Supplementary Material: Federated Learning under Heterogeneous and Correlated Client Availability

Appendix A
Proof of Theorem 1

Theorem 1 (Decomposing the total error)**.** *Let $\kappa := L/\mu$. Under Assumptions 2–4, the optimization error of the target global objective $\epsilon = F(\boldsymbol{w}) - F^*$ can be bounded as follows:*

$$\epsilon \leq 2\kappa^2 (\underbrace{F_B(\boldsymbol{w}) - F_B^*}_{:=\epsilon_{opt}} + \underbrace{F(\boldsymbol{w}_B^*) - F^*}_{:=\epsilon_{bias}}). \tag{10}$$

Moreover, let $\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}} := \sum_{k=1}^N (\alpha_k - p_k)^2/p_k$. Then:

$$\epsilon_{bias} \leq \kappa^2 \cdot \underbrace{\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}} \cdot \Gamma}_{:=\epsilon_{bias}}. \tag{11}$$

The proof of Theorem 1 employs well-established techniques from convex optimization. It is based on the proof presented in [33, Theorem 2].

Proof of Theorem 1. By leveraging the L-smoothness and μ-strong convexity properties of F, we obtain:

$$\begin{aligned}
F(\boldsymbol{w}) - F^* &\leq \frac{1}{2\mu} \|\nabla F(\boldsymbol{w})\|^2 && (21)\\
&\leq \frac{L^2}{2\mu} \|\boldsymbol{w} - \boldsymbol{w}^*\|^2 && (22)\\
&\leq \frac{L^2}{\mu} (\|\boldsymbol{w} - \boldsymbol{w}_B^*\|^2 + \|\boldsymbol{w}_B^* - \boldsymbol{w}^*\|^2) && (23)\\
&\leq \frac{2L^2}{\mu^2} \Big(\underbrace{F_B(\boldsymbol{w}) - F_B^*}_{:=\epsilon_{\text{opt}}} + \underbrace{F(\boldsymbol{w}_B^*) - F^*}_{:=\epsilon_{\text{bias}}} \Big), && (24)
\end{aligned}$$

where the inequality in (21) follows from Assumption 4 and is commonly referred to as the *Polyak-Lojasiewicz inequality*; the inequality in (22) is derived using the fact that $\nabla F(\boldsymbol{w}^*) = 0$ (Assumption 2) and the definition of L-Lipschitz continuous gradient for F (Assumption 3); the inequality in (23) is based on $(a+b)^2 \leq 2(a^2+b^2)$; lastly, the inequality in (24) follows from the μ-strong convexity of both F_B and F (Assumptions 4), and uses $\nabla F_B(\boldsymbol{w}_B^*) = 0$ and $\nabla F(\boldsymbol{w}^*) = 0$ (Assumption 2). The obtained results complete the first part of the proof, establishing the bound in (10).

Next, to prove the relation in (11), we proceed by bounding the term ϵ_{bias} as follows:

$$\epsilon_{\text{bias}} := (F(\boldsymbol{w}_B^*) - F^*) \leq \frac{1}{2\mu} \|\nabla F(\boldsymbol{w}_B^*)\|^2, \tag{25}$$

where the inequality in (25) directly follows from the Polyak-Lojasiewicz inequality (Assumption 4).

Furthermore, we bound the term $\|\nabla F(\boldsymbol{w}_B^*)\|$ as follows:

$$\begin{aligned}
\|\nabla F(\boldsymbol{w}_B^*)\| &= \left\| \sum_{k=1}^N (\alpha_k - p_k) \nabla F_k(\boldsymbol{w}_B^*) \right\| && (26)\\
&\leq \sum_{k=1}^N |\alpha_k - p_k| \, \|\nabla F_k(\boldsymbol{w}_B^*)\| && (27)\\
&\leq L \sum_{k=1}^N |\alpha_k - p_k| \, \|\boldsymbol{w}_B^* - \boldsymbol{w}_k^*\| && (28)\\
&\leq L \sqrt{\frac{2}{\mu}} \sum_{k=1}^N |\alpha_k - p_k| \sqrt{(F_k(\boldsymbol{w}_B^*) - F_k^*)}, && (29)
\end{aligned}$$

where, in (26), we use $\nabla F_B(\boldsymbol{w}_B^*) = 0$ (Assumption 2) and apply the definitions of F and F_B given in (1) and (4), respectively. The bound in (27) follows from the triangle inequality. Next, the inequality in (28) uses $\nabla F_k(\boldsymbol{w}_k^*) = 0$ (Assumption 2) and

the L-smoothness of F_k (Assumption 3). Finally, the inequality in (29) leverages the μ-strong convexity of F_k (Assumption 4) and $\nabla F_k(\boldsymbol{w}_k^*) = 0$ (Assumption 2), and follows multiplying and dividing by $\sqrt{p_k}$.

By squaring both sides of Equation (29), we obtain:

$$\|\nabla F(\boldsymbol{w}_B^*)\|^2 \leq \frac{2L^2}{\mu}\left(\sum_{k=1}^{N}\frac{|\alpha_k - p_k|}{\sqrt{p_k}}\sqrt{p_k(F_k(\boldsymbol{w}_B^*) - F_k^*)}\right)^2 \tag{30}$$

$$\leq \frac{2L^2}{\mu}\left(\sum_{k=1}^{N}\frac{(\alpha_k - p_k)^2}{p_k}\right)\left(\sum_{k=1}^{N} p_k(F_k(\boldsymbol{w}_B^*) - F_k^*)\right) \tag{31}$$

$$\leq \frac{2L^2}{\mu}\cdot\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}}\cdot\Gamma, \tag{32}$$

where the inequality in (31) follows from the Cauchy-Schwarz inequality. Furthermore, the inequality in (32) holds because:

$$\sum_{k=1}^{N} p_k(F_k(\boldsymbol{w}_B^*) - F_k^*) = F_B^* - \sum_{k=1}^{N} p_k F_k^* \tag{33}$$

$$\leq F_B(\boldsymbol{w}^*) - \sum_{k=1}^{N} p_k F_k^* \tag{34}$$

$$= \sum_{k=1}^{N} p_k(F_k(\boldsymbol{w}^*) - F_k^*) \tag{35}$$

$$\leq \max_{k\in\mathcal{K}}\{F_k(\boldsymbol{w}^*) - F_k^*\} \coloneqq \Gamma. \tag{36}$$

We remark that the inequality in (34) only holds if $\boldsymbol{w}_B^*$ is the global minimizer of F_B, as guaranteed by Assumption 2. By replacing (32) into (25), we have:

$$\epsilon_{\text{bias}} \leq \frac{1}{2\mu}\|\nabla F(\boldsymbol{w}_B^*)\|^2 \leq \frac{L^2}{\mu^2}\cdot\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}}\cdot\Gamma, \tag{37}$$

which concludes the proof of Equation (11), and therefore, of Theorem 1. □

APPENDIX B
PROOF OF THEOREM 2

B1. Algorithm Overview and Supplementary Notation

Let $\boldsymbol{w}_{t,j}^k$ represent the model parameter maintained by the k-th client during the t-th global communication round and the j-th local step. The t-th global communication round can be described as follows: 1) The server broadcasts the model parameter $\boldsymbol{w}_{t,0}$ to the active clients, which adopt it as their local model, i.e., $\boldsymbol{w}_{t,0}^k = \boldsymbol{w}_{t,0}$ for $k \in \mathcal{A}_t$; 2) Each active client $k \in \mathcal{A}_t$ generates a sequence of local models $\{\boldsymbol{w}_{t,j}^k\}_{j=1}^E$ using the local-SGD update rule defined in (2); 3) The active clients send their model updates $\Delta_t^k \coloneqq \boldsymbol{w}_{t,E}^k - \boldsymbol{w}_{t,0}$ back to the server; 4) The server aggregates the model updates using the aggregation rule specified in (3), resulting in the new global model parameter $\boldsymbol{w}_{t+1,0}$.

$$\begin{cases} \boldsymbol{w}_{t,j+1}^k = \boldsymbol{w}_{t,j}^k - \eta_t \nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k) & \text{for } j = 0,\ldots,E-1; \quad (2)\\ \boldsymbol{w}_{t+1,0} = \mathbf{Proj}_W(\boldsymbol{w}_{t,0} + \sum_{k\in\mathcal{A}_t} q_k\left(\boldsymbol{w}_{t,E}^k - \boldsymbol{w}_{t,0}\right)) & \text{for } j = E. \quad (3) \end{cases}$$

The projection operator in (3) ensures that the current iterate $\boldsymbol{w}_{t+1,0}$ in the optimization algorithm defined by (2) and (3) remains within the feasible region W.

Sources of randomness: In the system, we model two sources of randomness. The first arises from the availability of random clients, which follows a Markov process as stated in Assumption 1. The second source of randomness originates from the random sampling of batches for computing stochastic gradients. Remember that $\mathcal{A}_t$ denotes the random set of clients available at the t-th communication round and that $\mathcal{B}_{t,j}^k$ denotes the random batch independently sampled from client-k's local dataset at round t, local iteration j. For the analysis, we introduce the following additional notation:

- $\mathcal{A}_{i:j} \coloneqq \{\mathcal{A}_i, \ldots, \mathcal{A}_j\}$: the family of random sets of clients available from the i-th to the j-th communication rounds, $i<j$;
- $\mathcal{B}_t^k \coloneqq \{\mathcal{B}_{t,j}^k\}_{j=0}^{E-1}$: the set of random batches sampled by the k-th client at the t-th communication round;
- $\mathcal{B}_t \coloneqq \{\mathcal{B}_t^k\}_{k\in\mathcal{A}_t}$: the set of random batches sampled by the available clients ($\mathcal{A}_t$) in the t-th communication round;
- $\mathcal{B}_{t,i:j}^k \coloneqq \{\mathcal{B}_{t,i}^k, \ldots, \mathcal{B}_{t,j}^k\}$: the set of random batches sampled by the k-th client at the t-th communication round between the i-th and the j-th local iterations, $i < j$;
- $\mathcal{B}_{i:j} \coloneqq \{\mathcal{B}_i, \ldots, \mathcal{B}_j\}$: the set of random batches sampled by the available clients ($\mathcal{A}_{i:j}$) between the i-th and j-th communication rounds, $i < j$.

With this notation established, the randomness in the t-th communication round, which starts with the initial model $\boldsymbol{w}_{t,0}$ and yields the updated model $\boldsymbol{w}_{t+1,0}$, is fully determined by the sets $\mathcal{A}_t$ and $\mathcal{B}_t$. This implies that the evolution of the algorithm, governed by the update rules in (2) and (3), from round 0 to round t can be completely described by the tuple:

$$\mathcal{H}_t \coloneqq (\mathcal{A}_0, \ldots, \mathcal{A}_{t-1}; \mathcal{B}_0, \ldots, \mathcal{B}_{t-1}), \tag{38}$$

which represents the historical information up to the t-th communication round.

We introduce the following additional quantities for our analysis:

$$\boldsymbol{g}_t(\mathcal{A}_t, \mathcal{B}_t) \coloneqq \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k), \tag{39}$$

and

$$\bar{\boldsymbol{g}}_t(\mathcal{A}_t, \mathcal{B}_t) \coloneqq \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \nabla F_k(\boldsymbol{w}_{t,j}^k), \tag{40}$$

where $\boldsymbol{g}_t(\mathcal{A}_t, \mathcal{B}_t)$ denotes the global pseudo-gradient computed at communication round t, aggregated from the active clients in $\mathcal{A}_t$, and $\bar{\boldsymbol{g}}_t(\mathcal{A}_t, \mathcal{B}_t)$ denotes its expected value with respect to the choices of the random batches $\mathcal{B}_{t,j}^k$, for all $j = 0, \ldots, E-1$ and $k \in \mathcal{A}_t$. With this notation established, the global update rule for the t-th communication round can be expressed as:

$$\boldsymbol{w}_{t+1,0} = \mathbf{Proj}_W(\boldsymbol{w}_{t,0} - \eta_t \boldsymbol{g}_t(\mathcal{A}_t, \mathcal{B}_t)). \tag{41}$$

B2. Supporting Lemmas

In this section, we introduce several lemmas that are instrumental in proving Theorem 2. Firstly, we prove Lemma 1, introduced in Section III-A. Its proof relies on the convexity and compactness of the hypothesis class W (Assumption 2), on the L-smoothness of the functions $\{F_k\}_{k\in\mathcal{K}}$ (Assumption 3), and on the bounded variance of the stochastic gradients (Assumption 5).

Lemma 1. *Under Assumptions 2, 3, and 5, there exist constants D, G, and $H > 0$, such that, for $\boldsymbol{w} \in W$ and $k \in \mathcal{K}$, we have:*

$$\|\nabla F_k(\boldsymbol{w})\| \leq D, \tag{6}$$
$$\mathbb{E}\,\|\nabla F_k(\boldsymbol{w}, \xi)\|^2 \leq G^2, \tag{7}$$
$$|F_k(\boldsymbol{w}) - F_k(\boldsymbol{w}_B^*)| \leq H. \tag{8}$$

Proof of Lemma 1. The boundedness of the hypothesis class W (Assumption 2) provides a bound on the sequence $(\boldsymbol{w}_{t,0})_{t\geq 0}$ generated by the scheme defined in Equations (2) and (3). Moreover, since $\boldsymbol{w}_k^*$ minimizes $\nabla F_k(\boldsymbol{w})$, we have $\nabla F_k(\boldsymbol{w}_k^*) = 0$. Furthermore, the L-smoothness of $\{F_k\}_{k\in\mathcal{K}}$ (Assumption 3) leads to the following inequality:

$$\|\nabla F_k(\boldsymbol{w})\| = \|\nabla F_k(\boldsymbol{w}) - \nabla F_k(\boldsymbol{w}_k^*)\| \leq L\,\|\boldsymbol{w} - \boldsymbol{w}_k^*\| \coloneqq D < +\infty. \tag{42}$$

The bound in (6) is directly derived from (42), while the bound in (8) follows from the continuity of $\{F_k\}_{k\in\mathcal{K}}$ over the compact set W (Assumption 2). Finally, the inequality in (7) requires a bound on the variance of the stochastic gradients (Assumption 5). In particular, it holds that:

$$\mathbb{E}\,\|\nabla F_k(\boldsymbol{w},\xi)\|^2 \leq D^2 + \max_{k\in\mathcal{K}}\{\sigma_k^2\} \coloneqq G^2. \tag{43}$$

□

The following lemma proves that the global pseudo-gradient $\boldsymbol{g}_t(\mathcal{A}_t,\mathcal{B}_t)$ is an unbiased estimator of $\bar{\boldsymbol{g}}_t(\mathcal{A}_t,\mathcal{B}_t)$. A similar result has been used in previous works, specifically in [33, Appendix C1]. Here, we provide a comprehensive proof for this result.

Lemma 2. *Let $\boldsymbol{g}_t(\mathcal{A}_t,\mathcal{B}_t)$ and $\bar{\boldsymbol{g}}_t(\mathcal{A}_t,\mathcal{B}_t)$ be defined as in (39) and (40), respectively. The following equality holds:*

$$\mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left[\boldsymbol{g}_t(\mathcal{A}_t,\mathcal{B}_t)\right] = \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left[\bar{\boldsymbol{g}}_t(\mathcal{A}_t,\mathcal{B}_t)\right]. \tag{44}$$

Proof of Lemma 2.

$$\mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left[\boldsymbol{g}_t(\mathcal{A}_t,\mathcal{B}_t)\right] = \tag{45}$$

$$= \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left[\sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \nabla F_k(\boldsymbol{w}_{t,j}^k,\mathcal{B}_{t,j}^k)\right] \tag{46}$$

$$= \sum_{k\in\mathcal{A}_t} q_k \mathop{\mathbb{E}}_{\mathcal{B}_t^k}\left[\sum_{j=0}^{E-1} \nabla F_k(\boldsymbol{w}_{t,j}^k,\mathcal{B}_{t,j}^k)\right] \tag{47}$$

$$= \sum_{k\in\mathcal{A}_t} q_k \left[\mathop{\mathbb{E}}_{\mathcal{B}_{t,0}^k}\left[\nabla F_k(\boldsymbol{w}_{t,0},\mathcal{B}_{t,0}^k)\right] + \mathop{\mathbb{E}}_{\mathcal{B}_{t,0}^k,\mathcal{B}_{t,1}^k}\left[\nabla F_k(\boldsymbol{w}_{t,1}^k,\mathcal{B}_{t,1}^k)\right] + \cdots + \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:E-1}^k}\left[\nabla F_k(\boldsymbol{w}_{t,E-1}^k,\mathcal{B}_{t,E-1}^k)\right]\right] \tag{48}$$

$$= \sum_{k\in\mathcal{A}_t} q_k \left[\nabla F_k(\boldsymbol{w}_{t,0}) + \mathop{\mathbb{E}}_{\mathcal{B}_{t,0}^k}\left[\mathop{\mathbb{E}}_{\mathcal{B}_{t,1}^k|\mathcal{B}_{t,0}^k}\left[\nabla F_k(\boldsymbol{w}_{t,1}^k,\mathcal{B}_{t,1}^k)\right]\right] + \cdots + \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:E-2}^k}\left[\mathop{\mathbb{E}}_{\mathcal{B}_{t,E-1}^k|\mathcal{B}_{t,0:E-2}^k}\left[\nabla F_k(\boldsymbol{w}_{t,E-1}^k,\mathcal{B}_{t,E-1}^k)\right]\right]\right] \tag{49}$$

$$= \sum_{k\in\mathcal{A}_t} q_k \left[\nabla F_k(\boldsymbol{w}_{t,0}) + \mathop{\mathbb{E}}_{\mathcal{B}_{t,0}^k}\left[\nabla F_k(\boldsymbol{w}_{t,1}^k)\right] + \cdots + \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:E-2}^k}\left[\nabla F_k(\boldsymbol{w}_{t,E-1}^k)\right]\right] \tag{50}$$

$$= \sum_{k\in\mathcal{A}_t} q_k \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:E-2}^k}\left[\sum_{j=0}^{E-1} \nabla F_k(\boldsymbol{w}_{t,j}^k)\right] \tag{51}$$

$$= \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left[\sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \nabla F_k(\boldsymbol{w}_{t,j}^k)\right] = \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left[\bar{\boldsymbol{g}}_t(\mathcal{A}_t,\mathcal{B}_t)\right], \tag{52}$$

where, in (47), we considered that both the evolution of the local models $\{\boldsymbol{w}_{t,j}^k\}_{j=0}^{E-1}$ and the choices of the random batches $\{\mathcal{B}_{t,j}^k\}_{j=0}^{E-1}$ are independent among different clients $k\in\mathcal{A}_t$ within the same communication round $t\in\mathcal{T}$. □

For the sake of simplicity, we will henceforth denote $\boldsymbol{g}_t(\mathcal{A}_t,\mathcal{B}_t)$ and $\bar{\boldsymbol{g}}_t(\mathcal{A}_t,\mathcal{B}_t)$ as $\boldsymbol{g}_t$ and $\bar{\boldsymbol{g}}_t$, respectively. The following lemma decomposes the optimization error into multiple components, which we will bound separately in subsequent lemmas.

