Efficient Federated Learning against Heterogeneous and Non-stationary Client Unavailability

Ming Xiang[1] **Stratis Ioannidis**[1] **Edmund Yeh**[1] **Carlee Joe-Wong**[2] **Lili Su**[1]
[1]Northeastern University, Boston, MA [2]Carnegie Mellon University, Pittsburgh, PA
{xiang.mi,l.su}@northeastern.edu
{ioannidis,eyeh}@ece.neu.edu
cjoewong@andrew.cmu.edu

Abstract

Addressing intermittent client availability is critical for the real-world deployment of federated learning algorithms. Most prior work either overlooks the potential non-stationarity in the dynamics of client unavailability or requires substantial memory/computation overhead. We study federated learning in the presence of heterogeneous and non-stationary client availability, which may occur when the deployment environments are uncertain, or the clients are mobile. The impacts of heterogeneity and non-stationarity on client unavailability can be significant, as we illustrate using `FedAvg`, the most widely adopted federated learning algorithm. We propose `FedAWE`, which includes novel algorithmic structures that (i) compensate for missed computations due to unavailability with only $O(1)$ additional memory and computation with respect to standard `FedAvg`, and (ii) evenly diffuse local updates within the federated learning system through implicit gossiping, despite being agnostic to non-stationary dynamics. We show that `FedAWE` converges to a stationary point of even non-convex objectives while achieving the desired linear speedup property. We corroborate our analysis with numerical experiments over diversified client unavailability dynamics on real-world data sets.

1 Introduction

Federated learning is a distributed machine learning approach that enables training global models without disclosing raw local data [31, 20]. It has been adopted in commercial applications such as autonomous vehicles [6, 69, 40], the Internet of things [38], and natural language processing [62, 42].

Heterogeneous data and massive client populations are two of the defining characteristics of cross-device federated learning systems [31, 20]. Despite intensive efforts [31, 28, 67, 44, 20], several key challenges that arise from the involvement of large-scale client populations are often overlooked in the existing literature [41]. One of the primary hurdles is the issue of client unavailability. Intuitively, more active clients drive the global model to their local optima by overfitting their local data, which biases the training. In addition, the higher the uncertainty in client unavailability, the larger the performance degradation. Concrete examples that confirm these intuitions in the context of `FedAvg` - the most widely adopted federated learning algorithm - can be found in Section 4. Client unavailability issues can arise from internal factors such as different working schedules and heterogeneous hardware/software constraints. External factors, such as poor network coverage and frequent handovers of base stations due to fast movements, only exacerbate these problems [49, 56, 63, 3, 20]. The intricate interplay of internal and external factors results in the *non-stationarity* and *heterogeneity* of client unavailability.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Most prior work either assumes exact knowledge of the clients' available dynamics or requires their dynamics to be benignly stationary [31, 26, 41, 54, 53]. A related line of work studies asynchronous federated learning wherein clients are vulnerable to delays in message transmission and the reported model updates may be stale [58, 37, 48, 24]. The proposed methods therein assume the availability of all clients or uniformly sampled clients, making them inapplicable to our settings. A few recent works [43, 57] study non-stationary dynamics. Ribero et al. [43] consider the settings where the available probabilities follow a homogeneous Markov chain. Xiang et al. [57] require that clients be capable of continuous local optimization regardless of communication failures. A handful of other works [13, 59] memorize the old gradients of the unavailable clients to compensate for their unavailability. However, the added memory burdens the federated learning system with substantial memory proportional to the product of the number of clients and the model dimension.

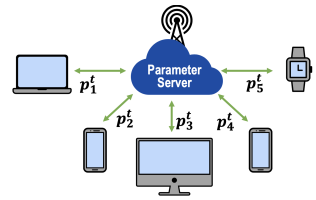


Figure 1: Client i's available probabilities p_i^t's are heterogeneous and are subject to *non-stationary* dynamics.

Contributions. In this work, we focus on stochastic client unavailability, where client i is available for federated learning model training with probability p_i^t at any time t. An illustration can be found in Fig. 1. Our contributions are four-fold:

- In Section 4, via constructing concrete examples, we demonstrate that both heterogeneity and non-stationarity of p_i^t will result in bias and thus significant performance degradation of `FedAvg`.
- In Section 5, we propose an algorithm named `FedAWE`, which features computational and memory efficiency: only $O(1)$ additional computation and memory per client will be used when compared with `FedAvg`. The design of `FedAWE` introduces two novel algorithmic structures: *adaptive innovation echoing* and *implicit gossiping*. At a high level, these novel algorithmic structures (i) help clients catch up on the missed computation, and (ii) simultaneously enable a balanced information mixture through implicit client-client gossip, which ultimately corrects the remaining bias. Notably, no direct neighbor information exchanges are used, and the client unavailability dynamics remains unknown to all clients and the parameter server.
- In Section 6, we show that `FedAWE` converges to a stationary point of even non-convex global objective and achieves the linear speedup property without conditions on second-order partial derivatives of the loss function in analysis.
- In Section 7, we validate our analysis with numerical experiments over diversified client unavailability dynamics on real-world data sets.

2 Related Work

Dynamical client availability. There is a recent surge of efforts to study time-varying client availability [44, 43, 7, 53, 43, 41, 57], which can be roughly classified into two categories depending on whether the parameter server can unilaterally determine the participating clients.

(i) *Controllable participation.* Earlier research [31, 28] presumes that, in each round, the parameter server could select a small set of clients either uniformly at random or in proportion to the volume of local data held by clients. More recently, Cho et al. [10] design adaptive and non-uniform client sampling to accelerate learning convergence, albeit at the cost of introducing a non-zero residual error. In another work, Cho et al. [8] study the convergence of `FedAvg` with cyclic client participation. Yet, the set of available clients is sampled uniformly at random per cyclic round and is decided unilaterally by the parameter server. Perazzone et al. [41] consider heterogeneous and time-varying response rates p_i^t under the assumptions that p_i^t is known a priori and that the stochastic gradients are bounded in expectation. Furthermore, the dynamics of p_i^t are determined by the parameter server by solving a stochastic optimization problem. Chen et al. [7] propose a client sampling scheme wherein only the clients with the most "important" updates communicate back to the parameter server. This sampling method can achieve performance comparable to that of full client participation, provided that p_i^t is globally known to both the parameter server and the clients. Departing from this line of literature, our setup neither assumes any side information or prior knowledge of the response rates p_i^t nor assumes that the parameter server has any influence on p_i^t.

(ii) *Uncontrollable participation.* There is a handful of work on building resilience against arbitrary client availability [43, 53, 59, 13, 61, 54]. Ribero et al. [43] consider random client availability whose underlying response rates are also heterogeneous and time-varying with unknown dynamics. However, the underlying dynamics of p_i^t in [43] are assumed to follow a homogeneous Markov chain. Wang et al. [53] propose a generalized `FedAvg` that amplifies parameter updates every P rounds for some carefully tuned P. Despite its elegant unified analysis and potential to accommodate non-independent unavailability dynamics, to reach a stationary point, p_i^t needs to satisfy some assumptions to ensure roughly equal availability of all clients over every P rounds. Yang et al. [61] analyze a setting where clients participate in the training at their will. Yet, their convergence is shown to be up to a non-zero residual error. The algorithms proposed in [13, 59] share the same idea of using the memorized latest updates from unavailable clients for global aggregation. Despite superior numerical performance, both algorithms demand a substantial amount of additional memory [54]. For non-convex objectives, both [59] and [13] require an absolute bounded inactive period, and share similar technical assumptions such as almost surely bounded stochastic gradients [59] or Lipschitz Hessian [13]. Though bounded inactive periods are relevant for applications wherein the sensors wake up on a periodic schedule, this assumption is not satisfied even for the simple stochastic setting when clients are selected uniformly at random. Wang and Ji consider unknown heterogeneous p_i' in a concurrent work [54]; however, p_i's are assumed to be fixed over time.

Asynchronous federated learning. Another related line of work is asynchronous federated learning. To the best of our knowledge, Xie et al. [58] initialize the study of asynchronous federated learning, wherein the parameter server revises the global model every time it receives an update from a client. Convergence is shown under some technical assumptions such as weakly-convex global objectives, bounded delay, and bounded stochastic gradients. Zakerinia et al. [68] propose QuAFL which is shown to be resilient to computation asynchronicity and quantized communication yet under the bounded and stationary delay assumption. Nguyen et al. [37] propose FedBuff, which uses additional memory to buffer asynchronous aggregation to achieve scalability and privacy. Convergence is shown under bounded gradients and bounded staleness assumptions. In fact, most convergence guarantees in the asynchronous federated learning literature rely on bounded staleness [58, 37, 48, 24], or bounded gradients [58, 37, 24]. Recently, arbitrary delay is considered in the context of distributed SGD with bounded stochastic gradients and $(0, \zeta)$-bounded inter-client heterogeneity [32] (see Assumption 4 for the definition). The convergence suffers from a non-zero residual term $O(\zeta^2)$. In contrast, our convergence guarantee is free from non-zero residual terms and does not require gradients to be bounded.

3 Problem Formulation

A federated learning system consists of a parameter server and m clients that collaboratively minimize

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} F(\boldsymbol{x}) \triangleq \frac{1}{m} \sum_{i=1}^{m} F_i(\boldsymbol{x}), \tag{1}$$

where $F_i(\boldsymbol{x}) \triangleq \mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[\ell_i(\boldsymbol{x}; \xi_i)\right]$ is the local objective and can be non-convex, $\mathcal{D}_i$ is the local distribution, ξ_i is a stochastic sample that client i has access to, ℓ_i is the local loss function, and d is the model dimension.

We use Assumption 1 to capture the uncertain *non-stationary* dynamics and heterogeneity. Let $\mathcal{A}^t$ denote the set of active clients, $\mathbb{1}_{\{\cdot\}}$ an indicator function, T the number of total training rounds.

Assumption 1. There exists a $\delta \in (0, 1]$ such that $p_i^t \triangleq \mathbb{E}[\mathbb{1}_{\{i \in \mathcal{A}^t\}}] \geq \delta$, where the events $\{i \in \mathcal{A}^t\}$ are independent across clients i and across rounds $t \in [T]$.

Assumption 1 subsumes uniform availability [26, 61] and stationary availability considered in [54]. Independent client unavailability is widely adopted by federated learning research [26, 28, 22, 60, 61, 54]. Analyzing non-independent unavailability, together with uncertain and non-stationary dynamics in Assumption 1, is in general challenging. Specifically, the involved entanglement of stochastic gradient and availability statistics fundamentally complicates the theoretical analysis. However, we conjecture that independence and strictly positive probabilities are only necessary for the technical convenience of our analysis. Our experiments in Section 7 suggest that our algorithm offers notable improvement even in the presence of non-independent and occasionally zero-valued probabilities. Future work will investigate how to provably accommodate correlated or zero-valued probabilities of arbitrary probabilistic trajectories.

4 Heterogeneity and Non-stationarity May Lead to Significant Bias

In this section, we illustrate the impacts of heterogeneity and non-stationarity of client availability under the classic FedAvg. We use two examples to showcase the significant bias incurred.

Example 1 (Heterogeneity)**.** Suppose that $m = 2$ and $p_i^t = p_i$ for $i \in [2]$. Let $F_i(x) \triangleq \|x - u_i\|_2^2/2$, where $x, u_i \in \mathbb{R}$. The global objective (1) is

$$F(x) = \frac{1}{2}(\|x - u_1\|_2^2 + \|x - u_2\|_2^2), \tag{2}$$

with unique minimizer $x^\star = (u_1 + u_2)/2$. Let $u_1 = 0$ and $u_2 = 100$. Fig. 2 illustrates how the heterogeneity in p_i affects the expected output of FedAvg.

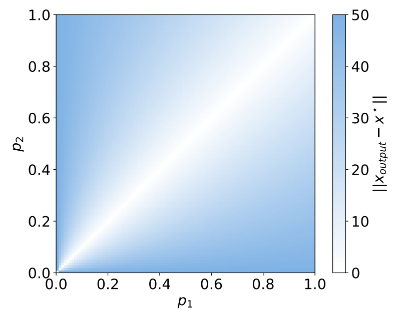


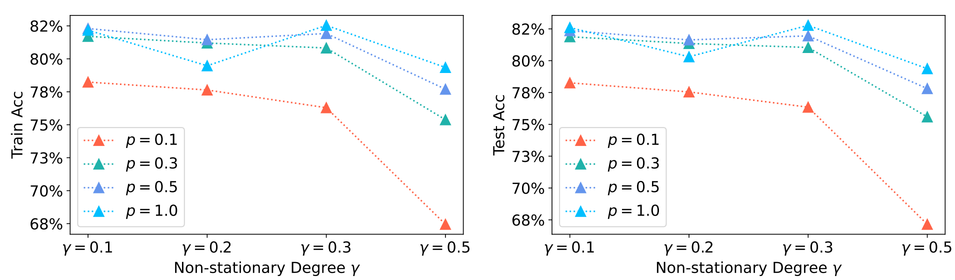
Figure 2: Let $x_{\text{output}} \triangleq \lim_{t\to\infty} \mathbb{E}[x^t]$. Under most of the choices of p_1, p_2, x_{output} is far from x^*.

Example 1 matches [54, Theorem 1], which shows that FedAvg leads to a biased global objective (3) under heterogeneous p_i's, and that (3) may be significantly away from (1) depending on p_i's.

$$\widetilde{F}(\boldsymbol{x}) \triangleq \sum_{i=1}^{m} \frac{p_i}{\sum_{j=1}^{m} p_j} F_i(\boldsymbol{x}). \tag{3}$$

When the probabilistic dynamics of p_i^t's is non-stationary, obtaining an exact biased objective similar to (3) in a neat analytical form becomes challenging, if not impossible, due to the unstructured non-stationary dynamics. Fortunately, Example 2 helps us confirm that the complex interplay between p_i^t's across rounds and clients will inevitably further degrade the performance of FedAvg algorithm.

Example 2 (Non-stationarity)**.** In Fig. 3, a total of $m = 100$ clients perform an image classification task on the SVHN dataset [36] under the FedAvg algorithm, whose local dataset distribution follows $\mathsf{Dirichlet}(0.1)$ [16]. Clients become available with probability $p_i^t = p \cdot [\gamma \cdot \sin(0.1\pi \cdot t) + (1 - \gamma)]$, $\forall i \in [m]$. The hyperparameter details are deferred to Appendix J. Observations can be found in the caption.



(a) Train accuracy. (b) Test accuracy

Figure 3: Train and test accuracy results in percentage (%). In particular, the parameter γ signifies the degree of non-stationary. Notice that, as the client availability becomes more non-stationary (a larger γ), FedAvg experiences a significant drop in accuracy. For example, both the train and test accuracies drop by over 10% when $p = 0.1$, and γ increases from 0.1 to 0.5.

5 Federated Agile Weight Re-Equalization (FedAWE)

To minimize (1), one natural idea is to have the entire client population performs the same number of local updates and mixes these updates carefully to ensure they are weighted equally. Unfortunately, when clients are available only intermittently, they will miss some rounds. A naive approach to equalizing the number of local updates is to have clients catch up by performing their missed local computations immediately when they become available. However, this approach requires a daunting amount of resources and may not be possible due to hardware/software constraints. Formally, recall that $\mathcal{A}^t$ is the set of available clients at time t. Let $\tau_i(t) \triangleq \{t' : t' < t \text{ and } i \in \mathcal{A}^{t'}\}$ denote the most recent (with respect to time t) round that client i is available. Compared with standard FedAvg, the

Algorithm 1: `FedAWE`

1 **Inputs:** T, s, η_l, η_g, $\boldsymbol{x}^0$.
2 **for** $i \in [m]$ **do** $\boldsymbol{x}_i^0 \leftarrow \boldsymbol{x}^0$ and $\tau_i(0) \leftarrow -1$;
3 **for** $t = 0, \cdots, T-1$ **do**
4   **for** $i \in \mathcal{A}^t$ **do**
5         $\boldsymbol{x}_i^{(t,0)} \leftarrow \boldsymbol{x}_i^t$;
6   **for** $k = 0, \cdots, s-1$ **do**
7             $\boldsymbol{x}_i^{(t,k+1)} \leftarrow$
8                 $\boldsymbol{x}_i^{(t,k)} - \eta_l \nabla \ell_i(\boldsymbol{x}_i^{(t,k)}; \xi_i^{(t,k)})$;
9   **end**
10         $\boldsymbol{G}_i^t \leftarrow \boldsymbol{x}_i^t - \boldsymbol{x}_i^{(t,s)}$;
11         $\boldsymbol{x}_i^{t\dagger} \leftarrow \boldsymbol{x}_i^{(t,0)} - \eta_g (t - \tau_i(t)) \boldsymbol{G}_i^t$;
12         $\tau_i(t+1) \leftarrow t$;
13   Report $\boldsymbol{x}_i^{t\dagger}$ to the parameter server;
14   **end**
15     $\boldsymbol{x}^{t+1} \leftarrow \frac{1}{|\mathcal{A}^t|} \sum_{i \in \mathcal{A}^t} \boldsymbol{x}_i^{t\dagger}$;
16   **for** $i \in [m]$ **do**
17   **if** $i \in \mathcal{A}^t$ **then**
18             $\boldsymbol{x}_i^{t+1} \leftarrow \boldsymbol{x}^{t+1}$;
19   **else if** $i \notin \mathcal{A}^t$ **then**
20             $\boldsymbol{x}_i^{t+1} \leftarrow \boldsymbol{x}_i^t$;
21             $\tau_i(t+1) \leftarrow \tau_i(t)$;
22   **end**
23 **end**

naive "catch-up" procedure will consume $(t - \tau_i(t) - 1) \cdot s$ local stochastic gradient descent updates and $(t - \tau_i(t) - 1)$ additional stochastic samples, where s is the number of local updates per round when a client is available in standard `FedAvg`.

In this work, we target computation-light algorithms that, compared with `FedAvg`, only take $O(1)$ additional computation without additional stochastic samples. We propose **Fed**erated **A**gile **W**eight Re-**E**qualization (`FedAWE`), which is formally described in Algorithm 1. It involves two novel algorithmic structures: *adaptive innovation echoing* and *implicit gossiping*. At a high level, these novel algorithmic structures (i) help clients catch up on the missed computation, and (ii) simultaneously enable a balanced information mixture through implicit client-client gossip, which ultimately corrects the remaining bias.

In Algorithm 1, each client keeps two local variables $\boldsymbol{x}_i$ and τ_i, along with a few auxiliary variables used in updating $\boldsymbol{x}_i$ and τ_i. The algorithm inputs are rather standard: total training rounds T, local and global learning rates η_l and η_g, the number of local updates per round s, and the initial model $\boldsymbol{x}^0$. In each round t, similar to `FedAvg`, an available client $i \in \mathcal{A}^t$ performs s steps of stochastic gradient descent on its local model $\boldsymbol{x}_i^t$ (lines 5-8), where $\nabla \ell_i(\cdot; \xi_i^{(t,k)})$ is the stochastic gradient of sample $\xi_i^{(t,k)}$. Next, we describe the two novel algorithmic structures used in `FedAWE`.

Adaptive innovation echoing. Departing from `FedAvg` wherein the local estimate $\boldsymbol{x}_i^t$ is updated as $\boldsymbol{x}_i^{t\dagger} \leftarrow \boldsymbol{x}_i^{(t,0)} - \eta_g \boldsymbol{G}_i^t$. In `FedAWE` (lines 10-11), we "echo" the local innovation $\boldsymbol{G}_i^t$ by multiplying it by $(t - \tau_i(t))$. Intuitively, this simple echoing helps us approximately equalize the number of local improvements, as formally stated in Proposition 1. It says that the total numbers of innovations echoing are the same for all active clients for any given round and allows the unavailable clients to catch up to the missed computations when they become available.

Proposition 1. *If* $\mathbb{1}_{\{i \in \mathcal{A}^{R-1}\}} = 1$, *it holds that* $\sum_{t=0}^{R-1} \mathbb{1}_{\{i \in \mathcal{A}^t\}} (t - \tau_i(t)) = R, \ \forall \ R \geq 1$.

Implicit gossiping. In `FedAWE`, the parameter server does not send the most recent global model to the active clients at the beginning of a round. Instead, the parameter server aggregates the locally updated models $\boldsymbol{x}_i^{t\dagger}$ and sends the new global model $\boldsymbol{x}^{t+1}$ to all active clients $\mathcal{A}^t$ (lines 14-15). By postponing multicasting the shared global model, the active clients in $\mathcal{A}^t$ *implicitly gossip* their updated local models with each other through the parameter server [57]. Though the postponed multi-cast brings in staleness, simple coupling argument show that the staleness is bounded (Lemma 2). In addition, our empirical results (Table 8 in Appendix J) suggest that there is no significant slowdown when compared to vanilla `FedAvg`. Gossip-type algorithms were originally proposed for peer-to-peer networks and are well-known for their agility to communication failures and asynchronous information exchange in achieving average consensus

$$W_{ij}^{(t)} \triangleq \begin{cases} \frac{1}{|\mathcal{A}^t|}, \text{ if } i, j \in \mathcal{A}^t; \\ 1, \text{ if } i = j \text{ and } i \notin \mathcal{A}^t; \\ 0, \text{ otherwise.} \end{cases} \tag{4}$$

[12, 4, 23, 15, 30, 35]. Intuitively, the clients' local estimates are eventually equally weighted in the final algorithm output. Note that, departing from the standard gossiping protocols therein [23, 45], information exchange in `FedAWE` does not involve client-client communication. The information mixing matrix under `FedAWE` is defined in (4), which is doubly stochastic. Let $M^{(t)} \triangleq \mathbb{E}[(W^{(t)})^2]$, $\rho(t) \triangleq \lambda_2(M^{(t)})$, $\mathbf{J} = \mathbb{1}\mathbb{1}^\top/m$, and $\rho \triangleq \max_t \rho(t)$, where $\lambda_2(\cdot)$ denotes the second largest eigenvalue. We next characterize the information mixing error, i.e., consensus error in Lemma 1.

Lemma 1 ([34, 33, 50]). *For any matrix $B \in \mathbb{R}^{d\times m}$, it holds that $\mathbb{E}_W[\|B\left(\prod_{r=1}^{t} W^{(r)} - \mathbf{J}\right)\|_{\mathrm{F}}^2] \le \rho^t \|B\|_{\mathrm{F}}^2$, where the expectation is taken with respect to randomness in W matrices.*

6 Convergence Analysis

In this section, we analyze the convergence of `FedAWE`. All missing proofs and intermediate results are deferred to the Appendix. Details can be found in Table of Contents.

6.1 Assumptions

We start by stating regulatory assumptions that are common in federated learning analysis [26, 51, 22].

Assumption 2. Each local objective function $\nabla F_i(\boldsymbol{x})$ is L-Lipschitz, i.e.,

$$\|\nabla F_i(\boldsymbol{x}_1) - \nabla F_i(\boldsymbol{x}_2)\|_2 \le L\,\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2\,,\ \forall \boldsymbol{x}_1,\ \boldsymbol{x}_2, \text{ and } \forall\, i \in [m].$$

Assumption 3. Stochastic gradients $\nabla \ell_i(\boldsymbol{x};\xi)$ are unbiased with bounded variance, i.e.,

$$\mathbb{E}\left[\nabla \ell_i(\boldsymbol{x};\xi) \mid \boldsymbol{x}\right] = \nabla F_i(\boldsymbol{x}) \text{ and } \mathbb{E}\left[\|\nabla \ell_i(\boldsymbol{x};\xi) - \nabla F_i(\boldsymbol{x})\|_2^2 \mid \boldsymbol{x}\right] \le \sigma^2,\ \forall\, i \in [m].$$

Assumption 4. The divergence between local and global gradients is bounded for $\beta,\ \zeta \ge 0$ such that

$$\frac{1}{m}\sum_{i=1}^{m} \|\nabla F_i(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\|_2^2 \le \beta^2 \|\nabla F(\boldsymbol{x})\|_2^2 + \zeta^2. \tag{5}$$

When the local data sets are homogeneous, $\nabla F_i(\boldsymbol{x}) = \nabla F(\boldsymbol{x})$ holds for any client $i \in [m]$, resulting in $\beta = \zeta = 0$. Assumption 4 and its variants in Table 1 are often referred to as bounded gradient dissimilarity assumption to account for data heterogeneity across clients. It can be easily checked that our Assumption 4 is more relaxed or equivalent to the variants therein.

