Distributed Non-Convex Optimization with Sublinear Speedup under Intermittent Client Availability

Yikai Yan [1] Chaoyue Niu [1] Yucheng Ding [1] Zhenzhe Zheng [1] Fan Wu [1] Guihai Chen [1] Shaojie Tang [2] Zhihua Wu [3]

Abstract

Federated learning is a new distributed machine learning framework, where a bunch of heterogeneous clients collaboratively train a model without sharing training data. In this work, we consider a practical and ubiquitous issue in federated learning: intermittent client availability, where the set of eligible clients may change during the training process. Such an intermittent client availability model would significantly deteriorate the performance of the classical Federated Averaging algorithm (FedAvg for short). We propose a simple distributed non-convex optimization algorithm, called Federated Latest Averaging (FedLaAvg for short), which leverages the latest gradients of all clients, even when the clients are not available, to jointly update the global model in each iteration. Our theoretical analysis shows that FedLaAvg attains the convergence rate of $O(1/(N^{1/4}T^{1/2}))$, achieving a sublinear speedup with respect to the total number of clients. We implement and evaluate FedLaAvg with the CIFAR-10 dataset. The evaluation results demonstrate that FedLaAvg indeed reaches a sublinear speedup and achieves 4.23% higher test accuracy than FedAvg.

1. Introduction

Federated Learning (FL) is a new paradigm of distributed machine learning (McMahan et al., 2017; Li et al., 2019a; Kairouz et al., 2019). It allows multiple clients to collaboratively train a global model without needing to upload local data to a centralized cloud server. In the FL setting, data are massively distributed over clients, with non-IID distribution (Hsieh et al., 2019) and unbalance in quantity (Mohri et al., 2019); in these ways, FL is distinguished from traditional distributed optimization (Li et al., 2014; Lian et al., 2018; Tang et al., 2018a;b; 2019; Yu & Jin, 2019). Furthermore, the agents participating in FL are typically heterogeneous clients with limited computation resources and unreliable communication links, resulting in a varying set of eligible clients during the training process. These new features pose challenges in designing and analyzing learning algorithms for FL.

One of the leading challenges in deploying FL systems is client availability, where the clients may not be available throughout the entire training process. Consider the typical FL scenario where Google's mobile keyboard Gboard polishes its language models among numerous mobile-device users (Bonawitz et al., 2019; Yang et al., 2018). To minimize the negative impact on user experience, only devices that meet certain requirements (e.g., charging, idle, and free Wi-Fi) are eligible for model training. These requirements are usually satisfied at night in local time, resulting in a diurnal pattern of client availability. Such intermittent client availability would introduce bias into training data. On one hand, as clients have diverse availability patterns, certain clients are more likely to be selected to participate in the training, and thus their data would be over-represented. On the other hand, if the criteria of client availability depend on latency, then clients with slower processors or delayed networks may be under-represented. Such bias gives rise to inconsistency between the training and test data distributions, thus degrading the generalization ability of FL algorithms. This inconsistency is also known as dataset shift (Quionero-Candela et al., 2009; Moreno-Torres et al., 2012), a notorious obstacle to the convergence of machine learning algorithms (Subbaswamy et al., 2019; Snoek et al., 2019), which also exists in FL.

Existing work in the literature has not considered the issue of intermittent client availability, and the convergence analysis of FL algorithms always requires all clients to be available throughout the training process. As shown in Table 1, much effort (Wang & Joshi, 2018; Yu et al., 2019b; Khaled et al., 2019; Stich, 2019; Stich & Karimireddy, 2019; Li et al., 2019b) has been expended in proving the convergence of the classical FedAvg algorithm (McMahan et al.,

[1]Shanghai Jiao Tong University, China [2]University of Texas at Dallas, USA [3]Alibaba Group, Hangzhou, China. Correspondence to: Fan Wu <fwu@cs.sjtu.edu.cn>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML).

Table 1. Convergence results in FL under different client availability assumptions.

Studies	Assumptions on Client Availability	Convergence Rate
Wang & Joshi (2018) Yu et al. (2019b) Khaled et al. (2019) Stich (2019) Stich & Karimireddy (2019)	All clients are available and participate in training.	$O(1/\sqrt{NT})$
Li et al. (2019b)	All clients are available, and a subset of clients participate in training.	$O(1/T)$
The current study	Each client is available at least once during any period with certain length.	$O(1/(N^{1/4}T^{1/2}))$

2017). However, this line of work assumed that all clients participate in each iteration of the training, to establish the $O(1/\sqrt{NT})$[1] convergence of FedAvg. Such a full client participation requirement would significantly increase the synchronization latency of the collaborative training process, and is hard to be satisfied in practical FL. One exception is Li et al. (2019b), who only required a subset of clients to participate in each iteration to obtain the $O(1/T)$ convergence of FedAvg. However, to guarantee such a convergence result, they assumed that the participating clients are selected either uniformly at random or with probabilities proportional to the volume of local data, which is possible only if all clients are available. As these studies adopted the full client availability model, there is no bias in the training data, which is an essential condition to obtain the positive convergence results of FedAvg in the literature.

In this study, we integrate the consideration of intermittent client availability into the design and analysis of the FL algorithm. We first formulate a practical model for intermittent client availability in FL; this model allows the set of available clients to follow any time-varying distribution, with the assumption that each client needs to be available at least once during any period with certain length. Under such a client availability model, FedAvg would diverge even in a simple learning scenario (shown in Subsection 3.1), because the training data are biased towards those highly available clients. For general distributed non-convex optimization, we propose a simple Federated Latest Averaging algorithm, namely FedLaAvg, to approximately balance the influence of each client's data on the global model training. Specifically, instead of averaging only the gradients collected from participating clients, FedLaAvg averages the latest gradients[2] of all clients. By setting appropriate parameters, we can prove an $O(1/(N^{1/4}T^{1/2}))$ convergence for FedLaAvg, implying that FedLaAvg can achieve a sublinear speedup with respect to the total number of clients. We summarize the contributions of this work as follows.

- To the best of our knowledge, we are the first to study the problem of intermittent client availability in FL, and present a formal formulation thereof. We also show the divergence of FedAvg in such a practical client availability model, and investigate the underlying reasons behind it.
- Under the intermittent client availability model, we propose the fast convergent algorithm FedLaAvg, which aggregates the latest gradients of all clients in each training iteration. Our theoretical analysis shows the $O(1/(N^{1/4}T^{1/2}))$ convergence of FedLaAvg for general distributed non-convex optimization.
- Using the CIFAR-10 dataset, we evaluate FedLaAvg and compare its performance with that of FedAvg. Our evaluation results demonstrate the effectiveness and efficiency of FedLaAvg, as it achieves 4.23% higher test accuracy than FedAvg and a sublinear speedup.

2. Problem Formulation

We consider a general distributed non-convex optimization scenario in which N clients collaboratively solve the following consensus optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^m} f(\mathbf{x}) \triangleq \sum_{i=1}^{N} w_i \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[F(\mathbf{x};\xi_i)] = \sum_{i=1}^{N} w_i \tilde{f}_i(\mathbf{x}).$$

Each client i holds training data $\xi_i \sim \mathcal{D}_i$, and w_i is the weight of this client (typically the proportion of client i's local data volume in the total data volume of the FL system (McMahan et al., 2017)). Function $F(\mathbf{x};\xi_i)$ is the training error of model parameters $\mathbf{x}$ over local data ξ_i, and $\tilde{f}_i(\mathbf{x})$ is the local generalization error, taking expectation over the randomness of local data. In iteration t, participating client i observes the local stochastic gradient:

$$\mathbf{g}_i^t = \nabla F(\mathbf{x}^{t-1};\xi_i^t),$$

[1] Notation N is the total number of clients, and T is the total number of iterations in the training.

[2] The latest gradient of a given client is the gradient calculated in her latest participating iteration. Please refer to Subsection 3.2 for detailed definition.

where $\mathbf{x}^{t-1}$ is the model parameters from the previous iteration and ξ_i^t is the local training data in this iteration. We note

$$\mathbb{E}\left[\mathbf{g}_i^t \mid \xi^{[t-1]}\right] = \nabla \tilde{f}_i(\mathbf{x}^{t-1}),$$

where $\xi^{[t-1]}$ is the historical training data from all clients before iteration t:

$$\xi^{[t-1]} \triangleq \{\xi_i^\tau | i \in \{1, 2, \cdots, N\}, \tau \in \{1, 2, \cdots, t-1\}\}.$$

To simplify the analysis of unbalanced data volume among clients, we use a scaling technique to obtain a revised local objective function:

$$f_i(\mathbf{x}) = w_i N \tilde{f}_i(\mathbf{x}).$$

Then, we can rewrite the global objective function as

$$f(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{x}). \quad (1)$$

In this study, we make three assumptions regarding the objective funtions as follows.

Assumption 1. *Local objective functions f_i are all $L-$ smooth:*

$$\| \nabla f_i(\mathbf{u}) - \nabla f_i(\mathbf{v}) \| \leq L \| \mathbf{u} - \mathbf{v} \|, \forall i, \mathbf{u}, \mathbf{v}.$$

The corollary is

$$f_i(\mathbf{v}) \leq f_i(\mathbf{u}) + \langle \mathbf{v} - \mathbf{u}, \nabla f_i(\mathbf{u}) \rangle + \frac{L}{2} \| \mathbf{v} - \mathbf{u} \|^2, \forall i, \mathbf{u}, \mathbf{v}.$$

Assumption 2. *Bounded variance: with $\sigma > 0$,*

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[\|\nabla F(\mathbf{x}; \xi_i) - \nabla f_i(\mathbf{x})\|^2\right] \leq \sigma^2, \forall i, \mathbf{x}.$$

Assumption 3. *Bounded gradient: with $G > 0$,*

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[\| \nabla F(\mathbf{x}; \xi_i) \|^2\right] \leq G^2, \forall i, \mathbf{x}$$

To model intermittent client availability, we use $\mathcal{C}^t$ to denote the set of available clients in iteration t. We formally introduce the following assumption regarding the model of intermittent client availability in FL.

Assumption 4. *Minimal availability: each client i is available at least once in any period with E successive iterations:*

$$\forall i, \forall t, \exists \tau \in \{t, t+1, \cdots, t+E-1\}, \text{ such that } i \in \mathcal{C}^\tau.$$

Assumption 1 is standard, and Assumptions 2 and 3 have also been widely made in the literature (Zhang et al., 2012; Stich et al., 2018; Yu et al., 2019b;a; Stich, 2019; Li et al., 2019b). Specifically, Yu et al. (2019b) worked with non-convex functions under Assumptions 1–3, and required all clients to be available and to participate in each iteration. Meanwhile, Li et al. (2019b) focused on convex functions while imposing the same full client availability requirement. The full client availability model in existing work is equivalent to the special case of our intermittent client availability model by setting $E = 1$ in Assumption 4. Furthermore, Assumption 4 regarding the intermittent client availability model is reasonable in practical FL. For example, as discussed earlier, clients are typically available at night, and thus Assumption 4 with E equal to the number of iterations in one day can describe such a client availability scenario.

3. Algorithm Design

In this section, we first show that the classical FedAvg algorithm produces arbitrarily poor-quality results in the presence of intermittent client availability. We investigate the underlying reasons for the divergence of FedAvg, and then propose a new algorithm called FedLaAvg, which converges at a fast rate under the intermittent client availability model.

3.1. Divergence of FedAvg

Example 1. We consider a distributed optimization problem with only two clients (denoted as 1 and 2) and a convex objective function. The goal is to learn the mean of one-dimensional data from these two clients. Following the problem formulation in Section 2, the local data distribution is $\xi_i \sim \mathcal{D}_i$ with mean $\mathbf{e}_i = \mathbb{E}[\xi_i]$. For simplicity, we assume the amounts of data from the two clients are balanced. We can formulate this simple learning problem as minimizing the following mean square error (MSE):

$$f(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^{2} f_i(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^{2} \mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[(\mathbf{x} - \xi_i)^2\right]$$
$$= \frac{1}{2} \sum_{i=1}^{2} (\mathbf{x} - \mathbf{e}_i)^2 + \frac{1}{2} \sum_{i=1}^{2} \mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[(\xi_i - \mathbf{e}_i)^2\right]. \quad (2)$$

For this example, we consider a specific intermittent client availability model: clients are available periodically and alternatelythat is, in each period, client 1 is available in the first t_1 iterations, and client 2 is available in the following t_2 iterations. Let k index the period; we then have

$$1 \in \mathcal{C}^{k(t_1+t_2)+i}, k \in \mathbb{N}, i \in \{1, 2, \cdots, t_1\};$$

$$2 \in \mathcal{C}^{k(t_1+t_2)+i}, k \in \mathbb{N}, i \in \{t_1+1, t_1+2 \cdots, t_1+t_2\}.$$

This model describes the client availability with a regular diurnal pattern. For example, clients around the world participate in FL at night. Clients 1 and 2 may correspond to clients from two different geographic regions, respectively.

Theorem 1. *Suppose each client computes the exact (not stochastic) gradient. In Example 1, even with a sufficiently* *low learning rate, the model parameters returned by FedAvg at the end of each period, i.e.,* $\mathbf{x}^{k(t_1+t_2)}$*, would converge to* $(t_1\mathbf{e}_1 + t_2\mathbf{e}_2)/(t_1+t_2)$*, which can be arbitrarily far away from the optimal solution* $\mathbf{x}^* = (\mathbf{e}_1 + \mathbf{e}_2)/2$.

Proof of Theorem 1. In Example 1, the training process of FedAvg is that the two clients train the global model using their own local data alternatively. Hence, after a certain number of training iterations, the global model parameters would be "pulled" in opposite directions when different clients are available, and would finally oscillate periodically around $(t_1\mathbf{e}_1 + t_2\mathbf{e}_2)/(t_1+t_2)$. For the detailed proof, please refer to Appendix A. □

3.2. Federated Latest Averaging

As shown in Subsection 3.1, intermittent client availability seriously affects the performance of FedAvg. In FL, the overall data distribution is an unbiased mixture of all clients' local data distributions. FedAvg can be proven to converge in the full client participation scenario (Yu et al., 2019b), because it uses the current gradients of all clients to update the global model. This makes the training data distribution in each iteration consistent with the overall data distribution. However, due to the intermittent client availability, some clients are selected to participate in the training process more frequently, introducing the bias into training data. To mitigate the bias problem, we imitate the full client participation scenario, and attempt to leverage the gradient information of all clients for model training in each iteration. The difficulty in employing this idea is that as some clients are absent from the training due to being either unavailable or unselected, we cannot obtain the current gradients of these clients. To resolve the lack of gradient information, we propose a natural and simple idea: *using the latest gradient of the client when her current gradient is not available.* By doing so, we can eliminate the bias in training data, and establish the convergence result.

We present in Algorithm 1 the detailed procedures of our approach FedLaAvg. In each iteration t, each selected client i locally calculates the gradient $\mathbf{g}_i^t$, and the cloud server maintains the average latest gradient $\mathbf{g}^t$ of all clients. The client selection principle in FedLaAvg is to *choose the* K *clients that are absent from the training process for the longest time* (Lines 5–7). Together with Assumption 4, we can guarantee that each client is selected at least once during any period with I successive iterations, where I is a function of parameters K, N, and E (please refer to Lemma 1 in Subsection 4.2 for the details). Based on this condition, we can establish an upper bound for the difference between each client's latest gradient and her current gradient, which would be critical for the convergence analysis of FedLaAvg in Subsection 4.2. To implement this principle, we use T_i^t to record the latest iteration before or at t in which client i

Algorithm 1 Federated Latest Averaging Algorithm

1: **Input:** initial model parameters $\mathbf{x}^0$; number of clients N; number of total iterations T; learning rate γ; proportion of selected clients β (i.e., the number of participating clients in each iteration is $K = \beta N$.)
2: Do initialization:
$$\mathbf{g}^0 \leftarrow \mathbf{0};\ \forall i \in \{1, 2, \cdots, N\},\ \mathbf{g}_i^0 \leftarrow \mathbf{0},\ T_i^0 \leftarrow 0.$$
3: **for** $t = 1$ **to** T **do**
4: $\quad\mathbf{g}^t \leftarrow \mathbf{g}^{t-1}$
5: $\quad\mathcal{C}^t \leftarrow$ the set of available clients
6: $\quad\mathcal{B}^t \leftarrow K$ clients from $\mathcal{C}^t$ with the lowest T_i^{t-1} values
7: $\quad$Update T_i^t values:
$$T_i^t \leftarrow t,\ \forall i \in \mathcal{B}^t;\ T_i^t \leftarrow T_i^{t-1},\ \forall i \notin \mathcal{B}^t. \tag{3}$$
8: $\quad$Each client $i \in \mathcal{B}^t$ calculates local gradient $\mathbf{g}_i^t$ and uploads gradient difference $\mathbf{g}_i^t - \mathbf{g}_i^{T_i^{t-1}}$ in parallel.
9: $\quad$Once receiving the gradient information from client i, the cloud server calculates the global gradient:
$$\mathbf{g}^t \leftarrow \mathbf{g}^t + \frac{1}{N}\left(\mathbf{g}_i^t - \mathbf{g}_i^{T_i^{t-1}}\right). \tag{4}$$
10: $\quad$The cloud server updates the global model parameters:
$$\mathbf{x}^t \leftarrow \mathbf{x}^{t-1} - \gamma \mathbf{g}^t. \tag{5}$$
11: **end for**

participates in the training process. During the aggregation procedure (Lines 8–9), to reduce the aggregation overhead, each selected client uploads the gradient difference: the difference between the gradients computed in the current participating iteration and the previous participating iteration, i.e., $\mathbf{g}_i^t - \mathbf{g}_i^{T_i^{t-1}}$, rather than the current gradient $\mathbf{g}_i^t$ as in the traditional FedAvg algorithm. Once the gradient difference from each client i is received, the cloud server would update the global gradient using (4). Following this aggregation method, the cloud server only needs to store the average latest gradient $\mathbf{g}^t$ and run K update operations. It can be proved by induction that at the end of each iteration t, the resulting gradient is indeed the average latest gradient:
$$\mathbf{g}^t = \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}_i^{T_i^t}. \tag{6}$$
Once the average latest gradient $\mathbf{g}^t$ is obtained, the cloud server uses it to update the global model parameters in (5).

4. Convergence Analysis

In this section, under the practical model of intermittent client availability, we show that FedLaAvg achieves an $O(1/(N^{1/4}T^{1/2}))$ convergence rate with a sublinear speedup in terms of the total number of clients.

4.1. Convergence on Example 1

We first demonstrate that FedLaAvg converges in Example 1, where FedAvg produces an arbitrarily poor-quality result. The convergence analysis of FedLaAvg for this simple example sheds light on the analysis for the case of general non-convex optimization in next subsection.

Theorem 2. *Suppose each client computes the exact (not stochastic) gradient. In Example 1, after T iterations, FedLaAvg with the learning rate $\gamma = 1/(2\sqrt{T})$ produces a solution $\hat{\mathbf{x}}$ that is within $O(1/\sqrt{T})$ range of the optimal solution $\mathbf{x}^*$:*

$$(\hat{\mathbf{x}} - \mathbf{x}^*)^2 = O\left(\frac{1}{\sqrt{T}}\right), \tag{7}$$

where we choose $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}^t} f(\mathbf{x}^t)$ as the output.

Proof of Theorem 2. We recall that

$$f(\mathbf{x}) = \left(\mathbf{x} - \frac{\mathbf{e}_1 + \mathbf{e}_2}{2}\right)^2 + \frac{(\mathbf{e}_1 - \mathbf{e}_2)^2}{4} + \frac{1}{2}\sum_{i=1}^{2} \mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[(\xi_i - \mathbf{e}_i)^2\right],$$

where the latter two terms are not associated with the variable $\mathbf{x}$. Hence, we only need to focus on the following part of the loss function:

$$\hat{f}(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^*)^2, \tag{8}$$

where $\mathbf{x}^* = (\mathbf{e}_1 + \mathbf{e}_2)/2$ is the optimal solution.

Note that

$$\begin{aligned}&\hat{f}(\mathbf{x}^t) - \hat{f}(\mathbf{x}^{t-1})\\ &= \left(\mathbf{x}^t - \mathbf{x}^{t-1}\right)^2 + 2\left(\mathbf{x}^{t-1} - \mathbf{x}^*\right)\left(\mathbf{x}^t - \mathbf{x}^{t-1}\right).\end{aligned} \tag{9}$$

We calculate the difference of $\mathbf{x}$ between two successive iterations:

$$\begin{aligned}\mathbf{x}^t - \mathbf{x}^{t-1} &= -\frac{\gamma}{2}\left(\mathbf{g}_1^{T_1^t} + \mathbf{g}_2^{T_2^t}\right)\\ &= -\gamma\left(\mathbf{x}^{T_1^t} - \mathbf{e}_1 + \mathbf{x}^{T_2^t} - \mathbf{e}_2\right)\\ &= -\gamma\left(\mathbf{x}^{T_1^t} + \mathbf{x}^{T_2^t} - 2\mathbf{x}^*\right),\end{aligned} \tag{10}$$

where T_i^t is defined in Subsection 3.2. Hence, we have

$$\begin{aligned}&2\left(\mathbf{x}^{t-1} - \mathbf{x}^*\right)\left(\mathbf{x}^t - \mathbf{x}^{t-1}\right)\\ &= -\gamma\left(2\mathbf{x}^{t-1} - 2\mathbf{x}^*\right)\left(\mathbf{x}^{T_1^t} + \mathbf{x}^{T_2^t} - 2\mathbf{x}^*\right)\\ &= -\frac{\gamma}{2}\left(2\mathbf{x}^{t-1} - 2\mathbf{x}^*\right)^2 - \frac{\gamma}{2}\left(\mathbf{x}^{T_1^t} + \mathbf{x}^{T_2^t} - 2\mathbf{x}^*\right)^2\\ &\quad + \frac{\gamma}{2}\left(2\mathbf{x}^{t-1} - \mathbf{x}^{T_1^t} - \mathbf{x}^{T_2^t}\right)^2.\end{aligned} \tag{11}$$

Substituting (10) and (11) into (9), we have

$$\begin{aligned}&\hat{f}(\mathbf{x}^t) - \hat{f}(\mathbf{x}^{t-1})\\ &= \left(\gamma^2 - \frac{\gamma}{2}\right)\left(\mathbf{x}^{T_1^t} + \mathbf{x}^{T_2^t} - 2\mathbf{x}^*\right)^2\\ &\quad - 2\gamma\left(\mathbf{x}^{t-1} - \mathbf{x}^*\right)^2 + \frac{\gamma}{2}\left(2\mathbf{x}^{t-1} - \mathbf{x}^{T_1^t} - \mathbf{x}^{T_2^t}\right)^2\\ &\overset{(a)}{\leq} - 2\gamma\left(\mathbf{x}^{t-1} - \mathbf{x}^*\right)^2 + \frac{\gamma}{2}\left(2\mathbf{x}^{t-1} - \mathbf{x}^{T_1^t} - \mathbf{x}^{T_2^t}\right)^2,\end{aligned} \tag{12}$$

where (a) follows from $0 < \gamma \leq 1/2$.

The algorithm starts from model parameters $\mathbf{x}^0$. When client 1 is available, $\mathbf{x}$ moves towards $\mathbf{e}_1$, and when client 2 is available, $\mathbf{x}$ moves towards $\mathbf{e}_2$. Hence, $\mathbf{x}$ is always within $G/2$ range of $\mathbf{x}^*$ in the training:

$$-\frac{G}{2} \leq \mathbf{x}^t - \mathbf{x}^* \leq \frac{G}{2}, \forall t \geq 0, \tag{13}$$

where $G = \max\left\{2\left(\mathbf{x}^0 - \mathbf{x}^*\right), |\mathbf{e}_1 - \mathbf{e}_2|\right\}$ is the the largest gradient norm during the training process. Substituting (13) into (10), we have

$$-\gamma G \leq \mathbf{x}^t - \mathbf{x}^{t-1} \leq \gamma G. \tag{14}$$

Referring to the specific client availability model in this example, we have

$$t - T_i^t \leq I,\ i = 1,\ 2, \tag{15}$$

where $I = \max\{t_1, t_2\}$. Therefore, when $t \geq T_i^t + 2$, summing (14) over iterations from $T_i^t + 1$ to $t - 1$, we have

$$-\gamma IG \leq \mathbf{x}^{t-1} - \mathbf{x}^{T_i^t} \leq \gamma IG,\ i = 1,\ 2. \tag{16}$$

Note that when $t = T_i^t$ or $t = T_i^t + 1$, the above formula also holds.

Substituting (16) into (12), we have

$$\hat{f}(\mathbf{x}^t) - \hat{f}(\mathbf{x}^{t-1}) \leq -2\gamma\left(\mathbf{x}^{t-1} - \mathbf{x}^*\right)^2 + 2\gamma^3 I^2 G^2.$$

Rearranging the above formula, we have

$$\left(\mathbf{x}^{t-1} - \mathbf{x}^*\right)^2 \leq \frac{1}{2\gamma}\left(\hat{f}(\mathbf{x}^{t-1}) - \hat{f}(\mathbf{x}^t)\right) + \gamma^2 I^2 G^2. \tag{17}$$

Summing (17) over iterations from 1 to T, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{x}^{t-1} - \mathbf{x}^*\right)^2 &\leq \frac{1}{2\gamma T}\left(\hat{f}(\mathbf{x}^0) - \hat{f}(\mathbf{x}^T)\right) + \gamma^2 I^2 G^2\\ &\leq \frac{1}{2\gamma T}\left(\hat{f}(\mathbf{x}^0) - \hat{f}(\mathbf{x}^*)\right) + \gamma^2 I^2 G^2.\end{aligned} \tag{18}$$

Substituting $\gamma = 1/(2\sqrt{T})$ into (18), we have

$$\frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{x}^{t-1}-\mathbf{x}^*\right)^2 \leq \frac{1}{\sqrt{T}}\left(\hat{f}(\mathbf{x}^0)-\hat{f}(\mathbf{x}^*)\right)+\frac{I^2G^2}{4T} = \frac{1}{\sqrt{T}}\left(f(\mathbf{x}^0)-f(\mathbf{x}^*)\right)+\frac{I^2G^2}{4T}. \tag{19}$$

Finally, we have

$$(\hat{\mathbf{x}}-\mathbf{x}^*)^2 \leq \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{x}^{t-1}-\mathbf{x}^*\right)^2 \leq \frac{1}{\sqrt{T}}\left(f(\mathbf{x}^0)-f(\mathbf{x}^*)\right)+\frac{I^2G^2}{4T} = O\left(\frac{1}{\sqrt{T}}\right). \tag{20}$$

□

4.2. Convergence on General Non-convex Functions

In this subsection, we show the $O(1/(N^{1/4}T^{1/2}))$ convergence of FedLaAvg on general non-convex functions.

We first introduce Lemma 1 about client participation.

Lemma 1. *Under Assumption 4, the client selection policy in FedLaAvg guarantees that for each client, the latest participating iteration is at most I iterations earlier than the current iteration:*

$$t - T_i^t \leq I, \forall t, \forall i, \text{ where } I = \left\lceil \frac{N}{K} \right\rceil E - 1 \tag{21}$$

Proof of Lemma 1. Due to space limitations, please refer to Appendix B for the detailed proof. □

With such a client participation condition, we can derive a key result for analyzing the convergence of FedLaAvg.

Theorem 3. *By setting $\gamma \leq 1/(2L)$ in FedLaAvg, we can derive the following bound on the average expected squared gradient norm under Assumptions 1–4:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1})\right\|^2\right] \leq \frac{4\gamma\sigma^2 L}{N}+\frac{2\gamma IL\left(G^2+\sigma^2\right)}{\sqrt{N}}+\frac{2\gamma^2I^2L^2G^2}{1-2\gamma L} + 4\gamma^2I^2L^2G^2+\frac{4}{\gamma T}\left(\mathbb{E}\left[f(\mathbf{x}^0)\right]-\mathbb{E}\left[f(\mathbf{x}^*)\right]\right),$$

where $\mathbf{x}^*$ is the the optimal solution for the general non-convex optimization problem.

Proof of Theorem 3. The basic idea is similar to the simple case discussed in Subsection 4.1. With Lemma 1, we first show that the difference between the latest gradient and the corresponding current gradient is bounded. Then the theorem follows with this bound and the smoothness of f_i. For the detailed proof, please refer to Appendix C. □

Before presenting our main result, we consider the full client participation setting discussed in Yu et al. (2019b), in which our FedLaAvg reduces to FedAvg. Since $K = N$, $E = 1$, and $I = \lceil N/K \rceil E - 1 = 0$ in this setting, the result in Theorem 3 becomes

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1})\right\|^2\right] \leq \frac{4\gamma\sigma^2 L}{N}+\frac{4}{\gamma T}\left(\mathbb{E}\left[f(\mathbf{x}^0)\right]-\mathbb{E}\left[f(\mathbf{x}^*)\right]\right).$$

Choosing $\gamma = \sqrt{N}/(2L\sqrt{T})$, when $T \geq N$, we can obtain the $O(1/\sqrt{NT})$ convergence, which is consistent with the linear speedup in terms of N as proven in Yu et al. (2019b).

For the intermittent client availability setting considered in this work, FedLaAvg achieves a sublinear speedup by choosing appropriate hyperparameters. For easy illustration, we define the loss difference between the initial solution $\mathbf{x}^0$ and the optimal solution $\mathbf{x}^*$ as $B = f(\mathbf{x}^0) - f(\mathbf{x}^*)$. In addition, we recall that $\beta = K/N$ is the proportion of the selected clients in each iteration.

Corollary 1. *By choosing the learning rate $\gamma = (\beta^{1/2}N^{1/4})/(2LE^{1/2}T^{1/2})$ and requiring $\gamma \leq 1/(4L)$ in FedLaAvg, we have the following convergence result:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1})\right\|^2\right] = O\left(\frac{E^{\frac{1}{2}}\left(G^2+\sigma^2+LB\right)}{N^{\frac{1}{4}}T^{\frac{1}{2}}\beta^{\frac{1}{2}}}+\frac{EG^2N^{\frac{1}{2}}}{T\beta}\right).$$

When $T \geq EN^{3/2}/\beta$, we further obtain the sublinear speedup with respect to the total number of clients:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1})\right\|^2\right] = O\left(\frac{E^{\frac{1}{2}}\left(G^2+\sigma^2+LB\right)}{N^{\frac{1}{4}}T^{\frac{1}{2}}\beta^{\frac{1}{2}}}\right) = O\left(\frac{1}{N^{\frac{1}{4}}T^{\frac{1}{2}}}\right).$$

Proof. Please refer to Appendix D for detailed proof. □

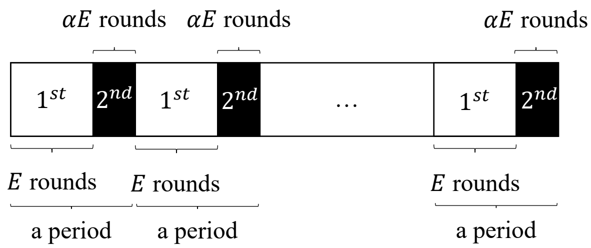


Figure 1. The diurnal pattern of client availability. Half of the clients are available in white grids, while the remaining clients are available in black grids.

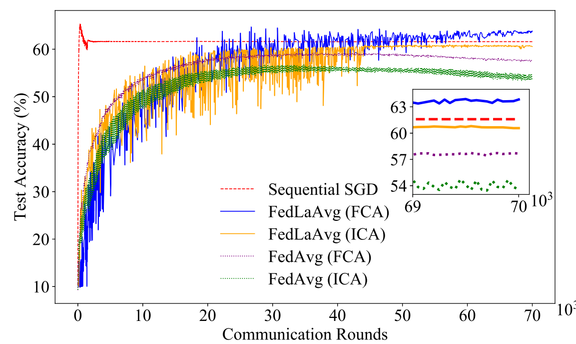


Figure 2. Performance of FedLaAvg, FedAvg, and sequential SGD under full client availability and intermittent client availability. We use ICA to abbreviate intermittent client availability and FCA to abbreviate full client availability.

Table 2. Best test accuracy of FedAvg and FedLaAvg achieved within 70000 rounds.

Algorithms	FCA	ICA
FedAvg	59.25%	56.75%
FedLaAvg	64.70%	60.98%

5. Experiments

5.1. Experiment Setting

In this section, we evaluate the performance of FedLaAvg in an image classification task over the CIFAR-10 dataset (Krizhevsky et al., 2009). The CIFAR-10 dataset consists of 60000 32×32 color images from 10 different classes, with 6000 images per class (5000 for training and 1000 for testing). We simulate the non-IID data distribution by setting each client to hold only images from one certain class and the number of clients from the same class to be $N/10$. To simulate the data unbalance, we let the number of samples on each client roughly follow a normal distribution with mean $5 \times 10^4/N$ and variance $(1 \times 10^4/N)^2$. For image classification, we take the deep learning model architecture from pytorch tutorial, with two convolutional layers followed by two fully connected layers and then a linear transformation layer to produce logits. The total parameter size of such a model is 62006.

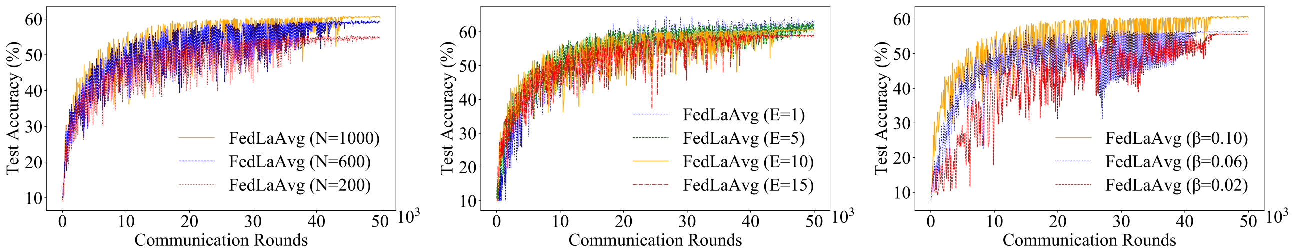
For simple illustration, in the previous discussion, we focus on the case in which participating clients upload gradient information in each iteration. In practical FL deployment, for communication efficiency, each participating client is allowed to perform multiple local training iterations before uploading the accumulated local model update (McMahan et al., 2017). FedLaAvg can be easily extended to this setting with the same performance guarantee, and the detailed design and convergence analysis are presented in Appendix E. To be consistent with practical FL deployment, we conduct experiments for FedLaAvg within the case of multiple local iterations in each communication round.

Figure 1 describes the intermittent client availability model with the diurnal pattern adopted in this experiment. In white grids, clients with the first five classes are available for E rounds, and in black grids, clients with the other five classes are available for the next αE rounds. The parameter α describes the degree of heterogeneous availability patterns among these two groups of clients.

In the default experiment setting, we set the total number of clients $N = 1000$, the period length $E = 10$, the ratio $\alpha = 0.5$, the proportion of selected clients in each round $\beta = 0.1$, and the number of local iterations $C = 10$. For hyperparameters, we let the learning rate γ decay in each communication round, and tune the initial learning rate and its decay for each experiment. The batch size for local iterations of each client is set to 5. For the detailed learning rate of each experiment, please refer to Appendix G.

5.2. Experimental Results

We compare FedLaAvg with FedAvg and sequential SGD, and show the experiment results in Figure 2. We run the standard SGD algorithm to train the global model using the whole dataset, and the total iterations in each round is $\beta NC = 1000$. The result of sequential SGD can be regarded as the optimal solution for the optimization problem. The test accuracy of sequential SGD decreases after reaching the peak and finally converges to 61.59%, due to the phenomenon of overfitting. The test accuracy of FedLaAvg with both intermittent and full client availability suffers from large oscillation at first but finally converges. This is because in the early stage of training, the model parameters change drastically, leading to a large difference between each client's latest gradient and her current gradient. As the training progresses, the model parameters would change smoothly, and this difference vanishes, indicating that the latest gradient better approximates the current gradient. In contrast with FedLaAvg, FedAvg has a smaller oscillation in



(a) Test accuracy with different N. (b) Test accuracy with different E. (c) Test accuracy with different β.

Figure 3. The performance of FedLaAvg with the variation of the total number of clients N, the period length E, and the proportion of selected clients β.

the early stage of training, as it always uses the current gradients to update the model parameters. Although FedAvg can finally converge under full client availability, it suffers from periodical oscillation under intermittent client availability even after a huge number of rounds (e.g., in communication rounds between 69×10^3 and 70×10^3). This is consistent with the divergence of FedAvg analyzed in Subsection 3.1. We can explore a useful trick in practice to combine the advantages of FedAvg and FedLaAvg: use FedAvg to train the model until it reaches the bottleneck, and then switch to FedLaAvg to further improve the performance.

We next show the performance of FedAvg and FedLaAvg in Table 2. We observe that FedLaAvg outperforms FedAvg in the two client availability models, and approaches the optimal solution. For the intermittent client availability, the best test accuracy of FedLaAvg is 4.23% higher than that of FedAvg. For the case of full client availability, FedLaAvg still achieves 5.45% higher test accuracy than FedAvg, because FedLaAvg leverages the gradient information of all clients for model training in each iteration, while FedAvg uses only the gradients of the selected clients. We also observe that FedLaAvg approaches the performance of sequential SGD. The best test accuracy of FedLaAvg under intermittent client availability is only 0.61% lower than the test accuracy that sequential SGD finally converges to. From Figure 2, we note that due to the overfitting, the convergent test accuracy of sequential SGD is slightly lower than the best test accuracy of FedLaAvg under full client availability.

The evaluation results in Figure 3 and Table 3 further validate the convergence result of FedLaAvg in Corollary 1. From Table 3, we see that FedLaAvg generally needs fewer training rounds to reach a certain test accuracy when either the total number of clients N or the proportion of selected clients β increases, or the period length E decreases. This result also validates the sublinear speedup of FedLaAvg with respect to N. In addition, as shown in Figure 3, FedLaAvg converges in diverse parameter settings, which is consistent with the convergence guarantee in Corollary 1. The different convergent test accuracies come from the non-convexity of the objective functions. From Subfigure 3(b), we further observe that with a larger E, the test accuracy of FedLaAvg oscillates more severely, because the latest gradient becomes an inaccurate approximation of the current gradient when clients are not available during a longer period.

Table 3. The number of training rounds needed to reach 55% test accuracy with different parameters.

N	200	400	600	800	1000
Rounds	33550	14800	12850	12700	10100

E	1	5	10	15
Rounds	10000	9950	10100	13400

β	0.02	0.04	0.06	0.08	0.10
Rounds	27800	32200	13450	9400	10100

6. Conclusion

In this work, we investigate intermittent client availability in federated learning and its impact on the convergence of the classical federated averaging algorithm. We use a collection of time-varying sets to represent the available clients in each training iteration, which can accurately model the intermittent client availability. Furthermore, we design a simple FedLaAvg algorithm with an $O(1/(N^{1/4}T^{1/2}))$ convergence guarantee for general distributed non-convex optimization problems. Empirical experiments with the CIFAR-10 dataset demonstrate the effectiveness and efficiency of FedLaAvg with a remarkable performance improvement and a sublinear speedup.

References

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konen, J., Mazzocchi, S., McMahan, H. B., Overveldt, T. V., Petrou, D., Ramage, D., and Roselander, J. Towards federated learning at scale: System design. In *Proceedings of MLSys*, 2019.

Hsieh, K., Phanishayee, A., Mutlu, O., and Gibbons, P. B. The non-IID data quagmire of decentralized machine learning. *CoRR*, abs/1910.00189, 2019.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning. *CoRR*, abs/1912.04977, 2019.

Khaled, A., Mishchenko, K., and Richtárik, P. First analysis of local GD on heterogeneous data. *CoRR*, abs/1909.04715, 2019.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. *Technique report*, 2009.

Li, M., Andersen, D. G., Park, J. W., Smola, A. J., Ahmed, A., Josifovski, V., Long, J., Shekita, E. J., and Su, B.-Y. Scaling distributed machine learning with the parameter server. In *Proceedings of OSDI*, pp. 583–598, 2014.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *CoRR*, abs/1908.07873, 2019a.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of FedAvg on non-IID data. *CoRR*, abs/1907.02189, 2019b.

Lian, X., Zhang, W., Zhang, C., and Liu, J. Asynchronous decentralized parallel stochastic gradient descent. In *Proceedings of ICML*, pp. 3043–3052, 2018.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of AISTATS*, pp. 1273–1282, 2017.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In *Proceedings of ICML*, pp. 4615–4625, 2019.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodrguez, R., Chawla, N. V., and Herrera, F. A unifying view on dataset shift in classification. *Pattern Recognition*, 45 (1):521–530, 2012.

Quionero-Candela, J., Sugiyama, M. S., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. The MIT Press, 2009. ISBN 0262170051.

Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., Dillon, J., Ren, J., and Nado, Z. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Proceedings of NeurIPS*, pp. 13969–13980, 2019.

Stich, S. U. Local SGD converges fast and communicates little. In *Proceedings of ICLR*, 2019.

Stich, S. U. and Karimireddy, S. P. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication. *CoRR*, abs/1909.05350, 2019.

Stich, S. U., Cordonnier, J.-B., and Jaggi, M. Sparsified SGD with memory. In *Proceedings of NeurIPS*, pp. 4447–4458, 2018.

Subbaswamy, A., Schulam, P., and Saria, S. Preventing failures due to dataset shift: Learning predictive models that transport. In *Proceedings of AISTATS*, pp. 3118–3127, 2019.

Tang, H., Gan, S., Zhang, C., Zhang, T., and Liu, J. Communication compression for decentralized training. In *Proceedings of NeurIPS*, pp. 7652–7662, 2018a.

Tang, H., Lian, X., Yan, M., Zhang, C., and Liu, J. D^2: Decentralized training over decentralized data. In *Proceedings of ICML*, pp. 4855–4863, 2018b.

Tang, H., Yu, C., Lian, X., Zhang, T., and Liu, J. `DoubleSqueeze`: Parallel stochastic gradient descent with double-pass error-compensated compression. In *Proceedings of ICML*, pp. 6155–6165, 2019.

Wang, J. and Joshi, G. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. *CoRR*, abs/1808.07576, 2018.

Yang, T., Andrew, G., Eichner, H., Sun, H., Li, W., Kong, N., Ramage, D., and Beaufays, F. Applied federated learning: Improving google keyboard query suggestions. *CoRR*, abs/1812.02903, 2018.

Yu, H. and Jin, R. On the computation and communication complexity of parallel SGD with dynamic batch sizes for stochastic non-convex optimization. In *Proceedings of ICML*, pp. 7174–7183, 2019.

Yu, H., Jin, R., and Yang, S. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In *Proceedings of ICML*, pp. 7184–7193, 2019a.

Yu, H., Yang, S., and Zhu, S. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of AAAI*, pp. 5693–5700, 2019b.

Zhang, Y., Wainwright, M. J., and Duchi, J. C. Communication-efficient algorithms for statistical optimization. In *Proceedings of NeurIPS*, pp. 1502–1510, 2012.

A. Proof of Theorem 1

Proof of Theorem 1. We first show that if $\gamma < \frac{1}{2}$, $\mathbf{x}^{k(t_1+t_2)}$ will converge to

$$X = \frac{(1-2\gamma)^{t_2}(\mathbf{e}_1 - \mathbf{e}_2) + \mathbf{e}_2 - \mathbf{e}_1(1-2\gamma)^{t_1+t_2}}{1-(1-2\gamma)^{t_1+t_2}}. \tag{22}$$

Note that for iterations where client 1 is available, we have

$$\forall t \in \{k(t_1+t_2)+i | k \in \mathbb{N}, i \in \{1, 2, \cdots, t_1\}\},\ \mathbf{x}^{t+1} = \mathbf{x}^t - 2\gamma\left(\mathbf{x}^t - \mathbf{e}_1\right),$$

where γ is the learning rate. Rearrange the equation, we have

$$\mathbf{x}^{t+1} - \mathbf{e}_1 = (1-2\gamma)\left(\mathbf{x}^t - \mathbf{e}_1\right),$$

which implies that $(\mathbf{x}^t - \mathbf{e}_1)$ is a geometric progression. Hence, we have

$$\mathbf{x}^{k(t_1+t_2)+t_1} = (1-2\gamma)^{t_1}\left(\mathbf{x}^{k(t_1+t_2)} - \mathbf{e}_1\right) + \mathbf{e}_1. \tag{23}$$

Applying the same analysis on iterations where client 2 is available, we have

$$\mathbf{x}^{(k+1)(t_1+t_2)} = (1-2\gamma)^{t_2}\left(\mathbf{x}^{k(t_1+t_2)+t_1} - \mathbf{e}_2\right) + \mathbf{e}_2. \tag{24}$$

Substituting (23) into (24), we have

$$\mathbf{x}^{(k+1)(t_1+t_2)} = (1-2\gamma)^{t_1+t_2}\left(\mathbf{x}^{k(t_1+t_2)} - \mathbf{e}_1\right) + (1-2\gamma)^{t_2}(\mathbf{e}_1 - \mathbf{e}_2) + \mathbf{e}_2. \tag{25}$$

Based on this recursion formula, we have

$$\mathbf{x}^{k(t_1+t_2)} = (1-2\gamma)^{(t_1+t_2)k}\mathbf{x}^0 + \left(1-(1-2\gamma)^{(t_1+t_2)k}\right)X. \tag{26}$$

Since $\gamma < \frac{1}{2}$, we have

$$\lim_{k\to+\infty} \mathbf{x}^{k(t_1+t_2)} = X.$$

Based on L'Hopital's rule, we have

$$\lim_{\gamma\to 0^+} X = \frac{t_1\mathbf{e}_1 + t_2\mathbf{e}_2}{t_1+t_2}.$$

The global minimization objective is

$$f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{2}(\mathbf{x}-\mathbf{e}_i)^2 + \frac{1}{2}\sum_{i=1}^{2}\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[(\xi_i - \mathbf{e}_i)^2\right],$$

which is obtained when $\mathbf{x} = \mathbf{x}^* = (\mathbf{e}_1 + \mathbf{e}_2)/2$. Note that $(\mathbf{e}_1+\mathbf{e}_2)/2 = (t_1\mathbf{e}_1 + t_2\mathbf{e}_2)/(t_1+t_2)$ only when $\mathbf{e}_1 = \mathbf{e}_2$ (data distributions are IID) or $t_1 = t_2$. Hence, FedAvg will produce arbitrarily poor-quality results without these inpractical assumptions. □

B. Proof of Lemma 1

Proof of Lemma 1. $\forall t, \forall i$, we focus on the training process from t (not included). In iteration $t+I+1$, under Assumption 4, client i has been available for at least $\lceil N/K \rceil$ times. Note the $\lceil N/K \rceil$ iterations as $\tau_1, \tau_2, \cdots, \tau_{\lceil N/K \rceil}$. We prove the lemma by contradiction. Suppose i is not selected in any of these iterations. Then we have $T_i^{\tau_{\lceil N/K \rceil}} = T_i^t$. In the $\lceil N/K \rceil$ iterations where client i is available, $\lceil N/K \rceil K$ clients have been selected. All these clients (noted as j) are with $T_j^\tau \le T_i^t$ for all iterations τ before she participates in the training process and $T_j^\tau > T_i^t$ for all iterations τ after participation. Hence, the

$\lceil N/K \rceil K$ clients are distinct. Including client i, the system has at least $\lceil N/K \rceil K + 1$ clients. However, the system has only $N \leq \lceil N/K \rceil K < \lceil N/K \rceil K + 1$ clients. This forms a contradiction. Therefore, for all t, the next iteration t_{next} where i participates in the training process after iteration t satisfies

$$t_{next} \leq t + I + 1. \tag{27}$$

For all client i, by setting t to iterations where client i is selected in (27), we can derive

$$\forall i,\ \forall t,\ t - T_i^t \leq I.$$

□

C. Proof of Theorem 3

Note that local gradient is not calculated in each iteration. In this subsection of the appendix, for mathematical analysis, we extend the definition $\mathbf{g}_i^t \triangleq \nabla F(\mathbf{x}^{t-1}, \xi_i^t)$. For iterations where client i does not participate, ξ_i^t is a random variable which follows $\xi_i^t \sim \mathcal{D}_i$.

Lemma 2. *Under Assumption 2 and 3, we have*

$$\mathbb{E}\left[\left\|\mathbf{g}_i^t\right\|^2\right] \leq G^2, \forall i, \forall t$$

and

$$\mathbb{E}\left[\left\|\mathbf{g}_i^t - \nabla f_i(\mathbf{x}^{t-1})\right\|^2\right] \leq \sigma^2, \forall i, \forall t.$$

Proof of Lemma 2. Our Assumptions 2 and 3 take the expectation over the randomness of one training iteration. But we care about the expectation taken over the randomness of the whole training process. This trivial lemma builds the gap.

For the gradient, we have

$$\mathbb{E}\left[\left\|\mathbf{g}_i^t\right\|^2\right] \overset{(a)}{=} \mathbb{E}\left[\mathbb{E}\left[\left\|\mathbf{g}_i^t\right\|^2 \mid \xi^{[t-1]}\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\nabla F\left(\mathbf{x}^{t-1}, \xi_i^t\right)\right\|^2 \mid \mathbf{x}^{t-1}\right]\right] \overset{(b)}{\leq} \mathbb{E}\left[G^2\right] = G^2, \tag{28}$$

where (a) follows from Law of Total Expectation $\mathbb{E}[\mathbb{E}[\mathbf{X}|\mathbf{Y}]] = \mathbb{E}[\mathbf{X}]$; (b) follows from Assumption 3.

For the variance, we have

$$\begin{aligned} & \mathbb{E}\left[\left\|\mathbf{g}_i^t - \nabla f_i(\mathbf{x}^{t-1})\right\|^2\right] \overset{(a)}{=} \mathbb{E}\left[\mathbb{E}\left[\left\|\mathbf{g}_i^t - \nabla f_i(\mathbf{x}^{t-1})\right\|^2 \mid \xi^{[t-1]}\right]\right] \\ = \ & \mathbb{E}\left[\mathbb{E}\left[\left\|\nabla F\left(\mathbf{x}^{t-1}; \xi_i^t\right) - \nabla f_i\left(\mathbf{x}^{t-1}\right)\right\|^2 \mid \mathbf{x}^{t-1}\right]\right] \overset{(b)}{\leq} \mathbb{E}\left[\sigma^2\right] = \sigma^2, \end{aligned} \tag{29}$$

where (a) follows from Law of Total Expectation $\mathbb{E}[\mathbb{E}[\mathbf{X}|\mathbf{Y}]] = \mathbb{E}[\mathbf{X}]$; (b) follows from Assumption 2. □

Lemma 3. $\forall i, \forall t$, *we have*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{N}\left(\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t - 1})\right)\right\|^2\right] = \sum_{i=1}^{N} \mathbb{E}\left[\left\|\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t - 1})\right\|^2\right].$$

Proof of Lemma 3. This lemma follows because training data are independent across clients. Specifically, note that

$$
\begin{aligned}
& \mathbb{E}\left[\left\|\sum_{i=1}^{N}\left(\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t-1})\right)\right\|^2\right] \\
= & \sum_{p=1}^{N}\sum_{q=1}^{N}\mathbb{E}\left[\left\langle \mathbf{g}_p^{T_p^t} - \nabla f_p(\mathbf{x}^{T_p^t-1}), \mathbf{g}_q^{T_q^t} - \nabla f_q(\mathbf{x}^{T_q^t-1})\right\rangle\right] \\
\overset{(a)}{=} & \sum_{p=1}^{N}\sum_{q=1}^{N}\mathbb{E}\left[\mathbb{E}\left[\left\langle \mathbf{g}_p^{T_p^t} - \nabla f_p(\mathbf{x}^{T_p^t-1}), \mathbf{g}_q^{T_q^t} - \nabla f_q(\mathbf{x}^{T_q^t-1})\right\rangle \mid \xi^{\left[\min\{T_p^t, T_q^t\}\right]}\right]\right] \\
\overset{(b)}{=} & \sum_{i=1}^{N}\mathbb{E}\left[\left\|\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t-1})\right\|^2\right],
\end{aligned}
\tag{30}
$$

where (a) follows from Law of Total Expectation $\mathbb{E}[\mathbb{E}[\mathbf{X}|\mathbf{Y}]] = \mathbb{E}[\mathbf{X}]$. Then we illustrate (b) case by case. Note that

$$
\mathbb{E}\left[\mathbb{E}\left[\left\langle \mathbf{g}_p^{T_p^t} - \nabla f_p(\mathbf{x}^{T_p^t-1}), \mathbf{g}_q^{T_q^t} - \nabla f_q(\mathbf{x}^{T_q^t-1})\right\rangle \mid \xi^{\left[\min\{T_p^t, T_q^t\}\right]}\right]\right]
$$

is equal to $\mathbb{E}\left[\left\|\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t-1})\right\|^2\right]$ when $p = q = i$. When $p \neq q$, without loss of generality, suppose $T_p^t \leq T_q^t$. Then it is equal to

$$
\begin{aligned}
& \mathbb{E}\left[\mathbb{E}\left[\left\langle \mathbf{g}_p^{T_p^t} - \nabla f_p(\mathbf{x}^{T_p^t-1}), \mathbf{g}_q^{T_q^t} - \nabla f_q(\mathbf{x}^{T_q^t-1})\right\rangle \mid \xi^{\left[T_p^t\right]}\right]\right] \\
= & \mathbb{E}\left[\left\langle \mathbf{g}_p^{T_p^t} - \nabla f_p(\mathbf{x}^{T_p^t-1}), \mathbb{E}\left[\mathbf{g}_q^{T_q^t} - \nabla f_q(\mathbf{x}^{T_q^t-1}) \mid \xi^{\left[T_p^t\right]}\right]\right\rangle\right]
\end{aligned}
\tag{31}
$$

because $\mathbf{g}_p^{T_p^t}$ and $f_q(\mathbf{x}^{T_p^t-1})$ are determined by $\xi^{[T_p^t]}$. When $T_p^t < T_q^t$, we have $\mathbb{E}[\mathbf{g}_q^{T_q^t} - \nabla f_q(\mathbf{x}^{T_q^t-1}) \mid \xi^{[T_p^t]}] = 0$. When $T_p^t = T_q^t$, we have

$$
\begin{aligned}
& \mathbb{E}\left[\left\langle \mathbf{g}_p^{T_p^t} - \nabla f_p(\mathbf{x}^{T_p^t-1}), \mathbb{E}\left[\mathbf{g}_q^{T_q^t} - \nabla f_q(\mathbf{x}^{T_q^t-1}) \mid \xi^{\left[T_p^t\right]}\right]\right\rangle\right] \\
= & \mathbb{E}\left[\left\langle \mathbf{g}_p^{T_p^t} - \nabla f_p(\mathbf{x}^{T_p^t-1}), \mathbf{g}_q^{T_p^t} - \nabla f_q(\mathbf{x}^{T_p^t-1})\right\rangle\right] \\
\overset{(a)}{=} & \mathbb{E}\left[\mathbb{E}\left[\left\langle \mathbf{g}_p^{T_p^t} - \nabla f_p(\mathbf{x}^{T_p^t-1}), \mathbf{g}_q^{T_p^t} - \nabla f_q(\mathbf{x}^{T_p^t-1})\right\rangle \mid \xi^{\left[T_p^t-1\right]}\right]\right] \\
\overset{(b)}{=} & 0,
\end{aligned}
\tag{32}
$$

where (a) follows from Law of Total Expectation $\mathbb{E}[\mathbb{E}[\mathbf{X}|\mathbf{Y}]] = \mathbb{E}[\mathbf{X}]$; (b) follows because $\xi_p^{T_p^t}$ adn $\xi_q^{T_p^t}$ are independent, and thus the covariance of $\mathbf{g}_p^{T_p^t}$ and $\mathbf{g}_q^{T_p^t}$ is 0. $\square$

Lemma 4. *Under Assumptions 1 and 3,* $\forall t, \forall t_0 \leq t, \forall i$*, we have*

$$
\mathbb{E}\left[\left\|\nabla f_i(\mathbf{x}^{t-1}) - \nabla f_i(\mathbf{x}^{t_0-1})\right\|^2\right] \leq (t - t_0)^2 L^2\gamma^2 G^2.
$$

Proof of Lemma 4. This lemma follows the intuition that the difference of $\mathbf{x}$ in two iterations is bounded by the number of iterations between them.

$$
\begin{aligned}
& \mathbb{E}\left[\left\|\nabla f_i(\mathbf{x}^{t-1}) - \nabla f_i(\mathbf{x}^{t_0-1})\right\|^2\right] \\
= \quad & \mathbb{E}\left[\left\|\sum_{\tau=t_0}^{t-1}\left(\nabla f_i(\mathbf{x}^{\tau}) - \nabla f_i(\mathbf{x}^{\tau-1})\right)\right\|^2\right] \\
\overset{(a)}{\leq} \quad & (t-t_0)\sum_{\tau=t_0}^{t-1}\mathbb{E}\left[\left\|\nabla f_i(\mathbf{x}^{\tau}) - \nabla f_i(\mathbf{x}^{\tau-1})\right\|^2\right] \\
\overset{(b)}{\leq} \quad & (t-t_0)\, L^2\sum_{\tau=t_0}^{t-1}\mathbb{E}\left[\left\|\mathbf{x}^{\tau} - \mathbf{x}^{\tau-1}\right\|^2\right] \\
\overset{(c)}{=} \quad & (t-t_0)\, L^2\gamma^2\sum_{\tau=t_0}^{t-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{j=1}^{N}\mathbf{g}_j^{T_j^{\tau}}\right\|^2\right] \\
\overset{(d)}{\leq} \quad & (t-t_0)\, L^2\gamma^2\frac{1}{N}\sum_{\tau=t_0}^{t-1}\sum_{j=1}^{N}\mathbb{E}\left[\left\|\mathbf{g}_j^{T_j^{\tau}}\right\|^2\right] \\
\overset{(e)}{\leq} \quad & (t-t_0)^2\, L^2\gamma^2 G^2,
\end{aligned} \tag{33}
$$

where (a) and (d) follows from the convexity of $\|\cdot\|^2$; (b) follows from Assumption 1; (c) follows from (5) and (6); (e) follows from Lemma 2. □

Corollary 2. *Corollary of Lemma 4:*

$$
\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{t_0-1})\right\|^2\right] \leq (t-t_0)^2 L^2\gamma^2 G^2.
$$

Proof of Corollary 2.

$$
\begin{aligned}
& \mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{t_0-1})\right\|^2\right] \\
\overset{(a)}{\leq} \quad & \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla f_i\left(\mathbf{x}^{t-1}\right) - \nabla f_i(\mathbf{x}^{t_0-1})\right\|^2\right] \\
\overset{(b)}{\leq} \quad & (t-t_0)^2 L^2\gamma^2 G^2,
\end{aligned} \tag{34}
$$

where (a) follows from the convexity of $\|\cdot\|^2$; (b) follows from Lemma 4. □

Main Proof of Theorem 3. From Assumption 1, local objective functions f_i are all $L-smooth$, and thus the global objective function f, which is the mean of them, is also $L-smooth$. Hence, fixing $t \geq 1$, we have

$$
\mathbb{E}\left[f(\mathbf{x}^t)\right] \leq \mathbb{E}\left[f(\mathbf{x}^{t-1})\right] + \frac{L}{2}\mathbb{E}\left[\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2\right] + \mathbb{E}\left[\left\langle\nabla f(\mathbf{x}^{t-1}), \mathbf{x}^t - \mathbf{x}^{t-1}\right\rangle\right], \tag{35}
$$

which corresponds to (9) in Example 1.

Corresponding to (10)–(12) in Example 1, we decompose the terms on the right. During the decomposition, we refer to Lemma 4 and Corollary 2, the proof of which corresponds to (13)–(16) in Example 1. Specifically, we first focuse on the second term on the right:

$$
\begin{aligned}
& \mathbb{E}\left[\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2\right] \\
\overset{(a)}{=} \quad & \gamma^2 \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} \mathbf{g}_i^{T_i^t}\right\|^2\right] \\
= \quad & \gamma^2 \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} \left(\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t-1})\right) + \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(\mathbf{x}^{T_i^t-1})\right\|^2\right] \\
\overset{(b)}{\leq} \quad & 2\gamma^2 \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} \left(\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t-1})\right)\right\|^2\right] + 2\gamma^2 \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} \nabla f_i(\mathbf{x}^{T_i^t-1})\right\|^2\right] \\
\overset{(c)}{=} \quad & \frac{2\gamma^2}{N^2}\sum_{i=1}^{N} \mathbb{E}\left[\left\|\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t-1})\right\|^2\right] + 2\gamma^2 \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} \nabla f_i(\mathbf{x}^{T_i^t-1})\right\|^2\right] \\
\overset{(d)}{\leq} \quad & \frac{2\gamma^2\sigma^2}{N} + 2\gamma^2 \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} \nabla f_i(\mathbf{x}^{T_i^t-1})\right\|^2\right],
\end{aligned} \tag{36}
$$

where (a) follows from (5) and (6); (b) follows from the convexity of $\|\cdot\|^2$; (c) follows from Lemma 3; (d) follows from Lemma 2.

Define $T^t \triangleq min_i\,(T_i^t)$. Focus on the third term in 35,

$$
\begin{aligned}
& \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1}), \mathbf{x}^t - \mathbf{x}^{t-1}\right\rangle\right] \\
\overset{(a)}{=} \quad & -\gamma \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1}), \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}_i^{T_i^t}\right\rangle\right] \\
= \quad & -\gamma \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{T^t-1}), \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}_i^{T_i^t}\right\rangle\right] - \gamma \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{T^t-1}), \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}_i^{T_i^t}\right\rangle\right] \\
= \quad & -\gamma \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{T^t-1}), \frac{1}{N}\sum_{i=1}^{N} \left(\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t-1})\right)\right\rangle\right] \\
& -\gamma \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{T^t-1}), \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(\mathbf{x}^{T_i^t-1})\right\rangle\right] \\
& -\gamma \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{T^t-1}), \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}_i^{T_i^t}\right\rangle\right],
\end{aligned} \tag{37}
$$

where (a) follows from (5) and (6). We further focus on the first term in (37):

$$\begin{aligned}
& -\gamma\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{T^t-1}), \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t-1})\right)\right\rangle\right] \\
= \quad & -\frac{\gamma^2 IL}{\sqrt{N}}\mathbb{E}\left[\left\langle \frac{1}{\gamma IL}\left(\nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{T^t-1})\right), \frac{\sqrt{N}}{N}\sum_{i=1}^{N}\left(\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t-1})\right)\right\rangle\right] \\
\overset{(a)}{\leq} \quad & \frac{\gamma^2 IL}{2\sqrt{N}}\mathbb{E}\left[\left\|\frac{1}{\gamma IL}\left(\nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{T^t-1})\right)\right\|^2\right] + \frac{\gamma^2 IL}{2\sqrt{N}}\mathbb{E}\left[\left\|\frac{\sqrt{N}}{N}\sum_{i=1}^{N}\left(\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t-1})\right)\right\|^2\right] \\
\overset{(b)}{\leq} \quad & \frac{\gamma^2 ILG^2}{2\sqrt{N}} + \frac{\gamma^2 IL}{2\sqrt{N}}\mathbb{E}\left[\left\|\frac{\sqrt{N}}{N}\sum_{i=1}^{N}\left(\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t-1})\right)\right\|^2\right] \\
\overset{(c)}{=} \quad & \frac{\gamma^2 ILG^2}{2\sqrt{N}} + \frac{\gamma^2 IL}{2N^{\frac{3}{2}}}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\left(\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t-1})\right)\right\|^2\right] \\
\overset{(d)}{\leq} \quad & \frac{\gamma^2 ILG^2}{2\sqrt{N}} + \frac{\gamma^2 IL\sigma^2}{2\sqrt{N}},
\end{aligned} \tag{38}$$

where (a) follows from CauchySchwarz inequality and AM-GM inequality; (b) follows from Corollary 2 with t_0 assigned as T^t and Lemma 1; (c) follows from Lemma 3; (d) follows from Lemma 2. Then we focus on the second term in 37 (Note that $\gamma < 1/(2L)$ and thus we can extract the root of $1 - 2\gamma L$):

$$\begin{aligned}
& -\gamma\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{T^t-1}), \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^{T_i^t-1})\right\rangle\right] \\
= \quad & -\gamma\mathbb{E}\left[\left\langle \frac{1}{\sqrt{1-2\gamma L}}\left(\nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{T^t-1})\right), \frac{1}{N}\sqrt{1-2\gamma L}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^{T_i^t-1})\right\rangle\right] \\
\overset{(a)}{\leq} \quad & \frac{\gamma}{2\left(1-2\gamma L\right)}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{T^t-1})\right\|^2\right] + \frac{\gamma\left(1-2\gamma L\right)}{2}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^{T_i^t-1})\right\|^2\right] \\
\overset{(b)}{\leq} \quad & \frac{\gamma^3 I^2 L^2 G^2}{2\left(1-2\gamma L\right)} + \frac{\gamma\left(1-2\gamma L\right)}{2}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^{T_i^t-1})\right\|^2\right],
\end{aligned} \tag{39}$$

where (a) follows from CauchySchwarz inequality and AM-GM inequality; (b) follows from Corollary 2 and Lemma 1. We finally focus on the third term in (37):

$$
\begin{aligned}
& \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{T^t-1}), \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i^{T_i^t}\right\rangle\right] \\
\overset{(a)}{=} \; & \mathbb{E}\left[\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{T^t-1}), \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i^{T_i^t}\right\rangle \mid \xi^{\left[T^t-1\right]}\right]\right] \\
= \; & \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{T^t-1}), \mathbf{g}_i^{T_i^t}\right\rangle \mid \xi^{\left[T^t-1\right]}\right]\right] \\
\overset{(b)}{=} \; & \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{T^t-1}), \mathbf{g}_i^{T_i^t}\right\rangle \mid \xi^{\left[T_i^t-1\right]}\right] \mid \xi^{\left[T^t-1\right]}\right]\right] \\
\overset{(c)}{=} \; & \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{T^t-1}), \nabla f_i(\mathbf{x}^{T_i^t-1})\right\rangle \mid \xi^{\left[T^t-1\right]}\right]\right] \\
\overset{(d)}{=} \; & \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{T^t-1}), \nabla f_i(\mathbf{x}^{T_i^t-1})\right\rangle\right] \\
= \; & \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{T^t-1}), \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^{T_i^t-1})\right\rangle\right] \\
\overset{(e)}{=} \; & \frac{1}{2}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{T^t-1})\right\|^2\right] + \frac{1}{2}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^{T_i^t-1})\right\|^2\right] - \frac{1}{2}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{T^t-1}) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^{T_i^t-1})\right\|^2\right], \quad (40)
\end{aligned}
$$

where (a), (b) and (d) follows from Law of Total Expectation $\mathbb{E}[\mathbb{E}[\mathbf{X}|\mathbf{Y}]] = \mathbb{E}[\mathbf{X}]$; (c) follows because $\forall i: T^t \leq T_i^t$, and thus $f(\mathbf{x}^{T^t-1})$ is determined by $\xi^{\left[T_i^t-1\right]}$; (e) follows from $\langle \mathbf{u}, \mathbf{v}\rangle = \frac{1}{2}(\| \mathbf{u} \|^2 + \| \mathbf{v} \|^2 - \| \mathbf{u} - \mathbf{v} \|^2)$. In (40), we further deal with the last term,

$$
\begin{aligned}
& \mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{T^t-1}) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^{T_i^t-1})\right\|^2\right] \\
= \; & \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\nabla f_i(\mathbf{x}^{T^t-1}) - \nabla f_i(\mathbf{x}^{T_i^t-1})\right)\right\|^2\right] \\
\overset{(a)}{\leq} \; & \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\left(\nabla f_i(\mathbf{x}^{T_i^t-1}) - \nabla f_i(\mathbf{x}^{T^t-1})\right)\right\|^2\right] \\
\overset{(b)}{\leq} \; & \frac{1}{N}\sum_{i=1}^{N}(T_i^t - T^t)^2 L^2\gamma^2 G^2 \\
\overset{(c)}{\leq} \; & I^2 L^2 \gamma^2 G^2, \quad (41)
\end{aligned}
$$

where (a) follows from the convexity of $\|\cdot\|^2$; (b) follows from Lemma 4 with t assigned as T_i^t and t_0 assigned as T^t; (c) follows from Lemma 1, $T_i^t \leq t$, and $T^t = \min_i (T_i^t)$. Substituting (41) into (40) and (38)–(40) into (37), we have:

$$
\begin{aligned}
& \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1}), \mathbf{x}^t - \mathbf{x}^{t-1}\right\rangle\right] \\
\leq \; & \frac{\gamma^2 ILG^2}{2\sqrt{N}} + \frac{\gamma^2 IL\sigma^2}{2\sqrt{N}} + \frac{\gamma^3 I^2 L^2 G^2}{2(1-2\gamma L)} + \frac{\gamma^3 I^2 L^2 G^2}{2} - \gamma^2 L\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^{T_i^t-1})\right\|^2\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{T^t-1})\right\|^2\right].
\end{aligned}
\quad (42)
$$

Further substituting (36) and (42) into (35), we have

$$\mathbb{E}\left[f(\mathbf{x}^t)\right] - \mathbb{E}\left[f(\mathbf{x}^{t-1})\right] \leq \frac{\gamma^2\sigma^2 L}{N} + \frac{\gamma^2 IL\left(G^2+\sigma^2\right)}{2\sqrt{N}} + \frac{\gamma^3 I^2 L^2 G^2}{2\left(1-2\gamma L\right)} + \frac{\gamma^3 I^2 L^2 G^2}{2} - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{T^t-1})\right\|^2\right]. \quad (43)$$

Corresponding to (17)–(20) in Example 1, we rearrange (43) with summation to obtain the convergence result. We first rearrange (43):

$$\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{T^t-1})\right\|^2\right] \leq \frac{2\gamma\sigma^2 L}{N} + \frac{\gamma IL\left(G^2+\sigma^2\right)}{\sqrt{N}} + \frac{\gamma^2 I^2 L^2 G^2}{1-2\gamma L} + \gamma^2 I^2 L^2 G^2 + \frac{2}{\gamma}\left(\mathbb{E}\left[f(\mathbf{x}^{t-1})\right] - \mathbb{E}\left[f(\mathbf{x}^t)\right]\right). \quad (44)$$

Summing (44) over iterations from 1 to T and deviding both sides by T, we have

$$\begin{aligned} & \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{T^t-1})\right\|^2\right] \\ \leq\ & \frac{2\gamma\sigma^2 L}{N} + \frac{\gamma IL\left(G^2+\sigma^2\right)}{\sqrt{N}} + \frac{\gamma^2 I^2 L^2 G^2}{1-2\gamma L} + \gamma^2 I^2 L^2 G^2 + \frac{2}{\gamma T}\left(\mathbb{E}\left[f(\mathbf{x}^0)\right] - \mathbb{E}\left[f(\mathbf{x}^*)\right]\right), \end{aligned} \quad (45)$$

where $\mathbf{x}^*$ is the optimal solution for the global objective function $f(\mathbf{x})$.

Finally, we build the gap between $\nabla f(\mathbf{x}^{t-1})$ and $\nabla f(\mathbf{x}^{T^t-1})$. Lemma 1 implies that $t - T^t \leq I$ since $T^t = \min_i T_i^t$. Hence, we have

$$\begin{aligned} & \mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1})\right\|^2\right] \\ =\ & \mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{T^t-1}) + \nabla f(\mathbf{x}^{T^t-1})\right\|^2\right] \\ \overset{(a)}{\leq}\ & 2\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{T^t-1})\right\|^2\right] + 2\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{T^t-1})\right\|^2\right] \\ \overset{(b)}{\leq}\ & 2\gamma^2 I^2 L^2 G^2 + 2\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{T^t-1})\right\|^2\right], \end{aligned} \quad (46)$$

where (a) follows from the convexity of $\|\cdot\|^2$; (b) follows from Corollary 2. Sum (46) over iterations from 1 to T, devide both sides by T, and substitute (45) into it. We then have

$$\begin{aligned} & \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1})\right\|^2\right] \\ \leq\ & \frac{4\gamma\sigma^2 L}{N} + \frac{2\gamma IL\left(G^2+\sigma^2\right)}{\sqrt{N}} + \frac{2\gamma^2 I^2 L^2 G^2}{1-2\gamma L} + 4\gamma^2 I^2 L^2 G^2 + \frac{4}{\gamma T}\left(\mathbb{E}\left[f(\mathbf{x}^0)\right] - \mathbb{E}\left[f(\mathbf{x}^*)\right]\right). \end{aligned} \quad (47)$$

□

D. Proof of Corollary 1

Proof of Corollary 1. We first summarize the $O(\cdot)$ form of Theorem 3:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1})\right\|^2\right] = O\left(\frac{\gamma IL\left(G^2+\sigma^2\right)}{\sqrt{N}} + \frac{\gamma^2 I^2 L^2 G^2}{1-2\gamma L} + \frac{B}{\gamma T}\right). \quad (48)$$

Substituting γ with $(\beta^{1/2}N^{1/4})/(2LE^{1/2}T^{1/2})$, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1})\right\|^2\right] &\overset{(a)}{=} O\left(\frac{\gamma IL\left(G^2+\sigma^2\right)}{\sqrt{N}}+I^2\gamma^2L^2G^2+\frac{B}{\gamma T}\right)\\
&= O\left(\frac{I\beta^{\frac{1}{2}}\left(G^2+\sigma^2\right)}{N^{\frac{1}{4}}E^{\frac{1}{2}}T^{\frac{1}{2}}}+\frac{I^2\beta G^2N^{\frac{1}{2}}}{ET}+\frac{BLE^{\frac{1}{2}}}{\beta^{\frac{1}{2}}N^{\frac{1}{4}}T^{\frac{1}{2}}}\right)\\
&\overset{(b)}{=} O\left(\frac{E^{\frac{1}{2}}\left(G^2+\sigma^2+BL\right)}{\beta^{\frac{1}{2}}N^{\frac{1}{4}}T^{\frac{1}{2}}}+\frac{EG^2N^{\frac{1}{2}}}{\beta T}\right),
\end{aligned} \tag{49}$$

where (a) follows because $\gamma \leq 1/(4L)$, and thus $1-2\gamma L > 1/2$; (b) follows because from Lemma 1, $I = \lceil N/K\rceil E = O(E/\beta)$.

When $T \geq EN^{3/2}/\beta$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1})\right\|^2\right] = O\left(\frac{E^{\frac{1}{2}}\left(G^2+\sigma^2+BL\right)}{\beta^{\frac{1}{2}}N^{\frac{1}{4}}T^{\frac{1}{2}}}\right) = O\left(\frac{1}{N^{\frac{1}{4}}T^{\frac{1}{2}}}\right), \tag{50}$$

where the final equation follows if we care only about N and T, and regard other parameters as constants. □

As shown in Section 1, FedAvg is proven to achieve $O(1/\sqrt{NT})$ convergence when all clients participate in each training iteration. However, we can prove only the $O(1/(N^{1/4}T^{1/2}))$ convergence for FedLaAvg because of the partial client participation as a result of the intermittent client availability. Specifically, this gap is introduced by (37). The randomness of the stochastic gradient $\mathbf{g}_i^{T_i^t}$ is an obstacle for the convergence analysis. With full client participation, we can reduce this randomness by the following equations:

$$\begin{aligned}
&\mathbb{E}\left[\left\langle\nabla f(\mathbf{x}^{t-1}), \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i^{T_i^t}\right\rangle\right] = \mathbb{E}\left[\left\langle\nabla f(\mathbf{x}^{t-1}), \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i^{t}\right\rangle\right]\\
=\;& \mathbb{E}\left[\mathbb{E}\left[\left\langle\nabla f(\mathbf{x}^{t-1}), \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i^{t}\right\rangle \mid \xi^{[t-1]}\right]\right] = \mathbb{E}\left[\left\langle\nabla f(\mathbf{x}^{t-1}), \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^{t-1})\right\rangle\right].
\end{aligned} \tag{51}$$

However, with partial client participation, (51) no longer holds. We analyze the gap between $\mathbb{E}\left[\left\langle\nabla f(\mathbf{x}^{t-1}), \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i^{T_i^t}\right\rangle\right]$ and $\mathbb{E}\left[\left\langle\nabla f(\mathbf{x}^{T^t-1}), \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^{T_i^t-1})\right\rangle\right]$ in (37). Then, we further study the upper bound for the absolute value of the first term of the gap, i.e., $\mathbb{E}\left[\left\langle\nabla f(\mathbf{x}^{t-1})-\nabla f(\mathbf{x}^{T^t-1}), \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{g}_i^{T_i^t}-\nabla f_i(\mathbf{x}^{T_i^t-1})\right)\right\rangle\right]$ in (38). This term is the inner product of two vectors. The norm of the second vector is bounded by $O(1/N)$, but the norm of the fisrt term is not related to N. Hence, the upper bound that we can obtain for the inner product is $O(1/\sqrt{N})$, while the $O(1/\sqrt{NT})$ convergence needs an upper bound in the order of $O(1/N)$. Whether the $O(1/(N^{1/4}T^{1/2}))$ convergence is a tight bound requires further studies.

E. Communication Round-Based FedLaAvg

We introduce some notations to represent the training process of the communication round-based FedLaAvg. Let $\hat{\mathcal{C}}^r \triangleq \bigcap_{t=(r-1)C+1}^{rC}\mathcal{C}^t$ be the set of available clients in round r. Each client i observes the stochastic gradient $\mathbf{g}_i^t$ on the local model parameters in each local iteration, and accumulates these stochastic gradients to obtain the local model update $\mathbf{u}_i^r \triangleq -\gamma\sum_{t=(r-1)C+1}^{rC}\mathbf{g}_i^t$ at round r. After collecting the local model updates until round r, in iteration $t = rC$, the cloud server calculates the global model parameters $\mathbf{x}^{rC}$. Similar to the notation T_i^t, we use R_i^r to denote the latest round where client i is available before or at round r. We define $r^t \triangleq \lfloor(t-1)/C\rfloor+1$ as the round that iteration t belongs to. The one-iteration-per-round scenario corresponds to the special case using the specific notations: $C=1$, $r^t=t$, $\hat{\mathcal{C}}^{r^t}=\mathcal{C}^t$, $\mathbf{u}_i^{r^t}=-\gamma\mathbf{g}_i^t$, and $R_i^{r^t}=T_i^t$. With these notations, we formally illustrate the communication round-based FedLaAvg in Algorithm 2.

We replace Assumption 4 with Assumption 5 to capture the client availability in the multi-iterations-per-round scenario.

Algorithm 2 The Communication Round-Based Federated Latest Averaging Algorithm

Input: initial model $\mathbf{x}^0$; number of clients N; number of total iterations T; learning rate γ; number of local iterations C; proportion of selected clients β (i.e., the number of participating clients in each iteration is $K = \beta N$.)

Do initialization:

$$\mathbf{u}^0 \leftarrow \mathbf{0},\ \forall i \in \{1, 2, \cdots, N\},\ \mathbf{u}_i^0 \leftarrow \mathbf{0},\ R_i^0 \leftarrow 0.$$

for $r = 1$ **to** R **do**

  $\mathbf{u}^r \leftarrow \mathbf{u}^{r-1}$

  $\hat{\mathcal{C}}^r \leftarrow$ the set of available clients

  $\hat{\mathcal{B}}^r \leftarrow K$ clients from $\hat{\mathcal{C}}^r$ with the lowest R_i^{r-1} values

  Update R_i^r values:

$$R_i^r \leftarrow r,\ \forall i \in \hat{\mathcal{B}}^r;\ R_i^r \leftarrow R_i^{r-1},\ \forall i \notin \hat{\mathcal{B}}^r. \tag{52}$$

  Each client $i \in \hat{\mathcal{B}}^r$ calculates the accumulated local model update $\mathbf{u}_i^r$ and uploads the update difference $\mathbf{u}_i^r - \mathbf{u}_i^{R_i^{r-1}}$ in parallel.

  Once receiving the update information from client i, the cloud server calculates the global update:

$$\mathbf{u}^r \leftarrow \mathbf{u}^r + \frac{1}{N}\left(\mathbf{u}_i^r - \mathbf{u}_i^{R_i^{r-1}}\right). \tag{53}$$

  The cloud server updates the global model parameters:

$$\mathbf{x}^{rC} \leftarrow \mathbf{x}^{rC-C} + \mathbf{u}^r. \tag{54}$$

end for

Assumption 5. *Minimal availability: each client i is available at least once in any period with E successive rounds:*

$$\forall i,\ \forall r,\ \exists \hat{r} \in \{r, r+1, \cdots, r+E-1\},\ \textit{such that } i \in \hat{\mathcal{C}}^{\hat{r}}.$$

Under such assumption, we can establish the following convergence result of the communication round-based FedLaAvg.

Theorem 4. *We recall that $B = f(\mathbf{x}^0) - f(\mathbf{x}^*)$ and $\beta = K/N$. Let the communication round-based FedLaAvg execute R rounds. By choosing the learning rate $\gamma = (\beta^{1/2}N^{1/4})/(2LCE^{1/2}R^{1/2})$ and requiring $\gamma \le 1/(4L)$, we have the following convergence result:*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{rC-1})\right\|^2\right] = O\left(\frac{E^{\frac{1}{2}}\left(G^2+\sigma^2+LB\right)}{N^{\frac{1}{4}}R^{\frac{1}{2}}\beta^{\frac{1}{2}}} + \frac{EG^2N^{\frac{1}{2}}}{R\beta}\right). \tag{55}$$

When $R \ge EN^{3/2}/\beta$, we further have

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{rC-1})\right\|^2\right] = O\left(\frac{E^{\frac{1}{2}}\left(G^2+\sigma^2\right)+BL}{\beta^{\frac{1}{2}}N^{\frac{1}{4}}R^{\frac{1}{2}}}\right) = O\left(\frac{1}{N^{\frac{1}{4}}R^{\frac{1}{2}}}\right). \tag{56}$$

Proof of Theorem 4. To make the proof more concise, we introduce an mathematically equivalent Algorithm 3 of Algorithm 2. Note that $\mathbf{x}^t$ (when t is not multiple of C) is intermediate variable for mathematical analysis. In addition, $\mathbf{g}_i^\tau$ ($\tau \le 0$) is extraly defined to avoid undefined symbols when $R_i^{r^t} = 0$ in (57). It can be proved by induction that all variables defined in Algorithm 2 are consistent with those in Algorithm 3.

With this equavalence and additional notations, we indroduce equavalent corresponding lemmas of Lemmas 1–4.

Lemma 5. *Under Assumption 5, following Algorithm 2, with $I = \lceil N/K \rceil E - 1$, $\forall r, \forall i$, we have*

$$r - R_i^r \le I. \tag{59}$$

Proof of Lemma 5. Replacing t with r and T_i^t with R_i^r, the proof is exactly the same with that of Lemma 1. □

Algorithm 3 An equivalent Algorithm of Algorithm 2

Input: Initial model $\mathbf{x}^0$
$\mathbf{g}_i^\tau \leftarrow \mathbf{0}$, $\forall i \in \{1, 2, \cdots, N\}$, $\tau \in \{0, -1, \cdots, 1-C\}$
$R_i^0 \leftarrow \mathbf{0}$, $\forall i \in \{1, 2, \cdots, N\}$
for $t = 1$ **to** RC **do**
  $r^t \leftarrow \lfloor (t-1)/C \rfloor + 1$
  if $t-1$ is a multiple of C **then**
    $\hat{\mathcal{C}}^{r^t} \leftarrow$ the set of available clients in round r^t
    $\hat{\mathcal{B}}^{r^t} \leftarrow K$ clients from $\hat{\mathcal{C}}^{r^t}$ with the lowest $R_i^{r^t-1}$ values
    Update $R_i^{r^t}$ values: $R_i^{r^t} \leftarrow r^t, \forall i \in \hat{\mathcal{B}}^{r^t}$; $R_i^{r^t} \leftarrow R_i^{r^t-1}, \forall i \notin \hat{\mathcal{B}}^{r^t}$.
    $\mathbf{x}_i^{t-1} \leftarrow \mathbf{x}^{t-1}, \forall i \in \hat{\mathcal{B}}^{r^t}$
  end if
  $\mathbf{g}_i^t \leftarrow \nabla F\left(\mathbf{x}_i^{t-1}; \xi_i^t\right), \forall i \in \hat{\mathcal{B}}^{r^t}$
  Update the global model parameters:

$$\mathbf{x}^t \leftarrow \mathbf{x}^{t-1} - \gamma \sum_{i=1}^{N} \mathbf{g}_i^{R_i^{r^t} C - r^t C + t}. \tag{57}$$

  Update the local model parameters:

$$\mathbf{x}_i^t \leftarrow \mathbf{x}_i^{t-1} - \gamma \mathbf{g}_i^t. \tag{58}$$

end for

Lemma 6. *Corresponding lemma of Lemma 2:*

$$\mathbb{E}\left[\left\|\mathbf{g}_i^t\right\|^2\right] \le G^2, \forall i, \forall t;$$

$$\mathbb{E}\left[\left\|\mathbf{g}_i^t - \nabla f_i(\mathbf{x}_i^{t-1})\right\|^2\right] \le \sigma^2, \forall i, \forall t.$$

Proof of Lemma 6. Replacing $\mathbf{x}^{t-1}$ with $\mathbf{x}_i^{t-1}$, the proof is exactly the same with that of Lemma 2. $\square$

Lemma 7. *Corresponding lemma of Lemma 3: $\forall i, \forall t$, we have*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{N}\left(\mathbf{g}_i^{R_i^{r^t} C - r^t C + t} - \nabla f_i(\mathbf{x}_i^{R_i^{r^t} C - r^t C + t - 1})\right)\right\|^2\right] = \sum_{i=1}^{N} \mathbb{E}\left[\left\|\mathbf{g}_i^{R_i^{r^t} C - r^t C + t} - \nabla f_i(\mathbf{x}_i^{R_i^{r^t} C - r^t C + t - 1})\right\|^2\right].$$

Proof of Lemma 7. Replacing $\mathbf{g}_i^{T_i^t}$ with $\mathbf{g}_i^{R_i^{r^t} C - r^t C + t}$ and $\nabla f_i(\mathbf{x}^{T_i^t - 1})$ with $\nabla f_i(\mathbf{x}_i^{R_i^{r^t} C - r^t C + t - 1})$, the proof is exactly the same with that of Lemma 3. $\square$

Note that Lemma 4 and Corollary 2 still hold. Their proof follows as well if we replace the relation $\mathbf{x}^\tau - \mathbf{x}^{\tau-1} = \sum_{j=1}^{N} \mathbf{g}_j^{T_j^\tau}$ with $\mathbf{x}^\tau - \mathbf{x}^{\tau-1} = \sum_{j=1}^{N} \mathbf{g}_j^{R_j^{r_\tau} C - r_\tau C + \tau}$.

Main proof of Theorem 4. The proof is similar to that of Theorem 3 and Corollary 1. We illustrate it in detail as follow.

Fix $t \ge 1$, by Assumption 1, we have

$$\mathbb{E}\left[f(\mathbf{x}^t)\right] \le \mathbb{E}\left[f(\mathbf{x}^{t-1})\right] + \frac{L}{2}\mathbb{E}\left[\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2\right] + \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1}), \mathbf{x}^t - \mathbf{x}^{t-1}\right\rangle\right]. \tag{60}$$

Focus on the second term on the right. Following the procedure of (36), we omit the intermediate results and show the final bound:

$$\mathbb{E}\left[\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2\right] \leq \frac{2\gamma^2\sigma^2}{N}+2\gamma^2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{R_i^{r^t}C-r^tC+t-1})\right\|^2\right]. \tag{61}$$

For simplicity, we define $\hat{T}^t$ as $\min_i \left(R_i^{r^t}C-r^tC+t\right)$. Focus on the third term in (60), we can separate it into 3 parts

$$\begin{aligned}
& \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1}), \mathbf{x}^t-\mathbf{x}^{t-1}\right\rangle\right] \\
= \; & -\gamma\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1})-\nabla f(\mathbf{x}^{\hat{T}^t-1}), \frac{1}{N}\sum_{i=1}^N\left(\mathbf{g}_i^{R_i^{r^t}C-r^tC+t}-\nabla f_i(\mathbf{x}_i^{R_i^{r^t}C-r^tC+t-1})\right)\right\rangle\right] \\
& -\gamma\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1})-\nabla f(\mathbf{x}^{\hat{T}^t-1}), \frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{R_i^{r^t}C-r^tC+t-1})\right\rangle\right] \\
& -\gamma\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{\hat{T}^t-1}), \frac{1}{N}\sum_{i=1}^N \mathbf{g}_i^{R_i^{r^t}C-r^tC+t}\right\rangle\right].
\end{aligned} \tag{62}$$

We further focus on the first term in (62). Following the procedure of (38), we have the following bound:

$$-\gamma\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1})-\nabla f(\mathbf{x}^{\hat{T}^t-1}), \frac{1}{N}\sum_{i=1}^N\left(\mathbf{g}_i^{R_i^{r^t}C-r^tC+t}-\nabla f_i(\mathbf{x}_i^{R_i^{r^t}C-r^tC+t-1})\right)\right\rangle\right] \leq \frac{\gamma^2 ICL\left(G^2+\sigma^2\right)}{2\sqrt{N}}. \tag{63}$$

Then we focus on the second term in (62). Following the procedure of (39), we have

$$\begin{aligned}
& \gamma\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1})-\nabla f(\mathbf{x}^{\hat{T}^t-1}), \frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{R_i^{r^t}C-r^tC+t-1})\right\rangle\right] \\
\leq \; & \frac{\gamma^3 I^2C^2L^2G^2}{2\left(1-2\gamma L\right)}+\frac{\gamma\left(1-2\gamma L\right)}{2}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{R_i^{r^t}C-r^tC+t-1})\right\|^2\right].
\end{aligned} \tag{64}$$

We finally focus on the third term in (62). Following the procedure of (40) and (41), we have

$$\begin{aligned}
& \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{\hat{T}^t-1}), \frac{1}{N}\sum_{i=1}^N \mathbf{g}_i^{R_i^{r^t}C-r^tC+t}\right\rangle\right] \\
\overset{(b)}{=} \; & \frac{1}{2}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{\hat{T}^t-1})\right\|^2\right]+\frac{1}{2}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{R_i^{r^t}C-r^tC+t-1})\right\|^2\right]-\frac{1}{2}I^2L^2C^2\gamma^2G^2.
\end{aligned} \tag{65}$$

Substituting (63)–(65) into (62), we have:

$$\begin{aligned}
& \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1}), \mathbf{x}^t-\mathbf{x}^{t-1}\right\rangle\right] \\
\leq \; & \frac{\gamma^2 ICL\left(G^2+\sigma^2\right)}{2\sqrt{N}}+\frac{\gamma^3 I^2C^2L^2G^2}{2\left(1-2\gamma L\right)}+\frac{\gamma^3 I^2C^2L^2G^2}{2} \\
& -\gamma^2 L\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{R_i^{r^t}C-r^tC+t-1})\right\|^2\right]-\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{\hat{T}^t-1})\right\|^2\right].
\end{aligned} \tag{66}$$

Further substituting (61) and (66) into (60), we have

$$\begin{aligned}
& \mathbb{E}\left[f(\mathbf{x}^t)\right]-\mathbb{E}\left[f(\mathbf{x}^{t-1})\right] \\
\leq \; & \frac{\gamma^2\sigma^2 L}{N}+\frac{\gamma^2 ICL\left(G^2+\sigma^2\right)}{2\sqrt{N}}+\frac{\gamma^3 I^2C^2L^2G^2}{2\left(1-2\gamma L\right)}+\frac{\gamma^3 I^2C^2L^2G^2}{2}-\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{\hat{T}^t-1})\right\|^2\right].
\end{aligned} \tag{67}$$

Rearrange the above equation and we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{\hat{T}^t-1})\right\|^2\right]\\ \leq\ & \frac{2\gamma\sigma^2L}{N}+\frac{\gamma ICL\left(G^2+\sigma^2\right)}{\sqrt{N}}+\frac{\gamma^2I^2C^2L^2G^2}{(1-2\gamma L)}+\gamma^2I^2C^2L^2G^2+\frac{2}{\gamma}\left(\mathbb{E}\left[f(\mathbf{x}^{t-1})\right]-\mathbb{E}\left[f(\mathbf{x}^t)\right]\right).\end{aligned} \tag{68}$$

Summing (68) over iterations from 1 to RC and deviding both sides by RC, we have

$$\begin{aligned}&\frac{1}{RC}\sum_{t=1}^{RC}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{\hat{T}^t-1})\right\|^2\right]\\ \leq\ & \frac{2\gamma\sigma^2L}{N}+\frac{\gamma ICL\left(G^2+\sigma^2\right)}{\sqrt{N}}+\frac{\gamma^2I^2C^2L^2G^2}{(1-2\gamma L)}+\gamma^2I^2C^2L^2G^2+\frac{2}{\gamma RC}\left(\mathbb{E}\left[f(\mathbf{x}^0)\right]-\mathbb{E}\left[f(\mathbf{x}^*)\right]\right),\end{aligned} \tag{69}$$

where $\mathbf{x}^*$ is the optimal value for the objective function $f(\mathbf{x})$.

Finally, we build the gap between $\nabla f(\mathbf{x}^{r^tC-1})$ and $\nabla f(\mathbf{x}^{\hat{T}^t-1})$. Lemma 5 implies that $t-\hat{T}^t\leq IC$, thus $r^tC-\hat{T}^t\leq(I+1)C$. Hence, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{r^tC-1})\right\|^2\right]\\ \leq\ & 2\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{r^tC-1})-\frac{1}{C}\sum_{\tau=(r_t-1)C+1}^{r_tC}\nabla f(\mathbf{x}^{\hat{T}^\tau-1})\right\|^2\right]+2\mathbb{E}\left[\left\|\frac{1}{C}\sum_{\tau=(r_t-1)C+1}^{r_tC}\nabla f(\mathbf{x}^{\hat{T}^\tau-1})\right\|^2\right]\\ \leq\ & 2\gamma^2\left(I+1\right)^2C^2L^2G^2+2\frac{1}{C}\sum_{\tau=(r_t-1)C+1}^{r_tC}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{\hat{T}^\tau-1})\right\|^2\right],\end{aligned} \tag{70}$$

which follows from the convexity of $\|\cdot\|^2$and Corollary 2.

Summing 70 over $t\in\{C,2C,\cdots,RC\}$, deviding both sides by R and substituting 69 into it, we have

$$\begin{aligned}&\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{rC-1})\right\|^2\right]\\ \leq\ & \frac{4\gamma\sigma^2L}{N}+\frac{2\gamma ICL\left(G^2+\sigma^2\right)}{\sqrt{N}}+\left(\frac{2I^2}{(1-2\gamma L)}+4I^2+4I+2\right)\gamma^2C^2L^2G^2+\frac{4}{\gamma RC}\left(\mathbb{E}\left[f(\mathbf{x}^0)\right]-\mathbb{E}\left[f(\mathbf{x}^*)\right]\right).\end{aligned} \tag{71}$$

Then, we write the $O(\cdot)$ expression of the above equation:

$$\frac{1}{R}\sum_{t=1}^{R}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{rC-1})\right\|^2\right]=O\left(\frac{\gamma ICL\left(G^2+\sigma^2\right)}{\sqrt{N}}+\frac{I^2\gamma^2C^2L^2G^2}{(1-2\gamma L)}+\frac{B}{\gamma RC}\right). \tag{72}$$

Substituting γ with $(\beta^{1/2}N^{1/4})/(2LCE^{1/2}R^{1/2})$, we have

$$\frac{1}{R}\sum_{t=1}^{R}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{rC-1})\right\|^2\right]=O\left(\frac{E^{\frac{1}{2}}\left(G^2+\sigma^2+BL\right)}{\beta^{\frac{1}{2}}N^{\frac{1}{4}}R^{\frac{1}{2}}}+\frac{EG^2N^{\frac{1}{2}}}{\beta R}\right). \tag{73}$$

If we further choose $R>EN^{3/2}/\beta$, we have

$$\frac{1}{R}\sum_{t=1}^{R}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{rC-1})\right\|^2\right]=O\left(\frac{E^{\frac{1}{2}}\left(G^2+\sigma^2\right)+BL}{\beta^{\frac{1}{2}}N^{\frac{1}{4}}R^{\frac{1}{2}}}\right)=O\left(\frac{1}{N^{\frac{1}{4}}R^{\frac{1}{2}}}\right). \tag{74}$$

The final equation follows if we care only about N and R, and regard other parameters as constants. □

Table 4. The initial learning rate and learning rate decay settings for each experiment. Experiment names are consistent with the legends in Figures 2 and 3.

Experiment	Initial Learning Rate	Learning Rate Decay
sequential SGD	0.00100	0.99990
FedAvg (FCA)	0.10000	0.99990
FedAvg (ICA)	0.10000	0.99990
FedLaAvg (FCA)	0.03000	0.99990
FedLaAvg (ICA)	0.01000	0.99990
FedLaAvg ($N = 200$)	0.00600	0.99990
FedLaAvg ($N = 400$)	0.00795	0.99990
FedLaAvg ($N = 600$)	0.00880	0.99990
FedLaAvg ($N = 800$)	0.00900	0.99990
FedLaAvg ($N = 1000$)	0.01000	0.99990
FedLaAvg ($E = 1$)	0.03000	0.99990
FedLaAvg ($E = 5$)	0.01410	0.99990
FedLaAvg ($E = 15$)	0.00816	0.99990
FedLaAvg ($\beta = 0.02$)	0.00250	0.99988
FedLaAvg ($\beta = 0.04$)	0.01000	0.99980
FedLaAvg ($\beta = 0.06$)	0.05000	0.99980
FedLaAvg ($\beta = 0.08$)	0.10000	0.99980
FedLaAvg ($\beta = 0.10$)	0.01000	0.99990

$\square$

FedAvg lets each client perform multiple local iterations to achieve communication efficiency. However, the theoretical analysis can not support the substantial improvement in communication efficiency for the communication round-based FedLaAvg. The main reason is that with C local iterations, the difference between the latest gradient and the current gradient is roughly C times larger, and we have to choose smaller γ to guarantee convergence. Since we finally choose $\gamma \propto 1/C$, increasing local iteration number can help reduce the variance term $4\gamma\sigma^2 L/N$ in (71), which is not the dominating term, however. To achieve substantial improvement in communication efficiency, some adjustment to the algorithm may be necessary. This is a promising direction for future research.

F. Complexity Analysis

We analyze the time and space complexity of Algorithms 1 and 2 in this appendix. We use P to denote the time complexity of one backpropagation and Q to denote the number of parameters in the deep learning model.

In each iteration of Algorithm 1, each client performs one backpropagation to obtain the local gradient and computes the gradient difference. This requires $O(P + Q)$ time complexity per client per iteration and $O(Q)$ space complexity to locally store the gradient calculated in the previous participating iteration. The cloud server selects K clients from $\mathcal{C}^t$ in each iteration t. Our implementation is sorting an array of T_i^{t-1} first and picking the K clients from $\mathcal{C}^t$ with the lowest T_i^{t-1} according to the sorted array. This requires $O(N \log N)$ time complexity to sort the array and $O(N)$ space complexity to store the array. Then, the cloud server aggregates the gradient difference to obtain the average latest gradient $\mathbf{g}^t$, and update the global model. This requires $O(KQ)$ time complexity and $O(Q)$ space complexity. To summarize, the time complexity of each iteration in Algorithm 1 is $O(P + Q)$ on each client and $O(N \log N + KQ)$ on the cloud server. The space complexity is $O(Q)$ on each client and $O(N + Q)$ on the cloud server. By similar analysis, the time complexity of each round in Algorithm 2 is $O(CP + Q)$ on each client and $O(N \log N + KQ)$ on the cloud server. The space complexity is $O(Q)$ on each client and $O(N + Q)$ on the cloud server.

G. Learning Rate Settings and Supplementary Notes for the Experiment Environment

Detailed learning rate settings are illustrated in Table 4. Note that FedLaAvg ($N = 1000$), FedLaAvg ($E = 10$), and FedLaAvg ($\beta = 0.1$) are the same experiment as FedLaAvg (ICA). In addition, FedLaAvg ($E = 1$) is the same experiment as FedLaAvg (FCA). We adopt exponential learning rate decay to avoid too elaborate tuning on the learning rate. In practice, multistep decay (learning rate decays only at specific rounds) is a better choice for higher test accuracy and faster convergence. Generally, with smaller N, larger E, or smaller β, smaller learning rate is better. This is consistent with Corollary 1. In some experiments, instead of smaller initial learning rate, smaller learning rate decay works better for FedLaAvg.

The CIFAR-10 dataset is available on `http://www.cs.toronto.edu/~kriz/cifar.html`, and can be downloaded automatically by our source code. The deep learning model architecture is taken from `https://pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html`. In addition, experiments are conducted on machines with operating system Ubuntu 18.04.3, CUDA version 10.1, and one NVIDIA GeForce RTX 2080Ti GPU.