Appendices

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

A Limitations

The limitations of our work are two-fold:

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

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

B Broader Impacts

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

C Nomenclatures

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

Table 3: Notation table

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

Missing definitions and equation formulas.

Table 4: Algorithmic nomenclature table

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

Table 5: Variable table

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

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

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

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

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

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

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

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

D Useful Inequalities

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

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

1. Jensen's inequality.

$$\left\| \frac{1}{m} \sum_{i=1}^m \boldsymbol{a}_i \right\|_2^2 \le \frac{1}{m} \sum_{i=1}^m \|\boldsymbol{a}_i\|_2^2 \quad \text{and} \quad \left\| \sum_{i=1}^m \boldsymbol{a}_i \right\|_2^2 \le m \sum_{i=1}^m \|\boldsymbol{a}_i\|_2^2 . \qquad (20)$$

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

$$\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle \le \frac{\|\boldsymbol{a}_1\|_2^2}{2\epsilon} + \frac{\epsilon \|\boldsymbol{a}_2\|_2^2}{2}, \quad \text{for any } \epsilon > 0. \qquad (21)$$

Equivalently, we have

$$\begin{aligned}
\|\boldsymbol{a}_1 + \boldsymbol{a}_2\|_2^2 &= \|\boldsymbol{a}_1\|_2^2 + \|\boldsymbol{a}_2\|_2^2 + 2 \langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle \\
&\le \left(1 + \frac{1}{\epsilon}\right) \|\boldsymbol{a}_1\|_2^2 + (1 + \epsilon) \|\boldsymbol{a}_2\|_2^2 , \quad \text{for any } \epsilon > 0. \qquad (22)
\end{aligned}$$

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

$$\begin{aligned}
F(\boldsymbol{a}_1) - F(\boldsymbol{a}_2) &= \left\langle \boldsymbol{a}_1 - \boldsymbol{a}_2, \int_0^1 \nabla F(\boldsymbol{a}_2 + \tau(\boldsymbol{a}_1 - \boldsymbol{a}_2)) \mathrm{d}\tau \right\rangle \\
&= \langle \nabla F(\boldsymbol{a}_2), \boldsymbol{a}_1 - \boldsymbol{a}_2 \rangle + \int_0^1 \langle \boldsymbol{a}_1 - \boldsymbol{a}_2, \nabla F(\boldsymbol{a}_2 + \tau(\boldsymbol{a}_1 - \boldsymbol{a}_2)) - \nabla F(\boldsymbol{a}_2) \rangle \,\mathrm{d}\tau \\
&\overset{(a)}{\le} \langle \nabla F(\boldsymbol{a}_2), \boldsymbol{a}_1 - \boldsymbol{a}_2 \rangle + L \int_0^1 \tau \|\boldsymbol{a}_1 - \boldsymbol{a}_2\|_2 \|(\boldsymbol{a}_1 - \boldsymbol{a}_2)\|_2 \,\mathrm{d}\tau \\
&\le \langle \nabla F(\boldsymbol{a}_2), \boldsymbol{a}_1 - \boldsymbol{a}_2 \rangle + \frac{L}{2} \|\boldsymbol{a}_1 - \boldsymbol{a}_2\|_2^2 , \qquad (23)
\end{aligned}$$

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

E Descent Lemma (Lemma 3)

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

E.1 Multi-step perturbation

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

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

Proof of Lemma 5. The proof shares a similar road map to [61, Lemma 2], but the objective is instead to show an upper bound with respect to $\|\nabla F_i(\boldsymbol{x}_i^t)\|_2^2$.

For $s \geq 1$, it holds that

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

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

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

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

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

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

Unroll the recursion, we have

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

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

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

□

E.2 Descent lemma

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

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

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

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

where equality (a) using the fact that $\boldsymbol{x}_i^k = \boldsymbol{x}_i^t$ for all k such that $\tau_i(t) + 1 \leq k \leq t$, 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 $\big(\mathbb{1}_{\{i \in \mathcal{A}^t\}} + \mathbb{1}_{\{i \notin \mathcal{A}^t\}}\big) = 1$.

