Toward Efficient and Reliable Wireless Federated Learning: System Design and Theoretical Analysis

A

Dissertation

Presented to
the faculty of the School of Engineering and Applied Science
University of Virginia

in partial fulfillment
of the requirements for the degree

Doctor of Philosophy

by

Xizixiang Wei

May 2024

APPROVAL SHEET

This

Dissertation

is submitted in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Author: Xizixiang Wei

This Dissertation has been read and approved by the examing committee:

Advisor: Cong Shen

Advisor:

Committee Member: Nikolaos Sidiropoulos

Committee Member: Stephen Wilson

Committee Member: Brad Campbell

Committee Member: Tianhao Wang

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

May 2024

Abstract

In recent years, the surge in data generation by wireless edge devices has propelled the integration of artificial intelligence (AI) across various domains, including computer vision and natural language processing. Traditionally, ML models are trained in a centralized fashion, where edge devices transmit their data to a central server. However, this centralized approach incurs substantial communication resource costs and poses significant privacy risks. Federated Learning (FL) emerges as an innovative solution, addressing these concerns by decentralizing the learning process. The quest for communication efficiency is central to the optimization of FL, as it represents a critical bottleneck that impacts its scalability and effectiveness. This dissertation concentrates on enhancing the reliability and efficiency of both uplink and downlink wireless transmissions within the FL framework.

This dissertation introduces a pioneering convergence analysis for FL under the condition of simultaneous noisy communications in both directions, establishing the criteria essential for ensuring FL's convergence over noisy channels. Followed by the theoretical results, two novel designs are proposed for both uplink and downlink transmissions in multiple-input multiple-output (MIMO) and single-input single-output (SISO) systems. For MIMO systems, the proposed "random orthogonalization" method, leveraging the massive MIMO characteristics of channel hardening and favorable propagation, facilitates natural over-the-air model aggregation without the need for channel state information at the transmitters (CSIT) in the uplink stage, and offers an efficient model broadcasting technique during the downlink stage. In the realm of SISO systems, our strategy also eliminates the necessity for CSIT through the use of orthogonal sequences, providing robust and flexible differential privacy (DP) guarantee at both the item and client levels.

Extensive numerical experiments conducted with real-world datasets affirm the effectiveness and efficiency of the proposed methods, highlighting their potential to significantly enhance the FL process by addressing its communication challenges.

To my parents, Yuqing Ruan and Yu Wei.

Acknowledgement

I wish to express my profound gratitude to Professor Cong Shen, my Ph.D. advisor, for his outstanding guidance, steadfast support, and invaluable mentorship. His deep knowledge, patience, and dedication have been crucial in shaping my academic and research journey. I am deeply appreciative of his role in my achievements.

My sincere thanks also go to my distinguished committee members, Professor Nikolaos Sidiropoulos, Professor Stephen Wilson, Professor Brad Campbell, and Professor Tianhao Wang. Their insightful feedback, constructive criticism, and contributions have played a significant role in refining my dissertation. Their dedication to academic rigor has greatly contributed to the depth and quality of my work.

I am deeply thankful to my collaborators, including Professor Vincent Poor, Professor Jing Yang, and Ruiquan Huang, for their collaborative spirit, intellectual contributions, and expertise. I also extend my heartfelt gratitude to my esteemed mentors from my master's program, Professor Xin Wang and Professor Yi Jiang, whose guidance and support have been instrumental in propelling me through this incredible academic odyssey. Working alongside such brilliant minds has greatly enriched my research, broadened its horizons, and led to notable progress. It has been a privilege to collaborate with them.

I also wish to thank my peers and friends at UVA, namely Chengshuai Shi, Kun Yang, Li Fan, Yujia Mu, Zihan Chen, and Wei Shen, for their stimulating discussions, support, and camaraderie. Their presence has greatly enhanced my academic journey and added significant value to my experience.

Lastly, but most importantly, I owe my deepest gratitude to my parents for their unwavering support and love. Their faith in me and their sacrifices have laid the groundwork for my path. Their inspiration and encouragement have been indispensable in my life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Overview

Federated learning (FL) [1, 2] is an emerging distributed machine learning paradigm that has many attractive properties which can address new challenges in machine learning (ML). In particular, FL is motivated by the growing trend that massive amount of the real-world data are *exogenously* generated at the edge devices. For better privacy protection, it is desirable to keep sensitive user data locally. FL is able to train a global ML model across all local datasets without the server having direct access to client data. The power of FL has been realized in commercial devices (e.g., Google Pixel 2 uses FL to train ML models to personalize user experience) and ML tasks (e.g., Gboard uses FL for keyboard prediction) [3]. Therefore, FL is considered as one of the potential key applications in 6th generation (6G) cellular communication systems [4].

Communication efficiency has been at the front and center of FL ever since its inception [2, 1], and it is widely regarded as one of its primary bottlenecks The local training nature of FL leads to massive communication costs, as an FL task consists of multiple learning rounds, each of which requires uplink and downlink model exchange between clients and the server. The limited communication resources in both uplink and downlink, combined with the dynamics in the communication links, severely impact the *efficiency* and *reliability* of FL tasks in a wireless communication system. Communication schemes for FL can be divided into two categories: digital communication and analog communication. Digital communication for FL is usually considered to incur a heavy burden for wireless networks, as it allocates different communication resources to the ML model parameters of each client. Analog communication reduces the communication overhead by allowing different clients to transmit FL models using shared resources.

To improve the communication efficiency of FL, early research has largely focused on either reducing the number of communication rounds [1, 5], or decreasing the size of the payload for transmission [6, 7, 8]. However, in most FL literature that deals with communication efficiency, it is often assumed that a perfect communication "tunnel" has been established, and the task of improving communication efficiency largely resides on the ML design that trades off computation and communication. More recent research starts to close this gap by focusing on the system design, particularly for wireless FL.

The primary difficulty in crafting and scrutinizing wireless FL systems is comprehending how the characteristics of wireless systems affect learning outcomes. Addressing diverse learning needs—ranging from accuracy and privacy assurances to dealing with the volatile nature of wireless connections, including noise, interference, and channel fading, alongside different system setups like single-input single-output (SISO) and multiple-input multiple-output (MIMO)—demands a thorough grasp of both learning and communication dynamics. The goal of this dissertation is to merge the design of learning and communication, thereby establishing an efficient and dependable FL framework. In particular, **this dissertation enhances the efficiency and reliability of wireless federated learning by delving into the optimal resource allocation and aggregation rules in noisy communication links. Leveraging the orthogonality provided by high-dimensional signals, it proposes two designs for MIMO and SISO systems, respectively, that improve upon existing over-the-air computation (AirComp) methods. The proposed designs do not require channel state information at transmitters (CSIT), simplifying system complexity while ensuring differential privacy guarantees.**

1.1 Challenges

Scalability. Compared with downlink broadcasting, uplink communication is more challenging in FL when communication is over the wireless medium [9, 6, 10, 11]. Due to the stringent power constraints at mobile devices, channel noise and fading have a more conspicuous impact on uplink communications. More importantly, significant scalability challenges arise from the large number of clients in FL versus limited uplink communication resources. Uplink communication is known to be one of the key bottlenecks of wireless federated learning [1, 12, 13, 14, 15].

A notable approach to addressing the scalability challenge of Federated Learning (FL) over wireless networks is through AirComp, as highlighted in several studies including [16, 17, 18, 19, 20]. Unlike the traditional method where individual local models from each client are decoded and then aggregated, AirComp enables the simultaneous transmission of uplink signals from multiple clients in a superposed manner, allowing the FL server to directly decode the aggregated global model. Nevertheless, managing power control and interference in AirComp presents significant challenges that are yet to be fully resolved.

CSIT requirements. In order to achieve AirComp, a common approach is to "invert" the fading channel at each transmitter [17, 20], so that the sum model can be obtained at the server. Variants and enhancements of AirComp have been studied, yet, a fundamental limitation of the existing methods is that they mostly require CSIT. Enabling CSIT in wireless communication systems is complicated and is substantially harder than obtaining the channel state information at the receiver (CSIR). Moreover, channel inversion based on CSIT is well known to “blow up” when one of the users’ channels is in deep fade [21]. Hence, exploring CSIT-free AirComp methods becomes attractive[22].

Privacy. Amidst a growing focus on data security, the importance of safeguarding individuals’ personal information has become increasingly emphasized. Although FL intuitively helps protect client privacy by keeping training data locally and never sharing it with the server, private information can still be leaked to some extent by analyzing the ML model parameters trained and uploaded by the clients [23, 24, 25]. To address the privacy concern, a natural way is to add (artificial) noise to ML model parameters in the upload phase of FL, whose privacy properties can be mathematically characterized using differential privacy (DP) [26]. Traditionally, adding artificial noise to achieve certain differential privacy during communications in FL will trade off with its convergence [27, 28]. In the context of FL within wireless systems, the strategy of adjusting system parameters to effectively harness the inherent randomness of wireless channels for DP becomes appealing. This approach is particularly attractive for its potential to achieve DP at no additional cost, essentially offering ”free” DP.

1.2 Dissertation Overview

The technical section of this dissertation is divided into three main parts. To address the identified challenges in FL, the study initially focuses on understanding how communication-induced noise during both the upload (uplink) and download (downlink) phases affects the convergence and accuracy of ML models. The insights gained from theoretical analyses inform the development of a signal-to-noise ratio (SNR) scaling strategy aimed at enhancing FL performance within a constrained total resource budget. Specifically, the proposed framework is designed to accommodate various scenarios, including both full and partial client participation, direct model updates and model differentials, and non-independent and identically distributed (non-IID) local datasets. This framework improves upon the commonly utilized algorithms for transmit power control and receive diversity combining, tailoring them to more effectively support FL in the presence of noisy communication channels.

The SNR scaling rules established in this dissertation act as a fundamental design guideline for wireless federated learning systems. Moving forward, the dissertation delves into the intricate physical layer design for various systems to adeptly tackle the scalability issue and the challenge posed by the requirement for CSIT in conventional AirComp setups. The suggested approach closely marries MIMO technology with FL, making deliberate use of the synergies between the two. By capitalizing on the distinct characteristics of channel hardening and favorable propagation characteristic

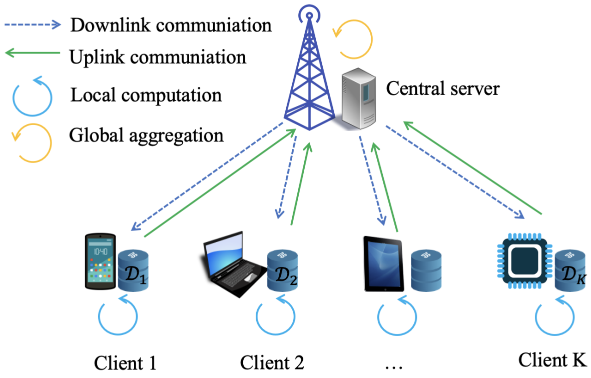


Figure 1.1: Federated learning pipeline.

of massive MIMO, the introduced concept of "random orthogonalization" enables the base station (BS) to compute the global model directly through a straightforward linear projection operation. This method significantly reduces the complexity and latency of uplink communication, presenting an effective solution to previously insurmountable challenges.

Inspired my random orthogonalization, we further propose FLORAS – Federated Learning using ORthogonAl Sequences, a novel uplink wireless physical layer design for FL by leveraging the properties of orthogonal sequences. FLORAS is an uplink communication scheme for SISO wireless FL systems. It enjoys all advantages of AirComp, yet without the CSIT requirement. Moreover, by the adjustment on the number of used orthogonal sequences in the system configuration, the novel signal processing techniques in FLORAS empowers flexible item-level and client-level DP guarantee. An interesting convergence characterizes the trade-off between the model convergence rate and the achievable DP levels is derived.

Extensive experiments were carried out to rigorously assess the effectiveness of the proposed solutions across various system configurations. This comprehensive evaluation aimed to showcase the versatility and robustness of the methodologies under consideration, focusing on different aspects such as system scalability, communication efficiency, and learning performance.

1.3 System Model and Assumptions

The FL problem setting studied throughout the this dissertation mostly follows the standard model, termed FEDAVG, in the original paper [1]. This section presents the FEDAVG framework and describe the main assumptions adopted in this dissertation. Consider an FL task with a central server and M total clients. Each client $k \in [M]$ stores a (disjoint) local dataset $\mathcal{D}_k$, with its size denoted by D_k. The size of the total data is $D \triangleq \sum_{k\in[M]} D_k$. We use $f_k(\mathbf{w})$ to denote the local loss function at client k, which measures how well a machine learning (ML) model with parameter $\mathbf{w} \in \mathbb{R}^d$ fits its local dataset. Therefore, the global objective function over all M clients can be denoted as

$$f(\mathbf{w}) = \sum_{k \in [M]} p_k f_k(\mathbf{w}),$$

where $p_k = \frac{D_k}{D}$ is the weight of each local loss function, and the purpose of FL is to distributively find the optimal model parameter $\mathbf{w}^*$ that minimizes the global loss function:

$$\mathbf{w}^* \triangleq \arg\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}).$$

Term $\Gamma \triangleq f^* - \sum_{k \in [M]} p_k f_k^*$ captures the non-i.i.d. degree of local datasets, where f^* and f_k^* are the minimums of global and local loss functions, respectively.

The FEDAVG [29] framework keeps clients' data locally, and the global model is converged at the central server by the composition of multiple learning rounds. One of the key characteristics of FL is *partial participation*, i.e., only a portion of clients are selected in a single learning round for model upload. Here, K of total M clients are uniformly randomly selected during each learning round for the FL task. To simplify the notation, this dissertation uses the subscript $k = 1, \cdots, K$ to indicate the K clients during a certain learning round, acknowledging that they could correspond to different clients in different rounds. A typical FL pipeline is illustrated in Fig. 1.1. Specifically, this pipeline iteratively executes the following steps for T learning rounds until the global model converges.

1. **Downlink communication.** The BS broadcasts the current global model $\mathbf{w}_{t-1}$ to all K selected devices over the downlink wireless channel.

2. **Local computation.** Each client uses its local data to train a local ML model improved upon the received global ML model. In this proposal, we assume that mini-batch SGD is used in the model training. Note that this is the most commonly adopted training method in modern ML tasks, e.g., deep neural networks, but its analysis is more complicated than gradient descent (GD) when communication noise is present. Specifically, mini-batch SGD operates by updating the weight iteratively (for E steps in each learning round) at client k as follows:

$$\begin{aligned}
\text{Initialization:} \quad & \mathbf{w}_{t,0}^k = \mathbf{w}_{t-1}^k, \\
\text{Iteration:} \quad & \mathbf{w}_{t,\tau}^k = \mathbf{w}_{t,\tau-1}^k - \eta_t \nabla F_k(\mathbf{w}_{t,\tau-1}^k, \xi_\tau^k), \forall \tau = 1, \cdots, E, \\
\text{Output:} \quad & \mathbf{w}_t^k = \mathbf{w}_{t,E}^k,
\end{aligned}$$

where ξ_τ^k is a batch of data points that are sampled independently and uniformly at random from the local dataset of client k in the τ-th iteration of mini-batch SGD.

3. **Uplink communication.** Each involved client uploads its latest local model to the server synchronously over the uplink wireless channel.

4. **Server Aggregation.** The BS aggregates the received local models $\mathbf{w}_t^k$ to generate a new global model. For simplicity, each local dataset is assumed to have a equal size. Therefore, the global model can be averaged by

$$\mathbf{w}_t = \Sigma_{k=1}^{K} \frac{1}{K} \mathbf{w}_t^k.$$

Throughout this dissertation, the following four assumptions are made on local loss functions $f_k \quad \forall k \in [M]$ to analyze the convergence performance theoretically of different methods. Note that they are all standard assumptions that are commonly adopted in the convergence analysis of FEDAVG and its variants; see [30, 31, 32, 6].

Assumption 1. *L-**smooth:** $\forall$ $\mathbf{v}$ and $\mathbf{w}$, $\|f_k(\mathbf{v}) - f_k(\mathbf{w})\| \leq L \|\mathbf{v} - \mathbf{w}\|$;*

Assumption 2. *μ-**strongly convex:** $\forall$ $\mathbf{v}$ and $\mathbf{w}$, $\langle f_k(\mathbf{v}) - f_k(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle \geq \mu \|\mathbf{v} - \mathbf{w}\|^2$;*

Assumption 3. ***Unbiased SGD:** $\forall k \in [M]$, $\mathbb{E}[\nabla \tilde{f}_k(\mathbf{w})] = \nabla f_k(\mathbf{w})$;*

Assumption 4. ***Uniformly bounded gradient:** $\forall k \in [M]$, $\mathbb{E} \left\| \nabla \tilde{f}_k(\mathbf{w}) \right\|^2 \leq H^2$ for all mini-batch data.*

In particular, Assumption 1 indicates that the gradient of f_k is Lipschitz continuous. The strongly convex loss function in Assumption 2 is a category of loss functions that are widely studied in the literature (see [30] and its follow-up works). Assumptions 3 and 4 imply that the mini-batch stochastic gradient and its variance are bounded [32].

Chapter 2

Federated Learning over Noisy Channels

While initial research has hinted at the benefits of refining communication protocols for FL, the crucial and more practical challenge posed by noisy communications—both in the uplink (where clients transmit local models to the parameter server) and the downlink (where the server dispatches the global model to clients)—remains underexplored. From an analytical standpoint, considering the impact of noise in both uplink and downlink simultaneously adds complexity to the analysis of FL convergence due to the propagation of noise through every round of communication in both directions. Moreover, the cumulative effect of these noisy communications in both the uplink and downlink phases plays a decisive role in determining the overall learning outcomes, necessitating a comprehensive approach to both the design and analysis of the system.

The goal in the research of this chapter is two-fold: To understand the impact of *communication-induced noise*, in both upload (uplink) and download (downlink) phases of FL, on the ML model convergence and accuracy performance, and to design communication algorithms to control the signal-to-noise ratio (SNR) to improve FL performance under a total resource budget.

The proposed solution is mainly focused on *analog* communications for model updates [19, 17, 18] and investigates SNR control in both uplink and downlink, which is especially crucial when the underlying ML method is stochastic gradient descent (SGD), because SGD is much more sensitive to noise than the (full) gradient descent [33, 32]. The treatment is novel because all prior works either study uplink-only [19, 17, 34, 35, 36] or downlink-only [37] noisy communications, but not both. Theoretically, the convergence analyses of the standard Federated Averaging (FEDAVG) scheme is presented, under non-IID datasets, full or partial clients participation, direct model or model differential transmissions, and simultaneous noisy downlink and uplink analog communications. These analyses are based on very general receive noise assumptions, and hence are broadly applicable to a variety of communication systems. The key insight of these theoretical results is a "flying under the radar" principle: SGD is inherently a noisy process, and as long as uplink/downlink channel noises do not dominate the SGD noise during model training (which is controlled by the time-varying learning rate), the scaling of convergence is not affected. This general principle is exemplified with two widely adopted communication techniques – *transmit power control* and *receive diversity combining* – by controlling the resulting post-processing SNR to satisfy the theoretical analyses under a fixed total budget constraint. Comprehensive numerical evaluations on three widely adopted ML tasks with increasing difficulties (MNIST, CIFAR-10 and Shakespeare) are carried out using these techniques. A sequence of experiments shows that adjusting transmit power and combining receive diversity, guided by theoretical analysis, can greatly surpass the baseline of maintaining equal SNR over time. In fact, in many experimental setups, this approach can come close to achieving the performance of ideal, noise-free communication. These results not only corroborate the theoretical conclusions, but also highlight the benefit of designing communication algorithms that are specifically tailored to the characteristics of FL.

To summarize, the main contributions of this chapter include the following.

- A novel convergence analyses for FL are presented with simultaneous uplink and downlink noisy analog communications, with full vs. partial clients participation, direct model vs. model differential, and non-IID local datasets. To the best of the authors' knowledge, this is the first time FL convergence analysis is carried out when both upload and download phases are over noisy communication channels, which introduce significant challenges because of the noise propagation in both directions.
- A *SNR scaling laws* is estabulished. In particular, the theoretical results prove that, in order to maintain the well-known $\mathcal{O}(1/T)$[1] convergence rate of FedAvg with noise-free communications, $\mathcal{O}(t^2)$ SNR scaling is needed for direct model and $\mathcal{O}(1)$ (i.e., constant) for model differential. This t^2-*vs*-1 *scaling law comparison* under the same communication environment is novel.
- The presented findings improve upon the commonly used algorithms for transmit power control and receive diversity combining, optimizing their use for FL over noisy channels. Through comprehensive numerical experiments, these enhancements have been shown to offer performance benefits over current leading methods, all while operating within the same total resource allocation.

The remainder of this chapter is organized as follows. Related works are surveyed in Section 2.1. The system model that captures the noisy channels in both uplink and downlink of FL is described in Section 2.2. Theoretical analyses are presented in Section 2.3 for three different FL configurations. These results inspire novel communication designs of transmit power control and receive diversity combining that are presented in Section 2.4. Experimental results are given in Section 2.5, followed by a summary in Section 2.6. All technical proofs are given in the Appendices.

[1] Notation $f = \mathcal{O}(g)$ denotes f is of order at most of g.

2.1 Related Works

Improve FL communication efficiency. The original FedAvg reduces the communication overhead by only periodically averaging the local models. Theoretical understanding of the communication-computation tradeoff has been actively pursued and, depending on the underlying assumptions (e.g., IID or non-IID local datasets, convex or non-convex loss functions, GD or SGD), rigorous analyses of the convergence behavior have been carried out [32, 38, 30]. For the approach of reducing the size of messages, general discussions on sparsification, subsampling, and quantization are given in [2]. There are also recent efforts in developing quantization and source coding to reduce the communication cost [39, 7, 19, 40, 6, 41, 8]. Nevertheless, they mostly do not consider the communication channel noise.

Communication design for FL. Recent years have also seen increased effort in the communication algorithm and system design for FL. Trade-off between local model update and global model aggregation is studied in [42] to optimize the transmission power/rate and training time. Various radio resource allocation and client selection policies [43, 44, 45, 46, 47, 48] have been proposed to minimize the learning loss or the training time. Joint communication and computation is investigated [18, 17, 39, 49]. In particular, the analog aggregation design [17, 50, 39] serves as one of our design examples in Section 2.4.

FL with imperfect/noisy communications. Existing literature is dominated by uplink-only noisy communications [51, 52, 18, 19, 17, 34, 35, 36]. There is very limited study on downlink-only noisy communications for FL; [37] proposes and analyzes downlink digital and analog transmissions while assuming an error-free uplink. On the other hand, existing literature that consider both upload and download imperfect communications focus only on how to modify the ML model training method. In particular, [53] changes the loss function of FL to accommodate the communication error. [54, 55, 56] propose to compress the gradients in order to tolerate both uplink and downlink bandwidth bottlenecks. Their methods are either error compensation, quantization or leveraging sparsity. None of these considers improving the communication design.

2.2 System Model for Learning over Noisy Communication

This chapter focus on a generic FL framework where *partial* client participation and *non-IID* local datasets, two critical features that separate FL from conventional distributed ML, are explicitly captured. Unlike the existing literature, the imperfect communications is captured and *both* the upload and download transmissions take place over noisy communication channels are fully considered. The overall system diagram is depicted in Fig. 2.1. In particular, the FL-over-noisy-channel pipeline works by iteratively executing the following steps at the t-th learning round, $\forall t \in [T]$.

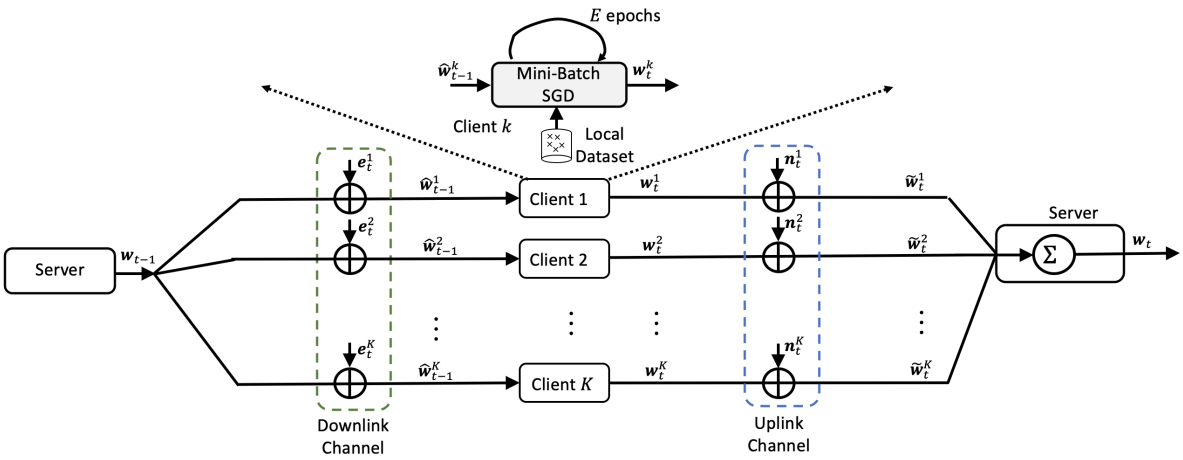


Figure 2.1: End-to-end FL system diagram in the t-th communication round. The impact of noisy channels in both uplink and downlink is captured.

Without loss of generality, the model transmitted at the transmitters $\mathbf{w}$ is assumed to be zero-mean and unit-variance elements[2], i.e., $\mathbb{E}||w_i||^2 = 1, \forall i \in [d]$.

(1) Downlink communication for global model download. The centralized server broadcasts the current global ML model, which is described by the latest weight vector $\mathbf{w}_{t-1}$ from the previous round, to a set of uniformly randomly selected clients[3] denoted as $\mathcal{S}_t$ with $|\mathcal{S}_t| = K$. Because of the imperfection introduced in communications, e.g., channel noise, imperfect channel estimation, and detection or estimation error, client k receives a noisy version of $\mathbf{w}_{t-1}$, which is written as

$$\hat{\mathbf{w}}_{t-1}^k = \mathbf{w}_{t-1} + \mathbf{e}_t^k, \tag{2.1}$$

where $\mathbf{e}_t^k = [e_{t,1}^k, \cdots, e_{t,d}^k]^T \in \mathbb{R}^d$ is the d-dimensional downlink *effective noise* vector at client k and time t. We assume that $\mathbf{e}_t^k$ is a zero-mean random vector consisting of IID elements with variance:

$$\mathbb{E}||e_{t,i}^k||^2 = \zeta_{t,k}^2 \quad \text{and} \quad \mathbb{E}||\mathbf{e}_t^k||^2 = d\zeta_{t,k}^2,\ \forall t \in [T], k \in \mathcal{S}_t, i \in [d]. \tag{2.2}$$

Effective noise and definition of SNR. In order to keep the problem general, a particular communication system for the actual downlink data transmission is not spcified here, and (2.1) only use the effective noise model. The same approach applies to the uplink. This is a conscious choice to keep the problem general, and to focus on analyzing the impact of communication-induced noise and further controlling the resulting SNR to improve FL performance. In this way, $\mathbf{e}_t^k$ shall be interpreted as the effective noise that captures all the processing components in a downlink

[2] The parameter normalization and de-normalization procedure in wireless FL can be found in the Appendix in [17]. Note that *weight normalization* is widely adopted in training deep neural networks [57].

[3] We note that for partial clients participation, we have $K < N$; in the case of full clients participation we have $K = N$.

communication phase in addition to the natural channel noise[4]. Because the variance of each scalar model parameter has been normalized , the (post-processing) receive SNR for the k-th client at the t-th communication round can be written as

$$\mathsf{SNR}^{\mathsf{L}}_{t,k} = \frac{\mathbb{E}||\mathbf{w}_{t-1}||^2}{\mathbb{E}||\mathbf{e}_t^k||^2} = \frac{1}{\zeta^2_{t,k}}. \tag{2.3}$$

Lastly, note that the noise assumption is very mild, because (2.2) only requires a *bounded* variance of the random noise, but does not limit to any particular distribution. In addition, the downlink communication model is very general in the sense that the effective noise variances, $\{\zeta^2_{t,k}\}$, can be different for different clients and at different rounds.

(2) Local computation. Each client uses its local data to train a local ML model improved upon the received global ML model. In this work, mini-batch SGD is used in the model training. Note that this is the most commonly adopted training method in modern ML tasks, e.g., deep neural networks, but its analysis is more complicated than gradient descent (GD) when communication noise is present.

Specifically, mini-batch SGD operates by updating the weight iteratively (for E steps in each learning round) at client k as follows:

$$\begin{aligned} \text{Initialization:} \quad & \mathbf{w}^k_{t,0} = \hat{\mathbf{w}}^k_{t-1}, \\ \text{Iteration:} \quad & \mathbf{w}^k_{t,\tau} = \mathbf{w}^k_{t,\tau-1} - \eta_t \nabla F_k(\mathbf{w}^k_{t,\tau-1}, \xi^k_\tau), \forall \tau = 1, \cdots, E, \\ \text{Output:} \quad & \mathbf{w}^k_t = \mathbf{w}^k_{t,E}, \end{aligned}$$

where ξ^k_τ is a batch of data points that are sampled independently and uniformly at random from the local dataset of client k in the τ-th iteration of mini-batch SGD.

(3) Uplink communication for local model upload. The K participating clients upload their latest local models to the server. More specifically, client k transmits a vector $\mathbf{x}^k_t$ to the server at the t-th round. Here, the practical case is again considered, where the server receives a noisy version of the individual weight vectors from each client in the uplink communications (e.g., channel noise, fading, transmitter and receiver distortion). The received vector for client k can be written as

$$\hat{\mathbf{x}}^k_t = \mathbf{x}^k_t + \mathbf{n}^k_t, \tag{2.4}$$

[4] As a simple example, if the downlink communication is over a standard Additive White Gaussian Noise (AWGN) channel, then the actual received signal at client k is $\mathbf{y}^k_{t-1} = \sqrt{P_{t-1}}\mathbf{w}_{t-1} + \mathbf{z}^k_t$ where $\mathbf{z}^k_t$ represents the AWGN and P_{t-1} is the downlink broadcast transmit power. The effective channel noise becomes $\mathbf{e}^k_t = \frac{1}{\sqrt{P_{t-1}}}\mathbf{z}^k_t$.

where $\mathbf{n}_t^k \in \mathbb{R}^d$ is the d-dimensional uplink *effective noise* vector for decoding client k's model at time t. We assume that $\mathbf{n}_t^k$ is a zero-mean random vector consisting of IID elements with bounded variance:

$$\mathbb{E}||n_{t,i}^k||^2 = \sigma_{t,k}^2 \quad \text{and} \quad \mathbb{E}||\mathbf{n}_t^k||^2 = d\sigma_{t,k}^2, \ \forall t \in [T], k \in \mathcal{S}_t, i \in [d]. \tag{2.5}$$

Once again, note that the uplink communication model in (2.5) is very general in the sense that (1) only bounded variance is assumed as opposed to the specific noise distribution; and (2) the effective noise variances, $\{\sigma_{t,k}^2\}$, can be different for different clients and at different rounds.

Unlike in the download phase where the model itself is transmitted to clients, two different choices of the vector $\mathbf{x}_t^k$ for model upload are considered in this paper.

1. **Model Transmission (MT).** The K participating clients upload the latest local models: $\mathbf{x}_t^k = \mathbf{w}_t^k$. Following (2.4), the server receives the updated local model of client k as

$$\tilde{\mathbf{w}}_t^k = \hat{\mathbf{x}}_t^k = \mathbf{w}_t^k + \mathbf{n}_t^k. \tag{2.6}$$

2. **Model Differential Transmission (MDT).** The K participating clients only upload the differences between the latest local model and the previously received (noisy) global model, i.e., $\mathbf{x}_t^k = \mathbf{d}_t^k \triangleq \mathbf{w}_t^k - \hat{\mathbf{w}}_{t-1}^k$. For MDT, the server uses $\mathbf{d}_t^k$ and the previously computed global model $\mathbf{w}_{t-1}$ to reconstruct the updated local model of client k as

$$\tilde{\mathbf{w}}_t^k = \mathbf{w}_{t-1} + \hat{\mathbf{x}}_t^k = \mathbf{w}_{t-1} + \mathbf{d}_t^k + \mathbf{n}_t^k = \mathbf{w}_t^k + \mathbf{n}_t^k - \mathbf{e}_t^k. \tag{2.7}$$

The SNR for these two models, however, has to be defined slightly differently because we have normalized the ML model parameter $\mathbf{w}$ to have unit-variance elements in Section 2.2. Thus, for MT, the received SNR at the server for k-th client's signal is written as

$$\mathsf{SNR}_{t,k}^{\text{S,MT}} = \frac{\mathbb{E}\left\|\mathbf{w}_t^k\right\|^2}{\mathbb{E}\left\|\mathbf{n}_t^k\right\|^2} = \frac{1}{\sigma_{t,k}^2}. \tag{2.8}$$

To keep the SNR expression general, the MDT SNR denotes:

$$\mathsf{SNR}_{t,k}^{\text{S,MDT}} = \frac{\mathbb{E}\left\|\mathbf{d}_t^k\right\|^2}{\mathbb{E}\left\|\mathbf{n}_t^k\right\|^2} = \frac{\mathbb{E}\left\|\mathbf{d}_t^k\right\|^2}{d\sigma_{t,k}^2}, \tag{2.9}$$

since the variance of model difference $\mathbf{d}_t^k$ is unknown *a priori* and also changes over time.

Differences between MT and MDT, and why they are both considered. The different choices of MT and MDT are not considered in most of the literature because with a perfect communication assumption, there is no difference between them from a pure learning perspective – as long as the server can reconstruct $\mathbf{w}_t^k$, this aspect does not impact the learning performance [1]. However, the choice becomes significant when communication noises are present. From a practical system point of view, both schemes can be useful in different use cases. For example, MDT in the uplink relies on the server keeping the previous global model $\mathbf{w}_{t-1}$, from which the new local models can be reconstructed. This, however, may not always be true if the server deletes intermediate model aggregation (after broadcast) for privacy preservation [3], which makes reconstruction from the model differential infeasible.

Note that the *download* phase, on the other hand, does not have these two choices – the server should always transmit the global model $\mathbf{w}_{t-1}$ itself. This is due to the limitation introduced by partial (and random) clients participation, where the set of clients participating the t-th round can be totally different from the $(t-1)$-th round, and they do not have the previous global model to reconstruct based on the model difference.

Noise propagation. Both uplink and downlink channel noises collectively impact the received local models at the server. This noise propagation effect is more prominent in MDT ((2.7) explicitly has both noise terms). However, this effect in fact exists in both cases, because the local model is trained using the previously received global model, which contains the downlink noise.

(4) Global aggregation. The server aggregates the received local models to generate a new global ML model, following the standard FEDAVG [1]: $\mathbf{w}_t = \sum_{k\in\mathcal{S}_t} \frac{D_k}{\sum_{i\in\mathcal{S}_t} D_i}\tilde{\mathbf{w}}_t^k$. The server then moves on to the $(t+1)$-th round. For ease of exposition and to simply the analysis, we assume in the remainder of the paper that the local dataset sizes at all clients are the same[5]: $D_i = D_j, \forall i, j \in [N]$, which leads to the following simplifications.

1. **MT.** The aggregation can be simplified as

$$\mathbf{w}_t = \frac{1}{K}\sum_{k\in\mathcal{S}_t}\tilde{\mathbf{w}}_t^k = \frac{1}{K}\sum_{k\in\mathcal{S}_t}\hat{\mathbf{x}}_t^k = \frac{1}{K}\sum_{k\in\mathcal{S}_t}\left(\mathbf{w}_t^k + \mathbf{n}_t^k\right). \tag{2.10}$$

2. **MDT.** The aggregation can be written as

$$\mathbf{w}_t = \frac{1}{K}\sum_{k\in\mathcal{S}_t}\tilde{\mathbf{w}}_t^k = \mathbf{w}_{t-1} + \frac{1}{K}\sum_{k\in\mathcal{S}_t}\hat{\mathbf{x}}_t^k = \frac{1}{K}\sum_{k\in\mathcal{S}_t}\left(\mathbf{w}_t^k + \mathbf{n}_t^k - \mathbf{e}_t^k\right). \tag{2.11}$$

[5] We emphasize that all the results of this paper can be extended to handle different local dataset sizes.

For the case of MT, the SNR for the global model (after aggregation) can be written as

$$\mathsf{SNR}_t^G = \frac{\mathbb{E}||\sum_{k\in\mathcal{S}_t}\mathbf{w}_t^k||^2}{\mathbb{E}||\sum_{k\in\mathcal{S}_t}\mathbf{n}_t^k||^2} = \frac{\mathbb{E}||\sum_{k\in\mathcal{S}_t}\mathbf{w}_t^k||^2}{d\sigma_t^2}, \tag{2.12}$$

and for MDT, the SNR for the global model can be written as

$$\mathsf{SNR}_t^G = \frac{\mathbb{E}||\sum_{k\in\mathcal{S}_t}\mathbf{w}_t^k||^2}{\mathbb{E}||\sum_{k\in\mathcal{S}_t}(\mathbf{n}_t^k-\mathbf{e}_t^k)||^2} = \frac{\mathbb{E}||\sum_{k\in\mathcal{S}_t}\mathbf{w}_t^k||^2}{d(\sigma_t^2+\zeta_t^2)}, \tag{2.13}$$

where $\sigma_t^2 \triangleq \sum_{k\in\mathcal{S}_t}\sigma_{t,k}^2$ and $\zeta_t^2 \triangleq \sum_{k\in\mathcal{S}_t}\zeta_{t,k}^2$ denote the total uplink and downlink effective noise power for participating clients, respectively.

In general, $\{\mathbf{w}_t^k\}$ are correlated across clients because the local model updates all start from (roughly) the same global model. Intuitively, once FL convergences, these models will largely be the same, leading to a signal power term of dK^2 for the numerator. On the other hand, under the assumption that these local models are independent across clients, which is reasonable in the early phases of FL with large local epochs, where the (roughly) same starting point has diminishing impact due to the long training period and non-IID nature of the data distribution, the signal power term of dK is not enjoyed. Nevertheless, since the SNR control can be realized by adjusting the effective noise power levels, we focus on the impact of σ_t^2 and ζ_t^2 on the FL performance in Section 2.3.

In this section, the main focus is placed on analog communication for FL, where model parameters are transmitted in an analog manner. Therefore, digital communication processing such as source coding, channel coding and modulation are not incorporated. The adopted zero-mean bounded random noise assumption is reasonable for this setting, because it does not require any specific distribution and thus can be applicable to a broad range of analog communication systems.

2.3 Convergence Analysis of FL over Noisy Channels

2.3.1 Convergence Analysis for Model Transmission for Full Clients Participation

The convergence of FEDAVG in the presence of both uplink and downlink communication noise is first analyzed, when direct model transmission (MT) is adopted for local model upload: $\mathbf{x}_t^k = \mathbf{w}_t^k$. To simplify the analysis and highlight the key techniques in deriving the convergence rate, the analysis is under the assumption $K = N$ in this subsection (i.e., *full* clients participation with $\mathcal{S}_t = [K] = [N]$), and leave the case of partial clients participation to Section 2.3.2.

The main convergence result of MT with full clients participation is presented in the following Theorem 1.

Theorem 1. *Define $\phi = L/\mu$, $\gamma = \max\{8\phi, E\}$. Set learning rate as $\eta_t = 2/(\mu(\gamma + t))$ and adopt a SNR control policy that scales the effective uplink and downlink noise power over t such that:*

$$\sigma_t^2 \leq \frac{4N^2}{\mu^2(\gamma + t - 1)^2} \sim \mathcal{O}\left(\frac{1}{t^2}\right) \tag{2.14}$$

$$\zeta_t^2 \leq \frac{4N^2}{\mu^2(\gamma + t)(\gamma + t - 2)} \sim \mathcal{O}\left(\frac{1}{t^2}\right). \tag{2.15}$$

where $\sigma_t^2 \triangleq \sum_{k \in [N]} \sigma_{t,k}^2$ and $\zeta_t^2 \triangleq \sum_{k \in [N]} \zeta_{t,k}^2$ denote the total uplink and downlink effective noise power, respectively. Then, under Assumption $1-4$, the convergence of FEDAVG *with non-IID datasets and full clients participation satisfies*

$$\mathbb{E}\left\|\mathbf{w}_T - \mathbf{w}^*\right\|^2 \leq \frac{8L + \mu E}{\mu(T + \gamma)}\left\|\mathbf{w}_0 - \mathbf{w}^*\right\|^2 + \frac{4D}{\mu^2(T + \gamma)} \tag{2.16}$$

with $D = \sum_{k=1}^{N} \delta_k^2/N^2 + 6L\Gamma + 8(E-1)^2H^2 + 2d$.

A few remarks about Theorem 1 and its proof are now in order.

Remark 1. A complete proof of Theorem 1 can be found in Appendix A. The core technique utilized in Appendix A is the *perturbed iterate framework* that was pioneered in [58], especially the virtual sequence construction that have been widely adopted in the distributed SGD analysis [32, 30, 7, 6]. The unique challenge of this proof, however, is how to handle *simultaneous* uplink and downlink noises, which cannot be isolated from the SGD iterations. Not only do we have to incorporate more virtual sequences in the proof, but they also have the "coupling" effect in that downlink noise is present in the SGD steps and further in the new local model for uplink, while the uplink noise is present in the next-round downlink model. A careful manipulation of these coupled noise components in the various virtual sequences is a key analytical novelty of the proof.

Remark 2. It is important to clarify that although the requirement of Theorem 1 is presented in terms of the effective noise power, what ultimately matters is the SNR defined in Section 2.2. Controlling the effective noise power to scale as $\mathcal{O}(1/t^2)$ is equivalent to scaling the SNR as $\mathcal{O}(t^2)$, and can be implemented by either increasing the signal power (e.g., transmit power control) or reducing the post-processing noise power (e.g., receive diversity combining) while satisfying a fixed total resource budget constraint. The design examples that realize the requirement of Theorem 1 will be established in Section 2.4.

Remark 3. It is not surprising to see that Theorem 1 requires the SNR to increase, which gradually suppresses the noise effect as the FL process converges. There are, however, two unique characteristics about this theorem:

1. It characterizes a sufficient condition for the SNR **scaling law** as $\mathcal{O}(t^2)$. As we will see in Section 2.5, choosing a SNR scaling that is slower than $\mathcal{O}(t^2)$ degrades the FL performance.
2. This $\mathcal{O}(t^2)$ scaling law can be realized under a fixed total budget constraint. In other words, the benefit of Theorem 1 does not come from using more communication resources, but rather is due to a more judicious allocation (following the scaling law) of the same resource budget.

Remark 4. Theorem 1 guarantees that even under *simultaneous* uplink and downlink noisy communications, the same $\mathcal{O}(1/T)$ convergence rate of FEDAVG with perfect communications can be achieved if we control the effective noise power of both uplink and downlink to scale at rate $\mathcal{O}(1/t^2)$ and choose the learning rate at $\mathcal{O}(1/t)$ over t. Note that the choice of η_t to scale as $\mathcal{O}(1/t)$ is well-known in distributed and federated learning [32, 38, 31, 30], which essentially controls the "SGD noise" that is inherent to the stochastic process in SGD to gradually shrink as the FL process converges. We also note that for other learning rate choices in SGD, the fundamental insight of Theorem 1, i.e., *controlling the "effective channel noise" to not dominate the "SGD noise"*, is still valid.

Remark 5. Lastly, it is important to point out that the scaling law in Theorem 1 should be viewed as an *average* SNR requirement that changes over learning rounds. The time scale of changing the average SNR is on the order of learning rounds, which is much slower[6] than the time scale of the time-varying wireless channel. Furthermore, the SNR scaling law can be used in conjunction with other "faster" resource allocation mechanisms, such as inner-loop power control, to handle wireless dynamics under the average SNR budget decided from Theorem 1. This will become clear in Section 2.4.1.

2.3.2 Convergence Analysis for Model Transmission for Partial Clients Participation

The convergence analysis for *partial* clients participation is generalized next, where K clients ($K < N$) are uniformly randomly selected to form a set $\mathcal{S}_t$ at round t, and carry out the FL process. This section mostly follows the FL system model described in Section 2.2, with the only simplification on *homogeneous* noise power levels at the uplink and downlink, i.e.,

$$\sigma_{t,k}^2 = \bar{\sigma}_t^2, \quad \text{and} \quad \zeta_{t,k}^2 = \bar{\zeta}_t^2, \quad \forall t \in [T], k \in [N]. \tag{2.17}$$

The main reason to introduce this simplification is due to the time-varying randomly participating clients: since $\mathcal{S}_t$ changes over t, the total power levels also vary over t if we insist on heterogeneous noise power for different clients. Furthermore, since clients are randomly selected, the total power level becomes a random variable as well, which

[6] This is particularly true when there are a large number of clients participating in the FL process, as the length of learning rounds is often dominated by the "straggler" [59].

significantly complicates the convergence analysis. Making this assumption would allow us to focus on the challenge with respect to the model update from partial clients participation.

Theorem 2. *Let ϕ, γ and η_t be the same as in Theorem 1. Adopt a SNR control policy that scales the effective uplink and downlink noise power over t such that:*

$$\bar{\sigma}_t^2 \leq \frac{4K}{\mu^2(\gamma+t-1)^2} \sim \mathcal{O}\left(\frac{1}{t^2}\right) \tag{2.18}$$

$$\bar{\zeta}_t^2 \leq \frac{4N}{\mu^2(\gamma+t)(\gamma+t-2)} \sim \mathcal{O}\left(\frac{1}{t^2}\right). \tag{2.19}$$

where $\bar{\sigma}_t^2$ and $\bar{\zeta}_t^2$ represent the individual client effective noise in the uplink and downlink, respectively, which are defined in (2.17). Then, under Assumption $1-4$, the convergence of FEDAVG *with non-IID datasets and partial clients participation has the same convergence rate expression as (2.16), with D being replaced as $D = \sum_{k=1}^{N} \delta_k^2/N^2 + 4(N-K)E^2H^2/(K(N-1)) + 6L\Gamma + 8(E-1)^2H^2 + 2d$.*

The proof of Theorem 2 is given in Appendix B. It is clear that partial clients participation does not fundamentally change the behavior of FL in the presence of uplink and downlink communication noises. However, unlike full client participation, the uplink effective noise depends on the number of active users, which makes it harder to satisfy (2.18) compared with (2.19). The reason behind this difference is that, in partial client participation, the downlink process remains the same as the fully client participation, while the number of participants in the uplink process reduces from N to K. Therefore, the effective uplink noise can only be controlled by K rather than N participants, which implies that each user needs to allocate more transmission power than the fully client participation case to achieve the desired noise-free convergence rate of FL. A practical example will be provided in Section 2.4.2 to handle this tighter upper bound.

2.3.3 Convergence Analysis for Model Differential Transmission

This section considers the model different transmission (MDT) scheme when the clients upload model parameters. Since only model differential is transmitted, the receiver must possess a copy of the "base" model to reconstruct the updated model. This precludes using MDT in the downlink for partial clients participation, because participating clients differ from round to round, and a newly participating client does not have the "base" model of the previous round to reconstruct the new global model. We thus only focus on MDT in the uplink and MT in the downlink with partial clients participation.

Theorem 3. *Let ϕ, γ and η_t be the same as in Theorem 1, and the effective noise follows (2.17). Adopt a SNR control policy that maintains a constant uplink SNR at each client over t:*

$$\mathsf{SNR}_{t,k}^{S,MDT} = \nu \sim \mathcal{O}\left(1\right), \tag{2.20}$$

and scales the effective downlink noise power at each client over t such that:

$$\bar{\zeta}_t^2 \leq \frac{\frac{4}{\mu^2}}{\frac{1}{N}(\gamma+t)(\gamma+t-2)+\frac{1}{K}\left(1+\frac{1}{\nu}\right)(\gamma+t)^2} \sim \mathcal{O}\left(\frac{1}{t^2}\right). \tag{2.21}$$

Then, under Assumption $1-4$, the convergence of FEDAVG *with non-IID datasets and partial clients participation for uplink MDT and downlink MT has the same convergence rate expression as (2.16), with D being replaced as $D = \sum_{k=1}^{N} \delta_k^2/N^2 + 4(N-K)E^2H^2/(K(N-1)) + 4E^2H^2/(K\nu) + 6L\Gamma + 8(E-1)^2H^2 + d$.*

The complete proof of Theorem 3 can be found in Appendix C. It is instrumental to note that unlike direct model transmission, only transmitting model differentials in the uplink allows us to remove the corresponding SNR scaling requirement. Instead, one can keep a *constant* SNR in uplink throughout the entire FL process. Intuitively, this is because the "scaling" already takes place in the model differential $\mathbf{d}_t^k$, which is the difference between the updated local model at client k after E epochs of training and the starting local model. As FL gradually converges, this differential becomes smaller. Thus, by keeping a constant communication SNR, we essentially scales down the effective noise power at the server.

Lastly, It is important to note that the constant SNR requirement of Theorem 3 enables very simple implementation given the MDT SNR expression in (2.9). The signal power in the numerator of (2.9) is unknown and varying over learning rounds. However, a constant SNR requirement means one can fix the transmit power and "scales" individual $\mathbf{d}_t^k$ to have the desired power, without prior knowledge of its true variance.

2.4 Communication Design Examples for FL in Noisy Channels

An immediate engineering question following the previous analyses is how one can realize the effective noise power (or equivalently the SNR) specified in the theorems. A natural approach is *transmit power control*, which has the flexibility of controlling the average receive SNR (and thus the effective noise power) while satisfying a total power constraint. Specially, for a FL task with T total communication rounds and a given total power budget of P over all rounds, it is straightforward to compute that

$$P_t = 6Pt^2/(T(T+1)(2T+1)), \forall t = 1, \cdots, T, \tag{2.22}$$

where P_t is the desired average transmit power of the communication round t.

Since the analog communication and aggregation is the main consideration of this work, it is necessary take the wireless channel fading into account. To combat the influence of channel fading on received power, two design examples are now proposed to demonstrate how the proposed $\mathcal{O}(t^2)$-power increased strategy could be adopted in both continuous and discrete average power allocation schemes.

2.4.1 Design Example I: Transmit Power Control for Analog Aggregation

The following design presents a power control policy for the analog aggregation FL framework in [17, 19, 50], as an example to demonstrate the system design for FL tasks in the presence of communication noise.

The analog aggregation method in [17, 19, 50]. Consider a communication system where several narrowband orthogonal channels (e.g., sub-carriers in orthogonal frequency-division multiplexing (OFDM), time slots in time division multiple access (TDMA)) are shared by K random selected clients in an uplink model upload phase of a communication round. Each element in the transmitted model $\mathbf{w} \in \mathbb{R}^d$ is allocated and transmitted in a narrowband channel and aggregated automatically over the air. Denote the received signal of each element $i = 1, \cdots, d$ in the t-th communication round as

$$y_{t,i} = \frac{1}{K} \sum_{k \in \mathcal{S}_t} r_{t,k}^{-\alpha/2} h_{t,k,i} \sqrt{p_{t,k,i}} w_{t,k,i} + n_{t,i} \;\; \forall k \in \mathcal{S}_t,$$

where $r_{t,k}^{-\alpha/2}$ and $h_{t,k,i} \sim \mathcal{CN}(0,1)$ are the large-scale and small-scale fading coefficients of the channel, respectively, $n_{t,i} \sim \mathcal{CN}(0,1)$ is the additive Gaussian white noise of the channel, and $p_{t,k,i}$ denotes the transmit power determined by the power control policy. By the assumption on perfect CSIT, the channel inversion rule is used as in [17], which leads to the following *instantaneous* transmit power of user k at time t for model weight element i:

$$p_{t,k,i} = \frac{\rho_t^{\mathrm{UL}}}{r_{t,k}^{-\alpha} |h_{t,k,i}|^2}, \tag{2.23}$$

where ρ_t^{UL} is a scalar that denotes the uplink average transmit power, which is to be optimized. Hence, the receive SNR of the global model can be written as

$$\mathsf{SNR}_t^G = \mathbb{E}\left\| \frac{1}{K} \sum_{i=1}^{d} \frac{\sqrt{\rho_t^{\text{UL}}} \sum_{k\in\mathcal{S}_t} w_{t,k,i}}{n_{t,i}} \right\|^2 = \frac{\rho_t^{\text{UL}}\mathbb{E}\|\sum_{k\in\mathcal{S}_t} \mathbf{w}_t^k\|^2}{dK^2}. \tag{2.24}$$

Transmit power control. The original analog aggregation framework in [17] assumes that ρ_t^{UL} is a constant over time t. However, our theoretical analysis in Section 2.3 suggests that this can be improved. Specifically, taking partial clients participation and MT as an example, and further assuming IID weight elements, the transmit power should follow

$$\rho_t^{\text{UL}} = \frac{K}{\bar{\sigma}_t^2} \geq \frac{\mu^2(\gamma+t-1)^2}{4} \sim \mathcal{O}(t^2), \tag{2.25}$$

which implies that ρ_t should be increased at the rate $\mathcal{O}(t^2)$ in the uplink to ensure the convergence of FedAvg. Similar policy can be derived for MDT and/or full clients participation, by invoking the corresponding theorems.

In the downlink case, when the server broadcasts the global model to K randomly selected clients, the receive signal of the i-th element for the n-th user in the t-th communication round is

$$y_{t,n,i} = r_{t,n}^{-\alpha/2} h_{t,n,i} \sqrt{\rho_t^{\text{DL}}} w_{t,i} + e_{t,n,i} \quad \forall n = 1 \cdots K,$$

where $e_{t,n,i} \in \mathcal{CN} \sim (0,1)$ is the additive Gaussian white noise, and ρ_t^{DL} is the transmitted power at the server. The downlink SNR for the n-th user is

$$\mathsf{SNR}_{t,n,i}^L = r_{t,n}^{-\alpha} |h_{t,n,i}|^2 \rho_t^{\text{DL}}. \tag{2.26}$$

Instead of keeping ρ_t^{DL} as a constant, we derive the following policy based on Theorem 3 to guarantee the convergence of FedAvg:

$$\rho_t^{\text{DL}} \geq \frac{r_{t,k}^{\alpha} \mu^2(\gamma+t)(\gamma+t-2)}{4N|h_{t,n,i}|^2} \sim \mathcal{O}(t^2). \tag{2.27}$$

Finally, by applying the power control policy defined in Eqns. (2.25) and (2.27), FL tasks are able to achieve better performances under the same energy budget. This is also numerically validated in the experiment.

Remarks. Note that the proposed transmit power control only changes the *average* transmit power at learning rounds. Such method is often referred to as the *outer-loop power control* (OLPC) [60], which operates at a very slow time scale and only relies on the large-scale, stationary information of the wireless FL system. In fact, this method can be used in conjunction with a faster *inner-loop power control*, such as the channel inversion power in (2.23) or any other methods that handle the fast fading component or interference, to determine the instantaneous transmit power of the sender.

Another minor note is that the pathloss component appears in (2.27) but not in (2.25). This is due to the broadcast nature of download. For upload, the pathloss is absorbed in the channel inversion expression (2.23).

2.4.2 Design Example II: Receive Diversity Combining for Analog Aggregation

Another technique that can benefit from our theoretical results is to control the *diversity order* of a receiver combining scheme, such as using multiple receive antennas, multiple time slots, or multiple frequency resources. Essentially, it leverages the repeated transmissions to reduce the effective noise power via receive diversity combining, and by only activating sufficient diversity branches as the progress over the learning rounds, resources can be more efficiently utilized.

Uplink diversity requirement. Denoting that the uploaded local model is independently received L_t times (over time, frequency, space, or some combination of them) in the t-th round; and reusing the notations and the channel inversion rule in (2.23), the L_t received signals for the i-th element can be given as

$$y_{t,i,l} = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\rho_{t,l}}w_{t,k,i} + n_{t,i,l} \;\; \forall k \in \mathcal{S}_t, \;\; \forall l = 1 \cdots L_t.$$

For simplicity, the average transmit power for each branch is assumed to be fixed: $\rho_{t,l} = \rho_0$, but this can be easily extended to incorporate power allocation over diversity branches [60]. The receive SNR of the global model after the diversity combining can be written as

$$\mathsf{SNR}_t^G = \mathbb{E}\left\|\sum_{i=1}^{d}\frac{\sum_{l=1}^{L_t}\frac{\sqrt{\rho_{t,l}}}{K}\sum_{k\in\mathcal{S}_t} w_{t,k,i}}{\sum_{l=1}^{L_t} n_{t,i,l}}\right\|^2 = L_t\frac{\rho_0\mathbb{E}||\sum_{k\in\mathcal{S}_t}\mathbf{w}_t^k||^2}{dK^2}.$$

Compared with the SNR of the power control policy in (2.24), we can derive the diversity requirement as

$$L_t = \left\lceil \rho_t^{\mathrm{UL}}/\rho_0 \right\rceil, \tag{2.28}$$

where $\lceil a \rceil$ denotes the ceiling operation on a.

Downlink diversity requirement. The server broadcasts the global weight for Q_t times (again it can be over time, frequency, space, or some combination of them) in the t-th round and each client combines the multiple independent copies of the received signals to achieve a higher SNR (i.e., lower effective noise power). The receive signal at client k can be written as

$$y_{t,k,i,q} = r_{t,k,i,q}^{-\alpha/2}h_{t,k,i,q}\sqrt{\rho_{t,q}}w_{t,i} + e_{t,k,i,q} \;\; \forall q = 1\cdots Q_t,$$

where $\rho_{t,q} = \rho_1$ is the (constant) transmit power at the server. The downlink SNR for the k-th user is $\mathsf{SNR}_{t,k}^{L} = r_{t,k}^{-\alpha} Q_t \rho_1$. Similarly, compared with the local SNR in (2.26), we can derive the diversity requirement as

$$Q_t = \left\lceil \rho_t^{\mathrm{DL}} / \rho_1 \right\rceil . \tag{2.29}$$

By applying the combining rules in Eqns. (2.28) and (2.29), the receive diversity combining policy that can guarantee the convergence of FL at rate $\mathcal{O}(1/T)$ is completed, under the transmit power constraints at both clients and server.

Remarks. Receive diversity combining is not as flexible as power control, because it can only achieve *discrete* effective noise power levels. This is also observed in the experiments. However, it can be useful in situations where adjusting the average transmit power is not feasible, e.g., no change at the transmitter is allowed. In addition, one can combine the transmit power control in Section 2.4.1 with the receive diversity combining in Section 2.4.2 in a straightforward manner. Note that there are other methods, such as increasing the precision of analog-to-digital converters (ADC), to implement the SNR control policy. The general design principles in Theorems 1 to 3 can be similarly realized.

2.5 Experiment Results

2.5.1 Experiment Setup

In experiments, various FL tasks are considered under noisy uplink and downlink communications. For simplicity, every channel use in the simulation has the same noise level, and both uplink and downlink have the same total energy budget $P = \sum_{t=1}^{T} P_t$, where P_t is the transmission power of the t-th round, $t = 1, \cdots, T$. However, note that the downlink energy is consumed only by the server (i.e., P_t), while the uplink budget is equally shared among all clients (i.e., P_t/N per transmitter), resulting in significantly smaller uplink transmit power per transmitter than the downlink. In each round of FL, the updated (locally or globally) ML model (or model differential when applicable) is transmitted over the noisy channel as described in Section 2.2. The following four schemes are considered in the experiments.

1. **Noise free.** This is the ideal case with no noise in either uplink or downlink. The accurate model parameters are perfectly received at the server and clients. This represents the best-case performance.
2. **Equal power allocation.** This corresponds to $P_t = P/T, \forall t = 1, \cdots, T$, as used in [17]. We adopte a normalized transmitted power $P_t = 1$ and the receive SNR of the model parameters is set as 10 dB in the experiments.
3. $\mathcal{O}(t^2)$**-increased power control policy.** Transmit power increases at the rate of $\mathcal{O}(t^2)$ with the round t but the overall energy consumption is kept constant as other methods, i.e., the receive SNR is increased and the effective

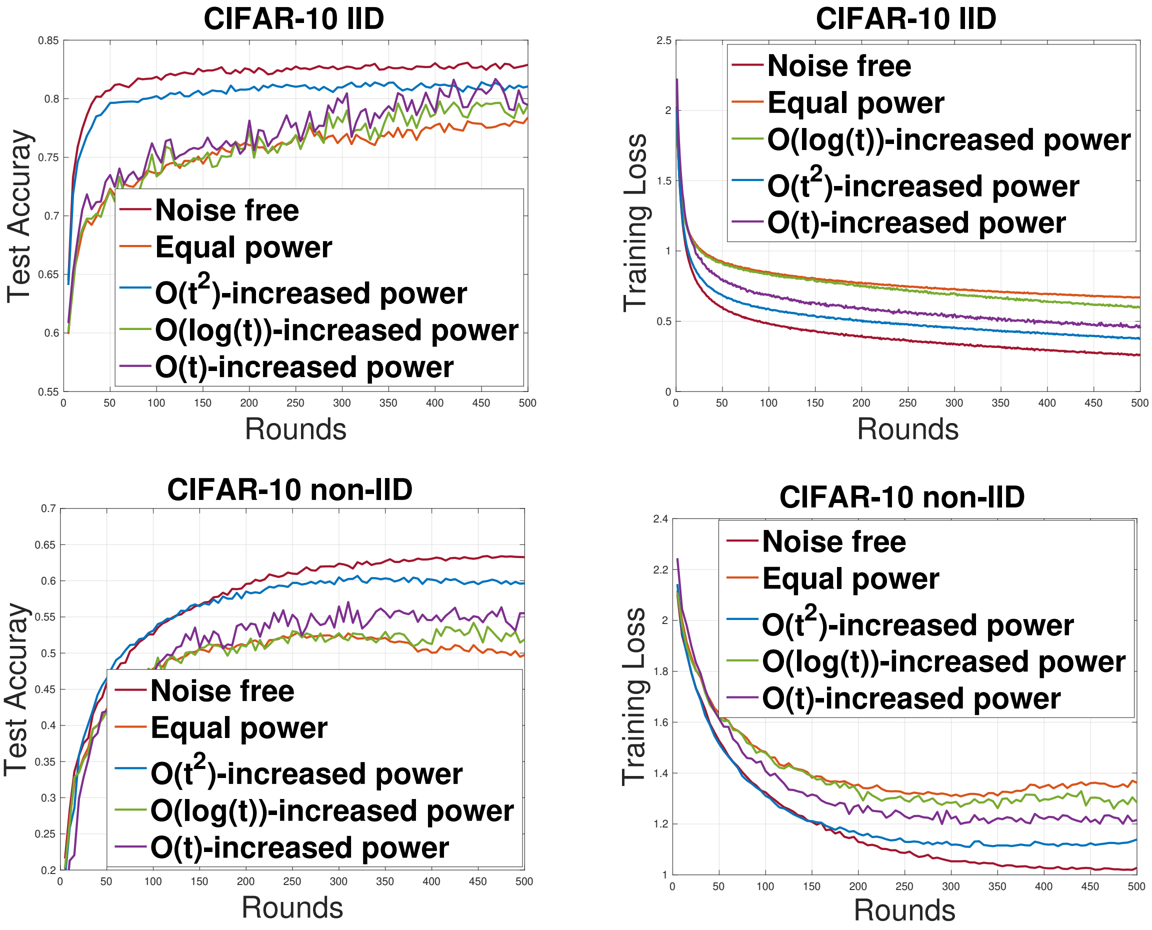


Figure 2.2: Comparing the performance of transmit power control to the baselines with full clients participation, model transmission, and both IID (left two) and non-IID (right two) FL on the CIFAR-10 dataset.

noise of the signal is decreased with the progress of FL. With the total budget P, (2.22) gives the power allocation solution.

4. $\mathcal{O}(t^2)$**-increased diversity combining policy.** The transmit power in both downlink and uplink remains the same as 2). However, the final models at the server and clients of each communication round are obtained by multiple repeated transmissions and the subsequent combining. The number of the repeated transmissions increases at the rate of $\mathcal{O}(t^2)$. For simple discretization, we use 1, 4, 9, 16 and 25 orders of receive diversity combining in both uplink and downlink model transmissions for 1st to 9th, 10th to 45th, 46th to 125th, 125th to 270th, and 270th to 500th communication round, respectively, of a 500-round task. Note that the total energy budget remains the same as the previous two methods.

The standard image classification and natural language processing FL tasks over different datasets are used to evaluate the performances of these schemes. The following three standard datasets are used in the experiments, which

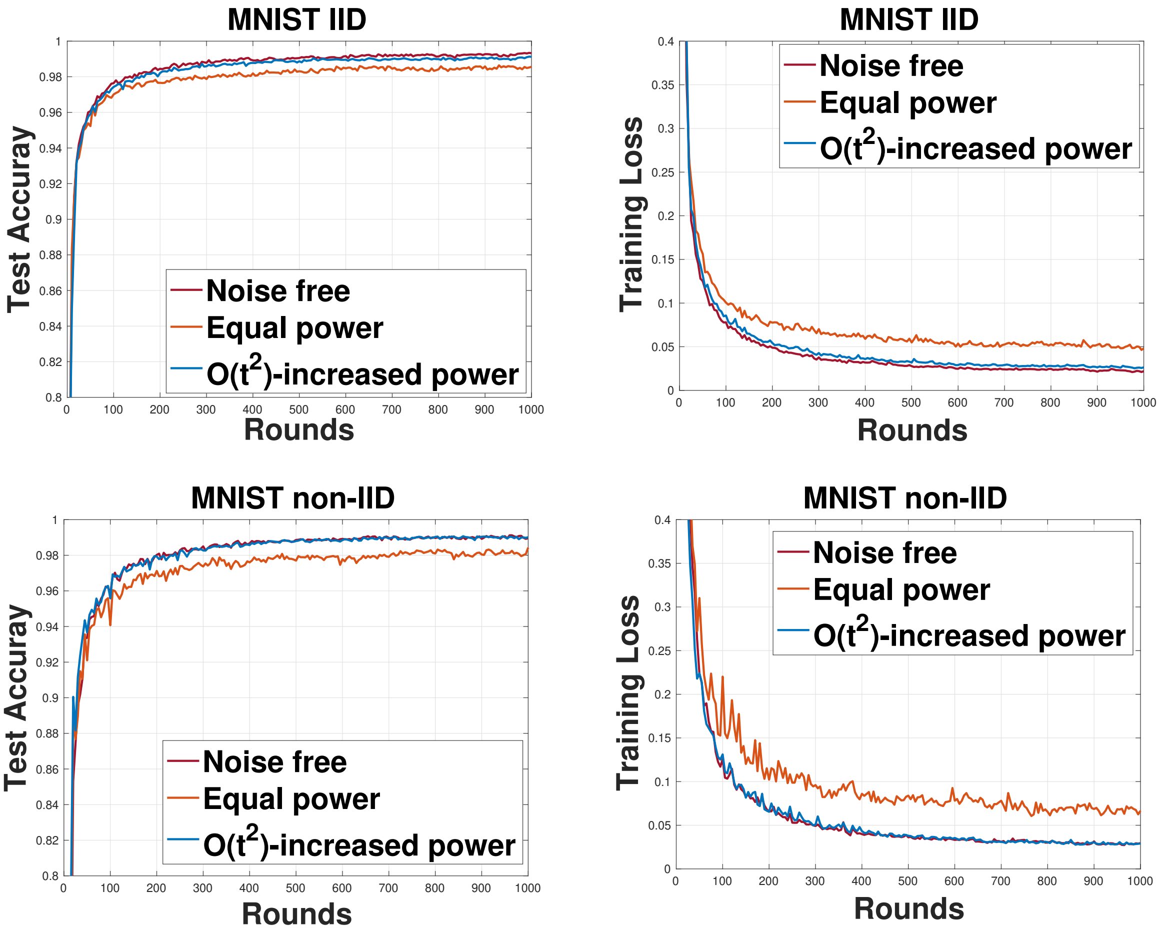


Figure 2.3: Comparing the performance of transmit power control to the baselines with partial clients participation, model transmission, and both IID (left two) and non-IID (right two) FL on the MNIST dataset.

are commonly accepted as the benchmark tasks to evaluate the performance of FL.

1. **MNIST.** The training sets contains 60000 examples. For the full clients participation case, the training sets are evenly distributed over $N = K = 10$ clients. For the partial clients participation case, the training sets are evenly partitioned over $N = 2000$ clients each containing 30 examples, and we set $K = 20$ per round (1% of total users). For the IID case, the data is shuffled and randomly assigned to each client, while for the non-IID case the data is sorted by labels and each client is then randomly assigned with 1 or 2 labels. The CNN model has two 5×5 convolution layers, a fully connected layer with 512 units and ReLU activation, and a final output layer with softmax. The first convolution layer has 32 channels while the second one has 64 channels, and both are followed by 2×2 max pooling. The following parameters are used for training: local batch size $BS = 5$, the number of local epochs $E = 1$, and learning rate $\eta = 0.065$.

2. **CIFAR-10.** We set $N = K = 10$ for the full clients participation case while $N = 100$ and $K = 10$ for the partial

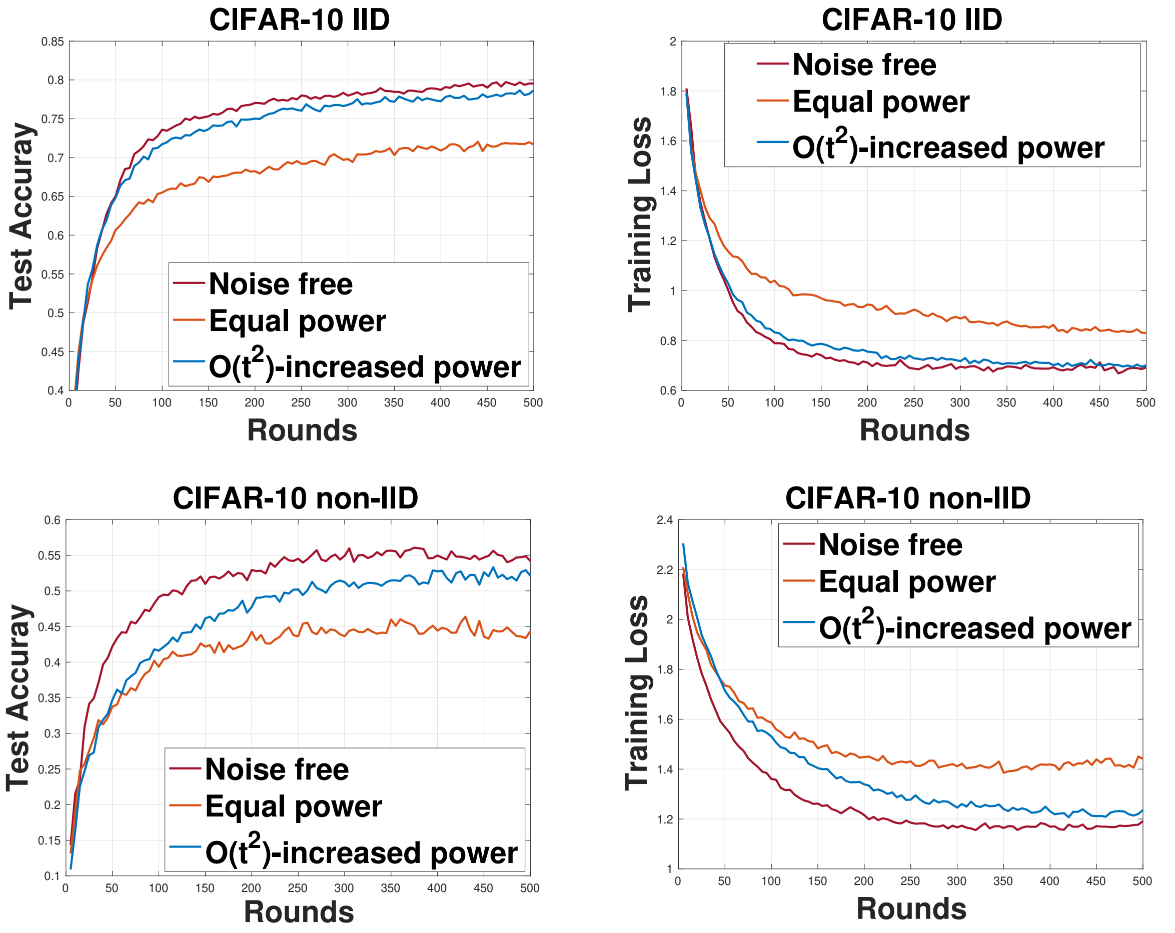


Figure 2.4: Comparing the performance of transmit power control to the baselines with partial clients participation, model transmission, and both IID (left two) and non-IID (right two) FL on the CIFAR-10 dataset.

clients participation case. We train a CNN model with two 5×5 convolution layers (both with 64 channels), two fully connected layers (384 and 192 units respectively) with ReLU activation and a final output layer with softmax. The two convolution layers are both followed by 2×2 max pooling and a local response norm layer. The training parameters are: (a) IID: $BS = 50$, $E = 5$, learning rate initially sets to $\eta = 0.15$ and decays every 10 rounds with rate 0.99; (b) non-IID: $BS = 100$, $E = 1$, $\eta = 0.1$ and decay every round with rate 0.992.

3. **Shakespeare.** This dataset is built from *The Complete Works of William Shakespeare* and each speaking role is viewed as a client. Hence, the dataset is naturally unbalanced and non-IID since the number of lines and speaking habits of each role vary significantly. There are totally 1129 roles in the dataset [61]. We randomly pick 300 of them and build a dataset with 794659 training examples and 198807 test examples. We also construct an IID dataset by shuffling the data and redistribute evenly to 300 roles and set $K = 10$. The ML task is the next-character prediction, and we use a classifier with an 8D embedding layer, two LSTM layers (each with 256 hidden units) and a softmax

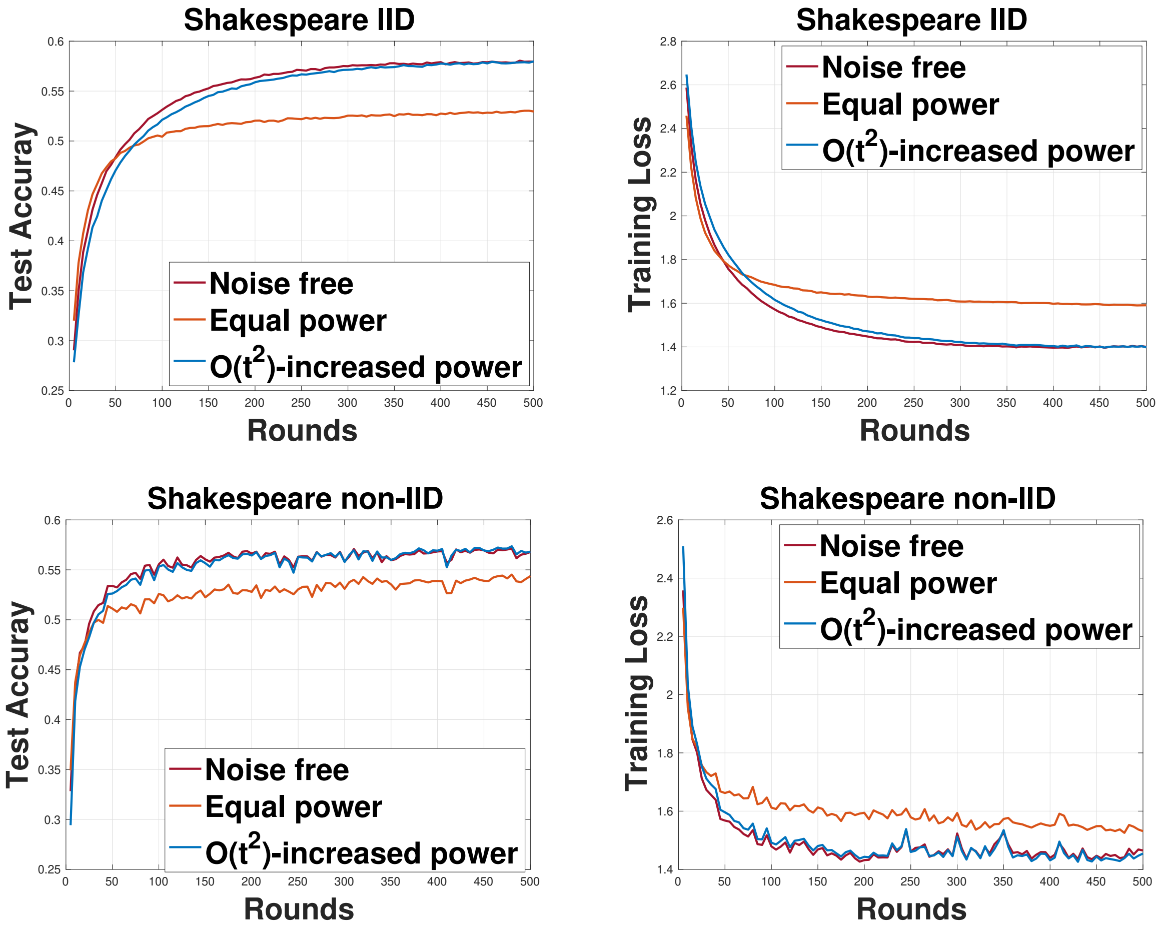


Figure 2.5: Comparing the performance of transmit power control to the baselines with partial clients participation, model transmission, and both IID (left two) and non-IID (right two) FL on the Shakespeare dataset.

output layer with 86 nodes. The training parameters are: $BS = 20$, $E = 1$, learning rate initially sets to $\eta = 0.8$ and decays every 10 rounds with rate 0.99.

The performance for all the aforementioned configurations is evaluated by comparing the test accuracies and training losses as functions of the communication rounds. All of the reported results are obtained by averaging over 5 independent runs. We also report the final test accuracy, which is averaged over the last 10 rounds, as the performance of the final global model.

2.5.2 Experiment Results for Transmit Power Control

The focus of the experiment is on partial clients participation under both MT and MDT, but results for full clients participation in CIFAR-10 is firstly reported, to highlight some common observations across all experiments.

Full clients participation. As seen from Fig. 2.2 that under the same total power budget, the $\mathcal{O}(t^2)$ power control policy performs better than the equal power allocation scheme and is very close to the noise-free ideal case. Specifically, $\mathcal{O}(t^2)$ power control policy achieves 81.1% and 59.6% final test accuracy in IID and non-IID data partitions on CIFAR-10, which is 2.6% and 9.8% better than that of the equal power allocation scheme. Note that the training loss (test accuracy) of equal power allocation scheme increases (decreases) during the late rounds (350th to 500th) in the non-IID case, implying that a non-increasing SNR may occur deterioration in the convergence of FL for more difficult ML tasks.

To further validate the $\mathcal{O}(t^2)$ scaling, experiments where power is increased as a slower rate of $\mathcal{O}(\log(t))$ and $\mathcal{O}(t)$ are carried out. The resulting performance is much worse than the $\mathcal{O}(t^2)$ scaling, and in fact has only very limited improvement over the equal power allocation.

Lastly, note that the early rounds of all methods have very similar performance. This is because although $\mathcal{O}(t^2)$ power control allocates less power than the equal power policy, both are dominated by the noise of SGD in early rounds and thus their performances are similar. This phenomenon is also observed in other experiments, which again highlights the benefits of adaptively "flying under the radar", to only allocate sufficient-but-not-excessive transmit power in each round. All of the aforementioned observations carry over to other tasks and different FL configurations.

Partial clients participation. The performance comparisons of the three schemes on MNIST, CIFAR-10 and Shakespeare datasets in both IID and non-IID configurations and MT are reported in Figs. 2.3, 2.4, and 2.5, respectively. Their final model accuracies (after T rounds of FL are complete) are also summarized in Table 2.1. First, as seen from Fig. 2.3, the proposed $\mathcal{O}(t^2)$-increased power allocation scheme achieves higher test accuracy and lower train loss than the equal power allocation scheme under the same energy budget on MNIST. In particular, $\mathcal{O}(t^2)$-increased power allocation scheme achieves 0.6% higher test accuracy than that of equal power allocation scheme in both IID and non-IID data partitions, respectively. It may seem that the gain is insignificant, but the reason is mostly due to that MNIST classification is a very simple task. In fact, the gain of power control is much more notable under the challenging CIFAR-10 and Shakespeare tasks as shown in Figs. 2.4 and Fig. 2.5, respectively. Compared with the equal power allocation scheme, which achieves 90.2% and 81.6% of the ideal (noise free) test accuracy in IID and non-IID data partitions under CIFAR-10 dataset respectively, the proposed $\mathcal{O}(t^2)$-increased power allocation achieves 99.2% (IID) and 95.9% (non-IID) of the ideal (noise free) test accuracy respectively after $T = 500$ communication rounds. Similarly, under Shakespeare dataset, the equal power allocation scheme achieves 91.5% (IID) and 95.8% (non-IID) of the ideal (noise free) test accuracy, while the proposed method improves 8.5% and 3.5%, respectively.

MDT. The experiment results of model differential transmission is presented next. Note that, by applying MDT, the uplink transmission power of the proposed scheme remains constant (recall that SNR is set as 10dB) while the

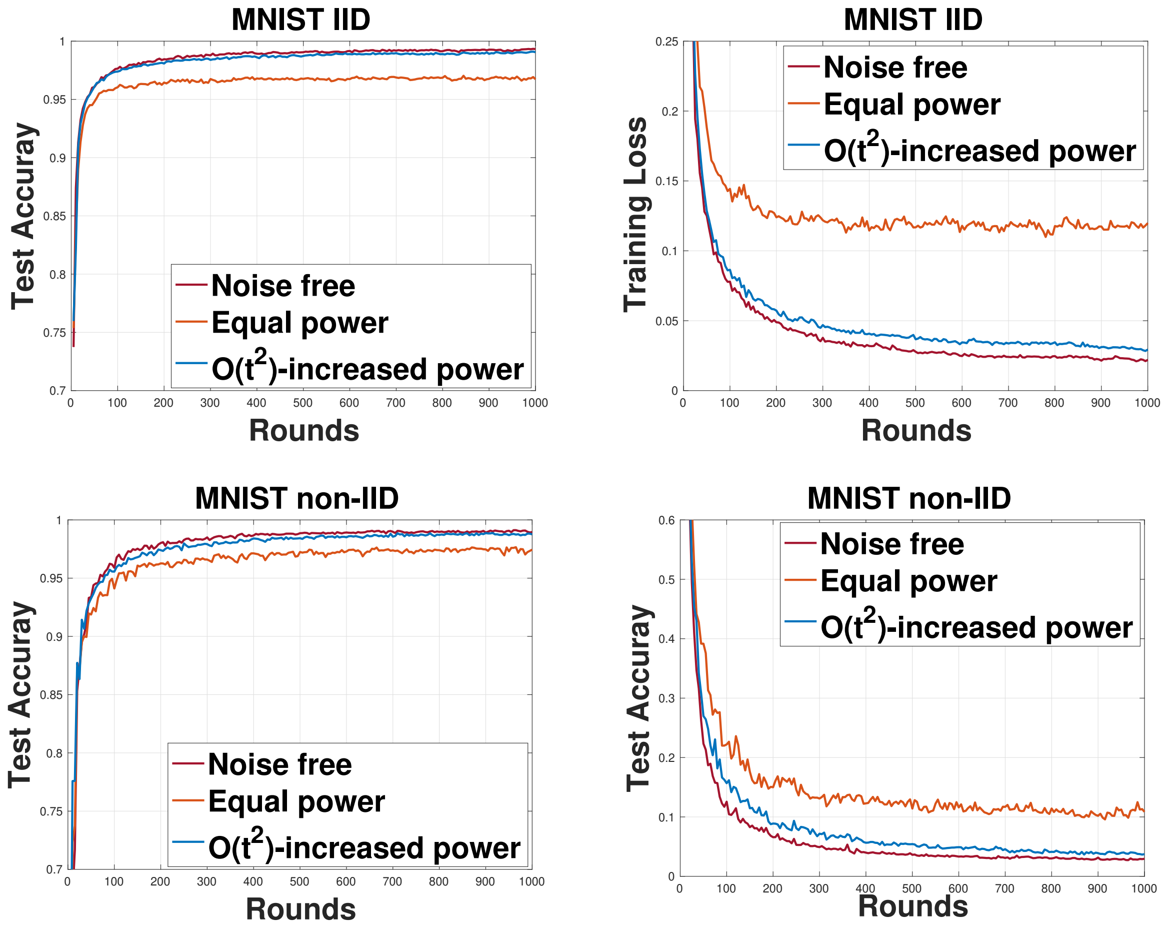


Figure 2.6: Comparing the performance of transmit power control to the baselines with partial clients participation, model differential transmission, and both IID (left two) and non-IID (right two) FL on the MNIST dataset.

downlink transmission power still increases at the rate of $\mathcal{O}(t^2)$. Figs. 2.6, 2.7 and 2.8 illustrate the test accuracies and training losses with MDT under MNIST, CIFAR-10 and Shakespeare datasets and the final model accuracies of the three schemes are summarized in Table 2.2. The proposed power control policy achieves 99.7% (99.7%), 99.2% (98.0%) and 100% (98.9%) of the ideal test accuracy in IID (non-IID) data setting under MNIST, CIFAR-10 and Shakespeare datasets, respectively, which significantly outperforms the baseline equal power allocation scheme.

2.5.3 Experiment Results for Receive Diversity Combining

The performance of receive diversity combining is evaluated on CIFAR-10 as the final part of the experiments. Fig. 2.9 captures the test accuracies and training losses of receive diversity combining together with noise free and equal power allocation schemes. Although receive diversity combining is less flexible than the (continuous) transmit power control policy, we can see that it still outperforms the baseline method and approaches the noise-free ideal case. Notice that

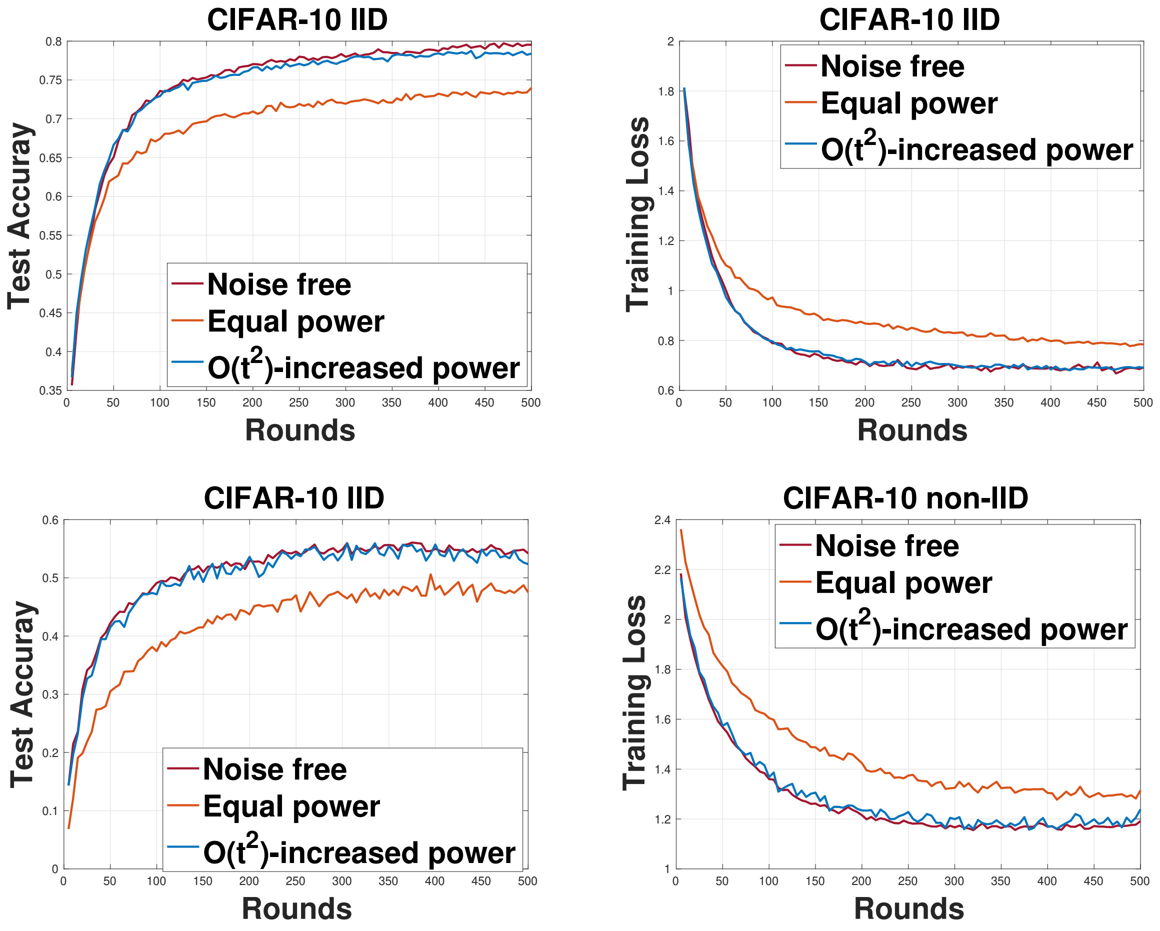


Figure 2.7: Comparing the performance of transmit power control to the baselines with partial clients participation, model differential transmission, and both IID (left two) and non-IID (right two) FL on the CIFAR-10 dataset.

Table 2.1: Performance Summary of MT

Dataset	Scheme	Accuracy	Percentage*	Accuracy	Percentage*
		IID		non-IID	
MNIST	Noise free	99.3%	100%	99.1%	100%
	Increased power	99.1%	99.8%	99.0%	99.9%
	Equal power	98.5%	99.2%	98.4%	99.3%
CIFAR-10	Noise free	79.5%	100%	54.3%	100%
	Increased power	78.9%	99.2%	52.1%	95.9%
	Equal power	71.7 %	90.2%	44.3%	81.6%
Shakespeare	Noise free	57.8%	100%	56.8%	100%
	Increased power	57.8%	100%	56.4%	99.3%
	Equal power	52.9 %	91.5%	54.4%	95.8%

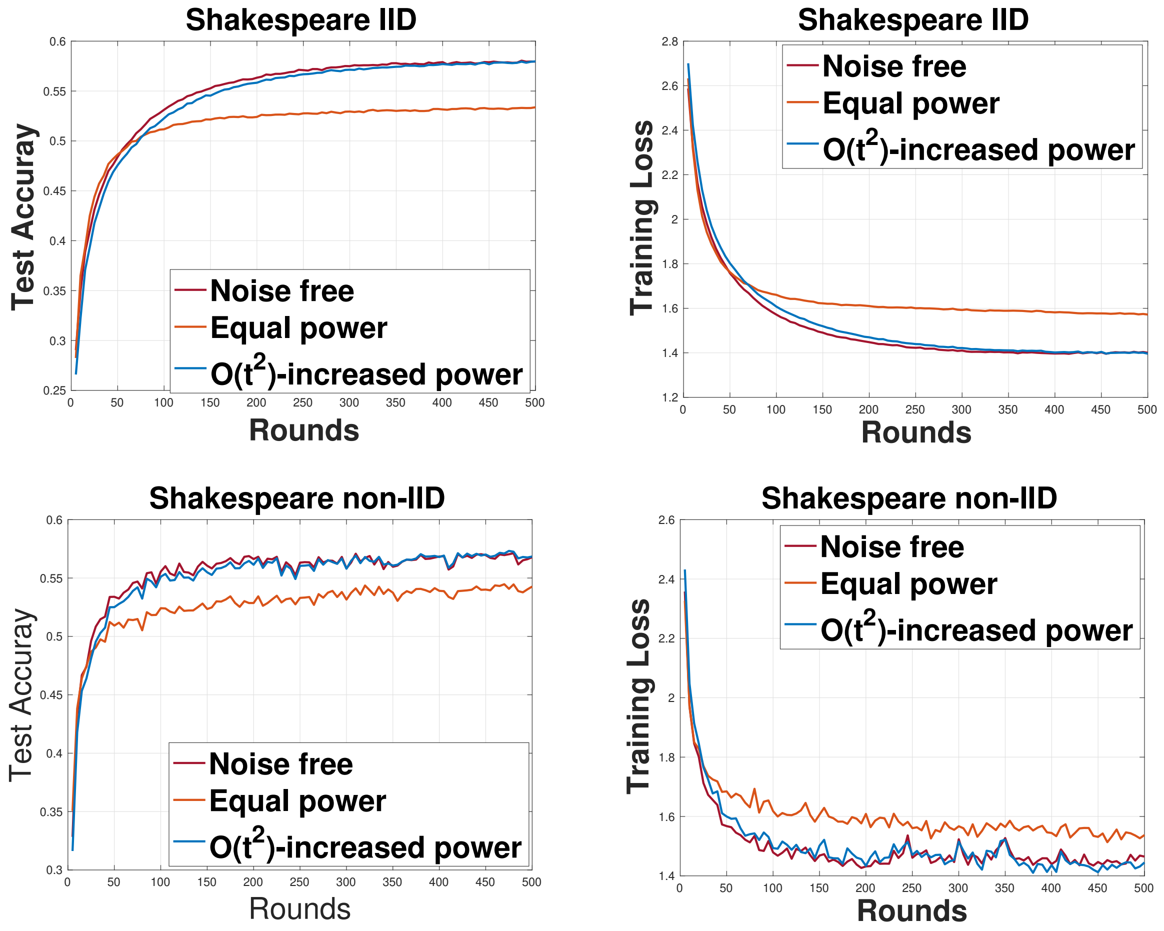


Figure 2.8: Comparing the performance of transmit power control to the baselines with partial clients participation, model differential transmission, and both IID (left two) and non-IID (right two) FL on the Shakespeare dataset.

Table 2.2: Performance Summary of MDT

Dataset	Scheme	Accuracy	Percentage*	Accuracy	Percentage*
		IID		non-IID	
MNIST	Noise free	99.3%	100%	99.1%	100%
	Increased power	99.0%	99.7%	98.8%	99.7%
	Equal power	96.7%	97.4%	97.5%	98.4%
CIFAR-10	Noise free	79.5%	100%	54.3%	100%
	Increased power	78.9%	99.2%	53.2%	98.0%
	Equal power	73.9 %	93.0%	47.7%	87.8%
Shakespeare	Noise free	57.8%	100%	56.8%	100%
	Increased power	57.8%	100%	56.2%	98.9%
	Equal power	53.3 %	92.2%	54.3%	95.6%

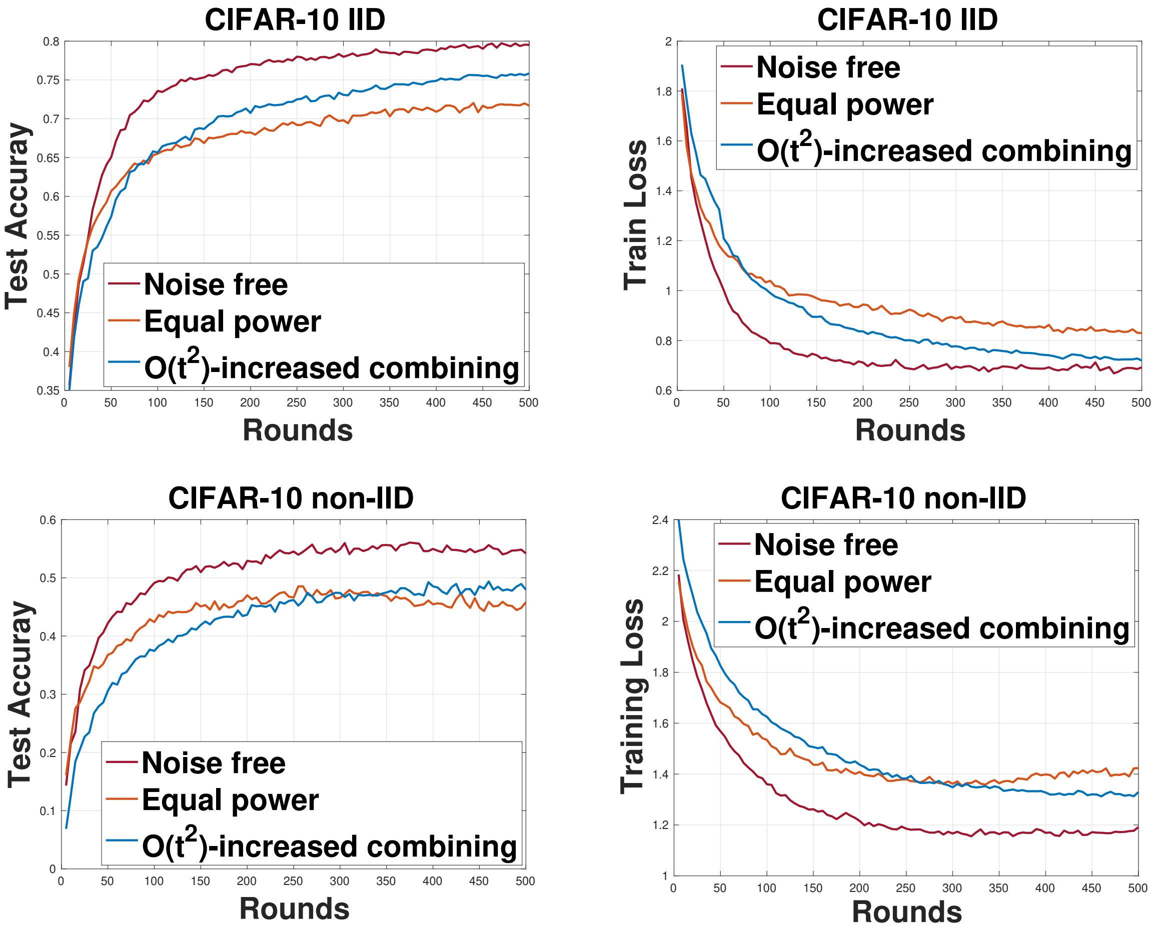


Figure 2.9: Comparing the performance of receive diversity combining to the baselines with partial clients participation, model differential transmission, and both IID (left two) and non-IID (right two) FL on the CIFAR-10 dataset.

the training losses of receive diversity combining are larger than those of the equal power allocation scheme at the beginning stage of convergence, but as the diversity branches increase, the training losses eventually reduce and the model converges to a better global one. Particularly, receive diversity combining achieves 75.6% and 47.8% test accuracies for IID and non-IID data partitions, which is 3.9% and 3.4% better than the equal power allocation scheme.

2.6 Summary

In this chapter, the exploration focuses on wireless FL conducted over noisy channels, examining a FedAvg pipeline impacted by both uplink and downlink communication noises. Through a rigorous theoretical analysis of model training convergence, it has been established that FedAvg can achieve the conventional $\mathcal{O}(1/T)$ convergence rate observed under ideal (noise-free) communication conditions. This outcome is feasible provided that the SNRs for uplink and downlink communications are appropriately managed at $\mathcal{O}(t^2)$ in scenarios of direct model transmission, and maintained at $\mathcal{O}(1)$ for differential model transmission across noisy channels. This chapter highlights the practical application of two prevalent communication strategies—transmit power control and receive diversity combining—to realize these theoretical insights. Comprehensive experimental investigations support the theoretical findings, showcasing the enhanced performance of these advanced strategies compared to conventional methods within the constraints of an identical total energy budget.

Chapter 3

Random Orthogonalization for Federated Learning in Massive MIMO Systems

The SNR scaling rules established in last chapter act as a fundamental design guideline for wireless FL systems. In this chapter, detailed physical layer design for AirComp FL will be proposed. As a potential solution to tickle the scalability challenge in wireless FL, the existing work on AirComp has several limitations. First, these methods often require CSIT for each individual client. The process of enabling individual CSIT is complicated – in a frequency division duplex (FDD) system, this involves the receiver estimating the channels and then sending back the estimates to the transmitters; in a time division duplex (TDD) system, one can benefit from channel reciprocity [21, 60], but there is still a need for an independent pilot for each client. In both cases, practical mechanisms to obtain individual CSIT do not scale with the number of clients. In addition, the precision of CSIT is often worse than CSIR. Second, most AirComp approaches in the literature require a channel inversion-type power control, which is well known to "blow up" when at least one of the channels is experiencing deep fading [21]. Third, AirComp approaches focus on improving the scalability and efficiency of the uplink communication phase in FL. How to address these challenges in the downlink communication phase remains underdeveloped.

Another important limitation is that the AirComp solution does not naturally extend to MIMO systems where the uplink and downlink channels become vectors. Compared with the studies in scalar channels, there are only a few recent papers that explore the potential of MIMO for wireless FL. MIMO beamforming design to optimize FL has been studied in [18, 62]. Coding, quantization, and compressive sensing over MIMO channels for FL have been studied in [63, 64]. Nevertheless, none of these works tightly incorporates the unique properties of MIMO to the FL communication design. On the other hand, if the unique characteristics of FL is ignored, MIMO can also be utilized in a straightforward manner. In the uplink phase, the conventional MIMO estimators such as zero-forcing (ZF) or minimum mean square error (MMSE) can be used to estimate each local model, and then the global model can be computed. In the downlink phase, MIMO precoders is designed to broadcast the global model. However, these approaches incur a large channel estimation overhead, especially when the channels have high dimensions. Moreover, matrix inversions in the ZF or MMSE estimators and the optimization algorithms for the precoding design are computationally demanding, in particular for massive MIMO. This increases the complexity and latency of the overall system. In addition, decoding individual local models also makes it easier for the server to sketch the data distribution of the clients, leading to potential privacy leakage.

This chapter aims at designing simple-yet-effective FL communication methods that can efficiently address the scalability challenge in FL for both uplink and downlink phases. The novelty comes from a tight integration of MIMO and FL – our design explicitly utilizes the characteristics of both components. The contributions are summarized as follows.

- A novel *Random Orthogonalization* design for *massive* MIMO is proposed. In uplink communications, by leveraging the unique channel hardening and favorable propagation properties of massive MIMO, the proposed framework only requires the BS to estimate a *summation channel* and allows it to directly compute the global model via a simple linear projection, which significantly alleviates the burden on channel estimation[1] and achieves extremely low complexity and low latency. In downlink communications, the proposed method leads to a simple but highly effective model broadcast method for FL. Moreover, our approach is agnostic to the number of clients, and thus improves the scalability of FL.

- Considering that conventional random orthogonalization techniques depend on channel hardening and favorable propagation to mitigate interference—conditions that may not always be present in practical scenarios, such as when the number of antennas is limited—we propose a refined strategy termed *enhanced random orthogonalization*. This advanced approach incorporates *channel echoes* as a means to offset the absence of channel hardening and favorable propagation. Consequently, the enhanced random orthogonalization design is adaptable to a broader range of MIMO systems, making it suitable for both uplink and downlink FL communications. This adaptability ensures more reliable and efficient communication across various network conditions, thereby extending the applicability of random orthogonalization techniques in practical MIMO deployments.

- To evaluate the efficacy of random orthogonalization strategies, the Cramer-Rao Lower Bounds (CRLBs) for the average model estimation errors were derived, establishing a theoretical benchmark for performance analysis.

[1] For example, a single pilot can be used by all clients as long as it is sent synchronously, regardless of the number of clients that participate in the current FL round.

Furthermore, considering both interference and noise, a novel convergence bound for FL across massive MIMO channels was formulated, elucidating the explicit relationship between the convergence rate, the number of clients, and the number of antennas. This analysis provides critical design insights for wireless FL systems. Extensive numerical experiments demonstrate the effectiveness and efficiency of the random orthogonalization approach in a broad range of FL and MIMO scenarios.

The remainder of this chapter is organized as follows. Related works are surveyed in Section 3.1. Section 3.2 introduces the FL pipeline and the wireless communication model. The proposed random orthogonalization principle is presented in Section 3.3, and then the enhanced design is proposed in Section 3.4. Analyses of the CRLB as well as the FL model convergence are given in Section 3.5. Experimental results are reported in Section 3.6, followed by a summary in Section 3.7.

3.1 Related Works

Improve FL communication efficiency. The original Federated Averaging (FedAvg) algorithm [1] reduces the communication overhead by only periodically averaging the local models. Theoretical understanding of the communication-computation tradeoff has been actively pursued and, depending on the underlying assumptions (e.g., independent and identically distributed (i.i.d.) or non-i.i.d. local datasets, convex or non-convex loss functions, gradient descent or stochastic gradient descent (SGD)), convergence analyses have been carried out [32, 30]. The approaches to reduce the payload size or communication frequency include sparsification [65, 66] and quantization [39, 40, 8]. There are also efforts to improve resource allocation [67, 68, 69].

AirComp for FL. As a special case of computing over multiple access channels [70], AirComp [17, 18, 19, 20] leverages the signal superposition properties in a wireless multiple access channel to efficiently compute the average ML model. This technique has attracted considerable interest, as it can reduce the uplink communication cost to be (nearly) agnostic to the number of participating clients. Client scheduling and various power and computation resource allocation methods have been investigated [71, 48, 72, 73, 74, 75]. The assumption of full CSIT is relaxed in [76] by only using the phase information of each individual channel. Convergence guarantees of Aircomp under different constraints are reported in [77, 78, 79, 80, 81].

Communication design for FL in MIMO systems. There are some recent studies on optimizing the communication efficiency and learning performance in MIMO systems for FL, including transmit power control [82, 83, 84], data rate allocation [85], and compression [64, 86]. Several beamforming designs have been proposed to improve the performance of wireless FL [87, 18, 88, 89, 34]. However, these methods require full CSIT and rely on complex optimization methods to design the beamformers, which becomes less attractive in massive MIMO due to the high communication and computation cost. Asymptotic analysis of the aggregation error in massive MIMO is provided in [87, 90, 91], which leads to beamformer designs that can relax the individual CSIT assumption in wireless FL. However, they only focus on the uplink communication phase.

3.2 System Model

Consider a MIMO TDD communication system equipped with M antennas at the BS (server) where K randomly-selected single-antenna devices (clients) are involved in the t-th round of the aforementioned FL task. Let $\mathbf{h}_k \in \mathbb{C}^{M\times 1}$ denote the uplink wireless channel between the k-th client and the BS. During the uplink communication phase, each client transmits the difference between the received global model and the newly computed local model

$$\mathsf{x}_t^k = \mathbf{w}_t - \mathbf{w}_{t+1}^k, \ \ \forall k \in [K] \tag{3.1}$$

to the BS, where $\mathsf{x}_t^k \triangleq [x_{1,t}^k, \cdots, x_{d,t}^k]^T \in \mathbb{R}^{d\times 1}$ denotes the d-dimensional model differential of client k at the t-th communication round. To simplify the notation, the index t is omitted by using $x_{k,i}$ instead of $x_{i,t}^k$ barring any confusion. Throughout this chapter, all active clients are assumed to be synchronized. This can be achieved in practice by having the BS send a beacon signal to initialize uplink transmissions. For more details on synchronization, please refer to [92]. Therefore, each client can transmit every element of the differential model $\{x_{k,i}\}_{i=1}^d$ via d shared time slots[2]. For a given element $x_{k,i}$, the received signal at the BS is $\mathbf{y}_i^{\mathsf{UL}} = \sqrt{P_{\mathsf{Client}}} \sum_{k\in[K]} \mathbf{h}_k x_{k,i} + \mathbf{n}_i, \forall i = 1, \cdots, d$, where P_{Client} is the maximum transmit power of each client, and $\mathbf{n}_i \in \mathbb{C}^{M\times 1}$ represents the uplink noise. Denoting $\mathbf{H} \triangleq [\mathbf{h}_1, \cdots, \mathbf{h}_K] \in \mathbb{C}^{M\times K}$ as the channel vectors from all K clients and $\mathbf{x}_i \triangleq [x_{1,i}, \cdots, x_{K,i}]^T \in \mathbb{R}^{K\times 1}, \forall i = 1, \cdots, d$ as the i-th dimension model differential from all K clients at the t-th learning round, the received signal[3] can be written as

$$\mathbf{y}_i^{\mathsf{UL}} = \sqrt{P_{\mathsf{Client}}} \mathbf{H} \mathbf{x}_i + \mathbf{n}_i. \tag{3.2}$$

It is easy to see that (3.2) is a standard MIMO communication model and traditional MIMO estimators can be adopted to estimate $\hat{\mathbf{x}}_i = [\hat{x}_{1,i}, \cdots, \hat{x}_{K,i}]^T$. However, as discussed before, decoding $\{x_{k,i}\}_{i=1}^d$ individually and obtaining the aggregated parameter $\tilde{x}_i \triangleq \sum_{k\in[K]} \hat{x}_{k,i}$ by a summation is inefficient. After the BS decoding all aggregated parameter

[2] In general, differential model parameters can be transmitted over any d shared orthogonal communication resources (e.g., time or frequency). For simplicity, we use d time slots here.

[3] For simplicity, real signals $\{x_{k,i}\}_{i=1}^d$ are assume to be transmitted. It can be easily extended to complex signals by stacking two real model parameters into a complex signal, so that the full degree of freedom (d.o.f.) is utilized.

$\tilde{\mathbf{x}}_t \triangleq [\tilde{x}_1, \cdots, \tilde{x}_d]^T$ in d slots, it can compute the new global model as

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \frac{1}{K}\tilde{\mathbf{x}}_t. \tag{3.3}$$

In the downlink, after the computation of the global model $\mathbf{w}_{t+1} = [w_{1,t+1}, \cdots, w_{d,t+1}]^T$, the BS broadcasts the global model to all clients via a precoder $\mathbf{f} \in \mathbb{C}^{M\times 1}$, and the received signal at client k is given by

$$y_i^{\mathsf{DL}} = \sqrt{P_{\mathsf{BS}}}\mathbf{h}_{k,t+1}^H \mathbf{f} w_{i,t+1} + z_i^k, \ \ \forall i = 1, \cdots, d, \tag{3.4}$$

where P_{BS} is the maximum transmit power of the BS and z_i^k denotes the downlink noise. We note that channel $\mathbf{h}_{k,t+1}^H \in \mathbb{C}^{1\times M}$ denotes the downlink vector channel that is reciprocal of the uplink channel in round $t+1$. Each client then computes an estimated global model and uses it as a new initial point for the next learning round after all d elements are received via (3.4). Traditionally, the precoder design of $\mathbf{f}$ belongs to broadcasting common messages (see [93] and the references therein). However, existing methods become impractical due to the difficulty in obtaining full CSI in massive MIMO systems, which motivates us to design $\mathbf{f}$ with only partial CSI. For mathematical simplicity, we assume a normalized symbol power[4], i.e., $\mathbb{E}\left\|x_{k,i}\right\|^2 = 1$ and $\mathbb{E}\left\|w_{i,t+1}\right\|^2 = 1$; normalized Rayleigh block fading channels[5] $\mathbf{h}_k \sim \mathcal{CN}(0, \frac{1}{M}\mathbf{I})$ in d slots; and i.i.d. Gaussian noise $\mathbf{n}_i \sim \mathcal{CN}(0, \frac{\sigma_{\mathsf{UL}}^2}{M}\mathbf{I})$ and $z_i^k \sim \mathcal{CN}(0, \sigma_{\mathsf{DL}}^2)$. We define the signal-to-noise ratio (SNR) as $\mathsf{SNR}_{\mathsf{UL}} \triangleq P_{\mathsf{Client}}/\sigma_{\mathsf{UL}}^2$ for uplink communications and $\mathsf{SNR}_{\mathsf{DL}} \triangleq P_{\mathsf{BS}}/\sigma_{\mathsf{DL}}^2$ for downlink communications, and without loss of generality (w.l.o.g.) we set $P_{\mathsf{Client}} = 1$ and $P_{\mathsf{BS}} = 1$.

3.3 Random Orthogonalization

In this section, the key ideas of random orthogonalization is presented. With this principle, the global model can be directly obtained at the BS via a simple operation in the uplink communications, and the global model can be broadcast to clients efficiently in the downlink communications. By exploring favorable propagation and channel hardening in massive MIMO, our proposed methods only require *partial* CSI, which significantly reduces the channel estimation overhead.

3.3.1 Uplink Communication Design

The designed framework contains the following three main steps in the uplink communications.

[4] The parameter normalization and de-normalization procedure in wireless FL follows the same as that in the Appendix of [17].

[5] The large-scale pathloss and shadowing effect is assumed to be taken care of by, e.g., open loop power control [94].

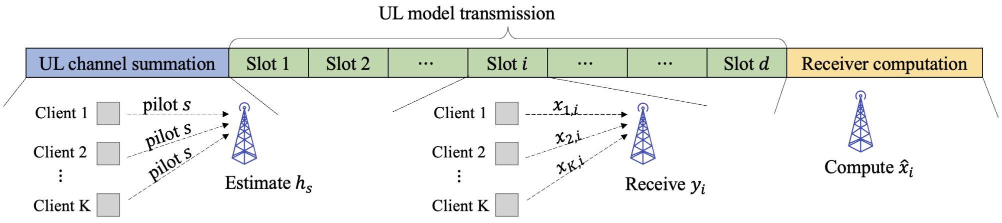


Figure 3.1: An illustration of the proposed uplink FL design with massive MIMO.

(U1) Uplink channel summation. The BS first schedules all clients participating in the current learning round to transmit a *common* pilot signal s synchronously. The received signal at the BS is

$$\mathbf{y}_s = \sum_{k\in[K]} \mathbf{h}_k s + \mathbf{n}_s, \tag{3.5}$$

and the BS can estimate the *summation* of channel vectors $\mathbf{h}_s \triangleq \sum_{k\in[K]} \mathbf{h}_k$ from the received signal $\mathbf{y}_s$. Given the pilot s, the estimates can be obtained via a maximum likelihood estimator $\arg\min_{\mathbf{h}_s} \|\mathbf{y}_s - \mathbf{h}_s s\|^2$ [95]. We can also adopt multiple pilots to improve the accuracy of channel estimation. Note that the complexity of this sum channel estimation does not scale with K. For the purpose of illustrating our key ideas, we assume perfect summation channel estimation at the BS for now. The channel estimation error of $\mathbf{h}_s$ will affect the effective SNR of the decoded model, and this impact will be evaluated in numerical experiments. Moreover, when the pilot SNR is sufficiently high, one can directly scale the received signal $\mathbf{y}_s$ (by $1/s$) to obtain the estimated summation channel.

(U2) Uplink model transmission. All selected clients transmit model differential parameters $\{x_{k,i}\}_{i=1}^d$ to the BS in d shared time slots. The received signal for each differential model element is $\mathbf{y}_i = \sum_{k\in[K]} \mathbf{h}_k x_{k,i} + \mathbf{n}_i, \forall i = 1, \cdots, d$.

(U3) Receiver computation. The BS estimates each aggregated model element via the following simple *linear projection* operation:

$$\begin{aligned}\tilde{x}_i = \mathbf{h}_s^H \mathbf{y}_i &= \sum_{k\in[K]} \mathbf{h}_k^H \sum_{k\in[K]} \mathbf{h}_k x_{k,i} + \sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i \\ &\overset{(a)}{=} \underbrace{\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{h}_k x_{k,i}}_{\text{Signal}} + \underbrace{\sum_{k\in[K]} \sum_{j\in[K],j\neq k} \mathbf{h}_k^H \mathbf{h}_j x_{j,i}}_{\text{Interference}} \\ &+ \underbrace{\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i}_{\text{Noise}} \overset{(b)}{\approx} \sum_{k\in[K]} x_{k,i}, \quad \forall i = 1, \cdots, d.\end{aligned} \tag{3.6}$$

The above three-step uplink communication procedure is illustrated in Fig. 3.1. Based on Eqn. (3.6), the BS then computes the global model via Eqn. (3.3) and begins the downlink global model broadcast.

As shown in (a) of Eqn. (3.6), inner product $\mathbf{h}_s^H \mathbf{y}_i$ can be viewed as the combination of three parts: signal, interference, and noise. We next show that, taking advantage of two fundamental properties of massive MIMO, the error-free approximation (b) in (3.6) is asymptotically accurate (as the number of BS antennas M goes to infinity).

Channel hardening. Since each element of $\mathbf{h}_k$ is i.i.d. complex Gaussian, by the law of large numbers, massive MIMO enjoys channel hardening [96]: $\mathbf{h}_k^H \mathbf{h}_k \to 1$, as $M \to \infty$. In practical systems, when M is large but finite, for the signal part of (3.6), we have

$$\mathbb{E}_{\mathbf{h}} \left[\sum_{k \in [K]} \mathbf{h}_k^H \mathbf{h}_k x_{k,i} \right] = \sum_{k \in [K]} x_{k,i}, \tag{3.7}$$

and

$$\mathbb{V}\mathrm{ar}_{\mathbf{h}} \left[\sum_{k \in [K]} \mathbf{h}_k^H \mathbf{h}_k x_{k,i} \right] = \frac{\sum_{k \in [K]} x_{k,i}^2}{M}. \tag{3.8}$$

Favorable propagation. Since channels between different users are independent random vectors, massive MIMO also offers favorable propagation [96]: $\mathbf{h}_k^H \mathbf{h}_j \to 0$, as $M \to \infty$, $\forall k \neq j$. Similarly, when M is finite, we have

$$\mathbb{E}_{\mathbf{h}} \left[\sum_{k \in [K]} \sum_{j \in [K], j \neq k} \mathbf{h}_k^H \mathbf{h}_j x_{j,i} \right] = 0, \tag{3.9}$$

and

$$\mathbb{V}\mathrm{ar}_{\mathbf{h}} \left[\sum_{k \in [K]} \sum_{j \in [K], j \neq k} \mathbf{h}_k^H \mathbf{h}_j x_{j,i} \right] = \frac{(K-1) \sum_{k \in [K]} x_{k,i}^2}{M}. \tag{3.10}$$

Furthermore, the expectation of the noise part in (3.6) is zero. Therefore, $\tilde{x}_i$ in (3.6) is an unbiased estimate of the average model. For a given K, the variances of both signal and interference decrease in the order of $\mathcal{O}(1/M)$, which shows that *massive MIMO offers* **random orthogonality** *for analog aggregation over wireless channels*. In particular, the asymptotic element-wise orthogonality of channel vector ensures channel hardening, and the asymptotic vector-wise orthogonality among different wireless channel vectors provides favorable propagation. Both properties render the linear projection operation $\mathbf{h}_s^H \mathbf{y}_i$ an ideal fit for the server model aggregation in FL.

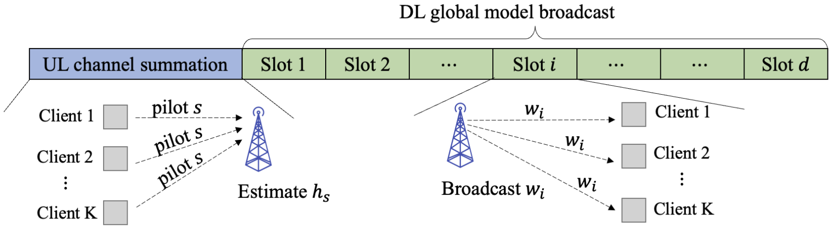


Figure 3.2: An illustration of the proposed downlink FL design with massive MIMO.

To gain some insight of random orthogonality, we approximate the average signal-to-interference-plus-noise-ratio (SINR) after the operation in (3.6) as

$$\begin{aligned} &\mathbb{E}[\text{SINR}_i] \approx \\ &\frac{\mathbb{E}_{\mathbf{h},x}\left\|\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_k x_{k,i}\right\|^2}{\mathbb{E}_{\mathbf{h},\mathbf{n},x}\left\|\sum_{k\in[K]}\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j x_{j,i}+\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{n}_i\right\|^2} \\ &=\frac{M}{K-1+1/\mathsf{SNR}}, \end{aligned} \tag{3.11}$$

which grows linearly with M for a fixed K. On the other hand, for a given number of antennas M, Eqn. (3.11) can be used to guide the choice of K in each communication round to satisfy an SINR requirement. More details on the scalability of clients via the convergence analysis will be provided in Section 3.5.2. Note that Eqn. (3.11) is an approximate expression for SINR but it sheds light into the relationship between K and M. This approximation, however, is not used in the convergence analysis of FL with random orthogonalization in Section 3.5.2.

The uplink random orthognalization design presented above is similar to that in [90], which also relies on orthogonality to directly compute the summation ML model at the server. This work, however, builds a more complete framework that has both uplink and downlink designs, for both massive MIMO and general MIMO. This will be elaborated in the following sections.

3.3.2 Downlink Communication Design

Inspired by the uplink communication design, the downlink design contains the following two steps.

(D1) Uplink channel summation. This step remains the same as **U1** in the uplink design. We similarly assume perfect sum channel estimation $\mathbf{h}_s = \sum_{k\in[K]}\mathbf{h}_k$ at the BS.

(D2) Downlink global model broadcast. The BS broadcasts global model $\{w_i\}$ to all users, using the estimated summation channel $\mathbf{h}_s$ as the precoder. Hence the received signal at the k-th user is

$$\begin{aligned} y_k = \mathbf{h}_k^H \mathbf{h}_s w_i + z_i^k \overset{(a)}{=} \underbrace{\mathbf{h}_k^H \mathbf{h}_k w_i}_{\text{Signal}} + \underbrace{\sum_{j \in [K], j \neq k} \mathbf{h}_k^H \mathbf{h}_j w_i}_{\text{Interference}} \\ + \underbrace{z_i^k}_{\text{Noise}} \overset{(b)}{\approx} w_i \;\; \forall i = 1, \cdots, d. \end{aligned} \tag{3.12}$$

The above two-step downlink communication procedure is illustrated in Fig. 3.2. Similar to the uplink case, the global model signal obtained at each client can also be regarded as the combination of three parts: signal, interference, and noise as shown in (3.12). Leveraging channel hardening and favorable propagation of massive MIMO channels as mentioned before, we have

$$\mathbb{E}_{\mathbf{h}} \left[\mathbf{h}_k^H \mathbf{h}_k w_i \right] = w_i \quad \text{and} \quad \mathbb{V}\text{ar}_{\mathbf{h}} \left[\mathbf{h}_k^H \mathbf{h}_k w_i \right] = \frac{w_i^2}{M}, \tag{3.13}$$

for the signal part of (3.12). Besides, we have

$$\mathbb{E}_{\mathbf{h}} \left[\sum_{j \in [K], j \neq k} \mathbf{h}_k^H \mathbf{h}_j w_i \right] = 0 \tag{3.14}$$

and

$$\mathbb{V}\text{ar}_{\mathbf{h}} \left[\sum_{j \in [K], j \neq k} \mathbf{h}_k^H \mathbf{h}_j w_i \right] = \frac{(K-1) w_i^2}{M}, \tag{3.15}$$

for the interference part. The above derivation demonstrates that, similar to the uplink design, received signals obtained via (3.12) are unbiased estimates of global model parameters whose variances decrease in the order of $\mathcal{O}(1/M)$ with the increase of BS antennas. Next, two remarks about the proposed uplink and downlink communication designs of FL with random orthogonalization are in order.

Remark 6. *In uplink communications, unlike the analog aggregation method in [17], the proposed random orthogonalization does not require any individual CSIT. On the contrary, it only requires partial CSIR, i.e., the estimation of a summation channel $\mathbf{h}_s$, which is $1/K$ of the channel estimation overhead compared with the AirComp method in [18] or the traditional MIMO estimators. In downlink communications, the traditional precoder design for common message broadcast requires CSIT for each client. By using the summation channel $\mathbf{h}_s$ as the precoder for global model broadcast, only partial CSIT is needed. In a TDD system, the downlink summation channel $\mathbf{h}_s$ can be estimated at a low cost utilizing channel reciprocity as shown in Step D1. Therefore, the proposed method is attractive in wireless FL due to its mild requirement of partial CSI. Moreover, the server obtains global models directly after a series of* *simple linear projections, which improves the privacy and reduces the system latency as a result of the extremely low computational complexity of random orthogonalization. The same applies to the downlink phase.*

Remark 7. *Note that although we assume i.i.d. Rayleigh fading channels across different clients, the proposed random orthogonalization method is still valid for other channel models as long as channel hardening and favorable propagation are offered. In massive MIMO millimeter-wave (mmWave) communications, Rayleigh fading channels and light-of-sight (LOS) channels represent two extreme cases: rich scattering and no scattering. It is shown in [96] that both channel models offer asymptotic channel hardening and favorable propagation. In practice, it is likely to have a scenario which lies in between these two cases. Therefore, even under some channel correlations, the MIMO channels can still provide certain level of channel hardening and favorable propagation. Generally speaking, the random orthogonalization method is still valid in MIMO channels with low to moderate correlations, albeit with increased interference in the decoded models.*

3.4 Enhanced Random Orthogonalization Design

The proposed random orthogonalization principle in Section 3.3 requires channel hardening and favorable propagation. Although these two properties are quite common in massive MIMO systems, in the case that they are not available (e.g., the number of BS antennas is small), the design philosophy can still be applied by introducing a novel **channel echo mechanism**. In this section, an enhanced design to the methods in Section 3.3 is presented by taking advantage of channel echos.

Channel echo refers to that the receiver sends whatever it receives back to the original transmitter as the data payload. Similar techniques have been proposed before, such as "Echo-MIMO" and "two-way training" in [97] and [98]. However, they are developed for cooperative beamforming and optimal power allocation, respectively, and only focus on the single-user case. The main purpose of channel echo in our setting, however, is to "cancel" channel fading for each of the involved clients. The enhanced design for the uplink communications contains the following four main steps, which is demonstrated in Fig. 3.3.

(EU1) Uplink channel summation. The first step of the enhanced design follows the same as the random orthogonalization method (U1 and D1), so that the BS has the estimate sum channel vector $\mathbf{h}_s = \sum_{j\in[K]} \mathbf{h}_k$.

(EU2) Downlink channel echo. The BS sends the previously estimated $\mathbf{h}_s$ (after normalization to satisfy the power constraint) to all clients. For the k-th client, the received signal is $\mathbf{y}_k = \frac{\mathbf{h}_k^H \mathbf{h}_s}{\sqrt{K}} + \mathbf{n}_k$, by which client k can estimate $g_k = \mathbf{h}_k^H \mathbf{h}_s = \mathbf{h}_k^H \sum_{j\in[K]} \mathbf{h}_k = \|\mathbf{h}_k\|^2 + \sum_{j\in[K],j\neq k}^{K} \mathbf{h}_k^H \mathbf{h}_j$. Note again that we assume a perfect estimation of $\mathbf{h}_s$.

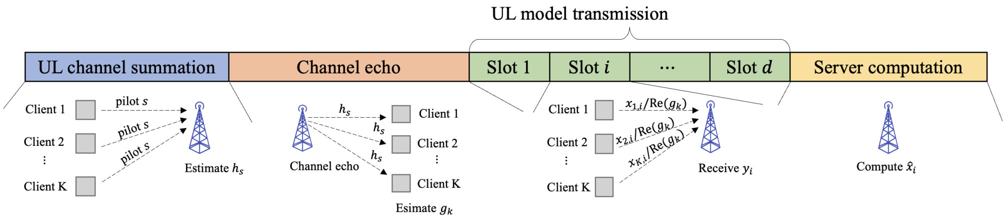


Figure 3.3: An illustration of the proposed enhanced uplink FL design with massive MIMO.

An additional error term will appear in the estimation of g_k when the summation channel estimation is imperfect, which will be discussed later.

(EU3) Uplink model transmission. All involved clients transmit local parameter $\{x_{k,i}/\mathrm{Re}(g_k)\}_{k\in[K]}$ to the BS synchronously in d shared time slots: $\mathbf{y}_i = \sum_{k\in[K]}^{K} \mathbf{h}_k \frac{x_{k,i}}{\mathrm{Re}(g_k)} + \mathbf{n}_i, \ \ \forall i = 1, \cdots, d$.

(EU4) Server computation. The BS obtains $\sum_{k\in[K]} x_{k,i}$ via the following operation:

$$\begin{aligned}
\tilde{x}_i = \mathrm{Re}(\mathbf{y}_i^H \mathbf{h}_s) &= \mathrm{Re}\left[\sum_{k\in[K]} \mathbf{h}_k^H \frac{x_{k,i}}{\mathrm{Re}(g_k)} \sum_{j\in[K]}^{K} \mathbf{h}_j + \mathbf{n}_i^H \sum_{j\in[K]} \mathbf{h}_j\right] \\
&= \sum_{k\in[K]} \frac{x_{k,i}}{\mathrm{Re}(g_k)} \mathrm{Re}\left[\mathbf{h}_k^H \sum_{j\in[K]} \mathbf{h}_j\right] + \mathrm{Re}\left[\mathbf{n}_i^H \sum_{j\in[K]} \mathbf{h}_j\right] \\
&= \sum_{k\in[K]} x_{k,i} + \mathrm{Re}\left[\sum_{j\in[K]} \mathbf{h}_j^H \mathbf{n}_i\right].
\end{aligned} \tag{3.16}$$

Similarly, as shown in Fig. 3.4, the enhanced design for the downlink communication contains the following four main steps.

(ED1-2) Uplink channel summation and downlink channel echo. The first two steps in the downlink design remain the same as Steps EU1 and EU2 in the uplink design, so that the BS can estimate channel vector summation $\mathbf{h}_s = \sum_{j\in[K]} \mathbf{h}_k$ and each client can estimate the parameter g_k.

(ED3) Downlink global model broadcast. The BS broadcasts global model $\{w_i\}$ to all clients using the estimated sum channel $\frac{\mathbf{h}_s}{\sqrt{K}}$ as the precoder. The received signal at the k-th client is $y_k = \mathbf{h}_k^H \frac{\mathbf{h}_s}{\sqrt{K}} w_i + \mathbf{n}_i = \frac{1}{\sqrt{K}} g_k w_i + z_i^k, \forall i = 1, \cdots, d$.

(ED4) Model parameter computation. Each user obtains the global model $\{w_i\}$ via the following calculation:

$$\mathrm{Re}\left[\frac{\sqrt{K} y_k}{g_k}\right] = w_i + \mathrm{Re}\left(\frac{\sqrt{K} z_i^k}{g_k}\right) \ \ \forall i = 1, \cdots, d. \tag{3.17}$$

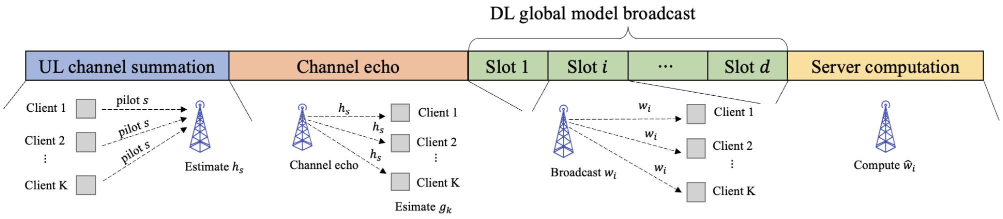


Figure 3.4: An illustration of the proposed enhanced downlink FL design with massive MIMO.

Note that the estimations of the aggregated signal and the global model in (3.16) and (3.17) are both *unbiased*, since $\mathbf{n}_j$, $\mathbf{h}_j$ and z_i^k are independent random variables with zero mean, and $\mathbb{E}[g_k] \neq 0$. Compared with the random orthogonalization method that offers *asymptotic* interference-free global model estimation, the received FL parameters obtained by the enhanced method are *completely interference-free* at both the server and the clients, as shown in (3.16) and (3.17). The extra channel echo steps (Step EU2 in uplink and Step ED1 in downlink) allow clients to obtain *partial* CSI g_k, so that they can pre-cancel and post-cancel channel interference among different user channels in uplink and downlink communications, respectively. Therefore, **this enhancement is valid even if channel hardening and favorable propagation are not present in wireless channels**, at a low cost of using one extra slot for the channel echo operation, and preserves all the other advantages of random orthogonalization.

Remark 8. *So far, both random orthogonalization and enhanced methods assume a perfect estimation of* $\mathbf{h}_s$. *In practical systems, to improve the accuracy of the estimate* $\hat{\mathbf{h}}_s$*, BS can use multiple pilots or multiple time slots for improved channel estimation, but summation channel estimation error will inevitably exist. In the following, taking uplink random orthogonalization as an example, the impact of imperfect summation channel estimation is evaluated. Denote the imperfect summation channel as* $\hat{\mathbf{h}}_s = \mathbf{h}_s + \boldsymbol{\epsilon}$*, where* $\boldsymbol{\epsilon} \sim \mathcal{CN}(0, \frac{\sigma_\epsilon^2}{M}\mathbf{I}_M)$ *is the summation channel estimation error that is modeled as a Gaussian random vector with i.i.d. elements. The decoded signal in* (3.6) *becomes*

$$\begin{aligned}
\tilde{x}_i = \hat{\mathbf{h}}_s^H \mathbf{y}_i = & \left[\sum_{k\in[K]} \mathbf{h}_k^H + \boldsymbol{\epsilon}^H\right] \sum_{k\in[K]} \mathbf{h}_k x_{k,i} \\
& + \sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i + \boldsymbol{\epsilon}^H \mathbf{n}_i = \underbrace{\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{h}_k x_{k,i}}_{\textit{Signal}} \\
& + \underbrace{\sum_{k\in[K]} \sum_{j\in[K], j\neq k} \mathbf{h}_k^H \mathbf{h}_j x_{j,i} + \sum_{k\in[K]} \boldsymbol{\epsilon}^H \mathbf{h}_k x_{k,i}}_{\textit{Effective interference}} \\
& + \underbrace{\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i + \boldsymbol{\epsilon}^H \mathbf{n}_i}_{\textit{Effective noise}} \approx \sum_{k\in[K]} x_{k,i}, \quad \forall i = 1, \cdots, d.
\end{aligned} \tag{3.18}$$

Note that estimation in (3.18) is still unbiased, since $\mathbb{E}\left[\sum_{k\in[K]}\boldsymbol{\epsilon}^H\mathbf{h}_k x_{k,i}\right] = 0$ *and* $\mathbb{E}\left[\boldsymbol{\epsilon}^H\mathbf{n}_i\right] = 0$*. Moreover, since*

$$\mathbb{V}ar_{\mathbf{h},\boldsymbol{\epsilon}}\left[\sum_{k\in[K]}\boldsymbol{\epsilon}^H\mathbf{h}_k x_{j,i}\right] = \frac{K\sigma_{\boldsymbol{\epsilon}}^2\sum_{k\in[K]}x_{k,i}^2}{M}, \tag{3.19}$$

and

$$\mathbb{V}ar\left[\boldsymbol{\epsilon}^H\mathbf{n}_i\right] = \frac{\sigma_{\mathsf{UL}}^2\sigma_{\boldsymbol{\epsilon}}^2}{M}. \tag{3.20}$$

it is equivalent to consider the presence of channel estimation error as a perfect estimation case with a larger effective interference and noise, which depend on the channel estimation quality $\sigma_{\boldsymbol{\epsilon}}^2$*. In addition, for the enhanced method, the estimation error of* $\mathbf{h}_s$ *itself will not affect the performance, since the imperfect estimated summation channel will cancel out in Step EU/ED*4*. Only the imperfect estimation of* g_k *will influence the results. We provide more details on the robustness of the proposed schemes over imperfect* $\hat{\mathbf{h}}_s$ *and* g_k *in the experiment results.*

Remark 9. *In the enhanced uplink design, each client pre-cancels the channel fading effect so that the global model can be directly obtained at the BS after simple operations. Note that the analog aggregation method in [17] also uses "channel inversion" to pre-cancel channel fading. However, the proposed design outperforms the method in [17] because the latter requires full CSIT, which leads to a large channel estimation overhead even with channel reciprocity in a TDD system. On the contrary, the proposed method only requires partial CSI, which can be efficiently obtained via channel echos. Moreover, channel-inversion-based methods do not naturally extend to MIMO systems when the uplink channels become vectors, which makes the inversion operations at the transmitters nontrivial.*

3.5 Performance Analyses

The performance of the proposed methods are analyzed from two aspects. On the communication performance side, the CRLBs on the estimates of model parameters in both uplink and downlink phases are derived as the theoretical benchmarks. On the machine learning side, the convergence analysis of FL is presnetd when the proposed communication designs are applied.

3.5.1 Cramer-Rao Lower Bounds

In the uplink communication, recall that the received signal is $\mathbf{y}_i = \mathbf{H}\mathbf{x}_i + \mathbf{n}_i$. Denoting $\boldsymbol{\mu}_{\mathsf{UL}} = \mathbf{H}\mathbf{x}_i$, it is clear that $\mathbf{y}_i \sim \mathcal{CN}(\boldsymbol{\mu}_{\mathsf{UL}}, \frac{1}{\mathsf{SNR}}\mathbf{I})$. To leverage CRLBs to evaluate the parameter estimation, one need to derive the Fisher information of $\mathbf{x}_i$. Based on Example 3.9 in [95], we can write the Fisher information matrix (FIM) of the estimation of

$\mathbf{x}_i$ as:

$$\mathbf{F}_{\mathsf{UL}} = 2 \cdot \mathsf{SNR} \cdot \mathrm{Re}\left[\frac{\partial^H \boldsymbol{\mu}_{\mathsf{UL}}(\mathbf{x}_i)}{\partial \mathbf{x}_i}\frac{\partial \boldsymbol{\mu}_{\mathsf{UL}}(\mathbf{x}_i)}{\partial \mathbf{x}_i}\right]. \tag{3.21}$$

After inserting $\frac{\partial \boldsymbol{\mu}_{\mathsf{UL}}(\mathbf{x}_i)}{\partial \mathbf{x}_i} = \mathbf{H}$ into the FIM, it is easy to have $\mathbf{F}_{\mathsf{UL}} = 2 \cdot \mathsf{SNR} \cdot \mathrm{Re}(\mathbf{H}^H \mathbf{H})$. Note that for the enhanced uplink design, one can absorb $\mathrm{Re}(g_k)$ into the effective channel as $\tilde{\mathbf{H}} \triangleq [\mathbf{h}_1/\mathrm{Re}(g_1), \cdots, \mathbf{h}_K/\mathrm{Re}(g_K)]$, and calculate FIM via $\mathbf{F}_{\mathsf{UL}} = 2 \cdot \mathsf{SNR} \cdot \mathrm{Re}(\tilde{\mathbf{H}}^H \tilde{\mathbf{H}})$.

In the downlink communication, since $y_k = \mathbf{h}_k^H \mathbf{h}_s w_i + \mathbf{n}_k$, by the definition of $\mu_{\mathsf{DL}} = \mathbf{h}_k^H \mathbf{h}_s w_i$, it is easy to see $y_k \sim \mathcal{CN}(\mu_{\mathsf{DL}}, \frac{1}{\mathsf{SNR}})$. The Fisher information of global model parameters is

$$F_{\mathsf{DL}} = 2 \cdot \mathsf{SNR} \cdot \mathrm{Re}\left[\frac{\partial^H \mu_{\mathsf{DL}}(w_i)}{\partial w_i}\frac{\partial \mu_{\mathsf{DL}}(w_i)}{\partial w_i}\right] = 2 \cdot \mathsf{SNR} \cdot \mathrm{Re}(\mathbf{h}_k^H \mathbf{h}_s \mathbf{h}_s^H \mathbf{h}_k). \tag{3.22}$$

The CRLBs of estimates are then given by the inverse of the Fisher information (matrix): $\mathbf{C}_{\hat{\mathbf{x}}_i} = \mathbf{F}_{\mathsf{UL}}^{-1}$ and $C_{\hat{w}_i} = 1/F_{\mathsf{DL}}$, respectively. CRLBs are the lower bounds on the variances of unbiased estimators, stating that the variance of any such estimator is at least as high as the inverse of the Fisher information (matrix). As shown that the proposed methods lead to unbiased estimations of the global model in both uplink and downlink communications, one can use the sum of all diagonal elements of $\mathbf{C}_{\hat{\mathbf{x}}}$ as the lower bound of the mean squared error (MSE) $\mathbb{E}\left\|\mathbf{x}_i - \hat{\mathbf{x}}_i\right\|^2$; and use $C_{\hat{w}_i}$ as the lower bound of MSE $\mathbb{E}\left\|w_i - \hat{w}_i\right\|^2$, to evaluate the performance of model estimation in both uplink and downlink communications. These bounds will be validated in the experiment results.

3.5.2 ML Model Convergence Analysis

This section delves into the machine learning model convergence performance of the proposed methods. Given that two distinct designs (basic and enhanced) have been introduced for uplink and downlink communications, this leads to four possible scenarios for convergence analysis. However, due to the similarity among these analyses, only one scenario is presented for brevity.

The focus here is on the convergence analysis of FL when random orthogonalization is employed for uplink communications, alongside the application of the enhanced design for downlink communications. It is important to note that, in contrast to uplink communications, the use of model differential is not feasible for downlink FL communications due to the selective participation of clients. To ensure FL convergence under these conditions, it is necessary to adhere to a guideline derived from prior research [99], which stipulates that the downlink transmit power should scale with the order of $\mathcal{O}(t^2)$. This prerequisite is vital for maintaining effective communication and ensuring the steady progress of FL in environments with partial client selection.

Theorem 4 (***Convergence for random orthogonalization in the uplink and enhanced method in the downlink***). *Consider a wireless FL task that applies random orthogonalization for the uplink communications and the enhanced method for the downlink communications. With Assumptions 1-4, for some $\gamma \geq 0$, if we set the learning rate as $\eta_t = \frac{2}{\mu(t+\gamma)}$ and downlink SNR scales as $\mathsf{SNR}_{\mathsf{DL}} \geq \frac{1-\mu\eta_t}{\eta_t^2}$ in round t, we have*

$$\mathbb{E}[f(\mathbf{w}_t)] - f^* \leq \frac{L}{2(t+\gamma)} \left[\frac{4B}{\mu^2} + (1+\gamma) \left\| \mathbf{w}_0 - \mathbf{w}^* \right\|^2 \right], \tag{3.23}$$

for any $t \geq 1$, where

$$\begin{aligned} B \triangleq & \sum_{k=1}^{N} \frac{H_k^2}{N^2} + 6L\Gamma + 8(E-1)^2 H^2 + \frac{N-K}{N-1} \frac{4}{K} E^2 H^2 \\ & + \frac{4}{K} \left(\frac{K}{M} + \frac{1}{\mathsf{SNR}_{\mathsf{UL}}} \right) E^2 H^2 + \frac{MK}{N^2(K+M)}. \end{aligned} \tag{3.24}$$

Proof. Proof of Theorem 4 is given in Appendix D.3. □

Theorem 4 shows that applying random orthogonalization in the uplink communications and enhanced method in the downlink communications preserves the $\mathcal{O}(1/T)$ convergence rate of vanilla SGD in FL tasks with perfect communications in both uplink and downlink phases. The factors that impact the convergence rate are captured entirely in the constant B, which come from multiple sources as explained below: $\frac{\sum_{k\in[N]} H_k^2}{N^2}$ comes from the variances of stochastic gradients; $6L\Gamma$ is introduced by the non-i.i.d. of local datasets; the choice of local computation steps and the fraction of partial client participation lead to $8(E-1)^2H^2$ and $\frac{N-K}{N-1}\frac{4}{K}E^2H^2$, respectively; and the interference and noise in uplink and downlink communications result in $\frac{4}{K}\left(\frac{K}{M} + \frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right)E^2H^2$ and $\left(d + \frac{dK}{M}\right)$, respectively. Note that the impact of the downlink noise, i.e., $\mathsf{SNR}_{\mathsf{DL}}$, is not explicit in B due to the requirement of $\mathsf{SNR}_{\mathsf{DL}} \geq \frac{1-\mu\eta_t}{\eta_t^2}$ to guarantee the convergence.

Remark 10. *Note that Theorem 4 considers random orthogonalization in the uplink and enhanced method in the downlink. When random orthogonalization is adopted in the downlink, the convergence bound in (3.23) will suffer from an additive constant term. This is because the interference cannot be effectively reduced when downlink power scales in the order of $\mathcal{O}(t^2)$, as required for direct model transmission [99]. This gap is also empirically observed in the experiments (see Section 3.6.2). However, we also note that this gap is inversely proportional to the number of antennas M. Hence, as M becomes large, it reduces to zero asymptotically*[6].

The analysis now turns to examining the relationship between the number of selected clients K and the number of antennas M at the BS, to gauge the scalability of multi-user MIMO for FL. This examination offers deeper insights for the practical design of such systems. For this purpose, a simplified scenario is considered where only random orthogonalization is configured for the uplink communications, with the assumption that downlink communications

[6] Due to the space limitation, the technical details for this remark are deferred to our technical report [100].

are error-free. This assumption is deemed reasonable in cases where the BS possesses significant transmit power. The assumptions on full client participation ($N = K$), one-step SGD at each device ($E = 1$), and i.i.d datasets across all clients ($\Gamma = 0$) are further made. For this special case, Corollary 5 is established as follows.

Corollary 5 (***Convergence for the simplified case*****).** *Consider a MIMO system that applies random orthogonalization for the uplink communications of FL with full client participation, one-step SGD at each device, and i.i.d datasets across all clients. Based on Assumptions 1-4 and choosing learning rate as $\eta_t = \frac{2}{\mu(t+\gamma)}$, $\forall t \in [T]$, the following inequality holds:*

$$\mathbb{E}[f(\mathbf{w}_t)] - f^* \leq \frac{L}{2(t+\gamma)} \left[\frac{4\tilde{B}}{\mu^2} + (1+\gamma) \|\mathbf{w}_0 - \mathbf{w}^*\|^2 \right] \tag{3.25}$$

for any $t \geq 1$, where

$$\tilde{B} \triangleq \left[1 + \frac{K}{M} + \frac{1}{\mathsf{SNR}} \right] \frac{H^2}{K}. \tag{3.26}$$

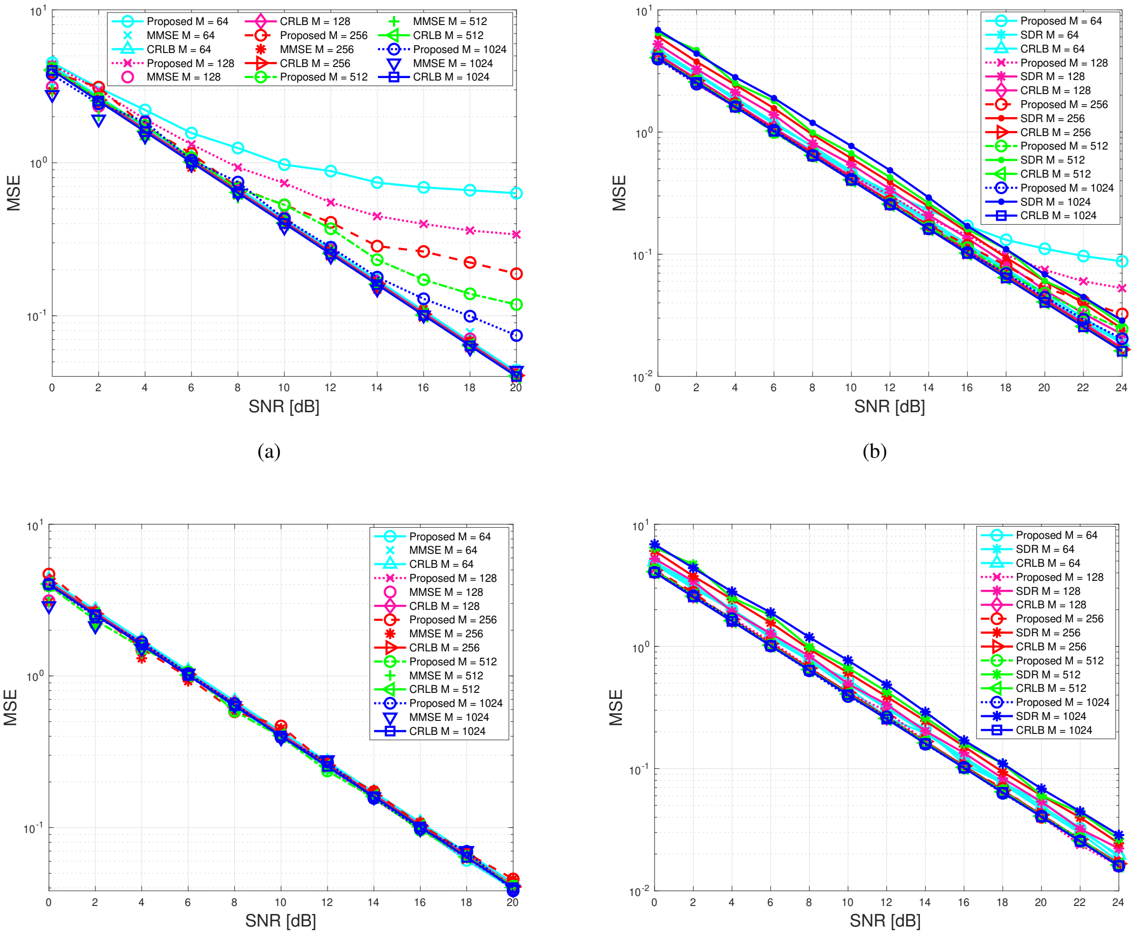
Proof. Corollary 5 comes naturally from Theorem 4 by setting $N = K, \Gamma = 0, E = 1$, omitting the $\left(d + \frac{dK}{M}\right)$ term due to the perfect downlink communications, and the fact that $\mathbb{E}\left\|\nabla f_k(\mathbf{w}) - \nabla \tilde{f}_k(\mathbf{w})\right\|^2 \leq \mathbb{E}\left\|\nabla \tilde{f}_k(\mathbf{w})\right\|^2 \leq H^2$. □

Corollary 5 shows that there are two main factors that impact the convergence rate of FL with MIMO: **variance reduction** and **channel interference and noise**. In particular, the definition of $\tilde{B}$ in (3.26), which appears in Corollary 5, captures the joint impact of both factors. The nature of distributed SGD suggests that, for a fixed mini-batch size at each client, involving K devices enjoys a $\frac{1}{K}$ variance reduction of stochastic gradient at each SGD iteration [101], which is captured by the $\frac{H^2}{K}$ term in (3.26). However, due to the existence of interference and noise, the convergence rate is determined by both factors, shown as $\frac{H^2}{K}$ and $\frac{(K/M+1/\mathsf{SNR})H^2}{K} \approx \frac{H^2}{M}$ terms in (3.26). This suggests that the desired variance reduction may be adversely impacted if channel interference and noise dominate the convergence bound. In particular, when $M \gg K$, we have $\frac{1}{K} \gg \frac{1}{M}$, and the system enjoys almost the same variance reduction as the interference-free and noise-free case. However, in the case of $K \gg M$, we have $\frac{(K/M+1/\mathsf{SNR})}{K} \approx \frac{1}{M} \gg \frac{1}{K}$, and $\frac{H^2}{M}$ dominates the convergence bound. In this case, it is unwise to blindly increase the number of clients, as it does not have the advantage of variance reduction.

Remark 11. *In massive MIMO, a BS is usually equipped with many (up to hundreds) antennas. Although there may be large number of users participating in FL, only a small number of them are simultaneously active [18]. Both factors indicate that $K \ll M$ often holds in typical massive MIMO systems. The analysis reveals that our proposed framework enjoys nearly the same interference-free and noise-free convergence rate with low communication and computation overhead in massive MIMO systems.*

3.6 Experiment Results

The performances of random orthogonalization and the enhanced method for uplink and downlink FL communications are evaluated through numerical experiments in this section. This evaluation includes a comparison of the proposed methods against classic MIMO estimators and precoders in terms of MSE from a communication performance perspective. Additionally, the computation time is compared to gauge the complexity of the various methods. The robustness of the proposed methods is also discussed, particularly in scenarios where the conditions of channel hardening and favorable propagation are not fully met, and where channel estimation is imperfect. The effectiveness of these methods is further validated through FL tasks using real-world datasets, providing a comprehensive assessment of their viability and efficiency in practical applications.



(a)

(b)

(c)

(d)

Figure 3.5: MSE of the received global ML model parameters versus SNR of random orthogonalization in uplink (a) and downlink (b) communications and of the enhanced method in uplink (c) and downlink (d) communications.

3.6.1 Communication Performance

Consider a massive MIMO BS with $M = 64, 128, 256, 512$, or 1024 antennas, supporting $K = 8$ active users engaged in an FL task. A Rayleigh fading channel model, i.e., $\mathbf{h}_k \sim \mathcal{CN}(0, \frac{1}{M}\mathbf{I})$, is assumed for each user. The system's performance is evaluated based on the MSE of the computed global model parameters in both uplink and downlink communications, providing a quantitative measure of effectiveness across various antenna configurations. All MSE results are obtained from 2000 Monte Carlo experiments. The CRLBs derived in Section 3.5.1 is used as the benchmark of the computed MSEs. More specifically, the benchmark corresponds to the mean CRLBs calculated via (3.21) and (3.22) using the channel realizations in the Monte Carlo simulation. In addition, the traditional MIMO MMSE estimator and the semidefinite relaxation based (SDR-based) precoder design method in [93] are chosen for performance comparisons of uplink and downlink communications, respectively.

Effectiveness. Fig. 3.5(a) and Fig. 3.5(b) compare the MSE performance of the random orthogonalization method in uplink and downlink communications with the traditional MIMO estimator/precoder as well as the CRLB under different system SNRs. As illustrated in these two plots, the proposed method performs nearly identically to the CRLB in low and moderate SNRs under different antenna configurations (see $\mathsf{SNR} \leq 12$ dB for uplink and $\mathsf{SNR} \leq 18$ dB for downlink, when $M = 128, 256$ and 1024). As the SNR increases, the dominant factor affecting system performances becomes the interference among different users. In the uplink communications, when $K \leq M$ and at high SNR, Eqn. (3.11) shows that for a given K and M, the proposed method has a fixed (approximate) $\mathsf{SIR} = \frac{K-1}{M}$ as $\mathsf{SNR} \to \infty$, which explains why the performance of the proposed scheme deteriorates compared with MMSE at high SNR. This phenomenon is more prominent when the number of antennas at the BS is relatively small ($M = 64$ and 128). However, this issue disappears naturally as the number of BS antennas increases. It can be seen in Fig. 3.5(a) that the performance gap between the proposed method and the CRLB reduces, from about 12 dB when $M = 64$ to about 2 dB when $M = 1024$ at $\mathsf{SNR} = 20$ dB, in uplink communications. Note that, although random orthogonalization produces higher MSEs than the MMSE estimator, the FL tasks have the same convergence rate under a constant SINR in uplink communications as indicated by the convergence analysis. This will be further validated in Section 3.6.2 by showing that random orthogonalization hardly slows the convergence of FL. Similar to uplink, random orthogonalization performs nearly identically to the CRLB in low and moderate SNRs in downlink, and only loses about $0.5 \sim 6$ dB under different antenna configurations at $\mathsf{SNR} = 24$ dB. Moreover, random orthogonalization outperforms the SDR-based method at almost all SNRs and antenna configurations. Due to its sub-optimality, the performance of SDR-based method deteriorates as the number of antennas increases. In particular, it has about 3 dB loss compared with the CRLB when $M = 1024$, which further highlights the strengths of our approach for large arrays in massive MIMO. It should be emphasized that our method only requires $1/K$ of the channel estimation overheard (partial CSI) compared with both

Table 3.1: Computation time comparison between the proposed methods and the MMSE/SDR Method

# antennas	Total CPU time (second)		Ratio	Total CPU time (second)		Ratio
(M)	RO-UL	MMSE	RO-UL/MMSE	Enhanced-UL	MMSE	Enhanced-UL/MMSE
256	0.0186	2.7141	0.68%	0.0203	2.9228	0.69%
512	0.0303	12.4155	0.24%	0.0469	16.3938	0.30%
1024	0.0448	82.3530	0.05%	0.0711	91.4117	0.07%
# antennas	Total CPU time (second)		Ratio	Total CPU time (second)		Ratio
(M)	RO-DL	SDR	RO-DL/SDR	Enhanced-DL	SDR	Enhanced-DL/SDR
256	0.0157	25.1492	0.062%	0.0163	28.8593	0.056%
512	0.0415	324.7349	0.012%	0.0592	492.9539	0.012%
1024	0.0571	4819.6221	0.0012%	0.0695	5925.9250	0.0011%

MMSE and the SDR-based method (full CSI), and this advantage is more pronounced when the BS is equipped with a larger number of antennas.

Similarly, Fig. 3.5(c) and Fig. 3.5(d) compare the MSE performance of the enhanced method in uplink and downlink communications with the MMSE estimator / SDR-based precoder. It is clear from both plots that the enhanced method achieves MSEs that are very close to the CRLBs. Furthermore, it performs nearly identically as the MMSE estimator in uplink, and outperforms the SDR-based method by about $0.5 - 3$ dB in downlink for different antenna configurations. Therefore, by introducing channel echos, the enhanced method achieves excellent performance while consuming relatively low additional resource. Unlike random orthogonalization, the enhanced method cancels out all the interference in the decoded signal. Therefore, it is more attractive for smaller antenna arrays when the MIMO channels are not sufficiently orthogonal, despite at an additional channel echo cost. Random orthogonalization and the enhancement hence supplement each other, and they jointly serve as an efficient framework for both uplink and downlink communications of FL under different array configurations. The machine learning model parameters we estimate in FL are *real* signals. This constraint leads to a biased estimator at very low SNR, which leads to MSE saturation. Therefore, the MMSE estimator achieves a lower MSE than the CRLB when SNR ≤ 4 dB. It is worth noting that we can obtain an unbiased estimator by using *complex* signals and keeping the imaginary part of the estimates. For more details on MSE saturation in low SNRs, please refer to Section VI-D in [102].

Efficiency. This part evaluates the low-latency advantage of the proposed methods, which originate from the low computational complexity. The complexity of both MMSE and SDR-based methods scale as $\tilde{O}(M^3)$, which is considerably higher compared with the $\tilde{O}(M)$ complexity of random orthogonalization and the enhanced method. To illustrate the benefit of low latency, the CPU time is reported as a reference for an intuitive demonstration. Table 3.1 compares the computational time of the proposed schemes with the MMSE estimator and the SDR-based precoder when $\mathsf{SNR} = 10$ dB in the uplink and downlink communications, respectively. The total CPU time is the *cumulative time* of each algorithm over 2000 Monte Carlo experiments. It can be seen that the time consumption of random orthogonalization and the enhanced method is much less than that of the MMSE estimator and the SDR-based precoder. Especially, when $M = 1024$, despite the 0.3 dB normalized MSE (NMSE) performance loss of random orthogonalization compared with the MMSE estimator in the uplink communications (as shown in Fig. 3.5(a)), the computation time of the former is only 0.05% of the latter. The proposed methods are even more computationally efficient for the downlink communications, as the total CPU time is less than 0.1% of the SDR-based method in all settings. All these results suggest that both random orthogonalization and its enhancement are attractive in massive MIMO systems, because they have promising MSE performances but require much less channel estimation overhead and achieve extremely lower system latency than the classic MIMO estimators and precoders.

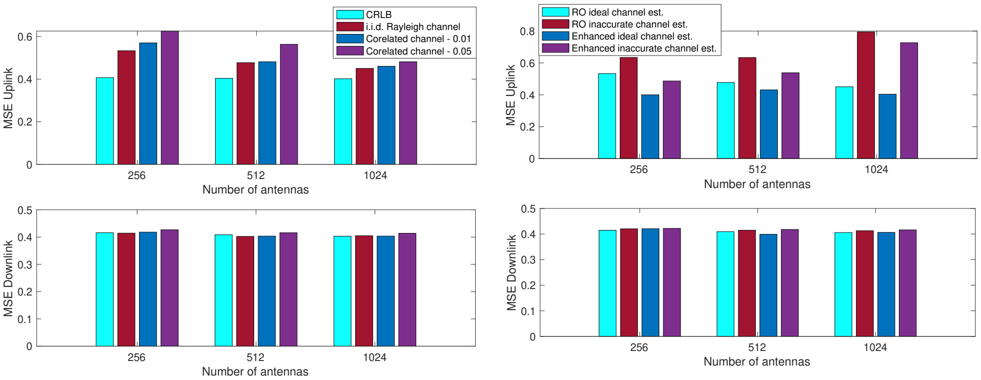
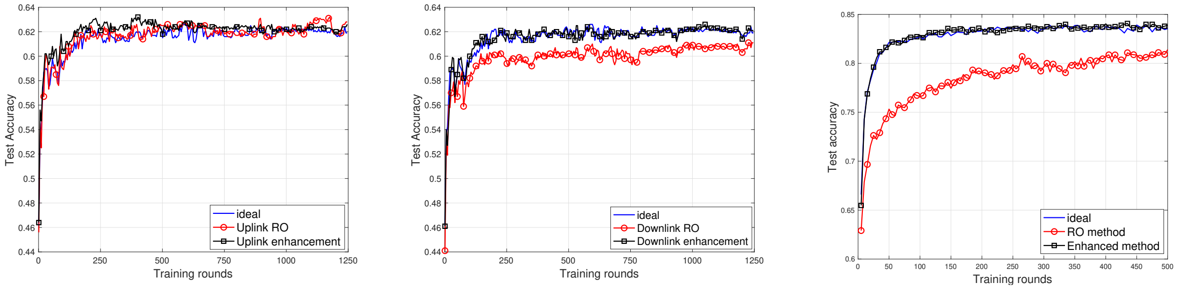


Figure 3.6: MSE comparison of the received global ML model parameters when channel hardening and favorable propagation are not fully offered (left) and channel estimation is imperfect (right).

Robustness. The subsection focuses on the robustness of the proposed methods, and evaluate the MSEs of the global model parameters obtained at $\mathsf{SNR} = 10$ dB through 2000 Monte Carlo experiments. Fig. 3.6 reports the achieved MSEs of the random orthogonalization method when the (approximate) channel hardening and favorable propagation are not strictly offered, i.e., the wireless channels are correlated. Two channel correlation models with covariance matrix elements equal to 1 on the diagonal and equal to 0.01 or 0.05 off the diagonal are considered, respectively. It is observed that when the off-diagonal elements are 0.01, random orthogonalization performs nearly identically as that in the ideal i.i.d. Rayleigh fading channel case. Even when the off-diagonal elements equal to 0.05, the achieved MSEs only increase by less than 1 dB in the worst case (when $M = 256$). The MSEs become closer to those of the i.i.d. Rayleigh channel cases when M increases, as larger antenna arrays offer higher orthogonality.

This subsection considers the performance when the estimation of the summation channel $\mathbf{h}_s$ (and g_k in the enhanced method) is imperfect. The right sub-figure of Fig. 3.6 compares the MSEs of both proposed methods when the channel estimation is obtained under $\mathsf{SNR} = 20$ dB. It reveals that the downlink communication is more robust than the uplink – the former achieves nearly identical MSEs as the ideal case even when the channel estimation is inaccurate. For the uplink, an imperfect channel estimation increases the MSEs by $1 \sim 3$ dB depending on the antenna configurations. However, we emphasize again that the FL tasks have the same convergence rate under a constant SINR in the uplink communications (thanks to the model differential transmission).



(a) MNIST uplink

(b) MNIST downlink

(c) CIFAR-10 uplink+downlink

Figure 3.7: Comparison of test accuracy. (a): a SVM FL task with an ideal uplink communication (interference and noise free), random orthogonalization, and the enhanced method; (b): a SVM FL task with an ideal downlink communication (interference and noise free), random orthogonalization, and the enhanced method; (c): a CIFAR classification FL task with ideal uplink and downlink communications (interference and noise free), random orthogonalization, and the enhanced method.

3.6.2 Learning Performance

To evaluate the learning performance, experiments of FL classification tasks are carried out using two widely adopted real-world datasets: MNIST and CIFAR-10, via a support vector machine (SVM) model and a convolution neural network (CNN) model, respectively. In the MNIST-SVM experiment, the proposed uplink and downlink design are evaluated separately, to identify their individual influence on the learning performance. Then, the CIFAR-CNN experiment jointly considers the uplink and downlink designs to evaluate the overall system performance of the proposed framework.

MNIST-SVM. The MNIST dataset collates small square 28×28 pixel gray-scale images of handwritten single digits between 0 and 9[103]. We implement a SVM to classify even and odd numbers in the MNIST dataset, with $d = 784$. Total clients are set as $N = 20$, the size of each local dataset is 500, the size of the test set is 2000, and $E = 1$. The local dataset can be regarded as non-i.i.d. since we only allocate data of one label to each client. We consider a massive MIMO cell with $M = 256$ antennas at the BS and $K = 8$ (out of 20) randomly selected clients are involved in each learning round. The channels between each client and the BS are assumed to be i.i.d. Rayleigh fading.

Fig. 3.7(a) reports the test accuracy when the uplink adopts the proposed method, and the downlink is assumed to be noise-free. The uplink SNR is set as 10 dB. We can see that both random orthogonalization and the enhanced method behave almost identically as the ideal case where both uplink and downlink communications are perfect. Note that although the global model received at the BS has noise and interference components, the actual learning performances of the two methods do not deteriorate. Due to the model differential transmission in the uplink communications, the effective SINR of the received global model gradually increases as the model converges, despite the presence of channel interference and noise. Fig. 3.7(b) demonstrates the learning performance when the proposed designs are applied to the downlink communications. Since the model differential transmission is infeasible, the initial downlink SNR is set to be 0 dB and scales at a rate of $\mathcal{O}(t^2)$ as the learning progresses (see [99]). Notice that the learning performance of the enhanced method is almost identical to that of the ideal case, while there is about 2% test accuracy loss for random orthogonalization. Note that, for downlink communications, because the BS only applies the normalized summation channel as the precoder, large-scale fading will result in different received SNRs for different clients. In this case, the above downlink SNR can be considered as the "worst" SNR among the involved clients (as the BS power control will need to target the worst-case user). Therefore, the reported result can be viewed as a lower bound of the actual performance.

CIFAR-CNN. The CIFAR-10 dataset consists of 32×32 color images in 10 classes, and we train a CNN model for the classification task. The CNN model consists of two 5×5 convolution layers (both with 64 channels), two fully connected layers (384 and 192 units respectively) with the ReLU activation, and a final output layer with the softmax activation. The two convolution layers are both followed by 2×2 max pooling and a local response norm layer. In the FL tasks, we set $N = K = 10$, and the size of each local dataset is 1000, with mini-batch size 50 and $E = 5$. The initial learning rate is $\eta = 0.15$ and decays every 10 rounds with rate 0.99. We consider a massive MIMO cell with $M = 1024$ antennas at the BS and the channels between each client and the BS are assumed to be i.i.d. Rayleigh fading. The uplink SNR is set as 10 dB and the initial downlink SNR is set as 0 dB, and scales at the rate of $\mathcal{O}(t^2)$.

Fig. 3.7(c) illustrates the training loss and test accuracy versus the learning rounds when *both* the uplink and downlink communications adopt the random orthogonalization method or the enhanced method, respectively. It is observed that the enhanced method achieves similar training loss and test accuracy as the ideal case. Due to the constant interference in the downlink communications, random orthogonalization incurs a test accuracy loss of about 3%.

To summarize, experiments on both datasets demonstrate that random orthogonalization suffers a slight performance degradation over the ideal case when it is applied to the downlink communications. As stated in Remark 10, unlike the enhanced method that cancels all interference in the received global model, the interference is constant in the global model obtained via random orthogonalization despite the increased SNR. Note that this gap can be reduced by increasing M. Therefore, downlink random orthogonalization is more attractive in systems with large number of antennas or severely limited resources.

3.7 Summary

Leveraging the distinctive properties of channel hardening and favorable propagation inherent to massive MIMO systems, a novel uplink communication method named *random orthogonalization* has been introduced. This approach notably diminishes the channel estimation overhead and facilitates natural over-the-air model aggregation without necessitating channel state information at the transmitter. This principle has been adapted to the downlink communication phase, resulting in a straightforward yet highly effective model broadcast technique for FL. Furthermore, to accommodate scenarios not strictly adhering to massive MIMO conditions, an enhanced random orthogonalization design employing channel echoes has been proposed. Comprehensive theoretical performance analyses have been conducted, encompassing both communication (CRLB) and machine learning (model convergence rate) aspects. Theoretical findings indicate that random orthogonalization can asymptotically achieve the same convergence rate as traditional FL under perfect communication conditions. These results have been corroborated through numerical experiments.

Chapter 4

Differentially Private Wireless Federated Learning Using Orthogonal Sequences

The exploration of physical layer design in MIMO systems has ventured into strategies that avoid the need for CSIT for AirComp. However, the application of such strategies to simpler SISO systems remains unclear.

Furthermore, while FL ostensibly enhances client privacy by localizing data training, the potential for privacy breaches persists through the analysis of ML model parameters shared by clients. To mitigate this privacy concern, introducing (artificial) noise to ML model parameters during the upload phase of FL has been proposed, with its privacy properties quantifiable through DP. Intriguingly, AirComp inherently offers the potential for DP assurance at no additional cost, owing to the natural noise present in wireless channels. It is suggested that varying levels of DP can be achieved by manipulating the SNR, thereby adjusting the effective channel noise at the receiver. Nonetheless, the literature on AirComp seldom provides a mathematically rigorous characterization of attainable privacy levels. A comprehensive review of related literature is presented in Section 4.1. It is also noted that current discussions predominantly cover item-level DP in wireless FL, whereas client-level DP (or user-level DP) presents a novel and crucial metric for FL that has not been adequately addressed.

To simultaneously remove the CSIT requirement of AirComp and address the privacy challenge, I propose `FLORAS` – Federated Learning using ORthogonAl Sequences, a novel uplink wireless physical layer design for FL by leveraging the properties of *orthogonal sequences*. On the communication design, `FLORAS` preserves all the advantages of AirComp while removing the CSIT requirement. From the perspective of privacy, `FLORAS` achieves desired DP guarantees (both item-level and client-level) by adjusting the number of used orthogonal sequences, making it much simpler and providing more flexibility to trade off privacy and utility.

The main contributions of this paper are summarized as follows:

- FLORAS is proposed for uplink communications in SISO wireless FL systems. FLORAS enjoys all the advantages of AirComp, yet without the CSIT requirement. In particular, orthogonal sequences allow the base station (BS) to obtain individual CSIR via a single pilot, by which global ML model parameters can be estimated via simple linear projections. Therefore, FLORAS significantly reduces the channel estimation overhead. Different from the channel inversion power control, FLORAS allows the transmit power to be independent of the channel realizations, which avoids increasing the dynamic range of the transmit signal and improves the power efficiency.
- By adjusting the number of orthogonal sequences in the system configuration, the novel signal processing technique in FLORAS produces Cauchy effective noise to the decoded global model, which empowers flexible item-level and client-level DP guarantees. Moreover, a new FL convergence bound based on the truncated Cauchy noise is derived, which allows us to characterize the tradeoff between the model convergence rate and the achievable DP levels.
- Extensive experiments are conducted based on real-world datasets to evaluate the performance of FLORAS. Numerical results demonstrate the performance advantages of FLORAS compared with the channel inversion method and validate our theoretical analysis by achieving tradeoffs between model convergence and DP.

The remainder of this chapter is organized as follows. Literature review is presented in Section 4.1. Section 4.2 introduces the FL pipeline and the uplink communication model. The proposed FLORAS design is detailed in Section 4.3. The DP guarantee and convergence analysis of FLORAS are presented in Section 4.4 and Section 4.5, respectively. Experimental results are reported in Section 4.6, followed by the summary in Section 4.7.

4.1 Related Work

AirComp for FL. The AirComp approach [17, 18, 19, 20] exploits the inherent signal superposition characteristics of a wireless multiple access channel to efficiently perform sum/average computations. This methodology can be considered as a special instance of computation over multiple access channels, as outlined by [70]. The approach has garnered significant attention due to its ability to minimize uplink communication costs, irrespective of the number of participating clients in FL. The exploration of client scheduling, along with various power and computation resource allocation techniques, has been investigated by [71, 48, 72, 74, 75]. Several studies have provided convergence guarantees for AirComp under diverse practical constraints and types of heterogeneity [77, 78, 79, 80, 81]. Efforts also have been made to reduce the CSIT requirement of AirComp. [76] relaxes full CSIT by utilizing only the phase information of the channel. Notably, [90] and [9] present CSIT-free AirComp methods that leverage channel orthogonality, although their effectiveness is limited to massive MIMO systems.

Differential Privacy for FL. DP-SGD has been widely regarded as a standard approach to train a differentially private ML model [104]. Along with its variants[105, 106], recent years have also witnessed increased efforts on DP for distributed learning systems, including the clipping technique in [107, 108], the subsampling principle in [109], and random quantization in [110]. In the category of exploiting the channel noise for differentially private FL, [111] proposes an AirComp design to achieve DP by adjusting the effective noise; [112, 113] investigate adding noise and power allocation in non-orthogonal multiple access (NOMA) systems; [114] considers DP amplification via user sampling and wireless aggregation; [115] applies it to personalized FL. [116] jointly optimizes the latency and DP requirements of FL, and [117] discusses the tradeoff among privacy, utility, and communication.

4.2 System Model

In each learning round, since there are K active clients, the uplink communication between clients (mobile devices) and the parameter server (base station) can be modeled as over a fading multiple access channel. Consider a cell with a single-antenna BS and K single-antenna users involved in one round of the aforementioned FL task. The communication system leverages orthogonal sequences for uplink transmissions. Note that one of the most popular implementations of the orthogonal sequence-based system is code-division multiple access (CDMA). We assume a spreading sequence set $\mathcal{A} = \{\mathbf{a}_1, \cdots, \mathbf{a}_k, \cdots, \mathbf{a}_N\}$ containing N unique spreading sequences ($N \geq K$), where each spreading sequence is denoted as $\mathbf{a}_k = [a_{1,k}, \cdots, a_{L,k}]^T$ and L is the length of the spreading sequence. Each user is (randomly) assigned with a unique spreading sequence $\mathbf{a}_k$ from $\mathcal{A}$ as its signature.

We assume that the BS only has knowledge of the entire spreading sequence set $\mathcal{A}$, *without knowing the specific signature of each user*. We emphasize that this restriction is consistent with our goal of guaranteeing user privacy – BS cannot identify users or decode individual models based on their spreading sequences. We will discuss the details of the spreading sequence assignment mechanism in Section 4.4.4.

In the uplink communication, each client transmits the differential between the received global model and the computed new local model:

$$\mathbf{x}_t^k = \mathbf{w}_t - \mathbf{w}_{t+1}^k \in \mathbb{R}^d, \ \ \forall k = 1, \cdots, K.$$

The BS aims at estimating $\mathbf{x}_t \triangleq \sum_{k=1}^K \mathbf{x}_t^k$.

Before the transmission of $\mathbf{x}_t^k$, client k will apply a normalization technique. We denote

$$\mathrm{x}_t^k \triangleq [x_{1,t}^k, \cdots, x_{i,t}^k, \cdots, x_{d,t}^k]^T \in \mathbb{R}^d$$

as the normalized transmit signal. The following normalization technique, adopted in [17], ensures $\mathbb{E}[\mathsf{x}_t^k] = 0$ and $\left\|\mathsf{x}_t^k\right\|_2^2 \leq C^2$:

$$\mathsf{x}_t^k \triangleq \frac{C(\mathbf{x}_t^k - \mu_k)}{C_{\max,t}},$$

where μ_k is the sample mean of d elements in $\mathbf{x}_t^k$. Note that normalization parameters μ_k, C and $C_{\max,t} \triangleq \max\{\left\|\mathbf{x}_t^k - \mu_k\right\|_2, \forall k\}$ will be determined by the BS and clients via a separate control channel as suggested in [17]. The l_2-norm bound guaranteed in the normalization not only provides the sensitivity of x_t^k for the DP analysis in Section 4.4, but also satisfies the practical requirement that each client has a limited transmit power. We note that the similar technique has also been applied in [107].

To simplify the notation, we omit index t and use x_k^i instead of $x_{i,t}^k$ barring any confusion. We assume that each client transmits every element of the model differential $\{x_k^1, \cdots, x_k^d\}$ via d shared time slots. In addition, *block fading channel* is assumed[1], i.e., the fading channel between each client and the BS h_k remains unchanged in d time slots. We emphasize that we do not make any specific assumption on the fading distribution throughout this paper. In the i-th slot, each client transmits symbol x_k^i spread by its uniquely assigned orthogonal sequence $\mathbf{a}_k$. The BS received signal can be written as

$$\mathbf{y}_i = \sum_{k=1}^{K} \mathbf{a}_k h_k x_k^i + \mathbf{n}_i, \ \ \forall i = 1, \cdots, d,$$

where $\mathbf{n}_i$ is the additive white Gaussian noise (AWGN) with mean zero and variance σ^2/L per dimension. Note that since the model differential parameters are real signals, we only need to consider the real parts of channel coefficients and noise. Although one-dimensional (real) modulation cannot fully leverage the channel degrees of freedom, it is consistent with the fact that binary phase-shift keying (BPSK) is the most common modulation scheme in CDMA systems [21]. In addition, focusing on the real dimension makes the subsequent discussion easier and highlights our contribution better. Also, as we detail later in Remark 13, we can leverage full channel gain at an affordable cost of partial CSIT.

We note that spreading sequences are *orthonormal*, i.e.,

$$\mathbf{a}_i^T \mathbf{a}_i = 1, \ \ \forall i; \quad \text{and} \quad \mathbf{a}_i^T \mathbf{a}_j = 0, \ \ \forall i \neq j. \tag{4.1}$$

[1] The large-scale pathloss and shadowing effect is assumed to be taken care of by, e.g., open loop power control [94], which is a common practice in real-world systems.

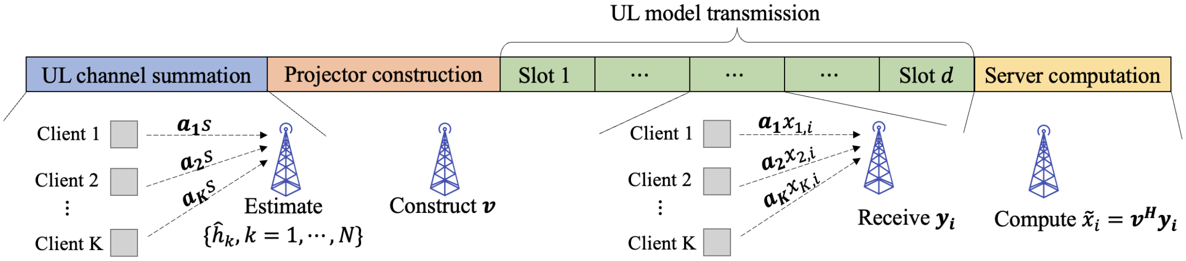


Figure 4.1: Illustration of the proposed uplink communication design of FLORAS.

At the BS, the receiver will decode the estimated aggregation parameter $\tilde{x}_i$, which is a noisy version of $x_i \triangleq \sum_{k=1}^K x_k^i$, and recover $\tilde{\mathrm{x}}_t \triangleq [\tilde{x}_1, \cdots, \tilde{x}_d]^T$ in d slots. After that, the BS can perform de-normalization:

$$\tilde{\mathbf{x}}_t \triangleq \frac{C_{\max,t}}{C} \tilde{\mathrm{x}}_t + \sum_{k=1}^K \mu_k$$

and compute the new global model as

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \frac{1}{K} \tilde{\mathbf{x}}_t. \tag{4.2}$$

Throughout the paper, we assume that all users are synchronized in frames, which can be achieved by the BS sending a beacon signal to initialize uplink transmissions [94].

4.3 FLORAS

We present the FLORAS design for uplink communications in wireless FL, and give some preliminary analysis.

4.3.1 Algorithm Design

FLORAS is a four-step protocol detailed as follows.

Step 1: Uplink channel estimation. The BS first schedules *all* participating users to transmit a common pilot s simultaneously. The received signal is

$$\mathbf{y}_s = \sum_{k=1}^K \mathbf{a}_k h_k s + \mathbf{n}_s.$$

The BS can utilize $\mathbf{y}_s$ and the *complete* set of orthogonal sequences $\mathcal{A} = \{\mathbf{a}_1, \cdots, \mathbf{a}_N\}$ to estimate the channel gain coefficients. For the K spreading sequences that are actually adopted by the user[2], we have

$$\hat{h}_k = \frac{\mathbf{a}_k^T \left[\sum_{i=1}^K \mathbf{a}_i h_i s + \mathbf{n}_s\right]}{s} = h_k + \frac{\mathbf{a}_k^T \mathbf{n}_s}{s}, \forall k = 1, \cdots, K.$$

For the $N - K$ spreading sequences that are not selected by any user, the BS obtains

$$\hat{h}_k = \frac{\mathbf{a}_k^T \left[\sum_{i=1}^K \mathbf{a}_i h_i s + \mathbf{n}_s\right]}{s} = \frac{\mathbf{a}_k^T \mathbf{n}_s}{s}, \forall k = K+1, \cdots, N.$$

We emphasize that the BS is not able to distinguish these two cases; all it has are N estimates $\{\hat{h}_1, \cdots, \hat{h}_N\}$.

Step 2: Projector construction. For simplicity, we assume $s = 1$ in the following discussion. After the channel estimation, the BS constructs the following vector based on *all* of the estimated channel coefficients:

$$\mathbf{v} = \sum_{k=1}^N \frac{1}{\hat{h}_k} \mathbf{a}_k = \sum_{k=1}^K \frac{\mathbf{a}_k}{h_k + \mathbf{a}_k^T \mathbf{n}_s} + \sum_{k=K+1}^N \frac{\mathbf{a}_k}{\mathbf{a}_k^T \mathbf{n}_s}.$$

We note that since the BS does not know which K of the total N spreading sequences are adopted by the users, it has to use all of $\{\hat{h}_1, \cdots, \hat{h}_N\}$ to construct the projector $\mathbf{h}_s$. This seemingly redundant design actually enables better privacy protection, which will be clear in Section 4.4.

Step 3: UL model transmission. All users transmit every element of the differentials via d shared time slots:

$$\mathbf{y}_i = \sum_{k=1}^K \mathbf{a}_k h_k x_k^i + \mathbf{n}_i, \quad \forall i = 1, \cdots, d.$$

Step 4: Sum model decoding. The BS applies the following linear projection to estimate each aggregated model differential $x_i, \forall i = 1, \cdots, d$:

$$\begin{aligned}
\tilde{x}_i = \mathbf{v}^T \mathbf{y}_i &= \sum_{k=1}^N \frac{1}{\hat{h}_k} \mathbf{a}_k \left[\sum_{k=1}^K \mathbf{a}_k h_k x_k^i + \mathbf{n}_i\right] \\
&= \left[\sum_{k=1}^K \frac{\mathbf{a}_k}{h_k + \mathbf{a}_k^T \mathbf{n}_s} + \sum_{k=K+1}^N \frac{\mathbf{a}_k}{\mathbf{a}_k^T \mathbf{n}_s}\right] \left[\sum_{k=1}^K \mathbf{a}_k h_k x_k^i + \mathbf{n}_i\right] \\
&= \sum_{k=1}^K \frac{h_k}{h_k + \mathbf{a}_k^T \mathbf{n}_s} x_k^i + \sum_{k=1}^K \frac{\mathbf{a}_k^T \mathbf{n}_i}{h_k + \mathbf{a}_k^T \mathbf{n}_s} + \sum_{k=K+1}^N \frac{\mathbf{a}_k^T \mathbf{n}_i}{\mathbf{a}_k^T \mathbf{n}_s}.
\end{aligned}$$

[2] Without loss of generality, we assume the first K spreading sequences from the set are selected. This assumption is made to ease the notation.

After obtaining $\{\tilde{x}_1, \cdots, \tilde{x}_d\}$, the BS can compute the new global model following (4.2) and start the next learning round. The above four-step procedure is illustrated in Fig. 4.1.

4.3.2 Preliminary Analysis

In the high signal-to-noise ratio (SNR) regime, where the channel fading effect dominates the noise, we have $\mathbb{E}[\|\mathbf{a}_k^T \mathbf{n}_s\|^2] \ll \mathbb{E}[\|h_k\|^2]$. Therefore, we can establish the following approximation for the estimated model in Step 4:

$$\tilde{x}_i \approx \sum_{k=1}^{K} x_k^i + \sum_{k=K+1}^{N} \frac{\mathbf{a}_k^T \mathbf{n}_i}{\mathbf{a}_k^T \mathbf{n}_s}, \ \ \forall i = 1, \cdots, d, \tag{4.3}$$

where $\sum_{k=K+1}^{N} \frac{\mathbf{a}_k^T \mathbf{n}_i}{\mathbf{a}_k^T \mathbf{n}_s}$ denotes the dominant noise of the received global model parameters[3]. The distribution of this post-processing noise is not straightforward, and we next present Lemma 1 to establish that the noise term is a *Cauchy random variable*.

Lemma 1. *Define $\gamma \triangleq N - K$. For IID Gaussian random vector $\mathbf{n}_i, \mathbf{n}_s \sim \mathcal{N}(0, \frac{\sigma^2}{L}\mathbf{I})$, random variable*

$$X \triangleq \sum_{k=K+1}^{N} \frac{\mathbf{a}_k^T \mathbf{n}_i}{\mathbf{a}_k^T \mathbf{n}_s} \sim \mathsf{Cauchy}(0, \gamma), \ \ \forall \mathbf{a}_k \in \mathcal{A},$$

with probability density function (PDF)

$$f_X(x) = \frac{1}{\pi} \frac{\gamma}{x^2 + \gamma^2}, \ \ x \in \mathbb{R}.$$

Proof. We first note that $\mathbf{a}_k^T \mathbf{n}_i$ and $\mathbf{a}_k^T \mathbf{n}_s$ are Gaussian random variables since they are linear combinations of IID Gaussian random variables. Let

$$Y \triangleq \left[\mathbf{a}_{K+1}^T \mathbf{n}_i, \mathbf{a}_{K+2}^T \mathbf{n}_i, \cdots, \mathbf{a}_N^T \mathbf{n}_i\right]^T$$

and

$$Z \triangleq \left[\mathbf{a}_{K+1}^T \mathbf{n}_s, \mathbf{a}_{K+2}^T \mathbf{n}_s, \cdots, \mathbf{a}_N^T \mathbf{n}_s\right]^T,$$

and it is straightforward to verify that Y and Z are IID Gaussian random vectors with distribution $\mathcal{N}(0, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a covariance matrix. According to [118], we have

$$\sum_{k=K+1}^{N} w_k \frac{Y_k}{Z_k} = \sum_{k=K+1}^{N} b_k \frac{\mathbf{a}_k^T \mathbf{n}_i}{\mathbf{a}_k^T \mathbf{n}_s} \sim \mathsf{Cauchy}(0, 1),$$

[3] Note that this approximation drops the minor noise term, which results in that the DP guarantee in the later discussion is a lower bound of the true DP level of `FLORAS`, i.e., we achieve better DP than that computed in this paper.

as long as w_k is independent of (Y, Z) and $\sum_{k=K+1}^{N} w_k = 1$. Letting $b_k = \frac{1}{\gamma}$ and using the fact that $kX \sim \mathsf{Cauchy}(0, |k|)$ if $X \sim \mathsf{Cauchy}(0, 1)$, we prove Lemma 1. □

Cauchy distribution is known as a "fat tail" distribution, as the tail of its PDF decreases proportionally with $1/x^2$. Lemma 1 suggests that, for a fixed K, a larger spreading sequence set will result in a heavier tail in the Cauchy noise. Therefore, we can adjust the size of the spreading sequence set to induce different additive Cauchy noise in (4.3). We will discuss the effect of Cauchy noise on DP and convergence in Sections 4.4 and 4.5, respectively.

A few remarks about `FLORAS` are now in order.

Remark 12 (Advantages over the channel inversion method)**.** *Compared with the widely studied channel inversion-based AirComp design,* `FLORAS` *does not require CSIT for uplink communications, which greatly reduces the communication overhead. This is especially attractive for Internet-of-Things (IoT) applications with massive devices. Moreover, in SISO systems, the maximum uplink transmit power of each user in the channel inversion-based methods is usually limited by the worst channel gain. As a result, the received SNR of the global model and the efficiency of power amplifiers (PAs) will significantly decrease, if one of the client channels experiences deep fading [21]. Thanks to the orthogonality of spreading sequences,* `FLORAS` *allows the transmit power to be independent of small-scale fading channel realizations, and thus avoids increasing the dynamic range of the transmit signal, which improves the power efficiency of PAs.*

Remark 13 (Leverage full channel degrees of freedom)**.** *As mentioned before, real modulation cannot exploit full degrees of freedom of the complex channel. To address this limitation, we can borrow the idea of [76]. Specifically, the fading channel between each client and the BS can be written as*

$$h_k = \tilde{h}_k e^{j\phi_k}, \quad \forall k = 1, \cdots, K,$$

where $\tilde{h}_k \triangleq |h_k| \in \mathbb{R}_+$ and $\phi_k \triangleq \angle h_k \in [0, 2\pi)$ represent channel amplitude (gain) and phase, respectively. For a given x_k^i, client k transmits symbol $x_k^i e^{-j\phi_k}$, where $e^{-j\phi_k}$ is the phase correction term as suggested in [76]. Hence the received signal at the BS can be written as

$$\mathbf{y}_i = \sum_{k=1}^{K} \mathbf{a}_k h_k e^{-j\phi_k} x_k^i + \mathbf{n}_i = \sum_{k=1}^{K} \mathbf{a}_k \tilde{h}_k x_k^i + \mathbf{n}_i, \forall i = 1, \cdots, d.$$

By phase correction, the imaginary part of h_k is projected to the real domain and the full channel gain can be leveraged without sacrificing any other advantages of `FLORAS`*. We note that same as in [76], this approach requires each client k to have the channel phase information ϕ_k, which is weaker than the complete CSIT h_k but stronger than the standard* `FLORAS`*.*

Note that we consider the imperfect channel estimation previously in Step 2 *for uplink communication. For simplicity, we assume that each client has perfect partial CSIT (channel phases) obtained from downlink channel estimation here. One reason for this different consideration is that the transmit power at the BS is usually much larger than those at devices, which naturally ensures more accurate downlink channel estimation than uplink. Additionally, even if the phase correction term suffers from estimation error, as long as the error is within* $[-\pi/2, \pi/2)$*, the proposed design is still valid [76], only at a cost of some channel gain lost.*

Remark 14 (Extensions on NOMA systems)**.** *Another limitation of* `FLORAS` *is that the number of orthogonal spreading sequences is fixed for a given sequence length. To expand the size of set* $\mathcal{A}$*, the system would need to adopt longer spreading sequences, which consumes more bandwidth. We first note that although there may exist a large number of clients in an FL task, the number of actively participating clients in each learning round is usually relatively small (due to client selection), which implies the bandwidth cost will not be too significant for our design. Second, as an alternative, the system can adopt non-orthogonal spreading sequences to improve the scalability without the cost of extra bandwidth. Applying non-orthogonal spreading sequences is consistent with non-orthogonal multiple access (NOMA) systems, which is an emerging technology for massive machine-type communications (mMTC) applications. For non-orthogonal spreading sequences, the requirement in* (4.1) *becomes*

$$\mathbb{E}\left[\mathbf{a}_i^T \mathbf{a}_i\right] = 1, \;\; \forall i \;\; \textit{and} \;\; \mathbb{E}\left[\mathbf{a}_i^T \mathbf{a}_j\right] = 0, \;\; \forall i \neq j,$$

which can be achieved by random Gaussian vectors. Note that Lemma 1 still holds for non-orthogonal spreading sequences, since it allows an arbitrary covariance matrix of random Gaussian vectors Y *and* Z*. Non-orthogonal spreading sequences will introduce inter-symbol interference besides noise when decoding the global model parameters in Step* 4*. Therefore, the convergence bound developed in Section 4.5 can be regarded as a lower bound for NOMA systems.*

4.4 Differential Privacy Analysis

In this section, we analyze the DP level achieved by `FLORAS`. We begin by introducing the basic concepts of DP in FL, and then prove that `FLORAS` achieves different levels of DP via the adjustment of the size of spreading sequence set N and the number of involved clients K. Both item-level and client-level DP are analyzed. The DP guarantee not only considers multiple sources of randomness in wireless FL, including random mini-batch in SGD, the Cauchy noise, and the random client participation, but also reveals the influence of multiple learning rounds.

4.4.1 Preliminaries

We first introduce the concept of *neighboring datasets*. We say that two datasets $\mathcal{D}$ and $\mathcal{D}'$ are neighboring, written as $\mathcal{D} \sim \mathcal{D}'$, if they differ in at most one sample. Based on this concept, we state the standard definition of (ϵ, δ)-DP as follows.

Definition 1 ((ϵ, δ)-DP [26]). *A randomized algorithm $\mathcal{M} : X^n \to \mathcal{R}$ provides (ϵ, δ)-DP with $0 \leq \delta < 1$, if for all pairs of neighboring datasets $\mathcal{D} \sim \mathcal{D}'$ and all measurable sets of outcomes $S \subseteq \mathcal{R}$, we have*

$$\mathbb{P}[\mathcal{M}(\mathcal{D}) \in S] \leq e^{\epsilon}\mathbb{P}[\mathcal{M}(\mathcal{D}') \in S] + \delta.$$

We say that a randomized algorithm achieves ϵ-DP (also known as *pure* DP) if it satisfies (ϵ, δ)-DP with $\delta = 0$. The common interpretation of δ is the "leakage probability", i.e., (ϵ, δ)-DP is ϵ-DP "except with probability δ".

We next introduce the definition of Rényi DP [119]. As a generalization of (ϵ, δ)-DP, it can help us obtain a tighter (ϵ, δ)-DP bound converted from its composition, which is particularly attractive in FL due to its multiple learning rounds.

Definition 2 ((α, ϵ)-Rényi DP [119]). *A randomized algorithm $\mathcal{M} : X^n \to \mathcal{R}$ is said to provide (α, ϵ)-Rényi DP, if for any pair of neighboring datasets $\mathcal{D} \sim \mathcal{D}'$ it holds that*

$$D_\alpha(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}')) \leq \epsilon,$$

where

$$D_\alpha(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}')) \triangleq \frac{1}{\alpha - 1} \log \int_{-\infty}^{+\infty} P(x)^\alpha Q(x)^{1-\alpha} dx$$

is the Rényi divergence. Specially, for $\alpha = +\infty$, we have

$$D_\infty(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}')) = \sup_{x \in suppQ} \log \frac{P(x)}{Q(x)}.$$

In the AirComp FL design, decoding the global model from the received signal can be regarded as a randomized mechanism on $\mathcal{D} = \bigcup_{k=1}^{M} \mathcal{D}_k$, where $\mathcal{D}$ is the union of the local datasets of all M clients throughout the whole FL task. We denote this randomized mechanism as

$$\mathcal{M}(\mathcal{D}) \triangleq \tilde{\mathbf{x}}_t = g(\mathcal{D}) + \mathbf{n}_t, \tag{4.4}$$

where $g(\mathcal{D}) = \mathbf{x}_t$ is the noise-free summation of model differentials at the t-th learning round, and $\mathbf{n}_t$ is a random noise following a certain distribution. The randomness of the mechanism comes from the random client participation, random mini-batch SGD, and the random noise.

To determine the DP level, we next define the *global sensitivity function* for the operator $g(\cdot)$ as

$$GS_g = \max_{\mathcal{D},\mathcal{D}'} \|g(\mathcal{D}) - g(\mathcal{D}')\|_2 \,, \tag{4.5}$$

where $\mathcal{D}'$ is a neighboring dataset of $\mathcal{D}$. As mentioned in Section 4.2, for uplink communications, we ensure that $\left\|\mathsf{x}_t^k\right\|_2 \leq C$. Therefore, we have

$$GS_g = \max_{\mathcal{D},\mathcal{D}'} \|g(\mathcal{D}) - g(\mathcal{D}')\|_2 = \max_{\mathsf{x}_t^k, \mathsf{x}_t^{k'}} \left\|\mathsf{x}_t^k - \mathsf{x}_t^{k'}\right\|_2 \leq \max_{\mathsf{x}_t^k, \mathsf{x}_t^{k'}} \left[\left\|\mathsf{x}_t^k\right\|_2 + \left\|\mathsf{x}_t^{k'}\right\|_2\right] \leq 2C, \tag{4.6}$$

where $\mathsf{x}_t^{k'}$ denotes the noise-free model differential from client k' whose local dataset is swapped in a random data sample.

4.4.2 Item-level Differential Privacy

The Rényi DP guarantee of `FLORAS` in a single learning round is given in Theorem 6.

Theorem 6. *Assume a spreading sequence set $\mathcal{A}$ containing N unique sequences and M total clients involved in an FL task. In each learning round, $K < \min\{N, M\}$ clients are independently and uniformly randomly selected to participate in FL. With local dataset $\mathcal{D}_k$ of size D and mini-batch size $d_{batch} \triangleq |\xi| < D$, `FLORAS` provides (α, ϵ)-Rényi DP for the global model, where*

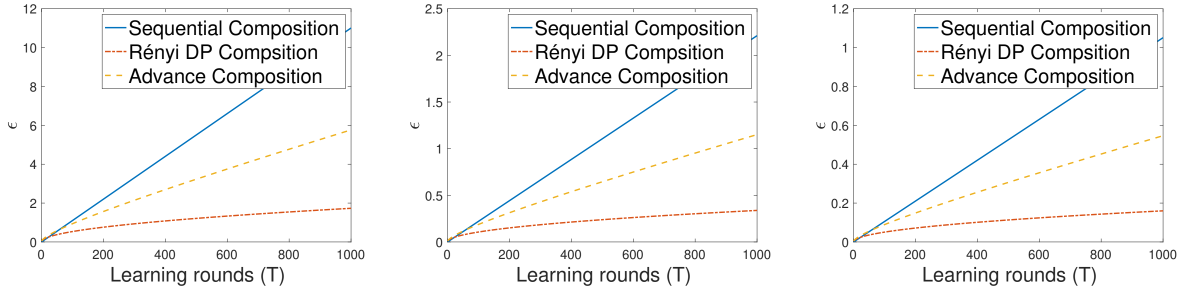
$$\epsilon = \frac{1}{2}\alpha \log^2 \left(1 + \frac{qp}{1+qp} \frac{2C\sqrt{C^2+\gamma^2} + 2C^2}{\gamma^2}\right) = O\left(\frac{\alpha q^2 p^2}{\gamma^2}\right), \tag{4.7}$$

with $q \triangleq \frac{d_{batch}}{D+1-d_{batch}}$, $p \triangleq \frac{K}{M}$, and $\gamma = N - K$.

Proof. See Appendix E.2. □

Based on the composition rule of Rényi DP, we next establish the (ϵ, δ)-DP guarantee for the overall uplink communications in an FL task of T rounds.

Theorem 7. *Consider a wireless FL task with T learning rounds, a spreading sequence set $\mathcal{A}$ containing N unique sequences, and $K < \min\{N, M\}$ clients are independently and uniformly randomly selected in each round. With local*



(a) $q = 0.05, \gamma = 5$ (b) $q = 0.01, \gamma = 5$ (c) $q = 0.01, \gamma = 10$

Figure 4.2: Comparison of sequential composition, Rényi DP composition, and advanced composition with different sizes of mini-batch SGD and orthogonal sequence set.

dataset $\mathcal{D}_k$ of size D and mini-batch size $d_{batch} < D$, `FLORAS` *provides (ϵ', δ)-DP for the entire FL task, where*

$$\epsilon' = \sqrt{2T\log(1/\delta)}\log\left(1 + \frac{qp}{1+qp}\frac{2C\sqrt{C^2+\gamma^2}+2C^2}{\gamma^2}\right) + \frac{1}{2}T\log^2\left(1 + \frac{qp}{1+qp}\frac{2C\sqrt{C^2+\gamma^2}+2C^2}{\gamma^2}\right) \sim \tilde{O}\left(\frac{\sqrt{T}qp}{\gamma}\right), \tag{4.8}$$

with $q \triangleq \frac{d_{batch}}{D+1-d_{batch}}$, $p \triangleq \frac{K}{M}$, and $\gamma = N - K$.

Proof. See Appendix E.3. □

Theorems 6 and 7 reveal that, for a given FL configuration, i.e., fixed number of selected clients K, mini-batch size ratio q, participation ratio p, and the number of learning rounds T, the expansion of spreading sequence set would achieve a higher level of DP per learning round and for the whole learning task, respectively. In particular, $\epsilon \propto \frac{1}{\gamma^2}$ in Rényi DP for each learning round, and $\epsilon' \propto \frac{1}{\gamma}$ in (ϵ, δ)-DP for the whole learning task. Since the BS (adversary) has no knowledge of which particular K out of the total N spreading sequences the clients have chosen, increasing the number of spreading sequences results in a heavier tail of the post-processing Cauchy noise, which achieves better privacy protection.

Guided by Theorem 7, we can adjust N to meet the privacy requirement of a practical FL task. Note that larger noise (better privacy protection) will affect the convergence rate of FL, and we will discuss this impact in detail in Section 4.5.

Remark 15 (Choice of DP metrics)**.** *Note that we use (ϵ, δ)-DP to evaluate the overall DP guarantee of the FL task. The reason why we choose to utilize Rényi DP to analyze the DP guarantee per learning round is the same as [120]: we can obtain a tighter (ϵ, δ)-DP guarantee from the composition of multiple Rényi-DP mechanisms. This advantage is empirically shown in Fig. 4.2, in which we compare the overall DP guarantees v.s. learning rounds of three different* *composition methods: i) vanilla sequential composition of (ϵ, δ)-DP; ii) advanced composition of (ϵ, δ)-DP; iii) (ϵ, δ)-DP converted from composition of Rényi DP. It clearly demonstrates that under different system configurations, Theorem 7 unanimously provides the tightest (ϵ, δ)-DP guarantee, which is consistent with the theoretical and experimental results in [119].*

4.4.3 Client-level Differential Privacy

So far, we have focused on the standard item-level DP that protects a single data sample of a certain local dataset. For FL, another DP concept called *client-level DP* (also known as user-level DP) is also important. The definition of client-level DP follows similarly from Definition 1, with a slight change that the neighboring dataset pair $\mathcal{D} \sim \mathcal{D}'$ differs in at most *all data samples of one single client*. As a result, client-level DP protects privacy when the entire data from a certain client is swapped. Intuitively, this guarantees that the participation of a client cannot be inferred by observing the received signals.

From the previous discussion of global sensitivity, the output x_t^k is always bounded even though the entire dataset of a certain client changes. Therefore, our method intuitively satisfies the client-level DP, and we formally establish the following guarantee.

Theorem 8. *Given a spreading sequence set $\mathcal{A}$ containing N unique sequences, `FLORAS` provides (α, ϵ)-client level Rényi DP, when K clients are independently and uniformly randomly selected from the total M clients in each learning round, where*

$$\epsilon = \frac{1}{2}\alpha \log^2\left(1 + p\frac{2C\sqrt{C^2+\gamma^2}+2C^2}{\gamma^2}\right) = O\left(\frac{\alpha p^2}{\gamma^2}\right). \tag{4.9}$$

Moreover, for total T learning rounds, `FLORAS` provides (ϵ'', δ)-client level DP for the entire FL task, where

$$\begin{aligned}\epsilon'' = &\sqrt{2T\log(1/\delta)}\log\left(1 + p\frac{2C\sqrt{C^2+\gamma^2}+2C^2}{\gamma^2}\right) \\ &+ \frac{1}{2}T\log^2\left(1 + p\frac{2C\sqrt{C^2+\gamma^2}+2C^2}{\gamma^2}\right) \sim \tilde{O}\left(\frac{\sqrt{T}p}{\gamma}\right).\end{aligned} \tag{4.10}$$

Proof. See Appendix E.4. □

Theorem 8 demonstrates that `FLORAS` provides the client-level DP guarantee in a manner that is similar to the item-level DP. However, since the entire local dataset of a certain client would be swapped, we cannot take advantage of the SGD randomness in the analysis of client-level DP. Therefore, q disappears from the client-level DP guarantee.

4.4.4 Spreading Sequence Assignment Mechanism

As we have discussed, the DP guarantee of FLORAS comes from the fact that the assignment of spreading sequences remains unknown to the BS (adversary). In traditional CDMA systems, the spreading sequence is assigned to each device by the BS. This mechanism becomes invalid in our setting, since we need to make sure that the BS only has knowledge of the spreading sequence set $\mathcal{A}$, not the individual assignment. In practice, we can resort to a trusted third-party to handle the assignment of orthogonal sequences. The design of such a mechanism belongs to the field of *secure multi-party computation (MPC)* [121, 122] and is out of the scope of this paper.

In the following, we provide a preliminary reference design based on a random permutation algorithm as illustrated in Fig. 4.3. The proposed design allows each user to autonomously choose a unique spreading code without collision. To better explain the mechanism, we first define a *fixed* matrix based on $\mathcal{A} = \{\mathbf{a}_k, k = 1, \cdots, N\} : \mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_N] \in \mathbb{R}^{L \times N}$. We assume that every device that will be involved in the FL task shares a common confidential `Key`. This key is *a priori* knowledge to clients, yet confidential to the BS (adversary). This assumption can be achieved via standard cryptography approaches, e.g., the key-exchange protocol [123]. When the BS schedules clients to participate in the current learning round, an `index` $\in \{1, \cdots, K\}$ will be assigned to each client. After that, every client leverages the confidential `Key` and the current system time `SystemTime()` to generate a random seed `RandSeed()`. Since the system has been synchronized, each client will obtain the same random seed, and generate a (common) random permutation matrix $\mathbf{P}_\pi \in \mathbb{R}^{N \times N}$ based on that. Followed by a random column permutation $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{P}_\pi$, each client can use the `index`-th column $\tilde{\mathbf{A}}$`[:,index]` as its spreading sequence for the current learning round. Note that as long as the confidential key is not leaked, the BS (adversary) will not know which K of the total N spreading sequences are adopted in the current learning round.

Remark 16. *We emphasize that the established DP guarantee is for the receiver processing proposed in Section 4.3.1. There may exist more sophisticated receiver algorithms, which could conceivably attempt to infer more information about the usage of the orthogonal sequences per learning round from the received signal. However, since we do not make any assumptions on the channel distribution, such approach would be difficult in general. Besides, it is impossible to always accurately identify all used sequences due to the channel noise. Therefore, even under this scenario, the proposed framework still provides privacy, although the DP guarantee dependency may decrease from $O(1/\gamma)$ to $O(1/(\alpha\gamma))$, where $0 < \alpha < 1$.*

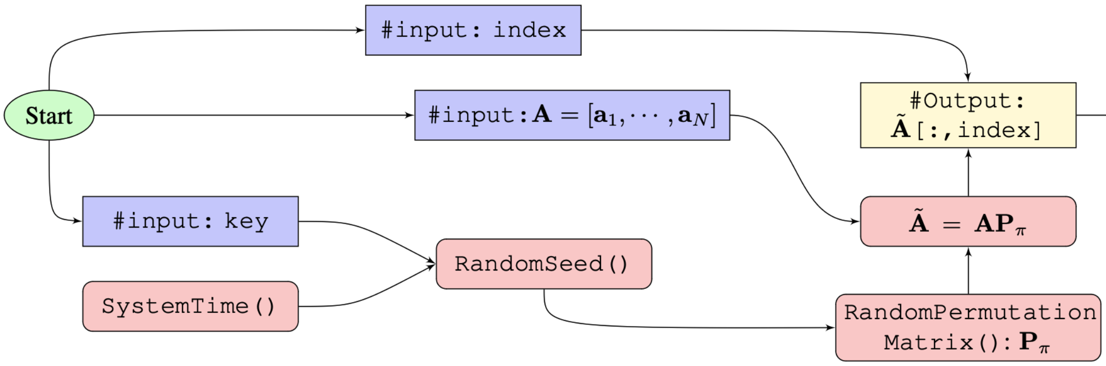

#input: index
Start
#input: $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_N]$
#Output: $\tilde{\mathbf{A}}$[:,index]
End
#input: key
$\tilde{\mathbf{A}} = \mathbf{A}\mathbf{P}_\pi$
SystemTime()
RandomSeed()
RandomPermutation Matrix(): $\mathbf{P}_\pi$

Figure 4.3: Flowchart of the proposed spreading sequence assignment reference design for any given client.

4.5 Convergence analysis

The main challenge, and hence the novelty of our analysis, lies in the Cauchy distributed post-processing noise. We note that a Cauchy distribution has uncertain (infinity) variance. To address this issue, the BS applies a *truncation* operation in the interval $[-B, B]$ on the decoded global parameters in (4.3), before de-normalization:

$$\tilde{x}_i \approx \max\left(\min\left(\sum_{k=1}^{K} x_k^i + \sum_{k=K+1}^{N} \frac{\mathbf{a}_k^T \mathbf{n}_i}{\mathbf{a}_k^T \mathbf{n}_s}, B\right), -B\right) \tag{4.11}$$

where $B \gg C$ and C is a normalization parameter defined in Section 4.2. Note that the truncation operation is universal (albeit sometimes implicit) in almost all practical systems, since the signal values in the processing units are always finite. As a post-processing procedure, the truncation operation has no impact on the DP guarantee of FLORAS. Unfortunately, it will introduce a bias in the estimate $\tilde{x}_i$, which impacts convergence. In particular, the equivalent noise after the limiting operation will be a *truncated Cauchy distribution* with support $[-B - \sum_{k=1}^{K} x_k^i, B - \sum_{k=1}^{K} x_k^i]$. Fortunately, as long as we ensure $B \gg C$ which requires a very mild truncation and thus is easy to satisfy, this operation will only have very limited impact on the received signal, hence does not significantly harm the convergence performance. We also note that biased gradients are very common in various DP-guaranteed SGD methods [105, 124, 125]. Moreover, after de-normalization, the biased term is effectively diminishing. Therefore, as we show in Theorem 9, FLORAS can still maintain a good convergence performance despite of the biased noise, which will also be numerically corroborated in Section 4.6.

Theorem 9. *With Assumptions 1-4 and $\mu > 2\sqrt{dD(\gamma)}EH/K$, for some $r \geq 0$, if we set the learning rate as $\eta_t = \frac{2}{\mu'(t+r)}$, a wireless system implementing* `FLORAS` *achieves*

$$\mathbb{E}[f(\mathbf{w}_t)] - f^* \leq \frac{L}{(t+r)} \left[\frac{4G}{\mu'^2} + (1+r) \left\| \mathbf{w}_0 - \mathbf{w}^* \right\|^2 \right],$$

for any $t \geq 1$, where $\mu' \triangleq \mu - 2\sqrt{dD(\gamma)}EH/K$,

$$G \triangleq \left(1 + \frac{2\sqrt{dD(\gamma)}}{K} EH\eta_1\right) \left(\sum_{k=1}^{M} \frac{H_k^2}{M^2} + 6L\Gamma + 8(E-1)^2H^2 + \frac{M-K}{M-1}\frac{4}{K}E^2H^2 \right) + \frac{4dD(\gamma)}{K^2}E^2H^2.$$

and

$$D(\gamma) = \frac{\gamma^2}{C^2 \arctan\left(\frac{B+C}{\gamma}\right)} \left[\frac{B}{\gamma} - \arctan\left(\frac{B+C}{\gamma}\right) \right].$$

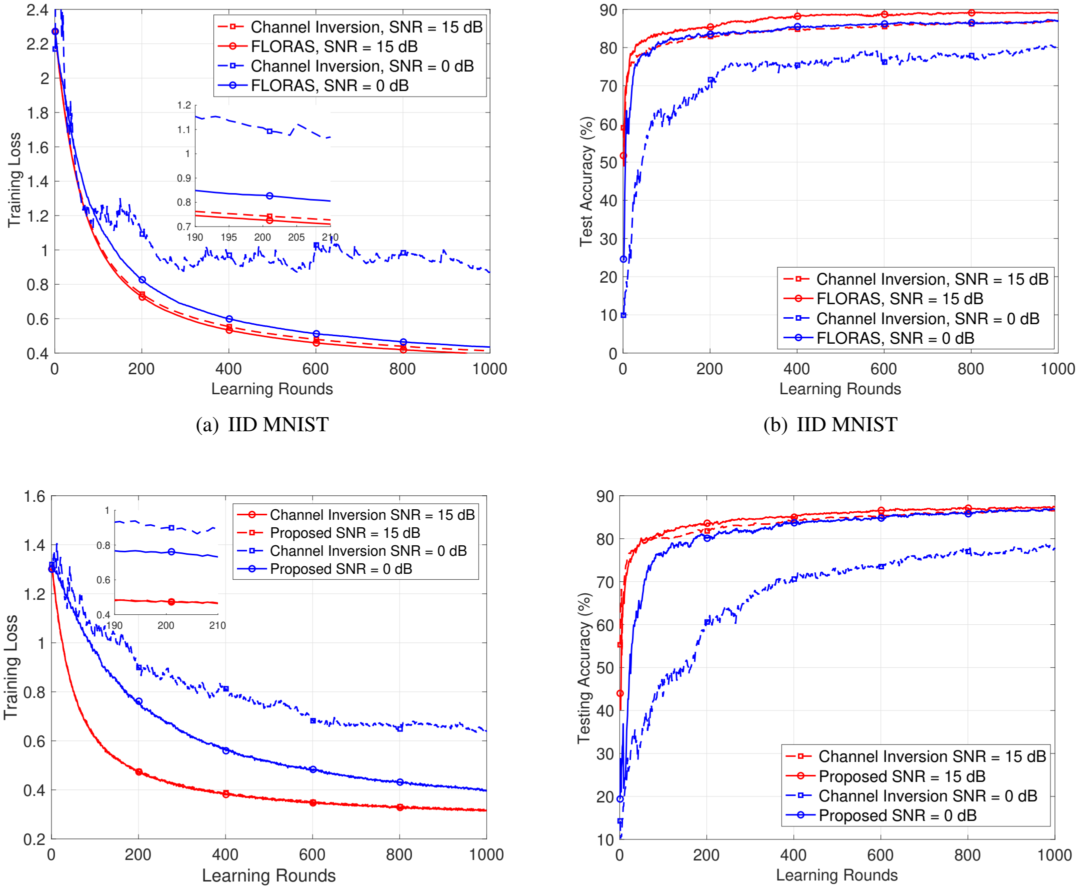
Proof. See Appendix F. □

Theorem 9 demonstrates that `FLORAS` preserves the $O(1/T)$ convergence rate of SGD for strongly convex loss functions (compared with the noise-free FL convergence). For a fixed initial point, there are multiple factors in the constant G that affect the convergence rate of FL. In particular, $\sum_{k=1}^{K} \frac{H_k^2}{K^2}$ reveals the *variance reduction* effect of SGD by involving more clients, and $6L\Gamma$ and $8(E-1)^2H^2$ capture the influence of non-IID dataset and the number of local epochs, respectively. We note that term $D(\gamma)$ in G demonstrates the impact from Cauchy noise, i.e. level of privacy protection. We also note that $D(\gamma)$ is increasing as γ becomes larger. It implies that a higher level of privacy protection will decrease the speed of convergence as constant $D(\gamma)$ becomes larger. Therefore, Theorem 9 establishes a tradeoff between privacy protection and convergence rate, which can guide the practical system design.

4.6 Experiments

In this section, we evaluate the performance of `FLORAS` through numerical experiments. We first compare the learning performance of `FLORAS` with the widely-investigated channel inversion AirComp method. Then, we evaluate the effect on the ML model convergence rate of various DP levels. In particular, we corroborate the theoretical results via real-world FL tasks on the MNIST dataset. The experiments demonstrate that `FLORAS` achieves superior performances under various SNRs and other system configurations.

Details on the setup of experiments are as follows. All convergence curves are the average of five independent Monte Carlo trials. The MNIST dataset contains multiple handwritten digit figures of 20×20 pixels. The training set contains $4,000$ examples and is evenly distributed over $K = 20$ clients. For the IID case, the data is shuffled and randomly assigned to each client; for the non-IID case, the data is sorted by labels and each client is



(a) IID MNIST

(b) IID MNIST

(c) non-IID MNIST

(d) non-IID MNIST

Figure 4.4: Training loss/test accuracy versus learning rounds of `FLORAS` and channel inversion for MNIST dataset under SNR $= 0$ and 15 dB.

randomly assigned with the data of one label. The test set size is $1,000$. We evaluate `FLORAS` and validate the theoretical results on a multinomial logistic regression task. Specifically, let $f(\mathbf{w}; x_i)$ denote the prediction model with the parameter $\mathbf{w} = (\mathbf{W}, \mathbf{b})$ and the form $f(\mathbf{w}; x_i) = \texttt{softmax}(\mathbf{W}x_i + \mathbf{b})$. The loss function is given by $\texttt{loss}(\mathbf{w}) = \frac{1}{D}\sum_{i=1}^{D} \texttt{CrossEntropy}(f(\mathbf{w}; x_i), \mathbf{y}_i) + \lambda \|\mathbf{w}\|^2$. We adopt the regularization parameter $\lambda = 0.01$ in the experiments.

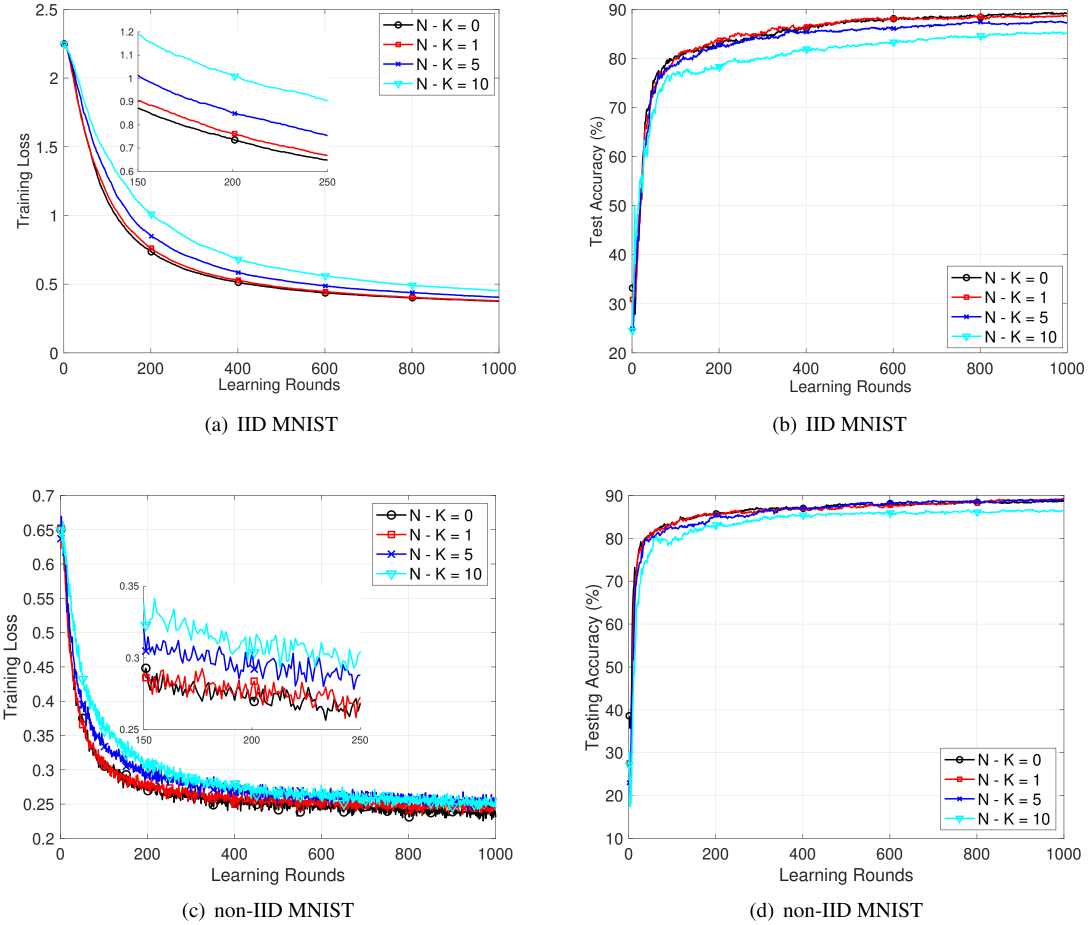
4.6.1 Communication Efficiency

We first evaluate the performance of `FLORAS` compared with the channel inversion method. We assume IID Raleigh block fading channel $h_k \sim \mathcal{CN}(0, 1), \forall k = 1, \cdots, K$. For channel inversion, we adopt a user admission threshold 0.01 for the fading channel gain to avoid deep fading. The following parameters are used for training: local batch size 50, the number of local epochs $E = 1$, and learning rate $\eta = 0.005$ and $\eta = 0.001$ for the IID and non-IID case, respectively.

Fig. 4.4-(a) and Fig. 4.4-(b) illustrate the training loss and test accuracy performance versus learning round of FLORAS and channel inversion in high (red line) and low (blue) SNR regimes, respectively. For high SNR (15 dB), FLORAS and channel inversion have similar performances. Although transmitter fading channel cancellation in channel inversion limits the maximum transmission power of each user, its performance does not deteriorate, since noise is not the dominant factor of the convergence. However, we note that, unlike FLORAS, channel inversion requires full CSIT at each client, which not only consumes larger communication overhead, but also increases the dynamic range of the signal, bringing higher hardware cost. The advantages of FLORAS become conspicuous in the low SNR regime (0 dB), in which noise becomes the dominant factor of the convergence rate. FLORAS allows all participated clients to make full use of transmit power and achieves significantly better performance. As shown in Fig. 4.4-(b), FLORAS achieves about 7.5% higher test accuracy compared with channel inversion at SNR = 0 dB. This phenomenon becomes more notable in the non-IID dataset as shown in Fig. 4.4-(c) and Fig. 4.4-(d), where the performance gap of test accuracy between FLORAS and channel inversion further grows to 10.2%.

4.6.2 Differential Privacy

We also evaluate the convergence performance versus different levels of DP. We keep SNR $=$ 20 dB and number of selected user $K = 20$, while changing the size of the orthogonal sequence set $\mathcal{A}$ from $20, 21, 25$ to 30, i.e., $\gamma = 0, 1, 5$ and 10. In each learning round, each user selects its own signature from set $\mathcal{A}$ via the method in Section 4.4.4. As discussed in Section 4.4, the larger size of set $\mathcal{A}$, the higher DP level FLORAS achieves. The following parameters are used for training: local batch size 20, the number of local epochs $E = 1$, and learning rate $\eta = 0.005$ and $\eta = 0.001$ for IID and non-IID cases, respectively. The training losses with different DP levels are illustrated in Fig. 4.5-(a) and Fig. 4.5-(c) of IID and non-IID cases, respectively. It is clear that although a higher privacy level decreases the convergence rate, the machine learning model can still converge with almost the same training loss as the no-differential-privacy case ($\gamma = 0$), which is consistent with the theoretical analysis in Section 4.5. The test accuracies in Fig. 4.5-(b) and Fig. 4.5-(d) further validate the effectiveness of FLORAS. We can see that the test accuracies for moderate DP levels ($\gamma = 1$ and 5) are almost the same as the case of $\gamma = 0$, i.e., we can achieve certain DP levels *almost for free*. Even when $\gamma = 10$, the test accuracy loss is tiny, with about 3.5% and 2.5% compared with the $\gamma = 0$ case in the IID and non-IID datasets, respectively.



(a) IID MNIST

(b) IID MNIST

(c) non-IID MNIST

(d) non-IID MNIST

Figure 4.5: Training loss/test accuracy versus learning rounds of FLORAS with different differential privacy levels for MNIST dataset under SNR $=$ 20 dB.

4.7 Summary

The FLORAS framework has been introduced as a differentially private AirComp FL solution. In comparison to the channel inversion method, FLORAS eliminates the need for CSIT and exhibits significantly enhanced robustness in low SNR environments, a factor of paramount importance for Internet of Things (IoT) applications. The framework's adaptability, through the adjustment of the orthogonal sequence set size, facilitates the management of both item-level and client-level DP assurances. Analysis of both the convergence and DP properties of FLORAS has illuminated the trade-off between convergence rate and privacy preservation, a finding substantiated by experimental results from real-world FL tasks.

Chapter 5

Conclusion and Future Work

This dissertation conducts an in-depth exploration of communication design strategies for wireless FL. The investigation spans various scenarios, environments, and settings to uncover how different factors and designs impact the efficiency and reliability of wireless FL.

Chapter 2 delves into the impact of communication-induced noise on the convergence of FL during both uplink and downlink phases. Through theoretical analysis, this study uncovers valuable insights that guide the formulation of a SNR scaling strategy. This strategy is pivotal in ensuring the convergence of FL within the limits of a total resource budget. The framework introduced is versatile, designed to support a range of conditions including full and partial client engagement, both direct model updates and differential model transmission, as well as the challenge of handling non-IID local datasets.

For MIMO systems studied in Chapter 3, by capitalizing on the distinct characteristics of channel hardening and favorable propagation characteristic of massive MIMO, “random orthogonalization” design was proposed. It enables an efficient uplink and downlink transmission method without CSIT. This method significantly reduces the overhead, computational complexity and latency of uplink communication, presenting an effective solution to previously insurmountable challenges.

For SISO systems studied in Chapter 4, `FLORAS` design was proposed. It is a novel uplink wireless physical layer design for FL by leveraging the properties of orthogonal sequences. `FLORAS` enjoys all advantages of AirComp, yet without the CSIT requirement. Moreover, by the adjustment on the number of used orthogonal sequences in the system configuration, the novel signal processing techniques in `FLORAS` empowers flexible item-level and client-level DP guarantee.

This dissertation makes a significant contribution to the burgeoning field of wireless FL by delivering in-depth insights into its efficiency, reliability, and privacy. By tackling the critical bottlenecks within FL, the research enhances the comprehension of wireless design principles for FL across a variety of systems, broadening the scope of knowledge and application in this dynamic and evolving field. There are several avenues for future research that can further extend the current knowledge and contribute to practical applications.

Heterogeneous Networks. Exploring the adaptation and optimization of FL in heterogeneous networks represents a promising direction for future research. These networks, characterized by a diverse array of devices with varying computational capabilities, network connections, and mobility, pose unique challenges for FL. Investigating strategies to efficiently integrate and manage FL processes across such varied environments can enhance scalability and inclusivity.

Robustness and Fault Tolerance. Ensuring the robustness and fault tolerance of FL systems in the face of device failures, network disruptions, and adversarial attacks is essential for their reliability and trustworthiness. Future work could investigate mechanisms to detect and mitigate such issues in real-time, enhancing the resilience of FL systems. This may involve the development of robust aggregation algorithms, secure communication channels, and anomaly detection techniques that can safeguard the FL process against various threats.

Personalized Differential Privacy. Wireless federated learning could significantly benefit from exploring the integration of personalized DP mechanisms. This entails developing adaptive privacy-preserving algorithms that can tailor the level of privacy guarantees to individual users' preferences or the sensitivity of their data. Such personalized DP approaches could leverage advancements in wireless communication technologies to optimize the trade-off between privacy, model accuracy, and communication efficiency. Investigating the impact of personalized DP on the convergence rates of FL models, especially in non-IID and resource-constrained environments, would be crucial. Moreover, future research could focus on designing novel cryptographic techniques or secure multi-party computation methods to support personalized DP in FL, ensuring robust privacy protection without compromising the collaborative learning process. This exploration would not only advance the state of FL but also open new avenues for creating more user-centric and privacy-aware machine learning models.

Bibliography

[1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proc. Artificial Intelligence and Statistics*, pages 1273–1282, Fort Lauderdale, FL, USA, Apr. 2017.

[2] Jakub Konecny, H. Brendan McMahan, Felix X. Yu, Peter Richtarik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. In *Proc. NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[3] Keith Bonawitz et al. Towards federated learning at scale: System design. In *The 2nd SysML Conference*, pages 1–15, 2019.

[4] Zhaohui Yang, Mingzhe Chen, Kai-Kit Wong, H Vincent Poor, and Shuguang Cui. Federated learning for 6G: Applications, challenges, and opportunities. *arXiv preprint arXiv:2101.01338*, 2021.

[5] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proc. the 3rd MLSys Conference*, 2020.

[6] Sihui Zheng, Cong Shen, and Xiang Chen. Design and analysis of uplink and downlink communications for federated learning. *IEEE J. Select. Areas Commun.*, 39(7):2150–2167, July 2021.

[7] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization. In *AISTATS*, 2020.

[8] Yuqing Du, Sheng Yang, and Kaibin Huang. High-dimensional stochastic gradient quantization for communication-efficient edge learning. *IEEE Trans. Signal Processing*, 68:2128–2142, 2020.

[9] Xizixiang Wei, Cong Shen, Jing Yang, and H. Vincent Poor. Random orthogonalization for federated learning in massive MIMO systems. *IEEE Trans. Wireless Commun.*, 23(3):2469–2485, 2024.

[10] Yujia Mu, Cong Shen, and Yonina C. Eldar. Optimizing federated averaging over fading channels. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, June 2022.

[11] Xizixiang Wei, Cong Shen, Jing Yang, and H. Vincent Poor. Random orthogonalization for federated learning in massive MIMO systems. In *Proc. IEEE Int. Conf. Commun.*, pages 3382–3387, 2022.

[12] Xizixiang Wei and Cong Shen. Federated learning over noisy channels. In *Proc. IEEE Int. Conf. Commun.*, June 2021.

[13] Xizixiang Wei, Cong Shen, Jing Yang, and H. V. Poor. Random orthogonalization for federated learning in massive MIMO systems. In *Proc. IEEE International Conference on Communications (ICC)*, pages 1–6, May 2022.

[14] Xizixiang Wei, Tianhao Wang, Ruiquan Huang, Cong Shen, Jing Yang, and H. Vincent Poor. FLORAS: Differentially private wireless federated learning using orthogonal sequences. In *Proc. IEEE Int. Conf. Commun.*, pages 3121–3126, May 2023.

[15] Xizixiang Wei, Tianhao Wang, Ruiquan Huang, Cong Shen, Jing Yang, and H Vincent Poor. Differentially private wireless federated learning using orthogonal sequences. *arXiv preprint arXiv:2306.08280*, 2023.

[16] Solmaz Niknam, Harpreet S Dhillon, and Jeffrey H Reed. Federated learning for wireless communications: Motivation, opportunities, and challenges. *IEEE Commun. Mag.*, 58(6):46–51, 2020.

[17] Guangxu Zhu, Yong Wang, and Kaibin Huang. Broadband analog aggregation for low-latency federated edge learning. *IEEE Trans. Wireless Commun.*, 19(1):491–506, 2020.

[18] Kai Yang, Tao Jiang, Yuanming Shi, and Zhi Ding. Federated learning via over-the-air computation. *IEEE Trans. Wireless Commun.*, 19(3):2022–2035, 2020.

[19] Mohammad Mohammadi Amiri and Deniz Gündüz. Federated learning over wireless fading channels. *IEEE Trans. Wireless Commun.*, 19(5):3546–3557, 2020.

[20] X. Cao, G. Zhu, J. Xu, and K. Huang. Optimized power control for over-the-air computation in fading channels. *IEEE Trans. Wireless Commun.*, 19(11):7498–7513, 2020.

[21] D. Tse and P. Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[22] Guangxu Zhu, Jie Xu, Kaibin Huang, and Shuguang Cui. Over-the-air computing for wireless data aggregation in massive IoT. *IEEE Wireless Commun.*, 28(4):57–65, 2021.

[23] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proc. 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1310–1321, 2015.

[24] Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *Proc. IEEE Int. Conf. Comput. Commun.*, pages 2512–2520, 2019.

[25] Chuan Ma, Jun Li, Ming Ding, Howard H Yang, Feng Shu, Tony QS Quek, and H Vincent Poor. On safeguarding privacy and security in the framework of federated learning. *IEEE Network*, 34(4):242–248, 2020.

[26] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[27] Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

[28] Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang, and Yi Zhou. A hybrid approach to privacy-preserving federated learning. In *Proc. the 12th ACM Workshop on Artificial Intelligence and Security*, pages 1–11, 2019.

[29] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[30] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-IID data. In *Proc. International Conference on Learning Representations*, 2020.

[31] Peng Jiang and Gagan Agrawal. A linear speedup analysis of distributed deep learning with sparse and quantized communication. In *Proc. Advances in Neural Information Processing Systems*, pages 2525–2536, 2018.

[32] Sebastian U Stich. Local SGD converges fast and communicates little. In *Proc. International Conference on Learning Representations*, 2018.

[33] Peter Kairouz et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[34] Shuhao Xia, Jingyang Zhu, Yuhan Yang, Yong Zhou, Yuanming Shi, and Wei Chen. Fast convergence algorithm for analog federated learning. In *Proc. IEEE Int. Conf. Commun.*, pages 1–6, June 2021.

[35] Tomer Sery, Nir Shlezinger, Kobi Cohen, and Yonina C Eldar. Over-the-air federated learning from heterogeneous data. *arXiv preprint arXiv:2009.12787*, 2020.

[36] H. Guo, A. Liu, and V. K. N. Lau. Analog gradient aggregation for federated learning over wireless networks: Customized design and convergence analysis. *IEEE Internet Things J.*, 8(1):197–210, 2021.

[37] Mohammad Mohammadi Amiri, Deniz Gunduz, Sanjeev R Kulkarni, and H Vincent Poor. Convergence of federated learning over a noisy downlink. *arXiv preprint arXiv:2008.11141*, 2020.

[38] Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. In *Proc. ICML Workshop on Coding Theory for Machine Learning*, 2019.

[39] Guangxu Zhu, Yuqing Du, Deniz Gündüz, and Kaibin Huang. One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis. *IEEE Trans. Wireless Commun.*, 20(3):2120–2135, 2021.

[40] Mohammad Mohammadi Amiri, Deniz Gunduz, Sanjeev R Kulkarni, and H Vincent Poor. Federated learning with quantized global model updates. *arXiv preprint arXiv:2006.10672*, 2020.

[41] Shengbo Chen, Cong Shen, Lanxue Zhang, and Yuanmin Tang. Dynamic aggregation for heterogeneous quantization in federated learning. *IEEE Trans. Wireless Commun.*, 2021. to appear.

[42] Xiaopeng Mo and Jie Xu. Energy-efficient federated edge learning with joint communication and computation design. *arXiv preprint arXiv:2003.00199*, 2020.

[43] Qunsong Zeng, Yuqing Du, Kin K Leung, and Kaibin Huang. Energy-efficient radio resource allocation for federated edge learning. *arXiv preprint arXiv:1907.06040*, 2019.

[44] Wenqi Shi, Sheng Zhou, and Zhisheng Niu. Device scheduling with fast convergence for wireless federated learning. *arXiv preprint arXiv:1911.00856*, 2019.

[45] Zhaohui Yang, Mingzhe Chen, Walid Saad, Choong Seon Hong, and Mohammad Shikh-Bahaei. Energy efficient federated learning over wireless communication networks. *arXiv preprint arXiv:1911.02417*, 2019.

[46] Howard H Yang, Zuozhu Liu, Tony QS Quek, and H Vincent Poor. Scheduling policies for federated learning in wireless networks. *IEEE Trans. Commun.*, 68(1):317–333, 2020.

[47] Mingzhe Chen, H Vincent Poor, Walid Saad, and Shuguang Cui. Convergence time optimization for federated learning over wireless networks. *arXiv preprint arXiv:2001.07845*, 2020.

[48] Jie Xu and Heqiang Wang. Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective. *IEEE Trans. Wireless Commun.*, 20(2):1188–1200, 2021.

[49] Mingzhe Chen, Zhaohui Yang, Walid Saad, Changchuan Yin, H Vincent Poor, and Shuguang Cui. A joint learning and communications framework for federated learning over wireless networks. *arXiv preprint arXiv:1909.07972*, 2019.

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

[51] Matthias Frey, Igor Bjelakovic, and Slawomir Stanczak. Over-the-air computation for distributed machine learning. *arXiv preprint arXiv:2007.02648*, 2020.

[52] R. Jiang and S. Zhou. Cluster-based cooperative digital over-the-air aggregation for wireless federated edge learning. In *IEEE/CIC International Conference on Communications in China (ICCC)*, pages 887–892, 2020.

[53] F. Ang, L. Chen, N. Zhao, Y. Chen, W. Wang, and F. R. Yu. Robust federated learning with noisy communication. *IEEE Trans. Commun.*, 68(6):3452–3464, 2020.

[54] Hanlin Tang, Chen Yu, Xiangru Lian, Tong Zhang, and Ji Liu. Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In *Proc. International Conference on Machine Learning*, pages 6155–6165. PMLR, 2019.

[55] Yue Yu, Jiaxiang Wu, and Longbo Huang. Double quantization for communication-efficient distributed optimization. *arXiv preprint arXiv:1805.10111*, 2018.

[56] Chia-Yu Chen et al. ScaleCom: Scalable sparsified gradient compression for communication-efficient distributed training. In *Proc. Advances in Neural Information Processing Systems*, pages 13551–13563, 2020.

[57] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Proc. Advances in Neural Information Processing Systems*, 2016.

[58] Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I. Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM J. Optim.*, 27(4):2202–2229, 2017.

[59] Amirhossein Reisizadeh, Isidoros Tziotis, Hamed Hassani, Aryan Mokhtari, and Ramtin Pedarsani. Straggler-resilient federated learning: Leveraging the interplay between statistical accuracy and system heterogeneity. *arXiv preprint arXiv:2012.14453*, 2020.

[60] Andrea Goldsmith. *Wireless Communications*. Cambridge University Press, 2005.

[61] Sebastian Caldas et al. LEAF: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

[62] Ahmet M Elbir and Sinem Coleri. Federated learning for hybrid beamforming in mm-Wave massive MIMO. *IEEE Commun. Letter*, 24(12):2795–2799, 2020.

[63] Tao Huang, Baoliu Ye, Zhihao Qu, Bin Tang, Lei Xie, and Sanglu Lu. Physical-layer arithmetic for federated learning in uplink MU-MIMO enabled wireless networks. In *Proc. IEEE Conference on Computer Communications (INFOCOM)*, pages 1221–1230, 2020.

[64] Yo-Seb Jeon, Mohammad Mohammadi Amiri, Jun Li, and H Vincent Poor. A compressive sensing approach for federated learning over massive MIMO communication systems. *IEEE Trans. Wireless Commun.*, 20(3):1990–2004, 2020.

[65] Kundjanasith Thonglek, Keichi Takahashi, Kohei Ichikawa, Chawanat Nakasan, Pattara Leelaprute, and Hajimu Iida. Sparse communication for federated learning. In *Proc. IEEE 6th International Conference on Fog and Edge Computing (ICFEC)*, pages 1–8, 2022.

[66] Yongjeong Oh, Yo-Seb Jeon, Mingzhe Chen, and Walid Saad. FedVQCS: Federated learning via vector quantized compressed sensing. *arXiv preprint arXiv:2204.07692*, 2022.

[67] Dingzhu Wen, Ki-Jun Jeon, Mehdi Bennis, and Kaibin Huang. Adaptive subcarrier, parameter, and power allocation for partitioned edge learning over broadband channels. *IEEE Trans. Wireless Commun.*, 20(12):8348–8361, 2021.

[68] Hao Chen, Shaocheng Huang, Deyou Zhang, Ming Xiao, Mikael Skoglund, and H Vincent Poor. Federated learning over wireless IoT networks with optimized communication and resources. *IEEE Internet Things J.*, 9(17):16592–16605, 2022.

[69] Sihua Wang, Mingzhe Chen, Changchuan Yin, Walid Saad, Choong Seon Hong, Shuguang Cui, and H Vincent Poor. Federated learning for task and resource allocation in wireless high-altitude balloon networks. *IEEE Internet Things J.*, 8(24):17460–17475, 2021.

[70] Bobak Nazer and Michael Gastpar. Computation over multiple-access channels. *IEEE Trans. Inf. Theory*, 53(10):3498–3516, 2007.

[71] Mingzhe Chen, H Vincent Poor, Walid Saad, and Shuguang Cui. Convergence time optimization for federated learning over wireless networks. *IEEE Trans. Wireless Commun.*, 20(4):2457–2471, 2020.

[72] Yuxuan Sun, Sheng Zhou, Zhisheng Niu, and Deniz Gündüz. Dynamic scheduling for over-the-air federated edge learning with energy constraints. *IEEE J. Select. Areas Commun.*, 40(1):227–242, 2021.

[73] Xiang Ma, Haijian Sun, Qun Wang, and Rose Qingyang Hu. User scheduling for federated learning through over-the-air computation. In *Proc. IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, pages 1–5, 2021.

[74] Hyun-Suk Lee and Jang-Won Lee. Adaptive transmission scheduling in wireless networks for asynchronous federated learning. *IEEE J. Select. Areas Commun.*, 39(12):3673–3687, 2021.

[75] Madhusanka Manimel Wadu, Sumudu Samarakoon, and Mehdi Bennis. Joint client scheduling and resource allocation under channel uncertainty in federated learning. *IEEE Trans. Commun.*, 69(9):5962–5974, 2021.

[76] Tomer Sery and Kobi Cohen. On analog gradient descent learning over multiple access fading channels. *IEEE Trans. Signal Processing*, 68:2897–2911, 2020.

[77] Zhenyi Lin, Xiaoyang Li, Vincent KN Lau, Yi Gong, and Kaibin Huang. Deploying federated learning in large-scale cellular networks: Spatial convergence analysis. *IEEE Trans. Wireless Commun.*, 21(3):1542–1556, 2021.

[78] Ozan Aygün, Mohammad Kazemi, Deniz Gündüz, and Tolga M Duman. Over-the-air federated learning with energy harvesting devices. In *Proc. IEEE Glob. Commun. Conf.*, pages 1942–1947. IEEE, 2022.

[79] Shuo Wan, Jiaxun Lu, Pingyi Fan, Yunfeng Shao, Chenghui Peng, and Khaled B Letaief. Convergence analysis and system design for federated learning over wireless networks. *IEEE J. Select. Areas Commun.*, 39(12):3622–3639, 2021.

[80] Tomer Sery, Nir Shlezinger, Kobi Cohen, and Yonina C Eldar. Over-the-air federated learning from heterogeneous data. *IEEE Trans. Signal Processing*, 69:3796–3811, 2021.

[81] Yuxuan Sun, Sheng Zhou, Zhisheng Niu, and Deniz Gündüz. Time-correlated sparsification for efficient over-the-air model aggregation in wireless federated learning. In *Proc. IEEE Int. Conf. Commun.*, pages 1–6, May 2022.

[82] Tung T Vu, Hien Q Ngo, Minh N Dao, Duy T Ngo, Erik G Larsson, and Tho Le-Ngoc. Energy-efficient massive MIMO for federated learning: Transmission designs and resource allocations. *arXiv preprint arXiv:2112.11723*, 2021.

[83] Rami Hamdi, Mingzhe Chen, Ahmed Ben Said, Marwa Qaraqe, and H Vincent Poor. Federated learning over energy harvesting wireless networks. *IEEE Internet Things J.*, 9(1):92–103, 2021.

[84] Tung T Vu, Duy T Ngo, Hien Quoc Ngo, Minh N Dao, Nguyen H Tran, and Richard H Middleton. Joint resource allocation to minimize execution time of federated learning in cell-free massive MIMO. *IEEE Internet Things J.*, Early access, 2022.

[85] Tung Thanh Vu, Duy Trong Ngo, Nguyen H Tran, Hien Quoc Ngo, Minh Ngoc Dao, and Richard H Middleton. Cell-free massive MIMO for wireless federated learning. *IEEE Trans. Wireless Commun.*, 19(10):6377–6392, 2020.

[86] Yuchen Mu, Navneet Garg, and Tharmalingam Ratnarajah. Communication-efficient federated learning for massive MIMO systems. In *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, pages 578–583, 2022.

[87] Chunmei Xu, Shengheng Liu, Zhaohui Yang, Yongming Huang, and Kai-Kit Wong. Learning rate optimization for federated learning exploiting over-the-air computation. *IEEE J. Select. Areas Commun.*, 39(12):3742–3756, 2021.

[88] Zhenyi Lin, Yi Gong, and Kaibin Huang. Distributed over-the-air computing for fast distributed optimization: Beamforming design and convergence analysis. *arXiv preprint arXiv:2204.06876*, 2022.

[89] Chenxi Zhong, Huiyuan Yang, and Xiaojun Yuan. Over-the-air federated multi-task learning over MIMO multiple access channels. *arXiv preprint arXiv:2112.13603*, 2021.

[90] Mohammad Mohammadi Amiri, Tolga M Duman, Deniz Gündüz, Sanjeev R Kulkarni, and H Vincent Poor. Blind federated edge learning. *IEEE Trans. Wireless Commun.*, 20(8):5129–5143, 2021.

[91] Busra Tegin and Tolga M Duman. Blind federated learning at the wireless edge with low-resolution ADC and DAC. *IEEE Transactions on Wireless Communications*, 20(12):7786–7798, 2021.

[92] Carsten Bockelmann, Nuno K Pratas, Gerhard Wunder, Stephan Saur, Monica Navarro, David Gregoratti, Guillaume Vivier, Elisabeth De Carvalho, Yalei Ji, Cedomir Stefanovic, et al. Towards massive connectivity support for scalable mmtc communications in 5g networks. *IEEE access*, 6:28969–28992, 2018.

[93] ND Sidiropoulos and TN Davidson. Broadcasting with channel state information. In *Proc. of 2004 Sensor Array and Multichannel Signal*, pages 489–493, 2004.

[94] S. Sesia, I. Toufik, and M. Baker. *LTE - The UMTS Long Term Evolution: From Theory to Practice*. Wiley, 2 edition, 2011.

[95] Steven M Kay. *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993.

[96] Hien Quoc Ngo, Erik G Larsson, and Thomas L Marzetta. Aspects of favorable propagation in massive MIMO. In *Proc. IEEE 22nd European Signal Processing Conference (EUSIPCO)*, pages 76–80, 2014.

[97] Lang P. Withers, Robert M. Taylor, and David M. Warme. Echo-MIMO: A two-way channel training method for matched cooperative beamforming. *IEEE Transactions on Signal Processing*, 56(9):4419–4432, 2008.

[98] Xiangyun Zhou, Tharaka A Lamahewa, Parastoo Sadeghi, and Salman Durrani. Two-way training: Optimal power allocation for pilot and data transmission. *IEEE transactions on wireless communications*, 9(2):564–569, 2010.

[99] Xizixiang Wei and Cong Shen. Federated learning over noisy channels: Convergence analysis and design examples. *IEEE Trans. Cogn. Commun. Netw.*, 8(2):1253–1268, 2022.

[100] Xizixiang Wei, Cong Shen, Jing Yang, and H. Vincent Poor. Technical report: Random orthogonalization for federated learning in massive MIMO systems. Technical report, University of Virginia, Aug. 2022.

[101] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Proc. Advances in Neural Information Processing Systems*, 26:315–323, 2013.

[102] Xiwen Jiang, Alexis Decurninge, Kalyana Gopala, Florian Kaltenberger, Maxime Guillaud, Dirk Slock, and Luc Deneire. A framework for over-the-air reciprocity calibration for tdd massive mimo systems. *IEEE Transactions on Wireless Communications*, 17(9):5975–5990, 2018.

[103] Li Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Process. Mag.*, 29(6):141–142, 2012.

[104] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proc. ACM Conf. Comput. Commun. Secur.*, pages 308–318, 2016.

[105] Jian Du, Song Li, Xiangyi Chen, Siheng Chen, and Mingyi Hong. Dynamic differential-privacy preserving SGD. *arXiv preprint arXiv:2111.00173*, 2021.

[106] Marten van Dijk, Phuong Ha Nguyen, Toan N Nguyen, and Lam M Nguyen. Generalizing DP-SGD with shuffling and batching clipping. *arXiv preprint arXiv:2212.05796*, 2022.

[107] Xinwei Zhang, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Jinfeng Yi. Understanding clipping for federated learning: Convergence and client-level differential privacy. In *Proc. International Conference on Machine Learning*, 2022.

[108] Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. *Proc. Advances in Neural Information Systems*, 34:17455–17466, 2021.

[109] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Proc. Advances in Neural Information Systems*, 31, 2018.

[110] Naman Agarwal, Ananda Theertha Suresh, Felix Xinnan X Yu, Sanjiv Kumar, and Brendan McMahan. cpSGD: Communication-efficient and differentially-private distributed SGD. *Proc. Advances in Neural Information Systems*, 31, 2018.

[111] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Trans. Inf. Forensics Secur.*, 15:3454–3469, 2020.

[112] Mohamed Seif, Ravi Tandon, and Ming Li. Wireless federated learning with local differential privacy. In *Proc. IEEE Int. Symp. Inf. Theory*, pages 2604–2609, 2020.

[113] Dongzhu Liu and Osvaldo Simeone. Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control. *IEEE J. Select. Areas Commun.*, 39(1):170–185, 2020.

[114] Mohamed Seif Eldin Mohamed, Wei-Ting Chang, and Ravi Tandon. Privacy amplification for federated learning via user sampling and wireless aggregation. *IEEE J. Select. Areas Commun.*, 39(12):3821–3835, 2021.

[115] Rui Hu, Yuanxiong Guo, Hongning Li, Qingqi Pei, and Yanmin Gong. Personalized federated learning with differential privacy. *IEEE Internet Things J.*, 7(10):9530–9539, 2020.

[116] Kang Wei, Jun Li, Chuan Ma, Ming Ding, Cailian Chen, Shi Jin, Zhu Han, and H Vincent Poor. Low-latency federated learning over wireless channels with differential privacy. *IEEE J. Select. Areas Commun.*, 40(1):290–307, 2021.

[117] Muah Kim, Onur Günlü, and Rafael F Schaefer. Federated learning with local differential privacy: Trade-offs between privacy, utility, and communication. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 2650–2654, 2021.

[118] Natesh S Pillai and Xiao-Li Meng. An unexpected encounter with Cauchy and Lévy. *Ann. Stat.*, 44(5):2089–2097, 2016.

[119] Ilya Mironov. Rényi differential privacy. In *Proc. IEEE Computer Security Foundations Symposium*, pages 263–275, 2017.

[120] Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled Gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

[121] Oded Goldreich. Secure multi-party computation. *Manuscript*, 1998.

[122] Wenliang Du and Mikhail J Atallah. Secure multi-party computation problems and their applications: a review and open problems. In *Proc. Workshop on New Security Paradigms*, pages 13–22, 2001.

[123] Whitfield Diffie and Martin E Hellman. New directions in cryptography. In *Democratizing Cryptography: The Work of Whitfield Diffie and Martin Hellman*, pages 365–390. ACM, 2022.

[124] Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private GLMs. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 2638–2646. PMLR, 2021.

[125] Xiangyi Chen, Steven Z Wu, and Mingyi Hong. Understanding gradient clipping in private SGD: A geometric perspective. *Proc. Advances in Neural Information Processing Systems*, 33:13773–13782, 2020.

[126] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Proc. Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[127] Tim Van Erven and Peter Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory*, 60(7):3797–3820, 2014.

Appendix

A Proof of Theorem 1

A.1 Preliminaries

With a slight abuse of notation, we change the timeline to be with respect to the overall SGD iteration time steps instead of the communication rounds, i.e.,

$$t = \underbrace{1, \cdots, E}_{\text{round } 1}, \underbrace{E+1, \cdots, 2E}_{\text{round } 2}, \cdots, \cdots, \underbrace{(T-1)E+1, \cdots, TE}_{\text{round } T}.$$

Note that the (noisy) global model $\mathbf{w}_t$ is only accessible at the clients for specific $t \in \mathcal{I}_E$, where $\mathcal{I}_E = \{nE \mid n = 1, 2, \dots\}$, i.e., the time steps for communication. The notations for η_t, σ_t and ζ_t are similarly adjusted to this extended timeline, but their values remain constant inside the same round.

As mentioned in Section 2.3.1, the key technique in the proof is the **perturbed iterate framework** in [58]. In particular, We first define the following variables

$$\mathbf{u}_{t+1}^k = \frac{1}{N} \sum_{i \in [N]} \mathbf{v}_{t+1}^i, \qquad \mathbf{p}_{t+1}^k = \mathbf{u}_{t+1}^k + \frac{1}{N} \sum_{i \in [N]} \mathbf{n}_{t+1}^i, \qquad \text{and } \ \mathbf{w}_{t+1}^k = \mathbf{p}_{t+1}^k + \mathbf{e}_{t+1}^k,$$

to summarize the aforementioned steps:

$$\mathbf{v}_{t+1}^k \triangleq \mathbf{w}_t^k - \eta_t \nabla F_k(\mathbf{w}_t^k, \xi_t^k);$$

$$\mathbf{u}_{t+1}^k \triangleq \begin{cases} \mathbf{v}_{t+1}^k & \text{if } t+1 \notin \mathcal{I}_E, \\ \frac{1}{N}\sum_{i\in[N]} \mathbf{v}_{t+1}^i & \text{if } t+1 \in \mathcal{I}_E; \end{cases}$$

$$\mathbf{p}_{t+1}^k \triangleq \begin{cases} \mathbf{v}_{t+1}^k & \text{if } t+1 \notin \mathcal{I}_E, \\ \mathbf{u}_{t+1}^k + \frac{1}{N}\sum_{i\in[N]} \mathbf{n}_{t+1}^i & \text{if } t+1 \in \mathcal{I}_E. \end{cases}$$

$$\mathbf{w}_{t+1}^k \triangleq \begin{cases} \mathbf{v}_{t+1}^k & \text{if } t+1 \notin \mathcal{I}_E, \\ \mathbf{p}_{t+1}^k + \mathbf{e}_{t+1}^k & \text{if } t+1 \in \mathcal{I}_E. \end{cases}$$

Then, we construct the following *virtual sequences*:

$$\overline{\mathbf{v}}_t = \frac{1}{N}\sum_{k=1}^N \mathbf{v}_t^k, \qquad \overline{\mathbf{u}}_t = \frac{1}{N}\sum_{k=1}^N \mathbf{u}_t^k, \qquad \overline{\mathbf{p}}_t = \frac{1}{N}\sum_{k=1}^N \mathbf{p}_t^k, \qquad \text{and } \overline{\mathbf{w}}_t = \frac{1}{N}\sum_{k=1}^N \mathbf{w}_t^k. \tag{1}$$

We also define $\overline{\mathbf{g}}_t = \frac{1}{N}\sum_{k=1}^N \nabla F_k(\mathbf{w}_t^k)$ and $\mathbf{g}_t = \frac{1}{N}\sum_{k=1}^N \nabla F_k(\mathbf{w}_t^k, \xi_t^k)$ for convenience. Therefore, $\overline{\mathbf{v}}_{t+1} = \overline{\mathbf{w}}_t - \eta_t \mathbf{g}_t$ and $\mathbb{E}\left[\mathbf{g}_t\right] = \overline{\mathbf{g}}_t$. After some manipulation, we can also write the specific formulations of these virtual sequences when $t+1 \in \mathcal{I}_E$ as follows:

$$\overline{\mathbf{u}}_{t+1} = \frac{1}{N}\sum_{i\in[N]} \mathbf{v}_{t+1}^i, \quad \overline{\mathbf{p}}_{t+1} = \overline{\mathbf{u}}_{t+1} + \frac{1}{N}\sum_{i\in[N]} \mathbf{n}_{t+1}^i, \quad \overline{\mathbf{w}}_{t+1} = \overline{\mathbf{p}}_{t+1} + \frac{1}{N}\sum_{k=1}^N \mathbf{e}_{t+1}^k. \tag{2}$$

Note that for $t+1 \notin \mathcal{I}_E$, all these virtual sequences are the same. In addition, the global model (at the server) $\overline{\mathbf{p}}_{t+1}$ is meaningful only at $t+1 \in \mathcal{I}_E$. We emphasize that when $t+1 \in \mathcal{I}_E$, Eqns. (2) and (2.10) indicate that $\overline{\mathbf{p}}_{t+1} = \mathbf{w}_{t+1}$. Thus it is sufficient to analyze the convergence of $\left\|\overline{\mathbf{p}}_{t+1} - \mathbf{w}^*\right\|^2$.

A.2 Lemmas

Lemma 2. *Let Assumption 1 hold, η_t is non-increasing, and $\eta_t \leq 2\eta_{t+E}$ for all $t \geq 0$. If $\eta_t \leq 1/(4L)$, we have*

$$\mathbb{E}\left\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\right\|^2 \leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t - \mathbf{w}^*\right\|^2 + \eta_t^2\left(\sum_{k=1}^N \delta_k^2/N^2 + 6L\Gamma + 8(E-1)^2H^2\right).$$

Lemma 2 establishes a bound for the one-step SGD. This result only concerns the local model update and is not impacted by the noisy communication. The derivation is similar to the technique in [32].

Lemma 3. *We have*

$$\mathbb{E}\left[\overline{\mathbf{p}}_{t+1}\right] = \overline{\mathbf{u}}_{t+1}, \mathbb{E}\left\|\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{p}}_{t+1}\right\|^2 = \frac{d\sigma_{t+1}^2}{N^2}; \quad \mathbb{E}\left[\overline{\mathbf{w}}_{t+1}\right] = \overline{\mathbf{p}}_{t+1}, \mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{p}}_{t+1}\right\|^2 = \frac{d\zeta_{t+1}^2}{N^2} \tag{3}$$

for $t+1 \in \mathcal{I}_E$*, where* $\sigma_{t+1}^2 \triangleq \sum_{k\in[N]} \sigma_{t+1,k}^2$ *and* $\zeta_{t+1}^2 \triangleq \sum_{k\in[N]} \zeta_{t+1,k}^2$*.*

Proof. For $t+1 \in \mathcal{I}_E$, we have

$$\mathbb{E}\left[\overline{\mathbf{p}}_{t+1} - \overline{\mathbf{u}}_{t+1}\right] = \frac{1}{K}\sum_{k\in[N]} \mathbb{E}[\mathbf{n}_{t+1}^k] = 0$$

and

$$\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1} - \overline{\mathbf{u}}_{t+1}\right\|^2 = \frac{1}{N^2}\mathbb{E}\left\|\sum_{k\in[N]} \mathbf{n}_{t+1}^k\right\|^2 = \frac{1}{N^2}\sum_{k\in[N]} \mathbb{E}\left\|\mathbf{n}_{t+1}^k\right\|^2 = \frac{d\sigma_{t+1}^2}{N^2}$$

from (2), because $\{\mathbf{n}_{t+1}^k, \forall k\}$ are independent variables. Similarly, according to (2), we have

$$\mathbb{E}\left[\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{p}}_{t+1}\right] = \frac{1}{N}\sum_{k\in[N]} \mathbb{E}[\mathbf{e}_{t+1}^k] = 0$$

and

$$\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{p}}_{t+1}\right\|^2 = \frac{1}{N^2}\mathbb{E}\left\|\sum_{k\in[N]} \mathbf{e}_{t+1}^k\right\|^2 = \frac{1}{N^2}\sum_{k\in[N]} \mathbb{E}\left\|\mathbf{e}_{t+1}^k\right\|^2 = \frac{\sum_{k\in[N]} d\zeta_{t+1,k}^2}{N^2} = \frac{d\zeta_{t+1}^2}{N^2}.$$

□

A.3 Proof of Theorem

We need to consider four cases for the analysis of the convergence of $\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1} - \mathbf{w}^*\right\|^2$.

1) If $t \notin \mathcal{I}_E$ and $t+1 \notin \mathcal{I}_E$, $\overline{\mathbf{w}}_t = \overline{\mathbf{p}}_t$ and $\overline{\mathbf{v}}_{t+1} = \overline{\mathbf{p}}_{t+1}$. Using Lemma 2, we have:

$$\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1} - \mathbf{w}^*\right\|^2 = \mathbb{E}\left\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\right\|^2 \leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{p}}_t - \mathbf{w}^*\right\|^2 + \eta_t^2\left[\sum_{k=1}^N \frac{\delta_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2\right].$$

2) If $t \in \mathcal{I}_E$ and $t+1 \notin \mathcal{I}_E$, we still have $\overline{\mathbf{v}}_{t+1} = \overline{\mathbf{p}}_{t+1}$. With $\overline{\mathbf{w}}_t = \overline{\mathbf{p}}_t + \frac{1}{N}\sum_{k=1}^N \mathbf{e}_t^k$, we have:

$$\left\|\overline{\mathbf{w}}_t - \mathbf{w}^*\right\|^2 = \left\|\overline{\mathbf{w}}_t - \overline{\mathbf{p}}_t + \overline{\mathbf{p}}_t - \mathbf{w}^*\right\|^2 = \left\|\overline{\mathbf{p}}_t - \mathbf{w}^*\right\|^2 + \underbrace{\left\|\overline{\mathbf{w}}_t - \overline{\mathbf{p}}_t\right\|^2}_{A_1} + \underbrace{2\left\langle \overline{\mathbf{w}}_t - \overline{\mathbf{p}}_t, \overline{\mathbf{p}}_t - \mathbf{w}^*\right\rangle}_{A_2}.$$

We first note that the expectation of A_2 over the noise randomness is zero since we have $\mathbb{E}\left[\overline{\mathbf{w}}_t - \overline{\mathbf{p}}_t\right] = \mathbf{0}$ (from (3)). Second, the expectation of A_1 can be bounded using Lemma 3. We then have

$$\begin{aligned}
&\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1} - \mathbf{w}^*\right\|^2 = \mathbb{E}\left\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\right\|^2 \\
&\leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{p}}_t - \mathbf{w}^*\right\|^2 + (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t - \overline{\mathbf{p}}_t\right\|^2 + \eta_t^2\left[\sum_{k=1}^{N}\frac{\delta_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2\right] \\
&\leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{p}}_t - \mathbf{w}^*\right\|^2 + (1-\eta_t\mu)\frac{d\zeta_t^2}{N^2} + \eta_t^2\left[\sum_{k=1}^{N}\frac{\delta_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2\right]. \qquad (4)
\end{aligned}$$

3) If $t \notin \mathcal{I}_E$ and $t+1 \in \mathcal{I}_E$, then we still have $\overline{\mathbf{w}}_t = \overline{\mathbf{p}}_t$. For $t+1$, we need to evaluate the convergence of $\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1} - \mathbf{w}^*\right\|^2$. We have

$$\begin{aligned}
\left\|\overline{\mathbf{p}}_{t+1} - \mathbf{w}^*\right\|^2 &= \left\|\overline{\mathbf{p}}_{t+1} - \overline{\mathbf{u}}_{t+1} + \overline{\mathbf{u}}_{t+1} - \mathbf{w}^*\right\|^2 \\
&= \underbrace{\left\|\overline{\mathbf{p}}_{t+1} - \overline{\mathbf{u}}_{t+1}\right\|^2}_{B_1} + \underbrace{\left\|\overline{\mathbf{u}}_{t+1} - \mathbf{w}^*\right\|^2}_{B_2} + \underbrace{2\left\langle\overline{\mathbf{p}}_{t+1} - \overline{\mathbf{u}}_{t+1}, \overline{\mathbf{u}}_{t+1} - \mathbf{w}^*\right\rangle}_{B_3}. \qquad (5)
\end{aligned}$$

We first note that the expectation of B_3 over the noise is zero since we have $\mathbb{E}\left[\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{p}}_{t+1}\right] = \mathbf{0}$ (from (3)). Second, the expectation of B_1 can be bounded using Lemma 3. Noticing that $\overline{\mathbf{u}}_{t+1} = \overline{\mathbf{v}}_{t+1}$ for B_2 and applying Lemma 2, we have:

$$\begin{aligned}
&\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1} - \mathbf{w}^*\right\|^2 \leq \mathbb{E}\left\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\right\|^2 + \frac{d\sigma_{t+1}^2}{K^2} \\
&\leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{p}}_t - \mathbf{w}^*\right\|^2 + \frac{d\sigma_{t+1}^2}{N^2} + \eta_t^2\left[\sum_{k=1}^{N}\frac{\delta_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2\right]. \qquad (6)
\end{aligned}$$

4) If $t \in \mathcal{I}_E$ and $t+1 \in \mathcal{I}_E$, $\overline{\mathbf{v}}_{t+1} \neq \overline{\mathbf{p}}_{t+1}$ and $\overline{\mathbf{w}}_t \neq \overline{\mathbf{p}}_t$. (Note that this is possible only for $E = 1$.) Combining the results from the previous two cases, we have

$$\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1} - \mathbf{w}^*\right\|^2 \leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{p}}_t - \mathbf{w}^*\right\|^2 + (1-\eta_t\mu)\frac{d\zeta_t^2}{N^2} + \frac{d\sigma_{t+1}^2}{N^2} + \eta_t^2\left[\sum_{k=1}^{N}\frac{\delta_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2\right]. \qquad (7)$$

Finally, we have that inequality (7) holds for all four cases. Denote $\Delta_t = \mathbb{E}\left\|\overline{\mathbf{p}}_t - \mathbf{w}^*\right\|^2$. If we set the effective noise power σ_{t+1}^2 and ζ_t^2 such that $\sigma_{t+1}^2 \leq N^2\eta_t^2$ and $\zeta_t^2 \leq N^2\frac{\eta_t^2}{1-\eta_t\mu}$, we always have $\Delta_{t+1} \leq (1-\eta_t\mu)\Delta_t + \eta_t^2 D$, where $D = \sum_{k=1}^{N}\frac{\delta_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2 + 2d$. We decay the learning rate as $\eta_t = \frac{\beta}{t+\gamma}$ for some $\beta \geq \frac{1}{\mu}$ and $\gamma \geq 0$ such that $\eta_1 \leq \min\{\frac{1}{\mu}, \frac{1}{4L}\} = \frac{1}{4L}$ and $\eta_t \leq 2\eta_{t+E}$. Now we prove that $\Delta_t \leq \frac{v}{\gamma+t}$ where $v = \max\{\frac{\beta^2 D}{\beta\mu-1}, (\gamma+1)\Delta_0\}$ by induction. First, the definition of v ensures that it holds for $t = 0$. Assume the conclusion holds for some $t > 0$. It then follows that

$$\Delta_{t+1} \leq (1-\eta\mu)\Delta_t + \eta_t^2 D = \left(1 - \frac{\beta\mu}{t+\gamma}\right)\frac{v}{t+\gamma} + \frac{\beta^2 D}{(t+\gamma)^2}$$
$$= \frac{t+\gamma-1}{(t+\gamma)^2}v + \left[\frac{\beta^2 D}{(t+\gamma)^2} - \frac{\mu\beta-1}{(t+\gamma)^2}v\right] \leq \frac{v}{t+\gamma+1}.$$

Then by the strong convexity of $F(\cdot)$, $\mathbb{E}\left[F(\overline{\mathbf{w}}_t)\right] - F^* \leq \frac{L}{2}\Delta_t \leq \frac{L}{2}\frac{v}{\gamma+t}$. Specially, if we choose $\beta = \frac{2}{\mu}$, $\gamma = \max\{8\frac{L}{\mu} - 1, E\}$ and denote $\phi = \frac{L}{\mu}$, then $\eta_t = \frac{2}{\mu}\frac{1}{\gamma+t}$. Using $\max\{a,b\} \leq a+b$, we have $v \leq \frac{4D}{\mu^2} + (\gamma+1)\Delta_0 \leq \frac{4D}{\mu^2} + (8\phi + E)\left\|\mathbf{w}_0 - \mathbf{w}^*\right\|^2$. Therefore, $\Delta_t \leq \frac{v}{\gamma+t} = \frac{1}{\gamma+t}\left[\frac{4D}{\mu^2} + (8\phi+E)\left\|\mathbf{w}_0 - \mathbf{w}^*\right\|^2\right]$. Setting $t = T$ concludes the proof.

B Proof of Theorem 2

The additional difficulty in proving Theorem 2 comes from partial clients participation. The approach we take is to study a "virtual" FL process where *all clients* receive the noisy downlink broadcast of the latest global model, and they all participate in the subsequent local model update phase. However, only the selected clients in $\mathcal{S}_{t+1}$ upload their updated local model to the server via the noisy uplink channel. It is clear that this "virtual" FL is equivalent to the original process in terms of the convergence – clients that are not selected do not contribute to the global model aggregation. This seemingly redundant process, however, circumvents the difficulty due to partial clients participation as can be seen in the analysis.

Before presenting the proof, we first elaborate on some necessary changes of notation. The notation defined in Appendix A.1 can be largely reused, with the notable distinction that now we have to separate the cases for K and for N. For $t+1 \in \mathcal{I}_E$, the variables of $\mathbf{u}_{t+1}^k$ and $\mathbf{p}_{t+1}^k$ are now defined as: $\mathbf{u}_{t+1}^k = \frac{1}{K}\sum_{i\in S_t}\mathbf{v}_{t+1}^i$ and $\mathbf{p}_{t+1}^k = \mathbf{u}_{t+1}^k + \frac{1}{K}\sum_{i\in\mathcal{S}_t}\mathbf{n}_{t+1}^i$. Note that Lemma 3 still holds with the following update: $\mathbb{E}\left\|\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{p}}_{t+1}\right\|^2 = \frac{d\bar{\sigma}_{t+1}^2}{K}$, $\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{p}}_{t+1}\right\|^2 = \frac{d\bar{\zeta}_{t+1}^2}{N}$. In addition, we need the following lemma that establishes the unbiased and variance-bounded client sampling..

Lemma 4. *Let Assumption $1-4$ hold. With $\eta_t \leq 2\eta_{t+E}$ for all $t \geq 0$ and $\forall t+1 \in \mathcal{I}_E$, we have $\mathbb{E}\left[\overline{\mathbf{u}}_{t+1}\right] = \overline{\mathbf{v}}_{t+1}$ and $\mathbb{E}\left\|\overline{\mathbf{v}}_{t+1} - \overline{\mathbf{u}}_{t+1}\right\|^2 \leq \frac{N-K}{N-1}\frac{4}{K}\eta_t^2 E^2 H^2$.*

Proof. Let $\mathcal{S}_{t+1}$ denote the set of chosen indexes. Note that the number of possible $\mathcal{S}_{t+1}$ is C_N^K and we denote the lth possible result as $\mathcal{S}_{t+1}^l = \{i_1^l, \ldots, i_K^l\}$, where $l = 1, \ldots, C_N^K$. Therefore,

$$\sum_{j=1}^{C_N^K} \sum_{k=1}^{K} \mathbf{v}_{t+1}^{i_k^l} = \frac{K \cdot C_N^K}{N} \sum_{i=1}^{N} \mathbf{v}_{t+1}^k = C_{N-1}^{K-1} \sum_{i=1}^{N} \mathbf{v}_{t+1}^k.$$

Since when $t+1 \in \mathcal{I}_E$,

$$\mathbf{u}_{t+1}^k = \frac{1}{K} \sum_{k \in St+1} \mathbf{v}_{t+1}^k$$

for all k, we have

$$\overline{\mathbf{u}}_{t+1} = \sum_{k=1}^{N} \mathbf{u}_{t+1}^k = \frac{1}{K} \sum_{k \in S_{t+1}} \mathbf{v}_{t+1}^k.$$

Then

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_t}\left[\overline{\mathbf{u}}_{t+1}\right] = \sum_{l=1}^{C_N^K} \mathbb{P}\left(\mathcal{S}_{t+1} = \mathcal{S}_{t+1}^l\right) \frac{1}{K} \sum_{k \in S_{t+1}^l} \mathbf{v}_{t+1}^k = \frac{1}{C_N^K} \frac{1}{K} \sum_{j=1}^{C_N^K} \sum_{k=1}^{K} \mathbf{v}_{t+1}^{i_k^l} = \frac{C_{N-1}^{K-1}}{C_N^K} \frac{1}{K} \sum_{k=1}^{N} \mathbf{v}_{t+1}^k \\
&= \frac{1}{N} \sum_{k=1}^{N} \mathbf{v}_{t+1}^k = \overline{\mathbf{v}}_{t+1}.
\end{aligned}$$

As for the variance, we have [30]

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_t}\left\|\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{v}}_{t+1}\right\|^2 = \mathbb{E}_{\mathcal{S}_t}\left\|\frac{1}{K} \sum_{i \in S_{t+1}} \mathbf{v}_{t+1}^i - \overline{\mathbf{v}}_{t+1}\right\| = \frac{1}{K^2} \mathbb{E}_{\mathcal{S}_t}\left\|\sum_{i=1}^{N} \mathbb{I}\left\{i \in S_t\right\}\left(\mathbf{v}_{t+1}^i - \overline{\mathbf{v}}_{t+1}\right)\right\|^2 \\
&= \frac{1}{K^2}\left[\sum_{i \in [N]} \mathbb{P}\left(i \in S_{t+1}\right)\left\|\mathbf{v}_{t+1}^i - \overline{\mathbf{v}}_{t+1}\right\|^2 + \sum_{i \neq j} \mathbb{P}\left(i, j \in S_{t+1}\right)\left\langle \mathbf{v}_{t+1}^i - \overline{\mathbf{v}}_{t+1}, \mathbf{v}_{t+1}^j - \overline{\mathbf{v}}_{t+1}\right\rangle\right] \\
&= \frac{1}{KN} \sum_{i=1}^{N}\left\|\mathbf{v}_{t+1}^i - \overline{\mathbf{v}}_{t+1}\right\|^2 + \sum_{i \neq j} \frac{K-1}{KN(N-1)}\left\langle \mathbf{v}_{t+1}^i - \overline{\mathbf{v}}_{t+1}, \mathbf{v}_{t+1}^j - \overline{\mathbf{v}}_{t+1}\right\rangle \\
&= \frac{1 - \frac{K}{N}}{K(N-1)} \sum_{i=1}^{N}\left\|\mathbf{v}_{t+1}^i - \overline{\mathbf{v}}_{t+1}\right\|^2
\end{aligned} \tag{8}$$

where we use the following results:

$$\mathbb{P}\left(i \in S_{t+1}\right) = \frac{K}{N}$$

and

$$\mathbb{P}\left(i, j \in S_{t+1}\right) = \frac{K(K-1)}{N(N-1)}$$

for all $i \neq j$, and

$$\sum_{i \in [N]} \left\| \mathbf{v}_{t+1}^i - \overline{\mathbf{v}}_{t+1} \right\|^2 + \sum_{i \neq j} \left\langle \mathbf{v}_{t+1}^i - \overline{\mathbf{v}}_{t+1}, \mathbf{v}_{t+1}^j - \overline{\mathbf{v}}_{t+1} \right\rangle = 0.$$

Since $t+1 \in \mathcal{I}_E$, we know that $t_0 = t - E + 1 \in \mathcal{I}_E$ is the communication time, implying that $\{\mathbf{u}_{t_0}^k\}_{k=1}^N$ are identical. Then

$$\begin{aligned}
&\sum_{i=1}^N \left\| \mathbf{v}_{t+1}^i - \overline{\mathbf{v}}_{t+1} \right\|^2 = \sum_{i=1}^N \left\| (\mathbf{v}_{t+1}^i - \overline{\mathbf{u}}_{t_0}) - (\overline{\mathbf{v}}_{t+1} - \overline{\mathbf{u}}_{t_0}) \right\|^2 \\
&= \sum_{i=1}^N \left\| \mathbf{v}_{t+1}^i - \overline{\mathbf{u}}_{t_0} \right\|^2 - 2 \left\langle \sum_{i=1}^N \mathbf{v}_{t+1}^i - \overline{\mathbf{u}}_{t_0}, \overline{\mathbf{v}}_{t+1} - \overline{\mathbf{u}}_{t_0} \right\rangle + \sum_{i=1}^N \left\| \overline{\mathbf{v}}_{t+1} - \overline{\mathbf{u}}_{t_0} \right\|^2 \\
&= \sum_{i=1}^N \left\| \mathbf{v}_{t+1}^i - \overline{\mathbf{u}}_{t_0} \right\|^2 - \sum_{i=1}^N \left\| \overline{\mathbf{v}}_{t+1} - \overline{\mathbf{u}}_{t_0} \right\|^2 \leq \sum_{i=1}^N \left\| \mathbf{v}_{t+1}^i - \overline{\mathbf{u}}_{t_0} \right\|^2
\end{aligned}$$

Taking expectation over the randomness of stochastic gradient on Eqn. (8), we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{K(N-1)} \left(1 - \frac{K}{N}\right) \sum_{k=1}^N \left\| \mathbf{v}_{t+1}^i - \overline{\mathbf{v}}_{t+1} \right\|^2 \right] \leq \frac{N-K}{K(N-1)} \frac{1}{N} \sum_{k=1}^N \mathbb{E} \left\| \mathbf{v}_{t+1}^i - \overline{\mathbf{u}}_{t_0} \right\|^2 \\
&\leq \frac{N-K}{K(N-1)} \frac{1}{N} \sum_{k=1}^N E \sum_{i=t_0}^t \mathbb{E} \left\| \eta_i \nabla F_k(\mathbf{u}_i^k, \xi_i^k) \right\|^2 \leq \frac{N-K}{K(N-1)} E^2 \eta_{t_0}^2 H^2 \leq \frac{N-K}{N-1} \frac{4}{K} E^2 \eta_t^2 H^2
\end{aligned}$$

where in the last line is because η_t is non-increasing and $\eta_{t_0} \leq 2\eta_t$. □

We can now similarly analyze the four cases as in Section A.3. Cases 1) and 2) remain the same as before. For Case 3) we need to consider $t \notin \mathcal{I}_E$ and $t+1 \in \mathcal{I}_E$. Note that (5) still holds, but we need to re-evaluate the expectation of B_2 because of partial clients participation. We have:

$$\begin{aligned}
\left\| \overline{\mathbf{u}}_{t+1} - \mathbf{w}^* \right\|^2 &= \left\| \overline{\mathbf{u}}_{t+1} - \overline{\mathbf{v}}_{t+1} + \overline{\mathbf{v}}_{t+1} - \mathbf{w}^* \right\|^2 \\
&= \underbrace{\left\| \overline{\mathbf{u}}_{t+1} - \overline{\mathbf{v}}_{t+1} \right\|^2}_{C_1} + \underbrace{\left\| \overline{\mathbf{v}}_{t+1} - \mathbf{w}^* \right\|^2}_{C_2} + \underbrace{2 \left\langle \overline{\mathbf{u}}_{t+1} - \overline{\mathbf{v}}_{t+1}, \overline{\mathbf{v}}_{t+1} - \mathbf{w}^* \right\rangle}_{C_3}.
\end{aligned} \tag{9}$$

When the expectation is taken over the random clients sampling, the expectation of C_3 is zero since we have $\mathbb{E}\left[\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{v}}_{t+1}\right] = \mathbf{0}$. The expectation of C_1 can be bounded using Lemma 4. Therefore we have $\mathbb{E}\left\| \overline{\mathbf{p}}_{t+1} - \mathbf{w}^* \right\|^2 \leq \mathbb{E}\left\| \overline{\mathbf{v}}_{t+1} - \mathbf{w}^* \right\|^2 + \frac{d\bar{\sigma}_{t+1}^2}{K} + \frac{N-K}{N-1}\frac{4}{K}\eta_t^2 E^2 H^2$. Using Lemma 2 and the new definition of D in Theorem 2, we have $\mathbb{E}\left\| \overline{\mathbf{p}}_{t+1} - \mathbf{w}^* \right\|^2 \leq \mathbb{E}\left\| \overline{\mathbf{v}}_{t+1} - \mathbf{w}^* \right\|^2 + \frac{d\bar{\sigma}_{t+1}^2}{K} + \frac{4\eta_t^2 E^2 H^2 (N-K)}{K(N-1)} \leq (1 - \eta_t \mu)\mathbb{E}\left\| \overline{\mathbf{p}}_t - \mathbf{w}^* \right\|^2 + \frac{d\bar{\sigma}_{t+1}^2}{K} + \eta_t^2 (D - 2d)$. Case 4) can be similarly updated based on the new result in Case 3). Finally, we have that

$$\mathbb{E}\left\| \overline{\mathbf{p}}_{t+1} - \mathbf{w}^* \right\|^2 \leq (1 - \eta_t \mu)\mathbb{E}\left\| \overline{\mathbf{p}}_t - \mathbf{w}^* \right\|^2 + (1 - \eta_t \mu)\frac{d\bar{\zeta}_t^2}{N} + \frac{d\bar{\sigma}_{t+1}^2}{K} + \eta_t^2 (D - 2d) \tag{10}$$

holds for all cases. If we set $\bar{\sigma}_{t+1}^2$ and $\bar{\zeta}_t^2$ such that $\bar{\sigma}_{t+1}^2 \leq K\eta_t^2$ and $\bar{\zeta}_t^2 \leq N\frac{\eta_t^2}{1-\eta_t\mu}$, the remaining proof follows the same way as in Appendix A.3.

C Proof of Theorem 3

For model differential transmission (MDT), if $t+1 \in \mathcal{I}_E$, the global aggregation is given in (2.11). Similar to Appendix A and B, we expand the timeline to be with respect to the overall SGD iteration time steps, and define the following variables to facilitate the proof. $\mathbf{v}_{t+1}^k \triangleq \mathbf{w}_t^k - \eta_t \nabla F_k(\mathbf{w}_t^k, \xi_t^k)$, $\mathbf{d}_{t+1}^k \triangleq \mathbf{v}_{t+1}^k - \mathbf{w}_{t+1-E}^k$. Furthermore, when $t+1 \notin \mathcal{I}_E$ we define $\mathbf{u}_{t+1}^k = \mathbf{p}_{t+1}^k = \mathbf{w}_{t+1}^k \triangleq \mathbf{v}_{t+1}^k$, and when $t+1 \in \mathcal{I}_E$ we define $\mathbf{u}_{t+1}^k \triangleq \frac{1}{K}\sum_{i\in \mathcal{S}_t} \mathbf{v}_{t+1}^i$, $\mathbf{p}_{t+1}^k \triangleq \mathbf{w}_{t+1-E} + \frac{1}{K}\sum_{i\in\mathcal{S}_t}[\mathbf{d}_{t+1}^i + \mathbf{n}_{t+1}^i]$, and $\mathbf{w}_{t+1}^k \triangleq \mathbf{p}_{t+1}^k + \mathbf{e}_{t+1}^k$. The virtual sequences $\overline{\mathbf{v}}_t$, $\overline{\mathbf{u}}_t$, $\overline{\mathbf{p}}_t$ and $\overline{\mathbf{w}}_t$ remain the same as (1). $\overline{\mathbf{g}}_t$ and $\mathbf{g}_t$ are also similarly defined. Note that the global model at the server is the same as $\overline{\mathbf{p}}_t$, i.e., $\mathbf{w}_{t+1} = \overline{\mathbf{p}}_{t+1}$.

We first establish the follow in lemma, which is instrumental in the proof of Theorem 3.

Lemma 5. *Let Assumption* $1-4$ *hold. Assume that* $\eta_t \leq 2\eta_{t+E}$ *for all* $t \geq 0$*, and further assume that the uplink communication adopts a constant SNR control policy:* $\mathsf{SNR}_{t,k}^{S,MDT} = \nu$*. Then, for* $t+1 \in \mathcal{I}_E$*, we have:* $\mathbb{E}\left[\overline{\mathbf{p}}_{t+1}\right] = \overline{\mathbf{u}}_{t+1}$ *and* $\mathbb{E}\left\|\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{p}}_{t+1}\right\|^2 \leq \left(1+\frac{1}{\nu}\right)\frac{d}{K}\bar{\zeta}_{t+1-E}^2 + \frac{4E^2}{K\nu}\eta_t^2 H^2$.

Proof. Note that if $t+1 \in \mathcal{I}_E$, so does $t+1-E$. Insert $\mathbf{d}_{t+1}^k = \mathbf{v}_{t+1}^k - \mathbf{w}_{t+1-E}^k$ into $\mathbf{p}_{t+1}^k$, we have $\mathbb{E}\left[\overline{\mathbf{p}}_{t+1}\right] = \overline{\mathbf{u}}_{t+1} + \frac{1}{K}\mathbb{E}\left[\sum_{k\in\mathcal{S}_t}\mathbf{n}_{t+1}^k\right] - \frac{1}{K}\mathbb{E}\left[\sum_{k\in\mathcal{S}_t}\mathbf{e}_{t+1-E}^k\right] = \mathbb{E}\left[\overline{\mathbf{u}}_{t+1}\right]$. As for the variance, we have

$$\mathbb{E}\left\|\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{p}}_{t+1}\right\|^2 = \frac{1}{K^2}\mathbb{E}\left\|\sum_{k\in\mathcal{S}_t}\mathbf{n}_{t+1}^k\right\|^2 + \frac{1}{K^2}\mathbb{E}\left\|\sum_{k\in\mathcal{S}_t}\mathbf{e}_{t+1-E}^k\right\|^2 = \frac{1}{K^2\nu}\mathbb{E}\left\|\sum_{k\in\mathcal{S}_t}\mathbf{d}_{t+1}^k\right\|^2 + \frac{d\bar{\zeta}_{t+1-E}^2}{K} \tag{11}$$

where the last equality comes from the constant uplink SNR control, (2.9), and the assumption that each client has the same downlink noise power $\bar{\zeta}_t^2$, $\forall k \in [N]$. We further have

$$\begin{aligned} &\mathbb{E}\left\|\sum_{k\in\mathcal{S}_t}\mathbf{d}_{t+1}^k\right\|^2 = \mathbb{E}\left\|\sum_{k\in\mathcal{S}_t}(\mathbf{v}_{t+1}^k - \mathbf{w}_{t+1-E})\right\|^2 + dK\bar{\zeta}_{t+1-E}^2 \\ &\leq \mathbb{E}_{\mathcal{S}_t}\left[\sum_{k\in\mathcal{S}_t}\mathbb{E}_{\text{SG}}\left\|\sum_{\tau=t+1-E}^{t}\eta_\tau\nabla F_k(\mathbf{w}_\tau^k, \xi_\tau^k)\right\|^2\right] + dK\bar{\zeta}_{t+1-E}^2 \leq 4E^2K\eta_t^2H^2 + dK\bar{\zeta}_{t+1-E}^2 \end{aligned} \tag{12}$$

using the Cauchy-Schwarz inequality, Assumption $1-4$, and $\eta_{t+1-E} < \eta_{t-E} \leq 2\eta_t$. Plugging (12) back to (11) gives

$$\mathbb{E}\left\|\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{p}}_{t+1}\right\|^2 = \frac{1}{K^2\nu}\mathbb{E}\left\|\sum_{k\in\mathcal{S}_{t+1}} \mathbf{d}_{t+1}^k\right\|^2 + \frac{d\bar{\zeta}_{t+1-E}^2}{K} \leq \left(1+\frac{1}{\nu}\right)\frac{d}{K}\bar{\zeta}_{t+1-E}^2 + \frac{4E^2}{K\nu}\eta_t^2 H^2,$$

which completes the proof. □

We are now ready to present the proof of Theorem 3, which is similar to that of Theorem 2. In particular, the analysis of four cases in Section B still hold, with the only change that (10) is updated to (13) below using Lemma 5 and the new definition of D in Theorem 3.

$$\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1} - \mathbf{w}^*\right\|^2 \leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{p}}_t - \mathbf{w}^*\right\|^2 + (1-\eta_t\mu)\frac{d}{N}\bar{\zeta}_t^2 + \left(1+\frac{1}{\nu}\right)\frac{d}{K}\bar{\zeta}_t^2 + \eta_t^2(D-d). \tag{13}$$

We note that the constant uplink SNR control is already used in Lemma 5 and (13). Then, by the definition of $\Delta_t = \mathbb{E}\left\|\overline{\mathbf{p}}_t - \mathbf{w}^*\right\|^2$ and controlling the downlink SNR such that $\bar{\zeta}_t^2 \leq \frac{NK\eta_t^2}{(1-\eta_t\mu)K+\left(1+\frac{1}{\nu}\right)N}$, we have $\Delta_{t+1} \leq (1-\eta_t\mu)\Delta_t + \eta_t^2 D$. The remaining proof follows using the same induction method.

D Proof of Theorem 4

D.1 Preliminaries

We first define the following local training variables for client k: $\mathbf{v}_{t+1}^k \triangleq \mathbf{p}_t^k - \eta_t\nabla\tilde{f}_k(\mathbf{p}_t^k)$; when $t+1 \notin \mathcal{I}_E$, we have $\mathbf{v}_{t+1}^k = \mathbf{u}_{t+1}^k = \mathbf{w}_{t+1}^k = \mathbf{p}_{t+1}^k$; when $t+1 \in \mathcal{I}_E$, we have: $\mathbf{u}_{t+1}^k = \frac{1}{K}\sum_{i\in[K]}\mathbf{v}_{t+1}^i$, $\mathbf{w}_{t+1}^k = \frac{1}{K}\sum_{i\in[K]}\mathbf{h}_s^H\mathbf{h}_k(\mathbf{v}_{t+1}^i - \mathbf{w}_{t+1-E}) + \frac{1}{K}\mathbf{N}_{t+1}\mathbf{h}_s + \mathbf{w}_{t+1-E}$ and

$$\mathbf{p}_{t+1}^k \triangleq \begin{cases} \mathbf{w}_{t+1}^k + \tilde{\mathbf{z}}_{t+1}^k & \text{if } k\in[K], \\ \mathbf{w}_{t+1}^k & \text{if } k\notin[K]; \end{cases}$$

where $\mathbf{N}_{t+1} \triangleq \left[\mathbf{n}_1,\cdots,\mathbf{n}_i,\cdots,\mathbf{n}_d\right]^H \in \mathbb{C}^{d\times M}$ is the stack of uplink noise in (3.5), and

$$\tilde{\mathbf{z}}_{t+1}^k \triangleq \begin{cases} \sqrt{K}\left[\mathrm{Re}(z_1^k/g_1),\cdots,\mathrm{Re}(z_d^k/g_d)\right]^H \in \mathbb{C}^{d\times 1} & \text{if } k\in[K], \\ 0 & \text{otherwise}, \end{cases}$$

are the downlink noise in (3.12), respectively. Then, we construct the following *virtual sequences*: $\overline{\mathbf{v}}_t = \frac{1}{N}\sum_{k=1}^N\mathbf{u}_t^k$, $\overline{\mathbf{u}}_t = \frac{1}{N}\sum_{k=1}^N\mathbf{v}_t^k$, $\overline{\mathbf{w}}_t = \frac{1}{N}\sum_{k=1}^N\mathbf{w}_t^k$, and $\overline{\mathbf{p}}_t = \frac{1}{N}\sum_{k=1}^N\mathbf{p}_t^k$. We also define $\overline{\mathbf{g}}_t = \frac{1}{N}\sum_{k=1}^N\nabla f_k(\mathbf{w}_t^k)$ and $\mathbf{g}_t = \frac{1}{N}\sum_{k=1}^N\nabla\tilde{f}_k(\mathbf{w}_t^k)$ for convenience. Therefore, $\overline{\mathbf{v}}_{t+1} = \overline{\mathbf{w}}_t - \eta_t\mathbf{g}_t$ and $\mathbb{E}\left[\mathbf{g}_t\right] = \overline{\mathbf{g}}_t$. Note that the global model $\mathbf{w}_{t+1}$ is only meaningful when $t+1 \in \mathcal{I}_E$, hence we have $\mathbf{w}_{t+1} \triangleq \frac{1}{K}\sum_{k\in[K]} \mathbf{w}_{t+1}^k = \frac{1}{N}\sum_{k=1}^N \mathbf{w}_{t+1}^k = \overline{\mathbf{w}}_{t+1}$. Thus it is sufficient to analyze the convergence of $\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2$ to evaluate random orthogonalization.

D.2 Lemmas

We first establish the following lemmas that are useful in the proof of Theorem 4.

Lemma 6. *Let Assumptions 1-4 hold. With $\eta_t \leq 2\eta_{t+E}$ for all $t \geq 0$ and $\forall t+1 \in \mathcal{I}_E$, we have $\mathbb{E}\left[\overline{\mathbf{w}}_{t+1}\right] = \overline{\mathbf{u}}_{t+1}$, and $\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{u}}_{t+1}\right\|^2 \leq \frac{4}{K}\left[\frac{K}{M} + \frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right]\eta_t^2 E^2 H^2$.*

Proof. We take expectation over randomness of fading channel and channel noise. As mentioned in Section 3.3, leveraging channel hardening and favorable propagation properties, we have

$$
\begin{aligned}
\mathbb{E}\left[\overline{\mathbf{w}}_{t+1}\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\mathbf{w}_{t+1}^k\right] = \mathbb{E}[\mathbf{w}_{t+1}^k] = \mathbb{E}\left[\frac{1}{K}\sum_{i\in[K]}\mathbf{h}_s^H\mathbf{h}_i(\mathbf{v}_{t+1}^i - \mathbf{w}_{t+1-E}) + \frac{1}{K}\mathbf{N}_{t+1}\mathbf{h}_s + \mathbf{w}_{t+1-E}\right] \\
&= \mathbb{E}\left[\frac{1}{K}\sum_{i\in[K]}\mathbf{h}_s^H\mathbf{h}_i(\mathbf{v}_{t+1}^i - \mathbf{w}_{t+1-E})\right] + \mathbb{E}\left[\frac{1}{K}\mathbf{N}_{t+1}\mathbf{h}_s\right] + \mathbb{E}\left[\mathbf{w}_{t+1-E}\right] \\
&= \frac{1}{K}\sum_{i\in[K]}\mathbb{E}\left[\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_i(\mathbf{v}_{t+1}^i - \mathbf{w}_{t+1-E})\right] + \mathbf{w}_{t+1-E} \\
&= \frac{1}{K}\sum_{i\in[K]}\mathbb{E}\left[\mathbf{h}_i^H\mathbf{h}_i(\mathbf{v}_{t+1}^i - \mathbf{w}_{t+1-E})\right] + \frac{1}{K}\sum_{i\in[K]}\mathbb{E}\left[\sum_{k\in[K],k\neq i}\mathbf{h}_k^H\mathbf{h}_i(\mathbf{v}_{t+1}^i - \mathbf{w}_{t+1-E})\right] + \mathbf{w}_{t+1-E} \\
&= \frac{1}{K}\sum_{i\in[K]}(\mathbf{v}_{t+1}^i - \mathbf{w}_{t+1-E}) + \mathbf{w}_{t+1-E} = \frac{1}{K}\sum_{i\in[K]}\mathbf{v}_{t+1}^i = \overline{\mathbf{u}}_{t+1}.
\end{aligned}
$$

We next evaluate the variance of $\overline{\mathbf{w}}_{t+1}$. Based on the facts that $\mathbb{E}[\mathbf{h}_i^H\mathbf{h}_i] = 1$, and $\forall i \neq j$, we have $\mathbb{E}[\mathbf{h}_i^H\mathbf{h}_j] = 0$, $\mathbb{V}\mathrm{ar}[\mathbf{h}_i^H\mathbf{h}_j] = \frac{1}{M}$, and $\mathbf{x}_i$ and $\mathbf{x}_j$ are independent, we have

$$
\begin{aligned}
&\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{u}}_{t+1}\right\|^2 = \mathbb{E}\left\|\frac{1}{K}\sum_{i\in[K]}\mathbf{h}_s^H\mathbf{h}_i(\mathbf{v}_{t+1}^i - \mathbf{w}_{t+1-E}) + \frac{1}{K}\mathbf{N}_{t+1}\mathbf{h}_s + \mathbf{w}_{t+1-E} - \frac{1}{K}\sum_{i\in[K]}\mathbf{v}_{t+1}^i\right\|^2 \\
&= \mathbb{E}\left\|\frac{1}{K}\sum_{k\in[K]}\hat{\mathbf{x}}_k - \frac{1}{K}\sum_{k\in[K]}\mathbf{x}_k\right\|^2 = \frac{1}{K^2}\mathbb{E}\left\|\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_k\mathbf{x}_k + \sum_{k\in[K]}\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j\mathbf{x}_j + \mathbf{N}_{t+1}\sum_{k\in[K]}\mathbf{h}_k - \sum_{k\in[K]}\mathbf{x}_k\right\|^2 \\
&= \frac{1}{K^2}\Bigg[\mathbb{E}\left\|\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_k\mathbf{x}_k\right\|^2 + \mathbb{E}\left\|\sum_{k\in[K]}\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j\mathbf{x}_j\right\|^2 + \mathbb{E}\left\|\mathbf{N}_{t+1}\sum_{k\in[K]}\mathbf{h}_k\right\|^2 + \mathbb{E}\left\|\sum_{k\in[K]}\mathbf{x}_k\right\|^2 \\
&+ 2\mathbb{E}\left[\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_k\mathbf{x}_k\sum_{k\in[K]}\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j\mathbf{x}_j\right] + 2\mathbb{E}\left[\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_k\mathbf{x}_k\mathbf{N}_{t+1}\sum_{k\in[K]}\mathbf{h}_k\right]
\end{aligned}
$$

$$
\begin{aligned}
&- 2\mathbb{E}\left[\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{h}_k \mathbf{x}_k \sum_{k\in[K]} \mathbf{x}_k\right] + 2\mathbb{E}\left[\sum_{k\in[K]}\sum_{j\in[K],j\neq k} \mathbf{h}_k^H \mathbf{h}_j \mathbf{x}_j \mathbf{N}_{t+1} \sum_{k\in[K]} \mathbf{h}_k\right] \\
&- 2\mathbb{E}\left[\sum_{k\in[K]}\sum_{j\in[K],j\neq k} \mathbf{h}_k^H\mathbf{h}_j\mathbf{x}_j \sum_{k\in[K]}\mathbf{x}_k\right] - 2\mathbb{E}\left[\mathbf{N}_{t+1}\sum_{k\in[K]}\mathbf{h}_k\sum_{k\in[K]}\mathbf{x}_k\right]\Bigg]\Bigg] \\
&= \frac{1}{K^2}\left[\left(1+\frac{1}{M}\right)\sum_{k\in[K]}\mathbb{E}\left\|\mathbf{x}_k\right\|^2 + \frac{K-1}{M}\sum_{k\in[K]}\mathbb{E}\left\|\mathbf{x}_k\right\|^2 + \frac{dK}{\mathsf{SNR}_{\mathsf{UL}}} + \sum_{k\in[K]}\mathbb{E}\left\|\mathbf{x}_k\right\|^2 - 2\sum_{k\in[K]}\mathbb{E}\left\|\mathbf{x}_k\right\|^2\right] \\
&= \frac{1}{K^2}\left[\frac{K}{M}\sum_{k\in[K]}\mathbb{E}\left\|\mathbf{x}_k\right\|^2 + \frac{\sum_{k\in[K]}\mathbb{E}\left\|\mathbf{x}_k\right\|^2}{\mathsf{SNR}_{\mathsf{UL}}}\right] = \frac{1}{K^2}\left[\frac{K}{M}+\frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right]\sum_{k\in[K]}\mathbb{E}\left\|\mathbf{x}_k\right\|^2 \\
&\le \frac{1}{K^2}\left[\frac{K}{M}+\frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right]\sum_{k\in[K]} E\sum_{i=t+1-E}^{t}\left\|\eta_i\nabla\tilde{f}_k(\mathbf{w}_i^k)\right\| \le \frac{1}{K}\left[\frac{K}{M}+\frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right]\eta_{t+1-E}^2E^2H^2 \\
&\le \frac{4}{K}\left[\frac{K}{M}+\frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right]\eta_t^2E^2H^2,
\end{aligned}
$$

where in the last inequality we use the fact that η_t is non-increasing and $\eta_{t+1-E} \le 2\eta_t$. □

Lemma 7. *Let Assumptions 1-4 hold and downlink SNR scales* $\mathsf{SNR}_{\mathsf{DL}} \ge \frac{1-\mu\eta_t}{\eta_t^2}$ *as learning round t.* $\forall t+1 \in \mathcal{I}_E$, *we have* $\mathbb{E}\left[\overline{\mathbf{p}}_{t+1}\right] = \overline{\mathbf{w}}_{t+1}$, *and* $\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1} - \overline{\mathbf{w}}_{t+1}\right\|^2 \le \left(\frac{dMK}{N^2(K+M)}\right)\frac{\eta_t^2}{1-\mu\eta_t}$.

Proof. We first show that

$$
\mathbb{E}\left[\mathrm{Re}\left(\frac{z_i^k}{g_k}\right)\right] = \mathrm{Re}\left(\mathbb{E}\left[z_i^k\right]\frac{1}{\mathbb{E}\left[g_k\right]}\right) = 0,
$$

and

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\mathrm{Re}\left(\frac{z_i^k}{g_k}\right)\right] &= \mathbb{E}\left[\mathrm{Re}\left(\frac{z_i^k}{g_k}\right)\mathrm{Re}\left(\frac{z_i^{k^*}}{g_{k^*}}\right)\right] = \mathbb{E}\left[\mathrm{Re}\left(\frac{z_i^k z_i^{k^*}}{g_k g_{k^*}}\right)\right] \le \frac{\mathbb{E}\left[\mathrm{Re}(z_i^{k^*} z_i^k)\right]}{\mathbb{E}\left[\mathrm{Re}(g_k^* g_k)\right]} \\
&= \frac{1/(2\mathsf{SNR}_{DL})}{1/2(1+K/M)} = \left(\frac{M}{K+M}\right)\frac{1}{\mathsf{SNR}_{\mathsf{DL}}},
\end{aligned}
$$

from which we can easily obtain $\mathbb{E}\left[\tilde{\mathbf{z}}_{t+1}^k\right] = \mathbf{0}$ and $\mathbb{V}\mathrm{ar}\left[\tilde{\mathbf{z}}_{t+1}^k\right] = \left(\frac{M}{K+M}\right)\frac{d}{\mathsf{SNR}_{\mathsf{DL}}}$. Therefore, we have

$$
\mathbb{E}\left[\overline{\mathbf{p}}_{t+1}\right] = \frac{1}{N}\sum_{k=1}^{N}\mathbf{w}_{t+1}^k + \frac{1}{N}\sum_{k\in[K]}\mathbb{E}\left[\tilde{\mathbf{z}}_{t+1}^k\right] = \overline{\mathbf{w}}_{t+1},
$$

and

$$
\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1} - \overline{\mathbf{w}}_{t+1}\right\|^2 = \mathbb{E}\left\|\frac{1}{N}\sum_{k\in[K]}\tilde{\mathbf{z}}_{t+1}^k\right\|^2 = \frac{1}{N^2}\sum_{k\in[K]}\mathbb{E}\left\|\tilde{\mathbf{z}}_{t+1}^k\right\|^2
$$

$$= \left(\frac{MK}{N^2(K+M)}\right)\frac{d}{\mathsf{SNR}_{\mathsf{DL}}} \leq \left(\frac{dMK}{N^2(K+M)}\right)\frac{\eta_t^2}{1-\mu\eta_t}.$$

□

D.3 Proof of Theorem 4

We need to consider four cases for the analysis of the convergence of $\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2$.

1) If $t \notin \mathcal{I}_E$ and $t+1 \notin \mathcal{I}_E$, $\overline{\mathbf{v}}_{t+1} = \overline{\mathbf{w}}_{t+1}$ and $\overline{\mathbf{p}}_t = \overline{\mathbf{w}}_t$. Using Lemma 2, we have:

$$\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1}-\mathbf{w}^*\right\|^2 = \mathbb{E}\left\|\overline{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|^2 \leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2 + \eta_t^2\mathbb{E}\left\|\mathbf{g}_t-\overline{\mathbf{g}}_t\right\|^2 + 6L\eta_t^2\Gamma$$
$$+ 2\mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\left\|\overline{\mathbf{w}}_t-\mathbf{w}_t^k\right\|^2\right] \leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{p}}_t-\mathbf{w}^*\right\|^2 + \eta_t^2\left[\sum_{k=1}^{N}\frac{\delta_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2\right].$$

2) If $t \in \mathcal{I}_E$ and $t+1 \notin \mathcal{I}_E$, we still have $\overline{\mathbf{v}}_{t+1} = \overline{\mathbf{w}}_{t+1}$. With $\overline{\mathbf{p}}_t = \overline{\mathbf{w}}_t + \frac{1}{N}\sum_{k=1}^{N}\tilde{\mathbf{z}}_t^k$, we have:

$$\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2 = \left\|\overline{\mathbf{p}}_t-\overline{\mathbf{w}}_t+\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2 = \left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2 + \underbrace{\left\|\overline{\mathbf{w}}_t-\overline{\mathbf{p}}_t\right\|^2}_{A_1} + \underbrace{2\left\langle\overline{\mathbf{w}}_t-\overline{\mathbf{p}}_t, \overline{\mathbf{p}}_t-\mathbf{w}^*\right\rangle}_{A_2}.$$

We first note that the expectation of A_2 over the noise and fading channel randomness is zero since we have $\mathbb{E}\left[\overline{\mathbf{w}}_t-\overline{\mathbf{p}}_t\right] = \mathbf{0}$. Second, the expectation of A_1 can be bounded using Lemma 7. We then have

$$\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2 = \mathbb{E}\left\|\overline{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|^2 \leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2 + (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t-\overline{\mathbf{p}}_t\right\|^2$$
$$+ \eta_t^2\left[\sum_{k=1}^{N}\frac{H_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2\right]$$
$$\leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2 + \eta_t^2\left[\sum_{k=1}^{N}\frac{H_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2 + \frac{MK}{N^2(K+M)}\right]. \tag{14}$$

3) If $t \notin \mathcal{I}_E$ and $t+1 \in \mathcal{I}_E$, then we still have $\overline{\mathbf{p}}_t = \overline{\mathbf{w}}_t$. For $t+1$, we need to evaluate the convergence of $\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2$. We have

$$\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2 = \left\|\overline{\mathbf{w}}_{t+1}-\overline{\mathbf{u}}_{t+1}+\overline{\mathbf{u}}_{t+1}-\mathbf{w}^*\right\|^2$$
$$= \underbrace{\left\|\overline{\mathbf{w}}_{t+1}-\overline{\mathbf{u}}_{t+1}\right\|^2}_{B_1} + \underbrace{\left\|\overline{\mathbf{u}}_{t+1}-\mathbf{w}^*\right\|^2}_{B_2} + \underbrace{2\left\langle\overline{\mathbf{w}}_{t+1}-\overline{\mathbf{u}}_{t+1}, \overline{\mathbf{u}}_{t+1}-\mathbf{w}^*\right\rangle}_{B_3}. \tag{15}$$

We first note that the expectation of B_3 over the noise is zero since we have $\mathbb{E}\left[\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{w}}_{t+1}\right] = \mathbf{0}$ and the expectation of B_1 can be bounded using Lemma 6. We next write B_2 into

$$\begin{aligned}\left\|\overline{\mathbf{u}}_{t+1} - \mathbf{w}^*\right\|^2 &= \left\|\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{v}}_{t+1} + \overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\right\|^2 \\ &= \underbrace{\left\|\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{v}}_{t+1}\right\|^2}_{C_1} + \underbrace{\left\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\right\|^2}_{C_2} + \underbrace{2\left\langle \overline{\mathbf{u}}_{t+1} - \overline{\mathbf{v}}_{t+1}, \overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\right\rangle}_{C_3}.\end{aligned} \tag{16}$$

Similarly, the expectation of C_3 over the noise is zero since we have $\mathbb{E}\left[\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{v}}_{t+1}\right] = \mathbf{0}$ and the expectation of C_1 can be bounded using Lemma 4. Therefore, we have

$$\begin{aligned}\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\right\|^2 &\leq \mathbb{E}\left\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\right\|^2 + \frac{4}{K}\left[\frac{K}{M} + \frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right]\eta_t^2 E^2 H^2 + \frac{N-K}{N-1}\frac{4}{K}\eta_t^2 E^2 H^2 \leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t - \mathbf{w}^*\right\|^2 \\ &+ \eta_t^2\left[\sum_{k=1}^N \frac{H_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2 + \frac{4}{K}\left(\frac{K}{M} + \frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right)E^2H^2 + \frac{N-K}{N-1}\frac{4}{K}E^2H^2\right].\end{aligned} \tag{17}$$

4) If $t \in \mathcal{I}_E$ and $t+1 \in \mathcal{I}_E$, $\overline{\mathbf{v}}_{t+1} \neq \overline{\mathbf{w}}_{t+1}$ and $\overline{\mathbf{p}}_t \neq \overline{\mathbf{w}}_t$. (Note that this is possible only for $E = 1$.) Combining the results from the previous two cases, we have

$$\begin{aligned}&\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\right\|^2 \leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t - \mathbf{w}^*\right\|^2 \\ &+ \left[\sum_{k=1}^N \frac{H_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2 + \frac{4}{K}\left(\frac{K}{M} + \frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right)E^2H^2 + \frac{N-K}{N-1}\frac{4}{K}E^2H^2 + \frac{MK}{N^2(K+M)}\right].\end{aligned} \tag{18}$$

Let $\Delta_t = \mathbb{E}\left\|\overline{\mathbf{w}}_t - \mathbf{w}^*\right\|^2$. From (F.2), (14), (17) and (18), it is clear that no matter whether $t+1 \in \mathcal{I}_E$ or $t+1 \notin \mathcal{I}_E$, we always have

$$\Delta_{t+1} \leq (1-\eta_t\mu)\Delta_t + \eta_t^2 B,$$

where

$$B = \sum_{k=1}^N \frac{H_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2 + \frac{4}{K}\left(\frac{K}{M} + \frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right)E^2H^2 + \frac{N-K}{N-1}\frac{4}{K}E^2H^2 + \frac{MK}{N^2(K+M)}.$$

Define $v \triangleq \max\{\frac{4B}{\mu^2}, (1+\gamma)\Delta_1\}$, by choosing $\eta_t = \frac{2}{\mu(t+\gamma)}$, we can prove $\Delta_t \leq \frac{v}{t+\gamma}$ by induction:

$$\begin{aligned}\Delta_{t+1} &\leq \left(1 - \frac{2}{t+\gamma}\right)\Delta_t + \frac{4B}{\mu^2(t+\gamma)^2} = \frac{t+\gamma-2}{(t+\gamma)^2}v + \frac{4B}{\mu^2(t+\gamma)^2} \\ &= \frac{t+\gamma-1}{(t+\gamma)^2}v + \left(\frac{4B}{\mu^2(t+\gamma)^2} - \frac{v}{(t+\gamma)^2}\right) \leq \frac{v}{t+\gamma+1}.\end{aligned}$$

By the L-smoothness of f and $v \leq \frac{4B}{\mu^2} + (1+\gamma)\Delta_1$, we can prove the result in (3.23).

E Proofs of Theorems 6, 7, and 8

E.1 Lemmas

We first establish the necessary lemmas for the proofs.

Lemma 8. *Let P and Q be two Cauchy probability distributions $\mathsf{Cauchy}(0,\gamma)$ and $\mathsf{Cauchy}(\theta,\gamma)$, with PDF $P(x) = \gamma/\pi(x^2+\gamma^2)$ and $Q(x) = \gamma/\pi[(x-\theta)^2+\gamma^2]$, respectively. $Q(x)/P(x) = (x^2+\gamma^2)/[(x-\theta)^2+\gamma^2]$ reaches its maximum at $x_{\max} = \frac{1}{2}(\theta + \sqrt{\theta^2+4\gamma^2})$ with $[Q(x)/P(x)]_{\max} = [\sqrt{\theta^2+4\gamma^2}+\theta]/[\sqrt{\theta^2+4\gamma^2}-\theta]$, and its minimum at $x_{\min} = \frac{1}{2}(\theta - \sqrt{\theta^2+4\gamma^2})$ with $[Q(x)/P(x)]_{\min} = [\sqrt{\theta^2+4\gamma^2}-\theta]/[\sqrt{\theta^2+4\gamma^2}+\theta]$.*

Proof. The results can be directly obtained by taking the first-order derivative of $\frac{Q(x)}{P(x)}$ and setting $\frac{\partial}{\partial x}\frac{Q(x)}{P(x)} = 0$. □

Lemma 9 (Proposition 3.3 in [126])**.** *Let P and Q be probability distributions satisfying $D_\infty(P||Q) \leq \epsilon$ and $D_\infty(Q||P) \leq \epsilon$. Then $D_\alpha(P||Q) \leq \frac{1}{2}\alpha\epsilon^2$ for all $\alpha > 1$.*

Lemma 10 (Proposition 3 in [119])**.** *If f is an (α,ϵ)-Rényi DP mechanism, it also satisfies $(\epsilon + \frac{\log(1/\delta)}{\alpha-1}, \delta)$-DP for any $0 < \delta < 1$.*

E.2 Proof of Theorem 6

Define neighboring datasets $\mathcal{D} = \bigcup_{i=1}^M \mathcal{D}_i$ and $\mathcal{D}' = \{\bigcup_{i=1, i\neq k}^M \mathcal{D}_i\} \cup \mathcal{D}'_k$, where $\mathcal{D}'_k \triangleq \mathcal{D}_k \cup \{z\}$ has one more element than $\mathcal{D}_k$. As we have assumed that all datasets $\{\mathcal{D}_i, i = 1, \cdots, M\}$ have the same size D, the size of $\mathcal{D}'_k$ is $D' \triangleq D+1$. We have $c = \begin{pmatrix} D \\ d_{\text{batch}} \end{pmatrix}$ and $c' = \begin{pmatrix} D' \\ d_{\text{batch}} \end{pmatrix}$ different mini-batches for dataset $\{\mathcal{D}_i, i = 1, \cdots, M\}$ and $\mathcal{D}'_k$, respectively. We further denote set $\Xi_i \triangleq \{\xi_{i,1}, \cdots, \xi_{i,c}\}$ as a collection of all the mini-batches corresponding to $\{\mathcal{D}_i, i = 1, \cdots, M\}$. We next calculate the number of mini-batches in dataset $\mathcal{D}'_k$. It is straightforward to verify that $c' = \begin{pmatrix} D+1 \\ d_{\text{batch}} \end{pmatrix} = \begin{pmatrix} D \\ d_{\text{batch}} \end{pmatrix} + \begin{pmatrix} D \\ d_{\text{batch}}-1 \end{pmatrix}$, which reveals that there are $\begin{pmatrix} D \\ d_{\text{batch}} \end{pmatrix}$ mini-batches in $\mathcal{D}'_k$ that are the same as those in Ξ_k, and the remaining $\begin{pmatrix} D \\ d_{\text{batch}}-1 \end{pmatrix}$ mini-batches are the ones that contain the data sample z. We denote $\Xi'_k = \{\xi'_{k,1}, \cdots, \xi'_{k,c'-c}\}$ as a collection of all the mini-batches that contain the data sample z. Therefore, all the possible mini-batches of $\mathcal{D}'_k$ are in $\Xi_k \cup \Xi'_k$.

We note that, in each learning round, K of total M clients are randomly selected. Therefore, there are total $\chi \triangleq \begin{pmatrix} M \\ K \end{pmatrix} c^K$ combinations of mini-batches in $\mathcal{D}$. We denote $\mathcal{S}_t = \{s_{t,1}, \cdots, s_{t,i}, \cdots, s_{t,K}\}$ as the index set for the K selected clients at t-th learning round, and it is easy to verify that we have $\begin{pmatrix} M \\ K \end{pmatrix}$ combinations for client selection per learning round. Therefore, we further denote each possible mini-batch combination as $\boldsymbol{\xi}_j \triangleq \{\xi_{s_{t,1},c_1}, \cdots, \xi_{s_{t,i},c_i}, \cdots, \xi_{s_{t,K},c_K}\}, j = 1, \cdots, \chi$, where $\xi_{s_{t,i},c_i} \in \Xi_{s_{t,i}}, \forall s_{t,i} \in \mathcal{S}_t$. As for $\mathcal{D}'$, there are total $\begin{pmatrix} M \\ K \end{pmatrix} c^K + \begin{pmatrix} M-1 \\ K-1 \end{pmatrix} (c'-c)c^{K-1}$ mini-batch combinations. Besides the same χ combinations in $\mathcal{D}$, there are additional $\chi' \triangleq \begin{pmatrix} M-1 \\ K-1 \end{pmatrix} (c'-c)c^{K-1}$ mini-batch combinations in $\mathcal{D}'$, which contains the mini-batch of the k-th client and is chosen from Ξ'_k. We denote them similarly as $\boldsymbol{\xi}'_j \triangleq \{\xi_{s_{t,1},c_1}, \cdots, \xi_{s_{t,i},c_i}, \cdots, \xi_{s_{t,K-1},c_{K-1}}, \xi'_{k,c'_k}\}, j = 1, \cdots \chi'$, where $\xi_{s_{t,i},c_i} \in \Xi_{s_{t,i}}, \forall s_{t,i} \in \mathcal{S}_t \backslash [k]$, and $\xi'_{k,c'_k} \in \Xi'_k$. In addition, we use $g(\boldsymbol{\xi}_j)$ to denote the noise-free global models calculated from mini-batch combination $\boldsymbol{\xi}_j$.

Building on these, the mechanism $\mathcal{M}(\mathcal{D})$ can be viewed as a random vector sampling from the following distribution:

$$\mathcal{M}(\mathcal{D}) = \sum_{j=1}^{\chi} \frac{1}{\chi} \mathsf{Cauchy}(g(\boldsymbol{\xi}_j), \gamma \mathbf{I}_d).$$

Similarly,

$$\mathcal{M}(\mathcal{D}') = \sum_{j=1}^{\chi} \frac{1}{\chi + \chi'} \mathsf{Cauchy}(g(\boldsymbol{\xi}_j), \gamma \mathbf{I}_d) + \sum_{j=1}^{\chi'} \frac{1}{\chi + \chi'} \mathsf{Cauchy}(g(\boldsymbol{\xi}'_j), \gamma \mathbf{I}_d). \tag{19}$$

We notice that $\frac{\chi'}{\chi} = \frac{K}{M} \frac{c'-c}{c}$ and $\frac{c'}{c} = 1 + \frac{d_{\text{batch}}}{D+1-d_{\text{batch}}}$. Defining $q \triangleq \frac{d_{\text{batch}}}{D+1-d_{\text{batch}}}$ and $p \triangleq \frac{K}{M}$, $\mathcal{M}(\mathcal{D}')$ can be re-written as

$$\begin{aligned}
\mathcal{M}(\mathcal{D}') =& \frac{1}{\chi + \chi'} \sum_{j=1}^{\chi} \Big[\mathsf{Cauchy}(g(\boldsymbol{\xi}_j), \gamma \mathbf{I}_d) + \frac{\chi'}{\chi} \mathsf{Cauchy}(g(\boldsymbol{\xi}'_{\lceil j * \frac{\chi'}{\chi} \rceil}), \gamma \mathbf{I}_d) \Big] \\
=& \frac{1}{\chi} \sum_{j=1}^{\chi} \Big[\frac{\chi}{\chi + \chi'} \mathsf{Cauchy}(g(\boldsymbol{\xi}_j), \gamma \mathbf{I}_d) + \frac{\chi}{\chi + \chi'} \frac{\chi'}{\chi} \mathsf{Cauchy}(g(\boldsymbol{\xi}'_{\lceil j * \frac{\chi'}{\chi} \rceil}, \gamma \mathbf{I}_d) \Big] \\
=& \sum_{j=1}^{\chi} \frac{1}{\chi} \left[\left(1 - \frac{pq}{1+pq} \right) \mathsf{Cauchy}(g(\boldsymbol{\xi}_j), \gamma \mathbf{I}_d) + \frac{pq}{1+pq} \mathsf{Cauchy}(\boldsymbol{\xi}'_{\lceil j * pq \rceil}, \gamma \mathbf{I}_d) \right],
\end{aligned}$$

where the first equality comes from replicating each of the χ' items in the second summation in Eqn. (19) $\frac{\chi'}{\chi}$ times, thus allowing the summation to be over the same range as the first summation.

We next bound the Rényi divergence for $\alpha = +\infty$. Since Rényi divergence is quasi-convex, we have

$$\begin{aligned}
&D_\infty(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}'))\\
\leq &\sup_j D_\infty\left[\mathsf{Cauchy}(g(\boldsymbol{\xi}_j),\gamma\mathbf{I}_d)||(1-\frac{pq}{1+pq})\mathsf{Cauchy}(g(\boldsymbol{\xi}_j),\gamma\mathbf{I}_d)+\frac{pq}{1+pq}\mathsf{Cauchy}(g(\boldsymbol{\xi}'_{\lceil j*pq\rceil}),\gamma\mathbf{I}_d)\right]\\
\leq &\sup_j D_\infty\left[\mathsf{Cauchy}(0,\gamma\mathbf{I}_d)||(1-\frac{pq}{1+pq})\mathsf{Cauchy}(0,\gamma\mathbf{I}_d)+\frac{pq}{1+pq}\mathsf{Cauchy}(g(\boldsymbol{\xi}'_{\lceil j*pq\rceil})-g(\boldsymbol{\xi}_j),\gamma\mathbf{I}_d)\right].
\end{aligned}$$

As shown in (4.6), $\left\|g(\boldsymbol{\xi}'_{\lceil j*pq\rceil})-g(\boldsymbol{\xi}_j)\right\|_2 \leq 2C$. Therefore, via a rotation, we have that $g(\boldsymbol{\xi}'_{\lceil j*pq\rceil})-g(\boldsymbol{\xi}_j) = c_\xi\mathbf{e}_1$ where $c_\xi \leq 2C$. By the additivity of Rényi divergence for product distributions [127], we have that

$$D_\infty(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}')) \leq D_\infty\left[\mathsf{Cauchy}(0,\gamma)||(1-\frac{pq}{1+pq})\mathsf{Cauchy}(0,\gamma)+\frac{pq}{1+pq}\mathsf{Cauchy}(2C,\gamma)\right].$$

We next bound $D_\infty(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}'))$ and $D_\infty(\mathcal{M}(\mathcal{D}')||\mathcal{M}(\mathcal{D}))$. Based on the results in Lemma 8, we have

$$\max\left[\frac{\mathsf{Cauchy}(2C,\gamma)}{\mathsf{Cauchy}(0,\gamma)}\right] = \frac{\sqrt{C^2+\gamma^2}+C}{\sqrt{C^2+\gamma^2}-C},$$

and

$$\min\left[\frac{\mathsf{Cauchy}(2C,\gamma)}{\mathsf{Cauchy}(0,\gamma)}\right] = \frac{\sqrt{C^2+\gamma^2}-C}{\sqrt{C^2+\gamma^2}+C}.$$

Therefore, $D_\infty(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}'))$ and $D_\infty(\mathcal{M}(\mathcal{D}')||\mathcal{M}(\mathcal{D}))$ can be bounded respectively as follows:

$$\begin{aligned}
&D_\infty\left[(1-\frac{pq}{1+pq})\mathsf{Cauchy}(0,\gamma)+\frac{pq}{1+pq}\mathsf{Cauchy}(2C,\gamma)||\mathsf{Cauchy}(0,\gamma)\right]\\
=&\sup\log\left(1-\frac{pq}{1+pq}+\frac{pq}{1+pq}\frac{\mathsf{Cauchy}(2C,\gamma)}{\mathsf{Cauchy}(0,\gamma)}\right)\\
\leq&\log\left(1+\frac{pq}{1+pq}\frac{2C\sqrt{C^2+\gamma^2}+2C^2}{\gamma^2}\right)
\end{aligned}$$

and

$$\begin{aligned}
&D_\infty\left[\mathsf{Cauchy}(0,\gamma)||(1-\frac{pq}{1+pq})\mathsf{Cauchy}(0,\gamma)+\frac{pq}{1+pq}\mathsf{Cauchy}(2C,\gamma)\right]\\
=&\sup\log\left(\frac{1}{1-\frac{pq}{1+pq}+\frac{pq}{1+pq}\frac{\mathsf{Cauchy}(2C,\gamma)}{\mathsf{Cauchy}(0,\gamma)}}\right)\leq\log\left(1+\frac{pq}{1+pq}\frac{2C\sqrt{C^2+\gamma^2}-2C^2}{\gamma^2}\right).
\end{aligned}$$

It is straightforward to verify that $D_\infty(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}')) \geq D_\infty(\mathcal{M}(\mathcal{D}')||\mathcal{M}(\mathcal{D})), \forall p, q,$ such that $0 < \frac{pq}{1+pq} < 1$. Therefore, both $D_\infty(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}'))$ and $D_\infty(\mathcal{M}(\mathcal{D}')||\mathcal{M}(\mathcal{D}))$ can be bounded by $\log\left(1+\frac{pq}{1+pq}\frac{2C\sqrt{C^2+\gamma^2}+2C^2}{\gamma^2}\right)$. Based on Lemma 9, we can guarantee $D_\alpha(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}')) \leq \frac{1}{2}\alpha\log^2\left(1+\frac{pq}{1+pq}\frac{2C\sqrt{C^2+\gamma^2}+2C^2}{\gamma^2}\right)$, which completes the proof.

E.3 Proof of Theorem 7

Based on Theorem 6 and the composition rule of Rényi DP as shown in Lemma 10, the Rényi DP guarantee for total T learning rounds is $\left[\frac{1}{2}T\alpha\log^2\left(1+\frac{pq}{1+pq}\frac{2C\sqrt{C^2+\gamma^2}+2C^2}{\gamma^2}\right), \alpha\right]$-Rényi DP. Therefore, the proposed design satisfies (ϵ', δ)-DP, where

$$\begin{aligned}
\epsilon' &= \min_{\alpha>1} \frac{1}{2}T\alpha\log^2\left(1+\frac{pq}{1+pq}\frac{2C\sqrt{C^2+\gamma^2}+2C^2}{\gamma^2}\right) + \frac{\log(1/\delta)}{\alpha-1} \\
&\geq \sqrt{2T\log(\frac{1}{\delta})}\log\left(1+\frac{pq}{1+pq}\frac{2C\sqrt{C^2+\gamma^2}+2C^2}{\gamma^2}\right) + \frac{1}{2}T\log^2\left(1+\frac{pq}{1+pq}\frac{2C\sqrt{C^2+\gamma^2}+2C^2}{\gamma^2}\right).
\end{aligned}$$

E.4 Proof of Theorem 8

Define neighboring datasets $\mathcal{D} = \bigcup_{i=1}^M \mathcal{D}_i$ and $\mathcal{D}' = \bigcup_{i=1}^M \mathcal{D}_i \cup \mathcal{D}'_k$, where $\mathcal{D}'_k$ is an arbitrary local dataset of a client k. For dataset $\mathcal{D}$, K of the total M clients are randomly scheduled for the task in each learning round. We denote $\mathcal{S}_j, \forall j = 1, \cdots, \psi$, where $\psi \triangleq \binom{M}{K}$, as the set of all the possible client combinations during a single learning round for dataset $\mathcal{D}$. For dataset $\mathcal{D}'$, beside the previous ψ combinations, there are additional $\psi' \triangleq \binom{M}{K-1}$ combinations that contain client k. We denote them as $\mathcal{S}'_j, \forall j = 1, \cdots, \psi'$. Therefore, the mechanism $\mathcal{M}(\mathcal{D})$ samples from the following distribution

$$\mathcal{M}(\mathcal{D}) = \frac{1}{\psi}\sum_{j=1}^{\psi}\mathsf{Cauchy}(g(\mathcal{S}_j), \gamma\mathbf{I}_d).$$

Similarly,

$$\begin{aligned}
\mathcal{M}(\mathcal{D}') &= \frac{1}{\psi+\psi'}\sum_{j=1}^{\psi}\mathsf{Cauchy}(g(\mathcal{S}_j), \gamma\mathbf{I}_d) + \frac{1}{\psi+\psi'}\sum_{j=1}^{\psi'}\mathsf{Cauchy}(g(\mathcal{S}'_j), \gamma\mathbf{I}_d) \\
&= \frac{1}{\psi}\left[\frac{\psi}{\psi+\psi'}\sum_{j=1}^{\psi}\mathsf{Cauchy}(g(\mathcal{S}_j), \gamma\mathbf{I}_d) + \frac{\psi}{\psi+\psi'}\frac{\psi'}{\psi}\mathsf{Cauchy}(g(\mathcal{S}'_j), \gamma\mathbf{I}_d)\right] \\
&= \frac{1}{\psi}\left[\left(1-\frac{K}{M+1}\right)\sum_{j=1}^{\psi}\mathsf{Cauchy}(g(\mathcal{S}_j), \gamma\mathbf{I}_d) + \frac{K}{M+1}\mathsf{Cauchy}(g(\mathcal{S}'_j), \gamma\mathbf{I}_d)\right].
\end{aligned}$$

By the definition of $p \triangleq \frac{K}{M+1}$ and following the similar techniques in Appendix E.2, $D_\infty(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}'))$ and $D_\infty(\mathcal{M}(\mathcal{D}')||\mathcal{M}(\mathcal{D}))$ can both be bounded by $\log\left(1+p\frac{2C\sqrt{C^2+\gamma^2}+2C^2}{\gamma^2}\right)$. Again, based on Lemma 9, we can guarantee $D_\alpha(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}')) \leq \frac{1}{2}\alpha\log^2\left(1+p\frac{2C\sqrt{C^2+\gamma^2}+2C^2}{\gamma^2}\right)$, which completes the single round DP guarantee. The proof of the client-level DP guarantee for the composition of T rounds follows the same as that in Appendix E.3.

F Proof of Theorem 9

With a slight abuse of notation, we change the timeline to be with respect to the overall SGD iteration time steps instead of the communication rounds, i.e.,

$$t = \underbrace{1,\cdots,E}_{\text{round 1}},\underbrace{E+1,\cdots,2E}_{\text{round 2}},\cdots,\cdots,\underbrace{(T-1)E+1,\cdots,TE}_{\text{round }T}.$$

Note that the global model $\mathbf{w}_t$ is only accessible at the clients for specific $t \in \mathcal{I}_E$, where $\mathcal{I}_E = \{nE \mid n = 1,2,\dots\}$, i.e., the time steps for communication. The notations for η_t are similarly adjusted to this extended timeline, but their values remain constant within the same round. The key technique in the proof is the *perturbed iterate framework* in [58]. In particular, we first define the following variables for client $k \in [M]$:

$$\mathbf{v}_{t+1}^k \triangleq \mathbf{w}_t^k - \eta_t \nabla \tilde{f}_k(\mathbf{w}_t^k);$$

$$\mathbf{u}_{t+1}^k \triangleq \begin{cases} \mathbf{v}_{t+1}^k & \text{if } t+1 \notin \mathcal{I}_E, \\ \frac{1}{K}\sum_{k=1}^K \mathbf{v}_{t+1}^i & \text{if } t+1 \in \mathcal{I}_E; \end{cases}$$

$$\mathbf{w}_{t+1}^k \triangleq \begin{cases} \mathbf{v}_{t+1}^k & \text{if } t+1 \notin \mathcal{I}_E, \\ \mathbf{u}_{t+1}^k + \frac{1}{K}\mathbf{n}_{t+1} & \text{if } t+1 \in \mathcal{I}_E; \end{cases}$$

where $\mathbf{n}_{t+1} \triangleq \frac{C_{\max,t}}{C}\left[\sum_{k=K+1}^N \frac{\mathbf{a}_k^T\mathbf{n}_1}{\mathbf{a}_k^T\mathbf{n}_s},\cdots,\sum_{k=K+1}^N \frac{\mathbf{a}_k^T\mathbf{n}_d}{\mathbf{a}_k^T\mathbf{n}_s}\right]^T \in \mathbb{C}^{d\times 1}$ is the effective noise vector after de-normalization. Note that $\mathbf{n}_{t+1}$ is a truncated Cauchy distribution vector with the following PDF:

$$f(n_{t+1,i}) = \frac{\gamma}{\left(n_{t+1,i}^2+\gamma^2\right)\left(\arctan\left[\frac{B-\sum_{k=1}^K x_k^i}{\gamma}\right]+\arctan\left[\frac{B+\sum_{k=1}^K x_k^i}{\gamma}\right]\right)},$$

where $n_{t+1,i} \in \left[-B - \sum_{k=1}^{K} x_k^i, B - \sum_{k=1}^{K} x_k^i\right], \forall i = 1, \cdots, d$. We further have

$$
\begin{aligned}
\mathbb{E}\left\|\frac{C_{\max,t}}{C} n_{t+1,i}\right\|^2 &= \frac{C_{\max,t}^2}{C^2}\left[\frac{2\gamma B}{\arctan\left[\frac{B-\sum_{k=1}^{K} x_k^i}{\gamma}\right] + \arctan\left[\frac{B+\sum_{k=1}^{K} x_k^i}{\gamma}\right]} - \gamma^2\right] \\
&\leq \frac{\gamma^2 C_{\max,t}^2}{C^2 \arctan\left(\frac{B+C}{\gamma}\right)}\left[\frac{B}{\gamma} - \arctan\left(\frac{B+C}{\gamma}\right)\right] \\
&\triangleq C_{\max,t}^2 D(\gamma).
\end{aligned}
$$

Then, we construct the following *virtual sequences*:

$$
\overline{\mathbf{v}}_t = \frac{1}{M}\sum_{k=1}^{M} \mathbf{v}_t^k, \quad \overline{\mathbf{u}}_t = \frac{1}{M}\sum_{k=1}^{M} \mathbf{u}_t^k, \quad \text{and} \quad \overline{\mathbf{w}}_t = \frac{1}{M}\sum_{k=1}^{M} \mathbf{w}_t^k.
$$

We also define $\overline{\mathbf{g}}_t = \frac{1}{M}\sum_{k=1}^{M} \nabla f_k(\mathbf{w}_t^k)$ and $\mathbf{g}_t = \frac{1}{M}\sum_{k=1}^{M} \nabla \tilde{f}_k(\mathbf{w}_t^k)$ for convenience. Therefore, $\overline{\mathbf{v}}_{t+1} = \overline{\mathbf{w}}_t - \eta_t \mathbf{g}_t$ and $\mathbb{E}\left[\mathbf{g}_t\right] = \overline{\mathbf{g}}_t$. Note that the global model $\mathbf{w}_{t+1}$ is only meaningful when $t+1 \in \mathcal{I}_E$. Hence, we have $\mathbf{w}_{t+1} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{w}_{t+1}^k = \mathbf{w}_{t+1}^k = \frac{1}{M}\sum_{k=1}^{K} \mathbf{w}_{t+1}^k = \overline{\mathbf{w}}_{t+1}$. Thus it is sufficient to analyze the convergence of $\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2$ to evaluate FLORAS.

F.1 Lemmas

Lemma 11. *Let Assumptions 1-4 hold. With $\eta_t \leq 2\eta_{t+E}$ for all $t \geq 0$ and $\forall t+1 \in \mathcal{I}_E$, we have*

$$
\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{u}}_{t+1}\right\|^2 \leq \frac{4dD(\gamma)}{K^2}\eta_t^2 E^2 H^2.
$$

Proof. As shown previously, $\mathbf{n}_{t+1}$ is a truncated Cauchy random vector. We have $\mathbb{E}[\mathbf{n}_{t+1}^H \mathbf{n}_{t+1}] \leq dC_{\max,t}^2 D(\gamma)$. Recall the definition that $C_{\max,t} \triangleq \max\{\left\|\mathbf{x}_t^k - \mu_k\right\|_2, \forall k\}$, which achieves its maximum at $k_{\max}$, and we use $\mu_{k_{\max}}$ to denote the element-wise mean of $\mathbf{x}_t^{k_{\max}}$. Since $\overline{\mathbf{w}}_{t+1} = \frac{1}{M}\sum_{k=1}^{M} \mathbf{w}_{t+1}^k = \overline{\mathbf{u}}_{t+1} + \frac{1}{K}\mathbf{n}_{t+1}$, we can bound $\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{u}}_{t+1}\right\|^2$ as

$$
\begin{aligned}
\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{u}}_{t+1}\right\|^2 = \mathbb{E}\left\|\frac{1}{K}\mathbf{n}_{t+1}\right\|^2 &\leq \mathbb{E}\left[\frac{dC_{\max,t}^2 D(\gamma)}{K^2}\right] = \frac{dD(\gamma)}{K^2}\mathbb{E}\left\|\mathbf{x}_t^{k_{\max}} - \mu_{k_{\max}}\right\|^2 \\
&\leq \frac{dD(\gamma)}{K^2}\mathbb{E}\left\|\mathbf{x}_t^{k_{\max}}\right\|^2 \leq \frac{dD(\gamma)}{K^2} E \sum_{i=t+1-E}^{t} \mathbb{E}\left\|\eta_i \nabla \tilde{f}_{k_{\max}}(\mathbf{w}_i^{k_{\max}})\right\|^2 \\
&\leq \frac{dD(\gamma)}{K^2}\eta_{t+1-E}^2 E^2 H^2 \leq \frac{4dD(\gamma)}{K^2}\eta_t^2 E^2 H^2,
\end{aligned}
$$

where in the last inequality we use the fact that η_t is non-increasing and $\eta_{t+1-E} \leq 2\eta_t$. □

F.2 Proof of Theorem 9

We next consider the convergence of $\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2$.

1) If $t+1 \notin \mathcal{I}_E$, $\overline{\mathbf{v}}_{t+1}=\overline{\mathbf{w}}_{t+1}$. Using Lemma 2, we have:

$$\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2=\mathbb{E}\left\|\overline{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|^2 \leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2+\eta_t^2\left[\sum_{k=1}^{M}\frac{H_k^2}{M^2}+6L\Gamma+8(E-1)^2H^2\right].$$

2) If $t+1 \in \mathcal{I}_E$, to evaluate the convergence of $\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2$, we establish

$$\begin{aligned}\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2 &= \left\|\overline{\mathbf{w}}_{t+1}-\overline{\mathbf{u}}_{t+1}+\overline{\mathbf{u}}_{t+1}-\mathbf{w}^*\right\|^2 \\ &= \underbrace{\left\|\overline{\mathbf{w}}_{t+1}-\overline{\mathbf{u}}_{t+1}\right\|^2}_{A_1}+\underbrace{\left\|\overline{\mathbf{u}}_{t+1}-\mathbf{w}^*\right\|^2}_{A_2}+\underbrace{2\left\langle\overline{\mathbf{w}}_{t+1}-\overline{\mathbf{u}}_{t+1},\overline{\mathbf{u}}_{t+1}-\mathbf{w}^*\right\rangle}_{A_3}.\end{aligned}$$

The expectation of A_1 can be bounded using Lemma 11. We then bound the expectation of A_3 by the Cauchy–Schwarz inequality:

$$\mathbb{E}\left[2\left\langle\overline{\mathbf{w}}_{t+1}-\overline{\mathbf{u}}_{t+1},\overline{\mathbf{u}}_{t+1}-\mathbf{w}^*\right\rangle\right] \leq \frac{2\sqrt{dD(\gamma)}}{K}\eta_t EH\mathbb{E}\left\|\overline{\mathbf{u}}_{t+1}-\mathbf{w}^*\right\|,$$

and it is now related to A_2. Finally, we can write A_2 as

$$\begin{aligned}\left\|\overline{\mathbf{u}}_{t+1}-\mathbf{w}^*\right\|^2 &= \left\|\overline{\mathbf{u}}_{t+1}-\overline{\mathbf{v}}_{t+1}+\overline{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|^2 \\ &= \underbrace{\left\|\overline{\mathbf{u}}_{t+1}-\overline{\mathbf{v}}_{t+1}\right\|^2}_{B_1}+\underbrace{\left\|\overline{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|^2}_{B_2}+\underbrace{2\left\langle\overline{\mathbf{u}}_{t+1}-\overline{\mathbf{v}}_{t+1},\overline{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\rangle}_{B_3}.\end{aligned}$$

Based on Lemma 4, the expectation of B_3 over random user selection is zero, since we have $\mathbb{E}\left[\overline{\mathbf{u}}_{t+1}-\overline{\mathbf{v}}_{t+1}\right]=\mathbf{0}$. The expectation of B_1 can be bounded also by Lemma 4. Therefore, we have

$$\begin{aligned}\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2 \leq & \left(1+\frac{2\sqrt{dD(\gamma)}}{K}\eta_t EH\right)\mathbb{E}\left\|\overline{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|^2 \\ &+\left(1+\frac{2\sqrt{dD(\gamma)}}{K}\eta_t EH\right)\frac{M-K}{M-1}\frac{4}{K}\eta_t^2E^2H^2+\frac{4dD(\gamma)}{K^2}\eta_t^2E^2H^2 \\ &\leq (1-\eta_t\mu')\mathbb{E}\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2+\eta_t^2\left[\left(1+\frac{2\sqrt{dD(\gamma)}}{K}\eta_t EH\right)\right. \\ &\times\left.\left(\sum_{k=1}^{M}\frac{H_k^2}{M^2}+6L\Gamma+8(E-1)^2H^2+\frac{M-K}{M-1}\frac{4}{K}E^2H^2\right)+\frac{4dD(\gamma)}{K^2}E^2H^2\right],\end{aligned}$$

where $\mu' \triangleq \mu - 2\sqrt{dD(\gamma)}EH/K$. Let $\Delta_t \triangleq \mathbb{E}\left\|\overline{\mathbf{w}}_t - \mathbf{w}^*\right\|^2$. No matter whether $t+1 \in \mathcal{I}_E$ or $t+1 \notin \mathcal{I}_E$, we always have

$$\Delta_{t+1} \leq (1-\eta_t\mu')\Delta_t + \eta_t^2 G,$$

where

$$G \triangleq \left(1 + \frac{2\sqrt{dD(\gamma)}}{K}EH\eta_1\right)\left(\sum_{k=1}^{M}\frac{H_k^2}{M^2} + 6L\Gamma + 8(E-1)^2H^2 + \frac{M-K}{M-1}\frac{4}{K}E^2H^2\right) + \frac{4dD(\gamma)}{K^2}E^2H^2.$$

Define $v \triangleq \max\{\frac{4G}{\mu'^2}, (1+r)\Delta_1\}$. By choosing $\eta_t = \frac{2}{\mu'(t+r)}$, we can prove $\Delta_t \leq \frac{v}{t+r}$ by induction:

$$\begin{aligned}\Delta_{t+1} &\leq \left(1 - \frac{2}{t+r}\right)\Delta_t + \frac{4G}{\mu'^2(t+r)^2} = \frac{t+r-2}{(t+r)^2}v + \frac{4G}{\mu'^2(t+r)^2}\\ &\leq \frac{t+r-1}{(t+r)^2}v + \frac{4G}{\mu'^2(t+r)^2} - \frac{v}{(t+r)^2} \leq \frac{v}{t+r+1}.\end{aligned}$$

By the L-smoothness of f and $v \leq \frac{4G}{\mu'^2} + (1+r)\Delta_1$, we prove Theorem 9.