Table 1: Popular variant assumptions on gradient dissimilarity.

Bounded Gradient Dissimilarity	**References**
$\max_{\boldsymbol{x}} \|\nabla F_i(\boldsymbol{x})\|_2^2 \le \zeta^2,\ \forall\, i \in [m]$	[28, 65, 9, 10, 59]
$\frac{1}{m}\sum_{i=1}^{m} \|\nabla F_i(\boldsymbol{x})\|_2^2 \le \beta^2 \|\nabla F(\boldsymbol{x})\|_2^2$	[26, 27]
$\frac{1}{m}\sum_{i=1}^{m} \|\nabla F_i(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\|_2^2 \le \zeta^2$	[52, 64, 17, 55, 1, 21, 53, 61]
$\frac{1}{m}\sum_{i=1}^{m} \|\nabla F_i(\boldsymbol{x})\|_2^2 \le \beta^2 \|\nabla F(\boldsymbol{x})\|_2^2 + \zeta^2$	[22, 67, 51, 50, 13]

6.2 Auxiliary/Imaginary update sequence construction.

Directly analyzing the evolution of $\boldsymbol{x}^t$ and $\boldsymbol{x}_i^t$ is challenging due to the fact that different clients update at different rounds, and that different active clients echo their local innovation $\boldsymbol{G}_i^t$ (line 9 in Algorithm 1) with different strength $(t - \tau_i)$. As such, we construct an auxiliary/imaginary update sequence $\boldsymbol{z}_i^t$ for client $i \in [m]$, whose evolution is closely coupled with $\boldsymbol{x}^t$ and $\boldsymbol{x}_i^t$ but is easier to analyze. Note that the auxiliary/imaginary update sequence is never actually computed by clients but acts as a necessary tool in building up the analysis.

Definition 1. *The auxiliary sequence $\{\boldsymbol{z}_i^t\}$ of client $i \in [m]$ is defined as*

$$\boldsymbol{z}_i^t \ \triangleq\ \boldsymbol{x}_i^t - \eta_l\eta_g s(t - \tau_i(t) - 1)\nabla F_i(\boldsymbol{x}_i^{\tau_i(t)+1}),\ \forall\, i \in [m]. \tag{6}$$

Recall that $\tau_i(0) = -1$. Thus, by definition, $\boldsymbol{z}_i^0 = \boldsymbol{x}_i^0$ according to (6). For general t, when client $i \in \mathcal{A}^{t-1}$, we simply have $\tau_i(t) = t-1$ and thus $t-1-\tau_i(t) = t-1-(t-1) = 0$. That is, the auxiliary model $\boldsymbol{z}_i^t$ and the real model $\boldsymbol{x}_i^t$ are *identical* whenever the client i becomes available in the previous round.

- When $i \in \mathcal{A}^{t-1}$, the iterate of $\boldsymbol{z}_i$ is a bit more involved:

$$\boldsymbol{z}_i^t \overset{(7.a)}{=} \boldsymbol{x}_i^t \overset{(7.b)}{=} \frac{\sum_{j\in\mathcal{A}^{t-1}}}{|\mathcal{A}^{t-1}|}\left(\boldsymbol{z}_j^{t-1} + \underbrace{(\boldsymbol{x}_j^{t-1} - \boldsymbol{z}_j^{t-1})}_{(7.c)} - \eta_l\eta_g(t-1-\tau_j(t-1))\boldsymbol{G}_j^{t-1}\right), \quad (7)$$

  where $(7.a)$ holds because of Definition 1 and $i \in \mathcal{A}^{t-1}$, $(7.b)$ because of line 10 in Algorithm 1, addition and subtraction. $(7.c)$ can be expanded by (6). We defer the simplified form of (7) to (18) in Appendix C for a tidy presentation.
- When $i \notin \mathcal{A}^{t-1}$, $\boldsymbol{z}_i^t$ has a simple iterative relation:

$$\boldsymbol{z}_i^t \ = \ \boldsymbol{z}_i^{t-1} - \eta_l\eta_g s\nabla F_i(\boldsymbol{x}_i^{\tau_i(t-1)+1}). \quad (8)$$

  At a high level, the sequence $\boldsymbol{z}_i^t$ approximately mimics the ideal descent evolution at a client as if the client performs local optimizations on its local model $\boldsymbol{x}_i$ per round regardless of its availability. Mathematically, the idea is that, if the progress per iteration of the auxiliary sequence $\boldsymbol{z}_i^t$ is bounded, we can show the convergence of $\boldsymbol{x}_i^t$ when $\boldsymbol{x}_i^t$ and $\boldsymbol{z}_i^t$ are close to each other.

It is worth noting that auxiliary sequences are used in peer-to-peer distributed learning literature [46, 2, 29, 66, 47, 33]. Yet, existing constructions are not applicable to our problem due to (1) the non-convexity of the global objectives, (2) multiple local updates per round, (3) possibly unbounded gradients, and (4) the general form of bounded gradient dissimilarity. Departing from the use of staled stochastic gradients for auxiliary updates therein, we adopt the true gradient $\nabla F_i(\cdot)$ to avoid the complications from the involved interplay between randomness in stochastic samples and randomness in $\tau_i(t)$. On the technical front, it follows from Definition 1 that $\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\|_2^2 \le \eta_l^2\eta_g^2 s^2(t-\tau_i(t)-1)^2\|\nabla F_i(\boldsymbol{x}_i^{\tau_i(t)+1})\|_2^2$, whose bound appears to be quite challenging to derive due to the coupling of different realizations of $\tau_i(t)$ and gradients. As such, we bound the average of $\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\|^2$ across clients and rounds in Proposition 2.

Lemma 2 (Unavailability statistics)**.** *Under Assumption 1 and δ defined therein. It holds for $t \ge 0$ that $\mathbb{E}\left[t-\tau_i(t)\right] \le 1/\delta$ and $\mathbb{E}\left[(t-\tau_i(t))^2\right] \le 2/\delta^2$.*

Lemma 2 yields an upper bound on the first and second moments of a client i's unavailable duration despite the unstructured nature of clients' non-stationary and heterogeneous unavailability. In the special case where we have clients available with the same probability δ, the duration simply follows a homogeneous geometric distribution. It can be easily checked that our bounds trivially hold. However, the duration becomes a more challenging *non-homogeneous* geometric random variable under our non-stationary unavailability dynamics. Lemma 2 can be derived by using a simple coupling argument and by using tools from probability theory [14].

6.3 Main results.

Let $\bar{\boldsymbol{z}}_t \triangleq \frac{1}{m}\sum_{i=1}^m \boldsymbol{z}_i^t$, $F^\star \triangleq \min_{\boldsymbol{x}} F(\boldsymbol{x})$, and $\delta_{\max} \triangleq \max_{i\in[m],t\in[T]} p_i^t$.

Lemma 3 (Descent Lemma)**.** *Let $\mathcal{F}^t$ define the sigma algebra generated by randomness up to round t. Suppose Assumptions 2, 3 hold and $\eta_l\eta_g \le 9/(100sL)$, it holds that*

$$\begin{aligned}\mathbb{E}\left[F(\bar{\boldsymbol{z}}^{t+1}) - F(\bar{\boldsymbol{z}}^t) \mid \mathcal{F}^t\right] \le -\frac{\eta_l\eta_g s}{4}\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 \\ + \frac{2\eta_l\eta_g sL\sigma^2\left(\eta_l\eta_g\delta_{\max} + 4.5m\eta_l^2 sL\right)}{m^2}\sum_{i=1}^m (t-\tau_i(t))^2 \\ + \frac{35\eta_g\eta_l^3 s^3L^2}{m}\sum_{i=1}^m (t-\tau_i(t))^2\left\|\nabla F_i(\boldsymbol{x}_i^{\tau_i(t)+1})\right\|_2^2 \\ + \frac{2.2\eta_l\eta_g sL^2}{m}\sum_{i=1}^m \underbrace{\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2}_{\text{Approximation Error}} + \frac{\eta_l\eta_g sL^2}{2m}\sum_{i=1}^m \underbrace{\left\|\boldsymbol{z}_i^t - \bar{\boldsymbol{z}}^t\right\|_2^2}_{\text{Consensus Error}}.\end{aligned}$$

The proof of Lemma 3 follows from the standard analysis for non-convex smooth objectives but with non-trivial adaptation to account for *adaptive innovation echoing* and *implicit gossiping*. In particular, it highlights two terms unique in our derivation: the approximation error from the auxiliary sequence and the consensus error from the implicit gossiping procedure.

Proposition 2 (Approximation error)**.** *Given Assumptions 2 and 4, it holds that*

$$\frac{1}{mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2\right]\leq\frac{6\eta_l^2\eta_g^2s^2}{\delta^2}\left(\beta^2+1\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]+\frac{6\eta_l^2\eta_g^2s^2}{\delta^2}\zeta^2$$
$$+\frac{6L^2\eta_l^2\eta_g^2s^2}{\delta^2}\frac{1}{m}\sum_{i=1}^{m}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right]. \tag{9}$$

The proof of Proposition 2 starts from Definition 1. Although in general it is difficult to bound the error, Assumptions 2 and 4 allow us to break down the problem into bounding the averaged gradient norm of $\bar{\boldsymbol{z}}^t$ and the consensus error over all randomness instead. Next, we analyze the consensus error. Note that although implicit gossiping takes place in Algorithm 1 for $\boldsymbol{x}_i^t$, its analysis is technically challenging as discussed before. So, we adopt the auxiliary $\boldsymbol{z}_i^t$ as an intermediary and apply Young's inequality to bound the actual consensus error. Details will be specified next. Formally, the auxiliary models can be expressed in a compact matrix form as $\boldsymbol{Z}^{(t)}\triangleq[\boldsymbol{z}_1^t,\ldots,\boldsymbol{z}_m^t]$. Their local parameter innovation matrix $\widetilde{\boldsymbol{G}}^t$ is formulated by combing (7) and (8). We refer the interested readers to (19) in Appendix C for the exact formula. Unrolling the recursion, the consensus error can be expanded as

$$\frac{1}{m}\left\|\left(\boldsymbol{Z}^{(t-1)}-\eta_l\eta_g\widetilde{\boldsymbol{G}}^{(t-1)}\right)W^{(t-1)}\left(\mathbf{I}-\mathbf{J}\right)\right\|_{\mathrm{F}}^2\overset{(10.a)}{=}\frac{\eta_l^2\eta_g^2}{m}\left\|\sum_{q=0}^{t-1}\widetilde{\boldsymbol{G}}^{(q)}\left(\prod_{l=q}^{t-1}W^{(q)}-\mathbf{J}\right)\right\|_{\mathrm{F}}^2, \tag{10}$$

where equality $(10.a)$ holds because all clients are initiated at the same weight.

Lemma 4 ([57])**.** *Under Assumption 1, it holds that* $\rho\leq1-\frac{\delta^4(1-(1-\delta)^m)^2}{8}$.

Recall that ρ bounds the expected spectral norm of the information mixing matrix $W^{(t)}$. It is important to have $\rho<1$ for an exponential decay of the consensus error (see Lemma 1). We now proceed to present the convergence rates. In the sequel, we assume it holds for η_g and η_l that

$$\eta_l\eta_g\leq\frac{\left(1-\sqrt{\rho}\right)\delta}{80s(L+1)\left(\sqrt{\rho}+1\right)\sqrt{\left(\beta^2+1\right)\left(1+L^2\right)}};\ \eta_l\leq\frac{\delta}{200sL\sqrt{\left(\beta^2+1\right)\left(1+L^2\right)}}. \tag{11}$$

The proof of the consensus error borrows insights from the analysis of the gossip algorithm [34, 52] but with substantial adaptation to accommodate the novel auxiliary formulation and multi-step local updates. Under the learning rate conidtions in (11) and Assumptions 1, 2, 3 and 4, we can show that

$$\frac{1}{mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2\right]\asymp\frac{1}{mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right]\asymp\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]. \tag{12}$$

It remains to bound the full convergence error of $\boldsymbol{z}_i^t$, which is presented in Theorem 1.

Theorem 1 (Convergence error of $\boldsymbol{z}_i^t$)**.** *Suppose that Assumptions 1, 2, 3 and 4 hold. Choose learning rates η_l and η_g such that the conditions in* (11) *are met for $T\geq1$, it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\lesssim\frac{\left(F(\bar{\boldsymbol{z}}^0)-F^\star\right)}{\eta_l\eta_g sT}+\frac{\eta_l\eta_g L\sigma^2}{m}\frac{\delta_{\max}}{\delta^2}+\eta_l^2\eta_g^2s^2L^2\left(\frac{\sigma^2+\zeta^2}{\delta^2(1-\sqrt{\rho})^2}\right). \tag{13}$$

By addition, subtraction, and Young's inequality, (14) and (15) hold under Assumption 2.

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\bar{\boldsymbol{x}}^t\right\|_2^2\right]\asymp\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2\right]+\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right]; \tag{14}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right]\asymp\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2\right]+\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]. \tag{15}$$

Moreover, from (12), (14) and (15), it can be seen that (16) holds.

$$\frac{1}{mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\bar{\boldsymbol{x}}^t\right\|_2^2\right]\asymp\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\asymp\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right]. \tag{16}$$

Combining (12), (13), (14) and (15), we are ready for Corollary 1.

Corollary 1 (Convergence rate of $\boldsymbol{x}_i^t$)**.** *Suppose that Assumptions 1, 2, 3 and 4 hold. Choose learning rates as* $\eta_l=\frac{1}{\sqrt{T}sL}$*,* $\eta_g=\sqrt{s\delta m}$ *such that the conditions in* (11) *are met for* $T\geq 1$*, it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right]\lesssim\frac{L\left(F(\bar{\boldsymbol{x}}^0)-F^\star\right)}{\sqrt{s\delta mT}}+\frac{\delta_{\max}}{\delta^{\frac{3}{2}}\sqrt{smT}}\sigma^2+\frac{sm}{T}\left(\frac{\sigma^2+\zeta^2}{\delta(1-\sqrt{\rho})^2}\right). \tag{17}$$

Corollary 1 establishes the full convergence rate for `FedAWE` algorithm. It can be seen that the first and second terms dominate when T is sufficiently large, which relate to initial suboptimality gap and stochastic gradient noise σ^2, respectively. The non-stationary client unavailability results in the third term, which relates to gradient divergence ζ^2 and also to σ^2. The proof of Corollary 1 follows from (15) by plugging in Proposition 2 and Theorem 1. In the special case where k clients participate uniformly at random, we simply have $\delta_{\max}=\delta=k/m$. Our convergence bound attains the rate of $O(1/\sqrt{skT})$. In other words, we achieve the desired linear speedup property with respect to the number of local steps s and the number of active clients k, matching the established literature [60, 53, 64, 65]. The linear speedup property enables a large cross-device federated learning system to take advantage of a massive scale of parallelism. Notice that the consensus error (16) and the convergence rate (17) have the same asymptotic order with respect to the parameters therein. Hence, the consensus error also enjoys the desired linear speedup property when T is sufficiently large.

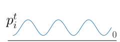
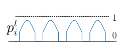
7 Numerical Experiments

Overview. In this section, we evaluate `FedAWE` on real-world data sets to corroborate our analysis and compare it with the other state-of-the-art algorithms. The missing specifications and additional results can be found in Appendix J. Specifically, we consider a federated learning system of one parameter server and $m=100$ clients, wherein clients become available intermittently. The image classification tasks use CNNs and are based on SVHN [36], CIFAR-10 [25] and CINIC-10 [11] data sets. All of them include 10 classes of images of different categories. To emulate a highly heterogeneous local data distribution, the image class distribution $\nu_i\sim\mathsf{Dirichlet}(\alpha=0.1)$ at client i [16, 53, 54].

Non-stationary client unavailability. A total of four unavailable dynamics are evaluated in Table 2, including stationary and *non*-stationary with staircase, sine and interleaved sine trajectories, with their visualizations available in the same table. The classification tasks become more challenging as the list progresses due to the growing complexity in the non-stationary dynamics. Furthermore, our choices of the non-stationary dynamics are motivated by real-world federated learning participation statistics, for example, sine trajectory [3], and by generalizing the existing participation patterns such as cyclic participation [8, 54]. In particular, the interleaved sine dynamics is more challenging than the vanilla cyclic availability dynamics since clients become available during each active period with probability that is less than 1 and non-stationary simultaneously. Formally, client i's dynamics is defined as $p_i^t=p_i\cdot f_i(t)$, where $f_i(t)$ is a time-dependent function under non-stationary dynamics but $f_i(t)=1$ when stationary, and $p_i=\langle\nu_i,\phi\rangle$. ϕ characterizes the unbalanced contribution of different image classes to the generated probabilities. Each element of $[\phi]_c$ is drawn from $\mathsf{Uniform}(0,\boldsymbol{\Phi}_c)$, where a smaller $\boldsymbol{\Phi}_c$ leads to a less significant contribution of that image class.

Correlating the local data distribution and the probability of client availability is a common practice in the prior literature. For example, Gu et al. in [13] experiment with a formula for p_i so that clients that hold images of smaller digits participate less frequently. Wang and Ji in [54] construct p_i as an inner product of the clients' local data distribution ν_i and an external distribution $\boldsymbol{\Phi}'$. It is immediately clear that the coupling of local data distribution ($\nu_i\sim\mathsf{Dirichlet}(\alpha=0.1)$) and class contribution ϕ leads to *non-independent* p_i's. In addition, Assumption 1 will not hold in the case of interleaved sine non-stationary dynamics since p_i^t's occasionally reach 0. Although being agnostic to the challenging client unavailability dynamics not covered by our analysis, we observe that `FedAWE` retains its outperformance. Comparisons will be specified next.

Table 2: Results and comparisons on real-world datasets in the form of mean accuracy ± standard deviation and are obtained over 3 repetitions in different random seeds. Results are averaged over the last 50 rounds. The total number of global rounds is 2000 for SVHN, CIFAR-10 and CINIC-10. Algorithms are categorized into two groups: (1) ones **not** aided by memory or known statistics; (2) ones assisted by memory or known statistics. For a fair competition, we **boldface** the best accuracy in the first group, while the second best is underlined.

Unavailable Dynamics	Datasets / Algorithms	SVHN Train	SVHN Test	CIFAR-10 Train	CIFAR-10 Test	CINIC-10 Train	CINIC-10 Test
Stationary 	`FedAWE` (**ours**)	**86.5** ± 0.7 %	**86.1** ± 0.7 %	**68.1** ± 1.4 %	**66.3** ± 1.1 %	**47.9** ± 2.1 %	**47.3** ± 2.0 %
	`FedAvg` over *active*	82.6 ± 1.0 %	82.4 ± 1.1 %	64.1 ± 1.9 %	62.9 ± 1.4 %	43.6 ± 2.4 %	43.1 ± 2.4 %
	`FedAvg` over *all*	76.1 ± 2.1 %	76.1 ± 2.4 %	55.8 ± 2.1 %	55.4 ± 1.8 %	38.4 ± 2.1 %	38.0 ± 2.1 %
	`FedAU`	83.4 ± 1.0 %	83.2 ± 1.0 %	65.4 ± 1.4 %	64.1 ± 1.0 %	45.6 ± 1.5 %	45.2 ± 1.5 %
	`F3AST`	83.2 ± 0.7 %	83.2 ± 0.7 %	64.4 ± 1.1 %	63.5 ± 0.9 %	45.3 ± 1.2 %	44.8 ± 1.2 %
	`FedAvg` with *known* p_i's	86.1 ± 0.5 %	85.6 ± 0.5 %	65.4 ± 1.0 %	63.1 ± 0.9 %	45.0 ± 1.2 %	44.6 ± 1.1 %
	`MIFA` (*memory aided*)	84.2 ± 0.5 %	84.1 ± 0.6 %	66.6 ± 0.8 %	65.3 ± 0.6 %	47.5 ± 0.5 %	46.9 ± 0.5 %
	`FedVARP` (*memory aided*)	84.6 ± 0.2 %	84.3 ± 0.1 %	67.5 ± 0.2 %	66.3 ± 0.3 %	47.8 ± 0.2 %	47.2 ± 0.2 %
Non-stationary **(Staircase)** 	`FedAWE` (**ours**)	**85.9** ± 0.8 %	**85.6** ± 1.0 %	**67.7** ± 1.3 %	**66.0** ± 1.2 %	**47.5** ± 2.0 %	**46.9** ± 2.0 %
	`FedAvg` over *active*	82.5 ± 1.0 %	82.4 ± 0.9 %	64.2 ± 1.8 %	63.0 ± 1.4 %	43.7 ± 2.0 %	42.3 ± 2.2 %
	`FedAvg` over *all*	75.9 ± 2.1 %	75.9 ± 2.3 %	55.7 ± 2.1 %	55.4 ± 1.8 %	38.4 ± 2.0 %	37.9 ± 2.0 %
	`FedAU`	83.6 ± 0.8 %	83.4 ± 0.8 %	65.2 ± 1.7 %	63.9 ± 1.5 %	45.7 ± 1.5 %	45.1 ± 1.5 %
	`F3AST`	83.1 ± 0.6 %	83.1 ± 0.6 %	64.3 ± 1.1%	63.3 ± 0.9 %	45.2 ± 1.2 %	44.8 ± 1.2 %
	`FedAvg` with *known* p_i^t's	85.8 ± 0.8 %	85.2 ± 0.9 %	68.0 ± 1.6 %	66.1 ± 1.8 %	45.0 ± 1.1 %	44.7 ± 1.0 %
	`MIFA` (*memory aided*)	84.2 ± 0.5 %	84.0 ± 0.5 %	66.7 ± 0.7 %	65.3 ± 0.5 %	47.5 ± 0.5 %	46.9 ± 0.5 %
	`FedVARP` (*memory aided*)	84.6 ± 0.2 %	84.3 ± 0.3 %	67.3 ± 0.3 %	66.1 ± 0.3 %	47.7 ± 0.2 %	47.2 ± 0.1 %
Non-stationary **(Sine)** 	`FedAWE` (**ours**)	**85.7** ± 0.9 %	**85.6** ± 0.9 %	**64.9** ± 1.9 %	**63.5** ± 2.0 %	**46.4** ± 2.4 %	**45.8** ± 2.4 %
	`FedAvg` over *active*	82.1 ± 1.1 %	82.0 ± 1.3 %	63.3 ± 1.9 %	62.1 ± 1.8 %	43.1 ± 2.5 %	42.6 ± 2.5 %
	`FedAvg` over *all*	71.3 ± 2.5 %	71.3 ± 2.8 %	52.2 ± 2.4 %	52.1 ± 2.2 %	36.4 ± 2.0 %	36.0 ± 1.9 %
	`FedAU`	82.5 ± 1.4 %	82.5 ± 1.3 %	64.2 ± 2.3 %	63.0 ± 1.9 %	44.4 ± 2.1 %	43.9 ± 2.1 %
	`F3AST`	82.3 ± 1.0 %	82.3 ± 1.0 %	63.1 ± 1.7 %	62.3 ± 1.5 %	44.1 ± 1.6 %	43.7 ± 1.6 %
	`FedAvg` with *known* p_i^t's	86.3 ± 1.0 %	86.0 ± 1.0 %	69.1 ± 1.2 %	67.3 ± 1.3 %	47.9 ± 1.5 %	47.4 ± 1.1 %
	`MIFA` (*memory aided*)	84.2 ± 0.4 %	84.1 ± 0.4 %	66.6 ± 0.8 %	65.5 ± 0.6 %	47.4 ± 0.5 %	46.9 ± 0.4 %
	`FedVARP` (*memory aided*)	84.5 ± 0.2 %	84.3 ± 0.1 %	67.4 ± 0.2 %	66.0 ± 0.3 %	47.7 ± 0.1 %	47.1 ± 0.2 %
Non-stationary **(Interleaved Sine)** 	`FedAWE` (**ours**)	**85.2** ± 1.6 %	**84.6** ± 1.6 %	**64.8** ± 3.1 %	**63.3** ± 2.7 %	**47.1** ± 2.7 %	**46.6** ± 2.7 %
	`FedAvg` over *active*	80.9 ± 1.7 %	80.7 ± 1.7 %	61.9 ± 2.4 %	60.7 ± 2.0 %	41.9 ± 2.7 %	41.5 ± 2.7 %
	`FedAvg` over *all*	69.5 ± 3.4 %	69.5 ± 4.1 %	51.3 ± 2.7 %	51.3 ± 2.7 %	35.9 ± 2.0 %	35.6 ± 2.0 %
	`FedAU`	82.6 ± 1.3 %	82.4 ± 1.1 %	63.9 ± 2.2 %	62.8 ± 1.8 %	44.2 ± 2.2 %	43.8 ± 2.1 %
	`F3AST`	81.3 ± 1.2 %	81.3 ± 1.2 %	62.2 ± 2.1 %	61.3 ± 1.7 %	43.1 ± 2.2 %	42.7 ± 2.2 %
	`FedAvg` with *known* p_i^t's	85.8 ± 1.2 %	85.2 ± 1.3 %	68.7 ± 2.1 %	66.5 ± 2.4 %	47.2 ± 2.3 %	46.8 ± 2.2 %
	`MIFA` (*memory aided*)	83.8 ± 0.9 %	83.7 ± 0.8 %	65.8 ± 1.9 %	64.6 ± 1.6 %	46.5 ± 1.8 %	45.9 ± 1.7 %
	`FedVARP` (*memory aided*)	84.5 ± 0.3 %	84.1 ± 0.5 %	67.3 ± 0.3 %	65.7 ± 0.2 %	47.7 ± 0.5 %	47.2 ± 0.3 %

Benchmark algorithms and discussions. We compare `FedAWE` with six baseline algorithms, including `FedAvg` over active clients [31], `FedAvg` over all clients, `FedAU` [54], `F3AST` [43], `FedAvg` with known p_i^t' [41], `MIFA` [13] and `FedVARP` [19]. The details of the algorithm and the additional results are deferred to Appendix J. It is observed that `FedAWE` consistently outperforms the algorithms not aided by memory or known statistics. Surprisingly, `FedAWE` occasionally beats `MIFA` and `FedVARP`, which are memory-heavy. We attribute it to reuse of stored gradients from the unavailable clients. Although `FedAWE` brings in stalenss due to implicit gossiping, our results (Table 8 in Appendix J) indicate that there is no significant slowdown for `FedAWE` when compared to vanilla `FedAvg`, where we study the first round to achieve a targeted accuracy by different algorithms. In addition, `FedAWE` attains competitive or even better performance than `FedAvg` with known probability, yet unknown to the underlying dynamics in client unavailability.

