Towards Understanding Federated Learning over Unreliable Networks

Chenyuan Feng, *Member, IEEE*, Ahmed Arafa, *Member, IEEE*, Zihan Chen, *Student Member, IEEE*, Mingxiong Zhao, *Member, IEEE*, Tony Q. S. Quek, *Fellow, IEEE*, and Howard H. Yang, *Member, IEEE*

***Abstract*—This paper studies the efficiency of training a statistical model among an edge server and multiple clients via Federated Learning (FL) – a machine learning method that preserves data privacy in the training process – over wireless networks. Due to unreliable wireless channels and constrained communication resources, the server can only choose a handful of clients for parameter updates during each communication round. To address this issue, analytical expressions are derived to characterize the FL convergence rate, accounting for key features from both communication and algorithmic aspects, including transmission reliability, scheduling policies, and momentum method. First, the analysis reveals that either delicately designed user scheduling policies or expanding higher bandwidth to accommodate more clients in each communication round can expedite model training in networks with reliable connections. However, these methods become ineffective when the connection is erratic. Second, it has been verified that incorporating the momentum method into the model training algorithm accelerates the rate of convergence and provides greater resilience against transmission failures. Last, extensive empirical simulations are provided to verify these theoretical discoveries and enhancements in performance.**

***Index Terms*—Federated learning, transmission failure, scheduling policy, momentum method, convergence analysis.**

I. Introduction

A de facto paradigm shift in machine learning models is being brought about by the surge in the processing capacity of terminal devices and the growing concern about data privacy. Complex computations previously exclusive to the cloud center are now shifting to the periphery of networks. The Federated Learning (FL) scheme is the result of the fusion of edge computing systems and artificial intelligence. It enables a swarm of terminal devices, i.e., the clients, and a global computing unit, i.e., the edge server, to collaboratively train a statistical model using datasets that are stored on the clients' devices while maintaining data privacy [2]–[8].

C. Feng is with the Department of Communication, EURECOM, 06410 Biot Sophia Antipolis Cedex, France (email: Chenyuan.Feng@eurecom.fr).

A. Arafa is with the Department of Electrical and Computer Engineering, University of North Carolina at Charlotte, NC 28223, USA (email: aarafa@uncc.edu).

Z. Chen and T. Q. S. Quek are with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore 487372 (e-mail: zihan_chen@mymail.sutd.edu.sg, tonyquek@sutd.edu.sg).

M. Zhao is with Engineering Research Center of Cyberspace, National Pilot School of Software, Yunnan University (email: mx_zhao@ynu.edu.cn).

H. H. Yang is with the ZJU-UIUC Institute, Zhejiang University, Haining 314400, China, the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310007, China, and the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA (email: haoyang@intl.zju.edu.cn).

Part of the work was presented at the IEEE 2021 Computing, Communications and IoT Applications (ComComAp) [1].

The corresponding author of this paper is H. H. Yang.

A. Related Works

FL brings the statistical models directly to the clients for local computing, in contrast to conventional machine learning approaches that aggregate all the data to a computing center for training. Here, only the obtained parameters are uploaded to the server for improvements to the global model, and the updated global model is fed back to the clients for another round of local training [2]. Such interactions between the server and clients will repeat for a sufficient number of rounds, after which the global model converges, and all the entities that participated in the training process can benefit from a better machine learning result. In light of this, FL highlights its trials of offering greater levels of privacy while significantly lowering communication overheads; this is especially pertinent to next-generation mobile networks [9]. Therefore, since the advent of this algorithm, it has attracted considerable attention from academia and industry alike. Nonetheless, the ultimate implementation of the FL system necessitates addressing novel problems that fundamentally diverge from the usual approaches developed for traditional machine learning environments [10]–[15].

Specifically, within the framework of FL, clients typically possess highly customized datasets, leading to a non-independent and identically distributed (i.i.d.) distribution of statistical data across the devices. System heterogeneity is also a result of the fact that various customers inside the network may differ significantly in terms of system attributes, such as processing power and/or connection quality. Heterogeneity is a feature that causes sluggish and even unstable convergence, and in order to solve this problem, new training techniques are required. Following this line, it is demonstrated that adding a proximal term to the global objective function can significantly improve the stability, as well as the overall accuracy, of the FL system [16]. Moreover, by means of variance reduction, a variate control scheme was proposed to rectify the drift in the local updates of clients so as to align the global gradient towards the optimal point and achieve faster convergence of the FL training [17]. Recognizing that historical parameters also contain useful information, it is suggested to either directly reuse the outdated gradients from clients in the global aggregation process [18] or leverage them to construct momentum terms [19], [20] so as to reduce the communication rounds in the process of model training. Aside from (stochastic) gradient descent, which is a

first-order training method, second-order-like training schemes such as the Newton method are also introduced to boost the convergence rate [21]. These works, while bringing about remarkable gains, have chiefly focused on the algorithmic perspective with little attention to the impacts of the communication aspects.

As wireless networks are expected to be the main FL deployment scenario, effective communication is also one of the main drivers of the system's implementation. In contrast to reliable connections offered by wired cables in a data center, the spectrum is an unstable medium where communication quality fluctuates over time. consequently, not every client can connect to the server with reliability after every global parameter aggregation. Furthermore, because spectral resources are typically scarce, the server can only choose a subset of clients to upload parameters for each communication round [22]. Joint optimization client scheduling and resource allocation has been studied as an enhancement of hierarchical federated edge learning [23]. Additionally, communications across the spectrum are frequently orders of magnitude slower than those using a chip. A series of recent research have been carried out [24]–[29] to address the communication bottleneck in the FL training. In particular, [24] examined the effects of three traditional client scheduling policies on the convergence property of FL systems and developed a theoretical framework to account for the communication conditions—quantified by the transmission success probability—in the convergence rate.

Acknowledging that the probability of transmission success differs for various clients and that the current scheduling techniques may result in a biased trained model, [25] proposed a scheduling policy that optimizes convergence rate by striking a balance between statistical bias and channel quality. Moreover, it is demonstrated that scheduling policy and resource allocation may be jointly designed to accelerate the training process, depending on the staleness [28], [29] or the importance [26], [27] of the customers' parameters. While these studies have alleviated the communication problems in FL, further research is needed to understand how channel quality, scheduling policies, and algorithmic improvements interact. Moreover, many previous works concentrated on strongly convex loss functions, which is inappropriate for the context of many popular machine learning models such as neural networks.

To study the convergence rate with partial client participation and non-i.i.d. datasets, [30] showed that linear speedup for convergence of FL is achievable and revealed that a large number of local training epochs can accelerate the convergence. To eliminate the bias caused by partial participants, [31] and [32] modified the model aggregation rule in FL to avoid waiting for straggling clients, where the server would re-use the memorized latest updates as the surrogate of the non-participating clients during each communication round. [33] derived convergence upper bounds for a wide range of non-stochastic and stochastic participation patterns, including regularized, ergodic, stationary, and strongly mixing (e.g., Markov process) and independent patterns. In contrast to the above-mentioned FL models, where the server and the clients are tightly coupled, [34] proposed a new paradigm in FL called Anarchic Federated Learning (AFL), where flexible client participation is allowed with cross-device and cross-silo settings. The author also provided convergence analysis and proved that the highly desirable linear speedup effect could be attained. The authors in [35] leveraged the concept of variance reduction from stochastic optimization. They proposed a novel bilayer FL algorithm to achieve a fast convergence rate in the setting where each client has an arbitrary probability of participating in each iteration.

B. Research Objectives and Contributions

In this paper, we aim to develop an analytical framework to study the impacts of different parameters, including both communication and algorithmic aspects, on the FL convergence rate. Specifically, we consider a network that consists of one server and multiple clients, connected to the server via wireless links. The task for the server and clients is to collaboratively learn a statistical model from the datasets residing on the clients' devices while preserving their data privacy, which is accomplished by means of federated computing. The server sends the objective function along with the model parameters to the clients, makes them train for a certain amount of time using their local datasets, and uploads only the resultant parameters, with which the server can improve the global model and feed it back to the clients for another round of local training.

During this process, owing to the time-varying nature of wireless channels, only a subset of the clients can establish reliable connections to the server upon each global aggregation round. Moreover, due to the scarcity of spectral resources, the server can only select a handful of clients in each communication round to participate in the FL training. In this respect, we investigate the efficacy of two scheduling policies, namely Random Scheduling (RS) and Age-Based Scheduling (ABS), in selecting the clients. To further accelerate the model training process, we adopt the momentum method in conjunction with the global aggregation on the server side. By invoking the tools from optimization theory, we derive analytical expressions to characterize the FL convergence rate in a general setting that accounts for the effects of channel quality, scheduling policies, and momentum method. The analysis allows us to grasp crisp insights into the impacts of different network parameters on the convergence performance of FL and obtain useful design guidelines. The results are expected to propel our understanding of FL and guide researchers in further research pursuits.

In our previous work [1], we investigated the convergence rate of partial client participation under the RS scheme in unreliable transmission networks. In this work, we extend our analysis to the ABS scheme by deriving the convergence analysis under FL and FML and conducting extensive experiments in both i.i.d. and non-i.i.d. datasets. The main contributions of this paper are summarized below.

- We develop a theoretical framework for understanding the convergence performance of FL algorithms run on wireless networks. Particularly, we derive analytical expressions for the convergence rates of FL under different

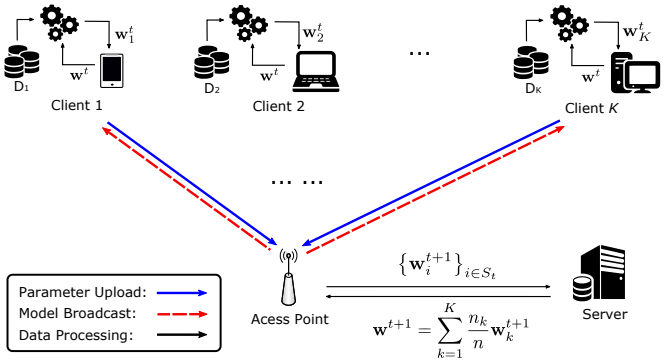


Fig. 1. An illustration of the FL process: (A) clients conduct local training based on their own dataset, (B) the server aggregates the received updates to improve the global model, (C) the new model is sent back to a subset of clients, and the process is repeated.

settings that encompass key features such as channel quality, spectral resources, scheduling policies, and integration with momentum methods, delivering a more comprehensive analysis. Different from [24]–[27], we do not assume convexity of the empirical loss function, making the results applicable to more general machine learning models.

- Based on the analysis, we find that when communication channels are reliable, the FL convergence rate can be boosted by (1) expanding the communication bandwidth to engage more clients in each communication round and (2) adopting a scheduling policy that reduces parameter staleness such as the ABS. On the other hand, when the communication channels are highly unreliable, the aforementioned approaches are not instrumental to enhancing the FL training efficiency.
- We also confirm, via analysis and simulation, that integrating momentum with the global aggregation speeds up the FL convergence rate and enhances its resilience against communication failures. It also implies that reliable communications shall be devised to keep pace with the growth of edge learning.
- We examine the convergence performance of the depicted FL system via extensive simulation experiments based on MNIST and CIFAR-10 datasets with different machine learning models such as Multi-Layer Perceptron (MLP) and Convolutional Neural Network (CNN). We also perform experimental comparisons between the FL and Federated Momentum Learning (FML). The simulation results validate the convergence of FML as well as its effectiveness in accelerating the convergence rate.

The remainder of this paper is organized as follows. We introduce the system model in Section II. In Section III, we analyze the convergence rate of FL training in wireless networks under both RS and ABS policies. We also present the convergence rate of running FL in tandem with momentum. Then, we show the simulation results in Section IV to compare the FL convergence performance amongst different circumstances and obtain the subsequent design insights. We conclude the paper in Section V.

Algorithm 1 Federated Learning Algorithm

Parameters: H = number of local steps per computation round, η = step size for stochastic gradient descent.
Initialize: $\boldsymbol{w}_0 \in \mathbb{R}^d$
for $t = 0, 1, 2, ..., T-1$ **do**
  The server selects a set S_t of at most N clients and broadcasts the global parameter $\boldsymbol{w}_t$ to them
  for each client $k \in S_t$ in parallel **do**
    Initialize $\boldsymbol{w}_{t,0}^{(k)} = \boldsymbol{w}_t$
    for s = 0 to $H-1$ **do**
      Sample $\xi_{k,s} \in \mathcal{D}_k$ uniformly at random, and update the local parameter $\boldsymbol{w}_t^{(k)}$ as follows:

$$\boldsymbol{w}_{t,s+1}^{(k)} = \boldsymbol{w}_{t,s}^{(k)} - \eta \nabla f_k(\boldsymbol{w}_{t,s}^{(k)}; \xi_{k,s}) \tag{4}$$

      in which ∇ represents the gradient operation
    end for
    Send the locally aggregated stochastic gradients $\sum_{s=0}^{H-1} \nabla f_k(\boldsymbol{w}_{t,s}^{(k)}; \xi_{k,s})$ to the server
  end for
  The server collects all the gradient parameters from the selected clients and assigns $\boldsymbol{g}_t^{(i)} = \sum_{s=0}^{H-1} \nabla f_i(\boldsymbol{w}_{t,s}^{(i)}; \xi_{k,s})$ for $i \in S_t$. Moreover, the server sets $\boldsymbol{g}_t^{(j)} = \boldsymbol{g}_{t-1}^{(j)}$ for $j \neq S_t$, and then updates the global parameter $\boldsymbol{w}_{t+1}$ as follows:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \sum_{k=1}^{K} p_k \boldsymbol{g}_t^{(k)} \tag{5}$$

end for
Output: $\boldsymbol{w}_T$

II. System Model

Let us consider the FL system depicted in Fig. 1, consisting of one server and K clients, where K is a large number. Each client k has a local dataset $\mathcal{D}_k = \{\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}\}_{i=1}^{n_k}$ of size $|\mathcal{D}_k| = n_k$. We assume the local datasets are statistically independent across the clients. The goal of the server is to learn a statistical model over the datasets residing on all the clients without sacrificing their privacy. More precisely, the server needs to fit a vector $\boldsymbol{w} \in \mathbb{R}^d$, commonly known as the model parameter, to minimize the following loss function without the explicit knowledge of $\mathcal{D} = \cup_{k=1}^{K} \mathcal{D}_k$:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{w}; \boldsymbol{x}_i, y_i) = \sum_{k=1}^{K} p_k f_k(\boldsymbol{w}) \tag{1}$$

where $n = \sum_{k=1}^{K} n_k$, $\ell(\cdot)$ is the loss function defined under some particular task, $p_k = n_k/n$, and $f_k(\boldsymbol{w})$ denotes the local empirical loss function of client k, given by

$$f_k(\boldsymbol{w}) = \frac{1}{n_k} \sum_{j=1}^{n_k} \ell(\boldsymbol{w}; \boldsymbol{x}_j, y_j). \tag{2}$$

We further define the optimal solution to (1) as

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} f(\boldsymbol{w}). \tag{3}$$

Since the server cannot access the individual client's datasets, the model training needs to be carried out by the server and the clients in an FL fashion [18]. The training procedure is detailed in Algorithm 1.

We consider the communications between the server and the clients to be conducted over a resource-limited spectrum

with time-varying channel gains. Specifically, we consider the maximum number of available channels for the parameter transmissions to be $N \ll K$. We assume that every client is able to establish a reliable connection to the server with probability p upon each global aggregation,[1] and that these connections vary independently across the communication rounds. We further assume that the server can obtain the information about the reliability of the clients' communication links at the beginning of each global aggregation, namely, the server has full knowledge about whether a typical client is connected to it or not before the global aggregation starts. The server, therefore, needs to select a subset S_t out of the available clients to participate in the FL training, where $|S_t| \leq N \ll K$. In this work, we consider two types of scheduling policies:

A. Random Scheduling (RS)

Under this policy, the server randomly samples a subset S_t out of the clients with reliable channels. Because the maximum number of available sub-channels is N, the scheduling of clients can encounter two different situations: (a) if the number of clients that have reliable channels is less than N, all of them will be selected for parameter update; and (b) when the number of clients with reliable sub-channels is larger than N, only N out of them will be selected (uniformly at random).

B. Age-Based Scheduling (ABS)

This approach aims to reduce the staleness in the clients' parameters during the training process. In order to quantify the staleness of each update, we leverage the information freshness and define a metric termed Age-of-Update (AoU) [28]. For a generic client k, its AoU evolves as follows:

$$A_k[t+1] = (A_k[t]+1)(1-S_k[t]), \quad S_k[t] \in \{0,1\} \tag{6}$$

where $A_k[0] = 0$, and $S_k[t]$ takes the value 1 if client k is selected by the server for update during communication round t and 0 otherwise.[2] By leveraging this metric, the scheduling policy is given by: (a) if the number of clients that have reliable channels is less than N, all of them will be selected for parameter update; and (b) otherwise, select the N clients with the highest AoU values, namely

$$\mathbf{S}^*[t] = \arg\max_{\mathbf{S}\subset\{1,2,...,K\}} \{A_1[t], A_2[t], ..., A_K[t]\} \tag{7}$$

where $\mathbf{S} = (S_1, \cdots, S_N)$ is a length-N vector and $\mathbf{S}^* = (S_1^*, \cdots, S_N^*)$ represents the indices of the selected clients. The appeal of this method is in (a) it does not require additional information and is as low-complexity as the RS, and (b) it has a potential to reduce the parameter staleness.

[1]We simplify this probability as a constant to represent a general case of clients checking in when ready to participate in training. It resembles the scenario in which each client uploads its local parameters based on the channel inversion scheme but has a certain probability of encountering a deep fade and suspending the current transmission. Note that following a similar vein as [24], [25], [36], one can adopt the notion of transmission success probability to account for effects from physical layer factors such as the fading, path loss, and interference.

[2]The AoU measures the time elapsed since the latest update is received by the server, and hence larger the AoU indicates there are higher degrees of staleness associated with the update.

Remark 1: *These two policies resemble the random scheduling and group round-robin in networks with unreliable connectivity, and they are unbiased client sampling approaches.*

Remark 2: *Different from the settings in [24], where the selected clients can experience transmission failures, the selection of clients in this work is performed on those with reliable connections and hence does not waste communication resources.*

III. Analysis

This section constitutes the main technical part of this paper, in which we derive analytical expressions for the convergence rate of FL over wireless networks.

A. Preliminaries

1) Parameter Staleness: Since only a subset of the clients can be selected to participate in the FL training during each round of communication, the parameters of the unselected clients become stale. To formally characterize this effect, we denote a random variable τ_k as the staleness associated with the global parameter possessed by the k-th client. Then, in accordance with (4) and (5), after the t-th communication round, the update of global parameters at the server side can be rewritten as follows:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \sum_{k=1}^{K} p_k \sum_{s=0}^{H-1} \nabla f_k(\boldsymbol{w}_{t-\tau_k,s}; \xi_{k,s}) \tag{8}$$

in which $\xi_{k,s}$ denotes an element sampled uniformly at random from the data set of the k-th client during the s-th local computing step. Note that *the distribution of τ_k is dependent on the client selection criteria.* In the sequel, we will leverage (8) to derive the convergence rate of FL under the RS and ABS policies, respectively, and obtain several insights based on the analyses.

2) Assumptions: To facilitate the analysis, we assume the following conditions.

Assumption 1: *The gradient of each function $f_k : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz continuous with a constant $L > 0$, namely, for any $\boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^d$ the following is satisfied:*

$$\|\nabla f_k(\boldsymbol{w}) - \nabla f_k(\boldsymbol{v})\| \leq L\|\boldsymbol{w} - \boldsymbol{v}\|, \ k \in \{1, 2, ..., K\}. \tag{9}$$

Assumption 2: *The gradients of f_k are upper bounded by a constant C, i.e., for any $\boldsymbol{w} \in \mathbb{R}^d$ the following is satisfied:*

$$\|\nabla f_k(\boldsymbol{w})\| \leq C, \quad k \in \{1, 2, ..., K\}. \tag{10}$$

It is worthwhile to stress that Assumption 2 holds in our setting because transmitting an arbitrarily large value over the wireless channel is not practical. Indeed, the excessively large gradients are usually clipped before being sent out [37].

Following the above assumptions, we have

$$\begin{aligned} \|\nabla f(\boldsymbol{w}) - \nabla f(\boldsymbol{v})\| &\leq \sum_{k=1}^{K} p_k \|\nabla f_k(\boldsymbol{w}) - \nabla f_k(\boldsymbol{v})\| \\ &\leq \sum_{k=1}^{K} p_k L \|\boldsymbol{w} - \boldsymbol{v}\| = L\|\boldsymbol{w} - \boldsymbol{v}\|, \end{aligned} \tag{11}$$

$$\beta = \Bigg\{ \binom{K-1}{N} p^N (1-p)^{K-1-N} \frac{N}{N+1} + \binom{K-1}{N+1} p^{N+1} (1-p)^{K-2-N} \frac{N}{N+2} + \cdots + \binom{K-1}{K-1} p^{K-1} \frac{N}{K} + \binom{K-1}{N-1} p^{N-1} (1-p)^{K-N} + \cdots + \binom{K-1}{0} (1-p)^{K-1} \Bigg\} \times p \tag{13}$$

and

$$\|\nabla f(\boldsymbol{w})\| \leq \sum_{k=1}^{K} p_k \|\nabla f_k(\boldsymbol{w})\| \leq C. \tag{12}$$

Notably, we do not assume convexity of the objective function and hence the result is more general and applicable to the context of, e.g., (deep) neural networks.

B. Convergence Analysis

We now focus on the FL convergence analysis. We will first analyze the FL convergence rate under the RS policy. Then, we explore the performance under the ABS policy.

1) FL under RS Policy: In a typical global iteration t, a generic client needs to satisfy two conditions to be able to partake in the FL training process: (i) there is a reliable connection between the client and the server, and (ii) the client is selected by the server. Therefore, the FL participation state of a typical client is a binary random variable, where the probability can be derived as follows.

Lemma 1: *In a typical communication round, under the RS policy, the probability, β, that a generic client can participate in the FL training is given by* (13)*.*

Proof: Please refer to Appendix A. □

Since the client selections are conducted in an i.i.d. manner across communication rounds, the parameter staleness of a typical client follows a geometric distribution, i.e.,

$$\mathbb{P}(\tau_R = l) = \beta(1-\beta)^l, \quad l = 0, 1, 2, ... \tag{14}$$

Using (14), we can derive the convergence rate under the RS policy, presented in the next theorem.

Theorem 1: *Under the RS policy, if the step size is chosen as $\eta = 1/H\sqrt{T}$, then after T rounds of communications, Algorithm 1 converges as follows:*

$$\min_{0 \leq t \leq T-1} \mathbb{E}\Big[\|\nabla f(\boldsymbol{w}_t)\|^2\Big] \leq \frac{f(\boldsymbol{w}_0) - f(\boldsymbol{w}^*) + \frac{LC^2}{H}\Big(\mathbb{E}[\tau_R] + \frac{3}{2}H\Big)}{\beta\sqrt{T}} + \frac{C^2}{\beta T}. \tag{15}$$

Proof: Please refer to Appendix B. □

From (15), it is clear that the client participation probability, β, plays a critical role in the FL convergence rate. Moreover, using (13), we can bound β as follows:

$$pN/K \leq \beta \leq p, \tag{16}$$

where the lower bound follows by replacing the terms $\frac{N}{i}$, $i \in \{N+1, \ldots, K\}$ in (13) by $\frac{N}{K}$, together with the fact that $\frac{N}{K}$ is a fraction. The above inequality allows us to obtain better insights into the convergence rate of FL over unreliable networks. Specifically, let us resort to the following two extremes:

a) When $p \to 1$, we have $\beta \approx N/K$. In this case, providing more communication channels monotonically increases the probability of participation at each client, which, in turn, bolsters faster convergence of the FL algorithm.
b) When $p \to 0$, we have $\beta \approx p$. In this case, increasing the number of communication channels cannot contribute to boosting up client participation probability and hence is not instrumental in speeding up the FL convergence.

2) FL under ABS Policy: Next, we study the convergence rate of FL under the ABS policy. Similar to the above, we commence with deriving the distribution of parameter staleness, denoted by τ_A in this case.

To do this, we rearrange the clients according to the ascending order of their AoU in each communication round. Such an operation results in client 1 having the lowest AoU and client K with the highest AoU. Then, as the FL training progresses, the position of a generic client i varies due to the dynamics of AoU of all the clients in the network. As depicted by Fig. 2, the probability that the client transits to other positions is dependent on its particular location. We detail the analysis in the sequel.

When $i < K - N + 1$, there are more than N clients that have AoU larger than client i. And the transition of client i's position can be summarized in Fig. 2 (a). We start with analyzing the event that client i is selected by the server, after which its AoU reduces to zero and it moves to position 1. This happens if (a) the client has a reliable channel to the server and (b) at most $N-1$ clients ahead of it can establish reliable connections. The corresponding probability is given by

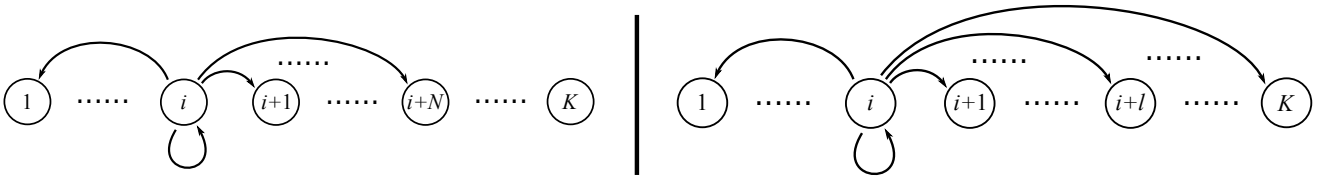
$$\mathbf{P}_{i,1} = p \sum_{m=0}^{N-1} \binom{K-i}{m} p^m (1-p)^{K-i-m}. \tag{17}$$

Next, we investigate the event that client i stays at its current position after a round of global iteration. This happens if all the clients in front of i, as well as client i itself, cannot connect to the server during the current communication round, which results in the following probability

$$\mathbf{P}_{i,i} = (1-p)^{K-i+1}. \tag{18}$$

The client may move forward from position i to $i+l$, whereas $1 \leq l \leq N-1$, given it is not connected to the server while there are l clients in front who have reliable channels for communications. As such, the probability can be written as

$$\mathbf{P}_{i,i+l} = (1-p)\binom{K-i}{l} p^l (1-p)^{K-i-l}, \ 1 \leq l \leq N-1. \tag{19}$$



(a) (b)

Fig. 2. An illustration of the position transition of a generic client i under the ABS policy.

Finally, we note that client i can also move to position $i+N$ if more than N clients ahead of it have reliable channels and hence are selected for parameter updating. This event happens with the following probability:

$$\mathbf{P}_{i,i+N} = \sum_{m=N}^{K-i} \binom{K-i}{m} p^m (1-p)^{K-i-m}. \tag{20}$$

On the other hand, when $i \geq K-N+1$, as illustrated in Fig. 2 (b), there are less than N clients standing before client i. In other words, client i is among the N candidates that have the highest AoU in the network. As such, it will be selected by the server as long as there is a reliable channel between them, which yields

$$\mathbf{P}_{i,1} = p. \tag{21}$$

The client will stay at its current position if, from position i onward, none of the clients is able to connect to the server reliably; this gives the following probability

$$\mathbf{P}_{i,i} = (1-p)^{K-i+1}. \tag{22}$$

Moreover, if client i cannot connect to the server in the present communication round but meanwhile, l clients in front of it are able to establish reliable connections, the client will transit to position $i+l$. It shall also be stressed that l is in the range of $[1, K-i]$ because client i cannot go beyond the end of the line, i.e., position K. Therefore, we have

$$\mathbf{P}_{i,i+l} = (1-p)\binom{K-i}{l} p^l (1-p)^{K-i-l}, \ 1 \leq l \leq K-i. \tag{23}$$

To this end, we can model the positions of the clients as the states of a Markov chain, where the transition matrix is given as $\mathbf{P} = [\,\mathbf{P}_{i,j}\,]_{1\leq i,j\leq K}$. This Markov chain is recurrent and irreducible and hence has a steady-state distribution. Let $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_K)$ be the steady state probability vector; then we can solve for the value of each entry via the following fixed-point equation:

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}. \tag{24}$$

As a result, we can characterize τ_A via the following.

Lemma 2: *The distribution of parameter staleness under the ABS policy is given by*

$$\mathbb{P}(\tau_A = l) = \sum_{i=1}^{K} \pi_i \left(\mathbf{P}_{(1)}^l \mathbf{P}\right)_{i,1}, \quad l = 0, 1, 2, ... \tag{25}$$

where $\mathbf{P}_{(1)}$ is a matrix obtained by replacing the first column of $\mathbf{P}$ with all zeros, and $(\mathbf{X})_{i,j}$ denotes the entry (i,j) of matrix $\mathbf{X}$.

Proof: At communication round t, if the parameter staleness of a typical client is $\tau_A = l$, it implies that beginning at $t-l-1$ (without loss of generality, we consider $l < t$), the client is not selected by the server for l consecutive global iterations and scheduled at the last communication round. Equivalently, this can be regarded as the client starting at state i and reaching state 1 for the first time after $l+1$ steps, which occurs with the following probability

$$\begin{aligned}&\mathbb{P}\,(\text{ Client reaches position 1 in } l \text{ steps from position } i\ )\\ &= \left(\mathbf{P}_{(1)}^l \mathbf{P}\right)_{i,1}.\end{aligned} \tag{26}$$

Because the probability of a typical client being in state i is given by π_i, the proof is complete by invoking the law of total probability. □

Following similar lines in Theorem 1, we can leverage (25) to derive the FL convergence rate under ABS policy. Nonetheless, the results may be too involved to offer useful insights. In that respect, we resort to the two extremes of communication conditions for better intuition. Particularly, when the communication channels are reliable, i.e., $p \approx 1$, the distribution of staleness τ_A can be approximated as follows:

$$\mathbb{P}(\tau_A = l) \approx \frac{1}{G}, \quad l = 0, 1, ..., G-1. \tag{27}$$

Armed with this result, we can derive the convergence rate of FL under ABS policy.

Theorem 2: *Under the ABS policy, when $p \approx 1$, if the step size is chosen as $\eta = 1/H\sqrt{T}$, then after T rounds of communications, Algorithm 1 converges as follows:*

$$\begin{aligned}&\min_{0\leq t\leq T-1} \mathbb{E}\Big[\|\nabla f(\boldsymbol{w}_t)\|^2\Big]\\ &\leq \frac{f(\boldsymbol{w}_0) - f(\boldsymbol{w}^*) + \frac{LC^2}{H}\Big(\mathbb{E}[\tau_A] + \frac{3}{2}H\Big)}{\sqrt{T}} + \frac{(G-1)C^2}{T}.\end{aligned} \tag{28}$$

Proof: Please refer to Appendix C. □

Putting together (15) and (28), we can see that for non-convex objective functions, the FL algorithm converges to stationary points in the order of $\mathcal{O}\left(1/\sqrt{T}\right)$, while scheduling policies affect the multiplicative factor. Moreover, using (27), we have $\mathbb{E}[\tau_A] = \frac{1}{2}(G-1) = \frac{1}{2}(K/N-1)$; and we have $\mathbb{E}[\tau_R] = \frac{1}{\beta}$ by (14), which indicates that $\mathbb{E}[\tau_R] \approx K/N > \mathbb{E}[\tau_A]$, for $p \approx 1$. Thus, we can conclude that when the communications are reliable, i.e., p is relatively large, using the ABS policy in the FL can achieve faster convergence than that achieved by using the RS policy. Intuitively, the gain is

Algorithm 2 Federated Momentum Learning Algorithm

Parameters: H = number of local steps per computation round, η = step size for stochastic gradient descent.
Initialize: $\boldsymbol{w}_0 \in \mathbb{R}^d$
for $t = 0, 1, 2, ..., T-1$ **do**
  The server selects a set S_t of at most N clients and broadcasts the global parameter $\boldsymbol{w}_t$ to them
  for each client $k \in S_t$ in parallel **do**
    Initialize $\boldsymbol{w}_{t,0}^{(k)} = \boldsymbol{w}_t$
    for s = 0 to $H-1$ **do**
      Sample $\xi_{k,s} \in \mathcal{D}_k$ uniformly at random, and update the local parameter $\boldsymbol{w}_t^{(k)}$ as follows:

$$\boldsymbol{w}_{t,s+1}^{(k)} = \boldsymbol{w}_{t,s}^{(k)} - \eta \nabla f_k(\boldsymbol{w}_{t,s}^{(k)}; \xi_{k,s}) \tag{29}$$

      in which ∇ represents the gradient operation
    end for
    Send the locally aggregated stochastic gradients $\sum_{s=0}^{H-1} \nabla f_k(\boldsymbol{w}_{t,s}^{(k)}; \xi_{k,s})$ to the server
  end for
  The server collects all the gradient parameters from the selected clients and assigns $\boldsymbol{g}_t^{(i)} = \sum_{s=0}^{H-1} \nabla f_i(\boldsymbol{w}_{t,s}^{(i)}; \xi_{k,s})$ for $i \in S_t$. Moreover, the server sets $\boldsymbol{g}_t^{(j)} = \boldsymbol{g}_{t-1}^{(j)}$ for $j \notin S_t$, and then updates the global parameter $\boldsymbol{w}_{t+1}$ as follows:

$$\boldsymbol{v}_t = \boldsymbol{v}_{t-1} + \gamma \sum_{k=1}^{K} p_k \boldsymbol{g}_t^{(k)} \tag{30}$$

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta\, \boldsymbol{v}_t \tag{31}$$

  where $\gamma \in [0, 1)$ is the control parameter
end for
Output: $\boldsymbol{w}_T$

mainly attributed to the fact that ABS accounts for the fairness in the channel access and leads to smaller parameter staleness.

On the other hand, when the clients are situated under poor communication environments in which the wireless connections are highly unreliable, namely, $p \ll 1$, we have $\tau_R \approx \tau_A$ in distribution. And this results in the following conclusion.

Corollary 1: *In networks with unreliable communication channels, i.e., $p \ll 1$, Algorithm 1 attains a similar convergence rate under both RS and ABS policies.*

Corollary 1 indicates that in the absence of reliable connections, neither providing more bandwidth nor leveraging better scheduling policies can enhance the FL convergence rate.

C. Federated Momentum Learning in Unreliable Networks

In this subsection, we detail the approach, and the efficacy, of adopting the momentum algorithm [38] in the training of (1), aiming to improve the performance. We term this method the Federated Momentum Learning (FML) and summarize the implementations in Algorithm 2. Particularly, the momentum term is introduced in the gradient updating step (30). It can be regarded as a "heavy ball" added in the update of parameters such that the values stay close to the current one.

The intuition behind this operation is that the update direction of SGD, while always along gradient descent, could cause an oscillating update path. Utilizing the momentum term can deviate the direction of the parameter update to the optimal decline and mitigate the possible oscillations caused by SGD. Since the clients in a wireless network are usually resource-constrained, algorithms that accelerate the convergence rate can attain higher resource utilization efficiency. The convergence rate of FML can be derived accordingly.

Theorem 3: *Under the RS policy, if the step size is chosen as $\eta = 1/H\sqrt{T}$, then after T rounds of communications, the Algorithm 2 converges as follows:*

$$\min_{0\le t\le T-1} \mathbb{E}\Big[\|\nabla f(\boldsymbol{w}_t)\|^2\Big] \le \frac{(1-\gamma)\big[f(\boldsymbol{w}_0) - f(\boldsymbol{w}^*)\big]}{\beta\sqrt{T}} + \frac{C^2}{\beta^2 T} + \frac{LC^2}{\beta\sqrt{T}}\Big(\frac{\mathbb{E}[\tau_R]}{H} + \frac{1}{2(1-\gamma)} + \frac{\gamma^2}{(1-\gamma)^2}\Big). \tag{32}$$

Proof: Please refer to Appendix D. □

By comparing (15) and (32), we can see that regardless of connection quality, by carefully choosing the control parameter, γ, in the FML, faster convergence can be attained. In that respect, it is confirmed that improving the FL from an algorithmic perspective is beneficial.

In a similar vein, we can derive the convergence rate of FML under the ABS policy.

Theorem 4: *Under the ABS policy, when $p \approx 1$, if the step size is chosen as $\eta = 1/H\sqrt{T}$, then after T rounds of communications, Algorithm 2 converges as follows:*

$$\min_{0\le t\le T-1} \mathbb{E}\Big[\|\nabla f(\boldsymbol{w}_t)\|^2\Big] \le \frac{(1-\gamma)\big[f(\boldsymbol{w}_0) - f(\boldsymbol{w}^*)\big]}{\sqrt{T}} + \frac{\mathbb{E}[\tau_A]C^2}{T} + \frac{LC^2}{\sqrt{T}}\Big(\frac{\mathbb{E}[\tau_A]}{H} + \frac{1}{2(1-\gamma)} + \frac{\gamma^2}{(1-\gamma)^2}\Big). \tag{33}$$

Proof: The proof is similar to that of Theorem 2 and 3 and hence omitted here. □

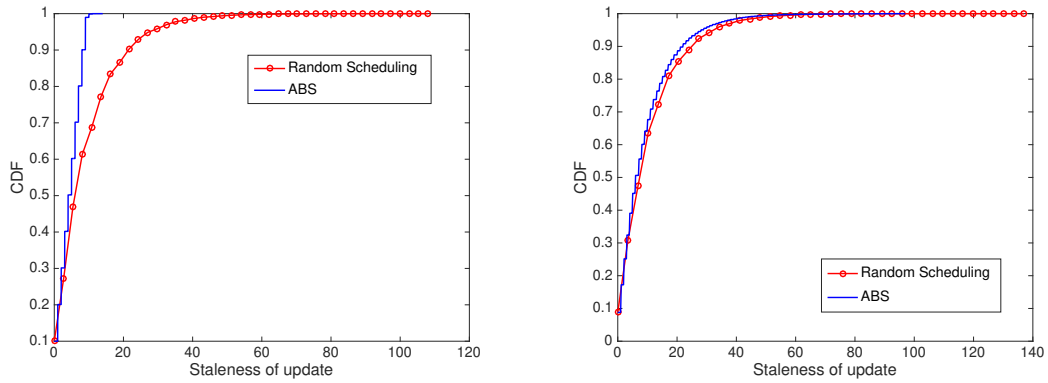
As in the case without momentum, we also have the following result for FML.

Corollary 2: *In networks with unreliable communication channels, i.e., $p \ll 1$, Algorithm 2 has similar convergence rate under both RS and ABS policies.*

From the above, we can see that faster convergence of FL can be achieved by introducing the momentum term and carefully choosing the corresponding parameters.

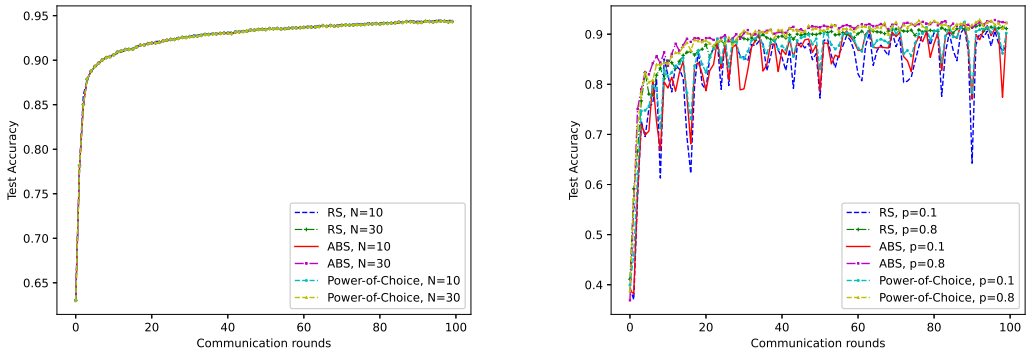
IV. Simulation Results

In this section, we conduct simulations to verify the analyses that have been developed. Specifically, we examine the efficiency of training FL on two different settings of machine learning models. One experiment is to train an MLP over the MNIST dataset. The MLP consists of 2 hidden layers, each having 64 units, and adopts the ReLU activations. The dataset contains 10,000 handwritten images of the numbers 0 to 9, where each digit has 1000 images. We take 9,000 data samples from the MNIST dataset for the training and allocate 1,000 samples for testing. The other experiment is to train a CNN on the CIFAR-10 dataset. This dataset consists of 60,000 color images in 10 classes, with 6000 images per class. The CNN has two convolutional layers with a max pooling, followed by two fully connected layers, and then a softmax output layer. We take 50,000 data samples from the CIFAR-10 dataset for



(a) p=0.8 (b) p=0.1

Fig. 3. Staleness under ABS and RS. In Fig. (a), the staleness of the parameter when the channels are reliable, i.e., $p = 0.8$. In Fig. (b), the staleness of parameters when the channels are unreliable, i.e., $p = 0.1$.



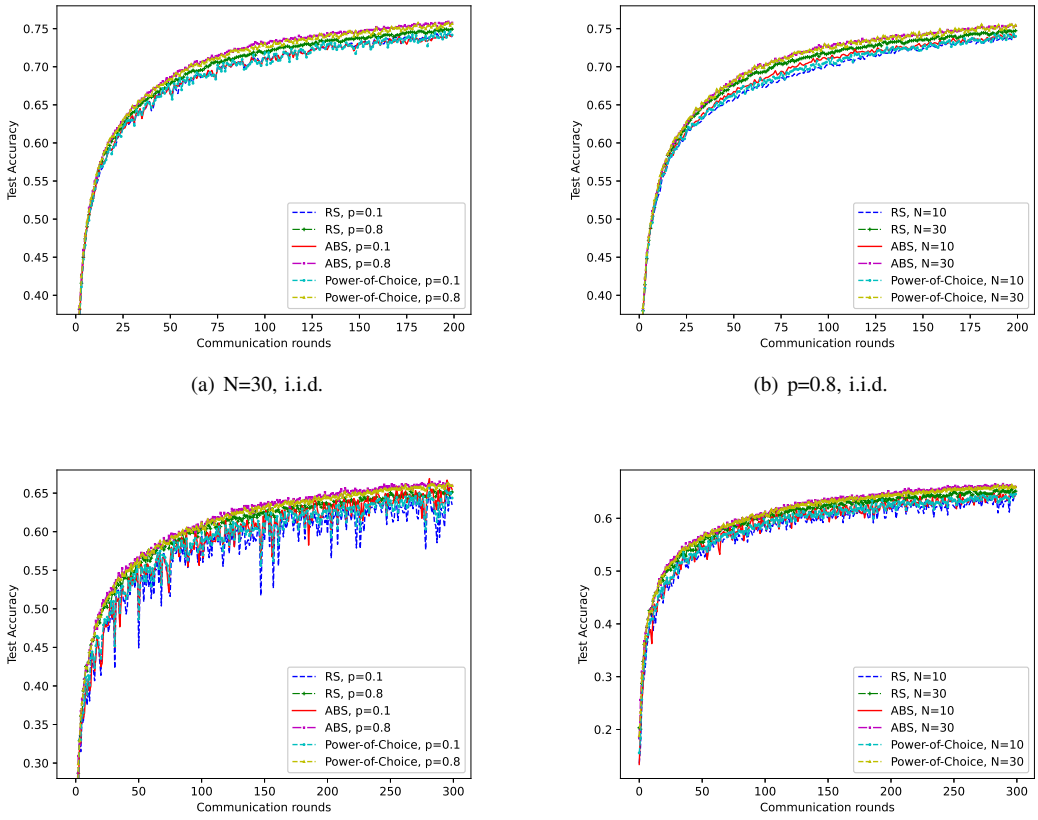
(a) p=0.1, i.i.d. (b) N=30, non-i.i.d.

Fig. 4. Convergence rate of training MLP on the MNIST dataset under various communication conditions and scheduling policies. Fig. (a), the transmission success probability is $p = 0.1$; the dataset is assigned to the clients in an i.i.d. manner. Fig. (b), number of communication channels is $N = 30$, the dataset is assigned to the clients in a non-i.i.d. manner.

training and assign 10,000 samples for testing. We partition the training dataset into 100 non-overlapped portions and assign them to $K = 100$ clients. In our experiments, we consider both i.i.d. and non-i.i.d. settings. For the i.i.d. local data partition, the whole dataset is uniformly distributed among all clients at random. For the non-i.i.d. data partition, we adopt a sort-and-partition scheme, where we sort all the data according to the labels and divide the data into 200 shards. Each client is assigned two shards. We choose the learning rate as $\eta = 0.01$ and the momentum weight as $\gamma = 0.9$. The wireless channels are considered reliable when $p \geq 0.8$ and unreliable when $p \leq 0.1$. All the experiments are implemented with Pytorch and averaged over three trials.

Fig. 3 plots the Cumulative Distribution Function (CDF) of the parameters' staleness, τ, under different configurations of the wireless channel and scheduling policy. We can see that when each client has a relatively high probability of establishing a reliable channel to the server, ABS attains a smaller value of parameter staleness than RS, while the staleness of parameters under the two schemes are similar when the communication channels become unreliable. This observation confirms that scheduling policy can influence the staleness of parameters in the FL training.

Fig. 4 depicts the test accuracy of FL training on the MNIST dataset as a function of communication rounds under different scheduling policies as well as the reliability of the wireless channels. From Fig. 4 (a), we can see that when the channels are unreliable, i.e., $p = 0.1$, the convergence rate of FL remains unchanged regardless of the employed scheduling policy or number of available communication channels. This is mainly due to the fact that when channels are unreliable, only a few clients – the total number of them may be even less than the number of channels available for communications – can establish connections to the server in each communication round. Since only these clients can be selected for parameter updating, neither scheduling policy nor communication bandwidth can be instrumental in enhancing the performance. On the other hand, Fig. 4 (b) demonstrates that when the communication channels become reliable, ABS can attain a faster convergence rate than RS, as it asserts a higher level of fairness amongst the clients. Additionally, we observe that the curve of convergence rate under ABS is smoother than that under RS, as the model parameters of the clients are more aligned under ABS. Finally, we notice that compared



(a) N=30, i.i.d.

(b) p=0.8, i.i.d.

(c) N=30, non-i.i.d.

(d) p=0.8, non-i.i.d.

Fig. 5. Convergence rate of training CNN on the CIFAR-10 dataset under various communication conditions and scheduling policies.

to the unreliable channel case, running FL in networks that have reliable connectivity guarantees faster convergence of the model training. Therefore, it is of paramount importance to maintain a reliable communication infrastructure for the FL system.

We can observe similar phenomena from the convergence rate of training a more complicated ML model, i.e., the CNN, on the CIFAR-10 dataset, as illustrated in Fig. 5. Particularly, Fig. 5 (a) shows that the convergence rate under RS and ABS almost coincide with each other when the communication channels are unreliable, i.e., $p = 0.1$. In contrast, there is a marked speedup in the convergence rate of ABS over RS if the network has good communication channels (in this case, $p = 0.8$). Moreover, Fig. 5 (b) confirms that when the communication channels are reliable, a faster convergence rate can be achieved by providing more communication channels. Fig. 5 (c) and Fig. 5 (d) show the superiority of our proposed ABS scheduling when the dataset is distributed to the clients in a non-i.i.d. manner with varying value of transmission success probability and number of submission channels. These observations corroborate the conclusions we have drawn in Section III.

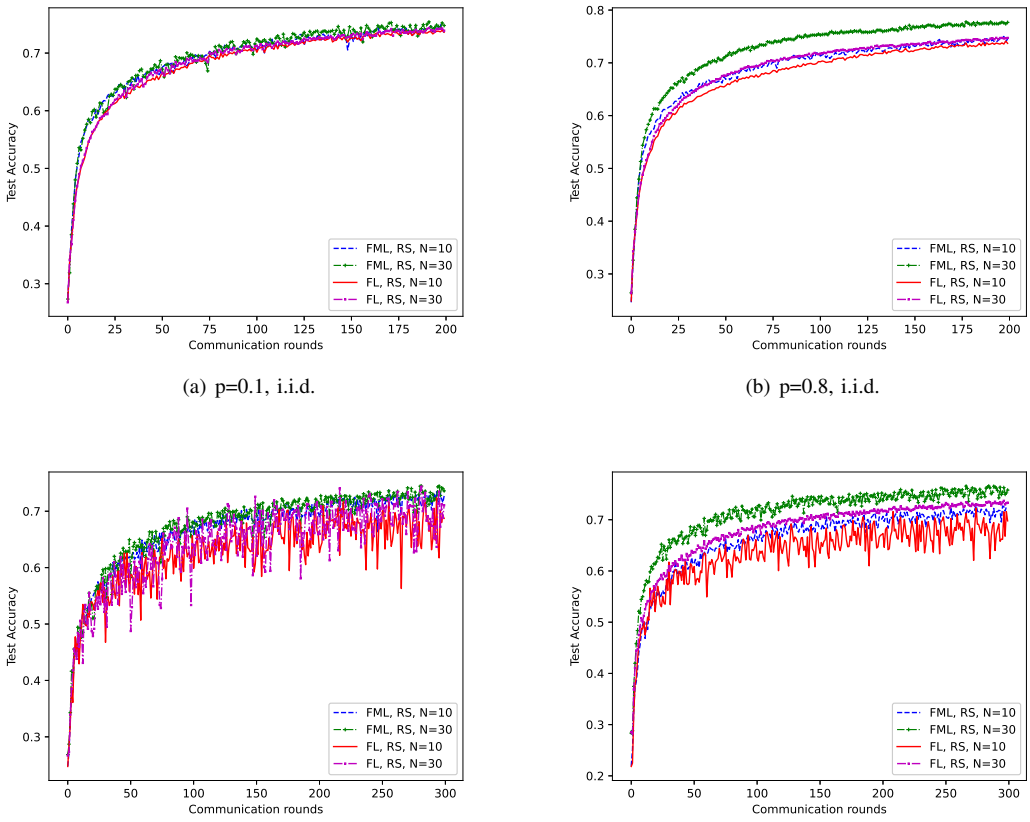
We now turn our attention to the convergence performance of FML. We concentrate on the task of training a CNN on the CIFAR-10 dataset. The experiments are conducted under the aforementioned settings, except that Algorithm 2 is adopted for the FL model training. The numerical results are summarized in Figures 6 and 7, which respectively illustrate the convergence rates under RS and ABS policies.

Particularly, Fig. 6 compares the FML convergence rate under different connectivity conditions of the network. We can see from Fig. 6 (a) that even when the wireless links are highly unreliable, i.e., $p = 0.1$, running FL in tandem with momentum results in a faster convergence rate for both i.i.d. and non-i.i.d. local datasets. Additionally, Fig. 6 (b) and Fig. 6 (d) show that the benefits conferred by momentum are more pronounced when communications are reliable. By comparing the convergence curve of FML with 10 channels against that under FL with 30 channels, we find that in order to speed up the model training, using the momentum method can be as effective as expanding the communication bandwidth.

We can draw similar conclusions in the context of running FML with ABS policy for both i.i.d. and non-i.i.d. user training data cases, as shown in Fig. 7. Notably, this figure demonstrates that FML with 10 communication channels can bring along faster convergence rate than FL with 30 channels, which discloses the importance of algorithm design in the FL training.

V. Conclusion

In this paper, we carried out an analytical study toward understanding the efficiency of training FL models over a



(a) p=0.1, i.i.d.

(b) p=0.8, i.i.d.

(c) p=0.1, non-i.i.d.

(d) p=0.8, non-i.i.d.

Fig. 6. Convergence rate of training CNN on the CIFAR-10 dataset under FML by using RS policy. In Fig. (a) and Fig. (c): $p = 0.1$. In Fig. (b) and Fig. (d): $p = 0.8$. The dataset is distributed to the clients in an i.i.d. manner in Fig. (a) and Fig. (c), and in a non-i.i.d. manner in Fig. (b) and Fig. (d).

wireless network. We established the FL convergence rate by taking into account key system parameters such as the probability of reliable transmissions, staleness of parameters, and scheduling method. Our analysis confirmed the importance of communication quality in the FL model training process. Specifically, if the clients can establish reliable connections to the server in each round of communication, then the model training can be accelerated by either adopting better scheduling policies or providing more communication bandwidth. But these methods become ineffectual when the connections are unreliable. We also demonstrated that the FL can be run in tandem with momentum, which can improve the convergence rate by appropriately tuning the momentum weight. These results advanced the understanding of the FL system and can be useful for researchers in their further research pursuit.

APPENDIX

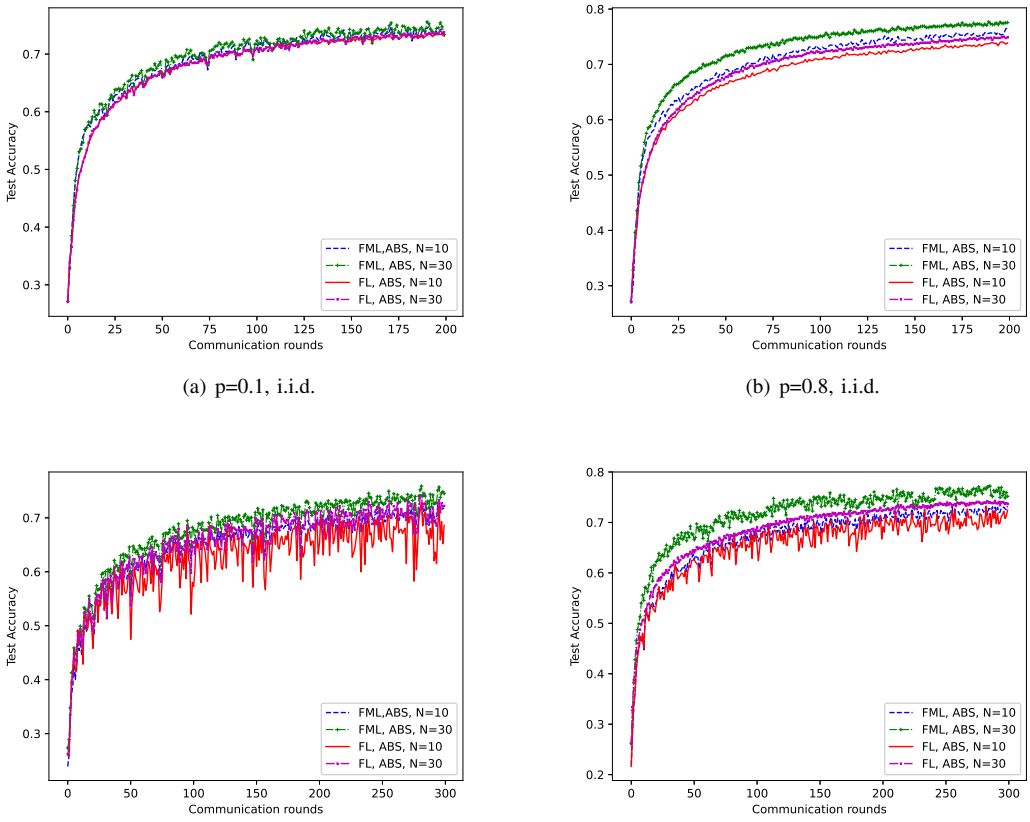
A. Proof of Lemma 1

In a typical communication round, we use a binary variable $R_k \in \{0, 1\}$ to indicate that client k has a reliable channel to the server (in this case, $R_k = 1$) or not (in this case, $R_k = 0$). Moreover, we denote $\tilde{N}$ as the number of reliable channels except for client k. As such, the probability β that the client attains successful parameter update can be written as

$$\begin{aligned}\beta &= \mathbb{P}\left(S_k[t] = 1 | R_k = 1\right) \times \mathbb{P}\left(R_k = 1\right) \\ &= p \times \Bigg(\underbrace{\mathbb{P}\left(S_k[t] = 1, \tilde{N} \geq N | R_k = 1\right)}_{Q_1} \\ &\qquad + \underbrace{\mathbb{P}\left(S_k[t] = 1, \tilde{N} < N | R_k = 1\right)}_{Q_2} \Bigg). \end{aligned} \tag{34}$$

Under the RS policy, when $\tilde{N} \geq N$, only N clients will be uniformly selected out at random. Therefore, upon noting that $\tilde{N}$ are R_k are independent, Q_1 can be calculated as

$$\begin{aligned} Q_1 &= \mathbb{P}(S_k[t] = 1 | \tilde{N} = N, R_k = 1) \times \mathbb{P}(\tilde{N} = N) \\ &\quad + \mathbb{P}(S_k[t] = 1 | \tilde{N} = N+1, R_k = 1) \times \mathbb{P}(\tilde{N} = N+1) \\ &\quad + \cdots + \mathbb{P}(S_k[t] = 1 | \tilde{N} = K, R_k = 1) \times \mathbb{P}(\tilde{N} = K) \\ &= \frac{N}{N+1} \times \binom{K-1}{N} \times p^N \times (1-p)^{K-1-N} \\ &\quad + \frac{N}{N+2} \times \binom{K-1}{N+1} \times p^{N+1} \times (1-p)^{K-1-(N+1)} \\ &\quad + \cdots + \frac{N}{K} \times \binom{K-1}{K-1} \times p^{K-1}. \end{aligned} \tag{35}$$



(a) p=0.1, i.i.d.

(b) p=0.8, i.i.d.

(c) p=0.1, non-i.i.d.

(d) p=0.8, non-i.i.d.

Fig. 7. Convergence rate of training CNN on the CIFAR-10 dataset under FML by using ABS policy. Fig (a) and Fig. (c): $p = 0.1$, Fig. (b) and Fig. (d): $p = 0.8$. The dataset is distributed to the clients in an i.i.d. manner in Fig (a) and Fig. (b) cases, and in a non-i.i.d. manner in Fig (c) and Fig. (d) cases.

In the situation that $\tilde{N} < N$, all the clients that have reliable channels will be selected for parameter update. We can thus compute Q_2 as follows:

$$\begin{aligned} Q_2 =& \mathbb{P}(S_k[t]=1|\tilde{N}=N-1, R_k=1)\times\mathbb{P}(\tilde{N}=N-1)\\ &+\mathbb{P}(S_k[t]=1|\tilde{N}=N-2, R_k=1)\times\mathbb{P}(\tilde{N}=N-2)\\ &+\cdots+\mathbb{P}(S_k[t]=1|\tilde{N}=0, R_k=1)\times\mathbb{P}(\tilde{N}=0)\\ =&\binom{K-1}{N-1}\times p^{N-1}\times(1-p)^{K-N}\\ &+\binom{K-1}{N-2}\times p^{N-2}\times(1-p)^{K-1-(N-2)}\\ &+\cdots+\binom{K-1}{0}\times(1-p)^{K-1}. \end{aligned} \tag{36}$$

The result then follows by substituting (35) and (36) into (34).

B. Proof of Theorem 1

According to (11), f is L-smooth and hence when the global parameter is updated from $\boldsymbol{w}^t$ to $\boldsymbol{w}^{t+1}$, the following relationship is satisfied:

$$\begin{aligned} f(\boldsymbol{w}_{t+1}) \leq& f(\boldsymbol{w}_t)+\langle \boldsymbol{w}_{t+1}-\boldsymbol{w}^t, \nabla f(\boldsymbol{w}_t)\rangle+\frac{L}{2}\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|^2\\ =& f(\boldsymbol{w}_t)-\eta\sum_{k=1}^{K}p_k\sum_{s=0}^{H-1}\langle\nabla f_k(\boldsymbol{w}_{t-\tau_k,s};\xi_{k,s}),\nabla f(\boldsymbol{w}_t)\rangle\\ &+\frac{L}{2}\left\|\eta\sum_{k=1}^{K}p_k\sum_{s=0}^{H-1}\nabla f_k(\boldsymbol{w}_{t-\tau_k,s};\xi_{k,s})\right\|^2. \end{aligned} \tag{37}$$

From the right-hand side of the above inequality, we can identify two sources of randomness in a generic round of communications: i) the random sampling of data points of the selected clients during the local computing stage, and ii) the staleness associated with the parameters of the unselected clients.

We thereby deal with these two aspects separately. First of all, by taking an expectation on both sides of (37) with respect to the data points, ξ, randomly sampled during the t-th round of FL training we have

$$\begin{aligned} &\mathbb{E}_\xi\left[f(\boldsymbol{w}_{t+1})\right]\\ \leq& f(\boldsymbol{w}_t)-\eta\sum_{k=1}^{K}p_k\sum_{s=0}^{H-1}\mathbb{E}_\xi\left[\langle\nabla f_k(\boldsymbol{w}_{t-\tau_k,s};\xi_{k,s}),\nabla f(\boldsymbol{w}_t)\rangle\right]\\ &+\frac{L}{2}\mathbb{E}_\xi\left[\left\|\eta\sum_{k=1}^{K}p_k\sum_{s=0}^{H-1}\nabla f_k(\boldsymbol{w}_{t-\tau_k,s};\xi_{k,s})\right\|^2\right]\\ \overset{(a)}{\leq}& f(\boldsymbol{w}_t)-\eta\sum_{k=1}^{K}p_k\sum_{s=0}^{H-1}\langle\nabla f_k(\boldsymbol{w}_{t-\tau_k,s}),\nabla f(\boldsymbol{w}_t)\rangle+\frac{L}{2}\eta^2H^2C^2\\ =& f(\boldsymbol{w}_t)-\eta\sum_{k=1}^{K}p_k\sum_{s=0}^{H-1}\langle\nabla f_k(\boldsymbol{w}_{t-\tau_k}),\nabla f(\boldsymbol{w}_t)\rangle+\frac{L}{2}\eta^2H^2C^2 \end{aligned}$$

$$
\begin{aligned}
&+\eta\sum_{k=1}^{K}p_k\sum_{s=0}^{H-1}\langle\nabla f_k(\boldsymbol{w}_{t-\tau_k})-\nabla f_k(\boldsymbol{w}_{t-\tau_k,s}),\nabla f(\boldsymbol{w}_t)\rangle\\
\overset{(b)}{\leq}&f(\boldsymbol{w}_t)-\eta\sum_{k=1}^{K}p_k\sum_{s=0}^{H-1}\langle\nabla f_k(\boldsymbol{w}_{t-\tau_k}),\nabla f(\boldsymbol{w}_t)\rangle+\frac{L}{2}\eta^2H^2C^2\\
&+\eta\sum_{k=1}^{K}p_k\sum_{s=0}^{H-1}\|\nabla f_k(\boldsymbol{w}_{t-\tau_k})-\nabla f_k(\boldsymbol{w}_{t-\tau_k,s})\|\cdot\|\nabla f(\boldsymbol{w}_t)\|\\
\overset{(c)}{\leq}&f(\boldsymbol{w}_t)-\eta H\sum_{k=1}^{K}p_k\langle\nabla f_k(\boldsymbol{w}_{t-\tau_k}),\nabla f(\boldsymbol{w}_t)\rangle+\frac{3L}{2}\eta^2H^2C^2
\end{aligned}
\tag{38}
$$

where (a) follows from $\mathbb{E}_\xi[\nabla f_k(\boldsymbol{w}_{t-\tau_k,s};\xi_{k,s})] = \nabla f_k(\boldsymbol{w}_{t-\tau_k,s})$, Jensen's inequality, and using (12); (b) by using the Cauchy-Schwartz inequality: $\langle\boldsymbol{x},\boldsymbol{y}\rangle\leq\|\boldsymbol{x}\|\cdot\|\boldsymbol{y}\|$; and (c) by using (11) and what follows:

$$
\begin{aligned}
&\|\nabla f_k(\boldsymbol{w}_{t-\tau_k})-\nabla f_k(\boldsymbol{w}_{t-\tau_k,s})\|\cdot\|\nabla f(\boldsymbol{w}_t)\|\\
\leq&LC\|\boldsymbol{w}_{t-\tau_k}-\boldsymbol{w}_{t-\tau_k,s}\|\\
\leq&LC\big(\|\boldsymbol{w}^{(k)}_{t-\tau_k,0}-\boldsymbol{w}^{(k)}_{t-\tau_k,1}\|+\|\boldsymbol{w}^{(k)}_{t-\tau_k,1}-\boldsymbol{w}^{(k)}_{t-\tau_k,2}\|\\
&\qquad+\cdots+\|\boldsymbol{w}^{(k)}_{t-\tau_k,s-1}-\boldsymbol{w}^{(k)}_{t-\tau_k,s}\|\big)\\
=&LC\eta\big(\|\nabla f_k(\boldsymbol{w}^{(k)}_{t-\tau_k,0})\|+\|\nabla f_k(\boldsymbol{w}^{(k)}_{t-\tau_k,1})\|\\
&\qquad+\cdots+\|\nabla f_k(\boldsymbol{w}^{(k)}_{t-\tau_k,H-1})\|\big)\\
\leq&\eta HLC^2.
\end{aligned}
\tag{39}
$$

Next, we take an expectation on both sides of (38) with respect to $\tau_k, k=1,2,...,K$. Because the random variables $\{\tau_k\}_{k=1}^K$ are i.i.d. $\sim\tau_R$, we arrive at the following

$$
\begin{aligned}
&\mathbb{E}_{\tau_k,k=1,2,...,K}\big[f(\boldsymbol{w}_{t+1})\big]\\
\leq&f(\boldsymbol{w}_t)-\eta H\sum_{k=1}^{K}p_k\mathbb{E}_{\tau_k}\left[\langle\nabla f_k(\boldsymbol{w}_{t-\tau_k}),\nabla f(\boldsymbol{w}_t)\rangle\right]\\
&\qquad+\frac{3L}{2}\eta^2H^2C^2\\
=&f(\boldsymbol{w}_t)\underbrace{-\eta H\mathbb{E}_{\tau_R}\big[\langle\nabla f(\boldsymbol{w}_{t-\tau_R}),\nabla f(\boldsymbol{w}_t)\rangle\big]}_{Q_3}+\frac{3L}{2}\eta^2H^2C^2.
\end{aligned}
\tag{40}
$$

For ease of exposition, let us denote $q_l=\mathbb{P}(\tau_R=l)$. Then, using (14), Q_3 can be calculated as follows:

$$
\begin{aligned}
Q_3=&-\eta H\sum_{l=0}^{t}q_l\langle\nabla f(\boldsymbol{w}_t),\nabla f(\boldsymbol{w}_{t-l})\rangle\\
&\qquad-\eta H(1-\beta)^{t+1}\langle\nabla f(\boldsymbol{w}_0),\nabla f(\boldsymbol{w}_t)\rangle\\
=&-\eta H\sum_{l=0}^{t}q_l\|\nabla f(\boldsymbol{w}_{t-l})\|^2\\
&\underbrace{-\eta H\sum_{l=0}^{t}q_l\langle\nabla f(\boldsymbol{w}_t)-\nabla f(\boldsymbol{w}_{t-l}),\nabla f(\boldsymbol{w}_{t-l})\rangle}_{Q_4}\\
&\qquad\underbrace{-\eta H(1-\beta)^{t+1}\langle\nabla f(\boldsymbol{w}_0),\nabla f(\boldsymbol{w}_t)\rangle}_{Q_5}.
\end{aligned}
\tag{41}
$$

By interchangeably using the Cauchy-Schwartz and AM-GM inequalities, we can bound Q_4 and Q_5 as follows:

$$
\begin{aligned}
Q_4=&\eta H\sum_{l=0}^{t}q_l\sum_{d=t-l}^{t-1}\langle\nabla f(\boldsymbol{w}_d)-\nabla f(\boldsymbol{w}_{d+1}),\nabla f(\boldsymbol{w}_{t-l})\rangle\\
\leq&\eta H\sum_{l=0}^{t}q_l\sum_{d=t-l}^{t-1}\|\nabla f(\boldsymbol{w}_d)-\nabla f(\boldsymbol{w}_{d+1})\|\cdot\|\nabla f(\boldsymbol{w}_{t-l})\|\\
\leq&\eta H\sum_{l=0}^{t}q_l\sum_{d=t-l}^{t-1}L\|\boldsymbol{w}_d-\boldsymbol{w}_{d+1}\|\cdot\|\nabla f(\boldsymbol{w}_{t-l})\|\\
=&\eta H\sum_{l=0}^{t}q_l\sum_{d=t-l}^{t-1}L\eta\,\|\nabla f(\boldsymbol{w}_d)\|\cdot\|\nabla f(\boldsymbol{w}_{t-l})\|\\
\leq&\eta^2H\sum_{l=0}^{t}q_lL\sum_{d=t-l}^{t-1}\frac{\|\nabla f(\boldsymbol{w}_d)\|^2+\|\nabla f(\boldsymbol{w}_{t-l})\|^2}{2}\\
\leq&\eta^2HL\sum_{l=0}^{t}q_l\times l\times C^2
\end{aligned}
\tag{42}
$$

and

$$
\begin{aligned}
Q_5=&\eta H(1-\beta)^{t+1}\langle-\nabla f(\boldsymbol{w}_0),\nabla f(\boldsymbol{w}_t)\rangle\\
\leq&\eta H(1-\beta)^{t+1}\|-\nabla f(\boldsymbol{w}_0)\|\cdot\|\nabla f(\boldsymbol{w}_t)\|\\
\leq&\eta H(1-\beta)^{t+1}\frac{\|\nabla f(\boldsymbol{w}_0)\|^2+\|\nabla f(\boldsymbol{w}_t)\|^2}{2}\\
\leq&\eta H(1-\beta)^{t+1}C^2.
\end{aligned}
\tag{43}
$$

By substituting (41), (42), and (43) into (38), and taking expectations to the corresponding random variables, we have

$$
\begin{aligned}
\mathbb{E}\big[f(\boldsymbol{w}_{t+1})\big]\leq&\mathbb{E}\big[f(\boldsymbol{w}_t)\big]-\eta H\sum_{l=0}^{t}q_l\mathbb{E}\big[\|\nabla f(\boldsymbol{w}_{t-l})\|^2\big]\\
&+\eta^2HLC^2\mathbb{E}\big[\tau_R\big]+\frac{3L}{2}\eta^2H^2C^2+\eta H(1-\beta)^{t+1}C^2.
\end{aligned}
\tag{44}
$$

Following the above inequality, we can rearrange the terms and telescope through t, which yields

$$
\begin{aligned}
&\min_{0\leq t\leq T-1}\mathbb{E}\big[\|\nabla f(\boldsymbol{w}_t)\|^2\big]\eta H\underbrace{\sum_{t=0}^{T-1}\sum_{l=0}^{t}q_l}_{Q_6}\\
\leq&\mathbb{E}\big[f(\boldsymbol{w}_0)-f(\boldsymbol{w}_T)\big]+\eta^2LC^2HT\Big(\mathbb{E}[\tau_R]+\frac{3}{2}H\Big)\\
&+\eta H\sum_{t=0}^{T-1}(1-\beta)^{t+1}C^2.
\end{aligned}
\tag{45}
$$

Using (14), we can further bound Q_6 as follows:

$$
\begin{aligned}
Q_6=&\sum_{t=0}^{T-1}\sum_{l=0}^{t}\beta(1-\beta)^l=\sum_{t=0}^{T-1}\big[1-(1-\beta)^{t+1}\big]\\
=&T-\frac{1-\beta}{\beta}\big[1-(1-\beta)^T\big]\\
\geq&T-\frac{1-\beta}{\beta}\big[1-(1-\beta T)\big]=\beta T.
\end{aligned}
\tag{46}
$$

Finally, one can also use the bound $\sum_{t=0}^{T-1}(1-\beta)^{t+1} \leq \frac{1}{\beta}$ on the last term in (45) and rewrite it as

$$\begin{aligned}&\beta\eta TH \min_{0\leq t\leq T-1} \mathbb{E}\big[\|\nabla f(\boldsymbol{w}_t)\|^2\big]\\ \leq& f(\boldsymbol{w}_0) - f(\boldsymbol{w}^*) + \eta^2 LC^2 HT\Big(\mathbb{E}[\tau_R] + \frac{3}{2}H\Big) + \frac{\eta HC^2}{\beta}.\end{aligned} \tag{47}$$

By further taking $\eta = 1/H\sqrt{T}$ we complete the proof.

C. Proof of Theorem 2

Following similar lines from the proof of Theorem 1 (cf. derivation of (40)), we can show that under the ABS policy, when the global parameter is updated from $\boldsymbol{w}^t$ to $\boldsymbol{w}^{t+1}$, the following relationship holds:

$$\begin{aligned}&\mathbb{E}_{\xi_k,\tau_k,k=1,2,\cdots,K}\big[f(\boldsymbol{w}_{t+1})\big]\\ \leq& f(\boldsymbol{w}^t) \underbrace{-\eta H \mathbb{E}_{\tau_A}\big[\langle \nabla f(\boldsymbol{w}_{t-\tau_A}), \nabla f(\boldsymbol{w}_t)\rangle\big]}_{Q_7} + \frac{3L}{2}\eta^2 H^2 C^2.\end{aligned} \tag{48}$$

Using the distribution of τ_A per (27), we can bound Q_7 as follows:

$$\begin{aligned}Q_7 &= -\eta H \sum_{l=0}^{G-1} \frac{1}{G} \left\langle \nabla f(\boldsymbol{w}^t), \nabla f(\boldsymbol{w}_{t-l})\right\rangle\\ &= -\eta H \sum_{l=0}^{G-1} \frac{1}{G} \left\|\nabla f(\boldsymbol{w}_{t-l})\right\|^2\\ &\quad + \eta H \sum_{l=0}^{G-1} \frac{1}{G} \left\langle \nabla f(\boldsymbol{w}_{t-l}) - \nabla f(\boldsymbol{w}_t), \nabla f(\boldsymbol{w}^{t-l})\right\rangle\\ &\leq -\eta H \sum_{l=0}^{G-1} \frac{1}{G} \left\|\nabla f(\boldsymbol{w}_{t-l})\right\|^2\\ &\quad + \eta H \sum_{l=0}^{G-1} \frac{1}{G} \Big(\frac{L\eta}{2} \sum_{d=t-l}^{t-1} \|\nabla f(\boldsymbol{w}_d)\|^2 + \frac{L\eta}{2} l \left\|\nabla f(\boldsymbol{w}_{t-l})\right\|^2\Big)\\ &\leq -\eta H \sum_{l=0}^{G-1} \frac{1}{G} \left\|\nabla f(\boldsymbol{w}_{t-l})\right\|^2 + \mathbb{E}\left[\tau_A\right] \eta H L \eta C^2.\end{aligned} \tag{49}$$

Putting (49) into (48), and taking expectation on both sides with respect to all the randomness up to communication round t, we have the following

$$\begin{aligned}\mathbb{E}\big[f(\boldsymbol{w}_{t+1})\big] \leq \mathbb{E}\big[f(\boldsymbol{w}_t)\big] - \eta H \sum_{l=0}^{G-1} \frac{1}{G} \mathbb{E}\left[\left\|\nabla f(\boldsymbol{w}_{t-l})\right\|^2\right]\\ + HL\eta^2 C^2 \left(\frac{3}{2}H + \mathbb{E}\left[\tau_A\right]\right).\end{aligned} \tag{50}$$

By rearranging the terms above and telescoping, we have the following

$$\begin{aligned}\eta H \sum_{t=0}^{T-1} \frac{1}{G} \sum_{l=0}^{G-1} \mathbb{E}\big[\|\nabla f(\boldsymbol{w}_{t-l})\|^2\big] \leq f(\boldsymbol{w}_0) - f(\boldsymbol{w}^*)\\ + HL\eta^2 C^2 \left(\frac{3}{2}H + \mathbb{E}\left[\tau_A\right]\right) T.\end{aligned} \tag{51}$$

The left hand side of the above inequality can be expressed as

$$\begin{aligned}&\eta H \sum_{t=0}^{T-1} \frac{1}{G} \sum_{l=0}^{G-1} \mathbb{E}\big[\|\nabla f(\boldsymbol{w}_{t-l})\|^2\big]\\ =&\eta H \sum_{t=G-1}^{T-1} \frac{1}{G} \sum_{l=0}^{G-1} \mathbb{E}\big[\|\nabla f(\boldsymbol{w}_{t-l})\|^2\big]\\ &+ \eta H \sum_{t=0}^{G-2} \frac{1}{G} \sum_{l=0}^{G-1} \mathbb{E}\big[\|\nabla f(\boldsymbol{w}_{t-l})\|^2\big].\end{aligned} \tag{52}$$

By jointly considering (51) and (52), we have

$$\begin{aligned}&\eta HT \min_{0\leq t\leq T-1} \mathbb{E}\big[\|\nabla f(\boldsymbol{w}_t)\|^2\big]\\ \leq& f(\boldsymbol{w}_0) - f(\boldsymbol{w}^*) + HL\eta^2 C^2 \left(\frac{3}{2}H + \mathbb{E}\left[\tau_A\right]\right) T\\ &+ \frac{\eta H}{G} \sum_{t=1}^{G-1} \sum_{l=G-t}^{G-1} \mathbb{E}\big[\|\nabla f(\boldsymbol{w}_l)\|^2\big]\\ \leq& f(\boldsymbol{w}_0) - f(\boldsymbol{w}^*) + HL\eta^2 C^2 \left(\frac{3}{2}H + \mathbb{E}\left[\tau_A\right]\right) T\\ &+ \mathbb{E}\left[\tau_A\right] \times \eta HC^2.\end{aligned} \tag{53}$$

By setting $\eta = 1/H\sqrt{T}$, the result follows.

D. Proof of Theorem 3

By substituting (8) into (30) and (31), the procedure of global parameter updates under momentum can be expressed as follows:

$$\boldsymbol{v}_{t+1} = \boldsymbol{w}_t - \eta \sum_{k=1}^{K} p_k \sum_{s=0}^{H-1} \nabla f_k(\boldsymbol{w}_{t-\tau_k,s}; \xi_k^s), \tag{54}$$

$$\boldsymbol{w}_{t+1} = \boldsymbol{v}_{t+1} + \gamma(\boldsymbol{v}_{t+1} - \boldsymbol{v}_t). \tag{55}$$

We denote by $\boldsymbol{g}_t = \eta \sum_{k=1}^{K} p_k \sum_{s=0}^{H-1} \nabla f_k(\boldsymbol{w}_{t-\tau_k,s}; \xi_k^s)$. Then, (54) can be rewritten as $\boldsymbol{v}_{t+1} = \boldsymbol{w}_t - \boldsymbol{g}_t$ and the update procedure (55) can be written as:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \boldsymbol{g}_t + \gamma(\boldsymbol{w}_t - \boldsymbol{g}_t - \boldsymbol{w}_{t-1} + \boldsymbol{g}_{t-1}). \tag{56}$$

Let us define an auxiliary term $\boldsymbol{u}_t$ as

$$\boldsymbol{u}_t = \frac{\gamma}{1-\gamma}\left(\boldsymbol{w}_t - \boldsymbol{w}_{t-1} + \boldsymbol{g}_{t-1}\right) \tag{57}$$

and establish the following relationship based on (56)

$$\boldsymbol{u}_{t+1} = \gamma \boldsymbol{u}_t - \frac{\gamma^2}{1-\gamma}\boldsymbol{g}_t. \tag{58}$$

Following the update process, one can use the above equations to express

$$\boldsymbol{w}_{t+1} + \boldsymbol{u}_{t+1} = \boldsymbol{w}_t + \boldsymbol{u}_t - \frac{1}{1-\gamma}\boldsymbol{g}_t. \tag{59}$$

We further denote $\boldsymbol{z}_t = \boldsymbol{w}_t + \boldsymbol{u}_t$ and rewrite (59) as

$$\boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \frac{1}{1-\gamma}\boldsymbol{g}_t. \tag{60}$$

Owing to the smoothness property of f, the following holds:

$$\begin{aligned} f\left(\boldsymbol{z}_{t+1}\right) \leq & f\left(\boldsymbol{z}_t\right)+\left\langle \boldsymbol{z}_{t+1}-\boldsymbol{z}_t, \nabla f\left(\boldsymbol{z}_t\right)\right\rangle+\frac{L}{2}\left\|\boldsymbol{z}_{t+1}-\boldsymbol{z}_t\right\|^2 \\ = & f\left(\boldsymbol{z}_t\right)-\frac{1}{1-\gamma}\left\langle \boldsymbol{g}_t, \nabla f\left(\boldsymbol{z}_t\right)\right\rangle+\frac{L}{2(1-\gamma)^2}\left\|\boldsymbol{g}_t\right\|^2 \\ = & f\left(\boldsymbol{z}_t\right)-\frac{\eta}{1-\gamma} \sum_{k=1}^K p_k \sum_{s=0}^{H-1}\left\langle \nabla f_k\left(\boldsymbol{w}_{t-\tau_k, s} ; \xi_k^s\right), \nabla f\left(\boldsymbol{z}_t\right)\right\rangle \\ & \underbrace{+\frac{\eta^2 L}{2(1-\gamma)^2}\left\|\sum_{k=1}^K p_k \sum_{s=0}^{H-1} \nabla f_k\left(\boldsymbol{w}_{t-\tau_k, s} ; \xi_k^s\right)\right\|^2}_{Q_8} . \end{aligned} \tag{61}$$

We can use Jensen's inequality to bound Q_8 as follows:

$$Q_8 \leq \frac{\eta^2 H^2 L C^2}{2(1-\gamma)^2}, \tag{62}$$

and write (61) in the following way:

$$\begin{aligned} & f\left(\boldsymbol{z}_{t+1}\right) \leq f\left(\boldsymbol{z}_t\right)+\frac{\eta^2 H^2 C^2 L}{2(1-\gamma)^2} \\ & -\frac{\eta}{1-\gamma} \sum_{k=1}^K p_k \sum_{s=0}^{H-1}\left\langle \nabla f_k\left(\boldsymbol{w}_{t-\tau_k, s} ; \xi_k^s\right), \nabla f\left(\boldsymbol{w}_t\right)\right\rangle \\ & -\frac{\eta}{1-\gamma} \sum_{k=1}^K p_k \sum_{s=0}^{H-1}\left\langle \nabla f_k\left(\boldsymbol{w}_{t-\tau_k, s} ; \xi_k^s\right), \nabla f\left(\boldsymbol{z}_t\right)-\nabla f\left(\boldsymbol{w}_t\right)\right\rangle . \end{aligned} \tag{63}$$

Note that $\left\langle \nabla f_k\left(\boldsymbol{w}_{t-\tau_k, s} ; \xi_k^s\right), \nabla f\left(\boldsymbol{w}_t\right)-\nabla f\left(\boldsymbol{z}_t\right)\right\rangle$ can be bounded using the Cauchy-Schwartz inequality and the smoothness of f as follows:

$$\begin{aligned} & \left\langle \nabla f_k\left(\boldsymbol{w}_{t-\tau_k, s} ; \xi_k^s\right), \nabla f\left(\boldsymbol{w}_t\right)-\nabla f\left(\boldsymbol{z}_t\right)\right\rangle \\ \leq & \left\|\nabla f_k\left(\boldsymbol{w}_{t-\tau_k, s}\right)\right\| \cdot\left\|\nabla f\left(\boldsymbol{w}_t\right)-\nabla f\left(\boldsymbol{z}_t\right)\right\| \\ \leq & C \times L \times\left\|\boldsymbol{w}_t-\boldsymbol{z}_t\right\|=C \times L \times\left\|\boldsymbol{u}_t\right\| \\ \stackrel{(a)}{=} & C \times L \times \frac{\gamma^2}{1-\gamma} \times\left\|\sum_{j=0}^{t-1} \gamma^j \boldsymbol{g}_{t-1-j}\right\| \\ \leq & \frac{\gamma^2 L C}{1-\gamma} \sum_{j=0}^{t-1} \gamma^j\left\|\boldsymbol{g}_{t-1-j}\right\| \\ \leq & \frac{\gamma^2 L C}{1-\gamma} \sum_{j=0}^{t-1} \gamma^j\left\|\eta \sum_{k=1}^K p_k \sum_{s=1}^H \nabla f_k\left(\boldsymbol{w}_{t-1-j-\tau_k, s} ; \xi_k^s\right)\right\| \\ \leq & \frac{\gamma^2}{1-\gamma} \sum_{j=0}^{t-1} \gamma^j L \eta H C^2 \\ = & \frac{\gamma^2}{1-\gamma} \times \frac{1-\gamma^t}{1-\gamma} \times L \eta H C^2, \end{aligned} \tag{64}$$

where (a) follows from solving the recurrence relation in (58). Therefore, we have

$$\begin{aligned} & f\left(\boldsymbol{z}_{t+1}\right) \\ \leq & f\left(\boldsymbol{z}_t\right)-\frac{\eta}{1-\gamma} \sum_{k=1}^K p_k \sum_{s=0}^{H-1}\left\langle \nabla f_k\left(\boldsymbol{w}_{t-\tau_k, s} ; \xi_k^s\right), \nabla f\left(\boldsymbol{z}_t\right)\right\rangle \\ & +\frac{\gamma^2}{(1-\gamma)^2} \times L \eta^2 H^2 C^2 \times \frac{1-\gamma^t}{1-\gamma}+\frac{L \eta^2 H^2 C^2}{2(1-\gamma)^2} . \end{aligned} \tag{65}$$

By taking an expectation on both sides of the above inequality, we have

$$\begin{aligned} \mathbb{E}_\xi\left[f\left(\boldsymbol{z}_{t+1}\right)\right] \leq & \mathbb{E}_\xi\left[f\left(\boldsymbol{z}_t\right)\right]+\frac{L \eta^2 H^2 C^2}{(1-\gamma)^2}\left(\frac{1}{2}+\gamma^2 \times \frac{1-\gamma^t}{1-\gamma}\right) \\ & -\frac{\eta H}{1-\gamma} \sum_{k=1}^K p_k \mathbb{E}_{\tau_k}\left[\left\langle \nabla f_k\left(\boldsymbol{w}_{t-\tau_k}\right), \nabla f\left(\boldsymbol{w}_t\right)\right\rangle\right], \end{aligned} \tag{66}$$

which results in the following

$$\begin{aligned} \mathbb{E}[f\left(\boldsymbol{z}_{t+1}\right)] \leq & \mathbb{E}[f\left(\boldsymbol{z}_t\right)] \underbrace{-\frac{\eta H}{1-\gamma} \mathbb{E}_{\tau_R}\left[\left\langle \nabla f\left(\boldsymbol{w}_{t-\tau_R}\right), \nabla f\left(\boldsymbol{w}_t\right)\right\rangle\right]}_{Q_9} \\ & +\frac{L \eta^2 H^2 C^2}{(1-\gamma)^2}\left(\frac{1}{2}+\gamma^2 \times \frac{1-\gamma^t}{1-\gamma}\right) . \end{aligned} \tag{67}$$

Under the RS policy, τ_R follows the distribution in (14), with which we can bound Q_9 as follows using the same set of tricks used in bounding the term Q_3 in (40):

$$\begin{aligned} Q_9= & -\frac{\eta H}{1-\gamma}\Big[\sum_{l=0}^t q_l\left\|\nabla f\left(\boldsymbol{w}_{t-l}\right)\right\|^2 \\ & +\sum_{l=0}^t q_l\left\langle \nabla f\left(\boldsymbol{w}_t\right)-\nabla f\left(\boldsymbol{w}_{t-l}\right), \nabla f\left(\boldsymbol{w}_{t-l}\right)\right\rangle \\ & +(1-\beta)^{t+1}\left\langle \nabla f\left(\boldsymbol{w}_0\right), \nabla f\left(\boldsymbol{w}_{t-l}\right)\right\rangle\Big] \\ \leq & -\frac{\eta H}{1-\gamma} \sum_{l=0}^t q_l\left\|\nabla f\left(\boldsymbol{w}_{t-l}\right)\right\|^2+\frac{H}{1-\gamma} \eta^2 L C^2 \mathbb{E}[\tau_R] \\ & +\eta H(1-\beta)^{t+1} C^2 \frac{1}{1-\gamma} . \end{aligned} \tag{68}$$

To this end, we can substitute (68) into (67) and obtain the following

$$\begin{aligned} \mathbb{E}[f\left(\boldsymbol{z}_{t+1}\right)] \leq & \mathbb{E}[f\left(\boldsymbol{z}_t\right)]-\frac{\eta H}{1-\gamma} \sum_{l=0}^t q_l \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{w}_{t-l}\right)\right\|^2\right] \\ & +\frac{L \eta^2 H^2 C^2}{(1-\gamma)^2}\left(\frac{1}{2}+\frac{\gamma^2}{1-\gamma} \times\left(1-\gamma^t\right)\right) \\ & +\frac{H L \eta^2 C^2}{1-\gamma} \mathbb{E}[\tau_R]+\frac{\eta H C^2}{1-\gamma}(1-\beta)^{t+1} . \end{aligned} \tag{69}$$

We telescope according to the above relationship and arrive at

$$\begin{aligned} & \frac{\eta H}{1-\gamma} \sum_{t=0}^{T-1} \sum_{l=0}^t q_l \min_{0 \leq t \leq T-1} \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{w}_t\right)\right\|^2\right] \\ \leq & f\left(\boldsymbol{w}_0\right)-f\left(\boldsymbol{w}^*\right)+\frac{H L \eta^2 C^2}{1-\gamma} \mathbb{E}[\tau_R] T+\frac{\eta H C^2}{1-\gamma} \sum_{t=0}^{T-1}(1-\beta)^{t+1} \\ & +\frac{L \eta^2 H^2 C^2}{(1-\gamma)^2}\left[\left(\frac{1}{2}+\frac{\gamma^2}{1-\gamma}\right) T-\sum_{t=0}^{T-1} \frac{\gamma^{t+2}}{1-\gamma}\right] \end{aligned} \tag{70}$$

and the result follows by substituting $\eta=1 / H \sqrt{T}$ to the above.

REFERENCES

[1] C. Feng, H. H. Yang, Z. Chen, D. Feng, Z. Wang, and T. Q. Quek, "On the convergence rate of federated learning over unreliable networks," in *2021 Computing, Communications and IoT Applications (ComComAp)*, Shenzhen, China, 2021, pp. 59–64.

[2] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *Workshop on Private Multi-Party Machine Learning*, Barcelona, Spain, Dec. 2016.

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

[4] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intelligent Syst. & Technology*, vol. 10, no. 2, pp. 1–19, Jan. 2019.

[5] J. Park, S. Samarakoon, A. Elgabli, J. Kim, M. Bennis, S.-L. Kim, and M. Debbah, "Communication-efficient and distributed learning over wireless networks: Principles and applications," *Proc. IEEE*, vol. 109, no. 5, pp. 796–819, May 2021.

[6] A. Imteaj, U. Thakker, S. Wang, J. Li, and M. H. Amini, "A survey on federated learning for resource-constrained iot devices," *IEEE Internet of Things J.*, vol. 9, no. 1, pp. 1–24, Jan. 2022.

[7] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Convergence of federated learning over a noisy downlink," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1422–1437, Mar. 2022.

[8] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6g: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, Nov. 2022.

[9] Z. Zhao, C. Feng, H. H. Yang, and X. Luo, "Federated learning-enabled intelligent fog-radio access networks: Fundamental theory, key techniques, and future trends," *IEEE Wireless Commun. Mag.*, vol. 27, no. 2, pp. 22–28, Apr. 2020.

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

[11] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, Mar. 2021.

[12] C. Feng, H. H. Yang, D. Hu, Z. Zhao, T. Q. S. Quek, and G. Min, "Mobility-aware cluster federated learning in hierarchical wireless networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8441–8458, Oct. 2022.

[13] Z. Chen, H. H. Yang, and T. Q. S. Quek, "Edge intelligence over the air: Two faces of interference in federated learning," *IEEE Commun. Mag.*, vol. 61, no. 12, pp. 62–68, Dec. 2023.

[14] C. Feng, H. H. Yang, S. Wang, Z. Zhao, and T. Q. S. Quek, "Hybrid learning: When centralized learning meets federated learning in the mobile edge computing systems," *IEEE Trans. Commun.*, vol. 71, no. 12, pp. 7008–7022, Dec. 2023.

[15] Z. Jia, Q. Wu, C. Dong, C. Yuen, and Z. Han, "Hierarchical aerial computing for internet of things via cooperation of haps and uavs," *IEEE Internet of Things J.*, vol. 10, no. 7, pp. 5676–5688, Apr. 2023.

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

[17] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *Int. Conf. Machine Learning (ICML)*, Vienna, Austria, Jul. 2020, pp. 5132–5143.

[18] T. Chen, G. B. Giannakis, T. Sun, and W. Yin, "LAG: Lazily aggregated gradient for communication-efficient distributed learning," in *Advance in Neural Information Process. Systems (NeurIPS)*, Montreal, Canada, Dec. 2018, pp. 1–6.

[19] W. Liu, L. Chen, Y. Chen, and W. Zhang, "Accelerating federated learning via momentum gradient descent," *IEEE Trans. Parallel & Dist. Syst.*, vol. 31, no. 8, pp. 1754–1766, Aug. 2020.

[20] H. H. Yang, Z. Liu, Y. Fu, T. Q. S. Quek, and H. V. Poor, "Federated Stochastic Gradient Descent Begets Self-induced Momentum," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Singapore, May 2022.

[21] K. Yang, T. Fan, T. Chen, Y. Shi, and Q. Yang, "A quasi-newton method based vertical federated learning framework for logistic regression," *Available as ArXiv:1912.00513*, Dec. 2019.

[22] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, Jun. 2019.

[23] W. Wen, Z. Chen, H. H. Yang, W. Xia, and T. Q. S. Quek, "Joint scheduling and resource allocation for hierarchical federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 5857–5872, Jan. 2022.

[24] 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.

[25] M. Salehi and E. Hossain, "Federated learning in unreliable and resource-constrained cellular wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5136–5151, May 2021.

[26] J. Ren, Y. He, D. Wen, G. Yu, K. Huang, and D. Guo, "Scheduling in cellular federated edge learning with importance and channel awareness," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7690–7703, Nov. 2020.

[27] 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.*, vol. 20, no. 6, pp. 3643–3658, June 2021.

[28] H. H. Yang, A. Arafa, T. Q. S. Quek, and H. V. Poor, "Age-based scheduling policy for federated learning in mobile edge networks," in *Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 8743–8747.

[29] M. M. Amiri and D. Gündüz, "Computation scheduling for distributed machine learning with straggling workers," *IEEE Trans. Signal Process.*, vol. 67, no. 24, pp. 6270–6284, Dec. 2019.

[30] H. Yang, M. Fang, and J. Liu, "Achieving linear speedup with partial worker participation in non-iid federated learning," *Available as ArXiv:2101.11203*, 2021.

[31] X. Gu, K. Huang, J. Zhang, and L. Huang, "Fast federated learning in the presence of arbitrary device unavailability," in *Advances in Neural Information Process. Systems*, vol. 34, New Orleans, USA, Dec. 2021, pp. 12 052–12 064.

[32] D. Jhunjhunwala, P. Sharma, A. Nagarkatti, and G. Joshi, "Fedvarp: Tackling the variance due to partial client participation in federated learning," in *Conf. Uncertainty Artificial Intelligence (UAI)*, ser. Proc. Machine Learning Research (PMLR), vol. 180, 01–05 Aug. 2022, pp. 906–916.

[33] S. Wang and M. Ji, "A unified analysis of federated learning with arbitrary client participation," in *Advances in Neural Information Process. Systems (NeurIPS)*, vol. 35, New Orleans, USA, Dec. 2022, pp. 19 124–19 137.

[34] H. Yang, X. Zhang, P. Khanduri, and J. Liu, "Anarchic federated learning," in *Int. Conf. Machine Learning (ICML)*, ser. Proc. Machine Learning Research, vol. 162. Proc. Machine Learning Research (PMLR), 17–23 Jul 2022, pp. 25 331–25 363.

[35] M. Rostami and S. S. Kia, "Federated learning using variance reduced stochastic gradient for probabilistically activated agents," in *2023 American Control Conference (ACC)*, San Diego, CA, USA, May 2023, pp. 861–866.

[36] J. Ren, Y. He, D. Wen, G. Yu, K. Huang, and D. Guo, "Scheduling for cellular federated edge learning with importance and channel awareness," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7690–7703, Aug. 2020.

[37] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Int. Conf. Mach. Learn. (ICML)*, Atlanta, GA, Jun. 2013, pp. 1310–1318.

[38] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.