Bounding (A)**.**

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

Bounding (A.I)

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

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

Bounding (A.II)

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

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

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

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

Bounding (A.III)**.**

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

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

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

Bounding (A.IV)

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

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

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

Putting (A) together,

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

Bounding (B)**.**

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

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

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

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,

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

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

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

Putting (B) together, we get

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

Now, everything:

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

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

F Intermediate Results

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

F.1 Bounding local and global dissimilarity

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

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

Proof of Proposition 3.

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

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

F.2 Weight re-equalization (Proposition 1)

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

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

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

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

Now, we have two cases:

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

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

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

□

F.3 Unavailable statistics (Lemma 2)

Proof of Lemma 2.

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

From [15, Section 12, Theorem 12.3 (i)], we know that

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

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

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

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

Therefore,

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

□

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

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

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

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

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

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

When $t = 0$,

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

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

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

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

Proof of Proposition 2. From Propositions 4, we have

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

Take expectation over all the randomness

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

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

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

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

F.5 Consensus error of the auxiliary sequence

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

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

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

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

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

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

Expanding $\boldsymbol{Z}$, we get

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

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

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

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

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

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

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

It follows that

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

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

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

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

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

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

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

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

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

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

Take expectation over the remaining randomness:

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

Therefore,

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

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

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

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

$$\sum_{q=0}^{t-1} \sum_{p=0, p \neq q}^{t-1} \frac{\sqrt{\rho}^{2t-p-q}}{2} \mathbb{E}\left[\|\Delta^p\|_{\mathrm{F}}^2 + \|\Delta^q\|_{\mathrm{F}}^2\right] \leq \sum_{q=0}^{t-1} \sum_{p=0}^{t-1} \frac{\sqrt{\rho}^{2t-p-q}}{2} \mathbb{E}\left[\|\Delta^p\|_{\mathrm{F}}^2 + \|\Delta^q\|_{\mathrm{F}}^2\right]$$

$$= \sum_{p=0}^{t-1} \frac{\sqrt{\rho}^{t-p}}{2} \mathbb{E}\left[\|\Delta^p\|_{\mathrm{F}}^2\right] \sum_{q=0}^{t-1} \sqrt{\rho}^{t-q} + \sum_{q=0}^{t-1} \frac{\sqrt{\rho}^{t-q}}{2} \mathbb{E}\left[\|\Delta^q\|_{\mathrm{F}}^2\right] \sum_{p=0}^{t-1} \sqrt{\rho}^{t-p}$$

$$= \frac{\sqrt{\rho} - \sqrt{\rho}^{t+1}}{1 - \sqrt{\rho}} \sum_{q=0}^{t-1} \sqrt{\rho}^{t-q} \mathbb{E}\left[\|\Delta^q\|_{\mathrm{F}}^2\right]. \tag{28}$$

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

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

$$\leq \frac{2\sqrt{\rho}}{1 - \sqrt{\rho}} \sum_{q=0}^{t-1} \sqrt{\rho}^{t-q} \mathbb{E}\left[\|\Delta^q\|_{\mathrm{F}}^2\right], \tag{29}$$

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

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

$$+ 16\eta_l^2 s^4 L^2 \sum_{i=1}^{m} \sum_{p=-1}^{q-1} \mathbb{1}_{\{\tau_i(q)=p\}} (q-p)^2 \left\|\nabla F_i(\boldsymbol{x}_i^q)\right\|_2^2,$$

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

$$\mathbb{E}\left[(\mathrm{II})\right] \leq \frac{2\sqrt{\rho}}{1 - \sqrt{\rho}} \sum_{q=0}^{t-1} \sqrt{\rho}^{t-q} \mathbb{E}\left[\|\Delta^q\|_{\mathrm{F}}^2\right]$$

$$\leq \frac{8\rho}{\left(1 - \sqrt{\rho}\right)^2} \left(\frac{2}{\delta^2}\right) \eta_l^2 s^3 L^2 m \sigma^2$$

$$+ \frac{32\sqrt{\rho}}{1 - \sqrt{\rho}} \left(\frac{2}{\delta^2}\right) \eta_l^2 s^4 L^2 \sum_{i=1}^{m} \sum_{q=0}^{t-1} \mathbb{E}\left[\|\nabla F_i(\boldsymbol{x}_i^q)\|_2^2\right] \sum_{k=1}^{T-1-t} \sqrt{\rho}^k$$

$$\leq \frac{8\rho}{\left(1 - \sqrt{\rho}\right)^2} \left(\frac{2}{\delta^2}\right) \eta_l^2 s^3 L^2 m \sigma^2 + \frac{32\rho}{\left(1 - \sqrt{\rho}\right)^2} \left(\frac{2}{\delta^2}\right) \eta_l^2 s^4 L^2 \sum_{i=1}^{m} \sum_{q=0}^{t-1} \mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^q)\right\|_2^2\right],$$

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

$$\frac{1}{mT} \sum_{t=1}^{T-1} \mathbb{E}\left[(\mathrm{II})\right] \leq \frac{8\rho}{\left(1 - \sqrt{\rho}\right)^2} \left(\frac{2}{\delta^2}\right) \eta_l^2 s^3 L^2 \sigma^2$$

$$+ \frac{32\rho}{\left(1 - \sqrt{\rho}\right)^2} \left(\frac{2}{\delta^2}\right) \eta_l^2 s^4 L^2 \frac{1}{T} \sum_{t=1}^{T-1} \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^t)\right\|_2^2\right]$$

$$\leq \frac{8\rho}{\left(1 - \sqrt{\rho}\right)^2} \left(\frac{2}{\delta^2}\right) \eta_l^2 s^3 L^2 \sigma^2 + \frac{64\rho}{\left(1 - \sqrt{\rho}\right)^2} \left(\frac{2}{\delta^2}\right) \eta_l^2 s^4 L^4 \frac{1}{T} \sum_{t=1}^{T-1} \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}\left[\left\|\boldsymbol{x}_i^t - \boldsymbol{z}_i^t\right\|_2^2\right]$$

$$+ \frac{64\rho}{\left(1 - \sqrt{\rho}\right)^2} \left(\frac{2}{\delta^2}\right) \eta_l^2 s^4 L^2 \frac{1}{T} \sum_{t=1}^{T-1} \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{z}_i^t)\right\|_2^2\right]$$

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

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

so that

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

Putting them together

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

Plug in Proposition 2.

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

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

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

It follows that

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

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

F.6 Spectral norm upper bound (Lemma 4)

Lemma 4 adapts from [58], we present its proof here for completeness.

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

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

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

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

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

- For $j \neq j'$, we have
$$W^2_{jj'} = \sum_{k=1}^{m} W_{jk} W_{j'k} = \frac{1}{|\mathcal{A}|} \mathbb{1}_{\{j \in \mathcal{A}\}} \mathbb{1}_{\{j' \in \mathcal{A}\}};$$
- For $j = j'$, we have
$$W^2_{jj} = \frac{1}{|\mathcal{A}|} \mathbb{1}_{\{j \in \mathcal{A}\}} + \left(1 - \mathbb{1}_{\{j \in \mathcal{A}\}}\right).$$

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

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

- For $j \neq j'$, it holds that
$$\mathbb{E}\left[W^2_{jj'} \mid \mathcal{A} \neq \emptyset\right] = \mathbb{E}\left[\frac{1}{|\mathcal{A}|} \mathbb{1}_{\{j \in \mathcal{A}\}} \mathbb{1}_{\{j' \in \mathcal{A}\}} \Big| \mathcal{A} \neq \emptyset\right] \overset{(a)}{\geq} \mathbb{E}\left[\frac{1}{m} \mathbb{1}_{\{j \in \mathcal{A}\}} \mathbb{1}_{\{j' \in \mathcal{A}\}} \Big| \mathcal{A} \neq \emptyset\right] = \frac{p_j p_{j'}}{m} \geq \frac{\delta^2}{m},$$
where inequality (a) holds because $|\mathcal{A}| \leq m$;
- For $j = j'$, it holds that
$$\begin{aligned}\mathbb{E}\left[W^2_{jj} \mid \mathcal{A} \neq \emptyset\right] &= \mathbb{E}\left[\frac{1}{|\mathcal{A}|} \mathbb{1}_{\{j \in \mathcal{A}\}} + \left(1 - \mathbb{1}_{\{j \in \mathcal{A}\}}\right) \Big| \mathcal{A} \neq \emptyset\right] \\ &\geq \mathbb{E}\left[\frac{1}{m}\left[\mathbb{1}_{\{j \in \mathcal{A}\}} + \left(1 - \mathbb{1}_{\{j \in \mathcal{A}\}}\right)\right] \Big| \mathcal{A} \neq \emptyset\right] = \frac{1}{m} \geq \frac{\delta^2}{m}.\end{aligned}$$