8 Conclusion

In this paper, we have shown that the impacts of heterogeneous and non-stationary client unavailability can be significant through concrete examples on `FedAvg`. To address this, we have proposed an algorithm `FedAWE`, which provably converges by adaptively echoing clients' local improvement and by evenly diffusing local updates through implicit gossiping. Theoretically, it achieves the desired linear speedup property. Experiments have validated the superiority of `FedAWE` over state-of-the-art algorithms under diversified non-stationary dynamics. Future work will investigate how to extend our analysis to broader unavailability dynamics such as non-independent and non-stationary unavailability and how to incorporate our findings into federated learning algorithms of different local optimization methods.

Acknowledgments and Disclosure of Funding

We gratefully acknowledge the support from the National Science Foundation under grants 2106891, 2107062, the National Science Foundation CAREER award under grant 2340482, and the Sony Faculty Innovation Award. The research was sponsored by the Army Research Laboratory under Cooperative Agreement Number W911NF-23-2-0014. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory, the National Science Foundation, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. We also thank Connor J. McLaughlin for valuable discussions and feedback on this work.

References

[1] Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, and John Stephan. Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pages 1232–1300. PMLR, 2023.

[2] Dmitrii Avdiukhin and Shiva Kasiviswanathan. Federated learning under arbitrary communication patterns. In *International Conference on Machine Learning*, pages 425–435. PMLR, 2021.

[3] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, Brendan McMahan, et al. Towards federated learning at scale: System design. *Proceedings of Machine Learning and Systems*, 1:374–388, 2019.

[4] Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006.

[5] Dan Busbridge, Jason Ramapuram, Pierre Ablin, Tatiana Likhomanenko, Eeshan Gunesh Dhekane, Xavier Suau Cuadros, and Russell Webb. How to scale your ema. *Advances in Neural Information Processing Systems*, 36, 2024.

[6] Jin-Hua Chen, Min-Rong Chen, Guo-Qiang Zeng, and Jia-Si Weng. Bdfl: a byzantine-fault-tolerance decentralized federated learning method for autonomous vehicle. *IEEE Transactions on Vehicular Technology*, 70(9):8639–8652, 2021.

[7] Wenlin Chen, Samuel Horváth, and Peter Richtárik. Optimal client sampling for federated learning. *Transactions on Machine Learning Research*, 2022.

[8] Yae Jee Cho, Pranay Sharma, Gauri Joshi, Zheng Xu, Satyen Kale, and Tong Zhang. On the convergence of federated averaging with cyclic client participation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 5677–5721. PMLR, 23–29 Jul 2023.

[9] Yae Jee Cho, Jianyu Wang, Tarun Chirvolu, and Gauri Joshi. Communication-efficient and model-heterogeneous personalized federated learning via clustered knowledge transfer. *IEEE Journal of Selected Topics in Signal Processing*, 2023.

[10] Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards understanding biased client selection in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 10351–10375. PMLR, 2022.

[11] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

[12] Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69(345):118–121, 1974.

[13] Xinran Gu, Kaixuan Huang, Jingzhao Zhang, and Longbo Huang. Fast federated learning in the presence of arbitrary device unavailability. *Advances in Neural Information Processing Systems*, 34:12052–12064, 2021.

[14] Allan Gut and Allan Gut. *Probability: a graduate course*, volume 200. Springer, 2006.

[15] John Hajnal and MS Bartlett. Weak ergodicity in non-homogeneous markov chains. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 54, pages 233–246. Cambridge Univ Press, 1958.

[16] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

[17] Xinmeng Huang, Yiming Chen, Wotao Yin, and Kun Yuan. Lower bounds and nearly optimal algorithms in distributed learning with communication compression. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[18] Mark Jerrum and Alistair Sinclair. Conductance and the rapid mixing property for markov chains: the approximation of permanent resolved. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, pages 235–244, 1988.

[19] Divyansh Jhunjhunwala, Pranay Sharma, Aushim Nagarkatti, and Gauri Joshi. Fedvarp: Tackling the variance due to partial client participation in federated learning. In *Uncertainty in Artificial Intelligence*, pages 906–916. PMLR, 2022.

[20] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[21] Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*. PMLR, 2022.

[22] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

[23] David Kempe, Alin Dobra, and Johannes Gehrke. Gossip-based computation of aggregate information. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 482–491. IEEE, 2003.

[24] Anastasiia Koloskova, Sebastian U Stich, and Martin Jaggi. Sharper convergence guarantees for asynchronous sgd for distributed and federated learning. *Advances in Neural Information Processing Systems*, 35:17202–17215, 2022.

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[26] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

[27] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smithy. Feddane: A federated newton-type method. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 1227–1231. IEEE, 2019.

[28] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020.

[29] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

[30] Nancy A. Lynch. *Distributed Algorithms*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1996.

[31] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[32] Konstantin Mishchenko, Francis Bach, Mathieu Even, and Blake E Woodworth. Asynchronous sgd beats minibatch sgd under arbitrary delays. *Advances in Neural Information Processing Systems*, 35:420–433, 2022.

[33] Angelia Nedić, Alex Olshevsky, and Michael G Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.

[34] Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

[35] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

[36] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[37] John Nguyen, Kshitiz Malik, Hongyuan Zhan, Ashkan Yousefpour, Mike Rabbat, Mani Malek, and Dzmitry Huba. Federated learning with buffered asynchronous aggregation. In *International Conference on Artificial Intelligence and Statistics*, pages 3581–3607. PMLR, 2022.

[38] Thien Duc Nguyen, Samuel Marchal, Markus Miettinen, Hossein Fereidooni, N Asokan, and Ahmad-Reza Sadeghi. Dïot: A federated self-learning anomaly detection system for iot. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 756–767. IEEE, 2019.

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[40] Muzi Peng, Jiangwei Wang, Dongjin Song, Fei Miao, and Lili Su. Privacy-preserving and uncertainty-aware federated trajectory prediction for connected autonomous vehicles. In *The 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2023)*. IEEE/RSJ, 2023.

[41] Jake Perazzone, Shiqiang Wang, Mingyue Ji, and Kevin S Chan. Communication-efficient device scheduling for federated learning using stochastic optimization. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 1449–1458. IEEE, 2022.

[42] Swaroop Ramaswamy, Rajiv Mathews, Kanishka Rao, and Françoise Beaufays. Federated learning for emoji prediction in a mobile keyboard. *arXiv preprint arXiv:1906.04329*, 2019.

[43] Mónica Ribero, Haris Vikalo, and Gustavo De Veciana. Federated learning under intermittent client availability and time-varying communication constraints. *IEEE Journal of Selected Topics in Signal Processing*, 17(1):98–111, 2022.

[44] Yichen Ruan, Xiaoxi Zhang, Shu-Che Liang, and Carlee Joe-Wong. Towards flexible device participation in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3403–3411. PMLR, 2021.

[45] Devavrat Shah et al. Gossip algorithms. *Foundations and Trends® in Networking*, 3(1):1–125, 2009.

[46] Artin Spiridonoff, Alex Olshevsky, and Ioannis Ch Paschalidis. Robust asynchronous stochastic gradient-push: Asymptotically optimal and network-independent performance for strongly convex functions. *Journal of Machine Learning Research*, 21(58), 2020.

[47] Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

[48] Mohammad Taha Toghani and César A Uribe. Unbounded gradients in federated learning with buffered asynchronous aggregation. In *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–8. IEEE, 2022.

[49] David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge university press, 2005.

[50] Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of local-update sgd algorithms. *The Journal of Machine Learning Research*, 22(1):9709–9758, 2021.

[51] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in Neural Information Processing Systems*, 33:7611–7623, 2020.

[52] Jianyu Wang, Anit Kumar Sahu, Gauri Joshi, and Soummya Kar. Matcha: A matching-based link scheduling strategy to speed up distributed optimization. *IEEE Transactions on Signal Processing*, 70:5208–5221, 2022.

[53] Shiqiang Wang and Mingyue Ji. A unified analysis of federated learning with arbitrary client participation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[54] Shiqiang Wang and Mingyue Ji. A lightweight method for tackling unknown participation statistics in federated averaging. In *The Twelfth International Conference on Learning Representations*, 2024.

[55] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems. *IEEE journal on selected areas in communications*, 37(6):1205–1221, 2019.

[56] Ming Wen, Chengchang Liu, and Yuedong Xu. Communication efficient distributed newton method over unreliable networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15832–15840, 2024.

[57] Ming Xiang, Stratis Ioannidis, Edmund Yeh, Carlee Joe-Wong, and Lili Su. Towards bias correction of fedavg over nonuniform and time-varying communications. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 6719–6724, 2023.

[58] Cong Xie, Sanmi Koyejo, and Indranil Gupta. Asynchronous federated optimization. *arXiv preprint arXiv:1903.03934*, 2019.

[59] Yikai Yan, Chaoyue Niu, Yucheng Ding, Zhenzhe Zheng, Shaojie Tang, Qinya Li, Fan Wu, Chengfei Lyu, Yanghe Feng, and Guihai Chen. Federated optimization under intermittent client availability. *INFORMS Journal on Computing*, 2023.

[60] Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-IID federated learning. In *International Conference on Learning Representations*, 2021.

[61] Haibo Yang, Xin Zhang, Prashant Khanduri, and Jia Liu. Anarchic federated learning. In *International Conference on Machine Learning*, pages 25331–25363. PMLR, 2022.

[62] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*, 2018.

[63] Hao Ye, Le Liang, and Geoffrey Ye Li. Decentralized federated learning with unreliable communications. *IEEE Journal of Selected Topics in Signal Processing*, 16(3):487–500, 2022.

[64] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pages 7184–7193. PMLR, 2019.

[65] Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5693–5700, 2019.

[66] Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

[67] Xiaotong Yuan and Ping Li. On convergence of fedprox: Local dissimilarity invariant bounds, non-smoothness and beyond. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[68] Hossein Zakerinia, Shayan Talaei, Giorgi Nadiradze, and Dan Alistarh. Communication-efficient federated learning with data and client heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pages 3448–3456. PMLR, 2024.

[69] Tengchan Zeng, Omid Semiari, Mingzhe Chen, Walid Saad, and Mehdi Bennis. Federated learning on the road autonomous controller design for connected and autonomous vehicles. *IEEE Transactions on Wireless Communications*, 21(12):10407–10423, 2022.

Appendices

Here, we provide an overview of the Appendix. In particular, the proofs of the main results are presented and backed by supporting lemmas and propositions.

A Limitations

The limitations of our work are two-fold:

1. The client unavailability dynamics are assumed to be independent and strictly positive across clients and rounds. While deriving guarantees is generally challenging without assuming independence and positivity (see Section 3), it is interesting to explore how to relax the client unavailability dynamics, where the probabilities can potentially have arbitrary trajectories.

2. Our study focuses on heterogeneous and non-stationary client unavailability in federated learning, which may vary greatly due to its inherent uncontrollable nature. Although we have shown `FedAWE` provably converges to a stationary point of even non-convex objectives, an interesting yet challenging future direction is to incorporate variance reduction techniques for a more robust update.

B Broader Impacts

Federated learning has become the main trend for distributed learning in recent years and has empowered commercial industries such as autonomous vehicles, the Internet of Things, and natural language processing. Our paper focuses on the practical implementation of federated learning systems in the real world and has significantly advanced the theory and algorithms for federated learning by bringing together insights from statistics, optimization, distributed computing and engineering practices. In addition, our research is important for federated learning systems to expand their outreach to more undesirable deployment environments. We are unaware of any potential negative social impacts of our work.

C Nomenclatures

In this section, we provide the notations and nomenclatures used throughout our proofs for a comprehensive presentation. However, it is worth noting that all notations have been properly introduced before their first use. We next articulate the missing definitions and equation formulas.

Table 3: Notation table

$\|\boldsymbol{v}\|_2$	The l_2 norm of a given vector $\boldsymbol{v}$.
$\|A\|_{\mathrm{F}}$	The Frobenius norm of a given matrix A.
$\mathcal{F}^t$	The sigma algebra generated by randomness up to round t.
$\lambda_2(A)$	The second largest eigenvalue of a square matrix A.
$\mathbb{R}^d$	A d-dimensional vector space, where d denotes the dimension.
$[m]$	A set $\{k \mid k \in \mathbb{N}, k \in [1, m]\}$.
$\mathbb{1}_{\{\mathcal{E}\}}$	An indicator function of event $\mathcal{E}$, i.e., $\mathbb{1}_{\{\mathcal{E}\}} = 1$ when event $\mathcal{E}$ occurs, but $\mathbb{1}_{\{\mathcal{E}\}} = 0$ otherwise.
$\lesssim$	$f(n) \lesssim g(n)$, if there exists a constant $c_o > 0$ and an integer $n_0 \in \mathbb{N}$, $f(n) \le c_o g(n)$ for all $n \ge n_0$.
$\asymp$	$f(n) \asymp g(n)$, if there exists a constant $c_\Theta > 0$ and an integer $n_0 \in \mathbb{N}$, $f(n) = c_\Theta g(n)$ for all $n \ge n_0$.

Missing definitions and equation formulas.

Table 4: Algorithmic nomenclature table

$\mathcal{A}^t$	The set of active clients in round t.
W^t	A doubly stochastic matrix to capture the information mixing error. Its definition can be found in (4).
$\tau_i(t)$	$\tau_i(t) \triangleq \sup\{t' \mid t' < t, i \in \mathcal{A}^{t'}\}$ defines client i's most recent active round. In particular, $\tau_i(0) = -1$ for all $i \in [m]$.
$\boldsymbol{x}_i^t$	The real model at client i at the **beginning** of round t in Algorithm 1.
$\boldsymbol{z}_i^t$	The auxiliary model at client i at the **beginning** of round t. Refer to Definition 1 for more details. The sequence is for analysis only and is not computed by any clients.
$\boldsymbol{x}^t$	The aggregated real model at the **end** of round $t-1$ in Algorithm 1.
$\boldsymbol{z}^t$	The auxiliary model at the **end** of round $t-1$.
$\boldsymbol{x}_i^{t\dagger}$, $\boldsymbol{z}_i^{t\dagger}$	The real model of an active client i, and auxiliary model of an active client i after s-step local computation in round t, respectively. Refer to Algorithm 1 for more details.
$\boldsymbol{x}_i^{(t,r)}$	The real model at client i after r-step local computation.
$\bar{\boldsymbol{x}}^t$, $\bar{\boldsymbol{z}}^t$	The real and auxiliary model mean over all clients in a distributed system and in round t, respectively.
$F_i(\boldsymbol{x})$	The local objective function at client i, which is assumed to be non-convex.
$F(\boldsymbol{x})$	The global objective function defined in (1): $F(\boldsymbol{x}) \triangleq \sum_{i=1}^m F_i(\boldsymbol{x})/m$.
$\nabla \ell_i(\boldsymbol{x})$	The local stochastic gradient function at client i taken with respect to $\boldsymbol{x}$.
$\nabla F_i(\boldsymbol{x})$	The local true gradient function at client i taken with respect to $\boldsymbol{x}$.
$\mathcal{D}_i$	Client i's local data distribution.
ξ_i	An **independent** stochastic sample drawn from client i's local distribution $\mathcal{D}_i$.

Table 5: Variable table

L	Lipschitz constant in Assumption 2.
σ^2	The upper bound of the stochastic gradient variance.
$(\beta,\ \zeta)$	Parameters that capture the averaged gradient dissimilarity between global and local objectives.
ρ	The spectral norm of a stochastic matrix in expectation.
s	The number of local computation steps.
m	The number of clients in the federated learning system.

The iterate of $\boldsymbol{z}_i$ when $i \in \mathcal{A}^{t-1}$.

$$\boldsymbol{z}_i^t = \frac{1}{|\mathcal{A}^{t-1}|} \sum_{j \in \mathcal{A}^{t-1}} \left(\boldsymbol{z}_j^{t-1} - \eta_l \eta_g \sum_{r=0}^{s-1} \nabla \ell_j(\boldsymbol{x}_j^{(t-1,r)}; \xi_i^{(t,r)}) \right)$$
$$+ \frac{\eta_l \eta_g}{|\mathcal{A}^{t-1}|} \sum_{j \in \mathcal{A}^{t-1}} (t - 2 - \tau_j(t-1)) \sum_{r=0}^{s-1} \left(\nabla F_j(\boldsymbol{x}_j^{\tau_j(t-1)+1}) - \nabla \ell_j(\boldsymbol{x}_j^{(t-1,r)}; \xi_i^{(t,r)}) \right). \quad (18)$$

Local parameter innovation $\widetilde{\boldsymbol{G}}^t$ of the auxiliary sequence.

$$
\begin{aligned}
\widetilde{\boldsymbol{G}}_i^t &\triangleq \mathbb{1}_{\{i\in\mathcal{A}^t\}}\left[(t-\tau_i(t))\sum_{r=0}^{s-1}\nabla\ell_i(\boldsymbol{x}_i^{(t,r)}) - s\,(t-1-\tau_i(t))\,\nabla F_i(\boldsymbol{x}_i^{\tau_i(t)+1})\right] \\
&\quad + \mathbb{1}_{\{i\notin\mathcal{A}^t\}} s\nabla F_i(\boldsymbol{x}_i^{\tau_i(t)+1}) \\
&= \mathbb{1}_{\{i\in\mathcal{A}^t\}}\,(t-\tau_i(t))\sum_{r=0}^{s-1}\Big(\nabla\ell_i(\boldsymbol{x}_i^{(t,r)}) - \nabla F_i(\boldsymbol{x}_i^t)\Big) + s\nabla F_i(\boldsymbol{x}_i^t),
\end{aligned}
\tag{19}
$$

where the last equality holds because $\boldsymbol{x}_i^t = \boldsymbol{x}_i^{\tau_i(t)+1}$ and re-grouping.

Decomposition in the Proof of Lemma 6. The local parameter innovation of the auxiliary sequence $\widetilde{\boldsymbol{G}}^t$ can be decomposed as $\widetilde{\boldsymbol{G}}^t \triangleq \widetilde{\Delta}^t + \Delta^t + s\nabla\boldsymbol{F}_{\boldsymbol{x}}^t$. Detailed definitions can be found below.

- $[\widetilde{\Delta}^t]_i \triangleq \mathbb{1}_{\{i\in\mathcal{A}^t\}}(t-\tau_i(t))\sum_{r=0}^{s-1}\Big(\nabla\ell_i(\boldsymbol{x}_i^{(t,r)};\xi_i^{(t,r)}) - \nabla F_i(\boldsymbol{x}_i^{(t,r)})\Big)$;
- $[\Delta^t]_i \triangleq \mathbb{1}_{\{i\in\mathcal{A}^t\}}(t-\tau_i(t))\sum_{r=0}^{s-1}\Big(\nabla F_i(\boldsymbol{x}_i^{(t,r)}) - \nabla F_i(\boldsymbol{x}_i^t)\Big)$;
- $[\nabla\boldsymbol{F}_{\boldsymbol{x}}^t]_i \triangleq \nabla F_i(\boldsymbol{x}_i^t)$.

D Useful Inequalities

For completeness and for ease of exposition, we present some common inequalities that will be frequently used in our proofs.

The followings hold for any $\boldsymbol{a}_i\in\mathbb{R}^d$ and any $i\in[m]$.

1. Jensen's inequality.
$$
\left\|\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_i\right\|_2^2 \le \frac{1}{m}\sum_{i=1}^m\|\boldsymbol{a}_i\|_2^2 \quad\text{and}\quad \left\|\sum_{i=1}^m \boldsymbol{a}_i\right\|_2^2 \le m\sum_{i=1}^m\|\boldsymbol{a}_i\|_2^2. \tag{20}
$$
2. Young's inequality (a.k.a. Peter-Paul inequality).
$$
\langle\boldsymbol{a}_1,\boldsymbol{a}_2\rangle \le \frac{\|\boldsymbol{a}_1\|_2^2}{2\epsilon} + \frac{\epsilon\,\|\boldsymbol{a}_2\|_2^2}{2}, \quad\text{for any } \epsilon>0. \tag{21}
$$
Equivalently, we have
$$
\begin{aligned}
\|\boldsymbol{a}_1+\boldsymbol{a}_2\|_2^2 &= \|\boldsymbol{a}_1\|_2^2 + \|\boldsymbol{a}_2\|_2^2 + 2\langle\boldsymbol{a}_1,\boldsymbol{a}_2\rangle \\
&\le \left(1+\frac{1}{\epsilon}\right)\|\boldsymbol{a}_1\|_2^2 + (1+\epsilon)\|\boldsymbol{a}_2\|_2^2, \quad\text{for any } \epsilon>0.
\end{aligned}
\tag{22}
$$
3. Smoothness corollary. *Given Assumption 2, it holds that*
$$
\begin{aligned}
F(\boldsymbol{a}_1) - F(\boldsymbol{a}_2) &= \left\langle \boldsymbol{a}_1-\boldsymbol{a}_2, \int_0^1 \nabla F(\boldsymbol{a}_2+\tau(\boldsymbol{a}_1-\boldsymbol{a}_2))\mathrm{d}\tau\right\rangle \\
&= \langle\nabla F(\boldsymbol{a}_2), \boldsymbol{a}_1-\boldsymbol{a}_2\rangle + \int_0^1 \langle \boldsymbol{a}_1-\boldsymbol{a}_2, \nabla F(\boldsymbol{a}_2+\tau(\boldsymbol{a}_1-\boldsymbol{a}_2)) - \nabla F(\boldsymbol{a}_2)\rangle\,\mathrm{d}\tau \\
&\overset{(a)}{\le} \langle\nabla F(\boldsymbol{a}_2), \boldsymbol{a}_1-\boldsymbol{a}_2\rangle + L\int_0^1 \tau\,\|\boldsymbol{a}_1-\boldsymbol{a}_2\|_2\,\|(\boldsymbol{a}_1-\boldsymbol{a}_2)\|_2\,\mathrm{d}\tau \\
&\le \langle\nabla F(\boldsymbol{a}_2), \boldsymbol{a}_1-\boldsymbol{a}_2\rangle + \frac{L}{2}\|\boldsymbol{a}_1-\boldsymbol{a}_2\|_2^2,
\end{aligned}
\tag{23}
$$
where (a) follows from Cauchy-Schwartz inequality and Assumption 2.

E Descent Lemma (Lemma 3)

In this section, we first present a bound on multi-step local computation. Then, we apply the bound to the analysis of descent lemma.

E.1 Multi-step perturbation

Lemma 5. *For $s \geq 1$ and under Assumption 2, 3 and $\eta_l \leq 1/(4sL)$, we have*

$$\mathbb{E}\left[\left\|\sum_{r=0}^{s-1} \nabla F_i(\boldsymbol{x}_i^{(t,r)}) - \nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] \leq 4\eta_l^2 s^3 L^2 \sigma^2 + 16\eta_l^2 s^4 L^2 \left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2$$

Proof of Lemma 5. The proof shares a similar road map to [60, Lemma 2], but the objective is instead to show an upper bound with respect to $\|\nabla F_i(\boldsymbol{x}_i^t)\|_2^2$.

For $s \geq 1$, it holds that

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{r=0}^{s-1} \nabla F_i(\boldsymbol{x}_i^{(t,r)}) - \nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] &\overset{(a)}{\leq} s\sum_{r=0}^{s-1} \mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^{(t,r)}) - \nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] \\ &\overset{(b)}{\leq} sL^2 \sum_{r=0}^{s-1} \mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t,r)} - \boldsymbol{x}_i^t\right\|_2^2 \,\middle|\, \mathcal{F}^t\right], \end{aligned} \tag{24}$$

where inequality (a) holds because of Jensen's inequality, inequality (b) holds because of Assumption 2. It remains to bound $\mathbb{E}[\|\boldsymbol{x}_i^{(t,r)} - \boldsymbol{x}_i^t\|^2 \mid \mathcal{F}^t]$. In what follows, we use $\nabla \ell_i^{(t,k)}$ to denote $\nabla \ell_i(\boldsymbol{x}_i^{(t,k)})$ and $\nabla F_i^{(t,k)}$ as $\nabla F_i(\boldsymbol{x}_i^{(t,k)})$, respectively, for ease of presentation.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t,r)} - \boldsymbol{x}_i^t\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] = \mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t,r-1)} - \boldsymbol{x}_i^t - \eta_l \nabla \ell_i^{(t,r-1)}\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] \\
&= \mathbb{E}\left[\left\|-\eta_l\left(\nabla \ell_i^{(t,r-1)} - \nabla F_i^{(t,r-1)}\right) + \boldsymbol{x}_i^{(t,r-1)} - \boldsymbol{x}_i^t - \eta_l\left(\nabla F_i^{(t,r-1)} - \nabla F_i^t + \nabla F_i^t\right)\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] \\
&\overset{(c)}{=} \eta_l^2 \mathbb{E}\left[\left\|\nabla \ell_i^{(t,r-1)} - \nabla F_i^{(t,r-1)}\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] + \mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t,r-1)} - \boldsymbol{x}_i^t - \eta_l\left(\nabla F_i^{(t,r-1)} - \nabla F_i^t + \nabla F_i^t\right)\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] \\
&\overset{(d)}{\leq} \eta_l^2 \mathbb{E}\left[\left\|\nabla \ell_i^{(t,r-1)} - \nabla F_i^{(t,r-1)}\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] \\
&\quad + \left(1 + \frac{1}{2s-1}\right)\mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t,r-1)} - \boldsymbol{x}_i^t\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] + 2s\eta_l^2 \mathbb{E}\left[\left\|\nabla F_i^{(t,r-1)} - \nabla F_i^t + \nabla F_i^t\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] \\
&\leq \eta_l^2 \mathbb{E}\left[\left\|\nabla \ell_i^{(t,r-1)} - \nabla F_i^{(t,r-1)}\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] \\
&\quad + \left(1 + \frac{1}{2s-1}\right)\mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t,r-1)} - \boldsymbol{x}_i^t\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] + 4s\eta_l^2 \mathbb{E}\left[\left\|\nabla F_i^{(t,r-1)} - \nabla F_i^t\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] + 4s\eta_l^2\left\|\nabla F_i^t\right\|_2^2 \\
&\overset{(e)}{\leq} \eta_l^2\sigma^2 + 4s\eta_l^2\left\|\nabla F_i^t\right\|_2^2 \\
&\quad + \left(1 + \frac{1}{2s-1}\right)\mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t,r-1)} - \boldsymbol{x}_i^t\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] + 4sL^2\eta_l^2 \mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t,r-1)} - \boldsymbol{x}_i^t\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] \\
&= \eta_l^2\sigma^2 + 4s\eta_l^2\left\|\nabla F_i^t\right\|_2^2 + \left(1 + \frac{1}{2s-1} + 4sL^2\eta_l^2\right)\mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t,r-1)} - \boldsymbol{x}_i^t\right\|_2^2 \,\middle|\, \mathcal{F}^t\right],
\end{aligned}$$

where equality (c) holds because $\nabla \ell_i^{(t,k)}$ is an unbiased estimator of $\nabla F_i^{(t,r)}$, inequality (d) holds because of Young's inequality, inequality (e) holds because of Assumption 2.

By $\eta_l \leq \frac{1}{4sL}$, it holds that

$$\frac{1}{2s-1} + 4sL^2\eta_l^2 \leq \frac{1}{2s-1} + \frac{1}{4s} \leq \frac{2}{2s-1}.$$

Unroll the recursion, we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t,r)}-\boldsymbol{x}_i^t\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] &\le \sum_{k=0}^{r-1}\left(1+\frac{2}{2s-1}\right)^k\left(\eta_l^2\sigma^2+4s\eta_l^2\left\|\nabla F_i^t\right\|_2^2\right)\\
&\le \sum_{k=0}^{s-1}\left(1+\frac{2}{2s-1}\right)^k\left(\eta_l^2\sigma^2+4s\eta_l^2\left\|\nabla F_i^t\right\|_2^2\right)\\
&= \frac{2s-1}{2}\left[\left(1+\frac{2}{2s-1}\right)^{s-\frac{1}{2}}\left(1+\frac{2}{2s-1}\right)^{\frac{1}{2}}-1\right]\left(\eta_l^2\sigma^2+4s\eta_l^2\left\|\nabla F_i^t\right\|_2^2\right)\\
&\overset{(f)}{\le} \left(s-\frac{1}{2}\right)\left[\sqrt{3}e-1\right]\left(\eta_l^2\sigma^2+4s\eta_l^2\left\|\nabla F_i^t\right\|_2^2\right)\\
&\overset{(g)}{\le} 4s\eta_l^2\sigma^2+16s^2\eta_l^2\left\|\nabla F_i^t\right\|_2^2,
\end{aligned}
$$

where inequality (f) holds because of $(1+1/x)^x<\exp(1)$, inequality (g) holds because of $\sqrt{3}\exp(1)-1<4$. Plug it back into (24), we have the desired result

$$
\mathbb{E}\left[\left\|\sum_{r=0}^{s-1}\nabla F_i(\boldsymbol{x}_i^{(t,r)})-\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2 \,\middle|\, \mathcal{F}^t\right] \le 4\eta_l^2 s^3L^2\sigma^2+16\eta_l^2 s^4L^2\left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2.
$$

□

E.2 Descent lemma

Proof of Lemma 3. By Assumption 2 and inequality (23), we have

$$
F(\bar{\boldsymbol{z}}^{t+1})-F(\bar{\boldsymbol{z}}^t)\le \underbrace{\left\langle \nabla F(\bar{\boldsymbol{z}}^t),\bar{\boldsymbol{z}}^{t+1}-\bar{\boldsymbol{z}}^t\right\rangle}_{\text{(A)}}+\underbrace{\frac{L}{2}\left\|\bar{\boldsymbol{z}}^{t+1}-\bar{\boldsymbol{z}}^t\right\|_2^2}_{\text{(B)}}.
$$

The one-round innovation of $\bar{\boldsymbol{z}}$ can be rewritten as

$$
\begin{aligned}
&\bar{\boldsymbol{z}}^{t+1}-\bar{\boldsymbol{z}}^t=\frac{1}{m}\sum_{i\in\mathcal{A}^t}\left(\boldsymbol{z}_i^{t\dagger}-\boldsymbol{z}_i^t\right)+\frac{1}{m}\sum_{i\notin\mathcal{A}^t}\left(\boldsymbol{z}_i^{t+1}-\boldsymbol{z}_i^t\right)\\
&=\frac{1}{m}\sum_{i=1}^m\mathbb{1}_{\{i\in\mathcal{A}^t\}}\left(\eta_l\eta_g s\sum_{k=\tau_i(t)+1}^{t-1}\nabla F_i(\boldsymbol{x}_i^k)-\eta_l\eta_g(t-\tau_i(t))\sum_{r=0}^{s-1}\nabla\ell_i(\boldsymbol{x}_i^{(t,r)};\xi_i^{(t,r)})\right)\\
&\quad-\frac{\eta_l\eta_g s}{m}\sum_{i=1}^m\mathbb{1}_{\{i\notin\mathcal{A}^t\}}\nabla F_i(\boldsymbol{x}_i^t)\\
&\overset{(a)}{=}\frac{1}{m}\sum_{i=1}^m\mathbb{1}_{\{i\in\mathcal{A}^t\}}\eta_l\eta_g s(t-1-\tau_i(t))\nabla F_i(\boldsymbol{x}_i^t)-\frac{1}{m}\sum_{i=1}^m\mathbb{1}_{\{i\in\mathcal{A}^t\}}\eta_l\eta_g(t-\tau_i(t))\sum_{r=0}^{s-1}\nabla\ell_i(\boldsymbol{x}_i^{(t,r)};\xi_i^{(t,r)})\\
&\quad-\frac{\eta_l\eta_g s}{m}\sum_{i=1}^m\mathbb{1}_{\{i\notin\mathcal{A}^t\}}\nabla F_i(\boldsymbol{x}_i^t)\\
&\overset{(b)}{=}\frac{\eta_l\eta_g}{m}\sum_{i=1}^m\mathbb{1}_{\{i\in\mathcal{A}^t\}}(t-\tau_i(t))\sum_{r=0}^{s-1}\left(\nabla F_i(\boldsymbol{x}_i^{(t,r)})-\nabla\ell_i(\boldsymbol{x}_i^{(t,r)};\xi_i^{(t,r)})\right)\\
&\quad+\frac{\eta_l\eta_g}{m}\sum_{i=1}^m\mathbb{1}_{\{i\in\mathcal{A}^t\}}(t-\tau_i(t))\sum_{r=0}^{s-1}\left(\nabla F_i(\boldsymbol{x}_i^t)-\nabla F_i(\boldsymbol{x}_i^{(t,r)})\right)\\
&\quad-\frac{\eta_l\eta_g s}{m}\sum_{i=1}^m\nabla F_i(\boldsymbol{x}_i^t),
\end{aligned}
$$

where equality (a) using the fact that $\boldsymbol{x}_i^k = \boldsymbol{x}_i^t$ for all k such that $\tau_i(t) + 1 \le k \le t$, and equality (b) is obtained by adding and subtracting $\nabla \ell_i(\boldsymbol{x}_i^t; \xi_i^{(t,r)})$ and by the fact that $\left(\mathbb{1}_{\{i\in\mathcal{A}^t\}} + \mathbb{1}_{\{i\notin\mathcal{A}^t\}}\right) = 1$.

Bounding (A)**.**

$$
\begin{aligned}
\text{(A)} &= \left\langle \nabla F(\bar{\boldsymbol{z}}^t), \bar{\boldsymbol{z}}^{t+1} - \bar{\boldsymbol{z}}^t \right\rangle \\
&= \underbrace{\eta_l \eta_g \left\langle \nabla F(\bar{\boldsymbol{z}}^t), \frac{1}{m} \sum_{i=1}^m \mathbb{1}_{\{i\in\mathcal{A}^t\}} \sum_{p=-1}^{t-1} \mathbb{1}_{\{\tau_i(t)=p\}} (t-p) \sum_{r=0}^{s-1} \left(\nabla F_i(\boldsymbol{x}_i^{(t,r)}) - \nabla \ell_i(\boldsymbol{x}_i^{(t,r)}; \xi_i^{(t,r)}) \right) \right\rangle}_{\text{(A.I)}} \\
&\quad + \underbrace{\frac{\eta_l \eta_g}{m} \sum_{i=1}^m \mathbb{1}_{\{i\in\mathcal{A}^t\}} \sum_{p=-1}^{t-1} \mathbb{1}_{\{\tau_i(t)=p\}} \left\langle \nabla F(\bar{\boldsymbol{z}}^t), (t-p) \sum_{r=0}^{s-1} \left(\nabla F_i(\boldsymbol{x}_i^t) - \nabla F_i(\boldsymbol{x}_i^{(t,r)}) \right) \right\rangle}_{\text{(A.II)}} \\
&\quad + \underbrace{\frac{\eta_l \eta_g s}{m} \sum_{i=1}^m \left\langle \nabla F(\bar{\boldsymbol{z}}^t), \nabla F_i(\boldsymbol{z}_i^t) - \nabla F_i(\boldsymbol{x}_i^t) \right\rangle}_{\text{(A.III)}} - \underbrace{\eta_l \eta_g s \left\langle \nabla F(\bar{\boldsymbol{z}}^t), \frac{1}{m} \sum_{i=1}^m \nabla F_i(\boldsymbol{z}_i^t) \right\rangle}_{\text{(A.IV)}}.
\end{aligned}
$$

Bounding (A.I)

$$
\begin{aligned}
&\mathbb{E}\left[\text{(A.I)} \middle| \mathcal{F}^t \right] \\
&\overset{(a)}{=} \eta_l \eta_g \mathbb{E}\left[\mathbb{E}\left[\left\langle \nabla F(\bar{\boldsymbol{z}}^t), \frac{1}{m} \sum_{i=1}^m \mathbb{1}_{\{i\in\mathcal{A}^t\}} \sum_{p=-1}^{t-1} \mathbb{1}_{\{\tau_i(t)=p\}} (t-p) \sum_{r=0}^{s-1} \left(\nabla F_i(\boldsymbol{x}_i^{(t,r)}) - \nabla \ell_i(\boldsymbol{x}_i^{(t,r)}; \xi_i^{(t,r)}) \right) \right\rangle \middle| \boldsymbol{x}_i^{(t,r)}, \mathcal{F}^t \right] \middle| \mathcal{F}^t \right] \\
&\overset{(b)}{=} \eta_l \eta_g \langle \nabla F(\bar{\boldsymbol{z}}^t), \\
&\qquad \frac{1}{m} \sum_{i=1}^m \mathbb{E}\left[\mathbb{1}_{\{i\in\mathcal{A}^t\}} \middle| \mathcal{F}^t \right] \sum_{p=-1}^{t-1} \mathbb{1}_{\{\tau_i(t)=p\}} (t-p) \sum_{r=0}^{s-1} \mathbb{E}\left[\mathbb{E}\left[\left(\nabla F_i(\boldsymbol{x}_i^{(t,r)}) - \nabla \ell_i(\boldsymbol{x}_i^{(t,r)}; \xi_i^{(t,r)}) \right) \middle| \boldsymbol{x}_i^{(t,r)}, \mathcal{F}^t \right] \middle| \mathcal{F}^t \right] \Bigg\rangle \\
&= 0,
\end{aligned}
$$

where equality (a) holds because of the law of total expectation, equality (b) holds because $\mathbb{1}_{\{i\in\mathcal{A}^t\}}$ is by definition independent of others and Assumption 3.

Bounding (A.II)

$$
\begin{aligned}
\text{(A.II)} &\overset{(c)}{\le} \frac{\eta_l \eta_g}{m} \sum_{i=1}^m \mathbb{1}_{\{i\in\mathcal{A}^t\}} \sum_{p=-1}^{t-1} \mathbb{1}_{\{\tau_i(t)=p\}} \left(\frac{s}{8} \left\| \nabla F(\bar{\boldsymbol{z}}^t) \right\|_2^2 + \frac{2(t-p)^2}{s} \left\| \sum_{r=0}^{s-1} \nabla F_i(\boldsymbol{x}_i^t) - \nabla F_i(\boldsymbol{x}_i^{(t,r)}) \right\|_2^2 \right) \\
&= \frac{\eta_l \eta_g s}{8m} \sum_{i=1}^m \mathbb{1}_{\{i\in\mathcal{A}^t\}} \left\| \nabla F(\bar{\boldsymbol{z}}^t) \right\|_2^2 \\
&\quad + \frac{\eta_l \eta_g}{m} \sum_{i=1}^m \mathbb{1}_{\{i\in\mathcal{A}^t\}} \sum_{p=-1}^{t-1} \mathbb{1}_{\{\tau_i(t)=p\}} \frac{2(t-p)^2}{s} \left\| \sum_{r=0}^{s-1} \nabla F_i(\boldsymbol{x}_i^t) - \nabla F_i(\boldsymbol{x}_i^{(t,r)}) \right\|_2^2,
\end{aligned}
$$

where inequality (c) holds because of Young's inequality. It follows that

$$\mathbb{E}\left[(\text{A.II})\middle|\mathcal{F}^t\right] \overset{(d)}{\leq} \frac{\eta_l\eta_g s}{8}\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 + \frac{8\eta_g\eta_l^3 s^2L^2\sigma^2}{m}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2$$

$$+ \frac{32\eta_g\eta_l^3 s^3L^2}{m}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2\left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2$$

$$= \frac{\eta_l\eta_g s}{8}\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 + \frac{8\eta_g\eta_l^3 s^2L^2\sigma^2}{m}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2$$

$$+ \frac{32\eta_g\eta_l^3 s^3L^2}{m}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2,$$

where inequality (d) holds because of Lemma 5, the last equality using the fact that $\boldsymbol{x}_i^k = \boldsymbol{x}_i^t$ for all k such that $\tau_i(t)+1 \leq k \leq t$.

Bounding (A.III)**.**

$$(\text{A.III}) = \frac{\eta_l\eta_g s}{m}\sum_{i=1}^m\left\langle\nabla F(\bar{\boldsymbol{z}}^t), \nabla F_i(\boldsymbol{z}_i^t) - \nabla F_i(\boldsymbol{x}_i^t)\right\rangle \overset{(e)}{\leq} \frac{\eta_l\eta_g s}{8}\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 + \frac{2\eta_l\eta_g sL^2}{m}\sum_{i=1}^m\left\|\boldsymbol{z}_i^t - \boldsymbol{x}_i^t\right\|_2^2,$$

where inequality (e) follows from Young's inequality and Assumption 2. It holds that,

$$\mathbb{E}\left[(\text{A.III})\middle|\mathcal{F}^t\right] \leq \frac{\eta_l\eta_g s}{8}\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 + \frac{2\eta_l\eta_g sL^2}{m}\sum_{i=1}^m\left\|\boldsymbol{z}_i^t - \boldsymbol{x}_i^t\right\|_2^2.$$

Bounding (A.IV)

$$(\text{A.IV}) = \frac{\eta_l\eta_g s}{2}\left(\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 + \left\|\frac{1}{m}\sum_{i=1}^m\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2 - \left\|\nabla F(\bar{\boldsymbol{z}}^t) - \frac{1}{m}\sum_{i=1}^m\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right),$$

where the equality follows from the identity in Appendix D (3). It holds that

$$\mathbb{E}\left[(\text{A.IV})\middle|\mathcal{F}^t\right] = \frac{\eta_l\eta_g s}{2}\left(\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 + \left\|\frac{1}{m}\sum_{i=1}^m\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2 - \left\|\frac{1}{m}\sum_{i=1}^m\nabla F_i(\bar{\boldsymbol{z}}^t) - \frac{1}{m}\sum_{i=1}^m\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right)$$

$$\geq \frac{\eta_l\eta_g s}{2}\left(\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 + \left\|\frac{1}{m}\sum_{i=1}^m\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2 - \frac{L^2}{m}\sum_{i=1}^m\left\|\bar{\boldsymbol{z}}^t - \boldsymbol{z}_i^t\right\|_2^2\right).$$

Putting (A) together,

$$\mathbb{E}\left[(\text{A})\middle|\mathcal{F}^t\right] \leq -\frac{\eta_l\eta_g s}{4}\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 + \frac{8\eta_g\eta_l^3 s^2L^2\sigma^2}{m}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2$$

$$+ \frac{2\eta_l\eta_g sL^2}{m}\sum_{i=1}^m\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2 + \frac{\eta_l\eta_g sL^2}{2m}\sum_{i=1}^m\left\|\bar{\boldsymbol{z}}^t - \boldsymbol{z}_i^t\right\|_2^2$$

$$- \frac{\eta_l\eta_g s}{2}\left\|\frac{1}{m}\sum_{i=1}^m\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2 + \frac{32\eta_g\eta_l^3 s^3L^2}{m}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2.$$

Bounding (B)**.**

$$
\begin{aligned}
(\mathrm{B}) \le\ & \underbrace{2L\frac{\eta_l^2\eta_g^2}{m^2}\left\|\sum_{i=1}^m \mathbb{1}_{\{i\in\mathcal{A}^t\}}(t-\tau_i(t))\sum_{r=0}^{s-1}\left(\nabla F_i(\boldsymbol{x}_i^{(t,r)})-\nabla \ell_i(\boldsymbol{x}_i^{(t,r)};\xi_i^{(t,r)})\right)\right\|_2^2}_{(\mathrm{B.I})}\\
& + \underbrace{2L\frac{\eta_l^2\eta_g^2}{m^2}m\sum_{i=1}^m \mathbb{1}_{\{i\in\mathcal{A}^t\}}(t-\tau_i(t))^2\left\|\sum_{r=0}^{s-1}\left(\nabla F_i(\boldsymbol{x}_i^{t})-\nabla F_i(\boldsymbol{x}_i^{(t,r)})\right)\right\|_2^2}_{(\mathrm{B.II})}\\
& + \underbrace{2L\frac{\eta_l^2\eta_g^2 s^2}{m^2}m\sum_{i=1}^m \left\|\nabla F_i(\boldsymbol{x}_i^t)-\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2}_{(\mathrm{B.III})} + \underbrace{2L\eta_l^2\eta_g^2 s^2\left\|\frac{1}{m}\sum_{i=1}^m \nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2}_{(\mathrm{B.IV})}
\end{aligned}
$$

Bounding (B.I) Recall that $\delta_{\max} \triangleq \sup_{i\in[m],t\in[T]} p_i^t$. It holds that,

$$
\begin{aligned}
\mathbb{E}\left[(\mathrm{B.I})\middle|\mathcal{F}^t\right] &\overset{(f)}{=} 2L\frac{\eta_l^2\eta_g^2}{m^2}\sum_{i=1}^m \mathbb{E}\left[\mathbb{1}_{\{i\in\mathcal{A}^t\}}\middle|\mathcal{F}^t\right](t-\tau_i(t))^2\sum_{r=0}^{s-1}\mathbb{E}\left[\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^{(t,r)})-\nabla \ell_i(\boldsymbol{x}_i^{(t,r)};\xi_i^{(t,r)})\right\|_2^2\middle|\boldsymbol{x}_i^{(t,r)},\mathcal{F}^t\right]\middle|\mathcal{F}^t\right]\\
&\overset{(g)}{\le} \frac{2\eta_l^2\eta_g^2 sL\delta_{\max}\sigma^2}{m^2}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2,
\end{aligned}
$$

where equality (f) holds by the law of total expectation and by the independence of event $\{i\in\mathcal{A}^t\}$, inequality (g) holds because of Assumption 3 and by definition $p_i^t \le \delta_{\max}$.

Bounding (B.II) We have,

$$
\begin{aligned}
\mathbb{E}\left[(\mathrm{B.II})\middle|\mathcal{F}^t\right] &\le 2L\frac{\eta_l^2\eta_g^2}{m}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 4\eta_l^2 s^3L^2\sigma^2\\
&\quad + 2L\frac{\eta_l^2\eta_g^2}{m}\sum_{i=1}^m \mathbb{1}_{\{\tau_i(t)=p\}}\sum_{p=-1}^{t-1}(t-p)^2 16\eta_l^2 s^4 L^2\left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2\\
&= \frac{8\eta_g^2\eta_l^4 s^3L^3\sigma^2}{m}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 + \frac{32\eta_g^2\eta_l^4 s^4L^3}{m}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2,
\end{aligned}
$$

where the last equality using the fact that $\boldsymbol{x}_i^k = \boldsymbol{x}_i^t$ for all k such that $\tau_i(t)+1\le k\le t$.

Bounding (B.III)**.** $\mathbb{E}\left[(\mathrm{B.III})\middle|\mathcal{F}^t\right] \le \frac{2\eta_l^2\eta_g^2 s^2L^3}{m}\sum_{i=1}^m\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2$.

