Federated Learning Under Heterogeneous and Correlated Client Availability

Angelo Rodio, Francescomaria Faticanti, Othmane Marfoq, Giovanni Neglia, Emilio Leonardi

Federated Learning under Heterogeneous and Correlated Client Availability

Angelo Rodio*, Francescomaria Faticanti*, Othmane Marfoq*†, Giovanni Neglia*, Emilio Leonardi‡
*Inria, Université Côte d'Azur, France. Email: {firstname.lastname}@inria.fr,
†Accenture Labs, Sophia-Antipolis, France. Email: {firstname.lastname}@accenture.com,
‡Politecnico di Torino, Turin, Italy. Email: {firstname.lastname}@polito.it

***Abstract*—In Federated Learning (FL), devices – also referred to as clients – can exhibit heterogeneous availability patterns, often correlated over time and with other clients. This paper addresses the problem of heterogeneous and correlated client availability in FL. Our theoretical analysis is the first to demonstrate the negative impact of correlation on FL algorithms' convergence rate and highlights a trade-off between optimization error (related to convergence speed) and bias error (indicative of model quality). To optimize this trade-off, we propose Correlation-Aware FL (`CA-Fed`), a novel algorithm that dynamically balances the competing objectives of fast convergence and minimal model bias. `CA-Fed` achieves this by dynamically adjusting the aggregation weight assigned to each client and selectively excluding clients with high temporal correlation and low availability. Experimental evaluations on diverse datasets demonstrate the effectiveness of `CA-Fed` compared to state-of-the-art methods. Specifically, `CA-Fed` achieves the best trade-off between training time and test accuracy. By dynamically handling clients with high temporal correlation and low availability, `CA-Fed` emerges as a promising solution to mitigate the detrimental impact of correlated client availability in FL.**

***Index Terms*—Federated Learning, Correlated Client Availability, Markov Chains.**

I. Introduction

The enormous amount of data generated by mobile and IoT devices motivated the development of distributed machine learning training paradigms [2], [3]. Federated Learning (FL) [4]–[7] is an emerging framework where geographically distributed devices (or clients) participate in the training of a shared Machine Learning (ML) model without sharing their local data. FL was proposed to reduce the overall cost of collecting a large amount of data as well as to protect potentially sensitive users' private information. In the original Federated Averaging algorithm (`FedAvg`) [5], a central server selects a random subset of clients from the set of available clients and broadcasts them the shared model. The sampled clients perform a number of independent Stochastic Gradient Descent (SGD) steps over their local datasets and send their local model updates back to the server. Then, the server aggregates the received client updates to produce a new global model, and a new training round begins. At each iteration of `FedAvg`, the server typically samples randomly a few hundred devices to participate [8], [9].

This research was supported by the French government through the 3IA Côte d'Azur Investments in the Future project by the National Research Agency (ANR) with reference ANR-19-P3IA-0002, and by Groupe La Poste, sponsor of Inria Foundation, in the framework of FedMalin Inria Challenge.
A first version of this work was presented at IEEE INFOCOM 2023 [1].

In real-world scenarios, the availability of clients is dictated by exogenous factors that are beyond the control of the orchestrating server and hard to predict. For instance, only smartphones that are idle, under charge, and connected to broadband networks are commonly allowed to participate in the training process [5], [10]. These eligibility requirements can make the availability of devices correlated over time and space [8], [11]–[13]. For example, *temporal correlation* may origin from a smartphone being under charge for a few consecutive hours and then ineligible for the rest of the day. Similarly, the activity of a sensor powered by renewable energy may depend on natural phenomena intrinsically correlated over time (e.g., solar light). *Spatial correlation* refers instead to correlation across different clients, which often emerges as consequence of users' different geographical distribution. For instance, clients in the same time zone often exhibit similar availability patterns, e.g., due to time-of-day effects.

Temporal correlation in the data sampling procedure is known to negatively affect the performance of ML training even in the centralized setting [14], [15] and can potentially lead to *catastrophic forgetting*: the data used during the final training phases can have a disproportionate effect on the final model, "erasing" the memory of previously learned information [16], [17]. Catastrophic forgetting has also been observed in FL, where clients in the same geographical area have more similar local data distributions and clients' participation follows a cyclic daily pattern (leading also to spatial correlation) [8], [11], [12], [18]. Despite this evidence, a theoretical study of the convergence of FL algorithms under both temporally and spatially correlated client participation is still missing.

This paper presents the first convergence analysis of `FedAvg` [5] under heterogeneous and correlated client availability. We assume that the clients' availability follows an arbitrary finite-state Markov chain, modeling both temporal and spatial correlation while maintaining analytical tractability. Our theoretical analysis provides valuable insights by (i) quantitatively measuring the negative impact of correlation on the algorithm's convergence rate through a novel additional term that depends on the spectral properties of the Markov chain, and (ii) highlighting an important trade-off between two conflicting objectives: slow convergence to the optimal model and fast convergence to a biased model that minimizes a different objective function from the initial target. To leverage this trade-off, we propose `CA-Fed`,

an algorithm which achieves an optimal balance between maximizing convergence speed and minimizing model bias through dynamic adjustment of aggregation weights assigned to clients. Depending on their contribution to the learning process, `CA-Fed` can decide to exclude clients exhibiting low availability and high temporal correlation. Our experimental results demonstrate that excluding such clients is a simple, but effective approach to handle the heterogeneous and correlated client availability in FL. Across synthetic and real datasets, `CA-Fed` consistently outperforms the state-of-the-art methods `F3AST` [19] and `AdaFed` [20] in terms of test accuracy. These results underscore the importance of optimizing the training process to leverage available client resources effectively and mitigate the impact of less available and correlated clients, a task successfully accomplished by `CA-Fed`.

The remainder of this paper is organized as follows. Section II introduces the problem of correlated client availability in FL and discusses the main related works. Section III provides a convergence analysis of `FedAvg` under heterogeneous and correlated client availability. `CA-Fed`, our correlation-aware FL algorithm, is presented in Section IV. We evaluate `CA-Fed` in Section V, comparing it with state-of-the-art methods on synthetic and real-world data. Section VI concludes the paper. Supplementary material, comprising Appendices A–H, provides detailed proofs and further discussions on `CA-Fed` not included in the main text due to space constraints.

II. Background and Related Works

We consider a finite set $\mathcal{K}$ of N clients. Each client $k \in \mathcal{K}$ holds a local dataset D_k. Clients aim to jointly learn the parameters $\boldsymbol{w} \in W \subseteq \mathbb{R}^d$ of a global ML model (e.g., the weights of a neural network architecture). During training, the quality of the model with parameters $\boldsymbol{w}$ on a data sample $\xi \in D_k$ is measured by a loss function $f(\boldsymbol{w};\xi)$. The clients solve, under the orchestration of a central server, the following optimization problem:

$$\min_{\boldsymbol{w}\in W\subseteq\mathbb{R}^d} \left[F(\boldsymbol{w}) \coloneqq \sum_{k\in\mathcal{K}} \alpha_k F_k(\boldsymbol{w}) \right], \tag{1}$$

where $F_k(\boldsymbol{w}) \coloneqq \frac{1}{|D_k|}\sum_{\xi\in D_k} f(\boldsymbol{w};\xi)$ is the average loss computed on client k's local dataset, and $\boldsymbol{\alpha} = (\alpha_k)_{k\in\mathcal{K}}$ are positive coefficients such that $\sum_k \alpha_k = 1$. They represent the *target importance* assigned by the central server to each client k. Typically $(\alpha_k)_{k\in\mathcal{K}}$ are set proportional to the clients' dataset size $|D_k|$, such that the objective function F in (1) coincides with the average loss computed on the union of the clients' local datasets $D = \cup_{k\in\mathcal{K}} D_k$.

Under proper assumptions, precised in Section III, Problem (1) admits a unique solution. We use $\boldsymbol{w}^*$ (resp. F^*) to denote the minimizer (resp. the minimum value) of F. Moreover, for $k \in \mathcal{K}$, F_k admits a unique minimizer. We use $\boldsymbol{w}_k^*$ (resp. F_k^*) to denote the minimizer (resp. the minimum value) of F_k.

