UNDERSTANDING THE BIAS OF MARKOV CHAIN SGD WITH INTERMITTENT NODE AVAILABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies stochastic gradient descent when the sampling process follows a Markov chain (Markov chain SGD). Specifically, we consider a token-based setting for distributed learning, where a token carrying the shared model moves sequentially among the nodes according to a Markov chain, and the node holding the token performs one gradient update. In practical distributed systems, nodes may be intermittently unavailable because of resource constraints. We analyze the convergence of Markov chain SGD with intermittent node availability. The derived upper and lower bounds of convergence rates show that different node availability rates can cause a non-vanishing optimization bias with respect to the intended objective, and that correction weights can correct for the bias at the cost of slower convergence. Experiments on quadratic functions and neural networks support our theoretical findings.

1 INTRODUCTION

In this work, we consider the general optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \coloneqq \mathbb{E}_{s\sim\pi}[f_s(\mathbf{x})], \tag{1}$$

where d denotes the dimension of the parameter vector, π denotes a probability distribution over a set $\mathcal{S}$ and each $f_s : \mathbb{R}^d \to \mathbb{R}$ is assumed to be differentiable. One typical way to solve this problem is to apply Stochastic Gradient Descent (SGD) algorithms; that is,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \nabla f_{s_t}(\mathbf{x}_t),$$

where γ denotes the step size, and t denotes the index of the iteration. When the states $\{s_t\}_{t\geq 0}$ are independent and identically distributed (i.i.d.), the algorithm is called classical SGD (Bottou et al., 2018). When $\{s_t\}$ are generated by a Markov chain, whose stationary distribution is π, the algorithm is called Markov chain SGD (Markovian SGD) (Johansson et al., 2010; Even, 2023). Markovian SGD has been extensively studied across multiple fields, for example, reinforcement learning (Dorfman & Levy, 2022), and distributed learning (Even, 2023). In distributed learning, which is the setting considered in this work, Markovian SGD is a typical token-based algorithm, where a token carrying the shared model moves sequentially among the nodes according to a Markov chain, and the node holding the token performs one gradient update (Even, 2023). Accordingly, the states of Markovian SGD are also called nodes/machines/clients. Moreover, we assume that the size of the state space $\mathcal{S}$ is finite and denote its size by $|\mathcal{S}| = S$.

Node availability problem in distributed learning. In practical distributed systems, the nodes can be intermittently unavailable, possibly due to communication and computation constraints, especially on the device and edge sides. In federated learning (FL) (McMahan et al., 2017), the node availability problem has been studied extensively, for example, Yan et al. (2024) and Gu et al. (2021). Some works may assume that the node availability follows some specific patterns (Wang & Ji, 2022). However, in practical scenarios, the availability of nodes may be uncontrollable. In this case, for example, when the nodes participate in the training according to their intrinsic mechanisms, Wang & Ji (2024) showed that there is a bias between the effective optimization objective and the intended objective. In this work, we aim to investigate whether such a bias exists in Markovian SGD with intermittent node availability.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Markovian SGD with intermittent node availability. In consensus-based algorithms, the unavailability of some nodes does not affect the algorithm's execution, for example, FL (Wang & Ji, 2024) and decentralized parallel SGD (Koloskova et al., 2020; Zehtabi et al., 2025). However, in token-based algorithms, the unavailability of the node that holds the token can cause the failure of the algorithm (Chen et al., 2026). To circumvent this, we separate the tasks of each node into computation tasks and communication tasks. In each iteration, each node can be either available or unavailable for the computation task. However, all the nodes are still required to complete their communication tasks. In other words, we mainly consider intermittent computation availability, while assuming reliable communication among nodes. We call nodes that compute and communicate "active", those that only communicate "relay", and those that do neither "idle" (see Figure 1). These special relay nodes were adopted and studied in Sayed (2014, Chapter 13, Role of Informed Agents).

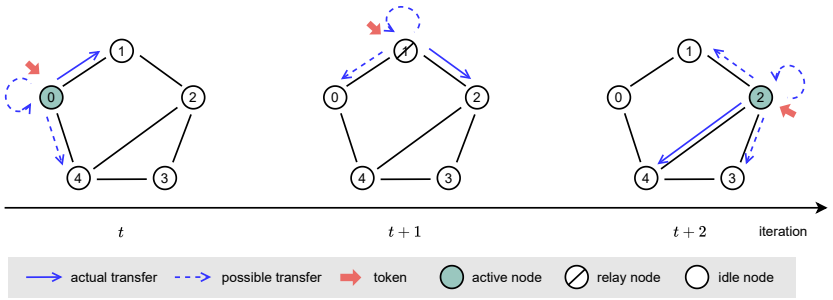


Figure 1: An illustrative example of Markovian SGD with intermittent node availability. At iteration t, node 0 holds the token ($s_t = 0$) and is available for the computation task ($\delta_t^{s_t} = 1$); it is active and others are idle. Node 0 can transfer the token to nodes 0, 1, or 4; in this example, it transfers the token to node 1. At iteration $t+1$, node 1 holds the token ($s_{t+1} = 1$) and is unavailable ($\delta_{t+1}^{s_{t+1}} = 0$); it is the relay node and others are idle. Although node 1 cannot compute, it is still required to transfer the token to nodes 0, 1, or 2. The case at iteration $t + 2$ is similar (to iteration t).

The detailed algorithm is presented in Algorithm 1, and a simple illustrative example is in Figure 1. The token (that is, the updated model parameters in our setting) is transferred according to a Markov chain (Line 2). Here, $s_t = i$ means that the token is held by node i at iteration t. The availability status is modeled by Bernoulli random variables (Line 3). If $\delta_t^{s_t} \sim \text{Bernoulli}(p_t^{s_t})$ equals 1, node s_t is available for computation at iteration t, and otherwise it is unavailable. When the node holds the token and is available for computation, it performs one step of update. When the node holds the token and is unavailable, it just passes the token to the next node. The correction weight $w_t^{s_t}$, which will be discussed in detail in Section 3.3, is used to balance the contributions of each node's update.

Algorithm 1: Markovian SGD with intermittent node availability

Input: $\mathbf{x}_0$; **Output**: $\{\mathbf{x}_t\}$
1 **for** $t = 0, 1, \ldots, T-1$ **do**
2 Sample s_t from a Markov chain
3 $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \delta_t^{s_t} w_t^{s_t} \nabla f_{s_t}(\mathbf{x}_t)$

Contributions. The main contributions are as follows:

- We derive upper bounds of convergence rates for Markovian SGD with intermittent node availability in Theorem 1, which identifies a non-vanishing availability bias term (as the total number of iterations T increases). We show that this availability bias is intrinsic by constructing a matching asymptotic lower bound in Theorem 2.
- The bounds reveal that heterogeneous node availability introduces a non-vanishing optimization bias, and correction weights can correct for the bias at the cost of slower convergence (Section 3.3). The experiments on quadratic functions and neural networks validate these theoretical findings (Section 4).
- Enabled by the permutation-based SGD analysis of Koloskova et al. (2024), our bound for Markovian SGD, unlike prior works (Even, 2023; Beznosikov et al., 2023), requires neither algorithmic modifications nor a large-iteration condition.

2 Related Works

This section reviews related works on FL with intermittent node availability and Markovian SGD.

2.1 FL with Intermittent Node Availability

The node availability (participation) problem has been studied from multiple perspectives in FL. Some works consider scenarios where node availability is highly coordinated, allowing node sampling strategies to be applied (for example, Cho et al., 2022; Chen et al., 2022). Some works consider scenarios where node availability follows some specific patterns (for example, Wang & Ji, 2022; Sun et al., 2025; Ying et al., 2025). In contrast, some works consider scenarios where node availability is uncoordinated, possibly due to heterogeneous local conditions across nodes (for example, Gu et al., 2021; Wang & Ji, 2024; Xiang et al., 2024; Zehtabi et al., 2025). In particular, Wang & Ji (2024) showed that a bias exists between the effective optimization objective and the intended objective when node availability is determined by uncoordinated node-specific processes. In this work, we show such a bias exists in Markovian SGD with intermittent node availability. Detailed comparisons are in Section 3.4.

In addition, we note that similar forms of bias were studied in other settings, beyond that induced by node availability (Wang et al., 2020; Horváth et al., 2022; Weng et al., 2026).

2.2 Markovian SGD

The upper bounds of Markovian SGD have been extensively studied (Ram et al., 2009; Johansson et al., 2010; Sun et al., 2018; Mao et al., 2020; Doan, 2023; Dorfman & Levy, 2022; Even, 2023; Beznosikov et al., 2023; Huynh et al., 2025). The lower bounds of Markovian SGD have recently been studied in Even (2023) and Sun & Wei (2025a;b). The biased gradient in Markovian SGD makes its analysis more challenging than unbiased SGD (for example, classical SGD). To the best of our knowledge, there are roughly two ways to achieve a vanishing bound with respect to T: (i) increasing the minibatch size (Dorfman & Levy, 2022; Beznosikov et al., 2023) and (ii) ensuring a sufficiently long burn-in period (Even, 2023). The former relies on an algorithmic variant of Markovian SGD (Beznosikov et al., 2023, Algorithm 2), while the latter requires T to be sufficiently large (Even, 2023, Theorem 2). In contrast, our analysis establishes a vanishing bound for Markovian SGD without requiring either algorithmic modifications or large-iteration requirements. Detailed comparisons are in Section 3.4.

3 Convergence Analysis

Notations. We use $\mathcal{O}(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ to denote upper, lower, and two-sided bounds up to universal constant factors, respectively. We further use the tilde notation $\tilde{\mathcal{O}}(\cdot)$ to hide polylogarithmic factors in the upper bounds. For a sequence of random variables a_n, $\liminf_{n\to\infty} a_n = \lim_{n\to\infty}(\inf_{m\geq n} a_m)$ (for example, see Durrett (2019)). For brevity, we write $\sup_{t\in\{0,1,\ldots,T-1\},s\in\mathcal{S}}$ as $\sup_{t,s}$ and write $\sup_{t\in\{0,1,\ldots,T-1\}}$ as $\sup_t$.

3.1 Assumptions

We impose Assumption 1 on the token-state process $\{s_t\}_{t\geq 0}$. Under Assumption 1, $\pi(s)$, the stationary probability of state s, is strictly positive for every $s \in \mathcal{S}$ (Sun et al., 2018).

Assumption 1. *The token-state process $\{s_t\}_{t\geq 0}$ is a time-homogeneous, irreducible, and aperiodic Markov chain on the finite state space $\mathcal{S}$. Its transition matrix M admits the target distribution π in Equation* (1) *as its stationary distribution, that is, $\pi M = \pi$.*

As done in Sun & Wei (2025a;b), we use the mixing time (Definition 1) to characterize the rate at which a Markov chain approaches its stationary distribution. Under Assumption 1, the mixing time τ_{mix} is finite (Levin & Peres, 2017, Theorem 4.9).

Definition 1 (Mixing time, Levin & Peres (2017))**.** *Given a Markov chain $\{s_t\}_{t\geq 0}$ with transition matrix M and stationary distribution π, for any $s, s' \in \mathcal{S}$, let $M(s,s') \coloneqq \mathbb{P}(s_{t+1} = s' \mid s_t = s)$*

and let $M^t(s,\cdot)$ denote the distribution of the chain after t steps starting from state s. The ε-mixing time is defined as

$$t_{\mathrm{mix}}(\varepsilon) \coloneqq \inf\left\{t \geq 1 : \max_{s\in\mathcal{S}} d_{\mathrm{TV}}\big(M^t(s,\cdot),\pi\big) \leq \varepsilon\right\},$$

where $d_{\mathrm{TV}}(\cdot,\cdot)$ denotes the total variation distance. We further define $\tau_{\mathrm{mix}} \coloneqq t_{\mathrm{mix}}\big(\frac{1}{4}\big)$.

Assumptions 2 and 3 are standard in non-convex optimization in Markovian SGD (Even, 2023; Sun & Wei, 2025a;b). Assumption 4 was adopted in Wang & Ji (2024) to model the node availability.

Assumption 2. *The local objective functions are L-smooth. That is, there exists a positive constant L such that for any $s \in \mathcal{S}$*

$$\|\nabla f_s(\mathbf{x}) - \nabla f_s(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|, \forall \mathbf{x},\mathbf{y}.$$

In addition, we assume that the global objective function f is lower bounded by f_* and define $F_0 \coloneqq f(\mathbf{x}_0) - f_*$.

Assumption 3. *For any $s \in \mathcal{S}$, there exists a positive constant ς_s such that*

$$\|\nabla f_s(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq \varsigma_s^2, \quad \forall \mathbf{x},$$

and we further define $\varsigma^2 \coloneqq \max_{s\in\mathcal{S}} \varsigma_s^2$.

Assumption 4. *We assume the availability status $\delta_t^s \sim \mathrm{Bernoulli}(p_t^s)$. Moreover, we assume that $\{\delta_t^s\}_{t,s}$ are mutually independent, and independent of the Markov chain $\{s_t\}_{t\geq 0}$.*

3.2 Main Results

The update of Markovian SGD with intermittent node availability in one iteration (Line 3 in Algorithm 1) can be written as

$$\begin{aligned}\mathbf{x}_{t+1} - \mathbf{x}_t &= -\gamma \delta_t^{s_t} w_t^{s_t} \nabla f_{s_t}(\mathbf{x}_t) \\ &= -\underbrace{\gamma \sum_{s\in\mathcal{S}} p_t^s w_t^s \pi(s)}_{\text{effective step size}} \underbrace{\frac{\delta_t^{s_t} w_t^{s_t}}{\sum_{s\in\mathcal{S}} p_t^s w_t^s \pi(s)} \nabla f_{s_t}(\mathbf{x}_t)}_{\text{normalized gradient}},\end{aligned} \tag{2}$$

where p_t^s denotes the mean of the availability status δ_t^s of node s at iteration t, w_t^s denotes the correction weight of node s at iteration t and $\pi(s)$ denotes the stationary probability of node s. Similar ideas can also be found in Wang et al. (2020). The key difference is that the randomness of the normalized gradient in Equation (2) additionally comes from node availability and the token transition process across nodes (specifically, $\delta_t^{s_t}$). This makes the analysis more challenging.

In this paper, we assume that p_t^s and w_t^s are positive and deterministic for any t and s. Define

$$P_t \coloneqq \sum_{s\in\mathcal{S}} p_t^s w_t^s \pi(s),$$

$$A_1 \coloneqq \sup_{t,s} \frac{p_t^s w_t^s}{P_t}, \qquad A_2 \coloneqq \sup_{t,s} \frac{p_t^s (w_t^s)^2}{P_t}, \qquad \text{and} \qquad A_3 \coloneqq \sup_{t,s} p_t^s w_t^s.$$

The upper bound is given in Theorem 1. The left-hand side of Equation (3) can be interpreted as the expected measure of a randomly selected iterate, where iteration t is sampled proportionally to its effective step size γP_t (with γ canceled), reflecting the unequal contributions of different iterations caused by node availability rates p_t^s and correction weights w_t^s. Similar measures have appeared in the analysis of SGD with iteration-varying step size (Ghadimi & Lan, 2013; Koloskova et al., 2022).

Theorem 1. *Under Assumptions 1, 2, 3 and 4, if $\gamma L A_1 A_2 \leq 2\times 10^{-7}$ and $\gamma L A_1 A_3 \tau_{mix} \leq 2\times 10^{-7}$, then the output of Algorithm 1 (Markovian SGD with intermittent node availability) satisfies*

$$\frac{1}{\sum_{t=0}^{T-1} P_t} \sum_{t=0}^{T-1} P_t \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 = \mathcal{O}\left(\frac{F_0}{\gamma \sum_{t=0}^{T-1} P_t} + \gamma L A_1 A_2 \varsigma^2 + \gamma L A_1 A_3 \tau_{mix} \varsigma^2 + \Phi^2\right). \tag{3}$$

The availability bias is

$$\Phi \coloneqq \sup_t \sum_{s\in\mathcal{S}} \left| \frac{p_t^s w_t^s \pi(s)}{\sum_{s\in\mathcal{S}} p_t^s w_t^s \pi(s)} - \pi(s) \right| \varsigma. \tag{4}$$

Proof ideas. Unlike Even (2023), whose proof draws on the analysis of asynchronous SGD (Mania et al., 2017) and biased SGD (Even et al., 2022), our proof is based on the analysis of permutation-based SGD (Koloskova et al., 2024; Islamov et al., 2024). Specifically, inspired by the epoch-wise analysis commonly used for permutation-based SGD, we partition the iterations into consecutive periods and perform period-wise analysis rather than iteration-wise analysis for Markovian SGD. This period-wise analysis enables the application of useful properties of Markov chains (specifically, Lemma 2). The detailed proof is in Appendix A. □

In the following, for ease of further analysis, we consider the iteration-invariant scenario. That is, we assume that $p_t^s = p^s$ and $w_t^s = w^s$ for any t, which implies that $P_t = P \coloneqq \sum_{s\in\mathcal{S}} p^s w^s \pi(s)$. After choosing an appropriate step size, we get Corollary 1. The upper bound in Corollary 1 consists of three terms, the optimization term, the error term and the bias term (from left to right). Among them, the bias term Φ does not vanish to 0 even as $T \to \infty$. This suggests that intermittent node availability can introduce a bias into Markovian SGD. This is analogous to the finding in Wang & Ji (2024) for FL with intermittent node availability.

Corollary 1. *If $p_t^s = p^s$ and $w_t^s = w^s$ in Theorem 1, then with an appropriate step size choice (see Equation* (38) *in Appendix A.4),*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 = \mathcal{O}\left(\frac{LF_0A_1\left(A_2+A_3\tau_{mix}\right)}{PT} + \left(\frac{LF_0A_1\left(A_2+A_3\tau_{mix}\right)\varsigma^2}{PT}\right)^{\frac{1}{2}} + \Phi^2\right).$$

The lower bound in Theorem 2 confirms that the bias term is unavoidable. Notably, the condition $\gamma T \to \infty$ is compatible with the upper bound: The step size selected in Corollary 1 satisfies $\gamma = \Theta(T^{-\frac{1}{2}})$ for sufficiently large T, and thus $\gamma T = \Theta(\sqrt{T}) \to \infty$.

Theorem 2. *For any time-invariant availability rates $p^s > 0$, correction weights $w^s > 0$, and distribution π with $\pi(s) > 0$ for $s \in \mathcal{S}$, there exist local objective functions satisfying Assumptions 2 and 3 and a Markov chain satisfying Assumption 1 such that, for any initialization $\mathbf{x}_0$, availability status $\delta_t^s \sim \text{Bernoulli}(p^s)$ satisfying Assumption 4, and step size $0 < \gamma \le \frac{1}{LP}$ with $\gamma T \to \infty$, the output of Algorithm 1 satisfies*

$$\liminf_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2 = \Omega\left(\Phi^2\right).$$

3.3 Implications

In this section, we discuss several implications of our bound. For ease of analysis, we consider the iteration-invariant scenario (that is, $p_t^s = p^s$ and $w_t^s = w^s$ for any t). Specifically, see Corollary 1.

Availability bias Φ. We begin by defining the effective distribution π_{eff} induced by the node availability and correction as

$$\pi_{\text{eff}}(s) \coloneqq \frac{p^s w^s \pi(s)}{\sum_{s\in\mathcal{S}} p^s w^s \pi(s)}, \quad s \in \mathcal{S}. \tag{5}$$

Then, the availability bias can be characterized by the total variation distance (a measure of the discrepancy between two distributions) between the effective distribution π_{eff} and the stationary distribution π:

$$\Phi = \sum_{s\in\mathcal{S}}\left|\frac{p^s w^s \pi(s)}{\sum_{s\in\mathcal{S}} p^s w^s \pi(s)} - \pi(s)\right|\varsigma = 2d_{\text{TV}}(\pi_{\text{eff}}, \pi)\varsigma.$$

This implies that the bias becomes smaller as the discrepancy between π_{eff} and π becomes smaller, and vice versa. In particular, the bias equals zero (that is, $\Phi = 0$) if and only if $p^s w^s$ is constant across $s \in \mathcal{S}$. Notably, the bias is governed by the collaborative capability induced by the effective distribution π_{eff}, rather than by the absolute availability probability itself.

In the following, we further assume that the availability bias Φ is zero; equivalently, $p^s w^s$ is constant across $s \in \mathcal{S}$. We first study the effects of the availability rates $\{p^s\}$ in the absence of the correction weights $\{w^s\}$, and then study the function of the correction weights $\{w^s\}$.

Effect of availability rates $\{p^s\}$. We consider the special case with zero availability bias and no correction weights; that is, $p^s = p$ and $w^s = 1$ for all $s \in \mathcal{S}$. In this case, the upper bound becomes

$$\mathcal{O}\left(\frac{LF_0\left(\frac{1}{p}+\tau_{\text{mix}}\right)}{T}+\left(\frac{LF_0\varsigma^2\left(\frac{1}{p}+\tau_{\text{mix}}\right)}{T}\right)^{\frac{1}{2}}\right).$$

It can be observed that the convergence rate improves (the upper bound decreases) as the availability rate p increases.

Effect of the correction weights $\{w^s\}$. In practical scenarios, the availability probabilities $\{p^s\}$ can be quite different due to the heterogeneity of nodes. In this case, the correction weights can be used to correct for the availability bias: setting $w^s = \frac{1}{p^s}$ yields $\Phi = 0$ and the following upper bound:

$$\mathcal{O}\left(\frac{LF_0\left(\max_{s\in\mathcal{S}}\frac{1}{p^s}+\tau_{\text{mix}}\right)}{T}+\left(\frac{LF_0\varsigma^2\left(\max_{s\in\mathcal{S}}\frac{1}{p^s}+\tau_{\text{mix}}\right)}{T}\right)^{\frac{1}{2}}\right).$$

This implies that although the correction weights can help correct for the availability bias, they also worsen the bound, especially for cases with low availability rates.

In practical scenarios, it would be better to adjust the weights w_t^s adaptively. See such examples in Wang & Ji (2024) and Sun et al. (2025). In this paper, we do not discuss these adaptive methods.

3.4 Comparisons with Related Works

This section provides detailed comparisons with existing works to show the tightness of our bound.