Lemma 3 (Decomposition of the error in a global communication round)**.** *Let Assumption 2 hold. We have:*

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left\|\boldsymbol{w}_{t+1,0}-\boldsymbol{w}_B^*\right\|^2 \leq \left\|\boldsymbol{w}_{t,0}-\boldsymbol{w}_B^*\right\|^2 &\underbrace{-2\eta_t \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\langle \boldsymbol{w}_{t,0}-\boldsymbol{w}_B^*,\bar{\boldsymbol{g}}_t\rangle}_{\text{bounded in Lemma 4}} + \underbrace{\eta_t^2 \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left\|\bar{\boldsymbol{g}}_t\right\|^2}_{\text{bounded in Lemma 5}} \\ &\underbrace{+2\eta_t \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\langle \boldsymbol{w}_{t,0}-\boldsymbol{w}_B^*-\eta_t\bar{\boldsymbol{g}}_t,\bar{\boldsymbol{g}}_t-\boldsymbol{g}_t\rangle}_{\text{bounded in Lemma 6}} + \underbrace{\eta_t^2 \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left\|\boldsymbol{g}_t-\bar{\boldsymbol{g}}_t\right\|^2}_{\text{bounded in Lemma 7}}.\end{aligned} \tag{53}$$

Proof of Lemma 3.

$$\left\|\boldsymbol{w}_{t+1,0}-\boldsymbol{w}_B^*\right\|^2 = \left\|\mathbf{Proj}_W(\boldsymbol{w}_{t,0}-\eta_t\boldsymbol{g}_t)-\mathbf{Proj}_W(\boldsymbol{w}_B^*)\right\|^2 \tag{54}$$

$$\leq \|\boldsymbol{w}_{t,0} - \eta_t \boldsymbol{g}_t - \boldsymbol{w}_B^* + \eta_t \bar{\boldsymbol{g}}_t - \eta_t \bar{\boldsymbol{g}}_t\|^2 \tag{55}$$

$$= \|\boldsymbol{w}_{t,0} - \boldsymbol{w}_B^* - \eta_t \bar{\boldsymbol{g}}_t\|^2 + 2\eta_t \langle \boldsymbol{w}_{t,0} - \boldsymbol{w}_B^* - \eta_t \bar{\boldsymbol{g}}_t, \bar{\boldsymbol{g}}_t - \boldsymbol{g}_t \rangle + \eta_t^2 \|\boldsymbol{g}_t - \bar{\boldsymbol{g}}_t\|^2 \tag{56}$$

$$= \|\boldsymbol{w}_{t,0} - \boldsymbol{w}_B^*\|^2 - 2\eta_t \langle \boldsymbol{w}_{t,0} - \boldsymbol{w}_B^*, \bar{\boldsymbol{g}}_t \rangle + \eta_t^2 \|\bar{\boldsymbol{g}}_t\|^2 + 2\eta_t \langle \boldsymbol{w}_{t,0} - \boldsymbol{w}_B^* - \eta_t \bar{\boldsymbol{g}}_t, \bar{\boldsymbol{g}}_t - \boldsymbol{g}_t \rangle + \eta_t^2 \|\boldsymbol{g}_t - \bar{\boldsymbol{g}}_t\|^2 , \tag{57}$$

where, in (54), we used Assumption 2; whereas, the inequality in (55) is due to the contracting property of projection. We observe that (55) does not hold in general if $\boldsymbol{w}_B^* \notin W$. □

In what follows, we present a series of lemmas to establish bounds for the error in (53).

Lemma 4. *Let Assumption 3 hold and the local functions $\{F_k\}_{k=1}^N$ be convex. We have:*

$$-2\eta_t \langle \boldsymbol{w}_{t,0} - \boldsymbol{w}_B^*, \bar{\boldsymbol{g}}_t \rangle \leq -2\eta_t (1 - \eta_t L) \sum_{k \in \mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left(F_k(\boldsymbol{w}_{t,j}^k) - F_k(\boldsymbol{w}_B^*) \right) + \underbrace{\sum_{k \in \mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left\| \boldsymbol{w}_{t,j}^k - \boldsymbol{w}_{t,0} \right\|^2}_{\text{bounded in Lemma 9}} + 2\eta_t^2 LE \underbrace{\sum_{k \in \mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_B^*) - F_k^* \right)}_{\text{bounded in Lemma 10}} . \tag{58}$$

Proof of Lemma 4. We decompose the term $-2\eta_t \langle \boldsymbol{w}_{t,0} - \boldsymbol{w}_B^*, \bar{\boldsymbol{g}}_t \rangle$, by adding and subtracting $\boldsymbol{w}_{t,j}^k$:

$$-2\eta_t \langle \boldsymbol{w}_{t,0} - \boldsymbol{w}_B^*, \bar{\boldsymbol{g}}_t \rangle = \underbrace{-2\eta_t \langle \boldsymbol{w}_{t,0} - \boldsymbol{w}_{t,j}^k, \bar{\boldsymbol{g}}_t \rangle}_{\text{developed in Eq. (60)}} \underbrace{-2\eta_t \langle \boldsymbol{w}_{t,j}^k - \boldsymbol{w}_B^*, \bar{\boldsymbol{g}}_t \rangle}_{\text{developed in Eq. (64)}} . \tag{59}$$

We bound the two terms separately. We bound the first term in (59) as:

$$-2\eta_t \langle \boldsymbol{w}_{t,0} - \boldsymbol{w}_{t,j}^k, \bar{\boldsymbol{g}}_t \rangle = -2\eta_t \sum_{k \in \mathcal{A}_t} q_k \sum_{j=0}^{E-1} \langle \nabla F_k(\boldsymbol{w}_{t,j}^k), \boldsymbol{w}_{t,0} - \boldsymbol{w}_{t,j}^k \rangle \tag{60}$$

$$\leq \eta_t^2 \sum_{k \in \mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left\| \nabla F_k(\boldsymbol{w}_{t,j}^k) \right\|^2 + \sum_{k \in \mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left\| \boldsymbol{w}_{t,j}^k - \boldsymbol{w}_{t,0} \right\|^2 \tag{61}$$

$$\leq 2\eta_t^2 L \sum_{k \in \mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left(F_k(\boldsymbol{w}_{t,j}^k) - F_k^* \right) + \sum_{k \in \mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left\| \boldsymbol{w}_{t,j}^k - \boldsymbol{w}_{t,0} \right\|^2 \tag{62}$$

$$= 2\eta_t^2 L \sum_{k \in \mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left(F_k(\boldsymbol{w}_{t,j}^k) - F_k(\boldsymbol{w}_B^*) \right) + \sum_{k \in \mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left\| \boldsymbol{w}_{t,j}^k - \boldsymbol{w}_{t,0} \right\|^2 + 2\eta_t^2 LE \sum_{k \in \mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_B^*) - F_k^* \right) , \tag{63}$$

where, in (61), we used $|\langle \boldsymbol{a}, \boldsymbol{b} \rangle| \leq \frac{1}{2} \|\boldsymbol{a}\|^2 + \frac{1}{2} \|\boldsymbol{b}\|^2$; in (62), we applied the L-smoothness of $\{F_k(\boldsymbol{w})\}_{k \in \mathcal{K}}$ (Assumption 3); in (63), we added and subtracted $F_k(\boldsymbol{w}_B^*)$.

We bound the second term in (59) as:

$$-2\eta_t \langle \boldsymbol{w}_{t,j}^k - \boldsymbol{w}_B^*, \bar{\boldsymbol{g}}_t \rangle = -2\eta_t \sum_{k \in \mathcal{A}_t} q_k \sum_{j=0}^{E-1} \langle \boldsymbol{w}_{t,j}^k - \boldsymbol{w}_B^*, \nabla F_k(\boldsymbol{w}_{t,j}^k) \rangle \tag{64}$$

$$\leq -2\eta_t \sum_{k \in \mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left(F_k(\boldsymbol{w}_{t,j}^k) - F_k(\boldsymbol{w}_B^*) \right) , \tag{65}$$

where, in (65), we use the convexity of $\{F_k(\boldsymbol{w})\}_{k \in \mathcal{K}}$.

By summing the bounds provided in (63) and (65), we conclude the proof. □

Lemma 5 (Bound on the squared norm of a global gradient step)**.** *Let Assumption 3 hold. We have:*

$$\eta_t^2 \|\bar{\boldsymbol{g}}_t\|^2 \leq 2\eta_t^2 LEQ \sum_{k \in \mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left(F_k(\boldsymbol{w}_{t,j}^k) - F_k(\boldsymbol{w}_B^*) \right) + 2\eta_t^2 LE^2 Q \underbrace{\sum_{k \in \mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_B^*) - F_k^* \right)}_{\text{bounded in Lemma 10}} . \tag{66}$$

Proof of Lemma 5.

$$\eta_t^2 \left\|\bar{\boldsymbol{g}}_t\right\|^2 = \eta_t^2 \left\| \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \nabla F_k(\boldsymbol{w}_{t,j}^k) \right\|^2 \tag{67}$$

$$\leq \eta_t^2 \sum_{k'\in\mathcal{A}_t} q_{k'} \sum_{k\in\mathcal{A}_t} q_k \left\| \sum_{j=0}^{E-1} \nabla F_k(\boldsymbol{w}_{t,j}^k) \right\|^2 \tag{68}$$

$$\leq \eta_t^2 QE \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left\| \nabla F_k(\boldsymbol{w}_{t,j}^k) \right\|^2 \tag{69}$$

$$\leq 2\eta_t^2 QLE \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left(F_k(\boldsymbol{w}_{t,j}^k) - F_k^* \right) \tag{70}$$

$$= 2\eta_t^2 LEQ \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left(F_k(\boldsymbol{w}_{t,j}^k) - F_k(\boldsymbol{w}_B^*) \right) + 2\eta_t^2 LE^2 Q \sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_B^*) - F_k^* \right), \tag{71}$$

where, in (68) and in (69), we applied the Jensen's inequality; in (69), we also observed that $\sum_{k\in\mathcal{A}_t} q_k \leq \sum_{k\in\mathcal{K}} q_k \coloneqq Q$; in (70), we used the L-smoothness of $\{F_k(\boldsymbol{w})\}_{k\in\mathcal{K}}$ (Assumption 3); in (71), we added and subtracted $F_k(\boldsymbol{w}_B^*)$ to the sum. □

Lemma 6. *Let Assumption 5 hold. We have:*

$$\begin{aligned} 2\eta_t \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \left[\langle \boldsymbol{w}_{t,0} - \boldsymbol{w}_B^* - \eta_t \bar{\boldsymbol{g}}_t, \bar{\boldsymbol{g}}_t - \boldsymbol{g}_t \rangle\right] \; &\leq 2\eta_t^2 LEQ \sum_{k\in\mathcal{A}_t} q_k \sum_{j=1}^{E-1} \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t} \left[F_k(\boldsymbol{w}_{t,j}^k) - F_k(\boldsymbol{w}_B^*)\right] \\ &\quad + \frac{1}{2}\eta_t^2 E(E-1) \sum_{k\in\mathcal{A}_t} q_k^2 \sigma_k^2 \\ &\quad + 2\eta_t^2 LE^2 Q \underbrace{\sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_B^*) - F_k^* \right)}_{\text{bounded in Lemma 10}}. \end{aligned} \tag{72}$$

Proof of Lemma 6. We decompose the term $\langle \boldsymbol{w}_{t,0} - \boldsymbol{w}_B^* - \eta_t \bar{\boldsymbol{g}}_t, \bar{\boldsymbol{g}}_t - \boldsymbol{g}_t \rangle$ in two parts:

$$2\eta_t \langle \boldsymbol{w}_{t,0} - \boldsymbol{w}_B^* - \eta_t \bar{\boldsymbol{g}}_t, \bar{\boldsymbol{g}}_t - \boldsymbol{g}_t \rangle = 2\eta_t \langle \boldsymbol{w}_{t,0} - \boldsymbol{w}_B^*, \bar{\boldsymbol{g}}_t - \boldsymbol{g}_t \rangle - 2\eta_t^2 \langle \bar{\boldsymbol{g}}_t, \bar{\boldsymbol{g}}_t - \boldsymbol{g}_t \rangle. \tag{73}$$

From Lemma 2, we conclude that $\mathbb{E}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \langle \boldsymbol{w}_{t,0} - \boldsymbol{w}_B^*, \bar{\boldsymbol{g}}_t - \boldsymbol{g}_t \rangle = 0$.

We now focus on:

$$-2\eta_t^2 \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \left[\langle \bar{\boldsymbol{g}}_t, \bar{\boldsymbol{g}}_t - \boldsymbol{g}_t \rangle\right] = \tag{74}$$

$$= -2\eta_t^2 \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \left[\sum_{k\in\mathcal{A}_t} \sum_{k'\in\mathcal{A}_t} q_k q_{k'} \sum_{j=0}^{E-1} \sum_{j'=0}^{E-1} \langle \nabla F_k(\boldsymbol{w}_{t,j}^k), \nabla F_{k'}(\boldsymbol{w}_{t,j'}^{k'}) - \nabla F_{k'}(\boldsymbol{w}_{t,j'}^{k'}, \mathcal{B}_{t,j'}^{k'}) \rangle \right] \tag{75}$$

$$= -2\eta_t^2 \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \left[\sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=0}^{E-1} \sum_{j'=0}^{E-1} \langle \nabla F_k(\boldsymbol{w}_{t,j}^k), \nabla F_k(\boldsymbol{w}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k, \mathcal{B}_{t,j'}^k) \rangle \right]$$

$$- 2\eta_t^2 \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \left[\sum_{k\in\mathcal{A}_t} \sum_{\substack{k'\in\mathcal{A}_t \\ k'\neq k}} q_k q_{k'} \sum_{j=0}^{E-1} \sum_{j'=0}^{E-1} \langle \nabla F_k(\boldsymbol{w}_{t,j}^k), \nabla F_{k'}(\boldsymbol{w}_{t,j'}^{k'}) - \nabla F_{k'}(\boldsymbol{w}_{t,j'}^{k'}, \mathcal{B}_{t,j'}^{k'}) \rangle \right] \tag{76}$$

$$= -2\eta_t^2 \sum_{k\in\mathcal{A}_t} q_k^2 \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t} \left[\sum_{j=0}^{E-1} \sum_{j'=0}^{E-1} \langle \nabla F_k(\boldsymbol{w}_{t,j}^k), \nabla F_k(\boldsymbol{w}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k, \mathcal{B}_{t,j'}^k) \rangle \right]$$

$$-2\eta_t^2 \sum_{k\in\mathcal{A}_t}\sum_{\substack{k'\in\mathcal{A}_t\\k'\neq k}} q_k q_{k'} \sum_{j=0}^{E-1}\sum_{j'=0}^{E-1} \Big\langle \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t}\left[\nabla F_k(\boldsymbol{w}_{t,j}^k)\right], \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:j'-1}^{k'}|\mathcal{A}_t,\mathcal{H}_t} \underbrace{\left[\mathop{\mathbb{E}}_{\mathcal{B}_{t,j'}^{k'}|\mathcal{B}_{t,0:j'-1}^{k'},\mathcal{A}_t,\mathcal{H}_t}\left[\nabla F_{k'}(\boldsymbol{w}_{t,j'}^{k'}) - \nabla F_{k'}(\boldsymbol{w}_{t,j'}^{k'},\mathcal{B}_{t,j'}^{k'})\right]\right]}_{=0}\Big\rangle, \tag{77}$$

where, in (75), we replaced the definitions of g_t and $\bar{\boldsymbol{g}}_t$ given in (39) and in (40), respectively; in (76), we consider the cases $k=k'$ and $k\neq k'$ separately; (77) follows from the consideration that local models of different clients evolve independently and then all the terms with $k'\neq k$ equal zero because $\nabla F_k(\boldsymbol{w},\mathcal{B})$ is an unbiased estimator of $\nabla F_k(\boldsymbol{w})$. It follows that:

$$-2\eta_t^2 \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left[\langle \bar{\boldsymbol{g}}_t, \bar{\boldsymbol{g}}_t - \boldsymbol{g}_t\rangle\right] = \tag{78}$$

$$= -2\eta_t^2 \sum_{k\in\mathcal{A}_t} q_k^2 \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t}\left[\sum_{j=0}^{E-1}\sum_{j'=0}^{E-1}\langle \nabla F_k(\boldsymbol{w}_{t,j}^k), \nabla F_k(\boldsymbol{w}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k,\mathcal{B}_{t,j'}^k)\rangle\right] \tag{79}$$

$$= -2\eta_t^2 \sum_{k\in\mathcal{A}_t} q_k^2 \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t}\left[\sum_{j=0}^{E-1}\sum_{\substack{j'=0\\j'<j}}^{E-1}\langle \nabla F_k(\boldsymbol{w}_{t,j}^k), \nabla F_k(\boldsymbol{w}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k,\mathcal{B}_{t,j'}^k)\rangle\right]$$

$$- 2\eta_t^2 \sum_{k\in\mathcal{A}_t} q_k^2 \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t}\left[\sum_{j=0}^{E-1}\sum_{\substack{j'=0\\j'\geq j}}^{E-1}\langle \nabla F_k(\boldsymbol{w}_{t,j}^k), \nabla F_k(\boldsymbol{w}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k,\mathcal{B}_{t,j'}^k)\rangle\right] \tag{80}$$

$$= -2\eta_t^2 \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=0}^{E-1}\sum_{\substack{j'=0\\j'<j}}^{E-1} \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t}\left[\langle \nabla F_k(\boldsymbol{w}_{t,j}^k), \nabla F_k(\boldsymbol{w}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k,\mathcal{B}_{t,j'}^k)\rangle\right]$$

$$- 2\eta_t^2 \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=0}^{E-1}\sum_{\substack{j'=0\\j'\geq j}}^{E-1} \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:j'-1}^k|\mathcal{A}_t,\mathcal{H}_t}\left[\mathop{\mathbb{E}}_{\mathcal{B}_{t,j'}^k|\mathcal{B}_{t,0:j'-1}^k,\mathcal{A}_t,\mathcal{H}_t}\left[\langle \nabla F_k(\boldsymbol{w}_{t,j}^k), \nabla F_k(\boldsymbol{w}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k,\mathcal{B}_{t,j'}^k)\rangle\right]\right] \tag{81}$$

$$= -2\eta_t^2 \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=0}^{E-1}\sum_{\substack{j'=0\\j'<j}}^{E-1} \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t}\left[\langle \nabla F_k(\boldsymbol{w}_{t,j}^k), \nabla F_k(\boldsymbol{w}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k,\mathcal{B}_{t,j'}^k)\rangle\right]$$

$$- 2\eta_t^2 \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=0}^{E-1}\sum_{\substack{j'=0\\j'\geq j}}^{E-1} \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:j'-1}^k|\mathcal{A}_t,\mathcal{H}_t}\left[\langle \nabla F_k(\boldsymbol{w}_{t,j}^k), \underbrace{\mathop{\mathbb{E}}_{\mathcal{B}_{t,j'}^k|\mathcal{B}_{t,0:j'-1}^k,\mathcal{A}_t,\mathcal{H}_t}\left[\nabla F_k(\boldsymbol{w}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k,\mathcal{B}_{t,j'}^k)\right]}_{=0}\rangle\right], \tag{82}$$

where, in (80), we consider the cases $j'<j$ and $j'\geq j$ separately; then, in (81) and in (82), we use the law of total expectation. Finally, we bound the remaining term in the right-hand side of (82) as follows:

$$-2\eta_t^2 \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left[\langle \bar{\boldsymbol{g}}_t, \bar{\boldsymbol{g}}_t - \boldsymbol{g}_t\rangle\right] = \tag{83}$$

$$= -2\eta_t^2 \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=1}^{E-1}\sum_{j'<j} \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t}\langle \nabla F_k(\boldsymbol{w}_{t,j}^k), \nabla F_k(\boldsymbol{w}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k,\mathcal{B}_{t,j'}^k)\rangle \tag{84}$$

