Blind Federated Edge Learning

Mohammad Mohammadi Amiri, *Student Member, IEEE*, Tolga M. Duman, *Fellow, IEEE*,
Deniz Gündüz, *Senior Member, IEEE*, Sanjeev R. Kulkarni, *Fellow, IEEE*, H. Vincent Poor, *Fellow, IEEE*

***Abstract*—We study federated edge learning (FEEL), where wireless edge devices, each with its own dataset, learn a global model collaboratively with the help of a wireless access point acting as the parameter server (PS). At each iteration, wireless devices perform local updates using their local data and the most recent global model received from the PS, and send their local updates to the PS over a wireless fading multiple access channel (MAC). The PS then updates the global model according to the signal received over the wireless MAC, and shares it with the devices. Motivated by the additive nature of the wireless MAC, we propose an analog 'over-the-air' aggregation scheme, in which the devices transmit their local updates in an uncoded fashion. However, unlike recent literature on over-the-air FEEL, here we assume that the devices do not have channel state information (CSI), while the PS has imperfect CSI. On the other hand, the PS is equipped with multiple antennas to alleviate the destructive effect of the channel, exacerbated due to the lack of perfect CSI. We design a receive beamforming scheme at the PS, and show that it can compensate for the lack of perfect CSI when the PS has a sufficient number of antennas. We also derive the convergence rate of the proposed algorithm highlighting the impact of the lack of perfect CSI, as well as the number of PS antennas. Both the experimental results and the convergence analysis illustrate the performance improvement of the proposed algorithm with the number of PS antennas, where the wireless fading MAC becomes deterministic despite the lack of perfect CSI when the PS has a sufficiently large number of antennas.**

***Index Terms*—Federated edge learning, fading multiple access channel, blind transmitters, multi-antenna parameter server.**

I. Introduction

With the growing prevalence of Internet of things (IoT) devices [2], constantly collecting information about various physical phenomena, and the growth in the number and processing capabilities of mobile edge devices (phones, tablets, smart watches and activity monitors), there is a growing interest in enabling machine learning (ML) to learn from data distributed across edge devices. Centralized ML techniques are often developed, assuming that the datasets are offloaded to a central processor. In the case of wireless devices, centralized ML techniques are not desirable, since offloading such massive amounts of data to a central cloud may be too costly in terms of energy and bandwidth, and may compromise data privacy. *Federated learning* (FL) has been developed to enable ML at the wireless edge by pushing the network intelligence to the edge by utilizing the processing capabilities of devices.

With FL, wireless devices train a global model collaboratively using their local datasets, which remain localized enhancing data privacy, with the help of a parameter server (PS) that keeps track of the model [3]. At each iteration of FL, the PS shares the current global model with the devices, and collects the local model updates from the devices to update the global model. This procedure continues until the global model converges, or the devices stop participating in the training because of hitting their limited power budget, or moving out of the coverage of the PS.

FL involves communications over unreliable wireless networks with limited resources, particularly in the device-to-PS direction where a large number of devices, each with limited bandwidth and power, communicate with the PS over a shared wireless medium. Therefore, it is vital to design communication-efficient protocols for the realization of an FL framework. Several approaches have been proposed in recent years to limit the communication requirements in the FL setting [3]–[9]. However, these works ignore the physical characteristics of the underlying communication channels for wireless edge learning and consider interference-and-error-free rate-limited communication links.

Recently there have been significant efforts to incorporate physical layer characteristics of wireless networks into FL system design [1], [10]–[37], referred to as *federated edge learning* (FEEL). Several studies have incorporated over-the-air computation into FEEL utilizing the superposition property of the wireless multiple access channel (MAC) for reliable transmission from the devices to the PS, where the MAC naturally provides the sum of the updates from the devices to the PS [1], [10]–[14], [17]–[19], [35]–[37]. Various device scheduling techniques for FEEL have been introduced in order to select a subset of devices sharing the limited wireless resources in each communication round [21]–[25]. Also, allocating resources to optimize a performance measure is another active research direction in FEEL [15], [27]–[30], [33]. Several studies have provided convergence guarantees of FEEL under different practical constraints and types of heterogeneity in a federated setting [8], [9], [30]–[32], [35]. Furthermore, beamforming techniques at the PS with multiple antennas have been designed to improve the quality of the estimated signal used for updating the global model [1], [12], [16]. In [12], a beamforming technique is used at the PS to maximize the number of devices participating in each communication round of training, while [16] introduces a

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).

T. M. Duman is with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey (e-mail: duman@ee.bilkent.edu.tr).

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).

Part of this work was presented at the IEEE Global Conference on Signal and Information Processing (GlobalSIP), Ottawa, ON, Canada, Nov. 2019 [1]. This work was supported in part by the U.S. National Science Foundation under Grants CCF-0939370 and CCF-1908308, the European Research Council (ERC) Starting Grant BEACON (grant agreement no. 677854), and CHIST-ERA-18-SDCDN-001grant (funded by UK EPSRC, EP/T023600/1).

nonlinear estimation method to recover the sum of updates sent from the devices using their sparsity.

In this paper, we study FEEL over a wireless fading MAC from the devices to the PS. In order to benefit from the over-the-air computation, we consider uncoded transmission of local model updates from the devices to the PS, whose advantages over digital transmission have been shown in [10]–[14]. Over-the-air computation over a wireless fading MAC requires each transmitting device to scale its transmission depending on the instantaneous channel state so that they arrive at the same power level at the PS. This, in turn, necessitates perfect channel state information (CSI) at the devices, acquisition of which would introduce additional delays and reduce the spectral efficiency. Alternatively, in this work, we consider FEEL with no CSI at the devices (i.e., the devices are blind about the CSI) and extend our previous work in [1] by studying imperfect CSI at the PS. To the best of our knowledge, this is the first paper in the FEEL literature to consider no CSI at the transmitters (CSIT) and imperfect CSI at the receiver for the device-to-PS transmission. We employ multiple antennas at the PS and design a receive beamformer to overcome the exacerbated negative impact of the underlying wireless fading MAC due to the lack of CSIT and perfect CSI at the PS. We analytically show that the proposed beamforming technique alleviates the destructive effects of the interference and noise terms at the PS thanks to the utilization of multiple antennas; and, in the limit, the fading MAC boils down to a deterministic channel with identical gains from all the devices, which is due to channel hardening [38]. Furthermore, extending our initial work in [1], we provide a convergence analysis of the proposed algorithm, and study the impact of the lack of CSIT and perfect CSI at the PS on the convergence rate. The convergence analysis shows how the increasing number of antennas at the PS remedies the lack of perfect CSI in the system. Numerical experiments on MNIST and CIFAR-10 datasets, corroborated by the analytical convergence results, illustrate the success of the proposed algorithm in combating the unavailability of perfect CSI in the system. It is shown that, despite the lack of CSI at the devices and perfect CSI at the PS, with sufficiently large number of PS antennas, the proposed algorithm can perform as well as having error-free communication links from the devices to the PS.

Notations: $\mathbb{R}$ and $\mathbb{C}$ represent the sets of real and complex values, respectively. We denote entry-wise complex conjugate of vector $\boldsymbol{x}$ by $(\boldsymbol{x})^*$, and $\mathrm{Re}\{\boldsymbol{x}\}$ and $\mathrm{Im}\{\boldsymbol{x}\}$ return entry-wise real and imaginary components of $\boldsymbol{x}$, respectively. For $\boldsymbol{x}$ and $\boldsymbol{y}$ with the same dimension, $\boldsymbol{x}\circ\boldsymbol{y}$ returns their element-wise product. We denote a zero-mean normal distribution with variance σ^2 by $\mathcal{N}\left(0,\sigma^2\right)$, and $\mathcal{CN}\left(0,\sigma^2\right)$ represents a circularly symmetric complex normal distribution with real and imaginary terms each distributed according to $\mathcal{N}\left(0,\sigma^2/2\right)$. We let $[i]\triangleq\{1,\ldots,i\}$. Notation $|\cdot|$ returns the cardinality of a set or the absolute value of a real number, and the l_2 norm of vector $\boldsymbol{x}$ is denoted by $\|\boldsymbol{x}\|_2$.

II. System Model

In FL the goal is to minimize a loss function, $F\left(\boldsymbol{\theta}\right)$, where $\boldsymbol{\theta}\in\mathbb{R}^d$ represents the model parameters to be optimized, collaboratively across M devices. We denote device m's local dataset by $\mathcal{B}_m$ with $B_m\triangleq|\mathcal{B}_m|$, for $m\in[M]$, and $B\triangleq\sum_{m=1}^{M}B_m$. We have

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

where $F_m\left(\boldsymbol{\theta}\right)$ represents the average empirical loss at device m with respect to model parameters $\boldsymbol{\theta}$; that is,

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

where $f\left(\boldsymbol{\theta},\boldsymbol{u}\right)$ denotes the empirical loss function at data sample $\boldsymbol{u}$ with respect to the model parameters $\boldsymbol{\theta}$ and is defined by the learning task. Devices perform stochastic gradient descent (SGD) to minimize the loss function $F_m\left(\boldsymbol{\theta}\right)$. During global iteration t, having received the model parameters $\boldsymbol{\theta}(t)$ from the PS, device m performs τ local iterations of SGD, with the following update during the i-th local iteration:

$$\boldsymbol{\theta}_m^{i+1}(t)=\boldsymbol{\theta}_m^{i}(t)-\eta_m^i(t)\nabla F_m\left(\boldsymbol{\theta}_m^{i}(t),\xi_m^i(t)\right),\ i\in[\tau], \tag{3}$$

where $\boldsymbol{\theta}_m^1(t)=\boldsymbol{\theta}(t)$, $\eta_m^i(t)$ 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 $\nabla F_m\left(\boldsymbol{\theta}_m^{i}(t),\xi_m^i(t)\right)$ provides an unbiased estimate of the actual gradient $\nabla F_m\left(\boldsymbol{\theta}_m^{i}(t)\right)$ with respect to the randomness of the stochastic gradient function; that is, $\forall i\in[\tau],\forall m\in[M],\forall t$,

$$\mathbb{E}_{\xi}\left[\nabla F_m\left(\boldsymbol{\theta}_m^{i}(t),\xi_m^i(t)\right)\right]=\nabla F_m\left(\boldsymbol{\theta}_m^{i}(t)\right). \tag{4}$$

After performing τ local updates, device m aims to send its local model update $\Delta\boldsymbol{\theta}_m(t)=\boldsymbol{\theta}_m^{\tau+1}(t)-\boldsymbol{\theta}(t)$ to the PS, $m\in[M]$. Ideally, having received the accurate local model updates from the devices, the PS updates the global model as

$$\boldsymbol{\theta}(t+1)=\boldsymbol{\theta}(t)+\Delta\boldsymbol{\theta}(t), \tag{5}$$

where we have defined

$$\Delta\boldsymbol{\theta}(t)\triangleq\frac{1}{M}\sum_{m=1}^{M}\Delta\boldsymbol{\theta}_m(t). \tag{6}$$

However, in our model, the devices transmit their local model updates over a wireless shared medium, which provides the PS with a noisy estimate of $\Delta\boldsymbol{\theta}(t)$. In the following, we describe the wireless channels from the devices to the PS, which is equipped with K antennas.

We model the shared wireless channel from the devices to the PS with K antennas as a wireless fading MAC, where OFDM is used to divide the available bandwidth into s subchannels, $s\leq d$ (in practice, we typically have $s\ll d$). We assume that N OFDM symbols can be transmitted over each subchannel at each global iteration. The received vector corresponding to the n-th OFDM symbol during global iteration t at the k-th antenna of the PS is given by

$$\boldsymbol{y}_k^n(t)=\sum_{m=1}^{M}\boldsymbol{h}_{m,k}^n(t)\circ\boldsymbol{x}_m^n(t)+\boldsymbol{z}_k^n(t),\quad k\in[K], \tag{7}$$

where $\boldsymbol{x}_m^n(t)$ is the n-th symbol of dimension s transmitted by device m, $\boldsymbol{h}_{m,k}^n(t) \in \mathbb{C}^s$ denotes the vector of channel gains from device m to the k-th PS antenna, $m \in [M]$, and $\boldsymbol{z}_k^n(t) \in \mathbb{C}^s$ represents the additive noise at the k-th antenna of the PS, $n \in [N]$. The i-th entry of channel vector $\boldsymbol{h}_{m,k}^n(t)$, denoted by $h_{m,k,i}^n(t)$, is distributed according to $\mathcal{CN}\left(0, \sigma_h^2\right)$, $i \in [s]$, and different entries of $\left[\boldsymbol{h}_{m,k}^1(t), ..., \boldsymbol{h}_{m,k}^N(t)\right]$ can be correlated, while the channel gains are assumed to be independent and identically distributed (iid) across PS antennas and wireless devices, $k \in [K]$, $n \in [N]$, $m \in [M]$. Similarly, different entries of $\left[\boldsymbol{z}_k^1(t), ..., \boldsymbol{z}_k^N(t)\right]$ can be correlated, and $z_{k,i}^n(t)$ denotes the i-th entry of $\boldsymbol{z}_k^n(t)$ and is distributed according to $\mathcal{CN}\left(0, \sigma_z^2\right)$, $i \in [s]$, $k \in [K]$, $n \in [N]$. Noise vectors are also assumed to be iid across PS antennas. We consider the following average power constraint at each device assuming a total of T global iterations:

$$\frac{1}{NT}\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\left[||\boldsymbol{x}_m^n(t)||_2^2\right] \leq \bar{P},\ \forall m \in [M], \tag{8}$$

where the expectation is taken with respect to the randomness of the communication channel.

We assume that the devices do not have CSI, and the PS has imperfect/noisy CSI about the wireless fading MAC. To be precise, we assume that the PS has only imperfect CSI about the sum of the channel gains from the devices to each PS antenna, i.e., $\sum_{m=1}^{M}\boldsymbol{h}_{m,k}^n(t)$, $\forall k \in [K]$, for $n \in [N]$ (please refer to [39] for details on channel estimation). We denote the CSI of $\sum_{m=1}^{M}\boldsymbol{h}_{m,k}^n(t)$ at the PS by $\widehat{\boldsymbol{h}}_k^n(t)$, where [40]

$$\widehat{\boldsymbol{h}}_k^n(t) = \sum_{m=1}^{M}\boldsymbol{h}_{m,k}^n(t) + \tilde{\boldsymbol{h}}_k^n(t),\ \forall n, k, t, \tag{9}$$

where $\tilde{\boldsymbol{h}}_k^n(t)$ represents the independent CSI estimation error vector with each entry an iid random variable with zero-mean and variance $\tilde{\sigma}_h^2$. At each global iteration, the goal at the PS is to estimate $\Delta\boldsymbol{\theta}(t)$, denoted by $\Delta\widehat{\boldsymbol{\theta}}\left(t\right)$, based on its received symbols $\boldsymbol{y}_k^n(t)$, and the CSI $\widehat{\boldsymbol{h}}_k^n(t)$, $\forall n, k$. The PS then updates the global model as

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) + \Delta\widehat{\boldsymbol{\theta}}\left(t\right), \tag{10}$$

and shares the new global model with the devices accurately.

Remark 1. *The CSI of $\sum_{m=1}^{M}\boldsymbol{h}_{m,k}^n(t)$, $\forall n, k$, dictates that the PS only needs to estimate the sum of the channel gains from all the devices to each antenna rather than each individual channel gain $\boldsymbol{h}_{m,k}^n(t)$. This significantly reduces the overhead of channel estimation, particularly for a relatively large number of devices M or number of PS antennas K, and this overhead does not increase with M.*

We note that the PS is interested in the average of the local model updates computed by the devices rather than each individual model update. Motivated by the additive nature of the wireless MAC, we consider an analog approach similarly to [10]–[12], [14], where the devices transmit their gradient estimates in an analog fashion without employing any channel coding.

III. Analog FEEL without CSIT

Next, we present the proposed analog FEEL scheme in the absence of CSIT at the devices. For the global model update at the PS, we first assume perfect CSI at the PS, in which case $\tilde{\sigma}_h^2 = 0$, and study the impact of the imperfect CSI at the PS in the next subsection.

At the global iteration t, device m aims to transmit its local update $\Delta\boldsymbol{\theta}_m\left(t\right) \in \mathbb{R}^d$ over $N = \lceil d/2s \rceil$ OFDM symbols across s subchannels in an uncoded manner, $m \in [M]$. We denote the i-th entry of $\Delta\boldsymbol{\theta}_m\left(t\right)$ by $\Delta\theta_{m,i}\left(t\right)$, $i \in [d]$, and define, for $n \in [N]$, $m \in [M]$,

$$\Delta\boldsymbol{\theta}_m^{n,\mathrm{re}}\left(t\right) \triangleq \left[\Delta\theta_{m,2(n-1)s+1}\left(t\right), \ldots, \Delta\theta_{m,(2n-1)s}\left(t\right)\right]^T, \tag{11a}$$

$$\Delta\boldsymbol{\theta}_m^{n,\mathrm{im}}\left(t\right) \triangleq \left[\Delta\theta_{m,(2n-1)s+1}\left(t\right), \ldots, \Delta\theta_{m,2ns}\left(t\right)\right]^T, \tag{11b}$$

$$\Delta\boldsymbol{\theta}_m^{n}\left(t\right) \triangleq \Delta\boldsymbol{\theta}_m^{n,\mathrm{re}}\left(t\right) + j\Delta\boldsymbol{\theta}_m^{n,\mathrm{im}}\left(t\right), \tag{11c}$$

where $j \triangleq \sqrt{-1}$, and we zero-pad $\Delta\boldsymbol{\theta}_m\left(t\right)$ to have length $2sN$. The i-th entry of $\Delta\boldsymbol{\theta}_m^n\left(t\right)$ is then given by, for $i \in [s]$, $n \in [N]$, $m \in [M]$,

$$\Delta\theta_{m,i}^n\left(t\right) = \Delta\theta_{m,2(n-1)s+i}\left(t\right) + j\Delta\theta_{m,(2n-1)s+i}\left(t\right). \tag{12}$$

According to (11), we have

$$\Delta\boldsymbol{\theta}_m\left(t\right) = \big[\Delta\boldsymbol{\theta}_m^{1,\mathrm{re}}\left(t\right), \Delta\boldsymbol{\theta}_m^{1,\mathrm{im}}\left(t\right), \ldots, \Delta\boldsymbol{\theta}_m^{N,\mathrm{re}}\left(t\right), \Delta\boldsymbol{\theta}_m^{N,\mathrm{im}}\left(t\right)\big]^T, \tag{13}$$

with $N = \lceil d/2s \rceil$. At the n-th OFDM symbol of iteration t, device m sends

$$\boldsymbol{x}_m^n(t) = \alpha_t\Delta\boldsymbol{\theta}_m^n(t),\ n \in [N], m \in [M], \tag{14}$$

where α_t is the scaling factor that will be chosen according to the power constraint. Accordingly, the average transmit power depends on α_t, and is evaluated as follows:

$$\frac{1}{NT}\sum_{t=1}^{T}\alpha_t^2\sum_{n=1}^{N}||\Delta\boldsymbol{\theta}_m^n(t)||_2^2 \leq \bar{P}. \tag{15}$$

The PS observes the following signal at its k-th antenna, for $k \in [K], n \in [N]$:

$$\boldsymbol{y}_k^n(t) = \alpha_t\sum_{m=1}^{M}\boldsymbol{h}_{m,k}^n(t) \circ \Delta\boldsymbol{\theta}_m^n(t) + \boldsymbol{z}_k^n(t). \tag{16}$$

A. Perfect CSI at the PS

In this subsection, we assume that the PS has access to perfect CSI, i.e., it has access to the sum of the channel gains from all the devices to each of its antennas, i.e., $\tilde{\sigma}_h^2 = 0$ and $\widehat{\boldsymbol{h}}_k^n(t) = \sum_{m=1}^{M}\boldsymbol{h}_{m,k}^n(t)$, $\forall n, k, t$. Having access to perfect CSI, the PS combines the signals at different antennas in the following form:

$$\boldsymbol{y}^n(t) \triangleq \frac{1}{K}\sum_{k=1}^{K}\Big(\sum_{m=1}^{M}\boldsymbol{h}_{m,k}^n(t)\Big)^* \circ \boldsymbol{y}_k^n(t), \tag{17}$$

whose i-th entry is given by

$$y_i^n(t) = \frac{1}{K}\sum_{k=1}^{K}\sum_{m=1}^{M}\left(h_{m,k,i}^n(t)\right)^* y_{k,i}^n(t), \tag{18}$$

where $y_{k,i}^n(t)$ denotes the i-th entry of $\boldsymbol{y}_k^n(t)$, $i \in [s]$, $n \in [N]$. We employed the maximal-ratio combining (MRC) detector in (17), which is a linear detector with a relatively small computational complexity and achieves the optimal performance when $K \to \infty$. We note that the computational complexity of the detector used in (17) is equivalent to a matrix-vector multiplication of dimensions $d \times K$ and K, respectively. By substituting $y_{k,i}^n(t)$, given in (16), it follows that

$$y_i^n(t) = \underbrace{\alpha_t \sum_{m=1}^{M} \Big(\frac{1}{K} \sum_{k=1}^{K} \left| h_{m,k,i}^n(t) \right|^2 \Big) \Delta\theta_{m,i}^n(t)}_{\text{signal term}} + \underbrace{\frac{\alpha_t}{K} \sum_{m=1}^{M} \sum_{m'=1, m' \neq m}^{M} \sum_{k=1}^{K} \left(h_{m,k,i}^n(t) \right)^* h_{m',k,i}^n(t) \Delta\theta_{m',i}^n(t)}_{\text{interference term}} + \underbrace{\frac{1}{K} \sum_{m=1}^{M} \sum_{k=1}^{K} \left(h_{m,k,i}^n(t) \right)^* z_{k,i}^n(t)}_{\text{noise term}}. \quad (19)$$

As we can see in (19), $y_i^n(t)$ consists of three terms, specified as the signal, interference, and noise terms, respectively. Following the law of large numbers, as the number of antennas at the PS $K \to \infty$, the signal term approaches

$$y_{i,\text{sig}}^n(t) \triangleq \alpha_t \sigma_h^2 \sum_{m=1}^{M} \Delta\theta_{m,i}^n(t), \ i \in [s], n \in [N], \quad (20)$$

from which the PS can recover

$$\frac{1}{M} \sum_{m=1}^{M} \Delta\theta_{m,2(n-1)s+i}(t) = \frac{\text{Re}\left\{ y_{i,\text{sig}}^n(t) \right\}}{\alpha_t M \sigma_h^2}, \quad (21a)$$

$$\frac{1}{M} \sum_{m=1}^{M} \Delta\theta_{m,(2n-1)s+i}(t) = \frac{\text{Im}\left\{ y_{i,\text{sig}}^n(t) \right\}}{\alpha_t M \sigma_h^2}. \quad (21b)$$

However, the interference term in (19) does not allow the exact recovery of $\frac{1}{M}\sum_{m=1}^{M} \Delta\theta_{m,2(n-1)s+i}(t)$ and $\frac{1}{M}\sum_{m=1}^{M} \Delta\theta_{m,(2n-1)s+i}(t)$ from $y_i^n(t)$, which is observed at the PS. To analyze the interference term, defined as $y_{i,\text{itf}}^n(t)$, we rewrite it as follows, for $i \in [s], n \in [N]$:

$$y_{i,\text{itf}}^n(t) = \alpha_t \sum_{m=1}^{M} \Big(\frac{1}{K} \sum_{k=1}^{K} h_{m,k,i}^n(t) \sum_{m'=1, m' \neq m}^{M} \left(h_{m',k,i}^n(t) \right)^* \Big) \Delta\theta_{m,i}^n(t). \quad (22)$$

We then define, for $m \in [M], i \in [s], n \in [N]$,

$$\mathfrak{h}_{m,i}^n(t) \triangleq \frac{1}{K} \sum_{k=1}^{K} h_{m,k,i}^n(t) \sum_{m'=1, m' \neq m}^{M} \left(h_{m',k,i}^n(t) \right)^*, \quad (23)$$

and is easy to verify that the mean and the variance of $\mathfrak{h}_{m,i}^n(t)$ are given by

$$\mathbb{E}\left[\mathfrak{h}_{m,i}^n(t) \right] = 0, \quad \mathbb{E}\left[\left| \mathfrak{h}_{m,i}^n(t) \right|^2 \right] = \frac{(M-1)\sigma_h^4}{K}, \quad (24)$$

respectively. We note that the local updates computed at each iteration are independent of the channel realizations experienced during the same iteration. From the analysis in (24), we conclude that the interference term in (19) has zero-mean and M terms, each with a variance that scales with $(M-1)/K$. Thus, for a fixed number of wireless devices M, the variance of the interference term in (19) approaches zero as $K \to \infty$. In practice, it is feasible to employ a sufficiently large number of antennas at the PS exploiting massive multiple-input multiple-output (MIMO) systems [41]. Numerical results with a finite number of antennas will be presented in Section V.

