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

Antesh Upadhyay and Abolfazl Hashemi

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. This analysis demonstrates 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 hard-to-verify assumption of *bounded client dissimilarity*, and secondly, it only requires smooth non-convex loss functions. We also perform extensive empirical analysis to verify the validity of our theoretical findings.

I. Introduction

The advent of edge devices, such as smartphones, sensors, wearables, etc., generates a massive amount of real-world data. Traditional Collaborative optimization and Machine Learning (ML), where we store the entire data on a central server and then process it to derive an inference, is not preferable because of privacy concerns and the ever-increasing communication requirements. This led to an advancement in distributed and collaborative ML paradigms such as consensus-based distributed optimization and Federated learning (FL) [1]–[6], where the latter is the focus of this paper. The algorithmic basis of FL, i.e., Federated Averaging (`FedAvg`) was presented originally in [7]. 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 — as opposed to their data in the traditional optimization and ML paradigm — 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 show the importance of communication vis-à-vis, `FedAvg` and other FL and consensus-based methods, and it has been an active area of research in terms of improving the efficiency and resiliency of such algorithms [8]–[14].

Several recent results, where improving communication efficiency is the core, focus mainly on reducing the number of communication rounds [7], [15], or the size of information during transmission [16]–[21]. 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 [22]–[26]. However, only a few articles in the literature deal with the impact of only downlink noise or both noises [27]. 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 [22], 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 focus on modification of the training of ML models. Reference [28] aims to counter the effect of noise by modifying the loss function to consider the addition of noise as a regularizer. Similarly [19], [29]–[31] 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. Different from these works, 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 required by existing works.

Recently, [27] studies the impact of both uplink and downlink noises with restrictive assumptions of strong-convexity and Bounded Client Dissimilarity (BCD) [32]. To avoid the client-drift [32], 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 [27] 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. Unlike the previous work, [27], 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

Antesh Upadhyay and Abolfazl Hashemi are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA.

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 2, 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 $\Omega(\sqrt{k})$. These scaling rates ensure that the noise appears as a higher-order term, not as a dominant term.

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

- We provide improved analysis of FL under the presence of uplink and downlink noise without using any constraining assumption which results in tighter convergence analysis.
- Following the analysis, we propose an improved SNR control strategy to diminish the adverse effect of uplink and downlink noises.
- We provide empirical results on both synthetic and real-world deep learning experiments to establish the validity of the proposed technique.

II. Preliminaries and System-Model

The setting of the problem follows the traditional FL scenario presented in [7] (see also Figure 1). In a standard FL setting, we have a central server and a set of n clients, each having their local training data. The i^{th} 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 $\boldsymbol{w} \in \mathbb{R}^d$. Then, $\boldsymbol{f}_i(\boldsymbol{w})$ is the expected loss over a sample $\boldsymbol{x}$ drawn from $\mathcal{D}_i$ with respect to a loss function ℓ for the i^{th} client. The primary objective of the central server is to minimize the loss $\boldsymbol{f}(\boldsymbol{w})$ over n clients, i.e.,

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

Also, to emulate a 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 unbiased stochastic gradient of their individual losses which is denoted by $\widetilde{\nabla} f_i(\boldsymbol{w};\mathcal{B})$ computed at $\boldsymbol{w}$ over a batch of samples $\mathcal{B}$. In addition, $\boldsymbol{K}$ denotes the communication rounds, and $\boldsymbol{E}$ represents the number of local iterations for each communication round.

The FL process can be thought of as a repetitive, 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, $\boldsymbol{w}_k$, 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.,

$$\boldsymbol{w}_{k,0}^{(i)} = \boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)}, \tag{2}$$

where $\boldsymbol{\nu}_k^{(i)} \in \mathbb{R}^d$ is the zero mean random downlink noise and $\boldsymbol{w}_{k,0}^{(i)}$ is the received model to the i^{th} client. Subsequently, the SNR we get for the i^{th} client for k^{th} downlink communication round can be written as,

$$\mathrm{SNR}_{k,(i)}^{D} = \frac{\mathbb{E}[||\boldsymbol{w}_k||^2]}{\mathbb{E}[||\boldsymbol{\nu}_k^{(i)}||^2]}. \tag{3}$$

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

$$\mathbb{E}\left[\left\|\boldsymbol{\nu}_k^{(i)}\right\|^2\right] = \boldsymbol{N}_{k,i}^2. \tag{4}$$

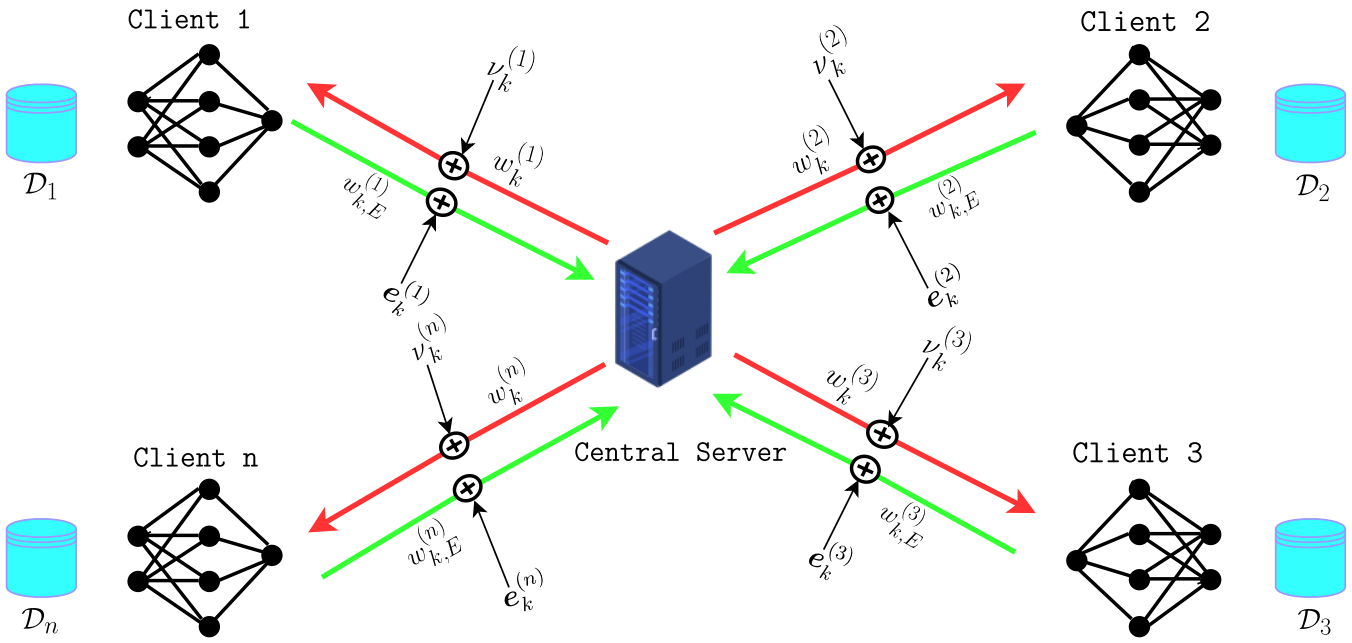


Fig. 1: Problem Setting: an FL system with uplink and downlink communication noise.

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,

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

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,

$$\boldsymbol{w}_{k,0}^{(i)} - \boldsymbol{w}_{k,E}^{(i)} + \boldsymbol{e}_k^{(i)}, \tag{6}$$

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

$$\text{SNR}_{k,(i)}^{U} = \frac{\mathbb{E}[||\boldsymbol{w}_{k,0}^{(i)} - \boldsymbol{w}_{k,E}^{(i)}||^2]}{\mathbb{E}[||\boldsymbol{e}_k^{(i)}||^2]}. \tag{7}$$

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

$$\mathbb{E}\left[\left\|\boldsymbol{e}_k^i\right\|^2\right] = \boldsymbol{E}_{k,i}^2. \tag{8}$$

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

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \frac{1}{r}\sum_{i \in \mathcal{S}_k} (\boldsymbol{w}_{k,0}^{(i)} - \boldsymbol{w}_{k,E}^{(i)} + \boldsymbol{e}_k^{(i)}), \tag{9}$$

and the process continues again for all the communication rounds.

Algorithm 1 `Noisy-FedAvg`

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.
for $k = 0, \ldots, K-1$ **do**
  Server sends $\boldsymbol{w}_k$ to a set $\mathcal{S}_k$ of r clients chosen uniformly at random without replacement.
  for client $i \in \mathcal{S}_k$ **do**
    Downlink communication: Broadcasting $\boldsymbol{w}_k$ through a noisy downlink communication channel having zero mean. Set $\boldsymbol{w}_{k,0}^{(i)} = \boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)}$, where $\boldsymbol{\nu}_k^{(i)}$ is the downlink noise.
    for $\tau = 0, \ldots, E-1$ **do**
      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. $\widetilde{\nabla} f_i(\boldsymbol{w}_{k,\tau}^{(i)}; \mathcal{B}_{k,\tau}^{(i)})$.
      Update $\boldsymbol{w}_{k,\tau+1}^{(i)} = \boldsymbol{w}_{k,\tau}^{(i)} - \eta_k \widetilde{\nabla} f_i(\boldsymbol{w}_{k,\tau}^{(i)}; \mathcal{B}_{k,\tau}^{(i)})$.
    end for
    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.
  end for
  Update $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \frac{1}{r}\sum_{i \in \mathcal{S}_k}(\boldsymbol{w}_{k,0}^{(i)} - \boldsymbol{w}_{k,E}^{(i)} + \boldsymbol{e}_k^{(i)})$.
end for

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 [14], [32]. Assumption 4 is referred to as **Noise model** is also used in [27].