$$= \eta_t^2 \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=1}^{E-1}\sum_{j'<j} \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t}\left[\left\|\nabla F_k(\boldsymbol{w}_{t,j}^k)\right\|^2 + \left\|\nabla F_k(\boldsymbol{w}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k,\mathcal{B}_{t,j'}^k)\right\|^2\right] \tag{85}$$

$$= \eta_t^2 \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=1}^{E-1}\sum_{j'<j} \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t}\left[\left\|\nabla F_k(\boldsymbol{w}_{t,j}^k)\right\|\right] +$$

$$+ \eta_t^2 \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=1}^{E-1}\sum_{j'<j} \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:j'-1}^k|\mathcal{A}_t,\mathcal{H}_t} \underbrace{\left[\mathop{\mathbb{E}}_{\mathcal{B}_{t,j'}^k|\mathcal{B}_{t,0:j'-1}^k,\mathcal{A}_t,\mathcal{H}_t}\left\|\nabla F_k(\boldsymbol{w}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k,\mathcal{B}_{t,j'}^k)\right\|^2\right]}_{\text{bounded with Assumption 5}} \tag{86}$$

$$\leq \eta_t^2 \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=1}^{E-1} \sum_{j'<j} \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t} \left\|\nabla F_k(\boldsymbol{w}_{t,j}^k)\right\|^2 + \frac{1}{2}\eta_t^2 E(E-1) \sum_{k\in\mathcal{A}_t} q_k^2\sigma_k^2 \tag{87}$$

$$\leq \eta_t^2 L(E-1) \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=1}^{E-1} \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t} \left[\left(F_k(\boldsymbol{w}_{t,j}^k) - F_k^*\right)\right] + \frac{1}{2}\eta_t^2 E(E-1) \sum_{k\in\mathcal{A}_t} q_k^2\sigma_k^2 \tag{88}$$

$$\begin{aligned}= \eta_t^2 L(E-1) &\sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=1}^{E-1} \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t} \left[\left(F_k(\boldsymbol{w}_{t,j}^k) - F_k(\boldsymbol{w}_B^*)\right)\right] \\ &+ \eta_t^2 LE(E-1) \sum_{k\in\mathcal{A}_t} q_k^2 \left(F_k(\boldsymbol{w}_B^*) - F_k^*\right) + \frac{1}{2}\eta_t^2 E(E-1) \sum_{k\in\mathcal{A}_t} q_k^2\sigma_k^2 \end{aligned} \tag{89}$$

$$\begin{aligned}\leq \eta_t^2 L(E-1)Q &\sum_{k\in\mathcal{A}_t} q_k \sum_{j=1}^{E-1} \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t} \left[\left(F_k(\boldsymbol{w}_{t,j}^k) - F_k(\boldsymbol{w}_B^*)\right)\right] \\ &+ \eta_t^2 LE(E-1)Q \underbrace{\sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_B^*) - F_k^*\right)}_{\text{bounded in Lemma 10}} + \frac{1}{2}\eta_t^2 E(E-1) \sum_{k\in\mathcal{A}_t} q_k^2\sigma_k^2, \end{aligned} \tag{90}$$

where, in (85), we used $|\langle \boldsymbol{a}, \boldsymbol{b}\rangle| \leq \frac{1}{2}\|\boldsymbol{a}\|^2 + \frac{1}{2}\|\boldsymbol{b}\|^2$; in (87), we applied Assumption 5; in (88), we used the L-smoothness of $\{F_k(\boldsymbol{w})\}_{k\in\mathcal{K}}$; in (89), we added and subtracted $F_k(\boldsymbol{w}_B^*)$ from the sum; finally, in (90), we used $\sum_{k\in\mathcal{A}_t} q_k^2 f(k) \leq (\sum_{k\in\mathcal{A}_t} q_k)(\sum_{k\in\mathcal{A}_t} q_k f(k))$ and $\sum_{k\in\mathcal{A}_t} q_k \leq \sum_{k=1}^N q_k := Q$. Noting that $E-1 < 2E$ concludes the proof of Lemma 6. □

Lemma 7 (Bound on the variance of the stochastic gradients)**.** *Let Assumption 5 hold. Similarly to [23, Lemma 2], we have:*

$$\eta_t^2 \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \|\boldsymbol{g}_t - \bar{\boldsymbol{g}}_t\|^2 \leq \eta_t^2 E \sum_{k\in\mathcal{A}_t} q_k^2\sigma_k^2. \tag{91}$$

Proof of Lemma 7.

$$\mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \|\boldsymbol{g}_t - \bar{\boldsymbol{g}}_t\|^2 = \tag{92}$$

$$= \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \left\| \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left(\nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k) - \nabla F_k(\boldsymbol{w}_{t,j}^k)\right) \right\|^2 \tag{93}$$

$$\begin{aligned}= &\sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=0}^{E-1} \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t} \left\|\nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k) - \nabla F_k(\boldsymbol{w}_{t,j}^k)\right\|^2 \\ &+ \sum_{k\in\mathcal{A}_t} q_k^2 \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t} \left[\sum_{j=0}^{E-1} \sum_{\substack{j'=0 \\ j'\neq j}}^{E-1} \langle \nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k) - \nabla F_k(\boldsymbol{w}_{t,j}^k), \nabla F_k(\boldsymbol{w}_{t,j'}^k, \mathcal{B}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k)\rangle\right] \\ &+ \sum_{k\in\mathcal{A}_t} \sum_{\substack{k'\in\mathcal{A}_t \\ k'\neq k}} q_k q_{k'} \sum_{j=0}^{E-1} \Big\langle \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:j-1}^k|\mathcal{A}_t,\mathcal{H}_t} \underbrace{\left[\mathop{\mathbb{E}}_{\mathcal{B}_{t,j}^k|\mathcal{B}_{t,0:j-1}^k,\mathcal{A}_t,\mathcal{H}_t} \left[\nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k) - \nabla F_k(\boldsymbol{w}_{t,j}^k)\right]\right]}_{=0}, \\ &\qquad \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:j-1}^{k'}|\mathcal{A}_t,\mathcal{H}_t} \underbrace{\left[\mathop{\mathbb{E}}_{\mathcal{B}_{t,j}^{k'}|\mathcal{B}_{t,0:j-1}^{k'},\mathcal{A}_t,\mathcal{H}_t} \left[\nabla F_{k'}(\boldsymbol{w}_{t,j}^{k'}, \mathcal{B}_{t,j}^{k'}) - \nabla F_{k'}(\boldsymbol{w}_{t,j}^{k'})\right]\right]}_{=0} \Big\rangle \\ &+ \sum_{k\in\mathcal{A}_t} \sum_{\substack{k'\in\mathcal{A}_t \\ k'\neq k}} q_k q_{k'} \sum_{j=0}^{E-1} \sum_{\substack{j'=0 \\ j'\neq j}}^{E-1} \Big\langle \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:j-1}^k|\mathcal{A}_t,\mathcal{H}_t} \underbrace{\left[\mathop{\mathbb{E}}_{\mathcal{B}_{t,j}^k|\mathcal{B}_{t,0:j-1}^k,\mathcal{A}_t,\mathcal{H}_t} \left[\nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k) - \nabla F_k(\boldsymbol{w}_{t,j}^k)\right]\right]}_{=0}, \\ &\qquad \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:j'-1}^{k'}|\mathcal{A}_t,\mathcal{H}_t} \underbrace{\left[\mathop{\mathbb{E}}_{\mathcal{B}_{t,j'}^{k'}|\mathcal{B}_{t,0:j'-1}^{k'},\mathcal{A}_t,\mathcal{H}_t} \left[\nabla F_{k'}(\boldsymbol{w}_{t,j'}^{k'}, \mathcal{B}_{t,j'}^{k'}) - \nabla F_{k'}(\boldsymbol{w}_{t,j'}^{k'})\right]\right]}_{=0} \Big\rangle \end{aligned} \tag{94}$$

$$= \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=0}^{E-1} \underbrace{\mathop{\mathbb{E}}_{\mathcal{B}_{t,j}^k|\mathcal{A}_t,\mathcal{H}_t} \left\|\nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k) - \nabla F_k(\boldsymbol{w}_{t,j}^k)\right\|^2}_{\text{bounded with Assumption 5}}$$
$$+ \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=0}^{E-1} \sum_{\substack{j'=0\\ j'<j}}^{E-1} \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:j-1}^k|\mathcal{A}_t,\mathcal{H}_t} \left[\mathop{\mathbb{E}}_{\mathcal{B}_{t,j}^k|\mathcal{B}_{t,0:j-1}^k,\mathcal{A}_t,\mathcal{H}_t} \left[\langle \nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k) - \nabla F_k(\boldsymbol{w}_{t,j}^k), \nabla F_k(\boldsymbol{w}_{t,j'}^k, \mathcal{B}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k)\rangle\right]\right]$$
$$+ \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=0}^{E-1} \sum_{\substack{j'=0\\ j'>j}}^{E-1} \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:j'-1}^k|\mathcal{A}_t,\mathcal{H}_t} \left[\mathop{\mathbb{E}}_{\mathcal{B}_{t,j'}^k|\mathcal{B}_{t,0:j'-1}^k,\mathcal{A}_t,\mathcal{H}_t} \left[\langle \nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k) - \nabla F_k(\boldsymbol{w}_{t,j}^k), \nabla F_k(\boldsymbol{w}_{t,j'}^k, \mathcal{B}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k)\rangle\right]\right] \tag{95}$$

$$= \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=0}^{E-1} \underbrace{\mathop{\mathbb{E}}_{\mathcal{B}_{t,j}^k|\mathcal{A}_t,\mathcal{H}_t} \left\|\nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k) - \nabla F_k(\boldsymbol{w}_{t,j}^k)\right\|^2}_{\text{bounded with Assumption 5}}$$
$$+ \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=0}^{E-1} \sum_{\substack{j'=0\\ j'<j}}^{E-1} \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:j-1}^k|\mathcal{A}_t,\mathcal{H}_t} \left[\underbrace{\mathop{\mathbb{E}}_{\mathcal{B}_{t,j}^k|\mathcal{B}_{t,0:j-1}^k,\mathcal{A}_t,\mathcal{H}_t} \left[\nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k) - \nabla F_k(\boldsymbol{w}_{t,j}^k)\right]}_{=0}, \nabla F_k(\boldsymbol{w}_{t,j'}^k, \mathcal{B}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k)\rangle\right]$$
$$+ \sum_{k\in\mathcal{A}_t} q_k^2 \sum_{j=0}^{E-1} \sum_{\substack{j'=0\\ j'>j}}^{E-1} \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:j'-1}^k|\mathcal{A}_t,\mathcal{H}_t} \left[\langle \nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k) - \nabla F_k(\boldsymbol{w}_{t,j}^k), \underbrace{\mathop{\mathbb{E}}_{\mathcal{B}_{t,j'}^k|\mathcal{B}_{t,0:j'-1}^k,\mathcal{A}_t,\mathcal{H}_t} \left[\nabla F_k(\boldsymbol{w}_{t,j'}^k, \mathcal{B}_{t,j'}^k) - \nabla F_k(\boldsymbol{w}_{t,j'}^k)\right]}_{=0}\rangle\right] \tag{96}$$
$$\leq E \sum_{k\in\mathcal{A}_t} q_k^2 \sigma_k^2, \tag{97}$$

where, in (94), (95), and (96), we used the law of total expectation; in (97), we applied Assumption 5. Multiplying both sides of (97) by η_t^2 completes the proof of Lemma 7. □

Lemma 8. *Let Assumption 3 hold and let the local functions $\{F_k\}_{k=1}^N$ be convex. Define $\gamma_t := 2\eta_t(1-\eta_t L(1+2EQ))$. For a diminishing step-size $0 < \eta_t \leq \frac{1}{2L(1+2EQ)}$, satisfying $\gamma_t > 0$, we have:*

$$-\gamma_t \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left(F_k(\boldsymbol{w}_{t,j}^k) - F_k(\boldsymbol{w}_B^*)\right) \leq -\frac{1}{2}\eta_t E \sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_{t,0}) - F_k(\boldsymbol{w}_B^*)\right)$$
$$+ \underbrace{\sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left\|\boldsymbol{w}_{t,j}^k - \boldsymbol{w}_{t,0}\right\|^2}_{\text{bounded in Lemma 9}} + 2\eta_t^2 LE \underbrace{\sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_B^*) - F_k^*\right)}_{\text{bounded in Lemma 10}}, \tag{98}$$

Proof of Lemma 8. In the following, we require $\gamma_t > 0$.

$$-\gamma_t \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left(F_k(\boldsymbol{w}_{t,j}^k) - F_k(\boldsymbol{w}_B^*)\right) \tag{99}$$
$$= -\gamma_t \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left(F_k(\boldsymbol{w}_{t,j}^k) - F_k(\boldsymbol{w}_{t,0})\right) - \gamma_t \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left(F_k(\boldsymbol{w}_{t,0}) - F_k(\boldsymbol{w}_B^*)\right) \tag{100}$$
$$\leq -\gamma_t \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \langle \nabla F_k(\boldsymbol{w}_{t,0}), \boldsymbol{w}_{t,j}^k - \boldsymbol{w}_{t,0}\rangle - \gamma_t E \sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_{t,0}) - F_k(\boldsymbol{w}_B^*)\right) \tag{101}$$
$$\leq \gamma_t \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \frac{1}{2}\left[\eta_t \left\|\nabla F_k(\boldsymbol{w}_{t,0})\right\|^2 + \frac{1}{\eta_t}\left\|\boldsymbol{w}_{t,j}^k - \boldsymbol{w}_{t,0}\right\|^2\right] - \gamma_t E \sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_{t,0}) - F_k(\boldsymbol{w}_B^*)\right) \tag{102}$$

$$\leq \gamma_t \eta_t LE \sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_{t,0}) - F_k^*\right) + \frac{\gamma_t}{2\eta_t} \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left\|\boldsymbol{w}_{t,j}^k - \boldsymbol{w}_{t,0}\right\|^2 - \gamma_t E \sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_{t,0}) - F_k(\boldsymbol{w}_B^*)\right) \tag{103}$$

$$\leq -\gamma_t E(1-\eta_t L) \sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_{t,0}) - F_k(\boldsymbol{w}_B^*)\right) + \frac{\gamma_t}{2\eta_t} \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left\|\boldsymbol{w}_{t,j}^k - \boldsymbol{w}_{t,0}\right\|^2 + \gamma_t \eta_t LE \sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_B^*) - F_k^*\right) \tag{104}$$

where, in (100), we added and subtracted $F_k(\boldsymbol{w}_{t,0})$ to the sum; in (101), we used the convexity of $\{F_k(\boldsymbol{w})\}_{k\in\mathcal{K}}$; note that (101) also requires $\gamma_t > 0$; in (102), we used the inequality $|\langle \boldsymbol{a}, \boldsymbol{b}\rangle| \leq \frac{1}{2}\|\boldsymbol{a}\|^2 + \frac{1}{2}\|\boldsymbol{b}\|^2$; in (103), we applied the L-smoothness of $\{F_k(\boldsymbol{w})\}_{k\in\mathcal{K}}$ (Assumption 3); finally, in (104), we added and subtracted $F_k(\boldsymbol{w}_B^*)$ to the sum.

In particular, for $\gamma_t := 2\eta_t(1-\eta_t L(1+2EQ)) > 0$, since $0 < \eta_t \leq \frac{1}{2L(1+2EQ)}$, we further obtain:

$$\begin{aligned}
&-\gamma_t \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left(F_k(\boldsymbol{w}_{t,j}^k) - F_k(\boldsymbol{w}_B^*)\right) \\
&\leq -\frac{1}{2}\eta_t E \sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_{t,0}) - F_k(\boldsymbol{w}_B^*)\right) + \underbrace{\sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left\|\boldsymbol{w}_{t,j}^k - \boldsymbol{w}_{t,0}\right\|^2}_{\text{bounded in Lemma 9}} + 2\eta_t^2 LE \underbrace{\sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_B^*) - F_k^*\right)}_{\text{bounded in Lemma 10}},
\end{aligned} \tag{105}$$

where, in (105), we used $0 < \eta_t \leq \frac{1}{2L(1+2EQ)}$, which gives $-\gamma_t E(1-\eta_t L) = -2\eta_t E\,(1-\eta_t L(1+2EQ))\,(1-\eta_t L) \leq -\frac{1}{2}\eta_t E$. Moreover, since $\gamma_t \leq 2\eta_t$, we also used $\gamma_t\eta_t \leq 2\eta_t^2$, and $\frac{\gamma_t}{2\eta_t} \leq 1$. □

Lemma 9 (Bound on the divergence of local models)**.** *Let Assumption 2, 3, and 5 hold, the local functions $\{F_k\}_{k=1}^N$ be convex and G be defined as in Lemma 1, Equation (7). Similarly to [23, Lemma 3], we obtain the following inequality:*

$$\mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \left[\sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left\|\boldsymbol{w}_{t,j}^k - \boldsymbol{w}_{t,0}\right\|^2\right] \leq \frac{1}{2}\eta_t^2 E^3 G^2 \left(\sum_{k\in\mathcal{A}_t} q_k\right). \tag{106}$$

Proof of Lemma 9.

$$\mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \left[\sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \left\|\boldsymbol{w}_{t,j}^k - \boldsymbol{w}_{t,0}\right\|^2\right] = \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \left[\sum_{k\in\mathcal{A}_t} q_k \sum_{j=1}^{E-1} \eta_t^2 \left\|\sum_{j'=0}^{j-1} \nabla F_k(\boldsymbol{w}_{t,j'}^k, \mathcal{B}_{t,j'}^k)\right\|^2\right] \tag{107}$$

$$\leq \eta_t^2 \sum_{k\in\mathcal{A}_t} q_k \sum_{j=1}^{E-1} j \sum_{j'=0}^{j-1} \mathop{\mathbb{E}}_{\mathcal{B}_t^k|\mathcal{A}_t,\mathcal{H}_t} \left[\left\|\nabla F_k(\boldsymbol{w}_{t,j'}^k, \mathcal{B}_{t,j'}^k)\right\|^2\right] \tag{108}$$

$$\leq \eta_t^2 G^2 \left(\sum_{j=1}^{E-1} j^2\right) \left(\sum_{k\in\mathcal{A}_t} q_k\right) \tag{109}$$

$$= \frac{1}{6}\eta_t^2 E(E-1)(2E-1) G^2 \left(\sum_{k\in\mathcal{A}_t} q_k\right), \tag{110}$$

where, in (108), we used the triangle and the Jensen's inequalities; in (109), we applied the bound in Lemma 1, Equation (7); finally, in (110), we developed the sum of sequence of squares $\sum_{j=1}^{E-1} j^2 = \frac{1}{6}E(E-1)(2E-1) \leq \frac{1}{2}E^3$ since $E \geq 1$. □

Lemma 10 (Bound on the dissimilarity of local functions)**.** *Let Assumption 1 hold and $(\mathcal{A}_t)_{t\geq 0}$ defined therein. We have:*

$$\mathbb{E}\left[\sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_B^*) - F_k^*\right)\right] \leq \left(\sum_{k=1}^N \pi_k q_k\right) \Gamma, \tag{111}$$

where Γ is defined in (9).

Proof of Lemma 10.

$$\mathbb{E}\left[\sum_{k\in\mathcal{A}_t} q_k\left(F_k(\boldsymbol{w}_B^*)-F_k^*\right)\right] = \sum_{k=1}^N \pi_k q_k\left(F_k(\boldsymbol{w}_B^*)-F_k^*\right) \tag{112}$$

$$= \left(\sum_{k'=1}^N \pi_{k'}q_{k'}\right)\sum_{k=1}^N p_k\left(F_k(\boldsymbol{w}_B^*)-F_k^*\right) \tag{113}$$

$$\leq \left(\sum_{k'=1}^N \pi_{k'}q_{k'}\right)\sum_{k=1}^N p_k\left(F_k(\boldsymbol{w}^*)-F_k^*\right) \tag{114}$$

$$\leq \left(\sum_{k'=1}^N \pi_{k'}q_{k'}\right)\underbrace{\max_{k\in\mathcal{K}}\{(F_k(\boldsymbol{w}^*)-F_k^*)\}}_{:=\Gamma} = \left(\sum_{k=1}^N \pi_k q_k\right)\Gamma, \tag{115}$$

where, in (112), we solved the total expectation, observing that $\mathbb{E}\left[\sum_{k\in\mathcal{A}_t} q_k f(k)\right] = \sum_{k=1}^N \pi_k q_k f(k)$ (Assumption 1); in (113), we applied $p_k \coloneqq \frac{\pi_k q_k}{\sum_{k'=1}^N \pi_{k'}q_{k'}}$; in (114), we used $F_B(\boldsymbol{w}) \coloneqq \sum_{k=1}^N p_k F_k(\boldsymbol{w})$ and we observed $F_B(\boldsymbol{w}_B^*) \leq F_B(\boldsymbol{w}^*)$; finally, in (115), we used $\sum_{k=1}^N p_k = 1$ and $\Gamma \coloneqq \max_{k\in\mathcal{K}}\{(F_k(\boldsymbol{w}^*)-F_k^*)\}$. □

Lemma 11 (Convergence results under heterogeneous client availability)**.** *Let Assumptions 1–3 and 5 hold and the functions $\{F_k\}_{k=1}^N$ be convex. For a diminishing step-size $0<\eta_t\leq\frac{1}{2L(1+2EQ)}$ satisfying $\sum_{t=1}^{+\infty}\eta_t^2<+\infty$, for any $t_0\leq T$, we have:*

$$\begin{aligned}\sum_{t=t_0}^T \eta_t\,\mathbb{E}\left[\sum_{k\in\mathcal{A}_t} q_k\big(F_k(\boldsymbol{w}_{t,0})-F_k(\boldsymbol{w}_B^*)\big)\right] &\leq \frac{2}{E}\operatorname{diam}(W)^2 + (E+1)\left(\sum_{k=1}^N \pi_k q_k^2\sigma_k^2\right)\left(\sum_{t=1}^{+\infty}\eta_t^2\right)\\ &\quad + 2E^2G^2\left(\sum_{k=1}^N \pi_k q_k\right)\left(\sum_{t=1}^{+\infty}\eta_t^2\right)\\ &\quad + 4L(1+EQ)\Gamma\left(\sum_{k=1}^N \pi_k q_k\right)\left(\sum_{t=1}^{+\infty}\eta_t^2\right)\\ &:= C_0 < +\infty.\end{aligned} \tag{116}$$

Proof of Lemma 11. We take expectation over $\mathcal{B}_t \mid \mathcal{A}_t, \mathcal{H}_t$ on Lemma 3:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left\|\boldsymbol{w}_{t+1,0}-\boldsymbol{w}_B^*\right\|^2 \leq \left\|\boldsymbol{w}_{t,0}-\boldsymbol{w}_B^*\right\|^2 &\underbrace{-2\eta_t \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\langle \boldsymbol{w}_{t,0}-\boldsymbol{w}_B^*, \bar{\boldsymbol{g}}_t\rangle}_{\text{bounded in Lemma 4}} + \underbrace{\eta_t^2 \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left\|\bar{\boldsymbol{g}}_t\right\|^2}_{\text{bounded in Lemma 5}}\\ &+ \underbrace{2\eta_t \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\langle \boldsymbol{w}_{t,0}-\boldsymbol{w}_B^*-\eta_t\bar{\boldsymbol{g}}_t, \bar{\boldsymbol{g}}_t-\boldsymbol{g}_t\rangle}_{\text{bounded in Lemma 6}} + \underbrace{\eta_t^2 \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left\|\boldsymbol{g}_t-\bar{\boldsymbol{g}}_t\right\|^2}_{\text{bounded in Lemma 7}}.\end{aligned} \tag{117}$$

Replacing Lemmas 4–7 in (117), we obtain:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left\|\boldsymbol{w}_{t+1,0}-\boldsymbol{w}_B^*\right\|^2 \leq \left\|\boldsymbol{w}_{t,0}-\boldsymbol{w}_B^*\right\|^2 &+ 2\eta_t^2 LE(1+2EQ)\mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left[\sum_{k\in\mathcal{A}_t} q_k\left(F_k(\boldsymbol{w}_B^*)-F_k^*\right)\right]\\ &\underbrace{-\underbrace{2\eta_t(1-\eta_t L(1+2EQ))}_{\gamma_t}\mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left[\sum_{k\in\mathcal{A}_t} q_k\sum_{j=1}^{E-1}\left(F_k(\boldsymbol{w}_{t,j}^k)-F_k(\boldsymbol{w}_B^*)\right)\right]}_{\text{bounded in Lemma 8}}\\ &+ \frac{1}{2}\eta_t^2 E(E+1)\sum_{k\in\mathcal{A}_t} q_k^2\sigma_k^2 + \underbrace{\mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left[\sum_{k\in\mathcal{A}_t} q_k\sum_{j=0}^{E-1}\left\|\boldsymbol{w}_{t,j}^k-\boldsymbol{w}_{t,0}\right\|^2\right]}_{\text{bounded in Lemma 9}}\end{aligned} \tag{118}$$

We apply Lemmas 8 and 9 to (118) with $\gamma_t \coloneqq 2\eta_t(1-\eta_t L(1+2EQ))$. We observe that $\gamma_t > 0$ because:

$$0 \leq \eta_t \leq \frac{1}{2L(1+2EQ)}. \tag{119}$$

We obtain:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \|\boldsymbol{w}_{t+1,0} - \boldsymbol{w}_B^*\|^2 \leq \|\boldsymbol{w}_{t,0} - \boldsymbol{w}_B^*\|^2 &- \frac{1}{2}\eta_t E \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t}\left[\sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_{t,0}) - F_k(\boldsymbol{w}_B^*)\right)\right] \\ &+ \frac{1}{2}\eta_t^2 E(E+1)\sum_{k\in\mathcal{A}_t} q_k^2\sigma_k^2 + \eta_t^2 E^3 G^2 \sum_{k\in\mathcal{A}_t} q_k \\ &+ 4\eta_t^2 LE(1+EQ)\left[\sum_{k\in\mathcal{A}_t} q_k\left(F_k(\boldsymbol{w}_B^*) - F_k^*\right)\right].\end{aligned} \tag{120}$$

Computing the total expectation on (120), we have:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{A}_t,\mathcal{B}_t,\mathcal{H}_t} \|\boldsymbol{w}_{t+1,0} - \boldsymbol{w}_B^*\|^2 \leq \mathop{\mathbb{E}}_{\mathcal{H}_t} \|\boldsymbol{w}_{t,0} - \boldsymbol{w}_B^*\|^2 &- \frac{1}{2}\eta_t E \mathop{\mathbb{E}}_{\mathcal{A}_t,\mathcal{B}_t,\mathcal{H}_t}\left[\sum_{k\in\mathcal{A}_t} q_k \left(F_k(\boldsymbol{w}_{t,0}) - F_k(\boldsymbol{w}_B^*)\right)\right] \\ &+ \frac{1}{2}\eta_t^2 E(E+1)\mathop{\mathbb{E}}_{\mathcal{A}_t,\mathcal{H}_t}\left[\sum_{k\in\mathcal{A}_t} q_k^2\sigma_k^2\right] + \eta_t^2 E^3 G^2 \mathop{\mathbb{E}}_{\mathcal{A}_t,\mathcal{H}_t}\left[\sum_{k\in\mathcal{A}_t} q_k\right] \\ &+ 4\eta_t^2 LE(1+EQ)\underbrace{\mathop{\mathbb{E}}_{\mathcal{A}_t,\mathcal{H}_t}\left[\sum_{k\in\mathcal{A}_t} q_k\left(F_k(\boldsymbol{w}_B^*) - F_k^*\right)\right]}_{\text{bounded in Lemma 10}}\end{aligned} \tag{121}$$

Applying Lemma 10 to (121) and considering $\mathbb{E}\left[\sum_{k\in\mathcal{A}_t} a_k\right] = \sum_{k=1}^N \pi_k a_k$ (Assumption 1), the following inequality holds:

$$\begin{aligned}\mathbb{E}\|\boldsymbol{w}_{t+1,0} - \boldsymbol{w}_B^*\|^2 \leq \mathbb{E}\|\boldsymbol{w}_{t,0} - \boldsymbol{w}_B^*\|^2 &- \frac{1}{2}\eta_t E\,\mathbb{E}\left[\sum_{k\in\mathcal{A}_t} q_k\left(F_k(\boldsymbol{w}_{t,0}) - F_k(\boldsymbol{w}_B^*)\right)\right] \\ &+ \frac{1}{2}\eta_t^2 E(E+1)\left(\sum_{k=1}^N \pi_k q_k^2\sigma_k^2\right) + \eta_t^2 E^3 G^2\left(\sum_{k=1}^N \pi_k q_k\right) + 4\eta_t^2 LE(1+EQ)\Gamma\left(\sum_{k=1}^N \pi_k q_k\right).\end{aligned} \tag{122}$$

Rearranging and summing over $t = t_0, \ldots, T$, we obtain the following inequality:

$$\begin{aligned}\sum_{t=t_0}^T \eta_t\,\mathbb{E}\left[\sum_{k\in\mathcal{A}_t} q_k\big(F_k(\boldsymbol{w}_{t,0}) - F_k(\boldsymbol{w}_B^*)\big)\right] \leq \frac{2}{E}\sum_{t=t_0}^T \mathbb{E}\left[\left(\|\boldsymbol{w}_{t,0} - \boldsymbol{w}_B^*\|^2 - \|\boldsymbol{w}_{t+1,0} - \boldsymbol{w}_B^*\|^2\right)\right] &\\ + (E+1)\left(\sum_{k=1}^N \pi_k q_k^2\sigma_k^2\right)\left(\sum_{t=t_0}^T \eta_t^2\right) &\\ + 2E^2G^2\left(\sum_{k=1}^N \pi_k q_k\right)\left(\sum_{t=t_0}^T \eta_t^2\right) &\\ + 4L(1+EQ)\Gamma\left(\sum_{k=1}^N \pi_k q_k\right)\left(\sum_{t=t_0}^T \eta_t^2\right).&\end{aligned} \tag{123}$$

The first term in the right-hand side of (123) is a telescoping sum and we remove the negative term $-\mathbb{E}\|\boldsymbol{w}_{T+1,0} - \boldsymbol{w}_B^*\|^2$:

$$\begin{aligned}\sum_{t=t_0}^T \eta_t\,\mathbb{E}\left[\sum_{k\in\mathcal{A}_t} q_k\big(F_k(\boldsymbol{w}_{t,0}) - F_k(\boldsymbol{w}_B^*)\big)\right] \leq \frac{2}{E}\mathbb{E}\|\boldsymbol{w}_{t_0,0} - \boldsymbol{w}_B^*\|^2 + (E+1)\left(\sum_{k=1}^N \pi_k q_k^2\sigma_k^2\right)\left(\sum_{t=t_0}^T \eta_t^2\right) &\\ + 2E^2G^2\left(\sum_{k=1}^N \pi_k q_k\right)\left(\sum_{t=t_0}^T \eta_t^2\right)&\end{aligned}$$

$$+ 4L(1+EQ)\Gamma \left(\sum_{k=1}^{N} \pi_k q_k\right)\left(\sum_{t=t_0}^{T} \eta_t^2\right). \tag{124}$$

Finally, by noting that $\|\boldsymbol{w}_{t_0,0} - \boldsymbol{w}_B^*\| \leq \operatorname{diam}(W)$ and $\sum_{t=t_0}^{T} \eta_t^2 \leq \sum_{t=1}^{+\infty} \eta_t^2 < +\infty$, we complete the proof of Lemma 11. □

Lemma 12. *Let Assumptions 2 and 3 hold, and the local functions $\{F_k\}_{k=1}^N$ be convex. We have:*

$$|F_k(\boldsymbol{v}) - F_k(\boldsymbol{w})| \leq D \cdot \|\boldsymbol{v} - \boldsymbol{w}\|, \ \forall \boldsymbol{v}, \boldsymbol{w} \in W \tag{125}$$

Proof of Lemma 12. In Lemma 1, under Assumptions 2 and 3, we have already proved that:

$$\|\nabla F_k(\boldsymbol{w})\| \leq D. \tag{6}$$

Moreover, from the convexity of $\{F_k\}_{k\in\mathcal{K}}$, it follows that:

$$\langle \nabla F_k(\boldsymbol{v}), \boldsymbol{v} - \boldsymbol{w}\rangle \leq F_k(\boldsymbol{v}) - F_k(\boldsymbol{w}) \leq \langle \nabla F_k(\boldsymbol{w}), \boldsymbol{v} - \boldsymbol{w}\rangle. \tag{126}$$

The Cauchy–Schwarz inequality completes the proof of Lemma 12:

$$|F_k(\boldsymbol{v}) - F_k(\boldsymbol{w})| \leq \max\{\|\nabla F_k(\boldsymbol{v})\|, \|\nabla F_k(\boldsymbol{w})\|\} \cdot \|\boldsymbol{v} - \boldsymbol{w}\| \leq D \cdot \|\boldsymbol{v} - \boldsymbol{w}\|. \tag{127}$$

□

Lemma 13. *Let Assumptions 2, 3, and 5 hold. We have:*

$$\mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \|\boldsymbol{w}_{t+1,0} - \boldsymbol{w}_{t,0}\| \leq \eta_t EG \left(\sum_{k\in\mathcal{A}_t} q_k\right). \tag{128}$$

Proof of Lemma 13. The proof is based on [15, Proposition 1.4].

$$\mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \|\boldsymbol{w}_{t+1,0} - \boldsymbol{w}_{t,0}\| = \mathop{\mathbb{E}}_{\mathcal{B}_t|\mathcal{A}_t,\mathcal{H}_t} \left\| -\eta_t \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k) \right\| \tag{129}$$

$$\leq \eta_t \sum_{k\in\mathcal{A}_t} q_k \sum_{j=0}^{E-1} \mathop{\mathbb{E}}_{\mathcal{B}_{t,0:j-1}^k|\mathcal{A}_t,\mathcal{H}_t} \left[\mathop{\mathbb{E}}_{\mathcal{B}_{t,j}^k|\mathcal{B}_{t,0:j-1}^k,\mathcal{A}_t,\mathcal{H}_t} \left[\nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k)\right]\right] \tag{130}$$

$$\leq \eta_t EG \left(\sum_{k\in\mathcal{A}_t} q_k\right), \tag{131}$$

where, in (130), we used the triangle inequality and the law of total expectation; in (131), we applied Lemma 1, Equation (7). □

Similarly to [15, Theorem 1], we provide the following definition.

Definition 1. For communication round $t \geq 1$, denote the positive integer $\mathcal{J}_t$ as follows:

$$\mathcal{J}_t := \min\left\{\max\left\{\left\lceil \frac{\ln(2C_P H t)}{\ln(1/\lambda(\boldsymbol{P}))}\right\rceil, T_P\right\}, t\right\}. \tag{132}$$

The parameter $\mathcal{J}_t$ is crucial in our analysis: it represents the communication rounds needed to bound the stationary distribution convergence of the Markov process $(\mathcal{A}_t)_{t>0}$. It will play a key role in Lemmas 14–18 and in the proof of Theorem 2. We remark that, by definition: $T_P \leq \mathcal{J}_t \leq t$.

Our definition of $\mathcal{J}_t$ corrects a typo in [15, (6.27)], which considered $\ln(t/(2C_P H))$ rather than $\ln(2C_P H t)$. In fact, we observe that [15, (6.28)] and consequently [15, (6.35)] do not hold when $\mathcal{J}_t$ is defined as in [15, (6.27)].

Lemma 14 (Convergence results under heterogeneous and correlated client availability after $\mathcal{J}_t$ communication rounds)**.** *Let Assumptions 1–3, and 5 hold, the local functions $\{F_k\}_{k=1}^N$ be convex, and the parameter $\mathcal{J}_t \leq t$ be as in Definition 1. For a diminishing step-size $\{\eta_t\}_{t\geq 1}$ satisfying $\sum_{t=1}^{+\infty} \ln(t)\cdot\eta_t^2$, for any $t_0 \leq T$, we have:*

$$\sum_{t=t_0}^{T} \eta_t \mathbb{E}\left[\sum_{k\in\mathcal{A}_t} q_k \big(F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_{t,0})\big)\right] \leq \frac{C_1}{\ln(1/\lambda(\boldsymbol{P}))} < +\infty, \tag{133}$$

where:

$$C_1 \coloneqq EDGQ \left(\sum_{k=1}^{N} \pi_k q_k \right) \left(\sum_{t=1}^{+\infty} \ln \left(2 C_P H t \right) \eta_{t-\mathcal{J}_t}^2 \right). \tag{134}$$

Proof of Lemma 14. This proof is based on [15, Equation (6.31)].

$$\sum_{t=t_0}^{T} \eta_t \, \mathbb{E} \left[\sum_{k \in \mathcal{A}_t} q_k (F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_{t,0})) \right] \leq Q \sum_{t=t_0}^{T} \eta_t \, \mathbb{E} \left[\max_{k \in \mathcal{K}} \{ F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_{t,0}) \} \right] \tag{135}$$

$$\leq DQ \sum_{t=t_0}^{T} \eta_t \, \mathbb{E} \left\| \boldsymbol{w}_{t-\mathcal{J}_t,0} - \boldsymbol{w}_{t,0} \right\| \tag{136}$$

$$\leq DQ \sum_{t=t_0}^{T} \eta_t \sum_{d=t-\mathcal{J}_t}^{t-1} \mathop{\mathbb{E}}_{\mathcal{A}_d, \mathcal{H}_d} \left[\mathop{\mathbb{E}}_{\mathcal{B}_d | \mathcal{A}_d, \mathcal{H}_d} \left\| \boldsymbol{w}_{d,0} - \boldsymbol{w}_{d+1,0} \right\| \right] \tag{137}$$

$$\leq EDGQ \sum_{t=t_0}^{T} \sum_{d=t-\mathcal{J}_t}^{t-1} \eta_t \eta_d \, \mathbb{E} \left[\sum_{k \in \mathcal{A}_d} q_k \right] \tag{138}$$

$$\leq EDGQ \left(\sum_{k=1}^{N} \pi_k q_k \right) \sum_{t=t_0}^{T} \sum_{d=t-\mathcal{J}_t}^{t-1} \eta_t \eta_d \tag{139}$$

$$\leq \frac{EDGQ}{2} \left(\sum_{k=1}^{N} \pi_k q_k \right) \sum_{t=t_0}^{T} \sum_{d=t-\mathcal{J}_t}^{t-1} \left(\eta_t^2 + \eta_d^2 \right) \tag{140}$$

$$\leq EDGQ \left(\sum_{k=1}^{N} \pi_k q_k \right) \sum_{t=t_0}^{T} \mathcal{J}_t \eta_{t-\mathcal{J}_t}^2, \tag{141}$$

where, in (135), we used $\sum_{k \in \mathcal{A}_t} q_k a_k \leq \sum_{k=1}^{N} q_k a_k \leq (\sum_{k=1}^{N} q_k) \cdot \max_{k \in \mathcal{K}} \{a_k\} = Q \cdot \max_{k \in \mathcal{K}} \{a_k\}$; in (136), we applied Lemma 12; in (137), we used the triangle inequality and the law of total expectation; in (138), we applied Lemma 13 and again the law of total expectation; in (139), we observed that $\mathbb{E} \left[\sum_{k \in \mathcal{A}_d} q_k \right] = \sum_{k=1}^{N} \pi_k q_k$ (Assumption 1); in (140), we used $2ab \leq a^2 + b^2$; finally, in (141), we applied $\eta_t < \eta_d \leq \eta_{t-\mathcal{J}_t}$ due to the diminishing learning rate.

We apply then the definition of $\mathcal{J}_t$ in (132) and we observe that $\sum_{t=t_0}^{T} \ln(t) \eta_{t-\mathcal{J}_t}^2 \leq \sum_{t=1}^{+\infty} \ln(t) \eta_{t-\mathcal{J}_t}^2$:

$$\sum_{t=t_0}^{T} \eta_t \, \mathbb{E} \left[\sum_{k \in \mathcal{A}_t} q_k (F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_{t,0})) \right] \leq EDGQ \left(\sum_{k=1}^{N} \pi_k q_k \right) \left(\sum_{t=t_0}^{T} \frac{\ln \left(2 C_P H t \right)}{\ln(1/\lambda(\boldsymbol{P}))} \eta_{t-\mathcal{J}_t}^2 \right) \tag{142}$$

$$\leq EDGQ \left(\sum_{k=1}^{N} \pi_k q_k \right) \left(\sum_{t=1}^{+\infty} \frac{\ln \left(2 C_P H t \right)}{\ln(1/\lambda(\boldsymbol{P}))} \eta_{t-\mathcal{J}_t}^2 \right) = \frac{C_1}{\ln(1/\lambda(\boldsymbol{P}))}. \tag{143}$$

Finally, we conclude that C_1 is finite. To this purpose, we observe that $\mathcal{J}_t \leq a \ln(t) + b$, for opportune positive values a and b. Let t' be a positive integer such that $t \geq a \ln(t) + b$ for any $t \geq t'$. Then:

$$\sum_{t=t'}^{T} \ln(t) \cdot \eta_{t-\mathcal{J}_t}^2 = \sum_{t=t'-\mathcal{J}_t}^{T-\mathcal{J}_t} \ln(t + \mathcal{J}_t) \cdot \eta_t^2 \tag{144}$$

$$\leq \sum_{t=1}^{+\infty} \ln(t + a \ln t + b) \cdot \eta_t^2 \tag{145}$$

$$\leq \sum_{t=1}^{+\infty} \ln \left((1 + a + b) \, t \right) \cdot \eta_t^2 < +\infty. \tag{146}$$

□

Lemma 15. *Let Assumptions 2, 3 and 5 hold, the local functions $\{F_k\}_{k=1}^N$ be convex, and $\mathcal{J}_t \leq t$ be as in Definition 1. Let the step-size be decreasing and satisfy: $\sum_{t=1}^{+\infty} \ln(t) \cdot \eta_t^2 < +\infty$. For any $t_0 \leq T$, we have:*

$$\left(\sum_{k=1}^{N} \pi_k q_k\right) \sum_{t=t_0}^{T} \eta_t \, \mathbb{E}\left[F_B(\boldsymbol{w}_{t,0}) - F_B(\boldsymbol{w}_{t-\mathcal{J}_t,0})\right] \leq \frac{C_1}{\ln\left(1/\lambda(\boldsymbol{P})\right)} < +\infty, \tag{147}$$

where:

$$C_1 \coloneqq EDGQ \left(\sum_{k=1}^{N} \pi_k q_k\right) \left(\sum_{t=1}^{+\infty} \ln\left(2C_P H t\right) \eta_{t-\mathcal{J}_t}^2\right). \tag{148}$$

Proof of Lemma 15. This proof is based on [15, Equation (6.38)].

$$\left(\sum_{k=1}^{N} \pi_k q_k\right) \sum_{t=t_0}^{T} \eta_t \, \mathbb{E}\left[F_B(\boldsymbol{w}_{t,0}) - F_B(\boldsymbol{w}_{t-\mathcal{J}_t,0})\right] = \sum_{t=t_0}^{T} \eta_t \sum_{k=1}^{N} \pi_k q_k \, \mathbb{E}\left[F_k(\boldsymbol{w}_{t,0}) - F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0})\right] \tag{149}$$

$$\leq D \left(\sum_{k=1}^{N} \pi_k q_k\right) \sum_{t=t_0}^{T} \eta_t \, \mathbb{E}\left\|\boldsymbol{w}_{t-\mathcal{J}_t,0} - \boldsymbol{w}_{t,0}\right\| \tag{150}$$

$$\leq D \left(\sum_{k=1}^{N} \pi_k q_k\right) \sum_{t=t_0}^{T} \eta_t \sum_{d=t-\mathcal{J}_t}^{t-1} \mathbb{E}_{\mathcal{A}_d, \mathcal{H}_d} \left[\mathbb{E}_{\mathcal{B}_d | \mathcal{A}_d, \mathcal{H}_d} \left\|\boldsymbol{w}_{d,0} - \boldsymbol{w}_{d+1,0}\right\|\right] \tag{151}$$

$$\leq DEGQ \left(\sum_{k=1}^{N} \pi_k q_k\right) \sum_{t=t_0}^{T} \sum_{d=t-\mathcal{J}_t}^{t-1} \eta_t \eta_d \tag{152}$$

$$\leq \frac{DEGQ}{2} \left(\sum_{k=1}^{N} \pi_k q_k\right) \sum_{t=t_0}^{T} \sum_{d=t-\mathcal{J}_t}^{t-1} \left(\eta_t^2 + \eta_d^2\right) \tag{153}$$

$$\leq DEGQ \left(\sum_{k=1}^{N} \pi_k q_k\right) \sum_{t=t_0}^{T} \mathcal{J}_t \cdot \eta_{t-\mathcal{J}_t}^2 \tag{154}$$

$$\leq EDGQ \left(\sum_{k=1}^{N} \pi_k q_k\right) \left(\sum_{t=1}^{+\infty} \frac{\ln\left(2C_P H t\right)}{\ln(1/\lambda(\boldsymbol{P}))} \eta_{t-\mathcal{J}_t}^2\right) = \frac{C_1}{\ln(1/\lambda(\boldsymbol{P}))}, \tag{155}$$

where, in (149), we applied $F_B(\boldsymbol{w}) = \sum_{k=1}^N p_k F_k(\boldsymbol{w})$, where $p_k = \frac{\pi_k q_k}{\sum_{h=1}^N \pi_h q_h}$; in (150), we applied Lemma 12; in (151), we applied the triangle inequality and the law of total expectation; in (152), we applied Lemma 13; in (153), we used $2ab \leq a^2 + b^2$; in (154), we observed that $\eta_t^2 + \eta_d^2 \leq 2\eta_{t-\mathcal{J}_t}^2$ due to the diminishing learning rate; finally, in (155), we applied the definition of $\mathcal{J}_t$ given in (132) and we observed that $\sum_{t=t_0}^{T} \ln(t) \eta_{t-\mathcal{J}_t}^2 \leq \sum_{t=1}^{+\infty} \ln(t) \eta_{t-\mathcal{J}_t}^2 < +\infty$ and then $C_1 < +\infty$. □

Lemma 16 (Bound on the distance dynamics between the current and the stationary distributions of the Markov process)**.** *Let Assumption 1 hold, and $\boldsymbol{P}$, $\boldsymbol{\rho}$ defined therein. The following inequality holds:*

$$\max_{i,j \in [M]} \left|[\boldsymbol{P}^t]_{i,j} - \rho_j\right| \leq C_P \cdot \lambda(\boldsymbol{P})^t, \quad \textit{for } t \geq T_P, \tag{5}$$

where C_P and T_P are positive constants defined as:

$$C_P \coloneqq \left(\sum_{i=2}^{d} n_i^2\right)^{\frac{1}{2}} \cdot \|\boldsymbol{U}\|_F \|\boldsymbol{U}^{-1}\|_F, \tag{156}$$

$$T_P \coloneqq \max\left\{\max_{1 \leq i \leq d} \left\{\left\lceil \frac{2n_i(n_i-1)(\ln(\frac{2n_i}{\ln \lambda(\boldsymbol{P})/|\bar{\lambda}_2(\boldsymbol{P})|}) - 1)}{(n_i+1)\ln(\lambda(\boldsymbol{P})/|\bar{\lambda}_2(\boldsymbol{P})|)} \right\rceil\right\}, \; 0\right\}. \tag{157}$$

Here, d, n_i, and $\boldsymbol{U}$ are quantities related to the Jordan canonical form of $\boldsymbol{P}$. Specifically, $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{J}\boldsymbol{U}^{-1}$, where $\boldsymbol{J}$ denotes the Jordan $M \times M$ matrix with d blocks $\boldsymbol{J}_i, i = 2, \ldots, d$. Each block $\boldsymbol{J}_i$, $i = 2, 3, \ldots, d$, has a dimension $n_i \geq 1$, and $\sum_{i=1}^d n_i = M$. Moreover, $|\boldsymbol{U}|_F$ denotes the Frobenius norm of the matrix $\boldsymbol{U}$.

Furthermore, let Assumptions 2 and 3 hold, H be defined as in Lemma 1, Equation (8), and $T_P \leq \mathcal{J}_t \leq t$ be defined in (132). We obtain the additional inequality:

$$\left|[\boldsymbol{P}^{\mathcal{J}_t}]_{i,j} - \rho_j\right| \leq C_P \cdot \lambda(\boldsymbol{P})^t \leq C_P \lambda(\boldsymbol{P})^{\mathcal{J}_t} = \frac{1}{2Ht}, \quad \forall i,j \in [M] \text{ and } \forall t \geq T_P. \tag{158}$$

Proof of Lemma 16. The inequality in (5) is proven in [15, Lemma 1] and holds for any $t \geq T_P$. Here, T_P is a constant dependent on the transition matrix $\boldsymbol{P}$ of the Markov chain $(\mathcal{A}_t)_{t\geq 0}$ defined in Assumption 1. To prove (158), we further observe that $0 < \lambda(\boldsymbol{P}) \leq 1$ and $T_P \leq \mathcal{J}_t \leq t$. The last inequality in (158) follows from the definition of $\mathcal{J}_t$ in (132). □

We remark that the bounds in [15, Lemma 1], and consequently our (158), require $t \geq T_P$. Therefore, the derivations in [15, (6.28)] and [15, (6.35)–(6.37)] are not accurate, since they hold for $t \geq T_P$. We address this problem with Lemmas 17 and 18.

Lemma 17. *Let Assumptions 1–3 hold, and T_P be defined as in (157). The following inequality holds:*

$$\left(\sum_{k=1}^{N} \pi_k q_k\right) \sum_{t=1}^{T_P-1} \eta_t \, \mathbb{E}\left[F_B(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_B^*\right] \leq C_2 < +\infty, \tag{159}$$

where:

$$C_2 \coloneqq H \left(\sum_{t=1}^{T_P-1} \eta_t\right) \left(\sum_{k=1}^{N} \pi_k q_k\right) < +\infty. \tag{160}$$

Proof of Lemma 17.

$$\left(\sum_{k=1}^{N} \pi_k q_k\right) \sum_{t=1}^{T_P-1} \eta_t \, \mathbb{E}\left[F_B(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_B^*\right] = \sum_{t=1}^{T_P-1} \eta_t \sum_{k=1}^{N} \pi_k q_k \, \mathbb{E}\left[F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*)\right] \tag{161}$$

$$\leq H \left(\sum_{t=1}^{T_P-1} \eta_t\right) \left(\sum_{k=1}^{N} \pi_k q_k\right) \coloneqq C_2 < +\infty, \tag{162}$$

where, in (161), we used the definition of F_B from (4), and in (162), we applied Lemma 1, Equation (8), which holds for any $\boldsymbol{w} \in W$. Lastly, it is worth noting that C_2 is a sum of finite elements, and is therefore finite. □

Lemma 18. *Let Assumptions 1–3 and 5 hold, and $\{F_k\}_{k=1}^N$ be convex. Recall the definitions of $\mathcal{J}_t$ and T_P in (132) and in (157), respectively. Let the step-size $(\eta_t)_{t\geq 1}$ decrease and satisfy $\eta_1 \leq \frac{1}{2L(1+2EQ)}$, $\sum_{t=1}^{+\infty} \eta_t^2 < +\infty$, and $\sum_{t=1}^{+\infty} \ln(t) \cdot \eta_t^2 < +\infty$. For $t \geq T_P$, we have:*

$$\left(\sum_{k=1}^{N} \pi_k q_k\right) \sum_{t=T_P}^{T} \eta_t \, \mathbb{E}\left[F_B(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_B^*\right] \leq \frac{C_1}{\ln\left(1/\lambda(\boldsymbol{P})\right)} + C_3 < +\infty, \tag{163}$$

where:

$$C_1 \coloneqq EDGQ \left(\sum_{k=1}^{N} \pi_k q_k\right) \left(\sum_{t=1}^{+\infty} \ln(2C_P H t) \cdot \eta_{t-\mathcal{J}_t}^2\right) < +\infty. \tag{164}$$

$$C_3 \coloneqq C_0 + \frac{MQ}{4} \sum_{t=1}^{+\infty} \left(\eta_t^2 + \frac{1}{t^2}\right) < +\infty; \tag{165}$$

Proof of Lemma 18. Assume $t \geq T_P$. With a similar proof technique to [15, (6.35)], we derive the following lower bound:

$$\mathop{\mathbb{E}}_{\mathcal{A}_t | \mathcal{A}_{t-\mathcal{J}_t}, \mathcal{H}_{t-\mathcal{J}_t}} \left[\sum_{k \in \mathcal{A}_t} q_k\big(F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*)\big)\right] =$$

$$= \sum_{a \in \mathcal{M}} \mathbb{P}(\mathcal{A}_t = a \mid \mathcal{A}_{t-\mathcal{J}_t}, \mathcal{H}_{t-\mathcal{J}_t}) \sum_{k \in a} q_k\big(F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*)\big) \tag{166}$$

$$= \sum_{a \in \mathcal{M}} \left[\boldsymbol{P}^{\mathcal{J}_t}\right]_{\mathcal{A}_{t-\mathcal{J}_t}, a} \sum_{k \in a} q_k\big(F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*)\big) \tag{167}$$

$$\geq \sum_{a \in \mathcal{M}} \left(\rho_a - \frac{1}{2Ht}\right) \sum_{k \in a} q_k\big(F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*)\big) \tag{168}$$

$$= \sum_{k=1}^{N} \mathbb{E}\left[\mathbb{1}_{k\in\mathcal{A}_t}\right] q_k\big(F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*)\big) - \frac{1}{2Ht}\sum_{a\in\mathcal{M}}\sum_{k\in a} q_k\big(F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*)\big) \tag{169}$$

$$\geq \sum_{k=1}^{N} \pi_k q_k\big(F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*)\big) - \frac{MQ}{2Ht}\max_{k\in\mathcal{K}}\left\{F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*)\right\} \tag{170}$$

$$\geq \left(\sum_{k=1}^{N}\pi_k q_k\right)\cdot\big(F_B(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_B^*\big) - \frac{MQ}{2t}, \tag{171}$$

where, in (166), we applied the definition of expected value to the random variable $\mathcal{A}_t$, with a representing a realization of $\mathcal{A}_t$, that is a state in the state space $\mathcal{M}$, and $\mathbb{P}(\mathcal{A}_t = a \mid \mathcal{A}_{t-\mathcal{J}_t}, \mathcal{H}_{t-\mathcal{J}_t})$ denoting the conditional probability of the event $\mathcal{A}_t = a$ given $(\mathcal{A}_{t-\mathcal{J}_t}, \mathcal{H}_{t-\mathcal{J}_t})$; in (167), we applied the Markov property (Assumption 1), observing that $\mathbb{P}(\mathcal{A}_t = a \mid \mathcal{A}_{t-\mathcal{J}_t}) = [\boldsymbol{P}^{\mathcal{J}_t}]_{\mathcal{A}_{t-\mathcal{J}_t},a}$, where $[\boldsymbol{P}^k]_{i,j}$ denotes the (i,j)-th element of the k-th power of the transition matrix $\boldsymbol{P}$; in (168), we applied Lemma 16, Equation (158); for the first term in (169), we used $\sum_{a\in\mathcal{M}}\rho_a\sum_{k\in a} f(k) = \sum_{a\in\mathcal{M}}\rho_a\sum_{k=1}^{N}\mathbb{1}_{\{k\in a\}}f(k) = \sum_{k=1}^{N} f(k)\sum_{a\in\mathcal{M}}\rho_a\mathbb{1}_{k\in a} = \sum_{k=1}^{N} f(k)\,\mathbb{E}\left[\mathbb{1}_{k\in\mathcal{A}_t}\right]$, where $\mathbb{1}_{k\in\mathcal{A}_t}$ is the indicator function that equals 1 if and only if $k \in \mathcal{A}_t$; in (170), we used $\mathbb{E}\left[\mathbb{1}_{k\in\mathcal{A}_t}\right] = \mathbb{P}(k \in \mathcal{A}_t) \coloneqq \pi_k$ for the first term, and $\sum_{k\in a} q_k f(k) \leq \sum_{k=1}^{N} q_k f(k) \leq (\sum_{k=1}^{N} q_k)(\max_{k\in\mathcal{K}} f(k)) = Q\max_{k\in\mathcal{K}} f(k)$ and $\sum_{a\in\mathcal{M}} 1 = M$ for the second term; finally, in (171), we used the definition of F_B in (4) for the first term, and we used Lemma 1, Equation (8) for the second term.

Our derivations in (170) and (171) correct a typo in [15, (6.35)], which considered $Q/(2t)$ instead of $(MQ)/(2t)$. In (171), the dimension (M) of the state space $(\mathcal{M})$ of the Markov chain $(\mathcal{A}_t)_{t\geq 0}$ appears in the numerator of the second term.

Note that the steps in (168)–(171) require $t \geq T_P$. Multiplying by η_t and summing for $t = T_P, \ldots, T$, rearranging, and computing the total expectation, we obtain the following inequality:

$$\left(\sum_{k=1}^{N}\pi_k q_k\right)\sum_{t=T_P}^{T}\eta_t\,\mathbb{E}\left[F_B(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_B^*\right] \leq \sum_{t=T_P}^{T}\eta_t\,\mathbb{E}\left[\sum_{k\in\mathcal{A}_t} q_k\big(F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*)\big)\right] + \frac{MQ}{2}\sum_{t=T_P}^{T}\frac{\eta_t}{t} \tag{172}$$

$$\leq \underbrace{\sum_{t=T_P}^{T}\eta_t\,\mathbb{E}\left[\sum_{k\in\mathcal{A}_t} q_k\big(F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*)\big)\right]}_{\text{bounded with Lemma 11 + Lemma 14}} + \frac{MQ}{4}\sum_{t=1}^{T}\left(\eta_t^2 + \frac{1}{t^2}\right), \tag{173}$$

where, in (173), we used $2ab \leq a^2 + b^2$ and we observed that $\sum_{t=T_P}^{T}\left(\eta_t^2 + \frac{1}{t^2}\right) \leq \sum_{t=1}^{T}\left(\eta_t^2 + \frac{1}{t^2}\right)$ since $t > 0$ and $\eta_t > 0$.