Problem (1) is commonly solved through iterative algorithms [5], [9] requiring multiple communication rounds between the server and the clients. At round $t > 0$, the server broadcasts the latest estimate of the global model $\boldsymbol{w}_{t,0}$ to the set of available clients ($\mathcal{A}_t$). Client $k \in \mathcal{A}_t$ updates the global model with its local data through $E \geq 1$ steps of local Stochastic Gradient Descent (SGD):

$$\boldsymbol{w}_{t,j+1}^k = \boldsymbol{w}_{t,j}^k - \eta_t \nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k) \quad j = 0,\dots,E-1, \tag{2}$$

where $\eta_t > 0$ is an appropriately chosen learning rate, referred to as *local learning rate*; $\mathcal{B}_{t,j}^k$ is a random batch sampled from client-k's local dataset at round t and step j; $\nabla F_k(\cdot,\mathcal{B}) \coloneqq \frac{1}{|\mathcal{B}|}\sum_{\xi\in\mathcal{B}} \nabla f(\cdot,\xi)$ is an unbiased estimator of the local gradient ∇F_k. Then, each client sends its local model update $\Delta_t^k \coloneqq \boldsymbol{w}_{t,E}^k - \boldsymbol{w}_{t,0}^k$ to the server. The server computes $\Delta_t \coloneqq \sum_{k\in\mathcal{A}_t} q_k \cdot \Delta_t^k$, a weighted average of the clients' local updates with non-negative *aggregation weights* $\boldsymbol{q} = (q_k)_{k\in\mathcal{K}}$. The choice of the aggregation weights defines an aggregation strategy (we will discuss different aggregation strategies later). The aggregated update Δ_t can be interpreted as a proxy for $-\nabla F(\boldsymbol{w}_{t,0})$; the server applies it to the global model:

$$\boldsymbol{w}_{t+1,0} = \mathbf{Proj}_W(\boldsymbol{w}_{t,0} + \bar{\eta}\cdot\Delta_t), \tag{3}$$

where $\mathbf{Proj}_W(\cdot)$ denotes the projection over the set W, and $\bar{\eta} > 0$ is an appropriately chosen learning rate, referred to as the *server learning rate*.[1]

The aggregate update Δ_t is generally a biased estimator of the pseudo-gradient $-\nabla F(\boldsymbol{w}_{t,0})$, to which each client k contributes proportionally to its frequency of appearance in the set $\mathcal{A}_t$ and its aggregation weight q_k. More specifically, under proper assumptions specified in Section III, we will prove in Theorem 2 that the update rule described by (2) and (3) converges to the unique minimizer of a biased global objective F_B. This objective function depends depends both on the clients' availability (i.e., on the sequence $(\mathcal{A}_t)_{t>0}$) and on the aggregation strategy (i.e., on $\boldsymbol{q} = (q_k)_{k\in\mathcal{K}}$):

$$F_B(\boldsymbol{w}) \coloneqq \sum_{k=1}^N p_k F_k(\boldsymbol{w}), \text{ with } p_k \coloneqq \frac{\pi_k q_k}{\sum_{h=1}^N \pi_h q_h}, \tag{4}$$

where π_k represents the asymptotic availability of client k, defined as $\pi_k \coloneqq \lim_{t\to+\infty}\mathbb{P}(k \in \mathcal{A}_t)$. We denote $\boldsymbol{\pi} = (\pi_k)_{k\in\mathcal{K}}$. Moreover, the coefficients $\boldsymbol{p} = (p_k)_{k\in\mathcal{K}}$ in (4) can be interpreted as the *biased importance* the server is giving to each client k during training, in general different from the *target importance* $\boldsymbol{\alpha}$. In what follows, $\boldsymbol{w}_B^*$ (resp. F_B^*) denotes the minimizer (resp. the minimum value) of F_B.

In some large-scale FL applications, like training Google keyboard next-word prediction models, each client participates in training at most for one round. The orchestrator usually selects a few hundred clients at each round for a few thousand rounds (e.g., see [6, Table 2]), but the available set of clients may include hundreds of millions of Android devices. In this scenario, it is difficult to address the potential bias unless there is some a-priori information about each client's availability.

[1] The aggregation rule (3) has been considered also in other works, e.g., [9], [21], [22]. In other FL algorithms, the server computes an average of clients' local models. This aggregation rule can be obtained with minor changes to (3).

Anyway, FL can be used by service providers with access to a much smaller set of clients (e.g., smartphone users that have installed a specific app). In this case, a client participates multiple times in training: the orchestrating server may keep track of each client's availability and try to compensate for the potentially dangerous heterogeneity in their participation.

Much previous effort on federated learning [5], [18]–[20], [23]–[26] considered this problem and, under different assumptions on the clients' availability (i.e., on $(\mathcal{A}_t)_{t>0}$), designed aggregation strategies that unbias Δ_t through an appropriate choice of $\boldsymbol{q}$. Reference [23] provides the first analysis of `FedAvg` on non-iid data under clients' partial participation. Their analysis covers both the case when active clients are sampled uniformly at random without replacement from $\mathcal{K}$ and assigned aggregation weights equal to their target importance (as assumed in [5]), and the case when active clients are sampled iid with replacement from $\mathcal{K}$ with probabilities $\boldsymbol{\alpha}$ and assigned equal weights (as assumed in [24]). However, references [5], [23], [24] ignore the variance induced by the clients stochastic availability. The authors of [25] reduce such variance by considering only the clients with important updates, as measured by the value of their norm. References [18] and [26] reduce the aggregation variance through clustered and soft-clustered sampling, respectively.

Some recent works [19], [20], [27] do not actively pursue the optimization of the unbiased objective. Instead, they derive bounds for the convergence error and propose heuristics to minimize those bounds, potentially introducing some bias. Our work follows a similar development: we compare our algorithm with `F3AST` from [19] and `AdaFed` from [20].

The novelty of our study is in considering the spatial and temporal correlation in clients' availability dynamics. As discussed in the introduction, such correlations are also introduced by clients' eligibility criteria, e.g., smartphones being under charge and connected to broadband networks. The effect of correlation has been ignored until now, probably due to the additional complexity in studying FL algorithms' convergence. To the best of our knowledge, the only exception is [19], which scratches the issue of spatial correlation by proposing two different algorithms for the case when clients' availabilities are uncorrelated and for the case when they are positively correlated (there is no smooth transition from one algorithm to the other as a function of the degree of correlation).

The effect of temporal correlation on *centralized* stochastic gradient methods has been addressed in [13]–[15], [28]: these works study a variant of stochastic gradient descent where samples are drawn according to a Markov chain. Reference [13] extends its analysis to a FL setting where each client draws samples according to a Markov chain. In contrast, our work does not assume a correlation in the data sampling but rather in the client's availability. Nevertheless, some of our proof techniques are similar to those used in this line of work and, in particular, we rely on some results in [15].

III. Analysis

A. Main assumptions

We consider a time-slotted system where a slot corresponds to a single FL communication round. We assume that clients' availability over the timeslots $t \in \mathbb{N}$ follows a discrete-time Markov chain $(\mathcal{A}_t)_{t\geq 0}$.[2]

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 $t = 0$.*

Markov chains have already been used in the literature to model the dynamics of stochastic networks where some nodes or edges in the graph can switch between active and inactive states [29], [30]. The previous Markovian assumption, while allowing a great degree of flexibility, still guarantees the analytical tractability of the system. The distance dynamics between the current and the stationary distributions of the Markov process can be characterized in terms of the spectral properties of its transition matrix $\boldsymbol{P}$ [31]. Let $\bar{\lambda}_2(\boldsymbol{P})$ denote the the second largest module of the eigenvalues of $\boldsymbol{P}$. Previous work [15] has shown that:

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

where the parameters $\lambda(\boldsymbol{P}) \coloneqq (\bar{\lambda}_2(\boldsymbol{P}) + 1)/2$, C_P, and T_P are positive constants whose values are defined in [15, Lemma 1] and reported for completeness in Appendix B2, Lemma 16.[3] Note that $\lambda(\boldsymbol{P})$ quantifies the correlation of the Markov process $(\mathcal{A}_t)_{t\geq 0}$: the closer $\lambda(\boldsymbol{P})$ is to one, the slower the Markov chain converges to its stationary distribution.

In our analysis, we make the following additional assumptions.

Assumption 2. *The hypothesis class W is convex and compact with diameter $\operatorname{diam}(W)$, and contains the minimizers $\boldsymbol{w}^*, \boldsymbol{w}_B^*, \boldsymbol{w}_k^*$ in its interior.*

The following assumptions concern clients' local objective functions $\{F_k\}_{k\in\mathcal{K}}$. Assumptions 3 and 4 are standard in the literature on convex optimization [32, Sections 4.1, 4.2]. Assumption 5 is a standard hypothesis in the analysis of federated optimization algorithms [9, Section 6.1].

