Federated Learning Over Noisy Channels: Convergence Analysis and Design Examples

Xizixiang Wei and Cong Shen, *Senior Member, IEEE*

Abstract—Does Federated Learning (FL) work when *both* uplink and downlink communications have errors? How much communication noise can FL handle and what is its impact on the learning performance? This work is devoted to answering these practically important questions by explicitly incorporating both uplink and downlink noisy channels in the FL pipeline. We present several novel convergence analyses of FL over simultaneous uplink and downlink noisy communication channels, which encompass full and partial clients participation, direct model and model differential transmissions, and non-independent and identically distributed (IID) local datasets. These analyses characterize the sufficient conditions for FL over noisy channels to have the same convergence behavior as the ideal case of no communication error. More specifically, in order to maintain the $\mathcal{O}(1/T)$ convergence rate of FEDAVG with perfect communications, the uplink and downlink signal-to-noise ratio (SNR) for direct model transmissions should be controlled such that they scale as $\mathcal{O}(t^2)$ where t is the index of communication rounds, but can stay $\mathcal{O}(1)$ (i.e., constant) for model differential transmissions. The key insight of these theoretical results is a "flying under the radar" principle – stochastic gradient descent (SGD) is an inherent noisy process and uplink/downlink communication noises can be tolerated as long as they do not dominate the time-varying SGD noise. We exemplify these theoretical findings with two widely adopted communication techniques – transmit power control and receive diversity combining – and further validate their performance advantages over the standard methods via numerical experiments using several real-world FL tasks.

Index Terms—Federated learning, convergence analysis, noisy communications, power allocation optimization.

I. INTRODUCTION

FEDERATED learning (FL) [2], [3] 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 a massive amount of the real-world data is *exogenously* generated at the edge devices and is considered as one of the potential key applications in 6th generation (6G) cellular communication systems [4].

The authors are with the Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904 USA (e-mail: xw8cw@virginia.edu; cong@virginia.edu).

Communication efficiency has been at the front and center of FL ever since its inception [2], [3], and it is widely regarded as one of its primary bottlenecks [5]–[7]. 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. Early research has largely focused on either reducing the number of communication rounds [2], [8], or decreasing the size of the payload for transmission [9]–[11]. 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; see Section II for an overview. Nevertheless, the focus has been on bandwidth allocation, device selection, or either uplink or downlink (but not both) cellular system designs.

While the early studies provide a glimpse of the potential of optimizing the communication design for FL, the important and more practical issue of noisy communications for both uplink (clients send local models to the parameter server) and downlink (server sends the global model to clients) has not been well investigated. Analytically speaking, joint consideration of both noisy uplink and downlink complicates the convergence analysis because of *noise propagation* in both directions of every communication round. Furthermore, all of these noisy uplink and downlink communications collectively determine the final learning performance, which requires a holistic design and analysis.

The goal of this paper is two-fold: we want to first 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 then design communication algorithms to control the signal-to-noise ratio (SNR) to improve FL performance under a total resource budget. We focus on *analog* communications for model updates [12]–[14] and investigate SNR control in both uplink and downlink, which is especially crucial when the underlying ML method is stochastic gradient descent (SGD) as considered in this work, because SGD is much more sensitive to noise than the (full) gradient descent [6], [15].

Our treatment is novel because all prior works either study uplink-only [12], [13], [16]–[18] or downlink-only [19] noisy communications, but not both. We present novel convergence analyses of the standard Federated Averaging (FEDAVG) scheme 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. We carry out a series of experiments to demonstrate that the fine-tuned transmit power control and receive diversity combining that are guided by the theoretical analyses can significantly outperform the equal-SNR-over-time baseline, and in fact can approach the ideal noise-free communication performance in many of the experiment settings.

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

- We present novel convergence analyses for FL 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 introduces significant challenges because of the noise propagation in both directions.

- We establish *SNR scaling laws*. In particular, we prove that in order to maintain the well established $\mathcal{O}(1/T)$[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.

- We enhance the widely adopted transmit power control and receive diversity combining algorithms to better serve FL over noisy channels, and validate their performance advantages over the state-of-the-art methods under the same total resource budget via extensive numerical experiments.

The remainder of this paper is organized as follows. Related works are surveyed in Section II. The system model that captures the noisy channels in both uplink and downlink of FL is described in Section III. Theoretical analyses are presented in Section IV for three different FL configurations. These results inspire novel communication designs of transmit power control and receive diversity combining that are presented in Section V. Experimental results are given in Section VI, followed by the conclusions in Section VII. All technical proofs are given in the Appendices.

II. 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 [15], [20], [21]. For the approach of reducing the size of messages, general discussions on sparsification, subsampling, and quantization are given in [3]. There are also recent efforts in developing quantization and source coding to reduce the communication cost [9]–[12], [22]–[24]. 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. The trade-off between local model update and global model aggregation is studied in [25] to optimize the transmission power/rate and training time. Various radio resource allocation and client selection policies [26]–[31] have been proposed to minimize the learning loss or the training time. Joint communication and computation is investigated [13], [14], [22], [32]. In particular, the analog aggregation design [13], [22], [33] serves as one of our design examples in Section V.

FL With Imperfect/Noisy Communications: Existing literature is dominated by uplink-only noisy communications [12]–[14], [16]–[18], [34], [35]. There is very limited study on downlink-only noisy communications for FL; [19] 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, [36] changes the loss function of FL to accommodate the communication error. References [37]–[39] 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.

III. SYSTEM MODEL FOR LEARNING AND NOISY COMMUNICATION

We first introduce the FL problem formulation, and then describe the FL pipeline where both local model upload (uplink) and global model download (downlink) take place over noisy channels.

[1] Notation $f = \mathcal{O}(g)$ denotes f is of order at most of g.
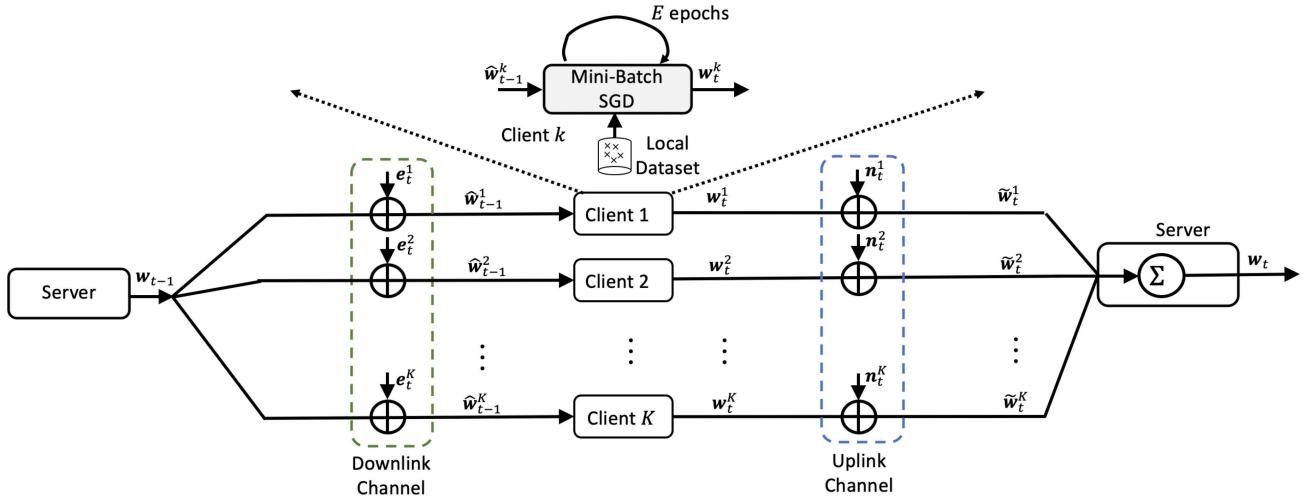
Fig. 1. End-to-end FL system diagram in the t-th communication round. The impact of noisy channels in both uplink and downlink is captured.

A. FL Problem Formulation

The federated learning problem setting studied in this paper mostly follows the standard model in the original paper [2]. In particular, we consider a FL system with one central parameter server (e.g., base station) and a set of at most N clients (e.g., mobile devices). Client $k \in [N] \triangleq \{1, 2, \ldots, N\}$ stores a local dataset $\mathcal{D}_k = \{z_i\}_{i=1}^{D_k}$, with its size denoted by D_k, that never leaves the client. Datasets across clients are assumed to be non-IID and disjoint. The maximum data size when all clients participate in FL is $D_{\text{tot}} = \sum_{k=1}^{N} D_k$. Each data sample z is given as an input-output pair $\{\mathbf{x}, \mathbf{y}\}$. The loss function $f(\mathbf{w}, z)$ measures how well a ML model with parameter $\mathbf{w} \in \mathbb{R}^d$ fits a single data sample z. Without loss of generality, we assume that $\mathbf{w}$ has zero-mean and unit-variance elements,[2] i.e., $\mathbb{E}\|w_i\|^2 = 1, \forall i \in [d]$. For the k-th client, its local loss function $F_k(\cdot)$ is defined by $F_k(\mathbf{w}) \triangleq \frac{1}{D_k} \sum_{\mathbf{z} \in \mathcal{D}_k} f(\mathbf{w}, z)$, and we further use $\nabla F_k(\mathbf{w}, \xi)$ to denote the SGD operation with model $\mathbf{w}$ and data sample ξ at client k. The goal of FL is to learn a global machine learning (ML) model at the parameter server based on the distributed local datasets at the N clients, by coordinating and aggregating the training processes at individual clients without allowing the server to access the raw data. Specifically, the global optimization objective over all N clients is given by $F(\mathbf{w}) \triangleq \sum_{k=1}^{N} \frac{D_k}{D_{\text{tot}}} F_k(\mathbf{w}) = \frac{1}{D_{\text{tot}}} \sum_{k=1}^{N} \sum_{\mathbf{z} \in \mathcal{D}_k} f(\mathbf{w}, z)$. The global loss function measures how well the model fits the entire corpus of data on average. The learning objective is to find the best model parameter $\mathbf{w}^*$ that minimizes the global loss function: $\mathbf{w}^* = \arg \min_{\mathbf{w}} F(\mathbf{w})$. Let F^* and F_k^* be the minimum value of F and F_k, respectively. Then, $\Gamma = F^* - \sum_{k=1}^{N} \frac{D_k}{D_{\text{tot}}} F_k^*$ quantifies the degree of non-IID as shown in [21].

B. FL Over Noisy Uplink and Downlink Channels

We study 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, we focus on imperfect communications and consider that *both* the upload and download transmissions take place over noisy communication channels. The overall system diagram is depicted in Fig. 1. In particular, the FL-over-noisy-channel pipeline works by iteratively executing the following steps at the t-th learning round, $\forall t \in [T]$.

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

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

where $\mathbf{e}_t^k = [e_{t,1}^k, \ldots, e_{t,d}^k]^T \in \mathbb{R}^d$ 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:

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

Effective Noise and Definition of SNR: In order to keep the problem general, we do not specify a particular communication system for the actual downlink data transmission, and only use the effective noise model in (1). The same approach applies to the uplink. This is a conscious choice to keep the problem general, and we want to focus on analyzing the impact

---

[2]The parameter normalization and de-normalization procedure in wireless FL can be found in the Appendix in [13]. We further note that *weight normalization* is widely adopted in training deep neural networks [40].

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

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

$$\mathrm{SNR}_{t,k}^{\mathrm{L}} = \frac{\mathbb{E}\|\mathbf{w}_{t-1}\|^2}{\mathbb{E}\|\mathbf{e}_t^k\|^2} = \frac{1}{\zeta_{t,k}^2}. \tag{3}$$

Lastly, we note that the noise assumption is very mild, because (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, $\{\zeta_{t,k}^2\}$, 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, 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:

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

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 for Local Model Upload: The K participating clients upload their latest local models to the server. More specifically, client k transmits a vector $\mathbf{x}_t^k$ to the server at the t-th round. We again consider the practical case 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

$$\hat{\mathbf{x}}_t^k = \mathbf{x}_t^k + \mathbf{n}_t^k, \tag{4}$$

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

---

[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 $\mathbf{y}_{t-1}^k = \sqrt{P_{t-1}}\mathbf{w}_{t-1} + \mathbf{z}_t^k$ where $\mathbf{z}_t^k$ represents the AWGN and P_{t-1} is the downlink broadcast transmit power. The effective channel noise becomes $\mathbf{e}_t^k = \frac{1}{\sqrt{P_{t-1}}}\mathbf{z}_t^k$.

$\mathbf{n}_t^k$ is a zero-mean random vector consisting of IID elements with bounded variance:

$$\mathbb{E}\|n_{t,i}^k\|^2 = \sigma_{t,k}^2 \quad \text{and} \quad \mathbb{E}\|\mathbf{n}_t^k\|^2 = d\sigma_{t,k}^2, \\ \forall t \in [T], k \in \mathcal{S}_t, i \in [d]. \tag{5}$$

We again note that the uplink communication model in (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 (4), the server receives the updated local model of client k as

$$\tilde{\mathbf{w}}_t^k = \hat{\mathbf{x}}_t^k = \mathbf{w}_t^k + \mathbf{n}_t^k. \tag{6}$$

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

$$\tilde{\mathbf{w}}_t^k = \mathbf{w}_{t-1} + \hat{\mathbf{x}}_t^k = \mathbf{w}_{t-1} + \mathbf{d}_t^k + \mathbf{n}_t^k = \mathbf{w}_t^k + \mathbf{n}_t^k - \mathbf{e}_t^k. \tag{7}$$

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 III-A. Thus, for MT, we can write the receive SNR at the server for k-th client's signal as

$$\mathrm{SNR}_{t,k}^{\mathrm{S,MT}} = \frac{\mathbb{E}\|\mathbf{w}_t^k\|^2}{\mathbb{E}\|\mathbf{n}_t^k\|^2} = \frac{1}{\sigma_{t,k}^2}. \tag{8}$$

For MDT, we keep the SNR expression general since the variance of model difference $\mathbf{d}_t^k$ is unknown *a priori* and also changes over time. We have:

$$\mathrm{SNR}_{t,k}^{\mathrm{S,MDT}} = \frac{\mathbb{E}\|\mathbf{d}_t^k\|^2}{\mathbb{E}\|\mathbf{n}_t^k\|^2} = \frac{\mathbb{E}\|\mathbf{d}_t^k\|^2}{d\sigma_{t,k}^2}. \tag{9}$$

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 [2]. 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 [5], which makes reconstruction from the model differential infeasible.

We also note that the *download* phase, on the other hand, does not have these two choices – we always transmit the global model $\mathbf{w}_{t-1}$ itself. This is because we have partial (and random) clients participation, where the set of clients participating in 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 ((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 [2]: $\mathbf{w}_t = \sum_{k \in \mathcal{S}_t} \frac{D_k}{\sum_{i \in \mathcal{S}_t} D_i} \tilde{\mathbf{w}}_t^k$. The server then moves on to the $(t+1)$-th round. For ease of exposition and to simplify 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

$$\mathbf{w}_t = \frac{1}{K} \sum_{k \in \mathcal{S}_t} \tilde{\mathbf{w}}_t^k = \frac{1}{K} \sum_{k \in \mathcal{S}_t} \hat{\mathbf{x}}_t^k = \frac{1}{K} \sum_{k \in \mathcal{S}_t} \left(\mathbf{w}_t^k + \mathbf{n}_t^k \right).$$
(10)

2) *MDT:* The aggregation can be written as

$$\mathbf{w}_t = \frac{1}{K} \sum_{k \in \mathcal{S}_t} \tilde{\mathbf{w}}_t^k = \mathbf{w}_{t-1} + \frac{1}{K} \sum_{k \in \mathcal{S}_t} \hat{\mathbf{x}}_t^k$$
$$= \frac{1}{K} \sum_{k \in \mathcal{S}_t} \left(\mathbf{w}_t^k + \mathbf{n}_t^k - \mathbf{e}_t^k \right).$$
(11)

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

$$\mathsf{SNR}_t^G = \frac{\mathbb{E} \left\| \sum_{k \in \mathcal{S}_t} \mathbf{w}_t^k \right\|^2}{\mathbb{E} \left\| \sum_{k \in \mathcal{S}_t} \mathbf{n}_t^k \right\|^2} = \frac{\mathbb{E} \left\| \sum_{k \in \mathcal{S}_t} \mathbf{w}_t^k \right\|^2}{d \sigma_t^2}, \quad (12)$$

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

$$\mathsf{SNR}_t^G = \frac{\mathbb{E} \left\| \sum_{k \in \mathcal{S}_t} \mathbf{w}_t^k \right\|^2}{\mathbb{E} \left\| \sum_{k \in \mathcal{S}_t} \left(\mathbf{n}_t^k - \mathbf{e}_t^k \right) \right\|^2} = \frac{\mathbb{E} \left\| \sum_{k \in \mathcal{S}_t} \mathbf{w}_t^k \right\|^2}{d \left(\sigma_t^2 + \zeta_t^2 \right)}, \quad (13)$$

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, if we assume that these local models are independent across clients, which is reasonable in the early phases of FL with large local epochs, where

---

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

---

the (roughly) same starting point has diminishing impact due to the long training period and non-IID nature of the data distribution, we can have a signal power term of dK. 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 IV.

In this paper, we mainly focus 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.

IV. CONVERGENCE ANALYSIS OF FL OVER NOISY CHANNELS

A. Convergence Analysis for Model Transmission for Full Clients Participation

We first analyze the convergence of FEDAVG in the presence of both uplink and downlink communication noise 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, we assume $K = N$ in this subsection (i.e., *full* clients participation with $\mathcal{S}_t = [K] = [N]$), and leave the case of partial clients participation to Section IV-B.

We make the following standard assumptions that are commonly adopted in the convergence analysis of FEDAVG and its variants; see [9], [10], [15], [21], [41]. In particular, Assumption 1-2) indicates that we focus on strongly convex $F_k(\cdot)$, which represents a category of loss functions that are widely studied in the literature.

Assumption 1: 1) *L-Smooth:* $\forall$ $\mathbf{v}$ and $\mathbf{w}$, $F_k(\mathbf{v}) \leq F_k(\mathbf{w}) + (\mathbf{v} - \mathbf{w})^T \nabla F_k(\mathbf{w}) + \frac{L}{2} \|\mathbf{v} - \mathbf{w}\|^2$.

2) *μ-Strongly Convex:* $\forall$ $\mathbf{v}$ and $\mathbf{w}$, $F_k(\mathbf{v}) \geq F_k(\mathbf{w}) + (\mathbf{v} - \mathbf{w})^T \nabla F_k(\mathbf{w}) + \frac{\mu}{2} \|\mathbf{v} - \mathbf{w}\|^2$.

3) *Bounded Variance for Unbiased Mini-Batch SGD:* The mini-batch SGD is unbiased: $\mathbb{E}[\nabla F_k(\mathbf{w}, \xi)] = \nabla F_k(\mathbf{w})$, and the variance of stochastic gradients is bounded: $\mathbb{E} \|\nabla F_k(\mathbf{w}, \xi) - \nabla F_k(\mathbf{w})\|^2 \leq \delta_k^2$, for mini-batch data ξ at client $k \in [N]$.

4) *Uniformly Bounded Gradient:* $\mathbb{E} \|\nabla F_k(\mathbf{w}, \xi)\|^2 \leq H^2$ for mini-batch data ξ at client $k \in [N]$.

We present the main convergence result of MT with full clients participation in 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 an SNR control policy that scales the effective uplink and downlink noise power over t such that:

$$\sigma_t^2 \leq \frac{4N^2}{\mu^2 (\gamma + t - 1)^2} \sim \mathcal{O}\left(\frac{1}{t^2}\right) \quad (14)$$

$$\zeta_t^2 \leq \frac{4N^2}{\mu^2 (\gamma + t)(\gamma + t - 2)} \sim \mathcal{O}\left(\frac{1}{t^2}\right) \quad (15)$$

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, the convergence of FEDAVG with non-IID datasets and full clients participation satisfies

$$\mathbb{E}\|\mathbf{w}_T - \mathbf{w}^*\|^2 \leq \frac{8L + \mu E}{\mu(T + \gamma)}\|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \frac{4D}{\mu^2(T + \gamma)} \tag{16}$$

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 [42], especially the virtual sequence construction that have been widely adopted in the distributed SGD analysis [9], [10], [15], [21]. 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: We make an important clarification 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 III-B. 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. We discuss design examples that realize the requirement of Theorem 1 in Section V.

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 VI, choosing an 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. We note that the choice of η_t to scale as $\mathcal{O}(1/t)$ is well-known in distributed and federated learning [15], [20], [21], [41], 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. We will investigate the convergence requirement that clients adopt different learning rates in future research.

Remark 5: Lastly, we note 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 V-A.

B. Convergence Analysis for Model Transmission for Partial Clients Participation

We now generalize the convergence analysis for full clients participation to *partial* clients participation, where we have a given $K < N$ and uniformly randomly select a set of clients $\mathcal{S}_t$ at round t to carry out the FL process. In this section, we mostly follow the FL system model described in Section III-B, with the only simplification that we consider *homogeneous* noise power levels at the uplink and downlink, i.e., we assume

$$\sigma_{t,k}^2 = \bar{\sigma}_t^2, \quad \text{and} \quad \zeta_{t,k}^2 = \bar{\zeta}_t^2, \quad \forall t \in [T], k \in [N]. \tag{17}$$

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 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 an SNR control policy that scales the effective uplink and downlink noise power over t such that:

$$\bar{\sigma}_t^2 \leq \frac{4K}{\mu^2(\gamma + t - 1)^2} \sim \mathcal{O}\left(\frac{1}{t^2}\right) \tag{18}$$

$$\bar{\zeta}_t^2 \leq \frac{4N}{\mu^2(\gamma + t)(\gamma + t - 2)} \sim \mathcal{O}\left(\frac{1}{t^2}\right) \tag{19}$$

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 (17). Then, under Assumption 1, the convergence of FEDAVG with non-IID datasets and partial clients participation has the same convergence rate expression as (16), with D being replaced as $D = \sum_{k=1}^{N} \delta_k^2/N^2 + 4(N-K)E^2H^2/(K(N-1)) + 6L\Gamma + 8(E-1)^2H^2 + 2d$.

---

[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" [43].

The proof of Theorem 2 is given in Appendix B. We can see 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 (18) compared with (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. We will provide a practical example to handle this tighter upper bound in Section V-B.

C. Convergence Analysis for Model Differential Transmission

In this section, we consider 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 (17). Adopt an SNR control policy that maintains a constant uplink SNR at each client over t:

$$\mathsf{SNR}_{t,k}^{\mathrm{S,MDT}} = \nu \sim \mathcal{O}(1), \tag{20}$$

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

$$\bar{\zeta}_t^2 \leq \frac{\frac{4}{\mu^2}}{\frac{1}{N}(\gamma + t)(\gamma + t - 2) + \frac{1}{K}\left(1 + \frac{1}{\nu}\right)(\gamma + t)^2} \sim \mathcal{O}\left(\frac{1}{t^2}\right). \tag{21}$$

Then, under Assumption 1, 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 (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 scale down the effective noise power at the server.

Lastly, we note that the constant SNR requirement of Theorem 3 enables very simple implementation given the MDT SNR expression in (9). The signal power in the numerator of (9) is unknown and varies 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.

V. COMMUNICATION DESIGN EXAMPLES FOR FL IN NOISY CHANNELS

An immediate engineering question following the previous analyses is how we 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 an FL task with T total communication rounds and a given total power budget of P over all rounds, it is straightforward to compute that

$$P_t = 6Pt^2/(T(T + 1)(2T + 1)), \ \forall t = 1, \ldots, T, \tag{22}$$

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

Since we consider analog communication and aggregation for FL, we also need to take the wireless channel fading into account. To combat the influence of channel fading on received power, we now propose two design examples to demonstrate how the proposed $\mathcal{O}(t^2)$-power increased strategy could be adopted in both continuous and discrete average power allocation schemes.

A. Design Example I: Transmit Power Control for Analog Aggregation

We first design a power control policy for the analog aggregation FL framework in [12], [13], [33], as an example to demonstrate the system design for FL tasks in the presence of communication noise.

The Analog Aggregation Method in [12], [13], [33]: 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 $i = 1, \ldots, d$ 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. We assume perfect channel state information at the transmitters (CSIT). Due to the aggregation requirement of federated learning, the channel inversion rule is used in [13], which leads to the following *instantaneous* transmit power of user k at time t for model weight element i:

$$p_{t,k,i} = \frac{\rho_t^{\mathrm{UL}}}{r_{t,k}^{-\alpha}|h_{t,k,i}|^2}, \tag{23}$$

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

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

$$= \frac{\rho_t^{\mathrm{UL}} \mathbb{E}\left\| \sum_{k \in \mathcal{S}_t} \mathbf{w}_t^k \right\|^2}{dK^2}. \tag{24}$$

Transmit Power Control: The original analog aggregation framework in [13] assumes that ρ_t^{UL} is a constant over time t. However, our theoretical analysis in Section IV suggests that this can be improved. Specifically, if we take partial clients participation and MT as an example, and further assume IID weight elements, we have

$$\rho_t^{\mathrm{UL}} = \frac{K}{\bar{\sigma}_t^2} \geq \frac{\mu^2(\gamma + t - 1)^2}{4} \sim \mathcal{O}\left(t^2\right), \tag{25}$$

by plugging in Theorem 2, 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

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

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

$$\mathsf{SNR}_{t,n,i}^L = r_{t,n}^{-\alpha}|h_{t,n,i}|^2 \rho_t^{\mathrm{DL}}. \tag{26}$$

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

$$\rho_t^{\mathrm{DL}} \geq \frac{r_{t,k}^{\alpha} \mu^2(\gamma + t)(\gamma + t - 2)}{4N|h_{t,n,i}|^2} \sim \mathcal{O}(t^2). \tag{27}$$

Finally, by applying the power control policy defined in Eqns. (25) and (27), FL tasks are able to achieve better performances under the same energy budget. This is also numerically validated in the experiment.

Remarks: We 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) [44], 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 (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 (27) but not in (25). This is due to the broadcast nature of download. For upload, the pathloss is absorbed in the channel inversion expression (23).

B. 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 we are leveraging the repeated transmissions to reduce the effective noise power via receive diversity combining, and by only activating sufficient diversity branches as we progress over the learning rounds, resources can be more efficiently utilized.

Uplink Diversity Requirement: We assume the uploaded local model is independently received L_t times (over time, frequency, space, or some combination of them) in the t-th round. Reusing the notations and the channel inversion rule in (23), the L_t received signals for the i-th element can be denoted as

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

For simplicity, we fixed the average transmit power for each branch: $\rho_{t,l} = \rho_0$, but this can be easily extended to incorporate power allocation over diversity branches [44]. The receive SNR of the global model after the diversity combining can be written as

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

$$= L_t \frac{\rho_0 \mathbb{E}\left\| \sum_{k \in \mathcal{S}_t} \mathbf{w}_t^k \right\|^2}{dK^2}.$$

Compared with the SNR of the power control policy in (24), we can derive the diversity requirement as

$$L_t = \left\lceil \rho_t^{\mathrm{UL}}/\rho_0 \right\rceil, \tag{28}$$

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

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

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 (26), we can derive the diversity requirement as

$$Q_t = \left\lceil \rho_t^{\mathrm{DL}}/\rho_1 \right\rceil. \qquad (29)$$

By applying the combining rules in Eqns. (28) and (29), we have the complete design for receive diversity combining that can guarantee the convergence of FL at rate $\mathcal{O}(1/T)$, 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 V-A with the receive diversity combining in Section V-B in a straightforward manner. We also 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.

VI. EXPERIMENT RESULTS

A. Experiment Setup

We consider noisy uplink and downlink communications to support various FL tasks. For simplicity, we assume that every channel use has the same noise level, and we also assume that 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, $t = 1, \ldots, T$. However, we 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 III-B. We consider the following four schemes 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 $P_t = P/T, \forall t = 1, \ldots, T$, as used in [13]. 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 noise of the signal is decreased with the progress of FL. With the total budget P, (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.

We use the standard image classification and natural language processing FL tasks to evaluate the performances of these schemes. The following three standard datasets are used in the experiments, which are commonly accepted as the benchmark tasks to evaluate the performance of FL.

1) *MNIST:* The training sets contain 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 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 varies significantly. There are totally 1129 roles in the dataset [45]. 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
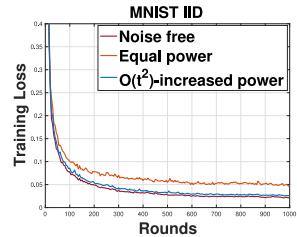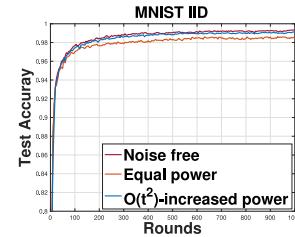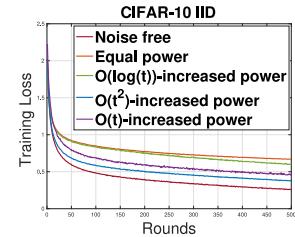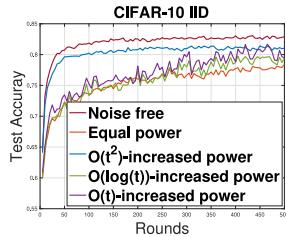
Fig. 2. Comparing the performance of transmit power control to the baselines with full clients participation, model transmission, and both IID (left two) and non-IID (right two) FL on the CIFAR-10 dataset.

Fig. 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.

with 256 hidden units) and a softmax 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.

We compare the test accuracies and training losses as functions of the communication rounds for all the aforementioned configurations. 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.

B. Experiment Results for Transmit Power Control

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

Full Clients Participation: We see from Fig. 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, we also carry out experiments where power is increased as a slower rate of $\mathcal{O}(\log(t))$ and $\mathcal{O}(t)$. 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, we 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
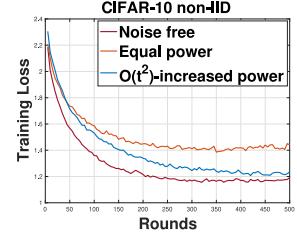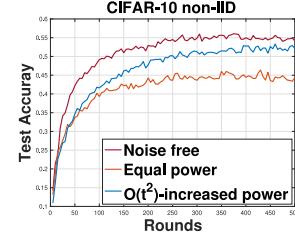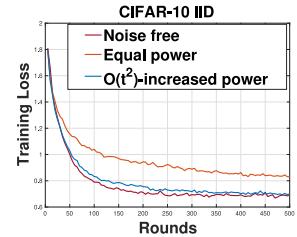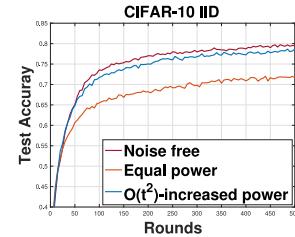

Fig. 4. Comparing the performance of transmit power control to the baselines with partial clients participation, model transmission, and both IID (left two) and non-IID (right two) FL on the CIFAR-10 dataset.

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. 3, 4, and 5, respectively. Their final model accuracies (after T rounds of FL are complete) are also summarized in Table I. First, we see from Fig. 3 that 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.
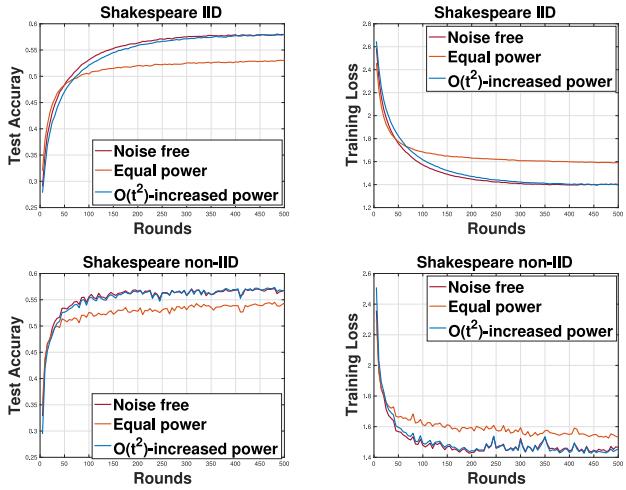
Fig. 5. Comparing the performance of transmit power control to the baselines with partial clients participation, model transmission, and both IID (left two) and non-IID (right two) FL on the Shakespeare dataset.

Fig. 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.

TABLE I
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%

TABLE II
PERFORMANCE SUMMARY OF MDT

| Dataset	Scheme	Accuracy	Percentage*	Accuracy	Percentage*
		IID		non-IID	
MNIST	Noise free	99.3%	100%	99.1%	100%
	Increased power	99.1%	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%

Fig. 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.

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. 4 and Fig. 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: We next present the experiment results of model differential transmission. Note that, by applying MDT, the uplink transmission power of the proposed scheme remains constant (recall that SNR is set as 10dB) while the downlink transmission power still increases at the rate of $\mathcal{O}(t^2)$. Figs. 6, 7 and 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 II. We see that 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.
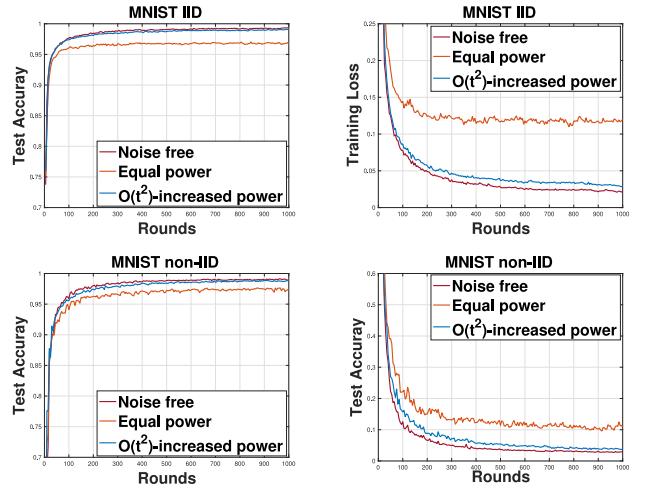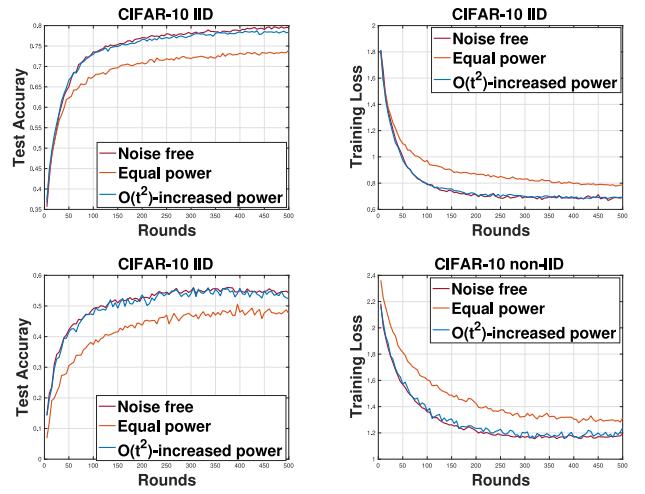
C. Experiment Results for Receive Diversity Combining

We next evaluate the performance of receive diversity combining. Due to space limitations, we only report the result for CIFAR-10, but similar conclusions hold for other tasks. Fig. 9 captures the test accuracies and training losses of receive diversity combining together with noise free and equal power
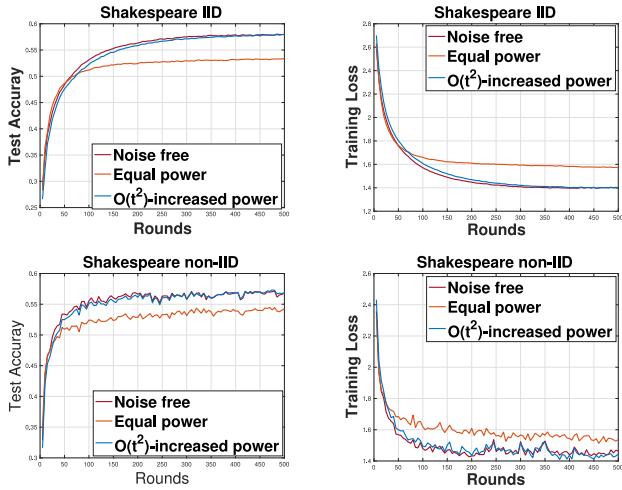
Fig. 8. Comparing the performance of transmit power control to the baselines with partial clients participation, model differential transmission, and both IID (left two) and non-IID (right two) FL on the Shakespeare dataset.
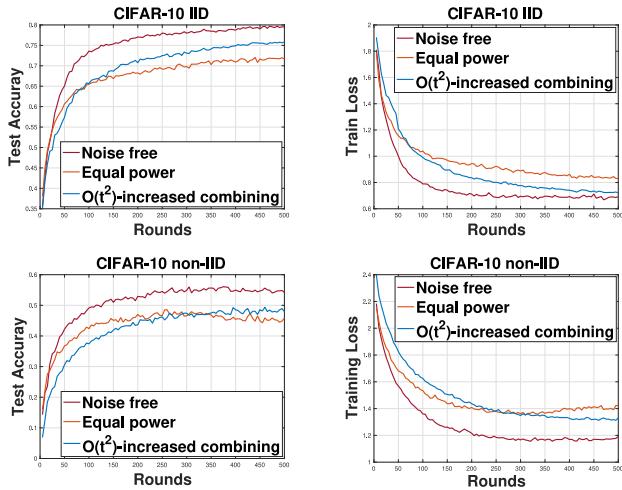

Fig. 9. Comparing the performance of receive diversity combining to the baselines with partial clients participation, model differential transmission, and both IID (left two) and non-IID (right two) FL on the CIFAR-10 dataset.

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. We notice that 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.

VII. CONCLUSION

In this paper, we have investigated federated learning over noisy channels, where a FEDAVG pipeline with both uplink and downlink communication noises was studied. By theoretically analyzing the model training convergence, we have

proved that the same $\mathcal{O}(1/T)$ convergence rate of FEDAVG under perfect (noise-free) communications can be maintained if the uplink and downlink SNRs are controlled as $\mathcal{O}(t^2)$ over noisy channels for direct model transmission, and $\mathcal{O}(1)$ for model differential transmission. We have showcased two widely used communication methods – transmit power control and receive diversity combining – to implement these theoretical results. Extensive experimental results have corroborated the theoretical analysis and demonstrated the performance superiority of the advanced designs over baseline methods under the same total energy budget. Future research directions include relaxing the assumption of strongly convex loss functions to a broader class (e.g., convex, non-convex), and removing the fixed T assumption to develop an *any-time* version of the proposed design.

APPENDIX A
PROOF OF THEOREM 1

A. 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.,

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

Note that the (noisy) global model $\mathbf{w}_t$ is only accessible at the clients for specific $t \in \mathcal{I}_E$, where $\mathcal{I}_E = \{nE \mid n = 1, 2, \ldots\}$, 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 IV-A, the key technique in the proof is the **perturbed iterate framework** in [42]. In particular, We first define the following variables

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

and $\mathbf{w}_{t+1}^k = \mathbf{p}_{t+1}^k + \mathbf{e}_{t+1}^k,$

to summarize the aforementioned steps:

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

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

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

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

Then, we construct the following *virtual sequences*:

$$\overline{\mathbf{v}}_t = \frac{1}{N} \sum_{k=1}^N \mathbf{v}_t^k, \quad \overline{\mathbf{u}}_t = \frac{1}{N} \sum_{k=1}^N \mathbf{u}_t^k,$$

$$\overline{\mathbf{p}}_t = \frac{1}{N} \sum_{k=1}^N \mathbf{p}_t^k, \quad \text{and } \overline{\mathbf{w}}_t = \frac{1}{N} \sum_{k=1}^N \mathbf{w}_t^k. \tag{30}$$

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

$$\bar{\mathbf{u}}_{t+1} = \frac{1}{N}\sum_{i\in[N]}\mathbf{v}_{t+1}^i, \quad \bar{\mathbf{p}}_{t+1} = \bar{\mathbf{u}}_{t+1} + \frac{1}{N}\sum_{i\in[N]}\mathbf{n}_{t+1}^i,$$

$$\bar{\mathbf{w}}_{t+1} = \bar{\mathbf{p}}_{t+1} + \frac{1}{N}\sum_{k=1}^{N}\mathbf{e}_{t+1}^k. \tag{31}$$

Note that for $t+1 \notin \mathcal{I}_E$, all these virtual sequences are the same. In addition, the global model (at the server) $\bar{\mathbf{p}}_{t+1}$ is meaningful only at $t+1 \in \mathcal{I}_E$. We emphasize that when $t+1 \in \mathcal{I}_E$, Eqns. (31) and (10) indicate that $\bar{\mathbf{p}}_{t+1} = \mathbf{w}_{t+1}$. Thus it is sufficient to analyze the convergence of $\|\bar{\mathbf{p}}_{t+1} - \mathbf{w}^*\|^2$.

B. Lemmas and Proofs

Lemma 1: Let Assumption 1 hold, η_t is non-increasing, and $\eta_t \le 2\eta_{t+E}$ for all $t \ge 0$. If $\eta_t \le 1/(4L)$, we have $\mathbb{E}\|\bar{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 \le (1 - \eta_t\mu)\mathbb{E}\|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \eta_t^2(\sum_{k=1}^{N}\delta_k^2/N^2 + 6L\Gamma + 8(E-1)^2H^2)$.

Lemma 1 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 [15].

Lemma 2: We have

$$\mathbb{E}[\bar{\mathbf{p}}_{t+1}] = \bar{\mathbf{u}}_{t+1}, \mathbb{E}\|\bar{\mathbf{u}}_{t+1} - \bar{\mathbf{p}}_{t+1}\|^2 = \frac{d\sigma_{t+1}^2}{N^2};$$

$$\mathbb{E}[\bar{\mathbf{w}}_{t+1}] = \bar{\mathbf{p}}_{t+1}, \mathbb{E}\|\bar{\mathbf{w}}_{t+1} - \bar{\mathbf{p}}_{t+1}\|^2 = \frac{d\zeta_{t+1}^2}{N^2} \tag{32}$$

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 $\mathbb{E}[\bar{\mathbf{p}}_{t+1} - \bar{\mathbf{u}}_{t+1}] = \frac{1}{K}\sum_{k\in[N]}\mathbb{E}[\mathbf{n}_{t+1}^k] = 0$ and $\mathbb{E}\|\bar{\mathbf{p}}_{t+1} - \bar{\mathbf{u}}_{t+1}\|^2 = \frac{1}{N^2}\mathbb{E}\|\sum_{k\in[N]}\mathbf{n}_{t+1}^k\|^2 = \frac{1}{N^2}\sum_{k\in[N]}\mathbb{E}\|\mathbf{n}_{t+1}^k\|^2 = \frac{d\sigma_{t+1}^2}{N^2}$ from (31), because $\{\mathbf{n}_{t+1}^k, \forall k\}$ are independent variables. Similarly, according to (31), we have $\mathbb{E}[\bar{\mathbf{w}}_{t+1} - \bar{\mathbf{p}}_{t+1}] = \frac{1}{N}\sum_{k\in[N]}\mathbb{E}[\mathbf{e}_{t+1}^k] = 0$ and $\mathbb{E}\|\bar{\mathbf{w}}_{t+1} - \bar{\mathbf{p}}_{t+1}\|^2 = \frac{1}{N^2}\mathbb{E}\|\sum_{k\in[N]}\mathbf{e}_{t+1}^k\|^2 = \frac{1}{N^2}\sum_{k\in[N]}\mathbb{E}\|\mathbf{e}_{t+1}^k\|^2 = \frac{\sum_{k\in[N]}d\zeta_{t+1,k}^2}{N^2} = \frac{d\zeta_{t+1}^2}{N^2}$. ∎

C. Proof of Theorem

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

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

$$\mathbb{E}\|\bar{\mathbf{p}}_{t+1} - \mathbf{w}^*\|^2 = \mathbb{E}\|\bar{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 \le (1 - \eta_t\mu)\mathbb{E}\|\bar{\mathbf{p}}_t - \mathbf{w}^*\|^2$$

$$+ \eta_t^2\left[\sum_{k=1}^{N}\frac{\delta_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2\right].$$

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

$$\|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2 = \|\bar{\mathbf{w}}_t - \bar{\mathbf{p}}_t + \bar{\mathbf{p}}_t - \mathbf{w}^*\|^2$$

$$= \|\bar{\mathbf{p}}_t - \mathbf{w}^*\|^2 + \underbrace{\|\bar{\mathbf{w}}_t - \bar{\mathbf{p}}_t\|^2}_{A_1}$$

$$+ \underbrace{2\langle\bar{\mathbf{w}}_t - \bar{\mathbf{p}}_t, \bar{\mathbf{p}}_t - \mathbf{w}^*\rangle}_{A_2}.$$

We first note that the expectation of A_2 over the noise randomness is zero since we have $\mathbb{E}[\bar{\mathbf{w}}_t - \bar{\mathbf{p}}_t] = \mathbf{0}$ (from (32)). Second, the expectation of A_1 can be bounded using Lemma 2. We then have

$$\mathbb{E}\|\bar{\mathbf{p}}_{t+1} - \mathbf{w}^*\|^2 = \mathbb{E}\|\bar{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2$$

$$\le (1 - \eta_t\mu)\mathbb{E}\|\bar{\mathbf{p}}_t - \mathbf{w}^*\|^2$$

$$+ (1 - \eta_t\mu)\mathbb{E}\|\bar{\mathbf{w}}_t - \bar{\mathbf{p}}_t\|^2 \tag{33}$$

$$+ \eta_t^2\left[\sum_{k=1}^{N}\frac{\delta_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2\right]$$

$$\le (1 - \eta_t\mu)\mathbb{E}\|\bar{\mathbf{p}}_t - \mathbf{w}^*\|^2 + (1 - \eta_t\mu)\frac{d\zeta_t^2}{N^2} \tag{34}$$

$$+ \eta_t^2\left[\sum_{k=1}^{N}\frac{\delta_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2\right]. \tag{35}$$

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

$$\|\bar{\mathbf{p}}_{t+1} - \mathbf{w}^*\|^2 = \|\bar{\mathbf{p}}_{t+1} - \bar{\mathbf{u}}_{t+1} + \bar{\mathbf{u}}_{t+1} - \mathbf{w}^*\|^2$$

$$= \underbrace{\|\bar{\mathbf{p}}_{t+1} - \bar{\mathbf{u}}_{t+1}\|^2}_{B_1} + \underbrace{\|\bar{\mathbf{u}}_{t+1} - \mathbf{w}^*\|^2}_{B_2}$$

$$+ \underbrace{2\langle\bar{\mathbf{p}}_{t+1} - \bar{\mathbf{u}}_{t+1}, \bar{\mathbf{u}}_{t+1} - \mathbf{w}^*\rangle}_{B_3}. \tag{36}$$

We first note that the expectation of B_3 over the noise is zero since we have $\mathbb{E}[\bar{\mathbf{u}}_{t+1} - \bar{\mathbf{p}}_{t+1}] = \mathbf{0}$ (from (32)). Second, the expectation of B_1 can be bounded using Lemma 2. Noticing that $\bar{\mathbf{u}}_{t+1} = \bar{\mathbf{v}}_{t+1}$ for B_2 and applying Lemma 1, we have:

$$\mathbb{E}\|\bar{\mathbf{p}}_{t+1} - \mathbf{w}^*\|^2 \le \mathbb{E}\|\bar{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 + \frac{d\sigma_{t+1}^2}{K^2}$$

$$\le (1 - \eta_t\mu)\mathbb{E}\|\bar{\mathbf{p}}_t - \mathbf{w}^*\|^2 + \frac{d\sigma_{t+1}^2}{N^2} \tag{37}$$

$$+ \eta_t^2\left[\sum_{k=1}^{N}\frac{\delta_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2\right]. \tag{38}$$

4) If $t \in \mathcal{I}_E$ and $t+1 \in \mathcal{I}_E$, $\bar{\mathbf{v}}_{t+1} \ne \bar{\mathbf{p}}_{t+1}$ and $\bar{\mathbf{w}}_t \ne \bar{\mathbf{p}}_t$. (Note that this is possible only for $E = 1$.) Combining the

results from the previous two cases, we have

$$
\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1} - \mathbf{w}^*\right\|^2 \leq (1 - \eta_t\mu)\mathbb{E}\|\overline{\mathbf{p}}_t - \mathbf{w}^*\|^2
$$
$$
+ (1 - \eta_t\mu)\frac{d\zeta_t^2}{N^2} + \frac{d\sigma_{t+1}^2}{N^2}
$$
$$
+ \eta_t^2\left[\sum_{k=1}^{N}\frac{\delta_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2\right].
$$
(39)

Finally, we have that inequality (39) holds for all four cases. Denote $\Delta_t = \mathbb{E}\|\overline{\mathbf{p}}_t - \mathbf{w}^*\|^2$. 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 $F(\cdot)$, $\mathbb{E}[F(\overline{\mathbf{w}}_t)] - F^* \leq \frac{L}{2}\Delta_t \leq \frac{L}{2}\frac{v}{\gamma+t}$. 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 $v \leq \frac{4D}{\mu^2} + (\gamma+1)\Delta_0 \leq \frac{4D}{\mu^2} + (8\phi+E)\|\mathbf{w}_0 - \mathbf{w}^*\|^2$. Therefore, $\Delta_t \leq \frac{v}{\gamma+t} = \frac{1}{\gamma+t}\left[\frac{4D}{\mu^2} + (8\phi+E)\|\mathbf{w}_0 - \mathbf{w}^*\|^2\right]$. Setting $t = T$ concludes the proof.

APPENDIX 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-A 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 2 still holds with the following update:

$\mathbb{E}\|\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{p}}_{t+1}\|^2 = \frac{d\bar{\sigma}_{t+1}^2}{K}$, $\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{p}}_{t+1}\|^2 = \frac{d\bar{\zeta}_{t+1}^2}{N}$. In addition, we need the following lemma, whose proof is available on a complete online version of this paper [46].

Lemma 3: 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 $\mathbb{E}[\overline{\mathbf{u}}_{t+1}] = \overline{\mathbf{v}}_{t+1}$ and $\mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \overline{\mathbf{u}}_{t+1}\|^2 \leq \frac{N-K}{N-1}\frac{4}{K}\eta_t^2 E^2 H^2$.

We can now similarly analyze the four cases as in Section A-C. 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 (36) still holds, but we need to re-evaluate the expectation of B_2 because of partial clients participation. We have:

$$
\|\overline{\mathbf{u}}_{t+1} - \mathbf{w}^*\|^2 = \|\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{v}}_{t+1} + \overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2
$$
$$
= \underbrace{\|\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{v}}_{t+1}\|^2}_{C_1} + \underbrace{\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2}_{C_2}
$$
$$
+ \underbrace{2\langle\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{v}}_{t+1}, \overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\rangle}_{C_3}. \quad (40)
$$

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

$$
\mathbb{E}\|\overline{\mathbf{p}}_{t+1} - \mathbf{w}^*\|^2 \leq (1 - \eta_t\mu)\mathbb{E}\|\overline{\mathbf{p}}_t - \mathbf{w}^*\|^2
$$
$$
+ (1 - \eta_t\mu)\frac{d\bar{\zeta}_t^2}{N} + \frac{d\bar{\sigma}_{t+1}^2}{K} + \eta_t^2(D - 2d)
$$
(41)

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-C.

APPENDIX C
PROOF OF THEOREM 3

For model differential transmission (MDT), if $t+1 \in \mathcal{I}_E$, the global aggregation is given in (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 $\mathbf{u}_{t+1}^k \triangleq \frac{1}{K}\sum_{i \in S_t}\mathbf{v}_{t+1}^i$, $\mathbf{p}_{t+1}^k \triangleq \mathbf{w}_{t+1-E} + \frac{1}{K}\sum_{i \in \mathcal{S}_t}[\mathbf{d}_{t+1}^i + \mathbf{n}_{t+1}^i]$, 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 (30). $\overline{\mathbf{g}}_t$ 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 4: 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: $\text{SNR}_{t,k}^{\text{S,MDT}} = \nu$. Then, for $t+1 \in \mathcal{I}_E$, we have: $\mathbb{E}[\overline{\mathbf{p}}_{t+1}] = \overline{\mathbf{u}}_{t+1}$ and $\mathbb{E}\|\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{p}}_{t+1}\|^2 \leq (1+\frac{1}{\nu})\frac{d}{K}\bar{\zeta}_{t+1-E}^2 + \frac{4E^2}{K\nu}\eta_t^2 H^2$.

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 $\mathbb{E}[\overline{\mathbf{p}}_{t+1}] = \overline{\mathbf{u}}_{t+1} + \frac{1}{K}\mathbb{E}[\sum_{k\in\mathcal{S}_t}\mathbf{n}_{t+1}^k] - \frac{1}{K}\mathbb{E}[\sum_{k\in\mathcal{S}_t}\mathbf{e}_{t+1-E}^k] = \mathbb{E}[\overline{\mathbf{u}}_{t+1}]$. As for the variance, we have

$$\mathbb{E}\|\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{p}}_{t+1}\|^2 = \frac{1}{K^2}\mathbb{E}\left\|\sum_{k\in\mathcal{S}_t}\mathbf{n}_{t+1}^k\right\|^2$$
$$+ \frac{1}{K^2}\mathbb{E}\left\|\sum_{k\in\mathcal{S}_t}\mathbf{e}_{t+1-E}^k\right\|^2$$
$$= \frac{1}{K^2\nu}\mathbb{E}\left\|\sum_{k\in\mathcal{S}_t}\mathbf{d}_{t+1}^k\right\|^2 + \frac{d\bar{\zeta}_{t+1-E}^2}{K} \quad (42)$$

where the last equality comes from the constant uplink SNR control, (9), and the assumption that each client has the same downlink noise power $\bar{\zeta}_t^2$, $\forall k \in [N]$. We further have

$$\mathbb{E}\left\|\sum_{k\in\mathcal{S}_t}\mathbf{d}_{t+1}^k\right\|^2 = \mathbb{E}\left\|\sum_{k\in\mathcal{S}_t}\left(\mathbf{v}_{t+1}^k - \mathbf{w}_{t+1-E}^k\right)\right\|^2 + dK\bar{\zeta}_{t+1-E}^2$$
$$\leq \mathbb{E}_{\mathcal{S}_t}\left[\sum_{k\in\mathcal{S}_t}\mathbb{E}_{\text{SG}}\left\|\sum_{\tau=t+1-E}^t \eta_\tau \nabla F_k(\mathbf{w}_\tau^k, \xi_\tau^k)\right\|^2\right]$$
$$+ dK\bar{\zeta}_{t+1-E}^2$$
$$\leq 4E^2 K\eta_t^2 H^2 + dK\bar{\zeta}_{t+1-E}^2 \quad (43)$$

using the Cauchy-Schwarz inequality, Assumption 1-4), and $\eta_{t+1-E} < \eta_{t-E} \leq 2\eta_t$. Plugging (43) back to (42) gives

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

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 (41) is updated to (44) below using Lemma 4 and the new definition of D in Theorem 3.

$$\mathbb{E}\|\overline{\mathbf{p}}_{t+1} - \mathbf{w}^*\|^2 \leq (1-\eta_t\mu)\mathbb{E}\|\overline{\mathbf{p}}_t - \mathbf{w}^*\|^2 + (1-\eta_t\mu)\frac{d}{N}\bar{\zeta}_t^2$$
$$+ \left(1+\frac{1}{\nu}\right)\frac{d}{K}\bar{\zeta}_t^2 + \eta_t^2(D-d). \quad (44)$$

We note that the constant uplink SNR control is already used in Lemma 4 and (44). Then, by the definition of $\Delta_t = \mathbb{E}\|\overline{\mathbf{p}}_t - \mathbf{w}^*\|^2$ and controlling the downlink SNR such that $\bar{\zeta}_t^2 \leq \frac{NK\eta_t^2}{(1-\eta_t\mu)K+(1+\frac{1}{\nu})N}$, 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.

REFERENCES

[1] X. Wei and C. Shen, "Federated learning over noisy channels," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2021, pp. 1–6.

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

[3] J. Konecný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *Proc. NIPS Workshop Private Multi-Party Mach. Learn.*, 2016.

[4] Z. Yang, M. Chen, K.-K. Wong, H. V. Poor, and S. Cui, "Federated learning for 6G: Applications, challenges, and opportunities," 2021, *arXiv:2101.01338*.

[5] K. Bonawitz *et al.*, "Towards federated learning at scale: System design," in *Proc. 2nd SysML Conf.*, 2019, pp. 1–15.

[6] P. Kairouz *et al.*, "Advances and open problems in federated learning," 2019, *arXiv:1912.04977*.

[7] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[8] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. 3rd MLSys Conf.*, 2020.

[9] S. Zheng, C. Shen, and X. Chen, "Design and analysis of uplink and downlink communications for federated learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2150–2167, Jul. 2021.

[10] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization," in *Proc. AISTATS*, 2020, pp. 2021–2031.

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

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

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

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

[15] S. U. Stich, "Local SGD converges fast and communicates little," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[16] S. Xia, J. Zhu, Y. Yang, Y. Zhou, Y. Shi, and W. Chen, "Fast convergence algorithm for analog federated learning," 2020, *arXiv:2011.06658*.

[17] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Over-the-air federated learning from heterogeneous data," 2020, *arXiv:2009.12787*.

[18] 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.*, vol. 8, no. 1, pp. 197–210, Jan. 2021.

[19] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Convergence of federated learning over a noisy downlink," 2020, *arXiv:2008.11141*.

[20] J. Wang and G. Joshi, "Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms," in *Proc. ICML Workshop Coding Theory Mach. Learn.*, 2019.

[21] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[22] G. Zhu, Y. Du, D. Gündüz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," 2020, *arXiv:2001.05713*.

[23] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Federated learning with quantized global model updates," 2020, *arXiv:2006.10672*.

[24] S. Chen, C. Shen, L. Zhang, and Y. Tang, "Dynamic aggregation for heterogeneous quantization in federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6804–6819, Oct. 2021.

[25] X. Mo and J. Xu, "Energy-efficient federated edge learning with joint communication and computation design," 2020, *arXiv:2003.00199*.

[26] Q. Zeng, Y. Du, K. K. Leung, and K. Huang, "Energy-efficient radio resource allocation for federated edge learning," 2019, *arXiv:1907.06040*.

[27] W. Shi, S. Zhou, and Z. Niu, "Device scheduling with fast convergence for wireless federated learning," 2019, *arXiv:1911.00856*.

[28] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," 2019, *arXiv:1911.02417*.

[29] H. H. Yang, Z. Liu, T. Q. S. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 317–333, Jan. 2020.

[30] M. Chen, H. V. Poor, W. Saad, and S. Cui, "Convergence time optimization for federated learning over wireless networks," 2020, *arXiv:2001.07845*.

[31] J. Xu and H. Wang, "Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1188–1200, Feb. 2021.

[32] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," 2019, *arXiv:1909.07972*.

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

[34] M. Frey, I. Bjelakovic, and S. Stanczak, "Over-the-air computation for distributed machine learning," 2020, *arXiv:2007.02648*.

[35] R. Jiang and S. Zhou, "Cluster-based cooperative digital over-the-air aggregation for wireless federated edge learning," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Chongqing, China, 2020, pp. 887–892.

[36] F. Ang, L. Chen, N. Zhao, Y. Chen, W. Wang, and F. R. Yu, "Robust federated learning with noisy communication," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3452–3464, Jun. 2020.

[37] H. Tang, C. Yu, X. Lian, T. Zhang, and J. Liu, "DoubleSqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6155–6165.

[38] Y. Yu, J. Wu, and L. Huang, "Double quantization for communication-efficient distributed optimization," 2018, *arXiv:1805.10111*.

[39] C.-Y. Chen *et al.*, "ScaleCom: Scalable sparsified gradient compression for communication-efficient distributed training," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., Inc., 2020, pp. 13551–13563.

[40] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Advances in Neural Information Processing Systems*. Red Hook, NY, UA: Curran Assoc., Inc., 2016.

[41] P. Jiang and G. Agrawal, "A linear speedup analysis of distributed deep learning with sparse and quantized communication," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran 2018, pp. 2525–2536.

[42] H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M. I. Jordan, "Perturbed iterate analysis for asynchronous stochastic optimization," *SIAM J. Optim.*, vol. 27, no. 4, pp. 2202–2229, 2017.

[43] A. Reisizadeh, I. Tziotis, H. Hassani, A. Mokhtari, and R. Pedarsani, "Straggler-resilient federated learning: Leveraging the interplay between statistical accuracy and system heterogeneity," 2020, *arXiv:2012.14453*.

[44] A. Goldsmith, *Wireless Communications*. New York, NY, USA: Cambridge Univ. Press, 2005.

[45] S. Caldas *et al.*, "LEAF: A benchmark for federated settings," 2018, *arXiv:1812.01097*.

[46] X. Wei and C. Shen, "Federated learning over noisy channels: Convergence analysis and design examples," 2021, *arXiv:2101.02198*.

Xizixiang Wei received the B.E. degree in engineering from Tongji University, Shanghai, China, in 2017, and the M.E. degree in engineering from Fudan University, Shanghai, in 2020. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Virginia. His research interests are mainly in the area of federated learning, wireless communication, and signal processing.

Cong Shen (Senior Member, IEEE) received the B.S. and M.S. degrees from the Department of Electronic Engineering, Tsinghua University, China, and the Ph.D. degree in electrical engineering from the University of California, Los Angeles. He is currently an Assistant Professor with the Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia. He also has extensive industry experience, having worked for Qualcomm, SpiderCloud Wireless, Silvus Technologies, and Xsense.ai, in various full time and consulting roles. His general research interests are in the area of wireless communications and machine learning. He currently serves as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, and IEEE WIRELESS COMMUNICATIONS LETTERS.