Moreover, if the step-size $(\eta_t)_{t\geq 1}$ decreases and satisfies $\eta_1 \leq \frac{1}{2L(1+2EQ)}$, $\sum_{t=1}^{+\infty}\eta_t^2 < +\infty$, and $\sum_{t=1}^{+\infty}\ln(t)\cdot\eta_t^2 < +\infty$, we can further bound the first term in (173) by combining Lemma 11 and Lemma 14 for $t_0 = T_P$, and we obtain:

$$\sum_{t=T_P}^{T}\eta_t\,\mathbb{E}\left[\sum_{k\in\mathcal{A}_t} q_k\big(F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*)\big)\right] \leq C_0 + \frac{C_1}{\ln\left(1/\lambda(\boldsymbol{P})\right)} < +\infty, \tag{174}$$

where:

$$\begin{aligned} C_0 \coloneqq\ & \frac{2}{E}\operatorname{diam}(W)^2 + (E+1)\left(\sum_{k=1}^{N}\pi_k q_k^2\sigma_k^2\right)\left(\sum_{t=1}^{+\infty}\eta_t^2\right) \\ & + 2E^2G^2\left(\sum_{k=1}^{N}\pi_k q_k\right)\left(\sum_{t=1}^{+\infty}\eta_t^2\right) \\ & + 4L(1+EQ)\Gamma\left(\sum_{k=1}^{N}\pi_k q_k\right)\left(\sum_{t=1}^{+\infty}\eta_t^2\right). \end{aligned} \tag{175}$$

Finally, plugging (174) into (173), observing that $\sum_{t=1}^{T}\left(\eta_t^2 + \frac{1}{t^2}\right) \leq \sum_{t=1}^{+\infty}\left(\eta_t^2 + \frac{1}{t^2}\right) < +\infty$ because $\sum_{t=1}^{+\infty}\eta_t^2 < +\infty$ and $\sum_{t=1}^{+\infty}\frac{1}{t^2} = \frac{\pi}{6} < +\infty$, and denoting $C_3 \coloneqq C_0 + \frac{MQ}{4}\sum_{t=1}^{+\infty}\left(\eta_t^2 + \frac{1}{t^2}\right) < +\infty$, we conclude the proof of Lemma 18. □

B3. Proof of Theorem 2

Theorem 2 (Convergence of the optimization error ϵ_{opt})**.** *Let Assumptions 1–3 and 5 hold and the functions $\{F_k\}_{k=1}^N$ be convex. Recall the constants M, L, D, G, H, Γ, $\sigma_k, C_P, T_P, \mathcal{J}_t$, and $\lambda(\boldsymbol{P})$ defined above. Let $Q = \sum_{k\in\mathcal{K}} q_k$.*
Let the step-size $\eta_t > 0$ decrease and satisfy:

$$\eta_1 \leq \frac{1}{2L(1+2EQ)}, \qquad \sum_{t=1}^{+\infty} \eta_t = +\infty, \qquad \sum_{t=1}^{+\infty} \ln(t)\cdot \eta_t^2 < +\infty. \tag{12}$$

Let T denote the total communication rounds.
For $T \geq T_P$, the expected optimization error $\mathbb{E}[F_B(\bar{\boldsymbol{w}}_{T,0}) - F_B^*]$ can be bounded as follows:

$$\mathbb{E}[F_B(\bar{\boldsymbol{w}}_{T,0}) - F_B^*] \leq \frac{\frac{\frac{1}{2}\boldsymbol{q}^\intercal \boldsymbol{\Sigma} \boldsymbol{q} + \upsilon}{\boldsymbol{\pi}^\intercal \boldsymbol{q}} + \psi + \frac{\phi}{\ln(1/\lambda(\boldsymbol{P}))}}{(\sum_{t=1}^T \eta_t)}, \tag{13}$$

where $\bar{\boldsymbol{w}}_{T,0} = \frac{\sum_{t=1}^T \eta_t \boldsymbol{w}_{t,0}}{\sum_{t=1}^T \eta_t}$, and:

$$\boldsymbol{\Sigma} \coloneqq \operatorname{diag}\left(2\left(E+1\right)\pi_k \sigma_k^2 \sum_{t=1}^{+\infty} \eta_t^2\right); \tag{176}$$

$$\upsilon \coloneqq \frac{2}{E}\operatorname{diam}(W)^2 + \frac{MQ}{4}\sum_{t=1}^{+\infty}\left(\eta_t^2 + \frac{1}{t^2}\right); \tag{177}$$

$$\psi \coloneqq 4L(1+EQ)\Gamma\left(\sum_{t=1}^{+\infty}\eta_t^2\right) + 2E^2G^2\left(\sum_{t=1}^{+\infty}\eta_t^2\right) + H\left(\sum_{t=1}^{T_P-1}\eta_t\right); \tag{178}$$

$$\phi \coloneqq 2EDGQ\left(\sum_{t=1}^{+\infty}\ln(2C_PHt)\cdot\eta_{t-\mathcal{J}_t}^2\right). \tag{179}$$

Proof of Theorem 2. The proof involves three main steps.

Step 1: From Lemma 15, observe that:

$$\left(\sum_{k=1}^N \pi_k q_k\right)\sum_{t=1}^T \eta_t\, \mathbb{E}[F_B(\boldsymbol{w}_{t,0}) - F_B(\boldsymbol{w}_{t-\mathcal{J}_t,0})] \leq \frac{C_1}{\ln(1/\lambda(\boldsymbol{P}))} < +\infty, \tag{180}$$

where:

$$C_1 \coloneqq EDGQ\left(\sum_{k=1}^N \pi_k q_k\right)\left(\sum_{t=1}^{+\infty}\ln\left(2C_PHt\right)\cdot\eta_{t-\mathcal{J}_t}^2\right) < +\infty. \tag{181}$$

Step 2: By combining Lemma 17 and Lemma 18, we obtain:

$$\left(\sum_{k=1}^N \pi_k q_k\right)\sum_{t=1}^T \eta_t\, \mathbb{E}[F_B(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_B^*)] \leq \frac{C_1}{\ln(1/\lambda(\boldsymbol{P}))} + C_2 + C_3 < +\infty, \tag{182}$$

where C_1 is defined in (181), and:

$$C_2 \coloneqq H\left(\sum_{t=1}^{T_P-1}\eta_t\right)\left(\sum_{k=1}^N \pi_k q_k\right) < +\infty; \tag{183}$$

$$\begin{aligned} C_3 \coloneqq{}& \frac{2}{E}\operatorname{diam}(W)^2 + (E+1)\left(\sum_{k=1}^N \pi_k q_k^2 \sigma_k^2\right)\left(\sum_{t=1}^{+\infty}\eta_t^2\right) \\ &+ 2E^2G^2\left(\sum_{k=1}^N \pi_k q_k\right)\left(\sum_{t=1}^{+\infty}\eta_t^2\right) \\ &+ 4L(1+EQ)\Gamma\left(\sum_{k=1}^N \pi_k q_k\right)\left(\sum_{t=1}^{+\infty}\eta_t^2\right) + \frac{MQ}{4}\sum_{t=1}^{+\infty}\left(\eta_t^2 + \frac{1}{t^2}\right) < +\infty. \end{aligned} \tag{184}$$

Step 3: By summing the results from Steps 1 and 2, given in (180) and (182), respectively, we have:

$$\left(\sum_{k=1}^{N}\pi_k q_k\right)\sum_{t=1}^{T}\eta_t\,\mathbb{E}[F_B(\boldsymbol{w}_{t,0})-F_B^*]\leq\frac{2C_1}{\ln(1/\lambda(\boldsymbol{P}))}+C_2+C_3<+\infty. \tag{185}$$

With the convexity of $F_B(\cdot)$, applying the Jensen's inequality, we complete Step 3:

$$\left(\sum_{t=1}^{T}\eta_t\right)\left(\sum_{k=1}^{N}\pi_k q_k\right)\mathbb{E}[F_B(\bar{\boldsymbol{w}}_{T,0})-F_B^*]\leq\left(\sum_{k=1}^{N}\pi_k q_k\right)\sum_{t=1}^{T}\eta_t\,\mathbb{E}[F_B(\boldsymbol{w}_{t,0})-F_B^*] \tag{186}$$

$$\leq\frac{2C_1}{\ln(1/\lambda(\boldsymbol{P}))}+C_2+C_3<+\infty, \tag{187}$$

where $\bar{\boldsymbol{w}}_{T,0}:=\frac{\sum_{t=1}^{T}\eta_t\boldsymbol{w}_{t,0}}{\sum_{t=1}^{T}\eta_t}$, and the constants C_1, C_2, and C_3 are defined in (181), (183), and (184), respectively.

By dividing (186) and (187) by $\left(\sum_{t=1}^{T}\eta_t\right)\cdot\left(\sum_{k=1}^{N}\pi_k q_k\right)$, we obtain the expression for Theorem 2 given in (13). □

Appendix C
Proof of Theorem 3

Theorem 3 (An alternative bound on the bias error ϵ_{bias})**.** *Under the same assumptions of Theorem 1, define* $\Gamma'\;:=\;\max_k\{F_k(\boldsymbol{w}_B^*)-F_k^*\}$. *The following result holds:*

$$\epsilon_{bias}\leq 4\kappa^2\cdot\underbrace{d_{TV}^2(\boldsymbol{\alpha},\boldsymbol{p})\cdot\Gamma'}_{:=\bar{\epsilon}'_{bias}}, \tag{15}$$

where $d_{TV}(\boldsymbol{\alpha},\boldsymbol{p}):=\frac{1}{2}\sum_{k=1}^{N}|\alpha_k-p_k|$ *denotes the total variation distance between the probability distributions* $\boldsymbol{\alpha}$ *and* $\boldsymbol{p}$.

Proof of Theorem 3. The proof follows the same steps as in Theorem 1, proceeding from (29) as follows:

$$\|\nabla F(\boldsymbol{w}_B^*)\|\leq L\sqrt{\frac{2}{\mu}}\sum_{k=1}^{N}|\alpha_k-p_k|\sqrt{(F_k(\boldsymbol{w}_B^*)-F_k^*)} \tag{29}$$

$$\leq 2L\sqrt{\frac{2}{\mu}}d_{TV}(\boldsymbol{\alpha},\boldsymbol{p})\sqrt{\Gamma'}, \tag{188}$$

where, in (188), we applied the definitions of $d_{TV}(\boldsymbol{\alpha},\boldsymbol{p}):=\frac{1}{2}\sum_{k=1}^{N}|\alpha_k-p_k|$ and $\Gamma':=\max_k\{F_k(\boldsymbol{w}_B^*)-F_k^*\}$.

Squaring (188), we obtain the following expression:

$$\|\nabla F(\boldsymbol{w}_B^*)\|^2\leq\frac{8L^2}{\mu}d_{TV}^2(\boldsymbol{\alpha},\boldsymbol{p})\Gamma'. \tag{189}$$

Then, replacing (189) in (25), we obtain:

$$\epsilon_{\text{bias}}:=(F(\boldsymbol{w}_B^*)-F^*)\leq\frac{1}{2\mu}\|\nabla F(\boldsymbol{w}_B^*)\|^2\leq 4\frac{L^2}{\mu^2}\underbrace{d_{TV}^2(\boldsymbol{\alpha},\boldsymbol{p})\Gamma'}_{:=\bar{\epsilon}'_{\text{bias}}}, \tag{190}$$

which concludes the proof of Theorem 3. □

Appendix D
Convexity of $\bar{\epsilon}_{\text{OPT}}+\bar{\epsilon}_{\text{BIAS}}$

For the proof of the convexity of $\bar{\epsilon}_{\text{opt}}(\boldsymbol{q})$, please refer to Appendix E1. To prove that $\bar{\epsilon}_{\text{bias}}(\boldsymbol{q})$ is also convex, we need to study the convexity of $\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}}:=\sum_{k=1}^{N}(\alpha_k-p_k)^2/p_k$ in $\boldsymbol{q}\in\{q_k>0\;\forall k,\|\boldsymbol{q}\|_1=Q>0\}$. To this purpose, we define the following functions:

$$h_k:\mathbb{R}_{\geq 0}^N\setminus\{\mathbf{0}\}\to\mathbb{R}_{\geq 0},\qquad h_k(\boldsymbol{q}):=\frac{\pi_k q_k}{\sum_{k'=1}^{N}\pi_{k'}q_{k'}}; \tag{191}$$

$$g_k:\mathbb{R}_{>0}\to\mathbb{R}_{\geq 0},\qquad g_k(p_k):=\frac{(p_k-\alpha_k)^2}{p_k}. \tag{192}$$

Finally, we write the chi-square divergence $\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}}$ between the target and biased probability distributions $\boldsymbol{\alpha}$ and $\boldsymbol{p}$ as:

$$\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}}(\boldsymbol{q}) = \sum_{k=1}^{N} (g_k \circ h_k)(\boldsymbol{q}) = \sum_{k=1}^{N} g_k(h_k(\boldsymbol{q})). \tag{193}$$

We observe that:

- $h_k(\boldsymbol{q})$ is a particular case of linear-fractional functions [40, Example 3.32, p. 97];
- $g_k(\cdot)$ is a convex in p_k over $\mathbb{R}_{>0}$ because sum of convex functions;
- each $g_k \circ h_k$ is quasi-convex in $\boldsymbol{q} \in \mathbb{R}^N_{>0}$ because composition of a convex function (g_k) and a linear-fractional function (h_k) [40, p. 102].

However, note that the sum of quasi-convex functions is not necessarily quasi-convex.

Proposition 1. *The function $\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}}(\boldsymbol{q})$ is not convex over $\mathbb{R}^N_{>0}$.*

Proof of Proposition 1. To analyze the convexity of $\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}}(\boldsymbol{q}) = \sum_{k=1}^{N} (g_k \circ h_k)(\boldsymbol{q})$ over $\mathbb{R}^N_{>0}$, a possible approach is to check whether each function $(g_k \circ h_k)(\boldsymbol{q})$ is convex over $\mathbb{R}^N_{>0}$. In what follows, we show that $(g_k \circ h_k)$ is not convex over $\mathbb{R}^N_{>0}$.

Consider the case when $\pi_k = 1 \ \forall k \in \mathcal{K}$. We can rewrite $(g_k \circ h_k)(\boldsymbol{q})$ as follows:

$$(g_k \circ h_k)(\boldsymbol{q}) = \frac{\left(\frac{q_k}{\|\boldsymbol{q}\|_1} - \alpha_k\right)^2}{\frac{q_k}{\|\boldsymbol{q}\|_1}}. \tag{194}$$

We show that this function fails to satisfy the definition of convexity, i.e., $\exists\ \boldsymbol{q}, \boldsymbol{q'} \in \mathbb{R}^N_{>0}, \zeta \in [0,1]$ such that:

$$(g_k \circ h_k)\left(\zeta \boldsymbol{q} + (1-\zeta)\boldsymbol{q'}\right) > \zeta\,(g_k \circ h_k)\,(\boldsymbol{q}) + (1-\zeta)\,(g_k \circ h_k)\,(\boldsymbol{q'}). \tag{195}$$

The left-hand side (LHS) of (195) is:

$$(g_k \circ h_k)\left(\zeta \boldsymbol{q} + (1-\zeta)\boldsymbol{q'}\right) = \frac{\left(\frac{\zeta q_k + (1-\zeta) q'_k}{\zeta \|\boldsymbol{q}\|_1 + (1-\zeta)\|\boldsymbol{q'}\|_1} - \alpha_k\right)^2}{\frac{\zeta q_k + (1-\zeta) q'_k}{\zeta \|\boldsymbol{q}\|_1 + (1-\zeta)\|\boldsymbol{q'}\|_1}}. \tag{196}$$

If we take $\boldsymbol{q} : \|\boldsymbol{q}\|_1 = 1$, $q_k = \alpha_k$, $\zeta = \frac{1}{2}$, $\boldsymbol{q'} = \frac{Q}{N}\mathbf{1}$, and we let $Q \to +\infty$, then the LHS in (196) converges to:

$$\lim_{Q \to +\infty} \frac{\left(\frac{\frac{1}{2}\alpha_k + \frac{1}{2}\frac{Q}{N}}{\frac{1}{2}1 + \frac{1}{2}Q} - \alpha_k\right)^2}{\frac{\frac{1}{2}\alpha_k + \frac{1}{2}\frac{Q}{N}}{\frac{1}{2}1 + \frac{1}{2}Q}} = \frac{\left(\frac{1}{N} - \alpha_k\right)^2}{\frac{1}{N}}. \tag{197}$$

On the other hand, for the same choices of q_k, $\boldsymbol{q}$, $\boldsymbol{q'}$, and ζ, and if we let $Q \to +\infty$, the right-hand side (RHS) of (195) is:

$$\zeta\,(g_k \circ h_k)\,(\boldsymbol{q}) + (1-\zeta)\,(g_k \circ h_k)\,(\boldsymbol{q'}) = 0 + \frac{1}{2}\frac{\left(\frac{1}{N} - \alpha_k\right)^2}{\frac{1}{N}}. \tag{198}$$

Finally, comparing (197) and (198), we conclude that, for Q large enough, the LHS in (195) is larger than the RHS. □

Proposition 2. *The function $\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}}(\boldsymbol{q})$ is convex over $\mathbb{R}^N_{>0} \cap \{\boldsymbol{q} : \|\boldsymbol{q}\|_1 = Q > 0\}$.*

Proof of Proposition 2. To verify the convexity of $\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}}(\boldsymbol{q}) = \sum_{k=1}^{N} (g_k \circ h_k)(\boldsymbol{q})$ over $\mathbb{R}^N_{>0} \cap \{\boldsymbol{q} : \|\boldsymbol{q}\|_1 = Q > 0\}$, one possible approach is to demonstrate the convexity of each function $(g_k \circ h_k)(\boldsymbol{q})$ over the set $\mathbb{R}^N_{>0} \cap \{\boldsymbol{q} : \|\boldsymbol{q}\|_1 = Q > 0\}$.

We prove this result for a more general case. We show that, if

$$\tilde{g} \text{ is a convex function over its domain } \mathcal{D}_g \tag{199}$$

and

$$\tilde{h}(\boldsymbol{q}) = \frac{\boldsymbol{A}\boldsymbol{q} + b}{\boldsymbol{c}^\intercal \boldsymbol{q} + d}, \tag{200}$$

then

$$\tilde{g} \circ \tilde{h} \text{ is convex over } \mathcal{D} = \mathbb{R}^N_{>0} \cap \{\boldsymbol{q} : \boldsymbol{c}^\intercal \boldsymbol{q} + d = Q > 0, \frac{\boldsymbol{A}\boldsymbol{q} + b}{\boldsymbol{c}^\intercal \boldsymbol{q} + d} \in \mathcal{D}_g\}. \tag{201}$$

It is then sufficient to apply this result to each pair (g_k, h_k) to conclude that $(g_k \circ h_k)$ is convex and then $\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}}(\boldsymbol{q})$ is convex. By direct inspection, for all $\boldsymbol{q}, \boldsymbol{q'} \in \mathcal{D},\ \forall\ \zeta \in [0,1]$, the following equality holds:

$$\left(\tilde{g} \circ \tilde{h}\right)(\zeta \boldsymbol{q} + (1-\zeta)\boldsymbol{q'}) = \tilde{g}\left(\tilde{h}\left(\zeta \boldsymbol{q} + (1-\zeta)\boldsymbol{q'}\right)\right) = \tilde{g}\left(\zeta' \frac{\boldsymbol{A}\boldsymbol{q} + b}{\boldsymbol{c}^\intercal \boldsymbol{q} + d} + (1-\zeta')\frac{\boldsymbol{A}\boldsymbol{q'} + b}{\boldsymbol{c}^\intercal \boldsymbol{q'} + d}\right), \tag{202}$$

where:

$$\zeta' = \frac{\zeta\left(\boldsymbol{c}^\intercal \boldsymbol{q} + d\right)}{\zeta\left(\boldsymbol{c}^\intercal \boldsymbol{q} + d\right) + (1-\zeta)\left(\boldsymbol{c}^\intercal \boldsymbol{q'} + d\right)} \in [0,1]. \tag{203}$$

Applying the convexity of $\tilde{g}$, we bound Equation (202) as follows:

$$\tilde{g}\left(\zeta' \frac{\boldsymbol{A}\boldsymbol{q} + b}{\boldsymbol{c}^\intercal \boldsymbol{q} + d} + (1-\zeta')\frac{\boldsymbol{A}\boldsymbol{q'} + b}{\boldsymbol{c}^\intercal \boldsymbol{q'} + d}\right) \overset{\text{convexity of } \tilde{g}}{\leq} \zeta' \tilde{g}\left(\frac{\boldsymbol{A}\boldsymbol{q} + b}{\boldsymbol{c}^\intercal \boldsymbol{q} + d}\right) + (1-\zeta')\tilde{g}\left(\frac{\boldsymbol{A}\boldsymbol{q'} + b}{\boldsymbol{c}^\intercal \boldsymbol{q'} + d}\right) \tag{204}$$

$$= \zeta'\left(\tilde{g} \circ \tilde{h}\right)(\boldsymbol{q}) + (1-\zeta')\left(\tilde{g} \circ \tilde{h}\right)(\boldsymbol{q'}). \tag{205}$$

Finally, to conclude the proof, we show that $\zeta' = \zeta$. This is true because, for any $\boldsymbol{q}$ and $\boldsymbol{q'} \in \mathcal{D}$, $\boldsymbol{c}^\intercal \boldsymbol{q} + d = \boldsymbol{c}^\intercal \boldsymbol{q'} + d = Q > 0$. In fact, by using this condition in Equation (203), we have that:

$$\zeta' = \frac{\zeta Q}{\zeta Q + (1-\zeta)Q} = \zeta, \tag{206}$$

which establishes the convexity of $\tilde{g} \circ \tilde{h}$ by definition. □

APPENDIX E
MINIMIZING $\bar{\epsilon}_{\text{OPT}}$

Equation (13) can be rewritten as:

$$\left(\sum_{t=1}^{T} \eta_t\right) \mathbb{E}\left[F_B(\bar{\boldsymbol{w}}_{T,0}) - F_B^*\right] \leq \frac{\frac{1}{2}\boldsymbol{q}^\intercal \boldsymbol{\Sigma} \boldsymbol{q} + \upsilon}{\boldsymbol{\pi}^\intercal \boldsymbol{q}} + \psi + \frac{\phi}{\ln(1/\lambda(\boldsymbol{P}))} \tag{207}$$

$$= \frac{\frac{1}{2}\boldsymbol{q}^\intercal \boldsymbol{A} \boldsymbol{q} + B}{\boldsymbol{\pi}^\intercal \boldsymbol{q}} + C \coloneqq J(\boldsymbol{q}), \tag{208}$$

where:

$$\boldsymbol{A} \coloneqq \boldsymbol{\Sigma} = \operatorname{diag}\left(2\left(E+1\right)\pi_k \sigma_k^2 \sum_{t=1}^{+\infty} \eta_t^2\right); \tag{209}$$

$$B \coloneqq \upsilon = \frac{2}{E}\operatorname{diam}(W)^2 + \frac{MQ}{4}\sum_{t=1}^{+\infty}\left(\eta_t^2 + \frac{1}{t^2}\right); \tag{210}$$

$$C \coloneqq \psi + \frac{\phi}{\ln(1/\lambda(\boldsymbol{P}))} = \left(4L(1+EQ)\Gamma + 2E^2G^2\right)\left(\sum_{t=1}^{+\infty}\eta_t^2\right) + 2EDGQ\left(\sum_{t=1}^{+\infty}\mathcal{J}_t \cdot \eta_{t-\mathcal{J}_t}^2\right) + H\left(\sum_{t=1}^{T_P-1}\eta_t\right). \tag{211}$$

The minimization of (208), defines the following optimization problem:

$$\underset{\boldsymbol{q}}{\text{minimize}} \quad J(\boldsymbol{q}) \coloneqq \frac{\frac{1}{2}\boldsymbol{q}^\intercal \boldsymbol{A} \boldsymbol{q} + B}{\boldsymbol{\pi}^\intercal \boldsymbol{q}} + C; \tag{212a}$$

$$\text{subject to} \quad \boldsymbol{q} \geq 0, \tag{212b}$$

$$\boldsymbol{\pi}^\intercal \boldsymbol{q} > 0, \tag{212c}$$
$$\|\boldsymbol{q}\|_1 = Q. \tag{212d}$$

Remark. *In Problem (212a)–(212d), when setting some q_k to zero, we do not consider the possibility of redefining the Markov chain $(\mathcal{A}_t)_{t\geq 0}$ in Assumption 1 by considering the reduced state space of clients with $q_k > 0$. In this case, the redefined Markov chain would have a different transition matrix $\boldsymbol{P'} \neq \boldsymbol{P}$ with $\lambda(\boldsymbol{P'}) \neq \lambda(\boldsymbol{P})$, resulting in C no longer being constant.*

E1. The optimization problem in (212a)–(212d) is convex

Let us rewrite the problem by adding a variable $s \coloneqq 1/\boldsymbol{\pi}^\intercal \boldsymbol{q}$ and then replacing $\boldsymbol{y} \coloneqq s\boldsymbol{q}$. We have:

$$J(\boldsymbol{y}, s) = s\left(\frac{1}{2}\frac{\boldsymbol{y}^\intercal}{s}\boldsymbol{A}\frac{\boldsymbol{y}}{s} + B\right) + C = s \cdot K\left(\frac{\boldsymbol{y}}{s}\right) + C, \tag{213}$$

where $K : \mathbb{R}^N \to \mathbb{R}$, $K(\boldsymbol{q}) \coloneqq \frac{1}{2}\boldsymbol{q}^\intercal \boldsymbol{A}\boldsymbol{q} + B$ is a (strictly) convex function, and:

$$\underset{\boldsymbol{y},\, s}{\text{minimize}} \quad J(\boldsymbol{y}, s) = \frac{1}{2s}\boldsymbol{y}^\intercal \boldsymbol{A}\boldsymbol{y} + Bs + C \tag{214a}$$
$$\text{subject to} \quad \boldsymbol{y} \geq 0, \tag{214b}$$
$$s > 0, \tag{214c}$$
$$\boldsymbol{\pi}^\intercal \boldsymbol{y} = 1, \tag{214d}$$
$$\|\boldsymbol{y}\|_1 = Qs. \tag{214e}$$

Note that the objective function $J(\boldsymbol{y}, s) : \mathbb{R}^{N+1} \to \mathbb{R}$, $J(\boldsymbol{y}, s) = s \cdot K(\boldsymbol{y}/s) + C$ in (213) is the perspective of the convex function $K(\boldsymbol{q}) + C$, and is therefore convex [40, pp. 89–90]. Moreover, the constraints in (214b)–(214e) define a convex set, and then the optimization problem defined by (214a)–(214e) is convex. We solve it with the method of Lagrange multipliers.

E2. Support for Guideline A (Section III)

The Lagrangian function $\mathcal{L}$ is as follows:

$$\mathcal{L}(\boldsymbol{y}, s, \iota, \theta, \boldsymbol{\omega}) = \frac{1}{2s}\boldsymbol{y}^\intercal \boldsymbol{A}\boldsymbol{y} + Bs + C + \iota(1 - \boldsymbol{\pi}^\intercal \boldsymbol{y}) + \theta(\|\boldsymbol{y}\|_1 - Qs) - \boldsymbol{\omega}^\intercal \boldsymbol{y}. \tag{215}$$

Since the constraint $s > 0$ defines an open set, the set defined by the constraints in (214b)–(214e) is not closed. However, the solution of the optimization problem defined by (214a)–(214e) is never on the boundary $s = 0$ because $\mathcal{L} \to +\infty$ as $s \to 0^+$, therefore we can consider $s \geq 0$. Moreover, strong duality holds for the Slater's constraint qualification for convex problems.

The KKT conditions read:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial s}(\boldsymbol{y}^*, s^*, \iota^*, \theta^*, \boldsymbol{\omega}^*) = 0, & (216) \\ \nabla_{\boldsymbol{y}} \mathcal{L}(\boldsymbol{y}^*, s^*, \iota^*, \theta^*, \boldsymbol{\omega}^*) = 0, & (217) \\ \boldsymbol{\pi}^\intercal \boldsymbol{y}^* - 1 = 0, & (218) \\ \|\boldsymbol{y}^*\|_1 - Qs = 0, & (219) \\ \boldsymbol{\omega}^{*\intercal} \boldsymbol{y}^* = 0, & (220) \\ \boldsymbol{y}^*, \boldsymbol{\omega}^* \geq 0. & (221) \end{cases}$$

In particular, the KKT condition for $\boldsymbol{y}^*$ read:

$$\nabla_{\boldsymbol{y}} \mathcal{L}(\boldsymbol{y}^*, s^*, \iota^*, \theta^*, \boldsymbol{\omega}^*) = \frac{1}{s^*}\boldsymbol{A}\boldsymbol{y}^* - \iota^* \boldsymbol{\pi} + \theta^* \mathbf{1} - \boldsymbol{\omega}^* = 0, \tag{222}$$

which is satisfied when:

$$\frac{\partial \mathcal{L}}{\partial y_k^*} = \frac{1}{s^*} A_{kk} y_k^* - \iota^* \pi_k + \theta^* - \omega_k^* = 0, \ \forall k \in \mathcal{K}, \tag{223}$$

where A_{ij} denotes the element on the i-th row and the j-th column of matrix $\boldsymbol{A}$.

Furthermore, the Complementary Slackness conditions in (220) and (221) present two cases:

1) If $y_k^* > 0$ (and $q_k^* > 0$), then $\omega_k^* = 0$ and:

$$y_k^* = \frac{s^*}{A_{kk}}(\iota^* \pi_k - \theta^*), \quad q_k^* = \frac{1}{A_{kk}}(\iota^* \pi_k - \theta^*); \tag{224}$$

2) $y_k^* = q_k^* = 0$ otherwise.

By replacing the equality constraint (214d) in Problem (214a)–(214e) with the inequality constraint $\boldsymbol{\pi}^\intercal \boldsymbol{y} \geq 1$, we establish an equivalent optimization problem. The equivalence holds because, for any feasible solution $\boldsymbol{y}'$ with $\boldsymbol{\pi}^\intercal \boldsymbol{y}' > 1$, we can consider the solution $\boldsymbol{y}'' = \frac{\boldsymbol{y}'}{\boldsymbol{\pi}^\intercal \boldsymbol{y}'} < \boldsymbol{y}'$, leading to a lower objective function value. Additionally, the new problem states that the Lagrange multiplier (ι^*) associated with the inequality constraint must be non-negative. By considering $A_{kk} \geq 0$ and $\iota^* \geq 0$ in Equation (224), we conclude that q_k^* increases with π_k, providing analytical support for Guideline A.

E3. Closed-form solution of the optimization problem in (212a)–(212d)

The solution of the optimization problem in (212a)–(212d) is not of practical utility because its constants (e.g., L, ω, Γ, C_P) are in general problem-dependent and difficult to estimate during training. In particular, Γ poses particular difficulties as it is defined in terms of the minimizer of the target objective F, but the FL algorithm generally minimizes the biased function F_B. Nevertheless, we include the closed-formed solution of the optimization problem in (212a)–(212d) for completeness.

We use the active-set method: let $\mathcal{X}$ be the set of coordinates corresponding to the active inequalities, i.e., $\mathcal{X} = \{k \mid y_k^* = 0\}$.

From the KKT condition in (218), we derive a relation between ι^* and θ^*:

$$\boldsymbol{\pi}^\intercal \boldsymbol{y}^* = \sum_{k \notin \mathcal{X}} \pi_k y_k^* = \sum_{k \notin \mathcal{X}} \pi_k \frac{s^*}{\boldsymbol{A}_{kk}} (\iota^* \pi_k - \theta^*) = \iota^* s^* \sum_{k \notin \mathcal{X}} \frac{\pi_k^2}{\boldsymbol{A}_{kk}} - \theta^* s^* \sum_{k \notin \mathcal{X}} \frac{\pi_k}{\boldsymbol{A}_{kk}} = 1. \tag{225}$$

We use the KKT condition in (219) to derive another relation between ι^* and θ^*:

$$\|\boldsymbol{y}^*\|_1 = \sum_{k \notin \mathcal{X}} y_k^* = \sum_{k \notin \mathcal{X}} \frac{s^*}{\boldsymbol{A}_{kk}} (\iota^* \pi_k - \theta^*) = Qs \quad \Leftrightarrow \quad \iota^* = \frac{Q + \theta^* \sum_{k \notin \mathcal{X}} \frac{1}{\boldsymbol{A}_{kk}}}{\sum_{k \notin \mathcal{X}} \frac{\pi_k}{\boldsymbol{A}_{kk}}}, \tag{226}$$

and, replacing (226) in (225), we derive the closed-form solution for θ^*:

$$\theta^* = \frac{\sum_{k \notin \mathcal{X}} \frac{\pi_k}{\boldsymbol{A}_{kk}} - Qs^* \sum_{k \notin \mathcal{X}} \frac{\pi_k^2}{\boldsymbol{A}_{kk}}}{s^* \left[\left(\sum_{k \notin \mathcal{X}} \frac{1}{\boldsymbol{A}_{kk}} \right) \cdot \left(\sum_{k \notin \mathcal{X}} \frac{\pi_k^2}{\boldsymbol{A}_{kk}} \right) - \left(\sum_{k \notin \mathcal{X}} \frac{\pi_k}{\boldsymbol{A}_{kk}} \right)^2 \right]}. \tag{227}$$

APPENDIX F
BACKGROUND ON MARKOV CHAINS

F1. Markov Chain for the Analysis (Section III)

We recall some existing results [15], [31] for the Markov chain $(\mathcal{A}_t)_{t \geq 0}$ used in our analysis (Assumption 1).

Assumption 1. *The Markov chain $(\mathcal{A}_t)_{t \geq 0}$ on the M-finite state space $\mathcal{M}$ is time-homogeneous, irreducible, and aperiodic. It has transition matrix $\boldsymbol{P}$, stationary distribution $\boldsymbol{\rho}$, and has state distribution $\boldsymbol{\rho}$ at time $t = 0$.*

Let $\boldsymbol{\rho}^{(t)} = [\rho_1^{(t)}, \rho_2^{(t)}, \ldots, \rho_M^{(t)}], \sum_{i=1}^M \rho_i^{(t)} = 1$ be the state probability distribution on the Markov chain $(\mathcal{A}_t)_{t \geq 0}$ at time step t. Assumption 1 guarantees the existence of a stationary distribution $\boldsymbol{\rho} = \lim_{t \to +\infty} \boldsymbol{\rho}^{(t)} = [\rho_1, \rho_2, \ldots, \rho_M]$ with $\min_i \{\rho_i\} > 0$ and $\boldsymbol{\rho}^\intercal \boldsymbol{P} = \boldsymbol{\rho}^\intercal$. Then $\boldsymbol{\rho}$ is a left eigenvector relative to the eigenvalue 1, which is the largest eigenvalue of the matrix $\boldsymbol{P}$.

For the transition matrix $\boldsymbol{P}$, we label its eigenvalues in decreasing order:

$$1 = \lambda_1(\boldsymbol{P}) > \lambda_2(\boldsymbol{P}) \geq \cdots \geq \lambda_M(\boldsymbol{P}). \tag{228}$$

We define:

$$\bar{\lambda}_2(\boldsymbol{P}) := \max \{|\lambda_2(\boldsymbol{P})|, |\lambda_M(\boldsymbol{P})|\} \quad \text{and} \quad \lambda(\boldsymbol{P}) := \frac{\bar{\lambda}_2(\boldsymbol{P}) + 1}{2}. \tag{229}$$

The second largest absolute eigenvalue $\bar{\lambda}_2(\boldsymbol{P})$ of the transition matrix $\boldsymbol{P}$ characterizes the mixing time of a Markov chain. The absolute spectral gap $\gamma := 1 - \bar{\lambda}_2(\boldsymbol{P})$ and its reciprocal, the relaxation time $t_{\text{rel}} := \frac{1}{\gamma}$, play a role in this relationship. To quantify the convergence of the Markov chain towards stationarity, we use the parameter $d(t) := \max_{a \in \mathcal{M}} \|[\boldsymbol{P}^t]_{a,\cdot} - \boldsymbol{\rho}\|_{TV}$, which measures the maximum distance between the distribution $[\boldsymbol{P}^t]_{a,\cdot}$ and the stationary distribution $\boldsymbol{\rho}$ for all initial states $a \in \mathcal{M}$. The mixing time $t_{\text{mix}}(\varepsilon)$ is defined as the minimum time at which the distance $d(t)$ becomes less than or equal to a given threshold ε: $t_{\text{mix}}(\varepsilon) := \min \{t : d(t) \leq \varepsilon\}$. Upper and lower bounds exist for the mixing time based on the relaxation time and the stationary distribution: $(t_{\text{rel}} - 1) \log \left(\frac{1}{2\varepsilon}\right) \leq t_{\text{mix}}(\varepsilon) \leq \log \left(\frac{1}{\varepsilon \rho_{\min}}\right) t_{\text{rel}}$, where $\rho_{\min} := \min_{a \in \mathcal{M}} \rho_a$ [31, pp. 154–156].

F2. Markov Chain for Guideline B (Section IV)

In Section III-D (Guideline B), we examine a specific scenario where the availability of each client k follows an independent Markov chain $(\mathcal{A}_t^k)_{t\geq 0}$ with transition probability matrix $\boldsymbol{P}_k$. This setup allows us to model the aggregate process as a product of independent Markov chains, known as a Product Chain [31, Section 12.4].

