Convergence of Federated Learning over a Noisy Downlink

Mohammad Mohammadi Amiri, Deniz Gündüz, Sanjeev R. Kulkarni, H. Vincent Poor

Abstract

We study federated learning (FL), where power-limited wireless devices utilize their local datasets to collaboratively train a global model with the help of a remote parameter server (PS). The PS has access to the global model and shares it with the devices for local training using their datasets, and the devices return the result of their local updates to the PS to update the global model. The algorithm continues until the convergence of the global model. This framework requires downlink transmission from the PS to the devices and uplink transmission from the devices to the PS. The goal of this study is to investigate the impact of the bandwidth-limited shared wireless medium in both the downlink and uplink on the performance of FL with a focus on the downlink. To this end, the downlink and uplink channels are modeled as fading broadcast and multiple access channels, respectively, both with limited bandwidth. For downlink transmission, we first introduce a digital approach, where a quantization technique is employed at the PS followed by a capacity achieving channel code to transmit the global model update over the wireless broadcast channel at a common rate such that all the devices can decode it. Next, we propose analog downlink transmission, where the global model is broadcast by the PS in an uncoded manner. We consider analog transmission over the uplink in both cases, since its superiority over digital transmission for uplink has been well studied in the literature. We further analyze the convergence behavior of the proposed analog transmission approach over the downlink assuming that the uplink transmission is error-free. Numerical experiments show that the analog downlink approach provides significant improvement over the digital one, despite a significantly lower transmit power at the PS, with a more notable improvement when the data

M. Mohammadi Amiri, S. R. Kulkarni, and H. V. Poor are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA (e-mail: {mamiri, kulkarni, poor}@princeton.edu).

D. Gündüz is with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: d.gunduz@imperial.ac.uk).

This work was supported in part by the U.S. National Science Foundation under Grant CCF-0939370, and by the European Research Council (ERC) Starting Grant BEACON (grant agreement no. 677854).

distribution across the devices is not independent and identically distributed. The experimental results corroborate the convergence results, and show that a smaller number of local iterations should be used when the data distribution is more biased, and also when the devices have a better estimate of the global model in the analog downlink approach.

I. Introduction

Wireless devices, such as mobile phones, wearables, and Internet-of-things (IoT) devices, continuously generate massive amounts of data. This massive data can be processed to infer the state of a system, or to anticipate its future states with applications in autonomous driving, unmanned aerial vehicles (UAVs), or extended reality (XR) technologies. Due to the growing storage and computational capabilities of wireless edge devices, it is increasingly attractive to store and process the data locally by shifting network computations to the edge. Also, in contrast to traditional machine learning (ML) solutions, it is not desirable to offload such massive amounts of data available at the wireless edge devices to a cloud server for centralized processing due to latency, bandwidth, and power constraints in wireless networks, as well as privacy concerns of users. *Federated learning* (FL) has emerged as an alternative method enabling ML at the wireless network edge by utilizing wireless edge computational capabilities to process data locally.

In FL the goal is to fit a global model to data generated and stored locally at the wireless devices by exploiting edge processing capabilities collaboratively with the help of a remote parameter server (PS) [1]. The PS keeps track of the global model, which is updated using the local model updates received from the participating devices, and shares it with the devices for training using their local data. When FL is employed at the wireless edge, the PS can be a wireless access point or a base station, and the communication between the PS and the devices takes place over the shared wireless medium with limited energy and bandwidth. There have been several studies to develop distributed ML techniques with communication constraints [1]–[11]. However, these studies focus on limiting the uplink communication from the devices to the PS by assuming rate-limited error-free links, and do not take into consideration the physical layer characteristics of the wireless medium.

Recently there have been efforts to develop a federated edge learning (FEEL) framework considering the physical layer aspects of the underlying wireless medium. FL over power- and bandwidth-limited multiple access channel (MAC) for the uplink is studied in [12], and

novel digital and analog transmission techniques at the wireless devices are proposed. While the former employs gradient sparsification followed by quantization and channel coding for digital transmission, the latter utilizes the superposition property of the underlying wireless MAC, and introduces a novel bandwidth-efficient transmission technique employing sparsification and linear projection. FL over a broadband wireless fading MAC is studied in [13], where the devices have channel state information (CSI) to perform channel inversion, while [14] proposes analog transmission over the wireless fading MAC without any power control. The extension of the approach introduced in [12] to the wireless fading MAC studied in [15], [16], which combines the linear projection idea of [12] with power control. Furthermore, FL over wireless networks with a multi-antenna PS is studied in [17]–[20], where beamforming techniques are used for efficient gradient aggregation at the PS. In [21] digital transmission over a Gaussian MAC from the devices to the PS is considered with quantization based on the channel qualities, and [22] studies digital transmission using the over-the-air aggregation property of the wireless MAC. Various device scheduling policies are studied for FEEL aiming to select a subset of the devices sharing the limited wireless resources efficiently, including frequency of participation in the training [23], minimizing the training delay [24], link qualities of the devices [25], energy consumption [26], and importance of the model update along with the channel quality [27]. Resource allocation for FEEL is formulated as an optimization problem to speed up training [28], to minimize the empirical loss function [29], and to minimize the total energy consumption [30]. Also, convergence of FEEL with limited bandwidth from the devices to the PS is analyzed in [31].

All the aforementioned works assume an error-free PS-to-devices shared link, and availability of an accurate global model at the devices for local training. In this paper, we consider a bandwidth-limited wireless fading broadcast channel from the PS to the devices with limited transmit power at the PS. We introduce *digital* and *analog* transmission approaches over the downlink. In the digital downlink, the PS employs quantization followed by channel coding to broadcast the quantized global model update over the wireless fading broadcast channel, at a rate targeting the device with the worst channel, so that all the devices can successfully receive the global model. On the other hand, with the analog downlink approach, the PS broadcasts the global model vector in an analog/uncoded manner over the wireless fading broadcast channel, and the devices receive different noisy versions of it. We model the uplink from the devices to the PS, over which the devices send their model updates, as

a bandwidth-limited fading MAC. We follow the existing works highlighting the efficiency of the analog transmission over the uplink fading MAC for FEEL [12], [13], [16], and consider analog communications. The convergence analysis of the proposed digital downlink approach is provided in [32]. Here, we provide the convergence analysis of the analog downlink approach, where for ease of analysis we assume error-free uplink transmission and focus on the impact of a noisy downlink transmission on the convergence behavior. Our theoretical analysis is complemented with numerical experiments on the MNIST dataset, which clearly illustrate the significant advantages of the analog downlink approach compared to its digital counterpart. We observe that the improvement is more significant when the data is not independent and identically distributed (iid) across the devices. The performance of both approaches improve with the number of devices thanks to the additional power introduced by each device. Our numerical results corroborate the analytical convergence analysis, showing that reducing the number of local iterations provides the best performance when introducing bias in the data distribution across the devices. Also, both analytical and experimental results show that, for non-iid data distribution, the number of local iterations at the devices should reduce when the transmit power at the PS increases.

Imperfect downlink transmission in FL is also treated in [33] and [34]. In [33], the shared link from the PS to the devices is assumed to be rate-limited without taking into account the physical layer characteristics of the wireless medium; the PS sends a compressed version of the current global model to the devices through quantization. The efficiency of quantizing the global model diminishes significantly since the peak-to-average ratio of the parameters is high. Therefore, [33] proposes employing a linear projection at the PS to first spread the information of the global model vector more evenly across its dimensions, and the devices perform the inverse of the linear projection to estimate the global model vector. Instead, in our proposed digital downlink approach, the PS broadcasts the quantized global model update, with respect to the global model estimate at the devices, and the devices recover an estimate of the current global model using their knowledge of the last global model. We highlight that the global model update has significantly less variability/variance than the global model itself. Hence, compared to the proposed digital downlink approach, the approach in [33] requires significantly higher computation overhead at the PS and the devices due to the linear projection and its inverse, respectively, and this overhead grows with the number of model parameters. Moreover, the results in both [33] and [34] are limited to

simulations, where [34] illustrates the advantages of analog transmission in the downlink but does not provide a convergence result. In this paper, we provide an in-depth analysis of the impact of a noisy downlink on the performance of FEEL through extensive experimental results together with theoretical convergence analysis.

The rest of this paper is organized as follows. In Section II, we present the system model. The digital and analog downlink approaches are introduced in Section III and Section IV, respectively. In Section V, we provide the convergence results of the analog downlink approach. Numerical results are presented in Section VI. Finally, we conclude the paper in Section VII, and provide a detailed proof of the main theorem in the Appendices.

Notation: We denote the set of real, natural and complex numbers by $\mathbb{R}$, $\mathbb{N}$ and $\mathbb{C}$, respectively. For $i \in \mathbb{N}$, we let $[i] \triangleq \{1, \ldots, i\}$. We denote a circularly symmetric complex Gaussian distribution with real and imaginary components with variance $\sigma/2$ by $\mathcal{CN}(0, \sigma)$. For vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ with the same dimension, $\boldsymbol{x} \circ \boldsymbol{y}$ returns their Hadamard/entry-wise product. Also, $\mathrm{Re}\{\boldsymbol{x}\}$ and $\mathrm{Im}\{\boldsymbol{x}\}$ return entry-wise real and imaginary components of $\boldsymbol{x}$, respectively, and $(\boldsymbol{x})^{-1}$ represents entry-wise inverse of vector $\boldsymbol{x}$. The notation $|\cdot|$ represents the cardinality of a set, the l_2-norm of vector $\boldsymbol{x}$ is denoted by $\|\boldsymbol{x}\|_2$, and $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ denotes the inner product of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. The imaginary unit is represented by j.

II. System Model

We consider FEEL where M wireless devices collaboratively train a model parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$ with the help of a remote parameter server (PS). Device m has access to B_m local data samples, the set of which is denoted by $\mathcal{B}_m$, i.e., $B_m = |\mathcal{B}_m|$, $m \in [M]$, and we define $B \triangleq \sum_{m=1}^{M} B_m$. The goal is to minimize loss function

$$F(\boldsymbol{\theta}) = \sum\nolimits_{m=1}^{M} \frac{B_m}{B} F_m(\boldsymbol{\theta}), \tag{1}$$

where $F_m(\boldsymbol{\theta})$ denotes the loss function at device m,

$$F_m(\boldsymbol{\theta}) = \frac{1}{B_m} \sum\nolimits_{\boldsymbol{u} \in \mathcal{B}_m} f(\boldsymbol{\theta}, \boldsymbol{u}), \quad m \in [M], \tag{2}$$

where $f(\cdot, \cdot)$ is an empirical loss function defined by the learning task. Device m performs multiple iterations of stochastic gradient descent (SGD) algorithm based on its local dataset and the global model parameter vector shared by the PS to minimize $F_m(\boldsymbol{\theta})$, $m \in [M]$.

FEEL involves iterative communications between the wireless devices and the PS until the model parameter vector converges to its optimum, minimizing loss function $F(\boldsymbol{\theta})$. It consists of *downlink* and *uplink* wireless transmissions, where in the downlink the PS shares the global model parameter vector with the devices for local training, and in the uplink the devices transmit their local model updates to the PS, which updates the global model parameter vector accordingly.

During the t-th global iteration, the PS broadcasts the global model parameter vector, denoted by $\boldsymbol{\theta}(t)$, to the devices over the downlink channel. We model the downlink wireless channel as a fading broadcast channel, where OFDM with n^{dl} subchannels is employed for transmission. We denote the length-n^{dl} channel input by the PS at the global iteration t by $\boldsymbol{x}^{\mathrm{dl}}(t) \in \mathbb{C}^{n^{\mathrm{dl}}}$, and consider a transmit power constraint P^{dl} at the PS at any global iteration. The received signal at device m is given by

$$\boldsymbol{y}_m^{\mathrm{dl}}(t) = \boldsymbol{h}_m^{\mathrm{dl}}(t) \circ \boldsymbol{x}^{\mathrm{dl}}(t) + \boldsymbol{z}_m^{\mathrm{dl}}(t), \quad \text{for } m \in [M], \tag{3}$$

where $\boldsymbol{h}_m^{\mathrm{dl}}(t) \in \mathbb{C}^{n^{\mathrm{dl}}}$ is the downlink channel gain vector from the PS to device m with each entry iid according to $\mathcal{CN}(0, \sigma^{\mathrm{dl}})$, and $\boldsymbol{z}_m^{\mathrm{dl}}(t) \in \mathbb{C}^{n^{\mathrm{dl}}}$ is the downlink additive noise vector at device m with each entry iid according to $\mathcal{CN}(0,1)$. We assume that device m has channel state information (CSI) about the downlink channel, and denote the noisy estimate of the global model parameter vector $\boldsymbol{\theta}(t)$ at device m by $\widehat{\boldsymbol{\theta}}_m(t)$, $m \in [M]$.

