Delay Sensitive Hierarchical Federated Learning With Stochastic Local Updates

Abdulmoneam Ali [ID], *Graduate Student Member, IEEE*, and Ahmed Arafa [ID], *Member, IEEE*

Abstract—The impact of local averaging on the performance of federated learning (FL) systems is studied in the presence of *communication delay* between the clients and the parameter server. To minimize the effect of delay, clients are assigned into different groups, each having its own *local* parameter server (LPS) that aggregates its clients' models. The groups' models are then aggregated at a *global* parameter server (GPS) that only communicates with the LPSs. Such setting is known as *hierarchical* FL (HFL). Unlike most works in the literature, the number of local and global communication rounds in our work is randomly determined by the (different) delays experienced by each group of clients. Specifically, the number of local averaging rounds is tied to a wall-clock time period coined the *sync time S*, after which the LPSs synchronize their models by sharing them with the GPS. Such sync time S is then reapplied until a global wall-clock time is exhausted. First, an upper bound on the deviation between the updated model at each LPS with respect to that available at the GPS is derived. This is then used as a tool to derive the convergence analysis of our proposed delay-sensitive HFL algorithm, first at each LPS individually, and then at the GPS. Our theoretical convergence bound showcases the effects of the whole system's parameters, including the number of groups, the number of clients per group, and the value of S. Our results show that the value of S should be carefully chosen, especially since it implicitly governs how the delay statistics affect the performance of HFL in situations where training time is restricted.

Index Terms—Hierarchical federated learning, limited-time training, delay-sensitive learning, random global iterations, random local iterations, convergence analysis.

I. Introduction

FEDERATED Learning (FL) is a distributed machine learning training system in which edge devices (clients) collaboratively train a model of interest based on their locally stored datasets. A central node (parameter server) orchestrates the learning process by collecting the clients' parameters for aggregation [2]. Due to its data privacy preserving and bandwidth saving nature, FL has attracted a lot of attention and has been used in diverse applications including healthcare and mobile services.

Challenges and Related Work. In order to successfully deploy FL in communication networks, lots of challenges should be addressed. These include: the computing capabilities of the clients; the communication overhead between the clients and the parameter server; and the system heterogeneity, whether in the clients' communication channels or their data statistics.

One way to overcome the resource-constrained capabilities of clients and the limited channel bandwidth is to employ quantization, sparsification, and compression when the size of the learning model is too large [3], [4]. Another challenge often addressed in the literature is related to the limited available spectrum that hinders the simultaneous participation of all clients. For that, client scheduling and its consequences on the system's performance becomes crucial [5], [6], [7].

Among all challenges, communication remains to be the bottleneck issue, and various solutions have been proposed in the literature to mitigate it. One of these solutions is to conduct several local updates at the clients' side before communicating with the parameter server [8], [9], [10], [11]. Another solution is to introduce intermediate parameter servers, denoted as local parameter servers (LPSs), between the clients and the (now) global parameter server (GPS). Such setting of FL is known in the literature as the *hierarchical* FL (HFL) setting [12]. The main advantage of having LPSs close to clients is to reduce the latency and energy required to communicate with the GPS [13]. In [14], a joint resource allocation and client association problem is formulated in an HFL setting, and then solved by an iterative algorithm. Moreover, the authors in [15] utilize the bottleneck time at the GPS by allowing the devices to perform extra local updates while they are waiting for the new global model from the GPS. To overcome the dispersion between the weights evaluated based on the latest global model and those at the stale model, the authors propose a linear global-local model combiner scheme to address this challenge.

In these mentioned works, the authors analyze their systems while assuming a fixed number of local iterations and global communication rounds. However, in more realistic scenarios, the number of local iterations may vary from one global communication round to another, depending on the dynamic nature of the (wireless) communication channel and the different computational capabilities of the edge devices. Moreover, the number of global communication rounds can also vary if the training time is restricted. One scenario in which this is the case is when model training has to be conducted during non-congested periods of the network.

Contributions. Motivated by the aforementioned endeavors and to cope with the very low latency service requirements in 6G networks (and beyond), in this paper we focus on HFL for delay-sensitive communication networks. We study FL settings that have an additional requirement of conducting training within a predefined deadline. Such scenario is relevant for, e.g., energy-limited clients whose availability for long times is not always guaranteed. To enforce the system to abide by this constraint, the number of local training updates will be determined by a wall-clock time. Specifically, we define a *sync time S* within which the LPSs are allowed to aggregate the parameters they receive from their groups' clients. Each local iteration consumes a random group-specific delay, and hence the total number of local updates within S will also be random, and could possibly be different across groups. This dissimilarity in the delay statistics is introduced to capture, e.g., the effects of wireless channels and different computational resources among different group clients. Following the deadline S, the LPSs forward their models to the GPS.

We set another time constraint at the GPS denoted the *system time T*. This is the total time allowed for the overall HFL system to perform the training. Different values of S and T will lead to a different number of local and global updates. By controlling S, we also control how many times the clients communicate with the GPS: more local iterations lead to fewer global ones and vice versa. This is different from the existing works that assume that the global communication rounds are predefined independently from the number of local updates. Thus, one of the main questions we address in this work is:

When would it be preferable for an HFL system to perform more local iterations than global ones, with local iterations being relatively faster yet leading to possibly lower model accuracy, and vice versa?

We present a thorough theoretical convergence analysis for the proposed HFL setting for non-convex loss functions. Our results show how the different system parameters affect the accuracy, namely the wall clock times S and T, the number of groups, and the number of clients per group. Various experiments are performed on different models and datasets to show how to optimize the sync time S based on the other system parameters. Our system model is depicted in Fig. 1.

Our work relaxes the assumption in [12] regarding the non-existence of communication constraints. In particular, different from [12], we incorporate the delay induced by communication and computation processing between LPSs and their associated clients in addition to that in between LPSs and the GPS. Moreover, we point out the trade-off between the number of local iterations and global communication rounds, and propose an approach to balance between them by optimizing the sync time S. We also add another lens when comparing HFL with centralized FL (which is different from the results in [13]). Specifically, depending on the system parameters and delay statistics, we show that there are scenarios in which direct communication with the GPS (centralized FL) can outperform communicating with the LPSs (HFL).

We now summarize our main contributions and novelty:
- To the best of our knowledge, this is the first work that considers stochastic local and global iterations in HFL
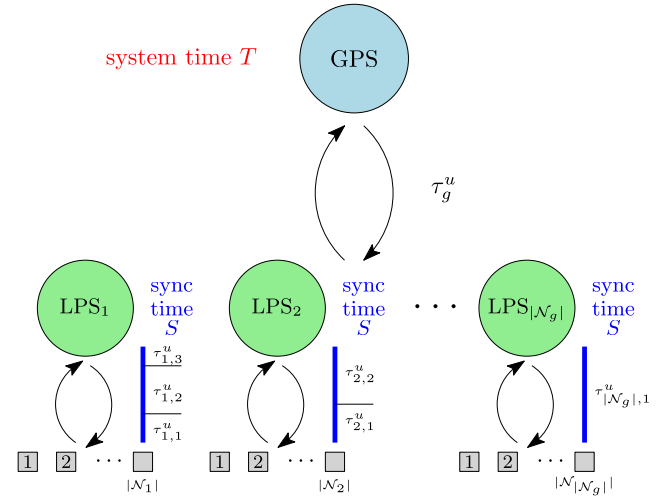

Fig. 1. System model of delay sensitive HFL.

(as opposed to being predefined prior to training); the number of iterations is determined by the induced delay between the LPSs and their associated clients, in addition to that in between the LPSs and the GPS.
- We show that there is an intrinsic trade-off between local and global iterations in delay-sensitive HFL settings; the more time you spend in local iterations, the less you spend in global ones, and vice versa. Hence, we provide insights on how to characterize the effect of such a trade-off on the training accuracy in HFL. This is captured by carefully selecting the proposed sync time S across LPSs.
- Our theoretical results characterize the consequences of considering a wall-clock training time, together with random local iterations and global communication rounds, on the training accuracy of non-convex loss functions in HFL settings.
- Instead of relying on similarity and dissimilarity assumptions about the updated models at the LPSs and the GPS (as done in most of the literature), we derive our own bound to capture the divergence between them as a function of the system parameters. Such a bound plays an instrumental role in proving our convergence results and reveals insights about how the system parameters can be designed to alleviate heterogeneity between groups.
- In addition to theoretical analyses, we carry out extensive experiments on various datasets and different scenarios to show the effects of choosing the sync time S, along with the remaining system parameters, on delay-sensitive HFL settings.

Notation and Organization: $\mathbb{R}$ denotes the real number field; $\|\cdot\|$ denotes the Euclidean norm; $\langle x, y \rangle$ denotes the inner product between two vectors x and y; $\mathbb{E}$ denotes statistical expectation, while $\mathbb{E}_{|x} \|\cdot\|$ represents the conditional expectation given x.

The remainder of the paper is organized as follows. Section II presents the system model and our proposed HFL algorithm. Theoretical convergence analyses are derived in Section III, and verified via extensive simulations in Section IV. Section V concludes the paper.

II. SYSTEM MODEL

We consider an HFL system with a global PS (GPS) and a set of local PSs (LPSs), $\mathcal{N}_g$, that serve a number of clients, see Fig. 1. Clients are distributed across different LPSs to form clusters (groups), in which a client can only belong to one group, and may only communicate with one specific LPS. Denoting by $\mathcal{N}_i$ the set of clients in group i, the total number of clients in the system is $\sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|$, with $|\cdot|$ denoting cardinality. Each client has its own dataset, and the data is independently and identically distributed (i.i.d.) among clients. The empirical loss function at the LPS of group $i \in \mathcal{N}_g$ is defined as follows:

$$f_i(x) \triangleq \frac{1}{|\mathcal{N}_i|}\sum_{k\in\mathcal{N}_i} F_{i,k}(x), \quad i \in \mathcal{N}_g, \tag{1}$$

where $F_{i,k}(x)$ is the loss function at client k in group i. The goal of the HFL system is to minimize a global loss function:

$$f7(x) \triangleq \frac{1}{\sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|}\sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|f_i(x)$$
$$= \frac{1}{\sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|}\sum_{i\in\mathcal{N}_g}\sum_{k\in\mathcal{N}_i} F_{i,k}(x). \tag{2}$$