Definition 2 (Product Chain)**.** Let $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ be transition matrices on state spaces $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively, with corresponding stationary distributions $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$. We consider a Markov Chain on the state space $\mathcal{M}_1 \times \mathcal{M}_2$ that moves independently in the first and second coordinates according to $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ respectively. The transition matrix of this Markov Chain is the Kronecker product $\widetilde{\boldsymbol{P}} = \boldsymbol{P}_1 \otimes \boldsymbol{P}_2$, defined as:

$$\widetilde{\boldsymbol{P}}((x,y),(z,w)) = \boldsymbol{P}_1(x,z)\boldsymbol{P}_2(y,w). \tag{230}$$

Proposition 3. *The stationary distribution of the Markov chain defined by $\widetilde{\boldsymbol{P}} = \boldsymbol{P}_1 \otimes \boldsymbol{P}_2$ is the Kronecker product $\widetilde{\boldsymbol{\rho}} = \boldsymbol{\pi}_1 \otimes \boldsymbol{\pi}_2$.*

Proof. We can observe the following:

$$\widetilde{\boldsymbol{\rho}}^\intercal \widetilde{\boldsymbol{P}} = (\boldsymbol{\pi}_1 \otimes \boldsymbol{\pi}_2)^\intercal \cdot (\boldsymbol{P}_1 \otimes \boldsymbol{P}_2) = (\boldsymbol{\pi}_1^\intercal \boldsymbol{P}_1) \otimes (\boldsymbol{\pi}_2^\intercal \boldsymbol{P}_2) = \boldsymbol{\pi}_1^\intercal \otimes \boldsymbol{\pi}_2^\intercal = \widetilde{\boldsymbol{\rho}}^\intercal, \tag{231}$$

where, in (231), we used the mixed-product property of the Kronecker product in the second step, and in the third step, we noted that $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are the stationary distributions for $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, respectively. For a comprehensive list of properties that the Kronecker product satisfies, please refer to [41, p. 597]. □

Proposition 4 ([31, Exercise 12.6])**.** *Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be eigenvectors of $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, respectively, with eigenvalues λ and μ. Then $\boldsymbol{u} \otimes \boldsymbol{v}$ is an eigenvector of $\boldsymbol{P}_1 \otimes \boldsymbol{P}_2$ with eigenvalue $\lambda\mu$.*

Proof. We can verify the following:

$$(\boldsymbol{u} \otimes \boldsymbol{v})^\intercal (\boldsymbol{P}_1 \otimes \boldsymbol{P}_2) = (\boldsymbol{u}^\intercal \boldsymbol{P}_1) \otimes (\boldsymbol{v}^\intercal \boldsymbol{P}_2) = (\lambda \boldsymbol{u}^\intercal) \otimes (\mu \boldsymbol{v}^\intercal) = \lambda\mu (\boldsymbol{u} \otimes \boldsymbol{v})^\intercal. \tag{232}$$

In (232), we used the mixed-product property and the associativity of the scalar multiplication with the Kronecker product. □

In general, let $\boldsymbol{P}_1$ be a $m \times m$ matrix with eigenvalues $\lambda_1, ..., \lambda_m$, and $\boldsymbol{P}_2$ be a $n \times n$ matrix with eigenvalues $\mu_1, ..., \mu_n$. The complete eigen-decomposition of $\boldsymbol{P}_1 \otimes \boldsymbol{P}_2$ depends on the Kronecker product structure and involves combinations of the eigenvalues and eigenvectors of $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$.

Proposition 5 (Spectrum of the Kronecker product, [41, Exercise 7.8.11])**.** *Let the eigenvalues of $\boldsymbol{P}_1 \in \mathbb{R}^{m\times m}$ be denoted by λ_i and let the eigenvalues of $\boldsymbol{P}_2 \in \mathbb{R}^{n\times n}$ be denoted by μ_j. The eigenvalues of $\boldsymbol{P}_1 \otimes \boldsymbol{P}_2$ are the mn numbers $\{\lambda_i\mu_j\}_{i=1,j=1}^{m,n}$.*

Proof. Let $\boldsymbol{J}_1 = \boldsymbol{A}_1^{-1}\boldsymbol{P}_1\boldsymbol{A}_1$ and $\boldsymbol{J}_2 = \boldsymbol{A}_2^{-1}\boldsymbol{P}_2\boldsymbol{A}_2$ be the respective Jordan forms for $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$. We use the mixed-product property and the inverse property of the Kronecker product to show that $\boldsymbol{P}_1 \otimes \boldsymbol{P}_2$ is similar to $\boldsymbol{J}_1 \otimes \boldsymbol{J}_2$:

$$\boldsymbol{J}_1 \otimes \boldsymbol{J}_2 = (\boldsymbol{A}_1^{-1}\boldsymbol{P}_1\boldsymbol{A}_1) \otimes (\boldsymbol{A}_2^{-1}\boldsymbol{P}_2\boldsymbol{A}_2) = (\boldsymbol{A}_1^{-1} \otimes \boldsymbol{A}_2^{-1})(\boldsymbol{P}_1 \otimes \boldsymbol{P}_2)(\boldsymbol{A}_1 \otimes \boldsymbol{A}_2) = (\boldsymbol{A}_1 \otimes \boldsymbol{A}_2)^{-1}(\boldsymbol{P}_1 \otimes \boldsymbol{P}_2)(\boldsymbol{A}_1 \otimes \boldsymbol{A}_2). \tag{233}$$

Consequently, the eigenvalues of $\boldsymbol{P}_1 \otimes \boldsymbol{P}_2$ coincide with those of $\boldsymbol{J}_1 \otimes \boldsymbol{J}_2$. Since $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$ are upper triangular with $\{\lambda_i\}_{i=1}^m$ and $\{\mu_j\}_{j=1}^n$ on the diagonals, respectively, $\boldsymbol{J}_1 \otimes \boldsymbol{J}_2$ is also upper triangular with diagonal entries given by $\{\lambda_i\mu_j\}_{i=1,j=1}^{m,n}$. □

Proposition 6. *Let $\bar{\lambda}_2(\boldsymbol{P}_k)$ denote the second largest eigenvalue in absolute value of the transition matrix $\boldsymbol{P}_k$ associated with the k-th client, and define $\lambda(\boldsymbol{P}_k) := \frac{\bar{\lambda}_2(\boldsymbol{P}_k)+1}{2}$. For the product chain defined by $\boldsymbol{P} = \bigotimes_{k\in\mathcal{K}} \boldsymbol{P}_k$, the second largest eigenvalue in absolute value $\bar{\lambda}_2(\boldsymbol{P})$ and $\lambda(\boldsymbol{P}) := \frac{\bar{\lambda}_2(\boldsymbol{P})+1}{2}$ satisfy:*

$$\bar{\lambda}_2(\boldsymbol{P}) = \max_{k\in\mathcal{K}} \bar{\lambda}_2(\boldsymbol{P}_k) \quad \text{and} \quad \lambda(\boldsymbol{P}) = \max_{k\in\mathcal{K}} \lambda(\boldsymbol{P}_k). \tag{234}$$

The proof of Proposition 6 follows a similar structure to the one in [31, Corollary 12.13].

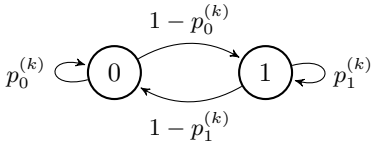
Proof. From Proposition 5, we know that the eigenvalues of $\boldsymbol{P} = \bigotimes_{k\in\mathcal{K}} \boldsymbol{P}_k$ are given by:

$$\left\{ \prod_{k\in\mathcal{K}} \lambda_i(\boldsymbol{P}_k) : \lambda_i(\boldsymbol{P}_k) \text{ an eigenvalue of } \boldsymbol{P}_k \right\}. \tag{235}$$

Recall that $\bar{\lambda}_2(\boldsymbol{P}_k)$ is the second largest eigenvalue of $\boldsymbol{P}_k$ in absolute value. If k^* denotes the index such that $\bar{\lambda}_2(\boldsymbol{P}_{k^*}) = \max_{k\in\mathcal{K}} \bar{\lambda}_2(\boldsymbol{P}_k)$, the second largest eigenvalue in module of $\boldsymbol{P}$ is the product of $\bar{\lambda}_2(\boldsymbol{P}_{k^*})$ for the k^*-th client and $\lambda_1(\boldsymbol{P}_j) = 1$ for the remaining clients $j \neq k^*$. The second result in (234) follows from the definitions of $\lambda(\boldsymbol{P})$ and $\lambda(\boldsymbol{P}_k)$. □

F3. Markov Chain for the Experiments (Section V)

In the experiments (Section V-A), we consider a scenario where the activity of each client $k \in \mathcal{K}$ follows a two-state homogeneous Markov process. The state space $\mathcal{M}$ consists of two states: "inactive" (with value 0) and "active" (with value 1):



We provide detailed expressions of the transition matrix $\boldsymbol{P}_k$, stationary distribution $\boldsymbol{\pi}^{(k)}$, and the second eigenvalue $\lambda_2(\boldsymbol{P}_k)$ used in the experiments for each client $k \in \mathcal{K}$:

$$\boldsymbol{P}_k = \begin{bmatrix} p_0^{(k)} & 1-p_0^{(k)} \\ 1-p_1^{(k)} & p_1^{(k)} \end{bmatrix} = \begin{bmatrix} 1-(1-\lambda_2(\boldsymbol{P}_k))\pi_k & (1-\lambda_2(\boldsymbol{P}_k))\pi_k \\ (1-\lambda_2(\boldsymbol{P}_k))(1-\pi_k) & \lambda_2(\boldsymbol{P}_k)+(1-\lambda_2(\boldsymbol{P}_k))\pi_k \end{bmatrix}. \tag{236}$$

$$\boldsymbol{\pi}^{(k)} = [1-\pi_k, \pi_k] = \left[\frac{1-p_1^{(k)}}{2-p_0^{(k)}-p_1^{(k)}}, \frac{1-p_0^{(k)}}{2-p_0^{(k)}-p_1^{(k)}}\right]. \tag{237}$$

$$\lambda_2(\boldsymbol{P}_k) = p_0^{(k)} + p_1^{(k)} - 1. \tag{238}$$

APPENDIX G
EXPERIMENTAL EVALUATION

G1. Details on Experimental Setup

A. Datasets and Models: In this section, we provide a detailed description of the datasets and models used in our experiments. We considered a total of $N = 100$ clients. We tested `CA-Fed` on the benchmark synthetic LEAF dataset [36] for regularized logistic regression tasks, which satisfy Assumptions 3-4. Additionally, we incorporated two "real-world" datasets: MNIST [37] for handwritten digit recognition and CIFAR-10 [38] for image recognition. Detailed descriptions of the datasets and the models used for each of them are provided below.

a) Synthetic LEAF dataset: Synthetic data provides us with precise control over heterogeneity. The Synthetic LEAF dataset achieves this by using parameters γ and δ, where γ determines the degree of variation among local models and δ determines the variability in the local data across different devices. The generation process follows the setup described in [23], [24]:

1) For each client $k \in \mathcal{K}$, sample the model parameters $\boldsymbol{W}_k \in \mathbb{R}^{10\times 60}$ and $\boldsymbol{b}_k \in \mathbb{R}^{10}$ from a normal distribution with mean μ_k and standard deviation 1, where μ_k is sampled from $\mathcal{N}(0,\gamma)$.
2) For each client $k \in \mathcal{K}$, generate the client's input data $\boldsymbol{X}_k \in \mathbb{R}^{n_k\times 60}$ as follows: sample each element $(x_k)_j$ from a normal distribution with mean v_k and standard deviation $\frac{1}{j^{1.2}}$, where v_k is sampled from $\mathcal{N}(B_k,1)$ and B_k is sampled from $\mathcal{N}(0,\delta)$.
3) Generate synthetic samples $(\boldsymbol{X}_k, \boldsymbol{Y}_k)$, where $\boldsymbol{Y}_k \in \mathbb{R}^{n_k}$, according to the model $y = \arg\max(\text{softmax}(\boldsymbol{W}_k\boldsymbol{x} + \boldsymbol{b}_k))$, where $\boldsymbol{x} \in \mathbb{R}^{60}$.

The distribution of samples $n_k = |D_k|$ among the clients follows a power law, resulting in an imbalanced data distribution. We refer to the synthetic dataset with parameters γ and δ as synthetic(γ,δ). We set (γ,δ) values to $(0,0)$, $(0.25,0.25)$, $(0.5,0.5)$, $(0.75,0.75)$, and $(1,1)$ to investigate various levels of heterogeneity in the data.

TABLE I: Average computation time and used CPU/GPU for each dataset.

Dataset	CPU/GPU	Simulation time
Binary Synthetic	Intel(R) Xeon(R) CPU	10min
Synthetic LEAF	Intel(R) Xeon(R) CPU	6min
MNIST [37]	GeForce GTX 1080 Ti	42min
CIFAR10 [38]	GeForce GTX 1080 Ti	2h37min

TABLE II: Learning rates η and $\bar{\eta}$ used for the experiments in Figure 1.

Dataset	`Unbiased`	`More available`	`CA-Fed` ($\bar{\kappa} = 1$)	`AdaFed` [20]	`F3AST` [19]
Synthetic LEAF	2.0/2.0	1.0/7.0	2.0/3.0	1.0/1.0	2.0/2.0
MNIST	0.03/1.0	0.1/4.0	0.1/1.0	0.03/1.0	0.1/0.3
CIFAR10	0.03/1.0	0.03/3.0	0.03/1.0	0.03/1.0	0.03/0.3

b) MNIST: To classify handwritten digits in the MNIST dataset, we employ multinomial logistic regression. The model takes a flattened 784-dimensional (28 × 28) image as input and predicts a class label from 0 to 9 as output. To introduce heterogeneity in the data distribution, we distribute the dataset among $N = 100$ clients using a Dirichlet allocation method [39] with parameter ς. This allocation scheme allows for varying proportions of the dataset to be assigned to each client, contributing to the heterogeneous nature of our experimental setting.

c) CIFAR-10: The CIFAR-10 dataset consists of 60,000 input images, sourced from a collection of 80 million tiny images, with 10 distinct labels. To partition the CIFAR-10 dataset among $N = 100$ clients, we employ a Dirichlet allocation [39] with parameter ς. For this particular dataset, we train a shallow neural network comprising two convolutional layers followed by one fully connected layer. This network architecture is designed to capture relevant features from the CIFAR-10 images and facilitate accurate classification.

B. Implementation Details:

a) Machines: The experiments were conducted on a CPU/GPU cluster, utilizing various available GPUs such as Nvidia Tesla V100, GeForce GTX 1080 Ti, and Quadro RTX 8000. The majority of experiments involving Synthetic datasets were executed on an Intel(R) Xeon(R) CPU E5-1660 v3 @ 3.00GHz. On the other hand, experiments involving MNIST and CIFAR-10 datasets were performed using GeForce GTX 1080 Ti cards. For each dataset, we conducted approximately 50 experiments, excluding the time dedicated to development and debugging. Due to the usage of a train batch size of 32 samples, the experiments with MNIST and CIFAR-10 datasets exhibited slower execution times. Table I provides the average duration required to execute one simulation for each dataset. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

b) Libraries: We extensively employed the PyTorch deep learning framework throughout our experiments. PyTorch provided us with a comprehensive set of tools and functionalities for model construction, training, and evaluation. It allowed us to efficiently implement and optimize various neural network architectures, including the multinomial logistic regression model for the MNIST dataset and the shallow neural network for the CIFAR-10 dataset. To simplify the data preparation process, we utilized Torchvision, a PyTorch package designed for computer vision tasks. Torchvision facilitated seamless dataset management, including the download and pre-processing of MNIST and CIFAR-10, enabling us to transform the raw image data into a suitable format for training and evaluation.

c) Hyper-parameters: For each method and task, we performed a grid search to determine the optimal learning rates η and $\bar{\eta}$. For the MNIST and CIFAR-10 datasets, we explored the grids $\eta = \{2.0, 1.0, 0.3, 0.1, 0.03, 0.01\}$ and $\bar{\eta} = \{5.0, 4.0, 3.0, 2.0, 1.0, 0.3, 0.1\}$. For the Synthetic LEAF dataset, we shifted the grid to $\bar{\eta} = \{8.0, 7.0, 6.0, 5.0, 4.0, 3.0, 2.0, 1.0\}$. Table II reports the learning rates η and $\bar{\eta}$ corresponding to the results in Figure 1 for each dataset and method. For `CA-Fed`, we use the hyper-parameters $\beta = \tau = 0$. In the case of `AdaFed`, we set full device participation, where the parameter server samples all active clients ($|\mathcal{S}_t| = |\mathcal{A}_t|$). To ensure a fair comparison, we set the number of clients sampled by `F3AST` to the average number of clients included by `CA-Fed`, which is 45 on average. Furthermore, we set the smoothness parameter β of `F3AST` to be $\mathcal{O}(1/T)$, as suggested by the authors in [19, Appendix D].

APPENDIX H
FURTHER DISCUSSION ABOUT CA-FED

H1. CA-Fed's computation/communication cost

CA-Fed aims to improve training convergence and not to reduce its computation and communication overhead. Nevertheless, excluding some available clients reduces the overall training cost, as we will discuss in this section referring, for the sake of concreteness, to neural networks' training.

In terms of computation, the available clients not selected for training are only requested to evaluate their local loss on the current model once on a single batch instead than performing E gradient updates, which would require roughly $2 \times E - 1$ more calculations (because of the forward and backward pass). The selected clients have no extra computation cost as computing the loss corresponds to the forward pass they should, in any case, perform during the first local gradient update.

In terms of communication, the excluded clients only transmit the loss, a single scalar, much smaller than the model update. Conversely, participating clients transmit the local loss and the model update. Still, this additional overhead is negligible and likely fully compensated by the communication savings for the excluded clients.

H2. CA-Fed and Client Sampling

In cross-device FL, a common practice is to employ client sampling, where a small subset of clients (denoted as $\mathcal{S}_t$) is uniformly selected at random from the set of active clients ($\mathcal{A}_t$) during each communication round of model training. This is primarily done to mitigate communication overhead and enhance scalability.

In our analysis, based on Assumption 1, we assume that spatial and temporal correlations primarily concern clients' availability dynamics and we consider, for simplicity, $\mathcal{S}_t = \mathcal{A}_t$. However, our findings have a noteworthy implication: while the set of available clients $\mathcal{A}_t$ exhibits correlation, the client sampling in $\mathcal{S}_t$ can be designed to make clients' participation dynamics independent over time and among clients. A promising direction for future research is to extend our work in this context and derive a refined bound similar to our result in Theorem 2 which quantifies the impact of client sampling on $\lambda(\boldsymbol{P})$.

Consistent with our analysis, we have designed our algorithm to align with the assumption $\mathcal{S}_t = \mathcal{A}_t$. By design, CA-Fed excludes clients with large temporal correlation and low availability and activates, in each communication round, only clients satisfying $\{k \in \mathcal{A}_t;\ q_k^{(t)} > 0\}$ (line 8 in Algorithm 1). However, when only a small fraction of clients is excluded, CA-Fed seamlessly integrates with client sampling. This only involves replacing $\mathcal{A}_t$ with $\mathcal{S}_t$ in Equation (17) and Algorithm 1 (server estimates for clients' local losses ($\hat{\boldsymbol{F}}^{(t)} = (\hat{F}_k^{(t)})_{k\in\mathcal{K}}$) are now updated from the sampled clients' losses ($\boldsymbol{F}^{(t)} = (F_k^{(t)})_{k\in\mathcal{S}_t}$)).

H3. About CA-Fed's fairness

Strategies that exclude clients from the training phase, such as CA-Fed, may raise concerns about fairness. The concept of *fairness* in federated learning does not have a unified definition in the literature [42, Chapter 8]. Fairness goals can be established by appropriately selecting the target weights $\boldsymbol{\alpha} = \{\alpha_k\}_{k\in\mathcal{K}}$ in the definition of the global target objective (1). For instance, *per-client fairness* can be achieved by setting α_k to be equal for every client (i.e., $\alpha_k = 1/N$), while *per-sample fairness* can be accomplished by setting α_k proportional to the local dataset size $|D_k|$ (i.e, $\alpha_k = |D_k|/|D|$).

Assuming that the global objective in (1) truly reflects fairness concerns, then CA-Fed can be considered intrinsically fair. This is because CA-Fed continually focuses on minimizing the total error $\epsilon \coloneqq F(\boldsymbol{w}_T) - F^*$, which guarantees that the performance objective of the learned model is as close as possible to its optimal value at every time. Although CA-Fed occasionally excludes clients with low availability and high temporal correlation, the optimization problem (1) is carefully designed to ensure that the learned model performs well for these clients. As a result, CA-Fed effectively learns a model that is consistently accurate and fair across all clients, regardless of their availability or temporal correlation.