Having estimated $\widehat{\boldsymbol{\theta}}_m(t)$, device m, $m \in [M]$, updates the model by running SGD τ steps locally, for some $\tau \in \mathbb{N}$. The i-th SGD step at device m during global iteration t is given by

$$\boldsymbol{\theta}_m^{i+1}(t) = \boldsymbol{\theta}_m^i(t) - \eta_m^i(t) \nabla F_m \left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t) \right), \quad i \in [\tau], \tag{4}$$

where $\boldsymbol{\theta}_m^1(t) = \widehat{\boldsymbol{\theta}}_m(t)$, $\eta_m^i(t)$ represents the learning rate, and $\nabla F_m \left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t) \right)$ denotes the stochastic gradient estimate with respect to $\boldsymbol{\theta}_m^i(t)$ and the local mini-batch sample $\xi_m^i(t)$, chosen uniformly at random from the local dataset $\mathcal{B}_m$, for $m \in [M]$. We highlight that

$$\mathbb{E}_\xi \left[\nabla F_m \left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t) \right) \right] = \nabla F_m \left(\boldsymbol{\theta}_m^i(t) \right), \quad \forall i \in [\tau], \forall m \in [M], \forall t, \tag{5}$$

where $\mathbb{E}_\xi$ denotes expectation with respect to the randomness of the stochastic gradient function. After performing the local SGD algorithm, device m aims to transmit the local model update $\Delta\boldsymbol{\theta}_m(t) = \boldsymbol{\theta}_m^{\tau+1}(t) - \boldsymbol{\theta}_m^1(t)$ to the PS over the uplink channel, $m \in [M]$.

We model the uplink channel as a fading MAC, where, similarly to the downlink, OFDM is employed for transmission. We assume $n^{\rm up}$ subchannels are available to each device in the uplink with transmit power constraint $P^{\rm up}$ during each global iteration. The length-$n^{\rm up}$ channel input by device m at the global iteration t is denoted by $\boldsymbol{x}_m^{\rm up}(t) \in \mathbb{C}^{n^{\rm up}}$, for $m \in [M]$. The channel output received at the PS during the global iteration t is given by

$$\boldsymbol{y}^{\rm up}(t) = \sum\nolimits_{m=1}^{M} \boldsymbol{h}_m^{\rm up}(t) \circ \boldsymbol{x}_m^{\rm up}(t) + \boldsymbol{z}^{\rm up}(t), \tag{6}$$

where $\boldsymbol{h}_m^{\rm up}(t) \in \mathbb{C}^{n^{\rm up}}$ is the uplink channel gain vector from device m to the PS with each entry iid according to $\mathcal{CN}(0, \sigma^{\rm up})$, and $\boldsymbol{z}_m^{\rm up}(t) \in \mathbb{C}^{n^{\rm up}}$ is the uplink additive noise vector at the PS with each entry iid according to $\mathcal{CN}(0, 1)$. We assume that the PS knows all the channel gains, while each device knows the states of its own subchannels. The PS's goal is to recover the average of the local model updates, $\frac{1}{M}\sum_{m=1}^{M} \Delta\boldsymbol{\theta}_m(t)$, whose estimate at the PS is denoted by $\Delta\widehat{\boldsymbol{\theta}}(t)$, which is then used to obtain the updated global model parameter vector, $\boldsymbol{\theta}(t+1)$.

In this paper, we study the impact of noisy downlink transmission on the performance of FEEL. For this purpose, we consider digital and analog transmission approaches over the downlink channel. When performing digital transmission, we assume that the PS has CSI about the downlink wireless channels, while for the analog transmission, no CSI about the downlink channels at the PS is needed. On the other hand, following the results in [12], [13], [16], which have shown the superiority of analog transmission for the uplink transmission over a wireless MAC, here we only consider analog transmission over the uplink.

III. Digital Downlink Approach

In this section, we present a digital approach for the downlink transmission of the global model update to the devices.

A. Downlink Channel Capacity

At the global iteration t, the PS aims to transmit vector $\boldsymbol{x}^{\rm dl}(t)$, containing information about the global model vector $\boldsymbol{\theta}(t)$, to all the devices using digital transmission with transmit power $P^{\rm dl}$ over the bandwidth-limited wireless channel. The PS broadcasts $\boldsymbol{x}^{\rm dl}(t)$ at a "common rate" such that all the devices can decode it. The downlink is a parallel fading broadcast channel with $n^{\rm dl}$ subchannels, where CSI is known at both the transmitter and

the receivers. In the following, we provide an upper bound on the maximum common rate of broadcasting over this $n^{\rm dl}$ parallel fading channels. Given an average transmission power $P^{\rm dl}$ at global iteration t, the maximum common rate of downlink transmission over $n^{\rm dl}$ parallel Gaussian channels, denoted by $C^{\rm dl}(t)$, is the solution of the following optimization problem [35], [36]:

$$\begin{aligned} &\max_{P_1,\ldots,P_{n^{\rm dl}}} \min_{m\in[M]} \sum_{i=1}^{n^{\rm dl}} \log_2\left(1+P_{m,i}^{\rm dl}(t)\left|h_{m,i}^{\rm dl}(t)\right|^2\right), \\ &\text{subject to } \sum_{i=1}^{n^{\rm dl}} P_{m,i}^{\rm dl}(t) = P^{\rm dl},\ \forall m\in[M]. \end{aligned} \tag{7}$$

The above problem is a convex optimization problem which can be efficiently solved by the minimax hypothesis testing approach [35]–[37]. Note that this rate would be achievable by coding across infinitely many realizations of the $n^{\rm dl}$ parallel Gaussian channels under consideration, and will serve as an upper bound on the rate transmitted over a single realization.

B. Compression Technique

In the following, we present the compression technique employed by the PS for transmitting information about the global model over the bandwidth-limited downlink channel, where we adopt the scheme introduced in [38] with a slight modification. Assume that vector $\boldsymbol{x}(t)\in\mathbb{R}^d$, whose i-th entry is denoted by $x_i(t)$, $i\in[d]$, is to be quantized and transmitted over the downlink channel by the PS. The PS first sparsifies $\boldsymbol{x}(t)$ by setting all but s entries of $\boldsymbol{x}(t)$ with the highest magnitudes to zero, for some integer $s\leq d$. We denote the set of s indices of the resultant sparse vector with non-zero entries by $\mathcal{S}(t)$. We also denote the resultant vector with dimension s after removing the zeroed entries due to the sparsification by $\boldsymbol{x}_s(t)$, whose i-th entry is denoted by $x_{s,i}(t)$, for $i\in[s]$. Then the PS quantizes the entries of $\boldsymbol{x}_s(t)$, and transmits the quantized values along with their locations in $\boldsymbol{x}(t)$, which are available in set $\mathcal{S}(t)$. We define

$$x_{\max} \triangleq \max_{i\in[s]} \left\{|x_{s,i}(t)|\right\}, \tag{8a}$$

$$x_{\min} \triangleq \min_{i\in[s]} \left\{|x_{s,i}(t)|\right\}. \tag{8b}$$

Given a quantization level $q(t)$, which will be determined later, we define the compression technique applied to the i-th entry of $\boldsymbol{x}_s(t)$, for $i \in [s]$, as

$$Q\left(x_{s,i}(t)\right) \triangleq \operatorname{sign}\left(x_{s,i}(t)\right) \cdot \left(x_{\min} + \left(x_{\max} - x_{\min}\right) \cdot \varphi\Big(\frac{|x_{s,i}(t)| - x_{\min}}{x_{\max} - x_{\min}}, q(t)\Big)\right), \tag{9a}$$

where, for $x \in \mathbb{R}$,

$$\operatorname{sign}(x) \triangleq \begin{cases} 1, & \text{if } x \geq 0, \\ -1, & \text{otherwise}, \end{cases} \tag{9b}$$

and $\varphi(\cdot,\cdot)$ is a quantization function defined in the following. For $0 \leq x \leq 1$ and some integer $q \geq 1$, let $l \in \{0, 1, \ldots, q-1\}$ be an integer such that $x \in [l/q, (l+1)/q)$. We then define

$$\varphi(x, q) \triangleq \begin{cases} l/q, & \text{with probability } 1 - (xq - l), \\ (l+1)/q, & \text{with probability } xq - l. \end{cases} \tag{9c}$$

We denote the compressed version of $x_i(t)$ by $S\left(x_i(t)\right)$, for $i \in [d]$, which is given by

$$S\left(x_i(t)\right) = \begin{cases} Q\left(x_i(t)\right), & \text{if } i \in \mathcal{S}(t), \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

and represent $\boldsymbol{S}\left(\boldsymbol{x}(t)\right) = \Big[S\left(x_1(t)\right), \ldots, S\Big(x_d(t)\Big)\Big]^T$. Note that we normalize the entries of $\boldsymbol{x}_s(t)$ with $x_{\max} - x_{\min}$ rather than $\|\boldsymbol{x}_s(t)\|_2$ as introduced in [38].

With the above compression technique, the PS needs to transmit

$$R^{\mathrm{dl}}(t) = 64 + s\left(1 + \log_2(q(t)+1)\right) + \log_2 \binom{d}{s} \text{ bits} \tag{11}$$

over the wireless broadcast channel to each of the devices, where 64 bits are used to represent the real numbers $x_{\max}$ and $x_{\min}$, s bits for presenting $\operatorname{sign}\left(x_{s,i}(t)\right)$, $\forall i \in [s]$, $s\log_2(q(t)+1)$ bits are used for $\varphi\left(\left(|x_{s,i}(t)| - x_{\min}\right)/(x_{\max} - x_{\min}), q\right)$, $\forall i \in [s]$, and $\log_2 \binom{d}{s}$ bits represent the indices of $\boldsymbol{x}(t)$ in set $\mathcal{S}(t)$. We set $q(t)$ to the largest integer satisfying $R^{\mathrm{dl}}(t) \leq C^{\mathrm{dl}}(t)$.

C. Model Update

Here we present the model update scheme including the global model update broadcasting from the PS to the devices and aggregation of the local updates via uplink transmission from

the devices to the PS.

Downlink transmission. We first elaborate on the downlink transmission. We highlight that, for the digital downlink approach, all the devices have the same estimate of $\boldsymbol{\theta}(t)$ during global iteration t, denoted by $\widehat{\boldsymbol{\theta}}(t)$, i.e., $\widehat{\boldsymbol{\theta}}_m(t) = \widehat{\boldsymbol{\theta}}(t)$, $\forall m \in [M]$. In the downlink, at the global iteration t, the PS wants to broadcast the global model update $\boldsymbol{\theta}(t) - \widehat{\boldsymbol{\theta}}(t-1)$ to all the devices. We define

$$\Delta\widehat{\boldsymbol{\theta}}(t-1) \triangleq \boldsymbol{\theta}(t) - \widehat{\boldsymbol{\theta}}(t-1) \in \mathbb{R}^d. \tag{12}$$

The PS first quantizes $\Delta\widehat{\boldsymbol{\theta}}(t-1)$ using the compression technique described in Section III-B, obtaining $\boldsymbol{S}\left(\Delta\widehat{\boldsymbol{\theta}}(t-1)\right) = \boldsymbol{S}\left(\boldsymbol{\theta}(t) - \widehat{\boldsymbol{\theta}}(t-1)\right)$, which results in $R^{\mathrm{dl}}(t)$ bits as given in (11). The PS then broadcasts these bits to all the devices using a capacity achieving channel code, where $q(t)$ is set to the largest integer satisfying $R^{\mathrm{dl}}(t) \leq C^{\mathrm{dl}}(t)$, where $C^{\mathrm{dl}}(t)$ given as the solution of (7). After decoding $\boldsymbol{S}\Big(\boldsymbol{\theta}(t) - \widehat{\boldsymbol{\theta}}(t-1)\Big)$, each device computes $\widehat{\boldsymbol{\theta}}(t)$ as

$$\widehat{\boldsymbol{\theta}}(t) = \widehat{\boldsymbol{\theta}}(t-1) + \boldsymbol{S}\Big(\boldsymbol{\theta}(t) - \widehat{\boldsymbol{\theta}}(t-1)\Big), \tag{13}$$

which is equivalent to

$$\widehat{\boldsymbol{\theta}}(t) = \boldsymbol{\theta}(0) + \textstyle\sum_{i=1}^{t} \boldsymbol{S}\Big(\boldsymbol{\theta}(i) - \widehat{\boldsymbol{\theta}}(i-1)\Big), \tag{14}$$

where we have assumed that $\widehat{\boldsymbol{\theta}}(0) = \boldsymbol{\theta}(0)$. Having knowledge about the compressed vector $\boldsymbol{S}\Big(\boldsymbol{\theta}(i) - \widehat{\boldsymbol{\theta}}(i-1)\Big)$, $\forall i \in [t]$, the PS can also recover $\widehat{\boldsymbol{\theta}}(t)$, which is used at the devices to compute the local updates.

Uplink transmission. For ease of presentation, we assume that $n^{\mathrm{up}} = d/2$, and we will discuss the generalization of the prorposed approach. Device m, $m \in [M]$, performs τ local SGD steps, where the i-th step is given by

$$\boldsymbol{\theta}_m^{i+1}(t) = \boldsymbol{\theta}_m^{i}(t) - \eta_m^i(t) \nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right), \quad i \in [\tau], \tag{15}$$

where $\boldsymbol{\theta}_m^1(t) = \widehat{\boldsymbol{\theta}}(t)$. It then transmits the local model update $\Delta\boldsymbol{\theta}_m(t) = \boldsymbol{\theta}_m^{\tau+1}(t) - \widehat{\boldsymbol{\theta}}(t)$ in an analog (uncoded) fashion. We define

$$\Delta\boldsymbol{\theta}_{m,\mathrm{re}}(t) \triangleq [\Delta\theta_{m,1}(t), \ldots, \Delta\theta_{m,d/2}(t)]^T, \tag{16a}$$

$$\Delta\boldsymbol{\theta}_{m,\mathrm{im}}(t) \triangleq [\Delta\theta_{m,d/2+1}(t), \ldots, \Delta\theta_{m,d}(t)]^T, \tag{16b}$$

where $\Delta\theta_{m,i}(t)$ denotes the i-th entry of $\Delta\boldsymbol{\theta}_m(t)$, for $i \in [d]$, $m \in [M]$, and we have $\Delta\boldsymbol{\theta}_m(t) = \left[\Delta\boldsymbol{\theta}_{m,\mathrm{re}}(t)^T, \Delta\boldsymbol{\theta}_{m,\mathrm{im}}(t)^T\right]^T$. Device m, $m \in [M]$, transmits

$$\boldsymbol{x}_m^{\mathrm{ul}}(t) = \boldsymbol{\alpha}_m^{\mathrm{ul}}(t) \circ \left(\Delta\boldsymbol{\theta}_{m,\mathrm{re}}(t) + j\Delta\boldsymbol{\theta}_{m,\mathrm{im}}(t)\right), \tag{17}$$

where $\boldsymbol{\alpha}_m^{\mathrm{ul}}(t) \in \mathbb{C}^{d/2}$ is the power allocation vector, whose i-th entry, $i \in [d/2]$, is set as

$$\alpha_{m,i}^{\mathrm{ul}}(t) = \begin{cases} \frac{\gamma_m(t)}{h_{m,i}^{\mathrm{ul}}(t)}, & \text{if } |h_{m,i}^{\mathrm{ul}}(t)| \geq \lambda_{\mathrm{thr}}(t), \\ 0, & \text{otherwise}, \end{cases} \tag{18}$$

for some $\gamma_m(t), \lambda_{\mathrm{thr}}(t) \in \mathbb{R}$, which are set to satisfy the transmit power constraint $\|\boldsymbol{x}_m^{\mathrm{ul}}(t)\|_2^2 \leq P^{\mathrm{ul}}$. We assume that device m first transmits the scaling factor $\gamma_m(t)$ to the PS in an error-free fashion, $m \in [M]$. The PS receives the following signal:

$$\boldsymbol{y}^{\mathrm{ul}}(t) = \sum\nolimits_{m=1}^{M} \boldsymbol{\alpha}_m^{\mathrm{ul}}(t) \circ \left(\Delta\boldsymbol{\theta}_{m,\mathrm{re}}(t) + j\Delta\boldsymbol{\theta}_{m,\mathrm{im}}(t)\right) \circ \boldsymbol{h}_m^{\mathrm{ul}}(t) + \boldsymbol{z}^{\mathrm{ul}}(t), \tag{19}$$

whose i-th entry, $i \in [d/2]$, is given by

$$y_i^{\mathrm{ul}}(t) = \sum\nolimits_{m \in \mathcal{M}_i(t)} \gamma_m(t) \left(\Delta\theta_{m,i}(t) + j\Delta\theta_{m,d/2+i}(t)\right) + z_i^{\mathrm{ul}}(t), \tag{20}$$

where we have defined

$$\mathcal{M}_i(t) \triangleq \left\{m \in [M] : \left|h_{m,i}^{\mathrm{ul}}(t)\right| \geq \lambda_{\mathrm{thr}}(t)\right\}. \tag{21}$$

With the knowledge of the channel state, and consequently $\mathcal{M}_i(t)$, $\forall i \in [d/2]$, the PS estimates $\frac{1}{|\mathcal{M}_i(t)|}\sum_{m \in \mathcal{M}_i(t)} \Delta\theta_{m,i}(t)$ and $\frac{1}{|\mathcal{M}_i(t)|}\sum_{m \in \mathcal{M}_i(t)} \Delta\theta_{m,d/2+i}(t)$ with

$$\Delta\hat{\theta}_i(t) = \begin{cases} \frac{\mathrm{Re}\left\{y_i^{\mathrm{ul}}(t)\right\}}{\bar{\gamma}(t)|\mathcal{M}_i(t)|}, & \text{if } |\mathcal{M}_i(t)| \neq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{22a}$$

$$\Delta\hat{\theta}_{d/2+i}(t) = \begin{cases} \frac{\mathrm{Im}\left\{y_i^{\mathrm{ul}}(t)\right\}}{\bar{\gamma}(t)|\mathcal{M}_i(t)|}, & \text{if } |\mathcal{M}_i(t)| \neq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{22b}$$

respectively, where we have defined $\bar{\gamma}(t) \triangleq \frac{1}{M}\sum_{m=1}^{M} \gamma_m(t)$. The estimated vector $\Delta\widehat{\boldsymbol{\theta}}(t) \triangleq [\Delta\hat{\theta}_1(t), \ldots, \Delta\hat{\theta}_d(t)]^T$ is used to update the global model parameter vector as

$$\boldsymbol{\theta}(t+1) = \widehat{\boldsymbol{\theta}}(t) + \Delta\widehat{\boldsymbol{\theta}}(t). \tag{23}$$

Algorithm 1 Digital Downlink Approach

Initialize $\boldsymbol{\theta}(0)$
for $t = 0, \ldots, T-1$ **do**
- **Downlink transmission:**

PS broadcasts $\boldsymbol{S}\Big(\boldsymbol{\theta}(t) - \widehat{\boldsymbol{\theta}}(t-1)\Big)$
$\widehat{\boldsymbol{\theta}}(t) = \widehat{\boldsymbol{\theta}}(t-1) + \boldsymbol{S}\Big(\boldsymbol{\theta}(t) - \widehat{\boldsymbol{\theta}}(t-1)\Big)$
- **Uplink transmission:**

for $m = 1, \ldots, M$ in parallel **do**
$\boldsymbol{x}_m^{\mathrm{ul}}(t) = \boldsymbol{\alpha}_m^{\mathrm{ul}}(t) \circ (\Delta\boldsymbol{\theta}_{m,\mathrm{re}}(t) + j\Delta\boldsymbol{\theta}_{m,\mathrm{im}}(t))$
$\alpha_{m,i}^{\mathrm{ul}}(t) = \begin{cases} \frac{\gamma_m(t)}{h_{m,i}^{\mathrm{ul}}(t)}, & \text{if } \left|h_{m,i}^{\mathrm{ul}}(t)\right| \geq \lambda_{\mathrm{thr}}(t), \\ 0, & \text{otherwise} \end{cases}, \quad \text{for } i \in [d/2]$
end for
$\boldsymbol{\theta}(t+1) = \widehat{\boldsymbol{\theta}}(t) + \Delta\widehat{\boldsymbol{\theta}}(t)$
end for

We remark here that for $n^{\mathrm{up}} < d/2$, we carry out the uplink transmission in $\lceil d/(2n^{\mathrm{up}}) \rceil$ time slots, where in each time slot we perform the above transmission.

Algorithm 1 summarizes the downlink and uplink transmissions for the digital downlink approach employing the compression technique presented in Section III-B.

IV. Analog Downlink Approach

In this section, we propose that the PS broadcasts the global model parameter vector $\boldsymbol{\theta}(t)$ in an analog (uncoded) manner. For ease of presentation, we consider $n^{\mathrm{dl}} = d/2$, and we will argue that the proposed approach can be readily extended to the general case.

Downlink transmission. We define

$$\boldsymbol{\theta}_{\mathrm{re}}(t) \triangleq [\theta_1(t), \ldots, \theta_{d/2}(t)]^T, \tag{24a}$$

$$\boldsymbol{\theta}_{\mathrm{im}}(t) \triangleq [\theta_{d/2+1}(t), \ldots, \theta_d(t)]^T, \tag{24b}$$

where $\boldsymbol{\theta}(t) = \left[\boldsymbol{\theta}_{\mathrm{re}}(t)^T, \boldsymbol{\theta}_{\mathrm{im}}(t)^T\right]^T$. At the global iteration t, the PS broadcasts $\boldsymbol{x}^{\mathrm{dl}}(t) = \alpha^{\mathrm{dl}}(t)\left(\boldsymbol{\theta}_{\mathrm{re}}(t) + j\boldsymbol{\theta}_{\mathrm{im}}(t)\right)$ in an uncoded manner, where $\alpha^{\mathrm{dl}}(t)$ is set to satisfy $\|\boldsymbol{x}^{\mathrm{dl}}(t)\|_2^2 \leq P^{\mathrm{dl}}$. Before broadcasting $\boldsymbol{x}^{\mathrm{dl}}(t)$, we assume that the PS shares $\alpha^{\mathrm{dl}}(t)$ with the devices in an error-free fashion. The received signal at device m is given by

$$\boldsymbol{y}_m^{\mathrm{dl}}(t) = \alpha^{\mathrm{dl}}(t)\boldsymbol{h}_m^{\mathrm{dl}}(t) \circ (\boldsymbol{\theta}_{\mathrm{re}}(t) + j\boldsymbol{\theta}_{\mathrm{im}}(t)) + \boldsymbol{z}_m^{\mathrm{dl}}(t), \quad m \in [M]. \tag{25}$$

Device m, $m \in [M]$, performs the following descaling:

$$\widehat{\boldsymbol{y}}_m^{\rm dl}(t) \triangleq \left(\frac{1}{\alpha^{\rm dl}(t)}\right)\boldsymbol{y}_m^{\rm dl}(t) \circ \left(\boldsymbol{h}_m^{\rm dl}(t)\right)^{-1} = \boldsymbol{\theta}_{\rm re}(t) + j\boldsymbol{\theta}_{\rm im}(t) + \left(\frac{1}{\alpha^{\rm dl}(t)}\right)\boldsymbol{z}_m^{\rm dl}(t) \circ \left(\boldsymbol{h}_m^{\rm dl}(t)\right)^{-1}, \quad (26)$$

and uses $\widehat{\boldsymbol{y}}_m^{\rm dl}(t)$ to recover the global model parameter vector $\boldsymbol{\theta}(t)$ as

$$\widehat{\boldsymbol{\theta}}_m(t) \triangleq \left[\text{Re}\{\widehat{\boldsymbol{y}}_m^{\rm dl}(t)\}^T, \text{Im}\{\widehat{\boldsymbol{y}}_m^{\rm dl}(t)\}^T\right]^T. \quad (27)$$

We highlight that the proposed approach can be extended for any number of subchannels $n^{\rm dl}$ through transmission over different time slots.