The global loss function is minimized over a number of global communication rounds between the GPS and the LPSs. At the beginning of the uth global round, the GPS broadcasts the global model, $x^u \in \mathbb{R}^d$, with d representing the model dimension, to the LPSs. The LPSs then forward x^u to their associated clients, which is used to run a number of SGD steps based on their own local datasets. After each SGD step, the clients share their models with their LPS, which aggregates them and broadcasts them back locally to its clients. We call this local round trip a local iteration. We further illustrate how the global rounds and local iterations interact as follows. Let $x_i^{u,l}$ denote the model available at LPS i after local iteration l during global round u, and let $x_{i,k}^{u,l}$ denote the corresponding local model of client k of group i. We now have the following equations that build up the models:

$$x_i^{u,0} = x^u, \quad \forall i \in \mathcal{N}_g, \tag{3}$$
$$x_{i,k}^{u,0} = x_i^{u,0}, \tag{4}$$
$$x_{i,k}^{u,l} = x_i^{u,l-1} - \alpha\,\tilde{g}_{i,k}\left(x_i^{u,l-1}\right), \forall k \in \mathcal{N}_i, \tag{5}$$

where α is the learning rate, and $\tilde{g}_{i,k}$ is an unbiased stochastic gradient evaluated at $x_i^{u,l-1}$. After the lth SGD step, LPS i collects $\{x_{i,k}^{u,l}\}$ from its associated clients and aggregates them to get the lth local model,

$$x_i^{u,l} = \frac{1}{|\mathcal{N}_i|}\sum_{k\in\mathcal{N}_i} x_{i,k}^{u,l}, \tag{6}$$

which is shared with its clients to initialize SGD step $l+1$. Each local iteration takes a random time to be completed. This includes the time for broadcasting the local model by the LPS to its clients, the SGD computation time, and the aggregation time. Let $\tau_{i,l}^u$ denote the wall-clock time elapsed during local iteration l for group i in global round u. The time
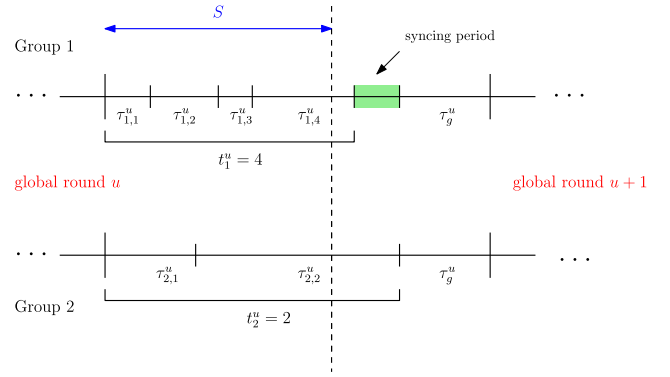

Fig. 2. Example sample path of global rounds and local iterations of 2 groups with wall-clock times considerations.

$\tau_{i,l}^u$ encompasses the training, communication, and aggregation time between the LPS and its associated clients. We assume that $\tau_{i,l}^u$'s are i.i.d. across local iterations l and global rounds u, but may not be identical across groups i. This is motivated by the different channel delay statistics that each group may experience when communicating with its LPS. In addition to that, each group may have clients with heterogeneous computational capabilities. These two factors together hinder one group to (statistically) do an identical number of local updates like other groups. We define a *sync time, S*, that represents the allowed local training time for all groups. After the sync time, the LPSs need to report their local models to the GPS, and thereby ending the current global round. During global round u, and within the sync time S, group i will then conduct a random number of local iterations given by

$$t_i^u \triangleq \min\left\{ n : \sum_{l=1}^{n} \tau_{i,l}^u \geq S \right\}, \quad i \in \mathcal{N}_g. \tag{7}$$

Observe that the statistics of t_i^u's are not identical across groups, see Fig. 2 for an example sample path during global round u. Based on (7), it is clear that the value of S determines the number of local iterations that each group conducts. After the t_i^u local iterations are done, and using (3)–(6), LPS i will have acquired the following model:

$$x_i^{u,t_i^u} = x_i^{u,0} - \frac{\alpha}{|\mathcal{N}_i|}\sum_{l=0}^{t_i^u-1}\sum_{k\in\mathcal{N}_i} \tilde{g}_{i,k}\left(x_i^{u,l}\right). \tag{8}$$

We consider a synchronous setting in which the GPS waits for all the LPSs to finish their local iterations before a global aggregation. Since LPSs incur different wall-clock times to collect their models, some of them may need to stay idle, waiting for others to finish. The GPS therefore starts aggregating the models after

$$\max_{i\in\mathcal{N}_g}\left\{ \sum_{l=1}^{t_i^u} \tau_{i,l}^u \right\} \tag{9}$$

time units from the start of the local iterations in global round u. We denote this period by the syncing period (see Fig. 2), which captures the straggler's effect in synchronous FL. When updating the GPS, LPS i sends the difference

between its final and initial models divided by the number of its local iterations performed [16], [17], i.e., it sends

$$\frac{1}{t_i^u}\left(x_i^{u,t_i^u} - x_i^{u,0}\right) = \frac{-\alpha}{|\mathcal{N}_i|t_i^u}\sum_{l=0}^{t_i^u-1}\sum_{k\in\mathcal{N}_i}\tilde{g}_{i,k}\left(x_i^{u,l}\right), i\in\mathcal{N}_g.$$
(10)

We note that the purpose of diving by t_i^u is to avoid *biasing* the global model and to avoid the objective inconsistency [17]. Moreover, normalization by the number of local iterations forces the aggregated model update to be the result of an *equal* contribution from all groups. To see this, observe that (cf. Assumption 2)

$$\mathbb{E}_{|\boldsymbol{t}_i^u}\frac{1}{t_i^u}\left(x_i^{u,t_i^u} - x_i^{u,0}\right)$$

$$= \frac{-\alpha}{|\mathcal{N}_i|t_i^u}\sum_{l=0}^{t_i^u-1}\sum_{k\in\mathcal{N}_i}\nabla F_{i,k}\left(x_i^{u,l}\right)$$

$$= \frac{-\alpha}{t_i^u}\sum_{l=0}^{t_i^u-1}\nabla f_i\left(x_i^{u,l}\right),$$
(11)

where $\mathbb{E}_{|\boldsymbol{t}_i^u}$ denotes the conditional expectation given the vector $\boldsymbol{t}_i^u \triangleq \{t_i^{u'}\}_{u'=1}^u$.

The GPS then updates its global model as

$$x^{u+1} = x^u + \sum_{i\in\mathcal{N}_g}\frac{|\mathcal{N}_i|}{\sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|}\frac{1}{t_i^u}\left(x_i^{u,t_i^u} - x_i^{u,0}\right)$$

$$= x^u - \frac{\alpha}{\sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|}\sum_{i\in\mathcal{N}_g}\frac{1}{t_i^u}\sum_{l=0}^{t_i^u-1}\sum_{k\in\mathcal{N}_i}\tilde{g}_{i,k}\left(x_i^{u,l}\right),$$
(12)

and broadcasts x^{u+1} to the LPSs to begin global round $u + 1$. We assume that the global aggregation and broadcasting processes consume i.i.d. τ_g^u's wall-clock times. An example of the HFL setting considered is depicted in Fig. 2.

The overall HFL training process stops after a total system time T. The value of T represents the time budget allowed for training in applications sensitive to delays. Within T, the total number of global rounds will be given by

$$\mathcal{U} \triangleq \min\left\{m : \sum_{u=1}^m\left(\max_{i\in\mathcal{N}_g}\left\{\sum_{l=1}^{t_i^u}\tau_{i,l}^u\right\} + \tau_g^u\right) \geq T\right\}.$$
(13)

Notably in (13), the number of global iterations $\mathcal{U}$, will be dominated by the number of local iterations $\{t_i^u\}$, the delay at the GPS τ_g^u, and the training time constraint T. We coin the proposed algorithm *delay sensitive HFL* which is summarized in Algorithm 1. We also include a list of the key notations in Table I for better presentation.

In the sequel, we analyze the performance of Algorithm 1 in terms of the wall-clock times, number of clients and other parameters of the system. We then discuss how to choose the sync time S to ensure better learning outcomes.

TABLE I
KEY NOTATIONS AND SYSTEM PARAMETERS

Symbol	Definition
$\mathcal{N}_g$	Set of local PSs (LPSs)
$\mathcal{N}_i$	Set of clients in group i
t_i^u	Number of local iterations (averaging) performed by group i within global round u
x^u	Global model after u global iterations
$x_{i,k}^{u,l}$	Local model of client k of group i at local iteration l within global round u
$\tau_{i,l}^u$	Wall-clock time elapsed during local iteration l for group i in global round u
τ_g^u	Wall-clock time elapsed at the GPS to update the global model in global round u
S	Sync time; allowed local training time for all groups
T	System time; overall allowed training time
$\mathcal{U}$	Number of global iterations

Algorithm 1 Delay Sensitive HFL

1: **Input:** learning rate α, system time T, sync time S
2: **Output:** global aggregated model $x^{\mathcal{U}}$
3: **Initialization:** $\bar{T}, u \leftarrow 0$
4: **while** $\bar{T} \leq T$ **do**
5:   Global Broadcast: $x_i^{u,0} \leftarrow x^u, \forall i\in\mathcal{N}_g$
6:   **for** $i\in\mathcal{N}_g$ **do**
7:     $t_i^u \leftarrow 0, \bar{t} \leftarrow 0$
8:     **while** $\bar{t} \leq S$ **do**
9:       **for** $k\in\mathcal{N}_i$ **do**
10:         SGD Update: $x_{i,k}^{u,l} = x_i^{u,l-1} - \alpha\tilde{g}_{i,k}(x_i^{u,l-1})$
11:       **end for**
12:       Local Aggregation: $x_i^{u,l} = \frac{1}{|\mathcal{N}_i|}\sum_{k\in\mathcal{N}_i}x_{i,k}^{u,l}$
13:       Group Broadcast: $x_{i,k}^{u,l} = x_i^{u,l}$
14:       Local updates increment: $t_i^u \leftarrow t_i^u + 1,$
15:       $\bar{t} \leftarrow \bar{t} + \tau_i^u$
16:     **end while**
17:     Upload: $\frac{1}{t_i^u}(x_i^{u,t_i^u} - x_i^{u,0})$
18:   **end for**
19:   Global Update:
20:     $x^{u+1} = x^u + \sum_{i\in\mathcal{N}_g}\frac{|\mathcal{N}_i|}{\sum_{j\in\mathcal{N}_g}|\mathcal{N}_j|}\frac{1}{t_i^u}(x_i^{u,t_i^u} - x_i^{u,0})$
21:   System Time Update: $\bar{T} \leftarrow \max_i\{\sum_{j=1}^{(t_i^u)}\tau_{i,j}^u\} + \tau_g,$
22:   $u \leftarrow u + 1$
23: **end while**

III. MAIN RESULTS

In this section, we present the convergence analysis for the proposed HFL setting. We have the following typical assumptions about the loss function and SGD [13]:

Assumption 1 (Smoothness): Loss functions are L-smooth: $\exists L > 0$ such that, $\forall x, y \in \mathbb{R}^d$:

$$F_{i,k}(y) \leq F_{i,k}(x) + \langle\nabla F_{i,k}(x), y - x\rangle + \frac{L}{2}\|y - x\|^2, \forall i, k.$$
(14)

Assumption 2 (Unbiased Gradient): The gradient estimate at each client satisfies

$$\mathbb{E}\tilde{g}_{i,k}(x) = \nabla F_{i,k}(x), \quad \forall i, k. \tag{15}$$

Assumption 3 (Bounded Gradient): There exists a constant $G > 0$ such that the stochastic gradient's second moment is bounded as

$$\mathbb{E}\left\|\tilde{g}_{i,k}(x)\right\|^2 \le G^2, \quad \forall i, k. \tag{16}$$

Assumption 4 (Bounded Variance): There exists a constant $\sigma > 0$, such that the variance of the stochastic gradient is bounded as

$$\mathbb{E}\left\|\tilde{g}_{i,k}(x) - \nabla F_{i,k}(x)\right\|^2 \le \sigma^2, \quad \forall i, k. \tag{17}$$

It is worth noting that we conduct our analysis *without* assuming convexity of the loss function at any entity in the system. According to our proposed algorithm, after each global round, the group clients will resume their local training from the aggregated global model instead of the their latest local one. Hence, we need to quantify the *deviation* between the two parameter models through the following lemma (the proof is in Appendix B):

Lemma 1: For $0 \le \alpha \le \frac{1}{L}$, the delay sensitive HFL algorithm satisfies the following, $\forall u, i$:

$$\mathbb{E}_{|t_i^u}\left\|x^{u+1} - x_i^{u,t_i^u}\right\|^2 \le$$
$$2\alpha^2\left((t_i^u)^2 + \frac{|\mathcal{N}_g|\sum_{j\in\mathcal{N}_g}|\mathcal{N}_j|^2}{\left(\sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|\right)^2}\right)G^2. \tag{18}$$

Remark 1: The first term in the bound in Lemma 1 represents the contribution of group i while the second one reflects the impact of all groups in the deviation between the parameter models. It is obvious that more local iterations lead to more deviation between the local and the global models. Note that local iterations are the sole determinant of the deviation in case of having one group only (e.g., when there is no hierarchy); having two or more groups carries an additive effect on the deviation as seen in the second term.