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 $f(\boldsymbol{w})$.*

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 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 samples in client i for τ^{th} local iteration.*

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

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. [27], [32], [33] is the BCD assumption, i.e.,

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

where G is a large constant. Furthermore, in contrast to [27], [33], We do not make any assumption about the *strong convexity* of loss function.

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

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

where $\boldsymbol{e}_t$ and $\boldsymbol{\nu}_t$ 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 1 (Smooth non-convex case for `Noisy-SGD`). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a L-smooth non-convex function and $w^* = \text{argmin}_w f(w)$. Consider the noisy-SGD method with the update in eq.* (11)*. Let $\boldsymbol{e}_t$ and $\boldsymbol{\nu}_t$ satisfy Assumption 4 and the stochastic gradient satisfy Assumption 3. If, $\eta_t = \eta$ for all $t \in \{0, \ldots, T-1\}$ noisy-SGD satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[||\nabla f(\boldsymbol{w}_t)||^2]] \leq \frac{2(f(\boldsymbol{w}_0) - f(\boldsymbol{w}^*))}{T\eta} + \eta L\sigma^2 + \underbrace{\frac{L^2}{T}\sum_{t=0}^{T-1}\boldsymbol{N}_t^2}_{\textit{Term I}} + \underbrace{\frac{\eta L}{T}\sum_{t=0}^{T-1}\boldsymbol{E}_t^2}_{\textit{Term II}}. \tag{12}$$

Proof. Using L-smoothness assumption we can obtain,

$$f(\boldsymbol{w}_{t+1}) \leq f(\boldsymbol{w}_t) + \langle \nabla f(\boldsymbol{w}_t), \boldsymbol{w}_{t+1} - \boldsymbol{w}_t \rangle + \frac{L}{2}||\boldsymbol{w}_{t+1} - \boldsymbol{w}_t||^2. \tag{13}$$

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

$$f(\boldsymbol{w}_{t+1}) \leq f(\boldsymbol{w}_t) - \eta \langle \nabla f(\boldsymbol{w}_t), \boldsymbol{e}_t + \nabla \widetilde{f}(\boldsymbol{w}_t + \boldsymbol{\nu}_t; \mathcal{B}_t) \rangle + \frac{\eta^2 L}{2}||\boldsymbol{e}_t + \nabla \widetilde{f}(\boldsymbol{w}_t + \boldsymbol{\nu}_t; \mathcal{B}_t)||^2. \tag{14}$$

Now, taking expectation w.r.t data and $\boldsymbol{e}_t$ alongside the independence assumption of noises in eq. (14) we get

$$\mathbb{E}[f(\boldsymbol{w}_{t+1})] \leq \mathbb{E}[f(\boldsymbol{w}_t)] - \eta \mathbb{E}[\langle \nabla f(\boldsymbol{w}_t), \nabla f(\boldsymbol{w}_t + \boldsymbol{\nu}_t) \rangle] + \frac{\eta^2 L}{2}\mathbb{E}[||\boldsymbol{e}_t||^2] + \frac{\eta^2 L}{2}\sigma^2 + \frac{\eta^2 L}{2}\mathbb{E}[||\nabla f(\boldsymbol{w}_t + \boldsymbol{\nu}_t)||^2]. \tag{15}$$

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

$$-\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \frac{1}{2}(\|\boldsymbol{a} - \boldsymbol{b}\|^2 - \|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2). \tag{16}$$

Using this in eq. (15) we get

$$\mathbb{E}[f(\boldsymbol{w}_{t+1})] \leq \mathbb{E}[f(\boldsymbol{w}_t)] + \frac{\eta}{2}\mathbb{E}\Big[||\nabla f(\boldsymbol{w}_t) - \nabla f(\boldsymbol{w}_t + \boldsymbol{\nu}_t)||^2 - ||\nabla f(\boldsymbol{w}_t + \boldsymbol{\nu}_t)||^2 - ||\nabla f(\boldsymbol{w}_t)||^2\Big] + \frac{\eta^2 L}{2}(\boldsymbol{E}_t^2 + \sigma^2) + \frac{\eta^2 L}{2}\mathbb{E}[||\nabla f(\boldsymbol{w}_t + \boldsymbol{\nu}_t)||^2]. \tag{17}$$

If $\eta \leq \frac{1}{L}$, we can drop $\mathbb{E}[||\nabla f(\boldsymbol{w}_t + \boldsymbol{\nu}_t)||^2$ as it will appear with a negative sign in the RHS of eq. (17). Consequently, using L-smoothness yields

$$\mathbb{E}[f(\boldsymbol{w}_{t+1})] \leq \mathbb{E}[f(\boldsymbol{w}_t)] + \frac{\eta}{2}\mathbb{E}\Big[L^2||(\boldsymbol{\nu}_t)||^2 - ||\nabla f(\boldsymbol{w}_t)||^2\Big] + \frac{\eta^2 L}{2}(\boldsymbol{E}_t^2 + \sigma^2). \tag{18}$$

Taking expectation w.r.t. $\boldsymbol{\nu}_t$ we have

$$\mathbb{E}[f(\boldsymbol{w}_{t+1})] \leq \mathbb{E}[f(\boldsymbol{w}_t)] + \frac{\eta}{2}L^2\boldsymbol{N}_t^2 - \frac{\eta}{2}\mathbb{E}[||\nabla f(\boldsymbol{w}_t)||^2] + \frac{\eta^2 L}{2}(\boldsymbol{E}_t^2 + \sigma^2). \tag{19}$$

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

Corollary 1.1. *If $\eta \leq \frac{1}{L}$ and $\eta = \mathcal{O}(\frac{1}{\sqrt{T}})$ and a SNR control strategy is employed such that $\frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{E}_t^2 = \mathcal{O}(T^{-\delta_1})$ and $\frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{N}_t^2 = \mathcal{O}(\frac{1}{T^{0.5+\delta_2}})$ 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.*

In what follows, we build upon Theorem 1 to present Theorem 2, which holds for both partial and full client participation, IID, and non-IID data distribution.

Theorem 2 (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 $\zeta := 8\eta^2L^2E^2\Big(\frac{(n-r)}{r(n-1)} + \frac{2\eta LE}{3}\Big)$. Sample k^* from $\mathbb{P}$ uniformly. Then, for $K \geq \max\Big(\frac{1024r^3}{9\gamma^2}(\frac{1}{\gamma^2-16})^2, \frac{4r}{\gamma^2}\Big)$,*

$$\mathbb{E}[\|\nabla f(\boldsymbol{w}_{k^*})\|^2] \leq \frac{8\gamma L f(\boldsymbol{w}_0)}{\sqrt{rK}} + \frac{4}{\gamma E}\sqrt{\frac{r}{K}}\Big(\frac{1}{\gamma n}\sqrt{\frac{r}{K}}(1 + \frac{2nE}{3} + n) + \frac{1}{r} + \frac{(n-r)}{r(n-1)}\Big)\sigma^2 + \frac{4}{\gamma E^2 K\sqrt{rK}}\sum_{k=0}^{K-1}\boldsymbol{E}_k^2 + \frac{4L^2}{EK}\Big(1 + 4E + \frac{2}{\gamma E}\sqrt{\frac{r}{K}}(1 + 2E^2\{\frac{3}{\gamma E^2}\sqrt{\frac{r}{K}} + 2(2 + \frac{3}{\gamma^2E^2}\frac{r}{K})(\frac{2}{3\gamma}\sqrt{\frac{r}{K}} + \frac{(n-r)}{r(n-1)})\})\Big)\sum_{k=0}^{K-1}\boldsymbol{N}_k^2. \tag{20}$$

From Theorem 2, mirroring the result of Theorem 1, 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 2 is verified numerically as well by showing that uplink noise's impact is not as detrimental as downlink noise. For detailed proof of Theorem 2, refer Appendix I.

IV. Theory guided SNR control

Theorem 2 provides us with an 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. [27] and the references therein). In this section, we aim to establish an improved SNR control strategy following the improved analysis presented in Theorem 2. First, we stated the following corollary for an easier-to-interpret result.

Corollary 2.1. *Instate the notation and hypotheses of Theorem 2. Also, let* $\boldsymbol{E}_k^2 \leq \boldsymbol{e}^2$ *and* $\boldsymbol{N}_k^2 \leq \boldsymbol{\nu}^2$ *to be the maximum bounded variances for all* k*. Then, if* $K = \Omega(r^3)$*,*

$$\mathbb{E}[\|\nabla f(\boldsymbol{w}_{k^*})\|^2] = \mathcal{O}\Big(\frac{1}{E}\sqrt{\frac{r}{K}}\sigma^2 + \frac{1}{E^2\sqrt{rK}}\boldsymbol{e}^2 + \boldsymbol{\nu}^2\Big). \tag{21}$$

Following Corollary 2.1, 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 $\mathcal{O}(1)$. 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; more specifically, 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 $\epsilon-$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. [27]. Putting the requirement of strong convexity aside, the proposed policy in [27], while consuming more power resources[1] 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.

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 on the MNIST dataset.

A. Synthetic experiment

In the synthetic experiment, we train a linear regression model with m = 15000 samples. The samples $\{(\boldsymbol{x_j}, y_j)_{j=1}^m\}$ are generated based on the model $y_j = \langle \theta^*, \boldsymbol{x_j}\rangle + c_j$, where $\boldsymbol{\theta}^* \in \mathbb{R}^{60}$, the j^{th} input $\boldsymbol{x_j} \sim \mathcal{N}(0, I_{60})$, and noise $c_j \sim \mathcal{N}(0, 0.05)$. This dataset is generated such that the $(samples \times features)$ 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 2), $L = 1$, $E = 5$ and $K = 100$ and BS (local batch size) $= 16$. In each round, 20% of the clients participate based on random selection, which leads to $r = 10$. Now from Theorem 2, we have $\eta_k = \frac{1}{\gamma LE}\sqrt{\frac{r}{K}}$, i.e. $\eta_k = 0.0035$.

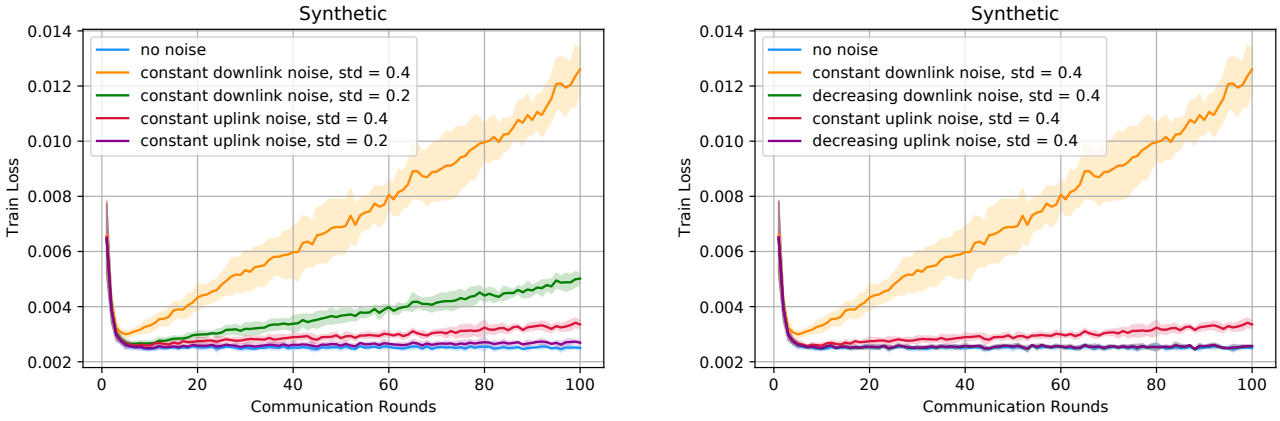
We first consider a constant noise setting where we add both uplink and downlink noise to the communicated messages where the noises are sampled from a Gaussian distribution having zero mean i.e., $\boldsymbol{e}_k^{(i)} \sim \mathcal{N}(0, \upsilon^2)$ and $\boldsymbol{\nu}_k^{(i)} \sim \mathcal{N}(0, \upsilon^2)$. We consider two choices of $\upsilon = 0.2$ and $\upsilon = 0.4$. We visualize the impact of adding noises in Figure 2a; as the figure demonstrates, consistent with the result of Theorem 2, the effect of downlink noise is more severe than the uplink noise and results in model 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 2 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 $\Omega(\frac{1}{E^2k})$ and $\Omega(\frac{1}{\sqrt{k}})$, respectively. We can observe the results from Fig. 2b and see how it converges almost in tandem with the noise-free case, as predicted by Corollary 2.1.

B. Deep learning experiment

To check the validity of our theory on a real-world dataset, we run deep learning experiments on the MNIST dataset, both for IID and non-IID settings. For this purpose, 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.

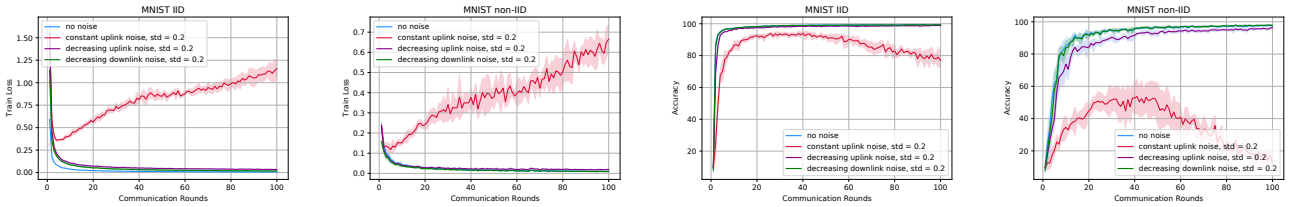
[1]Considering non-convex, L-smooth problems, the result in [27] seems to require $\Omega(k)$ scaling for both uplink and downlink noises.



(a) Constant noise

(b) Decreasing noise

Fig. 2: Linear regression: Comparing the impact of uplink and downlink noise with (left) and without (right) SNR control.



(a) IID MNIST: Train loss

(b) non-IID MNIST: Train loss

(c) IID MNIST: Accuracy

(d) non-IID MNIST: Accuracy

Fig. 3: Train loss (top two figures) and test accuracy (bottom two figures) in the presence of uplink and downlink noises, with and without SNR control. Observe that these plots do not have any graphs for constant downlink noise. In theorem 2 we established that the effect of downlink noise is more detrimental than uplink noise, which in the case of these examples caused gradient explosion.

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, $BS = 16$, and learning rate, $\eta_k = 0.005$.

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 $\sim \mathcal{N}(0, \upsilon^2)$, where $\upsilon = 0.2$. Again, to imitate the noisy transmission channels, we add both uplink and downlink noise to the communicated messages. The results are shown in Figure 3. We can visualize from the figure that the effect of noises inhibits the model from converging. One noteworthy observation is the gradient explosion due to downlink noise, which shows that the effect of downlink noise has more adverse effects than uplink noise. 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 I
PROOF OF NOISY-FEDAVG

Theorem 3 (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 $k \in \{0, \dots, K-1\}$ such that $\mathbb{P}(k) = \frac{(1+\zeta)^{(K-1-k)}}{\sum_{k=0}^{K-1}(1+\zeta)^k}$ where $\zeta := 8\eta^2L^2E^2\Big(\frac{(n-r)}{r(n-1)} + \frac{2\eta LE}{3}\Big)$. Sample k^* from $\mathbb{P}$ uniformly. Then, for $K \geq \max\Big(\frac{1024r^3}{9\gamma^2}(\frac{1}{\gamma^2-16})^2, \frac{4r}{\gamma^2}\Big)$:*

$$\mathbb{E}[\|\nabla f(\boldsymbol{w}_{k^*})\|^2] \leq \frac{8\gamma L f(\boldsymbol{w}_0)}{\sqrt{rK}} + \frac{4}{\gamma E}\sqrt{\frac{r}{K}}\Big(\frac{1}{\gamma n}\sqrt{\frac{r}{K}}(1 + \frac{2nE}{3} + n) + \frac{1}{r} + \frac{(n-r)}{r(n-1)}\Big)\sigma^2 + \frac{4}{\gamma E^2 K\sqrt{rK}}\sum_{k=0}^{K-1}\boldsymbol{E}_k^2$$
$$+\frac{4L^2}{EK}\Big(1 + 4E + \frac{2}{\gamma E}\sqrt{\frac{r}{K}}(1 + 2E^2\{\frac{3}{\gamma E^2}\sqrt{\frac{r}{K}} + 2(2+\frac{3}{\gamma^2E^2}\frac{r}{K})(\frac{2}{3\gamma}\sqrt{\frac{r}{K}} + \frac{(n-r)}{r(n-1)})\})\Big)\sum_{k=0}^{K-1}\boldsymbol{N}_k^2. \quad (22)$$

Proof. Using Lemma 1, for $\eta_k LE \leq \frac{1}{2}$, we can bound the per-round progress as:

$$\mathbb{E}[f(\boldsymbol{w}_{k+1})] \leq \mathbb{E}[f(\boldsymbol{w}_k)] - \frac{\eta_k(E-1)}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2] + 4\eta_k^2LE^2\Big(\frac{(n-r)}{r(n-1)} + \frac{2}{3}\eta_k LE\Big)\Big(\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2]\Big)$$
$$+\eta_k^2LE\Big(\frac{\eta_k LE}{n}\Big(1 + \frac{2nE}{3} + n\Big) + \frac{1}{r} + \frac{(n-r)}{r(n-1)}\Big)\sigma^2 + \frac{\eta_k^2L}{2r}\frac{1}{n}\sum_{i\in[n]}\boldsymbol{E}_{k,i}^2 + \frac{\eta_k L^2}{2}\Big(1 + 2\eta_k L + 4E\{1 + 3\eta_k^2L^2$$
$$+2\eta_k LE(2 + 3\eta_k^2L^2)(\frac{2}{3}\eta_k LE + \frac{(n-r)}{r(n-1)})\}\Big)\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2. \quad (23)$$