Uplink transmission. After recovering $\widehat{\boldsymbol{\theta}}_m(t)$, device m, $m \in [M]$, performs τ local SGD steps as in (15), where $\boldsymbol{\theta}_m^1(t) = \widehat{\boldsymbol{\theta}}_m(t)$. It then transmits the local model update $\Delta\boldsymbol{\theta}_m(t) = \boldsymbol{\theta}_m^{\tau+1}(t) - \widehat{\boldsymbol{\theta}}_m(t)$ in an analog (uncoded) fashion over the wireless MAC, $m \in [M]$. The uplink transmission follows the same steps as the one presented in Section III-C for the digital downlink approach. However, the PS recovers $\Delta\widehat{\boldsymbol{\theta}}(t)$, given in (22), and updates the global model parameter vector as $\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) + \Delta\widehat{\boldsymbol{\theta}}(t)$.

Remark 1. *We highlight that with the independent random noise added to the model parameter vector in the downlink at different devices, the analog downlink approach inherently introduces additional data privacy for the FL framework.*

V. Convergence Analysis of Analog Downlink Approach

Here we analyze convergence behavior of the analog downlink approach presented in Section IV. For simplicity of the convergence analysis, we assume that the device-to-PS transmission is error-free, and focus on the impact of noisy downlink transmission on the convergence performance. We first present the preliminaries and assumptions, and then the convergence result for the analog downlink approach, whose proof is provided in the Appendix.

A. Preliminaries

We define the optimal solution of minimizing $F(\boldsymbol{\theta})$ as

$$\boldsymbol{\theta}^* \triangleq \arg\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}), \quad (28)$$

and the minimum loss as $F^* \triangleq F(\boldsymbol{\theta}^*)$. We also denote the minimum value of $F_m(\cdot)$, the local loss function at device m, by F_m^*, $m \in [M]$. We then define

$$\Gamma \triangleq F^* - \sum\nolimits_{m=1}^{M} \frac{B_m}{B} F_m^*, \tag{29}$$

where $\Gamma \geq 0$, and its magnitude indicates the bias in the data distribution across devices. We note that for i.i.d. data distribution, given a large enough number of local data samples, Γ approaches zero.

According to (26) and (27), we have

$$\widehat{\boldsymbol{\theta}}_m(t) = \boldsymbol{\theta}(t) + \tilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t), \quad \text{for } m \in [M], \tag{30}$$

where, for ease of presentation, we have defined

$$\tilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t) \triangleq \left(\frac{1}{\alpha^{\mathrm{dl}}(t)}\right) \left[\mathrm{Re}\Big\{\boldsymbol{z}_m^{\mathrm{dl}}(t) \circ \Big(\boldsymbol{h}_m^{\mathrm{dl}}(t)\Big)^{-1}\Big\}^T, \mathrm{Im}\Big\{\boldsymbol{z}_m^{\mathrm{dl}}(t) \circ \Big(\boldsymbol{h}_m^{\mathrm{dl}}(t)\Big)^{-1}\Big\}^T\right]^T. \tag{31}$$

For simplicity of the convergence analysis, we consider $\eta_m^i(t) = \eta(t)$, $\forall m, i$. Thus, the i-th step local SGD at device m is given by

$$\boldsymbol{\theta}_m^{i+1}(t) = \boldsymbol{\theta}_m^i(t) - \eta(t) \nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right), \quad i \in [\tau], m \in [M], \tag{32}$$

where $\boldsymbol{\theta}_m^1(t) = \widehat{\boldsymbol{\theta}}_m(t)$, given in (30). Thus, we have

$$\boldsymbol{\theta}_m^{\tau+1}(t) = \boldsymbol{\theta}_m^1(t) - \eta(t) \sum\nolimits_{i=1}^{\tau} \nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right), \quad \text{for } m \in [M]. \tag{33}$$

Device m transmits the local model update $\Delta\boldsymbol{\theta}_m(t) = -\eta(t) \sum_{i=1}^{\tau} \nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right)$, $m \in [M]$. After receiving the local model updates from all the devices, $\Delta\boldsymbol{\theta}_m(t)$, $\forall m \in [M]$, the PS updates the global model parameter vector as

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) + \sum_{m=1}^{M} \frac{B_m}{B} \Delta\boldsymbol{\theta}_m(t) = \boldsymbol{\theta}(t) - \eta(t) \sum_{m=1}^{M} \sum_{i=1}^{\tau} \frac{B_m}{B} \nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right). \tag{34}$$

Assumption 1. *The loss functions $F_1, \ldots, F_M$ are all L-smooth; that is, $\forall \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^d$,*

$$F_m(\boldsymbol{v}) - F_m(\boldsymbol{w}) \leq \langle \boldsymbol{v} - \boldsymbol{w}, \nabla F_m(\boldsymbol{w}) \rangle + \frac{L}{2} \left\| \boldsymbol{v} - \boldsymbol{w} \right\|_2^2, \quad \forall m \in [M]. \tag{35}$$

Assumption 2. *The loss functions $F_1, \ldots, F_M$ are all μ-strongly convex; that is, $\forall \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^d$,*

$$F_m(\boldsymbol{v}) - F_m(\boldsymbol{w}) \geq \langle \boldsymbol{v} - \boldsymbol{w}, \nabla F_m(\boldsymbol{w}) \rangle + \frac{\mu}{2} \|\boldsymbol{v} - \boldsymbol{w}\|_2^2, \quad \forall m \in [M]. \tag{36}$$

Assumption 3. *The expectation of the squared l_2-norm of the stochastic gradients are bounded; that is,*

$$\mathbb{E}_\xi \left[\left\| \nabla F_m \left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t) \right) \right\|_2^2 \right] \leq G^2, \quad \forall i \in [\tau], \forall m \in [M], \ \forall t. \tag{37}$$

Assumption 4. *We assume*

$$\mathbb{E}\left[\left\| \sum_{m=1}^M \frac{B_m}{B} \left(\nabla F_m(\boldsymbol{\theta}(t) + \tilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t), \xi_m^1(t)) - \nabla F_m(\boldsymbol{\theta}(t), \xi_m^1(t)) \right) \right\|^2 \right] \leq \frac{Z^2}{M \sigma^{\mathrm{dl}} P^{\mathrm{dl}}}, \tag{38}$$

for some $Z \in \mathbb{R}$, where the upper bound reduces with the variance of the downlink channel gains, the downlink transmit power, and the number of devices, M. We have assumed that the effect of the downlink noise is alleviated by averaging over the devices.

B. Convergence Rate

Here we provide the convergence rate for the analog downlink approach introduced in Section IV assuming that the devices can send their local model updates accurately.

Theorem 1. *Let $0 < \eta(t) \leq \min\left\{ \frac{\mu}{\mu+1}, \frac{1}{\mu\tau} \right\}$, $\forall t$. For the analog downlink approach, we have*

$$\mathbb{E}\left[\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*\|_2^2 \right] \leq \left(\prod_{i=0}^{t-1} A(i) \right) \|\boldsymbol{\theta}(0) - \boldsymbol{\theta}^*\|_2^2 + \sum_{j=0}^{t-1} B(j) \prod_{i=j+1}^{t-1} A(i), \tag{39a}$$

where

$$A(i) \triangleq 1 - \mu\eta(i) \left(\tau - \eta(i)(\tau - 1 + 1/\mu) \right), \tag{39b}$$

$$\begin{aligned} B(i) \triangleq & \frac{Z^2}{M \sigma^{\mathrm{dl}} P^{\mathrm{dl}}} + \left(1 + \mu(1 - \eta(i)) \right) \eta^2(i) G^2 \frac{\tau(\tau-1)(2\tau-1)}{6} \\ & + \left(\tau - 1 + \eta^2(i)(\tau^2 + \tau - 1) \right) G^2 + 2\eta(i)(\tau - 1)\Gamma, \end{aligned} \tag{39c}$$

and the expectation is with respect to the stochastic gradient function and the randomness of the underlying wireless channel.

Proof. See Appendix A. □

Corollary 1. *From the L-smoothness of function $F(\cdot)$, after T global iterations of the analog downlink scheme, for $0<\eta(t)\leq \min\left\{\frac{\mu}{\mu+1},\frac{1}{\mu\tau}\right\}$, $\forall t$, we have*

$$\begin{aligned}\mathbb{E}\left[F(\boldsymbol{\theta}(T))\right]-F^*\leq&\frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}(T)-\boldsymbol{\theta}^*\|_2^2\right]\\ \leq&\frac{L}{2}\left(\prod_{i=0}^{T-1}A(i)\right)\|\boldsymbol{\theta}(0)-\boldsymbol{\theta}^*\|_2^2+\frac{L}{2}\sum_{j=0}^{T-1}B(j)\prod_{i=j+1}^{T-1}A(i),\end{aligned}\tag{40}$$

where the last inequality follows from (39a).

Remark 2. *We remark that $A(i)$ is a decreasing function of τ, while $B(i)$ increases with τ. Therefore, the impact of τ on the convergence performance in the general case is not evident, since it depends also on other parameters. However, for a more biased data distribution across devices, which results in a higher Γ and G, the destructive effect of increasing τ on $B(i)$ is more significant, while the reduction in $A(i)$ is the same as having a less biased data distribution. We note that $A(i)$ is not a function of the data distribution; therefore, for a less diverse data distribution, designing an efficient τ is more critical. This corroborates our intuitive understanding of convergence in this problem, where for a more biased data distribution, increasing the number of local iterations excessively leads to a more divergent local updates with a less chance of convergence.*

Remark 3. *The two terms, $\frac{Z^2}{M\sigma^{\mathrm{dl}}P^{\mathrm{dl}}}$ and $(\tau-1)G^2$ in $B(i)$, are not scaled with the learning rate, $\eta(i)$. Therefore, even for a decreasing learning rate, where $\lim_{t\to\infty}\eta(t)=0$, we have $\lim_{t\to\infty}B(t)=\frac{Z^2}{M\sigma^{\mathrm{dl}}P^{\mathrm{dl}}}+(\tau-1)G^2\neq 0$, which shows that $\lim_{t\to\infty}\mathbb{E}\left[F(\boldsymbol{\theta}(t))\right]-F^*\neq 0$. We highlight that having these two terms is the result of the noisy downlink transmission, where $\frac{Z^2}{M\sigma^{\mathrm{dl}}P^{\mathrm{dl}}}$ and $(\tau-1)G^2$ have appeared in the convergence analysis in inequalities* (47) *and* (55)*, respectively, in the appendices.*

VI. Numerical Experiments

Here we compare the performance of the proposed digital and analog downlink approaches for image classification on the MNIST dataset [39] with 60000 training and 10000 test samples. We train a convolutional neural network (CNN) with 6 layers including two 5×5 convolutional layers with ReLU activation and the same padding, where the first and the second layers have 32 and 64 channels, respectively, each with stride 1, and followed by a 2×2 max pooling layer with stride 2. Also, the CNN has a fully connected layer with 1024

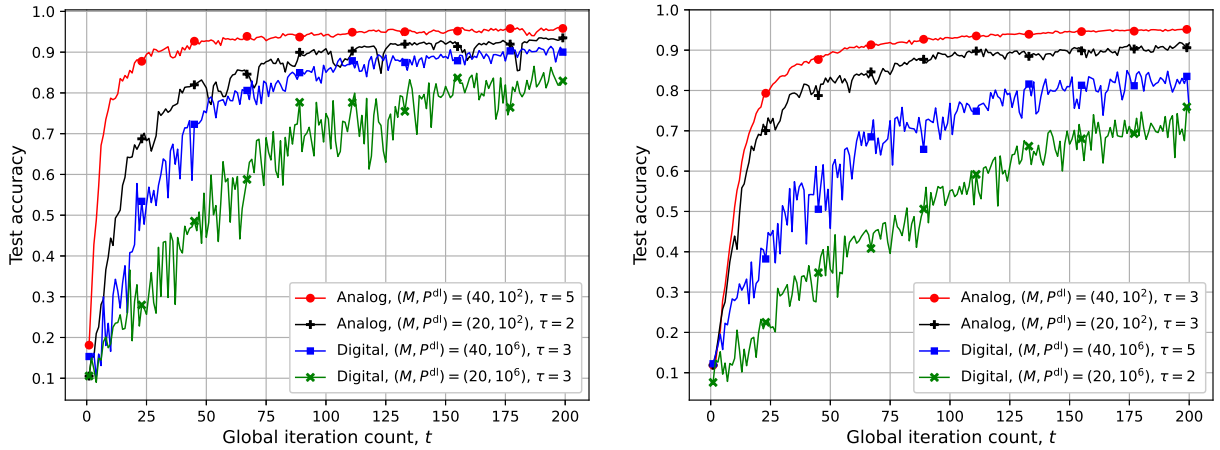
units and ReLU activation with dropout 0.8 followed by a softmax output layer. We utilize ADAM optimizer [40] to train the CNN.