Remark 2: In case of having only one group in the system, one gets a strictly smaller upper bound than that in [18], which is given by $4\alpha^2(t_i^u)^2 G^2$ (almost two times the bound in (18) for $|\mathcal{N}_g| = 1$ for large values of t_i^u).

Remark 3: The lemma suggests a way to overcome the effects of the heterogeneity of the number of local iterations conducted by each group on the divergence. By setting a different learning rate for each group that is inversely proportional to the number of local iterations (e.g., $\alpha^2 \le \frac{1}{2((t_i^u)^2+|\mathcal{N}_g|)G^2}$ for group i), one can achieve the same divergence bound (≤ 1) across all groups.

Lemma 1 serves as a building block for our main convergence theorems of the proposed delay sensitive HFL. These are mentioned next (with proofs in Appendices C and D).

Theorem 1 (Convergence Analysis per Group): For $0 \le \alpha \le \frac{1}{L}$, the delay sensitive HFL algorithm achieves the following group i bound for a given $\mathcal{U}$:

$$\frac{1}{\sum_{u=1}^{\mathcal{U}} t_i^u} \sum_{u=1}^{\mathcal{U}} \sum_{l=1}^{t_i^u} \mathbb{E}_{|t_i^u}\left\|\nabla f_i\left(x_i^{u,l-1}\right)\right\|^2$$
$$\le \frac{2}{\alpha \sum_{u=1}^{\mathcal{U}} t_i^u}\left(\mathbb{E}_{|t_i^u} f_i\left(x_i^1\right) - \mathbb{E}_{|t_i^u} f_i\left(x_i^{\mathcal{U},t_i^{\mathcal{U}}}\right)\right) + \frac{\alpha L \sigma^2}{|\mathcal{N}_i|}$$
$$+ \left(\frac{1}{\alpha \sum_{u=1}^{\mathcal{U}} t_i^u} + \frac{2(L+1)\kappa\alpha}{\sum_{u=1}^{\mathcal{U}} t_i^u}\right)(\mathcal{U}-1) G^2$$
$$+ \frac{2(L+1)\alpha}{\sum_{u=1}^{\mathcal{U}} t_i^u} \sum_{u=1}^{\mathcal{U}-1} (t_i^u)^2 G^2, \tag{19}$$

where the term κ is given by $\kappa \triangleq \frac{|\mathcal{N}_g|}{(\sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|)^2}\sum_{j\in\mathcal{N}_g}|\mathcal{N}_j|^2$.

Remark 4: Notably, setting $\mathcal{U} = 1$ means that the groups will work *individually*. The result of Theorem 1 shows that convergence is still guaranteed in this isolated case by choosing $0 \le \alpha \le \min\{\frac{1}{L}, \frac{1}{\sqrt{t_i^1}}\}$.

Theorem 2 (Global Convergence Analysis): For $0 \le \alpha \le \frac{1}{L}$, the delay sensitive HFL algorithm achieves the following global bound for a given $\mathcal{U}$:

$$\frac{1}{\mathcal{U}} \sum_{u=1}^{\mathcal{U}} \mathbb{E}_{|t_i^u}\|\nabla f(x^u)\|^2$$
$$\le \frac{2}{\alpha}\frac{1}{\mathcal{U}}\left(\mathbb{E}_{|t_i^u} f\left(x^1\right) - \mathbb{E}_{|t_i^u} f\left(x^{\mathcal{U}+1}\right)\right)$$
$$+ \frac{12\alpha^2 L^2 G^2 |\mathcal{N}_g|^2}{\left(\sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|\right)^4}\left(\sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|^2\right)^2$$
$$+ \frac{\alpha L |\mathcal{N}_g| \sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|^2 \sigma^2}{\left(\sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|\right)^2}$$
$$+ \frac{1}{\mathcal{U}}\sum_{u=1}^{\mathcal{U}} \frac{12\alpha^2 L^2 |\mathcal{N}_g|}{\left(\sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|\right)^2} \sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|^2\left(t_i^{u-1}\right)^2$$
$$+ \frac{1}{\mathcal{U}}\sum_{u=1}^{\mathcal{U}} \frac{4\alpha^2 L^2 G^2 |\mathcal{N}_g|}{\left(\sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|\right)^2} \sum_{i\in\mathcal{N}_g}|\mathcal{N}_i|^2\frac{1}{t_i^u}\sum_{l=0}^{t_i^u-1} l^2. \tag{20}$$

Remark 5: The significance of controlling S is reflected in the values of $\{t_i^u\}$ and $\mathcal{U}$: increasing S allows more local iterations, $\{t_i^u\}$, between the LPSs and their associated clients, at the expense of performing less global communication rounds, $\mathcal{U}$, as a consequence of the system time T being limited, and vice versa. The value of S affects the convergence bound in (20), since all the terms except the second one are functions of t_i^u and $\mathcal{U}$.

Remark 6: Since the terms including $\{t_i^u\}$ and $\mathcal{U}$ are in the form: $\sum_{u=1}^{\mathcal{U}} t_i^u$, this gives us an insight that one can tune
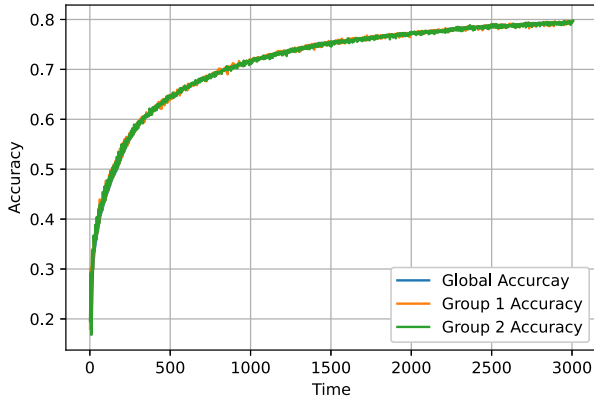
(a) Performance over the whole training time.
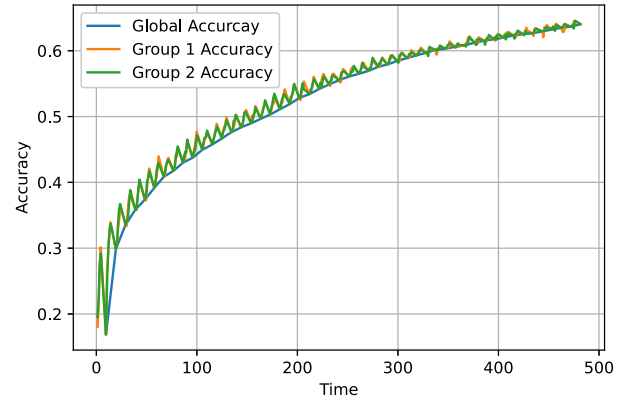
(b) Performance at the beginning of training.

Fig. 3. HFL: 10 clients per group, parameters $\{0.09, 0.1, 0.009, 0.01, 1, 3, 0.05, 0.1\}$, and $S = 5$.

S to obtain similar performances under different individual values of $\{t_i^u\}$ and $\mathcal{U}$. This is discussed in more detail in our experimental results.

Remark 7: Although we obtain a convergence bound that is a function of S, reflected in $\{t_i^u\}$ and $\mathcal{U}$, its optimal value is not directly obtainable analytically from the bound. In our experiments, we show how different system scenarios lead to different choices of S to optimize the performance.

Now that we have seen how the sync time S controls the upper bounds in the theorems above by statistically controlling the number of local iterations, and in order to characterize the convergence rate, let us assume that there exists a *minimum* local iteration time for group i:

$$\tau_{i,l}^u \geq c_i, \ \text{a.s.,} \ \forall l, u. \tag{21}$$

Then, one gets a *maximum* number of local iterations:

$$t_i^u \leq t_i^{\max} \triangleq \lceil * \rceil \frac{S}{c_i}, \ \text{a.s.,} \ \forall u. \tag{22}$$

Based on this, one can get the following convergence rate (the proof is in Appendix E):

Corollary 1 (Speed of Convergence): For a given $\{t_i^{\max}\}$, setting $\alpha = \min\{\frac{1}{\sqrt{\mathcal{U}}}, \frac{1}{L}\}$, the delay sensitive HFL algorithm achieves $\frac{1}{\mathcal{U}} \sum_{u=1}^{\mathcal{U}} \mathbb{E}\|\nabla f(x^u)\|^2 \leq \mathcal{O}(\frac{1}{\sqrt{\mathcal{U}}})$.

Therefore, for a finite sync time S, as the training time T increases, the number of the global communication rounds $\mathcal{U}$ also increases, and hence Corollary 1 shows that the gradient converges to 0 sublinearly.

IV. EXPERIMENTS

In this section, we present our simulation results for the proposed delay sensitive HFL algorithm to verify the findings from the theoretical analysis.

Datasets and Models. We adopt a variety of data sets to demonstrate our findings and emphasize the effectiveness of our proposed algorithm. We consider an image classification supervised learning task on the CIFAR-10 dataset [19], MNIST dataset and the Federated Extended MNIST dataset [20]. We also consider a word prediction task using the Shakespeare dataset [20]. We employ different
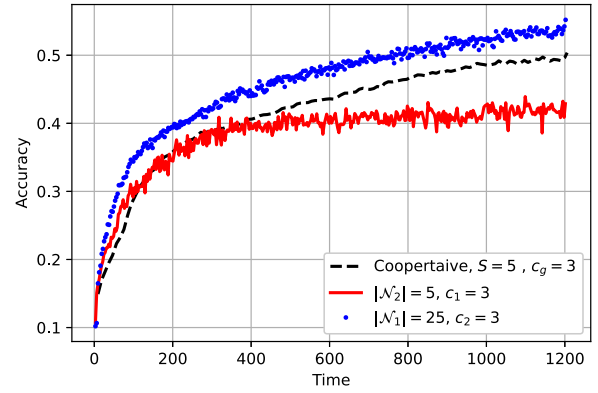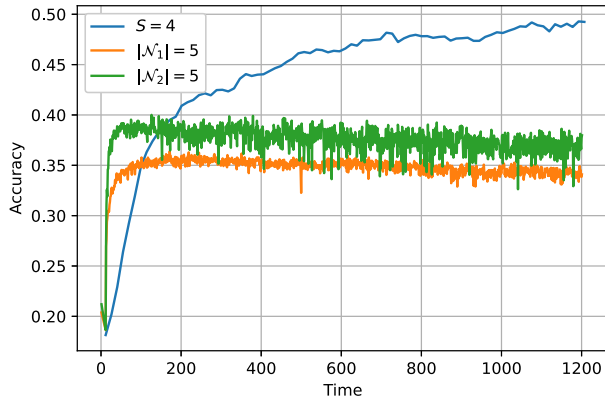

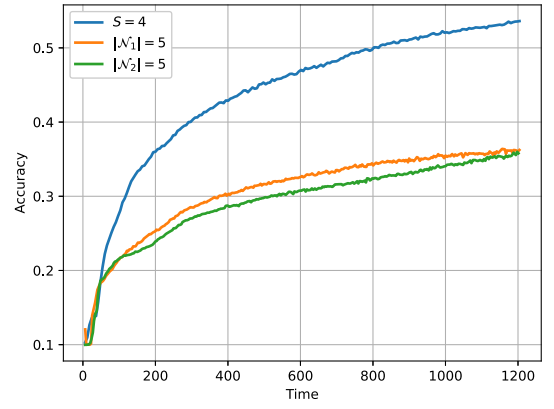Fig. 4. Significance of group cooperation under non-i.i.d data.

machine learning and deep learning models: (1) a convolution neural network (CNN) is adopted with two 5x5 convolution layers, two 2x2 max pooling layers, two fully connected layers with 120 and 84 units, respectively, ReLu activation, a final softmax output layer and cross entropy loss; (2) a multi-layer perceptron (MLP) with one hidden layer consisting of 50 neurons, with a logistic regression model; and (3) for the Shakespeare dataset, we train an recurrent neural network (RNN) for next-character-prediction [21].

Data Partitioning. In order to validate our proposed HFL algorithm, we test its performance under different data partitioning methods. These are a baseline data partitioning method, where data is distributed in an i.i.d. manner across clients, and another method, where data is distributed in a non-i.i.d. manner with various levels of data heterogeneity. The data partitioning is conducted based on the Dirichlet distribution and it follows the same data partitioning as in [20].

Linear Delay Model. We consider shifted exponential delays [22], [23]: $\tau_{i,l}^u \sim \exp(c_i, \lambda_i)$ and $\tau_g^u \sim \exp(c_g, \lambda_g)$. That is, $\mathbb{E}[\tau_{i,l}^u] = c_i + \frac{1}{\lambda_i}$ and $\mathbb{E}[\tau_g^u] = c_g + \frac{1}{\lambda_g}$. To capture the relation between the number of clients and the number of groups with the delay, we model the shift parameter for group i as a linearly increasing function in the number of clients in that group: $c_i = d \times |\mathcal{N}_i| + b, \ \forall i \in \mathcal{N}_g$, for some constants d and b. Similarly, we model the exponential rate as $1/\lambda_i = e \times |\mathcal{N}_i| + f, \ \forall i \in \mathcal{N}_g$. Thus, the intra-group delay increases as

(a) No common labels between two groups.

(b) Data distribution follows Dirichlet(0.1).

Fig. 5. Significance of group cooperation.

the number of its clients increases. We adopt a similar model for the delay between LPSs and the GPS: $c_g = d_g \times |\mathcal{N}_g| + b_g$ and $1/\lambda_g = e_g \times |\mathcal{N}_g| + f_g$. Thus, the more LPSs a system has, the more intense the communication bottleneck becomes while communicating with the GPS. We will state the delay model values in this order: $\{d, b, e, f, d_g, b_g, e_g, f_g\}$.

A. HFL Incentive and Motivation

We first start with an HFL setting with $|\mathcal{N}_g| = 2$ groups, with 10 clients per group training the CNN model described above on CIFAR-10. In Fig. 3, we show the evolution of both groups' accuracies and the global accuracy across time. The zoomed-in version in Fig. 3 shows the high (SGD) variance in the performance of the two groups especially during the earlier phase of training. Then, after more averaging with the GPS, the variance is reduced.

In Fig. 4, using the same CNN architecture on CIFAR-10, the significance of collaborative learning is emphasized. We change the number of clients per group, and run three experiments: one for each group in an isolated fashion, i.e., $S = T$, and one under the HFL setting ($S = 5$). First, while we do not conduct our theoretical analysis under heterogeneous data distribution, we consider a non-i.i.d. data distribution among the two groups in this setting (Dirichlet(0.1)), and we see that our proposed algorithm still *converges*. Second, it is clear that the performance of the group with fewer clients under heterogeneous data distribution and isolated learning will be deteriorated. However, aided by HFL, its performance improves, while the performance of the other group does not suffer greatly, promoting *fairness* among the groups.

To provide further insights on the significance of cooperation, we consider an extreme data heterogeneity setting in which each client only has two labels, with no common labels being shared among the two groups. In Fig. 5, we plot our results and show that HFL *converges* even when the groups are isolated (non-cooperative setting, i.e., $S = T$) yet since each group only has a strict subset of the total number of labels, they both converge to a lower accuracy compared to the one they get from cooperation. Further, in Fig. 5, cooperation is shown to be significant in a non-extreme scenario in which the data distribution is Dirichlet(0.1).
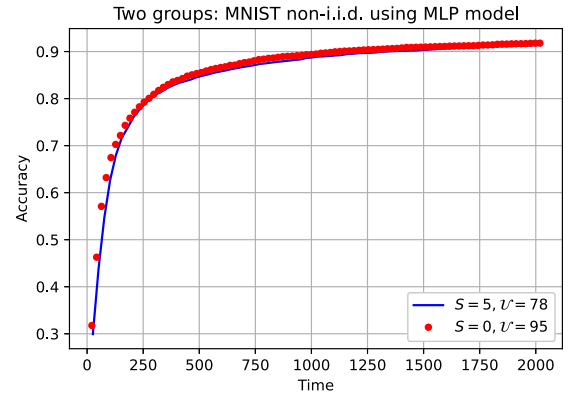
Fig. 6. Achieving the same performance as the baseline ($S = 0$) with fewer global rounds.

Different from Fig. 4, the cooperative result in Fig. 5 beats both groups' isolated training results.

Hence, from the previous figures we can summarize that our proposed HFL setting is able to enhance FL performance by reducing variance, improving fairness, and learning more collaboratively by offering indirect access to datasets that are otherwise unavailable to clients.

B. Significance of Choosing the Sync Time S

In this section, we show how choosing S can overcome the heterogeneity in the experienced delays among different groups, and can sustain a better performance for the whole system. Furthermore, we show that carefully choosing S can outperform the baseline of setting $S = 0$ (which corresponds to a centralized system (non-hierarchical)), with fewer visits to the GPS.

In Fig. 6, we set the linear delay model parameters to $\{10^{-3}, 0.5, 10^{-3}, 0.05, 5, 10, 10^{-3}, 0.05\}$, and train an MLP over the MNIST dataset. In this setting, we have two groups with 500 clients each. We can see that setting $S = 5$ time units achieves the same performance as that of the baseline ($S = 0$) yet with a *fewer number of global communication rounds*. The main reason behind this is that the cost of communicating with the GPS is relatively high in terms of the delay experienced in this setting.
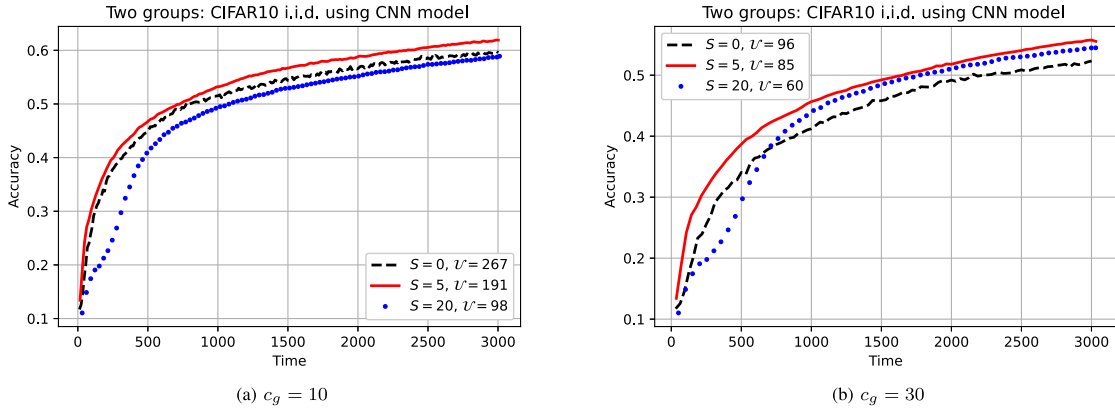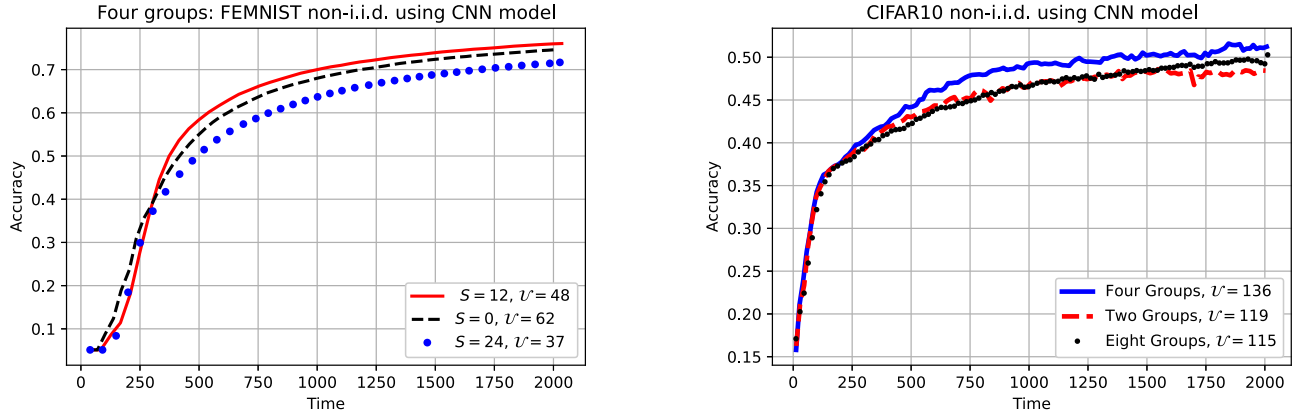
(a) $c_g = 10$

(b) $c_g = 30$

Fig. 7.   Impact of the global shift parameter c_g on choosing the sync time S.

Fig. 8.   Tuning S to beat the baseline in a four group setting.

Fig. 9.   Effect of the number of groups on the performance of HFL.

Next, we show that we can *outperform* that baseline ($S = 0$) by optimizing S. We train the CNN model on the CIFAR-10 dataset for 30 clients divided among 2 groups, together with the challenge of having a non-i.i.d. data distribution. In Fig. 7, we set the delay parameters to $\{0.01, 0.85, 10^{-3}, 0.085, 4, 2, 0.4, 0.2\}$. We see that $S = 5$ beats the baseline with a notable gap and with fewer number of global rounds. In addition, we also see that in case there were no strict system time constraints, one may set $S = 20$ which eventually approaches the baseline while saving 37% in terms of global rounds. We next change c_g to 30 and rerun our experiments and show the results in Fig. 7. We notice that $S = 5$ is still the optimum choice, but in case the system has an additional constraint on communicating with the GPS, $S = 20$ will be a better option, especially that the accuracy gain will not be sacrificed much. It is also worth noting that the system time T (training time budget) plays a significant role in choosing S: $S = 0$ (always communicate with the GPS) outperforms $S = 20$ as long as $T \lessapprox 750$, and the opposite is true afterwards. This means that in some scenarios, the hierarchical setting may not be the optimal setting (which is different from the findings in [13]). For instance, if the system has a stringent time constraint on learning, it would be better to communicate directly with GPS more frequently to get the advantage of learning the resulting models from different data.

To emphasize the significance of choosing S in more challenging system setups, we train the CNN model on the Federated E-MNIST dataset under non-i.i.d. data distribution. We consider a large number of clients, 3400, divided equally among $|\mathcal{N}_g| = 4$ groups, with delay parameters $\{0.1, 1.4, 0.01, 0.14, 4, 2, 0.5, 1\}$. In Fig. 8, the baseline $S = 0$ beats $S = 12$ and $S = 24$ for $T \lessapprox 300$, while eventually $S = 12$ achieves the best performance. We note that one reason for the relatively larger choices of S compared to previous cases is that the delay parameters are relatively larger. This emphasizes how S depends on the intra- and inter-group delays.

C. Clustering: Effect of the Number of Groups

Adopting the linear delay model presents an interesting phenomenon that we investigate in this section. That is, increasing the number of groups results in increasing the delay between the GPS and the LPSs (i.e., c_g and $1/\lambda_g$ will increase), but at the same time the delay between an LPS and its associated clients will decrease since there would be a smaller number of clients in a group (i.e., c_i and $1/\lambda_i$ will decrease). In Fig. 9, we consider a system of 64 clients divided equally between 2, 4, and 8 groups. The delay system parameters are given by $\{0.1, 1.4, 0.01, 0.14, 0.5, 5, 0.05, 0.5\}$. We show that under the same choice of $S = 6$, clustering clients into 4 groups is better than clustering them into 2 or 8 groups. We justify this since in the 4-group setting, the system strikes a balance between two extremes. That is, in the 2-group setting, the relatively high intra-group delay reduces
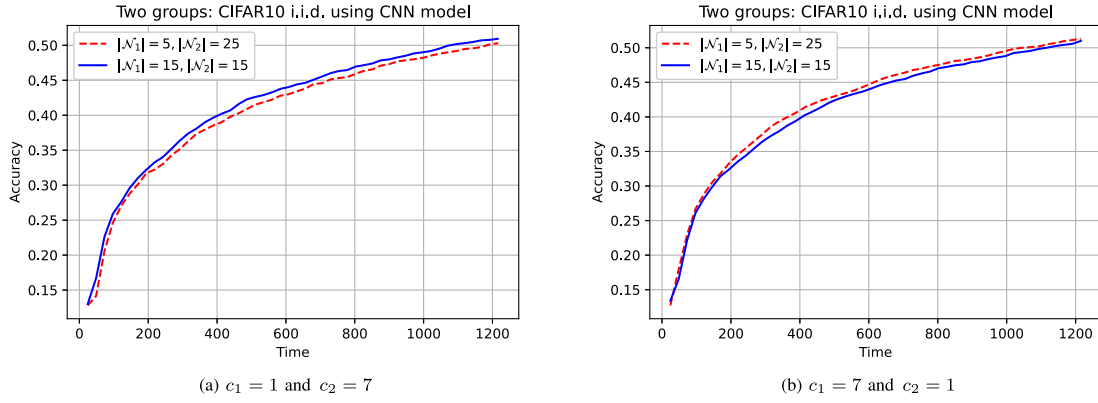
(a) $c_1 = 1$ and $c_2 = 7$

(b) $c_1 = 7$ and $c_2 = 1$

Fig. 10.   Impact of the groups' shift parameters c_1 and c_2 on the group-client association under $S = 8$ and $c_g = 10$.

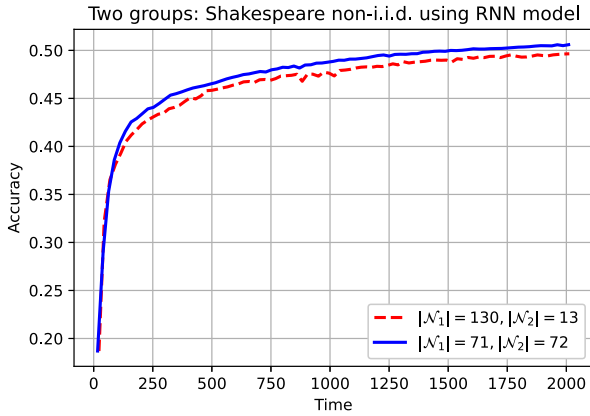Fig. 11.   The symmetric and asymmetric user association under linear delay model.

Fig. 12.   The effect of global shift parameter c_g under $S = 10$.

the number of local iterations (t_i^u), while in the 8-group setting the relatively high LPS-to-GPS delay reduces the number of global rounds $(\mathcal{U})$.

D. Effects of Client Association and Global Delay

In this section, we study the effects of the shift parameters on the overall performance of HFL, with a focus on client association and LPS-to-GPS delay.

In Fig. 10, the effect of the groups' shift parameters c_1 and c_2 on determining the optimal group-client association is investigated. The results show that it is not always optimal to cluster the clients evenly among the groups. In Fig. 10 for instance, we see that assigning less clients to a group with a relatively smaller shift parameter performs better than an equal assignment of clients among both groups; this observation is reversed in Fig. 10, in which a larger number of clients is assigned to the relatively slower LPS.

In Fig. 11, we train the RNN model on the Shakespeare dataset under non-i.i.d. data distribution with $S = 12$. The delay parameters are set to $\{0.1, 0.7, 0.001, 0.13, 3, 3, 0.01, 0.2\}$. We observe that the symmetric setting outperforms the non-symmetric one. The main reason behind this is that in the asymmetric case, the second group with fewer clients has to idly wait for the other group until it finishes its iterations, and hence the straggler effect becomes dominant.
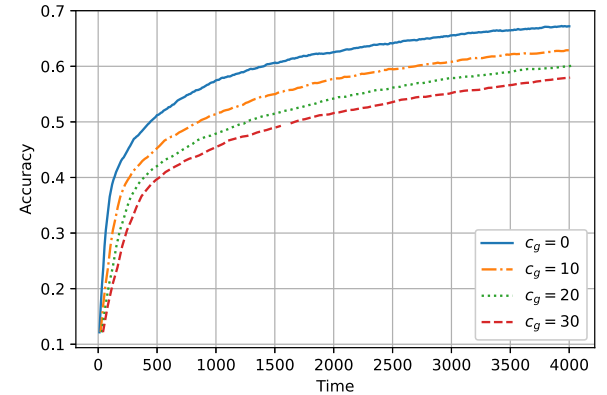
Finally, in Fig. 12, the impact of the global shift parameter c_g on the global accuracy is shown. As the global shift delay parameter increases, the performance gets worse. This is mainly because the number of global communication rounds with the GPS, $\mathcal{U}$, is reduced, preventing clients from getting the benefit of accessing other clients' learning models.

Future Investigation. Guided by our understanding from the convergence bounds and the simulation results, we observe that it is better to make the parameter S *variable* especially during the first global communication rounds. For instance, instead of fixing $S = 5$, we allow S to increase gradually with each round from 1 to 5, and then fix it at 5 for the remaining rounds. Our reasoning behind this is that the clients' models need to be *directed* towards global optimum, and not their local optima. Since this direction is done through the GPS, it is reasonable to communicate with it more frequently at the beginning of learning to push the local models towards the optimum direction. To investigate this setting, we train a logistic regression model over the MNIST dataset, and distribute it in a non-iid fashion over 500 clients per group. As shown in Fig. 13, the variable S approach achieves a higher accuracy than the fixed one, with the effect more pronounced as S increases.

V. CONCLUSION

A delay sensitive HFL algorithm has been proposed, in which the effects of wall-clock times and delays on the accuracy of FL are investigated. A sync time S governs how
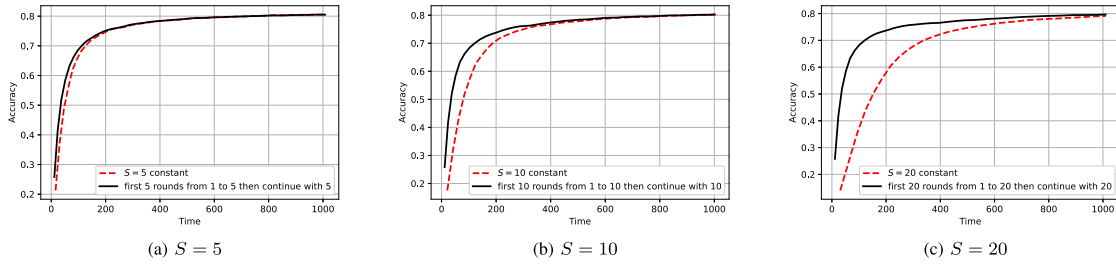
Fig. 13. Comparison between variable and fixed S with respect to the global learning accuracy.

many local iterations are allowed at LPSs before forwarding to the GPS, and a system time T constrains the overall training period. Our findings reveal that the optimal S depends on different factors such as the delays at the LPSs and the GPS, the number of clients per group, and the value of T. Multiple insights are drawn on the performance of HFL in time-restricted training settings.

Mitigating the impact of stragglers in delay sensitive HFL could be further mitigated via effective scheduling, or through changing the system into asynchronous learning. Both directions will be investigated in future work.

APPENDIX A
PRELIMINARIES

We will use the following relationships throughout our proofs without explicit reference:

- For any $x, y \in \mathbb{R}^n$, we have: $\langle x, y \rangle \leq \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2$.
- By Jensen's inequality, for $x_i \in \mathbb{R}^n$, $i \in \{1, 2, 3, \ldots, N\}$, we have $\|\frac{1}{N}\sum_{i=1}^N x_i\|^2 \leq \frac{1}{N}\sum_{i=1}^N \|x_i\|^2$, which implies $\|\sum_{i=1}^N x_i\|^2 \leq N\sum_{i=1}^N \|x_i\|^2$.

APPENDIX B
PROOF OF LEMMA 1

Conditioning on the number of local updates of group i up to and including global round u, $\boldsymbol{t}_i^u$, we evaluate the expected difference between the aggregated global model and the latest local model at group i, by the end of global round u. Based on (8) and (12), the following holds:

$$
\mathbb{E}_{|t_i^u}\left\|x^{u+1} - x_i^{u,t_i^u}\right\|^2 = \alpha^2
$$

$$
\times \mathbb{E}_{|t_i^u}\left\|\frac{\sum_{k \in \mathcal{N}_i}\sum_{l=0}^{t_i^u-1}\tilde{g}_{i,k}\left(x_i^{u,l}\right)}{|\mathcal{N}_i|} - \frac{\sum_{i \in \mathcal{N}_g}\frac{1}{t_i^u}\sum_{k \in \mathcal{N}_i}\sum_{l=0}^{t_i^u-1}\tilde{g}_{i,k}\left(x_i^{u,l}\right)}{\sum_{i \in \mathcal{N}_g}|\mathcal{N}_i|}\right\|^2
$$

$$
\leq 2\alpha^2\mathbb{E}_{|t_i^u}\left\|\frac{1}{|\mathcal{N}_i|}\sum_{k \in \mathcal{N}_i}\sum_{l=0}^{t_i^u-1}\tilde{g}_{i,k}\left(x_i^{u,l}\right)\right\|^2
$$

$$
+ 2\alpha^2\mathbb{E}_{|t_i^u}\left\|\frac{1}{\sum_{i \in \mathcal{N}_g}|\mathcal{N}_i|}\sum_{i \in \mathcal{N}_g}\frac{1}{t_i^u}\sum_{k \in \mathcal{N}_i}\sum_{l=0}^{t_i^u-1}\tilde{g}_{i,k}\left(x_i^{u,l}\right)\right\|^2
$$

$$
\leq 2\alpha^2\left(\frac{1}{|\mathcal{N}_i|^2} + \frac{|\mathcal{N}_g|}{\left(\sum_{i \in \mathcal{N}_g}|\mathcal{N}_i|\right)^2(t_i^u)^2}\right)
$$

$$
\times \mathbb{E}_{|t_i^u}\left\|\sum_{k \in \mathcal{N}_i}\sum_{l=0}^{t_i^u-1}\tilde{g}_{i,k}\left(x_i^{u,l}\right)\right\|^2 + \frac{2\alpha^2|\mathcal{N}_g|}{\left(\sum_{i \in \mathcal{N}_g}|\mathcal{N}_i|\right)^2}
$$

$$
\times \sum_{j \in \mathcal{N}_g\backslash\{i\}}\frac{1}{\left(t_j^u\right)^2}\mathbb{E}_{|t_i^u}\left\|\sum_{k \in \mathcal{N}_j}\sum_{l=0}^{t_j^u-1}\tilde{g}_{j,k}\left(x_j^{u,l}\right)\right\|^2
$$

$$
\leq \left(\frac{2\alpha^2}{|\mathcal{N}_i|} + \frac{2\alpha^2|\mathcal{N}_g||\mathcal{N}_i|}{\left(\sum_{i \in \mathcal{N}_g}|\mathcal{N}_i|\right)^2(t_i^u)^2}\right)\sum_{k \in \mathcal{N}_i}\mathbb{E}_{|t_i^u}\left\|\sum_{l=0}^{t_i^u-1}\tilde{g}_{i,k}\left(x_i^{u,l}\right)\right\|^2
$$

$$
+ \frac{2\alpha^2|\mathcal{N}_g|}{\left(\sum_{i \in \mathcal{N}_g}|\mathcal{N}_i|\right)^2}\sum_{j \in \mathcal{N}_g\backslash\{i\}}\frac{|\mathcal{N}_j|}{\left(t_j^u\right)^2}\sum_{k \in \mathcal{N}_j}\mathbb{E}_{|t_i^u}\left\|\sum_{l=0}^{t_j^u-1}\tilde{g}_{j,k}\left(x_j^{u,l}\right)\right\|^2.
$$

$$(23)$$

The proof now follows by applying Assumption 3 (bounded gradient in (16)) to the bound above.

APPENDIX C
PROOF OF THEOREM 1

Based on the smoothness assumption of the loss function at LPS i, the SGD update rule in (5), and the local aggregation rule in (6), one can write

$$
\mathbb{E}_{|t_i^u}f_i\left(x_i^{u,l}\right) \leq \mathbb{E}_{|t_i^u}f_i\left(x_i^{u,l-1}\right) + \frac{L}{2}\mathbb{E}_{|t_i^u}\left\|x_i^{u,l} - x_i^{u,l-1}\right\|^2
$$

$$
+ \mathbb{E}_{|t_i^u}\langle\nabla f_i\left(x_i^{u,l-1}\right), x_i^{u,l} - x_i^{u,l-1}\rangle
$$

$$
= \mathbb{E}_{|t_i^u}f_i\left(x_i^{u,l-1}\right) + \frac{\alpha^2 L}{2}\mathbb{E}_{|t_i^u}\left\|\sum_{k \in \mathcal{N}_i}\frac{1}{|\mathcal{N}_i|}\tilde{g}_{i,k}\left(x_i^{u,l-1}\right)\right\|^2
$$

$$
+ \alpha\mathbb{E}_{|t_i^u}\langle\nabla f_i\left(x_i^{u,l-1}\right), \frac{-1}{|\mathcal{N}_i|}\sum_{k \in \mathcal{N}_i}\tilde{g}_{i,k}\left(x_i^{u,l-1}\right)\rangle. \quad (24)
$$

For the inner product term above, we have

$$
2\mathbb{E}_{|\boldsymbol{t}_i^u}\langle\nabla f_i\left(x_i^{u,l-1}\right), \frac{-1}{|\mathcal{N}_i|}\sum_{k \in \mathcal{N}_i}\tilde{g}_{i,k}\left(x_i^{u,l-1}\right)\rangle \overset{(i)}{=}
$$

$$
\mathbb{E}_{|\boldsymbol{t}_i^u}\left\|\nabla f_i\left(x_i^{u,l-1}\right) - \frac{1}{|\mathcal{N}_i|}\sum_{k \in \mathcal{N}_i}\nabla F_{i,k}\left(x_i^{u,l-1}\right)\right\|^2
$$

$$
- \mathbb{E}_{|\boldsymbol{t}_i^u}\left\|\nabla f_i\left(x_i^{u,l-1}\right)\right\|^2
$$

$$
- \mathbb{E}_{|\boldsymbol{t}_i^u}\left\|\frac{1}{|\mathcal{N}_i|}\sum_{k \in \mathcal{N}_i}\nabla F_{i,k}\left(x_i^{u,l-1}\right)\right\|^2, \quad (25)
$$

where (i) follows from Assumption 2 (unbiased stochastic gradient in (15)), and then applying $\langle x, y \rangle = \frac{1}{2}(\|x + y\|^2 - \|x\|^2 - \|y\|^2)$. Regarding second term in (24), the following holds:

$$
\mathbb{E}_{|\boldsymbol{t}_i^u}\left\|\frac{1}{|\mathcal{N}_i|}\sum_{k \in \mathcal{N}_i}\tilde{g}_{i,k}\left(x_i^{u,l-1}\right)\right\|^2
$$

$$\overset{(ii)}{=} \frac{1}{|\mathcal{N}_i|^2} \sum_{k \in \mathcal{N}_i} \mathbb{E}_{|\boldsymbol{t}_i^u} \left\| \tilde{g}_{i,k}\left(x_i^{u,l-1}\right) - \nabla F_{i,k}\left(x_i^{u,l-1}\right) \right\|^2$$

$$+ \mathbb{E}_{|\boldsymbol{t}_i^u} \left\| \frac{1}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} \nabla F_{i,k}\left(x_i^{u,l-1}\right) \right\|^2$$

$$\leq \frac{\sigma^2}{|\mathcal{N}_i|} + \mathbb{E}_{|\boldsymbol{t}_i^u} \left\| \frac{1}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} \nabla F_{i,k}\left(x_i^{u,l-1}\right) \right\|^2, \quad (26)$$

where (ii) follows from adding and subtracting $\nabla F_{i,k}(x_i^{u,l-1})$, and that each kth term $\tilde{g}_{i,k}(x_i^{u,l-1}) - \nabla F_{i,k}(x_i^{u,l-1})$ has zero mean and the overall $|\mathcal{N}_i|$ terms are independent across different clients. Substituting (25) and (26) into (24):

$$\mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^{u,l}\right) \leq \mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^{u,l-1}\right) + \frac{\alpha^2 L \sigma^2}{2|\mathcal{N}_i|}$$

$$+ \frac{\alpha^2 L}{2} \mathbb{E}_{|\boldsymbol{t}_i^u} \left\| \frac{1}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} \nabla F_{i,k}\left(x_i^{u,l-1}\right) \right\|^2 + \frac{\alpha}{2}$$

$$\times \left(\mathbb{E}_{|\boldsymbol{t}_i^u} \left\| \nabla f_i\left(x_i^{u,l-1}\right) - \frac{1}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} \nabla F_{i,k}\left(x_i^{u,l-1}\right) \right\|^2 \right.$$

$$- \mathbb{E}_{|\boldsymbol{t}_i^u} \left\| \nabla f_i\left(x_i^{u,l-1}\right) \right\|^2$$

$$\left. - \mathbb{E}_{|\boldsymbol{t}_i^u} \left\| \frac{1}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} \nabla F_{i,k}\left(x_i^{u,l-1}\right) \right\|^2 \right)$$

$$= \mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^{u,l-1}\right) - \frac{\alpha}{2} \mathbb{E}_{|\boldsymbol{t}_i^u} \left\| \nabla f_i\left(x_i^{u,l-1}\right) \right\|^2 + \frac{\alpha^2 L \sigma^2}{2|\mathcal{N}_i|}$$

$$- \left(\frac{\alpha}{2} - \frac{\alpha^2 L}{2} \right) \mathbb{E}_{|\boldsymbol{t}_i^u} \left\| \frac{1}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} \nabla F_{i,k}\left(x_i^{u,l-1}\right) \right\|^2 \quad (27)$$

$$\leq \mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^{u,l-1}\right) - \frac{\alpha}{2} \mathbb{E}_{|\boldsymbol{t}_i^u} \left\| \nabla f_i\left(x_i^{u,l-1}\right) \right\|^2 + \frac{\alpha^2 L \sigma^2}{2|\mathcal{N}_i|}, \quad (28)$$

where (27) follows from (1), and the last inequality follows by choosing $0 < \alpha \leq \frac{1}{L}$.

Next, rearranging the terms above and summing over all local iterations until t_i^u, we have

$$\frac{\alpha}{2} \sum_{l=1}^{t_i^u} \mathbb{E}_{|\boldsymbol{t}_i^u} \left\| \nabla f_i\left(x_i^{u,l-1}\right) \right\|^2$$

$$\leq \sum_{l=1}^{t_i^u} \left[\mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^{u,l-1}\right) - \mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^{u,l}\right) \right] + t_i^u \frac{\alpha^2 L \sigma^2}{2|\mathcal{N}_i|}$$

$$= \left[\mathbb{E}_{|\boldsymbol{t}_i^u} f_i(x_i^u) - \mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^{u,t_i^u}\right) \right] + t_i^u \frac{\alpha^2 L \sigma^2}{2|\mathcal{N}_i|}. \quad (29)$$

Now taking the average over all global communication rounds yields

$$\frac{1}{\sum_{u=1}^{\mathcal{U}} t_i^u} \sum_{u=1}^{\mathcal{U}} \frac{\alpha}{2} \sum_{l=1}^{t_i^u} \mathbb{E}_{|\boldsymbol{t}_i^u} \left\| \nabla f_i\left(x_i^{u,l-1}\right) \right\|^2$$

$$\leq \frac{1}{\sum_{u=1}^{\mathcal{U}} t_i^u} \mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^1\right) - \mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^{\mathcal{U},t_i^{\mathcal{U}}}\right) + \frac{1}{\sum_{u=1}^{\mathcal{U}} t_i^u}$$

$$\times \sum_{u=1}^{\mathcal{U}-1} \mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^{u+1}\right) - \mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^{u,t_i^u}\right) + \frac{\alpha^2 L \sigma^2}{2|\mathcal{N}_i|}. \quad (30)$$

Now let us consider one of the summands in the equality above. We have

$$\mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^{u+1}\right) - \mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^{u,t_i^u}\right) \leq \frac{L}{2} \mathbb{E}_{|\boldsymbol{t}_i^u} \left\| x_i^{u+1} - x_i^{u,t_i^u} \right\|^2$$

$$+ \mathbb{E}_{|\boldsymbol{t}_i^u} \langle \nabla f_i\left(x_i^{u,t_i^u}\right), x_i^{u+1} - x_i^{u,t_i^u} \rangle$$

$$\leq \frac{1}{2} \mathbb{E}_{|\boldsymbol{t}_i^u} \left(\left\| \nabla f_i\left(x_i^{u,t_i^u}\right) \right\|^2 + \left\| x_i^{u+1} - x_i^{u,t_i^u} \right\|^2 \right)$$

$$+ \frac{L}{2} \mathbb{E}_{|\boldsymbol{t}_i^u} \left\| x_i^{u+1} - x_i^{u,t_i^u} \right\|^2 \leq \frac{G^2}{2} + (L+1)\alpha^2$$

$$\times \left((t_i^u)^2 + \frac{|\mathcal{N}_g|}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i| \right)^2} \sum_{j \in \mathcal{N}_g} |\mathcal{N}_j|^2 \right) G^2, \quad (31)$$

where the last inequality follows directly from Lemma 1 (note that each group restarts its model updates following each global iteration, and hence $x_i^{u+1,0} = x^{u+1}$). Finally, by substituting (31) into (30) we get

$$\frac{1}{\sum_{u=1}^{\mathcal{U}} t_i^u} \sum_{u=1}^{\mathcal{U}} \sum_{l=1}^{t_i^u} \mathbb{E}_{|\boldsymbol{t}_i^u} \left\| \nabla f_i\left(x_i^{u,l-1}\right) \right\|^2$$

$$\leq \frac{2}{\alpha \sum_{u=1}^{\mathcal{U}} t_i^u}$$

$$\times \left(\mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^1\right) - \mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^{\mathcal{U},t_i^{\mathcal{U}}}\right) \right) + \frac{\alpha L \sigma^2}{|\mathcal{N}_i|} + \frac{(\mathcal{U}-1)G^2}{\alpha \sum_{u=1}^{\mathcal{U}} t_i^u}$$

$$+ \frac{2G^2(L+1)\alpha}{\sum_{u=1}^{\mathcal{U}} t_i^u} \sum_{u=1}^{\mathcal{U}-1} \left((t_i^u)^2 + \underbrace{\frac{|\mathcal{N}_g|}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i| \right)^2} \sum_{j \in \mathcal{N}_g} |\mathcal{N}_j|^2}_{\triangleq \kappa} \right)$$

$$= \frac{2}{\alpha \sum_{u=1}^{\mathcal{U}} t_i^u} \left(\mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^1\right) - \mathbb{E}_{|\boldsymbol{t}_i^u} f_i\left(x_i^{\mathcal{U},t_i^{\mathcal{U}}}\right) \right) + \frac{\alpha L \sigma^2}{|\mathcal{N}_i|}$$

$$+ \left(\frac{1}{\alpha \sum_{u=1}^{\mathcal{U}} t_i^u} + \frac{2(L+1)\kappa\alpha}{\sum_{u=1}^{\mathcal{U}} t_i^u} \right)(\mathcal{U}-1)G^2$$

$$+ \frac{2(L+1)\alpha}{\sum_{u=1}^{\mathcal{U}} t_i^u} \sum_{u=1}^{\mathcal{U}-1} (t_i^u)^2 G^2. \quad (32)$$

APPENDIX D
PROOF OF THEOREM 2

We first use the smoothness assumption of the global loss function, together with the SGD update rule in (12) to get the following:

$$\mathbb{E}_{|\boldsymbol{t}_i^u} f\left(x^{u+1}\right) \leq \mathbb{E}_{|\boldsymbol{t}_i^u} f(x^u) + \frac{\alpha^2 L}{2 \left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i| \right)^2}$$

$$\times \mathbb{E}_{|\boldsymbol{t}_i^u} \left\| \sum_{i \in \mathcal{N}_g} \frac{1}{t_i^u} \sum_{l=0}^{t_i^u-1} \sum_{k \in \mathcal{N}_i} \tilde{g}_{i,k}\left(x_i^{u,l}\right) \right\|^2 + \alpha$$

$$\times \mathbb{E}_{|\boldsymbol{t}_i^u} \langle \nabla f(x^u), \frac{-1}{\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|} \sum_{i \in \mathcal{N}_g} \frac{1}{t_i^u} \sum_{l=0}^{t_i^u-1} \sum_{k \in \mathcal{N}_i} \tilde{g}_{i,k}\left(x_i^{u,l}\right) \rangle. \quad (33)$$

For the inner product term above, we have

$$\alpha \mathbb{E}_{|t_i^u} \langle \nabla f(x^u), \frac{-1}{\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|} \sum_{i \in \mathcal{N}_g} \frac{1}{t_i^u} \sum_{l=0}^{t_i^u - 1} \sum_{k \in \mathcal{N}_i} \tilde{g}_{i,k}\left(x_i^{u,l}\right) \rangle$$

$$= \frac{\alpha}{2} \mathbb{E}_{|t_i^u} \left\| \nabla f(x^u) - \frac{1}{\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|} \sum_{i \in \mathcal{N}_g} \frac{1}{t_i^u} \sum_{l=0}^{t_i^u - 1} \right.$$
$$\left. \times \sum_{k \in \mathcal{N}_i} \nabla F_{i,k}\left(x_i^{u,l}\right) \right\|^2 - \frac{\alpha}{2} \mathbb{E}_{|t_i^u} \|\nabla f(x^u)\|^2$$
$$- \frac{\alpha}{2} \mathbb{E}_{|t_i^u} \left\| \frac{1}{\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|} \sum_{i \in \mathcal{N}_g} \frac{1}{t_i^u} \sum_{l=0}^{t_i^u - 1} \sum_{k \in \mathcal{N}_i} \nabla F_{i,k}\left(x_i^{u,l}\right) \right\|^2 \quad (34)$$

For the second term in (33), we have

$$\mathbb{E}_{|t_i^u} \left\| \sum_{i \in \mathcal{N}_g} \frac{1}{t_i^u} \sum_{l=0}^{t_i^u - 1} \sum_{k \in \mathcal{N}_i} \tilde{g}_{i,k}\left(x_i^{u,l}\right) \right\|^2$$
$$\leq |\mathcal{N}_g| \sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|^2 \sigma^2$$
$$+ \mathbb{E}_{|t_i^u} \left\| \sum_{i \in \mathcal{N}_g} \frac{1}{t_i^u} \sum_{l=0}^{t_i^u - 1} \sum_{k \in \mathcal{N}_i} \nabla F_{i,k}\left(x_i^{u,l}\right) \right\|^2. \quad (35)$$

Substituting (34) and (35) into (33) and choosing $0 < \alpha \leq \frac{1}{L}$ yields

$$\mathbb{E}_{|t_i^u} f\left(x^{u+1}\right) \leq \mathbb{E}_{|t_i^u} f(x^u) + \frac{\alpha}{2}$$
$$\times \mathbb{E}_{|t_i^u} \left\| \nabla f(x^u) - \frac{1}{\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|} \sum_{i \in \mathcal{N}_g} \frac{1}{t_i^u} \sum_{l=0}^{t_i^u - 1} \sum_{k \in \mathcal{N}_i} \nabla F_{i,k}\left(x_i^{u,l}\right) \right\|^2$$
$$- \frac{\alpha}{2} \mathbb{E}_{|t_i^u} \|\nabla f(x^u)\|^2 + \frac{\alpha^2 L |\mathcal{N}_g| \sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|^2 \sigma^2}{2\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^2}. \quad (36)$$

Regarding the second term in (36), although the division by t_i^u fixes the bias issue of the cumulative gradient at the GPS, it does not make it not coincide with its theoretical definition in (2). Hence, different from the analogous step in (27) in the proof of Theorem 1, the term above requires more mathematical manipulations. Towards that end, we bound it as follows:

$$\mathbb{E}_{|t_i^u} \left\| \nabla f(x^u) - \frac{1}{\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|} \sum_{i \in \mathcal{N}_g} \frac{1}{t_i^u} \sum_{l=0}^{t_i^u - 1} \sum_{k \in \mathcal{N}_i} \nabla F_{i,k}\left(x_i^{u,l}\right) \right\|^2$$
$$\leq \frac{|\mathcal{N}_g|}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^2} \sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|$$
$$\times \sum_{k \in \mathcal{N}_i} \mathbb{E}_{|t_i^u} \left\| \nabla F_{i,k}(x^u) - \frac{1}{t_i^u} \sum_{l=0}^{t_i^u - 1} \nabla F_{i,k}\left(x_i^{u,l}\right) \right\|^2$$
$$\leq \frac{2L^2 |\mathcal{N}_g|}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^2} \sum_{i \in \mathcal{N}_g} |\mathcal{N}_i| \sum_{k \in \mathcal{N}_i} \frac{1}{t_i^u}$$
$$\times \sum_{l=0}^{t_i^u - 1} \mathbb{E}_{|t_i^u} \left\| x^u - x_i^{u-1,t_i^{u-1}} \right\|^2$$
$$+ \mathbb{E}_{|t_i^u} \left\| x_i^{u-1,t_i^{u-1}} - x_i^{u,l} \right\|^2. \quad (37)$$

For the last term above, we have

$$\mathbb{E}_{|t_i^u} \left\| x_i^{u-1,t_i^{u-1}} - x_i^{u,l} \right\|^2$$
$$\leq 2\mathbb{E}_{|t_i^u} \left\| x_i^{u-1,t_i^{u-1}} - x^u \right\|^2 + 2\mathbb{E}_{|t_i^u} \left\| x^u - x_i^{u,l} \right\|^2$$
$$\leq 2\mathbb{E}_{|t_i^u} \left\| x_i^{u-1,t_i^{u-1}} - x^u \right\|^2 + \frac{2\alpha^2}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} l^2 G^2. \quad (38)$$

Next, substituting the bound of (38) in (37), and then substituting (37) in (36) and rearranging, we get

$$\mathbb{E}_{|t_i^u} \|\nabla f(x^u)\|^2 \leq \frac{2}{\alpha} \left(\mathbb{E}_{|t_i^u} f(x^u) - \mathbb{E}_{|t_i^u} f\left(x^{u+1}\right) \right)$$
$$+ \frac{\alpha L |\mathcal{N}_g| \sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|^2 \sigma^2}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^2}$$
$$+ \frac{4L^2 |\mathcal{N}_g|}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^2} \sum_{i \in \mathcal{N}_g} |\mathcal{N}_i| \sum_{k \in \mathcal{N}_i} \frac{1}{t_i^u} \sum_{l=0}^{t_i^u - 1} \left(3\alpha^2 \left(t_i^{u-1}\right)^2 \right.$$
$$\left. + \frac{3\alpha^2 |\mathcal{N}_g|}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^2} \sum_{j \in \mathcal{N}_g} |\mathcal{N}_j|^2 G^2 + \alpha^2 l^2 G^2 \right). \quad (39)$$

Then, taking the average over global communication rounds $\mathcal{U}$ yields

$$\frac{1}{\mathcal{U}} \sum_{u=1}^{\mathcal{U}} \mathbb{E}_{|t_i^u} \|\nabla f(x^u)\|^2 \leq \frac{2}{\alpha} \frac{1}{\mathcal{U}} \left(\mathbb{E}_{|t_i^u} f\left(x^1\right) - \mathbb{E}_{|t_i^u} f\left(x^{\mathcal{U}+1}\right) \right)$$
$$+ \frac{\alpha L |\mathcal{N}_g| \sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|^2 \sigma^2}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^2}$$
$$+ \frac{1}{\mathcal{U}} \sum_{u=1}^{\mathcal{U}} \frac{4L^2 |\mathcal{N}_g|}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^2} \sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|$$
$$\times \sum_{k \in \mathcal{N}_i} \frac{1}{t_i^u} \sum_{l=0}^{t_i^u - 1} 3\alpha^2 \left(t_i^{u-1}\right)^2$$
$$+ \frac{12\alpha^2 L^2 G^2 |\mathcal{N}_g|^2}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^4} \sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|^2 \sum_{j \in \mathcal{N}_g} |\mathcal{N}_j|^2$$
$$+ \frac{1}{\mathcal{U}} \sum_{u=1}^{\mathcal{U}} \frac{4\alpha^2 L^2 G^2}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^2} |\mathcal{N}_g| \sum_{i \in \mathcal{N}_g} |\mathcal{N}_i| \sum_{k \in \mathcal{N}_i} \frac{1}{t_i^u} \sum_{l=0}^{t_i^u - 1} l^2. \quad (40)$$

Direct simplifications of the above expression give the result of the theorem.

APPENDIX E
PROOF OF COROLLARY 1

By Theorem 2, we have shown the convergence rate of the whole setting. Furthermore, bounding the local iteration time, and as a consequence the number of local iterations as stated in (21) and (22), one can show that the bound in Theorem 2 behaves as follows:

$$\frac{1}{\mathcal{U}} \sum_{u=1}^{\mathcal{U}} \mathbb{E}_{|t_i^u} \|\nabla f(x^u)\|^2 \leq \frac{2}{\alpha} \frac{1}{\mathcal{U}} \left(\mathbb{E}f\left(x^1\right) - \mathbb{E}f\left(x^{\mathcal{U}+1}\right) \right)$$

$$+ \frac{12\alpha^2 L^2 G^2 |\mathcal{N}_g|^2}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^4} \left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|^2\right)^2$$

$$+ \frac{\alpha L |\mathcal{N}_g| \sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|^2 \sigma^2}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^2}$$

$$+ \frac{\alpha^2}{\mathcal{U}} \sum_{u=1}^{\mathcal{U}} \frac{12 L^2 |\mathcal{N}_g|}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^2} \sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|^2 \left(t_i^{u-1}\right)^2$$

$$+ \frac{\alpha^2}{\mathcal{U}} \sum_{u=1}^{\mathcal{U}} \frac{4 L^2 |\mathcal{N}_g|}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^2}$$

$$\times \sum_{i \in \mathcal{N}_g} \frac{|\mathcal{N}_i|^2 \left(t_i^u - 1\right)\left(2 t_i^u - 1\right) G^2}{6}$$

$$\leq \frac{2}{\alpha} \frac{1}{\mathcal{U}} \left(\mathbb{E}_{|t_i^u} f\left(x^1\right) - \mathbb{E}_{|t_i^u} f\left(x^{\mathcal{U}+1}\right)\right)$$

$$+ \frac{12 L^2 G^2 \alpha^2 |\mathcal{N}_g|^2}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^4} \left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|^2\right)^2$$

$$+ \frac{\alpha L |\mathcal{N}_g| \sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|^2 \sigma^2}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^2}$$

$$+ \frac{12 L^2 \alpha^2 |\mathcal{N}_g|}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^2} \sum_{i \in \mathcal{N}_g} \left(|\mathcal{N}_i| t_i^{\max}\right)^2$$

$$+ \frac{4 L^2 \alpha^2 G^2 |\mathcal{N}_g|}{\left(\sum_{i \in \mathcal{N}_g} |\mathcal{N}_i|\right)^2} \sum_{i \in \mathcal{N}_g} \frac{\left(|\mathcal{N}_i| t_i^{\max}\right)^2}{3}. \tag{41}$$

Choosing $\alpha \leq \frac{1}{\sqrt{\mathcal{U}}}$ in the final bound above directly gives the required $\mathcal{O}(\frac{1}{\sqrt{\mathcal{U}}})$ bound.

REFERENCES

[1] A. Ali and A. Arafa, "Hierarchical federated learning in delay sensitive communication networks," in *Proc. Asilomar*, Oct. 2022, pp. 1180–1184.

[2] 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*, Apr. 2017, pp. 1273–1282.

[3] P. S. Bouzinis, P. D. Diamantoulakis, and G. K. Karagiannidis, "Wireless quantized federated learning: A joint computation and communication design," 2022, *arXiv:2203.05878*.

[4] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-IID data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3400–3413, Sep. 2020.

[5] Y. Jee Cho, J. Wang, and G. Joshi, "Towards understanding biased client selection in federated learning," in *Proc. AISTATS*, Mar. 2022, pp. 10351–10375.

[6] S. Wang and M. Ji, "A unified analysis of federated learning with arbitrary client participation," in *Proc. NeurIPS*, Nov. 2022, pp. 1–14.

[7] 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 *Proc. IEEE ICASSP*, May 2020, pp. 8743–8747.

[8] S. U. Stich, "Local SGD converges fast and communicates little," 2018, *arXiv:1805.09767*.

[9] B. Woodworth et al., "Is local SGD better than minibatch SGD?" in *Proc. ICML*, Jul. 2020, pp. 1–10.

[10] T. Lin, S. U. Stich, K. K. Patel, and M. Jaggi, "Don't use large mini-batches, use local SGD," 2018, *arXiv:1808.07217*.

[11] H. S. Ghadikolaei, S. Stich, and M. Jaggi, "LENA: Communication-efficient distributed learning with self-triggered gradient uploads," in *Proc. 24th AISTATS*, Apr. 2021, pp. 3943–3951.

[12] J. Wang, S. Wang, R.-R. Chen, and M. Ji, "Demystifying why local aggregation helps: Convergence analysis of hierarchical SGD," in *Proc. AAAI*, Jun. 2022, pp. 1–9.

[13] L. Liu, J. Zhang, S. H. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in *Proc. IEEE ICC*, Jun. 2020, pp. 1–6.

[14] S. Luo, X. Chen, Q. Wu, Z. Zhou, and S. Yu, "HFEL: Joint edge association and resource allocation for cost-efficient hierarchical federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6535–6548, Oct. 2020.

[15] F. P.-C. Lin, S. Hosseinalipour, N. Michelusi, and C. G. Brinton, "Delay-aware hierarchical federated learning," *IEEE Trans. Cogn. Commun. Netw.*, vol. 10, no. 2, pp. 674–688, Apr. 2024.

[16] D. Jhunjhunwala, P. Sharma, A. Nagarkatti, and G. Joshi, "FedVARP: Tackling the variance due to partial client participation in federated learning," in *Proc. 38th UAI*, Aug. 2022, pp. 1–28.

[17] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Proc. 34th NeurIPS*, Dec. 2020, pp. 7611–7623.

[18] H. Yu, S. Yang, and S. Zhu, "Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *Proc. AAAI*, Jan. 2019, pp. 1–11.

[19] A. Krizhevsky, "Learning multiple layers of features from tiny images," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Rep. TR-2009, 2009. [Online]. Available: https://www.cs.toronto.edu/ kriz/cifar.html

[20] H. Chaoyang et al., "FedML: A research library and benchmark for federated machine learning," 2020, *arXiv:2007.13518*.

[21] S. J. Reddi et al., "Adaptive federated optimization," in *Proc. ICLR*, May 2021, pp. 1–38.

[22] B. Buyukates and S. Ulukus, "Timely communication in federated learning," in *Proc. IEEE INFOCOM*, May 2021, pp. 1–6.

[23] G. Liang and U. C. Kozat, "TOFEC: Achieving optimal throughput-delay trade-off of cloud storage using erasure codes," in *Proc. IEEE INFOCOM*, Apr. 2014, pp. 826–834.

Abdulmoneam Ali (Graduate Student Member, IEEE) received the B.Sc. degree in electrical engineering from Zagazig University, Egypt, in 2012, and the M.S. degree in computer science from Nile University, Egypt, in 2021. He is currently pursuing the Ph.D. degree with the Electrical and Computer Engineering Department, University of North Carolina at Charlotte, USA. His research interests include federated learning, wireless communications, and stochastic geometry. He was a recipient of the Graduate Fellowship from Nile University, Egypt, from 2016 to 2019.

Ahmed Arafa (Member, IEEE) received the B.Sc. degree (with Highest Hons.) in electrical engineering from Alexandria University, Egypt, in 2010, the first M.Sc. degree in wireless technologies from the Wireless Intelligent Networks Center, Nile University, Egypt, in 2012, and the second M.Sc. and Ph.D. degrees in electrical engineering from the University of Maryland at College Park, College Park, MD, USA, in 2016 and 2017, respectively. He was with the Electrical Engineering Department, Princeton University as a Postdoctoral Research Associate from 2017 to 2019. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of North Carolina at Charlotte. His research interests are in communication theory, information theory, machine learning, and signal processing, with recent focus on federated learning, timely information processing and transfer (age-of-information), information-theoretic security and privacy, and energy harvesting communications. He was the recipient of the Distinguished Dissertation Award from the Department of Electrical and Computer Engineering, University of Maryland in 2017. In 2022, he received the NSF CAREER Award. He currently serves as an Editor for the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING.