Journal of Machine Learning Research 27 (2026) 1-73 Submitted 5/25; Published 8/26

Resilience Beyond Stationary Client Unavailability: Unlocking Efficient and Unbiased Federated Learning *

Ming Xiang[1] XIANG.MI@NORTHEASTERN.EDU
Stratis Ioannidis[1] E.IOANNIDIS@NORTHEASTERN.EDU
Edmund Yeh[1] E.YEH@NORTHEASTERN.EDU
Carlee Joe-Wong[2] CJOEWONG@ANDREW.CMU.EDU
Lili Su[1] L.SU@NORTHEASTERN.EDU
[1] *Northeastern University, Boston, MA USA*;
[2] *Carnegie Mellon University, Pittsburgh, PA USA*.

Editor: Zhaoran Wang

Abstract

Due to resource constraints or external and internal uncertainties, clients in real-world federated learning systems are often intermittently available edge devices. In highly dynamic environments, the parameter server lacks prior real-time knowledge of clients' availability, making it challenging to adapt traditional federated learning algorithms to be resilient to uncertainties in client availability. If not carefully addressed, complex client availability can introduce significant bias, potentially harming the performance of the trained model. Most prior work either fails to account for non-stationary client availability dynamics or demands significant memory and computational overhead. This paper aims to develop efficient federated learning algorithms that are provably resilient to heterogeneous and non-stationary stochastic client availability. We propose `FedSWE`, which admits novel algorithmic structures to (i) compensate for missed computations, (ii) stabilize and diffuse the global updates over rounds, and (iii) evenly mix the local updates through implicit gossiping, despite being agnostic to non-stationary dynamics. Compared with the standard `FedAvg`, `FedSWE` introduces light additional memory and computation overhead. We show that `FedSWE` converges to a stationary point of non-convex objectives while achieving the desired linear speedup property in certain special cases. We corroborate our analysis with numerical experiments over diversified client unavailability dynamics on real-world data sets.

Keywords: federated learning, non-convex optimization, heterogeneous data, client unavailability, fault-tolerance

1. Introduction

Federated learning is a distributed machine learning framework that enables training global models without disclosing raw local data (McMahan et al., 2017; Kairouz et al., 2021). It has been adopted in commercial applications such as autonomous vehicles (Chen et al., 2021; Zeng et al., 2022; Peng et al., 2023), internet of things (Nguyen et al., 2019), and natural language processing (Yang et al., 2018; Ramaswamy et al., 2019).

∗. A preliminary version of this work (Xiang et al., 2024) was presented at the 38th Annual Conference on Neural Information Processing Systems, Vancouver, Canada.

Heterogeneous data and massive client populations are two of the defining characteristics of cross-device federated learning systems (McMahan et al., 2017; Kairouz et al., 2021; McLaughlin and Su, 2024). Despite intensive efforts (McMahan et al., 2017; Li et al., 2020b; Yuan and Li, 2022; Ruan et al., 2021; Kairouz et al., 2021), several key challenges that arise from the involvement of large-scale client populations are often overlooked in the existing literature (Perazzone et al., 2022). One of the primary hurdles is the issue of intermittent client unavailability. Intuitively, more active clients drive the global model to their local optima, biasing the training. In addition, the higher the uncertainty in client unavailability, the larger the performance degradation. Concrete examples that confirm these intuitions 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 (Tse and Viswanath, 2005; Wen et al., 2024; Ye et al., 2022; Bonawitz et al., 2019; Kairouz et al., 2021). The intricate interplay of internal and external factors results in non-stationary and heterogeneous client unavailability.

There is a recent surge in the study of client unavailability (Li et al., 2020a; Yang et al., 2022; Wang and Ji, 2022, 2024; Cho et al., 2023a; Gu et al., 2021; Yan et al., 2023; Crawshaw and Liu, 2024). Despite their solid foundation in this direction, most prior work either assumes exact knowledge of the clients' availability or requires their dynamics to be benignly stationary (McMahan et al., 2017; Li et al., 2020a; Perazzone et al., 2022; Wang and Ji, 2022, 2024; Crawshaw and Liu, 2024). The non-stationarity in client unavailability remains largely underexplored. 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 (Xie et al., 2019; Nguyen et al., 2022; Toghani and Uribe, 2022; Koloskova et al., 2022). However, the proposed methods assume the availability of all clients or uniformly sampled clients, making them infeasible for dynamic and complex client availability in practice. A handful of other works (Gu et al., 2021; Jhunjhunwala et al., 2022; Yan et al., 2023) memorize the old gradients of unavailable clients. 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.

We adopt the commonly-used stochastic client unavailability model (McMahan et al., 2017; Wang and Ji, 2022; Jhunjhunwala et al., 2022; Perazzone et al., 2022; Wang and Ji, 2024), where each client i is available for federated learning training with probability p_i^t in round t. The p_i^t's are heterogeneous across clients and are subject to unknown and non-stationary dynamics. An example can be found in Fig. 1. (Xiang et al., 2025) marks our first move towards understanding heterogeneity and non-stationarity in p_i^t's but focuses on a significantly simpler problem where p_i^t's are used to describe the uplink communication failures—it requires that clients be capable of continuous local optimization regardless of failures. In addition, it imposes the assumption that $p_i^t \geq \delta$, where $\delta > 0$ is an absolute constant.

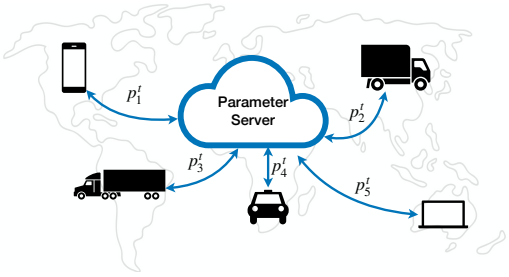


Figure 1: An illustration of heterogeneous and non-stationary availability.

Relaxing the requirement of continuous local computation, a preliminary version of this work (Xiang et al., 2024) studies client unavailability yet still imposes the technical assumption that $p_i^t \geq \delta$. In this extended version, we generalize the dynamics of p_i^t by allowing it to take zero values occasionally. Our generalized setup is motivated by the real-world scenario in which clients located in different geographical regions may experience availability issues due to time zone differences, naturally causing p_i^t's to drop to zero from time to time (Zhu et al., 2022). When clients participate independently, our generalized model of p_i^t covers some popular existing models as special cases such as $p_i^t = p$ (Li et al., 2020a; Yang et al., 2022), $p_i^t = p_i$ (Wang and Ji, 2024), regularized participation (Wang and Ji, 2022; Crawshaw and Liu, 2024) and cyclic participation (Cho et al., 2023a). Details can be found in Section 3.

Contributions. Our contributions are four-fold:

- We demonstrate in Section 4, using concrete examples in the context of `FedAvg` - the most widely adopted federated learning algorithm, that both heterogeneity and non-stationarity of p_i^t can result in bias and thus significant performance degradation of `FedAvg`.
- We propose a computational and memory-efficient algorithm `FedSWE` in Section 5. At a high level, the design of `FedSWE` introduces three novel algorithmic components:
  (i) **adaptive innovation echoing**, which helps clients catch up on missed computation;
  (ii) **global moving average**, which stabilizes and diffuses the global updates over rounds;
  (iii) **implicit gossiping**, which facilitates a balanced information mixture through implicit client-client gossip, ultimately correcting residual bias.

  Notably, no direct neighbor information exchanges are involved, and the client unavailability dynamics remain undisclosed to all clients and the parameter server.
- In Section 6, we show the convergence of `FedSWE`, which exhibits the desired linear speedup property in certain special cases.
- In Section 7, we validate our analysis with numerical experiments over diversified client unavailability dynamics on real-world data sets.

2. Related Work

2.1 Dynamical client availability

There is a recent surge of efforts to study client availability (Ruan et al., 2021; Ribero et al., 2022; Chen et al., 2022; Jhunjhunwala et al., 2022; Wang and Ji, 2022, 2024; Perazzone et al., 2022; Xiang et al., 2023; Crawshaw and Liu, 2024), which can be roughly classified into two categories depending on whether the parameter server can unilaterally determine the participating clients.

(i) **Controllable availability.** Earlier research (McMahan et al., 2017; Li et al., 2020b; Jhunjhunwala et al., 2022) presumes that, in each round, the parameter server could recruit 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. (2022) 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. (2023a) 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 chosen unilaterally and deliberately by the parameter server. Perazzone

et al. (2022) 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 values of p_i^t are determined by the parameter server by solving a stochastic optimization problem. Chen et al. (2022) 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 probability p_i^t nor assumes that the parameter server has any influence on p_i^t's.

(ii) **Uncontrollable availability.** There is a handful of work on building resilience against arbitrary client availability (Ribero et al., 2022; Wang and Ji, 2022; Yan et al., 2023; Gu et al., 2021; Yang et al., 2022; Wang and Ji, 2024; Crawshaw and Liu, 2024). Ribero et al. (2022) consider random client availability whose underlying probabilities are also heterogeneous and time-varying with unknown dynamics. However, the underlying dynamics of p_i^t's in (Ribero et al., 2022) are assumed to follow a homogeneous Markov chain. Wang and Ji (2022) 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. Sharing a similar spirit, Crawshaw and Liu (2024) propose a `SCAFFOLD` variant that amplifies global parameter and local gradient updates every P round. In spite of its communication efficiency, ability for correlated participation, and resilience to data heterogeneity, the rolling average of p_i^t over every P round is assumed to be the same constant for all clients. Yang et al. (2022) 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 (Gu et al., 2021; Yan et al., 2023) share the same idea of using the memorized latest gradient updates from unavailable clients for global aggregation. Despite superior numerical performance, both algorithms demand a substantial amount of additional memory (Wang and Ji, 2024). For non-convex objectives, both (Yan et al., 2023) and (Gu et al., 2021) require an absolute bounded inactive period, and share similar technical assumptions such as almost surely bounded stochastic gradients (Yan et al., 2023) or almost surely bounded gradient noise (Gu et al., 2021). 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. A recent work (Wang and Ji, 2024) considers unknown heterogeneous p_i's yet assumes p_i's are fixed over time. Another concurrent work (Sun et al., 2025) characterizes periodic client participation through the lens of the Markov chain. In spite of its resilience to non-uniform and correlated availability, p_i^t's are also assumed to be stationary over the window of participation.

2.2 Asynchronous federated learning

Another related line of work is asynchronous federated learning. To the best of our knowledge, Xie et al. (2019) initialize the study of asynchronous federated learning, wherein the parameter server adjusts 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. Nguyen et al. (2022) 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 (Xie et al., 2019; Nguyen et al., 2022; Toghani and Uribe, 2022; Koloskova et al., 2022), or bounded gradients (Xie et al., 2019; Nguyen et al., 2022; Koloskova et al., 2022). Recently, arbitrary delay is considered in the context of distributed SGD with bounded stochastic gradients and $(0, \zeta)$-bounded inter-client heterogeneity (Mishchenko et al., 2022) (see Assumption 4 therein 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 or staleness to be bounded.

Notations. Let $\|\boldsymbol{v}\|_2$, $\|A\|_{\mathrm{F}}$ and $\lambda_2(B)$ define the l_2 norm of a vector $\boldsymbol{v}$, the Frobenius norm of a matrix A, and the second largest eigenvalue of a squared matrix B, respectively. Denote $\mathcal{F}^t$ the sigma algebra generated by randomness up to round t, $\mathbb{R}^d$ a d-dimensional vector space, and $[m]$ a set $\{k : k \in \mathbb{N}, 1 \leq k \leq m]\}$. $\mathbb{1}_{\{\mathcal{E}\}}$ is an indicator function of an event $\mathcal{E}$, i.e., $\mathbb{1}_{\{\mathcal{E}\}} = 1$ when event $\mathcal{E}$ occurs, but $\mathbb{1}_{\{\mathcal{E}\}} = 0$ otherwise. For two functions $f(n)$ and $g(n)$, we have $f(n) \lesssim g(n)$, if there exists a constant $c_o > 0$ and an integer $n_o \in \mathbb{N}$ such that $f(n) \leq c_o g(n)$ for all $n \geq n_o$, while $f(n) \asymp g(n)$, if there exists a constant $c_\theta > 0$ and an integer $n_\theta \in \mathbb{N}$ such that $f(n) = c_\theta g(n)$ for all $n \geq n_\theta$.

3. Problem Formulation

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

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

where $F_i(\boldsymbol{x}) \triangleq \mathbb{E}_{\xi_i \sim \mathcal{D}_i} [\ell_i(\boldsymbol{x}; \xi_i)]$ is the non-convex local objective, $\mathcal{D}_i$ 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 formally describe the non-stationary and heterogeneous client availability that we consider in this paper. Let $\mathcal{A}^t$ denote the set of available clients, and T be the number of total training rounds.

Assumption 1. Define $p_i^t \triangleq \mathbb{E}[\mathbb{1}_{\{i \in \mathcal{A}^t\}}]$. For any given $\delta \in (0, 1]$, there exists a window of P rounds such that

$$\frac{1}{P} \sum_{t=nP}^{(n+1)P-1} p_i^t \geq \delta, \quad \forall n \in \mathbb{N}, \tag{2}$$

where the events $\{i \in \mathcal{A}^t\}$ are independent across clients i and across rounds t. Let $\mathcal{P}_\delta$ be the collection of all P's that satisfy (2), and $P_\delta \triangleq \min \mathcal{P}_\delta$.

Assumption 1 requires that the averaged p_i^t's over P consecutive rounds are non-trivially lower bounded away from zero while allowing p_i^t's occasionally drop to zeros. Note that if (2)

holds for some P_0, then it must also hold for $c \cdot P_0$, where $c \in \mathbb{N}$. Hence, we will focus on P_δ to eliminate ambiguity. Next, we elaborate on the generality of Assumption 1 in Remark 1.

Remark 1. Within the context that $\{i \in \mathcal{A}^t\}$ are independent across agents i and across rounds t, Assumption 1 generalizes many existing client availability assumptions.

- When $P_\delta = 1$, (2) reduces to $p_i^t \geq \delta$, which encompasses the cases of uniform availability $p_i^t = p$ (Li et al., 2020a; Yang et al., 2022), stationary availability $p_i^t = p_i \geq \min_{i\in[m]} p_i$ (Wang and Ji, 2024) and non-stationary availability $p_i^t \geq \delta$ (Xiang et al., 2024, 2025).
- When $P_\delta > 1$, Assumption 1 encompasses the following dynamics:

(i) *Regularized participation* (Wang and Ji, 2022; Crawshaw and Liu, 2024): There exists $\mu_{t_0} \geq 1/P_\delta$ such that

$$\frac{1}{P_\delta} \sum_{t=nP_\delta}^{(n+1)P_\delta - 1} \mathbb{P}\left\{i \in \mathcal{A}^t\right\} = \mu_n, \text{ for } i \in [m] \text{ and } \forall n \in \mathbb{N}, \tag{3}$$

where P_δ is some carefully chosen integer. (3) says that every client becomes available equally often within each window. By contrast, our Assumption 1 does not require such "balance" in μ_n. On the other hand, outside the restriction that $\{i \in \mathcal{A}^t\}$ are independent across i and t, the regularized participation assumption (Wang and Ji, 2022; Crawshaw and Liu, 2024) allows clients' participation to be correlated within P_δ consecutive rounds.

(ii) *Cyclic participation* (Cho et al., 2023a; Crawshaw and Liu, 2024): Suppose that the m clients can be equally partitioned into $\bar{K}$ groups. Denote the group index of client i as $G(i) \in \{1, \cdots, \bar{K}\}$. In each round t, p_i^t is determined as

$$p_i^t = \begin{cases} p, & \text{if } (t \bmod \bar{K}) = G(i); \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Cyclic participation characterizes the scenario where client groups become available in order. For example, if clients are populated in different time zones around the globe, their availability naturally exhibits cyclic patterns due to time differences. On the other hand, outside the restriction that $\{i \in \mathcal{A}^t\}$ are independent across i and t, cyclic participation also admits correlated participation. Sun et al. (2025) generalize (4) by allowing clients to have different p_i's rather than a homogeneous p.

Independent client unavailability is a widely adopted assumption in federated learning research (Li et al., 2020a,b; Karimireddy et al., 2020; Yang et al., 2021, 2022; Wang and Ji, 2024), and it is also the focus of this paper. We aim to extend our approach to address correlated client unavailability in future work. Analyzing non-independent availability with uncertain probabilistic trajectories in Assumption 1 is in general challenging. For example, the involved entanglement of stochastic gradient and availability statistics fundamentally complicates the theoretical analysis. We conjecture that independent participation may be only for the technical convenience of our analysis. Our experiments in Section 7 suggest that the proposed algorithms offer notable improvements even when the clients' participation is correlated.

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

$$F(x) = \frac{1}{2}(\|x - u_1\|_2^2 + \|x - u_2\|_2^2), \tag{5}$$

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`.

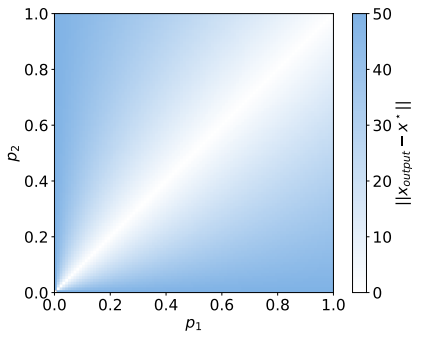
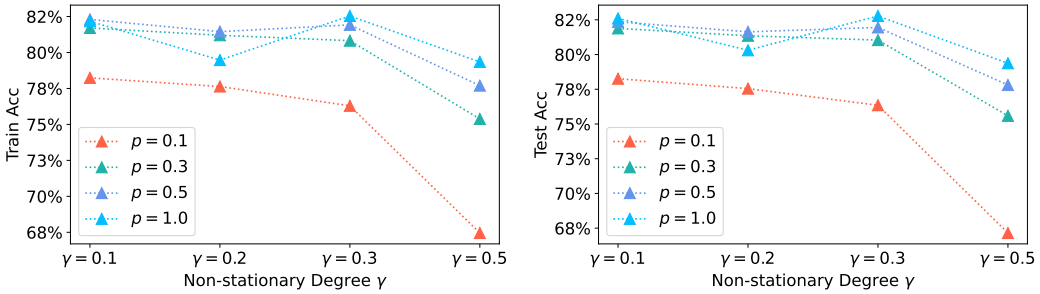


Figure 2: Let $x_{\text{output}} \triangleq \lim_{t\to\infty} \mathbb{E}[x^t]$. Under most of the choices of p_1, p_2, x_{output} is far from x^*.

Example 2 (Non-stationarity)**.** In Fig. 3, a total of $m = 100$ clients perform an image classification task on the SVHN data set (Netzer et al., 2011) under the `FedAvg` algorithm, whose local data set distribution follows Dirichlet(0.1) (Hsu et al., 2019). 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 H. 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 Stabilized Agile Weight Re-Equalization (`FedSWE`)

In order to minimize (1), it is natural to have the entire client population perform the same number of local updates and mix these updates carefully to ensure that they are weighted equally. However, due to intermittent availability, clients may miss computations in certain rounds and, as a result of the heterogeneity in p_i^t, are unable to contribute an equal number of local updates. An alternative 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 feasible due to hardware or software constraints. Formally, recall that $\mathcal{A}^t$ is the set of available clients at time t. Let $\tau_i(t) := \{t' : t' < t \text{ and } i \in \mathcal{A}^{t'}\}$ denote the most recent (with respect to time t) round that client i is available. Compared with standard `FedAvg`, the 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 global round when a client is available in standard `FedAvg`. In this work, we target computation-light algorithms that, compared with `FedAvg`, adjust local updates by $O(1)$ additional computation per client without additional stochastic samples.

We propose **Fed**erated **S**tabilized Agile **W**eight Re-**E**qualization (`FedSWE`), which is formally described in Algorithm 1. It involves three novel algorithmic structures: *adaptive innovation echoing*, *global moving average* and *implicit gossiping*. At a high level, these novel algorithmic structures (i) help clients catch up on the missed computation, (ii) stabilize and diffuse the global updates over rounds by interpolating between the fresh local updates and the most recent global update, and (iii) enable a balanced information mixture through implicit client-client gossip, ultimately correcting 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 server keeps tracking the most recent global update $\boldsymbol{x}^t$. The algorithm's 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, the interpolation coefficient k, and the initial model $\boldsymbol{x}^0$. In each round t, in lines 6-10, similar to `FedAvg`, an available client $i \in \mathcal{A}^t$ performs s steps of stochastic gradient descent on its local model $\boldsymbol{x}_i^t$, where $\nabla \ell_i(\cdot; \xi_i^{(t,k)})$ is the stochastic gradient of sample $\xi_i^{(t,k)}$. Next, we describe the novel algorithmic structures used in `FedSWE`.

5.1 Adaptive innovation echoing

Departing from `FedAvg`, wherein the local estimate $\boldsymbol{x}_i^t$ is updated as $\boldsymbol{x}_i^{t\dagger} \leftarrow \boldsymbol{x}_i^{(t,0)} - \eta_g \boldsymbol{G}_i^t$. `FedSWE` "echos" the local innovation $\boldsymbol{G}_i^t$ by multiplying it by $(t - \tau_i(t))$ (lines 12-13). Intuitively, this simple echoing helps approximately equalize the number of local improvements, as formally stated in Proposition 2. It says that the total number of innovations echoed is the same for all active clients for any given round.

Proposition 2. For any $R \in \mathbb{N}$, if $i \in \mathcal{A}^{R-1}$, then $\sum_{t=0}^{R-1} \mathbb{1}_{\{i \in \mathcal{A}^t\}} (t - \tau_i(t)) = R$.

5.2 Global moving average

When client availability is highly dynamic, the set of active clients $\mathcal{A}^t$ can vary greatly from round to round, leading to unsteady global updates. Our extensive experiments in Section 7.1 confirm that most of the state-of-the-art methods experience significant fluctuations during training. The early version of our algorithm (i.e., Algorithm 1 with $k = 0$) presented in the conference paper (Xiang et al., 2024) is no exception and has a similar level of variability.

When $k > 0$, Algorithm 1 adaptively interpolates between the fresh local updates $\sum_{i \in \mathcal{A}^t} \boldsymbol{x}_i^{t\dagger}$ and the most recent global update $\boldsymbol{x}^t$ in line 17, where the interpolation coefficients are jointly decided by the number of active clients $\left|\mathcal{A}^t\right|$ and the parameter k. For ease of

Algorithm 1: **Fed**erated **S**tabilized Agile **W**eight Re-**E**qualization (`FedSWE`)

Inputs: T, s, k, η_l, η_g, $\boldsymbol{x}^0$.
★ **Initializations.**
for $i \in [m]$ **do** $\boldsymbol{x}_i^0 \leftarrow \boldsymbol{x}^0$ and $\tau_i(0) \leftarrow -1$;
for $t = 0, \cdots, T-1$ **do**
  ★ **On the available clients.**
  for $i \in \mathcal{A}^t$ **do**
    $\boldsymbol{x}_i^{(t,0)} \leftarrow \boldsymbol{x}_i^t$;
    for $k = 0, \cdots, s-1$ **do**
      $\boldsymbol{x}_i^{(t,k+1)} \leftarrow \boldsymbol{x}_i^{(t,k)} - \eta_l \nabla \ell_i(\boldsymbol{x}_i^{(t,k)}; \xi_i^{(t,k)})$; ▷ Client local SGD
    end
    $\boldsymbol{G}_i^t \leftarrow \boldsymbol{x}_i^{(t,0)} - \boldsymbol{x}_i^{(t,s)}$;
    $\boldsymbol{x}_i^{t\dagger} \leftarrow \boldsymbol{x}_i^{(t,0)} - \eta_g(t - \tau_i(t))\boldsymbol{G}_i^t$; ▷ Client adaptive innovation echoing (Section 5.1)
    $\tau_i(t+1) \leftarrow t$; ▷ Client round index counter update (Section 5.1)
    Report $\boldsymbol{x}_i^{t\dagger}$ to the parameter server
  end
  ★ **On the parameter server.**
  $\boldsymbol{x}^{t+1} \leftarrow \frac{1}{|\mathcal{A}^t|+k}\left(\sum_{i\in\mathcal{A}^t} \boldsymbol{x}_i^{t\dagger} + k\boldsymbol{x}^t\right)$; ▷ Global moving average (Section 5.2)
  Multicast $\boldsymbol{x}^{t+1}$ to clients $i \in \mathcal{A}^t$; ▷ Postponed multicast (Section 5.3)
  ★ **On all clients.**
  for $i \in [m]$ **do**
    if $i \in \mathcal{A}^t$ **then**
      $\boldsymbol{x}_i^{t+1} \leftarrow \boldsymbol{x}^{t+1}$;
    else
      $\boldsymbol{x}_i^{t+1} \leftarrow \boldsymbol{x}_i^t$;
      $\tau_i(t+1) \leftarrow \tau_i(t)$;
  end
end

exposition, we restate the interpolation as follows:

$$\boldsymbol{x}^{t+1} = \frac{1}{|\mathcal{A}^t| + k} \sum_{i \in \mathcal{A}^t} \boldsymbol{x}_i^{t\dagger} + \left(1 - \frac{1}{|\mathcal{A}^t| + k}\right) \boldsymbol{x}^t. \tag{6}$$

Intuitively, as k increases, the interpolation produces a smoother curve by emphasizing more on the most recent global update. However, the budget for increasing k is not unlimited. In particular, when $k \to \infty$, the global update $\boldsymbol{x}^t$ duplicates the global model from the last round, preventing effective learning from occurring. Specifically, unrolling the recursion, we have $\boldsymbol{x}^{t+1} = \boldsymbol{x}^0$ for all $t \geq 0$, i.e., the global model $\boldsymbol{x}^t$ is not updated at all. In Section 7, we will show that $k = \Theta(m)$ is generally a reasonable empirical choice. We conjecture that discrete k's are only necessary for the technical convenience of our analysis, and leave it as a future work on how to analyze continuous k theoretically.

Furthermore, we want to note that (6) is closely related to global momentum and model exponential moving average, yet the interpolation coefficient is neither static nor decaying over rounds. Therefore, the existing theoretical analysis for momentum with static coefficient (Reddi et al., 2019; Li et al., 2023; Cheng et al., 2024) and model exponential moving average with decaying coefficient (Ahn and Cutkosky, 2024) is inapplicable to our problem. Beyond stabilizing training, we will show in Section 6 that interpolation is also necessary to strengthen global information diffusion across rounds. Details can be found therein.

5.3 Implicit gossiping

In `FedSWE`, the parameter server does not send the most recent global model to the active clients at the beginning of a global round. Instead, the parameter server aggregates the locally updated models $\boldsymbol{x}_i^{t\dagger}$ through (6) and sends the new global model $\boldsymbol{x}^{t+1}$ to all active clients $\mathcal{A}^t$ (lines 20-26). 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 (Xiang et al., 2023, 2024). Though the postponed multicasting brings in staleness, we will show that the staleness is bounded in Lemma 5. In addition, our empirical results (Table 7 in Appendix H) 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 (DeGroot, 1974; Boyd et al., 2006; Kempe et al., 2003; Hajnal and Bartlett, 1958; Lynch, 1996; Nedic and Ozdaglar, 2009). Intuitively, the clients' local estimates are eventually equally weighted in the final algorithm output. Note that, departing from the standard gossiping protocols therein (Kempe et al., 2003; Shah et al., 2009), information exchange in `FedSWE` does not involve direct client-client communication.

6. Convergence Analysis

6.1 Assumptions

In this section, we analyze the convergence of `FedSWE`. We start by stating regulatory assumptions that are common in federated learning analysis (Li et al., 2020a; Wang et al., 2020; Karimireddy et al., 2020).

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

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

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

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

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

$$\frac{1}{m} \sum_{i=1}^{m} \|\nabla F_i(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\|_2^2 \leq \beta^2 \|\nabla F(\boldsymbol{x})\|_2^2 + \zeta^2. \tag{7}$$

Table 1: Popular variant assumptions on gradient dissimilarity.

Bounded Gradient Dissimilarity	**References**
$\max_{\boldsymbol{x}} \|\nabla F_i(\boldsymbol{x})\|_2^2 \leq \zeta^2,\ \forall\ i \in [m]$	(Li et al., 2020b; Yu et al., 2019b; Cho et al., 2022, 2023b; Yan et al., 2023).
$\frac{1}{m}\sum_{i=1}^m \|\nabla F_i(\boldsymbol{x})\|_2^2 \leq \beta^2 \|\nabla F(\boldsymbol{x})\|_2^2$	(Li et al., 2019, 2020a)
$\frac{1}{m}\sum_{i=1}^m \|\nabla F_i(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\|_2^2 \leq \zeta^2$	(Yu et al., 2019a; Wang et al., 2019; Huang et al., 2022; Karimireddy et al., 2022; Wang and Ji, 2022, 2024; Yang et al., 2022; Wang et al., 2022; Allouah et al., 2023).
$\frac{1}{m}\sum_{i=1}^m \|\nabla F_i(\boldsymbol{x})\|_2^2 \leq \beta^2 \|\nabla F(\boldsymbol{x})\|_2^2 + \zeta^2$	(Karimireddy et al., 2020; Wang et al., 2020; Wang and Joshi, 2021; Gu et al., 2021; Yuan and Li, 2022).

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

6.2 Augmented Learning Systems

For ease of analysis, it is technically convenient to consider an augmented learning system with k virtual clients, and to show convergence of $\boldsymbol{x}$ through this system. Observing that the update in (6) can be rewritten as

$$\boldsymbol{x}^{t+1} = \frac{1}{|\mathcal{A}^t| + k} \sum_{i \in \mathcal{A}^t} \boldsymbol{x}_i^{t\dagger} + \frac{1}{|\mathcal{A}^t| + k} \sum_{i=m+1}^{m+k} \boldsymbol{x}^t.$$

More specifically, we construct the augmented learning system as follows: let $\mathcal{V} = \{m+1, \cdots, m+k\}$ with each element representing a virtual client; it holds that $\boldsymbol{x}_i^{t\dagger} = \boldsymbol{x}^t$ for each $i \in \mathcal{V}$. Let $F_i(\boldsymbol{x}) \triangleq 0$ for $i \in \mathcal{V}$ and $\forall \boldsymbol{x} \in \mathbb{R}^d$. It is easy to see that, with this local objective function, we have

$$\boldsymbol{x}^{t+1} = \frac{1}{|\mathcal{A}^t| + k} \sum_{i \in \mathcal{A}^t \cup \mathcal{V}} \boldsymbol{x}_i^{t\dagger},$$

In addition, unlike regular clients in $[m]$, each virtual client is always available, i.e., $\tau_i(t) = t-1$ for $i \in \mathcal{V}$. To distinguish, let $\mathcal{R} = [m]$ denote the regular clients and $M = m + k$. We define an auxiliary global objective function $\widetilde{F}$ as in (8):

$$\widetilde{F}(\boldsymbol{x}) \triangleq \frac{1}{M} \sum_{i \in \mathcal{R} \cup \mathcal{V}} F_i(\boldsymbol{x}) = \frac{1}{M} \sum_{i \in \mathcal{R}} F_i(\boldsymbol{x}) = \left(\frac{m}{M}\right) \frac{1}{m} \sum_{i \in \mathcal{R}} F_i(\boldsymbol{x}) = \left(\frac{m}{M}\right) F(\boldsymbol{x}). \tag{8}$$

Note that the auxiliary local and global objectives are only used to facilitate our analysis of Algorithm 1; they do not affect the computation at the regular clients. Hence, to show

that $\boldsymbol{x}$ converges to a stationary point of $F(\cdot)$ with $k > 0$ on the regular client population $\mathcal{R}$ is equivalent to showing that $\boldsymbol{x}$ converges to $\widetilde{F}(\boldsymbol{x})$ with $k = 0$ on the augmented client population $\mathcal{R} \cup \mathcal{V}$ up to rescaling. The latter case can be analyzed by adapting our road map from the conference version (Xiang et al., 2024), but with non-trivial characterizations to account for the generalized Assumption 1.

6.2.1 Information Mixing on the Augmented Learning System.

We construct a doubly stochastic information mixing matrix $W^{(t)}$ in (9) that characterizes the information diffusion in `FedSWE`.

$$W_{ij}^{(t)} \triangleq \begin{cases} \frac{1}{|\mathcal{A}^t|+k}, & \text{if } \{i,j \in \mathcal{A}^t \cup \mathcal{V}\}; \\ 1, & \text{if } \{i = j\} \text{ and } \{i \in \mathcal{R} \setminus \mathcal{A}^t\}; \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Let $W^{(n,P_\delta)} \triangleq \prod_{t=nP_\delta}^{(n+1)P_\delta-1} W^{(t)}$ and $\rho(n,k) \triangleq \lambda_2(\mathbb{E}[(W^{(n,P_\delta)})^2])$, where $\lambda_2(\cdot)$ denotes the second largest eigenvalue, $n \in \mathbb{Z}^+$, $\mathbf{J} = \mathbb{1}\mathbb{1}^\top/M$, and $\rho_k \triangleq \max_n \rho(n,k)$. The information mixing errors, i.e., consensus errors, are quantified through Lemma 3.

Lemma 3. (Boyd et al., 2005; Koloskova et al., 2020) For any matrix $B \in \mathbb{R}^{d\times M}$, it holds that $\mathbb{E}_W[\|B\left(W^{(n,P_\delta)} - \mathbf{J}\right)\|_{\mathrm{F}}^2] \leq \rho_k \|B(\mathbf{I} - \mathbf{J})\|_{\mathrm{F}}^2$, where the expectation is taken w.r.t. randomness in W matrices.

6.3 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 12 in Algorithm 1) with different strength $(t - \tau_i)$. As such, we construct an imaginary update sequence $\boldsymbol{z}_i^t$ for client $i \in [m]$, whose evolution is closely coupled with $\boldsymbol{x}^t$ and $\boldsymbol{x}_i^t$ yet is easier to analyze. Note that the imaginary update sequence is never actually computed by clients but acts as a necessary tool in building up the analysis.

Definition 4. The auxiliary sequence $\{\boldsymbol{z}_i^t\}$ of client $i \in \mathcal{R} \cup \mathcal{V}$ is defined as

$$\boldsymbol{z}_i^t \triangleq \begin{cases} \boldsymbol{x}_i^t, & \forall\ i \in \mathcal{V}; \\ \boldsymbol{x}_i^t - \eta_l \eta_g s(t - \tau_i(t) - 1)\nabla F_i(\boldsymbol{x}_i^{\tau_i(t)+1}), & \forall\ i \in \mathcal{R}. \end{cases} \tag{10}$$

We know that the virtual clients in $\mathcal{V}$ are (i) always available and (ii) with zero-valued local gradient updates since their local models are duplicates of the global model from the immediate previous round. Therefore, (10) can be simplified as (11) by using the convention that $\nabla F_i(\boldsymbol{x}_i^{\tau_i(t)+1}) = \mathbf{0}$ for any virtual client $i \in \mathcal{V}$ in any round $t \in [T]$:

$$\boldsymbol{z}_i^t \triangleq \boldsymbol{x}_i^t - \eta_l \eta_g s(t - \tau_i(t) - 1)\nabla F_i(\boldsymbol{x}_i^{\tau_i(t)+1}),\ \forall\ i \in \mathcal{R} \cup \mathcal{V}. \tag{11}$$

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

$$\boldsymbol{z}_i^t \overset{(12.a)}{=} \boldsymbol{x}_i^t \overset{(12.b)}{=} \frac{\sum_{j\in\mathcal{A}^{t-1}}}{|\mathcal{A}^{t-1}|}\left(\boldsymbol{z}_j^{t-1} + \underbrace{(\boldsymbol{x}_j^{t-1} - \boldsymbol{z}_j^{t-1})}_{(12.c)} - \eta_g(t-1-\tau_j(t-1))\boldsymbol{G}_j^{t-1}\right) \tag{12}$$

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

$$+ \frac{\eta_l\eta_g}{|\mathcal{A}^{t-1}|}\sum_{j\in\mathcal{A}^{t-1}}(t-2-\tau_j(t-1))\sum_{r=0}^{s-1}\left(\nabla F_j(\boldsymbol{x}_j^{\tau_j(t-1)+1}) - \nabla\ell_j(\boldsymbol{x}_j^{(t-1,r)};\xi_i^{(t,r)})\right), \tag{13}$$

where $(12.a)$ holds because of Definition 4 and $i \in \mathcal{A}^{t-1}$, $(12.b)$ because of line 12 in Algorithm 1, addition and subtraction, and we can get (13) by replacing $(12.c)$ with (10). Recall that, for client $j \in \mathcal{V}$, we have (i) $\boldsymbol{x}_j^t = \boldsymbol{z}_j^t$ and (ii) $\boldsymbol{G}_j^t = \boldsymbol{0}$.

When $i \in \mathcal{R}\setminus\mathcal{A}^{t-1}$, $\boldsymbol{z}_i^t$ has a simple iterative relation:

$$\boldsymbol{z}_i^t \;=\; \boldsymbol{z}_i^{t-1} - \eta_l\eta_g s\nabla F_i(\boldsymbol{x}_i^{\tau_i(t-1)+1}). \tag{14}$$

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 imaginary sequences are used in peer-to-peer distributed learning literature (Spiridonoff et al., 2020; Avdiukhin and Kasiviswanathan, 2021; Lian et al., 2017; Yuan et al., 2016; Stich, 2018; Nedić et al., 2018). Yet, existing constructions are not applicable to our problem due to the (i) non-convexity of the global objectives, (ii) multiple local updates per round, (iii) possibly unbounded gradients, and the (iv) 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 4 that $\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\|_2^2 \leq \eta_l^2\eta_g^2 s^2(t-\tau_i(t)-1)^2\|\nabla F_i(\boldsymbol{x}_i^{\tau_i(t)+1})\|_2^2$, 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 8.

Lemma 5 (Unavailability statistics)**.** Under Assumption 1 and δ defined therein. It holds for $t \geq 0$ that

$$\mathbb{E}\left[t - \tau_i(t)\right] \leq P_\delta + \frac{1}{\delta}; \tag{15}$$

$$\mathbb{E}\left[\left(t-\tau_i(t)\right)^2\right] \leq \frac{\left((P_\delta-1)\delta+1\right)^2 + (P_\delta-1)\delta^2 + 1}{\delta^2}. \tag{16}$$

Remark 6. Lemma 5 yields an upper bound on the first and second moments of a client i's unavailable duration. Its proof can be found in Appendix D.3, where we leverage the tools from probability theory (Gut, 2006). Here, we remark on some special cases:

- When $\delta = 1$, all clients are available during all training rounds, suggesting a static unavailable duration $t - \tau_i(t)$ of length 1 and thus a static second moment of value 1. Recall that we require P_δ to be the minimum window size when given a budget δ in Assumption 1, so it implies $P_\delta = 1$. In this case, both of our bounds are loose by only a constant offset 1.
- When $P_\delta = 1$, the dynamics (2) becomes $p_i^t \geq \delta$, which follows a heterogeneous geometric distribution. Our bounds (15) and (16) reduce to $1+1/\delta$ and $2/\delta^2$, respectively. Compared with (Xiang et al., 2024, Lemma 2), the first moment is loose by only a constant offset 1, while the second moment matches the result therein. When we further relax the condition and consider the special case where clients are available with the same probability δ, the unavailable duration $t - \tau_i(t)$ simply follows a homogeneous geometric distribution. It can be checked that our bound trivially holds.

6.4 Main Convergence Results

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

Lemma 7 (Descent Lemma)**.** Let $\mathcal{F}^t$ define the sigma algebra generated by randomness up to round t. Suppose that Assumptions 2, 3 hold and $\eta_l\eta_g \leq 1/(8sL)$. It holds that

$$\begin{aligned}
\mathbb{E}\left[\widetilde{F}(\bar{\boldsymbol{z}}^{t+1}) - \widetilde{F}(\bar{\boldsymbol{z}}^t) \mid \mathcal{F}^t\right] \leq & -\frac{\eta_l\eta_g s}{3}\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2 \\
& + \frac{2\eta_l\eta_g sL\sigma^2\left(\eta_l\eta_g\delta_{\max} + 9m\eta_l^2 sL\right)}{M^2}\sum_{i=1}^m (t-\tau_i(t))^2 \\
& + \frac{65\eta_g\eta_l^3 s^3L^2}{M}\sum_{i=1}^m (t-\tau_i(t))^2\left\|\nabla\widetilde{F}_i(\boldsymbol{x}_i^{\tau_i(t)+1})\right\|_2^2 \\
& + \frac{4\eta_l\eta_g sL^2}{M}\sum_{i=1}^m \underbrace{\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2}_{\text{Approximation Error}} + \frac{\eta_l\eta_g sL^2}{2M}\sum_{i=1}^m \underbrace{\left\|\boldsymbol{z}_i^t - \bar{\boldsymbol{z}}^t\right\|_2^2}_{\text{Consensus Error}}.
\end{aligned}$$

The proof of Lemma 7 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 8 (Approximation error)**.** Suppose that Assumptions 2 and 4 holds, we have

$$\begin{aligned}
&\frac{1}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^M \mathbb{E}\left[\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2\right] \overset{(17.a)}{=} \frac{1}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^m \mathbb{E}\left[\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2\right] \\
&\leq 3\eta_l^2\eta_g^2 s^2\frac{\left[(P_\delta-1)\delta+1\right]^2 + \left[(P_\delta-1)\delta^2+1\right]}{\delta^2}\frac{M\left(\beta^2+1\right)}{m}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] \\
&\quad + 3\eta_l^2\eta_g^2 s^2\frac{\left[(P_\delta-1)\delta+1\right]^2 + \left[(P_\delta-1)\delta^2+1\right]}{\delta^2}\left(\frac{m\zeta^2}{M} + \frac{L^2}{M}\sum_{i=1}^m\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t - \bar{\boldsymbol{z}}^t\right\|_2^2\right]\right), \qquad (17)
\end{aligned}$$

where $(17.a)$ holds because of (10).

The proof of Proposition 8 starts from Definition 4. 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. 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 $\widetilde{\boldsymbol{G}^t}$ (18) is formulated by combing (12) and (14).

$$\widetilde{\boldsymbol{G}}_i^t = \mathbb{1}_{\{i \in \mathcal{A}^t\}} \left(t - \tau_i(t)\right) \sum_{r=0}^{s-1} \left(\nabla \ell_i(\boldsymbol{x}_i^{(t,r)}) - \nabla F_i(\boldsymbol{x}_i^t) \right) + s \nabla F_i(\boldsymbol{x}_i^t). \tag{18}$$

Unrolling the recursion, it holds for the consensus error $\sum_{i=1}^M \|\boldsymbol{z}_i^t - \bar{\boldsymbol{z}}^t\|_2^2 / M$ that

$$\frac{1}{M} \left\| \left(\boldsymbol{Z}^{(t-1)} - \eta_l \eta_g \widetilde{\boldsymbol{G}}^{(t-1)} \right) W^{(t-1)} \left(\mathbf{I} - \mathbf{J} \right) \right\|_{\mathrm{F}}^2 \overset{(19.a)}{=} \frac{\eta_l^2 \eta_g^2}{M} \left\| \sum_{q=0}^{t-1} \widetilde{\boldsymbol{G}}^{(q)} \left(\prod_{l=q}^{t-1} W^{(q)} - \mathbf{J} \right) \right\|_{\mathrm{F}}^2, \tag{19}$$

where equality (19.a) holds because all clients are initiated at the same weight. Recall that ρ_k is the spectral norm of the information mixing matrix square over a rolling window of P_δ rounds in Lemma 3. To ensure an exponential decay of the consensus error, it is crucial to have $\rho_k < 1$, which is confirmed by Lemma 9.

Lemma 9. Suppose that Assumption 1 and $k > 0$.

When $P_\delta = 1$, it holds that

$$\rho_k \leq 1 - \frac{\left[m^2\delta + (k^2 + 2mk)\right]^2 \delta^2}{8(m+k)^2}. \tag{20}$$

When $P_\delta > 1$, it holds that

$$\rho_k \leq 1 - \frac{\left[m^2\delta + (k^2 + 2mk)\right]^2 \delta^2}{8(m+k)^{4P_\delta + 2}}. \tag{21}$$

Proof Proof Sketch. The full proof is deferred to Appendix D.5. Coarsely, we study the conductance of a hypothetical Markov chain, whose transition matrix is a matrix square $(W^{(n,P_\delta)})^2$, as we are interested in a sliding window of P_δ rounds. The unique challenge of our analysis arises from the involved information mixing between volatile regular clients in $\mathcal{R}$ and always-on virtual clients in $\mathcal{V}$. Instead, we represent the federated learning systems as a graph, where the nodes are clients, and the edges are defined by the mixing matrix W weights, capturing the interactions between clients. As such, we can bound the conductance of the hypothetical Markov chain by studying the product of the edge weights and use Cheeger's inequality to close the gap between the spectral norm and the conductance. ■

Remark 10. Lemma 9 is divided into two parts: $P_\delta = 1$ and $P_\delta > 1$, where (20) is tighter than (21) when $P_\delta = 1$. A similar dependence on the number of clients in the spectral norm bound has been noted in the prior fully decentralized learning literature, e.g., in (Nedić and Olshevsky, 2014), the clients therein are assumed to form a P_δ-strongly-connected graph. Mapping to our setup, it means that all clients are available at least once over P_δ consecutive rounds. In contrast, we only need the average available probability of each client to be non-zero across every P_δ rounds, which is more general and makes direct applications of their results inapplicable. It remains an open question whether the worst-case bound can be improved, and we would like to leave this as future work.

We now proceed to present the convergence rates. In the sequel, we assume it holds for η_g and η_l that

$$\eta_l\eta_g \leq \frac{\delta(1-\rho_k)\sqrt{m}}{96sL\sqrt{P_\delta M\left((P_\delta-1)^2\delta^2+1\right)(\beta^2+1)}}\ ;\ \eta_l \leq \frac{\delta}{216sL\sqrt{\left((P_\delta-1)^2\delta^2+1\right)(\beta^2+1)}}. \tag{22}$$

The proof of the consensus error borrows insights from the analysis of the gossip algorithm (Nedic et al., 2017; Wang et al., 2022) but with substantial adaptation to accommodate the novel auxiliary formulation and multi-step local updates. Under the learning rate conidtions in (22) and Assumptions 1, 2, 3 and 4, we can show that

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2\right] \asymp \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right] \asymp \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]. \tag{23}$$

It remains to bound the full convergence error of $\boldsymbol{z}_i^t$, which is presented in Theorem 11.

Theorem 11 (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 (22) are met for $T \geq 1$ and $k > 0$. It holds that

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] \lesssim& \left(\frac{m+k}{m}\right)\frac{\left(F(\bar{\boldsymbol{z}}^0)-F^\star\right)}{\eta_l\eta_g sT}+\left(\frac{m+k}{m}\right)\frac{\delta_{\max}\eta_l\eta_g L\sigma^2}{m+k}\left[(P_\delta-1)^2+\frac{1}{\delta^2}\right]\\ &+\left(\frac{m+k}{m}\right)\eta_l^2\eta_g^2 s^2L^2P_\delta\left(\sigma^2+\zeta^2\right)\left[(P_\delta-1)^2+\frac{1}{\delta^2}\right]\left[1+\frac{1}{(1-\rho_k)^2}\right].\end{aligned} \tag{24}$$

By addition, subtraction, and Young's inequality, (25) and (26) hold under Assumption 2.

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\bar{\boldsymbol{x}}^t\right\|_2^2\right] \asymp \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2\right]+\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right]; \tag{25}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right] \asymp \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2\right]+\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]. \tag{26}$$

Moreover, from (23), (25) and (26), it can be seen that (27) holds.

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\bar{\boldsymbol{x}}^t\right\|_2^2\right] \asymp \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] \asymp \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right]. \tag{27}$$

Combining (23), (24), (25) and (26), we are ready for Corollary 12.

Corollary 12 (Convergence rate of $\boldsymbol{x}_i^t$)**.** Suppose that Assumptions 1, 2, 3 and 4 hold. Choose learning rates as $\eta_l = \frac{1}{\sqrt{T}sL}$, $\eta_g = \sqrt{s\delta m}$. Let $T \geq 1$ be sufficiently large so that the conditions of η_l and η_g in (22) are satisfied. For $k > 0$, it holds that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right] \lesssim \left(\frac{m+k}{m}\right)\frac{L\left(F(\bar{\boldsymbol{x}}^0)-F^\star\right)}{\sqrt{s\delta mT}} + \frac{\delta_{\max}}{\delta^{\frac{3}{2}}\sqrt{smT}}\left[\left(P_\delta-1\right)^2\delta^2+1\right]\sigma^2 + \left(\frac{m+k}{m}\right)\frac{smP_\delta}{T}\left(\sigma^2+\zeta^2\right)\left[\left(P_\delta-1\right)^2\delta^2+1\right]\left[1+\frac{1}{(1-\rho_k)^2}\right]. \quad (28)$$

Remark 13 (Linear speedup)**.** Corollary 12 establishes the full convergence rate for `FedSWE` algorithm. It can be seen that the second term dominates when T is sufficiently large, which relates to stochastic gradient noise σ^2. 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 12 follows from (26) by plugging in Proposition 8 and Theorem 11.

In the special case where $k = \Theta(m)$ and $\mathcal{A}_t = [m]$, we simply have $\delta_{\max} = \delta = 1$ and $P_\delta = 1$. Our convergence bound reduces to $O(1/\sqrt{smT})$. In other words, we achieve the desired linear speedup property with respect to the number of local steps s and the number of clients m, matching rates in the established literature (Yu et al., 2019a,b; Yang et al., 2021; Wang and Ji, 2022, 2024). The linear speedup property enables a large cross-device federated learning system to take advantage of the massive scale of parallelism. Notice that the consensus error (27) and the convergence rate (28) have the same asymptotic order with respect to the parameters mentioned above. Hence, the consensus error also enjoys the desired linear speedup property when T is sufficiently large in this special case.

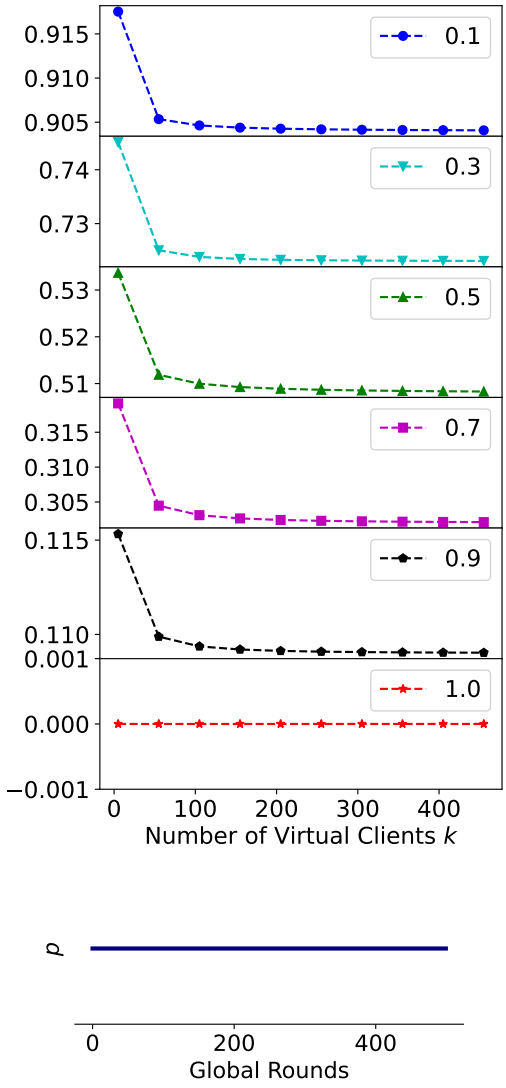
6.5 Impacts of k

In this section, we elaborate on the scaling of spectral norm ρ_k and convergence upper bound in (28) w.r.t. k. Next, we explore the necessity of interpolation ($k > 0$) under Assumption 1.

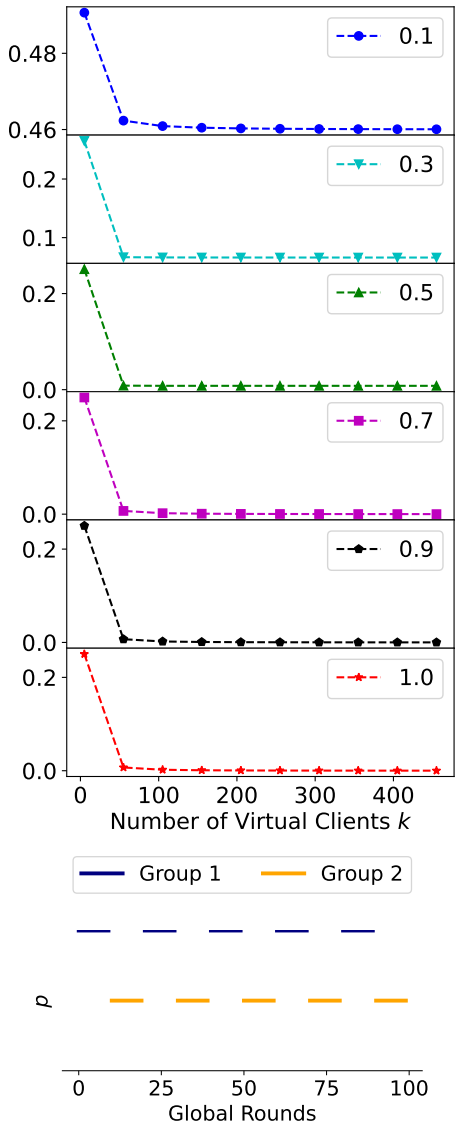
6.5.1 On the Scaling of Convergence Results

On spectral norm ρ_k. It is easy to show that the upper bound in (20) decreases monotonically in k by taking partial derivatives; yet, the monotonicity is not that straightforward in (21). On the other hand, the upper bounds characterize only the worst-case scenario, so they cannot directly inform the monotonicity analysis of ρ_k w.r.t. k. Intuitively, a greater k implies a more connected client population because virtual clients are always available. When $k \to \infty$, we have $\rho_k \approx 0$ since the W matrix will be dominated by the virtual clients, approaching a scaled all-one matrix. We hypothesize that the spectral norm ρ_k would decrease w.r.t. k, which is numerically demonstrated in Example 3 by explicit realizations of Assumption 1.

Example 3. In Fig. 4, a total of $m = 10$ clients are available under the dynamics shown in Fig. 4a with $P_\delta = 1$ and Fig. 4b with $P_\delta > 1$. We expect to see a smaller spectral norm when clients are more frequently available. The results match our hypothesis that the spectral norm ρ_k would decrease w.r.t. k. It is expected that a larger P_δ leads to worse information fusion, i.e., a smaller spectral norm ρ_k. Details can be found in the captions.



(a) Regular clients (i.e., clients in $\mathcal{R}$) are available with probability p in any global round. That is, the period is $P = 1$. The values of p can be found in the legends. y-axis is the calculated spectral norm of $(W^{(0,1)})^2$ over 1000 repetitions. We can see that the spectral norm decreases in k except when $p = 1$. In that case, clients are always available, leading to a zero-valued spectral norm.



(b) Regular clients (i.e., clients in $\mathcal{R}$) are divided into two groups. Each group i is available with probability p_i for 10 rounds. That is, the period is $P = 20$. y-axis is the calculated spectral norm of $(W^{(0,20)})^2$ over 1000 repetitions. We can see that the spectral norm decreases in k.

Figure 4: Calculated spectral norm results with varying numbers of k. We consider $m = 10$ clients with availability dynamics described in the plots. Note that the y-axis of each subplot is of a different range.

On convergence upper bound in (28)**.** We have shown in Lemma 9 that $\rho_k < 1$, which holds independently of k as long as $k > 0$. Intuitively, as we have discussed in Section 5.2,

there exists a sweet spot k^* for k that balances training stability and convergence speed. Yet, analytically obtaining the exact k^* is fundamentally challenging, if not impossible at all. Specifically, the value of k affects the convergence upper bound in (28) by (i) explicitly showing up in the numerator and by (ii) implicitly influencing the spectral norm ρ_k. Recall that we hypothesize in 3 that the spectral norm ρ_k monotonically decreases w.r.t. k. However, which term will ultimately dominate the monotonicity of (28) as k increases remains unclear. In Section 7, we empirically find that $k = \Theta(m)$ strikes the best balance between convergence speed and test accuracy.

6.5.2 Necessity of Interpolation ($k > 0$) when $P_\delta > 1$

Observe that we require the coefficient $k > 0$ in Lemma 9, which appears to be an artifact in our proofs to improve training stability. However, we show next in Proposition 14 that for our algorithm to hold under Assumption 1, it is necessary to have $k > 0$. To see this, we construct a counterexample via a similar quadratic function as in Example 1. Let client i's local objective $F_i(x) \triangleq \|x - u_i\|_2^2/2$, where $x,\ u_i \in \mathbb{R}$ and $i \in [m]$. The global objective is

$$F(x) = \frac{1}{2m} \sum_{i=1}^{m} \|x - u_i\|_2^2, \tag{29}$$

with unique global minimizer $x^* = (\sum_{i=1}^m u_i)/m$.

Proposition 14. For a global objective as per (29), let $\mathcal{M}_1$ the first half client population, $\mathcal{M}_2$ the remaining client population, and $\mathcal{M} \triangleq [m] = \mathcal{M}_1 \cup \mathcal{M}_2$. When all clients in $\mathcal{M}_1$ are available in even rounds only, while those in $\mathcal{M}_2$ are available in odd rounds only. We have

(i) Assumption 1 holds under $P_\delta = 2$ and $\delta = 0.5$;

(ii) It holds for the output x_{out}^t of Algorithm 1 without interpolation ($k = 0$) that

$$\lim_{t\to\infty} \mathbb{E}\left[\left\|\nabla F(x_{\text{out}}^t) - \nabla F(x^*)\right\|_2^2\right] = \frac{1}{m^2}\left\|\sum_{i\in\mathcal{M}_1} u_i - \sum_{j\in\mathcal{M}_2} u_j\right\|_2^2. \tag{30}$$

Proof Proof of Proposition 14. At a high level, our proof suggests an interesting client participation dynamics that prevents two groups of clients from properly mixing global updates.

Construction of p_i^t's. We assume that clients in $\mathcal{M}_1$ are available in even rounds only with $p_i^{2r} = 1$, while the rest clients in $\mathcal{M}_2$ are exclusively available in odd rounds with $p_i^{2r+1} = 1$, where $r \in \mathbb{Z}^+$. It is easy to extend the proof to the case where $p_i^t < 1$ for the same group of clients.

Global and local gradients. The global gradient $\nabla F(x)$ and local gradient $\nabla F_i(x)$ are

$$\nabla F(x) = x - \frac{1}{m}\sum_{i=1}^{m} u_i,\ \nabla F_i(x) = x - u_i.$$

Combining them together. Due to the postponed multi-cast procedure and the lack of interpolation, the parameter server cannot carry the aggregated global updates from one round to another. Hence, the available clients depend on their cohorts' updates in that round

to prepare their local models for the next availability. Since $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$ and non-overlap availability between two groups, the clients in $\mathcal{M}_1$ **cannot** exchange models with the clients in $\mathcal{M}_2$. Therefore, the global objective alternates between an average objective of clients in $\mathcal{M}_1$ and of clients in $\mathcal{M}_2$, i.e.,

$$\begin{cases} F_{\text{even}} & = \frac{1}{m} \sum_{i \in \mathcal{M}_1} \|x - u_i\|_2^2 ; \\ F_{\text{odd}} & = \frac{1}{m} \sum_{i \in \mathcal{M}_2} \|x - u_i\|_2^2 . \end{cases}$$

The expected output x_{out}^t follows that

$$\mathbb{E}\left[x_{\text{out}}^t\right] = \begin{cases} 2(\sum_{i \in \mathcal{M}_1} u_i)/m & \text{if } t = 2r; \\ 2(\sum_{i \in \mathcal{M}_2} u_i)/m & \text{if } t = 2r+1. \end{cases}$$

Consequently,

$$\mathbb{E}\left[\left\|\nabla F(x_{\text{out}}^t) - \nabla F(x^*)\right\|_2^2\right] = \frac{1}{m^2} \left\| \sum_{i \in \mathcal{M}_1} u_i - \sum_{j \in \mathcal{M}_2} u_j \right\|_2^2 .$$

■

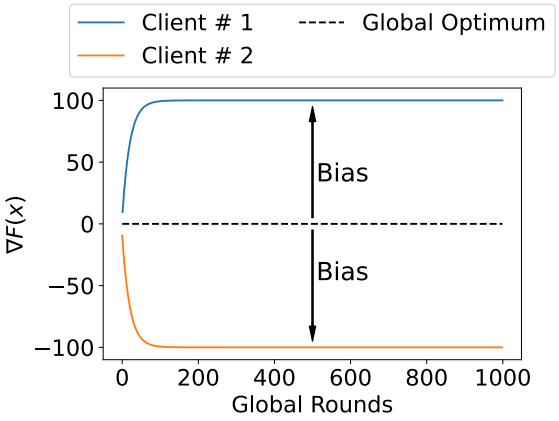
As is empirically verified by the scalar example in Fig. 5, $k > 0$ may correct the bias. Intuitively, interpolation mixes the global models over rounds and enables clients to share information through the parameter server with cohorts in the current round and those from previous rounds. Mathematically, we guarantee clients to exchange information with each other in expectation over every sliding window of P_δ rounds, provided $k > 0$. See the proof of Lemma 9 for details.

6.5.3 Special case when $P_\delta = 1$

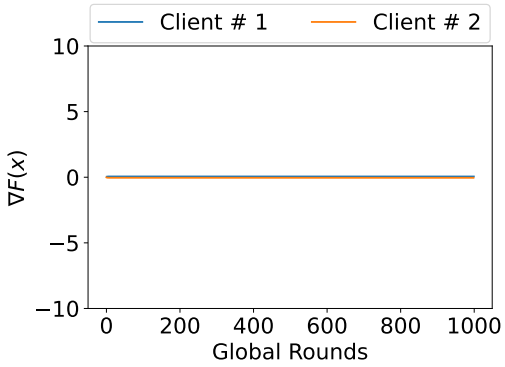
Our conference version (Xiang et al., 2024) studies a special case where $P_\delta = 1$ and interpolation becomes optional. Our discussions in Section 6.5.2 are consistent with our results therein. Informally, this is because clients in (Xiang et al., 2024) are never isolated into distinct groups when $P_\delta = 1$. On the technical front, all elements in the information mixing matrix $W^{(t)}$ are strictly positive in expectation in any round t under the unavailability dynamics therein. Hence, clients can evenly diffuse information with each other in expectation. Nevertheless, $k > 0$ still helps to reduce fluctuations of the trajectory of $\boldsymbol{x}^t$.

7. Numerical Experiments

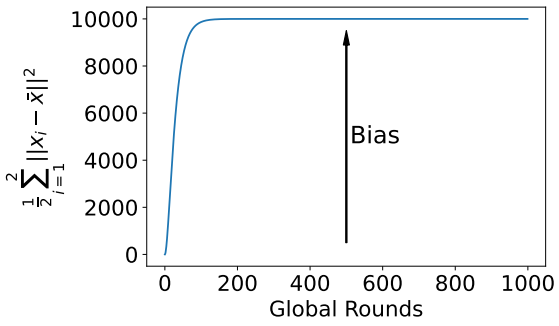
In this section, we evaluate `FedSWE` on real-world data sets to corroborate our analysis and compare the performance of `FedSWE` with the state-of-the-art algorithms. The missing specifications and additional numerical results can be found in Appendix H. Specifically, we consider a federated learning system of one parameter server and $m = 100$ clients. We focus on image classification tasks, and consider multiple real-world data sets (Netzer et al., 2011; Krizhevsky et al., 2009; Darlow et al., 2018). Each data set contains 10 image classes, but the categories differ.



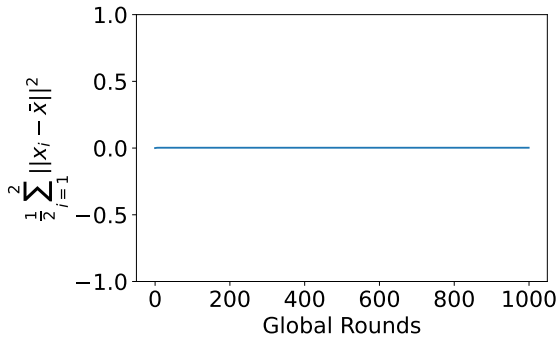
(a) A visualization of the scalar gradient of Algorithm 1 without interpolation ($k = 0$). It can be seen that the biases are significant.



(b) A visualization of the scalar gradient of Algorithm 1 **with** interpolation ($k = 1$). It can be seen that the gradient outputs converge to 0 under interpolation, recovering the global optimum.



(c) A visualization of the consensus of Algorithm 1 without interpolation ($k = 0$). It can be seen that two clients fail to reach a consensus.



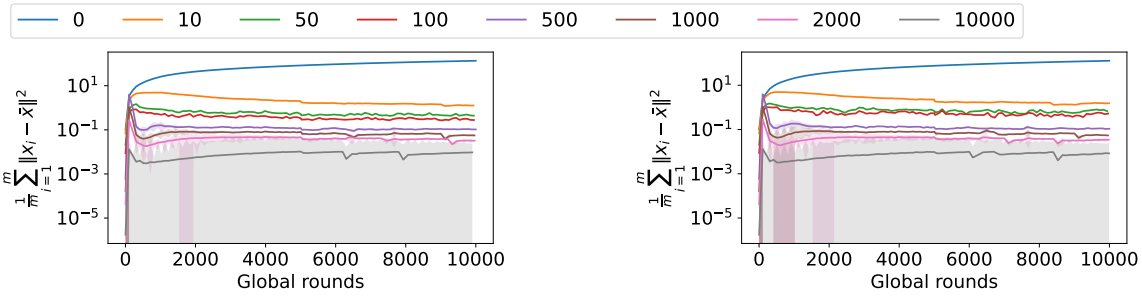
(d) A visualization of the consensus of Algorithm 1 **with** interpolation ($k = 1$). It can be seen that two clients can reach a consensus under interpolation.

Figure 5: Visualizations of Algorithm 1 with two clients and a global objective (29), whose $u_1 = 100$, $u_2 = -100$. The first client is available exclusively in even rounds, while the second client is available in odd rounds only. x-axis is the global round index, the details of y-axis can be found in the subplots. Figs. 5a and 5c plot the results of Algorithm 1 without interpolation. It can be seen that the outputs alternate between two clients' local optimums and can deviate far away from the true global minimizer $u_1 + u_2 = 0$. Moreover, the clients fail to reach a consensus. In sharp contrast, Figs. 5b and 5d indicate that Algorithm 1 recovers the global minimizer with interpolation ($k = 1$) and allows clients to reach a consensus.

7.1 Non-stationary and Heterogeneous Unavailability with $P_\delta > 1$

We start from a general case where the length of the sliding window $P_\delta > 1$ in Assumption 1.

Data sets and data heterogeneity. We perform the experiments on SVHN (Netzer et al., 2011) and CIFAR-10 (Krizhevsky et al., 2009) data sets. Similar to Proposition 14, we divide clients into two groups $\mathcal{M}_1$ and $\mathcal{M}_2$, each with 50 clients. Each group of clients collectively hold 5 classes of images from the original data set, non-overlapping with the other group. To emulate highly heterogeneous local data distributions within each group, the images are



(a) SVHN data set. Heterogeneous but stationary intra-period probability, detailed in Fig. 6e.

(b) SVHN data set. Heterogeneous and nonstationary intra-period p_i^t, detailed in Fig. 6f.

(c) CIFAR-10 data set. Heterogeneous but stationary intra-period probability, detailed in Fig. 6e.

(d) CIFAR-10 data set. Heterogeneous and nonstationary intra-period p_i^t, detailed in Fig. 6f.

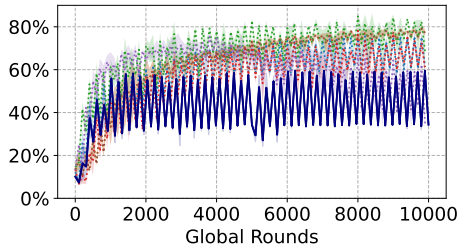
(e) An illustration of the client unavailability dynamics in Figs. 6a and 6c. The two client groups $\mathcal{M}_1$ and $\mathcal{M}_2$ are available alternately, where p_i's remain static within each available period.

(f) An illustration of the client unavailability dynamics in Figs. 6b and 6d. The two client groups are available alternately, where p_i^t's have a staircase pattern within each available window.

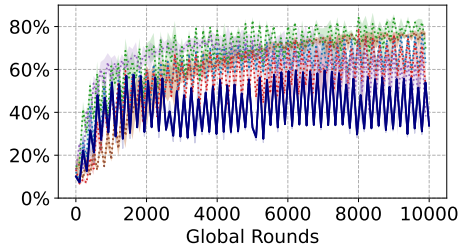
Figure 6: Plots of the consensus errors on a logarithmic scale with $m = 100$ clients. The 100 clients are evenly divided into two non-overlapping groups $\mathcal{M}_1$ and $\mathcal{M}_2$. The global rounds can be partitioned into periods, each with length 50 rounds. In all odd periods, the available probability $p_i^t > 0$ if $i \in \mathcal{M}_1$ while $p_i^t = 0$ otherwise. In all even periods, the available probability $p_i^t > 0$ if $i \in \mathcal{M}_2$ while $p_i^t = 0$ otherwise. We train CNN networks on the SVHN data set and CIFAR-10 data set for $T = 10000$ and $T = 20000$ rounds, respectively. When a client becomes available, it performs $s = 10$ steps of local computation. The results are obtained under 3 random seeds and sampled every P_δ round. The solid curves differ from each other by the values of k, and the shaded areas plot the standard deviation. It can be observed from the plots that clients fail to reach a consensus when $k = 0$.

assigned to individual clients i according to $\nu_i \sim \mathsf{Dirichlet}(\alpha = 0.1)$ (Hsu et al., 2019; Wang and Ji, 2022, 2024).

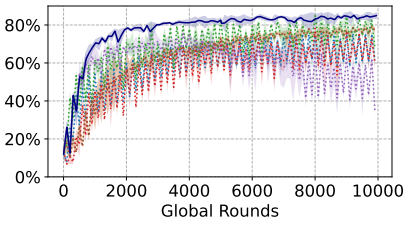
Non-stationary client unavailability with $P_\delta > 1$. We evaluate two non-stationary unavailable dynamics—static and staircase probabilistic trajectories—both with $P_\delta = 100$. Illustrative plots can be found in Figs. 6e and 6f. The non-stationary dynamics are motivated by real-world federated learning participation statistics and by generalizing the existing



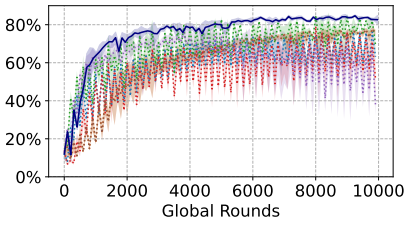
(a) No interpolation ($k = 0$). Heterogeneous but stationary intra-period probability p_i (Fig. 7e).



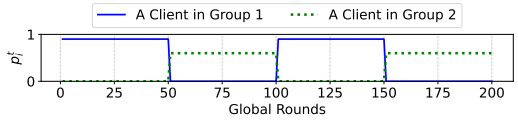
(b) No interpolation ($k = 0$). Heterogeneous and non-stationary intra-period probability p_i^t (Fig. 7f).



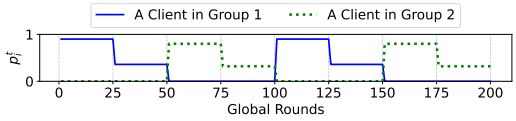
(c) With interpolation ($k = 100$). Heterogeneous but stationary intra-period probability p_i (Fig. 7e).



(d) With interpolation ($k = 100$). Heterogeneous and non-stationary intra-period probability p_i^t (Fig. 7f).



(e) The two client groups are available alternately; however, the clients' p_i's remain static within each available window.



(f) The two client groups are available alternately; however, the clients' p_i^t's show a staircase pattern within each available window.

Figure 7: Test accuracy results with $m = 100$ clients, who are divided into two non-overlapping but evenly sized groups. We train CNN networks on the SVHN data set for $T = 10000$ rounds. When a client becomes available, it performs $s = 10$ steps of local computation. The results are obtained under 3 random seeds and sampled every P_δ round. The curves plot the averaged results, while the shaded areas plot the standard deviation.

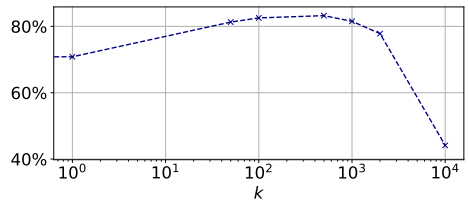
participation patterns, such as cyclic participation (Cho et al., 2023a; Wang and Ji, 2024). Formally, let $f_i(t)$ be a time-dependent function under the specific non-stationary dynamics, and $p_i = \langle \nu_i, \phi \rangle$, where $\nu_i \sim \mathsf{Dirichlet}(\alpha = 0.1)$, and ϕ characterizes the unbalanced contribution of different image classes to the generated probabilities. The unavailability dynamics of client group $\mathcal{M}_1$ and $\mathcal{M}_2$ are illustrated in (31) and in (32), respectively.

For $i \in \mathcal{M}_1$, we have

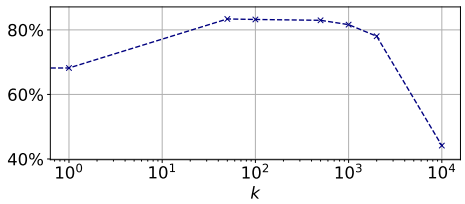
$$p_i^t = \begin{cases} p_i \cdot f_i(t), & \text{if } (t \bmod P_\delta) \leq \frac{P_\delta}{2}; \\ 0, & \text{otherwise.} \end{cases} \tag{31}$$

For $j \in \mathcal{M}_2$, we have

$$p_j^t = \begin{cases} p_j \cdot f_j(t), & \text{if } (t \bmod P_\delta) > \frac{P_\delta}{2}; \\ 0, & \text{otherwise.} \end{cases} \tag{32}$$

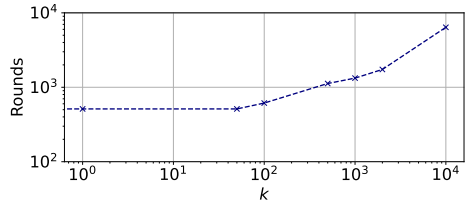


(a) Heterogeneous but stationary intra-period probability p_i (Fig. 7e).

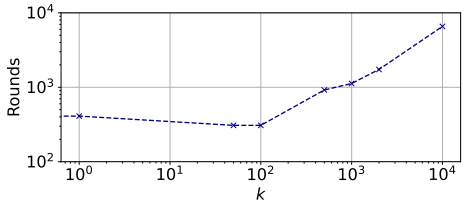


(b) Heterogeneous and non-stationary intra-period probability p_i^t (Fig. 7f).

Figure 8: Test accuracy of `FedSWE` with different k's on SVHN data set under different unavailability dynamics. The reported results are averaged over the last 500 rounds. Consistent with Remark 13, we observe an initial increase in test accuracy, followed by a decline, peaking around $k = m = 100$.



(a) Heterogeneous yet stationary intra-period probability p_i (Fig. 7e).



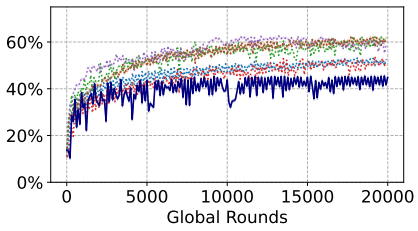
(b) Heterogeneous and non-stationary intra-period probability p_i^t (Fig. 7f).

Figure 9: The number of rounds needed to achieve 40% test accuracy of `FedSWE` with various k values on the SVHN data set under different unavailability dynamics. We can observe a similar trend as in Fig. 8 that a slight speedup in the beginning but a significant slowdown after $k = m$.

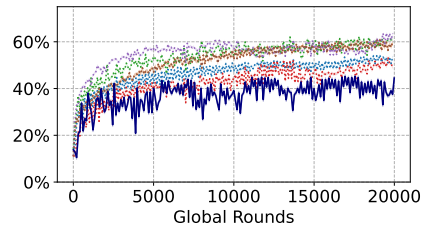
Each element of $[\phi]_c$ is drawn from $\mathsf{Uniform}(0, \mathbf{\Phi}_c)$, where a smaller $\mathbf{\Phi}_c$ leads to a less significant contribution of that image class. It is immediately clear that the coupling of local data distribution $\nu_i \sim \mathsf{Dirichlet}(\alpha = 0.1)$ and class contribution ϕ leads to *non*-independent p_i's. Although the non-independence setup violates our theoretical analysis, we observe that `FedSWE` retains its outperformance. Correlating the local data distribution and the probability of client availability is a common practice in the prior literature. For example, Gu et al. (2021) experiment with a formula for p_i so that clients that hold images of smaller digits participate less frequently. Wang and Ji (2024) construct p_i as an inner product of the clients' local data distribution ν_i and an external distribution $\mathbf{\Phi}'$.

We highlight that the periodic unavailability dynamics evaluated in our work are more challenging, e.g., than (Wang and Ji, 2022), where they select a fixed number of S clients out of the available client group to participate in each round t uniformly at random. In our work, we may have fewer than S available clients in any round t due to the heterogeneity in the base probability p_i and randomness in the Bernoulli sampling process; therefore, a fixed size of sampling clients in each round is not guaranteed.

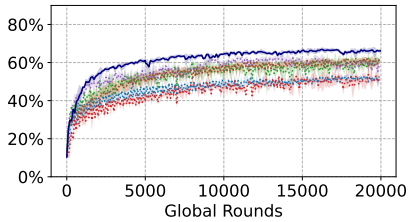
Benchmark algorithms. We compare `FedSWE` with six baseline algorithms, including `FedAvg` over active clients (McMahan et al., 2017), `gFedAvg` (Wang and Ji, 2022), `FedAvg` with known probability (`FedKnown`) (Perazzone et al., 2022), `FedAU` (Wang and Ji,



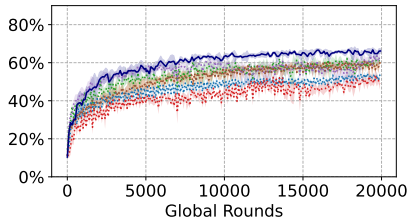
(a) No interpolation ($k = 0$). Heterogeneous but stationary intra-period probability p_i (Fig. 10e).



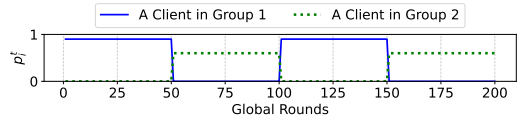
(b) No interpolation ($k = 0$). Heterogeneous and non-stationary intra-period probability p_i^t (Fig. 10f).



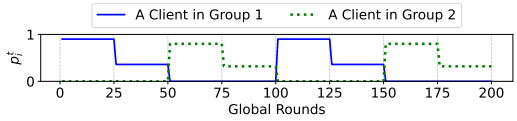
(c) With interpolation ($k = 100$). Heterogeneous but stationary intra-period probability p_i (Fig. 10e).



(d) With interpolation ($k = 100$). Heterogeneous and non-stationary intra-period probability p_i^t (Fig. 10f).



(e) The two client groups are available alternately; however, the clients' p_i's remain static within each available window.

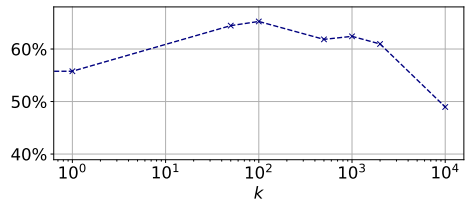


(f) The two client groups are available alternately; however, the clients' p_i^t's show a staircase pattern within each available window.

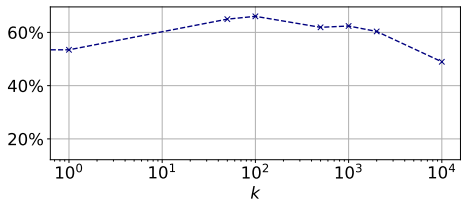
Figure 10: Test accuracy results with $m = 100$ clients, who are divided into two non-overlapping but evenly sized groups. Only a single client group is available in each training round. Inside an available window of P rounds, a client i in each group is available with probability p_i^t. We train CNN networks on the CIFAR-10 data set for $T = 20000$ rounds. When a client becomes available, it performs $s = 10$ steps of local computation. The results are obtained under 3 random seeds and sampled every P round. The curves plot the averaged results, while the shaded areas plot the standard deviation.

2024), `MIFA` (Gu et al., 2021) and `FedVARP` (Jhunjhunwala et al., 2022). The details of the algorithms are deferred to Appendix H.

Necessity of interpolation ($k > 0$). Recall that we show in Proposition 14 that interpolation is necessary for information diffusion over rounds under periodic unavailability. To validate such a claim, we show in Fig. 6 that clients fail to reach a consensus when $k = 0$. Specifically, instead of decaying, we observe that the consensus errors blow up in the plots. In contrast, the interpolation carries global updates from round to round and eventually allows clients to correct bias. Furthermore, as k increases, we observe a smaller consensus error,

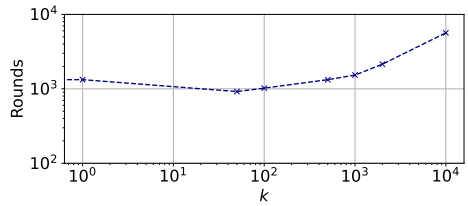


(a) Heterogeneous but stationary intra-period probability p_i (Fig. 10e).

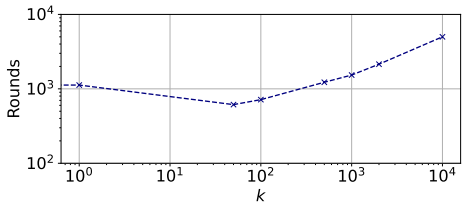


(b) Heterogeneous and non-stationary intra-period probability p_i^t (Fig. 10f).

Figure 11: Test accuracy of `FedSWE` with different k's on CIFAR-10 data set under different unavailability dynamics. The reported results are averaged over the last 500 rounds. Consistent with Remark 13, we observe an initial increase in test accuracy, followed by a decline, peaking around $k = m = 100$.



(a) Heterogeneous yet stationary intra-period probability p_i (Fig. 10e).



(b) Heterogeneous and non-stationary intra-period probability p_i^t (Fig. 10f).

Figure 12: The number of rounds needed to achieve 40% test accuracy of `FedSWE` with various k values on CIFAR-10 data set under different unavailability dynamics. We can observe a similar trend as in Fig. 11 that a slight speedup in the beginning but a significant slowdown after $k = m$.

which implies better client connectivity. Yet, as we will show next, excessively increasing k will inevitably lead to an unnecessary slowdown in convergence.

Performance discussions. We can observe from Figs. 7b, 7a, 10a and 10b that most of the algorithms, including our conference version: `FedSWE` without interpolation ($k = 0$), suffer from the challenging periodic non-stationary dynamics and experience high fluctuations. As we have illustrated in Proposition 14, `FedSWE` without interpolation fails to diffuse information across rounds. For baseline algorithms on the SVHN data set, `FedKnown` attains the best peak accuracy, while `FedVARP` and `gFedAvg` yield the smoothest curves yet with less accurate predictions. In sharp contrast, `FedSWE` with $k = 100$ generates relatively smooth trajectories and outperforms all the baseline algorithms on both SVHN (Figs. 7c and 7d) and CIFAR-10 data sets (Figs. 10c and 10d). For the SVHN data set, `FedSWE` with $k = 100$ attains comparable accuracy as the peak accuracy of `FedKnown` yet with more consistent performance. For the CIFAR-10 data set, `FedSWE` with $k = 100$ obtains even better accuracy than `FedKnown`. It is worth noting that, despite the impressive empirical performance of `FedKnown`, its analysis requires strictly positive probabilities (Perazzone et al., 2022), and its implementation adopts p_i^t's as a known priori. `FedAU` is provably robust to stationary dynamics; yet, we consider non-stationary availability here. It is a bit surprising

that FedSWE surpasses MIFA and FedVARP, which both require heavy memory of size $O(md)$; however, the interpolation in FedSWE only requires light memory of size $O(d)$. Under the periodic unavailability dynamics, the clients in the inactive client group can be unavailable for quite a long period of global rounds; therefore, their gradients from the most recent availability may not be a good approximation of their latest fresh gradients if they are available.

Effects of interpolation coefficient k. Figs. 8 and 11 compare the test accuracies of FedSWE under different choices of k. The results show that increasing k is not always beneficial. On the one hand, FedSWE converges faster as k approaches m, but it undergoes a substantial slowdown when k continues to rise, as shown by the number of rounds required to reach 40% test accuracy in Figs. 9 and 12. On the other hand, the final test accuracy increases and then drops in Figs. 8 and 11. This aligns with our discussions in Remark 13, where we show that the convergence upper bound minimizes at a unique point. However, the exact analytical form depends on parameters that are difficult, if not impossible, to obtain in practice. Based on the results, we empirically recommend $k = \Theta(m)$ to strike a balance between performance and convergence speed.

7.2 Non-stationary and Heterogeneous Unavailability with $P_\delta = 1$

In this section, we investigate the special case of Assumption 1, where $P_\delta = 1$ as in (Xiang et al., 2024), with interpolation providing an added benefit. Our numerical results are presented in Table 2, where we study additional availability dynamics and data sets. The results are partitioned into two parts, where the latter part includes algorithms aided by heavy memory or known statistics, such as MIFA, FedVARP and FedKnown.

Non-stationary client unavailability with $P_\delta = 1$. We study a total of three client unavailability dynamics in Table 2, including stationary, staircase, and sine probabilistic trajectories. Their visualizations are also available in the same table. Our choices of non-stationary dynamics are motivated by real-world federated learning participation statistics (Bonawitz et al., 2019; Ribero et al., 2022). The learning tasks become more challenging as the list progresses due to the growing complexity of the non-stationary dynamics.

Mathematically, similar to the construction—(31) and (32)—in Section 7.1 but without alternate client group availability, client i's dynamics is defined as $p_i^t = p_i \cdot f_i(t)$. The definitions of p_i^t and $f_i(t)$ can be found therein. Note that the p_i's remain to be *non*-independent across different clients, but we observe that FedSWE retains its outperformance.

Performance discussions. In addition to the baselines in Section 7.1, we include the evaluation results on FedAvg over all clients, F3AST algorithm (Ribero et al., 2022) and on CINIC-10 data set (Darlow et al., 2018). To understand the nuances in the performance of our conference version (Xiang et al., 2024) and this journal extension, we also compare the performance of FedSWE without ($k = 0$) and with interpolation ($k = 100$).

It is observed that FedSWE consistently outperforms the algorithms not aided by heavy memory or known statistics. In particular, FedSWE with interpolation attains better accuracies than its non-interpolation variant (Xiang et al., 2024) on almost all tasks, providing added benefits. In the only exception (stationary on SVHN data set), their performances are close, with accuracy differences of less than 1%. We also surprisingly observe that FedSWE occasionally beats MIFA, which is memory heavy. We attribute it to its reuse of stored gradients

Table 2: Results and comparisons on real-world data sets 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. For a fair competition, we **boldface** the best accuracy in the first group, while the second best is underlined.

Unavailable Dynamics	Data sets / Algorithms	SVHN Train	SVHN Test	CIFAR-10 Train	CIFAR-10 Test	CINIC-10 Train	CINIC-10 Test
Stationary p_i	FedSWE (**ours**, $k=0$)	**86.5** ± 0.7 %	**86.1** ± 0.7 %	68.1 ± 1.4 %	66.3 ± 1.1 %	47.9 ± 2.1 %	47.3 ± 2.0 %
	FedSWE (**ours**, $k=100$)	86.3 ± 1.1%	85.4 ± 1.0%	**68.7** ± 1.0 %	**66.9** ± 0.9 %	**48.3** ± 1.7 %	**47.9** ± 1.6 %
	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 %
	gFedAvg	83.5 ± 1.5 %	83.0 ± 1.2%	64.5 ± 1.7%	63.8 ± 1.7%	45.3 ± 1.4%	44.9 ± 1.2%
	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**) p_i^t	FedSWE (**ours**, $k=0$)	85.9 ± 0.8 %	85.6 ± 1.0 %	67.7 ± 1.3 %	66.0 ± 1.2 %	47.5 ± 2.0 %	46.9 ± 2.0 %
	FedSWE (**ours**, $k=100$)	**86.0** ± 1.2 %	**85.9** ± 0.7 %	**67.8** ± 1.0 %	**66.1** ± 1.2 %	**47.7** ± 1.5 %	**47.1** ± 1.4 %
	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 %
	gFedAvg	83.1 ± 1.3 %	83.0 ± 1.1 %	65.0 ± 1.6 %	64.9 ± 1.5 %	45.3 ± 1.4 %	44.9 ± 1.3 %
	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**) p_i^t	FedSWE (**ours**, $k=0$)	85.7 ± 0.9 %	85.6 ± 0.9 %	64.9 ± 1.9 %	63.5 ± 2.0 %	46.4 ± 2.4 %	45.8 ± 2.4 %
	FedSWE (**ours**, $k=100$)	**85.9** ± 1.2 %	**85.8** ± 0.8 %	**65.8** ± 1.8 %	**64.2** ± 1.9 %	**47.2** ± 2.0 %	**46.7** ± 1.8 %
	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 %
	gFedAvg	83.6 ± 1.3%	83.5 ± 1.1 %	62.4 ± 1.2 %	62.2 ± 1.2 %	43.3 ± 1.1 %	42.9 ± 1.0 %
	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 %

from the unavailable clients. Although FedSWE brings in staleness due to implicit gossiping, our results in Section 7.1 for $k \geq 0$ and Table 7 in Appendix H for $k = 0$ indicate that there is no significant slowdown for FedSWE when compared to the baseline algorithms. Furthermore, FedSWE attains competitive or even better performance than FedAvg with known probability, yet completely unknown to the underlying dynamics in client unavailability.

8. Conclusion

In this paper, we have shown that the significant impacts of heterogeneous and non-stationary client unavailability on learning performance through FedAvg. To address this, we have proposed an algorithm FedSWE, which provably converges to a stationary point of the global objective by adaptively echoing clients' local improvements, by interpolating updates across rounds via a global moving average, and by evenly diffusing local updates through implicit gossiping. Notably, it achieves the desired linear speedup property in certain special cases. Experiments have validated the superiority of FedSWE over state-of-the-art algorithms under diversified non-stationary dynamics. Future work will investigate how to relax the assumption of independence in client availability.

Acknowledgments

We gratefully acknowledge the support from the National Science Foundation under grants 2106891, 2107062, and the National Science Foundation CAREER award under grant 2340482. 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 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.

References

Kwangjun Ahn and Ashok Cutkosky. Adam with model exponential moving average is effective for nonconvex optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=v416YLOQuU.

Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, and John Stephan. Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity. *arXiv preprint arXiv:2302.01772*, 2023.

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

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.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Gossip algorithms: Design, analysis and applications. In *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, volume 3, pages 1653–1664. IEEE, 2005.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

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.

Wenlin Chen, Samuel Horváth, and Peter Richtárik. Optimal client sampling for federated learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=8GvRCWKHIL.

Ziheng Cheng, Xinmeng Huang, Pengfei Wu, and Kun Yuan. Momentum benefits non-iid federated learning simply and provably. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=TdhkAcXkRi.

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.

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 of *Proceedings of Machine Learning Research*, pages 5677–5721. PMLR, 23–29 Jul 2023a.

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*, 2023b.

Michael Crawshaw and Mingrui Liu. Federated learning under periodic client participation and heterogeneous data: A new communication-efficient algorithm and analysis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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.

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

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.

Allan Gut. *Probability: a graduate course*, volume 200. Springer, 2006.

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.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification, 2019.

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. URL `https://openreview.net/forum?id=Xm9iN3UsdpH`.

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.

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.

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.

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

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.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

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.

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

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36:52166–52196, 2023.

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.

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, 2020a.

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*, 2020b. URL `https://openreview.net/forum?id=HJxNAnVtDS`.

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.

Nancy A. Lynch. *Distributed Algorithms*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1996. ISBN 9780080504704.

Connor J. McLaughlin and Lili Su. Personalized federated learning via feature distribution adaptation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 77038–77059. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/8ce6c5450ccddbe6adee4b3749893587-Paper-Conference.pdf`.

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.

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.

Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2014.

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

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.

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.

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*, 2011. URL `http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf`.

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.

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.

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.

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.

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.

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

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

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.

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.

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

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.

Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Zhenyu Sun, ziyang zhang, Zheng Xu, Gauri Joshi, Pranay Sharma, and Ermin Wei. Debiasing federated learning with correlated client participation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=9h45qxXEx0`.

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.

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

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.

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.

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.

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. URL `https://openreview.net/forum?id=qSs7C7c4G8D`.

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. URL `https://openreview.net/forum?id=ZKEuFKfCKA`.

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.

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.

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. doi: 10.1109/CDC49753.2023.10383258.

Ming Xiang, Stratis Ioannidis, Edmund Yeh, Carlee Joe-Wong, and Lili Su. Efficient federated learning against heterogeneous and non-stationary client unavailability. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Ming Xiang, Stratis Ioannidis, Edmund Yeh, Carlee Joe-Wong, and Lili Su. Empowering federated learning with implicit gossiping: Mitigating connection unreliability amidst unknown and arbitrary dynamics. *IEEE Transactions on Signal Processing*, 73:766–780, 2025. doi: 10.1109/TSP.2025.3526782.

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

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.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203*, 2021.

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

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.

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. doi: 10.1109/JSTSP.2022.3152445.

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, 2019a.

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, 2019b.

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

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. URL `https://openreview.net/forum?id=_33ynl9VgCX`.

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.

Chen Zhu, Zheng Xu, Mingqing Chen, Jakub Konečnỳ, Andrew Hard, and Tom Goldstein. Diurnal or nocturnal? federated learning of multi-branch networks from periodically shifting distributions. In *International Conference on Learning Representations*, 2022.

Appendices

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

Appendix A. Nomenclature

In this section, we provide the notations and nomenclatures used throughout our proofs for a comprehensive presentation.

Table 3: Nomenclature table

Notation(s)	Definition
$\mathcal{A}^t$	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 (9).
p_i^t	The probability that a client i becomes available in round t.
$\tau_i(t)$	$\tau_i(t) \triangleq \sup\{t' \mid t' < t, i \in \mathcal{A}^{t'}\}$ 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 4 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 4: Variable table

δ	An absolute constant that is the lower bound on the client unavailability.
P_δ, P	The period in Assumption 1.
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.
k	The interpolation coefficient in the global moving average procedure.
m	The number of clients in the federated learning system.
M	$M = m + k$.

Appendix B. 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.

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{a}_i \right\|_2^2 \leq \frac{1}{m} \sum_{i=1}^{m} \|\boldsymbol{a}_i\|_2^2 \quad \text{and} \quad \left\| \sum_{i=1}^{m} \boldsymbol{a}_i \right\|_2^2 \leq m \sum_{i=1}^{m} \|\boldsymbol{a}_i\|_2^2 . \tag{33}$$

2. Young's inequality (a.k.a. Peter-Paul inequality).

$$\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle \leq \frac{\|\boldsymbol{a}_1\|_2^2}{2\epsilon} + \frac{\epsilon \|\boldsymbol{a}_2\|_2^2}{2}, \quad \text{for any } \epsilon > 0. \tag{34}$$

Equivalently, we have

$$\begin{aligned} \|\boldsymbol{a}_1 + \boldsymbol{a}_2\|_2^2 &= \|\boldsymbol{a}_1\|_2^2 + \|\boldsymbol{a}_2\|_2^2 + 2\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle \\ &\leq \left(1 + \frac{1}{\epsilon}\right) \|\boldsymbol{a}_1\|_2^2 + (1+\epsilon) \|\boldsymbol{a}_2\|_2^2, \quad \text{for any } \epsilon > 0. \end{aligned} \tag{35}$$

3. Smoothness corollary. *Given Assumption 2, it holds that*

$$\begin{aligned} F(\boldsymbol{a}_1) - F(\boldsymbol{a}_2) &= \left\langle \boldsymbol{a}_1 - \boldsymbol{a}_2, \int_0^1 \nabla F(\boldsymbol{a}_2 + \tau(\boldsymbol{a}_1 - \boldsymbol{a}_2)) \mathrm{d}\tau \right\rangle \\ &= \langle \nabla F(\boldsymbol{a}_2), \boldsymbol{a}_1 - \boldsymbol{a}_2 \rangle + \int_0^1 \langle \boldsymbol{a}_1 - \boldsymbol{a}_2, \nabla F(\boldsymbol{a}_2 + \tau(\boldsymbol{a}_1 - \boldsymbol{a}_2)) - \nabla F(\boldsymbol{a}_2) \rangle \, \mathrm{d}\tau \\ &\overset{(a)}{\leq} \langle \nabla F(\boldsymbol{a}_2), \boldsymbol{a}_1 - \boldsymbol{a}_2 \rangle + L \int_0^1 \tau \left\| \boldsymbol{a}_1 - \boldsymbol{a}_2 \right\|_2 \left\| (\boldsymbol{a}_1 - \boldsymbol{a}_2) \right\|_2 \mathrm{d}\tau \\ &\leq \langle \nabla F(\boldsymbol{a}_2), \boldsymbol{a}_1 - \boldsymbol{a}_2 \rangle + \frac{L}{2} \left\| \boldsymbol{a}_1 - \boldsymbol{a}_2 \right\|_2^2, \end{aligned} \tag{36}$$

where (a) follows from Cauchy-Schwartz inequality and Assumption 2.

Appendix C. Descent Lemma (Lemma 7)

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

C.1 Multi-step perturbation

Lemma 15. Suppose that Assumption 2 and Assumption 3 hold. For each regular client $i \in \mathcal{R}$, we have

$$\mathbb{E}\left[\left\| \sum_{r=0}^{s-1} \nabla F_i(\boldsymbol{x}_i^{(t,r)}) - \nabla F_i(\boldsymbol{x}_i^t) \right\|_2^2 \;\middle|\; \mathcal{F}^t \right] \leq 5\eta_l^2 s^3 L^2 \sigma^2 + 20\eta_l^2 s^4 L^2 \left\| \nabla F_i(\boldsymbol{x}_i^t) \right\|_2^2$$

Proof Proof of Lemma 15 The proof shares a similar road map to (Yang et al., 2021, Lemma 2), but the objective is instead to show an upper bound with respect to $\left\| \nabla F_i(\boldsymbol{x}_i^t) \right\|_2^2$.

It holds that

$$\begin{aligned} \mathbb{E}\left[\left\| \sum_{r=0}^{s-1} \nabla F_i(\boldsymbol{x}_i^{(t,r)}) - \nabla F_i(\boldsymbol{x}_i^t) \right\|_2^2 \right] &\overset{(a)}{\leq} s \sum_{r=0}^{s-1} \mathbb{E}\left[\left\| \nabla F_i(\boldsymbol{x}_i^{(t,r)}) - \nabla F_i(\boldsymbol{x}_i^t) \right\|_2^2 \;\middle|\; \mathcal{F}^t \right] \\ &\overset{(b)}{\leq} sL^2 \sum_{r=0}^{s-1} \mathbb{E}\left[\left\| \boldsymbol{x}_i^{(t,r)} - \boldsymbol{x}_i^t \right\|_2^2 \;\middle|\; \mathcal{F}^t \right], \end{aligned} \tag{37}$$

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)}{\le}\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]\\
&\le\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)}{\le}\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 $\eta_l \le \frac{1}{4sL}$, it holds that

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

Unroll the recursion, we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t,r)}-\boldsymbol{x}_i^t\right\|_2^2\,\middle|\,\mathcal{F}^t\right]&\le\sum_{k=0}^{r-1}\left(1+\frac{2}{2s-1}\right)^k\left(\eta_l^2\sigma^2+4s\eta_l^2\left\|\nabla F_i^t\right\|_2^2\right)\\
&\le\sum_{k=0}^{s-1}\left(1+\frac{2}{2s-1}\right)^k\left(\eta_l^2\sigma^2+4s\eta_l^2\left\|\nabla F_i^t\right\|_2^2\right)\\
&=\frac{2s-1}{2}\left[\left(1+\frac{2}{2s-1}\right)^{s-\frac{1}{2}}\left(1+\frac{2}{2s-1}\right)^{\frac{1}{2}}-1\right]\left(\eta_l^2\sigma^2+4s\eta_l^2\left\|\nabla F_i^t\right\|_2^2\right)\\
&\overset{(f)}{\le}\left(s-\frac{1}{2}\right)\left[\sqrt{3}e-1\right]\left(\eta_l^2\sigma^2+4s\eta_l^2\left\|\nabla F_i^t\right\|_2^2\right)\\
&\overset{(g)}{\le}5s\eta_l^2\sigma^2+20s^2\eta_l^2\left\|\nabla F_i^t\right\|_2^2,
\end{aligned}
$$

where inequality (f) holds because of $(1+1/x)^x < \exp(1)$, inequality (g) holds because of $\sqrt{3}\exp(1) - 1 < 5$. Plug it back into (37), we have the desired result

$$\mathbb{E}\left[\left\|\sum_{r=0}^{s-1}\nabla F_i(\boldsymbol{x}_i^{(t,r)}) - \nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2 \;\middle|\; \mathcal{F}^t\right] \leq 5\eta_l^2 s^3 L^2\sigma^2 + 20\eta_l^2 s^4 L^2 \left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2.$$

■

C.2 Descent lemma

Recall that we have defined an auxiliary global objective $\widetilde{F}(\boldsymbol{x}) = \frac{1}{M}\sum_{i\in\mathcal{R}\cup\mathcal{V}} F_i(\boldsymbol{x})$.

Proof Proof of Lemma 7 By Assumption 2 and inequality (36), we have

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

Recall that $M \triangleq m + k = |\mathcal{R}\cup\mathcal{V}|$. The one-round innovation of $\bar{\boldsymbol{z}}$ can be rewritten as

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

where equality (a) using the fact that $\boldsymbol{x}_i^k = \boldsymbol{x}_i^t$ for all k such that $\tau_i(t)+1 \leq k \leq t$, 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 $\mathbb{1}_{\{i\in\mathcal{A}^t\}} + \mathbb{1}_{\{i\in\mathcal{R}\setminus\mathcal{A}^t\}} = 1$ since a client $i \in \mathcal{V}$ does not update gradients.

Bounding (A)**.**

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

Bounding (A.I)

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

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)

$$
\begin{aligned}
(\mathrm{A.II}) &\overset{(c)}{\leq} \frac{\eta_l \eta_g}{M} \sum_{i=1}^m \mathbb{1}_{\{i\in\mathcal{A}^t\}} \sum_{p=-1}^{t-1} \mathbb{1}_{\{\tau_i(t)=p\}} \left(\frac{s}{12} \left\| \nabla \widetilde{F}(\bar{\boldsymbol{z}}^t) \right\|_2^2 + \frac{3(t-p)^2}{s} \left\| \sum_{r=0}^{s-1} \nabla F_i(\boldsymbol{x}_i^t) - \nabla F_i(\boldsymbol{x}_i^{(t,r)}) \right\|_2^2 \right) \\
&= \frac{\eta_l \eta_g s}{12M} \sum_{i=1}^m \mathbb{1}_{\{i\in\mathcal{A}^t\}} \left\| \nabla \widetilde{F}(\bar{\boldsymbol{z}}^t) \right\|_2^2 \\
&\quad + \frac{\eta_l \eta_g}{M} \sum_{i=1}^m \mathbb{1}_{\{i\in\mathcal{A}^t\}} \sum_{p=-1}^{t-1} \mathbb{1}_{\{\tau_i(t)=p\}} \frac{3(t-p)^2}{s} \left\| \sum_{r=0}^{s-1} \nabla F_i(\boldsymbol{x}_i^t) - \nabla F_i(\boldsymbol{x}_i^{(t,r)}) \right\|_2^2,
\end{aligned}
$$

where inequality (c) holds because of Young's inequality. We further have:

$$\begin{aligned}\mathbb{E}\left[(\text{A.II})\middle|\mathcal{F}^t\right] &\overset{(d)}{\leq} \frac{\eta_l\eta_g s}{12}\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2 + \frac{15\eta_g\eta_l^3 s^2L^2\sigma^2}{M}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 \\ &\quad + \frac{60\eta_g\eta_l^3 s^3L^2}{M}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2\left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2 \\ &= \frac{\eta_l\eta_g s}{12}\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2 + \frac{15\eta_g\eta_l^3 s^2L^2\sigma^2}{M}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 \\ &\quad + \frac{60\eta_g\eta_l^3 s^3L^2}{M}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2,\end{aligned}$$

where inequality (d) holds because of Lemma 15, 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)**.**

$$(\text{A.III}) = \frac{\eta_l\eta_g s}{M}\sum_{i=1}^m\left\langle\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t), \nabla F_i(\boldsymbol{z}_i^t) - \nabla F_i(\boldsymbol{x}_i^t)\right\rangle \overset{(e)}{\leq} \frac{\eta_l\eta_g s}{12}\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2 + \frac{3\eta_l\eta_g sL^2}{M}\sum_{i=1}^m\left\|\boldsymbol{z}_i^t - \boldsymbol{x}_i^t\right\|_2^2,$$

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

$$\mathbb{E}\left[(\text{A.III})\middle|\mathcal{F}^t\right] \leq \frac{\eta_l\eta_g s}{12}\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2 + \frac{3\eta_l\eta_g sL^2}{M}\sum_{i=1}^m\left\|\boldsymbol{z}_i^t - \boldsymbol{x}_i^t\right\|_2^2.$$

Bounding (A.IV)

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

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

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

where the first equality holds because a client $i \in \mathcal{V}$ does not update gradients, and it holds that

$$\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t) \triangleq \frac{1}{M}\sum_{i\in\mathcal{R}}\nabla F_i(\bar{\boldsymbol{z}}^t) + \frac{1}{M}\sum_{i\in\mathcal{V}}\nabla F_i(\bar{\boldsymbol{z}}^t) = \frac{1}{M}\sum_{i=1}^m\nabla F_i(\bar{\boldsymbol{z}}^t),$$

where we use the convention that $\nabla F_i(\bar{\boldsymbol{z}}^t) = \mathbf{0}$ for $i \in \mathcal{V}$. Putting (A) together,

$$\begin{aligned}\mathbb{E}\left[(\mathrm{A})\middle|\mathcal{F}^t\right] \leq & -\frac{\eta_l\eta_g s}{3}\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2 + \frac{15\eta_g\eta_l^3 s^2L^2\sigma^2}{M}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 \\ & + \frac{3\eta_l\eta_g sL^2}{M}\sum_{i=1}^m\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2 + \frac{\eta_l\eta_g sL^2}{2M}\sum_{i=1}^m\left\|\bar{\boldsymbol{z}}^t - \boldsymbol{z}_i^t\right\|_2^2 \\ & - \frac{\eta_l\eta_g s}{2}\left\|\frac{1}{M}\sum_{i=1}^m\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2 + \frac{60\eta_g\eta_l^3 s^3L^2}{M}\sum_{i=1}^m\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2.\end{aligned}$$

Bounding (B)**.**

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

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

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

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 $p_i^t \leq \delta_{\max}$.

Bounding (B.II) We have,

$$\begin{aligned}\mathbb{E}\left[(\text{B.II})\middle|\mathcal{F}^t\right] &\leq 2L\frac{\eta_l^2\eta_g^2}{M}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 5\eta_l^2 s^3 L^2\sigma^2 \\ &\quad + 2L\frac{\eta_l^2\eta_g^2}{M}\sum_{i=1}^{m}\mathbb{1}_{\{\tau_i(t)=p\}}\sum_{p=-1}^{t-1}(t-p)^2 20\eta_l^2 s^4 L^2\left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2 \\ &= \frac{10\eta_g^2\eta_l^4 s^3 L^3\sigma^2}{M}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 \\ &\quad + \frac{40\eta_g^2\eta_l^4 s^4 L^3}{M}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2,\end{aligned}$$

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

Bounding (B.III)**.**

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

Putting (B) together, we get

$$\begin{aligned}\mathbb{E}\left[(\text{B})\middle|\mathcal{F}^t\right] &\leq \frac{2\eta_l^2\eta_g^2 sL\sigma^2}{M^2}\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 + \frac{10\eta_g^2\eta_l^4 s^3 L^3\sigma^2}{M}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 \\ &\quad + \frac{40\eta_g^2\eta_l^4 s^4 L^3}{M}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2 \\ &\quad + \frac{2\eta_l^2\eta_g^2 s^2 L^3}{M}\sum_{i=1}^{m}\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2 + 2L\eta_l^2\eta_g^2 s^2\left\|\frac{1}{M}\sum_{i=1}^{m}\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2.\end{aligned}$$

Now, everything:

$$
\begin{aligned}
\mathbb{E}\left[\widetilde{F}(\bar{\boldsymbol{z}}^{t+1})-\widetilde{F}(\bar{\boldsymbol{z}}^{t})\middle|\mathcal{F}^{t}\right] \leq & -\frac{\eta_l\eta_g s}{3}\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^{t})\right\|_2^2 \\
& -\frac{\eta_l\eta_g s}{2}\left(1-4L\eta_l\eta_g s\right)\left\|\frac{1}{M}\sum_{i=1}^{m}\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2 \\
& +\frac{2\eta_l^2\eta_g^2 sL\delta_{\max}\sigma^2}{M^2}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 \\
& +\frac{5\eta_g\eta_l^3 s^2L^2\left(3+2\eta_g\eta_l sL\right)\sigma^2}{M}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 \\
& +\eta_l\eta_g sL^2\left(3+2\eta_l\eta_g sL\right)\frac{1}{M}\sum_{i=1}^{m}\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2+\frac{\eta_l\eta_g sL^2}{2M}\sum_{i=1}^{m}\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2 \\
& +20\eta_g\eta_l^3 s^3L^2\left(3+2\eta_g\eta_l sL\right)\frac{1}{M}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2 \\
\leq & -\frac{\eta_l\eta_g s}{3}\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^{t})\right\|_2^2+\frac{2\eta_l^2\eta_g^2 sL\delta_{\max}\sigma^2}{M^2}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 \\
& +\frac{17\eta_g\eta_l^3 s^2L^2\sigma^2}{M}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2 \\
& +4\eta_l\eta_g sL^2\frac{1}{M}\sum_{i=1}^{M}\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2+\frac{\eta_l\eta_g sL^2}{2M}\sum_{i=1}^{M}\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2 \\
& +65\eta_g\eta_l^3 s^3L^2\frac{1}{M}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{1}_{\{\tau_i(t)=p\}}(t-p)^2\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2,
\end{aligned}
$$

where the last inequality holds because $\eta_l\eta_g \leq \frac{1}{8sL}$ and that $\left\|\frac{1}{M}\sum_{i=1}^{m}\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2 \geq 0$. ■

Appendix D. Intermediate Results

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

D.1 Bounding local and global dissimilarity

Proposition 16. For any t, it holds that

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

Proof Proof of Proposition 16

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

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

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

where the equality holds because $\nabla F(\boldsymbol{x})=\frac{M}{m}\nabla\widetilde{F}(\boldsymbol{x})$. ■

D.2 Weight re-equalization (Proposition 2)

Proof Proof of Proposition 2 We show Proposition 2 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 $K+1$:

$$\sum_{t=0}^{K}\mathbb{1}_{\{i\in\mathcal{A}^t\}}\left(t-\tau_i(t)\right)=\sum_{t=0}^{K-1}\mathbb{1}_{\{i\in\mathcal{A}^t\}}\left(t-\tau_i(t)\right)+\mathbb{1}_{\{i\in\mathcal{A}^K\}}\left(K-\tau_i(K)\right). \tag{38}$$

Now, we have two cases:

- Suppose $i\in\mathcal{A}^{K-1}$, then we simply have $\tau_i(K)=K-1$. It follows that (38) $\overset{(a)}{=}K+1$, 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 $i\in\mathcal{A}^{\tau_i(K)}$.

■

D.3 Unavailable statistics (Lemma 5)

Proof Proof of Lemma 5

$$\begin{aligned}\mathbb{E}\left[t-\tau_i(t)\right] &= \sum_{r=0}^{t}\mathbb{P}\left\{t-\tau_i(t)>r\right\} \\ &= \sum_{r=0}^{t}\prod_{r_1=t-r}^{t-1}(1-p_i^{r_1}) = 1+\sum_{r=1}^{t}\prod_{r_1=t-r}^{t-1}(1-p_i^{r_1}) \le 1+\sum_{r=1}^{\infty}\prod_{r_1=t-r}^{t-1}(1-p_i^{r_1}) \\ &= 1+\lim_{\ell\to\infty}\sum_{r=1}^{\ell P}\prod_{r_1=t-r}^{t-1}(1-p_i^{r_1})\end{aligned} \tag{39}$$

Let $t_0 \in \{0, P_\delta, 2P_\delta, \ldots\}$, it holds that

$$\prod_{t=t_0}^{t_0+P_\delta-1}(1-p_i^t) \overset{(a)}{\le} \left(\frac{P_\delta - \sum_{t=t_0}^{t_0+P_\delta-1}p_i^t}{P_\delta}\right)^{P_\delta} \le (1-\delta)^{P_\delta},$$

where inequality (a) holds because of AM-GM inequality. Also, it trivially holds that

$$\prod_{t=l}^{l'}(1-p_i^t) \le 1,$$

where $t_0 \le l \le l' \le t_0+P_\delta-1$. In general, we have

- When $r \ge P_\delta$, it holds that

$$\prod_{r_1=t_0-r}^{t_0-1}(1-p_i^{r_1}) \le \prod_{r_0=(t_0-r)/P_\delta}^{\lfloor t_0/P_\delta\rfloor-1}\ \prod_{r_1=r_0P_\delta}^{(r_0+1)P_\delta-1}(1-p_i^{r_1}) \le (1-\delta)^{P_\delta(t_0/P_\delta-\lceil(t_0-r)/P_\delta\rceil)} \le (1-\delta)^{r-P_\delta}.$$

- When $r < P_\delta$, it holds that

$$\prod_{r_1=t_0-r}^{t_0-1}(1-p_i^{r_1}) \le 1.$$

Hence,

$$\prod_{r_1=t_0-r}^{t_0-1}(1-p_i^{r_1}) \le (1-\delta)^{(r-P_\delta)\mathbb{1}_{\{r\ge P_\delta\}}}. \tag{40}$$

It holds for (39) that

$$\mathbb{E}\left[t-\tau_i(t)\right] \le 1+\sum_{r=1}^{\infty}(1-\delta)^{(r-P_\delta)\mathbb{1}_{\{r\ge P_\delta\}}} = P_\delta+\frac{1}{\delta}.$$

From (Gut, 2006, Theorem 12.3 (1)), we know that

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

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

$$\begin{aligned}\mathbb{E}\left[(t-\tau_i(t))^2\right] &\overset{\text{(a)}}{\leq} 2\sum_{r=1}^{\infty} r\,(1-\delta)^{(r-P_\delta)\mathbb{1}_{\{r\geq P_\delta\}}} \\ &\leq 2\left(\frac{P_\delta(P_\delta-1)}{2}+\frac{P_\delta}{\delta}+\frac{1-\delta}{\delta^2}\right) \\ &= \frac{1}{\delta^2}\left((P_\delta-1)\delta+1\right)^2+\frac{1}{\delta^2}\left((P_\delta-1)\delta^2+1\right), \end{aligned} \tag{41}$$

where inequality (a) holds because of (40), For a neat presentation, we use (var) as a shorthand notation in the following proofs for the constant in (41). ■

D.4 Auxiliary sequence construction and properties (Proposition 8)

Proposition 17. For any $t \geq 0$, 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 Proof of Proposition 17 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 4 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 $\boldsymbol{x}^{t+1} = \left(1/\left|\mathcal{A}^t\right|\right)\sum_{i\in\mathcal{A}^t}\boldsymbol{x}_i^{t++} = \left(1/\left|\mathcal{A}^t\right|\right)\sum_{i\in\mathcal{A}^t}\boldsymbol{z}_i^{t++} = \boldsymbol{z}^{t+1}$. 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 Proof of Proposition 8 From Propositions 17, 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 17.

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

where inequality (c) follows from re-indexing, inequality (d) from Lemma 5. ■

D.5 Consensus error of the auxiliary sequence

Lemma 18 (Consensus error of $\boldsymbol{z}_i^t$)**.** Assuming that $\eta_l\leq\delta/(20sL)$, and $\eta_l\eta_g\leq\delta(1-\sqrt{\rho})/(10sL(\sqrt{\rho}+1))$, under Assumptions 2, 3 and 4, it holds that

$$
\begin{aligned}
\frac{1}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right]&\leq\frac{32\eta_l^2\eta_g^2sP_\delta}{(1-\rho_k)^2}\frac{\left((P_\delta-1)\delta+1\right)^2+\left((P_\delta-1)\delta^2+1\right)}{\delta^2}\left(\frac{m\sigma^2}{M}\right)\\
&\quad+\frac{24\eta_l^2\eta_g^2s^2P_\delta(\beta^2+1)}{(1-\rho_k)^2}\frac{\left((P_\delta-1)\delta+1\right)^2+\left((P_\delta-1)\delta^2+1\right)}{\delta^2}\left(\frac{M}{m}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\\
&\quad+\frac{24\eta_l^2\eta_g^2s^2P_\delta}{(1-\rho_k)^2}\frac{\left[(P_\delta-1)\delta+1\right]^2+\left[(P_\delta-1)\delta^2+1\right]}{\delta^2}\left(\frac{m\zeta^2}{M}\right).
\end{aligned}
$$

Proof Proof of Lemma 18 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 $t\geq1$, recall that $W^{(t)}$ is a doubly stochastic matrix to characterize the information mixture, and that $\widetilde{\boldsymbol{G}}^t$ in (42) captures the local parameter changes in each round. Specifically,

for a client $i \in \mathcal{R}$, it holds that

$$\begin{aligned}\widetilde{\boldsymbol{G}}_i^t &\triangleq \mathbb{1}_{\{i\in\mathcal{A}^t\}}\left[(t-\tau_i(t))\sum_{r=0}^{s-1}\nabla\ell_i(\boldsymbol{x}_i^{(t,r)}) - s\left(t-1-\tau_i(t)\right)\nabla F_i(\boldsymbol{x}_i^{\tau_i(t)+1})\right]\\ &\quad + \mathbb{1}_{\{i\notin\mathcal{A}^t\}} s\nabla F_i(\boldsymbol{x}_i^{\tau_i(t)+1})\\ &= \mathbb{1}_{\{i\in\mathcal{A}^t\}}\left(t-\tau_i(t)\right)\sum_{r=0}^{s-1}\left(\nabla\ell_i(\boldsymbol{x}_i^{(t,r)}) - \nabla F_i(\boldsymbol{x}_i^t)\right) + s\nabla F_i(\boldsymbol{x}_i^t), \end{aligned} \tag{42}$$

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

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

Define $n(t) = \lfloor \frac{t}{P_\delta} \rfloor - 1$. For ease of presentation, we drop the variable t. We have $t - 2P_\delta < nP_\delta \le t - P_\delta$; therefore, $(n+1)P_\delta \le t < (n+2)P_\delta$. Further, $P_\delta \le t - nP_\delta < 2P_\delta$. Expanding $\boldsymbol{Z}$, we get

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

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

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

Assuming an absolute constant $\widetilde{\rho} \in (\rho_k, 1)$, it holds that

$$\begin{aligned}&\mathbb{E}_W\left[\|\boldsymbol{Z}^{(t)}\left(\mathbf{I}-\mathbf{J}\right)\|_{\mathrm{F}}^2\right] = \mathbb{E}_W\left[\|\boldsymbol{Z}^{nP_\delta}\prod_{\ell=nP_\delta}^{t-1}W^{\ell}\left(\mathbf{I}-\mathbf{J}\right) - \eta_l\eta_g\sum_{q=nP_\delta}^{t-1}\widetilde{\boldsymbol{G}}^q\prod_{\ell=q}^{t-1}W^{(\ell)}\left(\mathbf{I}-\mathbf{J}\right)\|_{\mathrm{F}}^2\right]\\ &\overset{(a)}{\le}\left(1+\frac{1}{\gamma}\right)\rho_k\mathbb{E}_W\left[\|\boldsymbol{Z}^{nP_\delta} - \bar{\boldsymbol{Z}}^{nP_\delta}\|_{\mathrm{F}}^2\right] + (1+\gamma)\eta_l^2\eta_g^2\mathbb{E}_W\left[\|\sum_{q=nP_\delta}^{t-1}\widetilde{\boldsymbol{G}}^q\prod_{\ell=q}^{t-1}W^{(\ell)}\left(\mathbf{I}-\mathbf{J}\right)\|_{\mathrm{F}}^2\right]\\ &\overset{(b)}{\le}\widetilde{\rho}\|\boldsymbol{Z}^{nP_\delta} - \bar{\boldsymbol{Z}}^{nP_\delta}\|_{\mathrm{F}}^2 + \frac{\eta_l^2\eta_g^2}{\widetilde{\rho}-\rho_k}\mathbb{E}_W\left[\|\sum_{q=nP_\delta}^{t-1}\widetilde{\boldsymbol{G}}^q\prod_{\ell=q}^{t-1}W^{(\ell)}\left(\mathbf{I}-\mathbf{J}\right)\|_{\mathrm{F}}^2\right].\end{aligned}$$

where inequality (a) holds because Young's inequality and Lemma 3, inequality (b) holds by plugging in $\gamma = \frac{\rho_k}{\widetilde{\rho}-\rho_k}$. Unrolling the recursion, it holds that

$$\begin{aligned}
\|\boldsymbol{Z}^{(t)}(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2 &\leq \widetilde{\rho}\|\boldsymbol{Z}^{nP_\delta} - \bar{\boldsymbol{Z}}^{nP_\delta}\|_{\mathrm{F}}^2 + \frac{\eta_l^2\eta_g^2}{\widetilde{\rho}-\rho_k}\|\sum_{q=nP_\delta}^{t-1}\widetilde{\boldsymbol{G}}^q\prod_{\ell=q}^{t-1}W^{(\ell)}(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2 \\
&\leq \widetilde{\rho}^2\|\boldsymbol{Z}^{(n-1)P_\delta} - \bar{\boldsymbol{Z}}^{(n-1)P_\delta}\|_{\mathrm{F}}^2 + \frac{\eta_l^2\eta_g^2}{\widetilde{\rho}-\rho_k}\widetilde{\rho}\|\sum_{q=(n-1)P_\delta}^{nP_\delta-1}\widetilde{\boldsymbol{G}}^q\prod_{\ell=q}^{nP_\delta-1}W^{(\ell)}(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2 \\
&\qquad + \frac{\eta_l^2\eta_g^2}{\widetilde{\rho}-\rho_k}\|\sum_{q=nP_\delta}^{t-1}\widetilde{\boldsymbol{G}}^q\prod_{\ell=q}^{t-1}W^{(\ell)}(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2 \\
&\leq \frac{\eta_l^2\eta_g^2}{\widetilde{\rho}-\rho_k}\left(\|\sum_{q=nP_\delta}^{t-1}\widetilde{\boldsymbol{G}}^q\prod_{\ell=q}^{t-1}W^{(\ell)}(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2 + \sum_{j=0}^{n-1}\widetilde{\rho}^{n-j}\|\sum_{q=jP_\delta}^{(j+1)P_\delta-1}\widetilde{\boldsymbol{G}}^q\prod_{\ell=q}^{(j+1)P_\delta-1}W^{(\ell)}(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2\right).
\end{aligned}$$

Define $(A) = \|\sum_{q=a}^{b}\widetilde{\boldsymbol{G}}^q\prod_{\ell=q}^{b}W^{(\ell)}(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2$. It remains to bound (A):

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

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

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

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

$$\mathbb{E}\left[(A)\right] \le \mathbb{E}\left[\| \sum_{q=a}^{b} \widetilde{\Delta}^q \left(\prod_{\ell=q}^{b} W^{(\ell)} - \mathbf{J}\right) \|_{\mathrm{F}}^2\right] + \mathbb{E}\left[\| \sum_{q=a}^{b} (\Delta^q + \nabla \boldsymbol{F}_{\boldsymbol{x}}^q) \left(\prod_{\ell=q}^{t-1} W^{(\ell)} - \mathbf{J}\right) \|_{\mathrm{F}}^2\right]$$

$$\le \mathbb{E}\left[\underbrace{\| \sum_{q=a}^{b} \widetilde{\Delta}^q \left(\prod_{\ell=q}^{b} W^{(\ell)} - \mathbf{J}\right) \|_{\mathrm{F}}^2}_{\text{(I)}}\right] + 2\mathbb{E}\left[\underbrace{\| \sum_{q=a}^{b} \Delta^q \left(\prod_{\ell=q}^{b} W^{(\ell)} - \mathbf{J}\right) \|_{\mathrm{F}}^2}_{\text{(II)}}\right] + 2s^2\mathbb{E}\left[\underbrace{\| \sum_{q=a}^{b} \nabla \boldsymbol{F}_{\boldsymbol{x}}^q \left(\prod_{\ell=q}^{t-1} W^{(\ell)} - \mathbf{J}\right) \|_{\mathrm{F}}^2}_{\text{(III)}}\right].$$

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

$$\mathbb{E}\left[(\mathrm{I})\right] = \sum_{q=a}^{b} \mathbb{E}\left[\| \widetilde{\Delta}^q \left(\prod_{\ell=q}^{b} W^{(\ell)} - \mathbf{J}\right) \|_{\mathrm{F}}^2\right] \tag{43}$$

$$+ \sum_{q=a}^{b} \sum_{p\neq q} \mathbb{E}\left[\left\langle \widetilde{\Delta}^p \left(\prod_{\ell=p}^{t-1} W^{(\ell)} - \mathbf{J}\right), \widetilde{\Delta}^q \left(\prod_{\ell=q}^{t-1} W^{(\ell)} - \mathbf{J}\right) \right\rangle\right]$$

$$\overset{(c)}{\le} \sum_{q=a}^{b} \mathbb{E}\left[\|\widetilde{\Delta}^q\|_{\mathrm{F}}^2\right], \tag{44}$$

where inequality (c) holds because of independent and unbiased stochastic gradients. It remains to bound $\mathbb{E}\left[\|\widetilde{\Delta}^q\|_{\mathrm{F}}^2\right]$.

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

Further take expectation w.r.t. the randomness in stochastic gradients.

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

$$\le s\sigma^2 \sum_{i=1}^{m} \mathbb{1}_{\{i\in\mathcal{A}^q\}} \sum_{p=-1}^{q-1} \mathbb{1}_{\{\tau_i(t)=p\}} (q-p)^2.$$

Take expectation over the remaining randomness:

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

Recall that (var) refers to (41). Therefore, we have

$$\mathbb{E}\left[(\mathrm{I})\right] \le 2P_\delta ms\sigma^2 \left(\mathsf{var}\right).$$

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

$$\mathbb{E}[(\mathrm{II})] = \mathbb{E}\left[\| \sum_{q=a}^{b} \Delta^q \left(\prod_{\ell=q}^{b} W^{(\ell)} - \mathbf{J} \right) \|_{\mathrm{F}}^2 \right] \leq 2P_\delta \sum_{q=t-P_\delta}^{b} \mathbb{E}\left[\|\Delta^q \left(\prod_{\ell=q}^{b} W^{(\ell)} - \mathbf{J} \right) \|_{\mathrm{F}}^2 \right]$$
$$\leq 2P_\delta \sum_{q=a}^{b} \mathbb{E}\left[\|\Delta^q\|_{\mathrm{F}}^2 \right].$$

It remains to bound $\mathbb{E}\left[\|\Delta^q\|_{\mathrm{F}}^2\right]$. Take expectation with respect to randomness in stochastic gradients:

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

Next, we take expectation over the remaining randomness and plug back in:

$$\begin{aligned}
\mathbb{E}[(\mathrm{II})] &\leq 16\eta_l^2 s^3 L^2 P_\delta^2 \sum_{i=1}^{m} \sum_{p=-1}^{q-1} \mathbb{E}\left[\mathbb{1}_{\{\tau_i(q)=p\}}\right] (q-p)^2 \sigma^2 \\
&\quad + 32\eta_l^2 s^4 L^2 P_\delta \sum_{q=a}^{b} \sum_{i=1}^{m} \mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^q)\right\|_2^2\right] \sum_{p=-1}^{q-1} \mathbb{E}\left[\mathbb{1}_{\{\tau_i(q)=p\}}\right] (q-p)^2 \\
&\leq 16\eta_l^2 s^3 L^2 P_\delta^2 m\sigma^2 \left(\mathsf{var}\right) + 32\eta_l^2 s^4 L^2 P_\delta \sum_{q=a}^{b} \sum_{i=1}^{m} \mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^q)\right\|_2^2\right] \left(\mathsf{var}\right),
\end{aligned}$$

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

$$\mathbb{E}[(\mathrm{III})] = \mathbb{E}\left[\| \sum_{q=a}^{b} \nabla \boldsymbol{F}_{\boldsymbol{x}}^q \left(\prod_{\ell=q}^{b} W^{(\ell)} - \mathbf{J} \right) \|_{\mathrm{F}}^2 \right] \leq 2P_\delta \sum_{q=a}^{b} \sum_{i=1}^{m} \mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^q)\right\|_2^2\right].$$

Hence, we have

$$\mathbb{E}[(A)] \leq 2P_\delta \left(\mathsf{var}\right) \left(ms\sigma^2 \left(1 + 16\eta_l^2 s^2 L^2 P_\delta \left(\mathsf{var}\right)\right) + 2s^2 \left(1 + 16\eta_l^2 s^2 L^2\right) \sum_{q=a}^{b} \sum_{i=1}^{m} \mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^q)\right\|_2^2\right] \right)$$

It follows that

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{Z}^{(t)} (\mathbf{I} - \mathbf{J})\|_{\mathrm{F}}^2\right] \leq \frac{\eta_l^2 \eta_g^2}{\widetilde{\rho} - \rho_k} \left(\| \sum_{q=nP_\delta}^{t-1} \widetilde{\boldsymbol{G}}^q \prod_{\ell=q}^{t-1} W^{(\ell)} (\mathbf{I} - \mathbf{J}) \|_{\mathrm{F}}^2 + \sum_{j=0}^{n-1} \widetilde{\rho}^{n-j} \| \sum_{q=jP_\delta}^{(j+1)P_\delta - 1} \widetilde{\boldsymbol{G}}^q \prod_{\ell=q}^{(j+1)P_\delta - 1} W^{(\ell)} (\mathbf{I} - \mathbf{J}) \|_{\mathrm{F}}^2 \right) \\
&\leq 2P_\delta \left(\mathsf{var}\right) \frac{\eta_l^2 \eta_g^2}{\widetilde{\rho} - \rho_k} \left(ms\sigma^2 \left(1 + 16\eta_l^2 s^2 L^2 P_\delta \left(\mathsf{var}\right)\right) + 2s^2 \left(1 + 16\eta_l^2 s^2 L^2\right) \sum_{q=nP_\delta}^{t-1} \sum_{i=1}^{m} \mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^q)\right\|_2^2\right] \right) \\
&\quad + 2P_\delta \left(\mathsf{var}\right) \frac{\eta_l^2 \eta_g^2}{\widetilde{\rho} - \rho_k} \sum_{j=0}^{n-1} \widetilde{\rho}^{n-j} \left(ms\sigma^2 \left(1 + 16\eta_l^2 s^2 L^2 P_\delta \left(\mathsf{var}\right)\right) + 2s^2 \left(1 + 16\eta_l^2 s^2 L^2\right) \sum_{q=jP_\delta}^{(j+1)P_\delta - 1} \sum_{i=1}^{m} \mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^q)\right\|_2^2\right] \right).
\end{aligned}$$

Rearranging the terms, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{Z}^{(t)}\left(\mathbf{I}-\mathbf{J}\right)\right\|_{\mathrm{F}}^{2}\right] \leq 2P_{\delta}\left(\mathsf{var}\right)\frac{\eta_{l}^{2}\eta_{g}^{2}}{\widetilde{\rho}-\rho_{k}}\sum_{j=0}^{n}\widetilde{\rho}^{n-j}\left(ms\sigma^{2}\left(1+16\eta_{l}^{2}s^{2}L^{2}P_{\delta}\left(\mathsf{var}\right)\right)\right)\\
&\quad+4s^{2}\left(1+16\eta_{l}^{2}s^{2}L^{2}\right)P_{\delta}\left(\mathsf{var}\right)\frac{\eta_{l}^{2}\eta_{g}^{2}}{\widetilde{\rho}-\rho_{k}}\left(\sum_{q=nP_{\delta}}^{t-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_{i}(\boldsymbol{x}_{i}^{q})\right\|_{2}^{2}\right]+\sum_{j=0}^{n}\widetilde{\rho}^{n-j}\sum_{q=jP_{\delta}}^{(j+1)P_{\delta}-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_{i}(\boldsymbol{x}_{i}^{q})\right\|_{2}^{2}\right]\right)\\
&\leq 2P_{\delta}\left(\mathsf{var}\right)\frac{\eta_{l}^{2}\eta_{g}^{2}}{(\widetilde{\rho}-\rho_{k})(1-\widetilde{\rho})}\left(ms\sigma^{2}\left(1+16\eta_{l}^{2}s^{2}L^{2}P_{\delta}\left(\mathsf{var}\right)\right)\right)\\
&\quad+4s^{2}\left(1+16\eta_{l}^{2}s^{2}L^{2}\right)P_{\delta}\left(\mathsf{var}\right)\frac{\eta_{l}^{2}\eta_{g}^{2}}{\widetilde{\rho}-\rho_{k}}\left(\sum_{q=nP_{\delta}}^{t-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_{i}(\boldsymbol{x}_{i}^{q})\right\|_{2}^{2}\right]+\sum_{j=0}^{n}\widetilde{\rho}^{n-j}\sum_{q=jP_{\delta}}^{(j+1)P_{\delta}-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_{i}(\boldsymbol{x}_{i}^{q})\right\|_{2}^{2}\right]\right).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\frac{1}{MT}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{Z}^{(t)}\left(\mathbf{I}-\mathbf{J}\right)\right\|_{\mathrm{F}}^{2}\right] \leq 2P_{\delta}\left(\mathsf{var}\right)\frac{\eta_{l}^{2}\eta_{g}^{2}}{(\widetilde{\rho}-\rho_{k})(1-\widetilde{\rho})}\left(\frac{ms\sigma^{2}}{M}\left(1+16\eta_{l}^{2}s^{2}L^{2}P_{\delta}\left(\mathsf{var}\right)\right)\right)\\
&+4s^{2}\left(1+16\eta_{l}^{2}s^{2}L^{2}\right)P_{\delta}\left(\mathsf{var}\right)\frac{\eta_{l}^{2}\eta_{g}^{2}}{\widetilde{\rho}-\rho_{k}}\frac{1}{MT}\sum_{t=0}^{T-1}\left(\sum_{q=nP_{\delta}}^{t-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_{i}(\boldsymbol{x}_{i}^{q})\right\|_{2}^{2}\right]+\sum_{j=0}^{n}\widetilde{\rho}^{n-j}\sum_{q=jP_{\delta}}^{(j+1)P_{\delta}-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_{i}(\boldsymbol{x}_{i}^{q})\right\|_{2}^{2}\right]\right)\\
&\leq 4P_{\delta}\left(\mathsf{var}\right)\frac{\eta_{l}^{2}\eta_{g}^{2}}{(\widetilde{\rho}-\rho_{k})(1-\widetilde{\rho})}\left(\frac{ms\sigma^{2}}{M}\right)+16s^{2}P_{\delta}^{2}\left(\mathsf{var}\right)\frac{\eta_{l}^{2}\eta_{g}^{2}}{\widetilde{\rho}-\rho_{k}}\frac{1}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_{i}(\boldsymbol{x}_{i}^{t})\right\|_{2}^{2}\right]\left(\sum_{q=0}^{\infty}\widetilde{\rho}^{q}\right)\\
&\leq 4P_{\delta}\left(\mathsf{var}\right)\frac{\eta_{l}^{2}\eta_{g}^{2}}{(\widetilde{\rho}-\rho_{k})(1-\widetilde{\rho})}\frac{ms\sigma^{2}}{M}+16s^{2}P_{\delta}^{2}\left(\mathsf{var}\right)\frac{\eta_{l}^{2}\eta_{g}^{2}}{(\widetilde{\rho}-\rho_{k})(1-\widetilde{\rho})}\frac{1}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_{i}(\boldsymbol{x}_{i}^{t})\right\|_{2}^{2}\right].
\end{aligned}$$

Further, it holds that

$$\begin{aligned}
\frac{1}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_{i}(\boldsymbol{x}_{i}^{t})\right\|_{2}^{2}\right] &\leq \eta_{l}^{2}\eta_{g}^{2}s^{2}\left(\mathsf{var}\right)\frac{1}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_{i}(\boldsymbol{z}_{i}^{t})\right\|_{2}^{2}\right]\\
&\leq \frac{3M(\beta^{2}+1)}{m}\eta_{l}^{2}\eta_{g}^{2}s^{2}\left(\mathsf{var}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^{t})\right\|_{2}^{2}\right]\\
&\quad+\frac{3m\eta_{l}^{2}\eta_{g}^{2}s^{2}\left(\mathsf{var}\right)\zeta^{2}}{M}\\
&\quad+3\eta_{l}^{2}\eta_{g}^{2}s^{2}L^{2}\left(\mathsf{var}\right)\frac{1}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{z}_{i}^{t}-\bar{\boldsymbol{z}}^{t}\right\|_{2}^{2}\right].
\end{aligned}$$

Plug it back in, we have

$$\begin{aligned}\frac{1}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right] \leq\ & 4P_\delta\left(\mathsf{var}\right)\frac{\eta_l^2\eta_g^2}{(\widetilde{\rho}-\rho_k)(1-\widetilde{\rho})}\frac{sm\sigma^2}{M}\\ &+\frac{48M}{m}\frac{\eta_l^4\eta_g^4(\beta^2+1)s^4P_\delta^2\left(\mathsf{var}\right)^2}{(\widetilde{\rho}-\rho_k)(1-\widetilde{\rho})}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\\ &+48s^4P_\delta^2\left(\mathsf{var}\right)^2\frac{\eta_l^4\eta_g^4}{(\widetilde{\rho}-\rho_k)(1-\widetilde{\rho})}\left(\frac{m\zeta^2}{M}\right)\\ &+3\eta_l^2\eta_g^2s^2P_\delta\left(\mathsf{var}\right)\frac{1}{(\widetilde{\rho}-\rho_k)(1-\widetilde{\rho})}\frac{1}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{M}\left(\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right]\right).\end{aligned}$$

Under our choice of learning rate condition, it holds that

$$\begin{aligned}\frac{1}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right] \leq\ & \frac{8\eta_l^2\eta_g^2sP_\delta\left(\mathsf{var}\right)}{(\widetilde{\rho}-\rho_k)(1-\widetilde{\rho})}\frac{m\sigma^2}{M}\\ &+\frac{6\eta_l^2\eta_g^2s^2P_\delta\left(\mathsf{var}\right)(\beta^2+1)}{(\widetilde{\rho}-\rho_k)(1-\widetilde{\rho})}\left(\frac{M}{m}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\\ &+\frac{6\eta_l^2\eta_g^2s^2P_\delta\left(\mathsf{var}\right)}{(\widetilde{\rho}-\rho_k)(1-\widetilde{\rho})}\left(\frac{m\zeta^2}{M}\right).\end{aligned}$$

By taking $\widetilde{\rho}=\frac{1+\rho_k}{2}$, we get

$$\begin{aligned}\frac{1}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right] \leq\ & \frac{32\eta_l^2\eta_g^2sP_\delta\left(\mathsf{var}\right)}{(1-\rho_k)^2}\left(\frac{m\sigma^2}{M}\right)\\ &+\frac{24\eta_l^2\eta_g^2s^2P_\delta\left(\mathsf{var}\right)(\beta^2+1)}{(1-\rho_k)^2}\left(\frac{M}{m}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\\ &+\frac{24\eta_l^2\eta_g^2s^2P_\delta\left(\mathsf{var}\right)}{(1-\rho_k)^2}\left(\frac{m\zeta^2}{M}\right).\end{aligned}$$

■

D.6 Spectral norm upper bound (Lemma 9)

Proof Proof of Lemma 9 Before we dive into the concrete proof, let us first go over some additional notations and agree on a fact about matrix products.

Notations. $\prod_{t=1}^{k}W^{(t)}\triangleq W^{(1)}W^{(2)}\cdots W^{(k)}$ defines an ordered matrix product from $W^{(1)}$ to $W^{(k)}$, where $k\in\mathbb{Z}^+$, and an undirected graph $\mathcal{G}_t\triangleq\left\{\mathcal{V}_{\mathcal{G}_t},\mathcal{E}_t,W^{(t)}\right\}$, where $\mathcal{V}_{\mathcal{G}_t}\triangleq\mathcal{R}\cup\mathcal{V}$ defines the set of vertices for all $t\in[T]$, $W^{(t)}$ defines the weighted adjacency matrix, an edge $(i,j)\in\mathcal{E}_t$ if $W_{ij}^{(t)}>0$ for $i,j\in[M]$.

A fact about matrix product. By definition of matrix products, we have

$$\prod_{t=1}^{P_\delta} W_{ij}^{(t)} = \sum_{k_1=1}^{M} W_{ik_1}^{(1)} \cdots \sum_{k_{P_\delta-2}=1}^{M} W_{k_{P_\delta-3}k_{P_\delta-2}}^{(P_\delta-2)} \sum_{k_{P_\delta-1}=1}^{M} W_{k_{P_\delta-2}k_{P_\delta-1}}^{(P_\delta-1)} W_{k_{P_\delta-1}j}^{(P_\delta)}. \tag{45}$$

It is observed from (45) that $\prod_{t=1}^{P_\delta} W_{ij}^{(t)} > 0$ if there exists at least one

$$W_{ik_1}^{(1)} W_{k_1k_2}^{(2)} \cdots W_{k_{P_\delta-2}k_{P_\delta-1}}^{(P_\delta-1)} W_{k_{P_\delta-1}j}^{(P_\delta)} > 0,$$

where $k_1, \ldots, k_{P_\delta-1} \in \mathcal{R} \cup \mathcal{V}$. In the language of graph theory, $\prod_{t=1}^{P_\delta} W_{ij}^{(t)} > 0$ if there exists a directed path starting from node i of $\mathcal{G}_1$ and ends at node j of $\mathcal{G}_{P_\delta}$ through $\mathcal{G}_2$, $\mathcal{G}_3$, …, $\mathcal{G}_{P_\delta-1}$ sequentially. Therefore, to lower bound elements in $\prod_{t=1}^{P_\delta} W_t$, it suffices to find the paths of interest for the lower bound.

Element-wise lower bound of $W^{(t)}$ matrix for all $t \in [T]$. Recall the definition of a W matrix:

$$W_{ij}^{(t)} = \begin{cases} \frac{\mathbb{1}_{\{i\in\mathcal{A}^t\}}\mathbb{1}_{\{j\in\mathcal{A}^t\}}}{|\mathcal{A}^t|+k}, & \text{if } i \neq j \text{ and } i,j \in \mathcal{R}; \\ \frac{\mathbb{1}_{\{i\in\mathcal{A}^t\}}}{|\mathcal{A}^t|+k}, & \text{if } i \in \mathcal{R} \text{ and } j \in \mathcal{V}; \\ \frac{\mathbb{1}_{\{j\in\mathcal{A}^t\}}}{|\mathcal{A}^t|+k}, & \text{if } i \in \mathcal{V} \text{ and } j \in \mathcal{R}; \\ \frac{\mathbb{1}_{\{i\in\mathcal{A}^t\}}}{|\mathcal{A}^t|+k} + (1 - \mathbb{1}_{\{i\in\mathcal{A}^t\}}), & \text{if } i = j \text{ and } i \in \mathcal{R}; \\ \frac{1}{|\mathcal{A}^t|+k}, & \text{if } i = j \text{ and } i \in \mathcal{V}, \end{cases}$$

Next, we focus on the diagonal element in $W^{(t)}$, we have two cases:

- When $i \in \mathcal{R}$, we have

$$W_{ii}^{(t)} = \frac{\mathbb{1}_{\{i\in\mathcal{A}^t\}}}{|\mathcal{A}^t|+k} + (1 - \mathbb{1}_{\{i\in\mathcal{A}^t\}}) \geq \frac{\mathbb{1}_{\{i\in\mathcal{A}^t\}}}{|\mathcal{A}^t|+k} + \frac{(1 - \mathbb{1}_{\{i\in\mathcal{A}^t\}})}{|\mathcal{A}^t|+k} = \frac{1}{|\mathcal{A}^t|+k};$$

- When $i \in \mathcal{V}$, we have

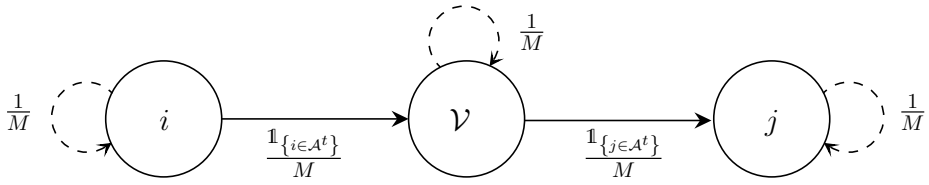
$$W_{ii}^{(t)} = \frac{1}{|\mathcal{A}^t|+k}.$$

For the edge weight of ij, where $i \neq j$, it holds that

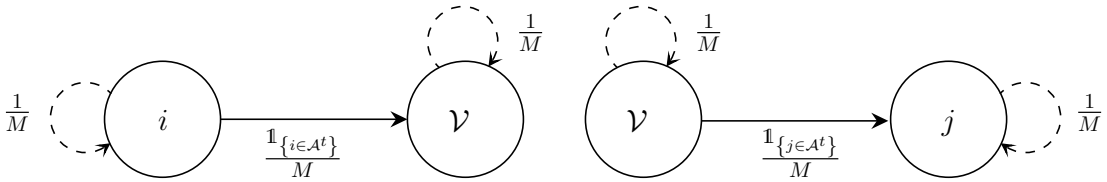
$$W_{ij}^{(t)} \geq \frac{\mathbb{1}_{\{i\in[M]\}}\mathbb{1}_{\{j\in[M]\}}}{|\mathcal{A}^t|+k}.$$

Element-wise lower bound of $\left(\prod_{t=t_0}^{t_0+P-1} W^{(t)}\right)^2$. Fig. 13 presents a simple example of a path from node i to j, which visualizes the transition between different nodes. We will use it to assist our proof. By definition, we have

$$\left(\prod_{t=t_0}^{t_0+P_\delta-1} W^{(t)}\right)^2 = \left(\prod_{t=t_0}^{t_0+P_\delta-1} W^{(t)}\right)^\top \left(\prod_{t=t_0}^{t_0+P_\delta-1} W^{(t)}\right) = \underbrace{\left(\prod_{t=t_0+P_\delta-1}^{t_0} W^{(t)}\right)}_{\text{(I)}} \underbrace{\left(\prod_{t=t_0}^{t_0+P_\delta-1} W^{(t)}\right)}_{\text{(II)}}.$$



(a) A directed path from node i to j through a virtual node.



(b) A directed path from node i to a virtual node. (c) A directed path from a virtual node to node j.

Figure 13: An illustration of a directed path from node i to j through an intermediate virtual node. The dashed arcs are self-loops. The weights next to the dashed arcs and solid lines are the lower bounds of each edge weight, respectively.

To lower bound $[\left(\prod_{t=t_0}^{t_0+P_\delta-1} W^{(t)}\right)^2]_{ij}$, where $i \neq j$. It suffices to let a node i take one path to an intermediate node h in (I) and then take another path from the intermediate node h back to node j in (II). We will return to the case $i = j$ later. Let $h \in \mathcal{V}$.

- In (I) (Fig. 13b), we start from node i in $W^{(t_0+P-1)}$ to node h in $W^{(t_0)}$ via a directed path. One possible path is to stay in self-loops as long as one can, and jump only once from node i to node h. It holds that

$$[(\mathrm{I})]_{ih} \geq \sum_{t=t_0}^{t_0+P_\delta-1} \left(\underbrace{\frac{1}{\prod_{t'\neq t} |\mathcal{A}^{t'}| + k}}_{\text{self-loop weights}} \cdot \underbrace{\frac{\mathbb{1}_{\{i\in\mathcal{A}^t\}}}{|\mathcal{A}^t| + k}}_{\text{one jump weight}} \right) = \frac{\sum_{t=t_0}^{t_0+P_\delta-1} \mathbb{1}_{\{i\in\mathcal{A}^t\}}}{\prod_{t'=t_0}^{t_0+P_\delta-1} (|\mathcal{A}^{t'}| + k)}. \quad (46)$$

- In (II) (Fig. 13c), we can show the results similarly in the following

$$[(\mathrm{II})]_{hj} \geq \sum_{t=t_0}^{t_0+P_\delta-1} \left(\underbrace{\frac{\mathbb{1}_{\{i\in\mathcal{A}^t\}}}{|\mathcal{A}^t| + k}}_{\text{one jump weight}} \cdot \underbrace{\frac{1}{\prod_{t'\neq t} |\mathcal{A}^{t'}| + k}}_{\text{self-loop weights}} \right) = \frac{\sum_{t=t_0}^{t_0+P_\delta-1} \mathbb{1}_{\{j\in\mathcal{A}^t\}}}{\prod_{t'=t_0}^{t_0+P_\delta-1} (|\mathcal{A}^{t'}| + k)}. \quad (47)$$

Hence, when $h \in \mathcal{V}$, we have

$$\sum_{h\in\mathcal{V}} [(\mathrm{I})]_{ih} \cdot [(\mathrm{II})]_{hj} \geq \frac{\left(\sum_{t=t_0}^{t_0+P_\delta-1} \mathbb{1}_{\{i\in\mathcal{A}^t\}}\right)\left(\sum_{t=t_0}^{t_0+P_\delta-1} \mathbb{1}_{\{j\in\mathcal{A}^t\}}\right)}{\left(\prod_{t'=t_0}^{t_0+P_\delta-1} (|\mathcal{A}^{t'}| + k)\right)^2}$$

Expected element-wise lower bound of $\left(\prod_{t=t_0}^{t_0+P_\delta-1} W^{(t)}\right)^2$. Recall that $\mathcal{F}^t$ defines the sigma algebra generated by the randomness up to round t, and that we assume independent availability. Note that we are interested only in the off-diagonal elements. By the law of total expectation, it holds for $i \neq j$ that

- $i \in \mathcal{R}$, $j \in \mathcal{R}$.

$$\begin{aligned}
&\mathbb{E}\left[\left[\left(\prod_{t=t_0}^{t_0+P_\delta-1} W^{(t)}\right)^2\right]_{ij} \Big| \mathcal{F}^{t_0}\right] \geq \mathbb{E}\left[\frac{\left(\sum_{t=t_0}^{t_0+P_\delta-1} \mathbb{1}_{\{i\in\mathcal{A}^t\}}\right)\left(\sum_{t=t_0}^{t_0+P_\delta-1} \mathbb{1}_{\{j\in\mathcal{A}^t\}}\right)}{M^{2P_\delta}} \Big| \mathcal{F}^{t_0}\right] \\
&= \frac{1}{M^{2P_\delta}}\mathbb{E}\left[\mathbb{E}\left[\left(\sum_{t=t_0}^{t_0+P_\delta-1} \mathbb{1}_{\{i\in\mathcal{A}^t\}}\right)\left(\sum_{t=t_0}^{t_0+P_\delta-1} \mathbb{1}_{\{j\in\mathcal{A}^t\}}\right) \Big| \mathcal{F}^{t_0+P_\delta-1}, \mathcal{F}^{t_0}\right] \Big| \mathcal{F}^{t_0}\right] \\
&\overset{(a)}{=} \frac{1}{M^{2P_\delta}}\mathbb{E}\left[\mathbb{E}\left[\left(\sum_{t=t_0}^{t_0+P_\delta-1} \mathbb{1}_{\{i\in\mathcal{A}^t\}}\right) \Big| \mathcal{F}^{t_0+P_\delta-1}\right]\mathbb{E}\left[\left(\sum_{t=t_0}^{t_0+P_\delta-1} \mathbb{1}_{\{j\in\mathcal{A}^t\}}\right) \Big| \mathcal{F}^{t_0+P_\delta-1}\right] \Big| \mathcal{F}^{t_0}\right] \\
&= \frac{1}{M^{2P_\delta}}\mathbb{E}\left[\left(\sum_{t=t_0}^{t_0+P_\delta-2} \mathbb{1}_{\{i\in\mathcal{A}^t\}} + \mathbb{P}\left\{i \in \mathcal{A}^{t_0+P_\delta-1}\right\}\right)\left(\sum_{t=t_0}^{t_0+P_\delta-2} \mathbb{1}_{\{j\in\mathcal{A}^t\}} + \mathbb{P}\left\{j \in \mathcal{A}^{t_0+P_\delta-1}\right\}\right) \Big| \mathcal{F}^{t_0}\right] \\
&= \cdots \\
&= \frac{\left(\sum_{t=t_0}^{t_0+P_\delta-1} \mathbb{P}\left\{i \in \mathcal{A}^t\right\}\right)\left(\sum_{t=t_0}^{t_0+P_\delta-1} \mathbb{P}\left\{j \in \mathcal{A}^t\right\}\right)}{M^{2P_\delta}} \geq \frac{\delta^2 {P_\delta}^2}{M^{2P_\delta}},
\end{aligned}$$

  where equality (a) holds because of independence.
- $i \in \mathcal{V}$, $j \in \mathcal{R}$. In this case, the node can move to an arbitrary virtual node (j included) in (I).

$$\begin{aligned}
&\mathbb{E}\left[\left[\left(\prod_{t=t_0}^{t_0+P_\delta-1} W^{(t)}\right)^2\right]_{ij} \Big| \mathcal{F}^{t_0}\right] \geq \frac{P_\delta}{M^{2P_\delta}}\mathbb{E}\left[\left(\sum_{t=t_0}^{t_0+P_\delta-1} \mathbb{1}_{\{i\in\mathcal{A}^t\}}\right) \Big| \mathcal{F}^{t_0}\right] \\
&= \frac{P_\delta}{M^{2P_\delta}}\mathbb{E}\left[\mathbb{E}\left[\left(\sum_{t=t_0}^{t_0+P_\delta-1} \mathbb{1}_{\{i\in\mathcal{A}^t\}}\right) \Big| \mathcal{F}^{t_0+P_\delta-1}\right] \Big| \mathcal{F}^{t_0}\right] \\
&= \frac{P_\delta\left(\sum_{t=t_0}^{t_0+P_\delta-1} \mathbb{P}\left\{i \in \mathcal{A}^t\right\}\right)}{M^{2P_\delta}} \geq \frac{\delta {P_\delta}^2}{M^{2P_\delta}}.
\end{aligned}$$

- $i \in \mathcal{R}$, $j \in \mathcal{V}$. Similar strategy as above.

$$\mathbb{E}\left[\left[\left(\prod_{t=t_0}^{t_0+P_\delta-1} W^{(t)}\right)^2\right]_{ij} \Big| \mathcal{F}^{t_0}\right] \geq \frac{\delta {P_\delta}^2}{M^{2P_\delta}}.$$

- $i \in \mathcal{V}$, $j \in \mathcal{V}$.

$$\mathbb{E}\left[\left[\left(\prod_{t=t_0}^{t_0+P_\delta-1} W^{(t)}\right)^2\right]_{ij} \Big| \mathcal{F}^{t_0}\right] \geq \frac{{P_\delta}^2}{M^{2P_\delta}} \geq \frac{\delta {P_\delta}^2}{M^{2P_\delta}}.$$

Cheeger's inequality. For ease of presentation, define $W^{t_0,P_\delta} \triangleq \mathbb{E}\left[\left(\prod_{i=t_0}^{t_0+P_\delta-1} {W^{(i)}}^2\right)_{ij}\right]$, and $\rho = \lambda_2(\mathcal{W}^{t_0,P_\delta})$. Suppose we have $|\mathcal{S}\cap\mathcal{V}| = a$. Accordingly, we have $|\mathcal{S}\cap\mathcal{R}| = |\mathcal{S}| - a$, $|\mathcal{V}\cap\mathcal{S}'| = k-a$ and $|\mathcal{R}\cap\mathcal{S}'| = |\mathcal{S}'| - (k-a)$. Let

$$f(a,\mathcal{S}) \triangleq \frac{\sum_{i\in\mathcal{S},j\notin\mathcal{S}} \pi_i W_{ij}^{t_0,P_\delta}}{\sum_{i\in\mathcal{S}}\pi_i} = \frac{\sum_{i\in\mathcal{S},j\notin\mathcal{S}} W_{ij}^{t_0,P_\delta}}{|\mathcal{S}|},$$

where the equality holds because $\pi_i = \frac{1}{m+k}$ for all $i \in \mathcal{R}\cup\mathcal{V}$. It follows that

$$\sum_{i\in\mathcal{S},j\notin\mathcal{S}} W_{ij}^{t_0,P_\delta} \geq \frac{{P_\delta}^2\delta^2}{M^{2P_\delta}}(|\mathcal{S}|-a)\left(\left|\mathcal{S}'\right|-(k-a)\right) + \frac{{P_\delta}^2\delta}{M^{2P_\delta}}\left(|\mathcal{S}|\left|\mathcal{S}'\right| - (|\mathcal{S}|-a)(\left|\mathcal{S}'\right|-(k-a))\right).$$

Next, we decompose the two coefficients, respectively.

$$\begin{aligned}
|\mathcal{S}|\left|\mathcal{S}'\right| - (|\mathcal{S}|-a)(\left|\mathcal{S}'\right|-(k-a)) &= |\mathcal{S}|(m+k-|\mathcal{S}|) - (|\mathcal{S}|-a)(m+k-|\mathcal{S}|-k+a)\\
&= m|\mathcal{S}| + k|\mathcal{S}| - |\mathcal{S}|^2 - m|\mathcal{S}| + |\mathcal{S}|^2 - a|\mathcal{S}| + am - a|\mathcal{S}| + a^2\\
&= a^2 + a(m-2|\mathcal{S}|) + k|\mathcal{S}|.
\end{aligned}$$

$$\begin{aligned}
(|\mathcal{S}|-a)\left(\left|\mathcal{S}'\right|-(k-a)\right) &= (|\mathcal{S}|-a)(m+k-|\mathcal{S}|-(k-a))\\
&= m|\mathcal{S}| - am - |\mathcal{S}|^2 - a^2 + 2a|\mathcal{S}|\\
&= -a^2 - a(m-2|\mathcal{S}|) + m|\mathcal{S}| - |\mathcal{S}|^2.
\end{aligned}$$

Therefore, f becomes:

$$\begin{aligned}
&\frac{\delta^2(|\mathcal{S}|-a)(|\mathcal{S}'|-(k-a)) + \delta(|\mathcal{S}||\mathcal{S}'| - (|\mathcal{S}|-a)(|\mathcal{S}'|-(k-a)))}{M^{2P_\delta}|\mathcal{S}|}\\
&= \frac{a^2(\delta-\delta^2) + a(m-2|\mathcal{S}|)(\delta-\delta^2) + k|\mathcal{S}|\delta P_\delta + |\mathcal{S}|(m-|\mathcal{S}|)\delta^2}{M^{2P_\delta}|\mathcal{S}|}.
\end{aligned}$$

It follows that,

- $m \geq 2|\mathcal{S}|$, i.e., $|\mathcal{S}| \leq \frac{m}{2}$. Hence, we have

$$\frac{f(a,\mathcal{S})}{k} \geq \frac{f(0,\mathcal{S})}{k} \geq \frac{(\frac{m}{2}\delta+k)\delta P_\delta}{M^{2P_\delta}}.$$

- $m < 2|\mathcal{S}|$, i.e., $|\mathcal{S}| > \frac{m}{2}$. In this case, the axis of symmetry is on the right-hand side. Hence, it holds that

$$\begin{aligned}
f &\geq \frac{a^2}{M^{2P_\delta}}\left(\frac{\delta-\delta^2}{|\mathcal{S}|}\right) + \frac{a}{M^{2P_\delta}}\left(\frac{m-2|\mathcal{S}|}{|\mathcal{S}|}\right)(\delta-\delta^2) + \frac{k\delta P_\delta + (m-|\mathcal{S}|)\delta^2 P_\delta}{M^{2P_\delta}}\\
&\geq \frac{k\delta + (m-|\mathcal{S}|)\delta^2 - \left(\frac{\delta-\delta^2}{4}\right)\left(2\sqrt{|\mathcal{S}|} - \frac{m}{\sqrt{|\mathcal{S}|}}\right)^2}{M^{2P}}.
\end{aligned} \tag{48}$$

It is easy to see that (48) is monotonic decreasing w.r.t. $|\mathcal{S}|$. By definition, we know that $|\mathcal{S}| \leq \frac{m+k}{2}$ and plug this in, we have

$$\begin{aligned} f &\geq \frac{k\delta + (\frac{m-k}{2})\delta^2 - \left(\frac{\delta-\delta^2}{4}\right)\left(4\left(\frac{m+k}{2}\right) + \frac{m^2}{\frac{m+k}{2}} - 4m\right)}{M^{2P_\delta}} \\ &= \frac{k\delta + (\frac{m-k}{2})\delta^2 - \left(\frac{\delta-\delta^2}{4}\right)\left(\frac{m^2}{\frac{m+k}{2}} - 2(m-k)\right)}{M^{2P_\delta}} \\ &= \frac{\delta\left(k - \frac{k^2}{2(m+k)}\right) + \delta^2\left(\frac{m-k}{2} + \frac{k^2}{2(m+k)}\right)}{(m+k)^{2P_\delta}} \\ &= \frac{m^2\delta^2 + (k^2+2mk)\delta}{2(m+k)^{2P_\delta+1}}. \end{aligned}$$

By comparing the above two lower bounds, we conclude that

$$\Phi(M) = \min_{\sum_{i\in\mathcal{S}}\pi_i \leq \frac{1}{2}} \frac{\sum_{i\in\mathcal{S}, j\notin\mathcal{S}} \pi_i W_{ij}^{P_\delta, t_0}}{\sum_{i\in\mathcal{S}}\pi_i} \geq \frac{m^2\delta^2 + (k^2+2mk)\delta}{2(m+k)^{2P_\delta+1}}.$$

From Cheeger's inequality, we know that $\frac{1-\lambda_2}{2} \leq \Phi(M) \leq \sqrt{2\left(1-\lambda_2\right)}$. Thus,

$$\rho(t) = \lambda_2 \leq 1 - \frac{\Phi^2\left(M\right)}{2} \leq 1 - \frac{\left[m^2\delta + (k^2+2mk)\right]^2}{8(m+k)^{4P+2}}\delta^2.$$

Special case where $P = 1$. By using a similar argument as above and adapting the results in (Xiang et al., 2024), it holds that

$$\rho(t) = \lambda_2 \leq 1 - \frac{\Phi^2\left(M\right)}{2} \leq 1 - \frac{\left[m^2\delta + (k^2+2mk)\right]^2}{8(m+k)^2}\delta^2,$$

which is decreasing in k. The monotonicity can be seen by taking the partial derivative w.r.t. k. Let

$$h(k) \triangleq \frac{m^2\delta + (2mk+k^2)}{(m+k)^2}.$$

It follows that

$$\begin{aligned} \frac{\partial h(k)}{\partial k} &= \frac{(2m+2k)(m+k)^2 - (2m+2k)(m^2\delta + 2mk + k^2)}{(m+k)^4} \\ &= 2\frac{(m+k)^2 - (m^2\delta + 2mk + k^2)}{(m+k)^2} \\ &= \frac{2m^2(1-\delta)}{(m+k)^2} > 0. \end{aligned}$$

■

Appendix E. Convergence Error of $\bar{\boldsymbol{z}}^t$ (Theorem 11)

In the sequel, we recall and assume the following learning rate conditions in (22):

$$\eta_l\eta_g \leq \frac{\delta(1-\rho_k)\sqrt{m}}{96sL\sqrt{P_\delta M\left[(P_\delta-1)^2\delta^2+1\right](\beta^2+1)}} \text{ and } \eta_l \leq \frac{\delta}{216sL\sqrt{\left[(P_\delta-1)^2\delta^2+1\right](\beta^2+1)}}.$$

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

Proof Proof of Theorem 11 Take expectation over all the randomness, plug in Lemma 18 and Proposition 8. By telescoping sum, it holds that

$$\begin{aligned}
\frac{\mathbb{E}\left[\widetilde{F}^\star - \widetilde{F}(\bar{\boldsymbol{z}}^0)\right]}{T} \leq & -\frac{\eta_l\eta_g s}{3}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] \\
& + \frac{2\eta_l^2\eta_g^2 sL\delta_{\max}\sigma^2}{M^2T}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{E}\left[\mathbb{1}_{\{\tau_i(t)=p\}}\right](t-p)^2 \\
& + \frac{17\eta_g\eta_l^3 s^2L^2\sigma^2}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{E}\left[\mathbb{1}_{\{\tau_i(t)=p\}}\right](t-p)^2 \\
& + \frac{4\eta_l\eta_g sL^2}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2\right] \qquad (49) \\
& + \frac{\eta_l\eta_g sL^2}{2MT}\sum_{t=0}^{T-1}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t - \bar{\boldsymbol{z}}^t\right\|_2^2\right] \qquad (50) \\
& + \frac{65\eta_g\eta_l^3 s^3L^2}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{E}\left[\mathbb{1}_{\{\tau_i(t)=p\}}\right](t-p)^2\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2\right]. \qquad (51)
\end{aligned}$$

Next, we bound (49), (50) and (51), respectively. First, we show that

$$\begin{aligned}
&\frac{1}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right] \leq \frac{3m\zeta^2}{M} + \frac{3M\left(\beta^2+1\right)}{m}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] + \frac{3L^2}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t - \bar{\boldsymbol{z}}^t\right\|_2^2\right] \\
&\leq 3\left[1+\frac{24\eta_l^2\eta_g^2 s^2P_\delta\left(\mathsf{var}\right)L^2}{(1-\rho_k)^2}\right]\left(\frac{m\zeta^2}{M}\right) \\
&\quad + \frac{3M\left(\beta^2+1\right)}{m}\left[1+\frac{24\eta_l^2\eta_g^2 s^2P_\delta\left(\mathsf{var}\right)L^2}{(1-\rho_k)^2}\right]\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] \\
&\quad + \frac{96\eta_l^2\eta_g^2 sP_\delta\left(\mathsf{var}\right)L^2}{(1-\rho_k)^2}\left(\frac{m\sigma^2}{M}\right), \qquad (52)
\end{aligned}$$

where the last inequality follows from Lemma 18.

Bounding (49)**.**

$$\begin{aligned}
&\frac{4\eta_l\eta_g sL^2}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\boldsymbol{z}_i^t\right\|_2^2\right]\le\frac{4\eta_l^3\eta_g^3 s^3L^2\left(\mathsf{var}\right)}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right]\\
&\le 12\eta_l^3\eta_g^3 s^3L^2\left(\mathsf{var}\right)\left[1+\frac{24\eta_l^2\eta_g^2 s^2P_\delta\left(\mathsf{var}\right)L^2}{(1-\rho_k)^2}\right]\left(\frac{m\zeta^2}{M}\right)\\
&\quad+12\eta_l^3\eta_g^3 s^3L^2\left(\mathsf{var}\right)\left(\beta^2+1\right)\left[1+\frac{24\eta_l^2\eta_g^2 s^2P_\delta\left(\mathsf{var}\right)L^2}{(1-\rho_k)^2}\right]\left(\frac{M}{m}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\\
&\quad+\frac{384\eta_l^5\eta_g^5 s^4P_\delta\left(\mathsf{var}\right)^2L^4}{(1-\rho_k)^2}\left(\frac{m\sigma^2}{M}\right)\\
&\le 16\eta_l^3\eta_g^3 s^3L^2\left(\mathsf{var}\right)\left(\frac{m\zeta^2}{M}\right)+4\eta_l^3\eta_g^3 s^3P_\delta\left(\mathsf{var}\right)L^2\left(\frac{m\sigma^2}{M}\right)\\
&\quad+16\eta_l^3\eta_g^3 s^3L^2\left(\mathsf{var}\right)\left(\beta^2+1\right)\left(\frac{M}{m}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right],
\end{aligned}$$

where the last inequality holds due to $\eta_l\eta_g\le(1-\rho_k)/(10sL\sqrt{P_\delta\cdot\mathsf{var}})$.

Bounding (50)**.**

$$\begin{aligned}
&\frac{\eta_l\eta_g sL^2}{2MT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\boldsymbol{z}_i^t-\bar{\boldsymbol{z}}^t\right\|_2^2\right]\\
&\le\frac{4\eta_l^3\eta_g^3 s^2L^2P_\delta\left(\mathsf{var}\right)}{(\widetilde{\rho}-\rho_k)(1-\widetilde{\rho})}\left(\frac{m\sigma^2}{M}\right)\\
&\quad+\frac{3\eta_l^3\eta_g^3 s^3L^2P_\delta\left(\mathsf{var}\right)\left(\beta^2+1\right)}{(\widetilde{\rho}-\rho_k)(1-\widetilde{\rho})}\left(\frac{M}{m}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]+\frac{3\eta_l^3\eta_g^3 s^3L^2P_\delta\left(\mathsf{var}\right)}{(\widetilde{\rho}-\rho_k)(1-\widetilde{\rho})}\left(\frac{m\zeta^2}{M}\right).
\end{aligned}$$

Bounding (51)**.**

$$\begin{aligned}
&65\eta_g\eta_l^3 s^3L^2\frac{1}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\sum_{p=-1}^{t-1}\mathbb{E}\left[\mathbb{1}_{\{\tau_i(t)=p\}}\right](t-p)^2\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^{p+1})\right\|_2^2\right]\\
&\le\frac{65\eta_l^3\eta_g s^3L^2\left(\mathsf{var}\right)}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2\right]\\
&\le 260\eta_l^3\eta_g s^3L^2\left(\mathsf{var}\right)\left(\frac{m\zeta^2}{M}\right)\\
&\quad+\left(\frac{M}{m}\right)\frac{260\eta_l^3\eta_g s^3L^2\left(\mathsf{var}\right)\left(\beta^2+1\right)}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]+65\eta_g\eta_l^3 s^3L^2\left(\frac{m\sigma^2}{M}\right),
\end{aligned}$$

where the last inequality holds due to $\eta_l\eta_g \leq (1-\rho_k)/(10sL\sqrt{P_\delta \cdot \mathsf{var}})$. Putting (49), (50) and (51) together and plugging them back into the telescoping sum, it holds that

$$
\begin{aligned}
&\frac{\mathbb{E}\left[\widetilde{F}^\star - \widetilde{F}(\bar{\boldsymbol{z}}^0)\right]}{T} \\
&\leq -\left(\frac{\eta_l\eta_g s}{3} - 16\eta_l^3\eta_g^3 s^3 L^2\left(\mathsf{var}\right)\left(\frac{M}{m}\right)\left(\beta^2+1\right) - \frac{12\eta_l^3\eta_g^3 s^3 L^2\left(\frac{M}{m}\right)P_\delta\left(\mathsf{var}\right)\left(\beta^2+1\right)}{(1-\rho_k)^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] \\
&\quad - \left(-260\eta_l^3\eta_g s^3 L^2\left(\mathsf{var}\right)\left(\frac{M}{m}\right)\left(\beta^2+1\right)\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] \\
&\quad + \frac{2\eta_l^2\eta_g^2 sL\delta_{\max}\left(\mathsf{var}\right)m\sigma^2}{M^2} + 17\eta_g\eta_l^3 s^2L^2\left(\mathsf{var}\right)\left(\frac{m\sigma^2}{M}\right) \\
&\quad + 4\eta_l^3\eta_g^3 s^3 P_\delta\left(\mathsf{var}\right)L^2\frac{m\sigma^2}{M} + \frac{16\eta_l^3\eta_g^3 s^2L^2P_\delta\left(\mathsf{var}\right)}{(1-\rho_k)^2}\frac{m\sigma^2}{M} + 65\eta_g\eta_l^3 s^3L^2\frac{m\sigma^2}{M} \\
&\quad + 16\eta_l^3\eta_g^3 s^3L^2\left(\mathsf{var}\right)\left(\frac{m\zeta^2}{M}\right) + \frac{12\eta_l^3\eta_g^3 s^3L^2P_\delta\left(\mathsf{var}\right)}{(1-\rho_k)^2}\left(\frac{m\zeta^2}{M}\right) + 260\eta_l^3\eta_g s^3L^2\left(\mathsf{var}\right)\left(\frac{m\zeta^2}{M}\right) \\
&\leq -\frac{\eta_l\eta_g s}{4}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] \\
&\quad + \frac{2\eta_l^2\eta_g^2 sL\delta_{\max}\left(\mathsf{var}\right)m\sigma^2}{M^2} + 17\eta_g\eta_l^3 s^2L^2\left(\mathsf{var}\right)\left(\frac{m\sigma^2}{M}\right) \\
&\quad + 4\eta_l^3\eta_g^3 s^3 P_\delta\left(\mathsf{var}\right)L^2\left(\frac{m\sigma^2}{M}\right) + \frac{16\eta_l^3\eta_g^3 s^2L^2P_\delta\left(\mathsf{var}\right)}{(1-\rho_k)^2}\left(\frac{m\sigma^2}{M}\right) + 65\eta_g\eta_l^3 s^3L^2\left(\frac{m\sigma^2}{M}\right) \\
&\quad + 16\eta_l^3\eta_g^3 s^3L^2\left(\mathsf{var}\right)\left(\frac{m\zeta^2}{M}\right) + \frac{12\eta_l^3\eta_g^3 s^3L^2P_\delta\left(\mathsf{var}\right)}{(1-\rho_k)^2}\left(\frac{m\zeta^2}{M}\right) + 260\eta_l^3\eta_g s^3L^2\left(\mathsf{var}\right)\left(\frac{m\zeta^2}{M}\right),
\end{aligned}
$$

where the last inequality holds because

$$
\eta_l\eta_g \leq \frac{(1-\rho_k)\sqrt{m}}{48sL\sqrt{P_\delta M\left(\mathsf{var}\right)\left(\beta^2+1\right)}} \text{ and } \eta_l \leq \frac{1}{108sL\sqrt{\left(\mathsf{var}\right)\left(\beta^2+1\right)}}.
$$

Combining the above and rearranging the terms, it holds that

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] &\leq \frac{4\left(\widetilde{F}(\bar{\boldsymbol{z}}^0) - \widetilde{F}^\star\right)}{\eta_l\eta_g sT} \\
&\quad + \frac{8\eta_l\eta_g L\delta_{\max}\left(\mathsf{var}\right)m\sigma^2}{M^2} + 68\eta_l^2 sL^2\left(\mathsf{var}\right)\left(\frac{m\sigma^2}{M}\right) \\
&\quad + 16\eta_l^2\eta_g^2 s^2 P_\delta\left(\mathsf{var}\right)L^2\left(\frac{m\sigma^2}{M}\right) + \frac{64\eta_l^2\eta_g^2 sL^2P_\delta\left(\mathsf{var}\right)}{(1-\rho_k)^2}\left(\frac{m\sigma^2}{M}\right) + 260\eta_l^2 s^2L^2\left(\mathsf{var}\right)\left(\frac{m\sigma^2}{M}\right) \\
&\quad + 64\eta_l^2\eta_g^2 s^2L^2\left(\mathsf{var}\right)\left(\frac{m\zeta^2}{M}\right) + \frac{48\eta_l^2\eta_g^2 s^2L^2P_\delta\left(\mathsf{var}\right)}{(1-\rho_k)^2}\left(\frac{m\zeta^2}{M}\right) + 1040\eta_l^2 s^2L^2\left(\mathsf{var}\right)\left(\frac{m\zeta^2}{M}\right).
\end{aligned}
$$

Expanding the second moment of the unavailable duration, it holds that

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{F}(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] \leq\ & \frac{4\left(\widetilde{F}(\bar{\boldsymbol{z}}^0)-\widetilde{F}^\star\right)}{\eta_l\eta_g sT}\\ &+\frac{\left[(P_\delta-1)\delta+1\right]^2+\left[(P_\delta-1)\delta^2+1\right]}{\delta^2}\left(\frac{8\eta_l\eta_g L\delta_{\max}}{M}+68\eta_l^2 sL^2\right)\left(\frac{m\sigma^2}{M}\right)\\ &+\frac{\left[(P_\delta-1)\delta+1\right]^2+\left[(P_\delta-1)\delta^2+1\right]}{\delta^2}\left(16\eta_l^2\eta_g^2 s^2P_\delta L^2+\frac{64\eta_l^2\eta_g^2 sL^2P_\delta}{(1-\rho_k)^2}+260\eta_l^2 s^2L^2\right)\left(\frac{m\sigma^2}{M}\right)\\ &+\frac{\left[(P_\delta-1)\delta+1\right]^2+\left[(P_\delta-1)\delta^2+1\right]}{\delta^2}\left(64\eta_l^2\eta_g^2 s^2L^2+\frac{48\eta_l^2\eta_g^2 s^2L^2P_\delta}{(1-\rho_k)^2}+1040\eta_l^2 s^2L^2\right)\left(\frac{m\zeta^2}{M}\right).\end{aligned}$$

It holds that

$$\begin{aligned}\frac{\left[(P_\delta-1)\delta+1\right]^2+\left[(P_\delta-1)\delta^2+1\right]}{\delta^2} &\leq \frac{2\left((P_\delta-1)^2\delta^2+1\right)+(P_\delta-1)^2\delta^2+1}{\delta^2}\\ &= 3\left((P_\delta-1)^2+\frac{1}{\delta^2}\right).\end{aligned}\tag{53}$$

Grouping the terms of the same order in terms of asymptotic, we have

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

where we use the convention that $\eta_g \geq 1$ for ease of presentation. Using the fact that $\nabla\widetilde{F}(\boldsymbol{x}) = (\frac{m}{M})\nabla F(\boldsymbol{x})$, it holds that

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

■

Appendix F. Convergence Rate of $\bar{\boldsymbol{x}}^t$ (Corollary 12)

F.1 Convergence error of Algorithm 1

Corollary 19 (Convergence error of $\boldsymbol{x}_i^t$)**.** Suppose learning rates conditions in (22) are met for η_l and η_g, and Assumptions 1, 2, 3 and 4 hold for $T \geq 1$, it holds that

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

Proof Proof of Corollary 19

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

where inequality (a) follows from Appendix B 2, inequality (b) follows from Assumption 2. Further plug in (52) and use the learning rate conditions, it holds that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}^t)\right\|_2^2\right] \leq \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right] + 12\left(\mathsf{var}\right)\eta_l^2\eta_g^2s^2L^2\left(\frac{M\zeta^2}{m}\right) + 3\left(\mathsf{var}\right)\eta_l^2\eta_g^2s^2L^2\left(\frac{M\sigma^2}{m}\right).$$

Grouping the terms of the same order in terms of asymptotic, we have

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

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

F.2 Convergence rate of Algorithm 1

Proof Proof of Corollary 12 Choose step-size as $\eta_l = \frac{1}{\sqrt{T}sL}$, $\eta_g = \sqrt{s\delta m}$ such that learning rate conditions in (22) are met, it holds that

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

■

Appendix G. Additional Results and Interpretations

G.1 Consensus error of Algorithm 1

Corollary 20 (Consensus error of $\boldsymbol{x}_i^t$)**.** Suppose learning rates conditions are met in (22) for η_l and η_g, and Assumptions 1, 2, 3 and 4 hold for $T \geq 1$, it holds that

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

Proof Proof of Corollary 20

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

where inequalities (a) and (b) follow from Jensen's inequality. It holds that

$$\begin{aligned}
\frac{1}{MT}\sum_{t=0}^{T-1}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{x}_i^t-\bar{\boldsymbol{x}}^t\right\|_2^2\right] \leq{}& \frac{4}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{z}}^t)\right\|_2^2\right]\\
&+\frac{96\eta_l^2\eta_g^2 sP_\delta\left(\mathsf{var}\right)}{(1-\rho_k)^2}\left(\frac{m\sigma^2}{M}\right)+6\left(\mathsf{var}\right)\eta_l^2\eta_g^2s^2L^2\left(\frac{m\sigma^2}{M}\right)\\
&+\frac{72\eta_l^2\eta_g^2 s^2P_\delta\left(\mathsf{var}\right)}{(1-\rho_k)^2}\left(\frac{m\zeta^2}{M}\right)+72\left(\mathsf{var}\right)\eta_l^2\eta_g^2s^2L^2\left(\frac{m\zeta^2}{M}\right).
\end{aligned}$$

where the last inequality holds because of learning rate condition in (22). Group the terms of the same order in terms of asymptotics, we have

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

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

G.2 Orders of the asymptotic rates

From Theorem 11, Corollary 19, Corollary 20, 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, 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 (23), (25) and (26) hold.

Appendix H. Numerical Experiments

H.1 Experimental 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 (Paszke et al., 2019) and Python 3.7.16.

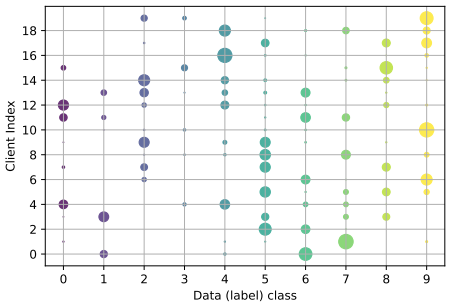


Figure 14: An example of data heterogeneity using Dirichlet($\alpha = 0.1$) distribution with 20 clients. x-axis denotes the categories of images, while y-axis denotes the client index. The size of a circle refers to the proportion of pictures in a given class. The color of a circle distinguishes images with different categories.

Neural Network and Hyperparameter Specifications. Table 5 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 6.

Table 5: Neural network architecture, loss function, learning rate scheduling, training steps and batch size specifications

Data sets	**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	$\eta_l = \eta_0$ in Section 7.1; $\eta_l = \frac{\eta_0}{\sqrt{t/10+1}}$ in Section 7.2, where t denotes the global round.		
Number of local steps s	10		
Number of global rounds T in Section 7.1	10000	20000	–
Number of global rounds T in Section 7.2	2000	2000	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).

Table 6: Initial learning rate η_0 and global learning rate η_g

Algorithms	`FedAvg` *active*		`FedAvg` *known*		`FedAvg` *all*		`FedAU`		`F3AST`		`FedSWE`		`MIFA`	
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
	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
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
	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
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
	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

Baseline Algorithm Details.

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` (Ribero et al., 2022), which is tuned over a grid of $\{0.1, 0.05, 0.01, 0.005, 0.001, 0.0005\}$. The amplification factor of `gFedAvg` in (Wang and Ji, 2022) is set as 10, and the period is set as 100. In addition, as recommended by (Wang and Ji, 2024), we choose $K = 50$ in `FedAU` without further specification. Fig. 3 adopts the same hyperparameter setups as the ones in Section 7.2, yet with only 1000 training rounds.

Data sets and Data Heterogeneity.

Data sets. All the data sets 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 random cropping and gradient clipping with a max norm of 0.5 to all data set trainings. Furthermore, random horizontal flipping is applied to CIFAR-10 and CINIC-10.

One full set of experiments in Section 7.2 takes about 6 hours on SVHN and CIFAR-10 data sets, while about 10 hours on CINIC-10 data set. The training time for experiments in Section 7.1 almost doubles.

- **SVHN (Netzer et al., 2011).** The data set contains 32×32 colored images of 10 different digits. In total, there are 73257 train images and 26032 test images.
- **CIFAR-10 (Krizhevsky et al., 2009).** The data set contains 32×32 colored images of 10 different objects. In total, there are 50000 train images and 10000 test images.
- **CINIC-10 (Darlow et al., 2018).** The data set contains 32×32 colored images of 10 different objects. In total, there are 90000 train images and 90000 test images.

Data heterogeneity. Fig. 14 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.

H.2 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 half image classes and $\mathbf{\Phi}_{c'} = 0.5$ for the remaining half image classes. Fig. 15 plots one resulting p_i's example, wherein p_i's are heterogeneous across clients.

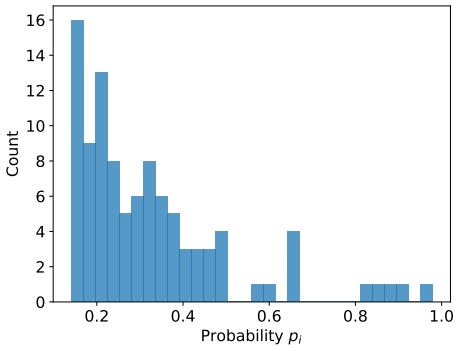


Figure 15: A histogram of one generated p_i's example with a total of $m = 100$ clients. It can be seen that the majority of p_i's are below 0.5.

Next, we formally introduce $f_i(t)$'s under each dynamic in Section 7.2.

- Stationary: $f_i(t) \triangleq 1$;
- Non-stationary with staircase trajectory:

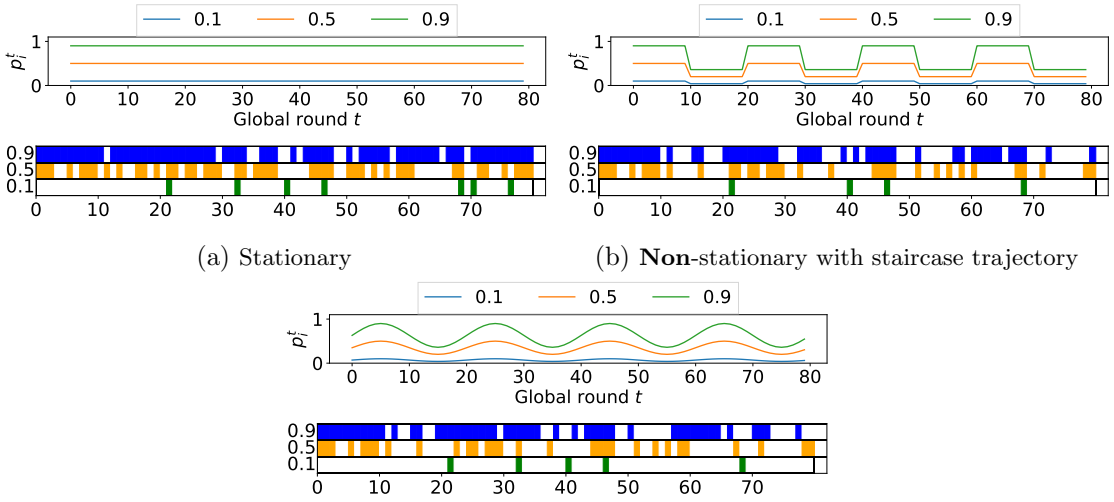
$$f_i(t) \triangleq \mathbb{1}_{\{t \in [t_0, t_0 + P/2)\}} + 0.4 \cdot \mathbb{1}_{\{t \in [t_0 + P/2, t_0 + P)\}},$$

  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.

We choose $\gamma = 0.3$ and $P = 20$ for all non-stationary dynamics in Section 7.2. Next, we visualize the probability trajectories and sampled client availability in Section 7.2 in Fig. 16.



(a) Stationary

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

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

Figure 16: 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 that 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 H.2

Table 7: 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 data set 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.

Data sets	**SVHN**				**CIFAR-10**				**CINIC-10**			
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%
`FedSWE` (**ours**, $k = 0$)	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

H.3 Additional results

In this section, we provide ablation results on `FedSWE` with $k = 0$.

Table 8: Results after different parameter γ. $p_i^t = p_i \cdot (\gamma \sin(2\pi/P \cdot t) + (1-\gamma))$.

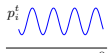
Unavailable Dynamics	Data sets / Algorithms	$\gamma = 0.3$ Train	$\gamma = 0.3$ Test	$\gamma = 0.2$ Train	$\gamma = 0.2$ Test	$\gamma = 0.1$ Train	$\gamma = 0.1$ Test
Non-stationary (**Sine**) 	FedSWE (**ours**, $k = 0$)	**85.7** $\pm$ 0.9 %	**85.6** $\pm$ 0.9 %	**85.7** $\pm$ 0.5 %	**85.7** $\pm$ 0.5 %	**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	82.5 $\pm$ 1.4 %	82.5 $\pm$ 1.3 %	83.5 $\pm$ 0.3 %	83.4 $\pm$ 0.4 %	83.7 $\pm$ 0.3 %	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 9: Results after different Dirichlet parameter α. $p_i^t = p_i(\gamma \sin(2\pi/P \cdot t) + (1-\gamma))$.

Unavailable Dynamics	Data sets / Algorithms	$\alpha = 0.05$ Train	$\alpha = 0.05$ Test	$\alpha = 0.1$ Train	$\alpha = 0.1$ Test	$\alpha = 1.0$ Train	$\alpha = 1.0$ Test
Non-stationary (**Sine**) 	FedSWE (**ours**, $k = 0$)	**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	79.5 $\pm$ 1.6 %	79.5 $\pm$ 1.7 %	82.5 $\pm$ 1.4 %	82.5 $\pm$ 1.3 %	88.4 $\pm$ 0.1 %	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 %

Staleness studies. Table 7 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), FedSWE slightly lags behind FedAvg over active clients. However, when reaching the final stage (the last quarter), FedSWE 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 data sets. The slowdown of FedSWE on CIFAR-10 data set is worth further investigation. In general, we arrive numerically at the conclusion that the staleness incurred by implicit gossiping in FedSWE is mild.

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 trajectory. The results are in Table 8 and Table 9. Overall, FedSWE keeps outperforming the algorithms not assisted by memories or known statistics.

In Table 9, clients' local data becomes more heterogeneous when α increases. We can see a clear increasing trend in accuracy. However, FedSWE remains to attain the best accuracies both train and test when compared to the algorithms not aided by heavy 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 FedSWE.