We consider two scenarios: in the *iid data distribution* scenario, we randomly split the 60000 training data samples to M disjoint subsets, and allocate each subset of data samples to a different device; while in the *non-iid data distribution* scenario, we split the training data samples with the same label (from the same class) to $M/5$ disjoint subsets (assume that M is divisible by 5). We then assign two subsets of the data samples, each from a different label/class selected at random, to each device, such that each subset of the data samples is assigned to a single device.

We assume $n^{\rm dl} = n^{\rm ul} = d/2$ subchannels, and a variance of $\sigma^{\rm dl} = \sigma^{\rm ul} = 1$ for the downlink and uplink channel gains. We set the transmit power constraint at the devices to $P^{\rm ul} = 10$, and the threshold on the uplink channel gains to $\lambda_{\rm thr}(t) = 10^{-4}$, $\forall t$. We also set the sparsity level of the digital downlink approach to $s = \lfloor d/50 \rfloor$ and the size of the local mini-batch sample for each local iteration to $|\xi_m^i(t)| = 500$, $\forall i, m, t$. We measure the performance as the accuracy with respect to the test samples, called *test accuracy*, versus the global iteration count, t.

For the analytical results on the convergence rate of the analog downlink approach, we set $\eta(t) = \frac{\min\left\{\frac{\mu}{\mu+1}, \frac{1}{\mu\tau}\right\}}{(10^{-3}t+1)}$, $\forall t$, and consider $M = 40$ devices. We assume that $\mu = 0.2$, $L = 10$, $\|\boldsymbol{\theta}(0) - \boldsymbol{\theta}^*\|_2^2 = 5{\times}10^3$, and $Z^2 = 2{\times}10^4$. We also model the iid and non-iid data distributions by setting $(G^2, \Gamma) = (10, 5)$ and $(G^2, \Gamma) = (100, 50)$, respectively, where we note that the non-iid scenario results in higher G and Γ values.

In Fig. 1 we compare the performance of the proposed digital and analog downlink approaches for both the iid and non-iid data distribution scenarios. We investigate the impact of the number of devices on the performance by considering $M \in \{20, 40\}$. For the analog downlink approach, we consider $P^{\rm dl} = 10^2$; while for the digital approach, we consider a significantly higher value for the downlink transmit power constraint at the PS, $P^{\rm dl} = 10^6$, which is to make sure that $q(t) \geq 1$, $\forall t$. For each experiment, whose result is illustrated in Fig. 1, we have found the number of local iterations, τ, which results in the best accuracy. Despite the significantly lower transmit power at the PS, we observe that the analog downlink scheme remarkably outperforms the digital one for both iid and non-iid scenarios with a notably larger gap between the two for the non-iid case. It can also be seen that the accuracy of the analog downlink approach is more stable than its digital

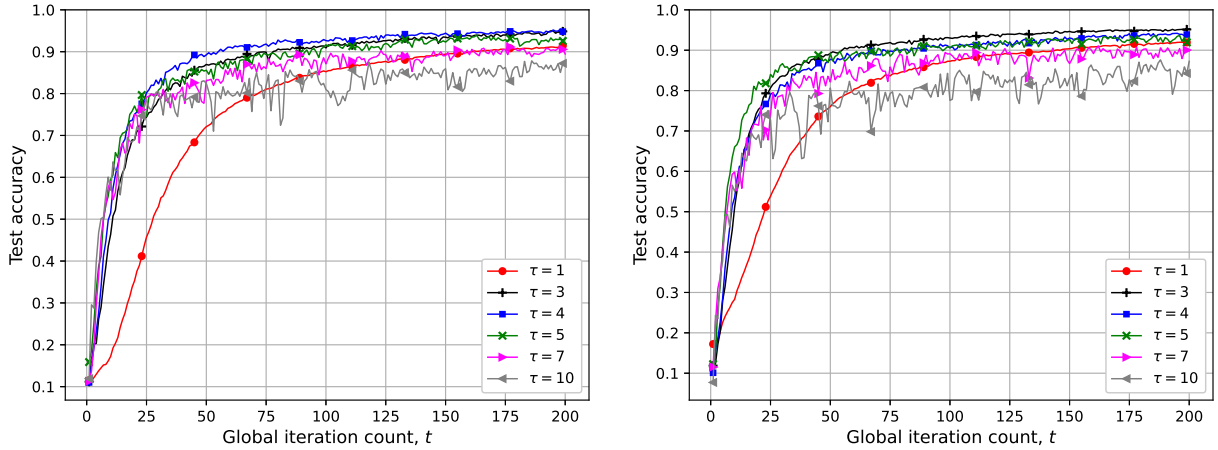


(a) Iid data distribution (b) Non-iid data distribution

Fig. 1: Accuracy of the digital and analog downlink approaches for $n^{\text{dl}} = n^{\text{ul}} = d/2$, $\sigma^{\text{dl}} = \sigma^{\text{ul}} = 1$, $P^{\text{ul}} = 10$, $\lambda_{\text{thr}}(t) = 10^{-4}$, $\forall t$, $s = \lfloor d/50 \rfloor$ for the digital approach, and $|\xi_m^i(t)| = 500$, $\forall i, m, t$.

counterpart, and the degradation in the performance of the analog approach due to the introduced bias in the non-iid data distribution is marginal. This shows that the analog approach is fairly robust against the heterogeneity of data distribution across devices. We highlight that with the analog downlink approach the destructive effect of the devices with relatively bad channel conditions, and consequently with a noisier/less accurate estimate of the global model, is alleviated with the devices with good channel conditions, since devices receive different estimates of the global model vector transmitted by the PS depending on their channel conditions. On the other hand, with the digital downlink approach the common rate at which the global model vector is delivered to the devices should be adjusted such that all the devices, including those with relatively bad channel conditions, can decode it. This limits the capacity of the devices with good channel conditions, and provides the same copy of the global model estimate to all the devices whose rate is adjusted to accommodate even the worst device. Another reason for the inferiority of the digital downlink approach is that it requires digitization/quantization of the model parameter vector to a limited number of bits, which provides a less accurate estimate of the global model vector to rely on for local training at the devices than the noisy estimate received from the analog downlink transmission. This is due to the limited capacity of the wireless broadcast channel.

The performance of both digital and analog downlink approaches improve with M for



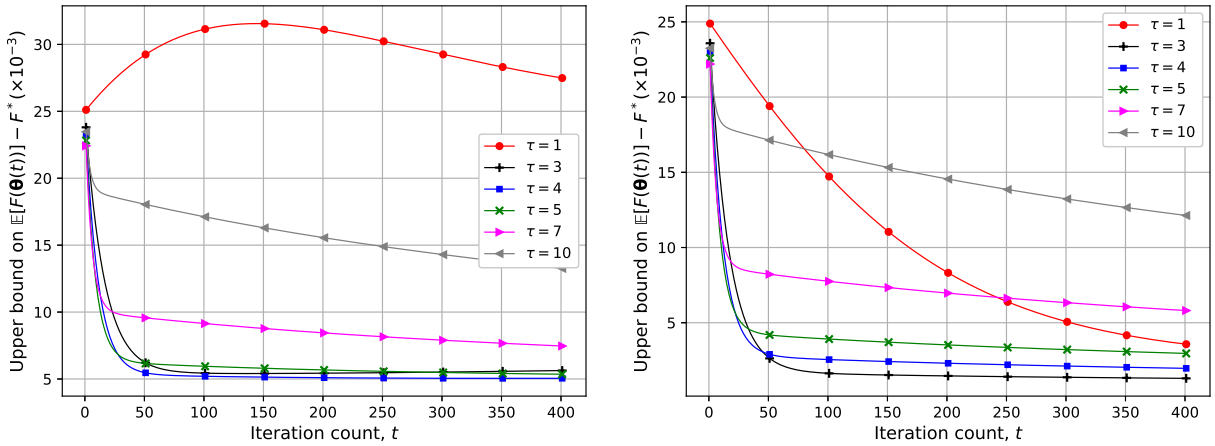
(a) $P^{\mathrm{dl}} = 10$ (b) $P^{\mathrm{dl}} = 10^2$

Fig. 2: Accuracy of analog downlink for the non-iid data distribution with $M = 40$, $n^{\mathrm{dl}} = n^{\mathrm{ul}} = d/2$, $\sigma^{\mathrm{dl}} = \sigma^{\mathrm{ul}} = 1$, $P^{\mathrm{ul}} = 10$, $\lambda_{\mathrm{thr}}(t) = 10^{-4}$, $\forall t$, and $|\xi_m^i(t)| = 500$, $\forall i, m, t$.

both iid and non-iid scenarios. This is mainly due to the uplink transmission. With more devices, each with its own power budget, analog transmission over the MAC is more robust against the noise, which is due to the additive nature of the MAC. However, the accuracy of the digital downlink approach is unstable in both iid and non-iid cases. This is due to the inaccurate model parameter vector estimate at the devices for the digital downlink approach, which leads to a more skewed/less similar local updates at the devices compared to the case of having the actual model parameter vector at the devices. This deficiency can be clearly seen for $M = 20$ in the iid scenario. By relying on the local updates from fewer devices, the chance of having more similar local updates (local updates with relatively small Euclidean distance) decreases, and it is less likely that the resultant vector recovered from the output of the MAC provides a good estimate of the gradient of the actual model parameter vector. Another interesting observation is about the best number of local iterations τ for each experiment. We observe that the best τ value for the analog downlink approach for $M = 40$ ($M = 20$) in the iid case is the same as that for the digital downlink approach for $M = 40$ ($M = 20$) in the non-iid scenario. The same observation can be made also for the performance of the digital downlink approach in the iid case and the analog downlink approach in the non-iid scenario. The reason for this opposite behavior is that, in contrast to the digital downlink approach, with the analog approach the devices have a relatively good estimate of $\boldsymbol{\theta}(t)$. For the analog downlink approach with sufficiently many devices,

i.e., $M = 40$, the best τ value for the iid case is larger than that for the non-iid case. This is intuitive since increasing τ excessively for the non-iid case provides biased local updates at the devices, which is due to the biased local datasets, with a relatively poor similarity. On the other hand, the digital downlink approach for $M = 40$ shows the opposite behavior, which is due to the relatively inaccurate estimate of $\boldsymbol{\theta}(t)$ at the devices. In this case, for the iid scenario, in which the local data is homogeneous, the inaccuracy of the model parameter vector estimate harms the performance when a relatively large number of local SGD iterations are performed for both M values. Whereas, for $M = 40$ in the non-iid scenario, a relatively small τ might not provide reliable local updates, since the local training dataset is biased and a relatively good estimate of $\boldsymbol{\theta}(t)$ is not available to rely on. On the other hand, for the digital approach with $M = 20$, where devices receive a more accurate estimate of $\boldsymbol{\theta}(t)$, due to the higher achievable common rate, a relatively small τ value provides a better performance. A similar observation is made for the analog downlink approach with $M = 20$ devices in the iid case, where a relatively small τ, $\tau = 2$, provides the best performance. This is due to the fact that, having less devices for training, where each device performs local updates using homogeneous local data and a distinct noisy version of the global model, the chance of having the noise in the local updates cancelled out at the aggregation phase at the PS reduces when a relatively large τ is used for local updates. We provide a more in-depth investigation of the impact of number of local SGD iterations on the performance of the analog downlink approach in Figures 2 and 3. We remark here that the randomness in the experiments also have an impact on the experimental results presented here.

