 2\eta^2G^2\left(1-\frac{1}{MC}\right) + \frac{\sigma_z^2N}{KC^3\sigma_h^2}\left(\frac{1}{P_{IS,t}} + \frac{(K'+1)}{KM^3P^2} + \frac{(C-1)}{P_{IS,t}^2M^2}\right) \\ &\approx \frac{\eta^2(t)G^2}{KM^2C^2} + 2\eta^2G^2 + \frac{\sigma_z^2N}{KM^3C^3\sigma_h^2P^2} \\ &\approx 2\eta^2(t)G^2 + \frac{\sigma_z^2N}{KM^3C^3\sigma_h^2P^2}, \end{aligned} \tag{4.27}$$

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

$$\begin{aligned} \mathbb{E}\left[F\big(\boldsymbol{\theta}_{PS}(T)\big)\right] - F^* \leq \frac{L}{2}\big(1-\mu\eta\big)^T\|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 \\ + \frac{L}{2\mu\eta}\Big(2\eta^2G^2 + \frac{\sigma_z^2N}{KM^3C^3\sigma_h^2P^2}\Big)\Big(1-\big(1-\mu\eta\big)^T\Big). \end{aligned} \tag{4.28}$$

Remark 2. As expected, it can be observed that the numbers of receive antennas, MUs and ISs have a positive effect on the convergence, whereas the model dimension has an adverse effect.

4.3 Numerical Examples

In this section, we evaluate and compare the performance of W-HFL with that of the conventional FL under different scenarios. Via different experiments, we observe the power consumption, as well as the convergence speed of the learning algorithm with different number of cluster aggregations, I. In our experiments, we use two different image classification datasets, MNIST [68] and CIFAR-10 [69]. 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 the convolutional neural network (CNN) architecture with $2N = 307498$ given in

Table 4.1. We employ Adam optimizer [70] for training both networks.
Table 4.1: CNN Architecture for CIFAR-10 Training

CIFAR-10
$3 \times 3 \times 32$ convolutional layer, same padding, batch normalization, ReLU
$3 \times 3 \times 32$ convolutional layer, same padding, batch normalization, ReLU
2×2 max pooling
Dropout with $p = 0.2$
$3 \times 3 \times 64$ convolutional layer, same padding, batch normalization, ReLU
$3 \times 3 \times 64$ convolutional layer, same padding, batch normalization, ReLU
2×2 max pooling
Dropout with $p = 0.3$
$3 \times 3 \times 128$ convolutional layer, same padding, batch normalization, ReLU
$3 \times 3 \times 128$ convolutional layer, same padding, batch normalization, ReLU
2×2 max pooling
Dropout with $p = 0.4$
Softmax activation with 10 output neurons

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 in each cluster, and a single PS. MUs in each cluster 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, where it is assumed that the conventional FL has $I = 1$, 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 $I = 1$.

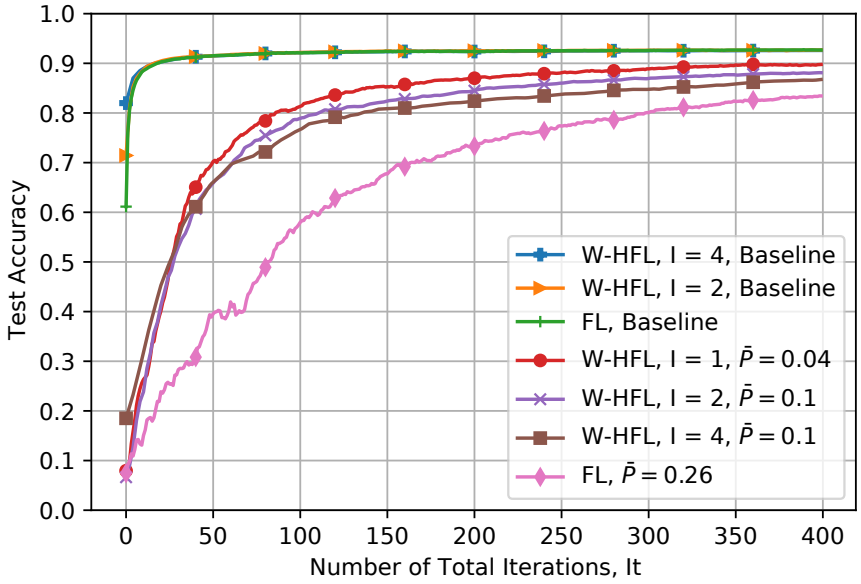

Figure 4.1: W-HFL Test accuracy for i.i.d. MNIST data with $\tau = 1$.

In Figs. 4.1-4.3, we present the performance of W-HFL with different number 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 $I = 1$, $I = 2$, and $I = 4$, as well as the conventional FL scheme, i.e., $I = 1$, with no IS in-between. To assess their performance, we compare the results with ideal baseline cases, where the model differences are assumed to be transmitted in an error-free manner. We can observe in Fig. 4.1 that W-HFL outperforms conventional FL 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

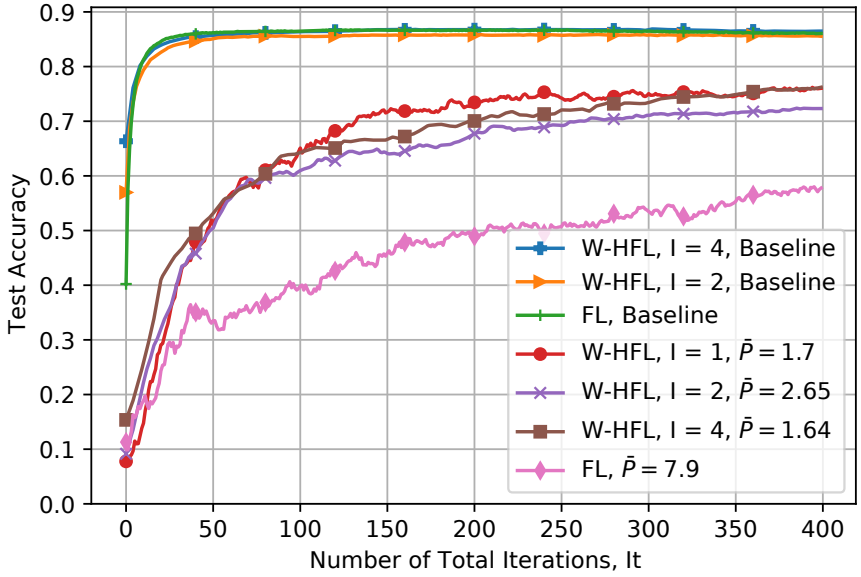

Figure 4.2: W-HFL Test accuracy for user non-i.i.d. MNIST data with $\tau = 3$.

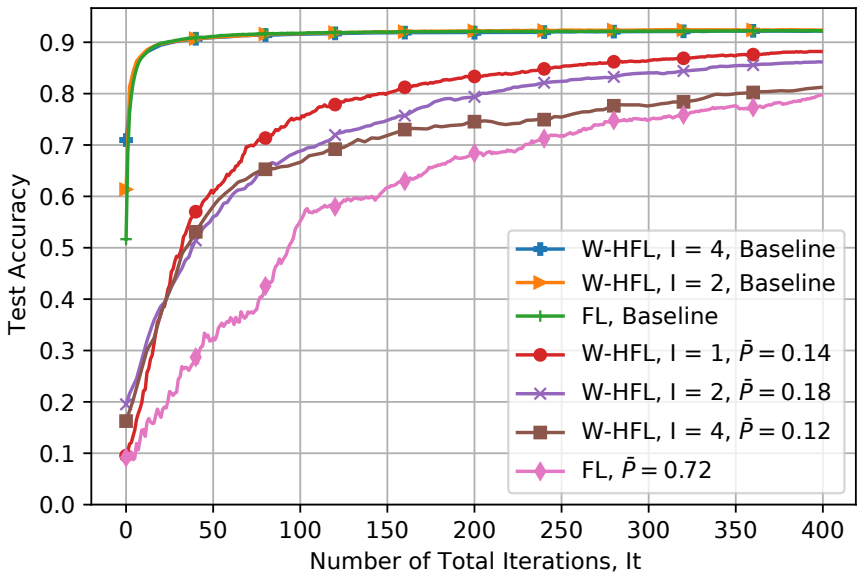

Figure 4.3: W-HFL Test accuracy for cluster non-i.i.d. with $\tau = 1$.

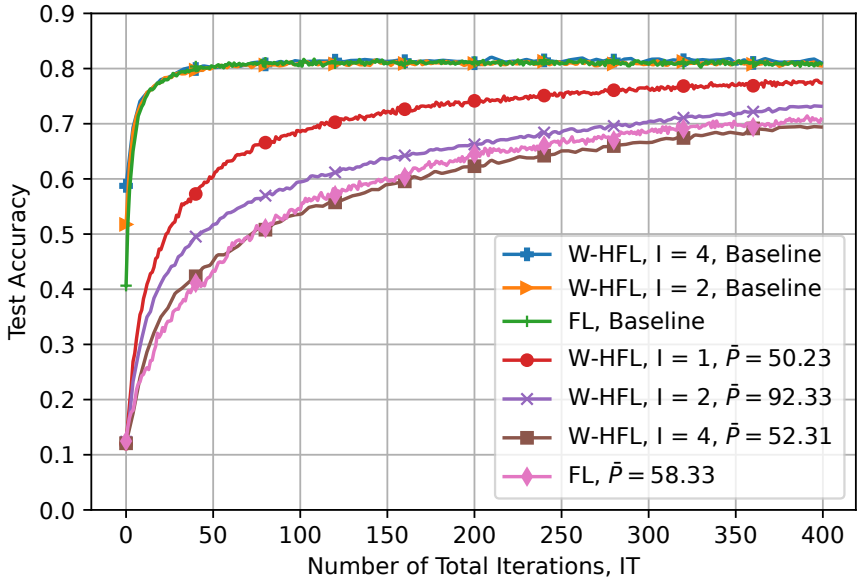

Figure 4.4: W-HFL Test accuracy for i.i.d. CIFAR-10 data with $\tau = 5$.

performance slightly deteriorates as I increases, while consuming less transmit power at the edge. The system performs better in i.i.d. distribution when the ISs perform less cluster aggregations, and the best performance is observed with $I = 1$, where the ISs just relay the received cluster updates. In Fig. 4.2, 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. 4.3. When the clusters are non-i.i.d, we observe a slight decrease in accuracies when compared to the i.i.d. data distribution.

In Fig. 4.4, 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 model contains more parameters in CIFAR-10 simulations to tackle with the more challenging dataset. It can be observed that using the ISs as relays gives the best performance

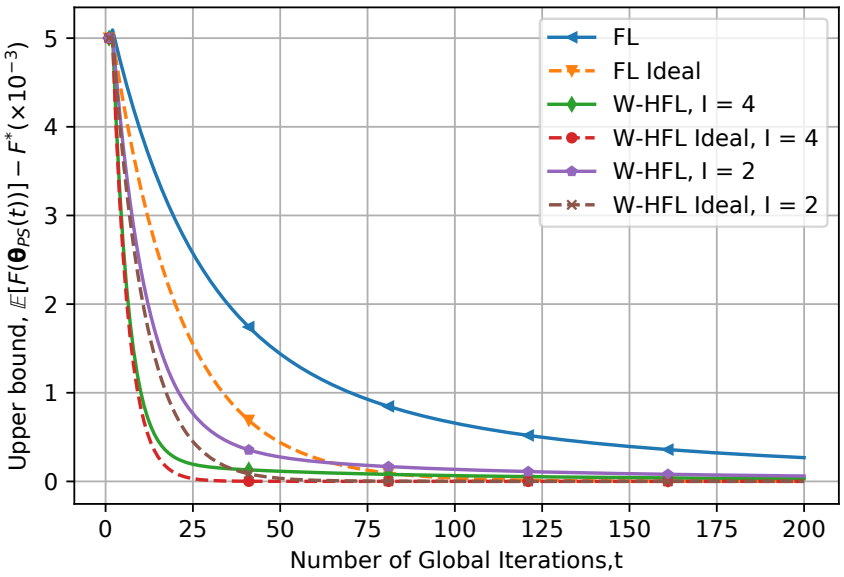

Figure 4.5: Convergence rate for Figure 4.1.

while using less transmit power. W-HFL with $I = 2$ uses the more transmit power than $I = 4$ since it performs more global iteration rounds with an increased P_t. We can also see that the gap between conventional FL and W-HFL is closed, and the main reason is that the transmit power is a lot higher than the noise variance since the more challenging datasets are more susceptible to wireless channel effects.

In Fig. 4.5, we numerically analyze the convergence rate of W-HFL, with the results presented in Corollary 2. The setting from MNIST i.i.d. training is used with $2N = 7850, L = 10, \mu = 1, G^2 = 1, \Gamma = 1, \eta(t) = 5 \cdot 10^{-2} - 2 \cdot 10^{-5} t, P_t = 1 + 10^{-2} t, P_{IS,t} = 10 P_t, \|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 = 10^3$. We can observe that W-HFL convergences faster than the conventional FL, and performs similar to the baseline.

4.4 Chapter Summary

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 global model for the next

iteration. We have considered the inter-cluster interference at the cluster aggregations, and introduced 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 and experimental analyses 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 less cluster aggregations in W-HFL can lead to higher accuracies, but with an increased cost of transmit power at the edge.

Chapter 5

Over-the-Air Federated Learning with Energy Harvesting Devices

To examine the performance of FL with energy harvesting in practical settings, we introduce OTA FL with energy harvesting MUs. In this setting, the participating MUs perform local SGD iterations and transmit their gradients using wireless links simultaneously over the same frequency band. Using OTA aggregation and combining techniques, the PS updates the global model based on the received signal, and the updated model is sent back to the users for the next global iteration. We compare the performance of our setup with the error-free scenarios and conventional FL using different energy arrival profiles. Numerical and experimental results show that even under energy harvesting limitations, the proposed algorithm can perform well under practical channel models with large number of users with convergence guarantees.

In addition to the case where the MUs are equipped with unit batteries, we also consider a system where the MUs can receive finite levels of energy which will be spent on computation of the gradients and their transmission, however, it cannot be saved for the next iteration. We propose two approaches where the excess energy is spent either on more local gradient computations or on amplifying the gradients for the transmission. We conduct a theoretical analysis of the

convergence rate for both cases and show that the emphasis on more computations gives a faster convergence rate and a better performance when compared to the case where the excess energy is spent on transmission.

The chapter is organized as follows. In Section 5.1, we study the communication model of OTA FL with MUs that have intermittent energy arrivals. In Section 5.2, we provide a convergence analysis of the energy harvesting FL under certain convexity assumptions on the loss function. We present our numerical results in Section 5.3, and conclude the paper in Section 5.4.

5.1 System Model

Since the objective and the wireless channel models are the same as those in Section 2.2, a detailed description is not repeated here.

5.1.1 System Model for Energy Harvesting Devices with Unit Battery

In the first scenario, we consider OTA FL with energy harvesting devices where each MU has a unit battery. The MUs harvest either unit energy, or no energy at all from various sources such as solar, kinetic, or RF energy in every global iteration. For simplicity, we assume that τ local SGD steps and the transmission of gradients to the PS cost a unit amount of energy.

We denote the binary energy arrival process of the m-th MU at the t-th global iteration as $F_m(t)$. If $F_m(t) = 1$, then the m-th MU receives enough energy to participate in the global iteration at iteration t. $F_m(t) = 0$, if no energy is harvested. We also define the elapsed time between the current iteration and the previous energy arrival as $\lambda_m(t) = \max_{t':t'<t,F_m(t')=1} t'$. Lastly, for a given t, we define a quantity called the cooldown multiplier as $c_m(t) = t - \lambda_m(t)$, which represents the number of iterations for which the m-th MU has not been

harvesting energy.