Assumption 3 (L-smoothness)**.** *The local functions $\{F_k\}_{k=1}^N$ have L-Lipschitz continuous gradients: $F_k(\boldsymbol{v}) \leq F_k(\boldsymbol{w}) + \langle \nabla F_k(\boldsymbol{w}), \boldsymbol{v} - \boldsymbol{w}\rangle + \frac{L}{2}\|\boldsymbol{v} - \boldsymbol{w}\|_2^2$, $\forall \boldsymbol{v}, \boldsymbol{w} \in W$.*

Assumption 4 (Strong convexity)**.** *The local functions $\{F_k\}_{k=1}^N$ are μ-strongly convex: $F_k(\boldsymbol{v}) \geq F_k(\boldsymbol{w}) + \langle \nabla F_k(\boldsymbol{w}), \boldsymbol{v} - \boldsymbol{w}\rangle + \frac{\mu}{2}\|\boldsymbol{v} - \boldsymbol{w}\|_2^2$, $\forall \boldsymbol{v}, \boldsymbol{w} \in W$.*

Assumption 5 (Bounded variance)**.** *The variance of stochastic gradients in each device is bounded: $\mathbb{E}\|\nabla F_k(\boldsymbol{w}, \mathcal{B}) - \nabla F_k(\boldsymbol{w})\|^2 \leq \sigma_k^2$, $k = 1, \ldots, N$.*

[2] In Section III-D we will focus on the case where this chain is the superposition of N independent Markov chains, one for each client.

[3] Note that (5) holds for different definitions of $\lambda(\boldsymbol{P})$ as long as $\lambda(\boldsymbol{P}) \in (\bar{\lambda}_2(\boldsymbol{P}), 1)$. The specific choice for $\lambda(\boldsymbol{P})$ changes the values of C_P and T_P.

Assumptions 2–5 imply the following properties for the local functions, described by Lemma 1 (proof in Appendix B).

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

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

$$\mathbb{E}\,\|\nabla F_k(\boldsymbol{w}, \mathcal{B})\|^2 \leq G^2, \tag{7}$$

$$|F_k(\boldsymbol{w}) - F_k(\boldsymbol{w}_B^*)| \leq H. \tag{8}$$

Similarly to other works [9], [23], [24], [33], we introduce a metric to quantify the heterogeneity of clients' local datasets, typically referred to as *statistical heterogeneity*:

$$\Gamma \coloneqq \max_{k \in \mathcal{K}} \{F_k(\boldsymbol{w}^*) - F_k^*\}. \tag{9}$$

If the local datasets are identical, the local functions $\{F_k\}_{k\in\mathcal{K}}$ coincide among them and with F, $\boldsymbol{w}^*$ is a minimizer of each local function, and $\Gamma = 0$. In general, Γ is smaller the closer the distributions the local datasets are drawn from.

B. Main theorems

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

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

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

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

Theorem 1 (proof in Appendix A) decomposes the error of the target objective (ϵ) as the sum of an optimization error for the biased objective (ϵ_{opt}) and a bias error (ϵ_{bias}). The term ϵ_{opt}, evaluated on the trajectory determined by scheme (3), quantifies the optimization error associated with the biased objective F_B and asymptotically vanishes (see Theorem 2 below). The non-vanishing bias error ϵ_{bias} captures the discrepancy between $F(\boldsymbol{w}_B^*)$ and F^*. This term is bounded by the chi-square divergence $\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}}$ between the target and biased probability distributions $\boldsymbol{\alpha} = (\alpha_k)_{k\in\mathcal{K}}$ and $\boldsymbol{p} = (p_k)_{k\in\mathcal{K}}$, and by Γ, that quantifies the degree of heterogeneity of the local functions. When all local functions are identical ($\Gamma = 0$), the bias term ϵ_{bias} also vanishes. For $\Gamma > 0$, the bias error can still be controlled by the aggregation weights assigned to the devices. In particular, the bias term vanishes when $q_k \propto \alpha_k/\pi_k, \forall k \in \mathcal{K}$. Since it asymptotically cancels the bias error, we refer to this choice as *unbiased aggregation strategy*.

However, in practice, FL training is limited to a finite number of iterations T (typically a few hundreds [6], [8]), and the previous asymptotic considerations may not apply. In this regime, the unbiased aggregation strategy can be sub-optimal, since the minimization of ϵ_{bias} not necessarily leads to the minimization of the total error $\epsilon \leq 2\kappa^2(\epsilon_{\text{opt}} + \epsilon_{\text{bias}})$. This motivates the analysis of the optimization error ϵ_{opt}.

Theorem 2 (Convergence of the optimization error ϵ_{opt})**.** *Let Assumptions 1–5 hold and the constants M, L, D, G, H, Γ, σ_k, C_P, T_P, and $\lambda(\boldsymbol{P})$ defined above. Let $Q \coloneqq \sum_{k\in\mathcal{K}} q_k$. We require a diminishing step-size $\eta_t > 0$ satisfying:*

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

Let T denote the total communication rounds. For $T \geq T_P$, the expected optimization error can be bounded as follows:

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

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

$$\boldsymbol{\Sigma} \coloneqq \operatorname{diag}(2(E+1)\sigma_k^2 \pi_k \textstyle\sum_{t=1}^{+\infty} \eta_t^2),$$

$$\upsilon \coloneqq \tfrac{2}{E} \operatorname{diam}(W)^2 + \tfrac{1}{4} MQ \textstyle\sum_{t=1}^{+\infty} (\eta_t^2 + \frac{1}{t^2}),$$

$$\psi \coloneqq (4L(1+EQ)\Gamma + 2E^2G^2) \textstyle\sum_{t=1}^{+\infty} \eta_t^2 + H(\sum_{t=1}^{T_P - 1} \eta_t),$$

$$\mathcal{J}_t \coloneqq \min\left\{\max\left\{\left\lceil \ln\left(2C_P H t\right)/\ln\left(1/\lambda(\boldsymbol{P})\right)\right\rceil, T_P\right\}, t\right\},$$

$$\phi \coloneqq 2EDGQ \textstyle\sum_{t=1}^{+\infty} \ln(2C_P H t) \eta_{t-\mathcal{J}_t}^2.$$

Theorem 2 (proof in Appendix B) proves convergence of the expected biased objective F_B to its minimum F_B^* under correlated client participation. Our bound (13) captures the effect of correlation through the factor $\ln(1/\lambda(\boldsymbol{P}))$: a high correlation worsens the convergence rate. In particular, we found that the numerator of (13) has a quadratic-over-linear fractional dependence on $\boldsymbol{q}$. Minimizing $\bar{\epsilon}_{\text{opt}}$ leads, in general, to a different choice of $\boldsymbol{q}$ than minimizing $\bar{\epsilon}_{\text{bias}}$.

C. Minimizing the total error $\epsilon \leq 2\kappa^2(\bar{\epsilon}_{opt} + \bar{\epsilon}_{bias})$

Our analysis points out a trade-off between minimizing $\bar{\epsilon}_{\text{opt}}$ or $\bar{\epsilon}_{\text{bias}}$. Our goal is to find the optimal aggregation weights $\boldsymbol{q}^*$ that minimize the upper bound on total error $\epsilon(\boldsymbol{q})$ in (10):

$$\begin{aligned} \underset{\boldsymbol{q}}{\text{minimize}} \quad & \bar{\epsilon}_{\text{opt}}(\boldsymbol{q}) + \bar{\epsilon}_{\text{bias}}(\boldsymbol{q}); \\ \text{subject to} \quad & \boldsymbol{q} \geq 0, \\ & \|\boldsymbol{q}\|_1 = Q. \end{aligned} \tag{14}$$

In Appendix D we prove that (14) is a convex optimization problem, which can be solved with the method of Lagrange multipliers. However, its solution lacks practical utility because the constants in (10) and (13) (e.g., L, μ, 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. Moreover, the bound in (10), as well as the bound in [33], diverges when setting some q_k values equal to 0, but this divergence is merely an artifact of the proof technique. For more practical considerations, we present the following result (proof in Appendix C):

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

The new constant Γ' is defined in terms of $\boldsymbol{w}_B^*$, and then it is easier to evaluate during training. However, Γ' depends on $\boldsymbol{q}$, because it is evaluated at the point of minimum of F_B. This dependence makes the minimization of the right-hand side of (15) more challenging (for example, the corresponding problem is not convex). We study the minimization of the two terms $\bar{\epsilon}_{\text{opt}}$ and $\bar{\epsilon}'_{\text{bias}}$ separately and learn some insights, which we use to design the new FL algorithm `CA-Fed`.

D. Minimizing $\bar{\epsilon}_{opt}$

The minimization of $\bar{\epsilon}_{\text{opt}}$ is still a convex optimization problem (Appendix E). In particular, at the optimum, non-negative weights are set accordingly to $q_k^* = a(\iota^*\pi_k - \theta^*)$ with a and ι^* positive constants (Appendix E2). It follows that clients with smaller availability get smaller weights in the aggregation. In particular, this suggests that clients with the smallest availability can be excluded from the aggregation, leading to the following guideline:

Guideline A: to accelerate convergence, we can exclude clients with low availability π_k *by setting* $q_k^* = 0$.

This guideline can be justified intuitively: updates from clients with low participation may be too sporadic to allow the FL algorithm to keep track of their local objectives. Their updates act as a noise slowing down the algorithm's convergence. It may then be advantageous to exclude these clients.

We observe that the choice of the aggregation weights $\boldsymbol{q}$ does not affect the clients' availability process and, in particular, $\lambda(\boldsymbol{P})$. However, if the algorithm excludes some clients, it is possible to consider the state space of the Markov chain that only specifies the availability state of the remaining clients, and this Markov chain may have different spectral properties. For the sake of concreteness, unless otherwise specified, we consider from now on the particular case when the availability of each client k evolves according to a Markov chain $(\mathcal{A}_t^k)_{t\geq 0}$ with transition probability matrix $\boldsymbol{P}_k$ and these Markov chains are all independent [31, Exercise 12.6]. In this case, the aggregate process is described by the product Markov chain $(\mathcal{A}_t)_{t\geq 0}$ with transition matrix $\boldsymbol{P} = \bigotimes_{k\in\mathcal{K}} \boldsymbol{P}_k$ and $\lambda(\boldsymbol{P}) = \max_{k\in\mathcal{K}} \lambda(\boldsymbol{P}_k)$, where $\boldsymbol{P}_i \bigotimes \boldsymbol{P}_j$ denotes the Kronecker product between matrices $\boldsymbol{P}_i$ and $\boldsymbol{P}_j$ (Appendix F2). In this setting, it is possible to redefine the Markov chain $(\mathcal{A}_t)_{t\geq 0}$ by taking into account the reduced state space defined by the clients with a non-null aggregation weight, i.e., $\boldsymbol{P'} = \bigotimes_{k'\in\mathcal{K}|q_{k'}>0} \boldsymbol{P}_{k'}$ and $\lambda(\boldsymbol{P'}) = \max_{k'\in\mathcal{K}|q_{k'}>0} \lambda(\boldsymbol{P}_{k'})$, which is potentially smaller w.r.t. the case when all clients participate to the aggregation. These considerations lead to the following guideline:

Guideline B: to accelerate convergence, we can exclude clients with high correlation (high $\lambda(\boldsymbol{P}_k)$*) by setting their* $q_k^* = 0$.

Intuition also supports this guideline. Clients with large $\lambda(\boldsymbol{P}_k)$ tend to be available or unavailable for long periods of time. Due to the well-known catastrophic forgetting problem affecting gradient methods [34], [35], these clients may unfairly steer the algorithm toward their local objective when they appear at the final stages of the training period. Moreover, their participation in the early stages may be useless, as their contribution will be forgotten during their long absence. The FL algorithm may benefit from directly neglecting such clients.

We observe that Guideline B strictly applies to this specific setting where clients' dynamics are independent (and there is no spatial correlation). We do not provide a corresponding guideline for the case when clients are spatially correlated (we leave this task for future research). However, in this more general setting, it is possible to ignore Guideline B but still draw on Guidelines A and C, or still consider Guideline B if the spatially correlated clients can be grouped in clusters, each cluster evolving as an independent Markov chain (see Section V-B, Paragraph e).

E. Minimizing $\bar{\epsilon}'_{bias}$

The bias error $\bar{\epsilon}'_{\text{bias}}$ in (15) vanishes when the total variation distance between the target importance $\boldsymbol{\alpha}$ and the biased importance $\boldsymbol{p}$ is zero, i.e., when $q_k \propto \alpha_k/\pi_k, \forall k \in \mathcal{K}$. Then, after excluding the clients that contribute the most to the optimization error and particularly slow down the convergence (Guidelines A and B), we can assign to the remaining clients an aggregation weight inversely proportional to their availability, such that the bias error $\bar{\epsilon}'_{\text{bias}}$ is minimized.

Guideline C: to minimize the bias error, we assign $q_k^* \propto \alpha_k/\pi_k$ *to the clients not excluded by the previous guidelines.*

IV. Proposed Algorithm

Guidelines A and B in Section III suggest that minimizing $\bar{\epsilon}_{\text{opt}}$ can lead to the exclusion of some available clients from the aggregation step (3), in particular those with low availability and/or high correlation. For the remaining clients, Guideline C proposes setting their aggregation weight inversely proportional to their availability to reduce the bias error $\bar{\epsilon}'_{\text{bias}}$. Motivated by these insights, we propose `CA-Fed`, a client aggregation strategy that considers the problem of correlated client availability in FL, described in Algorithm 1. `CA-Fed` learns during training which clients to exclude and how to set the aggregation weights of the remaining clients to achieve a good trade-off between $\bar{\epsilon}_{\text{opt}}$ and $\bar{\epsilon}'_{\text{bias}}$. While Guidelines A and B indicate which clients to remove, the exact number of clients to remove at round t is identified by minimizing $\epsilon^{(t)}$ as a proxy for the bounds in (10) and (15):

$$\epsilon^{(t)} := \underbrace{F_B(\boldsymbol{w}_{t,0}) - F_B^*}_{\epsilon_{\text{opt}}} + 4\bar{\kappa}^2 \cdot \underbrace{d_{TV}^2(\boldsymbol{\alpha}, \boldsymbol{p})\Gamma'}_{\bar{\epsilon}'_{\text{bias}}}, \tag{16}$$

where $\bar{\kappa}^2 \geq 0$ is a hyper-parameter that weights the relative importance of the optimization and bias error (see Sec. IV-C).

A. CA-Fed's core steps

At each communication round t, the server sends the current model $\boldsymbol{w}_{t,0}$ to all active clients and each client k sends back a noisy estimate $F_k^{(t)}$ of the current loss computed on a batch of samples $\mathcal{B}_{t,0}^k$, i.e., $F_k^{(t)} = \frac{1}{|\mathcal{B}_{t,0}^k|}\sum_{\xi\in\mathcal{B}_{t,0}^k} f(\boldsymbol{w}_{t,0},\xi)$ (line 3). The server uses these values and the information about the current set of available clients $\mathcal{A}_t$ to refine its own estimates of each client's loss ($\hat{\boldsymbol{F}}^{(t)} = (\hat{F}_k^{(t)})_{k\in\mathcal{K}}$), and each client's loss minimum value ($\hat{\boldsymbol{F}}^* = (\hat{F}_k^*)_{k\in\mathcal{K}}$), as well as of Γ', π_k, $\lambda(\boldsymbol{P}_k)$, and $\epsilon^{(t)}$, denoted as $\hat{\Gamma}'^{(t)}$, $\hat{\pi}_k^{(t)}$, $\hat{\lambda}_k^{(t)}$, and $\hat{\epsilon}^{(t)}$, respectively (possible estimators are described below) (line 4).

The server decides whether excluding clients whose availability pattern exhibits high correlation (high $\hat{\lambda}_k^{(t)}$) (line 6). First, the server considers all clients in descending order of $\hat{\boldsymbol{\lambda}}^{(t)}$ (line 14), and evaluates if, by excluding them (line 17), $\hat{\epsilon}^{(t)}$ appears to be decreasing by more than a threshold $\tau \geq 0$ (line 19). Then, the server considers clients in ascending order of $\hat{\boldsymbol{\pi}}^{(t)}$, and repeats the same procedure to possibly exclude some of the clients with low availability (low $\hat{\pi}_k^{(t)}$) (lines 7).

Once the participating clients (those with $q_k > 0$) have been selected, the server notifies them to proceed updating the current models (lines 9–10) according to (2), while the other available clients stay idle. Finally, model's updates are aggregated according to (3) (line 12).

Algorithm 1: CA-Fed (Correlation-Aware FL)

Input : $\boldsymbol{w}_{0,0}$, $\boldsymbol{\alpha}$, $\boldsymbol{q}^{(0)}$, $\{\eta_t\}_{t=1}^T$, $\bar{\eta}$, E, $\bar{\kappa}^2$, β, τ

1 Initialize $\hat{\boldsymbol{F}}^{(0)}$, $\hat{\boldsymbol{F}}^*$, $\hat{\Gamma}'^{(0)}$, $\hat{\boldsymbol{\pi}}^{(0)}$, and $\hat{\boldsymbol{\lambda}}^{(0)}$;
2 **for** $t = 1, \dots, T$ **do**
3 Receive set of active client $\mathcal{A}_t$, loss vector $\boldsymbol{F}^{(t)}$;
4 Update $\hat{\boldsymbol{F}}^{(t)}$, $\hat{\Gamma}'^{(t)}$, $\hat{\boldsymbol{\pi}}^{(t)}$, and $\hat{\boldsymbol{\lambda}}^{(t)}$;
5 Initialize $\boldsymbol{q}^{(t)} = \frac{\boldsymbol{\alpha}}{\hat{\boldsymbol{\pi}}^{(t)}}$;
6 $\boldsymbol{q}^{(t)} \leftarrow$ `get` $(\boldsymbol{q}^{(t)}, \boldsymbol{\alpha}, \hat{\boldsymbol{F}}^{(t)}, \hat{\boldsymbol{F}}^*, \hat{\Gamma}'^{(t)}, \hat{\boldsymbol{\pi}}^{(t)}, \hat{\boldsymbol{\lambda}}^{(t)})$;
7 $\boldsymbol{q}^{(t)} \leftarrow$ `get` $(\boldsymbol{q}^{(t)}, \boldsymbol{\alpha}, \hat{\boldsymbol{F}}^{(t)}, \hat{\boldsymbol{F}}^*, \hat{\Gamma}'^{(t)}, \hat{\boldsymbol{\pi}}^{(t)}, -\hat{\boldsymbol{\pi}}^{(t)})$;
8 **for** *client* $\{k \in \mathcal{A}_t;\ q_k^{(t)} > 0\}$, *in parallel* **do**
9 **for** $j = 0, \dots, E-1$ **do**
10 $\boldsymbol{w}_{t,j+1}^k = \boldsymbol{w}_{t,j}^k - \eta_t \nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k)$;
11 $\Delta_t^k \leftarrow \boldsymbol{w}_{t,E} - \boldsymbol{w}_{t,0}$;
12 $\boldsymbol{w}_{t+1,0} \leftarrow \mathbf{Proj}_W(\boldsymbol{w}_{t,0} + \bar{\eta}\sum_{k\in\mathcal{A}_t} q_k^{(t)} \cdot \Delta_t^k)$;
13 **Function** `get` $(\boldsymbol{q}, \boldsymbol{\alpha}, \boldsymbol{F}, \boldsymbol{F}^*, \Gamma, \boldsymbol{\pi}, \boldsymbol{\rho})$ **:**
14 Sort $\mathcal{K}$ by descending order in $\boldsymbol{\rho}$;
15 $\hat{\epsilon} \leftarrow \langle \boldsymbol{F} - \boldsymbol{F}^*, \boldsymbol{\pi}\tilde{\odot}\boldsymbol{q}\rangle + 4\bar{\kappa}^2 \cdot d_{TV}^2(\boldsymbol{\alpha}, \boldsymbol{\pi}\tilde{\odot}\boldsymbol{q})\Gamma$;
16 **for** $k \in \mathcal{K}$ **do**
17 $q_k^+ \leftarrow 0$;
18 $\hat{\epsilon}^+ \leftarrow \langle \boldsymbol{F} - \boldsymbol{F}^*, \boldsymbol{\pi}\tilde{\odot}\boldsymbol{q}^+\rangle + 4\bar{\kappa}^2 \cdot d_{TV}^2(\boldsymbol{\alpha}, \boldsymbol{\pi}\tilde{\odot}\boldsymbol{q}^+)\Gamma$;
19 **if** $\hat{\epsilon} - \hat{\epsilon}^+ \geq \tau$ **then**
20 $\hat{\epsilon} \leftarrow \hat{\epsilon}^+$;
21 $\boldsymbol{q} \leftarrow \boldsymbol{q}^+$;
22 **return** $\boldsymbol{q}$

B. Estimators

We now briefly discuss possible implementation of the estimators $\hat{F}_k^{(t)}$, $\hat{F}_k^*$, $\hat{\Gamma}'^{(t)}$, $\hat{\pi}_k^{(t)}$, and $\hat{\lambda}_k^{(t)}$. Server's estimates for the clients' local losses ($\hat{\boldsymbol{F}}^{(t)} = (\hat{F}_k^{(t)})_{k\in\mathcal{K}}$) can be obtained from the received active clients' losses ($\boldsymbol{F}^{(t)} = (F_k^{(t)})_{k\in\mathcal{A}_t}$) through an auto-regressive filter with parameter $\beta \in (0, 1]$:

$$\hat{\boldsymbol{F}}^{(t)} = (\mathbf{1} - \beta\mathbb{1}_{\mathcal{A}_t}) \odot \hat{\boldsymbol{F}}^{(t-1)} + \beta\mathbb{1}_{\mathcal{A}_t} \odot \boldsymbol{F}^{(t)}, \tag{17}$$

where $\odot$ denotes the component-wise multiplication between vectors, and $\mathbb{1}_{\mathcal{A}_t}$ is a N-dimensions binary vector whose k-th component equals 1 if and only if client k is active at round t, i.e., $k \in \mathcal{A}_t$. The server can estimate client-k's loss minimum value F_k^* as $\hat{F}_k^* = \min_{s\in[0,t]} \hat{F}_k^{(s)}$. The values of $F_B(\boldsymbol{w}_{t,0})$, F_B^*, Γ', and $\epsilon^{(t)}$ can be estimated as follows:

$$\hat{F}_B^{(t)} - \hat{F}_B^* = \langle \hat{\boldsymbol{F}}^{(t)} - \hat{\boldsymbol{F}}^*, \hat{\boldsymbol{\pi}}^{(t)}\tilde{\odot}\boldsymbol{q}^{(t)}\rangle, \tag{18}$$

$$\hat{\Gamma}'^{(t)} = \max_{k\in\mathcal{K}}(\hat{F}_k^{(t)} - \hat{F}_k^*), \tag{19}$$

$$\hat{\epsilon}^{(t)} = \hat{F}_B^{(t)} - \hat{F}_B^* + 4\bar{\kappa}^2 \cdot d_{TV}^2(\boldsymbol{\alpha}, \hat{\boldsymbol{\pi}}^{(t)}\tilde{\odot}\boldsymbol{q}^{(t)})\hat{\Gamma}'^{(t)}. \tag{20}$$

where $\boldsymbol{\pi}\tilde{\odot}\boldsymbol{q} \in \mathbb{R}^N$, such that $\left(\boldsymbol{\pi}\tilde{\odot}\boldsymbol{q}\right)_k := \frac{\pi_k q_k}{\sum_{h=1}^N \pi_h q_h}$, $k \in \mathcal{K}$.

For $\hat{\pi}_k^{(t)}$, the server can simply keep track of the total number of times client k was available up to time t and compute $\hat{\pi}_k^{(t)}$ using a Bayesian estimator with beta prior, i.e., $\hat{\pi}_k^{(t)} = (\sum_{s\leq t} \mathbb{1}_{k\in A_s} + n_k)/(t + n_k + m_k)$, where n_k and m_k are the initial parameters of the beta prior.

For $\hat{\lambda}_k^{(t)}$, the server can assume the client's availability evolves according to a Markov chain with two states (active and inactive), track the corresponding number of state transitions, and estimate the transition matrix $\hat{\boldsymbol{P}}_k^{(t)}$ through a Bayesian estimator similarly to what done for $\hat{\pi}_k^{(t)}$. Finally, $\hat{\lambda}_k^{(t)}$ is obtained computing the eigenvalues of $\hat{\boldsymbol{P}}_k^{(t)}$.

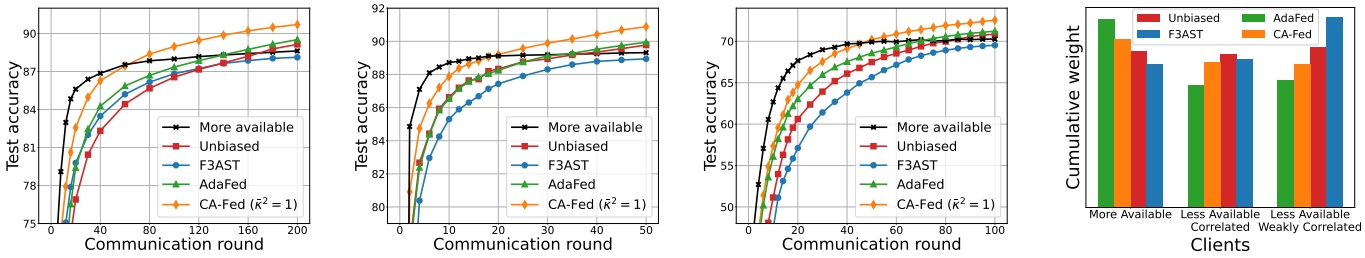
C. The role of the hyper-parameter $\bar{\kappa}^2$

Theorems 1 and 3 suggest that the condition number κ^2 has a significant impact on the minimization of the total error ϵ. Our algorithm uses a proxy ($\epsilon^{(t)}$) for the total error (see (16)). To account for the effect of κ^2, we introduced the hyper-parameter $\bar{\kappa}^2 \geq 0$, which weights the relative importance of the optimization and bias error in (16). In practice, $\bar{\kappa}^2$ controls the number of excluded clients by CA-Fed. A small value of $\bar{\kappa}^2$ penalizes the bias term in favor of the optimization error, resulting in a larger number of excluded clients. Conversely, the bias term dominates for large values of $\bar{\kappa}^2$, and CA-Fed tends to include more clients. Asymptotically, for $\bar{\kappa}^2 \to \infty$, CA-Fed reduces to the *unbiased aggregation strategy*.

V. Experimental Evaluation

A. Experimental Setup

a) Federated system simulator: In our experiments, we consider a population of $N = |\mathcal{K}| = 100$ clients. We model the activity of each client $k \in \mathcal{K}$ as a two-state homogeneous Markov process with state space $\mathcal{S} = \{$"*active*", "*inactive*"$\}$, characterized by a transition matrix $\boldsymbol{P}_k$, a stationary distribution $\boldsymbol{\pi}^{(k)}$, and a second largest absolute eigenvalue $\bar{\lambda}_2(\boldsymbol{P}_k)$ (see Appendix F3 for details). Our goal is to simulate realistic dynamics of federated systems featuring varying levels of



(a) Synthetic (b) MNIST (c) CIFAR-10 (d) Cumulative importance

Fig. 1: Average test accuracy among $N = 100$ clients achieved by the algorithms on the Synthetic, MNIST, and CIFAR-10 datasets. Cumulative importance assigned by the algorithms to the clients after $T = 200$ rounds on the Synthetic dataset.

clients' availability and correlation. To introduce heterogeneity in clients' availability patterns, we divide the population in two equally-sized classes: the "more available" clients with a steady-state probability of being active $\pi_{k,\text{active}} = 1/2+g$, and the "less available" clients with $\pi_{k,\text{active}} = 1/2 - g$. Here, the parameter $g \in (0, 1/2)$ controls the degree of heterogeneity in clients' availability. We furthermore divide each class of clients in two equally-sized sub-classes: clients exhibiting a largely correlated time behavior (in the following referred to as "correlated" clients) that tend to persist in the same state for rather long periods ($\lambda_k = \nu$ with values of ν close to 1), and clients exhibiting a weakly correlated time behavior (referred to as "weakly correlated" clients) that are almost as likely to keep as to change their state at every t ($\lambda_k \sim \mathcal{N}(0, \varepsilon^2)$, with ε close to 0). We use $g = 0.4$, $\nu = 0.9$, and $\varepsilon = 10^{-2}$.

b) Datasets and models: We conduct experiments on the LEAF Synthetic dataset [36], a benchmark for multinomial classification tasks, and on the real-world MNIST [37] and CIFAR-10 [38] datasets, respectively for handwritten digits and image recognition tasks. To simulate the statistical heterogeneity present in the federated learning system, we use common approaches in the literature. For the Synthetic dataset, we tune the parameters (γ, δ), which control data heterogeneity among clients [23]. For MNIST and CIFAR-10, we distribute samples from the same class across the clients according to a symmetric Dirichlet distribution with parameter ς, following the same approach as [39]. Unless otherwise indicated, we set $\gamma = \delta = \varsigma = 0.5$. We use the original training/test data split of MNIST and reserve 20% of the training dataset as the validation dataset. For Synthetic and MNIST, we use a linear classifier with a ridge penalization of parameter 10^{-2}, which corresponds to a strongly convex objective function. For CIFAR-10, we use a neural network with two convolutional and one fully connected layers.

c) Benchmarks: We compare `CA-Fed`, defined in Algorithm 1, with four baselines including two state-of-the-art FL algorithms discussed in Section II: 1) `Unbiased`, which aggregates the active clients $k \in \mathcal{A}_t$ with weights $q_k = \alpha_k/\pi_k$; 2) `More available`, which considers only the "more available" clients and always excludes the "less available" ones; 3) `AdaFed` [20], which, similarly to `Unbiased`, aggregates all active clients, but normalizes their aggregation weights (i.e., it considers $q_k = \frac{\alpha_k/\pi_k}{\sum_{k\in\mathcal{A}_t} \alpha_k/\pi_k}$); 4) `F3AST` [19], which, oppositely to `More available`, favors the "less available" clients. For all algorithms, we tuned the learning rates η, $\bar{\eta}$ via grid search. For `CA-Fed`, we use $\beta = \tau = 0$. Unless otherwise specified, we assume that the algorithms can access an oracle providing the true availability parameters for each client: in practice, all the algorithms rely on the exact knowledge of $\pi_{k,\text{active}}$; in addition, `CA-Fed` also receives $\lambda(\boldsymbol{P}_k)$. In Section V-B, Paragraph d, we will relax this assumption by considering the estimators $\hat{\pi}_k^{(t)}$ and $\hat{\lambda}_k^{(t)}$. The code for this paper is available at: https://github.com/arodio/CA-Fed.

B. Experimental Results

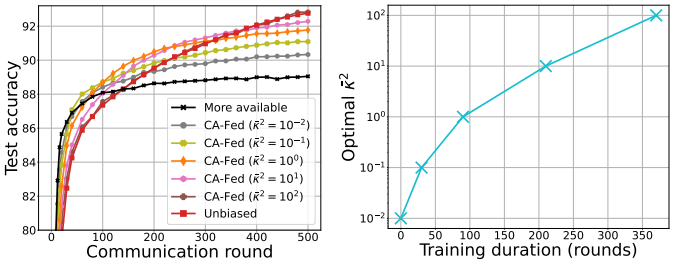
a) `CA-Fed` vs. baselines: Figure 1 compares the test accuracy achieved by `CA-Fed` ($\bar{\kappa}^2 = 1$) and the baselines on the Synthetic (Fig. 1a), MNIST (Fig. 1b), and CIFAR-10 (Fig. 1c) datasets over 10 different runs. Across all three datasets, `CA-Fed` consistently outperforms the baselines, achieving higher test accuracy (+1.56 pp on Synthetic; +0.94 pp on MNIST; +1.32 pp on CIFAR-10) compared to the second best performing method, `AdaFed`. These results demonstrate that `CA-Fed` achieves the best balance between convergence speed and test accuracy. For deeper insights into the algorithms' behavior, Figure 1d illustrates the cumulative aggregation weights $\{\frac{1}{T}\sum_{t=1}^{T} q_k^{(t)}\}_{k\in\mathcal{K}}$, representing the cumulative importance that the algorithms assigned to the clients at the end of the training. In Figure 1d, we grouped the clients into three categories: "more available", "less available, weakly correlated", and "less available, correlated". By setting the aggregation weights inversely proportional to the clients' availabilities, `Unbiased` equalizes the importance for all clients (see Fig. 1d), but achieves a slower convergence (as shown in Figs. 1a, 1b, and 1c). On the contrary, by excluding all the "less available" clients, `More available` achieves a faster convergence but introduces a non-vanishing bias error ϵ_{bias}, which, in practice, leads to poor accuracy performance. The state-of-the-art algorithm `AdaFed`, similarly to `Unbiased`, considers all the active clients, but normalizes their aggregation weights at each communication round. As a result, similarly to `CA-Fed`, `AdaFed` indeed prioritizes the "more available" clients (as shown in Fig. 1d), and then a convergence speed-up could be expected. However, `AdaFed`

does not exclude the "less available and correlated" clients, and therefore their presence causes a convergence slowdown. Finally, F3AST favors the "less available, correlated" clients and achieves a slower convergence with a non-vanishing bias error, which corresponds to lower accuracy performance. By opportunely excluding some of the "less available and correlated" clients, CA-Fed achieves the best test accuracy by the end of the training time.

b) Convergence speed vs. Bias error: The trade-off between ϵ_{opt} or ϵ_{bias} discussed in Section III is visible in our experiments. In particular, Figure 2a compares the test accuracy achieved by More available, Unbiased, and CA-Fed on the Synthetic dataset for $T = 500$ communication rounds. As expected, by targeting the minimization of ϵ_{opt} and thus excluding the "less available" clients, More available achieves the fastest convergence at the expense of a large non-vanishing bias error ϵ_{bias}. On the other hand, by targeting the minimization of ϵ_{bias} and thus equalizing the clients' importance, Unbiased asymptotically removes this error and ultimately achieves the highest test accuracy at communication round $T = 500$, but suffers from slower convergence due to the presence of the "correlated" clients. Our algorithm, CA-Fed, leverages the trade-off between convergence speed and model bias and achieves fast convergence to the neighborhood of the target objective. To explore this trade-off, in Figure 2a, we varied the value of the hyper-parameter $\bar{\kappa}^2$ in the range $\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$. CA-Fed tends to exclude more clients for low values of $\bar{\kappa}^2$ and achieves a similar convergence rate as More available for $\bar{\kappa}^2 = 10^{-2}$. For intermediate values of $\bar{\kappa}^2$, CA-Fed trades a small accuracy decrease for faster convergence (refer, for example, to the curves $\bar{\kappa}^2 = 10^{0}, 10^{1}$). For $\bar{\kappa}^2 = 10^{2}$, CA-Fed reduces to Unbiased (their curves overlap in Fig. 2a). Moreover, we observe that the optimal value of $\bar{\kappa}^2$ depends on the available time for training. Low values of $\bar{\kappa}^2$ speed-up convergence and then they can be beneficial for short training durations (e.g., CA-Fed ($\bar{\kappa} = 10^{-1}$) achieves a higher test accuracy of +2.8 pp with respect to Unbiased at communication round $t = 40$). For longer training periods, a larger value of $\bar{\kappa}^2$ may be preferable as it reduces the bias error and increases the test accuracy (e.g., CA-Fed ($\bar{\kappa} = 10^{2}$) improves of +3.8 pp with respect to More available at communication round $t = 500$). Figure 2b illustrates the optimal value of $\bar{\kappa}^2$ for different durations of the training period T.

c) Effect of statistical heterogeneity: The bias error bounds $\bar{\epsilon}_{\text{bias}}$ and $\bar{\epsilon}'_{\text{bias}}$ in Theorems 1 and 3 are influenced by the degree of heterogeneity among local functions, commonly known as *statistical heterogeneity*, characterized by the constants Γ and Γ' in (11) and (15), respectively. To control statistical heterogeneity, we manipulate the dissimilarity among the clients' local datasets, specifically through the parameters γ and δ in the case of the Synthetic dataset, as explained in Section V-A. Figure 3 illustrates the impact of γ and δ on the test accuracy achieved by CA-Fed after $T = 200$ communication rounds on the Synthetic dataset. As expected, in the extreme IID setting (when $\gamma = \delta = 0$), Γ and Γ' are small, and the bias error ϵ_{bias} is negligible. As a result, More available and CA-Fed ($\bar{\kappa}^2 = 10^{-2}$) reach the highest test accuracy, whereas CA-Fed ($\bar{\kappa}^2 = 10^{2}$) and Unbiased present slow convergence. Nevertheless, More available and CA-Fed ($\bar{\kappa}^2 = 10^{-2}$) perform poorly as the statistical heterogeneity increases (i.e., $\gamma = \delta \geq 0.25$). In the extreme non-IID setting (when $\gamma = \delta = 1$), Γ and Γ' are large, and ϵ_{bias} dominates. In this case, CA-Fed ($\bar{\kappa}^2 = 10^{2}$) and Unbiased should be preferred. For $\gamma = \delta = \{0.25, 0.5, 0.75\}$, CA-Fed (with $\bar{\kappa}^2 = 1$ or $\bar{\kappa}^2 = 10$) achieves the highest test accuracy (+1.6 pp, +1.2 pp, and +1.0 pp with respect to Unbiased).

d) Estimation of the clients' availability and correlation: In this experiment, CA-Fed utilizes estimators $\hat{\pi}_k^{(t)}$ and $\hat{\lambda}_k^{(t)}$ to estimate the clients' π_k and λ_k values. We employ a Bayesian estimator with a beta prior to estimate $\hat{\boldsymbol{P}}_k^{(t)}$, which we generate by observing the evolution of the Markov



(a) Test accuracy (b) Optimal $\bar{\kappa}^2$

Fig. 2: *Convergence speed vs. Model bias* trade-off for different values of $\bar{\kappa}^2$ on the Synthetic dataset, for $\gamma = \delta = 0.5$.

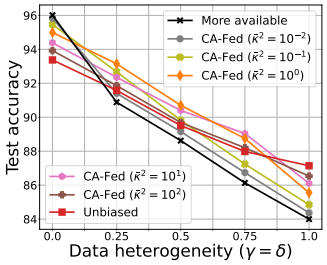
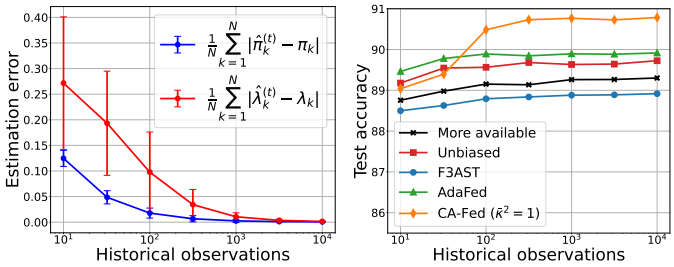
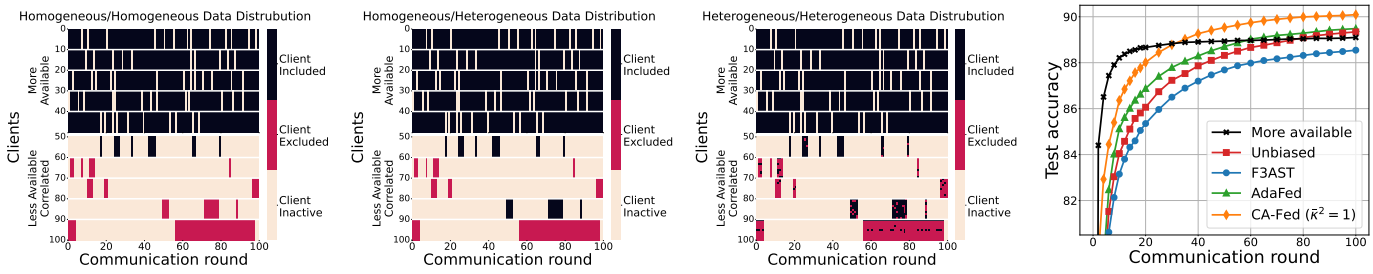


Fig. 3: Effects of *data heterogeneity* on the Synthetic dataset after $T = 200$ rounds.



(a) Estimation error (b) Test accuracy

Fig. 4: Estimation of the *clients' activities* $(\hat{\pi}_k^{(t)}, \hat{\lambda}_k^{(t)})$ for different priors $t \in \{10^{1}, 10^{1.5}, 10^{2}, 10^{2.5}, 10^{3}, 10^{3.5}, 10^{4}\}$ and test accuracy after $T = 50$ rounds on the MNIST dataset.



(a) Homogeneous/Homogeneous (b) Homogeneous/Heterogeneous (c) Heterogeneous/Heterogeneous (d) Test accuracy

Fig. 5: Clients' activities and CA-Fed's inclusion/exclusion decisions in the presence of *spatial correlation* for different degrees of *intra-cluster/inter-cluster* data distributions. Average test accuracy after $T = 100$ rounds on the MNIST dataset.

chain defined by $\boldsymbol{P}_k$ over t' time-steps. We compute $\hat{\pi}_k^{(t)}$ and $\hat{\lambda}_k^{(t)}$ analytically, following the methodology explained in Section IV-B and described in detail in Appendix F3. Figure 4a shows the estimation errors $\frac{1}{N}\sum_{k\in\mathcal{K}}|\hat{\pi}_k^{(t)} - \pi_k|$ and $\frac{1}{N}\sum_{k\in\mathcal{K}}|\hat{\lambda}_k^{(t)} - \lambda_k|$ as a function of the number of historical observations t'. As expected, both errors decrease with an increasing number of observations, and the estimation error for λ_k is larger than that for π_k. Furthermore, Figure 4b compares the final test accuracy obtained by CA-Fed and the baselines for varying numbers of historical observations $t' \in \{10^1, 10^{1.5}, 10^2, 10^{2.5}, 10^3, 10^{3.5}, 10^4\}$ when training for $T = 50$ rounds on the MNIST dataset. In this setting, CA-Fed outperforms the baselines for $t' \geq 100$. This value is reasonable, because estimating λ_k requires a number of observations comparable to the expected hitting time for the slowest Markov chain, which is given by $\max_{k\in\mathcal{K}} \frac{1}{(1-\lambda_k)\pi_k} = 100$.

e) CA-Fed with Spatial Correlation: Although CA-Fed is primarily designed to handle temporal correlation (as discussed in Section III-D), we also evaluate its performance in the presence of spatial correlation. In the considered spatially correlated scenario, clients are grouped into clusters, and each cluster $c \in \mathcal{C}$ is characterized by an underlying Markov chain that determines when all clients in the cluster are available or unavailable. The Markov chains of different clusters are independent. Let λ_c denote the second-largest eigenvalue in magnitude of cluster c's Markov chain. To reduce the eigenvalue of the aggregate Markov chain, CA-Fed needs to exclude all clients in the cluster $\bar{c} = \arg\max_{c\in\mathcal{C}} \lambda_c$. In this experiment, we consider a population of $N = 100$ clients grouped into $|\mathcal{C}| = 10$ clusters. We equally split the clients, or equivalently, the clusters, into two categories: "more available" with $\pi_c = 0.9$ and $\lambda_c = 0$ for $c = 0, \ldots, 4$, and "less available, correlated" with $\pi_c = 0.1$ and $\lambda_c = c/10$ for $c = 5, \ldots, 9$. In Figures 5a, 5b, and 5c, each pixel represents, for each client $k \in \mathcal{K}$ and for each communication round, the client's activity (active/inactive) and CA-Fed's decision (included/excluded in training). From the experiments, we observe that CA-Fed's decisions depend on the degree of statistical heterogeneity among clients within a cluster (i.e., *intra-cluster*) and among clusters (i.e., *inter-cluster*). When both the intra-cluster and inter-cluster clients' data distributions are homogeneous, CA-Fed starts considering the clients in cluster $\bar{c} = 9$ with $\lambda_{\bar{c}} = 0.9$, and sequentially excludes, in order, all clients from clusters $\{9, 8, 7, 6\}$ (as shown in Fig. 5a). When the clients' data distributions are homogeneous within clusters, but heterogeneous among clusters (Fig. 5b), CA-Fed still excludes all clients from clusters $c = \{9, 7, 6\}$, but decides to include clients from cluster $c = 8$. This is because these clients happen to have a lower value of $\hat{F}_k^{(t)} - \hat{F}_k^*$, and despite having a large λ_c, CA-Fed decides to include them. Finally, when both the intra-cluster and inter-cluster clients' data distributions are heterogeneous (Fig. 5c), CA-Fed can partially include clients from the more correlated clusters, even though their λ_c is large. Figure 5d compares the test accuracy achieved by CA-Fed and the baselines with spatial correlation in the same setting as in Figure 5c. The experimental results show that CA-Fed can operate correctly in the presence of spatial correlation and still outperforms the baselines (+0.6 pp w.r.t. AdaFed).