In Fig. 2 we study the impact of τ on the performance of the analog downlink approach focusing on the non-iid data distribution for two different transmit power levels $P^{\text{dl}} \in \{10, 10^2\}$ at the PS with $\tau \in \{1, 3, 4, 5, 7, 10\}$ and $M = 40$ devices. We note that with a higher P^{dl} the devices receive a better/less noisy estimate of $\boldsymbol{\theta}(t)$. Observe that, for a smaller P^{dl}, $P^{\text{dl}} = 10$, $\tau = 4$ provides the best performance, while for $P^{\text{dl}} = 10^2$, the best performance is achieved for $\tau = 3$. Therefore, for the non-iid scenario, when having a less accurate estimate of $\boldsymbol{\theta}(t)$ at the devices, a larger number of local SGD iterations should be performed compared to having a more accurate estimate of $\boldsymbol{\theta}(t)$ at the devices. As discussed for the performance of the digital downlink approach in Fig. 1, a relatively small τ value might not provide the most reliable local updates for the non-iid scenario when a good estimate



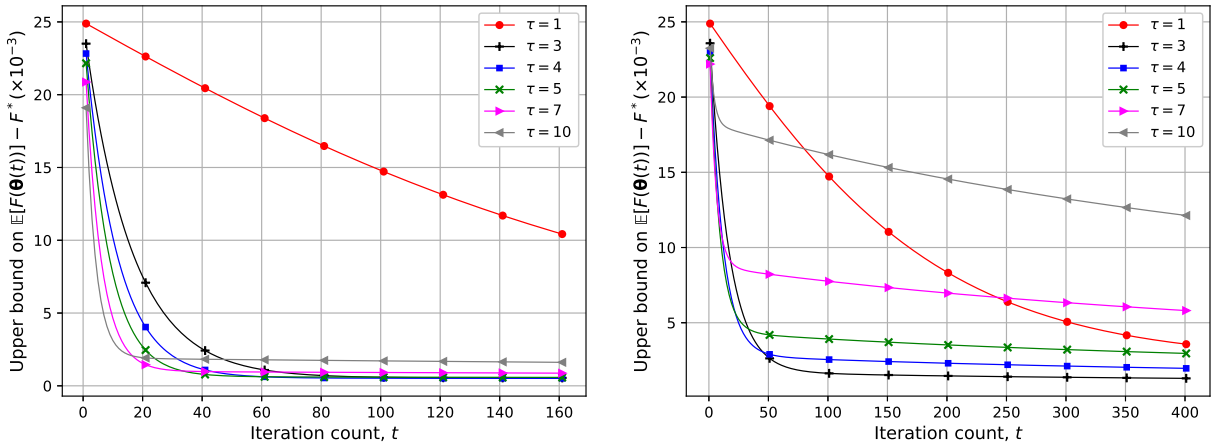
(a) $P^{\text{dl}} = 10$ (b) $P^{\text{dl}} = 10^2$

Fig. 3: Upper bound on $\mathbb{E}\left[F(\boldsymbol{\theta}(t))\right] - F^*$ for analog downlink for different τ values, $\tau \in \{1, 3, 4, 5, 7, 10\}$, considering non-iid data distribution with $(G^2, \Gamma) = (100, 50)$, for $\eta(t) = \frac{\min\left\{\frac{\mu}{\mu+1}, \frac{1}{\mu\tau}\right\}}{(10^{-3}t+1)}$, $\forall t$, $M = 40$, $\mu = 0.2$, $L = 10$, $\|\boldsymbol{\theta}(0) - \boldsymbol{\theta}^*\|_2^2 = 5 \times 10^3$, and $Z^2 = 2 \times 10^4$.

of $\boldsymbol{\theta}(t)$ is not available at the devices. This observation is corroborated in Fig. 3, which demonstrates the analytical results on the convergence rate bound of the analog downlink approach for the non-iid scenario for different τ values, $\tau \in \{1, 3, 4, 5, 7, 10\}$, with two P^{dl} values, $P^{\text{dl}} \in \{10, 10^2\}$. We observe in this figure that, for $P^{\text{dl}} = 10$, $\tau = 4$ provides the best performance in terms of the convergence speed and the final level of the average loss. Whereas, for $P^{\text{dl}} = 10^2$, $\tau = 3$ provides the lowest average loss, although it has a negligibly smaller convergence speed compared to $\tau = 4, 5, 7$.

In Fig. 4, we consider the analytical convergence result of the analog downlink approach for the iid and non-iid scenarios for various τ values, $\tau \in \{1, 3, 4, 5, 7, 10\}$. We observe that, for the iid scenario, considering both the convergence rate and the final average loss, $\tau = 5$ provides the best performance, although it has a slightly smaller convergence speed compared to $\tau = 7, 10$. On the other hand, we observe that a smaller τ value, $\tau = 3$, has the best performance in the non-iid scenario. This result corroborates the observation made in Fig. 1 for the analog downlink approach with $M = 40$ devices, in which a larger τ value should be used for a less biased data distribution to obtain the best performance. A relatively large τ for non-iid data results in a more biased/skewed local updates with less consensus.

There results suggest that a schedule for τ that depends on the iteration t might work well in a wide range of scenarios. Specifically, start with a larger τ and decrease it as t increases.



(a) Iid data distribution, $(G^2, \Gamma) = (10, 5)$ (b) Non-iid data distribution, $(G^2, \Gamma) = (100, 50)$

Fig. 4: Upper bound on $\mathbb{E}[F(\boldsymbol{\theta}(t))] - F^*$ for the analog downlink approach for different τ values, $\tau \in \{1, 3, 4, 5, 7, 10\}$, with $P^{\text{dl}} = 10^2$, for $\eta(t) = \frac{\min\left\{\frac{\mu}{\mu+1}, \frac{1}{\mu\tau}\right\}}{(10^{-3}t+1)}$, $\forall t$, $M = 40$, $\mu = 0.2$, $L = 10$, $\|\boldsymbol{\theta}(0) - \boldsymbol{\theta}^*\|_2^2 = 5 \times 10^3$, and $Z^2 = 2 \times 10^4$.

VII. Conclusions

We have studied FEEL, where the PS with a limited power budget transmits the model parameter vector to the wireless devices over a bandwidth-limited fading broadcast channel. We have proposed digital and analog transmission approaches for the PS-to-devices transmission. With the digital approach, the PS quantizes the global model update, with respect to the global model estimate at the devices, with the knowledge of the highest common rate sustainable over the downlink broadcast channel. For the analysis, we have utilized a capacity achieving channel code to broadcast the same estimate of the global model update to all the devices. On the other hand, with the analog approach, the PS broadcasts the global model vector in an uncoded manner without employing any channel code, and the devices receive different estimates of the global model through independent wireless connections. In both approaches, the devices perform multiple local SGD iterations with respect to their global model estimates utilizing their local datasets. The power-limited wireless devices then transmit their local model updates to the PS over a bandwidth-limited fading MAC in an analog fashion, whose superiority over digital transmission for the uplink has been shown in the literature [12], [13], [16]. We have also provided a convergence analysis for the analog downlink approach to study the impact of imperfect downlink transmission, leading to noisy estimates of the global model at the devices, on the performance of FL, where for

the ease of analysis we have assumed that the uplink transmission is error-free. Numerical experiments on the MNIST dataset have shown a significant improvement of the analog downlink approach over its digital counterpart, where the improvement is more pronounced for the non-iid data scenario. The analog downlink approach benefits from providing the devices with different estimates of the global model with the quality of these estimates depending on their downlink channel conditions, in which case the destructive effect of the devices with relatively worse channel conditions, and consequently less accurate estimates, can be alleviated by the devices with better channel conditions. However, with the digital downlink approach, the devices receive the same estimate of the model parameter vector with a common rate limited by the capacity of the worst device. Therefore, it is likely that all the devices perform local SGD iterations using an inaccurate estimate of the global model. Both the experimental and analytical results have shown that a smaller number of local SGD iterations should be performed to obtain the best performance of the analog downlink approach for non-iid data compared to iid data. Also, for non-iid data, by increasing the transmit power at the PS, which leads to a more accurate global model estimate at the devices, a smaller number of local SGD iterations should be performed at the devices.

Appendix A
Proof of Theorem 1

The global model parameter vector for the analog downlink approach is updated as

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) + \sum_{m=1}^{M} \frac{B_m}{B} \Delta\boldsymbol{\theta}_m(t). \tag{41}$$

We have

$$\mathbb{E}\left[\|\boldsymbol{\theta}(t+1) - \boldsymbol{\theta}^*\|_2^2\right] = \mathbb{E}\left[\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*\|_2^2\right] + \mathbb{E}\left[\left\|\sum_{m=1}^{M} \frac{B_m}{B} \Delta\boldsymbol{\theta}_m(t)\right\|_2^2\right] + 2\mathbb{E}\left[\langle \boldsymbol{\theta}(t) - \boldsymbol{\theta}^*, \sum_{m=1}^{M} \frac{B_m}{B} \Delta\boldsymbol{\theta}_m(t)\rangle\right]. \tag{42}$$

Next we bound the last two terms on the right hand side (RHS) of (42).

From the convexity of $\|\cdot\|_2^2$, it follows that

$$\mathbb{E}\left[\left\|\sum_{m=1}^{M} \frac{B_m}{B} \Delta\boldsymbol{\theta}_m(t)\right\|_2^2\right] \leq \sum_{m=1}^{M} \frac{B_m}{B} \mathbb{E}\left[\|\Delta\boldsymbol{\theta}_m(t)\|_2^2\right]$$

$$= \eta^2(t) \sum_{m=1}^{M} \frac{B_m}{B} \mathbb{E}\left[\left\| \sum_{i=1}^{\tau} \nabla F_m \left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t) \right) \right\|_2^2 \right]$$
$$\leq \eta^2(t) \tau \sum_{m=1}^{M} \sum_{i=1}^{\tau} \frac{B_m}{B} \mathbb{E}\left[\left\| \nabla F_m \left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t) \right) \right\|_2^2 \right] \overset{\text{(a)}}{\leq} \eta^2(t) \tau^2 G^2, \tag{43}$$

where (a) follows from Assumption 3.

We rewrite the third term on the RHS of (42) as follows:

$$2\mathbb{E}\left[\langle \boldsymbol{\theta}(t) - \boldsymbol{\theta}^*, \sum_{m=1}^{M} \frac{B_m}{B} \Delta \boldsymbol{\theta}_m(t) \rangle \right]$$
$$= 2\eta(t) \sum_{m=1}^{M} \frac{B_m}{B} \mathbb{E}\left[\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}(t), \sum_{i=1}^{\tau} \nabla F_m \left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t) \right) \rangle \right]$$
$$= 2\eta(t) \sum_{m=1}^{M} \frac{B_m}{B} \mathbb{E}\left[\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}(t), \nabla F_m \left(\boldsymbol{\theta}(t) + \tilde{\boldsymbol{z}}_m^{\text{dl}}(t), \xi_m^1(t) \right) \rangle \right]$$
$$+ 2\eta(t) \sum_{m=1}^{M} \frac{B_m}{B} \mathbb{E}\left[\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}(t), \sum_{i=2}^{\tau} \nabla F_m \left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t) \right) \rangle \right]. \tag{44}$$

We have

$$2\eta(t) \sum_{m=1}^{M} \frac{B_m}{B} \mathbb{E}\left[\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}(t), \nabla F_m \left(\boldsymbol{\theta}(t) + \tilde{\boldsymbol{z}}_m^{\text{dl}}(t), \xi_m^1(t) \right) \rangle \right]$$
$$= 2\eta(t) \sum_{m=1}^{M} \frac{B_m}{B} \mathbb{E}\left[\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}(t), \nabla F_m \left(\boldsymbol{\theta}(t), \xi_m^1(t) \right) \rangle \right]$$
$$+ 2\eta(t) \sum_{m=1}^{M} \frac{B_m}{B} \mathbb{E}\left[\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}(t), \nabla F_m \left(\boldsymbol{\theta}(t) + \tilde{\boldsymbol{z}}_m^{\text{dl}}(t), \xi_m^1(t) \right) - \nabla F_m \left(\boldsymbol{\theta}(t), \xi_m^1(t) \right) \rangle \right]. \tag{45}$$

In the following, we bound the two terms on the RHS of (45). We have

$$2\eta(t) \sum_{m=1}^{M} \frac{B_m}{B} \mathbb{E}\left[\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}(t), \nabla F_m \left(\boldsymbol{\theta}(t), \xi_m^1(t) \right) \rangle \right]$$
$$\overset{\text{(a)}}{=} 2\eta(t) \sum_{m=1}^{M} \frac{B_m}{B} \mathbb{E}\left[\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}(t), \nabla F_m \left(\boldsymbol{\theta}(t) \right) \rangle \right]$$
$$\overset{\text{(b)}}{\leq} 2\eta(t) \sum_{m=1}^{M} \frac{B_m}{B} \mathbb{E}\left[F_m(\boldsymbol{\theta}^*) - F_m\left(\boldsymbol{\theta}(t) \right) - \frac{\mu}{2} \left\| \boldsymbol{\theta}(t) - \boldsymbol{\theta}^* \right\|_2^2 \right]$$
$$= 2\eta(t) \left(F^* - \mathbb{E}\left[F\left(\boldsymbol{\theta}(t) \right) \right] - \frac{\mu}{2} \mathbb{E}\left[\left\| \boldsymbol{\theta}(t) - \boldsymbol{\theta}^* \right\|_2^2 \right] \right), \tag{46}$$

where (a) and (b) follow from (5) and Assumption 2, respectively. Also, from Cauchy-Schwarz inequality, we have

$$2\eta(t) \sum_{m=1}^{M} \frac{B_m}{B} \mathbb{E}\left[\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}(t), \nabla F_m \left(\boldsymbol{\theta}(t) + \tilde{\boldsymbol{z}}_m^{\text{dl}}(t), \xi_m^1(t) \right) - \nabla F_m \left(\boldsymbol{\theta}(t), \xi_m^1(t) \right) \rangle \right]$$
$$\leq \eta^2(t) \mathbb{E}\left[\left\| \boldsymbol{\theta}(t) - \boldsymbol{\theta}^* \right\|_2^2 \right] + \mathbb{E}\left[\left\| \sum_{m=1}^{M} \frac{B_m}{B} \left(\nabla F_m(\boldsymbol{\theta}(t) + \tilde{\boldsymbol{z}}_m^{\text{dl}}(t), \xi_m^1(t)) - \nabla F_m(\boldsymbol{\theta}(t), \xi_m^1(t)) \right) \right\|^2 \right]$$
$$\overset{\text{(a)}}{\leq} \eta^2(t) \mathbb{E}\left[\left\| \boldsymbol{\theta}(t) - \boldsymbol{\theta}^* \right\|_2^2 \right] + \frac{Z^2}{M \sigma^{\text{dl}} P^{\text{dl}}}, \tag{47}$$

where (a) follows from Assumption 4. Substituting (46) and (47) into (45) yields

$$2\eta(t)\sum_{m=1}^{M}\frac{B_m}{B}\mathbb{E}\left[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}(t),\nabla F_m\left(\boldsymbol{\theta}(t)+\tilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t),\xi_m^1(t)\right)\rangle\right]$$
$$\leq -\mu\eta(t)\left(1-\eta(t)/\mu\right)\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]+\frac{Z^2}{M\sigma^{\mathrm{dl}}P^{\mathrm{dl}}}+2\eta(t)\left(F^*-\mathbb{E}\left[F\left(\boldsymbol{\theta}(t)\right)\right]\right). \quad (48)$$

Lemma 1. *For $0<\eta(t)\leq\frac{\mu}{\mu+1}$, we have*

$$2\eta(t)\sum_{m=1}^{M}\frac{B_m}{B}\mathbb{E}\left[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}(t),\sum_{i=2}^{\tau}\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\right]$$
$$\leq -\mu\eta(t)(1-\eta(t))(\tau-1)\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]$$
$$+(1+\mu(1-\eta(t)))\eta^2(t)G^2\frac{\tau(\tau-1)(2\tau-1)}{6}+2\eta(t)(\tau-1)\Gamma$$
$$+\left(\eta^2(t)+1\right)(\tau-1)G^2+2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\left(F_m^*-\mathbb{E}\left[F_m(\boldsymbol{\theta}_m^i(t))\right]\right). \quad (49)$$

Proof. See Appendix B. □

By substituting (48) and (49) in (44), it follows that

$$2\mathbb{E}\left[\langle\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*,\sum_{m=1}^{M}\frac{B_m}{B}\Delta\boldsymbol{\theta}_m(t)\rangle\right]\leq-\mu\eta(t)\left(\tau-\eta(t)(\tau-1+1/\mu)\right)\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]$$
$$+\frac{Z^2}{M\sigma^{\mathrm{dl}}P^{\mathrm{dl}}}+(1+\mu(1-\eta(t)))\eta^2(t)G^2\frac{\tau(\tau-1)(2\tau-1)}{6}+\left(\eta^2(t)+1\right)(\tau-1)G^2$$
$$+2\eta(t)(\tau-1)\Gamma+2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\left(F_m^*-\mathbb{E}\left[F_m(\boldsymbol{\theta}_m^i(t))\right]\right)+2\eta(t)\left(F^*-\mathbb{E}\left[F\left(\boldsymbol{\theta}(t)\right)\right]\right), \quad (50)$$

which together with the inequality in (43), according to (42), the following upper bound on $\mathbb{E}\left[\|\boldsymbol{\theta}(t+1)-\boldsymbol{\theta}^*\|_2^2\right]$ is obtained:

$$\mathbb{E}\left[\|\boldsymbol{\theta}(t+1)-\boldsymbol{\theta}^*\|_2^2\right]\leq\left(1-\mu\eta(t)\left(\tau-\eta(t)(\tau-1+1/\mu)\right)\right)\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]+\frac{Z^2}{M\sigma^{\mathrm{dl}}P^{\mathrm{dl}}}$$
$$+(1+\mu(1-\eta(t)))\eta^2(t)G^2\frac{\tau(\tau-1)(2\tau-1)}{6}+\left(\tau-1+\eta^2(t)\left(\tau^2+\tau-1\right)\right)G^2$$
$$+2\eta(t)(\tau-1)\Gamma+2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\left(F_m^*-\mathbb{E}\left[F_m(\boldsymbol{\theta}_m^i(t))\right]\right)+2\eta(t)\left(F^*-\mathbb{E}\left[F\left(\boldsymbol{\theta}(t)\right)\right]\right)$$
$$\overset{(a)}{\leq}\left(1-\mu\eta(t)\left(\tau-\eta(t)(\tau-1+1/\mu)\right)\right)\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]+\frac{Z^2}{M\sigma^{\mathrm{dl}}P^{\mathrm{dl}}}$$
$$+(1+\mu(1-\eta(t)))\eta^2(t)G^2\frac{\tau(\tau-1)(2\tau-1)}{6}$$
$$+\left(\tau-1+\eta^2(t)\left(\tau^2+\tau-1\right)\right)G^2+2\eta(t)(\tau-1)\Gamma, \quad (51)$$

where (a) follows sine $F^* - F(\boldsymbol{\theta}(t)) \leq 0$, $\forall t$, and $F_m^* - F_m(\boldsymbol{\theta}_m^i(t)) \leq 0$, $\forall m, i, t$. It is trivial to prove Theorem 1 from the inequality in (51) for $0 < \eta(t) \leq \min\left\{\frac{\mu}{\mu+1}, \frac{1}{\mu\tau}\right\}$, $\forall t$.

Appendix B
Proof of Lemma 1

We have

$$\begin{aligned}
&2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\langle\boldsymbol{\theta}^* - \boldsymbol{\theta}(t), \nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right)\rangle\right] \\
&= 2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\langle\boldsymbol{\theta}_m^i(t) - \boldsymbol{\theta}(t), \nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right)\rangle\right] \\
&\quad + 2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\langle\boldsymbol{\theta}^* - \boldsymbol{\theta}_m^i(t), \nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right)\rangle\right]. \qquad (52)
\end{aligned}$$

For the first term on the RHS of (52), we have

$$\begin{aligned}
&2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\langle\boldsymbol{\theta}_m^i(t) - \boldsymbol{\theta}(t), \nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right)\rangle\right] \\
&\quad = 2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\langle\boldsymbol{\theta}_m^i(t) - \boldsymbol{\theta}_m^1(t), \nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right)\rangle\right] \\
&\quad\quad + 2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\langle\tilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t), \nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right)\rangle\right]. \qquad (53)
\end{aligned}$$

From Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\langle\boldsymbol{\theta}_m^i(t) - \boldsymbol{\theta}_m^1(t), \nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right)\rangle\right] \\
&\leq \eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\frac{1}{\eta(t)}\left\|\boldsymbol{\theta}_m^i(t) - \boldsymbol{\theta}_m^1(t)\right\|_2^2 + \eta(t)\left\|\nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right)\right\|_2^2\right] \\
&\overset{(a)}{\leq} \sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\left\|\boldsymbol{\theta}_m^i(t) - \boldsymbol{\theta}_m^1(t)\right\|_2^2\right] + \eta^2(t)\,(\tau-1)\,G^2, \qquad (54)
\end{aligned}$$

and

$$\begin{aligned}
&2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\langle\tilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t), \nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right)\rangle\right] \\
&\leq \eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\eta(t)\left\|\tilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t)\right\|_2^2 + \frac{1}{\eta(t)}\left\|\nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right)\right\|_2^2\right] \\
&\overset{(a)}{\leq} \eta^2(t)(\tau-1)\sum_{m=1}^{M}\frac{B_m}{B}\mathbb{E}\left[\left\|\tilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t)\right\|_2^2\right] + (\tau-1)\,G^2, \qquad (55)
\end{aligned}$$

where (a) follows from Assumption 3. Thus, the term on the left hand side (LHS) of (53) is bounded as

$$\begin{aligned}
&2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\langle\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}(t),\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\right]\\
&\quad\le\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}_m^1(t)\right\|_2^2\right]+\eta^2(t)(\tau-1)\sum_{m=1}^{M}\frac{B_m}{B}\mathbb{E}\left[\left\|\widetilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t)\right\|_2^2\right]\\
&\qquad+\left(\eta^2(t)+1\right)(\tau-1)G^2. \qquad (56)
\end{aligned}$$

From convexity of $\|\cdot\|_2^2$, we have

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}_m^1(t)\right\|_2^2\right]=\eta^2(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\left\|\sum_{j=1}^{i-1}\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\right\|_2^2\right]\\
&\le\eta^2(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}(i-1)\sum_{j=1}^{i-1}\mathbb{E}\left[\left\|\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\right\|_2^2\right]\overset{(a)}{\le}\eta^2(t)G^2\frac{\tau(\tau-1)(2\tau-1)}{6}, \qquad (57)
\end{aligned}$$

where (a) follows from Assumption 3. For the second term on the RHS of (52), we have

$$\begin{aligned}
&2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}_m^i(t),\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\right]\\
&\overset{(a)}{=}2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}_m^i(t),\nabla F_m\left(\boldsymbol{\theta}_m^i(t)\right)\rangle\right]\\
&\overset{(b)}{\le}2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[F_m(\boldsymbol{\theta}^*)-F_m(\boldsymbol{\theta}_m^i(t))-\frac{\mu}{2}\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}^*\right\|_2^2\right]\\
&=2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[F_m(\boldsymbol{\theta}^*)-F_m^*+F_m^*-F_m(\boldsymbol{\theta}_m^i(t))-\frac{\mu}{2}\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}^*\right\|_2^2\right]\\
&=2\eta(t)(\tau-1)\Gamma+2\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\left(F_m^*-\mathbb{E}\left[F_m(\boldsymbol{\theta}_m^i(t))\right]\right)\\
&\quad-\mu\eta(t)\sum_{m=1}^{M}\sum_{i=2}^{\tau}\frac{B_m}{B}\mathbb{E}\left[\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}^*\right\|_2^2\right], \qquad (58)
\end{aligned}$$

where (a) follows since $\mathbb{E}_\xi\left[\nabla F_m\left(\boldsymbol{\theta}(t),\xi_m^i(t)\right)\right]=\nabla F_m\left(\boldsymbol{\theta}(t)\right)$, $\forall i,m,t$, and (b) follows due to the fact that F_m is μ-strongly convex. We have

$$\begin{aligned}
&-\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}^*\right\|_2^2=-\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}_m^1(t)\right\|_2^2-\left\|\boldsymbol{\theta}_m^1(t)-\boldsymbol{\theta}^*\right\|_2^2-2\langle\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}_m^1(t),\boldsymbol{\theta}_m^1(t)-\boldsymbol{\theta}^*\rangle\\
&\overset{(a)}{\le}-\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}_m^1(t)\right\|_2^2-\left\|\boldsymbol{\theta}_m^1(t)-\boldsymbol{\theta}^*\right\|_2^2+\frac{1}{\eta(t)}\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}_m^1(t)\right\|_2^2+\eta(t)\left\|\boldsymbol{\theta}_m^1(t)-\boldsymbol{\theta}^*\right\|_2^2\\
&=-(1-\eta(t))\left\|\boldsymbol{\theta}_m^1(t)-\boldsymbol{\theta}^*\right\|_2^2+\left(\frac{1}{\eta(t)}-1\right)\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}_m^1(t)\right\|_2^2,\quad i\in[\tau],m\in[M], \qquad (59)
\end{aligned}$$

where (a) follows from Cauchy-Schwarz inequality. For $\eta(t) \leq 1$, we have

$$\begin{aligned}
&-(1-\eta(t))\mathbb{E}\left[\left\|\boldsymbol{\theta}_m^1(t)-\boldsymbol{\theta}^*\right\|_2^2\right] = -(1-\eta(t))\mathbb{E}\left[\left\|\boldsymbol{\theta}(t)+\widetilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t)-\boldsymbol{\theta}^*\right\|_2^2\right] \\
&\quad = -(1-\eta(t))\Big(\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]+\mathbb{E}\left[\left\|\widetilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t)\right\|_2^2\right]+\mathbb{E}\left[2\langle\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*,\widetilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t)\rangle\right]\Big) \\
&\quad \overset{\text{(a)}}{=} -(1-\eta(t))\left(\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]+\mathbb{E}\left[\left\|\widetilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t)\right\|_2^2\right]\right), \qquad (60)
\end{aligned}$$

where (a) follows since $\mathbb{E}\left[\widetilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t)\right]=\mathbf{0}$, and the fact that $\boldsymbol{\theta}(t)$ is independent of $\widetilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t)$, for $m\in[M]$. According to (59) and (60), it follows that, for $i\in[\tau]$, $m\in[M]$,

$$\begin{aligned}
-\mathbb{E}\left[\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}^*\right\|_2^2\right] \leq &-(1-\eta(t))\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]+\Big(\frac{1}{\eta(t)}-1\Big)\mathbb{E}\left[\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}_m^1(t)\right\|_2^2\right] \\
&-(1-\eta(t))\mathbb{E}\left[\left\|\widetilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t)\right\|_2^2\right]. \qquad (61)
\end{aligned}$$

Substituting (61) into (58) yields

$$\begin{aligned}
&2\eta(t)\textstyle\sum_{m=1}^M\sum_{i=2}^\tau\frac{B_m}{B}\mathbb{E}\left[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}_m^i(t),\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\right] \\
&\leq -\mu\eta(t)(1-\eta(t))(\tau-1)\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]+\mu(1-\eta(t))\eta^2(t)G^2\frac{\tau(\tau-1)(2\tau-1)}{6} \\
&\quad +2\eta(t)(\tau-1)\Gamma-\mu\eta(t)\left(1-\eta(t)\right)(\tau-1)\textstyle\sum_{m=1}^M\frac{B_m}{B}\mathbb{E}\left[\left\|\widetilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t)\right\|_2^2\right] \\
&\quad +2\eta(t)\textstyle\sum_{m=1}^M\sum_{i=2}^\tau\frac{B_m}{B}\left(F_m^*-\mathbb{E}\left[F_m(\boldsymbol{\theta}_m^i(t))\right]\right), \qquad (62)
\end{aligned}$$

where we have used the inequality in (57). Substituting (56) and (62) into (52) yields

$$\begin{aligned}
&2\eta(t)\textstyle\sum_{m=1}^M\sum_{i=2}^\tau\frac{B_m}{B}\mathbb{E}\left[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}(t),\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\right] \\
&\leq -\mu\eta(t)(1-\eta(t))(\tau-1)\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]+\left(1+\mu(1-\eta(t))\right)\eta^2(t)G^2\frac{\tau(\tau-1)(2\tau-1)}{6} \\
&\quad +2\eta(t)(\tau-1)\Gamma-\eta(t)(\tau-1)\left(\mu-\eta(t)(\mu+1)\right)\textstyle\sum_{m=1}^M\frac{B_m}{B}\mathbb{E}\left[\left\|\widetilde{\boldsymbol{z}}_m^{\mathrm{dl}}(t)\right\|_2^2\right] \\
&\quad +\left(\eta^2(t)+1\right)(\tau-1)G^2+2\eta(t)\textstyle\sum_{m=1}^M\sum_{i=2}^\tau\frac{B_m}{B}\left(F_m^*-\mathbb{E}\left[F_m(\boldsymbol{\theta}_m^i(t))\right]\right) \\
&\overset{\text{(a)}}{\leq} -\mu\eta(t)(1-\eta(t))(\tau-1)\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]+\left(1+\mu(1-\eta(t))\right)\eta^2(t)G^2\frac{\tau(\tau-1)(2\tau-1)}{6} \\
&\quad +2\eta(t)(\tau-1)\Gamma+\left(\eta^2(t)+1\right)(\tau-1)G^2+2\eta(t)\textstyle\sum_{m=1}^M\sum_{i=2}^\tau\frac{B_m}{B}\left(F_m^*-\mathbb{E}\left[F_m(\boldsymbol{\theta}_m^i(t))\right]\right), \\
&\qquad (63)
\end{aligned}$$

where (a) follows since $\eta(t)\leq\frac{\mu}{\mu+1}$. This completes the proof of Lemma 1.

References

[1] J. Konecny, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv:1610.05492v2 [cs.LG]*, Oct. 2017.

[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, 2017.

[3] B. McMahan and D. Ramage, "Federated learning: Collaborative machine learning without centralized training data," *[online]. Available. https://ai.googleblog.com/2017/04/federated-learning-collaborative.html*, Apr. 2017.

[4] J. Konecny and P. Richtarik, "Randomized distributed mean estimation: Accuracy vs communication," *arXiv:1611.07555 [cs.DC]*, Nov. 2016.

[5] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017.

[6] J. Konecny, B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," *arXiv:1511.03575 [cs.LG]*, Nov. 2015.

[7] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," *arXiv:1804.08333 [cs.NI]*, Oct. 2018.

[8] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," *arXiv:1806.00582 [cs.LG]*, Jun. 2018.

[9] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," *arXiv:1907.02189 [stat.ML]*, Feb. 2020.

[10] L. He, A. Bian, and M. Jaggi, "COLA: Decentralized linear learning," in *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, Montreal, Canada, 2018.

[11] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *Proc. International Conference on Machine Learning (ICML)*, Long Beach, CA, USA, 2019.

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

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

[14] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Signal Process.*, vol. 68, pp. 2897–2911, Apr. 2020.

[15] M. M. Amiri and D. Gündüz, "Over-the-air machine learning at the wireless edge," in *Proc. IEEE Int'l Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Cannes, France, Jul. 2019, pp. 1–5.

[16] ——, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, May 2020.

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

[18] T. T. Vu, D. T. Ngo, N. H. Tran, H. Q. Ngo, M. N. Dao, and R. H. Middleton, "Cell-free massive MIMO for wireless federated learning," *arXiv:1909.12567 [eess.SP]*, Dec. 2019.

[19] M. M. Amiri, T. M. Duman, and D. Gündüz, "Collaborative machine learning at the wireless edge with blind transmitters," in *Proc. IEEE Global Conference on Signal and Information Processing*, Ottawa, Canada, 2019.

[20] Y.-S. Jeon, M. M. Amiri, J. Li, and H. V. Poor, "Gradient estimation for federated learning over massive mimo communication systems," *arXiv:2003.08059 [eess.SP]*, Mar. 2020.

[21] W.-T. Chang and R. Tandon, "Communication efficient federated learning over multiple access channels," *arXiv:2001.08737 [cs.IT]*, Jan. 2020.

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

[23] H. H. Yang, A. Arafa, T. Q. S. Quek, and H. V. Poor, "Age-based scheduling policy for federated learning in mobile edge networks," *arXiv:1910.14648 [cs.IT]*, Oct. 2019.

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

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

[26] Y. Sun, S. Zhou, and D. Gündüz, "Energy-aware analog aggregation for federated learning with redundant data," *arXiv:1911.00188 [cs.IT]*, Nov. 2019.

[27] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Update aware device scheduling for federated learning at the wireless edge," in *Proc. IEEE Int'l Symp. on Inform. Theory (ISIT)*, Los Angeles, CA, USA, Jun. 2020.

[28] J. Ren, G. Yu, and G. Ding, "Accelerating DNN training in wireless federated edge learning system," *arXiv:1905.09712 [cs.LG]*, May 2019.

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

[30] C. Dinh, et al., "Federated learning over wireless networks: Convergence analysis and resource allocation," *arXiv:1910.13067 [cs.LG]*, Nov. 2019.

[31] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Convergence of update aware device scheduling for federated learning at the wireless edge," *arXiv:2001.10402 [cs.IT]*, May 2020.

[32] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Federated learning with quantized global model updates," *arXiv:2006.10672 [cs.IT]*, Jun. 2020.

[33] S. Caldas, J. Konecny, H. B. McMahan, and A. Talwalkar, "Expanding the reach of federated learning by reducing client resource requirements," *[Online]. https://arxiv.org/pdf/1812.07210.pdf*, Jan. 2019.

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

[35] Y. Liang, H. V. Poor, and S. Shamai, "Secure communication over fading channels," *IEEE Trans. Inform. Theory*, vol. 54, no. 6, pp. 2470–2492, Jun. 2008.

[36] V. L. Nir and B. Scheers, "Distributed power allocation for parallel broadcast channels with only common information in cognitive tactical radio networks," *EURASIP J. Wireless Commun. Netw.*, vol. 2010, no. 1, pp. 1–12, Jan. 2011.

[37] Y. Liang, V. V. Veeravalli, and H. V. Poor, "Resource allocation for wireless fading relay channels: Max-min solution," *IEEE Trans. Inform. Theory*, vol. 53, no. 10, pp. 3432–3453, Oct. 2007.

[38] D. Alistarh, D. Grubic, J. Z. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via randomized quantization and encoding," in *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, Dec. 2017, pp. 1709–1720.

[39] Y. LeCun, C. Cortes, and C. Burges, "The MNIST database of handwritten digits," *http://yann.lecun.com/exdb/mnist/*, 1998.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980v9 [cs.LG]*, Jan. 2017.