We investigate the use of MUs with stochastic energy arrival profiles, where the harvested energy has an underlying probability distribution, and the MUs have no prior information about the next energy arrival time. Note that the MUs do not know the underlying distribution of the stochastic process. We will consider two different stochastic energy arrival processes: Bernoulli and uniform energy arrivals.

Bernoulli Energy Arrivals

At the t-th global iteration, the m-th MU receives energy with probability $\alpha_m(t)$, i.e.,

$$F_m(t) = \begin{cases} 1 & \text{with probability } \alpha_m(t), \\ 0 & \text{with probability } 1-\alpha_m(t). \end{cases} \tag{5.1}$$

Uniform Energy Arrivals

In the uniform energy arrival model, the global iterations are divided into blocks of length T_m, and the m-th MU receives energy once for every T_m iterations. This means that with probability 1, an energy arrival is observed in $\{t, \ldots, t + T_m - 1\}$.

We now describe the proposed scheme for which the FL participants are energy harvesting devices and the gradients are sent through wireless channels using OTA aggregation. Since the mobile devices do not always have sufficient energy to perform local SGD computations or gradient transmissions, only the MUs that have harvested enough energy, i.e., those with $E_m(t) = 1$ can participate in the t-th global iteration. We define $\mathcal{S}(t)$ as the set of devices participating in the t-th global iteration.

Before each training round, the MUs receive the current global model $\boldsymbol{\theta}_{PS}(t)$ from the PS. If an MU is eligible to participate in the t-th iteration based on its energy status, the SGD calculations are performed. Then, based on the cooldown

multiplier of each MU, the weighted model differences are calculated as

$$\Delta\boldsymbol{\theta}_m^s(t) = C_m(t)\Delta\boldsymbol{\theta}_m(t), \tag{5.2}$$

where $C_m(t) = p_m(t)c_m(t)$, and $\Delta\boldsymbol{\theta}_m^s(t)$ denotes the scaled model differences for the m-th MU at the t-th global iteration. Considering error-free transmission of the scaled gradients, the PS performs the global update for the next iteration as

$$\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\boldsymbol{\theta}_{PS}(t), \tag{5.3}$$

where $\Delta\boldsymbol{\theta}_{PS}(t)$ is given as

$$\Delta\boldsymbol{\theta}_{PS}(t) = \frac{1}{C(t)} \sum_{m\in\mathcal{S}_t} \Delta\boldsymbol{\theta}_m^s(t), \tag{5.4}$$

with $C(t) = \sum_{m\in\mathcal{S}_t} C_m(t)$, which is assumed to be known by the PS [66]. The reader is referred to [71] and the references therein for related algorithms to estimate the number of participating users.

Next, we consider the OTA aggregation for the proposed model where the scaled local gradients are transmitted simultaneously over the same frequency band through the wireless channel. The PS receives a noisy target signal due to the channel effects and noise. As in the previous chapters, in the proposed scheme, we assume perfect channel state information at the receiver side and no CSI at the MUs.

For a more spectrally efficient approach, the model differences are written in terms of a complex signal $\Delta\boldsymbol{\theta}_m^{s,cx}(t) \in \mathbb{C}^N$ by grouping the symbols into its real and imaginary parts as

$$\Delta\boldsymbol{\theta}_m^{s,re}(t) \triangleq \left[\Delta\theta_m^{s,1}(t), \Delta\theta_m^{s,2}(t), \ldots, \Delta\theta_m^{s,N}(t)\right]^T, \tag{5.5a}$$

$$\Delta\boldsymbol{\theta}_m^{s,im}(t) \triangleq \left[\Delta\theta_m^{s,N+1}(t), \Delta\theta_m^{s,N+2}(t), \ldots, \Delta\theta_m^{s,2N}(t)\right]^T. \tag{5.5b}$$

For the k-th antenna, the PS receives the signal

$$\boldsymbol{y}_{PS,k}(t) = \sum_{m\in\mathcal{S}_t} \boldsymbol{h}_{m,k}(t) \circ \Delta\boldsymbol{\theta}_m^{s,cx}(t) + \boldsymbol{z}_{PS,k}(t), \tag{5.6}$$

where the terminology and the notation is the same as those adopted in Chapter 2.2.

Since a perfect CSI is available at the receiver side, the PS combines the received signal as (see [23])

$$\boldsymbol{y}_{PS}(t) = \frac{1}{K}\sum_{k=1}^{K}\Big(\sum_{m\in\mathcal{S}_t}\boldsymbol{h}_{m,k}(t)\Big)^{*}\circ\boldsymbol{y}_{PS,k}(t). \tag{5.7}$$

For the n-th symbol, the combined signal becomes

$$\begin{aligned} y_{PS}^{n}(t) = & \underbrace{\sum_{m\in\mathcal{S}_t}\Big(\frac{1}{K}\sum_{k=1}^{K}|h_{m,k}^{n}(t)|^{2}\Big)\Delta\theta_{m,s}^{n,cx}(t)}_{y_{PS}^{n,sig}(t)\text{ (signal term)}} \\ & + \underbrace{\frac{1}{K}\sum_{m\in\mathcal{S}_t}\sum_{\substack{m'\in\mathcal{S}_t\\ m'\neq m}}\sum_{k=1}^{K}(h_{m,k}^{n}(t))^{*}h_{m',k}^{n}(t)\Delta\theta_{m',s}^{n,cx}(t)}_{y_{PS}^{n,int}(t)\text{ (interference term)}} \\ & + \underbrace{\frac{1}{K}\sum_{m\in\mathcal{S}_t}\sum_{k=1}^{K}(h_{m,k}^{n}(t))^{*}z_{PS,k}^{n}(t)}_{y_{PS}^{n,noise}(t)\text{ (noise term)}}. \end{aligned} \tag{5.8}$$

We recover the aggregated model differences from the received signal as

$$\Delta\hat{\theta}_{PS}^{n}(t) = \frac{1}{C(t)\sigma_h^2\bar{\beta}}\,\mathrm{Re}\{y_{PS}^{n}(t)\}, \tag{5.9a}$$

$$\Delta\hat{\theta}_{PS}^{n+N}(t) = \frac{1}{C(t)\sigma_h^2\bar{\beta}}\,\mathrm{Im}\{y_{PS}^{n}(t)\}. \tag{5.9b}$$

Finally, the global update can be performed as

$$\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta\hat{\boldsymbol{\theta}}_{PS}(t), \tag{5.10}$$

where $\Delta\hat{\boldsymbol{\theta}}_{PS}(t) = \big[\Delta\hat{\theta}_{PS}^{1}(t)\ \Delta\hat{\theta}_{PS}^{2}(t)\ \cdots\ \Delta\hat{\theta}_{PS}^{2N}(t)\big]^{T}$.

5.1.2 System Model for Energy Harvesting Devices with Discrete Battery

We now consider MDs as energy harvesting devices with a discrete battery. We assume that each MD has a maximum battery level E_{max} and battery levels are

discrete, i.e., before the t-th global iteration, the m-th MD has a battery level of $E_m(t) \in \{0, 1, \ldots, E_{max}\}$. Energy arrival for the m-th user is modeled as a Poisson counting process with parameter $\lambda_{e,m}$. The time difference between two consecutive global iterations is t_g, the amount of energy harvested by the m-th MD in t_g amount of time at the t-th global iteration is denoted by the random variable $E_{m,in}(t)$, and the amount of energy spent is denoted by $Q_m(t)$.

At every global iteration t, mobile devices (MDs) with available energy perform $\tau_m(t)$ local SGD iterations with minimum and maximum numbers of SGD iterations that can be performed at each device as τ_{min} and τ_{max}, respectively.

We define the energy cost of one local SGD step as E_{sgd} and required energy for the local update transmission as $E_{tr,m}(t) = P_m(t)E_{tr,min}$ where $P_m(t) \geq 1$ is the power multiplier for the m-th MD at the t-th global iteration, and $E_{tr,min}$ is the minimum amount of required energy for the transmission. Therefore, we can write $Q_m(t) = \tau_m(t)E_{sgd} + E_{tr,m}(t)$. We further define the participation policy of the m-th MD at the t-th global iteration as $F_m(t)$, which is assigned based on its energy state. We say that $F_m(t) = 1$ if $E_m(t) \geq Q_m(t)$, and it is 0 otherwise. Moreover, the set of devices participating in the global iteration t is denoted as $\mathcal{S}_t = \{m \in [M] | F_m(t) = 1\}$.

We can calculate the battery level of the m-th MD at the t-th global iteration as

$$E_m(t) = \begin{cases} E_{max}, & \text{if } E_m(t-1)+E_{m,in}(t)-Q_m(t-1) \geq E_{max} \\ E_m(t-1)+E_{m,in}(t)-Q_m(t-1), & \text{if } Q_m(t) \leq E_m(t-1) \leq E_{max} \\ E_m(t-1)+E_{m,in}(t), & \text{if } E_m(t-1) \leq Q_m(t). \end{cases} \tag{5.11}$$

For this model, we define $\Delta\boldsymbol{\theta}_{m,s}(t) = X_m(t)\Delta\boldsymbol{\theta}_m(t) = p_m\pi_m(t)\Delta\boldsymbol{\theta}_m(t)$ where $p_m = \frac{|\mathcal{D}_m|}{\sum_{m\in\mathcal{S}_t}|\mathcal{D}_m|}$ and $\pi_m(t)$ is the multiplier variable based on the energy consumption policy. Instead of using $C_m(t)$ variable which comes from the cooldown variable, we now use $X_m(t)$, which is defined based on which energy policy is being applied. Similarly, we change $C(t)$ as $\bar{X}(t) = \sum_{m\in\mathcal{S}_t} X_m(t)$ to use the same system model with the case of unit battery. In the following, we explain different

power consumption policies for energy harvesting devices with discrete batteries.

Transmission-Greedy Over-the-Air Federated Learning with Energy Harvesting Devices (OFED)

The transmission greedy approach aims to utilize the excess energy for increasing the transmit power. This scenario works as follows: if an MD has enough energy to participate in t-th global iteration, it uses all the remaining energy in the battery for gradient amplification by fixing $\tau_m(t) = \tau$. The energy consumption for the m-th MD at the t-th global iteration can be written as

$$Q_m^{TG}(t) = \begin{cases} \tau E_{sgd} + P_m^{TG}(t) E_{tr,min}, & \text{if } \tau E_{sgd} + E_{tr,min} \leq E_m(t) \\ 0, & \text{otherwise,} \end{cases} \tag{5.12}$$

where $P_m^{TG}(t) = \frac{E_m(t) - \tau E_{sgd}}{E_{tr,min}}$. In this scenario, the normalization parameter $\pi_m(t)$ becomes $\pi_m(t) = P_m^{TG}(t)$.

SGD-Greedy OFED

SGD-greedy approach focuses on performing more local SGD computations with the excess energy in the battery. In this scenario, devices use as little energy as possible for the gradient transmission by equating $E_{tr,m}(t) = E_{tr,min}$, and try to perform as many local SGD steps as possible. For the m-th MD at the t-th global iteration, the energy consumption profile can be written as

$$Q_m^{SG}(t) = \begin{cases} \tau_m^{SG}(t) E_{sgd} + E_{tr,min}, & \text{if } E_{sgd} + E_{tr,min} \leq E_m(t) \\ 0, & \text{otherwise,} \end{cases} \tag{5.13}$$

where $\tau_m^{SG}(t) = \min\left(\left\lfloor \frac{E_m(t) - E_{tr,min}}{E_{sgd}} \right\rfloor, \tau_{max} \right)$ with τ_{max} being the maximum number of SGD iterations that each device can perform. For the normalization, $\pi_m(t) = \tau_m(t)$ is used.

5.2 Convergence Analysis

In this section, we analyze the convergence rate of the proposed algorithm by considering the global loss function, and providing some upper bounds on the difference between the global loss of the FL model and the optimal model with further iterations.

We denote the minimum local loss as F_m^*, the optimal weights of the model as $\boldsymbol{\theta}^* \triangleq \arg\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$, and the minimum total loss function is given as $F^* = F(\boldsymbol{\theta}^*)$. The dataset bias is defined as $\Gamma \triangleq F^* - \sum_{m=1}^{M} p_m F_m^* \geq 0$. Moreover, it is assumed that the learning rate remains unchanged among different MUs, i.e., $\eta_m^i(t) = \eta(t)$.

The following convergence rate analysis for OTA FL with energy harvesting employs similar analysis techniques as in the previous chapters and in [23]. The main difference is that we introduce randomness of the number of participants, which depends on the energy arrival parameters.

5.2.1 Convergence Analysis for OTA FL with Energy Harvesting Devices with Unit Battery

We first state the following two assumptions, which are similar to Assumptions 1 and 2 in Chapter 3.

Assumption 3. *Squared l_2 norm of the local stochastic gradients are bounded; i.e.,*

$$\mathbb{E}_{\xi}\Big[\, \|\nabla F_m(\boldsymbol{\theta}_m(t), \boldsymbol{\xi}_m(t))\|_2^2 \,\Big] \leq G^2, \tag{5.14}$$

which translates to $\forall n \in [2N]$, $\mathbb{E}_{\xi}[\nabla F_m(\theta_m^n, \xi_m^n(t))] \leq G$.

Assumption 4. *Local loss functions are assumed to be L-smooth and μ-strongly convex; i.e., $\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{2N}$, $\forall m \in [M]$,*

$$F_m(\boldsymbol{a}) - F_m(\boldsymbol{b}) \leq \langle \boldsymbol{a} - \boldsymbol{b}, \nabla F_m(\boldsymbol{b}) \rangle + \frac{L}{2}\|\boldsymbol{a} - \boldsymbol{b}\|_2^2, \tag{5.15}$$

$$F_m(\boldsymbol{a}) - F_m(\boldsymbol{b}) \geq \langle \boldsymbol{a} - \boldsymbol{b}, \nabla F_m(\boldsymbol{b}) \rangle + \frac{\mu}{2}\|\boldsymbol{a} - \boldsymbol{b}\|_2^2. \tag{5.16}$$

Theorem 3. *In energy harvesting OTA FL with Bernoulli energy arrivals $\alpha_m = \alpha$ and equal data distribution $p_m = p, \forall m \in [M]$, for $0 \leq \eta(t) \leq min\{1, \frac{1}{\tau\mu}\}$, we can upper bound the model difference between the global and the optimal weights as*

$$\mathbb{E}\big[\,\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\|_2^2\,\big] \leq \bigg(\prod_{a=1}^{t-1} X(a)\bigg)\|\boldsymbol{\theta}_{PS}(0)-\boldsymbol{\theta}^*\|_2^2+\sum_{b=1}^{t-1} Y(b)\prod_{a=b+1}^{t-1} X(a), \tag{5.17}$$

where $X(a) = (1-\mu\eta(a)\,(\tau-\eta(a)(\tau-1)))$ and

$$\begin{aligned} Y(a) =& \tau^2 G^2 \eta^2(a) \sum_{m_1\in\mathcal{S}_a}\sum_{m_2\in\mathcal{S}_a} A(m_1,m_2)\frac{\tau^2 G^2\eta^2(a)}{K\bar{\beta}^2}\sum_{m\in\mathcal{S}_a}\sum_{\substack{m'\in\mathcal{S}_a\\ m'\neq m}}\beta_m\beta_{m'} \\ &+\frac{\sigma_z^2 N}{p^2 K\sigma_h^2}\sum_{m\in\mathcal{S}_a}\frac{\beta_m}{\bar{\beta}^2}\left(1+\mu(1-\eta(t))\,\eta^2(t)G^2\frac{\tau(\tau-1)(2\tau-1)}{6}\right. \\ &+\eta^2(t)(\tau^2+\tau-1)G^2+2\eta(t)(\tau-1)\Gamma. \end{aligned} \tag{5.18}$$

with $A(m_1,m_2) = \Big(1-\frac{\beta_{m_1}}{\bar{\beta}}-\frac{\beta_{m_2}}{\bar{\beta}}+\frac{(M\alpha+1)(K+1)\beta_{m_1}\beta_{m_2}}{M\alpha K\bar{\beta}^2}\Big)$.

