CYCWIN-FL: CONVERGENCE FOR FEDERATED LEARNING WITH CYCLIC-WINDOW CLIENT AVAILABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

In resource-constrained federated learning (FL), client availability may be restricted by computing power, connectivity or privacy requirements, thus entailing *controlled* client participation. This prohibits direct application of existing convergence-analysis frameworks under the simplified assumption of full or uniform client sampling. Recent studies on convergence for FL with cyclic client participation schemes often considered restricted types of training objectives or specific cyclic patterns. In this paper, we investigate FL with a *general deterministic cyclic pattern*, i.e., *Cycwin-FL*, which applies FedAvg with uniform client sampling from within a window of B among M clients that cyclically advances by N_w clients per communication round. We provide convergence analysis for Cycwin-FL with Polyak–Łojasiewicz (PL)-conditioned and general smooth non-convex objectives, respectively. Specifically, if $\gcd(M, N_w) \mid B$ is satisfied, for PL-conditioned objectives, we reveal an $\widetilde{\mathcal{O}}(T^{-1})$ upper bound on the optimality gap leveraging window-induced bias cancellation, where T is the total number of communication rounds; and for general smooth non-convex objectives, we derive the same $\mathcal{O}(T^{-1/2})$ convergence rate as the vanilla FedAvg. In addition, we also quantify the impact of failure in meeting the condition $\gcd(M, N_w) \mid B$ on convergence. Finally, experiments validate our theoretical insights into the effect of the cyclic patterns.

1 INTRODUCTION

Federated learning (FL) trains a shared model from decentralized data through local optimization and server aggregation (McMahan et al., 2017; Bonawitz et al., 2019; Kairouz et al., 2021). Federated averaging (FedAvg) reduces communication through multiple local updates between server aggregations. Its analysis covers heterogeneous objectives and multiple local updates (Stich, 2019; Woodworth et al., 2020; Khaled et al., 2020; Li et al., 2020b). However, standard convergence analyses often assume full participation or partial participation with uniform client sampling in each communication round. Connectivity, mobility, computing power, and service policies can restrict such participation, thus requiring *controlled* client participation.

For example, satellites participating in FL in low Earth orbit (LEO) enter a ground station's visibility region, remain visible for several communication rounds, and leave as other satellites enter. This recurring client availability motivates a deterministic cyclic-window model, as do periodically shifting client populations (Zhu et al., 2022) and cyclic participation policies (Cho et al., 2023).

In FL with cyclic availability, communication schedules and training updates are closely coupled. This interaction is particularly relevant to satellite FL, where existing studies consider ground-satellite visibility constraints (Razmi et al., 2022), asynchronous aggregation (Elmahallawy & Luo, 2022; Huang et al., 2024), intra-plane model propagation (Elmahallawy & Luo, 2023), and on-orbit aggregation (Elmahallawy et al., 2023). However, convergence analyses based on full participation or uniform sampling from the entire client population do not directly account for the temporal dependence induced by cyclic availability. Motivated by this, we study FL under a general class of *deterministic cyclic patterns*, using a protocol termed *Cycwin-FL*. In each communication round, only a window of B among M clients is available, from which clients are sampled uniformly at random to perform FedAvg. The window then advances cyclically by N_w clients for the next round. We identify an exact condition under which the cyclic pattern is *balanced*. Under this condition, we derive a

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.

$\widetilde{\mathcal{O}}(T^{-1})$ upper bound on the optimality gap for Polyak–Łojasiewicz (PL)-conditioned objectives and an $\mathcal{O}(T^{-1/2})$ stationarity rate for general smooth non-convex objectives, where T is the total number of communication rounds. We further quantify the additional optimization error that can arise when the balance condition is violated.

1.1 Related work

FedAvg extensions address heterogeneous local training in different ways, including proximal regularization in FedProx (Li et al., 2020a), control variates to mitigate client drift in SCAFFOLD (Karimireddy et al., 2020), and normalization for unequal local computation in FedNova (Wang et al., 2020). Dynamic regularization and adaptive server optimization also modify local or global update rules (Acar et al., 2021; Reddi et al., 2021).

However, these existing analyses typically assume full participation or uniform client sampling (Li et al., 2020b). Most client-sampling schemes prioritize resource efficiency and data utility (Nishio & Yonetani, 2019; Lai et al., 2021), or improve representativeness and class balance (Fraboni et al., 2021; Zhang et al., 2023); biased-selection analyses quantify the resulting convergence effects (Cho et al., 2022). Such selection is constrained by which clients are available. This has motivated flexible participation and server assistance for inactive or incomplete updates (Ruan et al., 2021; Yang et al., 2024), as well as analyses of intermittent, arbitrary, or correlated availability (Gu et al., 2021; Ribero et al., 2023; Wang & Ji, 2022; Rodio et al., 2023; Xiang et al., 2024). Studies of periodically shifting populations (Zhu et al., 2022), cyclic participation (Cho et al., 2023), and probabilistic models of periodic participation (Crawshaw & Liu, 2024) further seek to capture participation structure across rounds. Within this line of work, we study how deterministic cyclic-window availability shapes the objective and sequential trajectory of FedAvg. Section 2 introduces the common FL workflow and compares the corresponding participation rules.

1.2 Contributions

In summary, our contributions are:

1. We derive exact per-client availability counts and show that every client is available equally often over a cycle if and only if $\gcd(M, N_w) \mid B$. This balance condition ensures that the cycle-averaged window gradient coincides with the global gradient when evaluated at a common model.
2. Under this balance condition, we exploit cancellation of window-induced biases to establish a $\widetilde{\mathcal{O}}(T^{-1})$ optimality-gap bound under the PL condition, where T is the number of communication rounds. We further extend the analysis to general smooth non-convex objectives, obtaining an $\mathcal{O}(T^{-1/2})$ rate for the average expected squared gradient norm, matching the order of vanilla FedAvg with uniform client sampling. To obtain these rates, we control sequential model drift, yielding bounds that separate the effects of sampling noise, local updates, and cycle length.
3. When the balance condition is violated, we characterize the induced client weights and their exact discrepancy from uniform weighting. Using this characterization, we derive PL and non-convex convergence bounds relative to the equally weighted global objective. The resulting bounds quantify how unequal cyclic weights interact with client heterogeneity and can introduce additional optimization error.

Notations. For a positive integer A, we denote a set $\{1, \ldots, A\}$ by $[A]$. We use $|\mathcal{A}|$ for set cardinality, $\|\cdot\|$ for the Euclidean norm, and $\mathbb{E}[\cdot]$ for expectation. For positive integers a and b, $\gcd(a, b)$ denotes their greatest common divisor; $a \mid b$ means that a divides b; and $b \bmod a$ is the remainder in $\{0, \ldots, a-1\}$ upon dividing b by a. For $x \in \mathbb{R}$, $\lfloor x \rfloor$ stands for the greatest integer not exceeding x. We write $A \triangleq B$ to define A as B.

2 Cycwin-FL: Training Protocol with a General Cyclic Pattern

In this section, we introduce the FL training protocol with a general cyclic pattern specified by Cycwin-FL.

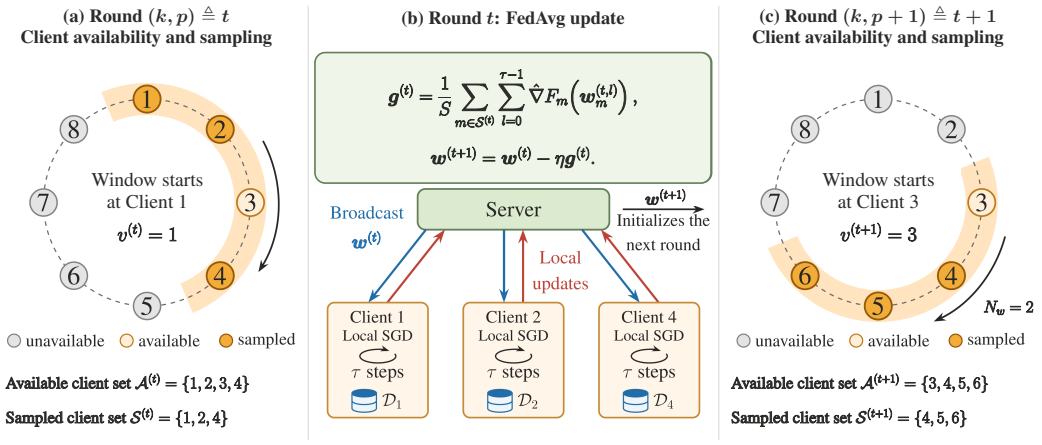


Figure 1: An illustration of Cycwin-FL with $M = 8$, $B = 4$, $S = 3$, and $N_w = 2$. (a) The current window determines the available client set $\mathcal{A}^{(t)}$. The server samples S clients uniformly at random from $\mathcal{A}^{(t)}$, composing the sampled client set $\mathcal{S}^{(t)}$. (b) The active clients in $\mathcal{S}^{(t)}$ perform FedAvg under the coordination of the server. (c) The window advances by N_w clients to repeat from (a) to (c) in the next round.

2.1 CYCWIN-FL

Consider an FL system consisting of M clients, where each client $m \in [M]$ holds a local dataset $\mathcal{D}_m$. All M clients collaboratively train a shared deep learning model parameterized by $\boldsymbol{w} \in \mathbb{R}^{d_{\mathrm{w}}}$ to solve an empirical loss minimization problem given by

$$\underset{\boldsymbol{w} \in \mathbb{R}^{d_{\mathrm{w}}}}{\texttt{Minimize}} \quad F(\boldsymbol{w}) \triangleq \frac{1}{M} \sum_{m=1}^{M} F_m(\boldsymbol{w}), \tag{1}$$

where $F_m(\boldsymbol{w}) = (1/|\mathcal{D}_m|) \sum_{\boldsymbol{\xi} \in \mathcal{D}_m} \ell(\boldsymbol{w}; \boldsymbol{\xi})$ is the local empirical loss function at client $m \in [M]$ with $\ell(\boldsymbol{w}; \boldsymbol{\xi})$ denoting the sample-wise loss evaluated by model $\boldsymbol{w}$ on data sample $\boldsymbol{\xi}$; and the optimal global empirical loss is defined as $F^* \triangleq \min_{\boldsymbol{w} \in \mathbb{R}^{d_{\mathrm{w}}}} F(\boldsymbol{w}) > -\infty$. To solve problem (1), we adopt a general federated averaging protocol, namely, *Cycwin-FL*, the client availability of which follows a *deterministic* cyclic pattern as elaborated in the sequel.

As illustrated in Fig. 1, at communication round $t \in \{0, 1, \ldots, T-1\}$, only a window of $B \leq M$ clients, starting from client $v^{(t)} \in [M]$, is available to be included in the available client set defined as

$$\mathcal{A}^{(t)} = \{((v^{(t)} - 1 + j) \bmod M) + 1 : j = 0, \ldots, B-1\}. \tag{2}$$

Then, conditional on the preceding training history, an active set $\mathcal{S}^{(t)} \subseteq \mathcal{A}^{(t)}$ of clients is sampled uniformly at random (without replacement) with the set size $S \triangleq |\mathcal{S}^{(t)}| = \varrho B$, where $\varrho \in (0, 1]$ is the ratio of the sampled active clients, to perform the FedAvg update (McMahan et al., 2017).

Specifically, at round t, each client $m \in \mathcal{S}^{(t)}$ initializes its local model as $\boldsymbol{w}_m^{(t,0)} \leftarrow \boldsymbol{w}^{(t)}$ and performs τ local SGD steps with a constant step size $\eta > 0$ using a mini-batch $\mathcal{D}_m^{(t,l)} \subseteq \mathcal{D}_m$ as

$$\boldsymbol{w}_m^{(t,l+1)} \leftarrow \boldsymbol{w}_m^{(t,l)} - \eta \hat{\nabla} F_m(\boldsymbol{w}_m^{(t,l)}), \qquad l = 0, \ldots, \tau - 1, \tag{3}$$

where

$$\hat{\nabla} F_m(\boldsymbol{w}_m^{(t,l)}) = \frac{1}{\left|\mathcal{D}_m^{(t,l)}\right|} \sum_{\boldsymbol{\xi} \in \mathcal{D}_m^{(t,l)}} \nabla \ell(\boldsymbol{w}_m^{(t,l)}; \boldsymbol{\xi}) \tag{4}$$

is a stochastic gradient approximating $\nabla F_m(\boldsymbol{w}_m^{(t,l)})$. At each local step, mini-batches are sampled uniformly and independently across clients conditional on the preceding training history. After clients

Algorithm 1 Cycwin-FL with cyclic-window client participation and local SGD.

Input: clients $[M]$, initial model $\boldsymbol{w}^{(0,0)}$, total number of rounds T, total number of cycles K, communication rounds per cycle P, step size η, number of local steps τ, sample size S, window size B, and window-advancing step N_w.
Initialize the first client $v^{(0,1)} \leftarrow 1$ that the window starts with.
for cycles $k = 0, \ldots, K-1$ **do**
  for communication rounds $p = 1, \ldots, P$ **do**
    Set the available client set $\mathcal{A}^{(k,p)} \leftarrow \{((v^{(k,p)} - 1 + j) \bmod M) + 1 : j = 0, \ldots, B-1\}$.
    Sample $\mathcal{S}^{(k,p)} \subseteq \mathcal{A}^{(k,p)}$ uniformly without replacement, with $|\mathcal{S}^{(k,p)}| = S$.
    Broadcast $\boldsymbol{w}^{(k,p-1)}$ to all clients in $\mathcal{S}^{(k,p)}$.
    for each client $m \in \mathcal{S}^{(k,p)}$ **in parallel do**
      Initialize $\boldsymbol{w}_m^{(k,p-1,0)} \leftarrow \boldsymbol{w}^{(k,p-1)}$.
      for $l = 0, \ldots, \tau - 1$ **do**
        Draw a mini-batch $\mathcal{D}_m^{(k,p,l)}$ from $\mathcal{D}_m$;
        $\boldsymbol{w}_m^{(k,p-1,l+1)} \leftarrow \boldsymbol{w}_m^{(k,p-1,l)} - \eta \hat{\nabla} F_m(\boldsymbol{w}_m^{(k,p-1,l)})$.
      end for
      Send the local update $\boldsymbol{w}_m^{(k,p-1,\tau)} - \boldsymbol{w}_m^{(k,p-1,0)}$ to the server.
    end for
    After receiving all the local updates from active clients, update the global model as $\boldsymbol{w}^{(k,p)} \leftarrow \boldsymbol{w}^{(k,p-1)} + (1/S) \sum_{m \in \mathcal{S}^{(k,p)}} \left(\boldsymbol{w}_m^{(k,p-1,\tau)} - \boldsymbol{w}_m^{(k,p-1,0)} \right)$.
    Advance the window: $v^{(k,p+1)} \leftarrow ((v^{(k,p)} - 1 + N_w) \bmod M) + 1$.
  end for
  Set $\boldsymbol{w}^{(k+1,0)} \leftarrow \boldsymbol{w}^{(k,P)}$ and $v^{(k+1,1)} \leftarrow v^{(k,P+1)}$.
end for
return $\boldsymbol{w}^{(T)} \triangleq \boldsymbol{w}^{(K,0)}$.