Now applying our earlier trick of 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 in eq. (23), we get for a constant learning rate of $\eta_k = \eta$, $\boldsymbol{E}_{k,i}^2 = \boldsymbol{E}_k^2$ and $\boldsymbol{N}_{k,i}^2 = \boldsymbol{N}_k^2$:

$$\mathbb{E}[f(\boldsymbol{w}_{k+1})] \leq \Big(1 + 8\eta^2L^2E^2\Big(\frac{(n-r)}{r(n-1)} + \frac{2\eta LE}{3}\Big)\Big)\mathbb{E}[f(\boldsymbol{w}_k)] - \frac{\eta(E-1)}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2] + \frac{\eta^2L}{2r}\frac{1}{n}\sum_{i\in[n]}\boldsymbol{E}_k^2$$
$$+\eta^2LE\Big(\frac{\eta LE}{n}\Big(1 + \frac{2nE}{3} + n\Big) + \frac{1}{r} + \frac{(n-r)}{r(n-1)}\Big)\sigma^2 + \frac{\eta L^2}{2}\Big(1 + 2\eta L + 4E\{1 + 3\eta^2L^2$$
$$+2\eta LE(2 + 3\eta^2L^2)(\frac{2\eta LE}{3} + \frac{(n-r)}{r(n-1)})\}\Big)\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_k^2. \quad (24)$$

For ease of notation, define $\zeta := 8\eta^2L^2E^2\Big(\frac{(n-r)}{r(n-1)} + \frac{2\eta LE}{3}\Big)$, $\zeta_2 := \Big(\frac{\eta LE}{n}\Big(1 + \frac{2nE}{3} + n\Big) + \frac{1}{r} + \frac{(n-r)}{r(n-1)}\Big)$ and $\zeta_3 := \Big(1 + 2\eta L + 4E\{1 + 3\eta^2L^2 + 2\eta LE(2 + 3\eta^2L^2)(\frac{2\eta LE}{3} + \frac{(n-r)}{r(n-1)})\}\Big)$. Then, unfolding the recursion of eq. (24) from $k = 0$ through to $k = K-1$, we get:

$$\mathbb{E}[f(\boldsymbol{w}_K)] \leq (1+\zeta)^K f(\boldsymbol{w}_0) - \frac{\eta(E-1)}{2}\sum_{k=0}^{K-1}(1+\zeta)^{(K-1-k)}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2] + \eta^2LE\zeta_2\sigma^2\sum_{k=0}^{K-1}(1+\zeta)^{(K-1-k)}$$
$$+\frac{\eta^2L}{2r}\sum_{k=0}^{K-1}\boldsymbol{E}_k^2(1+\zeta)^{(K-1-k)} + \frac{\eta L^2}{2}\zeta_3\sum_{k=0}^{K-1}\boldsymbol{N}_k^2(1+\zeta)^{(K-1-k)}. \quad (25)$$

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. (25) 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:

$$\sum_{k=0}^{K-1} p_k \mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2] \leq \frac{4(1+\zeta)^K f(\boldsymbol{w}_0)}{\eta E \sum_{k'=0}^{K-1}(1+\zeta)^{k'}} + 4\eta L \zeta_2 \sigma^2 + \frac{2\eta L}{rE} \frac{\sum_{k=0}^{K-1} \boldsymbol{E}_k^2 (1+\zeta)^{(K-1-k)}}{\sum_{k'=0}^{K-1}(1+\zeta)^{k'}} + \frac{2L^2\zeta_3}{E} \frac{\sum_{k=0}^{K-1} \boldsymbol{N}_k^2 (1+\zeta)^{(K-1-k)}}{\sum_{k'=0}^{K-1}(1+\zeta)^{k'}} \tag{26}$$

$$= \frac{4\zeta f(\boldsymbol{w}_0)}{\eta E(1-(1+\zeta)^{-K})} + 4\eta L \zeta_2 \sigma^2 + \frac{2\eta L}{rE} \frac{\zeta \sum_{k=0}^{K-1} \boldsymbol{E}_k^2}{(1+\zeta)-(1+\zeta)^{-K+1}} + \frac{2L^2\zeta_3}{E} \frac{\zeta \sum_{k=0}^{K-1} \boldsymbol{N}_k^2}{(1+\zeta)-(1+\zeta)^{-K+1}} \tag{27}$$

where the last step 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,

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

Also,

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

Plugging these with ζ_2 and ζ_3 in eq. (27), we have for $\zeta K < 1$:

$$\sum_{k=0}^{K-1} p_k \mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2] \leq \frac{4f(\boldsymbol{w}_0)}{\eta E K(1-\zeta K)} + 4\eta L E \Big(\frac{\eta L}{n} \Big(1 + \frac{2nE}{3} + n\Big) + \frac{1}{rE} + \frac{(n-r)}{r(n-1)E} \Big) \sigma^2 + \frac{2\eta L}{rE} \frac{\sum_{k=0}^{K-1} \boldsymbol{E}_k^2}{K(1-\zeta K)} + \frac{2L^2}{EK(1-\zeta K)} \Big(1 + 2\eta L + 4E\{1 + 3\eta^2 L^2 + 2\eta L E(2 + 3\eta^2 L^2) (\frac{2\eta L E}{3} + \frac{(n-r)}{r(n-1)})\} \Big) \sum_{k=0}^{K-1} \boldsymbol{N}_k^2. \tag{28}$$

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 $K \geq \frac{1024 r^3}{9\gamma^2}(\frac{1}{\gamma - 16})^2$. Thus, we should have $K \geq \max\Big(\frac{1024 r^3}{9\gamma^2}(\frac{1}{\gamma^2 - 16})^2, \frac{4r}{\gamma^2} \Big)$. Putting $\eta = \frac{1}{\gamma LE}\sqrt{\frac{r}{K}}$ in eq. (28) and also using $1 - \zeta K \geq \frac{1}{2}$, we get:

$$\sum_{k=0}^{K-1} p_k \mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2] \leq \frac{8\gamma L f(\boldsymbol{w}_0)}{\sqrt{rK}} + \frac{4}{\gamma E}\sqrt{\frac{r}{K}} \Big(\frac{1}{\gamma n}\sqrt{\frac{r}{K}}(1 + \frac{2nE}{3} + n) + \frac{1}{r} + \frac{(n-r)}{r(n-1)} \Big) \sigma^2 + \frac{4}{\gamma E^2 K \sqrt{rK}} \sum_{k=0}^{K-1} \boldsymbol{E}_k^2 + \frac{4L^2}{EK} \Big(1 + 4E + \frac{2}{\gamma E}\sqrt{\frac{r}{K}}(1 + 2E^2\{\frac{3}{\gamma E^2}\sqrt{\frac{r}{K}} + 2(2 + \frac{3}{\gamma^2 E^2}\frac{r}{K})(\frac{2}{3\gamma}\sqrt{\frac{r}{K}} + \frac{(n-r)}{r(n-1)})\}) \Big) \sum_{k=0}^{K-1} \boldsymbol{N}_k^2. \tag{29}$$

This finishes the proof. ■

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

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

Proof. Define

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

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

Then:

$$\nabla f(\boldsymbol{w}_k) = \frac{1}{n}\sum_{i\in[n]} \nabla f_i(\boldsymbol{w}_k) \tag{30}$$

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \eta_k \Big[\frac{1}{r}\sum_{i\in\mathcal{S}_k}\Big(\boldsymbol{e}_k^{(i)} + \sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)} + \nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\Big)\Big]. \tag{31}$$

$$\boldsymbol{w}_{k,\tau}^{(i)} = \boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)} - \eta_k\Big(\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)} + \nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\Big). \tag{32}$$

$$\overline{\boldsymbol{w}}_{k,\tau} = \boldsymbol{w}_k + \frac{1}{n}\sum_{i\in[n]}\boldsymbol{\nu}_k^{(i)} - \eta_k\Big[\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t} + \frac{1}{n}\sum_{i\in[n]}\Big(\nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\Big)\Big]. \tag{33}$$

$$\mathbb{E}_{\{\mathcal{B}_{k,\tau}^{(i)}\}_{i=1}^n}[\widehat{\boldsymbol{u}}_{k,\tau}] = \boldsymbol{u}_{k,\tau}. \tag{34}$$

$$\mathbb{E}\Big[\Big\|\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}\Big\|^2\Big] \le \tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|\boldsymbol{u}_{k,t}\|^2] + \frac{\tau\sigma^2}{n}. \tag{35}$$

$$\mathbb{E}\Big[\Big\|\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}^{(i)}\Big\|^2\Big] \le \tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2] + \tau\sigma^2. \tag{36}$$

$$\mathbb{E}\Big[\Big\|\boldsymbol{\nu}_k^i\Big\|^2\Big] = \boldsymbol{N}_{k,i}^2. \tag{37}$$

$$\mathbb{E}\Big[\Big\|\boldsymbol{e}_k^i\Big\|^2\Big] = \boldsymbol{E}_{k,i}^2. \tag{38}$$

Recall that σ^2 is the maximum variance of the local (client-level) stochastic gradients. In eq. (35), 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. (36), 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. (37) and eq. (38) follows since both downlink and uplink noises have zero mean.

Next, using the L-smoothness of f and eq. (31), we get

$$\mathbb{E}[f(\boldsymbol{w}_{k+1})] \le \mathbb{E}[f(\boldsymbol{w}_k)] \underbrace{- \mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k), \eta_k\Big[\frac{1}{r}\sum_{i\in\mathcal{S}_k}\Big(\boldsymbol{e}_k^{(i)} + \sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)} + \nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\Big)\Big]\Big\rangle\Big]}_{\text{(A)}} \tag{39}$$

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

Now using (A):

$$
\begin{aligned}
&-\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k),\eta_k\Big[\frac{1}{r}\sum_{i\in\mathcal{S}_k}\Big(\boldsymbol{e}_k^{(i)}+\sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}+\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\Big)\Big]\Big\rangle\Big]\\
&\qquad=-\eta_k\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k),\Big[\frac{1}{r}\sum_{i\in\mathcal{S}_k}\Big(\boldsymbol{e}_k^{(i)}+\sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}+\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\Big)\Big]\Big\rangle\Big]\\
&\qquad=\underbrace{-\eta_k\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k),\frac{1}{r}\sum_{i\in\mathcal{S}_k}\boldsymbol{e}_k^{(i)}\Big\rangle\Big]}_{(A_1)}\underbrace{-\eta_k\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k),\frac{1}{r}\sum_{i\in\mathcal{S}_k}\sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}\Big\rangle\Big]}_{(A_2)}\\
&\qquad\qquad\underbrace{-\eta_k\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k),\frac{1}{r}\sum_{i\in\mathcal{S}_k}(\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\Big\rangle\Big]}_{(A_3)}
\end{aligned}
$$

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

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

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

$$
\langle\boldsymbol{a},\boldsymbol{b}\rangle=\frac{1}{2}(\|\boldsymbol{a}\|^2+\|\boldsymbol{b}\|^2-\|\boldsymbol{a}-\boldsymbol{b}\|^2)\tag{40}
$$

Using this we will get A_2 as:

$$
-\eta_k\sum_{\tau=0}^{E-1}\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k),\boldsymbol{u}_{k,\tau}\Big\rangle\Big]=\frac{\eta_k}{2}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)-\boldsymbol{u}_{k,\tau}\|^2]-\frac{\eta_k E}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]-\frac{\eta_k}{2}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\boldsymbol{u}_{k,\tau}\|^2]\tag{41}
$$

Again using A_3 :

$$
\begin{aligned}
&-\eta_k\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k),\frac{1}{r}\sum_{i\in\mathcal{S}_k}(\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\Big\rangle\Big]\\
&\qquad=-\eta_k\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k),\frac{1}{n}\sum_{i\in[n]}(\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k)\Big\rangle\Big]\\
&\qquad=-\eta_k\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k),\frac{1}{n}\sum_{i\in[n]}(\nabla f_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)})\Big\rangle\Big]+\eta_k\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k),\frac{1}{n}\sum_{i\in[n]}\nabla f_i(\boldsymbol{w}_k)\Big\rangle\Big]\\
&=\underbrace{\frac{\eta_k}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)-\frac{1}{n}\sum_{i\in[n]}\nabla f_i(\boldsymbol{w}_k+\boldsymbol{n}_k^i)\|^2]}_{A_3'}-\frac{\eta_k}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]-\frac{\eta_k}{2}\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}\nabla f_i(\boldsymbol{w}_k+\boldsymbol{n}_k^i)\|^2]\\
&\qquad+\eta_k\mathbb{E}[\langle\nabla f(\boldsymbol{w}_k),\nabla f(\boldsymbol{w}_k)\rangle]
\end{aligned}
$$

The last step uses eq. (40) and eq. (30). Now let's reduce A_3' :

$$
\begin{aligned}
&\frac{\eta_k}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)-\frac{1}{n}\sum_{i\in[n]}\nabla f_i(\boldsymbol{w}_k+\boldsymbol{n}_k^i)\|^2]&(42)\\
&=\frac{\eta_k}{2}\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}\nabla f_i(\boldsymbol{w}_k)-\frac{1}{n}\sum_{i\in[n]}\nabla f_i(\boldsymbol{w}_k+\boldsymbol{n}_k^i)\|^2]&(43)\\
&\le\frac{\eta_k L^2}{2}\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2&(44)
\end{aligned}
$$

The eq. (43) follows by using eq. (30), while eq. (44) follows from the L-smoothness of f_i, eq. (37) and independence of noises. So, A_3 now becomes:

$$
\begin{aligned}
&-\eta_k\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k),\frac{1}{r}\sum_{i\in\mathcal{S}_k}(\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\Big\rangle\Big]\\
&\le\frac{\eta_k L^2}{2}\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2-\frac{\eta_k}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]-\frac{\eta_k}{2}\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}\nabla f_i(\boldsymbol{w}_k+\boldsymbol{n}_k^i)\|^2]+\eta_k\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]\\
&\le\frac{\eta_k L^2}{2}\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2+\frac{\eta_k}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]-\frac{\eta_k}{2}\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}\nabla f_i(\boldsymbol{w}_k+\boldsymbol{n}_k^i)\|^2]&(45)
\end{aligned}
$$

So, finally by combining eq. (41) and eq. (45), A becomes:

$$-\mathbb{E}\Big[\Big\langle\nabla f(\boldsymbol{w}_k),\eta_k\Big[\frac{1}{r}\sum_{i\in\mathcal{S}_k}\Big(\boldsymbol{e}_k^{(i)}+\sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}+\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\Big)\Big]\Big\rangle\Big]\leq\frac{\eta_k}{2}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)-\boldsymbol{u}_{k,\tau}\|^2]$$
$$-\frac{\eta_k E}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]-\frac{\eta_k}{2}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\boldsymbol{u}_{k,\tau}\|^2]+\frac{\eta_k L^2}{2}\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2+\frac{\eta_k}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]-\frac{\eta_k}{2}\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}\nabla f_i(\boldsymbol{w}_k+\boldsymbol{n}_k^i)\|^2]$$
$$\leq\frac{\eta_k}{2}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)-\boldsymbol{u}_{k,\tau}\|^2]-\frac{\eta_k(E-1)}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]-\frac{\eta_k}{2}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\boldsymbol{u}_{k,\tau}\|^2]+\frac{\eta_k L^2}{2}\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2$$
$$-\frac{\eta_k}{2}\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}\nabla f_i(\boldsymbol{w}_k+\boldsymbol{n}_k^i)\|^2] \quad (46)$$

Now using (B):

$$\frac{L}{2}\mathbb{E}\Big[\Big\|\eta_k\Big[\frac{1}{r}\sum_{i\in\mathcal{S}_k}\Big(\boldsymbol{e}_k^{(i)}+\sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}+\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\Big)\Big]\Big\|^2\Big]=\frac{L}{2}\eta_k^2\Big(\mathbb{E}[\|\frac{1}{r}\sum_{i\in\mathcal{S}_k}\boldsymbol{e}_k^{(i)}\|^2$$
$$+\mathbb{E}[\|\frac{1}{r}\sum_{i\in\mathcal{S}_k}\sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}\|^2+\mathbb{E}[\|\frac{1}{r}\sum_{i\in\mathcal{S}_k}\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\|^2+2\mathbb{E}[\langle\frac{1}{r}\sum_{i\in\mathcal{S}_k}\boldsymbol{e}_k^{(i)},\frac{1}{r}\sum_{i\in\mathcal{S}_k}\sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}\rangle]$$
$$+2\mathbb{E}[\langle\frac{1}{r}\sum_{i\in\mathcal{S}_k}\boldsymbol{e}_k^{(i)},\frac{1}{r}\sum_{i\in\mathcal{S}_k}(\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)}))\rangle]+2\mathbb{E}[\langle\frac{1}{r}\sum_{i\in\mathcal{S}_k}\sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)},$$
$$\frac{1}{r}\sum_{i\in\mathcal{S}_k}(\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)}))\rangle]\Big) \quad (47)$$

Using the fact that $\mathbb{E}[\|\boldsymbol{e}_k^i\|]=0$ and Young's Inequality in eq. (47), we get:

$$\frac{L}{2}\eta_k^2\mathbb{E}\Big[\Big\|\Big[\frac{1}{r}\sum_{i\in\mathcal{S}_k}\Big(\boldsymbol{e}_k^{(i)}+\sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}+\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\Big)\Big]\Big\|^2\Big]$$
$$\leq\frac{L}{2}\eta_k^2\Big(\underbrace{\mathbb{E}[\|\frac{1}{r}\sum_{i\in\mathcal{S}_k}\boldsymbol{e}_k^{(i)}\|^2]}_{B_1}+2\underbrace{\mathbb{E}[\|\frac{1}{r}\sum_{i\in\mathcal{S}_k}\sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}\|^2]}_{B_2}+2\underbrace{\mathbb{E}[\|\frac{1}{r}\sum_{i\in\mathcal{S}_k}\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\|^2]}_{B_3}\Big) \quad (48)$$

Starting withB_1 :

$$\mathbb{E}[\|\frac{1}{r}\sum_{i\in\mathcal{S}_k}\boldsymbol{e}_k^{(i)}\|^2]=\frac{n(r-1)}{r(n-1)}\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}\boldsymbol{e}_k^{(i)}\|^2]+\frac{(n-r)}{r(n-1)}\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\boldsymbol{e}_k^{(i)}\|^2] \quad (49)$$
$$=\frac{(r-1)}{nr(n-1)}\sum_{i\in[n]}\boldsymbol{E}_{k,i}^2+\frac{(n-r)}{nr(n-1)}\sum_{i\in[n]}\boldsymbol{E}_{k,i}^2 \quad (50)$$
$$=\frac{1}{nr}\sum_{i\in[n]}\boldsymbol{E}_{k,i}^2 \quad (51)$$

Here, eq. (49) follows due to expectation w.r.t $\mathcal{S}_k$ and eq. (50) follows due to expectation w.r.t uplink noise and it's independence. Now let's focus on B_2 :

$$2\mathbb{E}[\|\frac{1}{r}\sum_{i\in\mathcal{S}_k}\sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}\|^2]\leq\frac{2n(r-1)E}{r(n-1)}\Big(\sum_{\tau=0}^{E-1}\mathbb{E}[\|\boldsymbol{u}_{k,\tau}\|^2]+\frac{\sigma^2}{n}\Big)+\frac{2(n-r)E}{r(n-1)}\Big(\frac{1}{n}\sum_{i\in[n]}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,\tau}^i)\|^2]+\sigma^2\Big) \quad (52)$$

The eq. (52) follows due to expectation w.r.t $\mathcal{S}_k$, eq. (34), eq. (35) and eq. (36). Again, let's use B_3 :

$$2\mathbb{E}[\|\frac{1}{r}\sum_{i\in\mathcal{S}_k}\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\|^2]\leq2\mathbb{E}[\frac{1}{r}\sum_{i\in\mathcal{S}_k}\|\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\|^2] \quad (53)$$
$$\leq2L^2\mathbb{E}[\frac{1}{r}\sum_{i\in\mathcal{S}_k}\|\boldsymbol{\nu}_k^{(i)}\|^2] \quad (54)$$
$$\leq2L^2\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2 \quad (55)$$

Here we used Jensen's inequality to reach eq. (53). Again, eq. (54) follows due to the L-smoothness of f_i and eq. (55) follows due to expectation w.r.t $\mathcal{S}_k$ and eq. (37). So, finally by combining B_1, B_2 and B_3, (B) becomes:

$$\begin{aligned}
&\frac{L}{2}\mathbb{E}\Big[\Big\|\eta_k\Big[\frac{1}{r}\sum_{i\in\mathcal{S}_k}\Big(\boldsymbol{e}_k^{(i)}+\sum_{\tau=0}^{E-1}\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}+\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\Big)\Big]\Big\|^2\Big]\\
&\leq\frac{L\eta_k^2}{2}\Big(\frac{1}{nr}\sum_{i\in[n]}\boldsymbol{E}_{k,i}^2+\frac{2n(r-1)E}{r(n-1)}\Big(\sum_{\tau=0}^{E-1}\mathbb{E}[\|\boldsymbol{u}_{k,\tau}\|^2]+\frac{\sigma^2}{n}\Big)+\frac{2(n-r)E}{r(n-1)}\Big(\frac{1}{n}\sum_{i\in[n]}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,\tau}^i)\|^2]+\sigma^2\Big)+2L^2\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2\Big)\\
&\leq\frac{L\eta_k^2}{2nr}\sum_{i\in[n]}\boldsymbol{E}_{k,i}^2+\frac{L^3\eta_k^2}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2+\frac{\eta_k^2LE}{r}\sigma^2+\frac{n(r-1)}{r(n-1)}\eta_k^2LE\Big(\sum_{\tau=0}^{E-1}\mathbb{E}[\|\boldsymbol{u}_{k,\tau}\|^2]\Big)+\frac{(n-r)}{r(n-1)}\eta_k^2LE\Big(\frac{1}{n}\sum_{i\in[n]}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,\tau}^i)\|^2]\Big)
\end{aligned}\tag{56}$$

Now, by putting eq. (46) and eq. (56) in eq. (39) we will get:

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{w}_{k+1})]\leq\mathbb{E}[f(\boldsymbol{w}_k)]&-\frac{\eta_k(E-1)}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]-\frac{\eta_k}{2}\Big(1-\eta_kLE\frac{n(r-1)}{r(n-1)}\Big)\sum_{\tau=0}^{E-1}\mathbb{E}[\|\boldsymbol{u}_{k,\tau}\|^2]\\
&+\underbrace{\frac{\eta_k}{2}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)-\boldsymbol{u}_{k,\tau}\|^2]}_{(M)}+\underbrace{\eta_k^2LE\frac{(n-r)}{r(n-1)}\Big(\frac{1}{n}\sum_{i\in[n]}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,\tau}^{(i)})\|^2]\Big)}_{(N)}+\frac{\eta_k^2LE}{r}\sigma^2\\
&+\frac{\eta_k^2L}{2nr}\sum_{i\in[n]}\boldsymbol{E}_{k,i}^2+\frac{\eta_kL^2}{2}\Big(1+2\eta_kL\Big)\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2-\frac{\eta_k}{2}\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}\nabla f_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)})\|^2]
\end{aligned}\tag{57}$$

We upper bound (M) and (N) using Lemma 2 and Lemma 3, respectively. Plugging in these bounds and dropping the last term of eq. (57), we get:

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{w}_{k+1})]\leq\mathbb{E}[f(\boldsymbol{w}_k)]&-\frac{\eta_k(E-1)}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]-\frac{\eta_k}{2}\underbrace{\Big(1-\eta_kLE\frac{n(r-1)}{r(n-1)}-2\eta_k^2L^2E^2\Big)}_{(C)}\sum_{\tau=0}^{E-1}\mathbb{E}[\|\boldsymbol{u}_{k,\tau}\|^2]\\
&+4\eta_k^2LE^2\Big(\frac{(n-r)}{r(n-1)}+\frac{2}{3}\eta_kLE\Big)\Big(\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2]\Big)+\eta_k^2LE\Big(\frac{\eta_kLE}{n}\Big(1+\frac{2nE}{3}+n\Big)+\frac{1}{r}+\frac{(n-r)}{r(n-1)}\Big)\sigma^2\\
&+\frac{\eta_k^2L}{2r}\frac{1}{n}\sum_{i\in[n]}\boldsymbol{E}_{k,i}^2+\frac{\eta_kL^2}{2}\Big(1+2\eta_kL+4E\{1+3\eta_k^2L^2+2\eta_kLE(2+3\eta_k^2L^2)(\frac{2}{3}\eta_kLE+\frac{(n-r)}{r(n-1)})\}\Big)\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2.
\end{aligned}\tag{58}$$

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

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{w}_{k+1})]\leq\mathbb{E}[f(\boldsymbol{w}_k)]&-\frac{\eta_k(E-1)}{2}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)\|^2]+4\eta_k^2LE^2\Big(\frac{(n-r)}{r(n-1)}+\frac{2}{3}\eta_kLE\Big)\Big(\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2]\Big)\\
&+\eta_k^2LE\Big(\frac{\eta_kLE}{n}\Big(1+\frac{2nE}{3}+n\Big)+\frac{1}{r}+\frac{(n-r)}{r(n-1)}\Big)\sigma^2+\frac{\eta_k^2L}{2r}\frac{1}{n}\sum_{i\in[n]}\boldsymbol{E}_{k,i}^2+\frac{\eta_kL^2}{2}\Big(1+2\eta_kL+4E\{1+3\eta_k^2L^2\\
&+2\eta_kLE(2+3\eta_k^2L^2)(\frac{2}{3}\eta_kLE+\frac{(n-r)}{r(n-1)})\}\Big)\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2.
\end{aligned}\tag{59}$$

■

Lemma 2. *For $\eta_kLE\leq\frac{1}{2}$:*

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

Proof. We have:

$$\mathbb{E}[\|\nabla f(\boldsymbol{w}_k) - \boldsymbol{u}_{k,\tau}\|^2] = \mathbb{E}[\|\nabla f(\boldsymbol{w}_k) - \nabla f(\overline{\boldsymbol{w}}_{k,\tau}) + \nabla f(\overline{\boldsymbol{w}}_{k,\tau}) - \boldsymbol{u}_{k,\tau}\|^2] \tag{60}$$

$$\leq 2\mathbb{E}[\|\nabla f(\boldsymbol{w}_k) - \nabla f(\overline{\boldsymbol{w}}_{k,\tau})\|^2] + 2\mathbb{E}[\|\nabla f(\overline{\boldsymbol{w}}_{k,\tau}) - \boldsymbol{u}_{k,\tau}\|^2] \tag{61}$$

$$\leq \underbrace{2L^2\mathbb{E}[\|\boldsymbol{w}_k - \overline{\boldsymbol{w}}_{k,\tau}\|^2]}_{M_1} + \underbrace{2\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{i\in[n]}(\nabla f_i(\overline{\boldsymbol{w}}_{k,\tau}) - \nabla f_i(\boldsymbol{w}_{k,\tau}^{(i)}))\Big\|^2\Big]}_{M_2} \tag{62}$$

Using M_1 :

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

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

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

$$-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\}, -\frac{1}{n}\sum_{i\in[n]}\boldsymbol{\nu}_k^{(i)}\rangle] + 2\mathbb{E}[\langle-\frac{1}{n}\sum_{i\in[n]}\boldsymbol{\nu}_k^{(i)}, \eta_k\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}\rangle]\Big) \tag{63}$$

Simplifying eq. (63) using the fact that $\mathbb{E}[\|\boldsymbol{n}_k^{(i)}\|] = 0$ and Young's Inequality we will get:

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

$$+2\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}\boldsymbol{\nu}_k^{(i)}\|^2]\Big) \leq 2L^2\Big(2\eta_k^2\mathbb{E}[\|\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}\|^2] + 3\eta_k^2\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\|^2] \tag{64}$$

$$+\frac{2}{n^2}\mathbb{E}[\|\sum_{i\in[n]}\boldsymbol{\nu}_k^{(i)}\|^2]\Big) \leq 2L^2\Big(2\eta_k^2\tau(\sum_{t=0}^{\tau-1}\mathbb{E}[\|\boldsymbol{u}_{k,t}\|^2] + \frac{\sigma^2}{n}) + 3\eta_k^2L^2\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2 + \frac{2}{n^2}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2\Big) \tag{65}$$

Equation (64) follows due to Jensen's Inequality and eq. (65) follows from the L-smoothness of f_i, eq. (35), eq. (37) and independence of noise. Now let's use M_2 :

$$2\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}(\nabla f_i(\overline{\boldsymbol{w}}_{k,\tau}) - \nabla f_i(\boldsymbol{w}_{k,\tau}^{(i)}))\|^2] \leq 2L^2\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\boldsymbol{w}_{k,\tau}^{(i)} - \overline{\boldsymbol{w}}_{k,\tau}\|^2] \leq 2L^2\frac{1}{n}\sum_{i\in[n]}\Big(\mathbb{E}[\|(\boldsymbol{n}_k^{(i)} - \frac{1}{n}\sum_{i\in[n]}\boldsymbol{n}_k^{(i)}) \tag{66}$$

$$+\eta_k(\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t} - \sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}^{(i)}) + \eta_k(\{\frac{1}{n}\sum_{i\in[n]}(\nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)}))\} - \{\nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})$$

$$-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\})\|^2]\Big) \tag{67}$$

Here eq. (66) follows from Jensen's Inequality and the L-smoothness of f_i. Now here we are using the same simplification process as used to simplify eq. (63). So, we get:

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

$$\underbrace{+ 3\mathbb{E}[\|\eta_k(\{\frac{1}{n}\sum_{i\in[n]}(\nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)}))\} - \{\nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\})\|^2]}_{(Z)}\Big) \tag{68}$$

Using (X):

$$2\mathbb{E}[\|(\boldsymbol{n}_k^{(i)} - \frac{1}{n}\sum_{i\in[n]}\boldsymbol{n}_k^{(i)})\|^2] = 2\mathbb{E}[\|\boldsymbol{n}_k^{(i)}\|^2] + 2\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}\boldsymbol{n}_k^{(i)}\|^2] - 4\mathbb{E}[\langle \boldsymbol{n}_k^{(i)}, \frac{1}{n}\sum_{i\in[n]}\boldsymbol{n}_k^{(i)}\rangle] \tag{69}$$

$$= 2\boldsymbol{N}_{k,i}^2 + \frac{2}{n^2}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2 - \frac{4}{n}\boldsymbol{N}_{k,i}^2 \tag{70}$$

$$= 2(1-\frac{2}{n})\boldsymbol{N}_{k,i}^2 + \frac{2}{n^2}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2 \tag{71}$$

Here eq. (70) follows due to eq. (37) and independence of noise. So, now moving on to (Y):

$$2\mathbb{E}[\|\eta_k(\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t} - \sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}^{(i)})\|^2] = 2\eta_k^2\mathbb{E}[\|\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t} - \sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}^{(i)}\|^2] \tag{72}$$

$$\leq 2\eta_k^2\tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|\widehat{\boldsymbol{u}}_{k,t} - \widehat{\boldsymbol{u}}_{k,t}^{(i)}\|^2] \tag{73}$$

$$= 2\eta_k^2\tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|\widehat{\boldsymbol{u}}_{k,t}\|^2 + \|\widehat{\boldsymbol{u}}_{k,t}^{(i)}\|^2 - 2\langle\widehat{\boldsymbol{u}}_{k,t}, \widehat{\boldsymbol{u}}_{k,t}^{(i)}\rangle] \tag{74}$$

Equation (73) 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. (74) to:

$$2\mathbb{E}[\|\eta_k(\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t} - \sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}^{(i)})\|^2] \leq 2\eta_k^2\tau\sum_{t=0}^{\tau-1}(\mathbb{E}[\|\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}\|^2] - \mathbb{E}[\|\widehat{\boldsymbol{u}}_{k,t}\|^2]) \tag{75}$$

$$\leq 2\eta_k^2\tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|\widehat{\boldsymbol{u}}_{k,\tau}^{(i)}\|^2] \tag{76}$$

$$\leq 2\eta_k^2\tau\sum_{t=0}^{\tau-1}(\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2] + \sigma^2) \tag{77}$$

Now using Lemma 3 for $\boldsymbol{n}_k LE \leq \frac{1}{2}$ in eq. (77), we get:

$$2\mathbb{E}[\|\eta_k(\sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t} - \sum_{t=0}^{\tau-1}\widehat{\boldsymbol{u}}_{k,t}^{(i)})\|^2] \leq 2\eta_k^2\tau^2(4\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2] + 4L^2(2+3\eta_k^2L^2)\boldsymbol{N}_{k,i}^2) + 2\eta_k^2(\tau^2+\tau)\sigma^2 \tag{78}$$

Next, using (Z):

$$3\mathbb{E}[\|\eta_k(\{\frac{1}{n}\sum_{i\in[n]}(\nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)}))\} - \{\nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})\})\|^2] \tag{79}$$

$$= 3\eta_k^2\mathbb{E}\Big[\|\frac{1}{n}\sum_{i\in[n]}(\nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)}))\|^2 + \|(\nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)}))\|^2$$

$$- 2\langle\frac{1}{n}\sum_{i\in[n]}(\nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)})), (\nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)}))\rangle\Big] \tag{80}$$

$$= 3\eta_k^2(\mathbb{E}[\|(\nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)}))\|^2] - \mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}(\nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)}))\|^2]) \tag{81}$$

$$\leq 3\eta_k^2\mathbb{E}[\|(\nabla\widetilde{f}_i(\boldsymbol{w}_k + \boldsymbol{\nu}_k^{(i)}; \mathcal{B}_{k,0}^{(i)}) - \nabla\widetilde{f}_i(\boldsymbol{w}_k; \mathcal{B}_{k,0}^{(i)}))\|^2] \tag{82}$$

$$\leq 3\eta_k^2L^2\mathbb{E}[\|\boldsymbol{\nu}_k^{(i)}\|^2] \leq 3\eta_k^2L^2\boldsymbol{N}_{k,i}^2 \tag{83}$$

We simplified eq. (80) using the similar fact that we used to simplify eq. (74). Next, we used the L-smoothness of f_i. Now

putting the results of eq. (71), eq. (78) and eq. (83) in eq. (68) we get:

$$2\mathbb{E}[\|\frac{1}{n}\sum_{i\in[n]}(\nabla f_i(\overline{\boldsymbol{w}}_{k,\tau})-\nabla f_i(\boldsymbol{w}_{k,\tau}^{(i)}))\|^2] \leq 2L^2\frac{1}{n}\sum_{i\in[n]}\Big(2(1-\frac{2}{n})\boldsymbol{N}_{k,i}^2+\frac{2}{n^2}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2+2\eta_k^2\tau^2(4\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2] \\ +4L^2(2+3\eta_k^2L^2)\boldsymbol{N}_{k,i}^2)+2\boldsymbol{n}_k^2(\tau^2+\tau)\sigma^2+3\eta_k^2L^2\boldsymbol{N}_{k,i}^2\Big) \\ \leq 16\eta_k^2L^2\tau^2\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}k)\|^2]+2L^2(2-\frac{2}{n}+3\eta_k^2L^2)\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2+16\eta_k^2L^4\tau^2(2+3\eta_k^2L^2)\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2 \\ +4\eta_k^2L^2(\tau^2+\tau)\sigma^2 \quad (84)$$

So, eq. (62) becomes:

$$\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)-\boldsymbol{u}_{k,\tau}\|^2] \leq 4\eta_k^2L^2\tau\sum_{t=0}^{\tau-1}\mathbb{E}[\|\boldsymbol{u}_{k,t}\|^2]+16\eta_k^2L^2\tau^2\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}k)\|^2]+4\eta_k^2L^2(\tau^2+\frac{\tau}{n}+\tau)\sigma^2 \\ +4L^2(1+3\eta_k^2L^2)\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2+16\eta_k^2L^4\tau^2(2+3\eta_k^2L^2)\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2 \quad (85)$$

Now summing up eq. (85) for all $\tau \in \{0,\ldots,E-1\}$, we get:

$$\sum_{\tau=0}^{E-1}\mathbb{E}[\|\nabla f(\boldsymbol{w}_k)-\boldsymbol{u}_{k,\tau}\|^2] \leq 2\eta_k^2L^2E^2\sum_{\tau=0}^{E-1}\mathbb{E}[\|\boldsymbol{u}_{k,\tau}\|^2]+\frac{16}{3}\eta_k^2L^2E^3\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2] \\ +2\eta_k^2L^2E^2(\frac{1}{n}+\frac{2E}{3}+1)\sigma^2+4L^2E(1+3\eta_k^2L^2+\frac{4}{3}\eta_k^2L^2E^2(2+3\eta_k^2L^2))\frac{1}{n}\sum_{i\in[n]}\boldsymbol{N}_{k,i}^2. \quad (86)$$

■

Lemma 3. *For $\eta_kLE \leq \frac{1}{2}$, we have:*

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

Proof.

$$\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2] = \mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})-\nabla f_i(\boldsymbol{w}_k)+\nabla f_i(\boldsymbol{w}_k)\|^2] \\ \leq 2\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2]+2\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})-\nabla f_i(\boldsymbol{w}_k)\|^2] \\ \leq 2\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2]+2L^2\mathbb{E}[\|\boldsymbol{w}_{k,t}^{(i)}-\boldsymbol{w}_k\|^2]. \quad (87)$$

But:

$$\mathbb{E}[\|\boldsymbol{w}_{k,t}^{(i)}-\boldsymbol{w}_k\|^2] = \mathbb{E}\Big[\Big\|\boldsymbol{\nu}_k^{(i)}-\eta_k\Big(\sum_{t'=0}^{t-1}\widehat{\boldsymbol{u}}_{k,t'}^{(i)}+\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\Big)\Big\|^2\Big] \\ = \mathbb{E}[\|\boldsymbol{\nu}_k^{(i)}\|^2]+\eta_k^2\mathbb{E}[\|\sum_{t'=0}^{t-1}\widehat{\boldsymbol{u}}_{k,t'}^{(i)}\|^2]+\eta_k^2\mathbb{E}[\|\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\|^2]+2\mathbb{E}[\langle\boldsymbol{\nu}_k^{(i)},-\eta_k\sum_{t'=0}^{t-1}\widehat{\boldsymbol{u}}_{k,t'}^{(i)}\rangle] \\ +2\mathbb{E}[\langle\boldsymbol{\nu}_k^{(i)},-\eta_k\{\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\}\rangle]+2\mathbb{E}[\langle-\eta_k\sum_{t'=0}^{t-1}\widehat{\boldsymbol{u}}_{k,t'}^{(i)},-\eta_k\{\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\}\rangle] \quad (88)$$

Now using the fact that $\mathbb{E}[\|\boldsymbol{n}_k^i\|]=0$ and Young's Inequality in eq. (88), we get:

$$\mathbb{E}[\|\boldsymbol{w}_{k,t}^{(i)}-\boldsymbol{w}_k\|^2] \leq 2\mathbb{E}[\|\boldsymbol{\nu}_k^{(i)}\|^2]+2\eta_k^2\mathbb{E}[\|\sum_{t'=0}^{t-1}\widehat{\boldsymbol{u}}_{k,t'}^{(i)}\|^2]+3\eta_k^2\mathbb{E}[\|\nabla\widetilde{f}_i(\boldsymbol{w}_k+\boldsymbol{\nu}_k^{(i)};\mathcal{B}_{k,0}^{(i)})-\nabla\widetilde{f}_i(\boldsymbol{w}_k;\mathcal{B}_{k,0}^{(i)})\|^2] \\ \leq 2\boldsymbol{N}_{k,i}^2+2\eta_k^2t(\sum_{t'=0}^{t-1}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t'}^{(i)}\|^2])+\sigma^2)+3\eta_k^2L^2\boldsymbol{N}_{k,i}^2 \quad (89)$$

Now putting eq. (89) in eq. (87), we get:

$$\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2] \leq 2\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2] + 2L^2(2+3\eta_k^2L^2)\boldsymbol{N}_{k,i}^2 + 4\eta_k^2L^2t(\sum_{t'=0}^{t-1}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t'}^{(i)})\|^2] + \sigma^2) \tag{90}$$

Now summing up eq. (90) for all $t \in \{0, \ldots, \tau-1\}$, we get:

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

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

$$\sum_{t=0}^{\tau-1}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2] \leq 2\tau\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2] + 2L^2\tau(2+3\eta_k^2L^2)\boldsymbol{N}_{k,i}^2 + \frac{1}{2}\sum_{t=0}^{\tau-1}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2] + \frac{1}{2}\sigma^2 \tag{91}$$

Simplifying, we get:

$$\sum_{t=0}^{\tau-1}\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_{k,t}^{(i)})\|^2] \leq 4\tau\mathbb{E}[\|\nabla f_i(\boldsymbol{w}_k)\|^2] + 4L^2\tau(2+3\eta_k^2L^2)\boldsymbol{N}_{k,i}^2 + \sigma^2 \tag{92}$$

■

REFERENCES

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

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

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

[4] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 601–615, 2014.

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

[6] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

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

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

[9] R. Das, A. Hashemi, S. Sanghavi, and I. S. Dhillon, "Privacy-preserving federated learning via normalized (instead of clipped) updates," *arXiv preprint arXiv:2106.07094*, 2021.

[10] H. Tang, S. Gan, C. Zhang, T. Zhang, and J. Liu, "Communication compression for decentralized training," in *Advances in Neural Information Processing Systems*, pp. 7652–7662, 2018.

[11] 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 *Advances in neural information processing systems*, pp. 1509–1519, 2017.

[12] 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 *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 4035–4043, JMLR. org, 2017.

[13] 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 *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4685–4689, IEEE, 2021.

[14] 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 *Uncertainty in Artificial Intelligence*, pp. 496–506, PMLR, 2022.

[15] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

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

[17] Y. Du, S. Yang, and K. Huang, "High-dimensional stochastic gradient quantization for communication-efficient edge learning," *IEEE transactions on signal processing*, vol. 68, pp. 2128–2142, 2020.

[18] S. Zheng, C. Shen, and X. Chen, "Design and analysis of uplink and downlink communications for federated learning," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 2150–2167, 2020.

[19] 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 Transactions on Parallel and Distributed Systems*, vol. 33, no. 11, pp. 2727–2739, 2021.

[20] Y. Chen, A. Hashemi, and H. Vikalo, "Communication-efficient variance-reduced decentralized stochastic optimization over time-varying directed graphs," *IEEE Transactions on Automatic Control*, 2021.

[21] Y. Chen, A. Hashemi, and H. Vikalo, "Decentralized optimization on time-varying directed graphs under communication constraints," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3670–3674, IEEE, 2021.

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

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

[24] S. Xia, J. Zhu, Y. Yang, Y. Zhou, Y. Shi, and W. Chen, "Fast convergence algorithm for analog federated learning," in *ICC 2021-IEEE International Conference on Communications*, pp. 1–6, IEEE, 2021.
[25] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Over-the-air federated learning from heterogeneous data," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3796–3811, 2021.
[26] H. Guo, A. Liu, and V. K. Lau, "Analog gradient aggregation for federated learning over wireless networks: Customized design and convergence analysis," *IEEE Internet of Things Journal*, vol. 8, no. 1, pp. 197–210, 2020.
[27] X. Wei and C. Shen, "Federated learning over noisy channels: Convergence analysis and design examples," *IEEE Transactions on Cognitive Communications and Networking*, 2022.
[28] F. Ang, L. Chen, N. Zhao, Y. Chen, W. Wang, and F. R. Yu, "Robust federated learning with noisy communication," *IEEE Transactions on Communications*, vol. 68, no. 6, pp. 3452–3464, 2020.
[29] H. Tang, C. Yu, X. Lian, T. Zhang, and J. Liu, "Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression," in *International Conference on Machine Learning*, pp. 6155–6165, PMLR, 2019.
[30] Y. Yu, J. Wu, and L. Huang, "Double quantization for communication-efficient distributed optimization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[31] C.-Y. Chen, J. Ni, S. Lu, X. Cui, P.-Y. Chen, X. Sun, N. Wang, S. Venkataramani, V. V. Srinivasan, W. Zhang, *et al.*, "Scalecom: Scalable sparsified gradient compression for communication-efficient distributed training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13551–13563, 2020.
[32] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*, pp. 5132–5143, PMLR, 2020.
[33] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.