Comparison with FL with intermittent node availability (Wang & Ji, 2024). The upper bound of FL with intermittent node availability has been previously studied in Wang & Ji (2024, Theorem 2). The non-vanishing bias terms of Wang & Ji (2024) and this work are as follows:

$$\text{Wang \& Ji (2024)}: \quad \Psi^2 = \frac{1}{S}\sum_{s\in\mathcal{S}}(p^s w^s - 1)^2(G^2+\varsigma^2),$$

$$\text{This work}: \quad \Phi^2 = \left(\sum_{s\in\mathcal{S}}\left|\frac{p^s w^s \pi(s)}{\sum_{s\in\mathcal{S}} p^s w^s \pi(s)} - \pi(s)\right|\varsigma\right)^2.$$

Although the two results are derived for different algorithms, our bound has two advantages in characterizing the availability-induced bias: (i) Ours allows a more general zero-bias condition. The bias term of Wang & Ji (2024) vanishes when $p^s w^s = 1$ for all s, while ours vanishes as long as $p^s w^s$ is constant across nodes. (ii) Ours avoids restrictive additional assumptions. Wang & Ji (2024) depends on an additional assumption (Wang & Ji, 2024, Assumption 5), such as the bounded global gradient assumption (that is, $\|\nabla f(\mathbf{x})\| \le G, \forall \mathbf{x}$), while our bound does not.

Table 1: Comparisons of convergence rates of standard Markovian SGD. Note that standard Markovian SGD is a special case of Algorithm 1 when $p_t^s = 1$ and $w_t^s = 1$ for all t and s.

Algorithm	**Convergence rate**	**Constraint**
Markovian SGD (Even, 2023)	$\tilde{\mathcal{O}}\left(\frac{LF_0\tau'_{\text{mix}}+\varsigma^2}{T}+\left(\frac{LF_0\tau'_{\text{mix}}\varsigma_\pi^2}{T}\right)^{\frac{1}{2}}\right)^{(1)}$	$T = \Omega\left(\tau'_{\text{mix}}\right)^{(1)}$
(**This work**)	$\mathcal{O}\left(\frac{LF_0\tau_{\text{mix}}}{T}+\left(\frac{LF_0\tau_{\text{mix}}\varsigma^2}{T}\right)^{\frac{1}{2}}\right)$	–
Randomized Markovian SGD[(2)] (Beznosikov et al., 2023)	$\tilde{\mathcal{O}}\left(\frac{LF_0\tau_{\text{mix}}}{T}+\left(\frac{LF_0\tau_{\text{mix}}\varsigma^2}{T}\right)^{\frac{1}{2}}\right)$	–

[(1)] $\tau'_{\text{mix}} := t_{\text{mix}}(T^{-\frac{1}{2}}\min_{s\in\mathcal{S}}\pi(s))$ and $\varsigma_\pi^2 := \mathbb{E}_{s\sim\pi}[\varsigma_s^2]$.

[(2)] See Beznosikov et al. (2023, Algorithm 2).

Comparison with Markovian SGD (Even, 2023; Beznosikov et al., 2023; Sun & Wei, 2025b). When $p_t^s = 1$ and $w_t^s = 1$ for all t and s, Algorithm 1 reduces to standard Markovian SGD. Accordingly, upon specializing to Markovian SGD, our upper bound becomes

$$\mathcal{O}\left(\frac{LF_0\tau_{\text{mix}}}{T} + \left(\frac{LF_0\tau_{\text{mix}}\varsigma^2}{T}\right)^{\frac{1}{2}}\right). \tag{6}$$

Next, we compare our bound with the existing bounds for Markovian SGD, which are collected in Table 1. Our upper bound in Equation (6) matches the upper bound in Beznosikov et al. (2023, Corollary 2), whose result relies on a randomized batch size version of Markovian SGD (Beznosikov et al., 2023, Algorithm 2). In particular, our upper bound in Equation (6) is nearly matching in form with the lower bound $\Omega\big(\frac{LF_0\tau_{\text{hit}}}{T} + \big(\frac{LF_0\tau_{\text{hit}}\varsigma^2}{T}\big)^{\frac{1}{2}}\big)$ of Sun & Wei (2025b), modulo the difference between the mixing time τ_{mix} and the hitting time τ_{hit} (Sun & Wei, 2025b, Definition 2.1).

Compared with Even (2023, Theorem 2), a key advantage of our result is that Corollary 1 holds without a large-iteration requirement. Thus, it provides a convergence guarantee even when the computational budget is limited or the Markov chain mixes slowly, whereas the burn-in requirement in Even (2023) can make its guarantee inapplicable in such regimes. The trade-off is that the error term in Even (2023) depends on the stationary-distribution-weighted heterogeneity ς_π^2, which can be smaller than the worst-case heterogeneity ς^2 in our bound.

4 Experiments

We conduct experiments on quadratic functions and neural networks to validate the theoretical findings (see Section 3.3). Specifically, we focus on the iteration-invariant scenario (as in Corollary 1), and mainly validate the following findings:

- Finding 1. Heterogeneity of availability rates can cause the effective distribution of nodes to deviate from the intended distribution, thereby introducing a non-vanishing availability bias.
- Finding 2. Correction weights can help correct for the availability bias.
- Finding 3. Correction weights introduce a trade-off: while correcting for the availability bias, they can lead to slower convergence, especially under low availability rates.

The quadratic experiments isolate these effects in a controlled setting, while the neural network experiments examine whether the same behavior arises on standard learning tasks.

4.1 Quadratic Functions

Markov chain. We consider a Markov chain on S states with transition matrix $M = \rho\boldsymbol{I} + (1-\rho)\frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{S}$, where $\boldsymbol{I}$ denotes the identity matrix, $\mathbf{1}$ denotes the all-one vector, $\rho \in [0, 1)$ controls the persistence of the chain. More specifically, conditional on $s_t = i$, with probability ρ, the chain keeps the current state i; with probability $1-\rho$, the next state is sampled uniformly from all S states, including i. Unless stated otherwise, we set $\rho = 0.1$. This chain is irreducible and aperiodic and has the uniform stationary distribution $\pi(s) = \frac{1}{S}$ for every $s \in \mathcal{S}$.

Quadratic functions. We assume that S is even and divide the nodes into two equal-sized groups. We consider the following one-dimensional quadratic functions as the local objective functions:

$$f_s(\mathbf{x}) = \begin{cases} \frac{1}{2}\mathbf{x}^2 + \mathbf{x}, & 0 \le s \le \frac{S}{2} - 1, \\ \frac{1}{2}\mathbf{x}^2 - \mathbf{x}, & \frac{S}{2} \le s \le S - 1. \end{cases}$$

Under the uniform stationary distribution, the global objective is $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^2$, whose minimizer is $\mathbf{x}_* = 0$. Unless stated otherwise, we set $S = 1000$.

Availability rates. Nodes in the first group have availability rate $\mathfrak{p}_1 \in (0, 1/2]$, whereas nodes in the second group have rate $\mathfrak{p}_2 = 1 - \mathfrak{p}_1$. We use $\mathfrak{p}_1 \in \{0.5, 0.2, 0.1, 0.05\}$. Because the groups have

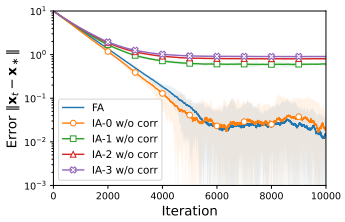
equal size, $\sum_{s\in\mathcal{S}} p^s\pi(s) = 1/2$ for every value of $\mathfrak{p}_1$. Thus, the expected total number of updates is held fixed while the availability heterogeneity varies. Without correction ($w^s = 1$), the bias is

$$\Phi = \sum_{s\in\mathcal{S}} \left| \frac{p^s\pi(s)}{\sum_{s\in\mathcal{S}} p^s\pi(s)} - \pi(s) \right| \varsigma = \frac{2}{S} \sum_{s\in\mathcal{S}} \left| p^s - \frac{1}{2} \right|.$$

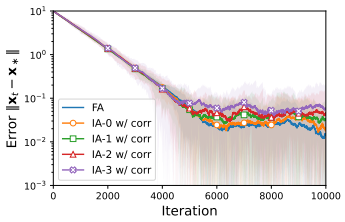
It is zero at $\mathfrak{p}_1 = 0.5$ and increases as $\mathfrak{p}_1$ decreases. With correction, we set $w^s = \frac{1}{p^s}$, which makes $p^s w^s$ constant and hence eliminates the availability bias. We additionally consider the full availability case in which $p^s = 1.0$ for all s (that is, each node is available at each iteration). As a summary, we collect the availability configurations in Table 2.

Table 2: Availability configurations of experiments. "FA" means the full availability case (there is no availability bias). "IA" means the intermittent availability cases. The availability bias is controlled by $\mathfrak{p}_1$. In particular, it is zero at $\mathfrak{p}_1 = 0.5$ and increases as $\mathfrak{p}_1$ decreases. In short, there is no bias in FA and IA-0 and the bias increases from IA-1 to IA-3.

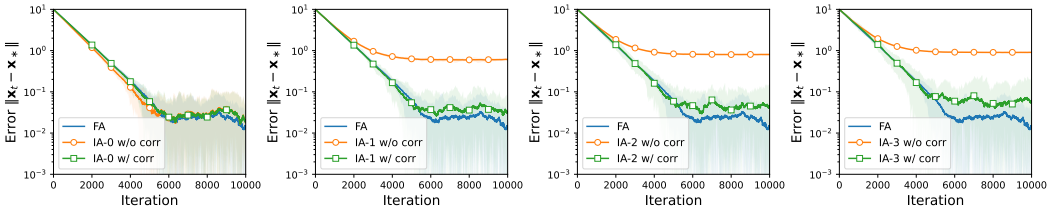
Config.	FA	IA-0	IA-1	IA-2	IA-3
Avail. rate	–	$\mathfrak{p}_1, \mathfrak{p}_2 = 0.5, 0.5$	$\mathfrak{p}_1, \mathfrak{p}_2 = 0.2, 0.8$	$\mathfrak{p}_1, \mathfrak{p}_2 = 0.1, 0.9$	$\mathfrak{p}_1, \mathfrak{p}_2 = 0.05, 0.95$



(a) Without correction.



(c) With correction.



(b) Comparison of corrected and uncorrected cases for each configuration.

Figure 2: Training curves of different availability configurations (Table 2). The shaded areas show the min-max values across 20 different random seeds. (a) is for Finding 1; (b) is for Finding 2; (c) is for Finding 3.

Experimental results. Figure 2 shows training curves of different availability configurations (Table 2). We measure convergence performance by the error $\|\mathbf{x}_t - \mathbf{x}_*\|$, with a smaller error indicating better performance. We have the following observations:

- Figure 2 (a): (i) IA-0 can achieve an error comparable to FA while other IAs cannot. This validates the existence of the availability bias. (ii) Comparing FA with IA-0, both of which have zero availability bias, FA achieves a smaller final error. This shows that a higher availability rate improves performance. (iii) Among IA-1, IA-2, and IA-3, the final error increases as the availability bias increases, showing that the bias predicts the performance. These validate Finding 1.
- Figure 2 (b): (i) In the zero-bias configuration IA-0, correction and no correction achieve nearly identical final errors. (ii) For IA-1 to IA-3, however, correction yields substantially smaller final errors, demonstrating its ability to mitigate the availability bias. These validate Finding 2.

- Figure 2 (c): (i) After correction, all IA configurations approach FA, confirming that their availability bias is removed. (ii) Nevertheless, their performance deteriorates from IA-0 to IA-3 as the minimum availability rate decreases. This demonstrates the trade-off introduced by correction: it corrects for availability bias but worsens convergence, particularly under low availability rates. These validate Finding 3.

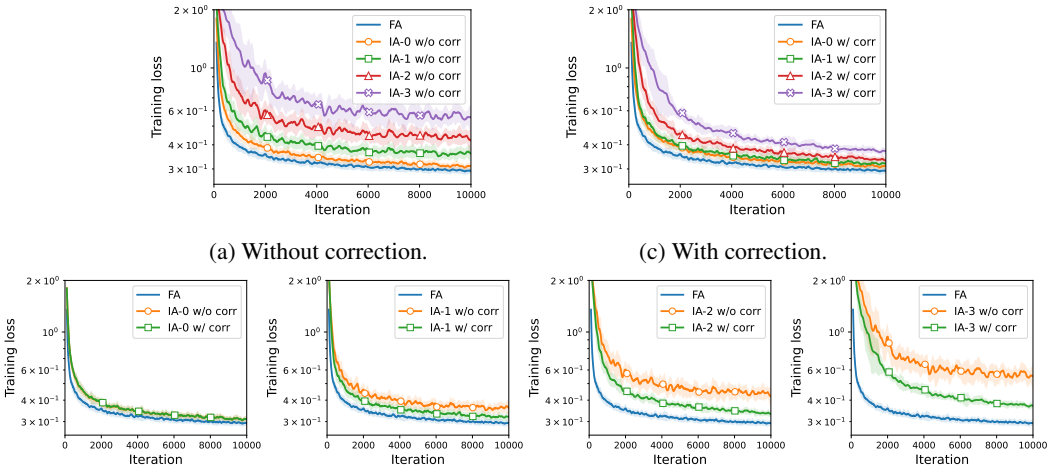
4.2 Neural Networks

We use the same Markov chain as in Section 4.1 with $\rho = 0.1$. More details are in Appendix C.

Models and datasets. We evaluate multinomial logistic regression on MNIST (LeCun et al., 1998) and convolutional neural networks on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009). For these datasets, we use the training sets to train models and compute training loss, and spare the test sets to compute test accuracy. We sort the training examples by their labels and split them into 1000 equal-sized consecutive partitions, which are assigned to nodes 0–999 in order. The local objective f_s is the training loss over the training examples assigned to node s.

Availability rates. We split the nodes into two groups, where nodes 0–499 form the first group and nodes 500–999 form the second group. Same as the availability configurations in Section 4.1, nodes in the first group have availability rate $\mathfrak{p}_1 \in (0, 1/2]$, whereas nodes in the second group have rate $\mathfrak{p}_2 = 1 - \mathfrak{p}_1$. In the following experiments, we use the same configurations as in Table 2. We compare training without correction, $w^s = 1$, with inverse-probability correction, $w^s = \frac{1}{p^s}$.

Experimental results. Figure 3 shows training curves of different availability configurations on MNIST. The results for CIFAR-10 and CIFAR-100 are deferred to Figure 6 and Figure 7 respectively in Appendix C. The observations are similar to those in Section 4.1.



(a) Without correction. (c) With correction.

(b) Comparison of corrected and uncorrected cases for each configuration.

Figure 3: Training curves of different availability configurations (Table 2) on MNIST. The shaded areas show the standard deviation across 10 random seeds. (a) is for Finding 1; (b) is for Finding 2; (c) is for Finding 3.

5 Conclusion

In this work, we analyze the convergence of Markovian SGD with intermittent node availability. The derived upper and lower bounds of convergence rates show that different node availability rates can cause a non-vanishing optimization bias with respect to the intended objective, and that correction weights can correct for the bias at the cost of slower convergence. Experiments on quadratic functions and neural networks support our theoretical findings.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, provide feedback on research methodology or experiments, implement methods, and assist with translation. We did not use generative AI tools to formulate mathematical claims, help develop theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design research methodology or experiments, or interpret results. Generating synthetic data sets, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis were not applicable to this work.

Additionally, we used generative AI tools to suggest experimental parameters, create or edit software code, draft parts of the research paper, summarize or analyse existing literature, support brainstorming, source or search for information, edit the paper to improve readability, suggest a structure for the paper, and propose a title or keywords for the paper.

We reviewed all AI-assisted work. The authors manually verified the AI-assisted portions of the proofs for logical correctness and consistency with the stated assumptions and mathematical claims. All mathematical claims, proof strategies, and critical ingredients of the proofs were developed and determined by the authors. AI-assisted feedback on the research methodology or experiments, as well as suggestions concerning experimental parameters, was evaluated by the authors before use. The research methodology and experiments were designed by the authors, and the experimental results were interpreted by the authors. AI-assisted implementations and software code were reviewed and tested for correctness. Literature summaries and sourced information were checked against the original sources. All AI-assisted drafting, translation, structural suggestions, and readability edits were reviewed to ensure technical accuracy and preservation of the intended meaning. The authors made all final decisions regarding the proofs, methodology, experiments, presentation, title, and keywords. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The problem setting and the proposed algorithm are described in Section 1, and the assumptions and theoretical results are stated in Section 3. Complete proofs of Theorems 1 and 2 are provided in Appendices A and B, respectively.

The experimental design, including the objective functions, Markov chain, node availability configurations, datasets, models, and data partitioning, is described in Section 4. Additional implementation details and further experimental results are provided in Appendix C. Code for reproducing the experiments on quadratic functions is included in the supplementary material.

REFERENCES

Aleksandr Beznosikov, Sergey Samsonov, Marina Sheshukova, Alexander Gasnikov, Alexey Naumov, and Eric Moulines. First order methods with Markovian noise: from acceleration to variational inequalities. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 44820–44835. Curran Associates, Inc., 2023. doi: 10.52202/075280-1941.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173.

Wenlin Chen, Samuel Horváth, and Peter Richtárik. Optimal client sampling for federated learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Xingran Chen, Parimal Parag, Rohit Bhagat, Zonghong Liu, and Salim El Rouayheb. Random walk learning and the Pac-Man attack. In *2026 IEEE International Symposium on Information Theory (ISIT)*, pp. 1–6, 2026. doi: 10.1109/ISIT62367.2026.11654115.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards understanding biased client selection in federated learning. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 10351–10375. PMLR, 28–30 Mar 2022.

Thinh T. Doan. Finite-time analysis of Markov gradient descent. *IEEE Transactions on Automatic Control*, 68(4):2140–2153, 2023. doi: 10.1109/TAC.2022.3172593.

Ron Dorfman and Kfir Yehuda Levy. Adapting to mixing time in stochastic optimization with Markovian data. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5429–5446. PMLR, 17–23 Jul 2022.

Rick Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5th edition, 2019.

Mathieu Even. Stochastic gradient descent under Markovian sampling schemes. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9412–9439. PMLR, 23–29 Jul 2023.

Mathieu Even, Laurent Massoulié, and Kevin Scaman. On sample optimality in personalized collaborative and federated learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 212–225. Curran Associates, Inc., 2022.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811.

Xinran Gu, Kaixuan Huang, Jingzhao Zhang, and Longbo Huang. Fast federated learning in the presence of arbitrary device unavailability. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12052–12064. Curran Associates, Inc., 2021.

Samuel Horváth, Maziar Sanjabi, Lin Xiao, Peter Richtárik, and Michael Rabbat. FedShuffle: Recipes for better use of local work in federated learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Khiem Huynh, Malcolm Egan, Giovanni Neglia, and Jean-Marie Gorce. Streaming federated learning with Markovian data. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 128137–128189. Curran Associates, Inc., 2025. doi: 10.52202/085713-4267.

Rustem Islamov, Mher Safaryan, and Dan Alistarh. AsGrad: A sharp unified analysis of asynchronous-SGD algorithms. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 649–657. PMLR, 02–04 May 2024.

Björn Johansson, Maben Rabi, and Mikael Johansson. A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization*, 20 (3):1157–1170, 2010. doi: 10.1137/08073038X.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5381–5393. PMLR, 13–18 Jul 2020.

Anastasia Koloskova, Nikita Doikov, Sebastian U Stich, and Martin Jaggi. On convergence of incremental gradient for non-convex smooth functions. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25058–25086. PMLR, 21–27 Jul 2024.

Anastasiia Koloskova, Sebastian U Stich, and Martin Jaggi. Sharper convergence guarantees for asynchronous SGD for distributed and federated learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17202–17215. Curran Associates, Inc., 2022. doi: 10.52202/068431-1251.

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

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

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

Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I. Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM Journal on Optimization*, 27(4):2202–2229, 2017. doi: 10.1137/16M1057000.

Xianghui Mao, Kun Yuan, Yubin Hu, Yuantao Gu, Ali H. Sayed, and Wotao Yin. Walkman: A communication-efficient random-walk algorithm for decentralized optimization. *IEEE Transactions on Signal Processing*, 68:2513–2528, 2020. doi: 10.1109/TSP.2020.2983167.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017.

S. Sundhar Ram, A. Nedić, and V. V. Veeravalli. Incremental stochastic subgradient algorithms for convex optimization. *SIAM Journal on Optimization*, 20(2):691–717, 2009. doi: 10.1137/080726380.

Ali H. Sayed. Adaptation, learning, and optimization over networks. *Foundations and Trends in Machine Learning*, 7(4-5):311–801, 07 2014. ISSN 1935-8237. doi: 10.1561/2200000051.

Tao Sun, Yuejiao Sun, and Wotao Yin. On Markov chain gradient descent. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Zhenyu Sun and Ermin Wei. Improved lower bounds for first-order stochastic non-convex optimization under Markov sampling. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57754–57772. PMLR, 13–19 Jul 2025a.

Zhenyu Sun and Ermin Wei. Learning from a single markovian trajectory: Optimality and variance reduction. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 20004–20033. Curran Associates, Inc., 2025b. doi: 10.52202/085713-0675.

Zhenyu Sun, Ziyang Zhang, Zheng Xu, Gauri Joshi, Pranay Sharma, and Ermin Wei. Debiasing federated learning with correlated client participation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 28564–28595, 2025.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7611–7623. Curran Associates, Inc., 2020.

Shiqiang Wang and Mingyue Ji. A unified analysis of federated learning with arbitrary client participation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 19124–19137. Curran Associates, Inc., 2022. doi: 10.52202/068431-1390.

Shiqiang Wang and Mingyue Ji. A lightweight method for tackling unknown participation statistics in federated averaging. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 854–893, 2024.

Shudi Weng, Chao Ren, Ming Xiao, and Mikael Skoglund. Heterogeneity-aware client sampling for optimal and efficient federated learning, 2026. URL https://arxiv.org/abs/2505.11304.

Ming Xiang, Stratis Ioannidis, Edmund Yeh, Carlee Joe-Wong, and Lili Su. Efficient federated learning against heterogeneous and non-stationary client unavailability. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 104281–104328. Curran Associates, Inc., 2024. doi: 10.52202/079017-3313.

Yikai Yan, Chaoyue Niu, Yucheng Ding, Zhenzhe Zheng, Shaojie Tang, Qinya Li, Fan Wu, Chengfei Lyu, Yanghe Feng, and Guihai Chen. Federated optimization under intermittent client availability. *INFORMS Journal on Computing*, 36(1):185–202, 2024.

Bicheng Ying, Zhe Li, and Haibo Yang. Exact and linear convergence for federated learning under arbitrary client participation is attainable. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 40156–40201. Curran Associates, Inc., 2025. doi: 10.52202/085713-1340.

Shahryar Zehtabi, Dong-Jun Han, Rohit Parasnis, Seyyedali Hosseinalipour, and Christopher Brinton. Decentralized sporadic federated learning: A unified algorithmic framework with convergence guarantees. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 59567–59636, 2025.

Appendix

A Proof of Theorem 1 **14**
A.1 Preliminaries 14
A.2 Supporting Lemmas 16
A.3 Proof of Theorem 1 26
A.4 Proof of Corollary 1 36

B Proof of Theorem 2 **36**

C Experiments **39**

A Proof of Theorem 1

This section presents the proof of Theorem 1.

A.1 Preliminaries

In this section, we present the preliminaries for the proof of Theorem 1.

Notations. In the following proofs, we adopt the convention that $\sum_{i=a}^{b} \cdot = 0$ whenever $a > b$. For vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, we use $\langle \mathbf{a}, \mathbf{b} \rangle$ to denote the Euclidean inner product and $\|\mathbf{a}\|$ to denote the induced Euclidean norm. For brevity, we write $\mathbb{E}\left[\langle \cdot, \cdot \rangle\right]$ as $\mathbb{E}\langle \cdot, \cdot \rangle$ and write $\mathbb{E}[\|\cdot\|^2]$ as $\mathbb{E}\|\cdot\|^2$. For brevity, we write $\sup_{t\in\{0,1,\ldots,T-1\},s\in\mathcal{S}}$ as $\sup_{t,s}$ and write $\sup_{t\in\{0,1,\ldots,T-1\}}$ as $\sup_t$. Recall that

$$P_t \coloneqq \sum_{s\in\mathcal{S}} p_t^s w_t^s \pi(s), \tag{7}$$

$$A_1 \coloneqq \sup_{t,s} \frac{p_t^s w_t^s}{P_t} \geq 1, \qquad A_2 \coloneqq \sup_{t,s} \frac{p_t^s (w_t^s)^2}{P_t}, \qquad \text{and} \qquad A_3 \coloneqq \sup_{t,s} p_t^s w_t^s. \tag{8}$$

Moreover, we define

$$\phi_t \coloneqq \sum_{s\in\mathcal{S}} \left| \frac{p_t^s w_t^s \pi(s)}{\sum_{s\in\mathcal{S}} p_t^s w_t^s \pi(s)} - \pi(s) \right| \varsigma. \tag{9}$$

Our proof is based on a period-wise analysis. We next introduce some period-related notations and describe how the periods are constructed.

Periods. The iterations are partitioned into multiple consecutive periods (see Figure 4 for an example). The algorithm of constructing these periods will be presented in the following paragraph. Here, we use l_q to denote the starting index of the period $q \in \{0, 1, \ldots, Q\}$ and use $\tau_q \coloneqq l_{q+1} - l_q$ to denote the length of the period $q \in \{0, 1, \ldots, Q-1\}$. In particular, for the last incomplete period Q, we further use $\tau_Q \coloneqq T - l_Q$ to denote its length. Notably, if $l_Q = T$, the period Q does not exist. We use $\mathcal{L}_T \coloneqq \{l_1, l_2, \ldots, l_Q\}$ to denote the set of all starting indices except l_0. We use $l(t)$ to denote the starting index of the period that t belongs to, and $\tau(t)$ to denote the length of the period that t belongs to. For example, in Figure 4, if $t = 5$, then $l(t) = 4$ and $\tau(t) = 3$.

In contrast to Koloskova et al. (2024), which considers a constant step size and sets the period length as $\Theta\big(\frac{1}{\gamma L}\big)$, we consider an iteration-varying effective step size. Accordingly, as in the proof of Theorem 1, we require each complete period to satisfy $\sum_{t=l_q}^{l_q+\tau_q-1} P_t = \Theta\big(\frac{1}{\gamma L A_1}\big)$. To achieve this, we use the following algorithm in Lemma 1 to construct the periods. The specific values of a and b in Lemma 1 will be specified later in the proof of Theorem 1 (see Appendix A.3).

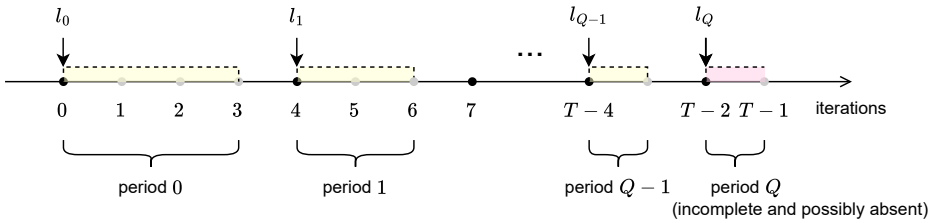


Figure 4: An example of the period partition.

Lemma 1. *Let $0 < a < b$ and suppose that $0 \le P_t \le b - a$ for all $t \in \{0, 1, \dots, T-1\}$. Let $l_0 = 0$, and recursively choose*

$$l_{q+1} = \min \left\{ l_q < l \le T : \sum_{t=l_q}^{l-1} P_t \ge a \right\},$$

whenever the set is nonempty. Then,

$$a \le \sum_{t=l_q}^{l_q+\tau_q-1} P_t \le b, \quad \textit{for } q = 0, 1, \dots, Q-1, \qquad \textit{and} \qquad \sum_{t=l_Q}^{l_Q+\tau_Q-1} P_t < a.$$

Proof. By the definition of l_{q+1},

$$\sum_{t=l_q}^{l_{q+1}-1} P_t \ge a.$$

Moreover, by the minimality of l_{q+1},

$$\sum_{t=l_q}^{l_{q+1}-2} P_t < a.$$

Since $P_{l_{q+1}-1} \le b - a$, it follows that

$$\sum_{t=l_q}^{l_{q+1}-1} P_t < a + (b - a) = b,$$

and hence the desired upper bound follows. Finally, since no l_{Q+1} exists, the defining set is empty. Thus,

$$\sum_{t=l_Q}^{T-1} P_t = \sum_{t=l_Q}^{l_Q+\tau_Q-1} P_t < a;$$

otherwise $l = T$ would belong to the defining set. This proves the claim. □

Following the perturbed iterate analysis of Koloskova et al. (2024), we construct virtual iterates based on the periods, and then analyze the convergence based on the virtual iterates.

Virtual iterates. The virtual iterates $\{\tilde{\mathbf{x}}_t\}$ are initialized with $\tilde{\mathbf{x}}_0 = \mathbf{x}_0$ and are defined by

$$\begin{aligned} \tilde{\mathbf{x}}_{t+1} &= \tilde{\mathbf{x}}_t - \gamma P_t \nabla f(\mathbf{x}_t), && \text{if } t+1 \notin \mathcal{L}_T, \\ \tilde{\mathbf{x}}_{t+1} &= \mathbf{x}_{t+1}, && \text{if } t+1 \in \mathcal{L}_T. \end{aligned}$$

By the definition of the virtual iterates $\{\tilde{\mathbf{x}}_t\}$, we can further obtain the following relations between $\tilde{\mathbf{x}}_{t+1}$ and $\tilde{\mathbf{x}}_t$, and between $\tilde{\mathbf{x}}_{t+1}$ and $\mathbf{x}_{t+1}$:

Case 1. For $t + 1 \notin \mathcal{L}_T$,

$$\tilde{\mathbf{x}}_{t+1} = \tilde{\mathbf{x}}_t - \gamma P_t \nabla f(\mathbf{x}_t) \tag{10}$$

$$= \mathbf{x}_{t+1} + \gamma \sum_{i=l(t)}^{t} \left(\delta_i^{s_i} w_i^{s_i} \nabla f_{s_i}(\mathbf{x}_i) - P_i \nabla f(\mathbf{x}_i) \right). \tag{11}$$

Case 2. For $t+1 \in \mathcal{L}_T$,

$$\tilde{\mathbf{x}}_{t+1} = \tilde{\mathbf{x}}_t - \gamma P_t \nabla f(\mathbf{x}_t) - \gamma \sum_{i=l(t)}^{t} (\delta_i^{s_i} w_i^{s_i} \nabla f_{s_i}(\mathbf{x}_i) - P_i \nabla f(\mathbf{x}_i)) \tag{12}$$

$$= \mathbf{x}_{t+1}. \tag{13}$$

A.2 Supporting Lemmas

Lemma 2. *Let M be the transition matrix of a Markov chain on the state space $\mathcal{S}$, and let π be its stationary distribution, that is, $\pi M = \pi$. Let τ_{mix} denote the mixing time of this Markov chain. Define*

$$d(t) \coloneqq \max_{s \in \mathcal{S}} d_{\mathrm{TV}}(M^t(s, \cdot), \pi).$$

Then,

$$\sum_{t=0}^{T} d(t) \le 3\tau_{\mathrm{mix}}, \quad \forall T \ge 0.$$

Proof. This lemma is from Sun & Wei (2025b, Lemma D.1). □

Lemma 3. *In the same settings as Theorem 1,*

$$\begin{aligned}
&\mathbb{E}\left\| \sum_{i=t}^{t+\tau-1} (p_i^{s_i} w_i^{s_i} \nabla f_{s_i}(\mathbf{x}_t) - P_i \nabla f(\mathbf{x}_t)) \right\|^2 \\
&\le 234 A_3^2 \tau_{\mathrm{mix}}^2 \varsigma^2 + 468 A_3^2 \tau_{\mathrm{mix}}^2 \mathbb{E} \left\| \nabla f(\mathbf{x}_t) \right\|^2 \\
&\quad + 75 \sum_{i=t}^{t+\tau-1} P_i A_3 \tau_{\mathrm{mix}} \varsigma^2 + 150 \sum_{i=t}^{t+\tau-1} P_i A_3 \tau_{\mathrm{mix}} \mathbb{E} \left\| \nabla f(\mathbf{x}_t) \right\|^2 + \sum_{i=t}^{t+\tau-1} P_i \sum_{i'=t}^{t+\tau-1} P_{i'} \phi_{i'}^2.
\end{aligned}$$

Proof. Throughout the proof, we use $\mathbb{E}_t$ to denote the expectation conditional on s_t and $\mathbf{x}_t$, and $\mathbb{E}$ to denote the unconditional expectation. For brevity, for $t \le i \le t+\tau-1$, we define

$$h_i^s \coloneqq p_i^s w_i^s \nabla f_s(\mathbf{x}_t) - P_i \nabla f(\mathbf{x}_t). \tag{14}$$

Given a Markov chain with the transition matrix M and stationary distribution π, for any $0 \le i \le j$ and $s \in \mathcal{S}$,

$$\sum_{s' \in \mathcal{S}} |\mathbb{P}(s_j = s' \mid s_i = s) - \pi(s')| = 2 d_{\mathrm{TV}}(M^{j-i}(s, \cdot), \pi), \tag{15}$$

which is due to the property of $d_{\mathrm{TV}}(\cdot, \cdot)$ (Levin & Peres, 2017, Proposition 4.2).

Using $\left\| \sum_{i=1}^n \mathbf{a}_i \right\|^2 = 2 \sum_{i=1}^n \sum_{j>i}^n \langle \mathbf{a}_i, \mathbf{a}_j \rangle + \sum_{i=1}^n \left\| \mathbf{a}_i \right\|^2$ for any $\mathbf{a}_i \in \mathbb{R}^d$ and $n \ge 1$, we expand the target term as follows

$$\mathbb{E}_t \left\| \sum_{i=t}^{t+\tau-1} h_i^{s_i} \right\|^2 = 2 \sum_{i=t}^{t+\tau-2} \sum_{j=i+1}^{t+\tau-1} \mathbb{E}_t \left\langle h_i^{s_i}, h_j^{s_j} \right\rangle + \sum_{i=t}^{t+\tau-1} \mathbb{E}_t \left\| h_i^{s_i} \right\|^2.$$

Using the property of Markov chain, we have

$$\begin{aligned}
\mathbb{E}_t \left\langle h_i^{s_i}, h_j^{s_j} \right\rangle = &\underbrace{\sum_{s \in \mathcal{S}} \big(\mathbb{P}\big(s_i = s \mid s_t\big) - \pi(s)\big) \sum_{s' \in \mathcal{S}} \big(\mathbb{P}\big(s_j = s' \mid s_i = s\big) - \pi(s')\big) \left\langle h_i^s, h_j^{s'} \right\rangle}_{\mathcal{T}_1} \\
&+ \underbrace{\sum_{s \in \mathcal{S}} \pi(s) \sum_{s' \in \mathcal{S}} \big(\mathbb{P}\big(s_j = s' \mid s_i = s\big) - \pi(s')\big) \left\langle h_i^s, h_j^{s'} \right\rangle}_{\mathcal{T}_2}
\end{aligned}$$

$$+\underbrace{\sum_{s\in\mathcal{S}}\big(\mathbb{P}\big(s_i=s\mid s_t\big)-\pi(s)\big)\sum_{s'\in\mathcal{S}}\pi(s')\left\langle h_i^s,h_j^{s'}\right\rangle}_{\mathcal{T}_3}$$
$$+\underbrace{\sum_{s\in\mathcal{S}}\pi(s)\sum_{s'\in\mathcal{S}}\pi(s')\left\langle h_i^s,h_j^{s'}\right\rangle}_{\mathcal{T}_4},$$

and

$$\mathbb{E}_t\left\|h_i^{s_i}\right\|^2=\underbrace{\sum_{s\in\mathcal{S}}\big(\mathbb{P}\big(s_i=s\mid s_t\big)-\pi(s)\big)\left\|h_i^s\right\|^2}_{\mathcal{T}_5}+\underbrace{\sum_{s\in\mathcal{S}}\pi(s)\left\|h_i^s\right\|^2}_{\mathcal{T}_6}.$$

Then, for brevity, the target term can be written as

$$\mathbb{E}_t\left\|\sum_{i=t}^{t+\tau-1}h_i^{s_i}\right\|^2=2\sum_{i=t}^{t+\tau-2}\sum_{j=i+1}^{t+\tau-1}\mathcal{T}_1+2\sum_{i=t}^{t+\tau-2}\sum_{j=i+1}^{t+\tau-1}\mathcal{T}_2+2\sum_{i=t}^{t+\tau-2}\sum_{j=i+1}^{t+\tau-1}\mathcal{T}_3+2\sum_{i=t}^{t+\tau-2}\sum_{j=i+1}^{t+\tau-1}\mathcal{T}_4$$
$$+\sum_{i=t}^{t+\tau-1}\mathcal{T}_5+\sum_{i=t}^{t+\tau-1}\mathcal{T}_6. \tag{16}$$

The term $\|h_i^s\|$ can be bounded as follows:

$$\begin{aligned}\|h_i^s\|&=\left\|p_i^s w_i^s\nabla f_s(\mathbf{x}_t)-\sum_{s\in\mathcal{S}}p_i^s w_i^s\pi(s)\nabla f(\mathbf{x}_t)\right\|\\&\le\|p_i^s w_i^s\nabla f_s(\mathbf{x}_t)-p_i^s w_i^s\nabla f(\mathbf{x}_t)\|+\left\|p_i^s w_i^s\nabla f(\mathbf{x}_t)-\sum_{s\in\mathcal{S}}p_i^s w_i^s\pi(s)\nabla f(\mathbf{x}_t)\right\|\\&\le p_i^s w_i^s\varsigma+p_i^s w_i^s\|\nabla f(\mathbf{x}_t)\|+\sum_{s\in\mathcal{S}}p_i^s w_i^s\pi(s)\|\nabla f(\mathbf{x}_t)\| \qquad (17)\\&\le A_3\varsigma+2A_3\|\nabla f(\mathbf{x}_t)\| \qquad (18)\end{aligned}$$

where the second inequality uses Assumption 3 and the third inequality uses Equation (8). Moreover, using Equation (17), we have

$$\sum_{s\in\mathcal{S}}\pi(s)\|h_i^s\|\le P_i\varsigma+2P_i\|\nabla f(\mathbf{x}_t)\|. \tag{19}$$

Next, we establish upper bounds on the terms on the right-hand side in Equation (16). For $2\sum_{i=t}^{t+\tau-2}\sum_{j=i+1}^{t+\tau-1}\mathcal{T}_1$, we have

$$\begin{aligned}&2\sum_{i=t}^{t+\tau-2}\sum_{j=i+1}^{t+\tau-1}\mathcal{T}_1\\&=2\sum_{i=t}^{t+\tau-2}\sum_{s\in\mathcal{S}}\big(\mathbb{P}\big(s_i=s\mid s_t\big)-\pi(s)\big)\sum_{s'\in\mathcal{S}}\sum_{j=i+1}^{t+\tau-1}\big(\mathbb{P}\big(s_j=s'\mid s_i=s\big)-\pi(s')\big)\left\langle h_i^s,h_j^{s'}\right\rangle\\&\le2\sum_{i=t}^{t+\tau-2}\sum_{s\in\mathcal{S}}\big|\mathbb{P}\big(s_i=s\mid s_t\big)-\pi(s)\big|\sum_{j=i+1}^{t+\tau-1}\sum_{s'\in\mathcal{S}}\big|\mathbb{P}\big(s_j=s'\mid s_i=s\big)-\pi(s')\big|\,\|h_i^s\|\left\|h_j^{s'}\right\|\\&\le2\sum_{i=t}^{t+\tau-2}\sum_{s\in\mathcal{S}}\big|\mathbb{P}\big(s_i=s\mid s_t\big)-\pi(s)\big|\sum_{j=i+1}^{t+\tau-1}\sum_{s'\in\mathcal{S}}\big|\mathbb{P}\big(s_j=s'\mid s_i=s\big)-\pi(s')\big|\,A_3^2\left(\varsigma+2\|\nabla f(\mathbf{x}_t)\|\right)^2\\&=4\sum_{i=t}^{t+\tau-2}\sum_{s\in\mathcal{S}}\big|\mathbb{P}\big(s_i=s\mid s_t\big)-\pi(s)\big|\sum_{j=i+1}^{t+\tau-1}d_{\mathrm{TV}}(M^{j-i}(s,\cdot),\pi)A_3^2\left(\varsigma+2\|\nabla f(\mathbf{x}_t)\|\right)^2\end{aligned}$$

$$
\begin{aligned}
&\leq 12 \sum_{i=t}^{t+\tau-2} \sum_{s\in\mathcal{S}} \left|\mathbb{P}\big(s_i = s \mid s_t\big) - \pi(s)\right| \tau_{\mathrm{mix}} A_3^2 \left(\varsigma + 2\left\|\nabla f(\mathbf{x}_t)\right\|\right)^2 \\
&= 24 \sum_{i=t}^{t+\tau-2} d_{\mathrm{TV}}(M^{i-t}(s_t,\cdot),\pi)\tau_{\mathrm{mix}} A_3^2 \left(\varsigma + 2\left\|\nabla f(\mathbf{x}_t)\right\|\right)^2 \\
&\leq 72\tau_{\mathrm{mix}}^2 A_3^2 \left(\varsigma + 2\left\|\nabla f(\mathbf{x}_t)\right\|\right)^2 \\
&\leq 216 A_3^2 \tau_{\mathrm{mix}}^2 \varsigma^2 + 432 A_3^2 \tau_{\mathrm{mix}}^2 \left\|\nabla f(\mathbf{x}_t)\right\|^2,
\end{aligned}
$$

where the second inequality uses Equation (18), the second and third equalities use Equation (15), the third and fourth inequalities use Lemma 2, and the last inequality uses $(a+2b)^2 \leq 3a^2 + 6b^2$ for $a, b \in \mathbb{R}$.

For $2\sum_{i=t}^{t+\tau-2}\sum_{j=i+1}^{t+\tau-1} \mathcal{T}_2$, we have

$$
\begin{aligned}
2 \sum_{i=t}^{t+\tau-2} \sum_{j=i+1}^{t+\tau-1} \mathcal{T}_2 &= 2 \sum_{i=t}^{t+\tau-2} \sum_{s\in\mathcal{S}} \pi(s) \sum_{j=i+1}^{t+\tau-1} \sum_{s'\in\mathcal{S}} \big(\mathbb{P}\big(s_j = s' \mid s_i = s\big) - \pi(s')\big) \left\langle h_i^s, h_j^{s'} \right\rangle \\
&\leq 2 \sum_{i=t}^{t+\tau-2} \sum_{s\in\mathcal{S}} \pi(s) \sum_{j=i+1}^{t+\tau-1} \sum_{s'\in\mathcal{S}} \left|\mathbb{P}\big(s_j = s' \mid s_i = s\big) - \pi(s')\right| \left\|h_i^s\right\| \left\|h_j^{s'}\right\| \\
&\leq 4 \sum_{i=t}^{t+\tau-2} \sum_{s\in\mathcal{S}} \pi(s) \left\|h_i^s\right\| \sum_{j=i+1}^{t+\tau-1} d_{\mathrm{TV}}(M^{j-i}(s,\cdot),\pi) A_3 \left(\varsigma + 2\left\|\nabla f(\mathbf{x}_t)\right\|\right) \\
&\leq 12 \sum_{i=t}^{t+\tau-2} \sum_{s\in\mathcal{S}} \pi(s) \left\|h_i^s\right\| \tau_{\mathrm{mix}} A_3 \left(\varsigma + 2\left\|\nabla f(\mathbf{x}_t)\right\|\right) \\
&\leq 12 \sum_{i=t}^{t+\tau-1} P_i \tau_{\mathrm{mix}} A_3 \left(\varsigma + 2\left\|\nabla f(\mathbf{x}_t)\right\|\right)^2 \\
&\leq 36 \sum_{i=t}^{t+\tau-1} P_i \tau_{\mathrm{mix}} A_3 \varsigma^2 + 72 \sum_{i=t}^{t+\tau-1} P_i \tau_{\mathrm{mix}} A_3 \left\|\nabla f(\mathbf{x}_t)\right\|^2,
\end{aligned}
$$

where the second inequality uses Equation (18) and then Equation (15), the third inequality uses Lemma 2, and the fourth inequality uses Equation (19).

For $2\sum_{i=t}^{t+\tau-2}\sum_{j=i+1}^{t+\tau-1} \mathcal{T}_3$, we have

$$
\begin{aligned}
2 \sum_{i=t}^{t+\tau-2} \sum_{j=i+1}^{t+\tau-1} \mathcal{T}_3 &= 2 \sum_{i=t}^{t+\tau-2} \sum_{s\in\mathcal{S}} \sum_{j=i+1}^{t+\tau-1} \big(\mathbb{P}\big(s_i = s \mid s_t\big) - \pi(s)\big) \sum_{s'\in\mathcal{S}} \pi(s') \left\langle h_i^s,\ h_j^{s'} \right\rangle \\
&\leq 2 \sum_{i=t}^{t+\tau-2} \sum_{s\in\mathcal{S}} \left|\mathbb{P}\big(s_i = s \mid s_t\big) - \pi(s)\right| \sum_{j=i+1}^{t+\tau-1} \sum_{s'\in\mathcal{S}} \pi(s') \left\|h_i^s\right\| \left\|h_j^{s'}\right\| \\
&\leq 2 \sum_{i=t}^{t+\tau-2} \sum_{s\in\mathcal{S}} \left|\mathbb{P}\big(s_i = s \mid s_t\big) - \pi(s)\right| \left\|h_i^s\right\| \sum_{j=t}^{t+\tau-1} P_j \left(\varsigma + 2\left\|\nabla f(\mathbf{x}_t)\right\|\right) \\
&\leq 4 \sum_{i=t}^{t+\tau-2} d_{\mathrm{TV}}(M^{i-t}(s_t,\cdot),\pi) \sum_{j=t}^{t+\tau-1} P_j A_3 \left(\varsigma + 2\left\|\nabla f(\mathbf{x}_t)\right\|\right)^2 \\
&\leq 12 \sum_{i=t}^{t+\tau-1} P_i \tau_{\mathrm{mix}} A_3 \left(\varsigma + 2\left\|\nabla f(\mathbf{x}_t)\right\|\right)^2 \\
&\leq 36 \sum_{i=t}^{t+\tau-1} P_i \tau_{\mathrm{mix}} A_3 \varsigma^2 + 72 \sum_{i=t}^{t+\tau-1} P_i \tau_{\mathrm{mix}} A_3 \left\|\nabla f(\mathbf{x}_t)\right\|^2,
\end{aligned}
$$

where the second inequality uses Equation (19), the third inequality uses Equation (18) and then Equation (15), and the fourth inequality uses Lemma 2.

For $2\sum_{i=t}^{t+\tau-2}\sum_{j=i+1}^{t+\tau-1}\mathcal{T}_4$, we have

$$
\begin{aligned}
&2\sum_{i=t}^{t+\tau-2}\sum_{j=i+1}^{t+\tau-1}\mathcal{T}_4 \\
&= 2\sum_{i=t}^{t+\tau-2}\sum_{j=i+1}^{t+\tau-1}\sum_{s\in\mathcal{S}}\pi(s)\sum_{s'\in\mathcal{S}}\pi(s')\left\langle h_i^s, h_j^{s'}\right\rangle \\
&= 2\sum_{i=t}^{t+\tau-2}\sum_{j=i+1}^{t+\tau-1}\left\langle\sum_{s\in\mathcal{S}}\pi(s)h_i^s, \sum_{s'\in\mathcal{S}}\pi(s')h_j^{s'}\right\rangle \\
&\le \left\|\sum_{i=t}^{t+\tau-1}\sum_{s\in\mathcal{S}}\pi(s)h_i^s\right\|^2 \\
&= \left\|\sum_{i=t}^{t+\tau-1}\left(\sum_{s\in\mathcal{S}}p_i^s w_i^s\pi(s)\nabla f_s(\mathbf{x}_t) - \sum_{s\in\mathcal{S}}p_i^s w_i^s\pi(s)\nabla f(\mathbf{x}_t)\right)\right\|^2 \\
&= \left\|\sum_{i=t}^{t+\tau-1}P_i\left(\sum_{s\in\mathcal{S}}\frac{p_i^s w_i^s\pi(s)}{P_i}\nabla f_s(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\right)\right\|^2 \\
&= \left\|\sum_{i=t}^{t+\tau-1}\sqrt{P_i}\left(\sqrt{P_i}\left(\sum_{s\in\mathcal{S}}\frac{p_i^s w_i^s\pi(s)}{P_i}\nabla f_s(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\right)\right)\right\|^2 \\
&\le \left(\sum_{i=t}^{t+\tau-1}P_i\right)\left(\sum_{i'=t}^{t+\tau-1}P_{i'}\left\|\sum_{s\in\mathcal{S}}\frac{p_{i'}^s w_{i'}^s\pi(s)}{P_{i'}}\nabla f_s(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\right\|^2\right) \\
&= \left(\sum_{i=t}^{t+\tau-1}P_i\right)\left(\sum_{i'=t}^{t+\tau-1}P_{i'}\left\|\sum_{s\in\mathcal{S}}\left(\frac{p_{i'}^s w_{i'}^s\pi(s)}{P_{i'}} - \pi(s)\right)\nabla f_s(\mathbf{x}_t)\right\|^2\right) \\
&= \left(\sum_{i=t}^{t+\tau-1}P_i\right)\left(\sum_{i'=t}^{t+\tau-1}P_{i'}\left\|\sum_{s\in\mathcal{S}}\left(\frac{p_{i'}^s w_{i'}^s\pi(s)}{P_{i'}} - \pi(s)\right)\left(\nabla f_s(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\right)\right\|^2\right) \\
&\le \left(\sum_{i=t}^{t+\tau-1}P_i\right)\left(\sum_{i'=t}^{t+\tau-1}P_{i'}\left(\sum_{s\in\mathcal{S}}\left|\frac{p_{i'}^s w_{i'}^s\pi(s)}{P_{i'}} - \pi(s)\right|\left\|\nabla f_s(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\right\|\right)^2\right) \\
&\le \left(\sum_{i=t}^{t+\tau-1}P_i\right)\left(\sum_{i'=t}^{t+\tau-1}P_{i'}\left(\sum_{s\in\mathcal{S}}\left|\frac{p_{i'}^s w_{i'}^s\pi(s)}{P_{i'}} - \pi(s)\right|\varsigma\right)^2\right), \\
&\le \sum_{i=t}^{t+\tau-1}P_i\sum_{i'=t}^{t+\tau-1}P_{i'}\phi_{i'}^2,
\end{aligned}
$$

where the first inequality uses $\left\|\sum_{i=1}^n \mathbf{a}_i\right\|^2 = 2\sum_{i=1}^n\sum_{j>i}^n\langle\mathbf{a}_i,\mathbf{a}_j\rangle + \sum_{i=1}^n\|\mathbf{a}_i\|^2$ for any $\mathbf{a}_i\in\mathbb{R}^d$ and $n\ge 1$, the third equality uses Equation (14), the fourth equality uses Equation (7) (definition of P_i), the second inequality uses Cauchy–Schwarz inequality, the fourth inequality uses Assumption 3, and the fifth inequality uses Equation (9) (definition of ϕ_t).

For $\sum_{i=t}^{t+\tau-1}\mathcal{T}_5$, we have

$$
\sum_{i=t}^{t+\tau-1}\mathcal{T}_5 = \sum_{i=t}^{t+\tau-1}\sum_{s\in\mathcal{S}}\left(\mathbb{P}\left(s_i = s \mid s_t\right) - \pi(s)\right)\|h_i^s\|^2
$$

$$
\begin{aligned}
&\leq \sum_{i=t}^{t+\tau-1} \sum_{s\in\mathcal{S}} \left|\mathbb{P}\big(s_i = s \mid s_t\big) - \pi(s)\right| \left\|h_i^s\right\|^2 \\
&\leq 2 \sum_{i=t}^{t+\tau-1} d_{\mathrm{TV}}(M^{i-t}(s_t,\cdot),\pi) A_3^2 \left(\varsigma + 2\left\|\nabla f(\mathbf{x}_t)\right\|\right)^2 \\
&\leq 6A_3^2\tau_{\mathrm{mix}} \left(\varsigma + 2\left\|\nabla f(\mathbf{x}_t)\right\|\right)^2 \\
&\leq 18A_3^2\tau_{\mathrm{mix}}\varsigma^2 + 36A_3^2\tau_{\mathrm{mix}}\left\|\nabla f(\mathbf{x}_t)\right\|^2,
\end{aligned}
$$

where the second inequality uses Equation (18) and then Equation (15), and the third inequality uses Lemma 2.

For $\sum_{i=t}^{t+\tau-1} \mathcal{T}_6$, we have

$$
\begin{aligned}
\sum_{i=t}^{t+\tau-1} \mathcal{T}_6 &= \sum_{i=t}^{t+\tau-1} \sum_{s\in\mathcal{S}} \pi(s) \left\|h_i^s\right\|^2 \\
&\leq \sum_{i=t}^{t+\tau-1} P_i A_3 \left(\varsigma + 2\left\|\nabla f(\mathbf{x}_t)\right\|\right)^2 \\
&\leq 3 \sum_{i=t}^{t+\tau-1} P_i A_3 \varsigma^2 + 6 \sum_{i=t}^{t+\tau-1} P_i A_3 \left\|\nabla f(\mathbf{x}_t)\right\|^2,
\end{aligned}
$$

where the first inequality uses Equations (18) and (19).

Substituting the preceding bounds for $\mathcal{T}_1$ to $\mathcal{T}_6$ into Equation (16), we obtain

$$
\begin{aligned}
&\mathbb{E}_t \left\| \sum_{i=t}^{t+\tau-1} \left(p_i^{s_i} w_i^{s_i} \nabla f_{s_i}(\mathbf{x}_t) - P_i \nabla f(\mathbf{x}_t)\right) \right\|^2 \\
&\leq 234A_3^2\tau_{\mathrm{mix}}^2\varsigma^2 + 468A_3^2\tau_{\mathrm{mix}}^2 \left\|\nabla f(\mathbf{x}_t)\right\|^2 \\
&\quad + 75 \sum_{i=t}^{t+\tau-1} P_i A_3 \tau_{\mathrm{mix}} \varsigma^2 + 150 \sum_{i=t}^{t+\tau-1} P_i A_3 \tau_{\mathrm{mix}} \left\|\nabla f(\mathbf{x}_t)\right\|^2 + \sum_{i=t}^{t+\tau-1} P_i \sum_{i'=t}^{t+\tau-1} P_{i'} \phi_{i'}^2.
\end{aligned}
$$

Taking unconditional expectation on both sides yields the claim. □

In Lemma 4, we examine the parameter deviation over any period. Notably, the universal constant c_1 in Equation (20) can be found as long as γ is small enough, which will be discussed in Step 5 in the proof of Theorem 1 (Appendix A.3).

Lemma 4. *In the same settings as Theorem 1, assuming that there exists one positive universal constant $c_1 > 1$ such that for any $q \in \{0, 1, \ldots, Q\}$*

$$
\begin{aligned}
&1 - 5\gamma^2 L^2 A_1^2 \left(\sum_{t=l_q}^{l_q+\tau_q-1} P_t\right)^2 - 5\gamma^2 L^2 \left(\sum_{t=l_q}^{l_q+\tau_q-1} P_t\right)^2 \\
&\quad - 4680\gamma^2 L^2 \tau_{\mathrm{mix}}^2 A_3^2 - 1500\gamma^2 L^2 \tau_{\mathrm{mix}} A_3 \sum_{t=l_q}^{l_q+\tau_q-1} P_t \geq \frac{1}{c_1},
\end{aligned} \tag{20}
$$

then for any $q \in \{0, 1, \ldots, Q\}$ and $l_q \leq t \leq l_q + \tau_q - 1$, we have

$$
\begin{aligned}
&\max_{l_q \leq i \leq t} \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2 \\
&\leq 10c_1\gamma^2 \sum_{i=l_q}^{t-1} P_i A_2 \varsigma^2 + 1170c_1\gamma^2 A_3^2 \tau_{\mathrm{mix}}^2 \varsigma^2 + 375c_1\gamma^2 \sum_{i=l_q}^{t-1} P_i A_3 \tau_{\mathrm{mix}} \varsigma^2 + 5c_1\gamma^2 \sum_{i=l_q}^{t-1} P_i \sum_{i'=l_q}^{t-1} P_{i'} \phi_{i'}^2 \\
&\quad + 10c_1\gamma^2 A_2 \sum_{i=l_q}^{t-1} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 + 5c_1\gamma^2 \sum_{i=l_q}^{t-1} P_i \sum_{i'=l_q}^{t-1} P_{i'} \mathbb{E}\|\nabla f(\mathbf{x}_{i'})\|^2
\end{aligned}
$$

$$+ 4680c_1\gamma^2 A_3^2\tau_{\text{mix}}^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + 1500c_1\gamma^2 A_3\tau_{\text{mix}}\sum_{i=l_q}^{t-1} P_i\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \tag{21}$$

and

$$\begin{aligned}
&\sum_{i=l_q}^{t} P_i\mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2 \\
&\quad\le 10c_1\gamma^2 A_2\left(\sum_{i=l_q}^{t} P_i\right)^2\varsigma^2 + 1170c_1\gamma^2\tau_{\text{mix}}^2 A_3^2\sum_{i=l_q}^{t} P_i\varsigma^2 \\
&\qquad + 375c_1\gamma^2\tau_{\text{mix}}A_3\left(\sum_{i=l_q}^{t} P_i\right)^2\varsigma^2 + 5c_1\gamma^2\left(\sum_{i=l_q}^{t} P_i\right)^2\sum_{i'=l_q}^{t} P_{i'}\phi_{i'}^2 \\
&\qquad + 10c_1\gamma^2 A_2\sum_{i=l_q}^{t} P_i\sum_{i'=l_q}^{t} P_{i'}\mathbb{E}\|\nabla f(\mathbf{x}_{i'})\|^2 + 5c_1\gamma^2\left(\sum_{i=l_q}^{t} P_i\right)^2\sum_{i'=l_q}^{t} P_{i'}\mathbb{E}\|\nabla f(\mathbf{x}_{i'})\|^2 \\
&\qquad + 4680c_1\gamma^2 A_3^2\tau_{\text{mix}}^2\sum_{i=l_q}^{t} P_i\mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 + 1500c_1\gamma^2 A_3\tau_{\text{mix}}\sum_{i=l_q}^{t} P_i\sum_{i'=l_q}^{t} P_{i'}\mathbb{E}\|\nabla f(\mathbf{x}_{i'})\|^2.
\end{aligned} \tag{22}$$

Proof. Given any period $q \in \{0, 1, \ldots, Q\}$, for $l_q + 1 \le i \le t \le l_q + \tau_q - 1$,

$$\begin{aligned}
\mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2 &= \mathbb{E}\left\|\gamma\sum_{j=l_q}^{i-1}\delta_j^{s_j}\, w_j^{s_j}\,\nabla f_{s_j}(\mathbf{x}_j)\right\|^2 \\
&\le \underbrace{5\mathbb{E}\left\|\gamma\sum_{j=l_q}^{i-1}\delta_j^{s_j} w_j^{s_j}\nabla f_{s_j}(\mathbf{x}_j) - \gamma\sum_{j=l_q}^{i-1} p_j^{s_j} w_j^{s_j}\nabla f_{s_j}(\mathbf{x}_j)\right\|^2}_{\mathcal{T}_1} \\
&\quad + \underbrace{5\mathbb{E}\left\|\gamma\sum_{j=l_q}^{i-1} p_j^{s_j} w_j^{s_j}\nabla f_{s_j}(\mathbf{x}_j) - \gamma\sum_{j=l_q}^{i-1} p_j^{s_j} w_j^{s_j}\nabla f_{s_j}(\mathbf{x}_{l_q})\right\|^2}_{\mathcal{T}_2} \\
&\quad + \underbrace{5\mathbb{E}\left\|\gamma\sum_{j=l_q}^{i-1} p_j^{s_j} w_j^{s_j}\nabla f_{s_j}(\mathbf{x}_{l_q}) - \gamma\sum_{j=l_q}^{i-1} P_j\nabla f(\mathbf{x}_{l_q})\right\|^2}_{\mathcal{T}_3} \\
&\quad + \underbrace{5\mathbb{E}\left\|\gamma\sum_{j=l_q}^{i-1} P_j\nabla f(\mathbf{x}_{l_q}) - \gamma\sum_{j=l_q}^{i-1} P_j\nabla f(\mathbf{x}_j)\right\|^2}_{\mathcal{T}_4} \\
&\quad + \underbrace{5\mathbb{E}\left\|\gamma\sum_{j=l_q}^{i-1} P_j\nabla f(\mathbf{x}_j)\right\|^2}_{\mathcal{T}_5}.
\end{aligned} \tag{23}$$

Next, we bound $\mathcal{T}_1$ to $\mathcal{T}_5$. For $\mathcal{T}_1$,

$$\mathcal{T}_1 = 5\mathbb{E}\left\|\gamma\sum_{j=l_q}^{i-1}\delta_j^{s_j} w_j^{s_j}\nabla f_{s_j}(\mathbf{x}_j) - \gamma\sum_{j=l_q}^{i-1} p_j^{s_j} w_j^{s_j}\nabla f_{s_j}(\mathbf{x}_j)\right\|^2$$

$$
\begin{aligned}
&= 5\gamma^2 \sum_{j=l_q}^{i-1} \mathbb{E}\big[(\delta_j^{s_j} - p_j^{s_j})^2 (w_j^{s_j})^2 \|\nabla f_{s_j}(\mathbf{x}_j)\|^2\big] \\
&\le 5\gamma^2 \sum_{j=l_q}^{i-1} \mathbb{E}\big[p_j^{s_j} (w_j^{s_j})^2 \|\nabla f_{s_j}(\mathbf{x}_j)\|^2\big] \\
&= 5\gamma^2 \sum_{j=l_q}^{i-1} P_j \, \mathbb{E}\left[\frac{p_j^{s_j} (w_j^{s_j})^2}{P_j} \|\nabla f_{s_j}(\mathbf{x}_j)\|^2\right] \\
&\le 5\gamma^2 A_2 \sum_{j=l_q}^{i-1} P_j \, \mathbb{E}\|\nabla f_{s_j}(\mathbf{x}_j)\|^2 \\
&\le 10\gamma^2 A_2 \sum_{j=l_q}^{i-1} P_j \, \mathbb{E}\|\nabla f_{s_j}(\mathbf{x}_j) - \nabla f(\mathbf{x}_j)\|^2 + 10\gamma^2 A_2 \sum_{j=l_q}^{i-1} P_j \, \mathbb{E}\|\nabla f(\mathbf{x}_j)\|^2 \\
&\le 10\gamma^2 A_2 \sum_{j=l_q}^{i-1} P_j \varsigma^2 + 10\gamma^2 A_2 \sum_{j=l_q}^{i-1} P_j \, \mathbb{E}\|\nabla f(\mathbf{x}_j)\|^2,
\end{aligned}
$$

where the first inequality uses $\delta_t^s \sim \text{Bernoulli}(p_t^s)$, the second inequality uses Equation (8) (the definition of A_2) and the fourth inequality uses Assumption 3. In particular, the second equality above follows because all cross terms have zero expectation. Indeed, consider any $j < k$ and condition on the state sequence up to iteration k and all availability variables sampled before iteration k. Given these variables, the iterate $\mathbf{x}_k$ and the term corresponding to iteration j are fixed. By Assumption 4, $\delta_k^{s_k}$ is independent of these variables and satisfies $\mathbb{E}[\delta_k^{s_k} - p_k^{s_k}] = 0$. Therefore,

$$
\mathbb{E}\left\langle (\delta_j^{s_j} - p_j^{s_j}) w_j^{s_j} \nabla f_{s_j}(\mathbf{x}_j), (\delta_k^{s_k} - p_k^{s_k}) w_k^{s_k} \nabla f_{s_k}(\mathbf{x}_k) \right\rangle = 0.
$$

Thus, all cross terms vanish, yielding

$$
\mathcal{T}_1 = 5\gamma^2 \sum_{j=l_q}^{i-1} \mathbb{E}\left[(\delta_j^{s_j} - p_j^{s_j})^2 (w_j^{s_j})^2 \left\|\nabla f_{s_j}(\mathbf{x}_j)\right\|^2\right].
$$

For $\mathcal{T}_2$,

$$
\begin{aligned}
\mathcal{T}_2 &= 5\mathbb{E}\left\|\gamma \sum_{j=l_q}^{i-1} p_j^{s_j} w_j^{s_j} \nabla f_{s_j}(\mathbf{x}_j) - \gamma \sum_{j=l_q}^{i-1} p_j^{s_j} w_j^{s_j} \nabla f_{s_j}(\mathbf{x}_{l_q})\right\|^2 \\
&= 5\gamma^2 \mathbb{E}\left\|\sum_{j=l_q}^{i-1} \frac{p_j^{s_j} w_j^{s_j}}{\sqrt{P_j}} \sqrt{P_j}\big(\nabla f_{s_j}(\mathbf{x}_j) - \nabla f_{s_j}(\mathbf{x}_{l_q})\big)\right\|^2 \\
&\le 5\gamma^2 \mathbb{E}\left[\left(\sum_{j=l_q}^{i-1} \frac{(p_j^{s_j} w_j^{s_j})^2}{P_j}\right)\left(\sum_{j'=l_q}^{i-1} P_{j'} \|\nabla f_{s_{j'}}(\mathbf{x}_{j'}) - \nabla f_{s_{j'}}(\mathbf{x}_{l_q})\|^2\right)\right] \\
&\le 5\gamma^2 A_1^2 \sum_{j=l_q}^{i-1} P_j \sum_{j'=l_q}^{i-1} P_{j'} \mathbb{E}\|\nabla f_{s_{j'}}(\mathbf{x}_{j'}) - \nabla f_{s_{j'}}(\mathbf{x}_{l_q})\|^2 \\
&\le 5\gamma^2 L^2 A_1^2 \sum_{j=l_q}^{i-1} P_j \sum_{j'=l_q}^{i-1} P_{j'} \mathbb{E}\|\mathbf{x}_{j'} - \mathbf{x}_{l_q}\|^2,
\end{aligned}
$$

where the first inequality uses Cauchy–Schwarz inequality, the second inequality uses Equation (8) (the definition of A_1) and the third inequality uses Assumption 2 (L-smoothness).