Proof. Define an auxiliary variable $\boldsymbol{v}(t+1)\triangleq\boldsymbol{\theta}_{PS}(t)+\Delta\boldsymbol{\theta}_{PS}(t)$, where $\Delta\boldsymbol{\theta}_{PS}(t)$ is defined in (5.4). Then, we have

$$\begin{aligned} \|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{\theta}^*\|_2^2 &= \|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)+\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\|_2^2 \\ &= \|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)\|_2^2+\|\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\|_2^2 \\ &\quad+2\langle\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1),\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\rangle. \end{aligned} \tag{5.19}$$

In the following lemmas, we provide upper bounds for (5.19).

Lemma 10. $\mathbb{E}\Big[\big\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)\big\|_2^2\Big]$

$$\begin{aligned} &\leq \tau^2 G^2\eta^2(t)\sum_{m_1\in\mathcal{S}_t}\sum_{m_2\in\mathcal{S}_t}A(m_1,m_2)+\frac{\sigma_z^2 N}{p^2K\sigma_h^2}\sum_{m\in\mathcal{S}_t}\frac{\beta_m}{\bar{\beta}^2} \\ &\quad+\frac{\tau^2G^2\eta^2(t)}{K\bar{\beta}^2}\sum_{m\in\mathcal{S}_t}\sum_{\substack{m'\in\mathcal{S}_t\\ m'\neq m}}\beta_m\beta_{m'}. \end{aligned} \tag{5.20}$$

Proof. We can write $\Delta\hat{\theta}_{PS}^n(t)=\sum_{p=1}^3\Delta\hat{\theta}_{PS,p}^n(t)$, for the n-th symbol using (5.8),

because of the i.i.d. of channel realizations, we obtain

$$\mathbb{E}\big[||\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)||_2^2\big] = \mathbb{E}\big[\big\|\Delta\hat{\boldsymbol{\theta}}_{PS}(t)-\Delta\boldsymbol{\theta}_{PS}(t)\big\|_2^2\big]$$
$$=\sum_{n=1}^{2N}(\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,1}(t)-\Delta\theta^n_{PS}(t)\big)^2\big]+\sum_{p=2}^{3}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,l}(t)\big)^2\big]. \tag{5.21}$$

Lemma 11. $\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,1}(t)-\Delta\theta^n_{PS}(t)\big)^2\big]$

$$\leq\sum_{n=1}^{2N}\sum_{m_1\in\mathcal{S}_t}\sum_{m_2\in\mathcal{S}_t}A(m_1,m_2)\mathbb{E}\big[\Delta\theta^n_{m_1}(t)\Delta\theta^n_{m_2}(t)\big]. \tag{5.22}$$

where $A(m_1,m_2)=\Big(1-\frac{\beta_{m_1}}{\bar{\beta}}-\frac{\beta_{m_2}}{\bar{\beta}}+\frac{2+(M\alpha-1)(K-1)\beta_{m_1}\beta_{m_2}}{M\alpha K\bar{\beta}^2}\Big)$.

Proof. For a single symbol, we can write

$$\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,1}(t)-\Delta\theta^n_{PS}(t)\big)^2\big]=\mathbb{E}\Big[\frac{1}{C(t)^2}\sum_{m_1\in\mathcal{S}_t}\sum_{m_2\in\mathcal{S}_t}C_{m_1}(t)C_{m_2}(t)\Delta\theta^n_{m_1}(t)\Delta\theta^n_{m_2}(t)$$
$$\times\Big(1-\frac{1}{K\sigma_h^2\bar{\beta}}\sum_{k_1=1}^{K}|h^n_{m_1,k_1}(t)|^2-\frac{1}{K\sigma_h^2\bar{\beta}}\sum_{k_2=1}^{K}|h^n_{m_2,k_2}(t)|^2$$
$$+\frac{1}{K^2\sigma_h^4\bar{\beta}^2}\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}|h^n_{m_1,k_1}(t)|^2|h^n_{m_2,k_2}(t)|^2\Big)\Big]. \tag{5.23}$$

Using $C_m(t)\leq p$ and $C^2(t)\leq p^2$ and utilizing the i.i.d. channel realizations result in (5.22). □

Lemma 12. $\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\big]\leq\sum_{m\in\mathcal{S}_t}\sum_{\substack{m'\in\mathcal{S}_t\\ m'\neq m}}\frac{\beta_m\beta_{m'}}{K\bar{\beta}^2}\mathbb{E}\big[\,\big\|\Delta\boldsymbol{\theta}_{m'}(t)\big\|_2^2\,\big]$.

Proof. For the real part, using the independence of channels for different m's and

k's, we obtain

$$\begin{aligned}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\big] &= \mathbb{E}\Big[\Big(\sum_{m\in\mathcal{S}_t}\sum_{\substack{m'\in\mathcal{S}_t\\ m'\neq m}}\frac{1}{C(t)\sigma_h^2\bar{\beta}}\\ &\quad\times\sum_{k=1}^{K}\mathrm{Re}\left\{\big(h^n_{m,k}(t)\big)^* h^n_{m',k}(t)C_{m'}(t)\Delta\theta^n_{m',c}(t)\right\}\Big)^2\Big]\\ &\leq \mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\sum_{\substack{m'\in\mathcal{S}_t\\ m'\neq m}}\frac{\beta_m\beta_{m'}}{2K\bar{\beta}^2}\big(\big(\Delta\theta^n_{m',c}(t)\big)^2+\big(\Delta\theta^{n+N}_{m'}(t)\big)^2\\ &\quad+\Delta\theta^n_m(t)\Delta\theta^n_{m'}(t)-\Delta\theta^{n+N}_m(t)\Delta\theta^{n+N}_{m'}(t)\big)\Big] \qquad (5.24)\end{aligned}$$

We obtain a similar expression for $N+1\leq n\leq 2N$, and summing the two parts concludes the lemma. □

Lemma 13. $\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,3}(t)\big)^2\big]\leq\frac{\sigma_z^2 N}{p^2K\sigma_h^2}\sum_{m\in\mathcal{S}_t}\frac{\beta_m}{\bar{\beta}^2}$.

Proof. The first half of the signal yields to

$$\begin{aligned}&\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,3}(t)\big)^2\big]\\ &=\mathbb{E}\Big[\Big(\sum_{m\in\mathcal{S}_t}\sum_{k=1}^{K}\frac{1}{C(t)K\sigma_h^2\bar{\beta}}\,\mathrm{Re}\left\{\big(h^n_{m,k}(t)\big)^* z^n_{PS,k}(t)\right\}\Big)^2\Big]\\ &\leq\frac{1}{p^2K^2\sigma_h^4\bar{\beta}^2}\mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\sum_{k=1}^{K}\big(\mathrm{Re}\left\{\big(h^n_{m,k}(t)\big)^* z^{i,n}_{PS,k}(t)\right\}\big)^2\Big]\\ &\overset{(a)}{=}\frac{\sigma_z^2}{2p^2K\sigma_h^2}\sum_{m\in\mathcal{S}_t}\frac{\beta_m}{\bar{\beta}^2}. \qquad (5.25)\end{aligned}$$

where (a) is obtained using the independence between the channel realizations and the noise. The result also holds for $N+1\leq n\leq 2N$. Summing with respect to all symbols completes the proof. □

The proof is completed using Assumption 3 and (2.2), and summing the results in Lemmas 11-13. □

Lemma 14. $\mathbb{E}\Big[\big\|v(t+1)-\boldsymbol{\theta}^*\big\|_2^2\Big]$

$$\begin{aligned}&\leq(1-\mu\eta(t)(\tau-\eta(t)(\tau-1)))\mathbb{E}\Big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\Big]\\&\quad+(1+\mu(1-\eta(t))\,\eta^2(t)G^2\frac{\tau(\tau-1)(2\tau-1)}{6}\\&\quad+\eta^2(t)(\tau^2+\tau-1)G^2+2\eta(t)(\tau-1)\Gamma.\end{aligned}\tag{5.26}$$

Proof. The proof follows the same line as in Lemma 2 in [23]. □

Lemma 15. $\mathbb{E}\left[\langle\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1),\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\rangle\right]=0.$

Proof. The derivation is the same as in Lemma 3 in [1] by using the independence between local updates and individual channel realizations. □

The proof of the theorem is concluded after applying recursion to the results of Lemmas 10, 14, and 15. □

Corollary 4. *Using Assumption 4, the global loss can be upper bounded after T global iterations as*

$$\begin{aligned}\mathbb{E}\left[F(\boldsymbol{\theta}_{PS}(T))-F^*\right]&\leq\frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}_{PS}(T)-\boldsymbol{\theta}^*\|_2^2\right]\\&\leq\frac{L}{2}\Big(\prod_{n=1}^{T-1}X(n)\Big)\|\boldsymbol{\theta}_{PS}(0)-\boldsymbol{\theta}^*\|_2^2+\frac{L}{2}\sum_{p=1}^{T-1}Y(p)\prod_{n=p+1}^{T-1}X(n).\end{aligned}\tag{5.27}$$

Assuming $\tau=1,\beta_m=1,\forall m\in[M],\eta(t)=\eta,\forall t$ and knowing that $K\gg M$, we get

$$\begin{aligned}\mathbb{E}\left[F\big(\boldsymbol{\theta}_{PS}(T)\big)\right]-F^*&\approx\frac{L}{2}\big(1-\mu\eta\big)^T\|\boldsymbol{\theta}_{PS}(0)-\boldsymbol{\theta}^*\|_2^2\\&\quad+\frac{L}{2\mu\eta}\Big(2\eta^2G^2+\frac{\sigma_z^2N}{p^2K\sigma_h^2}\Big)\Big(1-\big(1-\mu\eta\big)^T\Big).\end{aligned}\tag{5.28}$$

Remark. The noise term in $Y(t)$ does not depend on $\eta(t)$, so we have $\lim_{t\to\infty}\mathbb{E}[F(\boldsymbol{\theta}_{PS}(t))]-F^*\neq0$ even though $\lim_{t\to\infty}\eta(t)=0$. As expected, having more receive antennas and more data contribution from devices increases the convergence rate, whereas the model size and the noise variance have negative effects.

5.2.2 Convergence Analysis for OTA FL with Energy Harvesting Devices with Discrete Battery

Theorem 4. *In OFED, for* $0 \leq \eta(t) \leq min\{1, \frac{1}{\tau_{min}\mu}\}$*, the convergence rate can be upper-bounded as*

$$\mathbb{E}\big[\,\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\|_2^2\,\big] \leq \Big(\prod_{a=1}^{t-1}X(a)\Big)\|\boldsymbol{\theta}_{PS}(0)-\boldsymbol{\theta}^*\|_2^2+\sum_{b=1}^{t-1}Y(b)\prod_{a=b+1}^{t-1}X(a), \tag{5.29}$$

where $X(t)=\mathbb{E}\Big[\frac{1}{|\mathcal{S}_t|}\sum_{m\in\mathcal{S}_t}\Big((1-\mu\eta(t)(\tau_m(t)-\eta(t)(\tau_m(t)-1)))\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\Big]$ *and*

$$\begin{aligned}Y(t)=&\,\eta^2(t)G^2\mathbb{E}\Big[\frac{1}{|\mathcal{S}_t|^2}\sum_{m_1\in\mathcal{S}_t}\sum_{m_2\in\mathcal{S}_t}\tau_{m_1}(t)\tau_{m_2}(t)A(m_1,m_2)\Big]\\&+\eta^2(t)G^2\mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\sum_{\substack{m'\in\mathcal{S}_t\\m'\neq m}}\frac{\tau_{m'}^2(t)\beta_m\beta_{m'}X_{m'}^2(t)}{\bar{X}^2(t)K\bar{\beta}^2}\Big]+\frac{\sigma_z^2N}{K\sigma_h^2}\mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\frac{\beta_m}{\bar{X}^2(t)\bar{\beta}^2}\Big]\\&+\mathbb{E}\Big[\frac{1}{|\mathcal{S}_t|}\sum_{m\in\mathcal{S}_t}\Big((1+\mu(1-\eta(t))\,\eta^2(t)G^2\frac{\tau_m(t)(\tau_m(t)-1)(2\tau_m(t)-1)}{6}\\&+\eta^2(t)(\tau_m^2(t)+\tau_m(t)-1)G^2+2\eta(t)(\tau_m(t)-1)\Gamma\Big)\Big].\end{aligned} \tag{5.30}$$

Proof. Let us define an auxiliary variable $\boldsymbol{v}(t+1)\triangleq\boldsymbol{\theta}_{PS}(t)+\Delta\boldsymbol{\theta}_{PS}(t)$. Then, we can write

$$\begin{aligned}\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{\theta}^*\|_2^2&=\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)+\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\|_2^2\\&=\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)\|_2^2+\|\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\|_2^2\\&\quad+2\langle\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1),\boldsymbol{v}(t+1)-\boldsymbol{\theta}^*\rangle.\end{aligned} \tag{5.31}$$

We will be upper-bounding the three terms in (5.31) in the following Lemmas.

Lemma 16. $\mathbb{E}\Big[\big\|\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)\big\|_2^2\Big]$

$$\begin{aligned}\leq&\,\eta^2(t)G^2\mathbb{E}\Big[\frac{1}{|\mathcal{S}_t|^2}\sum_{m_1\in\mathcal{S}_t}\sum_{m_2\in\mathcal{S}_t}\tau_{m_1}(t)\tau_{m_2}(t)A(m_1,m_2)\Big]\\&+\eta^2(t)G^2\mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\sum_{\substack{m'\in\mathcal{S}_t\\m'\neq m}}\frac{\tau_{m'}^2(t)\beta_m\beta_{m'}X_{m'}^2(t)}{\bar{X}^2(t)K\bar{\beta}^2}\Big]+\frac{\sigma_z^2N}{K\sigma_h^2}\mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\frac{\beta_m}{\bar{X}^2(t)\bar{\beta}^2}\Big].\end{aligned} \tag{5.32}$$

where $A(m_1,m_2)=\Big(1-\frac{\beta_{m_1}}{\bar{\beta}}-\frac{\beta_{m_2}}{\bar{\beta}}+\frac{(K+1)\beta_{m_1}\beta_{m_2}}{K\bar{\beta}^2}\Big)$.

Proof. We can write $\Delta\hat{\theta}^n_{PS}(t)=\sum_{p=1}^{3}\Delta\hat{\theta}^n_{PS,p}(t)$, where the terms correspond to signal, interference, and noise terms. The first term in (5.31) can be written as

$$\begin{aligned}\mathbb{E}\big[||\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)||_2^2\big]&=\mathbb{E}\big[\big\|\Delta\hat{\boldsymbol{\theta}}_{PS}(t)-\Delta\boldsymbol{\theta}_{PS}(t)\big\|_2^2\big]\\ =\sum_{n=1}^{2N}(\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,1}(t)-\Delta\theta^n_{PS}(t)\big)^2\big]&+\sum_{p=2}^{3}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,l}(t)\big)^2\big].\end{aligned}\tag{5.33}$$

In the upcoming lemmas, we will provide some expressions for those terms.

Lemma 17. $\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,1}(t)-\Delta\theta^n_{PS}(t)\big)^2\big]$
$$\leq\eta^2(t)G^2\mathbb{E}\Big[\tfrac{1}{|\mathcal{S}_t|^2}\sum_{m_1\in\mathcal{S}_t}\sum_{m_2\in\mathcal{S}_t}\tau_{m_1}(t)\tau_{m_2}(t)A(m_1,m_2)\Big]$$

Proof. For a single symbol, we have

$$\begin{aligned}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,1}(t)-\Delta\theta^n_{PS}(t)\big)^2\big]&=\mathbb{E}\Big[\frac{1}{\bar{X}(t)^2}\sum_{m_1\in\mathcal{S}_t}\sum_{m_2\in\mathcal{S}_t}X_{m_1}(t)X_{m_2}(t)\Delta\theta^n_{m_1}(t)\Delta\theta^n_{m_2}(t)\\ &\times\Big(1-\frac{1}{K\sigma_h^2\bar{\beta}}\sum_{k_1=1}^{K}|h^n_{m_1,k_1}(t)|^2-\frac{1}{K\sigma_h^2\bar{\beta}}\sum_{k_2=1}^{K}|h^n_{m_2,k_2}(t)|^2\\ &+\frac{1}{K^2\sigma_h^4\bar{\beta}^2}\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}|h^n_{m_1,k_1}(t)|^2|h^n_{m_2,k_2}(t)|^2\Big)\Big],\\ =\mathbb{E}\Big[\frac{1}{\bar{X}(t)^2}\sum_{m_1\in\mathcal{S}_t}\sum_{m_2\in\mathcal{S}_t}&X_{m_1}(t)X_{m_2}(t)\Delta\theta^n_{m_1}(t)\Delta\theta^n_{m_2}(t)A(m_1,m_2)\Big],\end{aligned}\tag{5.34}$$

where $A(m_1,m_2)=\Big(1-\frac{\beta_{m_1}}{\bar{\beta}}-\frac{\beta_{m_2}}{\bar{\beta}}+\frac{(K+1)\beta_{m_1}\beta_{m_2}}{K\bar{\beta}^2}\Big)$. Combining for all the symbols, we get

$$\begin{aligned}\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,1}(t)-\Delta\theta^n_{PS}(t)\big)^2\big]&=\sum_{n=1}^{2N}\mathbb{E}\Big[\frac{1}{|\mathcal{S}_t|^2}\sum_{m_1\in\mathcal{S}_t}\sum_{m_2\in\mathcal{S}_t}\Delta\theta^n_{m_1}(t)\Delta\theta^n_{m_2}(t)A(m_1,m_2)\Big]\\ &\overset{(a)}{\leq}\eta^2(t)G^2\mathbb{E}\Big[\frac{1}{|\mathcal{S}_t|^2}\sum_{m_1\in\mathcal{S}_t}\sum_{m_2\in\mathcal{S}_t}\tau_{m_1}(t)\tau_{m_2}(t)A(m_1,m_2)\Big]\end{aligned}\tag{5.35}$$

where (a) is obtained using Assumption 4 and (2.2). □

Lemma 18. $\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\big] \le \eta^2(t)G^2\mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\sum_{\substack{m'\in\mathcal{S}_t\\ m'\neq m}}\frac{\tau^2_{m'}(t)\beta_m\beta_{m'}X^2_{m'}(t)}{\bar{X}^2(t)K\bar{\beta}^2}\Big]$

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

$$\begin{aligned}
&\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\big]\\
&=\mathbb{E}\Big[\Big(\sum_{m\in\mathcal{S}_t}\sum_{\substack{m'\in\mathcal{S}_t\\ m'\neq m}}\frac{1}{\bar{X}(t)K\sigma_h^2\bar{\beta}}\sum_{k=1}^{K}\mathrm{Re}\left\{\big(h^n_{m,k}(t)\big)^* h^n_{m',k}(t)X_{m'}(t)\Delta\theta^{n,cx}_{m'}(t)\right\}\Big)^2\Big]\\
&=\mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\sum_{\substack{m'\in\mathcal{S}_t\\ m'\neq m}}\frac{\beta_m\beta_{m'}}{2\bar{X}^2(t)K\bar{\beta}^2}\big(\big(X_{m'}(t)\Delta\theta^n_{m'}(t)\big)^2+\big(X_{m'}(t)\Delta\theta^{n+N}_{m'}(t)\big)^2\\
&\quad+X_m(t)X_{m'}(t)\Delta\theta^n_m(t)\Delta\theta^n_{m'}(t)-X_m(t)X_{m'}(t)\Delta\theta^{n+N}_m(t)\Delta\theta^{n+N}_{m'}(t)\big)\Big] \qquad (5.36)
\end{aligned}$$

For $N+1 \le n \le 2N$, similarly, we get

$$\begin{aligned}
\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\big]=&\mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\sum_{\substack{m'\in\mathcal{S}_t\\ m'\neq m}}\frac{\beta_m\beta_{m'}}{2\bar{X}^2(t)K\bar{\beta}^2}\big(\big(X_{m'}(t)\Delta\theta^n_{m'}(t)\big)^2+\big(X_{m'}(t)\Delta\theta^{n-N}_{m'}(t)\big)^2\\
&+X_m(t)X_{m'}(t)\Delta\theta^n_m(t)\Delta\theta^n_{m'}(t)-X_m(t)X_{m'}(t)\Delta\theta^{n-N}_m(t)\Delta\theta^{n-N}_{m'}(t)\big)\Big] \qquad (5.37)
\end{aligned}$$

Combining the two parts, we obtain

$$\begin{aligned}
\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\big] &= \sum_{n=1}^{N}\mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\sum_{\substack{m'\in\mathcal{S}_t\\ m'\neq m}}\frac{\beta_m\beta_{m'}X^2_{m'}(t)}{\bar{X}^2(t)K\bar{\beta}^2}\big(\big(\Delta\theta^n_{m'}(t)\big)^2+\big(\Delta\theta^{n+N}_{m'}(t)\big)^2\Big]\\
&= \mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\sum_{\substack{m'\in\mathcal{S}_t\\ m'\neq m}}\frac{\beta_m\beta_{m'}X^2_{m'}(t)}{\bar{X}^2(t)K\bar{\beta}^2}\big\|\Delta\boldsymbol{\theta}_{m'}(t)\big\|_2^2\Big]\\
&\overset{(a)}{\le} \eta^2(t)G^2\mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\sum_{\substack{m'\in\mathcal{S}_t\\ m'\neq m}}\frac{\tau^2_{m'}(t)\beta_m\beta_{m'}X^2_{m'}(t)}{\bar{X}^2(t)K\bar{\beta}^2}\Big] \qquad (5.38)
\end{aligned}$$

where (a) is obtained using Assumption 4 and (2.2). □

Lemma 19. $\sum_{n=1}^{2N}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,3}(t)\big)^2\big] = \frac{\sigma_z^2 N}{K\sigma_h^2}\mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\frac{\beta_m}{\bar{X}^2(t)\bar{\beta}^2}\Big]$

Proof. For $1 \leq n \leq N$, we get

$$\begin{aligned}\mathbb{E}\big[\big(\Delta\hat{\theta}^n_{PS,3}(t)\big)^2\big] &= \mathbb{E}\Big[\Big(\sum_{m\in\mathcal{S}_t}\sum_{k=1}^{K}\frac{1}{\bar{X}(t)K\sigma_h^2\bar{\beta}}\,\mathrm{Re}\left\{\big(h^n_{m,k}(t)\big)^* z^n_{PS,k}(t)\right\}\Big)^2\Big] \\ &= \mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\sum_{k=1}^{K}\frac{1}{\bar{X}^2(t)K^2\sigma_h^4\bar{\beta}^2}\big(\mathrm{Re}\left\{\big(h^n_{m,k}(t)\big)^* z^{i,n}_{PS,k}(t)\right\}\big)^2\Big] \\ &= \frac{\sigma_z^2}{2K\sigma_h^2}\mathbb{E}\Big[\sum_{m\in\mathcal{S}_t}\frac{\beta_m}{\bar{X}^2(t)\bar{\beta}^2}\Big] \qquad (5.39)\end{aligned}$$

The same expression can be obtained for $N+1 \leq n \leq 2N$. Conclusion of the lemma. □

The Lemma can be concluded by combining the results of Lemmas 17-19.

□

Lemma 20. $\mathbb{E}\Big[\big\|v(t+1)-\boldsymbol{\theta}^*\big\|_2^2\Big]$

$$\begin{aligned}&\leq \mathbb{E}\Big[\frac{1}{|\mathcal{S}_t|}\sum_{m\in\mathcal{S}_t}\Big(\big(1-\mu\eta(t)(\tau_m-\eta(t)(\tau_m-1))\big)\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2 \\ &\quad + \left(1+\mu(1-\eta(t))\,\eta^2(t)G^2\frac{\tau_m(\tau_m-1)(2\tau_m-1)}{6}\right. \\ &\quad \left.+\eta^2(t)(\tau_m^2+\tau_m-1)G^2+2\eta(t)(\tau_m-1)\Gamma\right)\Big]. \qquad (5.40)\end{aligned}$$

Proof. The proof follows the same line as in Lemma 2 in [23] when $|\mathcal{S}_t| = M$ and $\tau_m = \tau$. □

Lemma 21. $\mathbb{E}\left[\langle \boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{v}(t+1), \boldsymbol{v}(t+1) - \boldsymbol{\theta}^*\rangle\right] = 0.$

Proof. The derivation is the same as in Lemma 3 in [1] by using the independence between local updates and individual channel realizations. □

Writing the results of Lemmas 16-21 in a recursive manner concludes the theorem. □

Corollary 5. *Using L-smoothness, the global loss function after T global iterations can be upper bounded with*

$$\mathbb{E}\left[F(\boldsymbol{\theta}_{PS}(T)) - F^*\right] \leq \frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}_{PS}(T) - \boldsymbol{\theta}^*\|_2^2\right]$$
$$\leq \frac{L}{2}\Bigg(\prod_{n=1}^{T-1} X(n)\Bigg)\|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 + \frac{L}{2}\sum_{p=1}^{T-1} Y(p)\prod_{n=p+1}^{T-1} X(n). \tag{5.41}$$

5.3 Numerical Examples

5.3.1 Unit Battery Case

In this subsection, we will present several numerical results for OTA FL with energy harvesting devices equipped with unit battery. We consider both Bernoulli and uniform energy arrivals, and compare the results with some baseline schemes.

We consider an FL environment with $M = 40$ MUs and a PS with $K = 5M$ receive antennas. MUs are spread around the PS randomly in such a way that their distances to the PS is uniformly distributed between 0.5 and 2 units.

We use the CIFAR-10 dataset [69] with Adam optimizer [70], and consider an i.i.d. data distribution where the data samples are randomly and equally distributed among MUs. The same architecture as the one presented in Table 4.1 is used with $2N = 307498$.

We study the performance of conventional FL (without any communications constraints), OTA FL where all the MUs have available energy to participate at all iterations, and energy harvesting FL where MUs have intermittent energy arrivals with both error-free and OTA aggregation schemes. To make a comparison with the previous studies, we also consider the setup used in [66] with Bernoulli energy arrivals, which corresponds to the energy harvesting FL setup with no channel errors without any normalization at the PS with respect to the cooldown multipliers. Moreover, the MUs are divided into 4 equal-sized groups with different energy profiles. For Bernoulli energy arrivals, we have $\alpha_m(t) \in \{1, 1/5, 1/10, 1/20\}$, and

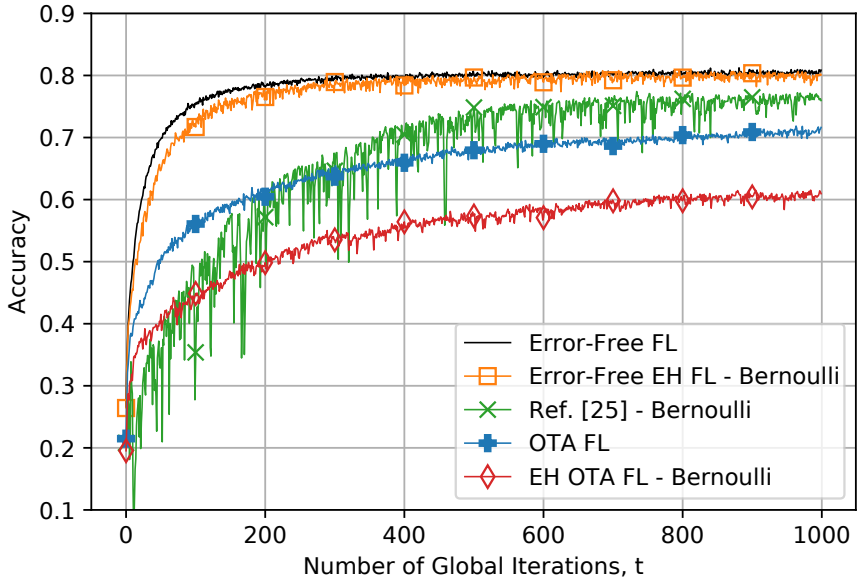

Figure 5.1: Energy harvesting OTA FL test accuracy for $\tau = 1$

for uniform energy arrivals, we have $T_m \in \{1, 5, 10, 20\}$ for MUs in 4 groups as in [66]. The training is performed for $T = 1000$ global iterations for $\tau = 1$, and $T = 400$ for $\tau = 3$ with mini-batch size $|\boldsymbol{\xi}^i_{m,c}(t)| = 128$, the path loss exponent $p = 4$, $\sigma_h^2 = 1$, and $\sigma_z^2 = 1$.

Accuracy plots for the case with Bernoulli energy arrival profiles with τ=1 and τ=3 are presented in Figs. 5.1 and 5.2, respectively. The results show that the energy harvesting FL with error-free links has a convergence rate close to that of FL with full participation, and that adding a normalization term with respect to the cooldown multipliers leads to a faster convergence and less fluctuations compared to the results in [66]. Moreover, OTA FL performance is very similar to the scenario used in [66] with error-free links. It can be seen that even though the links are wireless, the gap in the performance can be compensated as the number of global iterations increases. One reason is that the increased number of receive antennas at the PS can reduce the adverse affects of the small-scale fading and noise. Increasing τ achieves a better performance with faster convergence at the cost of making more computations at the edge. It can also be observed that the performance of Bernoulli arrivals is very close to the that of the uniform arrivals due to the similarities in the energy arrival profiles.

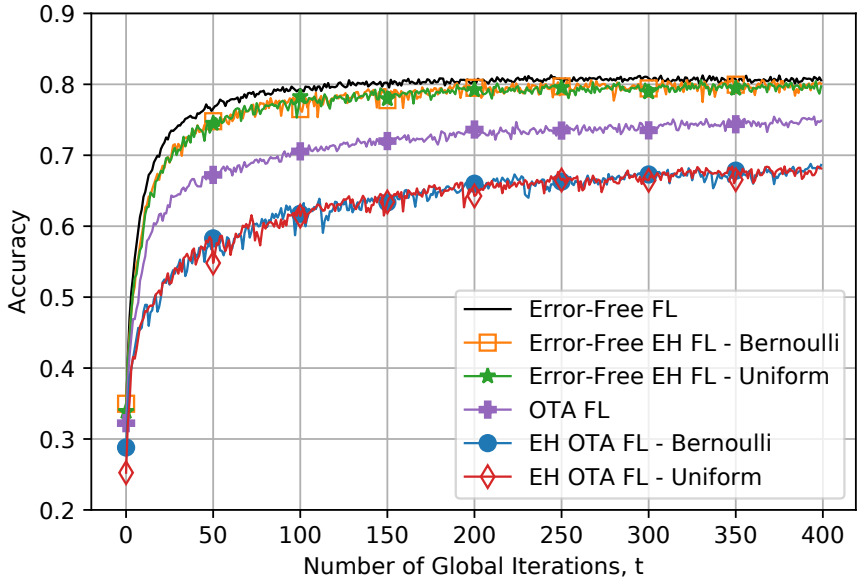

Figure 5.2: Energy harvesting OTA FL test accuracy for $\tau = 3$

In Fig. 5.3, we numerically evaluate the convergence rates of the scenarios considered in Fig. 5.2, using the expression in (5.27) with $M=40, 2N=307498, L=10, \mu=1, \tau=1, G^2=1, \eta(t)=10^{-2}-10^{-6}t, \sigma_z^2=5, \sigma_h^2=1, K=M, \|\boldsymbol{\theta}_{PS}(0)-\boldsymbol{\theta}^*\|_2^2=10^3$. We observe a close convergence rate between the conventional FL and the error-free energy harvesting FL as expected due to weighted averaging operation with respect to the cooldown multipliers. Energy harvesting FL with OTA aggregation has a slower convergence rate when compared to the others because of the wireless channel effects as well as the decreased number of participants at each iteration due to energy harvesting devices. We can observe that changing the energy arrival profiles and introducing MUs with less frequent energy arrivals affect $|\mathcal{S}_t|$ and $C(t)$, which are key reasons in the shifts and fluctuations of the convergence rates.