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

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

We note that $\rho(t) = \lambda_2(M)$, where λ_2 is the second largest eigenvalue of matrix M. A Markov chain with M as the transition matrix is ergodic as the chain is (1) *irreducible*: $M_{jj'} \geq \frac{\delta^2}{m}\left[1 - (1 - c)^m\right] > 0$ for $j, j' \in [m]$ and (2) *aperiodic* (it has self-loops). In addition, W matrix is by definition doubly-stochastic. Hence, M has a uniform stationary distribution $\pi = \mathbb{1}^\top / m$. Furthermore, the irreducible Markov chain is reversible since it holds for all the states that $\pi_i M_{ij} = \pi_j M_{ji}$. The conductance Φ of a reversible Markov chain [19] with a transition matrix M can be bounded by

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

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

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

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

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

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

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

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

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

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

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

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

where the last inequality follows from Lemma 6.

For (30), we have

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

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

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

For (32), we have

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

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

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

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

where the last inequality holds because of (11).

Combining the above and rearranging the terms, we get

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

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

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

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

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

H.1 Convergence error of Algorithm 1

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

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

Proof of Corollary 2.

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

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

Further plug in Proposition 3,

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

Finally, plug in Lemma 6.

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

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

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

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

H.2 Convergence rate of Algorithm 1

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

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

□

I Additional Results and Interpretations

I.1 Consensus error of Algorithm 1

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

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

Proof of Corollary 3.

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

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

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

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

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

Finally, we plug in Theorem 1

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

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

In terms of asymptotics, we have

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

□

I.2 Orders of the asymptotic rates

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

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

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

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

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

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

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

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

J Numerical Experiments

J.1 Code

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

J.2 Experimental setups

Hardware and Software Setups.

- **Hardware.** The simulations are performed on a private cluster with 64 CPUs, 500 GB RAM and 8 NVIDIA A5000 GPU cards.
- **Software.** We code the experiments based on PyTorch 1.13.1 [40] and Python 3.7.16.

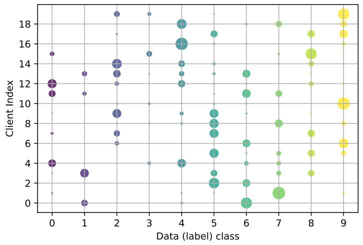


Figure 4: An example of data heterogeneity using Dirichlet($\alpha = 0.1$) distribution with 20 clients. x-axis denotes the categories of images, while y-axis denotes the client index. The size of a circle refers to the proportion of pictures in a given class. The color of a circle distinguishes images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

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

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` [44], which is tuned over a grid of $\{0.1, 0.05, 0.01, 0.005, 0.001, 0.0005\}$. In addition, as recommended by [55], we choose $K = 50$ in `FedAU` without further specification. We train CNNs on all datasets for 2000 rounds. Fig. 3 adopts the same hyperparameter setups, yet with only 1000 training rounds.

Datasets and Data Heterogeneity.

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

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

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

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

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

- **SVHN [37].** The dataset contains 32×32 colored images of 10 different number digits. In total, there are 73257 train images and 26032 test images.
- **CIFAR-10 [26].** The dataset contains 32×32 colored images of 10 different objects. In total, there are 50000 train images and 10000 test images.
- **CINIC-10[12].** The dataset contains 32×32 colored images of 10 different objects. In total, there are 90000 train images and 90000 test images.

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

J.3 Non-stationary client unavailability dynamics

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

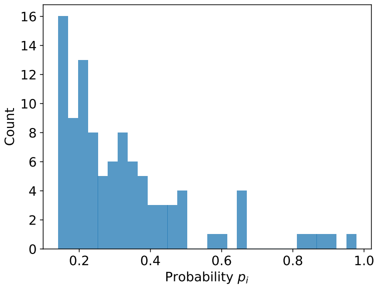


Figure 5: A histogram of one generated p_i's example with a total of $m = 100$ clients. It can be seen that the majority of p_i's are below 0.5.

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

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

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

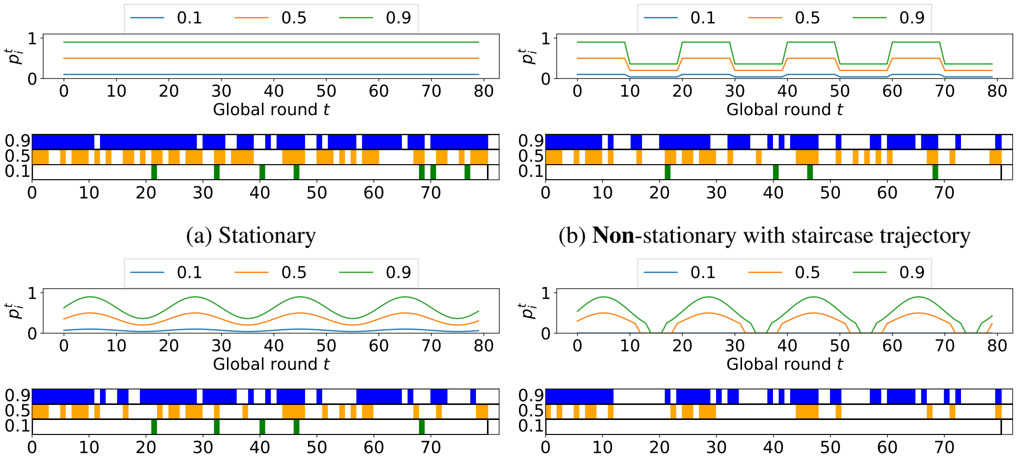
  where P defines a period, $t_0 \in \{0, P, 2P, 3P, \ldots\}$.
- Non-stationary with sine trajectory:

$$f_i(t) \triangleq \gamma \sin(2\pi / P \cdot t) + (1 - \gamma),$$

  where γ signifies the degree of non-stationary.
- Non-stationary with interleaved sine trajectory:

$$f_i(t) \triangleq g_i(t) \cdot \mathbb{1}_{\{p_i \cdot g_i(t) \geq \delta_0\}},$$

  where $g_i(t) \triangleq \gamma \sin(2\pi / P \cdot t) + (1 - \gamma)$ and $\delta_0 = 0.1$ defines a cutting-off lower bound. Specifically, δ_0 cuts off the sine curve and brings in a period of zero-valued probabilities. As different clients have different p_i's, the cut-off points are not synchronized among clients, leading to additional availability heterogeneity.



(a) Stationary

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

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

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

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

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

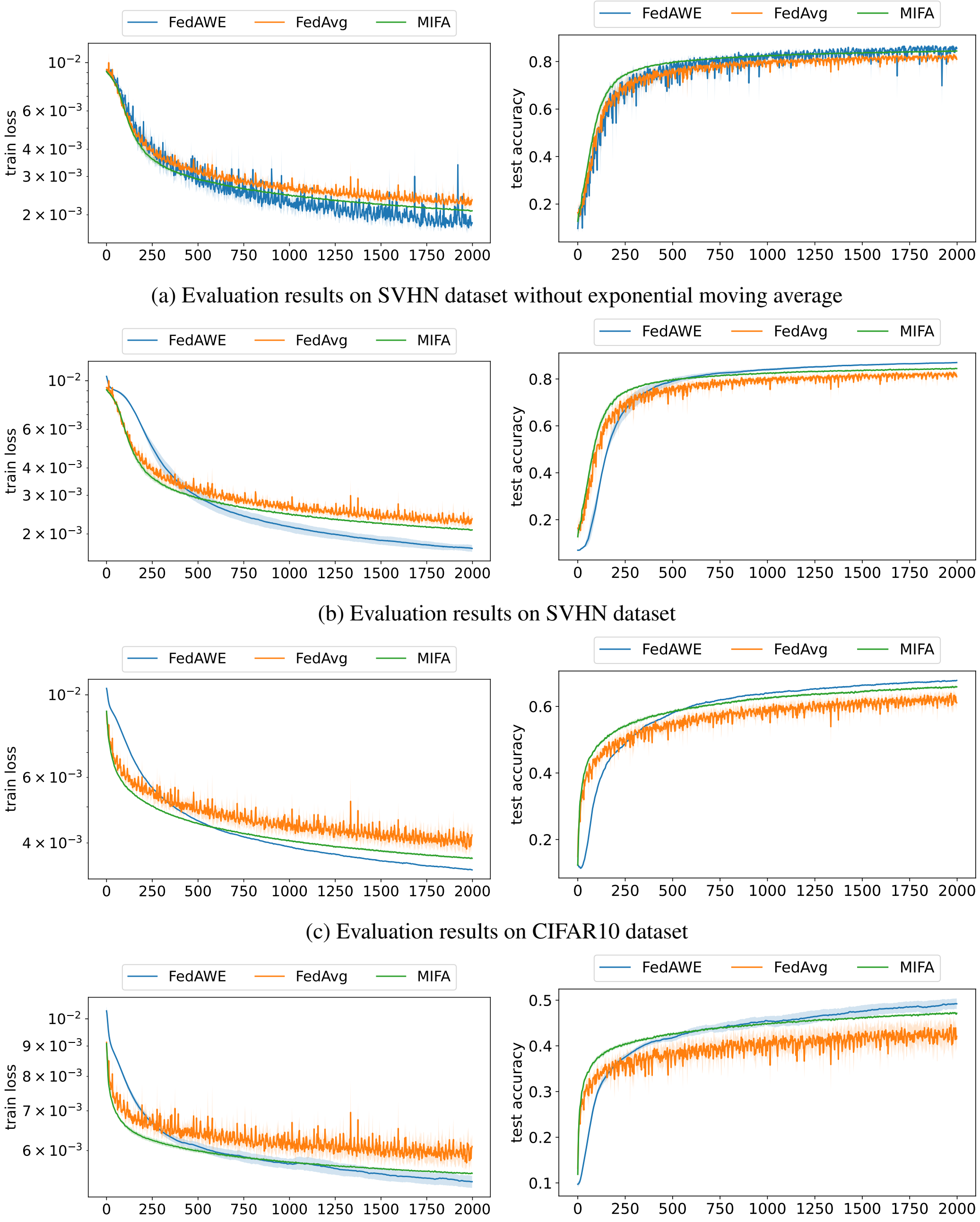
Datasets	**SVHN**				**CIFAR10**				**CINIC10**			
Quarters	**1/4**	**1/2**	**3/4**	**1**	**1/4**	**1/2**	**3/4**	**1**	**1/4**	**1/2**	**3/4**	**1**
Test accuracy	20%	40%	60%	80%	15%	30%	45%	60%	10%	20%	30%	40%
FedAWE (**ours**)	40	120	200	820	20	60	200	1360	0	20	120	540
FedAvg over *active* clients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

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

J.4 Additional results

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

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



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

(b) Evaluation results on SVHN dataset

(c) Evaluation results on CIFAR10 dataset

(d) Evaluation results on CINIC10 dataset

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

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

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

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

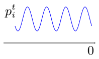
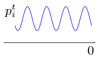
Unavailable Dynamics	Datasets	$\gamma = 0.3$		$\gamma = 0.2$		$\gamma = 0.1$	
	Algorithms	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Non-stationary (**Sine**) 	FedAWE (**ours**)	**85.7** ± 0.9 %	**85.6** ± 0.9 %	**85.7** ± 0.5 %	**85.7** ± 0.5 %	**85.8** ± 0.6 %	**85.7** ± 0.7 %
	FedAvg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

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

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

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

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

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Please refer to Appendix A for details.

3. **Theory Assumptions and Proofs**

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

   Answer: [Yes]

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

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

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

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

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

   Answer: [Yes]

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

6. **Experimental Setting/Details**

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

   Answer: [Yes]

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

7. **Experiment Statistical Significance**

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

   Answer: [Yes]

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

8. **Experiments Compute Resources**

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

Answer: [Yes]

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

9. **Code Of Ethics**

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

   Answer: [Yes]

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

10. **Broader Impacts**

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

    Answer: [Yes]

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: The paper poses no such risks

12. **Licenses for existing assets**

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

    Answer: [Yes]

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

13. **New Assets**

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

    Answer: [Yes]

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

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

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

    Answer: [NA]

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

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

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

    Answer: [NA]

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