Putting (B) together, we get

$$
\begin{aligned}
\mathbb{E}\left[(\mathrm{B})\middle|\mathcal{F}^t\right] \le\ & \frac{2\eta_l^2\eta_g^2 sL\delta_{\max}\sigma^2}{m^2}\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 + \frac{8\eta_g^2\eta_l^4 s^3L^3\sigma^2}{m}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2\\
& + \frac{32\eta_g^2\eta_l^4 s^4L^3}{m}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2\\
& + \frac{2\eta_l^2\eta_g^2 s^2L^3}{m}\sum_{i=1}^m\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2 + 2L\eta_l^2\eta_g^2 s^2\left\|\frac{1}{m}\sum_{i=1}^m\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2.
\end{aligned}
$$

Now, everything:

$$\begin{aligned}
\mathbb{E}\left[F(\bar{\boldsymbol{z}}^{t+1}) - F(\bar{\boldsymbol{z}}^t)\middle|\mathcal{F}^t\right] \leq &-\frac{\eta_l\eta_g s}{4}\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 \\
&- \frac{\eta_l\eta_g s}{2}\left(1-4L\eta_l\eta_g s\right)\left\|\frac{1}{m}\sum_{i=1}^m \nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2 \\
&+ \frac{2\eta_l^2\eta_g^2 sL\delta_{\max}\sigma^2}{m^2}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 \\
&+ \frac{8\eta_g\eta_l^3 s^2L^2\left(1+\eta_g\eta_l sL\right)\sigma^2}{m}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 \\
&+ 2\eta_l\eta_g sL^2\left(1+\eta_l\eta_g sL\right)\frac{1}{m}\sum_{i=1}^m\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2 + \frac{\eta_l\eta_g sL^2}{2m}\sum_{i=1}^m\left\|\boldsymbol{z}_i^t - \bar{\boldsymbol{z}}^t\right\|_2^2 \\
&+ 32\eta_g\eta_l^3 s^3L^2\left(1+\eta_g\eta_l sL\right)\frac{1}{m}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2 \\
\leq &-\frac{\eta_l\eta_g s}{4}\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 + \frac{2\eta_l^2\eta_g^2 sL\delta_{\max}\sigma^2}{m^2}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 \\
&+ \frac{9\eta_g\eta_l^3 s^2L^2\sigma^2}{m}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 \\
&+ 2.2\eta_l\eta_g sL^2\frac{1}{m}\sum_{i=1}^m\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2 + \frac{\eta_l\eta_g sL^2}{2m}\sum_{i=1}^m\left\|\boldsymbol{z}_i^t - \bar{\boldsymbol{z}}^t\right\|_2^2 \\
&+ 35\eta_g\eta_l^3 s^3L^2\frac{1}{m}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2,
\end{aligned}$$

where the last inequality holds because $\eta_l\eta_g \leq \frac{9}{100sL}$ and that $\left\|\frac{1}{m}\sum_{i=1}^m \nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2 \geq 0$. $\square$

F Intermediate Results

In this section, we present the intermediate results that serve as handy tools in building up our proofs afterwards.

F.1 Bounding local and global dissimilarity

Proposition 3. *For any t, it holds that*

$$\frac{1}{m}\sum_{i=1}^{m}\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2 \le \frac{3L^2}{m}\sum_{i=1}^{m}\left\|\boldsymbol{z}_i^t - \bar{\boldsymbol{z}}^t\right\|_2^2 + 3\left(\beta^2+1\right)\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 + 3\zeta^2.$$

Proof of Proposition 3.

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2 &= \frac{1}{m}\sum_{i=1}^{m}\left\|\nabla F_i(\boldsymbol{z}_i^t) - \nabla F_i(\bar{\boldsymbol{z}}^t) + \nabla F_i(\bar{\boldsymbol{z}}^t) - \nabla F(\bar{\boldsymbol{z}}^t) + \nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 \\
&\le \frac{3}{m}\sum_{i=1}^{m}\left\|\nabla F_i(\boldsymbol{z}_i^t) - \nabla F_i(\bar{\boldsymbol{z}}^t)\right\|_2^2 + \frac{3}{m}\sum_{i=1}^{m}\left\|\nabla F_i(\bar{\boldsymbol{z}}^t) - \nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 + 3\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 \\
&\overset{(a)}{\le} \frac{3L^2}{m}\sum_{i=1}^{m}\left\|\boldsymbol{z}_i^t - \bar{\boldsymbol{z}}^t\right\|_2^2 + 3\beta^2\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 + 3\zeta^2 + 3\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 \\
&= \frac{3L^2}{m}\sum_{i=1}^{m}\left\|\boldsymbol{z}_i^t - \bar{\boldsymbol{z}}^t\right\|_2^2 + 3\left(\beta^2+1\right)\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2 + 3\zeta^2,
\end{aligned}$$

where inequality (a) follows from Assumptions 2 and 4. □

F.2 Weight re-equalization (Proposition 1)

Proof of Proposition 1. We show Proposition 1 by induction.

When $T = 1$ and $i \in \mathcal{A}^0$, we have $\sum_{t=0}^{0}\mathbb{1}_{\{i\in\mathcal{A}^t\}}\left(t-\tau_i(t)\right) = \mathbb{1}_{\{i\in\mathcal{A}^0\}}\left(0-\tau_i(0)\right) = 1$. Therefore, the base case holds.

The induction hypothesis is that $\sum_{t=0}^{K-1}\mathbb{1}_{\{i\in\mathcal{A}^t\}}\left(t-\tau_i(t)\right) = K$ holds for $i \in \mathcal{A}^{K-1}$. Next, we focus on $K+1$:

$$\sum_{t=0}^{K}\mathbb{1}_{\{i\in\mathcal{A}^t\}}\left(t-\tau_i(t)\right) = \sum_{t=0}^{K-1}\mathbb{1}_{\{i\in\mathcal{A}^t\}}\left(t-\tau_i(t)\right) + \mathbb{1}_{\{i\in\mathcal{A}^K\}}\left(K-\tau_i(K)\right). \tag{25}$$

Now, we have two cases:

- Suppose $i \in \mathcal{A}^{K-1}$, then we simply have $\tau_i(K) = K-1$. It follows that Eq. (25) $\overset{(a)}{=} K+1$, where (a) follows from induction hypothesis.
- Suppose $i \notin \mathcal{A}^{K-1}$,

  $$\begin{aligned}
  \sum_{t=0}^{K}\mathbb{1}_{\{i\in\mathcal{A}^t\}}\left(t-\tau_i(t)\right) &\overset{(b)}{=} \sum_{t=0}^{\tau_i(K)}\mathbb{1}_{\{i\in\mathcal{A}^t\}}\left(t-\tau_i(t)\right) + \mathbb{1}_{\{i\in\mathcal{A}^K\}}\left(K-\tau_i(K)\right) \\
  &= \tau_i(K) + 1 + \left(K - \tau_i(K)\right) = K+1,
  \end{aligned}$$

  where (b) follows because $\mathbb{1}_{\{i\in\mathcal{A}^t\}} = 0$ for $\tau_i(K) \le t \le K-1$ and induction hypothesis that $\sum_{t=0}^{\tau_i(K)}\mathbb{1}_{\{i\in\mathcal{A}^t\}}\left(t-\tau_i(t)\right) = \tau_i(K)+1$ for $i \in \mathcal{A}^{\tau_i(K)}$.

□

F.3 Unavailable statistics (Lemma 2)

Proof of Lemma 2.

$$\mathbb{E}\left[t-\tau_i(t)\right] = \sum_{r=0}^{t}\mathbb{P}\left\{t-\tau_i(t) > r\right\} = \sum_{r=0}^{t}\prod_{r_1=t-r}^{t-1}\left(1-p_i^{r_1}\right) \le \sum_{r=0}^{t}(1-\delta)^r \le \frac{1}{\delta}.$$

From [14, Section 12, Theorem 12.3 (i)], we know that

$$\mathbb{E}\left[g(X)\right] = g(0) + \int_0^\infty g'(x)\mathbb{P}\left\{X > x\right\} \mathrm{d}x,$$

where X is a non-negative random variable, and g a non-negative strictly increasing differentiable function. It follows that,

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &\le 0 + 2\int_0^\infty x\mathbb{P}\left\{X > x\right\} \mathrm{d}x = 2\sum_{n=1}^{\infty}\int_{n-1}^{n} x\mathbb{P}\left\{X > x\right\} \mathrm{d}x \\
&\overset{(a)}{\le} 2\sum_{n=1}^{\infty} n\int_{n-1}^{n} \mathbb{P}\left\{X > x\right\} \mathrm{d}x \\
&\overset{(b)}{\le} 2\sum_{n=1}^{\infty} n\mathbb{P}\left\{X > n-1\right\} \int_{n-1}^{n} \mathrm{d}x = 2\sum_{n=1}^{\infty} n\mathbb{P}\left\{X > n-1\right\},
\end{aligned}$$

where inequality (a) holds because $x \le n$, $\forall x \in (n-1, n]$, inequality (b) holds because CCDF $\mathbb{P}\left\{X > x\right\}$ is non-increasing. In particular, for a discrete random variable, we have $\mathbb{P}\left\{X > n-1\right\} = \mathbb{P}\left\{X \ge n\right\}$.

Therefore,

$$\mathbb{E}\left[(t-\tau_i(t))^2\right] \le 2\sum_{n=1}^{\infty} n\mathbb{P}\left\{t - \tau_i(t) \ge n\right\} \le 2\sum_{n=1}^{\infty} n(1-\delta)^{n-1} \le \frac{2}{\delta^2}.$$

□

F.4 Auxiliary sequence construction and properties (Proposition 2)

Proposition 4. *For any $t \ge 0$, when $i \notin \mathcal{A}^t$, it holds that $\boldsymbol{x}_i^{t+1} - \boldsymbol{z}_i^{t+1} = \eta_l\eta_g s(t - \tau_i(t+1))\nabla F_i(\boldsymbol{x}_i^{\tau_i(t+1)+1})$; when $i \in \mathcal{A}^t$, it holds that $\boldsymbol{z}_i^{t\dagger} = \boldsymbol{x}_i^{t\dagger}$, $\boldsymbol{z}^{t+1} = \boldsymbol{x}^{t+1}$, and $\boldsymbol{z}_i^{t+1} = \boldsymbol{x}_i^{t+1}$.*

Proof of Proposition 4. The proof is divided into two parts: $i \notin \mathcal{A}^t$ and $i \in \mathcal{A}^t$,

When $i \notin \mathcal{A}^t$. It holds that

$$\begin{aligned}
\boldsymbol{x}_i^{t+1} - \boldsymbol{z}_i^{t+1} &= \boldsymbol{x}_i^{\tau_i(t+1)+1} - \left[\boldsymbol{z}_i^{\tau_i(t+1)+1} - \eta_l\eta_g s\sum_{k=\tau_i(t+1)+1}^{t}\nabla F_i(\boldsymbol{x}_i^{k})\right] \\
&\overset{(a)}{=} \boldsymbol{x}_i^{\tau_i(t+1)+1} - \left[\boldsymbol{x}_i^{\tau_i(t+1)+1} - \eta_l\eta_g s\sum_{k=\tau_i(t+1)+1}^{t}\nabla F_i(\boldsymbol{x}_i^{\tau_i(t+1)+1})\right] \\
&= \eta_l\eta_g s(t - \tau_i(t+1))\nabla F_i(\boldsymbol{x}_i^{\tau_i(t+1)+1}),
\end{aligned}$$

where equality (a) follows from Definition 1 for inactive clients.

When $i \in \mathcal{A}^t$. Note that if $\boldsymbol{z}_i^{t++} = \boldsymbol{x}_i^{t++}$ for each $i \in \mathcal{A}^t$, then by the aggregation rules, we know $\boldsymbol{x}^{t+1} = (1/|\mathcal{A}^t|)\sum_{i\in\mathcal{A}^t}\boldsymbol{x}_i^{t++} = (1/|\mathcal{A}^t|)\sum_{i\in\mathcal{A}^t}\boldsymbol{z}_i^{t++} = \boldsymbol{z}^{t+1}$. Then, we know that $\boldsymbol{x}_i^{t+1} = \boldsymbol{z}_i^{t+1}$, $\forall\, i \in \mathcal{A}^t$. Hence, to show the Proposition, it is sufficient to show $\boldsymbol{z}_i^{t++} = \boldsymbol{x}_i^{t++}$ holds for $i \in \mathcal{A}^t$, which can be shown by induction.

When $t = 0$,

$$\boldsymbol{z}_i^{0++} = \boldsymbol{z}_i^{0} + 0 - \left(\boldsymbol{x}_i^{(0,0)} - \boldsymbol{x}_i^{(0,s)}\right) = \boldsymbol{x}_i^{0} - \left(\boldsymbol{x}_i^{(0,0)} - \boldsymbol{x}_i^{(0,s)}\right) = \boldsymbol{x}_i^{0++}.$$

Thus, the base case holds. The induction hypothesis is that $\boldsymbol{z}_i^{t++} = \boldsymbol{x}_i^{t++}$, $\forall\, i \in \mathcal{A}^t$ is true for all $t \geq 0$. Now, we focus on $t+1$.

$$
\begin{aligned}
\boldsymbol{z}_i^{(t+1)++} &= \boldsymbol{z}_i^{t+1} + \eta_l \eta_g s \sum_{k=\tau_i(t+1)+1}^{t} \nabla F_i(\boldsymbol{x}_i^k) - (t+1-\tau_i(t+1)) \left(\boldsymbol{x}_i^{(t+1,0)} - \boldsymbol{x}_i^{(t+1,s)}\right) \\
&= \boldsymbol{z}_i^{t+1} + \eta_l \eta_g s (t-\tau_i(t+1)) \nabla F_i(\boldsymbol{x}_i^{\tau_i(t+1)+1}) - (t+1-\tau_i(t+1)) \left(\boldsymbol{x}_i^{(t+1,0)} - \boldsymbol{x}_i^{(t+1,s)}\right) \\
&\overset{(a)}{=} \boldsymbol{z}_i^{\tau_i(t+1)+1} - \eta_l \eta_g s (t-\tau_i(t+1)-1+1) \nabla F_i(\boldsymbol{x}_i^{\tau_i(t+1)+1}) \\
&\qquad + \eta_l \eta_g s (t-\tau_i(t+1)) \nabla F_i(\boldsymbol{x}_i^{\tau_i(t+1)+1}) - (t+1-\tau_i(t+1)) \left(\boldsymbol{x}_i^{(t+1,0)} - \boldsymbol{x}_i^{(t+1,s)}\right) \\
&= \boldsymbol{z}_i^{\tau_i(t+1)+1} - (t+1-\tau_i(t+1)) \left(\boldsymbol{x}_i^{(t+1,0)} - \boldsymbol{x}_i^{(t+1,s)}\right) \\
&\overset{(b)}{=} \boldsymbol{x}_i^{\tau_i(t+1)+1} - (t+1-\tau_i(t+1)) \left(\boldsymbol{x}_i^{(t+1,0)} - \boldsymbol{x}_i^{(t+1,s)}\right) \\
&= \boldsymbol{x}_i^{(t+1)++},
\end{aligned}
$$

where equality (a) follows from the auxiliary updates $\boldsymbol{z}_i$, and equality (b) holds because of the induction hypothesis and the fact that $\tau_i(t+1) < t+1$ and $i \in \mathcal{A}^{\tau_i(t+1)}$. □

Proof of Proposition 2. From Propositions 4, we have

$$
\begin{aligned}
\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2 &\leq \left\|\eta_l \eta_g s \left(t - \tau_i(t) - 1\right) \nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2 \\
&= \eta_l^2 \eta_g^2 s^2 \sum_{p=-1}^{t-1} \mathbb{1}_{\{\tau_i(t)=p\}} \left(t-p-1\right)^2 \left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2 .
\end{aligned}
$$

Take expectation over all the randomness

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2\right] &\overset{(a)}{\leq} \eta_l^2 \eta_g^2 s^2 \sum_{p=-1}^{t-1} \mathbb{E}\left[\mathbb{1}_{\{\tau_i(t)=p\}}\right] \left(t-p-1\right)^2 \mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2\right] \\
&\overset{(b)}{\leq} \eta_l^2 \eta_g^2 s^2 \sum_{p=-1}^{t-1} \left(t-p-1\right)^2 \mathbb{P}\left\{\tau_i(t)=p\right\} \cdot \mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^{p+1})\right\|_2^2\right],
\end{aligned}
$$

where inequality (a) follows because by definition $\mathbb{1}_{\{\tau_i(t)=p\}}$ is independent of $\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2$, inequality (b) follows because $\boldsymbol{x}_i^{p+1} = \boldsymbol{z}_i^{p+1}$ from Proposition 4.

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2\right] &= \eta_l^2 \eta_g^2 s^2 \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\sum_{p=-1}^{t-1} \mathbb{P}\left\{\tau_i(t)=p\right\}\left(t-p-1\right)^2 \mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^{p+1})\right\|_2^2\right] \\
&\overset{(c)}{\leq} \eta_l^2 \eta_g^2 s^2 \frac{1}{m}\sum_{i=1}^{m}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right]\left(\mathbb{E}\left[\left(t-\tau_i(t)\right)^2\right]\right) \\
&\overset{(d)}{\leq} \eta_l^2 \eta_g^2 s^2 \left(\frac{2}{\delta^2}\right)\frac{1}{m}\sum_{i=1}^{m}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right] \\
&\leq 3\eta_l^2 \eta_g^2 s^2 \left(\frac{2}{\delta^2}\right)\left(\beta^2+1\right)\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] + 3\eta_l^2 \eta_g^2 s^2 \left(\frac{2}{\delta^2}\right)\zeta^2 \\
&\quad + 3\eta_l^2 \eta_g^2 s^2 L^2 \left(\frac{2}{\delta^2}\right)\frac{1}{m}\sum_{i=1}^{m}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\boldsymbol{z}_i^t - \bar{\boldsymbol{z}}^t\right\|_2^2\right],
\end{aligned}
$$

where inequality (c) follows from re-indexing, inequality (d) from Lemma 2. □

F.5 Consensus error of the auxiliary sequence

Lemma 6 (Consensus error of z_i^t)**.** *Assuming that* $\eta_l \le \delta/(20sL)$*, and* $\eta_l\eta_g \le \delta(1-\sqrt{\rho})/(10sL(\sqrt{\rho}+1))$*, under Assumption 2, 3 and 4, it holds that*

$$\frac{1}{m}\sum_{i=1}^{m}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right] \le \frac{3\rho s\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2\delta^2}\sigma^2 + \frac{40\rho s^2\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2}\zeta^2 + \frac{40\rho s^2\eta_l^2\eta_g^2\left(\beta^2+1\right)}{(1-\sqrt{\rho})^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right].$$

Proof of Lemma 6. When $t=0$, $\boldsymbol{Z}^0=[\boldsymbol{z}^0,\cdots,\boldsymbol{z}^0]$, which immediately leads to

$$\boldsymbol{Z}^0\left(\mathbf{I}-\mathbf{J}\right)=[\boldsymbol{z}^0,\cdots,\boldsymbol{z}^0]-[\boldsymbol{z}^0,\cdots,\boldsymbol{z}^0]=\mathbf{0}.$$

For $t\ge 1$, recall that $W^{(t)}$ is a doubly stochastic matrix to characterize the information mixture, and $\widetilde{\boldsymbol{G}}^t$, defined in (19), captures the local parameter changes in each round. It can be seen that

$$\boldsymbol{Z}^{(t)}=\left(\boldsymbol{Z}^{(t-1)}-\eta_l\eta_g\widetilde{\boldsymbol{G}}^{t-1}\right)W^{(t-1)}.$$

Expanding $\boldsymbol{Z}$, we get

$$\begin{aligned}\boldsymbol{Z}^{(t)}\left(\mathbf{I}-\mathbf{J}\right)&=(\boldsymbol{Z}^{(t-1)}-\eta_l\eta_g\widetilde{\boldsymbol{G}}^{t-1})W^{(t-1)}\left(\mathbf{I}-\mathbf{J}\right)\\&=\boldsymbol{Z}^0\prod_{\ell=0}^{t-1}W^{\ell}\left(\mathbf{I}-\mathbf{J}\right)-\eta_l\eta_g\sum_{q=0}^{t-1}\widetilde{\boldsymbol{G}}^q\prod_{\ell=q}^{t-1}W^{(\ell)}\left(\mathbf{I}-\mathbf{J}\right).\end{aligned}$$

where the last follows from the fact that all clients are initiated at the same weights. Note that $\prod_{\ell=q}^{t-1}W^{(\ell)}\mathbf{I}=\prod_{\ell=q}^{t-1}W^{(\ell)}$ and $\prod_{\ell=q}^{t-1}W^{(\ell)}\mathbf{J}=\mathbf{J}$. Thus,

$$\boldsymbol{Z}^{(t)}\left(\mathbf{I}-\mathbf{J}\right)=\boldsymbol{Z}^0\left(\prod_{\ell=0}^{t-1}W^{\ell}-\mathbf{J}\right)-\eta_l\eta_g\sum_{q=0}^{t-1}\widetilde{\boldsymbol{G}}^q\left(\prod_{\ell=q}^{t-1}W^{(\ell)}-\mathbf{J}\right)=-\eta_l\eta_g\sum_{q=0}^{t-1}\widetilde{\boldsymbol{G}}^q\left(\prod_{\ell=q}^{t-1}W^{(\ell)}-\mathbf{J}\right),$$

where the last equality holds because that $\boldsymbol{Z}^0=[\boldsymbol{z}^0,\cdots,\boldsymbol{z}^0]$, which immediately leads to

$$\boldsymbol{Z}^0\left(\prod_{\ell=0}^{t-1}W^{\ell}-\mathbf{J}\right)=[\boldsymbol{z}^0,\cdots,\boldsymbol{z}^0]-[\boldsymbol{z}^0,\cdots,\boldsymbol{z}^0]=\mathbf{0}.$$

Let matrix notations $\widetilde{\Delta}^t$, Δ^t and $\nabla\boldsymbol{F}_{\boldsymbol{x}}^t$ define as follows:

$$\boldsymbol{G}_i^q=\underbrace{\mathbb{1}_{\{i\in\mathcal{A}^t\}}(t-\tau_i(t))\sum_{r=0}^{s-1}\left(\nabla\ell_i(\boldsymbol{x}_i^{(t,r)};\xi_i^{(t,r)})-\nabla F_i(\boldsymbol{x}_i^{(t,r)})\right)}_{[\widetilde{\Delta}^t]_i}+\underbrace{\mathbb{1}_{\{i\in\mathcal{A}^t\}}(t-\tau_i(t))\sum_{r=0}^{s-1}\left(\nabla F_i(\boldsymbol{x}_i^{(t,r)})-\nabla F_i(\boldsymbol{x}_i^{t})\right)}_{[\Delta^t]_i}+s\underbrace{\nabla F_i(\boldsymbol{x}_i^t)}_{[\nabla\boldsymbol{F}_{\boldsymbol{x}}^t]_i}.$$

It follows that

$$
\begin{aligned}
\|\boldsymbol{Z}^{(t)}(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^{2} &= \|\sum_{q=0}^{t-1}\left(\widetilde{\Delta}^{q}+\Delta^{q}+\nabla \boldsymbol{F}_{\boldsymbol{x}}^{q}\right)\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\|_{\mathrm{F}}^{2} \\
&= \|\sum_{q=0}^{t-1}\widetilde{\Delta}^{q}\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\|_{\mathrm{F}}^{2}+\|\sum_{q=0}^{t-1}\left(\Delta^{q}+\nabla \boldsymbol{F}_{\boldsymbol{x}}^{q}\right)\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\|_{\mathrm{F}}^{2} \\
&\quad +2\left\langle\sum_{q=0}^{t-1}\widetilde{\Delta}^{q}\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right), \sum_{q=0}^{t-1}\left(\Delta^{q}+\nabla \boldsymbol{F}_{\boldsymbol{x}}^{q}\right)\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\right\rangle_{\mathrm{F}}.
\end{aligned}
$$

Take expectation with respect to randomness in stochastic gradients, denote by $\mathbb{E}_{\xi}[\cdot]$:

$$
\begin{aligned}
\mathbb{E}_{\xi}\left[\|\boldsymbol{Z}^{(t)}(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^{2}\right] &= \mathbb{E}_{\xi}\left[\|\sum_{q=0}^{t-1}\widetilde{\Delta}^{q}\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\|_{\mathrm{F}}^{2}\right]+\mathbb{E}_{\xi}\left[\|\sum_{q=0}^{t-1}\left(\Delta^{q}+\nabla \boldsymbol{F}_{\boldsymbol{x}}^{q}\right)\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\|_{\mathrm{F}}^{2}\right] \\
&\quad +2\mathbb{E}_{\xi}\left[\left\langle\sum_{q=0}^{t-1}\widetilde{\Delta}^{q}\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right), \sum_{q=0}^{t-1}\left(\Delta^{q}+\nabla \boldsymbol{F}_{\boldsymbol{x}}^{q}\right)\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\right\rangle_{\mathrm{F}}\right] \\
&= \mathbb{E}_{\xi}\left[\|\sum_{q=0}^{t-1}\widetilde{\Delta}^{q}\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\|_{\mathrm{F}}^{2}\right]+\mathbb{E}_{\xi}\left[\|\sum_{q=0}^{t-1}\left(\Delta^{q}+\nabla \boldsymbol{F}_{\boldsymbol{x}}^{q}\right)\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\|_{\mathrm{F}}^{2}\right] \\
&\quad +2\left\langle\sum_{q=0}^{t-1}\mathbb{E}_{\xi}\left[\widetilde{\Delta}^{q}\right]\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right), \sum_{q=0}^{t-1}\left(\Delta^{q}+\nabla \boldsymbol{F}_{\boldsymbol{x}}^{q}\right)\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\right\rangle_{\mathrm{F}} \\
&\leq \mathbb{E}_{\xi}\left[\|\sum_{q=0}^{t-1}\widetilde{\Delta}^{q}\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\|_{\mathrm{F}}^{2}\right]+\mathbb{E}_{\xi}\left[\|\sum_{q=0}^{t-1}\left(\Delta^{q}+\nabla \boldsymbol{F}_{\boldsymbol{x}}^{q}\right)\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\|_{\mathrm{F}}^{2}\right],
\end{aligned}
$$

where the last inequality holds because $\mathbb{E}_{\xi}\left[\widetilde{\Delta}^{q}\right]=0$. Next, we take expectation over the remaining randomness.

$$
\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{Z}^{(t)}(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^{2}\right] &\leq \mathbb{E}\left[\|\sum_{q=0}^{t-1}\widetilde{\Delta}^{q}\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\|_{\mathrm{F}}^{2}\right]+\mathbb{E}\left[\|\sum_{q=0}^{t-1}\left(\Delta^{q}+\nabla \boldsymbol{F}_{\boldsymbol{x}}^{q}\right)\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\|_{\mathrm{F}}^{2}\right] \\
&\leq \eta_{l}^{2}\eta_{g}^{2}\underbrace{\|\sum_{q=0}^{t-1}\widetilde{\Delta}^{q}\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\|_{\mathrm{F}}^{2}}_{\text{(I)}} \\
&\quad +2\eta_{l}^{2}\eta_{g}^{2}\underbrace{\|\sum_{q=0}^{t-1}\Delta^{q}\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\|_{\mathrm{F}}^{2}}_{\text{(II)}} \\
&\quad +2\eta_{l}^{2}\eta_{g}^{2}s^{2}\underbrace{\|\sum_{q=0}^{t-1}\nabla \boldsymbol{F}_{\boldsymbol{x}}^{q}\left(\prod_{\ell=q}^{t-1} W^{(\ell)}-\mathbf{J}\right)\|_{\mathrm{F}}^{2}}_{\text{(III)}}.
\end{aligned}
\tag{26}
$$

Bounding $\mathbb{E}\left[(\mathrm{I})\right]$

$$
\begin{aligned}
\mathbb{E}\left[(\mathrm{I})\right] = & \sum_{q=0}^{t-1} \mathbb{E}\left[\|\widetilde{\Delta}^q \left(\prod_{\ell=q}^{t-1} W^{(\ell)} - \mathbf{J} \right) \|_{\mathrm{F}}^2 \right] \\
& + \sum_{q=0}^{t-1} \sum_{p=0, p\neq q}^{t-1} \mathbb{E}\left[\left\langle \widetilde{\Delta}^p \left(\prod_{\ell=p}^{t-1} W^{(\ell)} - \mathbf{J} \right), \widetilde{\Delta}^q \left(\prod_{\ell=q}^{t-1} W^{(\ell)} - \mathbf{J} \right) \right\rangle \right] \\
\overset{(a)}{\leq} & \sum_{q=0}^{t-1} \rho^{t-q} \mathbb{E}\left[\|\widetilde{\Delta}^q\|_{\mathrm{F}}^2 \right],
\end{aligned} \tag{27}
$$

where inequality (a) holds because of Assumption 3. It remains to bound $\mathbb{E}\left[\|\widetilde{\Delta}^q\|_{\mathrm{F}}^2\right]$.

$$
\|\widetilde{\Delta}^q\|_{\mathrm{F}}^2 = \sum_{i=1}^m \mathbb{1}_{\{i\in\mathcal{A}^q\}} \left\| \sum_{p=-1}^{q-1} \mathbb{1}_{\{\tau_i(t)=p\}} (q-p) \sum_{r=0}^{s-1} \left(\nabla \ell_i(\boldsymbol{x}_i^{(q,r)}; \xi_i^{(q,r)}) - \nabla F_i(\boldsymbol{x}_i^{(q,r)}) \right) \right\|_2^2 .
$$

$$
\begin{aligned}
\mathbb{E}_\xi \left[\|\widetilde{\Delta}^q\|_{\mathrm{F}}^2\right] &= \sum_{i=1}^m \mathbb{1}_{\{i\in\mathcal{A}^q\}} \sum_{p=-1}^{q-1} \mathbb{1}_{\{\tau_i(t)=p\}} (q-p)^2 \sum_{r=0}^{s-1} \mathbb{E}_\xi \left[\left\| \nabla \ell_i(\boldsymbol{x}_i^{(q,r)}; \xi_i^{(p,r)}) - \nabla F_i(\boldsymbol{x}_i^{(q,r)}) \right\|_2^2 \right] \\
&\leq s\sigma^2 \sum_{i=1}^m \mathbb{1}_{\{i\in\mathcal{A}^q\}} \sum_{p=-1}^{q-1} \mathbb{1}_{\{\tau_i(t)=p\}} (q-p)^2.
\end{aligned}
$$

Take expectation over the remaining randomness:

$$
\mathbb{E}\left[\|\widetilde{\Delta}^q\|_{\mathrm{F}}^2\right] = \mathbb{E}\left[\mathbb{E}_\xi\left[\|\widetilde{\Delta}^q\|_{\mathrm{F}}^2\right]\right] \leq s\sigma^2 \sum_{i=1}^m \mathbb{E}\left[\mathbb{1}_{\{i\in\mathcal{A}^q\}}\right] \sum_{p=-1}^{q-1} \mathbb{E}\left[\mathbb{1}_{\{\tau_i(t)=p\}}\right] (q-p)^2 \leq \frac{2ms\sigma^2}{\delta^2}
$$

Therefore,

$$
\frac{1}{mT} \sum_{i=1}^m \sum_{t=0}^{T-1} \mathbb{E}\left[(\mathrm{I})\right] \leq \frac{s\rho}{(1-\rho)} \left(\frac{2}{\delta^2}\right) \sigma^2.
$$

Bounding $\mathbb{E}\left[(\mathrm{II})\right]$

$$
\begin{aligned}
&\mathbb{E}\left[(\mathrm{II})\right] = \mathbb{E}\left[\|\sum_{q=0}^{t-1} \Delta^q \left(\prod_{\ell=q}^{t-1} W^{(\ell)} - \mathbf{J} \right) \|_{\mathrm{F}}^2 \right] \\
&= \sum_{q=0}^{t-1} \mathbb{E}\left[\|\Delta^q \left(\prod_{\ell=q}^{t-1} W^{(\ell)} - \mathbf{J} \right) \|_{\mathrm{F}}^2 \right] + \sum_{q=0}^{t-1} \sum_{p=0, p\neq q}^{t-1} \mathbb{E}\left[\left\langle \Delta^p \left(\prod_{\ell=p}^{t-1} W^{(\ell)} - \mathbf{J} \right), \Delta^q \left(\prod_{\ell=q}^{t-1} W^{(\ell)} - \mathbf{J} \right) \right\rangle \right] \\
&\leq \sum_{q=0}^{t-1} \rho^{t-q} \mathbb{E}\left[\|\Delta^q\|_{\mathrm{F}}^2\right] + \sum_{q=0}^{t-1} \sum_{p=0, p\neq q}^{t-1} \mathbb{E}\left[\|\Delta^p \left(\prod_{\ell=p}^{t-1} W^{(\ell)} - \mathbf{J} \right) \|_{\mathrm{F}} \|\Delta^q \left(\prod_{\ell=q}^{t-1} W^{(\ell)} - \mathbf{J} \right) \|_{\mathrm{F}} \right] \\
&\leq \sum_{q=0}^{t-1} \rho^{t-q} \mathbb{E}\left[\|\Delta^q\|_{\mathrm{F}}^2\right] + \sum_{q=0}^{t-1} \sum_{p=0, p\neq q}^{t-1} \mathbb{E}\left[\frac{\rho^{t-p}}{2\epsilon} \|\Delta^p\|_{\mathrm{F}}^2 + \frac{\epsilon\rho^{t-q}}{2} \|\Delta^q\|_{\mathrm{F}}^2 \right],
\end{aligned}
$$

Next, we bound the second term, choose $\epsilon = \rho^{\frac{q-p}{2}}$,

$$\begin{aligned}
&\sum_{q=0}^{t-1}\sum_{p=0,p\neq q}^{t-1}\frac{\sqrt{\rho}^{2t-p-q}}{2}\mathbb{E}\left[\|\Delta^p\|_{\mathrm{F}}^2+\|\Delta^q\|_{\mathrm{F}}^2\right] \leq \sum_{q=0}^{t-1}\sum_{p=0}^{t-1}\frac{\sqrt{\rho}^{2t-p-q}}{2}\mathbb{E}\left[\|\Delta^p\|_{\mathrm{F}}^2+\|\Delta^q\|_{\mathrm{F}}^2\right] \\
&= \sum_{p=0}^{t-1}\frac{\sqrt{\rho}^{t-p}}{2}\mathbb{E}\left[\|\Delta^p\|_{\mathrm{F}}^2\right]\sum_{q=0}^{t-1}\sqrt{\rho}^{t-q} + \sum_{q=0}^{t-1}\frac{\sqrt{\rho}^{t-q}}{2}\mathbb{E}\left[\|\Delta^q\|_{\mathrm{F}}^2\right]\sum_{p=0}^{t-1}\sqrt{\rho}^{t-p} \\
&= \frac{\sqrt{\rho}-\sqrt{\rho}^{t+1}}{1-\sqrt{\rho}}\sum_{q=0}^{t-1}\sqrt{\rho}^{t-q}\mathbb{E}\left[\|\Delta^q\|_{\mathrm{F}}^2\right]. \qquad (28)
\end{aligned}$$

Plugging the upper bound in (28) into (27), we get

$$\begin{aligned}
\mathbb{E}\left[(\mathrm{II})\right] &\leq \sum_{q=0}^{t-1}\left[\sqrt{\rho}^{t-q}+\frac{\sqrt{\rho}-\sqrt{\rho}^{t+1}}{1-\sqrt{\rho}}\right]\sqrt{\rho}^{t-q}\mathbb{E}\left[\|\Delta^q\|_{\mathrm{F}}^2\right] \overset{(b)}{\leq} \sum_{q=0}^{t-1}\left[\frac{\sqrt{\rho}+\sqrt{\rho}}{1-\sqrt{\rho}}\right]\sqrt{\rho}^{t-q}\mathbb{E}\left[\|\Delta^q\|_{\mathrm{F}}^2\right] \\
&\leq \frac{2\sqrt{\rho}}{1-\sqrt{\rho}}\sum_{q=0}^{t-1}\sqrt{\rho}^{t-q}\mathbb{E}\left[\|\Delta^q\|_{\mathrm{F}}^2\right], \qquad (29)
\end{aligned}$$

where inequality (b) follows because that $\sqrt{\rho}^{t-q} \leq \sqrt{\rho}$ for any $q \leq t-1$, and that $\sqrt{\rho}^{t+1} \geq 0$. It remains to bound $\mathbb{E}\left[\|\Delta^q\|_{\mathrm{F}}^2\right]$. Take expectation with respect to randomness in stochastic gradients:

$$\begin{aligned}
\mathbb{E}_\xi\left[\|\Delta^q\|_{\mathrm{F}}^2\right] \leq\ & 4\eta_l^2 s^3 L^2\sum_{i=1}^{m}\sum_{p=-1}^{q-1}\mathbb{1}_{\{\tau_i(q)=p\}}(q-p)^2\sigma^2 \\
&+ 16\eta_l^2 s^4 L^2\sum_{i=1}^{m}\sum_{p=-1}^{q-1}\mathbb{1}_{\{\tau_i(q)=p\}}(q-p)^2\left\|\nabla F_i(\boldsymbol{x}_i^q)\right\|_2^2,
\end{aligned}$$

where the inequality holds due to Lemma 5. Next, we take expectation over the remaining randomness and plug back into (29):

$$\begin{aligned}
&\mathbb{E}\left[(\mathrm{II})\right] \leq \frac{2\sqrt{\rho}}{1-\sqrt{\rho}}\sum_{q=0}^{t-1}\sqrt{\rho}^{t-q}\mathbb{E}\left[\|\Delta^q\|_{\mathrm{F}}^2\right] \\
&\leq \frac{8\rho}{\left(1-\sqrt{\rho}\right)^2}\left(\frac{2}{\delta^2}\right)\eta_l^2 s^3 L^2 m\sigma^2 \\
&\quad + \frac{32\sqrt{\rho}}{1-\sqrt{\rho}}\left(\frac{2}{\delta^2}\right)\eta_l^2 s^4 L^2\sum_{i=1}^{m}\sum_{q=0}^{t-1}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^q)\right\|_2^2\right]\sum_{k=1}^{T-1-t}\sqrt{\rho}^k \\
&\leq \frac{8\rho}{\left(1-\sqrt{\rho}\right)^2}\left(\frac{2}{\delta^2}\right)\eta_l^2 s^3 L^2 m\sigma^2 + \frac{32\rho}{\left(1-\sqrt{\rho}\right)^2}\left(\frac{2}{\delta^2}\right)\eta_l^2 s^4 L^2\sum_{i=1}^{m}\sum_{q=0}^{t-1}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^q)\right\|_2^2\right],
\end{aligned}$$

where the last inequality holds because of re-index and grouping. Therefore,

$$\begin{aligned}
&\frac{1}{mT}\sum_{t=1}^{T-1}\mathbb{E}\left[(\mathrm{II})\right] \leq \frac{8\rho}{\left(1-\sqrt{\rho}\right)^2}\left(\frac{2}{\delta^2}\right)\eta_l^2 s^3 L^2\sigma^2 \\
&\quad + \frac{32\rho}{\left(1-\sqrt{\rho}\right)^2}\left(\frac{2}{\delta^2}\right)\eta_l^2 s^4 L^2\frac{1}{T}\sum_{t=1}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2\right] \\
&\leq \frac{8\rho}{\left(1-\sqrt{\rho}\right)^2}\left(\frac{2}{\delta^2}\right)\eta_l^2 s^3 L^2\sigma^2 + \frac{64\rho}{\left(1-\sqrt{\rho}\right)^2}\left(\frac{2}{\delta^2}\right)\eta_l^2 s^4 L^4\frac{1}{T}\sum_{t=1}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2\right] \\
&\quad + \frac{64\rho}{\left(1-\sqrt{\rho}\right)^2}\left(\frac{2}{\delta^2}\right)\eta_l^2 s^4 L^2\frac{1}{T}\sum_{t=1}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right]
\end{aligned}$$

Bounding $\mathbb{E}\left[(\text{III})\right]$ Use a similar trick as in bounding $\mathbb{E}\left[(\text{II})\right]$, and we get

$$\mathbb{E}\left[(\text{III})\right] = \mathbb{E}\left[\|\sum_{q=0}^{t-1}\nabla \boldsymbol{F}_{\boldsymbol{x}}^q\left(\prod_{\ell=q}^{t-1}W^{(\ell)}-\mathbf{J}\right)\|_{\mathrm{F}}^2\right] \le \frac{2\sqrt{\rho}}{1-\sqrt{\rho}}\sum_{q=0}^{t-1}\sqrt{\rho}^{t-q}\mathbb{E}\left[\|\nabla \boldsymbol{F}_{\boldsymbol{x}}^q\|_{\mathrm{F}}^2\right],$$

so that

$$\begin{aligned}
\frac{1}{mT}\sum_{t=0}^{T-1}\mathbb{E}\left[(\text{III})\right] &\le \frac{2\sqrt{\rho}}{mT\left(1-\sqrt{\rho}\right)}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla \boldsymbol{F}_{\boldsymbol{x}}^t\|_{\mathrm{F}}^2\right]\sum_{q=1}^{T-1-t}\sqrt{\rho}^q \\
&\le \frac{2\rho}{\left(1-\sqrt{\rho}\right)^2}\frac{1}{mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2\right] \\
&\le \frac{4\rho L^2}{\left(1-\sqrt{\rho}\right)^2}\frac{1}{mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2\right] + \frac{4\rho}{\left(1-\sqrt{\rho}\right)^2}\frac{1}{mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right].
\end{aligned}$$

Putting them together

$$\begin{aligned}
\frac{1}{mT}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\boldsymbol{Z}^{(t)}\left(\mathbf{I}-\mathbf{J}\right)\|_{\mathrm{F}}^2\right] &\le \frac{s\rho\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2}\left(\frac{2}{\delta^2}\right)\left(1+16\eta_l^2s^2L^2\right)\sigma^2 \\
&+ \frac{8\rho s^2L^2\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2}\left(1+16\eta_l^2s^2L^2\left(\frac{2}{\delta^2}\right)\right)\frac{1}{T}\sum_{t=1}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2\right] \\
&+ \frac{8\rho s^2\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2}\left(1+16\eta_l^2s^2L^2\left(\frac{2}{\delta^2}\right)\right)\frac{1}{T}\sum_{t=1}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right].
\end{aligned}$$

Plug in Proposition 2.

$$\begin{aligned}
\frac{1}{mT}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\boldsymbol{Z}^{(t)}\left(\mathbf{I}-\mathbf{J}\right)\|_{\mathrm{F}}^2\right] &\le \frac{s\rho\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2}\left(\frac{2}{\delta^2}\right)\left(1+20\eta_l^2s^2L^2\right)\sigma^2 \\
&+ \frac{8\rho s^2\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2}\left(1+16\eta_l^2s^2L^2\left(\frac{2}{\delta^2}\right)\right)\left(1+\eta_l^2\eta_g^2s^2L^2\left(\frac{2}{\delta^2}\right)\right)\frac{1}{T}\sum_{t=1}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right] \\
&\le \frac{1.05\rho s\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2}\left(\frac{2}{\delta^2}\right)\sigma^2 + \frac{9\rho s^2\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2}\frac{1}{T}\sum_{t=1}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right],
\end{aligned}$$

where the last inequality holds because $\eta_l \le \delta/(20sL)$ and $\eta_l\eta_g \le \delta/(10sL)$. Next, plug in Proposition 3.

$$\begin{aligned}
\frac{1}{mT}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\boldsymbol{Z}^{(t)}\left(\mathbf{I}-\mathbf{J}\right)\|_{\mathrm{F}}^2\right] &\le \frac{1.05\rho s\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2}\left(\frac{2}{\delta^2}\right)\sigma^2 + \frac{27\rho s^2\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2}\zeta^2 \\
&+ \frac{27\rho s^2\eta_l^2\eta_g^2\left(\beta^2+1\right)}{(1-\sqrt{\rho})^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] + \frac{27\rho s^2L^2\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2}\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right].
\end{aligned}$$

It follows that

$$\begin{aligned}
\frac{1}{mT}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\boldsymbol{Z}^{(t)}\left(\mathbf{I}-\mathbf{J}\right)\|_{\mathrm{F}}^2\right] &\le \frac{3\rho s\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2\delta^2}\sigma^2 \\
&+ \frac{40\rho s^2\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2}\zeta^2 \\
&+ \frac{40\rho s^2\eta_l^2\eta_g^2\left(\beta^2+1\right)}{(1-\sqrt{\rho})^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right].
\end{aligned}$$

which is due to the fact that $\eta_l\eta_g \le \frac{1-\sqrt{\rho}}{10sL(\sqrt{\rho}+1)}$. □

F.6 Spectral norm upper bound (Lemma 4)

Lemma 4 adapts from [57], we present its proof here for completeness.

Proof of Lemma 4. For ease of exposition, in this proof we drop time index t. We first get the explicit expression for $\mathbb{E}\left[W_{jj'}^2 \mid \mathcal{A} \neq \emptyset\right]$. Suppose that $\mathcal{A} \neq \emptyset$. We have

$$W_{jj'}^2 = \sum_{k=1}^{m} W_{jk}W_{j'k} = W_{jj}W_{j'j} + W_{jj'}W_{j'j'} + \sum_{k\in[m]\setminus\{j,j'\}} W_{jk}W_{j'k}.$$

When $k \neq j$ and $k \neq j'$, we have

$$W_{jk}W_{j'k} = \frac{1}{|\mathcal{A}|^2}\mathbb{1}_{\{j\in\mathcal{A}\}}\mathbb{1}_{\{j'\in\mathcal{A}\}}\mathbb{1}_{\{k\in\mathcal{A}\}}.$$

In addition, we have $W_{jj}W_{j'j} = \frac{1}{|\mathcal{A}|^2}\mathbb{1}_{\{j\in\mathcal{A}\}}\mathbb{1}_{\{j'\in\mathcal{A}\}}$, and $W_{j'j'}W_{jj'} = \frac{1}{|\mathcal{A}|^2}\mathbb{1}_{\{j\in\mathcal{A}\}}\mathbb{1}_{\{j'\in\mathcal{A}\}}$. Thus,

- For $j \neq j'$, we have

$$W_{jj'}^2 = \sum_{k=1}^{m} W_{jk}W_{j'k} = \frac{1}{|\mathcal{A}|}\mathbb{1}_{\{j\in\mathcal{A}\}}\mathbb{1}_{\{j'\in\mathcal{A}\}};$$

- For $j = j'$, we have

$$W_{jj}^2 = \frac{1}{|\mathcal{A}|}\mathbb{1}_{\{j\in\mathcal{A}\}} + \left(1 - \mathbb{1}_{\{j\in\mathcal{A}\}}\right).$$

In the special case where $\mathcal{A} = \emptyset$, we simply have $W = \mathbf{I}$ by the algorithmic clauses. Therefore, $\mathbb{E}\left[W_{jj'} \mid \mathcal{A} = \emptyset\right] \geq 0$ holds for any pair of $j, j' \in [m]$. It follows, by the law of total expectation and for all $j, j' \in [m]$, that

$$\begin{aligned}\mathbb{E}\left[W_{jj'}\right] &= \mathbb{E}\left[W_{jj'} \mid \mathcal{A} = \emptyset\right]\mathbb{P}\left\{\mathcal{A} = \emptyset\right\} + \mathbb{E}\left[W_{jj'} \mid \mathcal{A} \neq \emptyset\right]\mathbb{P}\left\{\mathcal{A} \neq \emptyset\right\}\\ &\geq \mathbb{E}\left[W_{jj'} \mid \mathcal{A} \neq \emptyset\right]\mathbb{P}\left\{\mathcal{A} \neq \emptyset\right\}.\end{aligned}$$

- For $j \neq j'$, it holds that

$$\mathbb{E}\left[W_{jj'}^2 \mid \mathcal{A} \neq \emptyset\right] = \mathbb{E}\left[\frac{1}{|\mathcal{A}|}\mathbb{1}_{\{j\in\mathcal{A}\}}\mathbb{1}_{\{j'\in\mathcal{A}\}}\middle|\mathcal{A} \neq \emptyset\right] \overset{(a)}{\geq} \mathbb{E}\left[\frac{1}{m}\mathbb{1}_{\{j\in\mathcal{A}\}}\mathbb{1}_{\{j'\in\mathcal{A}\}}\middle|\mathcal{A} \neq \emptyset\right] = \frac{p_j p_{j'}}{m} \geq \frac{\delta^2}{m},$$

  where inequality (a) holds because $|\mathcal{A}| \leq m$;
- For $j = j'$, it holds that

$$\begin{aligned}\mathbb{E}\left[W_{jj}^2 \mid \mathcal{A} \neq \emptyset\right] &= \mathbb{E}\left[\frac{1}{|\mathcal{A}|}\mathbb{1}_{\{j\in\mathcal{A}\}} + \left(1 - \mathbb{1}_{\{j\in\mathcal{A}\}}\right)\middle|\mathcal{A} \neq \emptyset\right]\\ &\geq \mathbb{E}\left[\frac{1}{m}\left[\mathbb{1}_{\{j\in\mathcal{A}\}} + \left(1 - \mathbb{1}_{\{j\in\mathcal{A}\}}\right)\right]\middle|\mathcal{A} \neq \emptyset\right] = \frac{1}{m} \geq \frac{\delta^2}{m}.\end{aligned}$$

Recall that $M = \mathbb{E}\left[W^2\right]$. Next, we show that each element of M is lower bounded.

$$M_{jj'} \geq \mathbb{E}\left[W_{jj'}^2 \mid \mathcal{A} \neq \emptyset\right]\mathbb{P}\left\{\mathcal{A} \neq \emptyset\right\} \geq \frac{\delta^2}{m}\left[1 - (1-\delta)^m\right].$$

We note that $\rho(t) = \lambda_2(M)$, where λ_2 is the second largest eigenvalue of matrix M. A Markov chain with M as the transition matrix is ergodic as the chain is (1) *irreducible*: $M_{jj'} \geq \frac{\delta^2}{m}\left[1 - (1-c)^m\right] > 0$ for $j, j' \in [m]$ and (2) *aperiodic* (it has self-loops). In addition, W matrix is by definition doubly-stochastic. Hence, M has a uniform stationary distribution $\pi = \mathbb{1}^\top/m$. Furthermore, the irreducible Markov chain is reversible since it holds for all the states that $\pi_i M_{ij} = \pi_j M_{ji}$. The conductance Φ of a reversible Markov chain [18] with a transition matrix M can be bounded by

$$\Phi(M) = \min_{\sum_{i\in\mathcal{S}}\pi_i \leq \frac{1}{2}} \frac{\sum_{i\in\mathcal{S}, j\notin\mathcal{S}} \pi_i M_{ij}}{\sum_{i\in\mathcal{S}} \pi_i} \geq \frac{\left(\frac{\delta}{m}\right)^2\left[1 - (1-\delta)^m\right]|\mathcal{S}|\left|\bar{\mathcal{S}}\right|}{\frac{|\mathcal{S}|}{m}} = \frac{\delta^2\left[1 - (1-\delta)^m\right]}{m}\left|\bar{\mathcal{S}}\right|,$$

where $\left|\bar{\mathcal{S}}\right| = m - |\mathcal{S}| \geq \frac{m}{2}$. From Cheeger's inequality, we know that $\frac{1-\lambda_2}{2} \leq \Phi(M) \leq \sqrt{2\left(1-\lambda_2\right)}$. Finally, we have

$$\Phi(M) \geq \frac{\delta^2\left[1 - (1-\delta)^m\right]}{m}\left|\bar{\mathcal{S}}\right| \geq \frac{\delta^2\left[1 - (1-\delta)^m\right]}{2}.$$

Thus, $\rho(t) = \lambda_2 \leq 1 - \frac{\Phi^2(M)}{2} \leq 1 - \frac{\delta^4\left[1-(1-\delta)^m\right]^2}{8}$. $\square$

G Convergence Error of $\bar{z}^t$ (Theorem 1)

In the sequel, we recall and assume the following learning rate conditions in (11):

$$\eta_l \eta_g \leq \frac{\left(1-\sqrt{\rho}\right)\delta}{80s(L+1)\left(\sqrt{\rho}+1\right)\sqrt{(\beta^2+1)\left(1+L^2\right)}};\ \eta_l \leq \frac{\delta}{200sL\sqrt{(\beta^2+1)\left(1+L^2\right)}}.$$

Recall that $\delta_{\max} \triangleq \max_{i\in[m],t\in[T]} p_i^t$ and $F^\star \triangleq \min_{\boldsymbol{x}} F(\boldsymbol{x})$.

Proof of Theorem 1. Take expectation over all the randomness, plug in Lemma 6 and Proposition 2. By telescoping sum, it holds that

$$\begin{aligned}
\frac{\mathbb{E}\left[F^\star - F(\bar{\boldsymbol{z}}^0)\right]}{T} &\leq -\frac{\eta_l\eta_g s}{4}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] + \frac{2\eta_l^2\eta_g^2 sL\delta_{\max}\sigma^2}{m^2T}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{E}\left[\mathbb{1}_{\{\tau_i(t)=p\}}\right](t-p)^2 \\
&+ \frac{9\eta_g\eta_l^3 s^2L^2\sigma^2}{mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{E}\left[\mathbb{1}_{\{\tau_i(t)=p\}}\right](t-p)^2 \\
&+ 2.2\eta_l\eta_g sL^2\frac{1}{mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2\right] \qquad (30) \\
&+ \frac{\eta_l\eta_g sL^2}{2mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t - \bar{\boldsymbol{z}}^t\right\|_2^2\right] \qquad (31) \\
&+ \frac{35\eta_g\eta_l^3 s^3L^2}{mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{E}\left[\mathbb{1}_{\{\tau_i(t)=p\}}\right](t-p)^2\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2\right]. \qquad (32)
\end{aligned}$$

Next, we bound (30), (31) and (32), respectively. First, we show that

$$\begin{aligned}
&\frac{1}{mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right] \\
&\leq 3\zeta^2 + 3\left(\beta^2+1\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] + \frac{3L^2}{mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t - \bar{\boldsymbol{z}}^t\right\|_2^2\right] \\
&\leq 3\left[1+\frac{40\rho s^2\eta_l^2\eta_g^2L^2}{(1-\sqrt{\rho})^2}\right]\zeta^2 + 3\left(\beta^2+1\right)\left[1+\frac{40\rho s^2\eta_l^2\eta_g^2L^2}{(1-\sqrt{\rho})^2}\right]\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] + \frac{9\rho s\eta_l^2\eta_g^2L^2}{(1-\sqrt{\rho})^2\delta^2}\sigma^2, \qquad (33)
\end{aligned}$$

where the last inequality follows from Lemma 6.

For (30), we have

$$\begin{aligned}
2.2\eta_l\eta_g sL^2\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2\right] &\leq \frac{4.4\eta_l^3\eta_g^3s^3L^2}{\delta^2}\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right] \\
&\leq \frac{s^2\eta_l^3\eta_g^3L^2}{2\delta^2}\sigma^2 + \frac{14\eta_l^3\eta_g^3s^3L^2}{\delta^2}\left(1+\frac{40\eta_l^2\eta_g^2\rho s^2L^2}{(1-\sqrt{\rho})^2}\right)\zeta^2 \\
&+ \frac{14\eta_l^3\eta_g^3s^3L^2}{\delta^2}\left[\left(\beta^2+1\right)+\frac{40\eta_l^2\eta_g^2\rho s^2L^2}{(1-\sqrt{\rho})^2}\right]\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right],
\end{aligned}$$

where the last inequality holds due to (33). For (31), we similarly have

$$\begin{aligned}
\frac{\eta_l\eta_g sL^2}{2mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t - \bar{\boldsymbol{z}}^t\right\|_2^2\right] &\leq \frac{1.5\rho s^2\eta_l^3\eta_g^3L^2}{(1-\sqrt{\rho})^2\delta^2}\sigma^2 + \frac{20\rho s^3\eta_l^3\eta_g^3L^2}{(1-\sqrt{\rho})^2}\zeta^2 \\
&+ \frac{20\rho s^3\eta_l^3\eta_g^3L^2\left(\beta^2+1\right)}{(1-\sqrt{\rho})^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right].
\end{aligned}$$

For (32), we have

$$
\begin{aligned}
&35\eta_g\eta_l^3 s^3 L^2 \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{E}\left[\mathbb{1}_{\{\tau_i(t)=p\}}\right](t-p)^2\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2\right]\\
&\le \frac{70\eta_g\eta_l^3 s^3 L^2}{mT\delta^2}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2\right]\\
&\le \frac{140\eta_g\eta_l^3 s^3 L^4}{mT\delta^2}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2\right]+\frac{140\eta_g\eta_l^3 s^3 L^2}{mT\delta^2}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right]\\
&\le \left(1+\frac{2\eta_l^2\eta_g^2 s^2L^2}{\delta^2}\right)\left(\frac{2}{\delta^2}\right)\frac{70\eta_g\eta_l^3 s^3 L^2}{mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right]\\
&\overset{(a)}{\le} \left(\frac{2}{\delta^2}\right)\frac{71\eta_g\eta_l^3 s^3 L^2}{mT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right]\\
&\overset{(b)}{\le} \frac{426\eta_g\eta_l^3 s^3 L^2}{\delta^2}\left[1+\frac{40\rho s^2\eta_l^2\eta_g^2L^2}{(1-\sqrt{\rho})^2}\right]\zeta^2+\frac{426\eta_g\eta_l^3 s^3 L^2}{\delta^2}\left(\beta^2+1\right)\left[1+\frac{40\rho s^2\eta_l^2\eta_g^2L^2}{(1-\sqrt{\rho})^2}\right]\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\\
&\quad+\frac{\eta_g\eta_l^3 s^3 L^2\sigma^2}{2\delta^2},
\end{aligned}
$$

where inequality (a) holds because of (11), inequality (b) holds because of (33).

Putting (30), (31) and (32) together and plugging them back into the telescoping sum, it holds that

$$
\begin{aligned}
&\frac{\mathbb{E}\left[F^\star-F(\bar{\boldsymbol{z}}^0)\right]}{T}\\
&\le -\left(\frac{\eta_l\eta_g s}{4}-\frac{14\left(\beta^2+1\right)\eta_l^3\eta_g^3 s^3L^2\left(1+L^2\right)}{\delta^2}-\frac{20\rho s^3\eta_l^3\eta_g^3L^2\left(\beta^2+1\right)}{(1-\sqrt{\rho})^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\\
&\quad-\left(-\frac{426\eta_g\eta_l^3 s^3L^2\left(\beta^2+1\right)\left(1+L^2\right)}{\delta^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\\
&\quad+\frac{4\eta_l^2\eta_g^2 sL\delta_{\max}\sigma^2}{m\delta^2}+\left(\frac{\eta_l^3\eta_g^3 s^2L^2\sigma^2}{2\delta^2}+\frac{1.5\rho s^2\eta_l^3\eta_g^3L^2}{(1-\sqrt{\rho})^2\delta^2}\sigma^2+\frac{\eta_g\eta_l^3 s^3L^2\sigma^2}{2\delta^2}\right)\\
&\quad+\frac{15\eta_l^3\eta_g^3 s^3L^2\zeta^2}{\delta^2}+\frac{20\rho s^3\eta_l^3\eta_g^3L^2}{(1-\sqrt{\rho})^2}\zeta^2+\frac{430\eta_g\eta_l^3 s^3L^2\zeta^2}{\delta^2}\\
&\le -\frac{\eta_l\eta_g s}{6}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\\
&\quad+\frac{4\eta_l^2\eta_g^2 sL\delta_{\max}\sigma^2}{m\delta^2}+\left(\frac{\eta_l^3\eta_g^3 s^2L^2\sigma^2}{2\delta^2}+\frac{1.5\rho s^2\eta_l^3\eta_g^3L^2}{(1-\sqrt{\rho})^2\delta^2}\sigma^2+\frac{\eta_g\eta_l^3 s^3L^2\sigma^2}{2\delta^2}\right)\\
&\quad+\frac{15\eta_l^3\eta_g^3 s^3L^2\zeta^2}{\delta^2}+\frac{20\rho s^3\eta_l^3\eta_g^3L^2}{(1-\sqrt{\rho})^2}\zeta^2+\frac{430\eta_g\eta_l^3 s^3L^2\zeta^2}{\delta^2},
\end{aligned}
$$

where the last inequality holds because of (11).

Combining the above and rearranging the terms, we get

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] \le \frac{6\left(F(\bar{\boldsymbol{z}}^0)-F^\star\right)}{\eta_l\eta_g sT} \\
&+ \frac{24\eta_l\eta_g L\delta_{\max}\sigma^2}{m\delta^2} + \left(\frac{3\eta_l^2\eta_g^2 sL^2\sigma^2}{\delta^2} + \frac{9\rho s\eta_l^2\eta_g^2L^2}{(1-\sqrt{\rho})^2\delta^2}\sigma^2 + \frac{3\eta_l^2 s^2L^2\sigma^2}{\delta^2}\right) \\
&+ \frac{90\eta_l^2\eta_g^2 s^2L^2\zeta^2}{\delta^2} + \frac{120\rho s^2\eta_l^2\eta_g^2L^2}{(1-\sqrt{\rho})^2}\zeta^2 + \frac{2580\eta_l^2 s^2L^2\zeta^2}{\delta^2} \\
&\le \frac{6\left(F(\bar{\boldsymbol{z}}^0)-F^\star\right)}{\eta_l\eta_g sT} + \frac{24\eta_l\eta_g L\delta_{\max}\sigma^2}{m\delta^2} + \frac{15\eta_l^2\eta_g^2 s^2L^2\sigma^2}{(1-\sqrt{\rho})^2\delta^2} + \frac{2800\eta_l^2\eta_g^2 s^2L^2\zeta^2}{\delta^2(1-\sqrt{\rho})^2},
\end{aligned}
$$

where the last inequality holds because $\rho < 1$. In terms of asymptotics, we have

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] \lesssim \frac{\left(F(\bar{\boldsymbol{z}}^0)-F^\star\right)}{\eta_l\eta_g sT} + \frac{\eta_l\eta_g L\sigma^2}{m}\frac{\delta_{\max}}{\delta^2} + \eta_l^2\eta_g^2 s^2L^2\left(\frac{\sigma^2+\zeta^2}{\delta^2\left(1-\sqrt{\rho}\right)^2}\right),
$$

where we use the convention that $\eta_g \ge 1$ for ease of presentation. $\square$

H Convergence Rate of $\bar{x}^t$ (Corollary 1)

H.1 Convergence error of Algorithm 1

Corollary 2 (Convergence error of $\boldsymbol{x}_i^t$). *Suppose learning rates conditions in* (11) *are met for η_l and η_g, and Assumptions 1, 2, 3 and 4 hold for $T \geq 1$, it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right] \lesssim \frac{\left(F(\bar{\boldsymbol{x}}^0)-F^\star\right)}{\eta_l\eta_g sT} + \frac{\eta_l\eta_g L\sigma^2}{m}\frac{\delta_{\max}}{\delta^2} + \eta_l^2\eta_g^2 s^2 L^2\left(\frac{\sigma^2+\zeta^2}{\delta^2\left(1-\sqrt{\rho}\right)^2}\right),$$

Proof of Corollary 2.

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right] &\leq \frac{3}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)-\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] + \frac{3}{2T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\\
&\overset{(a)}{\leq} \frac{3L^2}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{\boldsymbol{x}}^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right] + \frac{3}{2T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\\
&\overset{(b)}{\leq} \frac{3L^2}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2\right] + \frac{3}{2T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\\
&\leq 3\left(\frac{2}{\delta^2}\right)\frac{\eta_l^2\eta_g^2 s^2L^2}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right] + \frac{3}{2T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right],
\end{aligned}$$

where inequality (a) follows from Appendix D 2, inequality (b) follows from Assumption 2.

Further plug in Proposition 3,

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right] &\leq \frac{3}{2T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] + 9\eta_l^2\eta_g^2s^2L^2\left(\frac{2}{\delta^2}\right)\left(\beta^2+1\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\\
&\quad + 9\eta_l^2\eta_g^2s^2L^4\left(\frac{2}{\delta^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right] + 9\eta_l^2\eta_g^2s^2L^2\left(\frac{2}{\delta^2}\right)\zeta^2.
\end{aligned}$$

Finally, plug in Lemma 6.

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right] &\leq \left(\frac{3}{2}+9\eta_l^2\eta_g^2s^2L^2\left(\frac{2}{\delta^2}\right)\left(\beta^2+1\right)\frac{90}{80^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\\
&\quad + \frac{9\times 8}{80^2}\eta_l^2\eta_g^2sL^2\left(\frac{2}{\delta^2}\right)\sigma^2 + 9\eta_l^2\eta_g^2s^2L^2\left(\frac{2}{\delta^2}\right)\zeta^2 + \frac{9\times 90}{200^2}\eta_l^2\eta_g^2s^2L^2\zeta^2\\
&\leq \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] + \frac{sL^2\eta_l^2\eta_g^2}{\delta^2}\sigma^2 + \frac{9\eta_l^2\eta_g^2s^2L^2}{\delta^2}\zeta^2 + s^2L^2\eta_l^2\eta_g^2\zeta^2\\
&\leq \frac{12\left(F(\bar{\boldsymbol{z}}^0)-F^\star\right)}{\eta_l\eta_g sT} + \frac{48\eta_l\eta_g L\delta_{\max}\sigma^2}{m\delta^2} + \frac{31\eta_l^2\eta_g^2s^2L^2\sigma^2}{(1-\sqrt{\rho})^2\delta^2} + \frac{5600\eta_l^2\eta_g^2s^2L^2\zeta^2}{(1-\sqrt{\rho})^2\delta^2},
\end{aligned}$$

where the last inequality holds because $\rho < 1$. In terms of asymptotics, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right] \lesssim \frac{\left(F(\bar{\boldsymbol{x}}^0)-F^\star\right)}{\eta_l\eta_g sT} + \frac{\eta_l\eta_g L\sigma^2}{m}\frac{\delta_{\max}}{\delta^2} + \eta_l^2\eta_g^2s^2L^2\left(\frac{\sigma^2+\zeta^2}{\delta^2(1-\sqrt{\rho})^2}\right),$$

where we use the convention that $\eta_g \geq 1$ for ease of presentation. □

H.2 Convergence rate of Algorithm 1

Proof of Corollary 1. Choose step-size as $\eta_l = \frac{1}{\sqrt{T}sL}$, $\eta_g = \sqrt{s\delta m}$ such that learning rate conditions in (11) are met, it holds that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right] \lesssim \frac{L\left(F(\bar{\boldsymbol{x}}^0) - F^\star\right)}{\sqrt{s\delta m T}} + \frac{\delta_{\max}}{\delta^{\frac{3}{2}}\sqrt{smT}}\sigma^2 + \frac{sm}{T}\left(\frac{\sigma^2 + \zeta^2}{\delta(1-\sqrt{\rho})^2}\right).$$

□

I Additional Results and Interpretations

I.1 Consensus error of Algorithm 1

Corollary 3 (Consensus error of $\boldsymbol{x}_i^t$)**.** *Suppose learning rates conditions are met in (11) for η_l and η_g, and Assumptions 1, 2, 3 and 4 hold for $T \geq 1$, it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\bar{\boldsymbol{x}}^t\right\|_2^2\right] \lesssim \frac{\left(F(\bar{\boldsymbol{x}}^0)-F^\star\right)}{\eta_l\eta_g sT}+\frac{\eta_l\eta_g L\sigma^2}{m}\frac{\delta_{\max}}{\delta^2}$$
$$+\eta_l^2\eta_g^2 s^2L^2\left(\frac{\sigma^2+\zeta^2}{\delta^2}\right)\left[1+\frac{\rho}{\left(1-\sqrt{\rho}\right)^2}\right],$$

Proof of Corollary 3.

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\left\|\boldsymbol{x}_i^t-\bar{\boldsymbol{x}}^t\right\|_2^2=\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t+\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t+\bar{\boldsymbol{z}}^t-\bar{\boldsymbol{x}}^t\right\|_2^2$$
$$\overset{(a)}{\leq}\frac{1}{T}\sum_{t=0}^{T-1}\frac{3}{m}\sum_{i=1}^{m}\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2+\frac{1}{T}\sum_{t=0}^{T-1}\frac{3}{m}\sum_{i=1}^{m}\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2+\frac{1}{T}\sum_{t=0}^{T-1}3\left\|\bar{\boldsymbol{z}}^t-\bar{\boldsymbol{x}}^t\right\|_2^2$$
$$\overset{(b)}{\leq}\frac{1}{T}\sum_{t=0}^{T-1}\frac{3}{m}\sum_{i=1}^{m}\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2+\frac{1}{T}\sum_{t=0}^{T-1}\frac{3}{m}\sum_{i=1}^{m}\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2+\frac{1}{T}\sum_{t=0}^{T-1}\frac{3}{m}\sum_{i=1}^{m}\left\|\boldsymbol{z}_i^t-\boldsymbol{x}_i^t\right\|_2^2$$
$$=\frac{1}{T}\sum_{t=0}^{T-1}\frac{6}{m}\sum_{i=1}^{m}\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2+\frac{1}{T}\sum_{t=0}^{T-1}\frac{3}{m}\sum_{i=1}^{m}\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2,$$

where inequalities (a) and (b) follow from Jensen's inequality. Plug in Proposition 2 and take expectation over all the randomness, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\bar{\boldsymbol{x}}^t\right\|_2^2\right]\leq\frac{36\eta_l^2\eta_g^2s^2}{\delta^2}\left(\beta^2+1\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]$$
$$+\frac{36\eta_l^2\eta_g^2s^2}{\delta^2}\zeta^2+\left(3+\frac{36\eta_l^2\eta_g^2s^2L^2}{\delta^2}\right)\frac{1}{m}\sum_{i=1}^{m}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right]$$
$$\leq\frac{36\eta_l^2\eta_g^2s^2}{\delta^2}\left(\beta^2+1\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]+\frac{36\eta_l^2\eta_g^2s^2}{\delta^2}\zeta^2$$
$$+\frac{4}{m}\sum_{i=1}^{m}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right],$$

where the last inequality holds because of learning rate condition in (11). Next, plug in Lemma 6:

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\bar{\boldsymbol{x}}^t\right\|_2^2\right]\leq\frac{36\eta_l^2\eta_g^2s^2}{\delta^2}\left(\beta^2+1\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]$$
$$+\frac{36\eta_l^2\eta_g^2s^2}{\delta^2}\zeta^2+\frac{4}{m}\sum_{i=1}^{m}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right]$$
$$\leq\frac{36\eta_l^2\eta_g^2s^2}{\delta^2}\left(\beta^2+1\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]+\frac{1}{4T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]$$
$$+\frac{36\eta_l^2\eta_g^2s^2}{\delta^2}\zeta^2+\frac{12\rho s\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2\delta^2}\sigma^2+\frac{160\rho s^2\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2}\zeta^2$$
$$\leq\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]+\frac{12\rho s\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2\delta^2}\sigma^2+\frac{36\eta_l^2\eta_g^2s^2}{\delta^2}\zeta^2+\frac{160\rho s^2\eta_l^2\eta_g^2}{(1-\sqrt{\rho})^2}\zeta^2.$$

Finally, we plug in Theorem 1

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\bar{\boldsymbol{x}}^t\right\|_2^2\right]\leq\frac{3\left(F(\bar{\boldsymbol{x}}^0)-F^\star\right)}{\eta_l\eta_g sT}+\frac{12\eta_l\eta_g L\delta_{\max}\sigma^2}{m\delta^2}+\frac{28s^2\eta_l^2\eta_g^2L^2}{\delta^2(1-\sqrt{\rho})^2}\sigma^2+\frac{1600\eta_l^2\eta_g^2s^2L^2}{\delta^2(1-\sqrt{\rho})^2}\zeta^2,$$

where we use the fact that $\bar{\boldsymbol{z}}^0=\bar{\boldsymbol{x}}^0$ and $\rho<1$, and the convention that $\eta_g\geq 1$ and $L\geq 1$ for ease of presentation.

In terms of asymptotics, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\bar{\boldsymbol{x}}^t\right\|_2^2\right]\lesssim\frac{\left(F(\bar{\boldsymbol{x}}^0)-F^\star\right)}{\eta_l\eta_g sT}+\frac{\eta_l\eta_g L\sigma^2}{m}\frac{\delta_{\max}}{\delta^2}+\eta_l^2\eta_g^2s^2L^2\left(\frac{\sigma^2+\zeta^2}{\delta^2(1-\sqrt{\rho})^2}\right).$$

□

I.2 Orders of the asymptotic rates

From Theorem 1, Corollary 2, Corollary 3, it is easy to see from the theorem statements that they are all of the same asymptotic order, i.e.,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{\boldsymbol{x}}^t)\|_2^2]\asymp\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\boldsymbol{x}_i^t-\bar{\boldsymbol{x}}^t\|_2^2]\asymp\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{\boldsymbol{z}}^t)\|_2^2].$$

In addition, by applying learning rate conditions in (11) to Lemma 6 and Proposition 2, we can also see that

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\|_2^2]\asymp\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\|_2^2]\asymp\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{\boldsymbol{z}}^t)\|_2^2].$$

Therefore, we conclude that (12), (14) and (15) hold.

Table 6: Neural network architecture, loss function, learning rate scheduling, training steps and batch size specifications

Datasets	**SVHN**	**CIFAR-10**	**CINIC-10**
Neural network	CNN	CNN	CNN
Model architecture*	C(3,32) – **R** – **M** – C(32,32) – **R** – **M** – **L**(128) – **R** – **L**(10)	C(3,32) – **R** – **M** – C(32,32) – **R** – **M** – **L**(256) – **R** – **L**(64) – **R** – **L**(10)	C(3,32) – **R** – **M** – C(32,32) – **R** – **M** – **D** – **L**(512) – **R** – **D** – **L**(256) – **R** – **D** – **L**(10)
Loss function	Cross-entropy loss		
Local learning rate η_l scheduling	$\eta_l = \frac{\eta_0}{\sqrt{t/10+1}}$, where t denotes the global round.		
Number of local steps s	10		
Number of global rounds T	2000		
Batch size	128		

* **C**(# in-channel, # out-channel): a 2D convolution layer (kernel size 3, stride 1, padding 1); **R**: ReLU activation function; **M**: a 2D max-pool layer (kernel size 2, stride 2); **L**: (# outputs): a fully-connected linear layer; **D**: a dropout layer (probability 0.2).

J Numerical Experiments

J.1 Code

The code for reproducing our experiments is available at https://github.com/mingxiang12/FedAWE.

J.2 Experimental setups

Hardware and Software Setups.

- **Hardware.** The simulations are performed on a private cluster with 64 CPUs, 500 GB RAM and 8 NVIDIA A5000 GPU cards.
- **Software.** We code the experiments based on PyTorch 1.13.1 [39] and Python 3.7.16.

Neural Network and Hyper-parameter Specifications. Table 6 specifies details of the structures of the convolutional neural network and training. We initialize CNNs using the Kaiming initialization. The initial local learning rate η_0 and the global learning rate η_g are searched, based on the best performance after 500 global rounds, over two grids $\{0.1, 0.05, 0.01, 0.005, 0.001, 0.0005\}$ and $\{0.5, 1, 1.5, 5, 10, 50\}$, respectively. The results are presented in Table 7.

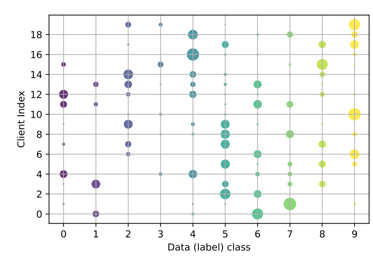


Figure 4: An example of data heterogeneity using $\mathsf{Dirichlet}(\alpha = 0.1)$ distribution with 20 clients. x-axis denotes the categories of images, while y-axis denotes the client index. The size of a circle refers to the proportion of pictures in a given class. The color of a circle distinguishes images with different categories.

The difference between `FedAvg` over active clients and `FedAvg` over all clients is that the latter counts the contributions of unavailable clients as **0**'s. We set $\beta = 0.001$ for `F3AST` [43], which is tuned over a grid of $\{0.1, 0.05, 0.01, 0.005, 0.001, 0.0005\}$. In addition, as recommended by [54], we choose $K = 50$ in `FedAU` without further specification. We train CNNs on all datasets for 2000 rounds. Fig. 3 adopts the same hyperparameter setups, yet with only 1000 training rounds.

Datasets and Data Heterogeneity.

Datasets. All the datasets we evaluate contain 10 classes of images. Some data enhancement tricks that are standard in training image classifiers are applied during training. Specifically, we apply

Table 7: Initial learning rate η_0 and global learning rate η_g

Algorithms	`FedAvg` *active*		`FedAvg` *known*		`FedAvg` *all*		`FedAU`		`F3AST`		`FedAWE`		`MIFA`		`FedVARP`	
SVHN	η_0	η_g	η_0	η_g	η_0	η_g	η_0	η_g	η_0	η_g	η_0	η_g	η_0	η_g	η_0	η_g
	0.05	1.0	0.1	1.0	0.05	1.0	0.05	1.0	0.05	1.0	0.1	1.0	0.05	1.0	0.05	1.0
CIFAR-10	η_0	η_g	η_0	η_g	η_0	η_g	η_0	η_g	η_0	η_g	η_0	η_g	η_0	η_g	η_0	η_g
	0.05	1.0	0.1	1.0	0.05	1.0	0.05	1.0	0.05	1.0	0.1	1.0	0.05	1.0	0.05	1.0
CINIC-10	η_0	η_g	η_0	η_g	η_0	η_g	η_0	η_g	η_0	η_g	η_0	η_g	η_0	η_g	η_0	η_g
	0.05	1.0	0.1	1.0	0.05	1.0	0.05	1.0	0.05	1.0	0.1	1.0	0.05	1.0	0.05	1.0

random cropping and gradient clipping with a max norm of 0.5 to all dataset trainings. Furthermore, random horizontal flipping is applied to CIFAR-10 and CINIC-10.

One full set of experiments takes about 6 hours on SVHN and CIFAR-10 datasets, while about 10 hours on CINIC-10 dataset.

- **SVHN [36].** The dataset contains 32×32 colored images of 10 different number digits. In total, there are 73257 train images and 26032 test images.
- **CIFAR-10 [25].** The dataset contains 32×32 colored images of 10 different objects. In total, there are 50000 train images and 10000 test images.
- **CINIC-10[11].** The dataset contains 32×32 colored images of 10 different objects. In total, there are 90000 train images and 90000 test images.

Data heterogeneity. Fig. 4 visualizes an example of 20 clients, the size of each circle corresponds to the relative proportion of images from a specific class. The larger the circle, the greater the share of images associated with that particular class. Moreover, α controls the heterogeneity of the data such that a greater α entails a more non-i.i.d. local data distribution and vice versa.

J.3 Non-stationary client unavailability dynamics

Client unavailability dynamics and visualizations. As specified in Section 7, we consider a total of four client unavailable dynamics in the form of $p_i^t = p_i \cdot f_i(t)$, where $p_i = \langle \nu_i, \phi \rangle$, $\nu_i \sim \mathsf{Dirichlet}(\alpha)$ and ϕ is the distribution to characterize the uneven contributions of each image class. In detail, each element $[\phi]_c$ is drawn from a uniform distribution $\mathsf{Uniform}(0, \mathbf{\Phi}_c)$. We set $\mathbf{\Phi}_c = 1$ for the first five image classes and $\mathbf{\Phi}_{c'} = 0.5$ for the remaining five image classes. Fig. 5 plots one resulting p_i's example, wherein p_i's are heterogeneous across clients.

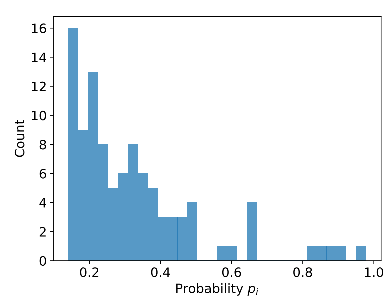
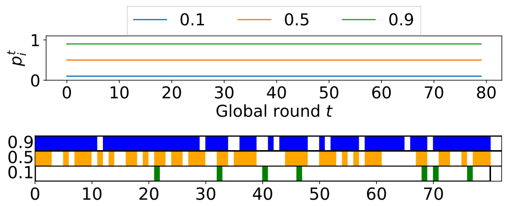


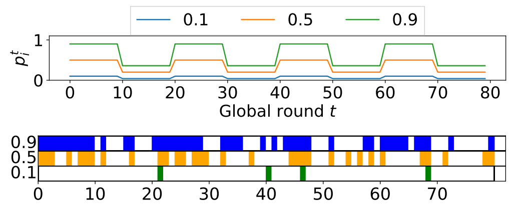
Figure 5: A histogram of one generated p_i's example with a total of $m = 100$ clients. It can be seen that the majority of p_i's are below 0.5.

Next, we formally introduce $f_i(t)$'s under each dynamic.

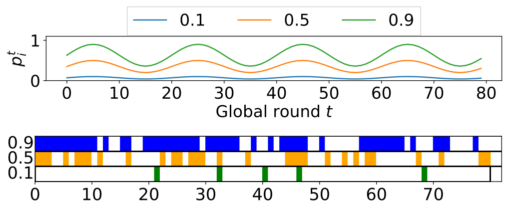
- Stationary: $f_i(t) \triangleq 1$;
- Non-stationary with staircase trajectory:
$$f_i(t) \triangleq \mathbb{1}_{\{t \in [t_0, t_0 + P/2)\}} + 0.4 \cdot \mathbb{1}_{\{t \in [t_0 + P/2, t_0 + P)\}},$$
where P defines a period, $t_0 \in \{0, P, 2P, 3P, \ldots\}$.
- Non-stationary with sine trajectory:
$$f_i(t) \triangleq \gamma \sin(2\pi / P \cdot t) + (1 - \gamma),$$
where γ signifies the degree of non-stationary.
- Non-stationary with interleaved sine trajectory:
$$f_i(t) \triangleq g_i(t) \cdot \mathbb{1}_{\{p_i \cdot g_i(t) \geq \delta_0\}},$$
where $g_i(t) \triangleq \gamma \sin(2\pi / P \cdot t) + (1 - \gamma)$ and $\delta_0 = 0.1$ defines a cutting-off lower bound. Specifically, δ_0 cuts off the sine curve and brings in a period of zero-valued probabilities. As different clients have different p_i's, the cut-off points are not synchronized among clients, leading to additional availability heterogeneity.



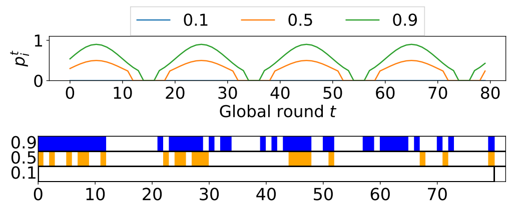
(a) Stationary



(b) **Non**-stationary with staircase trajectory



(c) **Non**-stationary with sine trajectory



(d) **Non**-stationary with interleaved sine trajectory

Figure 6: Examples of client unavailability with probabilistic trajectories. The first row in each sub-figure plots the probabilistic trajectory of each dynamics. The second row visualizes the simulated client availability by using a colored box to denote a client is available in that round. The y-axis is the base probability p_i to construct p_i^t. In other words, more blank space means that a client is more scarcely available. We simulate the cases where $p_i \in \{0.1, 0.5, 0.9\}$. The detailed construction of p_i^t can be found in Appendix J.3

Table 8: The first round to reach a targeted test accuracy under non-stationary of sine trajectory over 3 random seeds. We study the first round to reach $1/4$, $1/2$, $3/4$ and 1 of the best test accuracy of each dataset in Table 2, which is rounded up to the nearest 10% below for ease of presentation. In addition, we sample the mean of test accuracy every 20 global rounds to mitigate noisy progress. Some algorithms may never attain the targeted accuracy due to their inferior performance, where we use "–" as a placeholder.

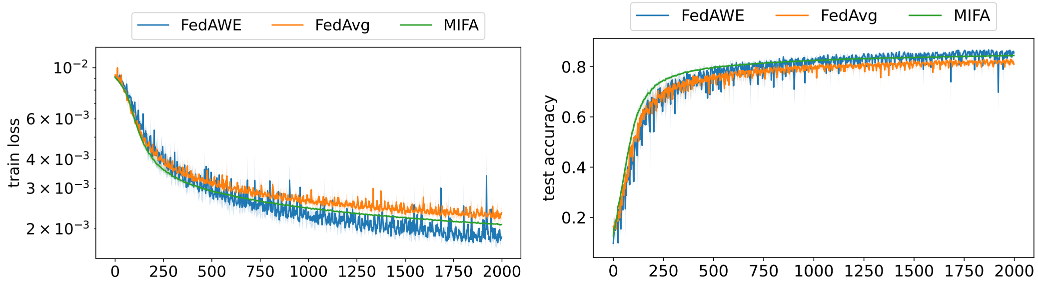
Datasets	**SVHN**				**CIFAR10**				**CINIC10**			
Quarters	**1/4**	**1/2**	**3/4**	**1**	**1/4**	**1/2**	**3/4**	**1**	**1/4**	**1/2**	**3/4**	**1**
Test accuracy	20%	40%	60%	80%	15%	30%	45%	60%	10%	20%	30%	40%
`FedAWE` (**ours**)	40	120	200	820	20	60	200	1360	0	20	120	540
`FedAvg` over *active* clients	20	80	160	900	10	20	120	1060	0	20	40	800
`FedAvg` over *all* clients	100	420	960	–	20	60	520	–	0	20	200	–
`FedAU`	60	100	160	840	10	20	100	960	0	20	80	460
`F3AST`	40	120	200	1080	20	40	160	1300	0	20	60	540
`FedAvg` with *known* p_i^t's	20	40	100	320	10	20	140	620	0	20	40	400
`MIFA` (*memory aided*)	20	80	140	600	10	20	80	700	0	20	40	240

We choose $\gamma = 0.3$ and $P = 20$ for all non-stationary dynamics. Next, we visualize the probability trajectories along with sampled client availability in Fig. 6. The plots confirm the intuition that interleaved dynamics is the most difficult one, e.g., no clients are available in the case of 0.1 therein.

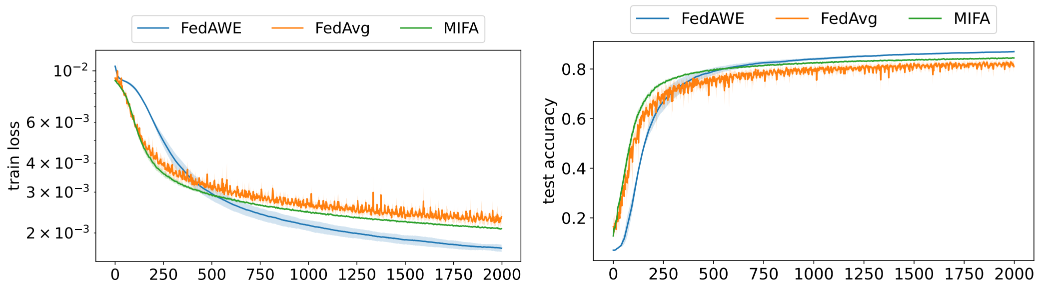
J.4 Additional results

Staleness studies. Table 8 illustrates the first round to reach a targeted test accuracy under non-stationary client availability with sine trajectory. Specifications can be found in the caption. It can be easily checked that, during the initial stage (the first three quarters), `FedAWE` slightly lags behind `FedAvg` over active clients. However, when reaching the final stage (the last quarter), `FedAWE` attains the target accuracy in a comparable or lower number of rounds to `FedAvg` over active clients in the evaluations on SVHN and CINIC-10 datasets. The slowdown of `FedAWE` on CIFAR-10 dataset is worth further investigation. In general, we arrive numerically at the conclusion that the staleness incurred by implicit gossiping in `FedAWE` is mild.

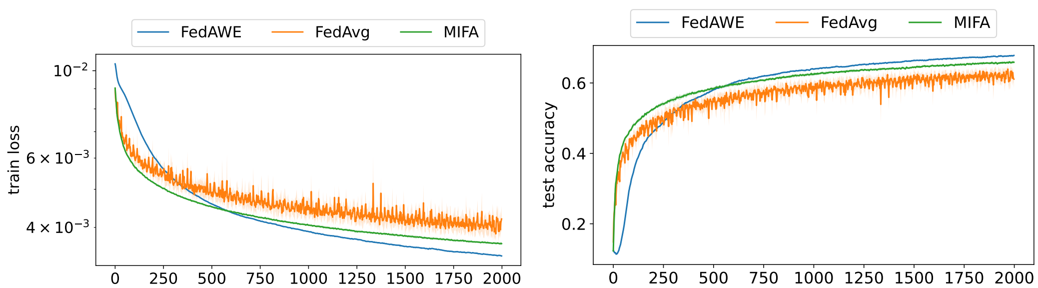
Training curves. In this part, we show the training curves of `FedAvg` over active clients, `FedAWE` and `MIFA`. In particular, the presented results of `FedAWE` are after exponential moving



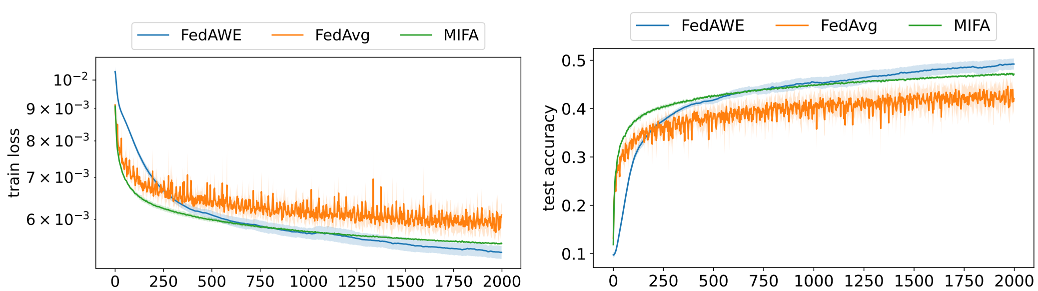
(a) Evaluation results on SVHN dataset without exponential moving average



(b) Evaluation results on SVHN dataset



(c) Evaluation results on CIFAR10 dataset



(d) Evaluation results on CINIC10 dataset

Figure 7: Missing training curves under non-stationary client unavailability dynamics with sine curve

average [5] under a parameter 0.99. Note that this is to ease down the noisy progress, and for a neat presentation only, the reported results in the main text and ablation studies are all from raw data. Fig. 7a plots the train loss and test accuracy from raw data. For example, when compared with Fig. 7b, EMA eases down the fluctuations but does not change either the trend or the order of algorithm performance results. All train losses are plotted on a logarithmic scale. The results are consistent with Table 2.

Impact of system-design parameters. In this part, we study the impact of system-design parameter including the degree of non-stationarity γ and data heterogeneity α under non-stationary with sine

Table 9: Results after different parameter γ. $p_i^t = p_i \cdot (\gamma \sin(2\pi/P \cdot t) + (1-\gamma))$.

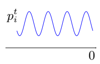
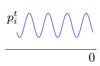
Unavailable Dynamics	Datasets	$\gamma = 0.3$		$\gamma = 0.2$		$\gamma = 0.1$	
	Algorithms	Train	Test	Train	Test	Train	Test
Non-stationary (**Sine**) 	`FedAWE` (**ours**)	**85.7** $\pm$ 0.9 %	**85.6** $\pm$ 0.9 %	**85.7** $\pm$ 0.5 %	**85.7** $\pm$ 0.5 %	**85.8** $\pm$ 0.6 %	**85.7** $\pm$ 0.7 %
	`FedAvg` over *active*	82.1 $\pm$ 1.1 %	82.0 $\pm$ 1.3 %	82.0 $\pm$ 1.2 %	81.9 $\pm$ 1.2 %	82.3 $\pm$ 0.9 %	82.2 $\pm$ 1.0 %
	`FedAvg` over *all*	71.3 $\pm$ 2.5 %	71.3 $\pm$ 2.8 %	73.2 $\pm$ 2.5 %	73.2 $\pm$ 2.8 %	74.0 $\pm$ 2.1 %	74.9 $\pm$ 2.4 %
	`FedAU`	$\underline{82.5}$ $\pm$ 1.4 %	$\underline{82.5}$ $\pm$ 1.3 %	$\underline{83.5}$ $\pm$ 0.3 %	$\underline{83.4}$ $\pm$ 0.4 %	$\underline{83.7}$ $\pm$ 0.3 %	$\underline{83.6}$ $\pm$ 0.3 %
	`F3AST`	82.3 $\pm$ 1.0 %	82.3 $\pm$ 1.0 %	82.3 $\pm$ 0.9 %	82.6 $\pm$ 0.8 %	82.9 $\pm$ 0.7 %	82.9 $\pm$ 0.6 %
	`FedAvg` with *known* p_i^t's	86.3 $\pm$ 1.0 %	86.0 $\pm$ 1.0 %	86.2 $\pm$ 1.2 %	86.0 $\pm$ 1.4 %	86.4 $\pm$ 0.9 %	86.0 $\pm$ 0.8 %
	`MIFA` (*memory aided*)	84.2 $\pm$ 0.4 %	84.1 $\pm$ 0.4 %	84.6 $\pm$ 0.1 %	84.5 $\pm$ 0.1 %	84.6 $\pm$ 0.1 %	84.4 $\pm$ 0.1 %

Table 10: Results after different Dirichlet parameter α. $p_i^t = p_i(\gamma \sin(2\pi/P \cdot t) + (1-\gamma))$.

Unavailable Dynamics	Datasets	$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 1.0$	
	Algorithms	Train	Test	Train	Test	Train	Test
Non-stationary (**Sine**) 	`FedAWE` (**ours**)	**82.5** $\pm$ 2.1 %	**82.5** $\pm$ 2.4 %	**85.7** $\pm$ 0.9 %	**85.6** $\pm$ 0.9 %	**90.6** $\pm$ 0.2 %	**89.7** $\pm$ 0.3 %
	`FedAvg` over *active*	78.9 $\pm$ 1.6 %	78.5 $\pm$ 1.8 %	82.1 $\pm$ 1.1 %	82.0 $\pm$ 1.3 %	88.3 $\pm$ 0.1 %	87.5 $\pm$ 0.1 %
	`FedAvg` over *all*	58.5 $\pm$ 3.0 %	58.5 $\pm$ 3.8 %	71.3 $\pm$ 2.5 %	71.3 $\pm$ 2.8 %	82.0 $\pm$ 0.7 %	81.9 $\pm$ 0.6 %
	`FedAU`	$\underline{79.5}$ $\pm$ 1.6 %	$\underline{79.5}$ $\pm$ 1.7 %	$\underline{82.5}$ $\pm$ 1.4 %	$\underline{82.5}$ $\pm$ 1.3 %	$\underline{88.4}$ $\pm$ 0.1 %	$\underline{87.6}$ $\pm$ 0.2 %
	`F3AST`	78.9 $\pm$ 1.3 %	78.9 $\pm$ 1.3 %	82.3 $\pm$ 1.0 %	82.3 $\pm$ 1.0 %	87.6 $\pm$ 0.1 %	87.0 $\pm$ 0.1 %
	`FedAvg` with *known* p_i^t's	84.2 $\pm$ 1.0 %	83.5 $\pm$ 1.0 %	86.3 $\pm$ 1.0 %	86.0 $\pm$ 1.0 %	91.5 $\pm$ 0.3 %	90.5 $\pm$ 0.1 %
	`MIFA` (*memory aided*)	82.6 $\pm$ 0.1 %	82.6 $\pm$ 0.0 %	84.2 $\pm$ 0.4 %	84.1 $\pm$ 0.4 %	88.4 $\pm$ 0.1 %	87.5 $\pm$ 0.1 %

trajectory. The results are in Table 9 and Table 10. Overall, `FedAWE` keeps outperforming the algorithms not assisted by memories or known statistics.

In Table 10, clients' local data becomes more heterogeneous when α increases. We can see a clear increase trend in accuracy. However, `FedAWE` remains to attain the best accuracies both train and test when compared to the algorithms not aided by memory or known statistics. Moreover, it outperforms `MIFA`, which consumes a lot of storage space, when $\alpha = 0.1$ and 1.0. The observations confirm the practicality of `FedAWE`.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have faithfully stated our contributions in both the abstract and introduction.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to Appendix A for details.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The regulatory assumptions are stated in Section 6. Due to space limitations, we are unable to present all the missing proofs and intermediate results in the main text. They are deferred to Appendix. Please refer to Table of Contents for details.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed experimental and the hyperparameter setups in Section 7 and Appendix J.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our evaluations are based on open-accessed datasets that are publically available. An official implementation code is provided through a GitHub link.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental setting/details are important parts of reproducing our results. We provide the details in Section 7 and Appendix J to the best of our ability.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our results are averaged over multiple random seeds and accompanied by error bars

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Please find the software/hardware specifications in Appendix J.2.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The NeurIPS code of ethics is strictly enforced throughout our research.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed broader impacts in Appendix B. We are unaware of any negative impacts.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The existing assets used in this paper has been adequately cited or credited to.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have documented the experiment details in Section 7 and Appendix J.2. In addition, we provide our code with clear details and examples.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects