RESEARCH ARTICLE

Improved Convergence Analysis and SNR Control Strategies for Federated Learning in the Presence of Noise

ANTESH UPADHYAY AND **ABOLFAZL HASHEMI**, (Member, IEEE)

School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA

Corresponding author: Antesh Upadhyay (antesh1414@gmail.com)

ABSTRACT We propose an improved convergence analysis technique that characterizes the distributed learning paradigm of federated learning (FL) with imperfect/noisy uplink and downlink communications. Such imperfect communication scenarios arise in the practical deployment of FL in emerging communication systems and protocols. The analysis developed in this paper demonstrates, for the first time, that there is an asymmetry in the detrimental effects of uplink and downlink communications in FL. In particular, the adverse effect of the downlink noise is more severe on the convergence of FL algorithms. Using this insight, we propose improved Signal-to-Noise (SNR) control strategies that, discarding the negligible higher-order terms, lead to a similar convergence rate for FL as in the case of a perfect, noise-free communication channel while incurring significantly less power resources compared to existing solutions. In particular, we establish that to maintain the $\mathcal{O}(\frac{1}{\sqrt{K}})$ rate of convergence like in the case of noise-free FL, we need to scale down the uplink and downlink noise by $\Omega(\sqrt{k})$ and $\Omega(k)$ respectively, where k denotes the communication round, $k = 1, \ldots, K$. Our theoretical result is further characterized by two major benefits: firstly, it does not assume the somewhat unrealistic assumption of *bounded client dissimilarity*, and secondly, it only requires smooth non-convex loss functions, a function class better suited for modern machine learning and deep learning models. We also perform extensive empirical analysis to verify the validity of our theoretical findings.

INDEX TERMS Deep learning, distributed systems, federated learning, optimization, signal-to-noise ratio.

I. INTRODUCTION

The advancements in the field of Machine Learning (ML) are attributable to the increasing ability to generate and process data from various edge devices such as sensors, mobile phones, and the internet of things (IoT) devices. However, traditional ML approaches that rely on storing data on a server can be problematic in terms of the privacy of the user and scalability. To address these issues, we see the eminent shift towards distributed and collaborative ML approaches, such as consensus-based distributed optimization and Federated learning (FL) [1], [2], [3], [4], [5], [6], [7], which allow for learning to occur without the need for

central data storage. In this paper, our primary focus is on the FL setting. The algorithm popularly known as Federated Averaging (`FedAvg`) presented in [8], provided the foundation for FL. Breaking the quintessential model of traditional ML, `FedAvg`, preserves the privacy of the agent (typically referred to as clients in FL) by allowing them to retain their data. In `FedAvg`, a set of agents (referred to as clients in FL), based on their local data, perform the Stochastic Gradient Descent (SGD) iteratively for a certain number of local steps and then transmits their updated model parameters to a central server, which then averages these updates and, in turn, updates the global model. Iterative communication between servers and clients and the collaborative nature of FL shows the importance of communication vis-à-vis, `FedAvg` and other FL and consensus-based

The associate editor coordinating the review of this manuscript and approving it for publication was Hao Wang.

methods, and it has been an active area of research in terms of improving the efficiency and resiliency of such algorithms [9], [10], [11], [12], [13], [14], [15], [16].

A. RELATED WORKS

Several recent results, where improving communication efficiency is the core, focus mainly on reducing the number of communication rounds [8], [17], or the size of information during transmission [18], [19], [20], [21], [22], [23]. However, in most of these studies, the process of communication from the server to clients (*downlink*) and then from clients to the server (*uplink*), a perfect communication link is often assumed. Now, some literature investigates the impact of having a noisy transmission channel but only studies the effect of noisy uplink transmission [24], [25], [26], [27], [28]. However, only a few articles in the literature deal with the impact of only downlink noise or both noises [29]. A major consideration among all these works is their somewhat restrictive assumptions that typically are not satisfied in practical settings or are hard to verify. For instance, in [24], where they analyze the effect of downlink noise, they assume a perfect uplink communication channel. Additionally, existing research that studies both uplink and downlink noise focuses on the modification of the training of ML models. Reference [30] aims to counter the effect of noise by modifying the loss function to consider the addition of noise as a regularizer. Similarly [21], [31], [32], [33] focus on compressing the gradients to counter the effect of noisy transmission channels. Since compression inherently adds noise to the message communicated, it poses an adverse impact on model convergence. The limitations of these works serve as our primary motivation as we aim to understand the impact of the uplink and downlink communication on the performance of `FedAvg` by developing an improved analysis to reduce the burden of intensive power consumption while relaxing the assumptions of convexity and bounded client dissimilarity (BCD) required by the existing works.

Recently, [29] studies the impact of both uplink and downlink noises with restrictive assumptions of strong-convexity and Bounded Client Dissimilarity (BCD) [34]. To avoid the client-drift [34], a standard assumption used in FL is BCD (refer to eq. (10)). This drift occurs due to multiple local SGD updates on clients with non-IID data distribution, which prohibits the algorithm from converging to the global optimum. Nevertheless, the result of [29] has an important shortcoming: the analysis is not tight due to which while the model converges, the dominant terms in the convergence error depend on noise characteristics and as a result, the Signal-to-Noise Ratio (SNR) scaling policy requires more power compared to our results.

B. CONTRIBUTIONS

The contributions of our work are motivated towards mitigating the restrictions imposed in previous literature. Unlike the previous work, [29], we propose an analysis of smooth non-convex `FedAvg` with noisy (both uplink and downlink) communication channels and without the BCD assumption. We leverage the non-negativity of the typical loss functions in optimization/ML and their smoothness in conjunction with a novel sampling technique to avoid using BCD while establishing our improved convergence results. We present the results of our analysis in Theorem 7 and Corollary 10, which shows that the effect of downlink noise, i.e., $\mathcal{O}(1)$, is more degrading than uplink noise, i.e., $\mathcal{O}(\frac{1}{\sqrt{K}})$ where K is the number of communication rounds. Hence, following these results, we draw an inference that as long as we control the effect of downlink and uplink noise such that they do not dominate the inherent noisy communication aspect of SGD, the convergence of the model can be achieved while limiting the adverse effect of noise to negligible higher-order terms. In particular, we demonstrate both theoretically and empirically that in order to maintain the convergence rate of $\mathcal{O}(\frac{1}{\sqrt{K}})$ for the case of noise-free `FedAvg`, we need to scale down the downlink noise by $\Omega(k)$ and uplink noise by $\Omega(\sqrt{k})$, or, equivalently scale the downlink noise by $\Omega(\frac{1}{k})$ and uplink noise by $\Omega(\frac{1}{\sqrt{k}})$. These scaling rates ensure that the noise appears as a higher-order term, not as a dominant term.[1]

To summarize, the contributions of this paper are as follows:

- We provide an improved analysis alongside the complete proof of FL under the presence of uplink and downlink noise without using any constraining assumption which results in tighter convergence analysis.
- In reference to Corollary 10, we provide empirical results that show the asymmetric effect of both uplink and downlink noises. We provide plots that verify that for a constant number of communication rounds, uplink noise scales as $\mathcal{O}(\frac{1}{E^2\sqrt{r}})$, while the term corresponding to downlink noise is $\mathcal{O}(1)$. Here r denotes the number of clients participating in each round.
- In this paper, we present a method for controlling the SNR ratio in order to mitigate the deleterious effects of noise on both the uplink and downlink communication channels. The proposed scaling policy is more accommodating and robust towards higher noise concentrations as well. Our analysis demonstrates that this approach leads to improved performance.
- We provide empirical results on both synthetic and real-world deep learning experiments on datasets such as MNIST, Fashion-MNIST, CIFAR-10, and FEMNIST datasets, to establish the efficacy and validity of the proposed technique.

II. PRELIMINARIES AND SYSTEM-MODEL

The setting of the problem follows the traditional FL scenario presented in [8] (see also Figure 1). In a standard FL setting, we have a central server and a set of n clients, each having

---

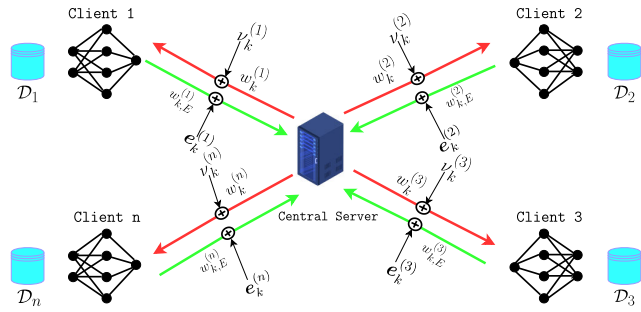[1]Refer Theorem 7 and Corollary 10 for more details about the effect of noises.

FIGURE 1. Problem setting: an FL system with uplink and downlink communication noise.

their local training data. The ith client stores their local data sampled from a distribution $\mathcal{D}_i$. The central server aims to train a machine learning model on the client's local data, parameterized by $w \in \mathbb{R}^d$. Then, $f_i(w)$ is the expected loss over a sample x drawn from $\mathcal{D}_i$ with respect to a loss function ℓ for the ith client. The primary objective of the central server is to minimize the loss $f(w)$ over n clients, i.e.,

$$f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w) \,\& \, f_i(w) = \mathbb{E}_{x \sim \mathcal{D}_i}[\ell(x, w)]. \quad (1)$$

Also, to emulate an FL setting in practice, we consider partial client participation, i.e., a set of r clients chosen uniformly at random without replacement from a set of n clients, whereas in the case of full participation, $r = n$. Such an assumption is motivated by the consideration that the clients may have limited communication capabilities and not all will be able to collaborate at every communication round. We assume that these clients have access to the unbiased stochastic gradient of their individual losses which is denoted by $\widetilde{\nabla} f_i(w; \mathcal{B})$ computed at w over a batch of samples $\mathcal{B}$. In addition, K denotes the communication rounds, and E represents the number of local iterations for each communication round.

The FL process can be thought of as an iterative, three-step pipeline: 1) global model update from the central server to the clients over a noisy channel, i.e., noisy downlink communication, 2) client-level computation, and 3) sending updated model parameters from the clients to the server over a noisy channel, i.e, noisy uplink communication. We will discuss these steps next.

A. NOISY DOWNLINK COMMUNICATION
The central server sends the global model parameter, w_k, to the set of r clients denoted by $\mathcal{S}_k$ chosen uniformly at random without replacement. Now due to disturbances and distortion in the communication channel, these clients receive a noisy version of the global model parameter, i.e.,

$$w_{k,0}^{(i)} = w_k + v_k^{(i)}, \quad (2)$$

where $v_k^{(i)} \in \mathbb{R}^d$ is the zero mean random downlink noise and $w_{k,0}^{(i)}$ is the received model to the ith client. Subsequently, the

SNR we get for the ith client for k^{th} downlink communication round can be written as,

$$\text{SNR}_{k,(i)}^{D} = \frac{\mathbb{E}[||w_k||^2]}{\mathbb{E}[||v_k^{(i)}||^2]}. \quad (3)$$

Since we assumed that $v_k^{(i)}$ is a zero mean noise, the variance can be written as:

$$N_{k,i}^2 := \mathbb{E}\left[\left\| v_k^{(i)} \right\|^2\right]. \quad (4)$$

B. CLIENT LEVEL COMPUTATION
Each client performs a local computation on its data using the updated noisy global model parameter. We use mini-batch SGD for training the model and updating the weights iteratively. This can be referenced from lines 6 to 9 in Algorithm 1 and written as,

$$w_{k,\tau+1}^{(i)} = w_{k,\tau}^{(i)} - \eta_k \widetilde{\nabla} f_i(w_{k,\tau}^{(i)}; \mathcal{B}_{k,\tau}^{(i)}),$$
$$\forall \tau = 0, 1, \ldots, E - 1, \quad (5)$$

where $\mathcal{B}_{k,\tau}^{(i)}$ represents the random batch of samples in client i for τ^{th} local iteration.

C. NOISY UPLINK COMMUNICATION
After the local computation, the clients in $\mathcal{S}_k$ send their local model to the central server. Similar to the downlink case, due to disturbances and distortion in the communication channel, the central server receives a noisy version of local weights which can be seen from line 10 in Algorithm 1 and is formulated as,

$$w_{k,0}^{(i)} - w_{k,E}^{(i)} + e_k^{(i)}, \quad (6)$$

where $e_k^{(i)} \in \mathbb{R}^d$ is a zero mean random noise. Subsequently, the SNR we get for the ith client for k^{th} uplink communication round can be written as,

$$\text{SNR}_{k,(i)}^{U} = \frac{\mathbb{E}[||w_{k,0}^{(i)} - w_{k,E}^{(i)}||^2]}{\mathbb{E}[||e_k^{(i)}||^2]}. \quad (7)$$

Since we assumed that $e_k^{(i)}$ is a zero mean noise, the variance can be depicted as:

$$U_{k,i}^2 := \mathbb{E}\left[\left\| e_k^i \right\|^2\right]. \quad (8)$$

Finally, the weights received from all the participating clients are aggregated and formulated in line 12 in Algorithm 1 as

$$w_{k+1} = w_k - \frac{1}{r} \sum_{i \in \mathcal{S}_k} (w_{k,0}^{(i)} - w_{k,E}^{(i)} + e_k^{(i)}), \quad (9)$$

and the process continues again for all the communication rounds.

D. MAIN ASSUMPTIONS

Before we start the analysis, the following is the set of assumptions that we make. Assumptions 1, 2, and 3 are standard used in analyzing FL setting [15], [34]. Assumption 4 is referred to as **Noise model** is also used in [29].

Assumption 1 (Smoothness): $\ell(\boldsymbol{x}, \boldsymbol{w})$ *is L-smooth with respect to* $\boldsymbol{w}$, *for all* $\boldsymbol{x}$. *Thus, each* $f_i(\boldsymbol{w})$ *($i \in [n]$) is L-smooth, and so is* $f(\boldsymbol{w})$.

$$\|\nabla f_i(\boldsymbol{w_1}) - \nabla f_i(\boldsymbol{w_2})\| \leq L\|\boldsymbol{w_1} - \boldsymbol{w_2}\|; \text{ for any } i, \boldsymbol{w_1}, \boldsymbol{w_2}.$$

Assumption 1 is a commonly used assumption in convergence analysis of gradient descent based algorithms, [35], [36] and it restricts the sudden change in gradients.

Assumption 2 (Non-Negativity): *Each* $f_i(\boldsymbol{w})$ *is non-negative and therefore,* $f_i^* \triangleq \min f_i(\boldsymbol{w}) \geq 0$.

Assumption 2 is a standard assumption made and is satisfied by most of the loss functions used in practice. However, if in case a loss function is negative the assumption can be achieved by simply adding a constant offset.

Assumption 3 (Bounded Variance): *The variance of the stochastic gradient for each client* i *is bounded:* $\mathbb{E}[\|\widetilde{\nabla} f_i(\boldsymbol{w}_{k,\tau}^{(i)}; \mathcal{B}_{k,\tau}^{(i)}) - \nabla f_i(\boldsymbol{w}_{k,\tau}^{(i)})\|^2] \leq \sigma^2$, $\forall i = 1, \ldots, n$, *where* $\mathcal{B}_{k,\tau}^{(i)}$ *represents the random batch of samples in client* i *for* τ^{th} *local iteration.*

The Assumption 3 is commonly used in analyzing the convergence of gradient descent-based algorithms, as seen in various works such as [29], [37], [38], and [39]. However, some other works have used a stricter assumption that assumes uniformly bounded stochastic gradients, i.e., $\mathbb{E}[\|\widetilde{\nabla} f_i(\boldsymbol{w}_{k,\tau}^{(i)}; \mathcal{B}_{k,\tau}^{(i)})\|^2] \leq \sigma^2$. This assumption is stronger than Assumption 3 and also does not hold true for convex loss functions as shown in [39].

Assumption 4 (Noise Model): *Both the downlink and uplink noise are independent and have zero mean i.e.,* $\mathbb{E}[\boldsymbol{v}_k^{(i)}] = 0$ *and* $\mathbb{E}[\boldsymbol{e}_k^i] = 0$ *and have a bounded variance i.e.,* $\mathbb{E}[\|\boldsymbol{e}_k^i\|^2] = \boldsymbol{U}_k^2 < \infty$ *and* $\mathbb{E}[\|\boldsymbol{v}_k^{(i)}\|^2] = \boldsymbol{N}_k^2 < \infty$.

In a noisy communication scenario, the modelling of noise as an additive white Gaussian noise (AWGN) is extremely common, refer [24], [30]. Here, Assumption 4 provides a weaker notion of AWGN and makes the problem more general and diverse.

III. NOISY-FedAvg: IMPROVED ANALYSIS

In this section, we describe the improved convergence analysis of the proposed algorithm. In addition to addressing complications arising from the simultaneous presence of both uplink and downlink noises, our analysis in this section is done without assuming *Bounded Client Dissimilarity* (BCD) that aims to limit the extent of client heterogeneity and is a frequently-used assumption in FL theory; see, e.g. [29], [34], [38] is the BCD assumption, i.e.,

$$\|\nabla f_i(\boldsymbol{w}) - \nabla f(\boldsymbol{w})\|^2 \leq G^2 \; \forall \, \boldsymbol{w}, i \in [n], \quad (10)$$

Algorithm 1 `Noisy-FedAvg`

1: **Input:** Initial point $\boldsymbol{w}_0$, # of communication rounds K, period E, learning rates $\{\eta_k\}_{k=0}^{K-1}$, and global batch size r.

2: **for** $k = 0, \ldots, K - 1$ **do**
3:      Server sends $\boldsymbol{w}_k$ to a set $\mathcal{S}_k$ of r clients chosen uniformly at random without replacement.
4:      **for** client $i \in \mathcal{S}_k$ **do**
5:          **Downlink communication:** Broadcasting $\boldsymbol{w}_k$ through a noisy downlink communication channel having zero mean. Set $\boldsymbol{w}_{k,0}^{(i)} = \boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}$, where $\boldsymbol{v}_k^{(i)}$ is the downlink noise.
6:          **for** $\tau = 0, \ldots, E - 1$ **do**
7:              Pick a random batch of samples in client i, $\mathcal{B}_{k,\tau}^{(i)}$. Compute the stochastic gradient of f_i at $\boldsymbol{w}_{k,\tau}^{(i)}$ over $\mathcal{B}_{k,\tau}^{(i)}$, viz. $\widetilde{\nabla} f_i(\boldsymbol{w}_{k,\tau}^{(i)}; \mathcal{B}_{k,\tau}^{(i)})$.
8:              Update $\boldsymbol{w}_{k,\tau+1}^{(i)} = \boldsymbol{w}_{k,\tau}^{(i)} - \eta_k \widetilde{\nabla} f_i(\boldsymbol{w}_{k,\tau}^{(i)}; \mathcal{B}_{k,\tau}^{(i)})$.
9:          **end for**
10:          **Uplink communication:** $(\boldsymbol{w}_k - \boldsymbol{w}_{k,E}^{(i)})$ goes to the server through a noisy uplink communication channel having zero mean. So, send $(\boldsymbol{w}_{k,0}^{(i)} - \boldsymbol{w}_{k,E}^{(i)} + \boldsymbol{e}_k^{(i)})$, where $\boldsymbol{e}_k^{(i)}$ is the uplink noise.
11:      **end for**
12:      Update $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \frac{1}{r} \sum_{i \in \mathcal{S}_k} (\boldsymbol{w}_{k,0}^{(i)} - \boldsymbol{w}_{k,E}^{(i)} + \boldsymbol{e}_k^{(i)})$.
13: **end for**

---

where G is a large constant. Furthermore, in contrast to [29] and [38], we do not make any assumption about the *strong convexity* of the loss function.

A. NOISY-SGD

To give more insight into the analysis of `Noisy-FedAvg` and its implications we first consider a fictitious scenario where a noisy version of SGD is employed to minimize a stochastic, non-convex, and L-smooth function with the following update

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta[\boldsymbol{e}_t + \widetilde{\nabla} f(\boldsymbol{w}_t + \boldsymbol{v}_t; \mathcal{B}_t)], \quad (11)$$

where $\boldsymbol{e}_t$ and $\boldsymbol{v}_t$ can be thought of as uplink and downlink noise, respectively. The purpose of this analysis is to shed light on the effect of noise on SGD-based FL algorithms.

Theorem 5 (Smooth Non-Convex Case for `Noisy-SGD`*):* Let $f : \mathbb{R}^d \to \mathbb{R}$ be a L-smooth non-convex function and $f^* := \inf_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w})$. Consider the noisy-SGD method with the update in eq. (11). Let $\boldsymbol{e}_t$ and $\boldsymbol{v}_t$ satisfy Assumption 4 and the stochastic gradient satisfy Assumption 3. If, $\eta_t = \eta$ and $\eta \leq \frac{1}{L}$, then for all $t \in \{0, \ldots, T - 1\}$ noisy-SGD satisfies

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_{\{\mathcal{B}_t, \boldsymbol{e}_t, \boldsymbol{v}_t\}_{t=0}^{T-1}}[\|\nabla f(\boldsymbol{w}_t)\|^2]]$$

$$\leq \frac{2\big(f(w_0) - f^*\big)}{T\eta}$$

$$+ \eta L\sigma^2 + \underbrace{\frac{L^2}{T}\sum_{t=0}^{T-1} N_t^2}_{\text{Term I}} + \underbrace{\frac{\eta L}{T}\sum_{t=0}^{T-1} U_t^2}_{\text{Term II}}. \quad (12)$$

Proof: Using L-smoothness assumption we can obtain,

$$f(w_{t+1}) \leq f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L}{2}||w_{t+1} - w_t||^2. \quad (13)$$

Using eq. (11) in eq. (13) yields

$$f(w_{t+1}) \leq f(w_t) - \underbrace{\eta\langle \nabla f(w_t), e_t + \widetilde{\nabla} f(w_t + v_t; \mathcal{B}_t)\rangle}_{(A)}$$

$$+ \underbrace{\frac{\eta^2 L}{2}||e_t + \widetilde{\nabla} f(w_t + v_t; \mathcal{B}_t)||^2}_{(B)}. \quad (14)$$

Using A:

$$A = \eta\langle \nabla f(w_t), e_t + \widetilde{\nabla} f(w_t + v_t; \mathcal{B}_t)\rangle$$
$$= \eta\langle \nabla f(w_t), e_t \rangle + \eta\langle \nabla f(w_t), \widetilde{\nabla} f(w_t + v_t; \mathcal{B}_t)\rangle \quad (15)$$

In eq. (15), taking expectation with respect to e_t, will result in eq. (16).

$$\mathbb{E}_{e_t}[A] = \eta\langle \nabla f(w_t), \widetilde{\nabla} f(w_t + v_t; \mathcal{B}_t)\rangle \quad (16)$$

Using B:

$$B = \frac{\eta^2 L}{2}||e_t + \widetilde{\nabla} f(w_t + v_t; \mathcal{B}_t) - \nabla f(w_t + v_t)$$
$$+ \nabla f(w_t + v_t)||^2 \quad (17)$$
$$= \frac{\eta^2 L}{2}||e_t||^2 + \frac{\eta^2 L}{2}||\widetilde{\nabla} f(w_t + v_t; \mathcal{B}_t) - \nabla f(w_t + v_t)||^2$$
$$+ \frac{\eta^2 L}{2}||\nabla f(w_t + v_t)||^2 + 2\langle e_t, \nabla f(w_t + v_t)\rangle$$
$$+ 2\langle e_t, \widetilde{\nabla} f(w_t + v_t; \mathcal{B}_t) - \nabla f(w_t + v_t)\rangle$$
$$+ 2\langle \widetilde{\nabla} f(w_t + v_t; \mathcal{B}_t) - \nabla f(w_t + v_t), \nabla f(w_t + v_t)\rangle \quad (18)$$

Taking the expectation with respect to e_t which has zero mean, alongside the independence assumption of noises in eq. (18), we can re-write B as

$$\mathbb{E}_{e_t}[B] = \frac{\eta^2 L}{2}||\widetilde{\nabla} f(w_t + v_t; \mathcal{B}_t) - \nabla f(w_t + v_t)||^2$$
$$+ 2\langle \widetilde{\nabla} f(w_t + v_t; \mathcal{B}_t) - \nabla f(w_t + v_t), \nabla f(w_t + v_t)\rangle$$
$$+ \frac{\eta^2 L}{2}U_t^2 + \frac{\eta^2 L}{2}||\nabla f(w_t + v_t)||^2 \quad (19)$$

Now, putting the results of eqs. (16) and (19) in eq. (14) along with taking the expectation with respect to data, we get

$$\mathbb{E}_{\mathcal{B}_t, e_t}[f(w_{t+1})] \leq \mathbb{E}_{\mathcal{B}_t, e_t}[f(w_t)] + \frac{\eta^2 L}{2}U_t^2 + \frac{\eta^2 L}{2}\sigma^2$$

$$- \eta\mathbb{E}_{\mathcal{B}_t, e_t}[\langle \nabla f(w_t), \nabla f(w_t + v_t)\rangle]$$
$$+ \frac{\eta^2 L}{2}\mathbb{E}_{\mathcal{B}_t, e_t}[||\nabla f(w_t + v_t)||^2]. \quad (20)$$

For any 2 vectors a and b, we have that

$$-\langle a, b \rangle = \frac{1}{2}(||a - b||^2 - ||a||^2 - ||b||^2). \quad (21)$$

Using this in eq. (20) we get

$$\mathbb{E}_{\mathcal{B}_t, e_t}[f(w_{t+1})] \leq \mathbb{E}_{\mathcal{B}_t, e_t}[f(w_t)] + \frac{\eta^2 L}{2}(U_t^2 + \sigma^2)$$
$$+ \frac{\eta}{2}\mathbb{E}_{\mathcal{B}_t, e_t}\Big[||\nabla f(w_t) - \nabla f(w_t + v_t)||^2$$
$$- ||\nabla f(w_t + v_t)||^2 - ||\nabla f(w_t)||^2\Big]$$
$$+ \frac{\eta^2 L}{2}\mathbb{E}_{\mathcal{B}_t, e_t}[||\nabla f(w_t + v_t)||^2]. \quad (22)$$

If $\eta \leq \frac{1}{L}$, we can drop $\mathbb{E}_{\mathcal{B}_t, e_t}[||\nabla f(w_t + v_t)||^2]$ as it will appear with a negative sign in the RHS of eq. (22). Consequently, using L-smoothness yields

$$\mathbb{E}_{\mathcal{B}_t, e_t}[f(w_{t+1})]$$
$$\leq \mathbb{E}_{\mathcal{B}_t, e_t}[f(w_t)] + \frac{\eta^2 L}{2}(U_t^2 + \sigma^2)$$
$$+ \frac{\eta}{2}\mathbb{E}_{\mathcal{B}_t, e_t}\Big[L^2||(v_t)||^2 - ||\nabla f(w_t)||^2\Big]. \quad (23)$$

Taking expectation w.r.t. v_t we have

$$\mathbb{E}_{\mathcal{B}_t, e_t, v_t}[f(w_{t+1})]$$
$$\leq \mathbb{E}_{\mathcal{B}_t, e_t, v_t}[f(w_t)] + \frac{\eta}{2}L^2 N_t^2$$
$$- \frac{\eta}{2}\mathbb{E}_{\mathcal{B}_t, e_t, v_t}[||\nabla f(w_t)||^2] + \frac{\eta^2 L}{2}(U_t^2 + \sigma^2). \quad (24)$$

Summing the above equation for $t = 0, 1, \ldots, T - 1$ and dividing both sides by $T\eta/2$ we obtain,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\{\mathcal{B}_t, e_t, v_t\}_{t=0}^{T-1}}[||\nabla f(w_t)||^2]$$
$$\leq \frac{2\big(f(w_0) - f(w_T)\big)}{T\eta}$$
$$+ \eta L\sigma^2 + \frac{L^2}{T}\sum_{t=0}^{T-1} N_t^2 + \frac{\eta L}{T}\sum_{t=0}^{T-1} U_t^2. \quad (25)$$

Now using the fact that $f^* \leq f(w_T)$ in the equation above, we obtain the stated result in eq. (12). ∎

The implication of eq. (12) is that the downlink noise (Term I) is more degrading than the uplink noise (Term II) given that the effect of the latter on the convergence can be controlled by η. That is, uplink noise slows the convergence rate while the downlink noise may inhibit the convergence. The following corollary describes a SNR control strategy that aims to recover the rate of noise-free SGD by pushing the noise-driven terms, i.e., Terms I and II in the RHS of eq. (12), to the higher-order term. In the context of this paper, a higher-order term deviant from the dominant term is one that does not control the order of convergence error.

Corollary 6: If $\eta \leq \frac{1}{L}$ and $\eta = \mathcal{O}(\frac{1}{\sqrt{T}})$ and a SNR control strategy is employed such that $\frac{1}{T}\sum_{t=0}^{T-1} U_t^2 = \mathcal{O}(T^{-\delta_1})$ and $\frac{1}{T}\sum_{t=0}^{T-1} N_t^2 = \mathcal{O}(\frac{1}{T^{0.5+\delta_2}})$ for some $\delta_1, \delta_2 > 0$, the dominant term in the convergence error of `Noisy-SGD` *will be $\mathcal{O}(\frac{1}{\sqrt{T}})$ which is independent of noise characteristics and hence similar to the noise-free case of SGD.*

B. NOISY-FedAvg

In what follows, we build upon Theorem 5 to present Theorem 7, which holds for both partial and full client participation, IID, and non-IID data distribution.

Theorem 7 (Smooth Non-Convex Case for `Noisy-FedAvg`*): Let Assumptions 1, 2, 3, 4 holds for* `Noisy-FedAvg` *(Algorithm 1). In* `Noisy-FedAvg`*, set $\eta_k = \frac{1}{\gamma LE}\sqrt{\frac{r}{K}}$ for all k, where $\gamma > 4$ is a universal constant. Define a distribution $\mathbb{P}$ for $k \in \{0, \ldots, K-1\}$ such that $\mathbb{P}(k) = \frac{(1+\zeta)^{(K-1-k)}}{\sum_{k=0}^{K-1}(1+\zeta)^k}$ where $\zeta := 8\eta^2 L^2 E^2\left(\frac{(n-r)}{r(n-1)} + \frac{2\eta LE}{3}\right)$. Sample k^* from $\mathbb{P}$ uniformly. Then, for $K \geq \max\left(\frac{1024 r^3}{9\gamma^2}(\frac{1}{\gamma^2-16})^2, \frac{4r}{\gamma^2}\right)$, eq. (26), as shown at the bottom of the page.*

In the Theorem 7 the expectation is w.r.t the choice of clients, data, uplink, and downlink noises. The Terms I and III in the theorem above depict the effects of uplink and downlink noises respectively. Furthermore, Term II is a direct consequence of Assumption 3, stemming from the stochastic gradients of the clients. Now, from Theorem 7, mirroring the result of Theorem 5, we can observe that the uplink noise is not dominant compared to the downlink noise. As we will discuss in Section V, our tight analysis in establishing 7 is verified numerically as well by showing that uplink noise's impact is not as detrimental as downlink noise.

Remark 8: Theorem 7 is derived without using the restrictive assumption of BCD, in eq. (10). Now, to clarify why this assumption is restrictive, let us consider a toy example of a univariate quadratic function. The following notations hold their usual meaning as defined in Section II.

$$f_i(w) = \frac{1}{2}(w^2), \forall i = 1, \ldots, n-1 \qquad (27)$$

$$f_n(w) = w^2 \qquad (28)$$

By using eq. (1), a global objective function can be formulated as,

$$f(w) = \frac{1}{2n}(n+1)w^2 \qquad (29)$$

For $i = n$, by taking the gradient of eqs. (28) and (29), and putting them back in eq. (10), we get

$$\|\nabla f_i(w) - \nabla f(w)\|^2$$
$$\leq G^2$$
$$\implies \|2w - \frac{n+1}{n}w\|^2 \leq G^2, \quad \text{where } i = n$$
$$\implies \|w\|^2\|1 - \frac{1}{n}\|^2 \leq G^2 \qquad (30)$$

We can infer from eq. (30) that the inequality does not hold for every $w \in \mathbb{R}$, given a fixed G and hence it makes the assumption restrictive and somewhat unrealistic.

Remark 9: Before we start with the proof of the Theorem 7, we would like to emphasize that the theorem provides an upper bound on the performance for the FedAvg algorithm, which depends on the noise characteristics. It is an interesting future work to investigate the effect of noise on the lower bound that essentially bounds the performance of any algorithm in this scenario and see if one could use such lower bounds towards SNR scaling (for more details on SNR scaling refer Section IV).

Proof: The proof is motivated by the approach taken in [15]. However, with the inclusion of downlink and uplink noises, the first local update (refer eq. (2) and the model update (refer eq. (6)) are considerably different which makes the analysis significantly different and more involved. As will be outlined shortly, the proof relies on careful treatment of the first local and model updates using new techniques.

To commence the proof, using Lemma 11, for $\eta_k LE \leq \frac{1}{2}$, we can bound the per-round progress as:

$$\mathbb{E}[f(w_{k+1})]$$
$$\leq \mathbb{E}[f(w_k)] - \frac{\eta_k(E-1)}{2}\mathbb{E}[\|\nabla f(w_k)\|^2]$$
$$+ 4\eta_k^2 LE^2\left(\frac{(n-r)}{r(n-1)} + \frac{2}{3}\eta_k LE\right)\left(\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\nabla f_i(w_k)\|^2]\right)$$

$$\mathbb{E}[\|\nabla f(w_{k^*})\|^2]$$
$$\leq \frac{8\gamma L f(w_0)}{\sqrt{rK}} + \underbrace{\frac{4}{\gamma E^2 K\sqrt{rK}}\sum_{k=0}^{K-1} U_k^2}_{\text{Term I}}$$
$$+ \underbrace{\frac{4}{\gamma E}\sqrt{\frac{r}{K}}\left(\frac{1}{\gamma n}\sqrt{\frac{r}{K}}(1 + \frac{2nE}{3} + n) + \frac{1}{r} + \frac{(n-r)}{r(n-1)}\right)\sigma^2}_{\text{Term II}}$$
$$+ \underbrace{\frac{4L^2}{EK}\left(1 + 4E + \frac{2}{\gamma E}\sqrt{\frac{r}{K}}(1 + 2E^2\{\frac{3}{\gamma E^2}\sqrt{\frac{r}{K}} + 2(2 + \frac{3}{\gamma^2 E^2}\frac{r}{K})(\frac{2}{3\gamma}\sqrt{\frac{r}{K}} + \frac{(n-r)}{r(n-1)})\})\right)\sum_{k=0}^{K-1} N_k^2}_{\text{Term III}}. \qquad (26)$$

$$+ \eta_k^2 LE\Big(\frac{\eta_k LE}{n}\Big(1 + \frac{2nE}{3} + n\Big) + \frac{1}{r} + \frac{(n-r)}{r(n-1)}\Big)\sigma^2$$
$$+ \frac{\eta_k^2 L}{2r}\frac{1}{n}\sum_{i\in[n]} U_{k,i}^2 + \frac{\eta_k L^2}{2}\Big(1 + 2\eta_k L + 4E\{1 + 3\eta_k^2 L^2$$
$$+ 2\eta_k LE(2 + 3\eta_k^2 L^2)(\frac{2}{3}\eta_k LE + \frac{(n-r)}{r(n-1)})\}\Big)\frac{1}{n}\sum_{i\in[n]} N_{k,i}^2.$$
(31)

Now using the L-smoothness and non-negativity of the f_i's, we get:

$$\sum_{i\in[n]}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2] \leq \sum_{i\in[n]} 2L(\mathbb{E}[f_i(\boldsymbol{w}_k)] - f_i^*)$$
$$\leq 2nL\mathbb{E}[f(\boldsymbol{w}_k)] - 2L\sum_{i\in[n]} f_i^*$$
$$\leq 2nL\mathbb{E}[f(\boldsymbol{w}_k)]$$

Putting this result in eq. (31), we get for a constant learning rate of $\eta_k = \eta$, $U_{k,i}^2 = U_k^2$ and $N_{k,i}^2 = N_k^2$:

$$\mathbb{E}[f(\boldsymbol{w}_{k+1})]$$
$$\leq \Big(1 + 8\eta^2 L^2 E^2\Big(\frac{(n-r)}{r(n-1)} + \frac{2\eta LE}{3}\Big)\Big)\mathbb{E}[f(\boldsymbol{w}_k)]$$
$$- \frac{\eta(E-1)}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2] + \eta^2 LE\Big(\frac{\eta LE}{n}(1 + \frac{2nE}{3}$$
$$+ n) + \frac{1}{r} + \frac{(n-r)}{r(n-1)}\Big)\sigma^2 + \frac{\eta^2 L}{2r}\frac{1}{n}\sum_{i\in[n]} U_k^2$$
$$+ \frac{\eta L^2}{2}\Big(1 + 2\eta L + 4E\{1 + 3\eta^2 L^2 + 2\eta LE(2 + 3\eta^2 L^2)$$
$$(\frac{2\eta LE}{3} + \frac{(n-r)}{r(n-1)})\}\Big)\frac{1}{n}\sum_{i\in[n]} N_k^2.$$

For ease of notation, define $\zeta := 8\eta^2 L^2 E^2\Big(\frac{(n-r)}{r(n-1)} + \frac{2\eta LE}{3}\Big)$, $\zeta_2 := \Big(\frac{\eta LE}{n}\Big(1 + \frac{2nE}{3} + n\Big) + \frac{1}{r} + \frac{(n-r)}{r(n-1)}\Big)$ and $\zeta_3 := \Big(1 + 2\eta L + 4E\{1 + 3\eta^2 L^2 + 2\eta LE(2 + 3\eta^2 L^2)(\frac{2\eta LE}{3} + \frac{(n-r)}{r(n-1)})\}\Big)$. Then, unfolding the recursion of the equation above from $k = 0$ through to $k = K - 1$, we get:

$$\mathbb{E}[f(\boldsymbol{w}_K)] \leq (1 + \zeta)^K f(\boldsymbol{w}_0)$$
$$- \frac{\eta(E-1)}{2}\sum_{k=0}^{K-1}(1 + \zeta)^{(K-1-k)}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]$$
$$+ \eta^2 LE\zeta_2\sigma^2\sum_{k=0}^{K-1}(1 + \zeta)^{(K-1-k)}$$
$$+ \frac{\eta^2 L}{2r}\sum_{k=0}^{K-1} U_k^2(1 + \zeta)^{(K-1-k)}$$
$$+ \frac{\eta L^2}{2}\zeta_3\sum_{k=0}^{K-1} N_k^2(1 + \zeta)^{(K-1-k)}.$$
(32)

Let us define $p_k := \frac{(1+\zeta)^{(K-1-k)}}{\sum_{k'=0}^{K-1}(1+\zeta)^{(K-1-k')}}$. Then, re-arranging eq. (32) and using the fact that $\mathbb{E}[f(\boldsymbol{w}_K)] \geq 0$ and

$\frac{\eta(E-1)}{2} > \frac{\eta E}{4}$, we get:

$$\sum_{k=0}^{K-1} p_k\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]$$
$$\leq \frac{4(1+\zeta)^K f(\boldsymbol{w}_0)}{\eta E\sum_{k'=0}^{K-1}(1+\zeta)^{k'}} + 4\eta L\zeta_2\sigma^2$$
$$+ \frac{2\eta L}{rE}\frac{\sum_{k=0}^{K-1} U_k^2(1+\zeta)^{(K-1-k)}}{\sum_{k'=0}^{K-1}(1+\zeta)^{k'}}$$
$$+ \frac{2L^2\zeta_3}{E}\frac{\sum_{k=0}^{K-1} N_k^2(1+\zeta)^{(K-1-k)}}{\sum_{k'=0}^{K-1}(1+\zeta)^{k'}}$$
(33)
$$= \frac{4\zeta f(\boldsymbol{w}_0)}{\eta E(1 - (1+\zeta)^{-K})} + 4\eta L\zeta_2\sigma^2$$
$$+ \frac{2\eta L}{rE}\frac{\zeta\sum_{k=0}^{K-1} U_k^2}{(1+\zeta) - (1+\zeta)^{-K+1}}$$
$$+ \frac{2L^2\zeta_3}{E}\frac{\zeta\sum_{k=0}^{K-1} N_k^2}{(1+\zeta) - (1+\zeta)^{-K+1}}$$
(34)

where the eq. (34) follows by using the fact that $\sum_{k'=0}^{K-1}(1 + \zeta)^{k'} = \frac{(1+\zeta)^K - 1}{\zeta}$ and Hölder's Inequality. Now,

$$(1 + \zeta)^{-K} < 1 - \zeta K + \zeta^2\frac{K(K+1)}{2} < 1 - \zeta K + \zeta^2 K^2$$
$$\implies 1 - (1 + \zeta)^{-K} > \zeta K(1 - \zeta K). \quad (35)$$

Also,

$$(1 + \zeta)^{-K+1} < 1 + \zeta(-K+1) + \zeta^2\frac{(-K)(-K+1)}{2}$$
$$< (1 + \zeta) - \zeta K + \zeta^2 K^2$$
$$\implies (1 + \zeta) - (1 + \zeta)^{-K+1} > \zeta K(1 - \zeta K).$$
(36)

Plugging eqs. (35) and (36) with ζ_2 and ζ_3 in eq. (34), we have for $\zeta K < 1$:

$$\sum_{k=0}^{K-1} p_k\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]$$
$$\leq \frac{4f(\boldsymbol{w}_0)}{\eta EK(1 - \zeta K)} + 4\eta LE\Big(\frac{\eta L}{n}\Big(1$$
$$+ \frac{2nE}{3} + n\Big) + \frac{1}{rE} + \frac{(n-r)}{r(n-1)E}\Big)\sigma^2 + \frac{2\eta L}{rE}\frac{\sum_{k=0}^{K-1} U_k^2}{K(1 - \zeta K)}$$
$$+ \frac{2L^2}{EK(1 - \zeta K)}\Big(1 + 2\eta L + 4E\{1 + 3\eta^2 L^2 + 2\eta LE(2$$
$$+ 3\eta^2 L^2)(\frac{2\eta LE}{3} + \frac{(n-r)}{r(n-1)})\}\Big)\sum_{k=0}^{K-1} N_k^2.$$
(37)

In this case, note that the optimal step size will be $\eta = \mathcal{O}(\frac{1}{LE\sqrt{K}})$, even for $r = n$.

So let us pick $\eta = \frac{1}{\gamma LE}\sqrt{\frac{r}{K}}$, where γ is some constant such that $\gamma > 4$. Note that we need to have $\eta LE \leq \frac{1}{2}$; this happens for $K \geq \frac{4r}{\gamma^2}$. Further, let us ensure $\zeta K < \frac{1}{2}$; this

happens for $K \geq \frac{1024r^3}{9\gamma^2}(\frac{1}{\gamma^2-16})^2$. Thus, we should have $K \geq$ $\max\left(\frac{1024r^3}{9\gamma^2}(\frac{1}{\gamma^2-16})^2, \frac{4r}{\gamma^2}\right)$. Putting $\eta = \frac{1}{\gamma LE}\sqrt{\frac{r}{K}}$ in eq. (37) and also using $1 - \zeta K \geq \frac{1}{2}$, we get:

$$
\sum_{k=0}^{K-1} p_k \mathbb{E}[\|\nabla f(\mathbf{w}_k)\|^2]
$$

$$
\leq \frac{8\gamma L f(\mathbf{w}_0)}{\sqrt{rK}}
$$

$$
+ \frac{4}{\gamma E}\sqrt{\frac{r}{K}}\left(\frac{1}{\gamma n}\sqrt{\frac{r}{K}}(1 + \frac{2nE}{3} + n) + \frac{1}{r} + \frac{(n-r)}{r(n-1)}\right)\sigma^2
$$

$$
+ \frac{4L^2}{EK}\left(1 + 4E + \frac{2}{\gamma E}\sqrt{\frac{r}{K}}(1 + 2E^2\{\frac{3}{\gamma E^2}\sqrt{\frac{r}{K}} + 2(2\right.
$$

$$
\left. + \frac{3}{\gamma^2 E^2}\frac{r}{K})(\frac{2}{3\gamma}\sqrt{\frac{r}{K}} + \frac{(n-r)}{r(n-1)})\}\right)\sum_{k=0}^{K-1} N_k^2
$$

$$
+ \frac{4}{\gamma E^2 K\sqrt{rK}}\sum_{k=0}^{K-1} U_k^2. \tag{38}
$$

This finishes the proof. ∎

IV. THEORY GUIDED SNR CONTROL

Theorem 7 provides us with actionable insight into the effect of uplink and downlink noise on model convergence, i.e., the inherent asymmetry of their adverse effect on the performance of FL algorithms. One strategy to improve the convergence properties of the model in noisy settings is to boost the SNR of the communicated messages (see, e.g. [29] and the references therein). In this section, we aim to establish an improved SNR control strategy following the improved analysis presented in Theorem 7. First, we stated the following corollary for an easier-to-interpret result.

Corollary 10: Instate the notation and hypotheses of Theorem 7. Also, let $U_k^2 \leq e^2$ and $N_k^2 \leq v^2$ to be the maximum bounded variances for all k. Then, if $K = \Omega(r^3)$,

$$
\mathbb{E}[\|\nabla f(\mathbf{w}_{k^*})\|^2] = \mathcal{O}\left(\frac{1}{E}\sqrt{\frac{r}{K}}\sigma^2 + \frac{1}{E^2\sqrt{rK}}e^2 + v^2\right). \tag{39}
$$

A. IMPLICATIONS

Following Corollary 10, we can observe that the term corresponding to uplink noise scale as $\mathcal{O}(\frac{1}{E^2\sqrt{rK}})$, while the term corresponding to downlink noise is $\mathcal{O}(1)$. We can visualize the impact of both uplink and downlink noises on the MNIST dataset (refer Section V-B for the details about the model architecture and other parameters) in Figure 2. From figs. 2b and 2d, we can confirm that any changes in the number of participating clients, i.e., r, and the local number of iterations, i.e., E does not make any difference and it scales as $\mathcal{O}(1)$. Similarly, from figs. 2a and 2c, we inspect that the term corresponding to uplink noise scales as $\mathcal{O}(\frac{1}{E^2\sqrt{rK}})$. This tells us that the impact of uplink and downlink noises on convergence errors is different. Using this result, we propose to employ SNR control strategies such that the effect of both

the noises appear as *higher-order terms*, not as dominant terms. Hence to curtail the impact of the noises, we want the order of the terms corresponding to the uplink and downlink noise to be $\mathcal{O}(\frac{1}{E^{1+\delta_1}K^{\frac{1}{2}+\delta_2}})$, for some $\delta_1, \delta_2 > 0$. For instance, in what follows we will adopt a strategy such that $\delta_1 = 1$ and $\delta_2 = 0.5$.

Since we have already established that the effects of both uplink and downlink noises are different, we need to employ different scaling policies for these noises. Hence, for the model to converge to an ϵ−stationary point like in `FedAvg`, we need to scale down the downlink noise by $\Omega(E^2k)$ and uplink noise by $\Omega(\sqrt{k})$. These scaling rates result in requiring considerably less power resources compared to the prior work, e.g. [29]. Putting the requirement of strong convexity aside, the proposed policy in [29], while consuming more power resources[2] ensures that the model converges, albeit the dominant term in the rate will depend on noise statistics. However, employing our strategy ensures that noise appears merely as a higher-order term, which means that for a large number of communication rounds, the difference between noisy and noise-free `FedAvg` will be negligible. In the discussions above we talk about the necessity of SNR scaling to ensure that the performance gracefully depends on the noise. There are works such as [20] and [24] and others that provide a practical perspective towards designing distributed systems and it is an interesting future work to implement the theoretical findings of this work in practical systems and come up with new design paradigms. A simple setting is a scenario where a set of UAVs (acting as clients in FL) need to communicate with the base station (acting as the central server in FL). Our paper's main findings indicate that the effect of the downlink channel noise is particularly detrimental, and thus, an effective strategy would involve boosting the power of the signal sent from the base station to the UAVs.

V. VERIFYING EXPERIMENTS

In this section, we demonstrate the efficacy and validity of the proposed theory through empirical analysis. For this purpose, we devise two categories of experiments, (i) synthetic experiments and (ii) deep learning experiments.

A. SYNTHETIC EXPERIMENT

In the synthetic experiment, we train a linear regression model with m = 15000 samples. The samples $\{(\mathbf{x}_j, y_j)\}_{j=1}^m$ are generated based on the model $y_j = \langle \theta^*, \mathbf{x}_j \rangle + c_j$, where $\theta^* \in \mathbb{R}^{60}$, the jth input $\mathbf{x}_j \sim \mathcal{N}(0, I_{60})$, and noise $c_j \sim \mathcal{N}(0, 0.05)$. This dataset is generated such that the (*samples × features*) matrix has the ℓ_2 norm of its Hessian equal to 1. These samples are then distributed over 50 clients resulting in 300 samples/client. Also, we use the mean squared error loss function. To conduct this numerical experiment we use $n = 50$ clients and set the values of $\gamma = 18$ (see Theorem 7),

---

[2]Considering non-convex, L-smooth problems, the result in [29] seems to require $\Omega(k)$ scaling for both uplink and downlink noises.
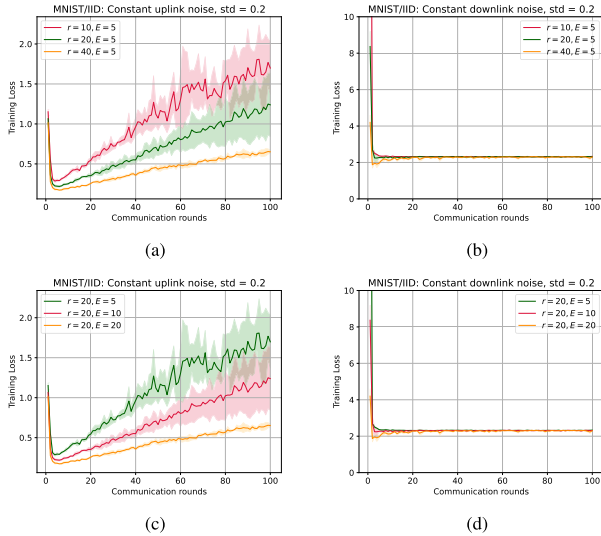
FIGURE 2. Effect of *r* (top 2 figures) and *E* (bottom 2 figures) on uplink and downlink noise.
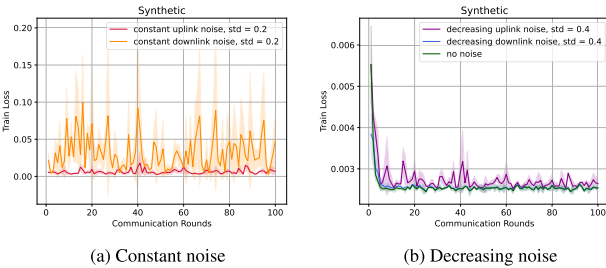

FIGURE 3. Linear regression: Comparing the impact of uplink and downlink noise with (left) and without (right) SNR control.
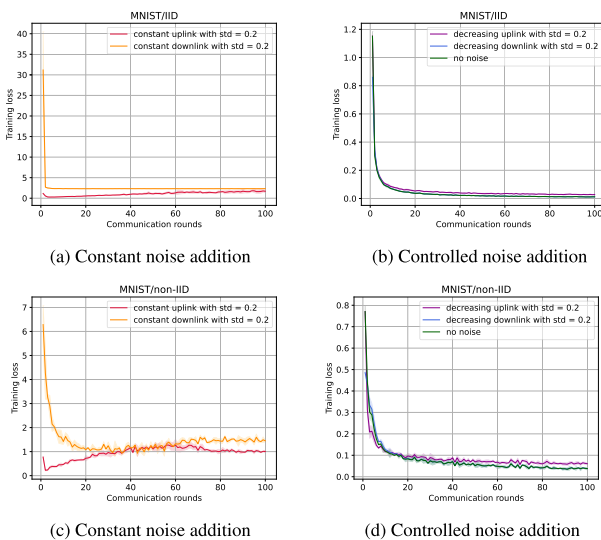

FIGURE 4. Train loss vs. communication rounds for the IID and non-IID data distribution of the MNIST dataset. The plots (4a) and (4c) present the constant noise addition setting. We can see that the theory is verified from the plots. Similarly, the plots (4b) and (4d) show the effect of the proposed SNR control strategy.

$L = 1$, $E = 5$ and $K = 100$ and BS (local batch size) = 16. In each round, 20% of the clients participate based on random
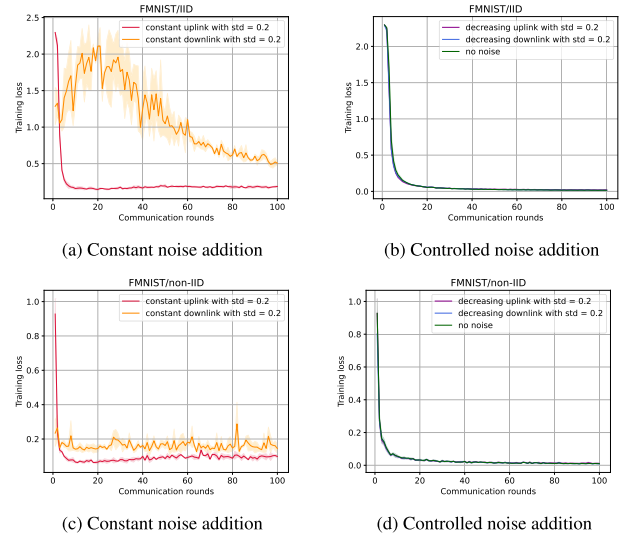

FIGURE 5. Train loss vs. communication rounds for the IID and non-IID data distribution of the FMNIST dataset. The plots (5a) and (5c) present the constant noise addition setting. We can see that the theory is verified from the plots. Similarly, the plots (5b) and (5d) show the effect of the proposed SNR control strategy.
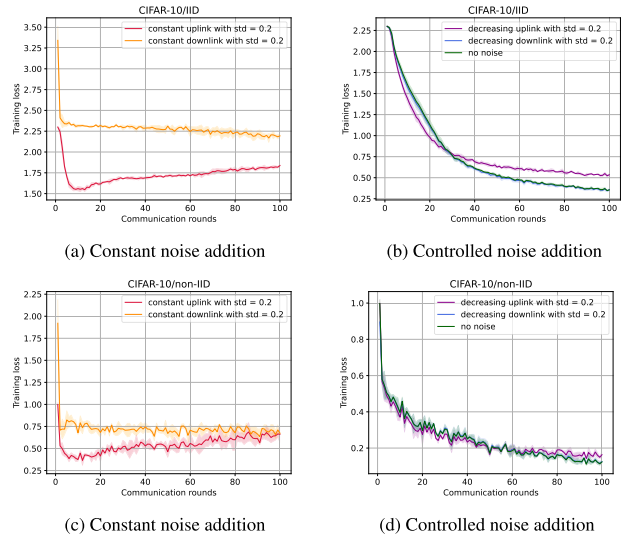

FIGURE 6. Train loss vs. communication rounds for the IID and non-IID data distribution of the CIFAR-10 dataset. The plots (6a) and (6c) present the constant noise addition setting. We can see that the theory is verified from the plots. Similarly, the plots (6b) and (6d) show the effect of the proposed SNR control strategy.

selection, which leads to $r = 10$. Now from Theorem 7, we have $\eta_k = \frac{1}{\gamma LE} \sqrt{\frac{r}{K}}$, i.e. $\eta_k = 0.0035$.

We first consider a constant noise setting where we add both uplink and downlink noise to the communicated messages where the noises are sampled from a Gaussian distribution having zero mean i.e., $e_k^{(i)} \sim \mathcal{N}(0, \upsilon^2)$ and $v_k^{(i)} \sim \mathcal{N}(0, \upsilon^2)$, where $\upsilon = 0.2$. We visualize the impact of adding noises in Figure 3a; as the figure demonstrates, consistent with the result of Theorem 7, the effect of downlink noise is more severe than the uplink noise and results in model
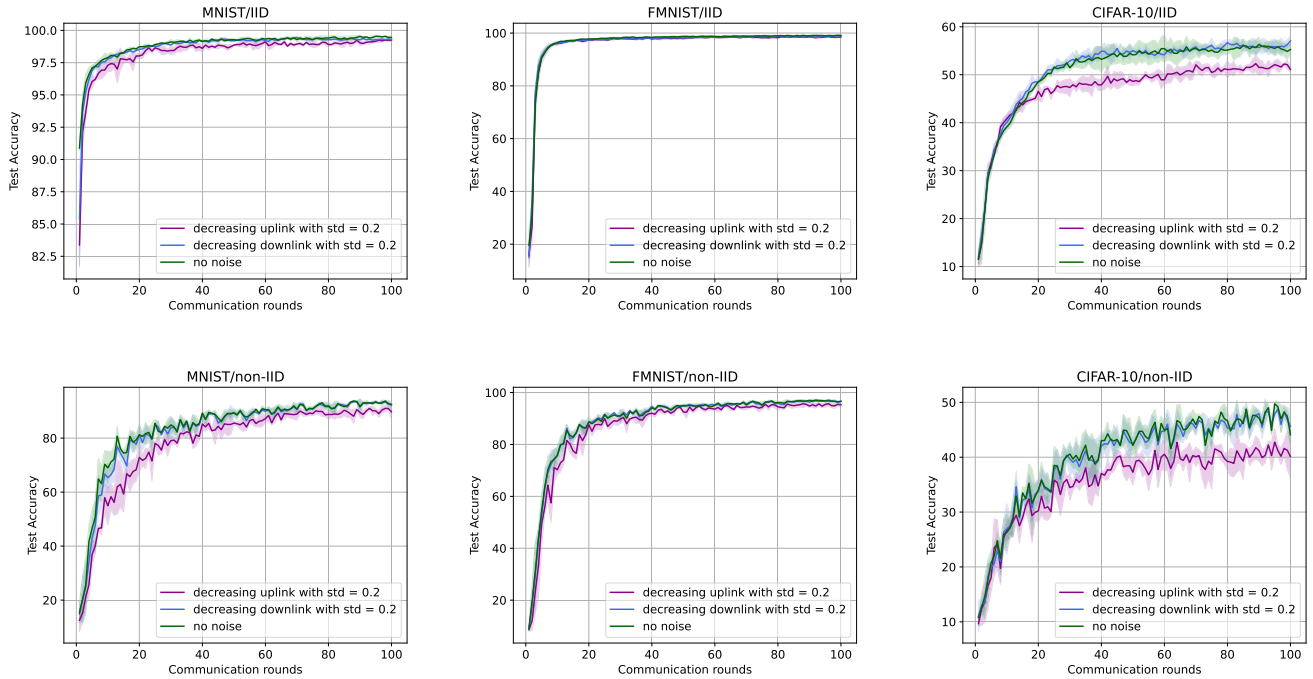
FIGURE 7. Test accuracy vs. communication rounds for the IID (top 3 figures) and non-IID (bottom 3 figures) data distribution of the MNIST, FMNIST, and CIFAR-10 datasets.

divergence. Furthermore, the adverse effect of noise increases as υ increases.

Now, we test the efficacy of the proposed SNR control strategy in Section IV. In particular, since we already established using Theorem 7 that the effect of downlink noise is more degrading than uplink noise, we can utilize different SNR scaling policies to save on power resources while alleviating the effect of noise. Hence, we scale the downlink and uplink noises by $\Omega(\frac{1}{E^2 k})$ and $\Omega(\frac{1}{\sqrt{k}})$, respectively. We can observe the results from Fig. 3b and see how it converges almost in tandem with the noise-free case, as predicted by Corollary 10.

B. DEEP LEARNING EXPERIMENT

To check the validity of our theory on real-world datasets, we run deep learning experiments on the MNIST, CIFAR-10, and Fashion-MNIST (FMNIST) datasets, both for IID and non-IID settings. Before we start with the different experimental setups we clarify that these experiments are not designed to provide benchmark results on the corresponding dataset but rather provide insights into the proposed theory.

1) MNIST

In this, we train a CNN model with 60000 samples, equally distributed over a set of $n = 100$ clients. In each round, 20% of the clients participate based on random selection, which leads to $r = 20$. To emulate the IID setting, the data is shuffled and then randomly assigned to each client resulting in 600 samples/client. For the non-IID setting, we assign 1 or 2 labels to each client randomly. The model in each client is two 5×5 convolution layers, having 32 and 64 channels,

respectively. Each of these layers is followed by a 2×2 max pooling. Finally, the output is fed to a fully connected layer with 512 units followed by a ReLU activation, and a final output layer with softmax. We also included a dropout layer having the dropout $= 0.2$. The following are the parameters used for the training: local number of iterations, $E = 5$, global communication rounds, $K = 100$, local batch size, $BS = 20$, and learning rate, $\eta_k = 0.01$ for IID case and $\eta_k = 0.001$ for the non-IID case.

2) FASHION-MNIST(FMNIST)

Here we train a CNN model with 60000 samples. The data is distributed over a set of $n = 100$ clients, where, $r = 20$ clients participate in each round randomly. The experimental premise is divided into the IID and non-IID settings based on the distribution of the dataset. In IID the dataset is equally distributed equally amongst all the clients and in the non-IID setting, we only share 1 or 2 labels with each client to maintain high data heterogeneity. In each client, the model consists of two convolutional layers and three fully-connected layers. These 5×5 convolution layers have 6 and 12 channels respectively. Each of these layers is followed by ReLU activation and a 2×2 max pooling layer. The following are the parameters used for the training: local number of iterations, $E = 5$, global communication rounds, $K = 100$, local batch size, $BS = 20$, and learning rate, $\eta_k = 0.01$.

3) CIFAR-10

We train a CNN model similar to MNIST and FMNIST. The 50000 samples are equally distributed over $n = 100$ clients

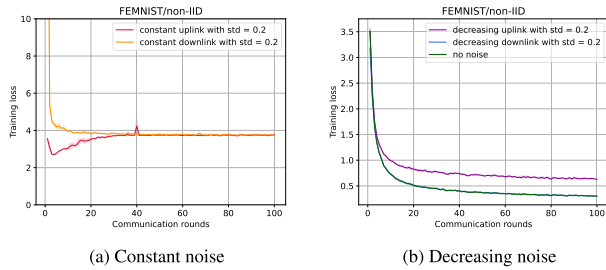(a) Constant noise   (b) Decreasing noise

FIGURE 8. Train loss vs. communication rounds for the non-IID data distribution of the FEMNIST dataset. The plots (8a) present the constant noise addition setting. We can see that the theory is verified from the plots. Similarly, the plots (8b) show the effect of the proposed SNR control strategy.

out of which 20%, i.e., $r = 20$, clients participate randomly in each round. In the case of the IID setting the data is shuffled and then randomly assigned to each client whereas in the case of the non-IID setting only 1 or 2 labels are assigned to each client randomly. The model architecture consists of two convolutional layers, followed by three fully-connected layers. The first convolutional layer has 3 input channels and 6 output channels, with a kernel size of 5. The second convolutional layer has 6 input channels and 16 output channels, also with a kernel size of 5. Between the convolutional layers, there is a max pooling layer with a kernel size of 2 and a stride of 2. After the convolutional layers, the output is flattened and passed through the three fully-connected layers, with 120, 84, and 10 units respectively. The forward function applies a ReLU activation function after each convolutional and fully-connected layer, except for the final fully-connected layer. The following are the parameters used for the training: local number of iterations, $E = 5$, global communication rounds, $K = 100$, local batch size, $BS = 20$, and learning rate, $\eta_k = 0.01$.

4) FEMNIST
In this, we train a CNN model with $n = 1000$ clients on the FEMNIST [40] dataset where in each round only 1%, i.e., 10 clients participate. This dataset consists of images that are highly heterogeneous since it has 62 classes consisting of digits and upper and lower case English characters. Additionally, the heterogeneity comes from the fact that these characters are written by different subjects. Thus, we only conduct experiments for the non-IID case. The model architecture consists of two convolutional layers followed by max-pooling, dropout, and fully-connected layers. The model architecture is adapted from [41]. The following are the parameters used for the training: local number of iterations, $E = 5$, global communication rounds, $K = 100$, local batch size, $BS = 50$, and learning rate, $\eta_k = 0.01$.

For all four (MNIST, FMNIST, CIFAR-10, and FEMNIST) experiments, we follow the same experimental setting as the synthetic experiment, i.e., the uplink and downlink noises are sampled from a zero mean Gaussian distribution $\sim \mathcal{N}(0, \upsilon^2)$. For our experiments, we choose $\upsilon = 0.2$. Again, to imitate
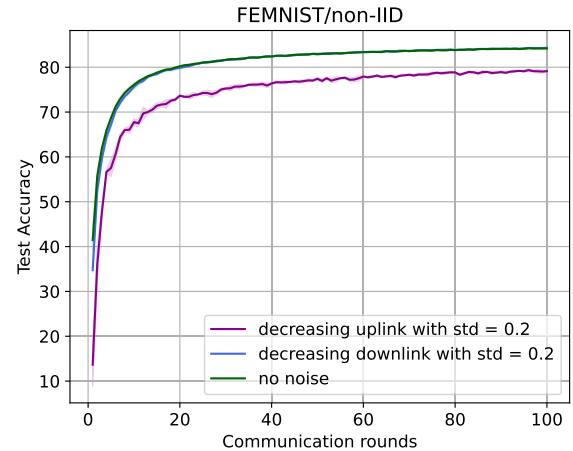

FIGURE 9. Test accuracy vs. communication rounds for the non-IID data distribution of FEMNIST dataset.

the noisy transmission channels, we add both uplink and downlink noise to the communicated messages. The results are shown in Figures 4 to 9 and are generated by averaging over 3 independent runs. We can visualize from the figure that the effect of noises inhibits the model from converging. One noteworthy point in the case of the FEMNIST dataset is the overlap of uplink and downlink noises after certain rounds. We hypothesize that this discrepancy goes back to the features of this experiment which is highly heterogeneous in nature. Also, the model and thus the objective function is not smooth and that is why we see this gap. So, essentially the effect of having a non-smooth function is exacerbated in this particular experiment due to the high level of heterogeneity. However, by employing our proposed SNR control policy in Section IV, we achieve model convergence for both noises with negligible convergence error with respect to the noise-free case.

VI. CONCLUSION
We studied the effects of having imperfect/noisy communication channels for federated learning. To the best of our knowledge, this paper is the first to establish the convergence analysis of FL where consideration has been made on both noisy transmission channels and smooth non-convex loss function without requiring the restrictive and hard-to-verify assumption of bounded client dissimilarity. By analyzing the convergence with these relaxed assumptions, we theoretically demonstrated that the effect of downlink noise is more detrimental than uplink noise. Using this insight, we proposed to employ SNR scaling policies for respective noisy channels that result in considerable savings in power consumption compared to existing approaches. We verified these theoretical findings via empirical results demonstrating the efficacy of the proposed analysis and its validity. Future work may involve investigating a parameter-free version of this scenario, i.e., an FL scheme that does not require the knowledge of parameters such as smoothness and analyzing its implication on the design of the system.

APPENDIX A
LEMMAS AND PROOFS

Lemma 11: For $\eta_k LE \leq \frac{1}{2}$, we have:

$\mathbb{E}[f(\boldsymbol{w}_{k+1})]$

$$\leq \mathbb{E}[f(\boldsymbol{w}_k)] - \frac{\eta_k(E-1)}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]$$

$$+ 4\eta_k^2 LE^2\Big(\frac{(n-r)}{r(n-1)} + \frac{2}{3}\eta_k LE\Big)\Big(\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2]\Big)$$

$$+ \eta_k^2 LE\Big(\frac{\eta_k LE}{n}\Big(1 + \frac{2nE}{3} + n\Big) + \frac{1}{r} + \frac{(n-r)}{r(n-1)}\Big)\sigma^2$$

$$+ \frac{\eta_k^2 L}{2r}\frac{1}{n}\sum_{i\in[n]}U_{k,i}^2 + \frac{\eta_k L^2}{2}\Big(1 + 2\eta_k L + 4E\{1 + 3\eta_k^2 L^2$$

$$+ 2\eta_k LE(2 + 3\eta_k^2 L^2)(\frac{2}{3}\eta_k LE + \frac{(n-r)}{r(n-1)})\}\Big)\frac{1}{n}\sum_{i\in[n]}N_{k,i}^2.$$

Proof: Define

$$\widehat{\boldsymbol{u}}_{k,\tau}^{(i)} := \widetilde{\nabla}f_i(\boldsymbol{w}_{k,\tau}^{(i)}; \mathcal{B}_{k,\tau}^{(i)}), \quad \widehat{\boldsymbol{u}}_{k,\tau} := \frac{1}{n}\sum_{i\in[n]}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)},$$

$$\boldsymbol{u}_{k,\tau} := \frac{1}{n}\sum_{i\in[n]}\nabla f_i(\boldsymbol{w}_{k,\tau}^{(i)}), \quad \overline{\boldsymbol{w}}_{k,\tau} := \frac{1}{n}\sum_{i\in[n]}\boldsymbol{w}_{k,\tau}^{(i)}.$$

Then:

$$\nabla f(\boldsymbol{w}_k) = \frac{1}{n}\sum_{i\in[n]}\nabla f_i(\boldsymbol{w}_k) \tag{40}$$

$\boldsymbol{w}_{k+1}$

$$= \boldsymbol{w}_k - \eta_k\Big[\frac{1}{r}\sum_{i\in\mathcal{S}_k}\Big(\boldsymbol{e}_k^{(i)} + \sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}$$

$$+ \widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\Big)\Big]. \tag{41}$$

$\boldsymbol{w}_{k,\tau}^{(i)}$

$$= \boldsymbol{w}_k + \boldsymbol{v}_k^{(i)} - \eta_k\Big(\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}$$

$$+ \widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\Big). \tag{42}$$

$\overline{\boldsymbol{w}}_{k,\tau}$

$$= \boldsymbol{w}_k + \frac{1}{n}\sum_{i\in[n]}\boldsymbol{v}_k^{(i)} - \eta_k\Big[\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}$$

$$+ \frac{1}{n}\sum_{i\in[n]}\Big(\widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\Big)\Big]. \tag{43}$$

$$\mathbb{E}_{\{\mathcal{B}_{k,\tau}^{(i)}\}_{i=1}^n}[\widehat{\boldsymbol{u}}_{k,\tau}] = \boldsymbol{u}_{k,\tau}. \tag{44}$$

$$\mathbb{E}\Big[\Big\|\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}\Big\|^2\Big]$$

$$\leq \tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|\boldsymbol{u}_{k,t}\|^2] + \frac{\tau\sigma^2}{n}. \tag{45}$$

$$\mathbb{E}\Big[\Big\|\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}^{(i)}\Big\|^2\Big]$$

$$\leq \tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2] + \tau\sigma^2. \tag{46}$$

$$N_{k,i}^2 := \mathbb{E}\Big[\Big\|\boldsymbol{v}_k^i\Big\|^2\Big]. \tag{47}$$

$$U_{k,i}^2 := \mathbb{E}\Big[\Big\|\boldsymbol{e}_k^i\Big\|^2\Big]. \tag{48}$$

Recall that σ^2 is the maximum variance of the local (client-level) stochastic gradients. In eq. (45), the expectation is w.r.t. $\{\mathcal{B}_{k,t}^{(i)}\}_{i=1,t=0}^{n,\tau-1}$ and it follows due to the independence of the noise in each local update of each client. Similarly, eq. (46), the expectation is w.r.t. $\{\mathcal{B}_{k,t}^{(i)}\}_{t=0}^{\tau-1}$ and it follows due to the independence of the noise in each local update. Also, eq. (47) and eq. (48) follows since both downlink and uplink noises have zero mean.

Next, using the L-smoothness of f and eq. (41), we get

$$\mathbb{E}[f(\boldsymbol{w}_{k+1})] \leq \mathbb{E}[f(\boldsymbol{w}_k)] + (A) + (B) \tag{49}$$

where

$$A = -\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k), \eta_k\Big[\frac{1}{r}\sum_{i\in\mathcal{S}_k}\Big(\boldsymbol{e}_k^{(i)} + \sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}$$

$$+ \widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\Big)\Big]\Big\rangle\Big] \tag{50}$$

and

$$B = \frac{L}{2}\mathbb{E}\Big[\Big\|\eta_k\Big[\frac{1}{r}\sum_{i\in\mathcal{S}_k}\Big(\boldsymbol{e}_k^{(i)} + \sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}$$

$$+ \widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\Big)\Big]\Big\|^2\Big] \tag{51}$$

Now using (A), eq. (52), as shown at the bottom of the next page.

A_1 will be zero since uplink noise has zero mean. Now, let's use A_2:

$$A_2 = -\eta_k\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k), \sum_{\tau=0}^{E-1}\boldsymbol{u}_{k,\tau}\Big\rangle\Big]$$

$$= -\eta_k\sum_{\tau=0}^{E-1}\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k), \boldsymbol{u}_{k,\tau}\Big\rangle\Big]$$

For any 2 vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, we have that:

$$\langle\boldsymbol{a}, \boldsymbol{b}\rangle = \frac{1}{2}(\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 - \|\boldsymbol{a} - \boldsymbol{b}\|^2) \tag{53}$$

Using this we will get A_2 as:

$$A_2 = \frac{\eta_k}{2} \sum_{\tau=0}^{E-1} \mathbb{E}[\|\nabla f(\boldsymbol{w}_k) - \boldsymbol{u}_{k,\tau}\|^2]$$
$$- \frac{\eta_k E}{2} \mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2] - \frac{\eta_k}{2} \sum_{\tau=0}^{E-1} \mathbb{E}[\|\boldsymbol{u}_{k,\tau}\|^2] \tag{54}$$

Again using A_3:

$$- \eta_k \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{w}_k), \frac{1}{n} \sum_{i\in[n]} (\widetilde{\nabla} f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}) - \widetilde{\nabla} f_i(\boldsymbol{w}_k)) \right\rangle\right]$$

$$= - \eta_k \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{w}_k), \frac{1}{n} \sum_{i\in[n]} (\nabla f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)})) \right\rangle\right]$$

$$+ \eta_k \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{w}_k), \frac{1}{n} \sum_{i\in[n]} \nabla f_i(\boldsymbol{w}_k) \right\rangle\right]$$

Using eq. (53) and eq. (40) in the equation above, we get

$$A_3 = \underbrace{\frac{\eta_k}{2} \mathbb{E}[\|\nabla f(\boldsymbol{w}_k) - \frac{1}{n} \sum_{i\in[n]} \nabla f_i(\boldsymbol{w}_k + \boldsymbol{n}_k^i)\|^2]}_{A_3'}$$

$$- \frac{\eta_k}{2} \mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2] - \frac{\eta_k}{2} \mathbb{E}[\|\frac{1}{n} \sum_{i\in[n]} \nabla f_i(\boldsymbol{w}_k + \boldsymbol{n}_k^i)\|^2]$$

$$+ \eta_k \mathbb{E}[\langle \nabla f(\boldsymbol{w}_k), \nabla f(\boldsymbol{w}_k) \rangle]$$

Reducing A_3':

$$A_3' = \frac{\eta_k}{2} \mathbb{E}[\|\frac{1}{n} \sum_{i\in[n]} \nabla f_i(\boldsymbol{w}_k) - \frac{1}{n} \sum_{i\in[n]} \nabla f_i(\boldsymbol{w}_k + \boldsymbol{n}_k^i)\|^2] \tag{55}$$

$$\leq \frac{\eta_k L^2}{2} \frac{1}{n} \sum_{i\in[n]} N_{k,i}^2 \tag{56}$$

The eq. (55) follows by using eq. (40), while eq. (56) follows from the L-smoothness of f_i, eq. (47) and independence of

noises. So, A_3 now becomes:

$$A_3 \leq \frac{\eta_k L^2}{2} \frac{1}{n} \sum_{i\in[n]} N_{k,i}^2 - \frac{\eta_k}{2} \mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]$$

$$- \frac{\eta_k}{2} \mathbb{E}[\|\frac{1}{n} \sum_{i\in[n]} \nabla f_i(\boldsymbol{w}_k + \boldsymbol{n}_k^i)\|^2] + \eta_k \mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]$$

$$\leq \frac{\eta_k L^2}{2} \frac{1}{n} \sum_{i\in[n]} N_{k,i}^2 + \frac{\eta_k}{2} \mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]$$

$$- \frac{\eta_k}{2} \mathbb{E}[\|\frac{1}{n} \sum_{i\in[n]} \nabla f_i(\boldsymbol{w}_k + \boldsymbol{n}_k^i)\|^2] \tag{57}$$

So, finally by combining eq. (54) and eq. (57), A becomes:

$$A \leq \frac{\eta_k}{2} \sum_{\tau=0}^{E-1} \mathbb{E}[\|\nabla f(\boldsymbol{w}_k) - \boldsymbol{u}_{k,\tau}\|^2]$$

$$- \frac{\eta_k(E-1)}{2} \mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2] - \frac{\eta_k}{2} \sum_{\tau=0}^{E-1} \mathbb{E}[\|\boldsymbol{u}_{k,\tau}\|^2]$$

$$+ \frac{\eta_k L^2}{2} \frac{1}{n} \sum_{i\in[n]} N_{k,i}^2 - \frac{\eta_k}{2} \mathbb{E}[\|\frac{1}{n} \sum_{i\in[n]} \nabla f_i(\boldsymbol{w}_k + \boldsymbol{n}_k^i)\|^2] \tag{58}$$

Now using (B):

$$B = \frac{L}{2} \eta_k^2 \Big(\mathbb{E}[\|\frac{1}{r} \sum_{i\in\mathcal{S}_k} \boldsymbol{e}_k^{(i)}\|^2] + \mathbb{E}[\|\frac{1}{r} \sum_{i\in\mathcal{S}_k} \sum_{\tau=0}^{E-1} \widehat{\boldsymbol{u}}_{k,\tau}^{(i)}\|^2]$$

$$+ \mathbb{E}[\|\frac{1}{r} \sum_{i\in\mathcal{S}_k} \widetilde{\nabla} f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla} f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\|^2]$$

$$+ 2\mathbb{E}[\langle \frac{1}{r} \sum_{i\in\mathcal{S}_k} \boldsymbol{e}_k^{(i)}, \frac{1}{r} \sum_{i\in\mathcal{S}_k} \sum_{\tau=0}^{E-1} \widehat{\boldsymbol{u}}_{k,\tau}^{(i)} \rangle] + 2\mathbb{E}[\langle \frac{1}{r} \sum_{i\in\mathcal{S}_k} \boldsymbol{e}_k^{(i)},$$

$$\frac{1}{r} \sum_{i\in\mathcal{S}_k} (\widetilde{\nabla} f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla} f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})) \rangle]$$

$$+ 2\mathbb{E}[\langle \frac{1}{r} \sum_{i\in\mathcal{S}_k} \sum_{\tau=0}^{E-1} \widehat{\boldsymbol{u}}_{k,\tau}^{(i)}, \frac{1}{r} \sum_{i\in\mathcal{S}_k} (\widetilde{\nabla} f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)})$$

$$- \widetilde{\nabla} f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)}) \rangle] \Big) \tag{59}$$

$$A = -\eta_k \underbrace{\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{w}_k), \frac{1}{r} \sum_{i\in\mathcal{S}_k} \boldsymbol{e}_k^{(i)} \right\rangle\right]}_{(A_1)}$$

$$-\eta_k \underbrace{\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{w}_k), \frac{1}{r} \sum_{i\in\mathcal{S}_k} \sum_{\tau=0}^{E-1} \widehat{\boldsymbol{u}}_{k,\tau}^{(i)} \right\rangle\right]}_{(A_2)}$$

$$-\eta_k \underbrace{\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{w}_k), \frac{1}{r} \sum_{i\in\mathcal{S}_k} \left(\widetilde{\nabla} f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla} f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\right) \right\rangle\right]}_{(A_3)} \tag{52}$$

Using the fact that $\mathbb{E}[e_k^i] = 0$ and Young's Inequality in eq. (59), we get:

B

$$
\leq \frac{L}{2}\eta_k^2 \Big(\underbrace{\mathbb{E}[\|\frac{1}{r}\sum_{i\in\mathcal{S}_k}e_k^{(i)}\|^2]}_{B_1} + \underbrace{2\mathbb{E}[\|\frac{1}{r}\sum_{i\in\mathcal{S}_k}\sum_{\tau=0}^{E-1}\widehat{u}_{k,\tau}^{(i)}\|^2]}_{B_2}
$$

$$
+ \underbrace{2\mathbb{E}[\|\frac{1}{r}\sum_{i\in\mathcal{S}_k}\widetilde{\nabla}f_i(w_k+v_k^{(i)};\mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(w_k;\mathcal{B}_{k,0}^{(i)})\|^2]}_{B_3} \Big)
$$

$$
\tag{60}
$$

Starting with B_1 :

$$
B_1 = \frac{n(r-1)}{r(n-1)}\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}e_k^{(i)}\|^2]
$$

$$
+ \frac{(n-r)}{r(n-1)}\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|e_k^{(i)}\|^2]
\tag{61}
$$

$$
= \frac{(r-1)}{nr(n-1)}\sum_{i\in[n]}U_{k,i}^2 + \frac{(n-r)}{nr(n-1)}\sum_{i\in[n]}U_{k,i}^2
\tag{62}
$$

$$
= \frac{1}{nr}\sum_{i\in[n]}U_{k,i}^2
\tag{63}
$$

Here, eq. (61) follows due to expectation w.r.t $\mathcal{S}_k$ and eq. (62) follows due to expectation w.r.t uplink noise and its independence. Now let's focus on B_2 :

$$
B_2 \leq \frac{2n(r-1)E}{r(n-1)}\Big(\sum_{\tau=0}^{E-1}\mathbb{E}[\|u_{k,\tau}\|^2] + \frac{\sigma^2}{n}\Big)
$$

$$
+ \frac{2(n-r)E}{r(n-1)}\Big(\frac{1}{n}\sum_{i\in[n]}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f_i(w_{k,\tau}^i)\|^2] + \sigma^2\Big)
$$

$$
\tag{64}
$$

The eq. (64) follows due to expectation w.r.t $\mathcal{S}_k$, eq. (44), eq. (45) and eq. (46). Again, let's use B_3 :

$$
B_3 \leq 2\mathbb{E}[\frac{1}{r}\sum_{i\in\mathcal{S}_k}\|\widetilde{\nabla}f_i(w_k+v_k^{(i)};\mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(w_k;\mathcal{B}_{k,0}^{(i)})\|^2]
$$

$$
\tag{65}
$$

$$
\leq 2L^2\mathbb{E}[\frac{1}{r}\sum_{i\in\mathcal{S}_k}\|v_k^{(i)}\|^2]
\tag{66}
$$

$$
\leq 2L^2\frac{1}{n}\sum_{i\in[n]}N_{k,i}^2
\tag{67}
$$

Here we used Jensen's inequality to reach eq. (65). Again, eq. (66) follows due to the L-smoothness of f_i and eq. (67) follows due to expectation w.r.t $\mathcal{S}_k$ and eq. (47). So, finally by combining B_1, B_2 and B_3, (B) becomes:

$$
B \leq \frac{L\eta_k^2}{2}\Big(\frac{1}{nr}\sum_{i\in[n]}U_{k,i}^2 + \frac{2n(r-1)E}{r(n-1)}\Big(\sum_{\tau=0}^{E-1}\mathbb{E}[\|u_{k,\tau}\|^2]
$$

$$
+ \frac{\sigma^2}{n}\Big) + \frac{2(n-r)E}{r(n-1)}\Big(\frac{1}{n}\sum_{i\in[n]}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f_i(w_{k,\tau}^i)\|^2] + \sigma^2\Big)
$$

$$
+ 2L^2\frac{1}{n}\sum_{i\in[n]}N_{k,i}^2\Big)
$$

$$
\leq \frac{L\eta_k^2}{2nr}\sum_{i\in[n]}U_{k,i}^2 + \frac{L^3\eta_k^2}{n}\sum_{i\in[n]}N_{k,i}^2 + \frac{\eta_k^2 LE}{r}\sigma^2
$$

$$
+ \frac{n(r-1)}{r(n-1)}\eta_k^2 LE\Big(\sum_{\tau=0}^{E-1}\mathbb{E}[\|u_{k,\tau}\|^2]\Big)
$$

$$
+ \frac{(n-r)}{r(n-1)}\eta_k^2 LE\Big(\frac{1}{n}\sum_{i\in[n]}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f_i(w_{k,\tau}^i)\|^2]\Big)
\tag{68}
$$

Now, by putting eq. (58) and eq. (68) in eq. (49) we will get:

$$
\mathbb{E}[f(w_{k+1})]
$$

$$
\leq \mathbb{E}[f(w_k)] - \frac{\eta_k(E-1)}{2}\mathbb{E}[\|\nabla f(w_k)\|^2]
$$

$$
- \frac{\eta_k}{2}\Big(1 - \eta_k LE\frac{n(r-1)}{r(n-1)}\Big)\sum_{\tau=0}^{E-1}\mathbb{E}[\|u_{k,\tau}\|^2]
$$

$$
- \frac{\eta_k}{2}\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}\nabla f_i(w_k+v_k^{(i)})\|^2] + \frac{\eta_k^2 LE}{r}\sigma^2
$$

$$
+ \frac{\eta_k^2 L}{2nr}\sum_{i\in[n]}U_{k,i}^2 + \frac{\eta_k L^2}{2}\Big(1+2\eta_k L\Big)\frac{1}{n}\sum_{i\in[n]}N_{k,i}^2
$$

$$
+ \underbrace{\frac{\eta_k}{2}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f(w_k) - u_{k,\tau}\|^2]}_{(M)}
$$

$$
+ \underbrace{\eta_k^2 LE\frac{(n-r)}{r(n-1)}\Big(\frac{1}{n}\sum_{i\in[n]}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f_i(w_{k,\tau}^{(i)})\|^2]\Big)}_{(N)}
\tag{69}
$$

We upper bound (M) and (N) using Lemma 12 and Lemma 13, respectively. Plugging in these bounds and dropping the last term of eq. (69), we get:

$$
\mathbb{E}[f(w_{k+1})]
$$

$$
\leq \mathbb{E}[f(w_k)] - \frac{\eta_k(E-1)}{2}\mathbb{E}[\|\nabla f(w_k)\|^2]
$$

$$
- \frac{\eta_k}{2}\underbrace{\Big(1 - \eta_k LE\frac{n(r-1)}{r(n-1)} - 2\eta_k^2 L^2 E^2\Big)}_{(C)}\sum_{\tau=0}^{E-1}\mathbb{E}[\|u_{k,\tau}\|^2]
$$

$$
+ 4\eta_k^2 LE^2\Big(\frac{(n-r)}{r(n-1)} + \frac{2}{3}\eta_k LE\Big)\Big(\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\nabla f_i(w_k)\|^2]\Big)
$$

$$
+ \eta_k^2 LE\Big(\frac{\eta_k LE}{n}\Big(1+\frac{2nE}{3}+n\Big) + \frac{1}{r} + \frac{(n-r)}{r(n-1)}\Big)\sigma^2
$$

$$+ \frac{\eta_k^2 L}{2r} \frac{1}{n} \sum_{i \in [n]} U_{k,i}^2 + \frac{\eta_k L^2}{2} \Big(1 + 2\eta_k L + 4E\{1 + 3\eta_k^2 L^2$$
$$+ 2\eta_k LE(2 + 3\eta_k^2 L^2)(\frac{2}{3}\eta_k LE + \frac{(n-r)}{r(n-1)})\}\Big) \frac{1}{n} \sum_{i \in [n]} N_{k,i}^2.$$

$$(70)$$

for $\eta_k LE \leq \frac{1}{2}$. Note that (C) ≥ 0 for $\eta_k LE \leq \frac{1}{2}$. Thus, for $\eta_k LE \leq \frac{1}{2}$, we have:

$$\mathbb{E}[f(\boldsymbol{w}_{k+1})]$$

$$\leq \mathbb{E}[f(\boldsymbol{w}_k)] - \frac{\eta_k(E-1)}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]$$
$$+ 4\eta_k^2 LE^2 \Big(\frac{(n-r)}{r(n-1)} + \frac{2}{3}\eta_k LE\Big)\Big(\frac{1}{n}\sum_{i \in [n]}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2]\Big)$$
$$+ \eta_k^2 LE\Big(\frac{\eta_k LE}{n}\Big(1 + \frac{2nE}{3} + n\Big) + \frac{1}{r} + \frac{(n-r)}{r(n-1)}\Big)\sigma^2$$
$$+ \frac{\eta_k^2 L}{2r}\frac{1}{n}\sum_{i \in [n]} U_{k,i}^2 + \frac{\eta_k L^2}{2}\Big(1 + 2\eta_k L + 4E\{1 + 3\eta_k^2 L^2$$
$$+ 2\eta_k LE(2 + 3\eta_k^2 L^2)(\frac{2}{3}\eta_k LE + \frac{(n-r)}{r(n-1)})\}\Big)\frac{1}{n}\sum_{i \in [n]} N_{k,i}^2.$$

$$(71)$$

∎

Lemma 12: For $\eta_k LE \leq \frac{1}{2}$:

$$\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k) - \boldsymbol{u}_{k,\tau}\|^2]$$

$$\leq 2\eta_k^2 L^2 E^2 \sum_{\tau=0}^{E-1}\mathbb{E}[\|\boldsymbol{u}_{k,\tau}\|^2]$$
$$+ \frac{16}{3}\eta_k^2 L^2 E^3 \frac{1}{n}\sum_{i \in [n]}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2]$$
$$+ 2\eta_k^2 L^2 E^2(\frac{1}{n} + \frac{2E}{3} + 1)\sigma^2$$
$$+ 4L^2 E(1 + 3\eta_k^2 L^2 + \frac{4}{3}\eta_k^2 L^2 E^2(2 + 3\eta_k^2 L^2))\frac{1}{n}\sum_{i \in [n]} N_{k,i}^2.$$

Proof: We have:

$$\mathbb{E}[\|\nabla f(\boldsymbol{w}_k) - \boldsymbol{u}_{k,\tau}\|^2]$$
$$= \mathbb{E}[\|\nabla f(\boldsymbol{w}_k) - \nabla f(\overline{\boldsymbol{w}}_{k,\tau}) + \nabla f(\overline{\boldsymbol{w}}_{k,\tau}) - \boldsymbol{u}_{k,\tau}\|^2]$$
$$\leq 2\mathbb{E}[\|\nabla f(\boldsymbol{w}_k) - \nabla f(\overline{\boldsymbol{w}}_{k,\tau})\|^2]$$
$$\quad + 2\mathbb{E}[\|\nabla f(\overline{\boldsymbol{w}}_{k,\tau}) - \boldsymbol{u}_{k,\tau}\|^2]$$
$$\leq \underbrace{2L^2 \mathbb{E}[\|\boldsymbol{w}_k - \overline{\boldsymbol{w}}_{k,\tau}\|^2]}_{M_1}$$
$$+ \underbrace{2\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{i \in [n]}(\nabla f_i(\overline{\boldsymbol{w}}_{k,\tau}) - \nabla f_i(\boldsymbol{w}_{k,\tau}^{(i)}))\Big\|^2\Big]}_{M_2} \quad (72)$$

Using M_1 :

$$2L^2\mathbb{E}[\|\boldsymbol{w}_k - \overline{\boldsymbol{w}}_{k,\tau}\|^2]$$

$$= 2L^2\mathbb{E}\Big[\Big\|\eta_k \sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t} + \frac{\eta_k}{n}\sum_{i \in [n]}\{\widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)})$$
$$\quad - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\} - \frac{1}{n}\sum_{i \in [n]}\boldsymbol{v}_k^{(i)}\Big\|^2\Big]$$

$$= 2L^2\Big(\mathbb{E}[\|\eta_k \sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}\|^2] + \mathbb{E}[\|\frac{1}{n}\sum_{i \in [n]}\boldsymbol{v}_k^{(i)}\|^2]$$
$$+ \mathbb{E}[\|\frac{\eta_k}{n}\sum_{i \in [n]}\{\widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\}\|^2]$$
$$+ 2\mathbb{E}[\langle \eta_k \sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}, \frac{\eta_k}{n}\sum_{i \in [n]}\{\widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)})$$
$$- \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\}\rangle]$$
$$+ 2\mathbb{E}[\langle \frac{\eta_k}{n}\sum_{i \in [n]}\{\widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\},$$
$$- \frac{1}{n}\sum_{i \in [n]}\boldsymbol{v}_k^{(i)}\rangle] + 2\mathbb{E}[\langle -\frac{1}{n}\sum_{i \in [n]}\boldsymbol{v}_k^{(i)}, \eta_k \sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}\rangle]\Big) \quad (73)$$

Simplifying eq. (73) using the fact that $\mathbb{E}[\|\boldsymbol{n}_k^{(i)}\|] = 0$ and Young's Inequality we will get:

$$2L^2\mathbb{E}[\|\boldsymbol{w}_k - \overline{\boldsymbol{w}}_{k,\tau}\|^2]$$

$$\leq 2L^2\Big(2\mathbb{E}[\|\eta_k \sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}\|^2]$$
$$+ 3\mathbb{E}[\|\frac{\eta_k}{n}\sum_{i \in [n]}\{\widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\}\|^2]$$
$$+ 2\mathbb{E}[\|\frac{1}{n}\sum_{i \in [n]}\boldsymbol{v}_k^{(i)}\|^2]\Big)$$

Using Jensen's Inequality in the equation above, we get

$$2L^2\mathbb{E}[\|\boldsymbol{w}_k - \overline{\boldsymbol{w}}_{k,\tau}\|^2]$$

$$\leq 2L^2\Big(2\eta_k^2\mathbb{E}[\|\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}\|^2]$$
$$+ 3\eta_k^2\frac{1}{n}\sum_{i \in [n]}\mathbb{E}[\|\widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\|^2]$$
$$+ \frac{2}{n^2}\mathbb{E}[\|\sum_{i \in [n]}\boldsymbol{v}_k^{(i)}\|^2]\Big)$$

$$(74)$$

In eq. (74) using the L-smoothness of f_i, eq. (45), eq. (47) and independence of noise, we get

$$2L^2\mathbb{E}[\|\boldsymbol{w}_k - \overline{\boldsymbol{w}}_{k,\tau}\|^2]$$

$$\leq 2L^2\Big(2\eta_k^2\tau(\sum_{t=0}^{\tau-1}\mathbb{E}[\|\boldsymbol{u}_{k,t}\|^2] + \frac{\sigma^2}{n})$$
$$+ 3\eta_k^2 L^2\frac{1}{n}\sum_{i \in [n]} N_{k,i}^2 + \frac{2}{n^2}\sum_{i \in [n]} N_{k,i}^2\Big)$$

Now let's use M_2:

$$2\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}(\nabla f_i(\overline{\boldsymbol{w}}_{k,\tau}) - \nabla f_i(\boldsymbol{w}_{k,\tau}^{(i)}))\|^2]$$
$$\leq 2L^2\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\boldsymbol{w}_{k,\tau}^{(i)} - \overline{\boldsymbol{w}}_{k,\tau}\|^2] \tag{75}$$

Here eq. (75) follows from Jensen's Inequality. Now using the definition of $\overline{\boldsymbol{w}}_{k,\tau}$ and the L-smoothness of f_i in conjunction with the same simplification process as used to simplify eq. (73), we get eq. (76), as shown at the bottom of the next page.

Using (X):

$$2\mathbb{E}[\|(\boldsymbol{n}_k^{(i)} - \frac{1}{n}\sum_{i\in[n]}\boldsymbol{n}_k^{(i)})\|^2]$$
$$= 2\mathbb{E}[\|\boldsymbol{n}_k^{(i)}\|^2]$$
$$+ 2\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}\boldsymbol{n}_k^{(i)}\|^2] - 4\mathbb{E}[\langle\boldsymbol{n}_k^{(i)}, \frac{1}{n}\sum_{i\in[n]}\boldsymbol{n}_k^{(i)}\rangle]$$
$$= 2N_{k,i}^2 + \frac{2}{n^2}\sum_{i\in[n]}N_{k,i}^2 - \frac{4}{n}N_{k,i}^2 \tag{77}$$
$$= 2(1 - \frac{2}{n})N_{k,i}^2 + \frac{2}{n^2}\sum_{i\in[n]}N_{k,i}^2 \tag{78}$$

Equation (77) follows due to eq. (47) and independence of noise. So, now moving on to (Y):

$$2\mathbb{E}[\|\eta_k(\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t} - \sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}^{(i)})\|^2]$$
$$= 2\eta_k^2\mathbb{E}[\|\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t} - \sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}^{(i)}\|^2]$$
$$\leq 2\eta_k^2\tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|\widehat{\boldsymbol{u}}_{k,t} - \widehat{\boldsymbol{u}}_{k,t}^{(i)}\|^2] \tag{79}$$
$$= 2\eta_k^2\tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|\widehat{\boldsymbol{u}}_{k,t}\|^2 + \|\widehat{\boldsymbol{u}}_{k,t}^{(i)}\|^2 - 2\langle\widehat{\boldsymbol{u}}_{k,t}, \widehat{\boldsymbol{u}}_{k,t}^{(i)}\rangle] \tag{80}$$

Equation (79) follows because of Jensen's Inequality and using the fact that $\widehat{\boldsymbol{u}}_{k,\tau} = \frac{1}{n}\sum_{i\in[n]}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}$, we can simplify eq. (80) to:

$$2\mathbb{E}[\|\eta_k(\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t} - \sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}^{(i)})\|^2]$$
$$\leq 2\eta_k^2\tau\sum_{t=0}^{\tau-1}(\mathbb{E}[\|\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}\|^2] - \mathbb{E}[\|\widehat{\boldsymbol{u}}_{k,t}\|^2])$$
$$\leq 2\eta_k^2\tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}\|^2] \tag{81}$$
$$\leq 2\eta_k^2\tau\sum_{t=0}^{\tau-1}(\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2] + \sigma^2) \tag{82}$$

Now using Lemma 13 for $\boldsymbol{n}_kLE \leq \frac{1}{2}$ in eq. (82), we get:

$$2\mathbb{E}[\|\eta_k(\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t} - \sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}^{(i)})\|^2]$$
$$\leq 2\eta_k^2\tau^2(4\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2]$$
$$+ 4L^2(2 + 3\eta_k^2L^2)N_{k,i}^2) + 2\eta_k^2(\tau^2 + \tau)\sigma^2 \tag{83}$$

Next, using (Z):

$$Z = 3\eta_k^2\mathbb{E}\Big[\|\frac{1}{n}\sum_{i\in[n]}(\widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)})$$
$$- \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)}))\|^2$$
$$+ \|(\widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)}))\|^2$$
$$- 2\langle\frac{1}{n}\sum_{i\in[n]}(\widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})),$$
$$(\widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)}))\rangle\Big] \tag{84}$$

We simplify eq. (84) using the similar fact that we used to simplify eq. (79). Subsequently, by using the L-smoothness of f_i, we get

$$Z = 3\eta_k^2(\mathbb{E}[\|(\widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)}))\|^2]$$
$$- \mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}(\widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)}))\|^2])$$
$$\leq 3\eta_k^2\mathbb{E}[\|(\widetilde{\nabla}f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla}f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)}))\|^2]$$
$$\leq 3\eta_k^2L^2\mathbb{E}[\|\boldsymbol{v}_k^{(i)}\|^2] \leq 3\eta_k^2L^2N_{k,i}^2 \tag{85}$$

Now putting the results of eq. (78), eq. (83) and eq. (85) in eq. (76) we get:

$$2\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}(\nabla f_i(\overline{\boldsymbol{w}}_{k,\tau}) - \nabla f_i(\boldsymbol{w}_{k,\tau}^{(i)}))\|^2]$$
$$\leq 2L^2\frac{1}{n}\sum_{i\in[n]}\Big(2(1 - \frac{2}{n})N_{k,i}^2 + \frac{2}{n^2}\sum_{i\in[n]}N_{k,i}^2$$
$$+ 2\eta_k^2\tau^2(4\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2] + 4L^2(2 + 3\eta_k^2L^2)N_{k,i}^2)$$
$$+ 2\eta_k^2(\tau^2 + \tau)\sigma^2 + 3\eta_k^2L^2N_{k,i}^2\Big)$$
$$\leq 16\eta_k^2L^2\tau^2\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}k)\|^2] + 2L^2(2 - \frac{2}{n}$$
$$+ 3\eta_k^2L^2)\frac{1}{n}\sum_{i\in[n]}N_{k,i}^2 + 16\eta_k^2L^4\tau^2(2 + 3\eta_k^2L^2)\frac{1}{n}\sum_{i\in[n]}N_{k,i}^2$$
$$+ 4\eta_k^2L^2(\tau^2 + \tau)\sigma^2. \tag{86}$$

So, by combining M_1 and M_2 in eq. (72), we get

$$\mathbb{E}[\|\nabla f(\boldsymbol{w}_k) - \boldsymbol{u}_{k,\tau}\|^2]$$
$$\leq 4\eta_k^2L^2\tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|\boldsymbol{u}_{k,t}\|^2]$$
$$+ 16\eta_k^2L^2\tau^2\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}k)\|^2]$$

$$+ 4\eta_k^2 L^2(\tau^2 + \frac{\tau}{n} + \tau)\sigma^2$$
$$+ 4L^2(1 + 3\eta_k^2 L^2)\frac{1}{n}\sum_{i\in[n]} N_{k,i}^2$$
$$+ 16\eta_k^2 L^4\tau^2(2 + 3\eta_k^2 L^2)\frac{1}{n}\sum_{i\in[n]} N_{k,i}^2. \tag{87}$$

Now summing up eq. (87) for all $\tau \in \{0, \ldots, E-1\}$, we get:

$$\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k) - \boldsymbol{u}_{k,\tau}\|^2] \leq 2\eta_k^2 L^2 E^2 \sum_{\tau=0}^{E-1}\mathbb{E}[\|\boldsymbol{u}_{k,\tau}\|^2]$$
$$+ \frac{16}{3}\eta_k^2 L^2 E^3 \frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2] + 2\eta_k^2 L^2 E^2(\frac{1}{n}$$
$$+ \frac{2E}{3} + 1)\sigma^2 + 4L^2 E\big(1 + 3\eta_k^2 L^2 + \frac{4}{3}\eta_k^2 L^2 E^2(2$$
$$+ 3\eta_k^2 L^2)\big)\frac{1}{n}\sum_{i\in[n]} N_{k,i}^2. \tag{88}$$

∎

Lemma 13: For $\eta_k LE \leq \frac{1}{2}$, we have:

$$\sum_{t=0}^{\tau-1}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2] \leq 4\tau\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2]$$
$$+ 4L^2\tau(2 + 3\eta_k^2 L^2)N_{k,i}^2 + \sigma^2.$$

Proof:

$$\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2]$$
$$= \mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)}) - \nabla f_i(\boldsymbol{w}_k) + \nabla f_i(\boldsymbol{w}_k)\|^2]$$
$$\leq 2\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2] + 2\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)}) - \nabla f_i(\boldsymbol{w}_k)\|^2]$$
$$\leq 2\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2] + 2L^2\mathbb{E}[\|\boldsymbol{w}_{k,t}^{(i)} - \boldsymbol{w}_k\|^2]. \tag{89}$$

But:

$$\mathbb{E}[\|\boldsymbol{w}_{k,t}^{(i)} - \boldsymbol{w}_k\|^2]$$
$$= \mathbb{E}\Big[\Big\|\boldsymbol{v}_k^{(i)} - \eta_k\Big(\sum_{t'=0}^{t-1}\widehat{\boldsymbol{u}}_{k,t'}^{(i)}$$

$$+ \widetilde{\nabla} f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla} f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\Big)\Big\|^2\Big]$$
$$= \mathbb{E}[\|\boldsymbol{v}_k^{(i)}\|^2] + \eta_k^2\mathbb{E}[\|\sum_{t'=0}^{t-1}\widehat{\boldsymbol{u}}_{k,t'}^{(i)}\|^2]$$
$$+ \eta_k^2\mathbb{E}[\|\widetilde{\nabla} f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla} f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\|^2]$$
$$+ 2\mathbb{E}[\langle\boldsymbol{v}_k^{(i)}, -\eta_k\sum_{t'=0}^{t-1}\widehat{\boldsymbol{u}}_{k,t'}^{(i)}\rangle] + 2\mathbb{E}[\langle\boldsymbol{v}_k^{(i)},$$
$$- \eta_k\{\widetilde{\nabla} f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla} f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\}\rangle]$$
$$+ 2\mathbb{E}[\langle -\eta_k\sum_{t'=0}^{t-1}\widehat{\boldsymbol{u}}_{k,t'}^{(i)}, -\eta_k\{\widetilde{\nabla} f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)})$$
$$- \widetilde{\nabla} f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\}\rangle] \tag{90}$$

Now using the fact that $\mathbb{E}[\boldsymbol{n}_k^i] = 0$ and Young's Inequality in eq. (90), we get:

$$\mathbb{E}[\|\boldsymbol{w}_{k,t}^{(i)} - \boldsymbol{w}_k\|^2]$$
$$\leq 2\mathbb{E}[\|\boldsymbol{v}_k^{(i)}\|^2] + 2\eta_k^2\mathbb{E}[\|\sum_{t'=0}^{t-1}\widehat{\boldsymbol{u}}_{k,t'}^{(i)}\|^2]$$
$$+ 3\eta_k^2\mathbb{E}[\|\widetilde{\nabla} f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla} f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\|^2]$$
$$\leq 2N_{k,i}^2 + 2\eta_k^2 t(\sum_{t'=0}^{t-1}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t'}^{(i)}\|^2]) + \sigma^2) + 3\eta_k^2 L^2 N_{k,i}^2 \tag{91}$$

Putting eq. (91) in eq. (89), we get:

$$\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2]$$
$$\leq 2\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2]$$
$$+ 2L^2(2 + 3\eta_k^2 L^2)N_{k,i}^2 + 4\eta_k^2 L^2 t(\sum_{t'=0}^{t-1}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t'}^{(i)})\|^2]$$
$$+ \sigma^2) \tag{92}$$

$$2\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}(\nabla f_i(\overline{\boldsymbol{w}}_{k,\tau}) - \nabla f_i(\boldsymbol{w}_{k,\tau}^{(i)}))\|^2]$$
$$\leq 2L^2\frac{1}{n}\sum_{i\in[n]}\Big(\underbrace{2\mathbb{E}[\|(\boldsymbol{n}_k^{(i)} - \frac{1}{n}\sum_{i\in[n]}\boldsymbol{n}_k^{(i)})\|^2]}_{(X)}$$
$$+ \underbrace{2\mathbb{E}[\|\eta_k(\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t} - \sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}^{(i)})\|^2]}_{(Y)}$$
$$+ \underbrace{3\mathbb{E}[\|\eta_k(\{\frac{1}{n}\sum_{i\in[n]}(\widetilde{\nabla} f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla} f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)}))\} - \{\widetilde{\nabla} f_i(\boldsymbol{w}_k + \boldsymbol{v}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \widetilde{\nabla} f_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\})\|^2]}_{(Z)}\Big) \tag{76}$$

Summing up eq. (92) for all $t \in \{0, \ldots, \tau - 1\}$, we get:

$$
\sum_{t=0}^{\tau-1} \mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2]
$$
$$
\leq 2\tau \mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2]
$$
$$
+ 2L^2\tau(2 + 3\eta_k^2 L^2)N_{k,i}^2 + 2\eta_k^2 L^2 \tau^2 \left(\sum_{t=0}^{\tau-1} \mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2] \right.
$$
$$
\left. + \sigma^2 \right) \tag{93}
$$

Let us set $\eta_k LE \leq 1/2$. Then:

$$
\sum_{t=0}^{\tau-1} \mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2]
$$
$$
\leq 2\tau \mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2]
$$
$$
+ 2L^2\tau(2 + 3\eta_k^2 L^2)N_{k,i}^2 + \frac{1}{2}\sum_{t=0}^{\tau-1} \mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2] + \frac{1}{2}\sigma^2 \tag{94}
$$

Simplifying, we get:

$$
\sum_{t=0}^{\tau-1} \mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2] \leq 4\tau \mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2]
$$
$$
+ 4L^2\tau(2 + 3\eta_k^2 L^2)N_{k,i}^2 + \sigma^2 \tag{95}
$$

∎

REFERENCES

[1] W. Ren and R. W. Beard, "Consensus seeking in multiagent systems under dynamically changing interaction topologies," *IEEE Trans. Autom. Control*, vol. 50, no. 5, pp. 655–661, May 2005.

[2] A. Jadbabaie, J. Lin, and A. S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Trans. Autom. Control*, vol. 48, no. 6, pp. 988–1001, Jun. 2003.

[3] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[4] A. Nedic and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Trans. Autom. Control*, vol. 60, no. 3, pp. 601–615, Mar. 2015.

[5] A. Nedic and A. Olshevsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Trans. Autom. Control*, vol. 61, no. 12, pp. 3936–3947, Dec. 2016.

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

[7] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, 3rd Quart., 2020.

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

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

[10] R. Das, A. Hashemi, S. Sanghavi, and I. S. Dhillon, "On the convergence of differentially private federated learning on non-Lipschitz objectives, and with normalized client updates," 2021, *arXiv:2106.07094*.

[11] H. Tang, S. Gan, C. Zhang, T. Zhang, and J. Liu, "Communication compression for decentralized training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7652–7662.

[12] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "TernGrad: Ternary gradients to reduce communication in distributed deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1509–1519.

[13] H. Zhang, J. Li, K. Kara, D. Alistarh, J. Liu, and C. Zhang, "ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 4035–4043.

[14] Y. Savas, A. Hashemi, A. P. Vinod, B. M. Sadler, and U. Topcu, "Physical-layer security via distributed beamforming in the presence of adversaries with unknown locations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 4685–4689.

[15] R. Das, A. Acharya, A. Hashemi, S. Sanghavi, I. S. Dhillon, and U. Topcu, "Faster non-convex federated learning via global and local momentum," in *Proc. Uncertainty Artif. Intell.*, 2022, pp. 496–506.

[16] S. Raja, G. Habibi, and J. P. How, "Communication-aware consensus-based decentralized task allocation in communication constrained environments," *IEEE Access*, vol. 10, pp. 19753–19767, 2022.

[17] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, vol. 2, 2020, pp. 429–450.

[18] 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. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 2021–2031.

[19] 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, 2020.

[20] 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.

[21] A. Hashemi, A. Acharya, R. Das, H. Vikalo, S. Sanghavi, and I. Dhillon, "On the benefits of multiple gossip steps in communication-constrained decentralized federated learning," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 11, pp. 2727–2739, Nov. 2022.

[22] Y. Chen, A. Hashemi, and H. Vikalo, "Communication-efficient variance-reduced decentralized stochastic optimization over time-varying directed graphs," *IEEE Trans. Autom. Control*, vol. 67, no. 12, pp. 6583–6594, Dec. 2022.

[23] Y. Chen, A. Hashemi, and H. Vikalo, "Decentralized optimization on time-varying directed graphs under communication constraints," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 3670–3674.

[24] 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.

[25] 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.

[26] S. Xia, J. Zhu, Y. Yang, Y. Zhou, Y. Shi, and W. Chen, "Fast convergence algorithm for analog federated learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2021, pp. 1–6.

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

[28] 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.

[29] X. Wei and C. Shen, "Federated learning over noisy channels: Convergence analysis and design examples," *IEEE Trans. Cognit. Commun. Netw.*, vol. 8, no. 2, pp. 1253–1268, Jun. 2022.

[30] 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.

[31] 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.

[32] Y. Yu, J. Wu, and L. Huang, "Double quantization for communication-efficient distributed optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[33] C.-Y. Chen, J. Ni, S. Lu, X. Cui, P.-Y. Chen, X. Sun, N. Wang, S. Venkataramani, V. Srinivasan, W. Zhang, and K. Gopalakrishnan, "ScaleCom: Scalable sparsified gradient compression for communication-efficient distributed training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 13551–13563.

[34] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5132–5143.

[35] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[36] J. Wang and G. Joshi, "Cooperative SGD: A unified framework for the design and analysis of local-update SGD algorithms," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 9709–9758, 2021.

[37] H. Yu, S. Yang, and S. Zhu, "Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 5693–5700.

[38] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," 2019, *arXiv:1907.02189*.

[39] L. Nguyen, P. H. Nguyen, M. Dijk, P. Richtárik, K. Scheinberg, and M. Takác, "SGD and hogwild! Convergence without the bounded gradients assumption," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3750–3758.

[40] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečnỳ, H. B. McMahan, V. Smith, and A. Talwalkar, "LEAF: A benchmark for federated settings," 2018, *arXiv:1812.01097*.

[41] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečnỳ, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," 2020, *arXiv:2003.00295*.

ANTESH UPADHYAY received the B.Tech. degree in electrical and electronics from the International Institute of Information Technology, Bhubaneswar, in 2020. He is currently pursuing the Ph.D. degree with the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. His research interests include optimization for distributed and decentralized algorithms.

ABOLFAZL HASHEMI (Member, IEEE) received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in July 2014, and the M.S.E. and Ph.D. degrees in electrical and computer engineering from The University of Texas at Austin, Austin, TX, USA, in May 2016 and August 2020, respectively. From August 2020 to August 2021, he was a Postdoctoral Scholar with the Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin. Since August 2021, he has been an Assistant Professor with the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. His research interests include distributed machine learning, optimization, signal processing, and control. He was the 2019 Schmidt Science Fellows Award Nominee from UT Austin. He was a recipient of the Iranian National Elite Foundation Fellowship and the Best Student Paper Award Finalist from the 2018 American Control Conference.

• • •