5.3.2 The Case With with Discrete Battery

In this subsection, we present our results on OTA FL with energy harvesting MDs, which receive finite levels of energy at each iteration, without any battery storage capabilities for subsequent iterations.

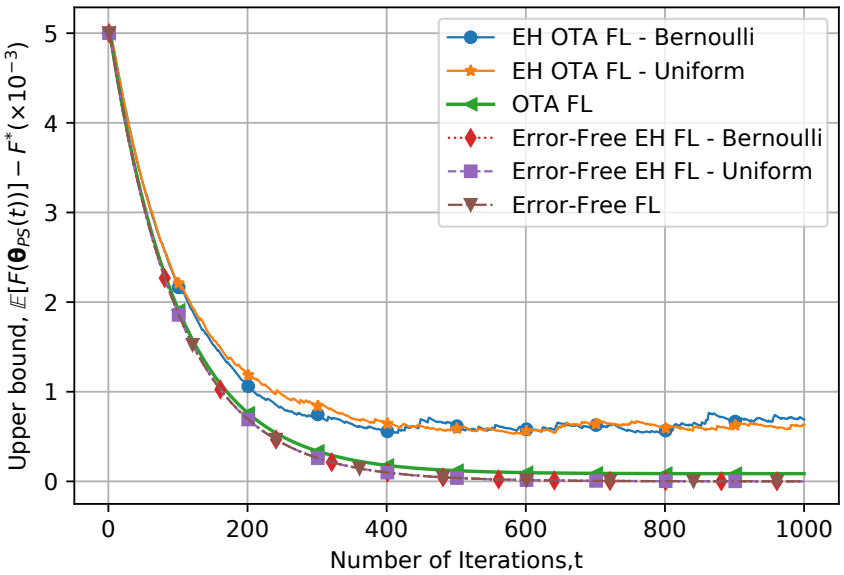

Figure 5.3: Upper bound on $\mathbb{E}\left[F(\boldsymbol{\theta}_{PS}(t)) - F^*\right]$

We consider an FL system with $M = 20$ MDs, each with a single antenna, and a PS with $K = 5M$ receive antennas. MDs are randomly placed on a circular area in such a way that their distances to the PS is between 0.5 and 2. Each MD is an energy harvesting device with no storage capability for the later iterations, $E_{tr,min} = 18$, $E_{sgd} = 3$, $\tau_{min} = 1$, $\tau_{max} = 3$ units, and all with the same energy arrival parameter $\lambda_m = \lambda = 20$ units.

We use the CIFAR-10 dataset [69] with Adam optimizer [70], and we examine the case with i.i.d. data distribution where the data samples are distributed randomly among users. As a network architecture, we use the convolutional neural network (CNN) architecture given in Table 4.1 with $2N = 307498$.

In the simulations, we consider the FL setup with full participation where the devices have the enough battery power to participate in all the iterations, transmission-greedy OFED with $\tau = \tau_{min}$, and SGD-greedy OFED. To make a comparison between the different transmission schemes, we consider these cases both with OTA aggregation and with error-free links.

In Fig. 5.4, we perform numerical evaluations of the upper bound obtained in Corollary 5 for the same location placement with $M = 20$, $K = M$, $2N =$

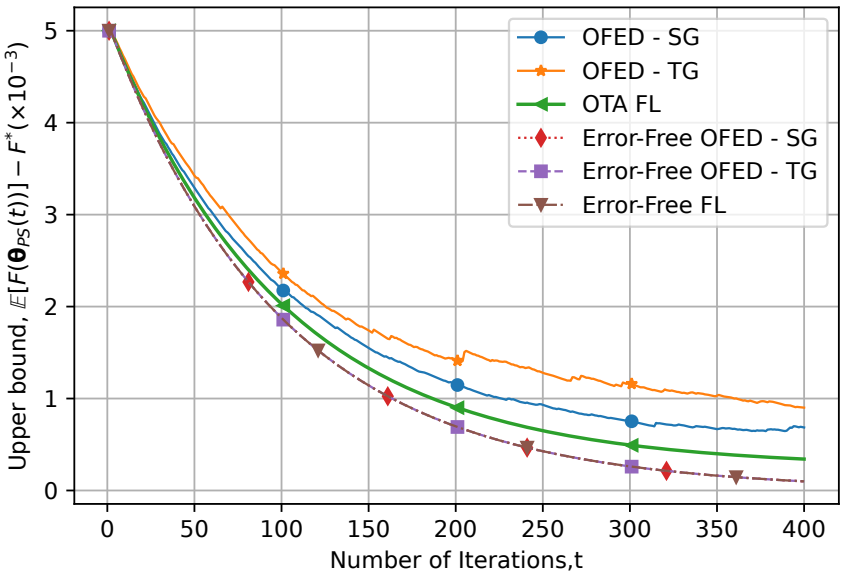

Figure 5.4: Upper bound on $\mathbb{E}\left[F(\boldsymbol{\theta}_{PS}(t)) - F^*\right]$

307498, $\sigma_z^2 = 5$, $\sigma_h^2 = 1$, $\mu = 1$, $L = 10$, $G^2 = 1$, $\Gamma = 1$, $\eta(t) = 10^{-2} - 10^{-6}t$, $p = 4$, $\|\boldsymbol{\theta}_{PS}(0) - \boldsymbol{\theta}^*\|_2^2 = 10^3$, and energy harvesting MDs with $E_{tr,min} = 18$, $E_{sgd} = 3$, $\tau_{min} = 1$, $\tau_{max} = 3$, and $\lambda_m = \lambda = 20$ units. The results show that the SGD-greedy OFED outperforms the transmission-greedy OFED in both error-free and wireless schemes. In the error-free cases, all the schemes give a convergence rate very close to the baseline. It can also be observed that the convergence rate of SGD-greedy OFED is very close to the full-participation FL with wireless transmission even though OFED performs global aggregation with partial participation of MDs.

5.4 Chapter Summary

In this chapter, we studied OTA FL with energy harvesting devices with intermittent and heterogeneous energy arrivals. Our framework consists of local SGD computations at the MUs that have available energy, and OTA aggregation of the gradients over a shared wireless medium. A comparison of the performance of the OTA FL with energy harvesting devices through neural network simulations and an analysis of its convergence rate are performed through numerical experiments.

The results with different energy profiles demonstrate that performing a weighted averaging using the latest energy arrival and dataset cardinality in energy harvesting FL can give a similar performance to the full-participation scheme in both error-free and OTA cases.

We also considered OTA FL using energy harvesting devices that require a certain amount of energy for both local SGD computations and gradient transmissions. In this setting, the MDs receive discrete levels of energy based on a point Poisson process at the beginning of every global iteration. We assume that they do not have the capability of storing energy for later iterations. The MDs with sufficient energy compute the local gradients and send them through a wireless channel for OTA aggregation. The PS combines the received signals and performs normalization to obtain a noisy estimate of the aggregated gradients. Through a theoretical analysis of the proposed schemes, we obtain upper bounds on their convergence rates. The results show that the SGD-greedy approach has a faster convergence rate than the transmission-greedy approach when the energy requirements are close to the energy arrival profiles of the MDs. A possible future research direction could be to optimize the energy allocation between the SGD computations and the transmit power so that the available energy can be used more efficiently.

Chapter 6

Conclusions and Future Directions

In this thesis, our main focus is on federated learning over wireless channels. Edge users called mobile users perform local computations using local data, and the transmitted gradients are combined at the receiver using over-the-air aggregation. We propose an FL structure with hierarchical clustering where intermediate servers (ISs) are employed around the areas where the MUs are more densely located. The MUs perform multiple stochastic gradient descent (SGD) iterations, and send their gradients using OTA aggregation to their corresponding cluster IS. After multiple cluster aggregations, ISs send their model updates to the parameter server (PS) through error-free links. Through numerical and experimental analysis, we observe that bringing server-side closer to the MUs provides a faster convergence and better performance.

We also extend our study on the hierarchical FL to a more practical scenario where both the MUs and ISs send their gradients through OTA aggregation, taking into account the effect of interference coming from other clusters on the cluster aggregations. Our numerical analysis shows that our wireless hierarchical FL setup gives a better performance while using less transmit power than the conventional FL setup where the MUs directly communicate with the PS. It

is also shown that global aggregation can be more important than the cluster aggregation depending on the data distribution of the MUs.

Finally, we study the OTA FL with energy harvesting devices, which harvest energy from their ambient environment in intermittent time arrivals. We consider a wireless setup for the energy harvesting FL where the participating users with the available energy transmit their gradients through OTA aggregation to the PS for the global aggregation. In order to compensate for heterogeneous energy arrival times among different users, we introduce a cooldown multiplier to the gradients to amplify them according to their importance. We show through experimental and numerical results with different energy arrival profiles that our proposed strategy performs better than the previous error-free approaches, and give a slightly worse performance than the OTA FL with full participation.

There are many future research directions that can be followed by building on the ideas developed in this thesis. For example, one can extend the idea of the hierarchical over-the-air FL into a more practical setup with multiple levels of hierarchy, i.e., by employing additional OTA communications at different levels. The main point of interest in this line of research could be the analysis of the optimal number of hierarchical layers for the best performance while providing the corresponding cost of adopting more layers. Moreover, it is important to analyze the heterogeneity among clusters where the number of MUs in each cluster is not fixed. Another idea regarding the hierarchical FL is to use the MUs as an IS since employing ISs will have an additional cost. A possible extension to this idea is to assign some users in every cluster as ISs, where the assignment can change based on the available energy and the distance to other MUs.

It is important to investigate the relationship between the required energy to transmit the gradients and the cooldown multipliers for the energy harvesting FL. Energy harvesting FL systems require an amplified participation from the devices that participate less frequently to prevent bias, however, the sporadically participating devices might not have enough energy to send amplified gradients. Therefore, it is important to obtain the optimal amplification coefficient for a certain energy arrival profile. In this framework, another possible direction for

future research may be to define the value of a local update before transmitting the gradient, and using this value to improve the performance of the algorithm. This approach may help mobile users with less valuable local gradients to save up energy to increase their chances to participate in the later iterations since their current gradients might not be as effective as the others to the global model accuracy. Then, the global aggregation can be performed with mobile users whose local gradients contribute the most to the global model.

Bibliography

[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] O. Aygün, M. Kazemi, D. Gündüz, and T. M. Duman, "Over-the-air federated learning with energy harvesting devices," *arXiv preprint arXiv:2205.12869*, 2022.

[3] O. Aygün, M. Kazemi, D. Gündüz, and T. M. Duman, "Over-the-air federated edge learning with hierarchical clustering," *arXiv preprint arXiv:2207.09232*, 2022.

[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*, pp. 1273–1282, PMLR, 2017.

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

[6] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[7] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International journal of computer vision*, vol. 40, no. 2, pp. 99–121, 2000.

[8] X. Yao, C. Huang, and L. Sun, "Two-stream federated learning: Reduce the communication costs," in *2018 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4, IEEE, 2018.

[9] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*, pp. 97–105, PMLR, 2015.

[10] Z. Tao and Q. Li, "{eSGD}: Communication efficient distributed deep learning on the edge," in *USENIX Workshop on Hot Topics in Edge Computing (HotEdge 18)*, 2018.

[11] W. Luping, W. Wei, and L. Bo, "Cmfl: Mitigating communication overhead for federated learning," in *2019 IEEE 39th international conference on distributed computing systems (ICDCS)*, pp. 954–964, IEEE, 2019.

[12] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 691–706, IEEE, 2019.

[13] C. Fung, C. J. Yoon, and I. Beschastnikh, "Mitigating sybils in federated learning poisoning," *arXiv preprint arXiv:1808.04866*, 2018.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[15] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: information leakage from collaborative deep learning," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pp. 603–618, 2017.

[16] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

[17] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*, pp. 265–284, Springer, 2006.

[18] Y. Aono, T. Hayashi, L. Wang, S. Moriai, *et al.*, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1333–1345, 2017.

[19] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[20] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," in *International Conference on Machine Learning*, pp. 634–643, PMLR, 2019.

[21] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International Conference on Artificial Intelligence and Statistics*, pp. 2938–2948, PMLR, 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] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2022–2035, 2020.

[25] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Cotaf: Convergent over-the-air federated learning," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pp. 1–6, IEEE, 2020.

[26] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, 2019.

[27] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18914–18919, 2009.

[28] 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), pp. 1–6, Jun. 2021.

[29] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "Uveqfed: Universal vector quantization for federated learning," *IEEE Transactions on Signal Processing*, vol. 69, pp. 500–514, 2020.

[30] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[31] 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*, pp. 2021–2031, PMLR, 2020.

[32] 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.

[33] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified sgd with memory," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[34] B. Tegin and T. M. Duman, "Blind federated learning at the wireless edge with low-resolution ADC and DAC," *IEEE Trans. Wireless Commun.*, 2021.

[35] 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), pp. 306–310, Sept. 2021.

[36] H. H. Yang, Z. Liu, T. Q. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE transactions on communications*, vol. 68, no. 1, pp. 317–333, 2019.

[37] 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.

[38] J. Ren, Y. He, D. Wen, G. Yu, K. Huang, and D. Guo, "Scheduling for cellular federated edge learning with importance and channel awareness," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7690–7703, 2020.

[39] H. H. Yang, A. Arafa, T. Q. Quek, and H. V. Poor, "Age-based scheduling policy for federated learning in mobile edge networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8743–8747, IEEE, 2020.

[40] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *ICC 2019-2019 IEEE international conference on communications (ICC)*, pp. 1–7, IEEE, 2019.

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

[42] B. Tegin and T. M. Duman, "Federated learning over time-varying channels," in *GLOBECOM 2021-2021 IEEE Global Communications Conference*, (Madrid, Spain), Dec. 2021.

[43] J.-H. Ahn, O. Simeone, and J. Kang, "Cooperative learning via federated distillation over fading channels," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8856–8860, IEEE, 2020.

[44] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, "Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data," *arXiv preprint arXiv:1811.11479*, 2018.

[45] 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.

[46] M. M. Amiri, T. M. Duman, and D. Gündüz, "Collaborative machine learning at the wireless edge with blind transmitters," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1–5, IEEE, 2019.

[47] 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), pp. 8866–8870, May 2020.

[48] 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), pp. 1–9, Sep. 2020.

[49] 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), pp. 1–6, Jun. 2020.

[50] 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.

[51] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Hierarchical quantized federated learning: Convergence analysis and system design," *arXiv preprint arXiv:2103.14272*, 2021.

[52] 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.

[53] T. Castiglia, A. Das, and S. Patterson, "Multi-level local sgd: Distributed sgd for heterogeneous hierarchical networks," in *International Conference on Learning Representations*, 2020.

[54] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," *arXiv preprint arXiv:1906.02243*, 2019.

[55] S. Ulukus, A. Yener, E. Erkip, O. Simeone, M. Zorzi, P. Grover, and K. Huang, "Energy harvesting wireless communications: A review of recent advances," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 3, pp. 360–381, 2015.

[56] J. A. Paradiso and T. Starner, "Energy scavenging for mobile and wireless electronics," *IEEE Pervasive computing*, vol. 4, no. 1, pp. 18–27, 2005.

[57] V. Sharma, U. Mukherji, V. Joseph, and S. Gupta, "Optimal energy management policies for energy harvesting sensor nodes," *IEEE Transactions on Wireless Communications*, vol. 9, no. 4, pp. 1326–1336, 2010.

[58] S. Luo, R. Zhang, and T. J. Lim, "Optimal save-then-transmit protocol for energy harvesting wireless transmitters," *IEEE Transactions on Wireless Communications*, vol. 12, no. 3, pp. 1196–1207, 2013.

[59] C. K. Ho and R. Zhang, "Optimal energy allocation for wireless communications powered by energy harvesters," in *2010 IEEE International Symposium on Information Theory*, pp. 2368–2372, IEEE, 2010.

[60] K. Tutuncuoglu and A. Yener, "Optimum transmission policies for battery limited energy harvesting nodes," *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 1180–1189, 2012.

[61] O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, and A. Yener, "Transmission with energy harvesting nodes in fading wireless channels: Optimal policies,"

IEEE Journal on selected areas in communications, vol. 29, no. 8, pp. 1732–1743, 2011.

[62] K. Tutuncuoglu and A. Yener, "Optimal power policy for energy harvesting transmitters with inefficient energy storage," in *2012 46th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, IEEE, 2012.

[63] O. Ozel, J. Yang, and S. Ulukus, "Optimal broadcast scheduling for an energy harvesting rechargeable transmitter with a finite capacity battery," *IEEE Transactions on Wireless Communications*, vol. 11, no. 6, pp. 2193–2203, 2012.

[64] B. Guler and A. Yener, "Sustainable federated learning," *arXiv preprint arXiv:2102.11274*, 2021.

[65] R. Hamdi, M. Chen, A. B. Said, M. Qaraqe, and H. V. Poor, "User scheduling in federated learning over energy harvesting wireless networks," in *2021 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, IEEE, 2021.

[66] B. Güler and A. Yener, "Energy-harvesting distributed machine learning," in *2021 IEEE Intl. Symp. Inf. Theory (ISIT)*, pp. 320–325, IEEE, 2021.

[67] R. Hamdi, M. Chen, A. B. Said, M. Qaraqe, and H. V. Poor, "Federated learning over energy harvesting wireless networks," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 92–103, 2021.

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

[69] A. Krizhevsky *et al.*, "Learning multiple layers of features from tiny images," 2009.

[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[71] L. Liu and W. Yu, "Massive connectivity with massive mimo—part i: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, 2018.

Appendix A

Proof of Lemma 5 & 8

We have

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

Where the second term can be bounded as

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

where (a) is due to the convexity of $\|\|_2^2$, (b) comes from utilizing (4.22), and

(c) is obtained using Assumption 2. Plugging in the result to (A.1), we have

$$\mathbb{E}\big[\big\|v(t+1)-\boldsymbol{\theta}^*\big\|_2^2\big] \leq \mathbb{E}\big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\big] + \eta^2(t) I \tau^2 G^2 + 2\mathbb{E}\big[\langle \boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*, \Delta\boldsymbol{\theta}_{PS}(t)\rangle\big]. \tag{A.3}$$

The last term of (A.3), we have

$$\begin{aligned}
2\mathbb{E}\big[\langle \boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*, \Delta\boldsymbol{\theta}_{PS}(t)\rangle\big] &= \frac{2}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\mathbb{E}\big[\langle \boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*, \Delta\boldsymbol{\theta}_{c,m}^{i}(t)\rangle\big] \\
&= \frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\mathbb{E}\Big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}_{PS}(t), \sum_{j=1}^{\tau}\nabla F_{c,m}\big(\boldsymbol{\theta}_{c,m}^{i,j}(t), \boldsymbol{\xi}_{c,m}^{i,j}(t)\big)\rangle\Big] \\
&= \frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\mathbb{E}\big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}_{PS}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}_{PS}(t), \boldsymbol{\xi}_{c,m}^{1,1}(t)\big)\rangle\big] \\
&\quad + \frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\mathbb{E}\Big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}_{PS}(t), \sum_{j=2}^{\tau}\nabla F_{c,m}\big(\boldsymbol{\theta}_{c,m}^{i,j}(t), \boldsymbol{\xi}_{c,m}^{i,j}(t)\big)\rangle\Big].
\end{aligned} \tag{A.4}$$

For the first term of (A.4), we have

$$\begin{aligned}
&\frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\mathbb{E}\big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}_{PS}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}_{PS}(t), \boldsymbol{\xi}_{c,m}^{1,1}(t)\big)\rangle\big] \\
&\quad\overset{(a)}{=} \frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\mathbb{E}\big[\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}_{PS}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}_{PS}(t)\big)\rangle\big] \\
&\quad\overset{(b)}{\leq} \frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\mathbb{E}\big[F_{c,m}(\boldsymbol{\theta}^*) - F_{c,m}(\boldsymbol{\theta}_{PS}(t)) - \frac{\mu}{2}\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\big] \\
&\quad= 2\eta(t) I\Big(F^* - \mathbb{E}\big[F(\boldsymbol{\theta}_{PS}(t))\big] - \frac{\mu}{2}\mathbb{E}\big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\big]\Big) \\
&\quad\overset{(c)}{\leq} -\eta(t) I\mu\mathbb{E}\big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\big\|_2^2\big],
\end{aligned} \tag{A.5}$$

where (a) comes from $\mathbb{E}_{\xi}\left[\nabla F_{m,c}\left(\boldsymbol{\theta}_{PS}(t), \boldsymbol{\xi}_{m,c}^{1,1}(t)\right)\right] = \nabla F_{m,c}(\boldsymbol{\theta}_{PS}(t))$, (b) holds due to the μ-strong convexity of $F_{m,c}$, and (c) follows since $F^* \leq F(\boldsymbol{\theta}(t))$. For the second term of (A.4), we have

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

$$= \frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\big[\langle \boldsymbol{\theta}_{c,m}^{i,j}(t) - \boldsymbol{\theta}_{PS}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}_{c,m}^{i,j}(t), \boldsymbol{\xi}_{c,m}^{i,j}(t)\big)\rangle\big]$$
$$+ \frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\big[\langle \boldsymbol{\theta}^{*} - \boldsymbol{\theta}_{c,m}^{i,j}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}_{c,m}^{i,j}(t), \boldsymbol{\xi}_{c,m}^{i,j}(t)\big)\rangle\big]. \tag{A.6}$$

Using Cauchy-Schwarz inequality, we obtain

$$\frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\big[\langle \boldsymbol{\theta}_{c,m}^{i,j}(t) - \boldsymbol{\theta}_{PS}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}_{c,m}^{i,j}(t), \boldsymbol{\xi}_{c,m}^{i,j}(t)\big)\rangle\big]$$
$$\leq \frac{\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\bigg[\frac{1}{\eta(t)}\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t) - \boldsymbol{\theta}_{PS}(t)\big\|_2^2 + \eta(t)\big\|\nabla F_{c,m}\big(\boldsymbol{\theta}_{c,m}^{i,j}(t), \boldsymbol{\xi}_{c,m}^{i,j}(t)\big)\big\|_2^2\bigg]$$
$$\overset{(a)}{\leq} \frac{1}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\bigg[\frac{1}{\eta(t)}\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t) - \boldsymbol{\theta}_{PS}(t)\big\|_2^2\bigg] + \eta^2(t)I(\tau-1)G^2, \tag{A.7}$$

where (a) is obtained using Assumption 2. The following lemma will give an upper bound on the second term of (A.6).

Lemma 22. $$\frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\big[\langle \boldsymbol{\theta}^{*} - \boldsymbol{\theta}_{c,m}^{i,j}(t), \nabla F_{c,m}\big(\boldsymbol{\theta}_{c,m}^{i,j}(t), \boldsymbol{\xi}_{c,m}^{i,j}(t)\big)\rangle\big]$$
$$= -\mu\eta(t)(1-\eta(t))I(\tau-1)\mathbb{E}\big[\big\|\boldsymbol{\theta}_{PS}(t) - \boldsymbol{\theta}^{*}\big\|_2^2\big]$$
$$+ \frac{\mu(1-\eta(t))}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\big[\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t) - \boldsymbol{\theta}_{PS}(t)\big\|_2^2\big] + 2\eta(t)I(\tau-1)\Gamma.$$

Proof. We have

$$\begin{aligned}
&\frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\big[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}_{c,m}^{i,j}(t),\nabla F_{c,m}\big(\boldsymbol{\theta}_{c,m}^{i,j}(t),\boldsymbol{\xi}_{c,m}^{i,j}(t)\big)\rangle\big]\\
&\overset{(a)}{\leq}\frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\big[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}_{c,m}^{i,j}(t),\nabla F_{c,m}\big(\boldsymbol{\theta}_{c,m}^{i,j}(t)\big)\rangle\big]\\
&\overset{(b)}{\leq}\frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\Big[F_{c,m}(\boldsymbol{\theta}^*)-F_{c,m}(\boldsymbol{\theta}_{c,m}^{i,j}(t))-\frac{\mu}{2}\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t)-\boldsymbol{\theta}^*\big\|_2^2\Big]\\
&=\frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\Big[F_{c,m}(\boldsymbol{\theta}^*)-F_{c,m}^*+F_{c,m}^*-F_{c,m}(\boldsymbol{\theta}_{c,m}^{i,j}(t))-\frac{\mu}{2}\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t)-\boldsymbol{\theta}^*\big\|_2^2\Big]\\
&=2\eta(t)I(\tau-1)\Big(F^*-\frac{1}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}F_{c,m}^*\Big)+\frac{2\eta}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\Big(F_{c,m}^*-\mathbb{E}\left[F_{c,m}(\boldsymbol{\theta}_{c,m}^{i,j})\right]\Big)\\
&\quad-\frac{\mu\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\big[\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t)-\boldsymbol{\theta}^*\big\|_2^2\big]\\
&\overset{(c)}{\leq}2\eta(t)I(\tau-1)\Gamma-\frac{\mu\eta}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\big[\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t)-\boldsymbol{\theta}^*\big\|_2^2\big],
\end{aligned}\tag{A.8}$$

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

$$\begin{aligned}
-\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t)-\boldsymbol{\theta}^*\big\|_2^2&=-\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t)-\boldsymbol{\theta}_{PS}(t)\big\|_2^2-\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\|_2^2-2\langle\boldsymbol{\theta}_{c,m}^{i,j}(t)-\boldsymbol{\theta}_{PS}(t),\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\rangle\\
&\overset{(a)}{\leq}-\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t)-\boldsymbol{\theta}_{PS}(t)\big\|_2^2-\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\|_2^2+\frac{1}{\eta(t)}\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t)-\boldsymbol{\theta}_{PS}(t)\big\|_2^2+\eta(t)\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\|_2^2\\
&=-(1-\eta(t))\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^*\|_2^2+\left(\frac{1}{\eta(t)}-1\right)\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t)-\boldsymbol{\theta}_{PS}(t)\big\|_2^2,
\end{aligned}\tag{A.9}$$

where (a) is due to Cauchy-Schwarz inequality. Plugging (A.8) and (A) concludes the Lemma. □

Using the results in (A.7) and , we can write (A.6) as

$$\begin{aligned}
&\frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\mathbb{E}\Big[\langle\boldsymbol{\theta}^{*}-\boldsymbol{\theta}_{PS}(t),\sum_{j=2}^{\tau}\nabla F_{c,m}\big(\boldsymbol{\theta}_{c,m}^{i,j}(t),\boldsymbol{\xi}_{c,m}^{i,j}(t)\big)\rangle\Big]\\
&\quad=-\mu\eta(t)(1-\eta(t))I(\tau-1)\mathbb{E}\big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^{*}\big\|_2^2\big]\\
&\qquad+\frac{(1+\mu(1-\eta(t)))}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\big[\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t)-\boldsymbol{\theta}_{PS}(t)\big\|_2^2\big]\\
&\qquad+\eta^2(t)I(\tau-1)G^2+2\eta(t)I(\tau-1)\Gamma. \qquad \text{(A.10)}
\end{aligned}$$

Also, we have

$$\begin{aligned}
&\frac{1}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\big[\big\|\boldsymbol{\theta}_{c,m}^{i,j}(t)-\boldsymbol{\theta}_{PS}(t)\big\|_2^2\big]\\
&\qquad=\frac{\eta^2}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=2}^{\tau}\mathbb{E}\big[\big\|\nabla F_{c,m}\big(\boldsymbol{\theta}_{c,m}^{i,j}(t),\boldsymbol{\xi}_{c,m}^{i,j}(t)\big)\big\|_2^2\big]\\
&\qquad\overset{(a)}{\leq}\eta^2IG^2\frac{\tau(\tau-1)(2\tau-1)}{6}, \qquad \text{(A.11)}
\end{aligned}$$

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

$$\begin{aligned}
&\frac{2\eta(t)}{MC}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{i=1}^{I}\mathbb{E}\Big[\langle\boldsymbol{\theta}^{*}-\boldsymbol{\theta}_{PS}(t),\sum_{j=2}^{\tau}\nabla F_{c,m}\big(\boldsymbol{\theta}_{c,m}^{i,j}(t),\boldsymbol{\xi}_{c,m}^{i,j}(t)\big)\rangle\Big]\\
&\quad\leq\mu\eta(t)(1-\eta(t))I(\tau-1)\mathbb{E}\big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^{*}\big\|_2^2\big]\\
&\qquad+\big(1+\mu(1-\eta(t))\big)\eta^2(t)IG^2\frac{\tau(\tau-1)(2\tau-1)}{6}\\
&\qquad+\eta^2(t)I(\tau-1)G^2+2\eta(t)I(\tau-1)\Gamma. \qquad \text{(A.12)}
\end{aligned}$$

Substituting the results in (A.5) and (A.12) into (A.4), we get

$$\begin{aligned}
2\mathbb{E}\big[\langle\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^{*},\Delta\boldsymbol{\theta}_{PS}(t)\rangle\big]&\leq\mu\eta(t)I(\tau-\eta(t)(\tau-1))\mathbb{E}\big[\big\|\boldsymbol{\theta}_{PS}(t)-\boldsymbol{\theta}^{*}\big\|_2^2\big]\\
&\quad+\big(1+\mu(1-\eta(t))\big)\eta^2(t)IG^2\frac{\tau(\tau-1)(2\tau-1)}{6}\\
&\quad+\eta^2(t)I(\tau-1)G^2+2\eta(t)I(\tau-1)\Gamma. \qquad \text{(A.13)}
\end{aligned}$$

Lemma is concluded by plugging (A.13) into (A.3).