According to the above analysis, the PS generates the estimates, for $i \in [s]$, $n \in [N]$,

$$\Delta\hat{\theta}_{2(n-1)s+i}(t) = \frac{\text{Re}\left\{ y_i^n(t) \right\}}{\alpha_t M \sigma_h^2}, \quad (25a)$$

$$\Delta\hat{\theta}_{(2n-1)s+i}(t) = \frac{\text{Im}\left\{ y_i^n(t) \right\}}{\alpha_t M \sigma_h^2}. \quad (25b)$$

It then utilizes the estimated vector $\Delta\widehat{\boldsymbol{\theta}}(t) \triangleq \left[\Delta\hat{\theta}_1(t), \ldots, \Delta\hat{\theta}_d(t) \right]^T$, which is an unbiased estimate of the average of the local model updates, to update the global model as

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) + \Delta\widehat{\boldsymbol{\theta}}(t). \quad (26)$$

B. Imperfect CSI at the PS

We now generalize the above beamforming technique by considering imperfect CSI at the PS. Let $\widehat{\boldsymbol{h}}_k^n(t) = [\hat{h}_{k,1}^n(t), \ldots, \hat{h}_{k,s}^n(t)]^T$ and $\tilde{\boldsymbol{h}}_k^n(t) = [\tilde{h}_{k,1}^n(t), \ldots, \tilde{h}_{k,s}^n(t)]^T$. In the case of imperfect CSI at the PS, the received signals at different PS antennas are combined as follows:

$$\begin{aligned} \boldsymbol{y}^n(t) =& \frac{1}{K} \sum_{k=1}^{K} \left(\widehat{\boldsymbol{h}}_k^n(t) \right)^* \circ \boldsymbol{y}_k^n(t) \\ =& \frac{1}{K} \sum_{k=1}^{K} \Big(\sum_{m=1}^{M} \boldsymbol{h}_{m,k}^n(t) \Big)^* \circ \boldsymbol{y}_k^n(t) \\ &+ \frac{1}{K} \sum_{k=1}^{K} \left(\tilde{\boldsymbol{h}}_k^n(t) \right)^* \circ \boldsymbol{y}_k^n(t), \end{aligned} \quad (27)$$

which generalizes the expression for the perfect CSI case given in (17). Accordingly, we have

$$\begin{aligned} y_i^n(t) =& \alpha_t \sum_{m=1}^{M} \Big(\frac{1}{K} \sum_{k=1}^{K} \left| h_{m,k,i}^n(t) \right|^2 \Big) \Delta\theta_{m,i}^n(t) \\ &+ \frac{\alpha_t}{K} \sum_{m=1}^{M} \sum_{m'=1, m' \neq m}^{M} \sum_{k=1}^{K} \left(h_{m,k,i}^n(t) \right)^* h_{m',k,i}^n(t) \Delta\theta_{m',i}^n(t) \\ &+ \frac{1}{K} \sum_{m=1}^{M} \sum_{k=1}^{K} \left(h_{m,k,i}^n(t) \right)^* z_{k,i}^n(t) \\ &+ \frac{\alpha_t}{K} \sum_{m=1}^{M} \sum_{k=1}^{K} \left(\tilde{h}_{k,i}^n(t) \right)^* h_{m,k,i}^n(t) \Delta\theta_{m,i}^n(t) \\ &+ \frac{1}{K} \sum_{k=1}^{K} \left(\tilde{h}_{k,i}^n(t) \right)^* z_{k,i}^n(t), \end{aligned} \quad (28)$$

where the last two terms on the right hand side (RHS) are due to having imperfect CSI at the PS, and for $\tilde{\sigma}_h^2 = 0$ the above expression is equivalent to (19). We denote the extra interference term introduced because of the lack of perfect CSI at the PS by $\tilde{y}^n_{i,\text{itf}}(t)$ given by, for $i \in [s], n \in [N]$,

$$\tilde{y}^n_{i,\text{itf}}(t) = \alpha_t \sum_{m=1}^{M} \left(\frac{1}{K} \sum_{k=1}^{K} \left(\tilde{h}^n_{k,i}(t) \right)^* h^n_{m,k,i}(t) \right) \Delta\theta^n_{m,i}(t). \tag{29}$$

We define, for $m \in [M], i \in [s], n \in [N]$,

$$\tilde{\mathfrak{h}}^n_{m,i}(t) \triangleq \frac{1}{K} \sum_{k=1}^{K} \left(\tilde{h}^n_{k,i}(t) \right)^* h^n_{m,k,i}(t), \tag{30}$$

where we have

$$\mathbb{E}\left[\tilde{\mathfrak{h}}^n_{m,i}(t) \right] = 0, \quad \mathbb{E}\left[\left| \tilde{\mathfrak{h}}^n_{m,i}(t) \right|^2 \right] = \frac{\tilde{\sigma}_h^2 \sigma_h^2}{K}. \tag{31}$$

Therefore, lack of perfect CSI at the PS introduces an extra interference term with zero-mean which includes M terms, each with a variance scaled with $1/K$. Similarly to the perfect CSI scenario, the PS estimates $\frac{1}{M}\sum_{m=1}^{M} \Delta\theta_{m,2(n-1)s+i}(t)$ and $\frac{1}{M}\sum_{m=1}^{M} \Delta\theta_{m,(2n-1)s+i}(t)$, for $i \in [s]$, $n \in [N]$, through

$$\Delta\hat{\theta}_{2(n-1)s+i}(t) = \frac{\text{Re}\{y^n_i(t)\}}{\alpha_t M \sigma_h^2}, \tag{32a}$$

$$\Delta\hat{\theta}_{(2n-1)s+i}(t) = \frac{\text{Im}\{y^n_i(t)\}}{\alpha_t M \sigma_h^2}, \tag{32b}$$

and uses $\Delta\widehat{\boldsymbol{\theta}}(t)$ to update the global model as in (26).

Remark 2. *We note that with SGD the empirical variances of the local model updates decay over time and approach zero asymptotically [10], [42]–[45]. Thus, for robust communication of the local model updates against noise at each global iteration, it is reasonable to increase the power allocation factor α_t over time.*

Remark 3. *We remark that the main focus in this paper is to develop techniques for FEEL with no CSIT, as well as imperfect CSI at the PS. Our approach to tackle this problem is to employ multiple antennas at the PS, which can help to mitigate the effect of fading, and, in the limit, align the received signals at the PS. We can further employ some of the existing schemes in the literature providing more efficient communication over the limited bandwidth wireless MAC, such as the idea of linear projection proposed in [10]. We leave the analysis of such combined techniques as future work.*

IV. Convergence Analysis

In this section, we provide a convergence analysis of the proposed analog FEEL scheme with no CSIT and imperfect CSI at the PS. For ease of presentation, we consider $N = 1$, i.e., $s = d/2$, and drop the dependency of all the variables on n. Accordingly, the received signal at the PS, given in (28), can be rewritten as follows:

$$y_i(t) = \sum_{l=1}^{5} y_{i,l}(t), \quad \text{for } i \in [d/2], \tag{33a}$$

where

$$y_{i,1}(t) \triangleq \alpha_t \sum_{m=1}^{M} \left(\frac{1}{K} \sum_{k=1}^{K} |h_{m,k,i}(t)|^2 \right) \left(\Delta\theta_{m,i}(t) + j\Delta\theta_{m,d/2+i}(t) \right), \tag{33b}$$

$$y_{i,2}(t) \triangleq \frac{\alpha_t}{K} \sum_{m=1}^{M} \sum_{m'=1, m' \neq m}^{M} \sum_{k=1}^{K} \left(h_{m,k,i}(t) \right)^* h_{m',k,i}(t) \left(\Delta\theta_{m',i}(t) + j\Delta\theta_{m',d/2+i}(t) \right), \tag{33c}$$

$$y_{i,3}(t) \triangleq \frac{1}{K} \sum_{m=1}^{M} \sum_{k=1}^{K} \left(h_{m,k,i}(t) \right)^* z_{k,i}(t), \tag{33d}$$

$$y_{i,4}(t) \triangleq \frac{\alpha_t}{K} \sum_{m=1}^{M} \sum_{k=1}^{K} \left(\tilde{h}_{k,i}(t) \right)^* h_{m,k,i}(t) \left(\Delta\theta_{m,i}(t) + j\Delta\theta_{m,d/2+i}(t) \right), \tag{33e}$$

$$y_{i,5}(t) \triangleq \frac{1}{K} \sum_{k=1}^{K} \left(\tilde{h}_{k,i}(t) \right)^* z_{k,i}(t). \tag{33f}$$

We further define, for $l \in [5]$,

$$\Delta\hat{\theta}_{i,l}(t) \triangleq \begin{cases} \frac{\text{Re}\{y_{i,l}(t)\}}{\alpha_t M \sigma_h^2}, & \text{if } 1 \leq i \leq d/2, \\ \frac{\text{Im}\{y_{i-d/2,l}(t)\}}{\alpha_t M \sigma_h^2}, & \text{otherwise}, \end{cases} \tag{34}$$

according to which the estimate of the average local updates at the PS can be rewritten as $\Delta\hat{\theta}_i(t) = \sum_{l=1}^{5} \Delta\hat{\theta}_{i,l}(t)$, $i \in [d]$.

A. Preliminaries

Let the optimal solution minimizing $F(\boldsymbol{\theta})$ be defined as $\boldsymbol{\theta}^* \triangleq \arg\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$, and we denote the minimum value of the loss function by $F^* = F(\boldsymbol{\theta}^*)$. We also denote the minimum value of the local loss function F_m by F_m^*, for $m \in [M]$. We further define $\Gamma \triangleq F^* - \sum_{m=1}^{M} \frac{B_m}{B} F_m^*$, where we note that $\Gamma \geq 0$ captures the amount of bias in the data distribution. Γ increases with the heterogeneity of data across the devices.

We use the same learning rate across different devices and local iterations, but allow it to change over different global iterations; that is, we assume $\eta^i_m(t) = \eta(t)$, $\forall m, i$. Accordingly, we have, for $i \in [\tau]$, $m \in [M]$,

$$\boldsymbol{\theta}^{i+1}_m(t) = \boldsymbol{\theta}^i_m(t) - \eta(t) \nabla F_m \left(\boldsymbol{\theta}^i_m(t), \xi^i_m(t) \right), \tag{35}$$

and

$$\boldsymbol{\theta}^{i+1}_m(t) - \boldsymbol{\theta}^1_m(t) = -\eta(t) \sum_{l=1}^{i} \nabla F_m \left(\boldsymbol{\theta}^l_m(t), \xi^l_m(t) \right). \tag{36}$$

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

$$F_m(\boldsymbol{v}) - F_m(\boldsymbol{w}) \leq \langle \boldsymbol{v} - \boldsymbol{w}, \nabla F_m(\boldsymbol{w}) \rangle + \frac{L}{2} \|\boldsymbol{v} - \boldsymbol{w}\|_2^2. \tag{37}$$

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

$$F_m(\boldsymbol{v}) - F_m(\boldsymbol{w}) \geq \langle \boldsymbol{v} - \boldsymbol{w}, \nabla F_m(\boldsymbol{w}) \rangle + \frac{\mu}{2} \|\boldsymbol{v} - \boldsymbol{w}\|_2^2. \tag{38}$$

Assumption 3. *The expected squared l_2-norm of the stochastic gradients are bounded; that is, $\forall i \in [\tau], \forall m \in [M], \forall t$,*

$$\mathbb{E}_{\xi}\left[\left\|\nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right)\right\|_2^2\right] \leq G^2. \tag{39}$$

B. Convergence Rate

Here we provide the convergence rate of the proposed analog FEEL scheme. The proof is provided in Appendix A.

Theorem 1. *Let $0 < \eta(t) \leq \min\left\{1, \frac{1}{\mu\tau}\right\}$, $\forall t$. We have*

$$\mathbb{E}\left[\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*\|_2^2\right] \leq \Big(\prod_{i=0}^{t-1} A(i)\Big)\|\boldsymbol{\theta}(0) - \boldsymbol{\theta}^*\|_2^2 + \sum_{j=0}^{t-1} B(j) \prod_{i=j+1}^{t-1} A(i), \tag{40a}$$

where

$$A(i) \triangleq 1 - \mu\eta(i)\left(\tau - \eta(i)(\tau-1)\right), \tag{40b}$$

$$B(i) \triangleq \frac{\left(1 + \frac{\tilde{\sigma}_h^2}{M\sigma_h^2}\right)\eta^2(i)\tau^2G^2}{K} + \frac{\left(1 + \frac{\tilde{\sigma}_h^2}{M\sigma_h^2}\right)\sigma_z^2 d}{2\alpha_t^2 KM\sigma_h^2} + \left(1 + \mu(1-\eta(i))\right)\eta^2(i)G^2\frac{\tau(\tau-1)(2\tau-1)}{6} + (\tau^2 + \tau - 1)\eta^2(i)G^2 + 2\eta(i)(\tau-1)\Gamma, \tag{40c}$$

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

Proof. See Appendix A. □

We highlight that the term $\prod_{i=0}^{t-1} A(i)$ in the upper bound (40a) indicates the decay rate of the distance to the optimum solution $\boldsymbol{\theta}^*$ at time t with respect to the initial model. On the other hand, the term $\sum_{j=0}^{t-1} B(j)\prod_{i=j+1}^{t-1} A(i)$ measures the residual distance of the current model to the optimal solution with respect to that of the initial model to the optimal solution. One expects that reducing $A(i)$ would result in a faster convergence; however, it is important to investigate how a reduction in $A(i)$ impacts $B(i)$ and, consequently, the residual term. To be specific, it is easy to verify that $A(i)$ decreases with τ, while $B(i)$ increases. This confirms the intuition that increasing τ may lead to faster convergence, while it may also lead to results far from the optimal solution, particularly for non-iid data.

Corollary 1. *Let $0 < \eta(t) \leq \min\left\{1, \frac{1}{\mu\tau}\right\}$, $\forall t$. Given a total number of T global iterations, the L-smoothness of loss function $F(\cdot)$ results in*

$$\mathbb{E}\left[F(\boldsymbol{\theta}(T))\right] - F^* \leq \frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}(T) - \boldsymbol{\theta}^*\|_2^2\right] \leq \frac{L}{2}\Big(\prod_{i=0}^{T-1} A(i)\Big)\|\boldsymbol{\theta}(0) - \boldsymbol{\theta}^*\|_2^2 + \frac{L}{2}\sum_{j=0}^{T-1} B(j)\prod_{i=j+1}^{T-1} A(i), \tag{41}$$

where the last inequality follows from (40a)*.*

Remark 4. *The second term in $B(i)$, $\frac{\left(1+\tilde{\sigma}_h^2/(M\sigma_h^2)\right)\sigma_z^2 d}{2\alpha_t^2 KM\sigma_h^2}$, which is the result of the additive noise over the MAC, is not scaled with $\eta(i)$. Therefore, even for a decreasing learning rate $\eta(t)$, i.e., $\lim_{t\to\infty}\eta(t) = 0$, we have $\lim_{t\to\infty} B(t) = \frac{\left(1+\tilde{\sigma}_h^2/(M\sigma_h^2)\right)\sigma_z^2 d}{2\alpha_t^2 KM\sigma_h^2} \neq 0$, which shows that $\lim_{t\to\infty}\mathbb{E}\left[F(\boldsymbol{\theta}(t))\right] - F^* \neq 0$. However, we note that the destructive effect of this term in the convergence rate reduces with the number of PS antennas, K. We further remark that $\frac{\tilde{\sigma}_h^2\eta^2(i)\tau^2G^2}{\sigma_h^2 KM} + \frac{\tilde{\sigma}_h^2\sigma_z^2 d}{2\alpha_t^2 KM^2\sigma_h^4}$ captures the impact of the imperfect CSI at the PS, which also reduces with K.*

Corollary 2. *Consider a simplified setting $\eta(t) = \eta$, $\forall t$, and $\tau = 1$. Accordingly, Corollary 1 can be simplified as*

$$\mathbb{E}\left[F(\boldsymbol{\theta}(T))\right] - F^* \leq \frac{L}{2}(1-\mu\eta)^T\|\boldsymbol{\theta}(0) - \boldsymbol{\theta}^*\|_2^2 + \frac{L}{2\mu\eta}\Big(\Big(1 + \frac{\tilde{\sigma}_h^2}{M\sigma_h^2}\Big)\Big(\frac{\eta^2G^2}{K} + \frac{\sigma_z^2 d}{2\alpha_t^2 MK\sigma_h^2}\Big) + \eta^2G^2\Big)\left(1 - (1-\mu\eta)^T\right). \tag{42}$$

Remark 5. *The proposed approach can be readily extended to accommodate partial device participation at each iteration. During iteration t, the PS shares the model parameter $\boldsymbol{\theta}(t)$ with a subset of devices participating in the training, denoted by $\mathcal{M}(t)$ with $M(t) \triangleq |\mathcal{M}(t)|$, and beamforming at the PS is employed with respect to the signals received from $M(t)$ devices in $\mathcal{M}(t)$ instead of all M devices. The convergence analysis can also be adapted depending on the device scheduling policy, i.e., random or other device scheduling policies, through techniques introduced in the relevant papers [8], [31].*

V. Numerical Experiments

Here we evaluate the performance of the proposed analog FEEL algorithm with no CSI available at the wireless devices. We are particularly interested in investigating the impact of the number of PS antennas on the performance. We perform image classification on MNIST [46] and CIFAR-10 datasets [47] using ADAM optimizer [48]. We train different convolutional neural networks (CNNs) whose architectures are described in Table I. The performance is measured as the accuracy with respect to the test dataset, known as the *test accuracy*, versus the global iteration count, t.

We consider two data distribution scenarios across the devices. In the non-iid data distribution scenario, we split the training data samples with the same label/class to $M/10$ disjoint groups (assuming that M is divisible by 10). Thus, having 10 labels/classes for both MNIST and CIFAR-10 datasets, this results in M disjoint training datasets, each consisting of samples with the same label/class, and we assign each group to a distinct device. On the other hand, in the iid data distribution scenario, we randomly split the training dataset into M disjoint datasets, and assign each of them to a distinct device. We set the local mini-batch sample size to $\left|\xi_m^i(t)\right| = 500$, $\forall m, i, t$, for each experiment.

We consider $M = 20$ devices in the system. For simplicity, we assume that the s channel gains associated with each OFDM symbol from each device to each PS antenna are iid, and $\sigma_h^2 = 1$. For each experiment, we measure the test accuracy for $T = 400$ global iterations, and we set the power allocation factor at the devices to $\alpha_t = 1 + 10^{-3}t$, $t \in [T]$.

TABLE I: CNN architecture for image classification on MNIST and CIFAR-10.

MNIST	CIFAR-10
5×5 convolutional layer, 32 channels, ReLU activation, same padding	3×3 convolutional layer, 32 channels, ReLU activation, same padding
	3×3 convolutional layer, 32 channels, ReLU activation, same padding
	2×2 max pooling
2×2 max pooling	dropout with probability 0.2
	3×3 convolutional layer, 64 channels, ReLU activation, same padding
5×5 convolutional layer, 64 channels, ReLU activation, same padding	3×3 convolutional layer, 64 channels, ReLU activation, same padding
	2×2 max pooling
	dropout with probability 0.3
2×2 max pooling	3×3 convolutional layer, 128 channels, ReLU activation, same padding
	3×3 convolutional layer, 128 channels, ReLU activation, same padding
fully connected layer with 1024 units, ReLU activation, dropout with probability 0.2	2×2 max pooling
	dropout with probability 0.4
softmax output layer with 10 units	

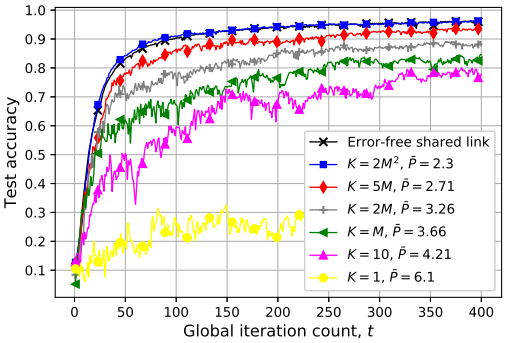
We further assume that $s = d/2$ resulting in $N = 1$. We note that, for a fixed power allocation factor α_t, $\forall t$, the value of s does not have any impact on the accuracy of the proposed analog FEEL scheme; instead, any change in s scales the average transmit power, whose value is proportional to N. We assume that the CSI estimation error at the PS, i.e., $\tilde{h}^n_{k,i}(t)$, is distributed according to $\mathcal{CN}\left(0, \tilde{\sigma}^2_h\right)$, $\forall k, i, n, t$.

For numerical comparison, we also consider a benchmark, in which the PS receives the average of the local model updates $\Delta\boldsymbol{\theta}(t) = \frac{1}{M}\sum_{m=1}^{M} \Delta\boldsymbol{\theta}_m(t)$ from the devices in an error-free manner, and updates the global model according to this noiseless observation. We refer to this as the *error-free shared link* scenario, and its accuracy can serve as an upper bound on the performance of the proposed analog FEEL scheme.

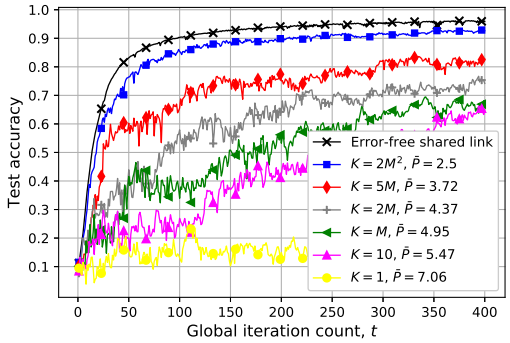
In Fig. 1 we illustrate the performance of the proposed analog FEEL scheme with no CSIT for increasing number of PS antennas, $K \in \{1, 10, M, 2M, 5M, 2M^2\}$, with non-iid MNIST data distributed across the devices, and number of local iterations $\tau = 3$. In Figs. 1a and 1b we assume perfect CSI at the PS, and investigate the performance for an increase in the noise variance from $\sigma_z^2 = 10$ to $\sigma_z^2 = 50$. We also include the performance of the error-free shared link scenario. As can be seen, both the final test accuracy and the convergence speed increases with the number of PS antennas, with the improvement significantly more noticeable when the noise level is higher. This is due to the fact that increasing K mitigates the effects of both the interference and noise terms, inferred from (19). Thus, the advantage of having more PS antennas is more pronounced when the channel is noisier. For example, for $\sigma_z^2 = 10$, the proposed scheme with $K = 2M^2$ antennas at the PS and average power $\bar{P} = 2.3$ performs as well as the error-free shared link scenario. On the other hand, further reducing the average signal-to-noise ratio $\bar{P}/\sigma_z^2$ by setting $\sigma_z^2 = 50$ results in a small performance gap between the error-free shared link scenario and the proposed scheme with $K = 2M^2$. These results illustrate the success of the proposed scheme with sufficient number of PS antennas in mitigating the noise term even when the average signal-to-noise ratio $\bar{P}/\sigma_z^2$ is as small as 0.05. Surprisingly, the accuracy improves drastically even with a few antennas at the PS, e.g., $K = 10$. We note that, with all the other parameters fixed, the required average transmit power reduces with K, which verifies a faster convergence rate with higher K resulting in a faster reduction in the empirical variances of the local model updates over time. The same observation is made by reducing σ_z^2 from 50 to 10 while all the other parameters are fixed.

Similar observations can be made in Figs. 1c and 1d considering imperfect CSI at the PS with $\tilde{\sigma}_h^2 = M\sigma_h^2/2$ and $\tilde{\sigma}_h^2 = M\sigma_h^2$, respectively. We observe the additional benefits of a large number of PS antennas in mitigating the adverse effects of imperfect CSI at the PS. Comparing the two figures, we can see that the benefits are more highlighted when the variance of the CSI estimation error is larger. Even when this variance is the same as that of the sum of channel gains from the devices, i.e., when $\tilde{\sigma}_h^2 = M\sigma_h^2$, the proposed scheme with a sufficient number of PS antennas performs almost as well as the error-free shared link scenario. Therefore, the proposed analog FEEL scheme can alleviate the negative effects of both the lack of CSIT and the imperfect CSI at the PS.

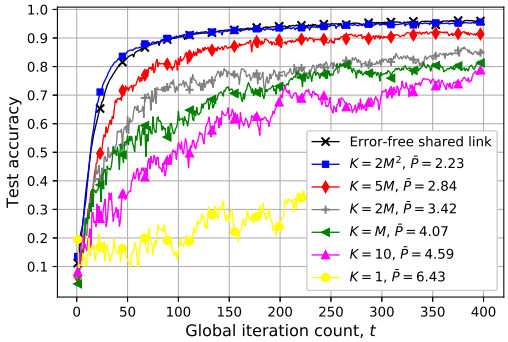
In Fig. 2, we investigate the performance of the proposed analog FEEL scheme for the more challenging CIFAR-10 dataset, distributed in an iid manner across the devices, considering different K values, $K \in \{M, 2M, 5M, 10M, 2M^2\}$, with $\tau = 5$ local iteration steps. Similarly to Fig. 1, we observe that the performance of the proposed scheme improves significantly with the number of PS antennas, and the improvement is more pronounced when the CSI is imperfect at the PS. In both cases under consideration, the proposed scheme with



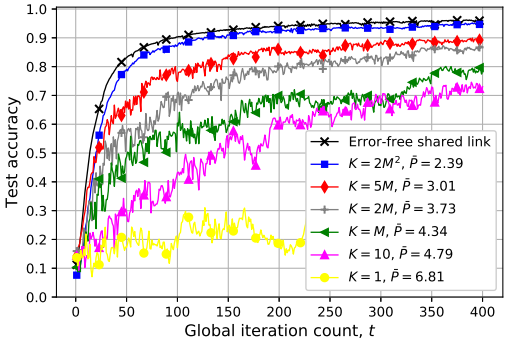
(a) Perfect CSI at PS, $\left(\sigma_z^2, \tilde{\sigma}_h^2\right) = (10, 0)$



(b) Perfect CSI at PS, $\left(\sigma_z^2, \tilde{\sigma}_h^2\right) = (50, 0)$



(c) Imperfect CSI at PS, $\left(\sigma_z^2, \tilde{\sigma}_h^2\right) = (10, M\sigma_h^2/2)$



(d) Imperfect CSI at PS, $\left(\sigma_z^2, \tilde{\sigma}_h^2\right) = (10, M\sigma_h^2)$

Fig. 1: Test accuracy of the proposed analog FEEL algorithm for non-iid MNIST dataset with different number of antennas $K \in \{1, 10, M, 2M, 5M, 2M^2\}$ for $M = 20$, $\sigma_h^2 = 1$ $\tau = 3$, and $\left|\xi_m^i(t)\right| = 500$, $\forall m, i, t$.

$K = 2M^2$ provides a performance as well as that of the error-free shared link scenario. However, the average required power in the experiments with CIFAR-10 dataset is higher than that for MNIST; this is mainly because of the larger network architecture required to reach reasonable accuracy levels for CIFAR-10, which leads to the gradients with higher norms, and consequently, resulting in higher empirical variance for the local model updates. Furthermore, the gaps between the performance of the proposed analog FEEL scheme for different K values are larger than those observed in Fig. 1, which indicates that the benefits of increasing K is even more when training larger models for more challenging learning tasks.

In Fig. 3, we illustrate the convergence rate of the proposed analog FEEL algorithm, presented in Corollary 1, for the setting considered in Fig. 2 for $K \in \{M, 2M, 5M, 10M, 2M^2\}$. The CNN for training on CIFAR-10, whose architecture is provided in Table I, has $d = 307498$ parameters, and we have $M = 20$ and $\sigma_z^2 = \sigma_h^2 = 1$. We set $\mu = 1$, $L = 5$, $G^2 = \Gamma = 1$, $\|\boldsymbol{\theta}(0) - \boldsymbol{\theta}^*\|_2^2 = 10^3$. We consider a decreasing learning rate $\eta(t) = \frac{1}{\mu\tau(10^{-4}t+1)}$, and $\alpha_t = 1+10^{-3}t$, $t \in [T]$. We also consider the convergence rate of the error-free shared link scenario given as follows:

$$\mathbb{E}\left[F(\boldsymbol{\theta}(T))\right] - F^* \leq \frac{L}{2}\Big(\prod_{i=0}^{T-1} A_{\text{ef}}(i)\Big)\left\|\boldsymbol{\theta}(0) - \boldsymbol{\theta}^*\right\|_2^2 + \frac{L}{2}\sum_{j=0}^{T-1} B_{\text{ef}}(j)\prod_{i=j+1}^{T-1} A_{\text{ef}}(i), \quad (43a)$$

where $0 < \eta(t) \leq \min\left\{1, \frac{1}{\mu\tau}\right\}$, and we have

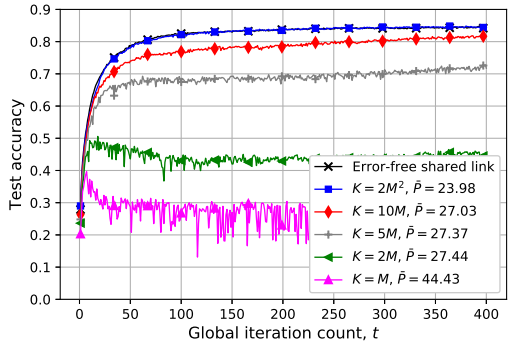
$$A_{\text{ef}}(i) \triangleq 1 - \mu\eta(i)\left(\tau - \eta(i)(\tau-1)\right), \quad (43b)$$

$$B_{\text{ef}}(i) \triangleq \left(1 + \mu(1-\eta(i))\right)\eta^2(i)G^2\frac{\tau(\tau-1)(2\tau-1)}{6} + (\tau^2+\tau-1)\eta^2(i)G^2 + 2\eta(i)(\tau-1)\Gamma, \quad (43c)$$

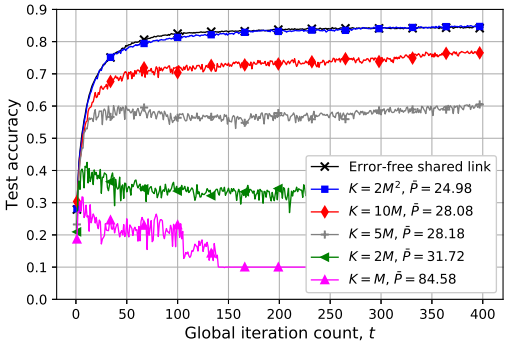
which can be obtained by following the procedure presented in the proof of Theorem 1. We investigate the convergence rate having perfect and imperfect CSI at the PS in Fig. 3a and Fig. 3b, respectively. We observe that the analytical results in Fig. 3 corroborate the experimental ones presented above, and the theoretical bound obtained for convex loss functions without CSIT approaches that of the perfect communication benchmark with the increasing number of PS antennas.

VI. Conclusions

We have studied FEEL, where colocated wireless devices collaboratively train a global model using their local datasets, and transmit their local updates to the PS over a wireless fading MAC. With the goal of recovering the average local model updates at the PS through over-the-air computation, we have considered analog transmission of the local updates from the devices over the wireless MAC. The current literature on over-the-air FEEL relies on perfect CSI both at the devices and the PS. However, acquiring perfect CSI in mobile wireless networks is typically not possible, and even imperfect CSI estimation can introduce delays and waste channel resources.

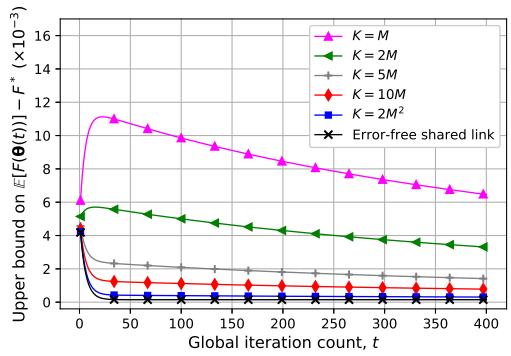


(a) Perfect CSI at PS, $\tilde{\sigma}_h^2 = 0$

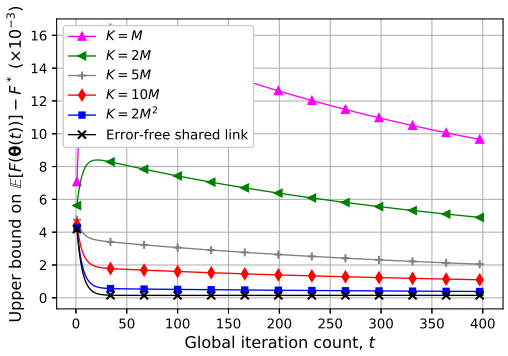


(b) Imperfect CSI at PS, $\tilde{\sigma}_h^2 = M\sigma_h^2/2$

Fig. 2: Test accuracy of the proposed analog FEEL algorithm for iid CIFAR-10 dataset with different number of antennas $K \in \{M, 2M, 5M, 10M, 2M^2\}$ for $M = 20$, $\sigma_h^2 = \sigma_z^2 = 1$, $\tau = 5$, and $\left|\xi_m^i(t)\right| = 500$, $\forall m, i, t$.



(a) Perfect CSI at PS, $\tilde{\sigma}_h^2 = 0$



(b) Imperfect CSI at PS, $\tilde{\sigma}_h^2 = M\sigma_h^2/2$

Fig. 3: Upper bound on $\mathbb{E}\left[F(\boldsymbol{\theta}(T))\right] - F^*$ for different number of antennas $K \in \{M, 2M, 5M, 10M, 2M^2\}$ with $d = 307498$ parameters, $M = 20$, $\sigma_z^2 = \sigma_h^2 = 1$, $\tau = 5$, $\mu = 1$, $L = 5$, $G^2 = \Gamma = 1$, $\|\boldsymbol{\theta}(0) - \boldsymbol{\theta}^*\|_2^2 = 10^3$, and $\eta(t) = \frac{1}{\mu\tau(10^{-4}t+1)}$.

Therefore, in this work, we have studied FEEL without any CSI at the devices and with imperfect CSI at the PS. To mitigate the effects of the time-varying channel without CSI, we assumed that the PS is equipped with multiple antennas, and designed a beamforming technique at the PS to estimate the computation result. We have derived the convergence rate of the proposed analog FEEL algorithm that highlights the impact of various system parameters on the performance. Experimental results on MNIST and CIFAR-10 datasets corroborated the theoretical convergence results, and illustrated that, with the proposed algorithm, increasing the number of PS antennas provides a better estimate of the average local model updates thanks to a better alignment of the desired signals, as well as the elimination of the interference and noise terms. Asymptotically, the proposed scheme guarantees that the wireless MAC becomes deterministic, despite the lack of CSIT and perfect CSI at the PS.

Appendix A
Proof of Theorem 1

We define an auxiliary variable $\boldsymbol{v}(t)$ given by

$$\boldsymbol{v}(t+1) \triangleq \boldsymbol{\theta}(t) + \Delta\boldsymbol{\theta}(t), \tag{44}$$

where $\Delta\boldsymbol{\theta}(t)$ is as defined in (6). We note that

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) + \Delta\widehat{\boldsymbol{\theta}}(t). \tag{45}$$

We have

$$\begin{aligned}\|\boldsymbol{\theta}(t+1) - \boldsymbol{\theta}^*\|_2^2 &= \|\boldsymbol{\theta}(t+1) - \boldsymbol{v}(t+1) + \boldsymbol{v}(t+1) - \boldsymbol{\theta}^*\|_2^2 \\ &= \|\boldsymbol{\theta}(t+1) - \boldsymbol{v}(t+1)\|_2^2 + \|\boldsymbol{v}(t+1) - \boldsymbol{\theta}^*\|_2^2 \\ &\quad + 2\langle\boldsymbol{\theta}(t+1) - \boldsymbol{v}(t+1), \boldsymbol{v}(t+1) - \boldsymbol{\theta}^*\rangle. \end{aligned} \tag{46}$$

Next, we bound the three terms on the RHS of (46).

Lemma 1. *We have*

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\theta}(t+1) - \boldsymbol{v}(t+1)\|_2^2\right] \leq &\frac{\left(1 + \frac{\tilde{\sigma}_h^2}{M\sigma_h^2}\right)\eta^2(t)\tau^2G^2}{K} \\ &+ \frac{\left(1 + \frac{\tilde{\sigma}_h^2}{M\sigma_h^2}\right)\sigma_z^2 d}{2\alpha_t^2 K M \sigma_h^2}. \end{aligned} \tag{47}$$

Proof. See Appendix B. □

Lemma 2. *We have*

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{v}(t+1) - \boldsymbol{\theta}^*\|_2^2\right] \leq &\\ &\left(1 - \mu\eta(t)\left(\tau - \eta(t)(\tau-1)\right)\right)\mathbb{E}\left[\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*\|_2^2\right] \\ &+ \left(1 + \mu(1-\eta(t))\right)\eta^2(t)G^2\frac{\tau(\tau-1)(2\tau-1)}{6} \\ &+ \eta^2(t)(\tau^2+\tau-1)G^2 + 2\eta(t)(\tau-1)\Gamma. \end{aligned} \tag{48}$$

Proof. See Appendix C. □

Lemma 3. *We have*

$$\mathbb{E}\big[\langle \boldsymbol{\theta}(t+1)-\boldsymbol{\upsilon}(t+1), \boldsymbol{\upsilon}(t+1)-\boldsymbol{\theta}^*\rangle\big]=0. \tag{49}$$

Proof. From the definition of $\boldsymbol{\upsilon}(t+1)$ in (44), it follows that

$$\mathbb{E}\big[\langle \boldsymbol{\theta}(t+1)-\boldsymbol{\upsilon}(t+1), \boldsymbol{\upsilon}(t+1)-\boldsymbol{\theta}^*\rangle\big] = \mathbb{E}\big[\langle \Delta\widehat{\boldsymbol{\theta}}(t)-\Delta\boldsymbol{\theta}(t), \boldsymbol{\theta}(t)+\Delta\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\rangle\big]. \tag{50}$$

From the independence of $h_{m,k,i}(t)$, $\tilde{h}_{m,k,i}(t)$, and $z_{k,i}(t)$, $\forall m\in[M],\forall k\in[K],\forall i\in[d]$, and (33) and (34), expectation of $\Delta\widehat{\boldsymbol{\theta}}(t)$ with respect to the channel gains and noise terms results in $\mathbb{E}\big[\Delta\widehat{\boldsymbol{\theta}}(t)\big]=\Delta\boldsymbol{\theta}(t)$. Since the local model updates at the global iteration t are independent of the channel characterizations during the same global iteration, it follows that

$$\mathbb{E}\big[\langle \Delta\widehat{\boldsymbol{\theta}}(t)-\Delta\boldsymbol{\theta}(t), \boldsymbol{\theta}(t)+\Delta\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\rangle\big]=0. \tag{51}$$

This completes the proof of Lemma 3. □

Substituting the results in Lemmas 1-3 into (46) yields

$$\begin{aligned}
&\|\boldsymbol{\theta}(t+1)-\boldsymbol{\theta}^*\|_2^2\\
&\le \left(1-\mu\eta(t)\left(\tau-\eta(t)(\tau-1)\right)\right)\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]\\
&\quad+\frac{\big(1+\frac{\tilde{\sigma}_h^2}{M\sigma_h^2}\big)\eta^2(t)\tau^2G^2}{K}+\frac{\big(1+\frac{\tilde{\sigma}_h^2}{M\sigma_h^2}\big)\sigma_z^2 d}{2\alpha_t^2KM\sigma_h^2}\\
&\quad+\left(1+\mu(1-\eta(t))\right)\eta^2(t)G^2\frac{\tau(\tau-1)(2\tau-1)}{6}\\
&\quad+\eta^2(t)(\tau^2+\tau-1)G^2+2\eta(t)(\tau-1)\Gamma,
\end{aligned} \tag{52}$$

and solving it recursively concludes Theorem 1.

Appendix B
Proof of Lemma 1

We have

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\theta}(t+1)-\boldsymbol{\upsilon}(t+1)\|_2^2\right]&=\mathbb{E}\left[\left\|\Delta\widehat{\boldsymbol{\theta}}(t)-\Delta\boldsymbol{\theta}(t)\right\|_2^2\right]\\
&=\sum_{i=1}^{d}\mathbb{E}\left[\left(\Delta\hat{\theta}_i(t)-\Delta\theta_i(t)\right)^2\right],
\end{aligned} \tag{53}$$

where $\Delta\theta_i(t)$ denotes the i-th entry of vector $\Delta\boldsymbol{\theta}(t)$, for $i\in[d]$. In the following, we bound $\mathbb{E}\left[\left(\Delta\hat{\theta}_i(t)-\Delta\theta_i(t)\right)^2\right]$, $\forall i$. Here we remind that $\Delta\hat{\theta}_i(t)=\sum_{l=1}^{5}\Delta\hat{\theta}_{i,l}(t)$, where $\Delta\hat{\theta}_{i,l}(t)$ is defined in (34). From the independence of $h_{m,k,i}(t)$, $\tilde{h}_{k,i}(t)$, and $z_{k,i}(t)$, $\forall m\in[M],\forall k\in[K],\forall i\in[d]$, and the fact that the local model updates at the global iteration t are independent of the channel realizations during the same global iteration, it is easy to verify that

$$\begin{aligned}
\mathbb{E}\left[\left(\Delta\hat{\theta}_i(t)-\Delta\theta_i(t)\right)^2\right]=&\mathbb{E}\left[\left(\Delta\hat{\theta}_{i,1}(t)-\Delta\theta_i(t)\right)^2\right]\\
&+\sum_{l=2}^{5}\mathbb{E}\left[\Delta\hat{\theta}_{i,l}^2(t)\right].
\end{aligned} \tag{54}$$

Lemma 4. *We have*

$$\sum_{i=1}^{d}\mathbb{E}\left[\left(\Delta\hat{\theta}_{i,1}(t)-\Delta\theta_i(t)\right)^2\right]=\frac{1}{KM^2}\sum_{m=1}^{M}\mathbb{E}\big[\,\|\Delta\boldsymbol{\theta}_m(t)\|_2^2\,\big]. \tag{55}$$

Proof. According to definition of $\Delta\hat{\theta}_{i,1}(t)$ in (34), we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\Delta\hat{\theta}_{i,1}(t)-\Delta\theta_i(t)\right)^2\right]\\
&=\mathbb{E}\Big[\Big(\frac{1}{M}\sum_{m=1}^{M}\Big(\frac{1}{K\sigma_h^2}\sum_{k=1}^{K}|h_{m,k,i}(t)|^2-1\Big)\Delta\theta_{m,i}(t)\Big)^2\Big]\\
&\overset{(a)}{=}\mathbb{E}\Big[\frac{1}{M^2}\sum_{m=1}^{M}\Big(-\frac{1}{K}+\frac{1}{K^2\sigma_h^4}\sum_{k=1}^{K}|h_{m,k,i}(t)|^4\Big)\Delta\theta_{m,i}^2(t)\Big]\\
&\overset{(b)}{=}\mathbb{E}\Big[\frac{1}{KM^2}\sum_{m=1}^{M}\Delta\theta_{m,i}^2(t)\Big],
\end{aligned} \tag{56}$$

where (a) follows from the independence of $h_{m,k,i}(t)$, $\forall m,k$, and (b) follows since $\mathbb{E}\big[\,|h_{m,k,i}(t)|^2\,\big]=\sigma_h^2$ and $\mathbb{E}\big[\,|h_{m,k,i}(t)|^4\,\big]=2\sigma_h^4$. Lemma 4 follows from (56). □

Lemma 5. *We have*

$$\sum_{i=1}^{d}\mathbb{E}\left[\Delta\hat{\theta}_{i,2}^2(t)\right]=\frac{(M-1)}{KM^2}\sum_{m=1}^{M}\mathbb{E}\big[\,\|\Delta\boldsymbol{\theta}_m(t)\|_2^2\,\big]. \tag{57}$$

Proof. We first consider $1\le i\le d/2$. By substituting $\Delta\hat{\theta}_{i,2}(t)$ from (34), it follows that

$$\begin{aligned}
&\mathbb{E}\left[\Delta\hat{\theta}_{i,2}^2(t)\right]\\
&=\mathbb{E}\Big[\Big(\frac{1}{KM\sigma_h^2}\sum_{m=1}^{M}\sum_{m'=1,m'\neq m}^{M}\sum_{k=1}^{K}\mathrm{Re}\left\{(h_{m,k,i}(t))^*h_{m',k,i}(t)\right.\\
&\qquad\left.\left(\Delta\theta_{m',i}(t)+j\Delta\theta_{m',d/2+i}(t)\right)\right\}\Big)^2\Big]\\
&\overset{(a)}{=}\mathbb{E}\Big[\frac{1}{K^2M^2\sigma_h^4}\sum_{m=1}^{M}\sum_{m'=1,m'\neq m}^{M}\sum_{k=1}^{K}\\
&\Big(\left(\mathrm{Re}\left\{(h_{m,k,i}(t))^*h_{m',k,i}(t)\left(\Delta\theta_{m',i}(t)+j\Delta\theta_{m',d/2+i}(t)\right)\right\}\right)^2\\
&+\mathrm{Re}\left\{(h_{m,k,i}(t))^*h_{m',k,i}(t)\left(\Delta\theta_{m',i}(t)+j\Delta\theta_{m',d/2+i}(t)\right)\right\}\\
&\mathrm{Re}\left\{(h_{m',k,i}(t))^*h_{m,k,i}(t)\left(\Delta\theta_{m,i}(t)+j\Delta\theta_{m,d/2+i}(t)\right)\right\}\Big)\Big]\\
&\overset{(b)}{=}\mathbb{E}\Big[\frac{1}{2KM^2}\sum_{m=1}^{M}\Big((M-1)\left(\Delta\theta_{m,i}^2(t)+\Delta\theta_{m,d/2+i}^2(t)\right)\\
&+\sum_{m'=1,m'\neq m}^{M}(\Delta\theta_{m,i}(t)\Delta\theta_{m',i}(t)\\
&\qquad-\Delta\theta_{m,d/2+i}(t)\Delta\theta_{m',d/2+i}(t))\Big)\Big],
\end{aligned} \tag{58}$$

where (a) and (b) follow from the independence of $h_{m,k,i}(t)$, $\forall m,k$, and $\mathbb{E}\big[\,|h_{m,k,i}(t)|^2\,\big]=\sigma_h^2$, respectively. Similarly, for $d/2+1\le i\le d$, it follows that

$$\begin{aligned}
&\mathbb{E}\left[\Delta\hat{\theta}_{i,2}^2(t)\right]\\
&=\mathbb{E}\Big[\Big(\frac{1}{KM\sigma_h^2}\sum_{m=1}^{M}\sum_{m'=1,m'\neq m}^{M}\sum_{k=1}^{K}\\
&\mathrm{Im}\left\{(h_{m,k,i}(t))^*h_{m',k,i}(t)\left(\Delta\theta_{m',i}(t)+j\Delta\theta_{m',d/2+i}(t)\right)\right\}\Big)^2\Big]\\
&=\mathbb{E}\Big[\frac{1}{2KM^2}\sum_{m=1}^{M}\Big((M-1)\left(\Delta\theta_{m,i-d/2}^2(t)+\Delta\theta_{m,i}^2(t)\right)
\end{aligned}$$

$$+\sum_{m'=1,m'\neq m}^{M}(\Delta\theta_{m,i}(t)\Delta\theta_{m',i}(t) -\Delta\theta_{m,i-d/2}(t)\Delta\theta_{m',i-d/2}(t))\Big)\Big]. \quad (59)$$

□

From (58) and (59), it follows that

$$\sum_{i=1}^{d}\mathbb{E}\left[\Delta\hat{\theta}_{i,2}^2(t)\right] = \mathbb{E}\Big[\frac{(M-1)}{KM^2}\sum_{m=1}^{M}\sum_{i=1}^{d/2}\Big(\Delta\theta_{m,i}^2(t)+\Delta\theta_{m,d/2+i}^2(t)\Big)\Big] = \frac{(M-1)}{KM^2}\sum_{m=1}^{M}\mathbb{E}\big[\,\|\Delta\boldsymbol{\theta}_m(t)\|_2^2\,\big]. \quad (60)$$

Lemma 6. *We have*

$$\sum_{i=1}^{d}\mathbb{E}\left[\Delta\hat{\theta}_{i,3}^2(t)\right] = \frac{\sigma_z^2 d}{2\alpha_t^2 KM\sigma_h^2}. \quad (61)$$

Proof. According to the definition of $\Delta\hat{\theta}_{i,3}(t)$, given in (34), for $1\leq i\leq d/2$ we have

$$\mathbb{E}\left[\Delta\hat{\theta}_{i,3}^2(t)\right] = \mathbb{E}\Big[\Big(\frac{1}{\alpha_t KM\sigma_h^2}\sum_{m=1}^{M}\sum_{k=1}^{K}\mathrm{Re}\left\{(h_{m,k,i}(t))^*\,z_{k,i}(t)\right\}\Big)^2\Big] \overset{(a)}{=} \mathbb{E}\Big[\frac{1}{\alpha_t^2K^2M^2\sigma_h^4}\sum_{m=1}^{M}\sum_{k=1}^{K}\left(\mathrm{Re}\left\{(h_{m,k,i}(t))^*\,z_{k,i}(t)\right\}\right)^2\Big] \overset{(b)}{=} \frac{\sigma_z^2}{2\alpha_t^2KM\sigma_h^2}, \quad (62)$$

where (a) follows from the independence of $h_{m,k,i}(t)$ and $z_{k,i}(t)$, $\forall m,k$, and (b) follows since $\mathbb{E}\big[\,|h_{m,k,i}(t)|^2\,\big]=\sigma_h^2$ and $\mathbb{E}\big[\,|z_{k,i}(t)|^2\,\big]=\sigma_z^2$. The same result can be obtained for $d/2+1\leq i\leq d$ by following the same procedure as above. It is straightforward to derive (61) from (62). □

Lemma 7. *We have*

$$\sum_{i=1}^{d}\mathbb{E}\left[\Delta\hat{\theta}_{i,4}^2(t)\right] = \frac{\tilde{\sigma}_h^2}{KM^2\sigma_h^2}\sum_{m=1}^{M}\mathbb{E}\big[\,\|\Delta\boldsymbol{\theta}_m(t)\|_2^2\,\big]. \quad (63)$$

Proof. By Substituting $\Delta\hat{\theta}_{i,4}(t)$ from (34), for $1\leq i\leq d/2$, we have

$$\mathbb{E}\left[\Delta\hat{\theta}_{i,4}^2(t)\right] = \mathbb{E}\Big[\Big(\frac{1}{KM\sigma_h^2}\sum_{m=1}^{M}\sum_{k=1}^{K}\mathrm{Re}\big\{\big(\tilde{h}_{k,i}(t)\big)^*h_{m,k,i}(t)\big(\Delta\theta_{m,i}(t)+j\Delta\theta_{m,d/2+i}(t)\big)\big\}\Big)^2\Big] \overset{(a)}{=} \mathbb{E}\Big[\frac{1}{K^2M^2\sigma_h^4}\sum_{m=1}^{M}\sum_{k=1}^{K}\Big(\mathrm{Re}\big\{\big(\tilde{h}_{k,i}(t)\big)^*h_{m,k,i}(t)\big(\Delta\theta_{m,i}(t)+j\Delta\theta_{m,d/2+i}(t)\big)\big\}\Big)^2\Big] \overset{(b)}{=} \mathbb{E}\Big[\frac{\tilde{\sigma}_h^2}{2KM^2\sigma_h^2}\sum_{m=1}^{M}\Big(\Delta\theta_{m,i}^2(t)+\Delta\theta_{m,d/2+i}^2(t)\Big)\Big], \quad (64)$$

where (a) follows from the independence of $\tilde{h}_{k,i}(t)$ and $h_{m,k,i}(t)$, $\forall m,k$, and (b) is the result of $\mathbb{E}\big[\big|\tilde{h}_{k,i}(t)\big|^2\big]=\tilde{\sigma}_h^2$ and $\mathbb{E}\big[\,|h_{m,k,i}(t)|^2\,\big]=\sigma_h^2$, $\forall i$. Similarly, for $d/2+1\leq i\leq d$, we can obtain

$$\mathbb{E}\left[\Delta\hat{\theta}_{i,4}^2(t)\right] = \mathbb{E}\Big[\frac{\tilde{\sigma}_h^2}{2KM^2\sigma_h^2}\sum_{m=1}^{M}\Big(\Delta\theta_{m,i-d/2}^2(t)+\Delta\theta_{m,d/2}^2(t)\Big)\Big]. \quad (65)$$

From (64) and (65), we have

$$\sum_{i=1}^{d}\mathbb{E}\left[\Delta\hat{\theta}_{i,4}^2(t)\right] = \mathbb{E}\Big[\frac{\tilde{\sigma}_h^2}{KM^2\sigma_h^2}\sum_{m=1}^{M}\sum_{i=1}^{d/2}\Big(\Delta\theta_{m,i}^2(t)+\Delta\theta_{m,d/2+i}^2(t)\Big)\Big] = \frac{\tilde{\sigma}_h^2}{KM^2\sigma_h^2}\sum_{m=1}^{M}\mathbb{E}\big[\,\|\Delta\boldsymbol{\theta}_m(t)\|_2^2\,\big]. \quad (66)$$

□

Lemma 8. *We have*

$$\sum_{i=1}^{d}\mathbb{E}\left[\Delta\hat{\theta}_{i,5}^2(t)\right] = \frac{\tilde{\sigma}_h^2\sigma_z^2 d}{2\alpha_t^2KM^2\sigma_h^4}. \quad (67)$$

Proof. From the definition of $\Delta\hat{\theta}_{i,5}(t)$, given in (34), for $1\leq i\leq d/2$ we have

$$\mathbb{E}\left[\Delta\hat{\theta}_{i,5}^2(t)\right] = \mathbb{E}\Big[\Big(\frac{1}{\alpha_t KM\sigma_h^2}\sum_{k=1}^{K}\mathrm{Re}\big\{\big(\tilde{h}_{k,i}(t)\big)^*z_{k,i}(t)\big\}\Big)^2\Big] \overset{(a)}{=} \mathbb{E}\Big[\frac{1}{\alpha_t^2K^2M^2\sigma_h^4}\sum_{k=1}^{K}\Big(\mathrm{Re}\big\{\big(\tilde{h}_{k,i}(t)\big)^*z_{k,i}(t)\big\}\Big)^2\Big] \overset{(b)}{=} \frac{\tilde{\sigma}_h^2\sigma_z^2}{2\alpha_t^2KM^2\sigma_h^4}, \quad (68)$$

where (a) follows from the independence of $\tilde{h}_{k,i}(t)$ and $z_{k,i}(t)$, $\forall k$, and (b) follows since $\mathbb{E}\big[\big|\tilde{h}_{k,i}(t)\big|^2\big]=\tilde{\sigma}_h^2$ and $\mathbb{E}\big[\,|z_{k,i}(t)|^2\,\big]=\sigma_z^2$. The same result can be obtained for $d/2+1\leq i\leq d$ by following the same procedure as above. The proof of Lemma 8 is completed from (68). □

By substituting the results of Lemmas 4-8 into (53), it follows that

$$\mathbb{E}\left[\|\boldsymbol{\theta}(t+1)-\boldsymbol{v}(t+1)\|_2^2\right] = \frac{\big(1+\frac{\tilde{\sigma}_h^2}{M\sigma_h^2}\big)}{KM}\sum_{m=1}^{M}\mathbb{E}\big[\,\|\Delta\boldsymbol{\theta}_m(t)\|_2^2\,\big]+\frac{\big(1+\frac{\tilde{\sigma}_h^2}{M\sigma_h^2}\big)\sigma_z^2 d}{2\alpha_t^2KM\sigma_h^2} \overset{(a)}{=} \frac{\Big(1+\frac{\tilde{\sigma}_h^2}{M\sigma_h^2}\Big)\eta^2(t)}{KM}\sum_{m=1}^{M}\mathbb{E}\Big[\Big\|\sum_{l=1}^{\tau}\nabla F_m\left(\boldsymbol{\theta}_m^l(t),\xi_m^l(t)\right)\Big\|_2^2\Big] + \frac{\big(1+\frac{\tilde{\sigma}_h^2}{M\sigma_h^2}\big)\sigma_z^2 d}{2\alpha_t^2KM\sigma_h^2}$$

$$\overset{\text{(b)}}{\leq} \frac{\left(1+\frac{\tilde{\sigma}_h^2}{M\sigma_h^2}\right)\eta^2(t)\tau}{KM}\sum_{m=1}^{M}\sum_{l=1}^{\tau}\mathbb{E}\Big[\Big\|\nabla F_m\left(\boldsymbol{\theta}_m^l(t),\xi_m^l(t)\right)\Big\|_2^2\Big] + \frac{\left(1+\frac{\tilde{\sigma}_h^2}{M\sigma_h^2}\right)\sigma_z^2 d}{2\alpha_t^2 KM\sigma_h^2}$$
$$\overset{\text{(c)}}{\leq} \frac{\left(1+\frac{\tilde{\sigma}_h^2}{M\sigma_h^2}\right)\eta^2(t)\tau^2 G^2}{K} + \frac{\left(1+\frac{\tilde{\sigma}_h^2}{M\sigma_h^2}\right)\sigma_z^2 d}{2\alpha_t^2 KM\sigma_h^2}, \tag{69}$$

where (a) follows by replacing $\Delta\boldsymbol{\theta}_m(t)$ from (36), (b) is due to the convexity of $\|\cdot\|_2^2$, and (c) follows from Assumption 3.

Appendix C
Proof of Lemma 2

We follow the same procedure as the one used to prove [31, Lemma 3]. We have

$$\mathbb{E}\left[\|\boldsymbol{\upsilon}(t+1)-\boldsymbol{\theta}^*\|_2^2\right] = \mathbb{E}\left[\|\boldsymbol{\theta}(t)+\Delta\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]$$
$$= \mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right] + \mathbb{E}\left[\|\Delta\boldsymbol{\theta}(t)\|_2^2\right] + 2\mathbb{E}\left[\langle\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*,\Delta\boldsymbol{\theta}(t)\rangle\right]. \tag{70}$$

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

$$\mathbb{E}\left[\|\Delta\boldsymbol{\theta}(t)\|_2^2\right] \leq \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\|\Delta\boldsymbol{\theta}_m(t)\|_2^2\right]$$
$$\overset{\text{(a)}}{=} \frac{\eta^2(t)}{M}\sum_{m=1}^{M}\mathbb{E}\Big[\Big\|\sum_{i=1}^{\tau}\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\Big\|_2^2\Big]$$
$$\leq \frac{\eta^2(t)\tau}{M}\sum_{m=1}^{M}\sum_{i=1}^{\tau}\mathbb{E}\left[\left\|\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\right\|_2^2\right]$$
$$\overset{\text{(b)}}{\leq} \eta^2(t)\tau^2 G^2, \tag{71}$$

where (a) follows by replacing $\Delta\boldsymbol{\theta}_m(t)$ from (36), and (b) follows from Assumption 3. Plugging the above inequality into (70) yields

$$\mathbb{E}\left[\|\boldsymbol{\upsilon}(t+1)-\boldsymbol{\theta}^*\|_2^2\right] \leq \mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right] + \eta^2(t)\tau^2G^2 + 2\mathbb{E}\left[\langle\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*,\Delta\boldsymbol{\theta}(t)\rangle\right]. \tag{72}$$

We bound the last term on the RHS of the above inequality. We have

$$2\mathbb{E}\left[\langle\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*,\Delta\boldsymbol{\theta}(t)\rangle\right] \overset{\text{(a)}}{=} \frac{2}{M}\sum_{m=1}^{M}\mathbb{E}\left[\langle\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*,\Delta\boldsymbol{\theta}_m(t)\rangle\right]$$
$$= \frac{2\eta(t)}{M}\sum_{m=1}^{M}\mathbb{E}\Big[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}(t),\sum_{i=1}^{\tau}\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\Big]$$
$$= \frac{2\eta(t)}{M}\sum_{m=1}^{M}\mathbb{E}\left[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}(t),\nabla F_m\left(\boldsymbol{\theta}(t),\xi_m^1(t)\right)\rangle\right] + \frac{2\eta(t)}{M}\sum_{m=1}^{M}\mathbb{E}\Big[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}(t),\sum_{i=2}^{\tau}\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\Big]. \tag{73}$$

Next we bound the two terms on the RHS of the above equality. We have

$$\frac{2\eta(t)}{M}\sum_{m=1}^{M}\mathbb{E}\left[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}(t),\nabla F_m\left(\boldsymbol{\theta}(t),\xi_m^1(t)\right)\rangle\right]$$
$$\overset{\text{(a)}}{=} \frac{2\eta(t)}{M}\sum_{m=1}^{M}\mathbb{E}\left[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}(t),\nabla F_m\left(\boldsymbol{\theta}(t)\right)\rangle\right]$$
$$\overset{\text{(b)}}{\leq} \frac{2\eta(t)}{M}\sum_{m=1}^{M}\mathbb{E}\big[F_m(\boldsymbol{\theta}^*)-F_m(\boldsymbol{\theta}(t))-\frac{\mu}{2}\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\big]$$
$$= 2\eta(t)\big(F^*-\mathbb{E}\left[F(\boldsymbol{\theta}(t))\right]-\frac{\mu}{2}\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]\big)$$
$$\overset{\text{(c)}}{\leq} -\mu\eta(t)\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right], \tag{74}$$

where (a) follows since $\mathbb{E}_\xi\left[\nabla F_m\left(\boldsymbol{\theta}(t),\xi_m^1(t)\right)\right] = \nabla F_m\left(\boldsymbol{\theta}(t)\right)$, (b) follows since F_m is μ-strongly convex, and (c) holds because $F^* \leq F(\boldsymbol{\theta}(t))$. For the second term on the RHS of (73), we have

$$\frac{2\eta(t)}{M}\sum_{m=1}^{M}\mathbb{E}\Big[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}(t),\sum_{i=2}^{\tau}\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\Big]$$
$$= \frac{2\eta(t)}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\left[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}(t),\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\right]$$
$$= \frac{2\eta(t)}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\left[\langle\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}(t),\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\right] + \frac{2\eta(t)}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\left[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}_m^i(t),\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\right]. \tag{75}$$

From Cauchy-Schwarz inequality, it follows that

$$\frac{2\eta(t)}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\left[\langle\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}(t),\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\right]$$
$$\leq \frac{\eta(t)}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\Big[\frac{1}{\eta(t)}\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}(t)\right\|_2^2 + \eta(t)\left\|\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\right\|_2^2\Big]$$
$$\overset{\text{(a)}}{\leq} \frac{1}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\big[\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}(t)\right\|_2^2\big] + \eta^2(t)\left(\tau-1\right)G^2, \tag{76}$$

where (a) follows from Assumption 3. Also, the following lemma presents an upper bound on the second term in the RHS of (75).

Lemma 9. *The second term on the RHS of* (75) *is upper bounded as follows:*

$$\frac{2\eta(t)}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\left[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}_m^i(t),\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\right]$$
$$\leq -\mu\eta(t)(1-\eta(t))(\tau-1)\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right] + \frac{\mu(1-\eta(t))}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\left[\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}(t)\right\|_2^2\right] + 2\eta(t)(\tau-1)\Gamma. \tag{77}$$

Proof. See Appendix D. □

Substituting the results in (76) and (77) into (75) yields

$$\begin{aligned}
&\frac{2\eta(t)}{M}\sum_{m=1}^{M}\mathbb{E}\Big[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}(t),\sum_{i=2}^{\tau}\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\Big]\\
&\leq -\mu\eta(t)(1-\eta(t))(\tau-1)\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]\\
&\quad+\frac{(1+\mu(1-\eta(t)))}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\big[\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}(t)\right\|_2^2\big]\\
&\quad+\eta^2(t)\left(\tau-1\right)+2\eta(t)(\tau-1)\Gamma. \qquad (78)
\end{aligned}$$

We have

$$\begin{aligned}
&\frac{1}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\big[\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}(t)\right\|_2^2\big]\\
&=\frac{\eta^2(t)}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\Big[\Big\|\sum\nolimits_{l=1}^{i}\nabla F_m\left(\boldsymbol{\theta}_m^l(t),\xi_m^l(t)\right)\Big\|_2^2\Big]\\
&\overset{(a)}{\leq}\eta^2(t)G^2\frac{\tau(\tau-1)(2\tau-1)}{6}, \qquad (79)
\end{aligned}$$

where (a) follows from the convexity of $\|\cdot\|_2^2$ and Assumption 3. For $\eta(t)\leq 1$, $\forall t$, it follows from (78) and (79) that

$$\begin{aligned}
&\frac{2\eta(t)}{M}\sum_{m=1}^{M}\mathbb{E}\Big[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}(t),\sum_{i=2}^{\tau}\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\Big]\\
&\leq -\mu\eta(t)(1-\eta(t))(\tau-1)\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]\\
&\quad+\left(1+\mu(1-\eta(t))\right)\eta^2(t)G^2\frac{\tau(\tau-1)(2\tau-1)}{6}\\
&\quad+\eta^2(t)\left(\tau-1\right)G^2+2\eta(t)(\tau-1)\Gamma. \qquad (80)
\end{aligned}$$

By substituting the results in (74) and (80) into (73), we obtain

$$\begin{aligned}
&2\mathbb{E}\left[\langle\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*,\Delta\boldsymbol{\theta}(t)\rangle\right]\\
&\leq -\mu\eta(t)\left(\tau-\eta(t)(\tau-1)\right)\mathbb{E}\left[\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\right]\\
&\quad+\left(1+\mu(1-\eta(t))\right)\eta^2(t)G^2\frac{\tau(\tau-1)(2\tau-1)}{6}\\
&\quad+\eta^2(t)\left(\tau-1\right)G^2+2\eta(t)(\tau-1)\Gamma. \qquad (81)
\end{aligned}$$

Plugging (81) into (72) completes the proof of Lemma 2.

Appendix D
Proof of Lemma 9

We have

$$\begin{aligned}
&\frac{2\eta(t)}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\left[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}_m^i(t),\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)\rangle\right]\\
&\overset{(a)}{\leq}\frac{2\eta(t)}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\left[\langle\boldsymbol{\theta}^*-\boldsymbol{\theta}_m^i(t),\nabla F_m\left(\boldsymbol{\theta}_m^i(t)\right)\rangle\right]\\
&\overset{(b)}{\leq}\frac{2\eta(t)}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\Big[F_m(\boldsymbol{\theta}^*)-F_m(\boldsymbol{\theta}_m^i(t))\\
&\qquad\qquad-\frac{\mu}{2}\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}^*\right\|_2^2\Big]\\
&=\frac{2\eta(t)}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\Big[F_m(\boldsymbol{\theta}^*)-F_m^*+F_m^*-F_m(\boldsymbol{\theta}_m^i(t))\\
&\qquad\qquad-\frac{\mu}{2}\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}^*\right\|_2^2\Big]\\
&=2\eta(t)(\tau-1)\big(F^*-\frac{1}{M}\sum\nolimits_{m=1}^{M}F_m^*\big)\\
&\quad+\frac{2\eta(t)}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\left(F_m^*-\mathbb{E}\left[F_m(\boldsymbol{\theta}_m^i(t))\right]\right)\\
&\quad-\frac{\mu\eta(t)}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\left[\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}^*\right\|_2^2\right]\\
&\overset{(c)}{\leq}2\eta(t)(\tau-1)\Gamma-\frac{\mu\eta(t)}{M}\sum_{m=1}^{M}\sum_{i=2}^{\tau}\mathbb{E}\left[\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}^*\right\|_2^2\right], \qquad (82)
\end{aligned}$$

where (a) follows since $\mathbb{E}_\xi\left[\nabla F_m\left(\boldsymbol{\theta}(t),\xi_m^i(t)\right)\right]=\nabla F_m\left(\boldsymbol{\theta}(t)\right)$, $\forall i,m,t$, (b) holds because F_m is μ-strongly convex, and (c) follows since $F_m^*\leq F_m(\boldsymbol{\theta}_m^i(t))$, $\forall m,i,t$. We have

$$\begin{aligned}
&-\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}^*\right\|_2^2=-\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}(t)\right\|_2^2-\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\\
&\quad-2\langle\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}(t),\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\rangle\\
&\overset{(a)}{\leq}-\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}(t)\right\|_2^2-\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\\
&\quad+\frac{1}{\eta(t)}\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}(t)\right\|_2^2+\eta(t)\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2\\
&=-(1-\eta(t))\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\|_2^2+\Big(\frac{1}{\eta(t)}-1\Big)\left\|\boldsymbol{\theta}_m^i(t)-\boldsymbol{\theta}(t)\right\|_2^2, \qquad (83)
\end{aligned}$$

where (a) follows from the Cauchy-Schwarz inequality. The proof of Lemma 9 is completed by substituting the result in (83) into (82).

References

[1] 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 (GlobalSIP)*, Ottawa, ON, Canada, Nov. 2019, pp. 1–5.

[2] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5g and beyond," *IEEE J. Sel. Areas Commun.*, Early Access, Sep. 2020.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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," *IEEE Trans. Wireless Commun.*, Early Access, Jun. 2020.

[16] Y.-S. Jeon, M. M. Amiri, J. Li, and H. V. Poor, "A compressive sensing approach for federated learning over massive MIMO communication systems," *IEEE Trans. Wireless Commun.*, Early Access, Dec. 2020.

[17] 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.

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

[19] 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.

[20] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *arXiv:2006.05459 [cs.IT]*, Jun. 2020.

[21] 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.

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

[23] 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.

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

[25] 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.

[26] 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.

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

[28] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, "Energy-efficient resource management for federated edge learning with CPU-GPU heterogeneous computing," *arXiv:2007.07122 [cs.IT]*, Jul. 2020.

[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," *IEEE Trans. Wireless Commun.*, Early Access, Jan. 2021.

[32] ——, "Convergence of federated learning over a noisy downlink," *arXiv:2008.11141 [cs.IT]*, Aug. 2020.

[33] W. Shi, S. Zhou, Z. Niu, M. Jiang, and L. Geng, "Joint device scheduling and resource allocation for latency constrained wireless federated learning," *IEEE Trans. Wireless Commun.*, Early Access, Sep. 2020.

[34] D. Gündüz et al., "Communicate to learn at the edge," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 14–19, Dec. 2020.

[35] H. Guo, A. Liu, and V. K. N. Lau, "Analog gradient aggregation for federated learning over wireless networks: Customized design and convergence analysis," *IEEE Internet of Things Journal*, Early Access, Jun. 2020.

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

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

[38] S. Gunnarsson, J. Flordelis, L. V. D. Perre, and F. Tufvesson, "Channel hardening in massive MIMO: Model parameters and experimental assessment," *IEEE Open Journal of the Commun. Society*, vol. 1, pp. 501–512, Apr. 2020.

[39] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[40] T. Weber, A. Sklavos, and M. Meurer, "Imperfect channel-state information in MIMO transmission," *IEEE Trans. Commun.*, vol. 54, no. 3, pp. 543–552, Mar. 2006.

[41] F. Rusek et al., "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[42] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, 2010, pp. 177–187.

[43] N. Strom, "Scalable distributed DNN training using commodity gpu cloud computing," in *INTERSPEECH*, 2015.

[44] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," in *ICML*, Jul. 2015.

[45] T. Lin, S. U. Stich, and M. Jaggi, "Don't use large mini-batches, use local SGD," *arXiv:1808.07217v3 [cs.LG]*, Oct. 2018.

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

[47] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," in *Technical Report, University of Toronto*, 2009.

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

Mohammad Mohammadi Amiri (S'16) received the B.Sc. and M.Sc. degrees in Electrical Engineering from Iran University of Science and Technology in 2011 and University of Tehran in 2014, respectively, both with highest rank in classes. He also obtained a Ph.D. degree at Imperial College London in 2019. He is currently a Postdoctoral Research Associate in the Department of Electrical Engineering at Princeton University. His research interests include information and coding theory, machine learning, wireless communications, and signal processing.

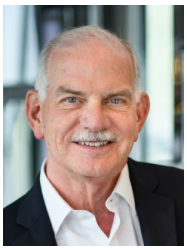
Tolga M. Duman (S'95–M'98–SM'03–F'11) received the B.S. degree from Bilkent University, Ankara, Turkey, in 1993, and the M.S. and Ph.D. degrees from Northeastern University, Boston, MA, USA, in 1995 and 1998, respectively, all in electrical engineering. He is currently a Professor with the Electrical and Electronics Engineering Department, Bilkent University. Prior to joining Bilkent University in 2012, he was a Professor with the School of ECEE at Arizona State University.

Dr. Duman's current research interests are in systems, with particular focus on communication and signal processing, including wireless and mobile communications, coding/modulation, coding for wireless communications, machine learning for communications, and distributed computing. He was a recipient of the National Science Foundation CAREER Award, the IEEE Third Millennium Medal, and a Distinguished Teaching Award from Bilkent University. Dr. Duman has served as editor and area editor for different journals and as technical program committee member for various conferences. He is currently the Editor-in-Chief of IEEE Transactions on Communications.

Deniz Gündüz (S'03-M'08-SM'13) received the B.S. degree in electrical and electronics engineering from METU, Turkey in 2002, and the M.S. and Ph.D. degrees in electrical engineering from NYU Tandon School of Engineering (formerly Polytechnic University) in 2004 and 2007, respectively. After his PhD, he served as a postdoctoral research associate at Princeton University, as a consulting assistant professor at Stanford University, and as a research associate at CTTC in Barcelona, Spain. ln Sep. 2012, he joined the Electrical and Electronic Engineering Department of Imperial College London, UK, where he is currently a Professor of Information Processing, and serves as the deputy head of the Intelligent Systems and Networks Group. He is also a part-time faculty member at the University of Modena and Reggio Emilia, Italy, and has held visiting positions at University of Padova (2018-2020) and Princeton University (2009-2012).

His research interests lie in the areas of communications and information theory, machine learning, and privacy. Dr. Gündüz is a Distinguished Lecturer for the IEEE Information Theory Society (2020-22). He is an Area Editor for the IEEE Transactions on Communications and the IEEE Journal on Selected Areas in Communications (JSAC). He also serves as an Editor of the IEEE Transactions on Wireless Communications. He is an organiser and co-chair of the 2021 London Symposium on Information Theory. He is the recipient of the IEEE Communications Society - Communication Theory Technical Committee (CTTC) Early Achievement Award in 2017, a Starting Grant of the European Research Council (ERC) in 2016, and several best paper awards.

Sanjeev R. Kulkarni (M'91, SM'96, F'04) received degrees in mathematics and electrical engineering from Clarkson University, the M.S. degree in E.E. from Stanford University, and the Ph.D. in E.E. from M.I.T. in 1991. From 1985 to 1991 he was a Member of the Technical Staff at M.I.T. Lincoln Laboratory. Since 1991 he has been with Princeton University, where he is currently the William R. Kenan, Jr., Professor of Electrical Engineering and the Dean of the Faculty. He also served as the Dean of the Graduate School at Princeton University from 2014-2017. Professor Kulkarni is an affiliated faculty member in the Department of Operations Research and Financial Engineering and the Department of Philosophy. His research interests include statistical pattern recognition, nonparametric estimation, learning and adaptive systems, information theory, and wireless networks.



H. Vincent Poor (S'72, M'77, SM'82, F'87) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990 he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor. During 2006 to 2016, he served as the dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory, machine learning and network science, and their applications in wireless networks, energy systems and related fields. Among his publications in these areas is the forthcoming book *Machine Learning and Wireless Communications*. (Cambridge University Press, 2021).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences and is a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal and a D.Eng. *honoris causa* from the University of Waterloo awarded in 2019.