For $\mathcal{T}_3$, applying Lemma 3, we obtain

$$
\mathcal{T}_3 = 5\mathbb{E}\left\|\gamma \sum_{j=l_q}^{i-1} p_j^{s_j} w_j^{s_j} \nabla f_{s_j}(\mathbf{x}_{l_q}) - \gamma \sum_{j=l_q}^{i-1} P_j \nabla f(\mathbf{x}_{l_q})\right\|^2
$$

$$\leq 1170\gamma^2\tau_{\text{mix}}^2 A_3^2\varsigma^2 + 2340\gamma^2\tau_{\text{mix}}^2 A_3^2\mathbb{E}\|\nabla f(\mathbf{x}_{l_q})\|^2 + 375\gamma^2\sum_{j=l_q}^{i-1} P_j A_3\tau_{\text{mix}}\varsigma^2 + 750\gamma^2\sum_{j=l_q}^{i-1} P_j A_3\tau_{\text{mix}}\mathbb{E}\|\nabla f(\mathbf{x}_{l_q})\|^2 + 5\gamma^2\sum_{j=l_q}^{i-1} P_j\sum_{j'=l_q}^{i-1} P_{j'}\phi_{j'}^2.$$

For $\mathcal{T}_4$,

$$\begin{aligned}\mathcal{T}_4 &= 5\mathbb{E}\left\|\gamma\sum_{j=l_q}^{i-1} P_j\nabla f(\mathbf{x}_{l_q}) - \gamma\sum_{j=l_q}^{i-1} P_j\nabla f(\mathbf{x}_j)\right\|^2 \\ &\leq 5\gamma^2\sum_{j=l_q}^{i-1} P_j\sum_{j'=l_q}^{i-1} P_{j'}\mathbb{E}\|\nabla f(\mathbf{x}_{l_q}) - \nabla f(\mathbf{x}_{j'})\|^2 \\ &\leq 5\gamma^2 L^2\sum_{j=l_q}^{i-1} P_j\sum_{j'=l_q}^{i-1} P_{j'}\mathbb{E}\|\mathbf{x}_{j'} - \mathbf{x}_{l_q}\|^2,\end{aligned}$$

where the first inequality uses Cauchy–Schwarz inequality and the second inequality uses Assumption 2 (L-smoothness).

For $\mathcal{T}_5$, similar to $\mathcal{T}_4$, using Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}\mathcal{T}_5 &= 5\mathbb{E}\left\|\gamma\sum_{j=l_q}^{i-1} P_j\nabla f(\mathbf{x}_j)\right\|^2 \\ &\leq 5\gamma^2\sum_{j=l_q}^{i-1} P_j\sum_{j'=l_q}^{i-1} P_{j'}\mathbb{E}\|\nabla f(\mathbf{x}_{j'})\|^2.\end{aligned}$$

Substituting the preceding bounds for $\mathcal{T}_1$ to $\mathcal{T}_5$ into Equation (23), we obtain, for $l_q + 1 \leq i \leq t \leq l_q + \tau_q - 1$,

$$\begin{aligned}&\mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2 \\ &\leq 10\gamma^2\sum_{j=l_q}^{i-1} P_j A_2\varsigma^2 + 1170\gamma^2 A_3^2\tau_{\text{mix}}^2\varsigma^2 + 375\gamma^2\sum_{j=l_q}^{i-1} P_j A_3\tau_{\text{mix}}\varsigma^2 + 5\gamma^2\sum_{j=l_q}^{i-1} P_j\sum_{j'=l_q}^{i-1} P_{j'}\phi_{j'}^2 \\ &\quad + 10\gamma^2 A_2\sum_{j=l_q}^{i-1} P_j\mathbb{E}\|\nabla f(\mathbf{x}_j)\|^2 + 5\gamma^2\sum_{j=l_q}^{i-1} P_j\sum_{j'=l_q}^{i-1} P_{j'}\mathbb{E}\|\nabla f(\mathbf{x}_{j'})\|^2 \\ &\quad + 2340\gamma^2 A_3^2\tau_{\text{mix}}^2\mathbb{E}\|\nabla f(\mathbf{x}_{l_q})\|^2 + 750\gamma^2 A_3\tau_{\text{mix}}\sum_{j=l_q}^{i-1} P_j\mathbb{E}\|\nabla f(\mathbf{x}_{l_q})\|^2 \\ &\quad + 5\gamma^2 L^2 A_1^2\sum_{j=l_q}^{i-1} P_j\sum_{j'=l_q}^{i-1} P_{j'}\mathbb{E}\|\mathbf{x}_{j'} - \mathbf{x}_{l_q}\|^2 + 5\gamma^2 L^2\sum_{j=l_q}^{i-1} P_j\sum_{j'=l_q}^{i-1} P_{j'}\mathbb{E}\|\mathbf{x}_{j'} - \mathbf{x}_{l_q}\|^2. \quad (24)\end{aligned}$$

Different treatments of the term $\mathbb{E}\|\nabla f(\mathbf{x}_{l_q})\|^2$ in Equation (24) lead to the following two useful results.

Proof of Equation (21). Substituting

$$\begin{aligned}\mathbb{E}\left\|\nabla f(\mathbf{x}_{l_q})\right\|^2 &\leq 2\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2 + 2\mathbb{E}\left\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{l_q})\right\|^2 \\ &\leq 2\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2 + 2L^2\mathbb{E}\left\|\mathbf{x}_t - \mathbf{x}_{l_q}\right\|^2,\end{aligned}$$

into Equation (24), we obtain, for $l_q + 1 \leq i \leq t \leq l_q + \tau_q - 1$,

$$\begin{aligned}&\mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2 \\ &\leq 10\gamma^2\sum_{j=l_q}^{i-1} P_j A_2\varsigma^2 + 1170\gamma^2 A_3^2\tau_{\text{mix}}^2\varsigma^2 + 375\gamma^2\sum_{j=l_q}^{i-1} P_j A_3\tau_{\text{mix}}\varsigma^2 + 5\gamma^2\sum_{j=l_q}^{i-1} P_j\sum_{j'=l_q}^{i-1} P_{j'}\phi_{j'}^2\end{aligned}$$

$$
\begin{aligned}
&+ 10\gamma^2 A_2 \sum_{j=l_q}^{i-1} P_j \mathbb{E}\|\nabla f(\mathbf{x}_j)\|^2 + 5\gamma^2 \sum_{j=l_q}^{i-1} P_j \sum_{j'=l_q}^{i-1} P_{j'} \mathbb{E}\|\nabla f(\mathbf{x}_{j'})\|^2 \\
&+ 4680\gamma^2 \tau_{\text{mix}}^2 A_3^2 \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + 1500\gamma^2 \tau_{\text{mix}} A_3 \sum_{j=l_q}^{i-1} P_j \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \\
&+ 4680\gamma^2 L^2 \tau_{\text{mix}}^2 A_3^2 \mathbb{E}\|\mathbf{x}_t - \mathbf{x}_{l_q}\|^2 + 1500\gamma^2 L^2 \tau_{\text{mix}} A_3 \sum_{j=l_q}^{i-1} P_j \mathbb{E}\|\mathbf{x}_t - \mathbf{x}_{l_q}\|^2 \\
&+ 5\gamma^2 L^2 A_1^2 \sum_{j=l_q}^{i-1} P_j \sum_{j'=l_q}^{i-1} P_{j'} \mathbb{E}\|\mathbf{x}_{j'} - \mathbf{x}_{l_q}\|^2 + 5\gamma^2 L^2 \sum_{j=l_q}^{i-1} P_j \sum_{j'=l_q}^{i-1} P_{j'} \mathbb{E}\|\mathbf{x}_{j'} - \mathbf{x}_{l_q}\|^2.
\end{aligned}
$$

Moreover, the preceding inequality holds trivially for $i = l_q$. Thus, we have

$$
\begin{aligned}
&\max_{l_q \le i \le t} \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2 \\
&\quad \le 10\gamma^2 \sum_{j=l_q}^{t-1} P_j A_2 \varsigma^2 + 1170\gamma^2 A_3^2 \tau_{\text{mix}}^2 \varsigma^2 + 375\gamma^2 \sum_{j=l_q}^{t-1} P_j A_3 \tau_{\text{mix}} \varsigma^2 + 5\gamma^2 \sum_{j=l_q}^{t-1} P_j \sum_{j'=l_q}^{t-1} P_{j'} \phi_{j'}^2 \\
&\qquad + 10\gamma^2 A_2 \sum_{j=l_q}^{t-1} P_j \mathbb{E}\|\nabla f(\mathbf{x}_j)\|^2 + 5\gamma^2 \sum_{j=l_q}^{t-1} P_j \sum_{j'=l_q}^{t-1} P_{j'} \mathbb{E}\|\nabla f(\mathbf{x}_{j'})\|^2 \\
&\qquad + 4680\gamma^2 \tau_{\text{mix}}^2 A_3^2 \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + 1500\gamma^2 \tau_{\text{mix}} A_3 \sum_{j=l_q}^{t-1} P_j \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \\
&\qquad + 4680\gamma^2 L^2 \tau_{\text{mix}}^2 A_3^2 \max_{l_q \le i \le t} \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2 + 1500\gamma^2 L^2 \tau_{\text{mix}} A_3 \sum_{j=l_q}^{t-1} P_j \max_{l_q \le i \le t} \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2 \\
&\qquad + 5\gamma^2 L^2 A_1^2 \left(\sum_{j=l_q}^{t-1} P_j \right)^2 \max_{l_q \le i \le t} \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2 + 5\gamma^2 L^2 \left(\sum_{j=l_q}^{t-1} P_j \right)^2 \max_{l_q \le i \le t} \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2.
\end{aligned}
$$

Rearranging and collecting the terms involving $\max_{l_q \le i \le t} \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2$ on the left hand side, we obtain, for $l_q \le i \le t \le l_q + \tau_q - 1$,

$$
\begin{aligned}
&\left(1 - 5\gamma^2 L^2 A_1^2 \left(\sum_{j=l_q}^{t-1} P_j \right)^2 - 5\gamma^2 L^2 \left(\sum_{j=l_q}^{t-1} P_j \right)^2 \right. \\
&\qquad \left. - 4680\gamma^2 L^2 \tau_{\text{mix}}^2 A_3^2 - 1500\gamma^2 L^2 \tau_{\text{mix}} A_3 \sum_{j=l_q}^{t-1} P_j \right) \max_{l_q \le i \le t} \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2 \\
&\quad \le 10\gamma^2 \sum_{j=l_q}^{t-1} P_j A_2 \varsigma^2 + 1170\gamma^2 A_3^2 \tau_{\text{mix}}^2 \varsigma^2 + 375\gamma^2 \sum_{j=l_q}^{t-1} P_j A_3 \tau_{\text{mix}} \varsigma^2 + 5\gamma^2 \sum_{j=l_q}^{t-1} P_j \sum_{j'=l_q}^{t-1} P_{j'} \phi_{j'}^2 \\
&\qquad + 10\gamma^2 A_2 \sum_{j=l_q}^{t-1} P_j \mathbb{E}\|\nabla f(\mathbf{x}_j)\|^2 + 5\gamma^2 \sum_{j=l_q}^{t-1} P_j \sum_{j'=l_q}^{t-1} P_{j'} \mathbb{E}\|\nabla f(\mathbf{x}_{j'})\|^2 \\
&\qquad + 4680\gamma^2 A_3^2 \tau_{\text{mix}}^2 \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + 1500\gamma^2 A_3 \tau_{\text{mix}} \sum_{j=l_q}^{t-1} P_j \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2.
\end{aligned}
$$

Using $t \le l_q + \tau_q - 1$ and the condition Equation (20), we obtain that the coefficient of $\max_{l_q \le i \le t} \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2$ on the left hand side is larger than $\frac{1}{c_1}$ ($c_1 > 1$). Then, dividing both sides by $\frac{1}{c_1}$, we obtain Equation (21).

Proof of Equation (22). Substituting

$$
\mathbb{E}\left\|\nabla f(\mathbf{x}_{l_q})\right\|^2 \le 2\mathbb{E}\left\|\nabla f(\mathbf{x}_i)\right\|^2 + 2\mathbb{E}\left\|\nabla f(\mathbf{x}_i) - \nabla f(\mathbf{x}_{l_q})\right\|^2
$$

$$\leq 2\mathbb{E}\left\|\nabla f(\mathbf{x}_i)\right\|^2 + 2L^2\mathbb{E}\left\|\mathbf{x}_i - \mathbf{x}_{l_q}\right\|^2,$$

into Equation (24), we obtain, for $l_q + 1 \leq i \leq t \leq l_q + \tau_q - 1$,

$$\begin{aligned}
&\mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2 \\
&\leq 10\gamma^2 A_2 \sum_{j=l_q}^{i-1} P_j \varsigma^2 + 1170\gamma^2\tau_{\text{mix}}^2 A_3^2 \varsigma^2 + 375\gamma^2\tau_{\text{mix}} A_3 \sum_{j=l_q}^{i-1} P_j \varsigma^2 + 5\gamma^2 \sum_{j=l_q}^{i-1} P_j \sum_{j'=l_q}^{i-1} P_{j'} \phi_{j'}^2 \\
&\quad + 10\gamma^2 A_2 \sum_{j=l_q}^{i-1} P_j \mathbb{E}\|\nabla f(\mathbf{x}_j)\|^2 + 5\gamma^2 \sum_{j=l_q}^{i-1} P_j \sum_{j'=l_q}^{i-1} P_{j'} \mathbb{E}\|\nabla f(\mathbf{x}_{j'})\|^2 \\
&\quad + 4680\gamma^2\tau_{\text{mix}}^2 A_3^2 \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 + 1500\gamma^2\tau_{\text{mix}} A_3 \sum_{j=l_q}^{i-1} P_j \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 \\
&\quad + 4680\gamma^2 L^2\tau_{\text{mix}}^2 A_3^2 \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2 + 1500\gamma^2 L^2\tau_{\text{mix}} A_3 \sum_{j=l_q}^{i-1} P_j \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2 \\
&\quad + 5\gamma^2 L^2 A_1^2 \sum_{j=l_q}^{i-1} P_j \sum_{j'=l_q}^{i-1} P_{j'} \mathbb{E}\|\mathbf{x}_{j'} - \mathbf{x}_{l_q}\|^2 + 5\gamma^2 L^2 \sum_{j=l_q}^{i-1} P_j \sum_{j'=l_q}^{i-1} P_{j'} \mathbb{E}\|\mathbf{x}_{j'} - \mathbf{x}_{l_q}\|^2.
\end{aligned}$$

Moreover, the preceding inequality holds trivially for $i = l_q$. Taking a weighted average with weights P_i over $l_q \leq i \leq t$, we obtain

$$\begin{aligned}
&\sum_{i=l_q}^{t} P_i \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2 \\
&\leq 10\gamma^2 A_2 \left(\sum_{j=l_q}^{t} P_j\right)^2 \varsigma^2 + 1170\gamma^2\tau_{\text{mix}}^2 A_3^2 \sum_{j=l_q}^{t} P_j \varsigma^2 \\
&\quad + 375\gamma^2\tau_{\text{mix}} A_3 \left(\sum_{j=l_q}^{t} P_j\right)^2 \varsigma^2 + 5\gamma^2 \left(\sum_{j=l_q}^{t} P_j\right)^2 \sum_{j'=l_q}^{t} P_{j'} \phi_{j'}^2 \\
&\quad + 10\gamma^2 A_2 \sum_{j=l_q}^{t} P_j \sum_{j'=l_q}^{t} P_{j'} \mathbb{E}\|\nabla f(\mathbf{x}_{j'})\|^2 + 5\gamma^2 \left(\sum_{j=l_q}^{t} P_j\right)^2 \sum_{j'=l_q}^{t} P_{j'} \mathbb{E}\|\nabla f(\mathbf{x}_{j'})\|^2 \\
&\quad + 4680\gamma^2 A_3^2 \tau_{\text{mix}}^2 \sum_{j=l_q}^{t} P_j \mathbb{E}\|\nabla f(\mathbf{x}_j)\|^2 + 1500\gamma^2 A_3 \tau_{\text{mix}} \sum_{j=l_q}^{t} P_j \sum_{j'=l_q}^{t} P_{j'} \mathbb{E}\|\nabla f(\mathbf{x}_{j'})\|^2 \\
&\quad + 4680\gamma^2 L^2 A_3^2 \tau_{\text{mix}}^2 \sum_{j=l_q}^{t} P_j \mathbb{E}\|\mathbf{x}_j - \mathbf{x}_{l_q}\|^2 + 1500\gamma^2 L^2 A_3 \tau_{\text{mix}} \sum_{j=l_q}^{t} P_j \sum_{j'=l_q}^{t} P_{j'} \mathbb{E}\|\mathbf{x}_{j'} - \mathbf{x}_{l_q}\|^2 \\
&\quad + 5\gamma^2 L^2 A_1^2 \left(\sum_{j=l_q}^{t} P_j\right)^2 \sum_{j'=l_q}^{t} P_{j'} \mathbb{E}\|\mathbf{x}_{j'} - \mathbf{x}_{l_q}\|^2 + 5\gamma^2 L^2 \left(\sum_{j=l_q}^{t} P_j\right)^2 \sum_{j'=l_q}^{t} P_{j'} \mathbb{E}\|\mathbf{x}_{j'} - \mathbf{x}_{l_q}\|^2.
\end{aligned}$$

Rearranging and collecting the terms involving $\sum_{i=l_q}^{t} P_i \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2$ on the left hand side, we obtain, for $l_q \leq i \leq t \leq l_q + \tau_q - 1$,

$$\begin{aligned}
&\Bigg(1 - 5\gamma^2 L^2 A_1^2 \left(\sum_{j=l_q}^{t} P_j\right)^2 - 5\gamma^2 L^2 \left(\sum_{j=l_q}^{t} P_j\right)^2 \\
&\qquad - 4680\gamma^2 L^2 \tau_{\text{mix}}^2 A_3^2 - 1500\gamma^2 L^2 \tau_{\text{mix}} A_3 \sum_{j=l_q}^{t} P_j\Bigg) \sum_{i=l_q}^{t} P_i \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2
\end{aligned}$$

$$
\begin{aligned}
&\leq 10\gamma^2 A_2 \left(\sum_{j=l_q}^{t} P_j\right)^2 \varsigma^2 + 1170\gamma^2 \tau_{\text{mix}}^2 A_3^2 \sum_{j=l_q}^{t} P_j \varsigma^2 \\
&\quad + 375\gamma^2 \tau_{\text{mix}} A_3 \left(\sum_{j=l_q}^{t} P_j\right)^2 \varsigma^2 + 5\gamma^2 \left(\sum_{j=l_q}^{t} P_j\right)^2 \sum_{j'=l_q}^{t} P_{j'} \phi_{j'}^2 \\
&\quad + 10\gamma^2 A_2 \sum_{j=l_q}^{t} P_j \sum_{j'=l_q}^{t} P_{j'} \mathbb{E}\|\nabla f(\mathbf{x}_{j'})\|^2 + 5\gamma^2 \left(\sum_{j=l_q}^{t} P_j\right)^2 \sum_{j'=l_q}^{t} P_{j'} \mathbb{E}\|\nabla f(\mathbf{x}_{j'})\|^2 \\
&\quad + 4680\gamma^2 A_3^2 \tau_{\text{mix}}^2 \sum_{j=l_q}^{t} P_j \mathbb{E}\|\nabla f(\mathbf{x}_j)\|^2 + 1500\gamma^2 A_3 \tau_{\text{mix}} \sum_{j=l_q}^{t} P_j \sum_{j'=l_q}^{t} P_{j'} \mathbb{E}\|\nabla f(\mathbf{x}_{j'})\|^2.
\end{aligned}
$$

Using $t \leq l_q + \tau_q - 1$ and the condition Equation (20), we obtain that the coefficient of $\sum_{i=l_q}^{t} P_i \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l_q}\|^2$ on the left hand side is larger than $\frac{1}{c_1}$ ($c_1 > 1$). Then, dividing both sides by $\frac{1}{c_1}$, we obtain Equation (22). □

A.3 Proof of Theorem 1

Proof of Theorem 1. For convenience, we introduce several universal constants c_1 to c_5 throughout the following proof and specify their values at the end.

Step 1: Overall decomposition.

According to the update rules (that is, Equation (10) and Equation (12)), using L-smoothness, we obtain that, if $t+1 \notin \mathcal{L}_T$ (Case 1),

$$
\begin{aligned}
\mathbb{E}[f(\tilde{\mathbf{x}}_{t+1}) - f(\tilde{\mathbf{x}}_t)] &\leq \mathbb{E}\langle \nabla f(\tilde{\mathbf{x}}_t), \tilde{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_t \rangle + \frac{L}{2}\mathbb{E}\|\tilde{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_t\|^2 \\
&= \mathbb{E}\langle \nabla f(\tilde{\mathbf{x}}_t), -\gamma P_t \nabla f(\mathbf{x}_t) \rangle + \frac{1}{2} L \mathbb{E}\|\gamma P_t \nabla f(\mathbf{x}_t)\|^2 \\
&\leq \underbrace{\mathbb{E}\langle \nabla f(\tilde{\mathbf{x}}_t), -\gamma P_t \nabla f(\mathbf{x}_t) \rangle + L\mathbb{E}\|\gamma P_t \nabla f(\mathbf{x}_t)\|^2}_{C_t^{\text{com}}},
\end{aligned}
$$

and that if $t+1 \in \mathcal{L}_T$ (Case 2),

$$
\begin{aligned}
\mathbb{E}\big[f(\tilde{\mathbf{x}}_{t+1}) - f(\tilde{\mathbf{x}}_t)\big] &\leq \mathbb{E}\langle \nabla f(\tilde{\mathbf{x}}_t), \tilde{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_t \rangle + \frac{L}{2}\mathbb{E}\|\tilde{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_t\|^2 \\
&\leq \underbrace{\mathbb{E}\langle \nabla f(\tilde{\mathbf{x}}_t), -\gamma P_t \nabla f(\mathbf{x}_t) \rangle + L\mathbb{E}\|\gamma P_t \nabla f(\mathbf{x}_t)\|^2}_{C_t^{\text{com}}} \\
&\quad + \mathbb{E}\left\langle \nabla f(\tilde{\mathbf{x}}_t), -\gamma \sum_{i=l(t)}^{t} (\delta_i^{s_i} w_i^{s_i} \nabla f_{s_i}(\mathbf{x}_i) - P_i \nabla f(\mathbf{x}_i)) \right\rangle \\
&\quad \underbrace{+ L\mathbb{E}\left\| \gamma \sum_{i=l(t)}^{t} (\delta_i^{s_i} w_i^{s_i} \nabla f_{s_i}(\mathbf{x}_i) - P_i \nabla f(\mathbf{x}_i)) \right\|^2}_{C_t^{\text{add}}}
\end{aligned}
$$

Next, we analyze the common part C_t^{com} in both cases and the additional part C_t^{add} in Case 2.

Step 2: Bounding the common part C_t^{com} in Cases 1 and 2.

For C_t^{com}, we have

$$
\begin{aligned}
C_t^{\text{com}} &= \mathbb{E}\langle \nabla f(\tilde{\mathbf{x}}_t), -\gamma P_t \nabla f(\mathbf{x}_t) \rangle + L\mathbb{E}\|\gamma P_t \nabla f(\mathbf{x}_t)\|^2 \\
&= -\frac{1}{2}\gamma P_t \mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 - \frac{1}{2}\gamma P_t \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2}\gamma P_t \mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}_t)\|^2 \\
&\quad + \gamma^2 L P_t^2 \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2
\end{aligned}
$$

$$\leq -\frac{1}{2}\gamma P_t \mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 - \frac{1}{2}\gamma P_t \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2}\gamma L^2 P_t \underbrace{\mathbb{E}\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2}_{\mathcal{T}_1}$$
$$+ \gamma^2 L P_t^2 \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2, \tag{25}$$

where the second equality uses $\langle \mathbf{a}, \mathbf{b}\rangle = \frac{1}{2}\|\mathbf{a}\|^2 + \frac{1}{2}\|\mathbf{b}\|^2 - \frac{1}{2}\|\mathbf{a}-\mathbf{b}\|^2$ for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, the first inequality uses Assumption 2 (L-smoothness). According to the update rules (specifically, Equation (11) and Equation (13)), we obtain, for any t,

$$\mathbf{x}_t - \tilde{\mathbf{x}}_t = -\gamma \sum_{i=l(t)}^{t-1} \left(\delta_i^{s_i} w_i^{s_i} \nabla f_{s_i}(\mathbf{x}_i) - P_i \nabla f(\mathbf{x}_i)\right), \tag{26}$$

which is because $l(t) = l(t-1)$ for $t \notin \mathcal{L}_T$ and $l(t) = t > t-1$ for $t \in \mathcal{L}_T$. Given that the case $t = l(t)$ is trivial, we next only consider the case $t \geq l(t)+1$. Following almost the same steps as those in Lemma 4 (specifically, see the derivation of Equation (24)), we obtain

$$\begin{aligned}
\mathcal{T}_1 &= \mathbb{E}\left\|\gamma \sum_{i=l(t)}^{t-1} \left(\delta_i^{s_i} w_i^{s_i} \nabla f_{s_i}(\mathbf{x}_i) - P_i \nabla f(\mathbf{x}_i)\right)\right\|^2 \\
&\leq 8\gamma^2 A_2 \sum_{i=l(t)}^{t-1} P_i \varsigma^2 + 936\gamma^2 \tau_{\text{mix}}^2 A_3^2 \varsigma^2 + 300\gamma^2 \tau_{\text{mix}} A_3 \sum_{i=l(t)}^{t-1} P_i \varsigma^2 + 4\gamma^2 \sum_{i=l(t)}^{t-1} P_i \sum_{i'=l(t)}^{t-1} P_{i'} \phi_{i'}^2 \\
&\quad + 8\gamma^2 A_2 \sum_{i=l(t)}^{t-1} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 \\
&\quad + 1872\gamma^2 \tau_{\text{mix}}^2 A_3^2 \mathbb{E}\|\nabla f(\mathbf{x}_{l(t)})\|^2 + 600\gamma^2 \tau_{\text{mix}} A_3 \sum_{i=l(t)}^{t-1} P_i \mathbb{E}\|\nabla f(\mathbf{x}_{l(t)})\|^2 \\
&\quad + 4\gamma^2 L^2 A_1^2 \sum_{i=l(t)}^{t-1} P_i \sum_{i'=l(t)}^{t-1} P_{i'} \mathbb{E}\|\mathbf{x}_{i'} - \mathbf{x}_{l(t)}\|^2 + 4\gamma^2 L^2 \sum_{i=l(t)}^{t-1} P_i \sum_{i'=l(t)}^{t-1} P_{i'} \mathbb{E}\|\mathbf{x}_{i'} - \mathbf{x}_{l(t)}\|^2.
\end{aligned}$$

Substituting

$$\begin{aligned}
\mathbb{E}\|\nabla f(\mathbf{x}_{l(t)})\|^2 &\leq 2\mathbb{E}\|\nabla f(\mathbf{x}_{l(t)}) - \nabla f(\mathbf{x}_t)\|^2 + 2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \\
&\leq 2L^2 \mathbb{E}\|\mathbf{x}_t - \mathbf{x}_{l(t)}\|^2 + 2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2,
\end{aligned}$$

into the preceding inequality, we obtain

$$\begin{aligned}
\mathcal{T}_1 &\leq 8\gamma^2 A_2 \sum_{i=l(t)}^{t-1} P_i \varsigma^2 + 936\gamma^2 \tau_{\text{mix}}^2 A_3^2 \varsigma^2 + 300\gamma^2 \tau_{\text{mix}} A_3 \sum_{i=l(t)}^{t-1} P_i \varsigma^2 + 4\gamma^2 \sum_{i=l(t)}^{t-1} P_i \sum_{i'=l(t)}^{t-1} P_{i'} \phi_{i'}^2 \\
&\quad + 8\gamma^2 A_2 \sum_{i=l(t)}^{t-1} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 \\
&\quad + 3744\gamma^2 \tau_{\text{mix}}^2 A_3^2 \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + 1200\gamma^2 \tau_{\text{mix}} A_3 \sum_{i=l(t)}^{t-1} P_i \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \\
&\quad + 3744\gamma^2 L^2 \tau_{\text{mix}}^2 A_3^2 \mathbb{E}\|\mathbf{x}_t - \mathbf{x}_{l(t)}\|^2 + 1200\gamma^2 L^2 \tau_{\text{mix}} A_3 \sum_{i=l(t)}^{t-1} P_i \mathbb{E}\|\mathbf{x}_t - \mathbf{x}_{l(t)}\|^2 \\
&\quad + 4\gamma^2 L^2 A_1^2 \sum_{i=l(t)}^{t-1} P_i \sum_{i'=l(t)}^{t-1} P_{i'} \mathbb{E}\|\mathbf{x}_{i'} - \mathbf{x}_{l(t)}\|^2 + 4\gamma^2 L^2 \sum_{i=l(t)}^{t-1} P_i \sum_{i'=l(t)}^{t-1} P_{i'} \mathbb{E}\|\mathbf{x}_{i'} - \mathbf{x}_{l(t)}\|^2.
\end{aligned}$$

By the definition of $l(t)$ and $\tau(t)$, we have $l(t) \leq t \leq l(t) + \tau(t) - 1$. This allows us to apply Equation (21) in Lemma 4 to the last four terms in the preceding inequality: Assuming that there

exists one universal constant c_2 (decided later) such that for any period $q \in \{0, 1, \ldots, Q\}$,

$$4\gamma^2 L^2 A_1^2 \left(\sum_{t=l_q}^{l_q+\tau_q-1} P_t \right)^2 + 4\gamma^2 L^2 \left(\sum_{t=l_q}^{l_q+\tau_q-1} P_t \right)^2 + 3744\gamma^2 L^2 \tau_{\text{mix}}^2 A_3^2 + 1200\gamma^2 L^2 \tau_{\text{mix}} A_3 \sum_{t=l_q}^{l_q+\tau_q-1} P_t \leq c_2, \tag{27}$$

then, we obtain

$$\begin{aligned}
\mathcal{T}_1 \leq\ & (8 + 10c_1c_2)\gamma^2 A_2 \sum_{i=l(t)}^{t-1} P_i \varsigma^2 \\
& + (936 + 1170c_1c_2)\gamma^2 \tau_{\text{mix}}^2 A_3^2 \varsigma^2 \\
& + (300 + 375c_1c_2)\gamma^2 \tau_{\text{mix}} A_3 \sum_{i=l(t)}^{t-1} P_i \varsigma^2 \\
& + (4 + 5c_1c_2)\gamma^2 \sum_{i=l(t)}^{t-1} P_i \sum_{i'=l(t)}^{t-1} P_{i'} \phi_{i'}^2 \\
& + (8 + 10c_1c_2)\gamma^2 A_2 \sum_{i=l(t)}^{t-1} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 \\
& + 5c_1c_2\gamma^2 \sum_{i=l(t)}^{t-1} P_i \sum_{i'=l(t)}^{t-1} P_{i'} \mathbb{E}\|\nabla f(\mathbf{x}_{i'})\|^2 \\
& + (3744 + 4680c_1c_2)\gamma^2 \tau_{\text{mix}}^2 A_3^2 \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \\
& + (1200 + 1500c_1c_2)\gamma^2 \tau_{\text{mix}} A_3 \sum_{i=l(t)}^{t-1} P_i \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2.
\end{aligned}$$

Substituting the bound of $\mathcal{T}_1$ (note that $\mathbb{E}\left\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\right\|^2 = \mathcal{T}_1$) into Equation (25), we obtain

$$\begin{aligned}
C_t^{\text{com}} \leq\ & -\frac{1}{2}\gamma P_t \mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 - \frac{1}{2}\gamma P_t \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \gamma^2 L P_t^2 \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \\
& + \frac{1}{2}(8 + 10c_1c_2)\gamma^3 L^2 A_2 P_t \sum_{i=l(t)}^{t-1} P_i \varsigma^2 \\
& + \frac{1}{2}(936 + 1170c_1c_2)\gamma^3 L^2 \tau_{\text{mix}}^2 A_3^2 P_t \varsigma^2 \\
& + \frac{1}{2}(300 + 375c_1c_2)\gamma^3 L^2 \tau_{\text{mix}} A_3 P_t \sum_{i=l(t)}^{t-1} P_i \varsigma^2 \\
& + \frac{1}{2}(4 + 5c_1c_2)\gamma^3 L^2 P_t \sum_{i=l(t)}^{t-1} P_i \sum_{i'=l(t)}^{t-1} P_{i'} \phi_{i'}^2 \\
& + \frac{1}{2}(8 + 10c_1c_2)\gamma^3 L^2 A_2 P_t \sum_{i=l(t)}^{t-1} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 \\
& + \frac{5}{2}c_1c_2\gamma^3 L^2 P_t \sum_{i=l(t)}^{t-1} P_i \sum_{i'=l(t)}^{t-1} P_{i'} \mathbb{E}\|\nabla f(\mathbf{x}_{i'})\|^2 \\
& + \frac{1}{2}(3744 + 4680c_1c_2)\gamma^3 L^2 \tau_{\text{mix}}^2 A_3^2 P_t \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2
\end{aligned}$$

$$+\frac{1}{2}(1200+1500c_1c_2)\gamma^3L^2\tau_{\text{mix}}A_3P_t\sum_{i=l(t)}^{t-1}P_i\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2. \tag{28}$$

Step 3: Bounding the additional part C_t^{add} in Case 2.

For C_t^{add}, we have

$$\begin{aligned}
C_t^{\text{add}} &= \mathbb{E}\left\langle \nabla f(\tilde{\mathbf{x}}_t), -\gamma\sum_{i=l(t)}^{t}(\delta_i^{s_i}w_i^{s_i}\nabla f_{s_i}(\mathbf{x}_i)-P_i\nabla f(\mathbf{x}_i))\right\rangle \\
&\quad + L\mathbb{E}\left\|\gamma\sum_{i=l(t)}^{t}(\delta_i^{s_i}w_i^{s_i}\nabla f_{s_i}(\mathbf{x}_i)-P_i\nabla f(\mathbf{x}_i))\right\|^2 \\
&\le \frac{1}{2c_3A_1L}\underbrace{\mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2}_{\mathcal{T}_2}+\left(\frac{1}{2}c_3A_1+1\right)L\underbrace{\mathbb{E}\left\|\gamma\sum_{i=l(t)}^{t}(\delta_i^{s_i}w_i^{s_i}\nabla f_{s_i}(\mathbf{x}_i)-P_i\nabla f(\mathbf{x}_i))\right\|^2}_{\mathcal{T}_3},
\end{aligned} \tag{29}$$

where the inequality uses Young's inequality (the value of the universal constant c_3 will be decided later).

For $\mathcal{T}_2$, we have

$$\begin{aligned}
\frac{1}{2}\mathcal{T}_2 &= \frac{1}{2}\frac{1}{\sum_{i=0}^{\tau(t)-1}P_{t-i}}\sum_{i=0}^{\tau(t)-1}P_{t-i}\mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 \\
&\le \frac{1}{\sum_{i=0}^{\tau(t)-1}P_{t-i}}\sum_{i=0}^{\tau(t)-1}P_{t-i}\mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_t)-\nabla f(\tilde{\mathbf{x}}_{t-i})\|^2 \\
&\quad + \frac{1}{\sum_{i=0}^{\tau(t)-1}P_{t-i}}\sum_{i=0}^{\tau(t)-1}P_{t-i}\mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_{t-i})\|^2 \\
&\le \frac{L^2}{\sum_{i=0}^{\tau(t)-1}P_{t-i}}\sum_{i=0}^{\tau(t)-1}P_{t-i}\mathbb{E}\|\tilde{\mathbf{x}}_t-\tilde{\mathbf{x}}_{t-i}\|^2 \\
&\quad + \frac{1}{\sum_{i=0}^{\tau(t)-1}P_{t-i}}\sum_{i=0}^{\tau(t)-1}P_{t-i}\mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_{t-i})\|^2 \\
&= \frac{L^2}{\sum_{i=0}^{\tau(t)-1}P_{t-i}}\sum_{i=0}^{\tau(t)-1}P_{t-i}\mathbb{E}\left\|\gamma\sum_{j=t-i}^{t-1}P_j\nabla f(\mathbf{x}_j)\right\|^2 \\
&\quad + \frac{1}{\sum_{i=0}^{\tau(t)-1}P_{t-i}}\sum_{i=0}^{\tau(t)-1}P_{t-i}\mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_{t-i})\|^2 \\
&\le \frac{\gamma^2L^2}{\sum_{i=0}^{\tau(t)-1}P_{t-i}}\sum_{i=0}^{\tau(t)-1}P_{t-i}\sum_{j=t-i}^{t-1}P_j\sum_{j'=t-i}^{t-1}P_{j'}\mathbb{E}\left\|\nabla f(\mathbf{x}_{j'})\right\|^2 \\
&\quad + \frac{1}{\sum_{i=0}^{\tau(t)-1}P_{t-i}}\sum_{i=0}^{\tau(t)-1}P_{t-i}\mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_{t-i})\|^2 \\
&\le \gamma^2L^2\sum_{i=0}^{\tau(t)-1}P_{t-i}\sum_{j'=t-i}^{t-1}P_{j'}\mathbb{E}\|\nabla f(\mathbf{x}_{j'})\|^2+\frac{1}{\sum_{i=0}^{\tau(t)-1}P_{t-i}}\sum_{i=0}^{\tau(t)-1}P_{t-i}\mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_{t-i})\|^2
\end{aligned}$$

$$\leq \gamma^2 L^2 \sum_{i=0}^{\tau(t)-1} P_{t-i} \sum_{i'=0}^{\tau(t)-1} P_{t-i'} \mathbb{E}\|\nabla f(\mathbf{x}_{t-i'})\|^2 + \frac{1}{\sum_{i=0}^{\tau(t)-1} P_{t-i}} \sum_{i=0}^{\tau(t)-1} P_{t-i} \mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_{t-i})\|^2,$$

where the second inequality uses Assumption 2 (L-smoothness), the second equality uses Equation (10), and the third inequality uses Cauchy–Schwarz inequality.

For $\mathcal{T}_3$, following almost the same steps as those in Lemma 4 (specifically, see the derivation of Equation (24)), we obtain

$$\begin{aligned}
\mathcal{T}_3 &= \mathbb{E}\left\|\gamma \sum_{i=l(t)}^{t} (\delta_i^{s_i} w_i^{s_i} \nabla f_{s_i}(\mathbf{x}_i) - P_i \nabla f(\mathbf{x}_i))\right\|^2 \\
&\leq 8\gamma^2 A_2 \sum_{i=l(t)}^{t} P_i \varsigma^2 + 936\gamma^2 \tau_{\text{mix}}^2 A_3^2 \varsigma^2 + 300\gamma^2 \tau_{\text{mix}} A_3 \sum_{i=l(t)}^{t} P_i \varsigma^2 + 4\gamma^2 \sum_{i=l(t)}^{t} P_i \sum_{i'=l(t)}^{t} P_{i'} \phi_{i'}^2 \\
&\quad + 8\gamma^2 A_2 \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 \\
&\quad + 1872\gamma^2 \tau_{\text{mix}}^2 A_3^2 \mathbb{E}\|\nabla f(\mathbf{x}_{l(t)})\|^2 + 600\gamma^2 \tau_{\text{mix}} A_3 \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\nabla f(\mathbf{x}_{l(t)})\|^2 \\
&\quad + 4\gamma^2 L^2 A_1^2 \sum_{i=l(t)}^{t} P_i \sum_{i'=l(t)}^{t} P_{i'} \mathbb{E}\|\mathbf{x}_{i'} - \mathbf{x}_{l(t)}\|^2 + 4\gamma^2 L^2 \sum_{i=l(t)}^{t} P_i \sum_{i'=l(t)}^{t} P_{i'} \mathbb{E}\|\mathbf{x}_{i'} - \mathbf{x}_{l(t)}\|^2.
\end{aligned}$$

Substituting

$$\begin{aligned}
&\mathbb{E}\|\nabla f(\mathbf{x}_{l(t)})\|^2 \\
&\quad = \frac{1}{\sum_{i=l(t)}^{t} P_i} \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\nabla f(\mathbf{x}_{l(t)})\|^2 \\
&\quad \leq 2\frac{1}{\sum_{i=l(t)}^{t} P_i} \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 + 2\frac{1}{\sum_{i=l(t)}^{t} P_i} \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\nabla f(\mathbf{x}_{l(t)}) - \nabla f(\mathbf{x}_i)\|^2 \\
&\quad \leq 2\frac{1}{\sum_{i=l(t)}^{t} P_i} \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 + 2L^2\frac{1}{\sum_{i=l(t)}^{t} P_i} \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l(t)}\|^2,
\end{aligned}$$

into the preceding inequality, we obtain

$$\begin{aligned}
\mathcal{T}_3 &\leq 8\gamma^2 A_2 \sum_{i=l(t)}^{t} P_i \varsigma^2 + 936\gamma^2 \tau_{\text{mix}}^2 A_3^2 \varsigma^2 + 300\gamma^2 \tau_{\text{mix}} A_3 \sum_{i=l(t)}^{t} P_i \varsigma^2 + 4\gamma^2 \sum_{i=l(t)}^{t} P_i \sum_{i'=l(t)}^{t} P_{i'} \phi_{i'}^2 \\
&\quad + 8\gamma^2 A_2 \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 \\
&\quad + 3744\gamma^2 \tau_{\text{mix}}^2 A_3^2 \frac{1}{\sum_{i=l(t)}^{t} P_i} \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 + 1200\gamma^2 \tau_{\text{mix}} A_3 \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 \\
&\quad + 3744\gamma^2 L^2 \tau_{\text{mix}}^2 A_3^2 \frac{1}{\sum_{i=l(t)}^{t} P_i} \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l(t)}\|^2 \\
&\quad + 1200\gamma^2 L^2 \tau_{\text{mix}} A_3 \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\mathbf{x}_i - \mathbf{x}_{l(t)}\|^2 \\
&\quad + 4\gamma^2 L^2 A_1^2 \sum_{i=l(t)}^{t} P_i \sum_{i'=l(t)}^{t} P_{i'} \mathbb{E}\|\mathbf{x}_{i'} - \mathbf{x}_{l(t)}\|^2 + 4\gamma^2 L^2 \sum_{i=l(t)}^{t} P_i \sum_{i'=l(t)}^{t} P_{i'} \mathbb{E}\|\mathbf{x}_{i'} - \mathbf{x}_{l(t)}\|^2.
\end{aligned}$$

Since $t+1 \in \mathcal{L}_T$, we have $t = l(t) + \tau(t) - 1$. This allows us to apply Equation (22) in Lemma 4 to the last four terms in the preceding inequality:

$$
\begin{aligned}
\mathcal{T}_3 \leq\ & (8 + 10c_1c_2)\gamma^2 A_2 \sum_{i=l(t)}^{t} P_i \varsigma^2 \\
& + (936 + 1170c_1c_2)\gamma^2 \tau_{\text{mix}}^2 A_3^2 \varsigma^2 \\
& + (300 + 375c_1c_2)\gamma^2 \tau_{\text{mix}} A_3 \sum_{i=l(t)}^{t} P_i \varsigma^2 \\
& + (4 + 5c_1c_2)\gamma^2 \sum_{i=l(t)}^{t} P_i \sum_{i'=l(t)}^{t} P_{i'} \phi_{i'}^2 \\
& + (8 + 10c_1c_2)\gamma^2 A_2 \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 \\
& + 5c_1c_2\gamma^2 \sum_{i=l(t)}^{t} P_i \sum_{i'=l(t)}^{t} P_{i'} \mathbb{E}\|\nabla f(\mathbf{x}_{i'})\|^2 \\
& + (3744 + 4680c_1c_2)\gamma^2 \tau_{\text{mix}}^2 A_3^2 \frac{1}{\sum_{i=l(t)}^{t} P_i} \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 \\
& + (1200 + 1500c_1c_2)\gamma^2 \tau_{\text{mix}} A_3 \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2,
\end{aligned}
$$

where we assume the existence of c_2 (see Equation (27)). Substituting the bounds for $\mathcal{T}_2$ and $\mathcal{T}_3$ into Equation (29), we obtain

$$
\begin{aligned}
C_t^{\text{add}} \leq\ & \frac{1}{c_3 A_1} \gamma^2 L \sum_{i=0}^{\tau(t)-1} P_{t-i} \sum_{i'=0}^{\tau(t)-1} P_{t-i'} \mathbb{E}\|\nabla f(\mathbf{x}_{t-i'})\|^2 \\
& + \frac{1}{c_3 L A_1} \frac{1}{\sum_{i=0}^{\tau(t)-1} P_{t-i}} \sum_{i=0}^{\tau(t)-1} P_{t-i} \mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_{t-i})\|^2 \\
& + \left(\frac{1}{2} c_3 A_1 + 1\right)(8 + 10c_1c_2)\gamma^2 L A_2 \sum_{i=l(t)}^{t} P_i \varsigma^2 \\
& + \left(\frac{1}{2} c_3 A_1 + 1\right)(936 + 1170c_1c_2)\gamma^2 L \tau_{\text{mix}}^2 A_3^2 \varsigma^2 \\
& + \left(\frac{1}{2} c_3 A_1 + 1\right)(300 + 375c_1c_2)\gamma^2 L \tau_{\text{mix}} A_3 \sum_{i=l(t)}^{t} P_i \varsigma^2 \\
& + \left(\frac{1}{2} c_3 A_1 + 1\right)(4 + 5c_1c_2)\gamma^2 L \sum_{i=l(t)}^{t} P_i \sum_{i'=l(t)}^{t} P_{i'} \phi_{i'}^2 \\
& + \left(\frac{1}{2} c_3 A_1 + 1\right)(8 + 10c_1c_2)\gamma^2 L A_2 \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 \\
& + \left(\frac{1}{2} c_3 A_1 + 1\right) 5c_1c_2\gamma^2 L \sum_{i=l(t)}^{t} P_i \sum_{i'=l(t)}^{t} P_{i'} \mathbb{E}\|\nabla f(\mathbf{x}_{i'})\|^2 \\
& + \left(\frac{1}{2} c_3 A_1 + 1\right)(3744 + 4680c_1c_2)\gamma^2 L \tau_{\text{mix}}^2 A_3^2 \frac{1}{\sum_{i=l(t)}^{t} P_i} \sum_{i=l(t)}^{t} P_i \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2
\end{aligned}
$$

$$+\left(\frac{1}{2}c_3A_1+1\right)(1200+1500c_1c_2)\gamma^2L\tau_{\text{mix}}A_3\sum_{i=l(t)}^{t}P_i\mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2. \tag{30}$$

Step 4: Combining Case 1 and Case 2.

Up to now, we have obtained

$$\begin{aligned}&\text{Case 1 } (t+1\notin\mathcal{L}_T): \quad \mathbb{E}\big[f(\tilde{\mathbf{x}}_{t+1})-f(\tilde{\mathbf{x}}_t)\big]\le C_t^{\text{com}},\\ &\text{Case 2 } (t+1\in\mathcal{L}_T): \quad \mathbb{E}\big[f(\tilde{\mathbf{x}}_{t+1})-f(\tilde{\mathbf{x}}_t)\big]\le C_t^{\text{com}}+C_t^{\text{add}}.\end{aligned}$$

In this step, we sum up the descent equations from all the iterations t. That is, we derive

$$\mathbb{E}\big[f(\tilde{\mathbf{x}}_T)-f(\tilde{\mathbf{x}}_0)\big]\le\sum_{t=0}^{T-1}C_t^{\text{com}}+\sum_{t=0}^{T-1}\mathbb{1}_{\{t+1\in\mathcal{L}_T\}}C_t^{\text{add}}.$$

Assume that there exist universal constants c_4 and $c_5\in(0,1)$ such that, for any t,

$$P_t\le(1-c_5)c_4\frac{1}{\gamma LA_1}. \tag{31}$$

Then, by applying the period construction algorithm in Lemma 1 with $a=c_5c_4\frac{1}{\gamma LA_1}$ and $b=c_4\frac{1}{\gamma LA_1}$, we obtain

$$c_5c_4\frac{1}{\gamma LA_1}\le\sum_{t=l_q}^{l_q+\tau_q-1}P_t\le c_4\frac{1}{\gamma LA_1},\ \text{for } q=0,1,\dots,Q-1,\quad\text{and}\quad\sum_{t=l_Q}^{T-1}P_t\le c_4\frac{1}{\gamma LA_1}. \tag{32}$$

For any non-negative sequence $\{x_t\}_{t=0}^{T-1}$, we have

$$\begin{aligned}\sum_{t=0}^{T-1}P_t\sum_{i=l(t)}^{t-1}P_ix_i&=\sum_{q=0}^{Q}\sum_{t=l_q}^{l_q+\tau_q-1}P_t\sum_{i=l_q}^{t-1}P_ix_i\\&\le\sum_{q=0}^{Q}\sum_{t=l_q}^{l_q+\tau_q-1}P_t\sum_{i=l_q}^{l_q+\tau_q-1}P_ix_i\\&\le c_4\frac{1}{\gamma LA_1}\sum_{q=0}^{Q}\sum_{i=l_q}^{l_q+\tau_q-1}P_ix_i\\&=c_4\frac{1}{\gamma LA_1}\sum_{t=0}^{T-1}P_tx_t,\end{aligned} \tag{33}$$

where the second inequality uses Equation (32). Using Equation (33) and Equation (28), for the common part, we obtain

$$\begin{aligned}\sum_{t=0}^{T-1}C_t^{\text{com}}\le&-\frac{1}{2}\gamma\sum_{t=0}^{T-1}P_t\mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2-\frac{1}{2}\gamma\sum_{t=0}^{T-1}P_t\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2\\&+(1-c_5)c_4\frac{1}{A_1}\gamma\sum_{t=0}^{T-1}P_t\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2\\&+\frac{1}{2}(8+10c_1c_2)c_4\frac{1}{A_1}\gamma^2LA_2\sum_{t=0}^{T-1}P_t\varsigma^2\\&+\frac{1}{2}(936+1170c_1c_2)\gamma^3L^2\tau_{\text{mix}}^2A_3^2\sum_{t=0}^{T-1}P_t\varsigma^2\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}(300+375c_1c_2)c_4\frac{1}{A_1}\gamma^3L^2\tau_{\text{mix}}A_3\frac{1}{\gamma L}\sum_{t=0}^{T-1}P_t\varsigma^2\\
&+\frac{1}{2}(4+5c_1c_2)c_4^2\frac{1}{A_1^2}\gamma\sum_{t=0}^{T-1}P_t\phi_t^2\\
&+\frac{1}{2}(8+10c_1c_2)c_4\frac{1}{A_1}\gamma^2LA_2\sum_{t=0}^{T-1}P_t\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2\\
&+\frac{5}{2}c_1c_2c_4^2\frac{1}{A_1^2}\gamma\sum_{t=0}^{T-1}P_t\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2\\
&+\frac{1}{2}(3744+4680c_1c_2)\gamma^3L^2\tau_{\text{mix}}^2A_3^2\sum_{t=0}^{T-1}P_t\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2\\
&+\frac{1}{2}(1200+1500c_1c_2)c_4\frac{1}{A_1}\gamma^2L\tau_{\text{mix}}A_3\sum_{t=0}^{T-1}P_t\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2.
\end{aligned}
$$

For any non-negative sequence $\{x_t\}_{t=0}^{T-1}$, we have

$$
\begin{aligned}
\sum_{t=0}^{T-1}\mathbb{1}_{\{t+1\in\mathcal{L}_T\}}\sum_{i=0}^{\tau(t)-1}P_{t-i}x_{t-i}&=\sum_{t=0}^{T-1}\mathbb{1}_{\{t+1\in\mathcal{L}_T\}}\sum_{i=l(t)}^{t}P_ix_i\\
&=\sum_{q=0}^{Q-1}\sum_{t=l_q}^{l_q+\tau_q-1}P_tx_t\\
&\le\sum_{q=0}^{Q-1}\sum_{t=l_q}^{l_q+\tau_q-1}P_tx_t+\sum_{t=l_Q}^{T-1}P_tx_t\\
&=\sum_{t=0}^{T-1}P_tx_t,
\end{aligned}
\tag{34}
$$

where the first equality is because $t+1\in\mathcal{L}_T$ and then $l(t)=t+1-\tau(t)$. Using Equation (34) and Equation (30), for the additional part in Case 2, we obtain

$$
\begin{aligned}
\sum_{t=0}^{T-1}\mathbb{1}_{\{t+1\in\mathcal{L}_T\}}C_t^{\text{add}}\le{}&\frac{c_4}{c_3}\frac{1}{A_1^2}\gamma\sum_{t=0}^{T-1}P_t\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2+\frac{1}{c_3c_4c_5}\gamma\sum_{t=0}^{T-1}P_t\mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2\\
&+\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)(8+10c_1c_2)\gamma^2LA_1A_2\sum_{t=0}^{T-1}P_t\varsigma^2\\
&+\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)(936+1170c_1c_2)\frac{1}{c_4c_5}\gamma^3L^2\tau_{\text{mix}}^2A_1^2A_3^2\sum_{t=0}^{T-1}P_t\varsigma^2\\
&+\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)(300+375c_1c_2)\gamma^2L\tau_{\text{mix}}A_1A_3\sum_{t=0}^{T-1}P_t\varsigma^2\\
&+\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)(4+5c_1c_2)c_4\gamma\sum_{t=0}^{T-1}P_t\phi_t^2\\
&+\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)(8+10c_1c_2)\gamma^2LA_1A_2\sum_{t=0}^{T-1}P_t\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2\\
&+\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)5c_1c_2c_4\gamma\sum_{t=0}^{T-1}P_t\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2
\end{aligned}
$$

$$+\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)(3744+4680c_1c_2)\frac{1}{c_4c_5}\gamma^3L^2\tau_{\text{mix}}^2A_1^2A_3^2\sum_{t=0}^{T-1}P_t\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2$$
$$+\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)(1200+1500c_1c_2)\gamma^2L\tau_{\text{mix}}A_1A_3\sum_{t=0}^{T-1}P_t\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2.$$

Combining $\sum_{t=0}^{T-1}C_t^{\text{com}}$ and $\sum_{t=0}^{T-1}\mathbb{1}_{\{t+1\in\mathcal{L}_T\}}C_t^{\text{add}}$, we obtain

$$\begin{aligned}
&\mathbb{E}\big[f(\tilde{\mathbf{x}}_T)-f(\tilde{\mathbf{x}}_0)\big]\\
&\le-\left(\frac{1}{2}-\frac{1}{c_3c_4c_5}\right)\gamma\sum_{t=0}^{T-1}P_t\mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2\\
&\quad-\Bigg(\frac{1}{2}-\frac{c_4}{c_3}\frac{1}{A_1^2}-(1-c_5)c_4\frac{1}{A_1}\\
&\qquad-\frac{1}{2}(8+10c_1c_2)c_4\frac{1}{A_1}\gamma LA_2\\
&\qquad-\frac{5}{2}c_1c_2c_4^2\frac{1}{A_1^2}\\
&\qquad-\frac{1}{2}(3744+4680c_1c_2)\gamma^2L^2\tau_{\text{mix}}^2A_3^2\\
&\qquad-\frac{1}{2}(1200+1500c_1c_2)c_4\frac{1}{A_1}\gamma L\tau_{\text{mix}}A_3\\
&\qquad-\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)(8+10c_1c_2)\gamma LA_1A_2\\
&\qquad-\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)5c_1c_2c_4\\
&\qquad-\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)(3744+4680c_1c_2)\frac{1}{c_4c_5}\gamma^2L^2\tau_{\text{mix}}^2A_1^2A_3^2\\
&\qquad-\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)(1200+1500c_1c_2)\gamma L\tau_{\text{mix}}A_1A_3\Bigg)\gamma\sum_{t=0}^{T-1}P_t\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2\\
&\quad+\left(\frac{1}{2}(8+10c_1c_2)c_4\frac{1}{A_1^2}+\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)(8+10c_1c_2)\right)\gamma^2LA_1A_2\sum_{t=0}^{T-1}P_t\varsigma^2\\
&\quad+\Bigg(\frac{1}{2}(936+1170c_1c_2)\frac{1}{A_1}\gamma L\tau_{\text{mix}}A_3+\frac{1}{2}(300+375c_1c_2)c_4\frac{1}{A_1^2}\\
&\qquad+\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)(936+1170c_1c_2)\frac{1}{c_4c_5}\gamma L\tau_{\text{mix}}A_1A_3\\
&\qquad+\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)(300+375c_1c_2)\Bigg)\gamma^2L\tau_{\text{mix}}A_1A_3\sum_{t=0}^{T-1}P_t\varsigma^2\\
&\quad+\left(\frac{1}{2}(4+5c_1c_2)c_4^2\frac{1}{A_1^2}+\left(\frac{1}{2}c_3+\frac{1}{A_1}\right)(4+5c_1c_2)c_4\right)\gamma\sum_{t=0}^{T-1}P_t\phi_t^2. \qquad (35)
\end{aligned}$$

Step 5: Deciding the universal constants c_1 to c_5.

It remains to determine the constraints on γ and choose the universal constants c_1 to c_5 so that Equation (35) yields a valid cumulative descent equation. Here, c_1 is introduced in Lemma 4, c_2 in Equation (27), c_3 in Equation (29), and c_4 and c_5 in Equation (31):

Indeed, to ensure Equation (35) is a valid cumulative descent equation, it requires ensuring the existence of $c_1 > 1$, and the coefficient of $\gamma \sum_{t=0}^{T-1} P_t \mathbb{E}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2$ is less than or equal to 0 and the coefficient of $\gamma \sum_{t=0}^{T-1} P_t \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2$ is strictly smaller than 0 in Equation (35).

Define $A_4 \coloneqq \sup_t P_t$. To simplify the calculations, we consider the following constraints on γ: There exist universal constants a_2 to a_4 such that

$$\begin{aligned}\gamma L A_1 A_2 &\le a_2,\\ \gamma L A_1 A_3 \tau_{\mathrm{mix}} &\le a_3,\\ \gamma L A_1 A_4 &\le a_4.\end{aligned}$$

Under these constraints, the preceding requirements are ensured by the following sufficient conditions:

$$\begin{aligned}
&\text{(i)} \quad 1 - 5c_4^2 - 5c_4^2 - 4680a_3^2 - 1500c_4a_3 > 0,\\
&\text{(ii)} \quad c_1 \ge \frac{1}{1 - 5c_4^2 - 5c_4^2 - 4680a_3^2 - 1500c_4a_3},\\
&\text{(iii)} \quad c_2 \ge 4c_4^2 + 4c_4^2 + 3744a_3^2 + 1200c_4a_3,\\
&\text{(iv)} \quad 0.5 - \frac{1}{c_3c_4c_5} \ge 0,\\
&\text{(v)} \quad 0.5 - \frac{c_4}{c_3} - (1-c_5)c_4 - (4+5c_1c_2)c_4a_2 - 2.5c_1c_2c_4^2\\
&\qquad - (1872 + 2340c_1c_2)a_3^2 - (600 + 750c_1c_2)c_4a_3\\
&\qquad - (0.5c_3+1)\,(8+10c_1c_2)a_2 - (0.5c_3+1)\,5c_1c_2c_4\\
&\qquad - (0.5c_3+1)\,(3744+4680c_1c_2)\frac{1}{c_4c_5}a_3^2\\
&\qquad - (0.5c_3+1)\,(1200+1500c_1c_2)a_3 \qquad > 0,
\end{aligned}$$

where (i) and (ii) follow from Lemma 4, (iii) follows from Equation (27), (iv) and (v) follow from Equation (35), and we choose $a_4 = c_4(1-c_5)$ according to Equation (31). Then, we numerically search a_2, a_3, c_2, c_3, $c_4 \in (0,\infty)$, $c_1 \in (1,\infty)$ and $c_5 \in (0,1)$ subject to (i) to (v).

After searching, we find that $a_2 = 2\times10^{-7}$, $a_3 = 2\times10^{-7}$, $a_4 = 2\times10^{-4}$, $c_1 = 1.00002$, $c_2 = 0.000009$, $c_3 = 2500$, $c_4 = 0.001$ and $c_5 = 0.8$ constitute an appropriate choice. Furthermore, by the definitions of A_3 and A_4, we have $A_4 \le A_3$, so the constraint of $\gamma L A_1 A_4 \le 2\times10^{-4}$ (since $a_4 = 2\times10^{-4}$) is redundant and can be omitted. As a result, this choice yields the final constraints

$$\gamma L A_1 A_2 \le 2\times10^{-7}, \tag{36}$$

$$\gamma L A_1 A_3 \tau_{\mathrm{mix}} \le 2\times10^{-7}, \tag{37}$$

and the valid cumulative descent equation

$$\begin{aligned}\mathbb{E}\big[f(\tilde{\mathbf{x}}_T) - f(\tilde{\mathbf{x}}_0)\big] \le -\frac{1}{10}\gamma \sum_{t=0}^{T-1} P_t \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + 10010\gamma^2 L A_1 A_2 \sum_{t=0}^{T-1} P_t \varsigma^2\\ + 375600\gamma^2 L A_1 A_3 \tau_{\mathrm{mix}} \sum_{t=0}^{T-1} P_t \varsigma^2 + 6\gamma \sum_{t=0}^{T-1} P_t \phi_t^2.\end{aligned}$$

Noting that $\tilde{\mathbf{x}}_0 = \mathbf{x}_0$ and $f(\tilde{\mathbf{x}}_T) \ge f_*$, we obtain

$$\begin{aligned}&\frac{1}{\sum_{t=0}^{T-1} P_t} \sum_{t=0}^{T-1} P_t \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2\\ &\quad\le \frac{10\,(f(\mathbf{x}_0) - f_*)}{\gamma \sum_{t=0}^{T-1} P_t} + 100100\gamma L A_1 A_2 \varsigma^2 + 3756000\gamma L A_1 A_3 \tau_{\mathrm{mix}} \varsigma^2 + 60 \frac{1}{\sum_{t=0}^{T-1} P_t} \sum_{t=0}^{T-1} P_t \phi_t^2.\end{aligned}$$

Using Equation (4) (the definition of Φ), we obtain the claimed bound. □

A.4 Proof of Corollary 1

Proof of Corollary 1. In the iteration-invariant scenario (that is, $p_t^s = p^s$ and $w_t^s = w^s$), we have $P_t = P = \sum_{s\in\mathcal{S}} p^s w^s \pi(s)$ for any $t \in \{0, 1, \ldots, T-1\}$. Applying Theorem 1, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 = \mathcal{O}\left(\frac{F_0}{\gamma PT} + \gamma L A_1 A_2 \varsigma^2 + \gamma L \tau_{\text{mix}} A_1 A_3 \varsigma^2 + \Phi^2\right).$$

Define $\tilde{\gamma} \coloneqq \gamma P$. Then we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 = \mathcal{O}\left(\frac{F_0}{\tilde{\gamma} T} + \frac{\tilde{\gamma} L A_1 A_2 \varsigma^2}{P} + \frac{\tilde{\gamma} L \tau_{\text{mix}} A_1 A_3 \varsigma^2}{P} + \Phi^2\right).$$

Using the constraints of γ (see Equations (36) and (37) and Theorem 1), we obtain $\tilde{\gamma} \leq \frac{2\times 10^{-7} P}{L A_1 \max\{A_2, A_3\tau_{\text{mix}}\}}$. Following Koloskova et al. (2024, Corollary E.1), we choose

$$\tilde{\gamma} = \min\left\{\frac{2\times 10^{-7} P}{L A_1 (A_2 + A_3\tau_{\text{mix}})}, \left(\frac{P F_0}{L A_1 (A_2 + A_3\tau_{\text{mix}})\varsigma^2 T}\right)^{\frac{1}{2}}\right\}, \tag{38}$$

and then obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 = \mathcal{O}\left(\frac{L F_0 A_1 (A_2 + A_3\tau_{\text{mix}})}{PT} + \left(\frac{L F_0 A_1 (A_2 + A_3\tau_{\text{mix}})\varsigma^2}{PT}\right)^{\frac{1}{2}} + \Phi^2\right).$$

□

B Proof of Theorem 2

Proof of Theorem 2. **First, choose the objective functions and Markov chain.** Define

$$a_s \coloneqq \frac{1}{2}\varsigma\left(\text{sign}\left(\pi_{\text{eff}}(s) - \pi(s)\right) - \sum_{s\in\mathcal{S}}\pi(s)\,\text{sign}\left(\pi_{\text{eff}}(s) - \pi(s)\right)\right), \tag{39}$$

where $\text{sign}(\cdot)$ denotes the sign function and $\pi_{\text{eff}}(s) = \frac{p^s w^s \pi(s)}{\sum_{s\in\mathcal{S}} p^s w^s \pi(s)}$ (see Equation (5)). Then, it follows that $\sum_{s\in\mathcal{S}}\pi(s)a_s = 0$. Choose

$$f_s(\mathbf{x}) = \frac{1}{2}L\mathbf{x}^2 - a_s\mathbf{x}, \qquad \text{and} \quad f(\mathbf{x}) = \sum_{s\in\mathcal{S}}\pi(s)f_s(\mathbf{x}) = \frac{1}{2}L\mathbf{x}^2.$$

Consequently,

$$\nabla f_s(\mathbf{x}) = L\mathbf{x} - a_s, \qquad \text{and} \quad \nabla f(\mathbf{x}) = L\mathbf{x}.$$

For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}$,

$$\|\nabla f_s(\mathbf{x}) - \nabla f_s(\mathbf{y})\| = L\|\mathbf{x} - \mathbf{y}\|,$$

so every f_s is L-smooth. Moreover, f is lower bounded by zero. These verify Assumption 2. In addition, we have

$$\|\nabla f_s(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 = |a_s|^2 \leq \varsigma^2,$$

which verifies Assumption 3.

Choose the Markov chain with transition probabilities $M(s, r) = \pi(r)$ for all $s, r \in \mathcal{S}$, and initialize $s_0 \sim \pi$. It can be verified that this chain satisfies Assumption 1. For any $T \geq 0$ and any $r_0, \ldots, r_T \in \mathcal{S}$, the Markov property gives

$$\mathbb{P}(s_0 = r_0, \ldots, s_T = r_T) = \pi(r_0)\prod_{t=0}^{T-1} M(r_t, r_{t+1}) = \prod_{t=0}^{T}\pi(r_t).$$

Thus, $\{s_t\}_{t\geq 0}$ is an i.i.d. sequence with common distribution π.

Second, derive $\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{x}_t\right]$**.** According to Algorithm 1, we obtain, for one iteration,

$$\begin{aligned}\mathbf{x}_{t+1} &= \mathbf{x}_t - \gamma \delta_t^{s_t} w^{s_t} \nabla f_{s_t}(\mathbf{x}_t) \\ &= \mathbf{x}_t - \gamma \delta_t^{s_t} w^{s_t} \left(L\mathbf{x}_t - a_{s_t}\right) \\ &= \left(1 - \gamma L \delta_t^{s_t} w^{s_t}\right) \mathbf{x}_t + \gamma \delta_t^{s_t} w^{s_t} a_{s_t}.\end{aligned}$$

Taking expectations in the recursion yields

$$\mathbb{E}\left[\mathbf{x}_{t+1}\right] = \left(1 - \gamma L \sum_{s\in\mathcal{S}} p^s w^s \pi(s)\right) \mathbb{E}\left[\mathbf{x}_t\right] + \gamma \sum_{s\in\mathcal{S}} p^s w^s \pi(s) a_s. \tag{40}$$

For $\gamma L \sum_{s\in\mathcal{S}} p^s w^s \pi(s) < 1$, unrolling Equation (40), we obtain

$$\begin{aligned}\mathbb{E}\left[\mathbf{x}_T\right] &= \left(1 - \gamma L \sum_{s\in\mathcal{S}} p^s w^s \pi(s)\right)^T \mathbf{x}_0 + \sum_{i=0}^{T-1} \left(1 - \gamma L \sum_{s\in\mathcal{S}} p^s w^s \pi(s)\right)^i \sum_{s\in\mathcal{S}} p^s w^s \pi(s) a_s \gamma \\ &= \left(1 - \gamma L \sum_{s\in\mathcal{S}} p^s w^s \pi(s)\right)^T \mathbf{x}_0 + \left(1 - \left(1 - \gamma L \sum_{s\in\mathcal{S}} p^s w^s \pi(s)\right)^T\right) \frac{\sum_{s\in\mathcal{S}} p^s w^s \pi(s) a_s}{L \sum_{s\in\mathcal{S}} p^s w^s \pi(s)}.\end{aligned}$$

Taking the average of both sides of Equation (40) over $t = 0, 1, \ldots, T-1$, and then rearranging the terms in the resulting equality, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{x}_t\right] &= \frac{\mathbf{x}_0 - \mathbb{E}[\mathbf{x}_T]}{\gamma L T \sum_{s\in\mathcal{S}} p^s w^s \pi(s)} + \frac{\sum_{s\in\mathcal{S}} p^s w^s \pi(s) a_s}{L \sum_{s\in\mathcal{S}} p^s w^s \pi(s)} \\ &= \frac{1 - \left(1 - \gamma L \sum_{s\in\mathcal{S}} p^s w^s \pi(s)\right)^T}{\gamma L T \sum_{s\in\mathcal{S}} p^s w^s \pi(s)} \left(\mathbf{x}_0 - \frac{\sum_{s\in\mathcal{S}} p^s w^s \pi(s) a_s}{L \sum_{s\in\mathcal{S}} p^s w^s \pi(s)}\right) + \frac{\sum_{s\in\mathcal{S}} p^s w^s \pi(s) a_s}{L \sum_{s\in\mathcal{S}} p^s w^s \pi(s)}.\end{aligned}$$

Note that if $\gamma L T \sum_{s\in\mathcal{S}} p^s w^s \pi(s) \to \infty$, then

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{x}_t\right] = \frac{\sum_{s\in\mathcal{S}} p^s w^s \pi(s) a_s}{L \sum_{s\in\mathcal{S}} p^s w^s \pi(s)} = \frac{1}{L}\sum_{s\in\mathcal{S}} \pi_{\text{eff}}(s) a_s.$$

From Equation (40), we can get that the preceding limit also holds for $\gamma L \sum_{s\in\mathcal{S}} p^s w^s \pi(s) = 1$.

Recall the definition of $P \coloneqq \sum_{s\in\mathcal{S}} p^s w^s \pi(s)$. In summary, in this step, we get that if $\gamma L P T \to \infty$ and $\gamma L P \leq 1$, then $\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{x}_t\right] = \frac{1}{L}\sum_{s\in\mathcal{S}} \pi_{\text{eff}}(s) a_s$.

Third, derive $\liminf_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2$**.**

$$\begin{aligned}\liminf_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2 &\geq \liminf_{T\to\infty} \mathbb{E}\left\|\frac{1}{T}\sum_{t=0}^{T-1} \nabla f(\mathbf{x}_t)\right\|^2 \\ &= \liminf_{T\to\infty} \mathbb{E}\left\|\frac{1}{T}\sum_{t=0}^{T-1} L\mathbf{x}_t\right\|^2 \\ &\geq \liminf_{T\to\infty} \left\|L\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\mathbf{x}_t]\right\|^2 \\ &= \lim_{T\to\infty} \left\|L\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\mathbf{x}_t]\right\|^2 \\ &= \left\|L \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\mathbf{x}_t]\right\|^2,\end{aligned}$$

where the first and second inequalities use Jensen's inequality, the first equality uses $\nabla f(\mathbf{x}) = L\mathbf{x}$, the second and third equalities hold because $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\mathbf{x}_t]$ exists and the map of $\mathbf{x} \mapsto \|\mathbf{x}\|^2$ is continuous for $\mathbf{x} \in \mathbb{R}$.

Next, substituting $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\mathbf{x}_t]$, we obtain

$$\begin{aligned}
\liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\mathbf{x}_t)\|^2 &\geq \left\| L \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\mathbf{x}_t] \right\|^2 \\
&= \left\| \sum_{s\in\mathcal{S}} \pi_{\text{eff}}(s) a_s \right\|^2 \\
&= \frac{1}{4}\varsigma^2 \left(\sum_{s\in\mathcal{S}} (\pi_{\text{eff}}(s) - \pi(s)) \operatorname{sign}(\pi_{\text{eff}}(s) - \pi(s)) \right)^2 \\
&= \frac{1}{4}\varsigma^2 \left(\sum_{s\in\mathcal{S}} |(\pi_{\text{eff}}(s) - \pi(s))| \right)^2 \\
&= \frac{1}{4}\Phi^2,
\end{aligned}$$

where the second equality uses Equation (39) (the definition of a_s) and the last equality uses Equation (4) (the definition of Φ). □

C EXPERIMENTS

Missing details for neural network experiments. The logistic regression model for MNIST is from Li et al. (2020), and the convolutional neural networks for CIFAR-10 and CIFAR-100 are from Wang & Ji (2024). The local optimizer is SGD with a constant step size, zero momentum, and zero weight decay. The mini-batch sizes are 60, 50, and 50 for MNIST, CIFAR-10, and CIFAR-100, respectively.

Grid search. We tune the constant step size separately for each dataset using a three-stage coarse-to-fine grid search. The complete fine grid is $\Gamma = \{10^{\frac{i}{3}} : i = -14, -13, \ldots, -1\}$. In practice, the decimal values are rounded to three significant digits, for example, $10^{-\frac{1}{3}} \approx 0.464$ and $10^{-\frac{2}{3}} \approx 0.215$. We use one seed in the grid search and select the step size that achieves the lowest average training loss over the last 30% of iterations. The procedure is the same for all three datasets; we describe MNIST as an example. In the first stage, we run IA-0 without correction on the coarse grid $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, which identifies 10^{-2} as the best coarse-scale step size. In the second stage, we refine the search by adding the two closest points in Γ on each side of 10^{-2}, namely $\{10^{-\frac{4}{3}}, 10^{-\frac{5}{3}}, 10^{-2}, 10^{-\frac{7}{3}}, 10^{-\frac{8}{3}}\}$. This refinement selects 10^{-2} for MNIST. In the third stage, for each dataset we form the final grid from the selected center, its three immediate neighbors to the left in Γ, and its two immediate neighbors to the right. We then tune every availability configuration both without correction and with correction. Thus, the final grid is $\{10^{-\frac{9}{3}}, 10^{-\frac{8}{3}}, 10^{-\frac{7}{3}}, 10^{-2}, 10^{-\frac{5}{3}}, 10^{-\frac{4}{3}}\}$ for MNIST. The selected step sizes are summarized in Table 3.

Table 3: The final third-stage step size grids and selected step sizes for MNIST, CIFAR-10 and CIFAR-100. A "✓" indicates the step size selected for the corresponding availability configuration.

MNIST: without correction

Step size	10^{-3}	$10^{-\frac{8}{3}}$	$10^{-\frac{7}{3}}$	10^{-2}	$10^{-\frac{5}{3}}$	$10^{-\frac{4}{3}}$
FA				✓		
IA-0				✓		
IA-1				✓		
IA-2				✓		
IA-3				✓		

MNIST: with correction

Step size	10^{-3}	$10^{-\frac{8}{3}}$	$10^{-\frac{7}{3}}$	10^{-2}	$10^{-\frac{5}{3}}$	$10^{-\frac{4}{3}}$
IA-0			✓			
IA-1			✓			
IA-2		✓				
IA-3	✓					

CIFAR-10: without correction

Step size	$10^{-\frac{10}{3}}$	10^{-3}	$10^{-\frac{8}{3}}$	$10^{-\frac{7}{3}}$	10^{-2}	$10^{-\frac{5}{3}}$
FA				✓		
IA-0				✓		
IA-1					✓	
IA-2					✓	
IA-3					✓	

CIFAR-10: with correction

Step size	$10^{-\frac{10}{3}}$	10^{-3}	$10^{-\frac{8}{3}}$	$10^{-\frac{7}{3}}$	10^{-2}	$10^{-\frac{5}{3}}$
IA-0			✓			
IA-1			✓			
IA-2			✓			
IA-3			✓			

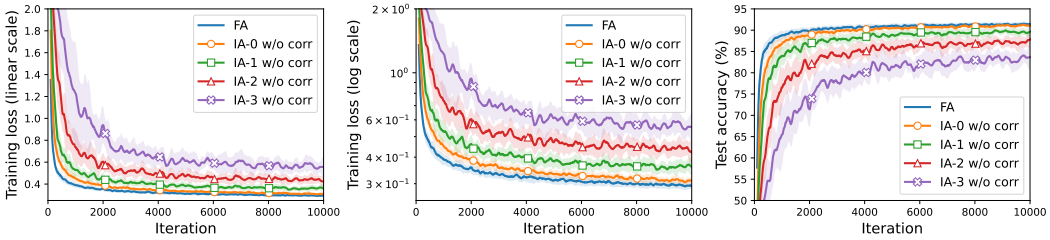
CIFAR-100: without correction

Step size	$10^{-\frac{10}{3}}$	10^{-3}	$10^{-\frac{8}{3}}$	$10^{-\frac{7}{3}}$	10^{-2}	$10^{-\frac{5}{3}}$
FA				✓		
IA-0				✓		
IA-1				✓		
IA-2				✓		
IA-3				✓		

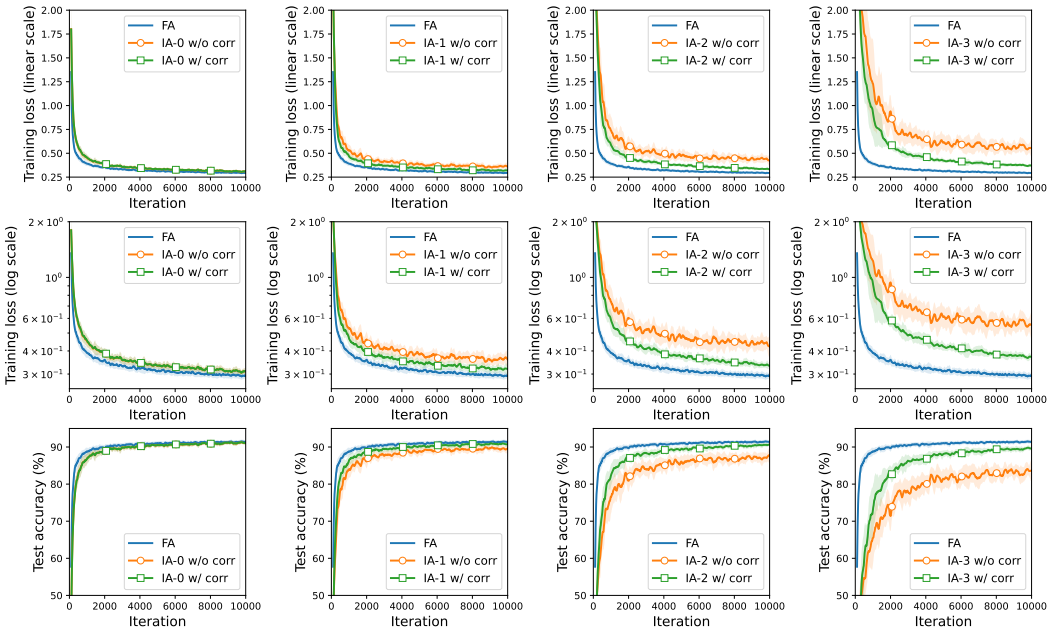
CIFAR-100: with correction

Step size	$10^{-\frac{10}{3}}$	10^{-3}	$10^{-\frac{8}{3}}$	$10^{-\frac{7}{3}}$	10^{-2}	$10^{-\frac{5}{3}}$
IA-0			✓			
IA-1			✓			
IA-2		✓				
IA-3		✓				

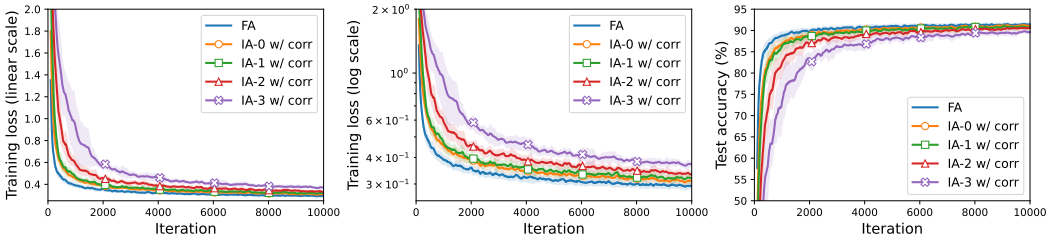
More experimental results. Figures 5, 6 and 7 show training curves of different availability configurations on MNIST, CIFAR-10 and CIFAR-100, respectively.



(a) Without correction.

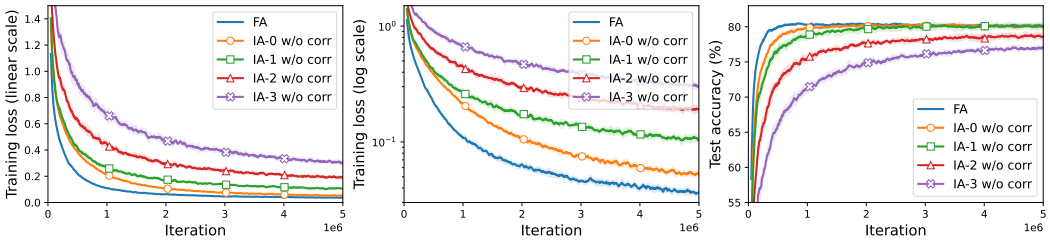


(b) Comparison of corrected and uncorrected cases for each configuration.

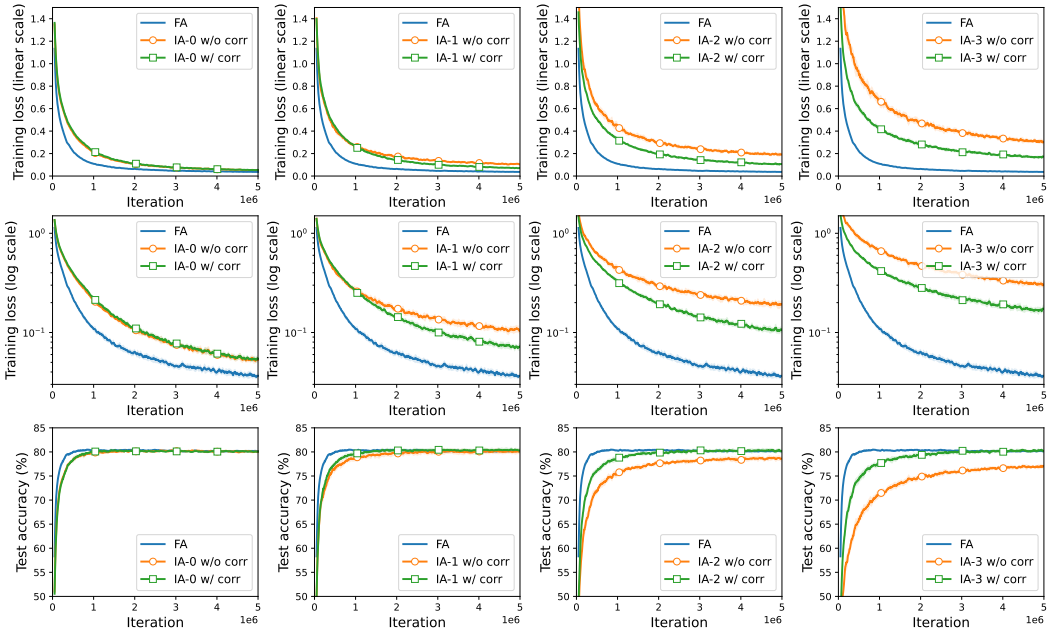


(c) With correction.

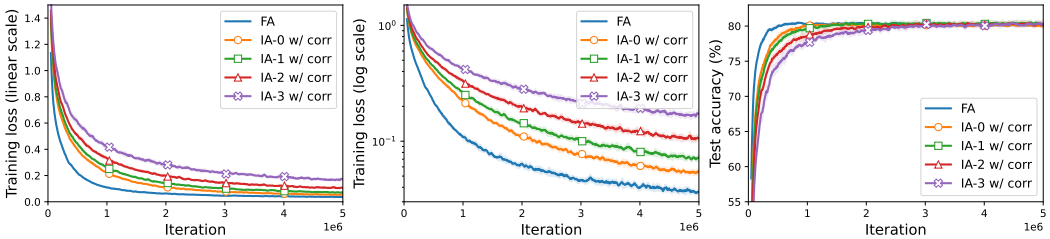
Figure 5: Training curves of different availability configurations (Table 2) on MNIST. The shaded areas show the standard deviation across 10 random seeds. (a) is for Finding 1; (b) is for Finding 2; (c) is for Finding 3.



(a) Without correction.

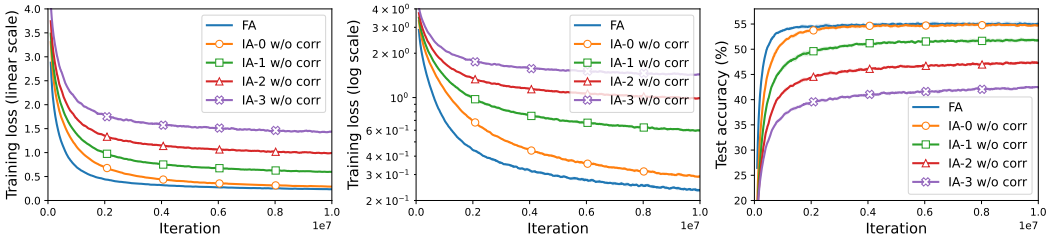


(b) Comparison of corrected and uncorrected cases for each configuration.

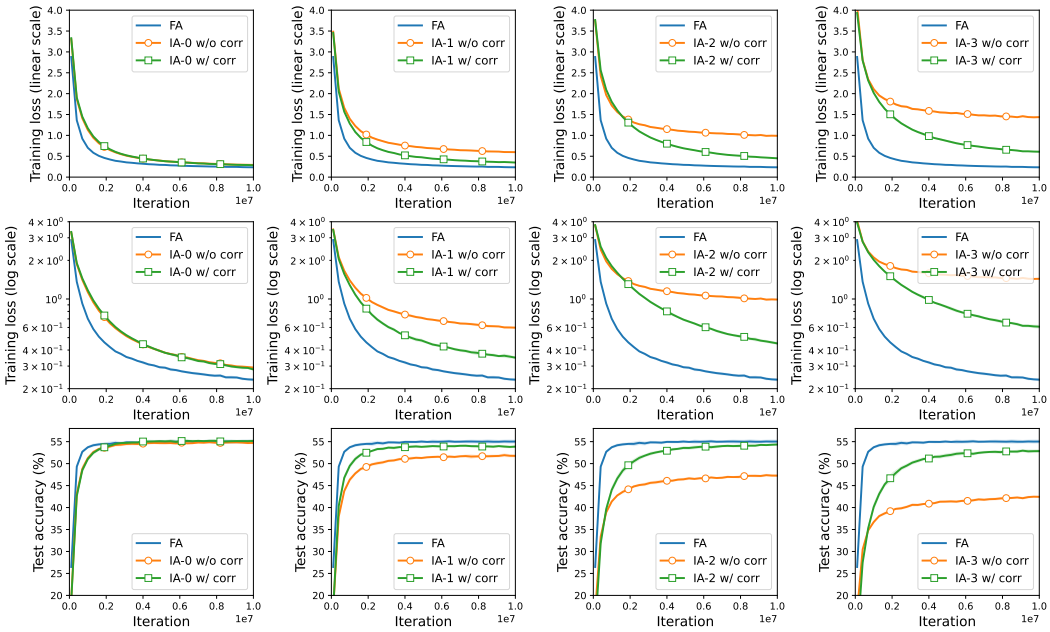


(c) With correction.

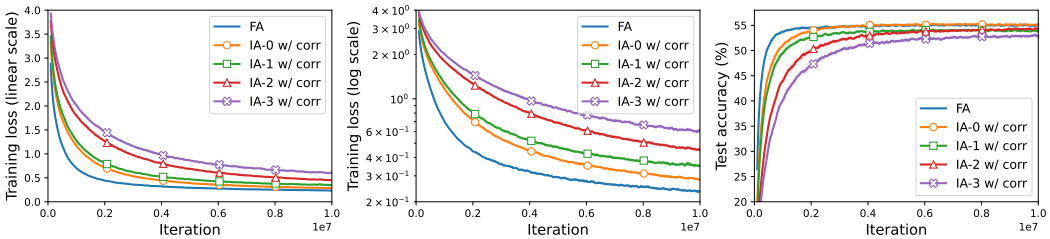
Figure 6: Training curves of different availability configurations (Table 2) on CIFAR-10. The shaded areas show the standard deviation across 3 random seeds. (a) is for Finding 1; (b) is for Finding 2; (c) is for Finding 3.



(a) Without correction.



(b) Comparison of corrected and uncorrected cases for each configuration.



(c) With correction.

Figure 7: Training curves of different availability configurations (Table 2) on CIFAR-100. The shaded areas show the standard deviation across 3 random seeds. (a) is for Finding 1; (b) is for Finding 2; (c) is for Finding 3.