Appendix B

Proof of Lemma 7

Using (4.25), we have

$$\mathbb{E}\left[||\boldsymbol{\theta}_{PS}(t+1)-\boldsymbol{v}(t+1)||_2^2\right]=\mathbb{E}\left[||\Delta\hat{\boldsymbol{\theta}}_{PS}(t)-\Delta\boldsymbol{\theta}_{PS}(t)||_2^2\right], \tag{B.1}$$

$$=\sum_{n=1}^{2N}\mathbb{E}\left[(\Delta\hat{\theta}^n_{PS}(t)-\Delta\theta^n_{PS}(t))^2\right]. \tag{B.2}$$

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 the noise, we can write

$$\mathbb{E}\left[||\Delta\hat{\theta}^n_{PS}(t)-\Delta\theta^n_{PS}(t)||_2^2\right]=\mathbb{E}\left[\left(\Delta\hat{\theta}^n_{PS,1}(t)-\Delta\theta^n_{PS}(t)\right)^2\right]+\sum_{l=2}^{9}\mathbb{E}\left[\left(\Delta\hat{\theta}^n_{PS,l}(t)\right)^2\right]. \tag{B.3}$$

Lemma 23. $\mathbb{E}\Big[\Delta\theta^{i_1,n}_{c_1,m_1}(t)\Delta\theta^{i_2,n}_{c_2,m_2}(t)\Big]\leq\eta^2(t)G^2\tau^2$

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

where (a) holds due to Assumption 2. □

Lemma 24. $\mathbb{E}\Big[\big\|\Delta\boldsymbol{\theta}^i_{c,m}(t)\big\|_2^2\Big]\leq\eta^2(t)G^2\tau^2$

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 25. $\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,1}(t)-\Delta\theta^n_{PS}(t)\big)^2\Big]$

$$\leq \frac{\eta^2(t)G^2I^2\tau^2}{M^2C^2}\sum_{c_1=1}^{C}\sum_{c_2=1}^{C}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}A(m_1,m_2,c_1,c_2).$$

Proof. Using (2.20) and (4.18), we have

$$\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,1}(t)-\Delta\theta^n_{PS}(t)\big)^2\Big]$$
$$=\mathbb{E}\Bigg[\bigg(\frac{1}{MC}\sum_{c,m,i}\Delta\theta^{i,n}_{c,m}(t)\Big(\Big(\frac{1}{KK'\sigma_h^4\bar{\beta}\bar{\beta}_c}\sum_{k,k'}|h^n_{IS,c,k'}(t)|^2|h^{i,n}_{c,m,c,k}(t)|^2\Big)-1\Big)\bigg)^2\Bigg],$$

$$=\frac{1}{M^2C^2}\mathbb{E}\Bigg[\sum_{\substack{c_1,c_2,m_1,\\ m_2,i_1,i_2}}\Delta\theta^{i_1,n}_{c_1,m_1}(t)\Delta\theta^{i_2,n}_{c_2,m_2}(t)\Big(1-\frac{2}{KK'\sigma_h^4\bar{\beta}\bar{\beta}_{c_1}}\sum_{k_1,k_1'}|h^n_{IS,c_1,k_1'}(t)|^2|h^{i_1,n}_{c_1,m_1,c_1,k_1}(t)|^2$$
$$+\frac{1}{K^2(K')^2\sigma_h^8\bar{\beta}^2\bar{\beta}_{c_1}\bar{\beta}_{c_2}}\sum_{\substack{k_1,k_1',\\ k_2,k_2'}}|h^n_{IS,c_1,k_1'}(t)|^2|h^n_{IS,c_2,k_2'}(t)|^2|h^{i_1,n}_{c_1,m_1,c_1,k_1}(t)|^2|h^{i_2,n}_{c_2,m_2,c_2,k_2}(t)|^2\Big)\Bigg],$$

$$=\frac{1}{M^2C^2}\mathbb{E}\Bigg[\sum_{\substack{c_1,c_2,m_1,\\ m_2,i_1,i_2}}\Delta\theta^{i_1,n}_{c_1,m_1}(t)\Delta\theta^{i_2,n}_{c_2,m_2}(t)\Big(1-\frac{\beta_{c_1,m_1,c_1}\beta_{IS,c_1}}{\bar{\beta}\bar{\beta}_{c_1}}-\frac{\beta_{c_2,m_2,c_2}\beta_{IS,c_2}}{\bar{\beta}\bar{\beta}_{c_2}}$$
$$+\frac{\beta_{c_1,m_1,c_1}\beta_{c_2,m_2,c_2}\beta_{IS,c_1}\beta_{IS,c_2}}{MCKK'I\bar{\beta}^2\bar{\beta}_{c_1}\bar{\beta}_{c_2}}\big(4+2(K'-1)+(M-1)(K-1)(I-1)(2+(K'-1)(C-1))\big)\Big)\Bigg],$$

$$=\mathbb{E}\Bigg[\frac{1}{M^2C^2}\sum_{c_1,c_2,m_1,m_2,i_1,i_2}\Delta\theta^{i_1,n}_{c_1,m_1}(t)\Delta\theta^{i_2,n}_{c_2,m_2}(t)A(m_1,m_2,c_1,c_2)\Bigg], \tag{B.4}$$

where $A(m_1,m_2,c_1,c_2)$ is given in Theorem 2. Combining for all symbols, we

have

$$\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,1}(t)-\Delta\theta^n_{PS}(t)\big)^2\Big]$$
$$=\frac{1}{M^2C^2}\sum_{n=1}^{2N}\sum_{c_1,c_2,m_1,m_2,i_1,i_2}A(m_1,m_2,c_1,c_2)\mathbb{E}\Big[\Delta\theta^{i_1,n}_{c_1,m_1}(t)\Delta\theta^{i_2,n}_{c_2,m_2}(t)\Big],$$
$$\overset{(a)}{\leq}\frac{\eta^2(t)G^2I^2\tau^2}{M^2C^2}\sum_{c_1=1}^{C}\sum_{c_2=1}^{C}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}A(m_1,m_2,c_1,c_2), \tag{B.5}$$

where (a) is obtained using Lemma 23. □

Lemma 26. $\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,2}(t)\big)^2\Big]\leq\frac{(K'+1)\eta^2(t)G^2I\tau^2}{KK'M^2C^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\frac{\beta^2_{IS,c}\beta_{c,m,c}\beta_{c,m',c}}{\bar{\beta}^2_c}$

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

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

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

$$=\frac{(K'+1)}{K^2K'M^2C^2\sigma_h^4\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{c,m,m'\neq m,i,k}\frac{\beta^2_{IS,c}}{\bar{\beta}^2_c}\Bigg(\mathrm{Re}\Big\{(h^{i,n}_{c,m,c,k}(t))^*h^{i,n}_{c,m',c,k}(t)\Delta\theta^{i,n,cx}_{c,m'}(t)\Big\}\Bigg)^2$$
$$+\,\mathrm{Re}\Big\{(h^{i,n}_{c,m,c,k}(t))^*h^{i,n}_{c,m',c,k}(t)\Delta\theta^{i,n,cx}_{c,m'}(t)\Big\}\,\mathrm{Re}\Big\{(h^{i,n}_{c,m',c,k}(t))^*h^{i,n}_{c,m,c,k}(t)\Delta\theta^{i,n,cx}_{c,m}(t)\Big\}\Bigg],$$

$$=\frac{(K'+1)}{2KK'M^2C^2\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{i=1}^{I}\frac{\beta^2_{IS,c}\beta_{c,m,c}\beta_{c,m',c}}{\bar{\beta}^2_c}$$
$$\times\left(\Big(\Delta\theta^{i,n}_{c,m'}(t)\Big)^2+\Big(\Delta\theta^{i,n+N}_{c,m'}(t)\Big)^2+\Delta\theta^{i,n}_{c,m}(t)\Delta\theta^{i,n}_{c,m'}(t)-\Delta\theta^{i,n+N}_{c,m}(t)\Delta\theta^{i,n+N}_{c,m'}(t)\right)\Bigg]. \tag{B.7}$$

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

$$= \frac{(K'+1)}{2KK'M^2C^2\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{i=1}^{I}\frac{\beta_{IS,c}^2\beta_{c,m,c}\beta_{c,m',c}}{\bar{\beta}_c^2}$$
$$\times\left(\left(\Delta\theta_{c,m'}^{i,n}(t)\right)^2+\left(\Delta\theta_{c,m'}^{i,n-N}(t)\right)^2+\Delta\theta_{c,m}^{i,n}(t)\Delta\theta_{c,m'}^{i,n}(t)-\Delta\theta_{c,m}^{i,n-N}(t)\Delta\theta_{c,m'}^{i,n-N}(t)\right)\Bigg]. \tag{B.8}$$

Combining the two cases, it becomes

$$\sum_{n=1}^{2N}\mathbb{E}\Big[\left(\Delta\hat{\theta}_{PS,2}^{n}(t)\right)^2\Big]=\frac{(K'+1)}{KK'M^2C^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{i=1}^{I}\frac{\beta_{IS,c}^2\beta_{c,m,c}\beta_{c,m',c}}{\bar{\beta}_c^2}\mathbb{E}\Big[\left\|\Delta\boldsymbol{\theta}_{c,m'}^{i}(t)\right\|_2^2\Big],$$
$$\overset{(a)}{\leq}\frac{(K'+1)\eta^2(t)G^2I\tau^2}{KK'M^2C^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\frac{\beta_{IS,c}^2\beta_{c,m,c}\beta_{c,m',c}}{\bar{\beta}_c^2}, \tag{B.9}$$

where (a) is obtained using Lemma 24. □

Lemma 27. $\sum_{n=1}^{2N}\mathbb{E}\Big[\left(\Delta\hat{\theta}_{PS,3}^{n}(t)\right)^2\Big]\leq\frac{(K'+1)\eta^2(t)G^2I\tau^2}{KK'M^2C^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{m=1}^{M}\sum_{m'=1}^{M}\frac{\beta_{IS,c}^2\beta_{c,m,c}\beta_{c,m',c'}}{\bar{\beta}_c^2}$

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

$$\mathbb{E}\Big[\left(\Delta\hat{\theta}_{PS,3}^{n}(t)\right)^2\Big]$$
$$= \frac{1}{K^2(K')^2M^2C^2\sigma_h^8\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{c_1,c_2}\sum_{c_1'\neq c_1,c_2'\neq c_2}\sum_{m_1,m_2}\sum_{m_1',m_2'}\sum_{i_1,i_2}\sum_{k_1,k_2,k_1',k_2'}$$
$$\times\frac{1}{\bar{\beta}_{c_1}\bar{\beta}_{c_2}}|h_{PS,c_1,k_1'}^{n}(t)|^2|h_{PS,c_2,k_2'}^{n}(t)|^2\,\mathrm{Re}\left\{(h_{c_1,m_1,c_1,k_1}^{i_1,n}(t))^*h_{c_1,m_1',c_1',k_1}^{i_1,n}(t)\Delta\theta_{c_1',m_1'}^{i_1,n,cx}(t)\right\}$$
$$\times\,\mathrm{Re}\left\{(h_{c_2,m_2,c_2,k_2}^{i_2,n}(t))^*h_{c_2,m_2',c_2',k_2}^{i_2,n}(t)\left(\Delta\theta_{c_2',m_2'}^{i_2,n,cx}(t)\right\}\right]. \tag{B.10}$$

In order 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

$$=\frac{(K'+1)}{K^2K'M^2C^2\sigma_h^4\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{\substack{c,c'\neq c,\\m,m',i,k}}\frac{\beta_{IS,c}^2}{\bar{\beta}_c^2}\left(\mathrm{Re}\left\{(h_{c,m,c,k}^{i,n}(t))^*h_{c,m',c',k}^{i,n}(t)\Delta\theta_{c',m'}^{i,n,cx}(t)\right\}\right)^2\Bigg],$$

$$= \frac{(K'+1)}{2KK'M^2C^2\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{\substack{c,c'\neq c,\\ m,m',i}} \frac{\beta_{IS,c}^2\beta_{c,m,c}\beta_{c,m',c'}}{\bar{\beta}_c^2}\Big(\big(\Delta\theta_{c',m'}^{i,n}(t)\big)^2+\big(\Delta\theta_{c',m'}^{i,n+N}(t)\big)^2\Big)\Bigg]. \tag{B.11}$$

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

$$\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}_{PS,3}^n(t)\big)^2\Big]$$
$$=\sum_{n=1}^{N}\frac{(K'+1)}{KK'M^2C^2\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{\substack{c,c'\neq c,\\ m,m',i}} \frac{\beta_{IS,c}^2\beta_{c,m,c}\beta_{c,m',c'}}{\bar{\beta}_c^2}\Big(\big(\Delta\theta_{c',m'}^{i,n}(t)\big)^2+\big(\Delta\theta_{c',m'}^{i,n+N}(t)\big)^2\Big)\Bigg],$$

$$\overset{(a)}{\leq} \frac{(K'+1)\eta^2(t)G^2I\tau^2}{KK'M^2C^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{\substack{c'=1\\ c'\neq c}}^{C}\sum_{m=1}^{M}\sum_{m'=1}^{M}\frac{\beta_{IS,c}^2\beta_{c,m,c}\beta_{c,m',c'}}{\bar{\beta}_c^2}, \tag{B.12}$$

where (a) is obtained using Lemma 24. □

Lemma 28. $\displaystyle\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}_{PS,4}^n(t)\big)^2\Big] = \frac{(K'+1)I\sigma_z^2N}{KK'M^2C^2P_t^2\sigma_h^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{\beta_{IS,c}^2\beta_{m,c,c}}{\bar{\beta}_c^2}$

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

$$\mathbb{E}\Big[\big(\Delta\hat{\theta}_{PS,4}^n(t)\big)^2\Big]=\frac{1}{P_{IS,t}^2K^2\big(K'\big)^2M^2C^2\sigma_h^8\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{c_1,c_2}\sum_{m_1,m_2,i_1,i_2}\sum_{k_1,k_2,k_1',k_2'}\frac{1}{\bar{\beta}_{c_1}\bar{\beta}_{c_2}}$$
$$\times |h_{PS,c_1,k_1'}^n(t)|^2|h_{PS,c_2,k_2'}^n(t)|^2$$
$$\times \mathrm{Re}\Big\{\big(h_{c_1,m_1,c_1,k_1}^{i_1,n}(t)\big)^* z_{c_1,k_1}^{i_1,n}(t)\Big\}\mathrm{Re}\Big\{\big(h_{c_2,m_2,c_2,k_2}^{i_2,n}(t)\big)^* z_{c_2,k_2}^{i_2,n}(t)\Big\}\Bigg]. \tag{B.13}$$

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

$$= \frac{(K'+1)}{P_{IS,t}^2K^2K'M^2C^2\sigma_h^4\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{c,m,i,k}\frac{\beta_{IS,c}^2}{\bar{\beta}_c^2}\Big(\mathrm{Re}\Big\{(h_{c,m,c,k}^{i,n}(t))^* z_{c,k}^{i,n}(t)\Big\}\Big)^2\Bigg],$$
$$= \frac{(K'+1)I\sigma_z^2}{2P_{IS,t}^2KK'M^2C^2\sigma_h^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{\beta_{IS,c}^2\beta_{c,m,c}}{\bar{\beta}_c^2}. \tag{B.14}$$

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

$$\sum_{n=1}^{2N} \mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,4}(t)\big)^2\Big] = \frac{(K'+1)I\sigma_z^2 N}{P^2_{IS,t}KK'M^2C^2\sigma_h^2\bar{\beta}^2} \sum_{c=1}^{C}\sum_{m=1}^{M} \frac{\beta^2_{IS,c}\beta_{c,m,c}}{\bar{\beta}_c^2}. \tag{B.15}$$

□

Lemma 29.
$$\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,5}(t)\big)^2\Big] \leq \frac{\big(2+(M-1)(C-2)(K-1)(I-1)\big)\eta^2(t)IG^2\tau^2}{K(K')M^3C^2(C-1)\bar{\beta}^2} \sum_{c=1}^{C}\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}\frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m_1,c'}\beta_{c',m_2,c'}}{\bar{\beta}^2_{c'}}$$