VI. Conclusion

This paper presents the first convergence analysis of a FedAvg-like federated learning (FL) algorithm in presence of heterogeneous and correlated client availability. The analysis reveals the detrimental effect of correlation on the convergence rate and highlights a fundamental trade-off between convergence speed and model bias. To navigate this tradeoff, we introduce CA-Fed, a novel FL algorithm, which adaptively manages the conflicting aims of enhancing convergence speed and reducing model bias, with the ultimate objective of maximizing model quality within the constraints of the training time available. CA-Fed achieves this goal by dynamically excluding clients who exhibit high temporal correlation and limited availability, contingent on their data distributions. Indeed, model updates from such clients may act as noise, increasing variance and slowing down the algorithm's convergence. CA-Fed disregards such clients unless their local datasets notably enhance the quality of the final model. The experimental results validate the effectiveness of our strategy, demonstrating that CA-Fed is a versatile and resilient FL algorithm, well-suited to address real-world scenarios characterized by heterogeneous and correlated client availability. Further discussions on the computation and communication costs, and fairness of CA-Fed can be found in Appendix H.

REFERENCES

[1] A. Rodio, F. Faticanti, O. Marfoq, G. Neglia, and E. Leonardi, "Federated Learning under Heterogeneous and Correlated Client Availability," in *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, May 2023, pp. 1–10.

[2] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen *et al.*, "A Survey on Distributed Machine Learning," *ACM Computing Surveys*, vol. 53, no. 2, pp. 30:1–30:33, Mar. 2020.

[3] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya *et al.*, "When Edge Meets Learning: Adaptive Control for Resource-Constrained Distributed Machine Learning," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, Apr. 2018, pp. 63–71.

[4] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh *et al.*, "Federated Learning: Strategies for Improving Communication Efficiency," in *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[5] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR, Apr. 2017, pp. 1273–1282.

[6] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis *et al.*, "Advances and Open Problems in Federated Learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, Jun. 2021.

[7] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, May 2020.

[8] H. Eichner, T. Koren, B. Mcmahan, N. Srebro, and K. Talwar, "Semi-Cyclic Stochastic Gradient Descent," in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, May 2019, pp. 1764–1773.

[9] J. Wang, Z. Charles, Z. Xu, G. Joshi, H. B. McMahan *et al.*, "A Field Guide to Federated Optimization," *arXiv:2107.06917*, Jul. 2021.

[10] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman *et al.*, "Towards Federated Learning at Scale: System Design," *Proceedings of Machine Learning and Systems*, vol. 1, pp. 374–388, Apr. 2019.

[11] Y. Ding, C. Niu, Y. Yan, Z. Zheng, F. Wu *et al.*, "Distributed Optimization over Block-Cyclic Data," *arXiv:2002.07454*, Feb. 2020.

[12] C. Zhu, Z. Xu, M. Chen, J. Konečnỳ, A. Hard *et al.*, "Diurnal or Nocturnal? Federated Learning from Periodically Shifting Distributions," in *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[13] T. T. Doan, "Local Stochastic Approximation: A Unified View of Federated Learning and Distributed Multi-Task Reinforcement Learning Algorithms," *arXiv:2006.13460*, Jun. 2020.

[14] T. T. Doan, L. M. Nguyen, N. H. Pham, and J. Romberg, "Convergence Rates of Accelerated Markov Gradient Descent with Applications in Reinforcement Learning," *arXiv:2002.02873*, Oct. 2020.

[15] T. Sun, Y. Sun, and W. Yin, "On Markov Chain Gradient Descent," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.

[16] M. McCloskey and N. J. Cohen, "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem," in *Psychology of Learning and Motivation*, G. H. Bower, Ed. Academic Press, 1989, vol. 24, pp. 109–165.

[17] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins *et al.*, "Overcoming Catastrophic Forgetting in Neural Networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017.

[18] M. Tang, X. Ning, Y. Wang, J. Sun, Y. Wang *et al.*, "FedCor: Correlation-Based Active Client Selection Strategy for Heterogeneous Federated Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[19] M. Ribero, H. Vikalo, and G. de Veciana, "Federated Learning Under Intermittent Client Availability and Time-Varying Communication Constraints," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 1, pp. 98–111, Jan. 2023.

[20] L. Tan, X. Zhang, Y. Zhou, X. Che, M. Hu *et al.*, "AdaFed: Optimizing Participation-Aware Federated Learning with Adaptive Aggregation Weights," *IEEE Transactions on Network Science and Engineering*, 2022.

[21] A. Nichol, J. Achiam, and J. Schulman, "On First-Order Meta-Learning Algorithms," *arXiv:1803.02999*, Oct. 2018.

[22] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush *et al.*, "Adaptive Federated Optimization," in *International Conference on Learning Representations*, 2021.

[23] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the Convergence of FedAvg on Non-IID Data," in *International Conference on Learning Representations*, 2019.

[24] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar *et al.*, "Federated Optimization in Heterogeneous Networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, Mar. 2020.

[25] W. Chen, S. Horváth, and P. Richtárik, "Optimal Client Sampling for Federated Learning," *Transactions on Machine Learning Research*, Aug. 2022.

[26] Y. Fraboni, R. Vidal, L. Kameni, and M. Lorenzi, "Clustered Sampling: Low-Variance and Improved Representativity for Clients Selection in Federated Learning," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 3407–3416.

[27] Y. Jee Cho, J. Wang, and G. Joshi, "Towards Understanding Biased Client Selection in Federated Learning," in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 10 351–10 375.

[28] T. T. Doan, L. M. Nguyen, N. H. Pham, and J. Romberg, "Finite-Time Analysis of Stochastic Gradient Descent under Markov Randomness," *arXiv:2003.10973*, Apr. 2020.

[29] A. Meyers and H. Yang, "Markov Chains for Fault-Tolerance Modeling of Stochastic Networks," *IEEE Transactions on Automation Science and Engineering*, 2021.

[30] H. Olle, P. Yuval, and E. S. Jeffrey, "Dynamical Percolation," in *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, vol. 33, no. 4. Elsevier, 1997, pp. 497–528.

[31] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times: Second Edition*. American Mathematical Soc., 2017, vol. 107.

[32] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization Methods for Large-Scale Machine Learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018.

[33] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 7611–7623.

[34] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks," *arXiv:1312.6211*, Mar. 2015.

[35] R. Kemker, M. McClure, A. Abitino, T. Hayes, and C. Kanan, "Measuring Catastrophic Forgetting in Neural Networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.

[36] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečný *et al.*, "LEAF: A Benchmark for Federated Settings," *arXiv:1812.01097*, Dec. 2019.

[37] L. Deng, "The MNIST Database of Handwritten Digit Images for Machine Learning Research," *IEEE Signal Processing Magazine*, 2012.

[38] A. Krizhevsky and G. Hinton, "Learning Multiple Layers of Features from Tiny Images," University of Toronto, Toronto, Tech. Rep., 2009.

[39] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated Learning with Matched Averaging," in *International Conference on Learning Representations*, 2020.

[40] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, Mar. 2004.

[41] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. SIAM, 2001.

[42] H. Ludwig and N. Baracaldo, *Federated Learning: A Comprehensive Overview of Methods and Applications*. Springer Cham, 2022.

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:

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

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

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

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

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 $\gamma_t \coloneqq 2\eta_t(1-\eta_t L(1+2EQ)) > 0$, 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)].

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

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

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

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

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

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

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

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].

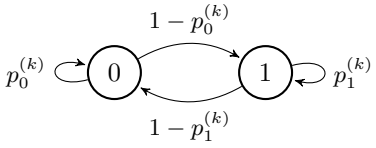
Proof. From Proposition 5, we know that the eigenvalues of $\boldsymbol{P} = \bigotimes_{k\in\mathcal{K}} \boldsymbol{P}_k$ are given by:

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

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.