Proof. For $1 \leq n \leq N$, the equation becomes

$$\begin{aligned}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,5}(t)\big)^2\Big] =& \frac{1}{K^2(K')^2M^2C^2\sigma_h^8\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{c_1,c_2}\sum_{c_1'\neq c_1,c_2'\neq c_2}\sum_{m_1,m_2,i_1,i_2}\sum_{k_1,k_2,k_1',k_2'} \\ &\times \frac{1}{\bar{\beta}_{c_1'}\bar{\beta}_{c_2'}}|h^{i_1,n}_{c_1',m_1,c_1',k_1}(t)|^2|h^{i_2,n}_{c_2',m_2,c_2',k_2}(t)|^2\,\mathrm{Re}\Big\{(h^n_{IS,c_1,k_1'}(t))^* h^n_{IS,c_1',k_1'}(t)\Delta\theta^{i_1,n,cx}_{c_1',m_1}(t)\Big\} \\ &\times \mathrm{Re}\Big\{(h^n_{IS,c_2,k_2'}(t))^* h^n_{IS,c_2',k_2'}(t)\Delta\theta^{i_2,n,cx}_{c_2',m_2}(t)\Big\}\Bigg]. \end{aligned}\tag{B.16}$$

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

$$\begin{aligned}=& \frac{\big(2+(M-1)(C-2)(K-1)(I-1)\big)}{4(K')M^3C^2(C-1)KI\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{c,c'\neq c,m_1,m_2,i_1,i_2}\beta_{IS,c}\beta_{IS,c'} \\ &\times\Bigg(\frac{\beta_{c',m_1,c'}\beta_{c',m_2,c'}}{\bar{\beta}^2_{c'}}\Big(\Delta\theta^{i_1,n}_{c',m_1}(t)\Delta\theta^{i_2,n}_{c',m_2}(t)+\Delta\theta^{i_1,n+N}_{c',m_1}(t)\Delta\theta^{i_2,n+N}_{c',m_2}(t)\Big) \\ &+\frac{\beta_{c',m_1,c'}\beta_{c,m_2,c}}{\bar{\beta}_c\bar{\beta}_{c'}}\Big(\Delta\theta^{i_1,n}_{c',m_1}(t)\Delta\theta^{i_2,n}_{c,m_2}(t)-\Delta\theta^{i_1,n+N}_{c',m_1}(t)\Delta\theta^{i_2,n+N}_{c,m_2}(t)\Big)\Bigg)\Bigg]. \end{aligned}\tag{B.17}$$

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

$$\sum_{n=1}^{2N} \mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,5}(t)\big)^2\Big] = \frac{\big(2+(M-1)(C-2)(K-1)(I-1)\big)}{2K(K')M^3C^2(C-1)I\bar{\beta}^2} \sum_{n=1}^{N} \sum_{c,c'\neq c,m_1,m_2,i_1,i_2}$$

$$\times \frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m_1,c'}\beta_{c',m_2,c'}}{\bar{\beta}^2_{c'}} \mathbb{E}\bigg[\Big(\Delta\theta^{i_1,n}_{c',m_1}(t)\Delta\theta^{i_2,n}_{c',m_2}(t) + \Delta\theta^{i_1,n+N}_{c',m_1}(t)\Delta\theta^{i_2,n+N}_{c',m_2}(t)\Big)\bigg],$$

$$\overset{(a)}{\leq} \frac{\big(2+(M-1)(C-2)(K-1)(I-1)\big)\eta^2(t)IG^2\tau^2}{K(K')M^3C^2(C-1)\bar{\beta}^2} \sum_{c=1}^{C} \sum_{\substack{c'=1\\c'\neq c}}^{C} \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} \frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m_1,c'}\beta_{c',m_2,c'}}{\bar{\beta}^2_{c'}}, \tag{B.18}$$

where (a) is obtained using Lemma 23. □

Lemma 30. $\displaystyle\sum_{n=1}^{2N} \mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,6}(t)\big)^2\Big] \leq \frac{\eta^2(t)IG^2\tau^2}{KK'M^2C^2\bar{\beta}^2} \sum_{c=1}^{C} \sum_{\substack{c'=1\\c'\neq c}}^{C} \sum_{m=1}^{M} \sum_{\substack{m'=1\\m'\neq m}}^{M} \frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m',c'}}{\bar{\beta}^2_{c'}}$

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

$$\begin{aligned}
&\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,6}(t)\big)^2\Big] \\
&= \frac{1}{K^2(K')^2M^2C^2\sigma_h^8\bar{\beta}^2} \mathbb{E}\Bigg[\sum_{c_1,c_2} \sum_{c'_1\neq c_1,c'_2\neq c_2} \sum_{m_1,m_2} \sum_{m'_1\neq m_1,m'_2\neq m_2} \sum_{i_1,i_2} \sum_{k_1,k_2,k'_1,k'_2} \\
&\quad \times \frac{1}{\bar{\beta}_{c'_1}\bar{\beta}_{c'_2}} \operatorname{Re}\Big\{(h^n_{PS,c_1,k'_1}(t))^* h^n_{PS,c'_1,k'_1}(t)(h^{i_1,n}_{c'_1,m_1,c'_1,k_1}(t))^* h^{i_1,n}_{c'_1,m'_1,c'_1,k_1}(t)\Delta\theta^{i_1,n,cx}_{c'_1,m'_1}(t)\Big\} \\
&\quad \times \operatorname{Re}\Big\{(h^n_{PS,c_2,k'_2}(t))^* h^n_{PS,c'_2,k'_2}(t)(h^{i_2,n}_{c'_2,m_2,c'_2,k_2}(t))^* h^{i_2,n}_{c'_2,m'_2,c'_2,k_2}(t)\Delta\theta^{i_2,n,cx}_{c'_2,m'_2}(t)\Big\}\Bigg].
\end{aligned} \tag{B.19}$$

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'.

$$
\begin{aligned}
&= \frac{1}{K^2(K')^2M^2C^2\sigma_h^8\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{c,c'\neq c,m,m'\neq m,i,k,k'}\frac{1}{\bar{\beta}_{c'}^2} \\
&\quad\times\Bigg(\Big(\mathrm{Re}\Big\{(h^n_{PS,c,k'}(t))^* h^n_{PS,c',k'}(t)(h^{i,n}_{c',m,c',k}(t))^* h^{i,n}_{c',m',c',k}(t)\Delta\theta^{i,n,cx}_{c',m'}(t)\Big\}\Big)^2 \\
&\qquad+\mathrm{Re}\Big\{(h^n_{PS,c,k'}(t))^* h^n_{PS,c',k'}(t)(h^{i,n}_{c',m,c',k}(t))^* h^{i,n}_{c',m',c',k}(t)\Delta\theta^{i,n,cx}_{c',m'}(t)\Big\} \\
&\qquad\times\mathrm{Re}\Big\{(h^n_{PS,c,k'}(t))^* h^n_{PS,c',k'}(t)(h^{i,n}_{c',m',c',k}(t))^* h^{i,n}_{c',m,c',k}(t)\Delta\theta^{i,n,cx}_{c',m}(t)\Big\}\Bigg)\Bigg], \\
&= \mathbb{E}\Bigg[\sum_{c=1}^{C}\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{i=1}^{I}\frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m',c'}}{2K(K')M^2C^2\bar{\beta}^2\bar{\beta}_{c'}^2}\Big((\Delta\theta^{i,n}_{c',m'}(t))^2+(\Delta\theta^{i,n+N}_{c',m'}(t))^2\Big)\Bigg].
\end{aligned}
\tag{B.20}
$$

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

$$
\begin{aligned}
\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,6}(t)\big)^2\Big] &= \sum_{c=1}^{C}\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\sum_{i=1}^{I}\frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m',c'}}{K(K')M^2C^2\bar{\beta}^2\bar{\beta}_{c'}^2}\mathbb{E}\Big[\big\|\Delta\boldsymbol{\theta}^i_{c',m'}(t)\big\|_2^2\Big], \\
&\overset{(a)}{\leq}\frac{\eta^2(t)IG^2\tau^2}{KK'M^2C^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m',c'}}{\bar{\beta}_{c'}^2},
\end{aligned}
\tag{B.21}
$$

where (a) is obtained using Lemma 24. □

Lemma 31. $\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,7}(t)\big)^2\Big]$

$$
\leq\frac{\eta^2(t)IG^2\tau^2}{KK'M^2C^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{\substack{c''=1\\c''\neq c'}}^{C}\sum_{m=1}^{M}\frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m,c''}}{\bar{\beta}_{c'}^2}
$$

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

$$\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,7}(t)\big)^2\Big] = \frac{1}{K^2(K')^2M^2C^2\sigma_h^8\bar{\beta}^2}$$
$$\times \mathbb{E}\Bigg[\sum_{c_1,c_2}\sum_{c_1'\neq c_1,c_2'\neq c_2}\sum_{c_1''\neq c_1',c_2''\neq c_2'}\sum_{m_1,m_2}\sum_{m_1',m_2'}\sum_{i_1,i_2}\sum_{k_1,k_2,k_1',k_2'}\frac{1}{\bar{\beta}_{c_1'}\bar{\beta}_{c_2'}}$$
$$\times \mathrm{Re}\Big\{(h^n_{PS,c_1,k_1'}(t))^* h^n_{PS,c_1',k_1'}(t)(h^{i_1,n}_{c_1',m_1,c_1',k_1}(t))^* h^{i_1,n}_{c_1',m_1',c_1'',k_1}(t)\Delta\theta^{i_1,n,cx}_{c_1'',m_1'}(t)\Big\}$$
$$\times \mathrm{Re}\Big\{(h^n_{PS,c_2,k_2'}(t))^* h^n_{PS,c_2',k_2'}(t)(h^{i_2,n}_{c_2',m_2,c_2',k_2}(t))^* h^{i_2,n}_{c_2',m_2',c_2'',k_2}(t)\Delta\theta^{i_2,n,cx}_{c_2'',m_2'}(t)\Big\}\Bigg]. \tag{B.22}$$

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

$$= \frac{1}{K^2(K')^2M^2C^2\sigma_h^8\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{c,c'\neq c,c''\neq c'}\sum_{m,m',i,k,k'}\frac{1}{\bar{\beta}^2_{c'}}$$
$$\times\Big(\mathrm{Re}\Big\{(h^n_{PS,c,k'}(t))^* h^n_{PS,c',k'}(t)(h^{i,n}_{c',m,c',k}(t))^* h^{i,n}_{c',m',c'',k}(t)\Delta\theta^{i,n,cx}_{c'',m}(t)\Big\}\Big)^2\Bigg],$$
$$= \mathbb{E}\Bigg[\sum_{c,c'\neq c,c''\neq c'}\sum_{m,m',i}\frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m',c''}}{2KK'M^2C^2\bar{\beta}^2\bar{\beta}^2_{c'}}\Big((\Delta\theta^{i,n}_{c'',m'}(t))^2 + (\Delta\theta^{i,n+N}_{c'',m'}(t))^2\Big)\Bigg]. \tag{B.23}$$

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

$$\sum_{n=1}^{2N}\mathbb{E}\Big[\big(\Delta\hat{\theta}^n_{PS,7}(t)\big)^2\Big]$$
$$= \sum_{n=1}^{N}\sum_{c,c'\neq c,c''\neq c'}\sum_{m,m',i}\frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m',c''}}{K(K')M^2C^2\bar{\beta}^2\bar{\beta}^2_{c'}}\mathbb{E}\Big[(\Delta\theta^{i,n}_{c'',m'}(t))^2 + (\Delta\theta^{i,n+N}_{c'',m'}(t))^2\Big],$$
$$\overset{(a)}{\le} \frac{\eta^2(t)IG^2\tau^2}{KK'M^2C^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{\substack{c''=1\\c''\neq c'}}^{C}\sum_{m=1}^{M}\frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\beta_{c',m,c''}}{\bar{\beta}^2_{c'}}. \tag{B.24}$$

where (a) is due to Lemma 24. □

Lemma 32. $\sum_{n=1}^{2N}\mathbb{E}\Big[\Big(\Delta\hat{\theta}^n_{PS,6}(t)\Big)^2\Big] = \frac{\sigma_z^2IN}{P^2_{IS,t}K(K')M^2C^2\sigma_h^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{m=1}^{M}\frac{\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}}{\bar{\beta}^2_{c'}}$

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

$$\mathbb{E}\Big[\Big(\Delta\hat{\theta}^n_{PS,6}(t)\Big)^2\Big] = \frac{1}{P^2_{IS,t}K^2(K')^2M^2C^2\sigma_h^8\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{c_1,c_2}\sum_{c_1'\neq c_1,c_2'\neq c_2}\sum_{m_1,m_2,i_1,i_2}\sum_{k_1,k_2,k_1',k_2'} \times \frac{1}{\bar{\beta}_{c_1'}\bar{\beta}_{c_2'}}\,\mathrm{Re}\Big\{\big(h^n_{PS,c_1,k_1'}(t)\big)^* h^n_{PS,c_1',k_1'}(t)\big(h^{i_1,n}_{c_1',m_1,c_1',k_1}(t)\big)^* z^{i_1,n}_{IS,c_1',k_1}(t)\Big\} \times \mathrm{Re}\Big\{\big(h^n_{PS,c_2,k_2'}(t)\big)^* h^n_{PS,c_2',k_2'}(t)\big(h^{i_2,n}_{c_2',m_2,c_2',k_2}(t)\big)^* z^{i_2,n}_{IS,c_2',k_2}(t)\Big\}\Bigg]. \quad \text{(B.25)}$$

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

$$\mathbb{E}\Big[\Big(\Delta\hat{\theta}^n_{PS,6}(t)\Big)^2\Big] = \frac{\sigma_z^2 I}{2P^2_{IS,t}K(K')M^2C^2\sigma_h^2\bar{\beta}^2}\sum_{c=1}^{C}\sum_{\substack{c'=1\\c'\neq c}}^{C}\sum_{m=1}^{M}\frac{\big(\beta_{IS,c}\beta_{IS,c'}\beta_{c',m,c'}\big)}{\bar{\beta}^2_{c'}}. \quad \text{(B.26)}$$

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

Lemma 33. $\displaystyle\sum_{n=1}^{2N}\mathbb{E}\Big[\Big(\Delta\hat{\theta}^n_{PS,7}(t)\Big)^2\Big] = \frac{\sigma_z^2 N}{P^2_{IS,t}(K')C^2\sigma_h^2\bar{\beta}^2}\sum_{c=1}^{C}\beta_{IS,c}$

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

$$\mathbb{E}\Big[\Big(\Delta\hat{\theta}^n_{PS,7}(t)\Big)^2\Big] = \frac{1}{P^2_{IS,t}(K')^2C^2\sigma_h^4\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{c_1=1}^{C}\sum_{c_2=1}^{C}\sum_{k_1'=1}^{K'}\sum_{k_2'=1}^{K'}\mathrm{Re}\Big\{\big(h^n_{PS,c_1,k_1'}(t)\big)^* z^n_{PS,k_1'}(t)\Big\} \times \mathrm{Re}\Big\{\big(h^n_{PS,c_2,k_2'}(t)\big)^* z^n_{PS,k_2'}(t)\Big\}\Bigg]. \quad \text{(B.27)}$$

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

$$\mathbb{E}\Big[\Big(\Delta\hat{\theta}^n_{PS,7}(t)\Big)^2\Big] = \frac{1}{P^2_{IS,t}(K')^2C^2\sigma_h^4\bar{\beta}^2}\mathbb{E}\Bigg[\sum_{c=1}^{C}\sum_{k'=1}^{K'}\Big(\mathrm{Re}\Big\{(h^n_{PS,c,k'}(t))^* z^n_{PS,k'}(t)\Big\}\Big)^2\Bigg], = \frac{\sigma_z^2}{2P^2_{IS,t}(K')C^2\sigma_h^2\bar{\beta}^2}\sum_{c=1}^{C}\beta_{IS,c}. \quad \text{(B.28)}$$

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

Combining Lemmas 25-33 completes the proof of Lemma 7.