$m \in \mathcal{S}^{(t)}$ send their model updates to the server, the server updates the global model by averaging over all active clients' local model updates as

$$\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} + \frac{1}{S} \sum_{m \in \mathcal{S}^{(t)}} \left(\boldsymbol{w}_m^{(t,\tau)} - \boldsymbol{w}_m^{(t,0)} \right). \tag{5}$$

Next, the window advances by a step of $N_w \leq B$ clients in a cyclic manner, starting from client $v^{(t+1)}$ for the next round of training, until the total number of communication rounds T is completed.

To characterize the periodic nature of the cyclic pattern, we define a *cycle* as the interval until the first client $v^{(t)} \in [M]$ in the window becomes the first client in the window again *for the first time* after P communication rounds, i.e., $v^{(t)} = v^{(t+P)}$, where P is the *period* of the cycle. It is thus easy to verify that $P = M/d$, where $d \triangleq \gcd(M, N_w)$ is the greatest common divisor (gcd) of M and N_w. Throughout this paper, we assume that T is a positive integer multiple of P, and define $K \triangleq T/P$ as the total number of cycles. To facilitate analysis of the training in the sequel, we re-index communication round $t \in \{0, 1, \ldots, T-1\}$ into a two-element tuple (k, p) by

$$t = kP + p - 1, \tag{6}$$

where $p = (t \bmod P) + 1$ and $k = 0, 1, \ldots, K-1$. Then, $\boldsymbol{w}^{(t)} = \boldsymbol{w}^{(k,p-1)}$ and $\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(k,p)}$. Using the renewed indexing, the training protocol of Cycwin-FL is summarized in Algorithm 1.

2.2 Comparison with FedAvg and CyCP

In terms of cyclic patterns, there are two typical special cases of Cycwin-FL in the literature. First, Cycwin-FL recovers FedAvg (McMahan et al., 2017) with uniform client sampling when $B = N_w = M$. Cycwin-FL also reduces to the CyCP scheme (Cho et al., 2023) when $B \mid M$ and $N_w = B$. In this case, the windows in $P = M/B$ rounds are disjoint and partition all clients in $[M]$. Appendix F gives a detailed comparison. However, note that in both cases the average window gradient over a complete cycle equals the average gradient over all clients when evaluated

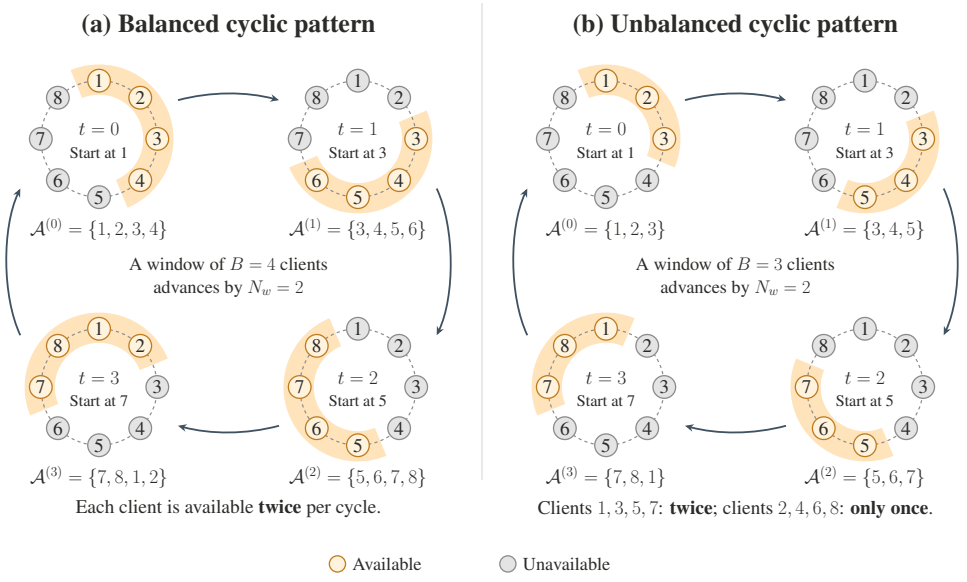


Figure 2: An illustration of balanced and unbalanced cyclic patterns with $M = 8$, $N_w = 2$, and period $P = 4$ in both cases. (a) A balanced cyclic pattern with $B = 4$. (b) An unbalanced cyclic pattern with $B = 3$.

at a common model, on which the convergence analysis builds. Section 3 exploits this equality under balanced cyclic patterns, while Section 4 accounts for its failure under unbalanced patterns. Moreover, our proof bounds gradient differences directly using Lipschitz continuity, without invoking the Hessian representations used in the CyCP analysis of Cho et al. (2023). We also establish convergence rates for both PL and general smooth non-convex objectives, whereas their analysis requires the PL condition.

3 Convergence Analysis under Balanced Cyclic Patterns

As shown in Fig. 2, Cycwin-FL generalizes two important classes of deterministic cyclic patterns, i.e., balanced and unbalanced cyclic patterns. A cyclic pattern is *balanced* if each client in $[M]$, regardless of whether it is included in the active set under uniform sampling, is available the same number of times per cycle, and *unbalanced* otherwise. While Section 4 considers unbalanced cyclic patterns, this section focuses on balanced cyclic patterns characterized by the following lemma.

Lemma 3.1 (Balanced cyclic pattern)**.** *For Cycwin-FL with M clients, window size B, and window-advancing step N_w, its cyclic pattern is balanced if and only if $d \mid B$, where $d \triangleq \gcd(M, N_w)$.*

Proof. See Appendix G.1. □

Assumption 1 (L-smoothness)**.** For each client $m \in [M]$, its local empirical loss function $F_m(\cdot)$ is L-smooth, i.e., for all $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^{d_w}$,

$$\|\nabla F_m(\boldsymbol{w}) - \nabla F_m(\boldsymbol{w}')\| \leq L \|\boldsymbol{w} - \boldsymbol{w}'\| . \tag{7}$$

Assumption 2 (Conditional stochastic gradients)**.** Let $\mathcal{F}_{k,p,l}$ denote the σ-algebra generated by the initial model, all client and mini-batch sampling in preceding rounds, the current selected client set, and the mini-batches sampled before local step l in round (k, p) (see Appendix A). For each $m \in \mathcal{A}^{(k,p)}$ and any $\mathcal{F}_{k,p,l}$-measurable model $\boldsymbol{w}$, the mini-batch gradient at local step l satisfies

$$\mathbb{E}\left[\hat{\nabla} F_m(\boldsymbol{w}) \middle| \mathcal{F}_{k,p,l}\right] = \nabla F_m(\boldsymbol{w}), \tag{8}$$

$$\mathbb{E}\left[\left\|\hat{\nabla} F_m(\boldsymbol{w}) - \nabla F_m(\boldsymbol{w})\right\|^2 \middle| \mathcal{F}_{k,p,l}\right] \leq \sigma^2. \tag{9}$$

Here $\hat{\nabla} F_m(\boldsymbol{w})$ uses the current mini-batch $\mathcal{D}_m^{(k,p,l)}$ in (4), with the batch indices suppressed.

Assumption 3 (Cyclic-window heterogeneity)**.** For each available client set $\mathcal{A}$, define the window-average gradient $\overline{\nabla F}_{\mathcal{A}}(\boldsymbol{w}) \triangleq (1/B)\sum_{m\in\mathcal{A}} \nabla F_m(\boldsymbol{w})$. Then, there exist $\gamma, \alpha \geq 0$ such that for each client $m \in \mathcal{A}$ and any $\boldsymbol{w} \in \mathbb{R}^{d_{\mathrm{w}}}$,

$$\left\|\nabla F_m(\boldsymbol{w}) - \overline{\nabla F}_{\mathcal{A}}(\boldsymbol{w})\right\| \leq \gamma, \tag{10}$$

$$\left\|\overline{\nabla F}_{\mathcal{A}}(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\right\| \leq \alpha. \tag{11}$$

Assumption 4 (Balanced cyclic pattern)**.** The cyclic-window schedule has a balanced cyclic pattern with the period $P \geq 2$.

To state the convergence bounds compactly, we collect the effects induced by within-window sampling, stochastic gradients, and sequential model drift into a few quantities. Define the combined heterogeneity and noise scale $H \triangleq \alpha + \gamma + \sigma$ and the normalized joint variance bound $\Sigma_{\mathrm{eff}}^2 \triangleq \omega_{B,S}\gamma^2 + \sigma^2/(\tau S)$, where $\omega_{B,S} \triangleq (B-S)/(S(B-1))$ for $B > 1$ and $\omega_{1,1} = 0$. In addition, $\mathcal{R}_{\mathrm{CW}} \triangleq P(\alpha + \Sigma_{\mathrm{eff}}) + (\tau - 1)H/\tau$ collects the model-drift contributions from communication rounds within a cycle and from local updates.

3.1 Convergence rates for PL objectives

Assumption 5 (Polyak–Łojasiewicz condition)**.** For some $\mu > 0$, the global empirical loss function satisfies

$$\frac{1}{2}\left\|\nabla F(\boldsymbol{w})\right\|^2 \geq \mu\big(F(\boldsymbol{w}) - F^*\big), \qquad \forall \boldsymbol{w} \in \mathbb{R}^{d_{\mathrm{w}}}. \tag{12}$$

Under this condition, we can derive the following convergence bound.

Theorem 3.1 (PL convergence under balanced cyclic patterns)**.** *Suppose Assumptions 1–5 hold. Let $T = KP$, $\Phi_0^{\mathrm{PL}} \triangleq \mathbb{E}[F(\boldsymbol{w}^{(0,0)})] - F^*$, and the step size $\eta_T^{\mathrm{PL}} = \min\{1/(64LP\tau), 4\log T/(\mu\tau T)\}$. If $T \geq 256(L/\mu)P\log T$, then, for universal constants $C_1, C_2 > 0$,*

$$\mathbb{E}[F(\boldsymbol{w}^{(K,0)})] - F^* \leq \frac{\Phi_0^{\mathrm{PL}}}{T^2} + C_1\frac{L\log T}{\mu^2 T}\Sigma_{\mathrm{eff}}^2 + C_2\frac{L^2(\log T)^2}{\mu^3 T^2}\mathcal{R}_{\mathrm{CW}}^2. \tag{13}$$

Proof sketch. The proof consists of three main steps. First, we decompose the window-induced gradient biases into terms evaluated at $\boldsymbol{w}^{(k,0)}$, which cancel over a balanced cycle, and residual terms bounded using smoothness. Second, analyzing the accumulated gradients over a full cycle yields recursive bounds on model displacement. To obtain these bounds, we unroll the first-moment displacement recursion and sum the resulting geometric series, while the restrictions on η_T^{PL} ensure that the recursive displacement term in the second-moment bound is absorbed. Third, conditional expectations given the history before each round jointly control client-sampling and mini-batch noise. These estimates yield a descent bound over each cycle $k = 0, \ldots, K-1$, which, together with the PL inequality, gives a recursive bound on the expected optimality gap. Unrolling this recursion with the prescribed step size gives (13). □

In (13), client-sampling and mini-batch noise contribute at order $\log T/T$, while the initial error and model drift decay at $\widetilde{\mathcal{O}}(T^{-2})$. Thus, balanced cyclic patterns leave no persistent heterogeneity term. With full-window participation $S = B$ and exact local gradients $\sigma = 0$, the leading variance term vanishes, yielding a $\widetilde{\mathcal{O}}(T^{-2})$ bound. Longer cycles increase $P(\alpha + \Sigma_{\mathrm{eff}})$ in $\mathcal{R}_{\mathrm{CW}}$, producing at most quadratic dependence on P in the drift term when other quantities are fixed. They also tighten the step-size bound $1/(64LP\tau)$ required for (13). Increasing τ reduces mini-batch variance but raises the local-drift factor $(\tau - 1)/\tau$ and further restricts the step size. These tradeoffs affect the finite-horizon bound, while its asymptotic order is unchanged for fixed system parameters.

3.2 Convergence rates for general smooth non-convex objectives

Without the PL condition, we measure stationarity by the average expected squared gradient norm at the round-start models $\boldsymbol{w}^{(k,p-1)}$ over all $T = KP$ communication rounds.

Theorem 3.2 (Non-convex convergence under balanced cyclic patterns)**.** *Suppose that Assumptions 1–4 hold. Let* $\Phi_0^{\mathrm{nc}} \triangleq \mathbb{E}[F(\boldsymbol{w}^{(0,0)})] - F^*$ *and choose the step size* $\eta_T^{\mathrm{nc}} = \min\{1/(64LP\tau), 1/(L\tau\sqrt{T})\}$. *If* $T \geq 4096P^2$*, so that* $\eta_T^{\mathrm{nc}} = 1/(L\tau\sqrt{T})$*, then, for universal constants* $C_3, C_4 > 0$*,*

$$\frac{1}{T}\sum_{k=0}^{K-1}\sum_{p=1}^{P}\mathbb{E}\left\|\nabla F(\boldsymbol{w}^{(k,p-1)})\right\|^2 \leq \frac{4L\Phi_0^{\mathrm{nc}}}{3\sqrt{T}} + \frac{C_3}{\sqrt{T}}\left[H^2 + \Sigma_{\mathrm{eff}}^2 + \alpha\mathcal{R}_{\mathrm{CW}}\right] + \frac{C_4}{T}\mathcal{R}_{\mathrm{CW}}^2. \tag{14}$$

Proof sketch. As in Theorem 3.1, we control window-induced gradient biases through cancellation at the cycle-initial model $\boldsymbol{w}^{(k,0)}$. To bound the average squared gradient norm over all rounds, we reverse the gradient comparison used in the PL proof: smoothness and squared-displacement estimates bound $\mathbb{E}\|\nabla F(\boldsymbol{w}^{(k,0)})\|^2$ in terms of the average $(1/P)\sum_{p=1}^{P}\mathbb{E}\|\nabla F(\boldsymbol{w}^{(k,p-1)})\|^2$ and a residual, as in (99). We also reformulate the descent recursion using conditional per-round inequalities to retain every round's squared gradient norm. Telescoping over cycles and substituting the prescribed step size gives (14). See Appendix E for the complete proof. □

For fixed system parameters, (14) recovers the standard $\mathcal{O}(T^{-1/2})$ stationarity rate (Karimireddy et al., 2020). Holding other quantities fixed, the dependence on P is at most linear in the $T^{-1/2}$ term and quadratic in the T^{-1} term; the horizon requirement also grows as P^2. Within-window sampling affects Σ_{eff}^2, while $\alpha\mathcal{R}_{\mathrm{CW}}$ couples window heterogeneity with trajectory drift. Even with $S = B$ and $\sigma = 0$, the initial-error and heterogeneity terms generally remain at order $T^{-1/2}$ in this bound.

4 Convergence Bounds under Unbalanced Cyclic Patterns

4.1 Cyclic weights and discrepancy

Note that periodicity alone does not necessarily imply a *balanced cyclic pattern*. To characterize possible imbalance, we first introduce the following quantities. Let $q_{\mathrm{cyc}} \triangleq \lfloor B/d \rfloor$ and $r \triangleq B - q_{\mathrm{cyc}}d$, where $0 \leq r < d$. For each client $m \in [M]$, let c_m denote the number of rounds in which the client is available within one basic cycle. Then, we have $c_m \in \{q_{\mathrm{cyc}}, q_{\mathrm{cyc}} + 1\}$, with exactly rP clients having $c_m = q_{\mathrm{cyc}} + 1$ and the remaining $(d - r)P$ clients having $c_m = q_{\mathrm{cyc}}$. Accordingly, we can define the cyclic-weighted objective $F_\pi(\boldsymbol{w}) \triangleq \sum_{m=1}^{M}\big(c_m/(PB)\big)F_m(\boldsymbol{w})$ and the cyclic discrepancy $\Delta_{\mathrm{cyc}} \triangleq \sum_{m=1}^{M}|c_m/(PB) - 1/M|$.

Lemma 4.1 (Unbalanced cyclic patterns and cyclic discrepancy)**.** *For any fixed model* $\boldsymbol{w}$ *and any complete cycle of window sets* $\{\mathcal{A}^{(k,p)}\}_{p=1}^{P}$ *generated by the prescribed cyclic-window schedule, the cycle-averaged window gradient coincides with the gradient of the cyclic-weighted objective, i.e.,*

$$\frac{1}{P}\sum_{p=1}^{P}\frac{1}{B}\sum_{m\in\mathcal{A}^{(k,p)}}\nabla F_m(\boldsymbol{w}) = \nabla F_\pi(\boldsymbol{w}), \qquad \forall k \in \{0, \ldots, K-1\}. \tag{15}$$

Moreover, the cyclic discrepancy is given exactly by

$$\Delta_{\mathrm{cyc}} = \frac{2r(d-r)}{Bd}. \tag{16}$$

When $d \nmid B$*, meaning that* d *does not divide* B*, the cyclic pattern is unbalanced and* $\Delta_{\mathrm{cyc}} > 0$*, which can introduce optimization error relative to* F*. The pattern is balanced with* $\Delta_{\mathrm{cyc}} = 0$ *if and only if* $d \mid B$*.*

The P window-start positions form a d-spaced grid on the client circle. A client is available whenever a window start falls within the length-B circular interval ending at that client, which yields the coverage counts above. Uniform sampling within each visible window preserves the induced client weights $c_m/(PB)$ in expectation over a cycle. Appendices G.1–G.2 provide the detailed proof.

4.2 Convergence bounds

We relax Assumption 4 while retaining $P \geq 2$. The cyclic gradient bias $\boldsymbol{b}_{\mathrm{cyc}}(\boldsymbol{w}) \triangleq \nabla F_\pi(\boldsymbol{w}) - \nabla F(\boldsymbol{w})$ satisfies $\|\boldsymbol{b}_{\mathrm{cyc}}(\boldsymbol{w})\| \leq (\alpha + \gamma)\Delta_{\mathrm{cyc}} \triangleq B_{\mathrm{cyc}}$ by Assumption 3 and the unit sum of cyclic

weights. Uniform within-window sampling preserves these weights in expectation at a common model. The following two theorems give convergence bounds relative to the equally weighted global objective F, including additional terms due to an unbalanced cyclic pattern.

Theorem 4.1 (PL convergence under an unbalanced cyclic pattern)**.** *Suppose $P \geq 2$ and that Assumptions 1–3 and Assumption 5 hold. Use the same step size and horizon condition as in Theorem 3.1. Then there exist universal constants $\widetilde{C}_1, \widetilde{C}_2, \widetilde{C}_3 > 0$ such that*

$$\mathbb{E}[F(\boldsymbol{w}^{(K,0)})] - F^* \leq \frac{\Phi_0^{\mathrm{PL}}}{T^2} + \widetilde{C}_1 \frac{L \log T}{\mu^2 T} \Sigma_{\mathrm{eff}}^2 + \widetilde{C}_2 \frac{L^2 (\log T)^2}{\mu^3 T^2} \mathcal{R}_{\mathrm{CW}}^2 + \frac{\widetilde{C}_3}{\mu} (\alpha + \gamma)^2 \Delta_{\mathrm{cyc}}^2. \quad (17)$$

Theorem 4.2 (Non-convex convergence under an unbalanced cyclic pattern)**.** *Suppose $P \geq 2$ and Assumptions 1–3 hold. Use the same step size, horizon condition, and universal constants C_3, C_4 as in Theorem 3.2. Then*

$$\frac{1}{T} \sum_{k=0}^{K-1} \sum_{p=1}^{P} \mathbb{E} \left\| \nabla F(\boldsymbol{w}^{(k,p-1)}) \right\|^2 \leq \frac{8 L \Phi_0^{\mathrm{nc}}}{3\sqrt{T}} + \frac{2 C_3}{\sqrt{T}} \left[H^2 + \Sigma_{\mathrm{eff}}^2 + \alpha \mathcal{R}_{\mathrm{CW}} \right] + \frac{2 C_4}{T} \mathcal{R}_{\mathrm{CW}}^2 + \frac{64}{3} (\alpha + \gamma)^2 \Delta_{\mathrm{cyc}}^2. \quad (18)$$

Proof sketch of Theorems 4.1 and 4.2. First, the cycle-initial window biases sum to $P \boldsymbol{b}_{\mathrm{cyc}}(\boldsymbol{w}^{(k,0)})$, while the within-cycle displacement bounds remain valid. Second, we include this bias in the full-cycle moment bounds and gradient coupling. Young's inequality absorbs the extra coupling into descent, leaving residuals proportional to B_{cyc}^2. Third, unrolling the PL recursion gives (17); the reverse gradient comparison and telescoping give (18). See Appendices G.4 and G.5 for details. □

The additional terms bound the effect of the gradient mismatch between F_π and F, scaling as B_{cyc}^2/μ for the PL objective gap and B_{cyc}^2 for non-convex stationarity. Full-window participation, exact gradients, and smaller step sizes do not change the cyclic weights and therefore cannot generally eliminate this mismatch. Unequal weights alone, however, need not cause persistent error: identical local gradients imply $\boldsymbol{b}_{\mathrm{cyc}} = \mathbf{0}$ even when $\Delta_{\mathrm{cyc}} > 0$. With heterogeneous objectives, a stationary point of F_π need not be stationary for F, so a persistent error relative to F may remain.

5 Experimental Validation

In this section, we compare Cycwin-FL under a balanced cyclic pattern with FedAvg on image-classification and controlled PL objectives. Meanwhile, ablation studies examine within-window sampling and an unbalanced cyclic pattern aligned with client heterogeneity. Appendix H gives the complete experimental configurations and provides additional curves across datasets, heterogeneity levels, and different system parameters.

5.1 Experimental setup

For CIFAR-10 (Krizhevsky, 2009), $M = 50$ clients train a ResNet-18 (He et al., 2016) for $T = 500$ rounds using batch size $b_{\mathrm{mb}} = 64$, $\tau = 15$, learning rate $\eta = 0.005$, and Dirichlet concentration $\alpha_{\mathrm{Dir}} = 1$. Cycwin-FL uses $N_w = 1$ and full-window participation $S = B \in \{1, 5, 10\}$, while FedAvg samples the same number of clients uniformly from all M clients. The PL test uses a finite-sum quadratic with dimension $d_{\mathrm{w}} = 20$, $\mu = 1$, $L = 20$, and a known optimum F^*. It uses $M = 50$, $\tau = 5$, $b_{\mathrm{mb}} = 16$, $\eta = 10^{-4}$, and $T = 5{,}000$. The sampling study fixes $B = N_w = 5$ and varies $S \in [5]$. The cyclic pattern study uses $B \in \{5, 3\}$ and $N_w \in \{1, 2, 3\}$. The choice $N_w = 2$ gives an unbalanced cyclic pattern with $d = 2$ and discrepancies $1/5$ and $1/3$, whereas $N_w \in \{1, 3\}$ gives a balanced cyclic pattern. Over-covered clients hold classes 0–4 and under-covered clients hold classes 5–9, making coverage bias observable while preserving population balance.

5.2 Results

Cycwin-FL versus FedAvg under a balanced cyclic pattern. At each matched participation budget, Figure 3 shows closely aligned Cycwin-FL and FedAvg curves. For example, at $B = 10$,

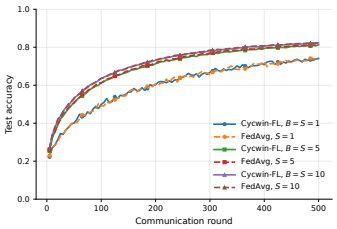


Figure 3: CIFAR-10: Cycwin-FL versus FedAvg.

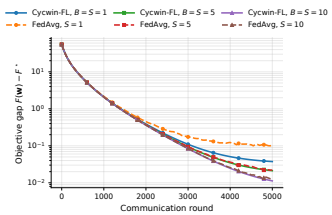


Figure 4: PL quadratic: objective gap.

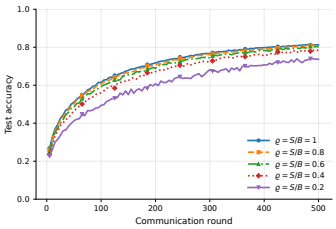


Figure 5: CIFAR-10: sampling-ratio ablation.

Table 1: Balanced versus unbalanced cyclic patterns on CIFAR-10. Final gaps are relative to the balanced $N_w = 1$ baseline at the same B; $\Delta_{\text{acc}} < 0$ or $\Delta_{\text{loss}} > 0$ indicates worse performance.

B	N_w	Cyclic pattern	Final accuracy ↑	Δ_{acc}	Final loss ↓	Δ_{loss}
5	1	Balanced	0.6801 ± 0.0131	Ref.	0.1882 ± 0.0108	Ref.
	2	Unbalanced	0.6441 ± 0.0233	-0.0360	0.2023 ± 0.0111	$+0.0141$
	3	Balanced	0.6849 ± 0.0093	$+0.0048$	0.1912 ± 0.0077	$+0.0030$
3	1	Balanced	0.6425 ± 0.0170	Ref.	0.1933 ± 0.0120	Ref.
	2	Unbalanced	0.5543 ± 0.0224	-0.0882	0.2154 ± 0.0166	$+0.0221$
	3	Balanced	0.6360 ± 0.0311	-0.0065	0.2142 ± 0.0606	$+0.0209$

these two methods achieve accuracies of 0.8228 ± 0.0041 and 0.8223 ± 0.0056, respectively, showing comparable empirical performance under balanced cyclic participation. Increasing B improves both methods. For Cycwin-FL with $S = B$, $\Sigma^2_{\text{eff}} = \sigma^2/(\tau B)$, consistent with reduced mini-batch noise as more client updates are averaged.

Controlled PL objective. As can be seen, Figure 4 shows nearly identical final objective gaps at $B = 10$, both approximately 1.1×10^{-2}. At $B = 1$, the Cycwin-FL and FedAvg objective gaps are 0.0374 ± 0.0001 and 0.1088 ± 0.1504, respectively. Cycwin-FL selects each client exactly once per 50-round cycle, reducing variability in cumulative client participation compared with random sampling. This effect is more pronounced at $B = 1$, where each round only uses one single client update. Full-window participation removes the client-sampling variance term $\omega_{B,S}\gamma^2$ in Σ^2_{eff}, while larger windows also reduce mini-batch variance by averaging more client updates.

Within-window sampling. For fixed B, the definition of Σ^2_{eff} shows that effective variance increases as ϱ decreases. Consistently, reducing ϱ from 1 to 0.2 in Figure 5 lowers final accuracy from 0.8127 ± 0.0086 to 0.7371 ± 0.0176 and increases the number of rounds required to reach 50% accuracy from 47.0 ± 4.0 to 88.8 ± 13.9. A smaller S reduces computation per round but increases finite-population variance and reduces the number of stochastic local updates averaged.

Unbalanced cyclic patterns. Table 1 compares balanced ($N_w = 1, 3$) and unbalanced ($N_w = 2$) cyclic patterns at each fixed B. The two balanced patterns achieve similar accuracies, whereas the unbalanced pattern reduces accuracy relative to $N_w = 1$ by 0.0360 for $B = 5$ and 0.0882 for $B = 3$. In the controlled data partition, over-covered and under-covered clients hold different classes. Unequal availability therefore gives greater weight to some classes, even though the overall population is class-balanced. The observed degradation is consistent with the additional optimization-error terms involving cyclic discrepancy and client heterogeneity in Theorem 4.2.

6 Conclusion

We established PL and non-convex convergence bounds for deterministic cyclic-window FedAvg, identifying the balance criterion and quantifying sampling noise, local drift, cycle order, and coverage bias relative to the equally weighted objective. Experiments support these findings. Future work includes irregular contacts, asynchronous updates, and adaptive windows.

AI USE STATEMENT

We used GPT-6 Astra and GPT-5.6 Sol to help check proof steps for logical gaps and calculation errors. These tools also suggested some mathematical tools, such as inequalities, to support the theoretical analysis. We used them to assist with developing and debugging the experimental code, searching for related work, and improving the presentation and grammar of the manuscript. The authors developed the main research ideas and proof arguments and designed and conducted the experiments. They carefully checked all AI-assisted work and take responsibility for the final content.

ETHICS STATEMENT

This work is theoretical and uses standard public image-classification benchmarks. It does not introduce new personal data or human-subject experiments. In deployment, structured participation may correlate with geography or device status; practitioners should therefore audit whether the induced coverage weights systematically underrepresent particular users or regions.

REPRODUCIBILITY STATEMENT

The appendix specifies the complete probability model and filtration, proves every cycle-level moment bound, verifies the explicit step-size ranges, and derives the PL, non-convex, and unbalanced cyclic pattern results. It also reports the experimental protocol and supplementary curves. The source code for this paper is available at `https://anonymous.4open.science/r/Cycwin-FL-main-52FF`.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konecny, Stefano Mazzocchi, H. Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards federated learning at scale: System design. In *Proceedings of Machine Learning and Systems*, volume 1, 2019.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards understanding biased client selection in federated learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pp. 10351–10375, 2022.

Yae Jee Cho, Pranay Sharma, Gauri Joshi, Zheng Xu, Satyen Kale, and Tong Zhang. On the convergence of federated averaging with cyclic client participation. In *International Conference on Machine Learning*, 2023. arXiv:2302.03109.

Michael Crawshaw and Mingrui Liu. Federated learning under periodic client participation and heterogeneous data: A new communication-efficient algorithm and analysis. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Mohamed Elmahallawy and Tie Luo. AsyncFLEO: Asynchronous federated learning for LEO satellite constellations with high-altitude platforms. In *IEEE International Conference on Big Data*, pp. 5478–5487, 2022.

Mohamed Elmahallawy and Tie Luo. Optimizing federated learning in LEO satellite constellations via intra-plane model propagation and sink satellite scheduling. In *IEEE International Conference on Communications*, pp. 3444–3449, 2023.

Mohamed Elmahallawy, Tie Luo, and Mohamed I. Ibrahem. Secure and efficient federated learning in LEO constellations using decentralized key generation and on-orbit model aggregation. In *IEEE Global Communications Conference*, pp. 5727–5732, 2023. doi: 10.1109/GLOBECOM54140.2023.10436841.

Yann Fraboni, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. Clustered sampling: Low-variance and improved representativity for clients selection in federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 3407–3416, 2021.

Xinran Gu, Kaixuan Huang, Jingzhao Zhang, and Longbo Huang. Fast federated learning in the presence of arbitrary device unavailability. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Yansong Huang, Xuan Li, Moke Zhao, Haiyan Li, and Mugen Peng. Asynchronous federated learning via over-the-air computation in LEO satellite networks. *IEEE Transactions on Wireless Communications*, 23(12):19885–19901, 2024.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2): 1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143, 2020.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtarik. Tighter theory for local SGD on identical and heterogeneous data. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 4519–4529, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Fan Lai, Xiangfeng Zhu, Harsha V. Madhyastha, and Mosharaf Chowdhury. Oort: Efficient federated learning via guided participant selection. In *15th USENIX Symposium on Operating Systems Design and Implementation*, pp. 19–35, 2021.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, 2020a.

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

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *IEEE International Conference on Communications*, pp. 1–7, 2019.

Nasrin Razmi, Bho Matthiesen, Armin Dekorsy, and Petar Popovski. Ground-assisted federated learning in LEO satellite constellations. *IEEE Wireless Communications Letters*, 11(4):717–721, 2022.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

Monica Ribero, Haris Vikalo, and Gustavo de Veciana. Federated learning under intermittent client availability and time-varying communication constraints. *IEEE Journal of Selected Topics in Signal Processing*, 17(1):98–111, 2023.

Angelo Rodio, Francescomaria Faticanti, Othmane Marfoq, Giovanni Neglia, and Emilio Leonardi. Federated learning under heterogeneous and correlated client availability. In *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, pp. 1–10, 2023.

Yichen Ruan, Xiaoxi Zhang, Shu-Che Liang, and Carlee Joe-Wong. Towards flexible device participation in federated learning. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pp. 3403–3411, 2021.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7611–7623, 2020.

Shiqiang Wang and Mingyue Ji. A unified analysis of federated learning with arbitrary client participation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Blake E. Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local SGD for heterogeneous distributed learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Ming Xiang, Stratis Ioannidis, Edmund Yeh, Carlee Joe-Wong, and Lili Su. Efficient federated learning against heterogeneous and non-stationary client unavailability. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Haibo Yang, Peiwen Qiu, Prashant Khanduri, Minghong Fang, and Jia Liu. Understanding server-assisted federated learning in the presence of incomplete client participation. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 56298–56318, 2024.

Jianyi Zhang, Ang Li, Minxue Tang, Jingwei Sun, Xiang Chen, Fan Zhang, Changyou Chen, Yiran Chen, and Hai Li. Fed-CBS: A heterogeneity-aware client sampling mechanism for federated learning via class-imbalance reduction. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 41354–41381, 2023.

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

A PRELIMINARIES

We use the model and cycle indexing of Section 2 and the assumptions in Section 3. This section introduces the additional update notation and probability histories needed for the proofs. The PL condition in Assumption 5 is invoked only for the PL results.

Cycle updates. At local step l of round (k,p), $\hat{\nabla}F_m(\boldsymbol{w})$ denotes the mini-batch gradient in (4) formed from $\mathcal{D}_m^{(k,p,l)}$, with the batch indices suppressed. Here $\mathcal{D}_m^{(k,p,l)} = \mathcal{D}_m^{(t,l)}$ for $t = kP + p - 1$. In cycle–window notation, define the aggregate stochastic direction by

$$\boldsymbol{g}^{(k,p)} \triangleq \frac{1}{S} \sum_{m \in \mathcal{S}^{(k,p)}} \sum_{l=0}^{\tau-1} \hat{\nabla} F_m\big(\boldsymbol{w}_m^{(k,p-1,l)}\big). \tag{19}$$

Then the server update (5) becomes

$$\boldsymbol{w}^{(k,p)} = \boldsymbol{w}^{(k,p-1)} - \eta \boldsymbol{g}^{(k,p)}. \tag{20}$$

For $T = KP$ communication rounds, the final model is $\boldsymbol{w}^{(T)} = \boldsymbol{w}^{(K,0)}$.

Balanced-cycle averaging. Under the condition in Lemma 3.1, each client belongs to B/d windows per cycle (Appendix G.1). Consequently, for arbitrary vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M \in \mathbb{R}^{d_{\mathrm{w}}}$,

$$\frac{1}{P} \sum_{p=1}^{P} \frac{1}{B} \sum_{m \in \mathcal{A}^{(k,p)}} \boldsymbol{x}_m = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{x}_m. \tag{21}$$

Uniform sampling within each window therefore gives each client an expected selection count of S/d per cycle. We use the finite-population factor $\omega_{B,S}$ defined in Section 3, including the convention $\omega_{1,1} = 0$. By Assumption 1, the global objective and every window-average objective are L-smooth.

Probability histories. All random variables are defined on a common probability space $(\Omega, \mathcal{H}, \mathbb{P})$, where Ω contains the possible outcomes, $\mathcal{H}$ is the sigma-algebra of events, and $\mathbb{P}$ assigns their probabilities. We use sigma-algebras to describe the information available during training. Here, $\sigma(Y)$ denotes the information generated by Y, while $\mathcal{G}_1 \vee \mathcal{G}_2$ combines two sigma-algebras. The notation $\bigvee_i \mathcal{G}_i$ combines an indexed collection.

For round (k,p), define

$$\begin{aligned} \mathcal{U}_{k,p} &\triangleq \sigma\Big(\mathcal{S}^{(k,p)}\Big), \\ \mathcal{X}_{k,p,l} &\triangleq \sigma\Big(\Big\{\mathcal{D}_m^{(k,p,j)} : m \in \mathcal{A}^{(k,p)},\ 0 \le j < l\Big\}\Big), \qquad 0 \le l \le \tau. \end{aligned}$$

Thus, $\mathcal{U}_{k,p}$ records the selected clients, and $\mathcal{X}_{k,p,l}$ records the mini-batches revealed before local step l. Set $\mathcal{X}_{k,p} = \mathcal{X}_{k,p,\tau}$. At $l = 0$, $\mathcal{X}_{k,p,0} = \{\varnothing, \Omega\}$, the trivial sigma-algebra containing only the impossible and certain events. Draws for unselected clients may be defined formally but are not used in the updates.

The cycle and round histories are

$$\begin{aligned} \mathcal{H}_k &\triangleq \sigma\Big(\boldsymbol{w}^{(0,0)}\Big) \vee \bigvee_{i=0}^{k-1} \bigvee_{s=1}^{P} (\mathcal{U}_{i,s} \vee \mathcal{X}_{i,s}), \\ \mathcal{H}_{k,p} &\triangleq \mathcal{H}_k \vee \bigvee_{s=1}^{p} (\mathcal{U}_{k,s} \vee \mathcal{X}_{k,s}), \qquad 0 \le p \le P. \end{aligned}$$

Empty joins are understood as $\{\varnothing, \Omega\}$. The history $\mathcal{H}_k$ contains the initial model and all random choices from earlier cycles; $\mathcal{H}_{k,p}$ additionally contains those from the first p rounds of cycle k. Consequently,

$$\mathcal{H}_{k,0} = \mathcal{H}_k, \qquad \mathcal{H}_{k,p} = \mathcal{H}_{k,p-1} \vee \mathcal{U}_{k,p} \vee \mathcal{X}_{k,p}, \qquad \mathcal{H}_{k,P} = \mathcal{H}_{k+1}.$$

The full training history is $\mathcal{H}_\infty \triangleq \bigvee_{k\geq 0} \mathcal{H}_k \subseteq \mathcal{H}$.

Immediately before round (k,p), the model $\boldsymbol{w}^{(k,p-1)}$ is $\mathcal{H}_{k,p-1}$-measurable, meaning that it is determined by this history. The available set $\mathcal{A}^{(k,p)}$ is deterministic. The client-sampling protocol in Section 2 is uniform conditional on this history: for every $\mathcal{C} \subseteq \mathcal{A}^{(k,p)}$ with $|\mathcal{C}| = S$,

$$\mathbb{P}\Big(\mathcal{S}^{(k,p)} = \mathcal{C} \,\Big|\, \mathcal{H}_{k,p-1}\Big) = \frac{1}{\binom{B}{S}} \qquad \text{almost surely.}$$

Before local step l, the relevant history is

$$\mathcal{F}_{k,p,l} \triangleq \mathcal{H}_{k,p-1} \vee \mathcal{U}_{k,p} \vee \mathcal{X}_{k,p,l}.$$

It includes the past rounds, the current selected clients, and the preceding local mini-batches. Hence, for each selected client m, $\boldsymbol{w}_m^{(k,p-1,l)}$ is $\mathcal{F}_{k,p,l}$-measurable before drawing $\mathcal{D}_m^{(k,p,l)}$. The index $p-1$ on the local model records its initialization from $\boldsymbol{w}^{(k,p-1)}$, whereas p on the mini-batch identifies the current round.

Assumption 2 imposes conditional unbiasedness and bounded noise variance. The mini-batch sampling protocol in Section 2 gives independence across clients at each local step conditional on $\mathcal{F}_{k,p,l}$. The proofs condition on $\mathcal{H}_k$ for cycle-level analysis, on $\mathcal{H}_{k,p-1}$ for round-level analysis, and on $\mathcal{F}_{k,p,l}$ for local-gradient calculations. These conditional assumptions do not require independence across communication rounds.

A.1 Round Error Decomposition

We express each round update as the global direction plus three residuals: window-induced gradient bias, local-update drift, and a centered joint innovation.

A.1.1 Window-Induced Gradient Bias

The current window may have a different average gradient from the full population. We measure this mismatch by

$$\boldsymbol{\epsilon}^{(k,p)} \triangleq \overline{\nabla F}_{\mathcal{A}^{(k,p)}}(\boldsymbol{w}^{(k,p-1)}) - \nabla F(\boldsymbol{w}^{(k,p-1)}). \tag{22}$$

By Assumption 3, $\|\boldsymbol{\epsilon}^{(k,p)}\| \leq \alpha$.

A.1.2 Local-Update Drift

Local steps move each selected client away from the round-start model. The resulting gradient drift is

$$\boldsymbol{\psi}^{(k,p)} \triangleq \frac{1}{S} \sum_{m\in\mathcal{S}^{(k,p)}} \sum_{l=0}^{\tau-1} \Big(\nabla F_m(\boldsymbol{w}_m^{(k,p-1,l)}) - \nabla F_m(\boldsymbol{w}^{(k,p-1)})\Big). \tag{23}$$

A.1.3 Joint Round-Level Random Innovation

Client sampling and mini-batch sampling are nested sources of randomness in the same stochastic round update. Their centered contributions are combined into the joint innovation

$$\begin{aligned}\boldsymbol{\nu}^{(k,p)} \triangleq \tau &\left[\frac{1}{S} \sum_{m\in\mathcal{S}^{(k,p)}} \nabla F_m(\boldsymbol{w}^{(k,p-1)}) - \overline{\nabla F}_{\mathcal{A}^{(k,p)}}(\boldsymbol{w}^{(k,p-1)})\right] \\ &+ \frac{1}{S} \sum_{m\in\mathcal{S}^{(k,p)}} \sum_{l=0}^{\tau-1} \Big[\hat{\nabla} F_m(\boldsymbol{w}_m^{(k,p-1,l)}) - \nabla F_m(\boldsymbol{w}_m^{(k,p-1,l)})\Big].\end{aligned} \tag{24}$$

The first term in (24) represents client-sampling noise at the round-start model, while the second represents stochastic-gradient noise along the local trajectories. The local-drift term $\boldsymbol{\psi}^{(k,p)}$ also depends on the sampled clients and mini-batches. Since it need not have zero conditional mean, we control it through separate first- and second-moment bounds. The finite-population sampling effect and stochastic-gradient variance together determine the conditional second moment of $\boldsymbol{\nu}^{(k,p)}$.

A.2 Conditional Expectations

Using the pre-round history defined in Appendix A, we write the round-level conditional expectation of any integrable random variable X as

$$\mathbb{E}_{k,p}[X] \triangleq \mathbb{E}[X \mid \mathcal{H}_{k,p-1}].$$

The subscript (k,p) identifies the round being averaged; the conditioning sigma-algebra is the information available immediately before that round. Thus, for a random variable measurable by the end of round (k,p), $\mathbb{E}_{k,p}$ jointly averages the current client sampling and all current mini-batch draws. When these two mechanisms must be exposed separately, first condition additionally on the client-sampling sigma-algebra $\mathcal{U}_{k,p}$ and apply the tower property:

$$\mathbb{E}_{k,p}[X] = \mathbb{E}\Big[\mathbb{E}[X \mid \mathcal{H}_{k,p-1} \vee \mathcal{U}_{k,p}] \;\Big|\; \mathcal{H}_{k,p-1}\Big]. \tag{25}$$

The inner expectation averages the mini-batch randomness after the selected set is fixed, and the outer expectation averages client sampling. The histories $\mathcal{F}_{k,p,l}$ in Assumption 2 refine this conditioning within each round. The tower property in (25) applies without unconditional independence between the two mechanisms, allowing both sources of randomness to be handled through $\mathbb{E}_{k,p}$.

For conditioning at the beginning of fundamental cycle k, we similarly write

$$\mathbb{E}_k[\,\cdot\,] \triangleq \mathbb{E}[\,\cdot \mid \mathcal{H}_k].$$

We call the cycle-initial model $\boldsymbol{w}^{(k,0)}$ the cycle anchor and define its gradient norm and the total cycle displacement by

$$G_k \triangleq \|\nabla F(\boldsymbol{w}^{(k,0)})\|, \qquad \boldsymbol{\delta}_k \triangleq \boldsymbol{w}^{(k+1,0)} - \boldsymbol{w}^{(k,0)}.$$

A.3 Per-Round and Fundamental-Cycle Decomposition

Adding and subtracting $\nabla F_m(\boldsymbol{w}^{(k,p-1)})$ and the available-window average in (19) yields

$$\boldsymbol{g}^{(k,p)} = \tau \nabla F(\boldsymbol{w}^{(k,p-1)}) + \tau \boldsymbol{\epsilon}^{(k,p)} + \boldsymbol{\psi}^{(k,p)} + \boldsymbol{\nu}^{(k,p)}. \tag{26}$$

Summing (20) over the P rounds of one fundamental cycle gives

$$\boldsymbol{\delta}_k = -\eta \sum_{p=1}^{P} \boldsymbol{g}^{(k,p)}.$$

To express this displacement in terms of the four update components, define the cycle sums by

$$\boldsymbol{h}_\Sigma^{(k)} \triangleq \sum_{p=1}^{P} \nabla F(\boldsymbol{w}^{(k,p-1)}), \quad \boldsymbol{\epsilon}_\Sigma^{(k)} \triangleq \sum_{p=1}^{P} \boldsymbol{\epsilon}^{(k,p)},$$
$$\boldsymbol{\psi}_\Sigma^{(k)} \triangleq \sum_{p=1}^{P} \boldsymbol{\psi}^{(k,p)}, \qquad \boldsymbol{\nu}_\Sigma^{(k)} \triangleq \sum_{p=1}^{P} \boldsymbol{\nu}^{(k,p)}.$$

With these cycle sums, the preceding displacement identity becomes

$$\boldsymbol{\delta}_k = -\eta\tau \boldsymbol{h}_\Sigma^{(k)} - \eta\tau \boldsymbol{\epsilon}_\Sigma^{(k)} - \eta \boldsymbol{\psi}_\Sigma^{(k)} - \eta \boldsymbol{\nu}_\Sigma^{(k)}. \tag{27}$$

B Technical Lemmas under a Balanced Cyclic Pattern

The cycle analysis uses three ingredients: a joint moment bound for client sampling and mini-batch noise, first- and second-moment bounds for local drift, and exact window-bias cancellation at a fixed model. Together, these lemmas separate the effects of randomness from the drift caused by local updates and sequential window visits.

B.1 Joint Random-Innovation Moment Bound

The first lemma quantifies the joint variance within each round and shows how it accumulates over a cycle.

Lemma B.1 (Joint client-sampling and stochastic-gradient randomness)**.** *With the finite-population factor $\omega_{B,S}$ defined in Section 3, set*

$$V_\nu \triangleq \tau^2 \omega_{B,S} \gamma^2 + \frac{\tau \sigma^2}{S}.$$

Under the joint round-level expectation,

$$\mathbb{E}_{k,p}[\boldsymbol{\nu}^{(k,p)}] = \mathbf{0}, \tag{28}$$

$$\mathbb{E}_{k,p}\|\boldsymbol{\nu}^{(k,p)}\|^2 \le V_\nu. \tag{29}$$

Moreover, for every integer $1 \le R \le P$,

$$\mathbb{E}_k \left\| \sum_{p=1}^{R} \boldsymbol{\nu}^{(k,p)} \right\|^2 \le R V_\nu. \tag{30}$$

Proof. Fix the pre-round history $\mathcal{H}_{k,p-1}$ and abbreviate $\boldsymbol{w} = \boldsymbol{w}^{(k,p-1)}$ and $\mathcal{A} = \mathcal{A}^{(k,p)}$. Write

$$\boldsymbol{u} \triangleq \tau \left[\frac{1}{S} \sum_{m \in \mathcal{S}^{(k,p)}} \nabla F_m(\boldsymbol{w}) - \frac{1}{B} \sum_{m \in \mathcal{A}} \nabla F_m(\boldsymbol{w}) \right]$$

for the client-sampling contribution. For each selected client and local step, define

$$\boldsymbol{\zeta}_{m,l}^{(k,p)} \triangleq \hat{\nabla} F_m\left(\boldsymbol{w}_m^{(k,p-1,l)}\right) - \nabla F_m\left(\boldsymbol{w}_m^{(k,p-1,l)}\right),$$

and write

$$\boldsymbol{z} \triangleq \frac{1}{S} \sum_{m \in \mathcal{S}^{(k,p)}} \sum_{l=0}^{\tau-1} \boldsymbol{\zeta}_{m,l}^{(k,p)}.$$

Thus, $\boldsymbol{\nu}^{(k,p)} = \boldsymbol{u} + \boldsymbol{z}$.

Condition first on $\mathcal{H}_{k,p-1} \vee \mathcal{U}_{k,p}$, so the sampled set and $\boldsymbol{u}$ are fixed. By Assumption 2 and the mini-batch sampling protocol in Section 2, the centered noises are martingale differences across local steps and are conditionally independent across clients within each local step. Iterated conditioning over $\mathcal{F}_{k,p,0}, \ldots, \mathcal{F}_{k,p,\tau}$ gives

$$\mathbb{E}[\boldsymbol{z} \mid \mathcal{H}_{k,p-1} \vee \mathcal{U}_{k,p}] = \mathbf{0}.$$

The same conditioning eliminates the cross terms between distinct client–step pairs, yielding

$$\begin{aligned} &\mathbb{E}\big[\|\boldsymbol{z}\|^2 \mid \mathcal{H}_{k,p-1} \vee \mathcal{U}_{k,p}\big] \\ &\quad = \frac{1}{S^2} \sum_{m \in \mathcal{S}^{(k,p)}} \sum_{l=0}^{\tau-1} \mathbb{E}\Big[\|\boldsymbol{\zeta}_{m,l}^{(k,p)}\|^2 \mid \mathcal{H}_{k,p-1} \vee \mathcal{U}_{k,p}\Big] \le \frac{\tau \sigma^2}{S}. \end{aligned}$$

Therefore,

$$\mathbb{E}\big[\|\boldsymbol{u} + \boldsymbol{z}\|^2 \mid \mathcal{H}_{k,p-1} \vee \mathcal{U}_{k,p}\big] = \|\boldsymbol{u}\|^2 + \mathbb{E}\big[\|\boldsymbol{z}\|^2 \mid \mathcal{H}_{k,p-1} \vee \mathcal{U}_{k,p}\big],$$

because the conditional cross term is zero. We next average over client sampling. Conditional on $\mathcal{H}_{k,p-1}$, uniform sampling without replacement ensures

$$\mathbb{E}[\boldsymbol{u} \mid \mathcal{H}_{k,p-1}] = \mathbf{0}.$$

It also yields the variance bound

$$\begin{aligned} \mathbb{E}[\|\boldsymbol{u}\|^2 \mid \mathcal{H}_{k,p-1}] &= \tau^2 \omega_{B,S} \frac{1}{B} \sum_{m \in \mathcal{A}} \big\|\nabla F_m(\boldsymbol{w}) - \overline{\nabla F}_{\mathcal{A}}(\boldsymbol{w})\big\|^2 \\ &\le \tau^2 \omega_{B,S} \gamma^2. \end{aligned}$$

Combining the two variance bounds through (25) yields

$$\mathbb{E}_{k,p}[\boldsymbol{\nu}^{(k,p)}] = \mathbf{0}, \qquad \mathbb{E}_{k,p}\|\boldsymbol{\nu}^{(k,p)}\|^2 \le \tau^2\omega_{B,S}\gamma^2 + \frac{\tau\sigma^2}{S} = V_\nu.$$

Finally, $\boldsymbol{\nu}^{(k,p)}$ is $\mathcal{H}_{k,p}$-measurable and satisfies

$$\mathbb{E}\Big[\boldsymbol{\nu}^{(k,p)} \mid \mathcal{H}_{k,p-1}\Big] = \mathbf{0}.$$

Consequently, the inner products between noise terms from different rounds have zero conditional expectation. For $1 \le s < p \le R$, $\boldsymbol{\nu}^{(k,s)}$ is $\mathcal{H}_{k,p-1}$-measurable, and the tower property gives

$$\mathbb{E}_k\left\langle \boldsymbol{\nu}^{(k,s)}, \boldsymbol{\nu}^{(k,p)} \right\rangle = 0.$$

Expanding the squared norm of the partial sum and applying (29) proves (30). □

B.2 Local-Drift First Moment

We next bound the gradient drift accumulated along the local trajectories. For the current round, write

$$C_\psi \triangleq L\eta\tau(\tau-1), \qquad G_{k,p} \triangleq \|\nabla F(\boldsymbol{w}^{(k,p-1)})\|.$$

Lemma B.2 (Local-drift first moment). *If $L\eta(\tau-1) \le 1/2$, then*

$$\mathbb{E}_{k,p}\|\boldsymbol{\psi}^{(k,p)}\| \le C_\psi\big(G_{k,p} + \alpha + \gamma + \sigma\big). \tag{31}$$

Proof. Let

$$\boldsymbol{\psi}_{m,l}^{(k,p)} \triangleq \nabla F_m(\boldsymbol{w}_m^{(k,p-1,l)}) - \nabla F_m(\boldsymbol{w}^{(k,p-1)}).$$

For each local step, define the expected drift averaged over the selected clients:

$$D_l \triangleq \mathbb{E}_{k,p}\left[\frac{1}{S}\sum_{m\in\mathcal{S}^{(k,p)}}\|\boldsymbol{\psi}_{m,l}^{(k,p)}\|\right], \qquad D_0 = 0.$$

Smoothness and the local recursion imply

$$D_l \le L\eta\sum_{j=0}^{l-1}\mathbb{E}_{k,p}\left[\frac{1}{S}\sum_{m\in\mathcal{S}^{(k,p)}}\|\hat{\nabla}F_m(\boldsymbol{w}_m^{(k,p-1,j)})\|\right].$$

To bound the stochastic-gradient norms under $\mathbb{E}_{k,p}$, we separate the round-start gradient, local drift, and mini-batch noise. Assumption 2 and Jensen's inequality give

$$\mathbb{E}_{k,p}\left[\frac{1}{S}\sum_{m\in\mathcal{S}^{(k,p)}}\|\hat{\nabla}F_m(\boldsymbol{w}_m^{(k,p-1,j)})\|\right] \le G_{k,p} + \alpha + \gamma + \sigma + D_j.$$

Here, Assumption 3 gives, for every available client,

$$\|\nabla F_m(\boldsymbol{w}^{(k,p-1)})\| \le G_{k,p} + \alpha + \gamma.$$

Defining $D_\Sigma \triangleq \sum_{l=0}^{\tau-1} D_l$ and summing the preceding recursion over l yields

$$D_\Sigma \le \frac{L\eta\tau(\tau-1)}{2}(G_{k,p} + \alpha + \gamma + \sigma) + L\eta(\tau-1)D_\Sigma.$$

Under the condition $L\eta(\tau-1) \le 1/2$,

$$D_\Sigma \le L\eta\tau(\tau-1)(G_{k,p} + \alpha + \gamma + \sigma) = C_\psi(G_{k,p} + \alpha + \gamma + \sigma).$$

Finally, the definition of $\boldsymbol{\psi}^{(k,p)}$ and the triangle inequality give $\mathbb{E}_{k,p}\|\boldsymbol{\psi}^{(k,p)}\| \le D_\Sigma$. Combining this with the preceding bound proves (31). □

B.3 LOCAL-DRIFT SECOND MOMENT

The squared local drift enters the smoothness remainder of the convergence proof. Define its coefficients by

$$C_{\psi,2} \triangleq \frac{L^2\eta^2\tau^2(\tau-1)(2\tau-1)}{3}, \qquad D_{\psi,2} \triangleq C_{\psi,2}\big[6(\alpha^2+\gamma^2)+\sigma^2\big].$$

Lemma B.3 (Local-drift second moment). *If $L^2\eta^2\tau(\tau-1) \le 1/2$, then*

$$\mathbb{E}_{k,p}\|\boldsymbol{\psi}^{(k,p)}\|^2 \le 6C_{\psi,2}G_{k,p}^2 + D_{\psi,2}. \tag{32}$$

Proof. Use the same $\boldsymbol{\psi}_{m,l}^{(k,p)}$ as in Lemma B.2 and define

$$Q_l \triangleq \mathbb{E}_{k,p}\left[\frac{1}{S}\sum_{m\in\mathcal{S}^{(k,p)}}\|\boldsymbol{\psi}_{m,l}^{(k,p)}\|^2\right], \qquad Q_0 = 0.$$

For $l \ge 1$, smoothness and Cauchy–Schwarz give

$$Q_l \le L^2\eta^2 l\sum_{j=0}^{l-1}\mathbb{E}_{k,p}\left[\frac{1}{S}\sum_{m\in\mathcal{S}^{(k,p)}}\|\hat{\nabla}F_m(\boldsymbol{w}_m^{(k,p-1,j)})\|^2\right].$$

Under the joint conditional expectation $\mathbb{E}_{k,p}$, Assumption 2 and

$$\nabla F_m(\boldsymbol{w}_m^{(k,p-1,j)}) = \nabla F_m(\boldsymbol{w}^{(k,p-1)}) + \boldsymbol{\psi}_{m,j}^{(k,p)}$$

give

$$\begin{aligned}
&\mathbb{E}_{k,p}\left[\frac{1}{S}\sum_{m\in\mathcal{S}^{(k,p)}}\|\hat{\nabla}F_m(\boldsymbol{w}_m^{(k,p-1,j)})\|^2\right] \\
&\le 2\mathbb{E}_{k,p}\left[\frac{1}{S}\sum_{m\in\mathcal{S}^{(k,p)}}\|\nabla F_m(\boldsymbol{w}^{(k,p-1)})\|^2\right] + 2Q_j + \sigma^2 \\
&\le 6G_{k,p}^2 + 6\alpha^2 + 6\gamma^2 + 2Q_j + \sigma^2,
\end{aligned}$$

where the last line uses Assumption 3. Let $Q_\Sigma \triangleq \sum_{l=0}^{\tau-1} Q_l$. Summing over l yields

$$\begin{aligned}
Q_\Sigma &\le L^2\eta^2(6G_{k,p}^2 + 6\alpha^2 + 6\gamma^2 + \sigma^2)\sum_{l=0}^{\tau-1} l^2 \\
&\quad + 2L^2\eta^2\sum_{l=0}^{\tau-1} l\sum_{j=0}^{l-1} Q_j.
\end{aligned}$$

Using the two elementary bounds

$$\sum_{l=0}^{\tau-1} l^2 = \frac{\tau(\tau-1)(2\tau-1)}{6}, \qquad \sum_{l=0}^{\tau-1} l\sum_{j=0}^{l-1} Q_j \le \frac{\tau(\tau-1)}{2}Q_\Sigma,$$

we obtain

$$\begin{aligned}
Q_\Sigma &\le L^2\eta^2\frac{\tau(\tau-1)(2\tau-1)}{6}(6G_{k,p}^2 + 6\alpha^2 + 6\gamma^2 + \sigma^2) \\
&\quad + L^2\eta^2\tau(\tau-1)Q_\Sigma.
\end{aligned}$$

Under the condition $L^2\eta^2\tau(\tau-1) \le 1/2$, rearranging and using the definition of $C_{\psi,2}$ gives

$$\tau Q_\Sigma \le 6C_{\psi,2}G_{k,p}^2 + 6C_{\psi,2}(\alpha^2+\gamma^2) + C_{\psi,2}\sigma^2.$$

Finally,

$$\mathbb{E}_{k,p}\|\boldsymbol{\psi}^{(k,p)}\|^2 \le \mathbb{E}_{k,p}\left[\frac{\tau}{S}\sum_{m\in\mathcal{S}^{(k,p)}}\sum_{l=0}^{\tau-1}\|\boldsymbol{\psi}_{m,l}^{(k,p)}\|^2\right] = \tau Q_\Sigma,$$

which proves (32). The factor $(2\tau-1)/3$ in $C_{\psi,2}$ comes from the exact sum $\sum_{l=0}^{\tau-1} l^2$. In particular, $C_{\psi,2} = 0$ when $\tau = 1$, since the sole local gradient is evaluated at the round-start model. □

B.4 Exact Window-Bias Cancellation at a Fixed Anchor

A balanced cyclic pattern cancels window mismatch at a common model. The following lemma relates this cancellation to the actual sequential iterates. For any fixed $\boldsymbol{w}$, define

$$\boldsymbol{\epsilon}^{(k,p)}(\boldsymbol{w}) \triangleq \frac{1}{B} \sum_{m \in \mathcal{A}^{(k,p)}} \nabla F_m(\boldsymbol{w}) - \nabla F(\boldsymbol{w}). \tag{33}$$

Here, $\boldsymbol{\epsilon}^{(k,p)} = \boldsymbol{\epsilon}^{(k,p)}(\boldsymbol{w}^{(k,p-1)})$ is the round-start bias defined in (22).

Lemma B.4 (Fundamental-cycle window-bias cancellation)**.** *Under Assumption 4,*

$$\sum_{p=1}^{P} \boldsymbol{\epsilon}^{(k,p)}(\boldsymbol{w}) = \mathbf{0} \qquad \textit{for every fixed } \boldsymbol{w}, \tag{34}$$

$$\|\boldsymbol{\epsilon}_{\Sigma}^{(k)}\| \le 2L \sum_{p=1}^{P} \|\boldsymbol{w}^{(k,p-1)} - \boldsymbol{w}^{(k,0)}\|. \tag{35}$$

Proof. Equation (21) applied to $\boldsymbol{x}_m = \nabla F_m(\boldsymbol{w})$ implies (34). For the actual iterates,

$$\boldsymbol{\epsilon}^{(k,p)} = \boldsymbol{\epsilon}^{(k,p)}(\boldsymbol{w}^{(k,0)}) + \left[\boldsymbol{\epsilon}^{(k,p)}(\boldsymbol{w}^{(k,p-1)}) - \boldsymbol{\epsilon}^{(k,p)}(\boldsymbol{w}^{(k,0)})\right].$$

The anchor terms cancel by (34). Smoothness gives

$$\left\|\boldsymbol{\epsilon}^{(k,p)}(\boldsymbol{w}^{(k,p-1)}) - \boldsymbol{\epsilon}^{(k,p)}(\boldsymbol{w}^{(k,0)})\right\| \le 2L\|\boldsymbol{w}^{(k,p-1)} - \boldsymbol{w}^{(k,0)}\|.$$

Summing proves (35). □

Equation (34) is a pathwise identity for the full-window means at a common model. The fluctuation caused by selecting only S clients per window is part of $\boldsymbol{\nu}^{(k,p)}$ and is controlled by Lemma B.1.

C Cycle-Level Moment Bounds

Using the cycle decomposition (27), we now bound the alignment of the cycle update with the anchor gradient and the squared size of that update. These estimates control the descent term and the smoothness remainder, respectively. Both rely on bounds for the models' displacement from the cycle anchor.

C.1 First Moment: Alignment with the Cycle-Anchor Gradient

The alignment bound depends on the distance from the cycle anchor and the gradient norm along the trajectory. Define

$$a_{k,p} \triangleq \mathbb{E}_k\|\boldsymbol{w}^{(k,p-1)} - \boldsymbol{w}^{(k,0)}\|, \tag{36}$$

$$b_{k,p} \triangleq \mathbb{E}_k\|\nabla F(\boldsymbol{w}^{(k,p-1)})\|. \tag{37}$$

Taking the conditional expectation of (27) and using Lemma B.1 gives

$$\mathbb{E}_k[\boldsymbol{\delta}_k] = -\eta\tau\mathbb{E}_k[\boldsymbol{h}_{\Sigma}^{(k)}] - \eta\tau\mathbb{E}_k[\boldsymbol{\epsilon}_{\Sigma}^{(k)}] - \eta\mathbb{E}_k[\boldsymbol{\psi}_{\Sigma}^{(k)}]. \tag{38}$$

Thus the joint innovation creates no first-order bias. Smoothness and Lemma B.4 control the first two terms in (38):

$$-\eta\tau\left\langle \nabla F(\boldsymbol{w}^{(k,0)}), \mathbb{E}_k[\boldsymbol{h}_{\Sigma}^{(k)}]\right\rangle \le -\eta\tau P G_k^2 + \eta\tau L G_k \sum_{p=1}^{P} a_{k,p},$$

$$-\eta\tau\left\langle \nabla F(\boldsymbol{w}^{(k,0)}), \mathbb{E}_k[\boldsymbol{\epsilon}_{\Sigma}^{(k)}]\right\rangle \le 2\eta\tau L G_k \sum_{p=1}^{P} a_{k,p}. \tag{39}$$

For the local-drift term, recall $H = \alpha + \gamma + \sigma$. Lemma B.2 gives

$$-\eta \left\langle \nabla F(\boldsymbol{w}^{(k,0)}), \mathbb{E}_k[\boldsymbol{\psi}_\Sigma^{(k)}] \right\rangle \le \eta C_\psi G_k \sum_{p=1}^{P} b_{k,p} + \eta C_\psi P H G_k.$$

Combining the global-gradient, window-mismatch, and local-drift bounds yields

$$\begin{aligned} \left\langle \nabla F(\boldsymbol{w}^{(k,0)}), \mathbb{E}_k[\boldsymbol{\delta}_k] \right\rangle \le -\eta\tau P G_k^2 + 3\eta\tau L G_k \sum_{p=1}^{P} a_{k,p} \\ + \eta C_\psi G_k \sum_{p=1}^{P} b_{k,p} + \eta C_\psi P H G_k. \end{aligned} \tag{40}$$

It remains to bound the two trajectory sums. For $p \ge 2$, summing (20) over the first $p-1$ rounds and substituting (26) gives

$$\begin{aligned} \boldsymbol{w}^{(k,p-1)} - \boldsymbol{w}^{(k,0)} = -\eta\tau \sum_{s=1}^{p-1} \nabla F(\boldsymbol{w}^{(k,s-1)}) - \eta\tau \sum_{s=1}^{p-1} \boldsymbol{\epsilon}^{(k,s)} \\ - \eta \sum_{s=1}^{p-1} \boldsymbol{\psi}^{(k,s)} - \eta \sum_{s=1}^{p-1} \boldsymbol{\nu}^{(k,s)}. \end{aligned} \tag{41}$$

Lemma B.1 and Jensen's inequality imply

$$\mathbb{E}_k \left\| \sum_{s=1}^{p-1} \boldsymbol{\nu}^{(k,s)} \right\| \le \sqrt{(p-1)V_\nu} \le (p-1)\sqrt{V_\nu}.$$

Together with the bound $\|\boldsymbol{\epsilon}^{(k,p)}\| \le \alpha$, Lemma B.2, and the triangle inequality, this gives

$$a_{k,p} \le \eta(\tau + C_\psi) \sum_{s=1}^{p-1} b_{k,s} + \eta(p-1)\big[\tau\alpha + C_\psi H + \sqrt{V_\nu}\big], \tag{42}$$

$$C_{\mathrm{dr},1} \triangleq \tau + C_\psi, \tag{43}$$

$$C_{\mathrm{dr},2} \triangleq \tau\alpha + C_\psi H + \sqrt{V_\nu}, \tag{44}$$

$$b_{k,s} \le G_k + L a_{k,s}, \tag{45}$$

$$a_{k,p} \le \eta(C_{\mathrm{dr},1} G_k + C_{\mathrm{dr},2})(p-1) + \eta C_{\mathrm{dr},1} L \sum_{s=1}^{p-1} a_{k,s}. \tag{46}$$

Consider the scalar recursion

$$x_1 = 0, \qquad x_p \le c_{\mathrm{rec}}(p-1) + \vartheta \sum_{s<p} x_s,$$

where $c_{\mathrm{rec}} \ge 0$ and $\vartheta > 0$. The sequence satisfies

$$x_p \le \frac{c_{\mathrm{rec}}}{\vartheta}\big((1+\vartheta)^{p-1} - 1\big). \tag{47}$$

The bound is exact at $p = 1$. For the inductive step, suppose it holds through $p-1$. Summing these bounds as a geometric series gives

$$\vartheta \sum_{s=1}^{p-1} x_s \le c_{\mathrm{rec}} \left[\frac{(1+\vartheta)^{p-1} - 1}{\vartheta} - (p-1) \right].$$

Adding the remaining $c_{\mathrm{rec}}(p-1)$ term then proves the bound at p, establishing (47) by induction. Applying (47) with $c_{\mathrm{rec}} = \eta(C_{\mathrm{dr},1} G_k + C_{\mathrm{dr},2})$ and $\vartheta = q \triangleq \eta L C_{\mathrm{dr},1}$ gives

$$q \triangleq \eta L C_{\mathrm{dr},1}, \tag{48}$$

$$a_{k,p} \le \frac{C_{\mathrm{dr},1} G_k + C_{\mathrm{dr},2}}{C_{\mathrm{dr},1} L}\big[(1+q)^{p-1} - 1\big]. \tag{49}$$

Introduce the geometric sums and their induced drift coefficients by

$$h_P(q) \triangleq \sum_{j=0}^{P-1}(1+q)^j = \frac{(1+q)^P - 1}{q}, \tag{50}$$

$$r_P(q) \triangleq h_P(q) - P, \tag{51}$$

$$\alpha_{\mathrm{d}} \triangleq \frac{r_P(q)}{L}, \tag{52}$$

$$\beta_{\mathrm{d}} \triangleq \frac{C_{\mathrm{dr},2}}{C_{\mathrm{dr},1}L} r_P(q), \tag{53}$$

$$\alpha_{\mathrm{g}} \triangleq h_P(q), \tag{54}$$

$$\beta_{\mathrm{g}} \triangleq \frac{C_{\mathrm{dr},2}}{C_{\mathrm{dr},1}} r_P(q). \tag{55}$$

Summing (49) and applying smoothness through (45) gives the two trajectory bounds:

$$\sum_{p=1}^{P} a_{k,p} \le \alpha_{\mathrm{d}} G_k + \beta_{\mathrm{d}}, \tag{56}$$

$$\sum_{p=1}^{P} b_{k,p} \le \alpha_{\mathrm{g}} G_k + \beta_{\mathrm{g}}. \tag{57}$$

Define

$$A_{\mathrm{align}}(\eta) \triangleq \eta\tau P - 3\eta\tau L\alpha_{\mathrm{d}} - \eta C_\psi \alpha_{\mathrm{g}}, \tag{58}$$

$$B_{\mathrm{align}}(\eta) \triangleq 3\eta\tau L\beta_{\mathrm{d}} + \eta C_\psi \beta_{\mathrm{g}} + \eta C_\psi P H. \tag{59}$$

Substituting (56)–(57) in (40) and using (58)–(59) completes the first-moment argument:

$$\left\langle \nabla F(\boldsymbol{w}^{(k,0)}), \mathbb{E}_k[\boldsymbol{\delta}_k] \right\rangle \le -A_{\mathrm{align}}(\eta) G_k^2 + B_{\mathrm{align}}(\eta) G_k. \tag{60}$$

For fixed system parameters, $A_{\mathrm{align}}(\eta) = \eta\tau P - \mathcal{O}(\eta^2)$ and $B_{\mathrm{align}}(\eta) = \mathcal{O}(\eta^2)$ as $\eta \to 0$.

C.2 Second Moment: Size of the Cycle Update

We now bound the squared cycle displacement by controlling the second moments of the same trajectory quantities. Define

$$u_{k,p} \triangleq \mathbb{E}_k \|\boldsymbol{w}^{(k,p-1)} - \boldsymbol{w}^{(k,0)}\|^2, \tag{61}$$

$$V_{k,p} \triangleq \mathbb{E}_k \|\nabla F(\boldsymbol{w}^{(k,p-1)})\|^2. \tag{62}$$

Applying $\| \sum_{i=1}^4 \boldsymbol{x}_i \|^2 \le 4 \sum_{i=1}^4 \|\boldsymbol{x}_i\|^2$ to (27) gives

$$\begin{aligned} \mathbb{E}_k \|\boldsymbol{\delta}_k\|^2 \le{}& 4\eta^2\tau^2 \mathbb{E}_k \|\boldsymbol{h}_\Sigma^{(k)}\|^2 + 4\eta^2\tau^2 \mathbb{E}_k \|\boldsymbol{\epsilon}_\Sigma^{(k)}\|^2 \\ &+ 4\eta^2 \mathbb{E}_k \|\boldsymbol{\psi}_\Sigma^{(k)}\|^2 + 4\eta^2 \mathbb{E}_k \|\boldsymbol{\nu}_\Sigma^{(k)}\|^2. \end{aligned} \tag{63}$$

Cauchy–Schwarz, Lemmas B.4 and B.3, and martingale orthogonality give the four estimates

$$\mathbb{E}_k \|\boldsymbol{h}_\Sigma^{(k)}\|^2 \le P \sum_{p=1}^{P} V_{k,p},$$

$$\mathbb{E}_k \|\boldsymbol{\epsilon}_\Sigma^{(k)}\|^2 \le 4L^2 P \sum_{p=1}^{P} u_{k,p},$$

$$\mathbb{E}_k \|\boldsymbol{\psi}_\Sigma^{(k)}\|^2 \le P \sum_{p=1}^{P} (6C_{\psi,2} V_{k,p} + D_{\psi,2}),$$

$$\mathbb{E}_k \|\boldsymbol{\nu}_\Sigma^{(k)}\|^2 \le P V_\nu.$$

With

$$C_{\delta,0} \triangleq 4\eta^2 P^2 D_{\psi,2} + 4\eta^2 P V_\nu,$$

substituting the four bounds above in (63) gives

$$\mathbb{E}_k\|\boldsymbol{\delta}_k\|^2 \le 4\eta^2 P(\tau^2 + 6C_{\psi,2}) \sum_{p=1}^{P} V_{k,p} + 16\eta^2\tau^2 L^2 P \sum_{p=1}^{P} u_{k,p} + C_{\delta,0}.$$

Let $\mathsf{U}_k \triangleq \sum_{p=1}^{P} u_{k,p}$. Smoothness yields

$$V_{k,p} \le 2G_k^2 + 2L^2 u_{k,p},$$
$$\sum_{p=1}^{P} V_{k,p} \le 2PG_k^2 + 2L^2 \mathsf{U}_k,$$

and hence

$$\mathbb{E}_k\|\boldsymbol{\delta}_k\|^2 \le 8\eta^2 P^2(\tau^2 + 6C_{\psi,2})G_k^2 + 8\eta^2 PL^2(3\tau^2 + 6C_{\psi,2})\mathsf{U}_k + C_{\delta,0}. \tag{64}$$

The partial-cycle decomposition (41) also bounds U_k. For $p \ge 2$, Cauchy–Schwarz for the gradient and drift sums, Assumption 3, Lemma B.3, and martingale orthogonality imply

$$\begin{aligned} u_{k,p} &\le 4\eta^2(\tau^2 + 6C_{\psi,2})(p-1)\sum_{s=1}^{p-1} V_{k,s} \\ &\quad + 4\eta^2(p-1)^2(\tau^2\alpha^2 + D_{\psi,2}) + 4\eta^2(p-1)V_\nu, \qquad (65) \\ &\le 8\eta^2(\tau^2 + 6C_{\psi,2})(p-1)^2 G_k^2 \\ &\quad + 8\eta^2 L^2(\tau^2 + 6C_{\psi,2})(p-1)\sum_{s=1}^{p-1} u_{k,s} \\ &\quad + 4\eta^2(p-1)^2(\tau^2\alpha^2 + D_{\psi,2}) + 4\eta^2(p-1)V_\nu. \qquad (66) \end{aligned}$$

For compactness, set

$$\alpha_{u,1} \triangleq 8\eta^2(\tau^2 + 6C_{\psi,2}), \tag{67}$$
$$\alpha_{u,0} \triangleq 4\eta^2(\tau^2\alpha^2 + D_{\psi,2}), \tag{68}$$
$$\beta_u \triangleq 8\eta^2 L^2(\tau^2 + 6C_{\psi,2}), \tag{69}$$
$$\gamma_u \triangleq 4\eta^2 V_\nu. \tag{70}$$

Then

$$u_{k,p} \le \alpha_{u,1}(p-1)^2 G_k^2 + \alpha_{u,0}(p-1)^2 + \beta_u(p-1)\sum_{s=1}^{p-1} u_{k,s} + \gamma_u(p-1). \tag{71}$$

For $P \ge 2$, let $\kappa_P \triangleq (P-2)(2P-3)/6$, and suppose

$$\beta_u\kappa_P \le \frac{1}{2}. \tag{72}$$

Define

$$C_U \triangleq 16\eta^2(\tau^2 + 6C_{\psi,2}), \tag{73}$$
$$D_U \triangleq 8\eta^2(\tau^2\alpha^2 + D_{\psi,2}) + 2\gamma_u. \tag{74}$$

We bound the squared displacement by a quadratic function of the round index. The inductive step uses

$$(p-1)\sum_{s=1}^{p-1}(s-1)^2 = (p-1)^2\frac{(p-2)(2p-3)}{6} \le \kappa_P(p-1)^2,$$

so inserting the induction hypothesis in (71), and using $\gamma_u(p-1) \le \gamma_u(p-1)^2$, gives

$$\begin{aligned} u_{k,p} &\le \big(\alpha_{u,1} + \beta_u\kappa_P C_U\big)(p-1)^2 G_k^2 \\ &\quad + \big(\alpha_{u,0} + \gamma_u + \beta_u\kappa_P D_U\big)(p-1)^2. \end{aligned}$$

Condition (72) and the definitions in (73)–(74) make the two coefficients at most C_U and D_U, respectively. Since $u_{k,1} = 0$, induction proves

$$u_{k,p} \le C_U(p-1)^2 G_k^2 + D_U(p-1)^2, \tag{75}$$

$$\mathsf{U}_k \le \frac{(P-1)P(2P-1)}{6}\big(C_U G_k^2 + D_U\big). \tag{76}$$

Finally, define

$$\mathcal{P}(\eta) \triangleq 8\eta^2 P^2(\tau^2 + 6C_{\psi,2}) + \frac{4\eta^2 P^2 L^2(3\tau^2 + 6C_{\psi,2})(P-1)(2P-1)}{3} C_U, \tag{77}$$

$$\mathcal{Q}(\eta) \triangleq \frac{4\eta^2 P^2 L^2(3\tau^2 + 6C_{\psi,2})(P-1)(2P-1)}{3} D_U + C_{\delta,0}. \tag{78}$$

Substituting (76) in (64) completes the second-moment argument:

$$\mathbb{E}_k\|\boldsymbol{\delta}_k\|^2 \le \mathcal{P}(\eta)G_k^2 + \mathcal{Q}(\eta). \tag{79}$$

D Proof of the PL Convergence Results

Under Assumptions 1–5, smoothness and the cycle-moment bounds yield descent over each cycle. The PL inequality converts this descent into an objective-gap recursion. We then verify a common step-size range and bound the residual terms to obtain Theorem 3.1.

D.1 Cycle Recursion and Fixed-Step Guarantee

Let

$$\Phi_k^{\mathrm{PL}} \triangleq \mathbb{E}\big[F(\boldsymbol{w}^{(k,0)}) - F^*\big], \tag{80}$$

$$D_{\mathrm{PL}}(\eta) \triangleq A_{\mathrm{align}}(\eta) - \frac{L}{2}\mathcal{P}(\eta).$$

Smoothness of F, followed by the cycle bounds (60) and (79), gives

$$\mathbb{E}_k[F(\boldsymbol{w}^{(k+1,0)})] \le F(\boldsymbol{w}^{(k,0)}) - D_{\mathrm{PL}}(\eta)G_k^2 + B_{\mathrm{align}}(\eta)G_k + \frac{L}{2}\mathcal{Q}(\eta). \tag{81}$$

When $D_{\mathrm{PL}}(\eta) > 0$, Young's inequality and the definition

$$B_{\mathrm{align}}(\eta)G_k \le \frac{D_{\mathrm{PL}}(\eta)}{2}G_k^2 + \frac{B_{\mathrm{align}}(\eta)^2}{2D_{\mathrm{PL}}(\eta)},$$

$$C_{\mathrm{PL}}(\eta) \triangleq \frac{B_{\mathrm{align}}(\eta)^2}{2D_{\mathrm{PL}}(\eta)} + \frac{L}{2}\mathcal{Q}(\eta)$$

reduce (81) to

$$\mathbb{E}_k[F(\boldsymbol{w}^{(k+1,0)})] \le F(\boldsymbol{w}^{(k,0)}) - \frac{D_{\mathrm{PL}}(\eta)}{2}G_k^2 + C_{\mathrm{PL}}(\eta).$$

Assumption 5 gives $G_k^2 \ge 2\mu(F(\boldsymbol{w}^{(k,0)}) - F^*)$, yielding the one-cycle recursion and its unrolling:

$$\Phi_{k+1}^{\mathrm{PL}} \le \big(1 - \mu D_{\mathrm{PL}}(\eta)\big)\Phi_k^{\mathrm{PL}} + C_{\mathrm{PL}}(\eta), \tag{82}$$

$$\Phi_K^{\mathrm{PL}} \le \big(1 - \mu D_{\mathrm{PL}}(\eta)\big)^K \Phi_0^{\mathrm{PL}} + \frac{C_{\mathrm{PL}}(\eta)}{\mu D_{\mathrm{PL}}(\eta)}, \qquad 0 < \mu D_{\mathrm{PL}}(\eta) \le 1. \tag{83}$$

D.2 A Common Admissible Step Size

We verify that one step-size cap satisfies the local-drift and cycle-drift conditions and ensures a positive descent coefficient. Assume $P \ge 2$ and set

$$\bar{\eta}_{\mathrm{PL}} \triangleq \frac{1}{64LP\tau}.$$

For $0 < \eta \le \bar{\eta}_{\mathrm{PL}}$, let $x \triangleq L\eta\tau P$ and $z \triangleq 6C_{\psi,2}/\tau^2$. Then

$$x \le \frac{1}{64}, \qquad L\eta(\tau-1) \le \frac{x}{P} \le \frac{1}{128}, \qquad 0 \le z \le 4\left(\frac{x}{P}\right)^2 \le \frac{1}{4096}.$$

These inequalities imply both local-drift conditions in Lemmas B.2 and B.3. They also give

$$\beta_u \kappa_P \le \frac{8}{3}x^2(1+z) < \frac{1}{2},$$

which verifies (72).

Next, $C_{\mathrm{dr},1} = \tau(1 + L\eta(\tau-1))$, and hence

$$qP = \eta L C_{\mathrm{dr},1} P \le \frac{1}{64}\left(1 + \frac{1}{128}\right) < \frac{1}{32}. \tag{84}$$

Using $r_P(q) \le P(P-1)qe^{qP}/2$ and $e^y \le \frac{1}{1-y}$ for $0 \le y < 1$, we obtain

$$\frac{r_P(q)}{P} < \frac{1}{120}, \qquad \frac{3L\alpha_{\mathrm{d}}}{P} < \frac{1}{40}, \qquad \frac{C_\psi \alpha_{\mathrm{g}}}{\tau P} = L\eta(\tau-1)\left(1 + \frac{r_P(q)}{P}\right) < \frac{1}{120}.$$

Substituting the above bounds in (58) gives

$$A_{\mathrm{align}}(\eta) > \eta\tau P\left(1 - \frac{1}{40} - \frac{1}{120}\right) = \frac{29}{30}\eta\tau P.$$

It remains to control the smoothness penalty. Since $\tau^2 + 6C_{\psi,2} = \tau^2(1+z)$, $3\tau^2 + 6C_{\psi,2} = \tau^2(3+z)$, $C_U = 16\eta^2\tau^2(1+z)$, and $(P-1)(2P-1) \le 2P^2$, the two components of $L\mathcal{P}(\eta)/(2\eta\tau P)$ satisfy

$$4x(1+z) < \frac{1}{15}, \qquad \frac{64}{3}x^3(3+z)(1+z) < \frac{1}{1000}.$$

Thus $L\mathcal{P}(\eta)/2 < \eta\tau P/8$, and therefore

$$D_{\mathrm{PL}}(\eta) \ge \left(\frac{29}{30} - \frac{1}{8}\right)\eta\tau P = \frac{101}{120}\eta\tau P \ge \frac{3}{4}\eta\tau P. \tag{85}$$

The contraction condition also holds throughout this range. The descent lemma gives

$$F(\boldsymbol{w}) - F^* \ge F(\boldsymbol{w}) - F\left(\boldsymbol{w} - \frac{1}{L}\nabla F(\boldsymbol{w})\right) \ge \frac{1}{2L}\|\nabla F(\boldsymbol{w})\|^2.$$

Together with Assumption 5, this implies $\mu \le L$ for every nonconstant objective. Consequently, throughout the same step-size range,

$$D_{\mathrm{PL}}(\eta) \le A_{\mathrm{align}}(\eta) \le \eta\tau P, \qquad 0 < \mu D_{\mathrm{PL}}(\eta) \le \frac{1}{64} < 1.$$

With $T = KP$ communication rounds, (83) and (85) yield

$$\Phi_K^{\mathrm{PL}} \le \exp\left(-\frac{3}{4}\mu\eta\tau T\right)\Phi_0^{\mathrm{PL}} + \frac{4C_{\mathrm{PL}}(\eta)}{3\mu\eta\tau P}. \tag{86}$$

D.3 Horizon-Dependent Rate and Consequences

To express the objective-gap residual in terms of the quantities used in the main paper, recall the effective variance and cyclic-window drift scale:

$$\Sigma_{\mathrm{eff}}^2 \triangleq \frac{V_\nu}{\tau^2} = \omega_{B,S}\gamma^2 + \frac{\sigma^2}{\tau S},$$
$$\mathcal{R}_{\mathrm{CW}} \triangleq P(\alpha + \Sigma_{\mathrm{eff}}) + \frac{\tau-1}{\tau}H.$$

The next lemma separates the variance contribution from the higher-order drift contribution.

Lemma D.1 (PL residual bound)**.** *There exist universal constants $C_{\mathrm{PL},1}, C_{\mathrm{PL},2} > 0$ such that, for every $0 < \eta \le \bar{\eta}_{\mathrm{PL}}$,*

$$\frac{C_{\mathrm{PL}}(\eta)}{\mu D_{\mathrm{PL}}(\eta)} \le C_{\mathrm{PL},1}\frac{L\eta V_\nu}{\mu\tau} + C_{\mathrm{PL},2}\frac{L^2\eta^2\tau^2}{\mu}\mathcal{R}_{\mathrm{CW}}^2. \tag{87}$$

Proof. Write

$$J_{\mathrm{CW}} \triangleq \tau\mathcal{R}_{\mathrm{CW}} = P(\tau\alpha + \sqrt{V_\nu}) + (\tau - 1)H.$$

Because $C_\psi = L\eta\tau(\tau - 1)$ and $L\eta\tau P \le 1/64$, $C_{\mathrm{dr},2} \le J_{\mathrm{CW}}/P$. Equations (52)–(55) and (84) consequently give, for a universal constant $c > 0$,

$$\begin{aligned} \alpha_{\mathrm{d}} &\le c\eta\tau P^2, & \beta_{\mathrm{d}} &\le c\eta P J_{\mathrm{CW}}, \\ \alpha_{\mathrm{g}} &\le cP, & \beta_{\mathrm{g}} &\le c\eta L P J_{\mathrm{CW}}. \end{aligned}$$

The first row bounds the displacement coefficients, and the second bounds the gradient coefficients.

Using these bounds in (59), followed by (85), gives

$$\frac{B_{\mathrm{align}}(\eta)}{D_{\mathrm{PL}}(\eta)} \le cL\eta J_{\mathrm{CW}}, \qquad \frac{B_{\mathrm{align}}(\eta)^2}{2\mu D_{\mathrm{PL}}(\eta)^2} \le c\frac{L^2\eta^2 J_{\mathrm{CW}}^2}{\mu}.$$

For the second-moment remainder, the definitions of $C_{\psi,2}$ and $D_{\psi,2}$ give

$$D_{\psi,2} \le cL^2\eta^2\tau^2(\tau - 1)^2H^2.$$

Moreover,

$$D_U = 8\eta^2\big(\tau^2\alpha^2 + D_{\psi,2} + V_\nu\big) \le c\eta^2\left(\frac{J_{\mathrm{CW}}}{P}\right)^2,$$

because $\tau\alpha + \sqrt{V_\nu} \le J_{\mathrm{CW}}/P$ and $L\eta\tau(\tau - 1)H \le J_{\mathrm{CW}}/P$. Substituting these estimates in the definition of $C_{\delta,0}$ and in (78) yields

$$\begin{aligned} \mathcal{Q}(\eta) &\le c\eta^2 P V_\nu + cL^2\eta^4\tau^2P^2J_{\mathrm{CW}}^2, \\ \frac{L\mathcal{Q}(\eta)}{2\mu D_{\mathrm{PL}}(\eta)} &\le c\frac{L\eta V_\nu}{\mu\tau} + c\frac{L^2\eta^2 J_{\mathrm{CW}}^2}{\mu}. \end{aligned}$$

In the second inequality, one remaining factor $L\eta\tau P$ is absorbed using the common step-size restriction. Combining these bounds with $J_{\mathrm{CW}} = \tau\mathcal{R}_{\mathrm{CW}}$ proves (87). □

For a prescribed horizon $T = KP \ge 2$, use the constant learning rate $\eta = \eta_T^{\mathrm{PL}}$, where

$$\eta_T^{\mathrm{PL}} = \min\left\{\bar{\eta}_{\mathrm{PL}}, \frac{4\log T}{\mu\tau T}\right\}. \tag{88}$$

When $T \ge 256(L/\mu)P\log T$, the horizon-dependent branch is active. Using this branch and the identity $V_\nu/\tau^2 = \Sigma_{\mathrm{eff}}^2$, we obtain the following rate. First, the contraction term in (86) is at most $T^{-3}\Phi_0^{\mathrm{PL}}$, which also satisfies the stated $T^{-2}\Phi_0^{\mathrm{PL}}$ bound. Combining this contraction estimate with (83) and the residual bound (87) yields universal constants $C_1, C_2 > 0$ such that

$$\mathbb{E}[F(\boldsymbol{w}^{(K,0)})] - F^* \le \frac{\Phi_0^{\mathrm{PL}}}{T^2} + C_1\frac{L\log T}{\mu^2 T}\Sigma_{\mathrm{eff}}^2 + C_2\frac{L^2(\log T)^2}{\mu^3T^2}\mathcal{R}_{\mathrm{CW}}^2. \tag{89}$$

This proves Theorem 3.1. The joint round innovation contributes the leading $\widetilde{\mathcal{O}}(T^{-1})$ term, while the cycle-drift and local-drift contributions appear at order $\widetilde{\mathcal{O}}(T^{-2})$.

Suppressing the logarithmic factors in (89) gives the following simplified PL rate.

Corollary D.1 (Simplified PL rates)**.** *For fixed system parameters,*

$$\mathbb{E}[F(\boldsymbol{w}^{(K,0)})] - F^* = \widetilde{\mathcal{O}}\left(\frac{\Sigma_{\mathrm{eff}}^2}{T} + \frac{\Phi_0^{\mathrm{PL}} + \mathcal{R}_{\mathrm{CW}}^2}{T^2}\right). \tag{90}$$

Deterministic full-window case. If $S = B$ and $\sigma = 0$, then $\omega_{B,S} = V_\nu = \Sigma_{\text{eff}} = 0$, and (89) reduces, for fixed system parameters, to

$$\mathbb{E}[F(\boldsymbol{w}^{(K,0)})] - F^* = \widetilde{\mathcal{O}}(T^{-2}). \tag{91}$$

When $\tau = 1$, $\mathcal{R}_{\text{CW}} = P\alpha$; hence the deterministic full-window bound contains no intra-window heterogeneity term γ, matching the cyclic Local GD specialization.

Effect of the sampling ratio. For $B > 1$, substituting $S = \varrho B$ into the finite-population factor gives $\omega_{B,S} = (1-\varrho)/[\varrho(B-1)]$. Inserting this expression into the effective variance gives the leading partial-sampling and stochastic contribution in (89):

$$\widetilde{\mathcal{O}}\left(\frac{1}{T}\left[\frac{1-\varrho}{\varrho(B-1)}\gamma^2 + \frac{\sigma^2}{\tau\varrho B}\right]\right),$$

with the first summand omitted when $B = 1$. The two summands quantify client-sampling variance and mini-batch noise, respectively, as components of the same joint conditional variance.

E Proof of the Non-Convex Convergence Result

We sum the per-round descent bounds over each cycle to bound stationarity without the PL condition. Cycle-wise analysis controls window-induced gradient biases, and trajectory bounds relate the resulting estimates to gradients at the actual server models. Under Assumptions 1–4, define the cycle-wise stationarity sum

$$\mathcal{V}_k \triangleq \sum_{p=1}^{P} \mathbb{E}\|\nabla F(\boldsymbol{w}^{(k,p-1)})\|^2. \tag{92}$$

For $T = KP$, the quantity to be bounded is

$$\frac{1}{T}\sum_{k=0}^{K-1}\sum_{p=1}^{P}\mathbb{E}\|\nabla F(\boldsymbol{w}^{(k,p-1)})\|^2.$$

E.1 Cycle-Summed Descent and Window-Bias Compensation

Smoothness and (20) give

$$F(\boldsymbol{w}^{(k,p)}) \le F(\boldsymbol{w}^{(k,p-1)}) - \eta\langle\nabla F(\boldsymbol{w}^{(k,p-1)}), \boldsymbol{g}^{(k,p)}\rangle + \frac{L\eta^2}{2}\|\boldsymbol{g}^{(k,p)}\|^2.$$

Taking $\mathbb{E}_{k,p}$, using the zero mean of $\boldsymbol{\nu}^{(k,p)}$, and abbreviating $\nabla F = \nabla F(\boldsymbol{w}^{(k,p-1)})$ inside this subsection yields

$$\begin{aligned}\mathbb{E}_{k,p}[F(\boldsymbol{w}^{(k,p)})] \le F(\boldsymbol{w}^{(k,p-1)}) &- \eta\tau\|\nabla F\|^2 - \eta\tau\langle\nabla F, \boldsymbol{\epsilon}^{(k,p)}\rangle \\ &- \eta\langle\nabla F, \mathbb{E}_{k,p}[\boldsymbol{\psi}^{(k,p)}]\rangle + \frac{L\eta^2}{2}\mathbb{E}_{k,p}\|\boldsymbol{g}^{(k,p)}\|^2.\end{aligned} \tag{93}$$

Lemma B.2 and Young's inequality bound the local-drift term in (93):

$$\begin{aligned}-\eta\langle\nabla F, \mathbb{E}_{k,p}[\boldsymbol{\psi}^{(k,p)}]\rangle &\le \eta\|\nabla F\|\,\mathbb{E}_{k,p}\|\boldsymbol{\psi}^{(k,p)}\| \\ &\le \eta C_\psi\left(\|\nabla F\|^2 + H\|\nabla F\|\right) \\ &\le \frac{3\eta C_\psi}{2}\|\nabla F\|^2 + \frac{\eta C_\psi H^2}{2}.\end{aligned}$$

For the smoothness remainder, the four-term decomposition (26) first yields

$$\|\boldsymbol{g}^{(k,p)}\|^2 \le 4\tau^2\|\nabla F\|^2 + 4\tau^2\|\boldsymbol{\epsilon}^{(k,p)}\|^2 + 4\|\boldsymbol{\psi}^{(k,p)}\|^2 + 4\|\boldsymbol{\nu}^{(k,p)}\|^2.$$

Taking $\mathbb{E}_{k,p}$, applying Assumption 3, and then invoking Lemmas B.3 and B.1 gives

$$\begin{aligned}\Gamma_{g,1} \triangleq 4\tau^2 + 24C_{\psi,2}, \qquad \Gamma_{g,0} \triangleq 4\tau^2\alpha^2 + 4D_{\psi,2} + 4V_\nu, \\ \frac{L\eta^2}{2}\mathbb{E}_{k,p}\|\boldsymbol{g}^{(k,p)}\|^2 \le \frac{L\eta^2}{2}\left(\Gamma_{g,1}\|\nabla F\|^2 + \Gamma_{g,0}\right).\end{aligned}$$

Collect these two remainders by defining

$$A_{\text{step}} \triangleq \eta\tau - \frac{3\eta C_\psi}{2} - \frac{L\eta^2\Gamma_{g,1}}{2}, \qquad C_{\text{step}} \triangleq \frac{\eta C_\psi H^2}{2} + \frac{L\eta^2\Gamma_{g,0}}{2}.$$

Equation (93) becomes

$$\begin{aligned}\mathbb{E}_{k,p}[F(\boldsymbol{w}^{(k,p)})] \le F(\boldsymbol{w}^{(k,p-1)}) - A_{\text{step}}\|\nabla F(\boldsymbol{w}^{(k,p-1)})\|^2 \\ - \eta\tau\langle \nabla F(\boldsymbol{w}^{(k,p-1)}), \boldsymbol{\epsilon}^{(k,p)}\rangle + C_{\text{step}}.\end{aligned} \tag{94}$$

Define the cycle-level gradient–window coupling

$$\mathcal{C}_{\epsilon,k} \triangleq \sum_{p=1}^{P} \mathbb{E}\Big[\Big\langle \nabla F(\boldsymbol{w}^{(k,p-1)}), \boldsymbol{\epsilon}^{(k,p)}\Big\rangle\Big]. \tag{95}$$

Summing (94) across the cycle gives

$$\mathbb{E}[F(\boldsymbol{w}^{(k+1,0)})] \le \mathbb{E}[F(\boldsymbol{w}^{(k,0)})] - A_{\text{step}}\mathcal{V}_k - \eta\tau\mathcal{C}_{\epsilon,k} + PC_{\text{step}}. \tag{96}$$

The window-coupling term can have either sign. To bound it using fixed-anchor cancellation, write

$$\begin{aligned}\boldsymbol{h}_p \triangleq \nabla F(\boldsymbol{w}^{(k,p-1)}), \qquad \boldsymbol{h}_0 \triangleq \nabla F(\boldsymbol{w}^{(k,0)}),\\ \boldsymbol{\epsilon}_p(\boldsymbol{w}) \triangleq \boldsymbol{\epsilon}^{(k,p)}(\boldsymbol{w}).\end{aligned}$$

Adding and subtracting the anchor quantities, and then using $\sum_p \boldsymbol{\epsilon}_p(\boldsymbol{w}^{(k,0)}) = \mathbf{0}$, gives

$$\begin{aligned}\mathcal{C}_{\epsilon,k} = \sum_{p=1}^{P} \mathbb{E}\Big[\Big\langle \boldsymbol{h}_p - \boldsymbol{h}_0, \boldsymbol{\epsilon}_p(\boldsymbol{w}^{(k,p-1)})\Big\rangle \\ + \Big\langle \boldsymbol{h}_0, \boldsymbol{\epsilon}_p(\boldsymbol{w}^{(k,p-1)}) - \boldsymbol{\epsilon}_p(\boldsymbol{w}^{(k,0)})\Big\rangle\Big].\end{aligned}$$

Smoothness, $\|\boldsymbol{\epsilon}_p(\boldsymbol{w}^{(k,p-1)})\| \le \alpha$, and the $2L$-Lipschitz property used in Lemma B.4 therefore imply

$$|\mathcal{C}_{\epsilon,k}| \le \mathbb{E}\left[(2LG_k + L\alpha)\sum_{p=1}^{P}\|\boldsymbol{w}^{(k,p-1)} - \boldsymbol{w}^{(k,0)}\|\right].$$

Because G_k is $\mathcal{H}_k$-measurable, the tower property and the conditional trajectory bound (56) yield

$$|\mathcal{C}_{\epsilon,k}| \le 2L\alpha_{\text{d}}\mathbb{E}[G_k^2] + (2L\beta_{\text{d}} + L\alpha\alpha_{\text{d}})\mathbb{E}[G_k] + L\alpha\beta_{\text{d}}. \tag{97}$$

For later use, define

$$\chi_2 \triangleq 2L\alpha_{\text{d}}, \qquad \chi_1 \triangleq 2L\beta_{\text{d}} + L\alpha\alpha_{\text{d}}, \qquad \chi_0 \triangleq L\alpha\beta_{\text{d}}.$$

E.2 From Cycle Anchors to Inner-Round Stationarity

The coupling bound depends on the anchor gradient norm G_k, while $\mathcal{V}_k$ measures stationarity along the actual server trajectory. We now reverse the gradient comparison used in the PL analysis: the squared-displacement bound in Appendix C.2 controls $\mathbb{E}[G_k^2]$ in terms of $\mathcal{V}_k$. Define

$$\Gamma_U \triangleq \frac{(P-1)P(2P-1)}{6}C_U, \qquad \Lambda_U \triangleq \frac{(P-1)P(2P-1)}{6}D_U.$$

Summing (75) and taking expectation gives the first bound below. To relate the anchor gradient to the inner-round gradients, smoothness gives

$$G_k^2 \le 2\|\nabla F(\boldsymbol{w}^{(k,p-1)})\|^2 + 2L^2\|\boldsymbol{w}^{(k,p-1)} - \boldsymbol{w}^{(k,0)}\|^2.$$

Summing this pointwise inequality over p, taking expectation, and using the first bound gives the second:

$$\begin{aligned}\sum_{p=1}^{P}\mathbb{E}\|\boldsymbol{w}^{(k,p-1)} - \boldsymbol{w}^{(k,0)}\|^2 \le \Gamma_U\mathbb{E}[G_k^2] + \Lambda_U,\\ \left(1 - \frac{2L^2\Gamma_U}{P}\right)\mathbb{E}[G_k^2] \le \frac{2}{P}\mathcal{V}_k + \frac{2L^2\Lambda_U}{P}.\end{aligned}$$

If

$$\frac{2L^2\Gamma_U}{P} \le \frac{1}{2}, \tag{98}$$

then, with the coefficients

$$\lambda_1 \triangleq \frac{4}{P}, \qquad \lambda_0 \triangleq \frac{4L^2\Lambda_U}{P},$$

we obtain

$$\begin{aligned} \mathbb{E}[G_k^2] &\le \lambda_1 \mathcal{V}_k + \lambda_0, \\ \mathbb{E}[G_k] &\le \sqrt{\lambda_1 \mathcal{V}_k + \lambda_0}. \end{aligned} \tag{99}$$

Young's inequality gives

$$\chi_1 \mathbb{E}[G_k] \le \frac{P}{32}\mathbb{E}[G_k^2] + \frac{8\chi_1^2}{P}.$$

Using this bound, define

$$\widehat{A}(\eta) \triangleq A_{\text{step}} - \eta\tau\left(\chi_2 + \frac{P}{32}\right)\lambda_1, \tag{100}$$

$$\widehat{C}(\eta) \triangleq PC_{\text{step}} + \eta\tau\left[\left(\chi_2 + \frac{P}{32}\right)\lambda_0 + \frac{8\chi_1^2}{P} + \chi_0\right]. \tag{101}$$

Substituting (97) into (96), and then using (99), produces the cycle descent inequality

$$\mathbb{E}[F(\boldsymbol{w}^{(k+1,0)})] \le \mathbb{E}[F(\boldsymbol{w}^{(k,0)})] - \widehat{A}(\eta)\mathcal{V}_k + \widehat{C}(\eta). \tag{102}$$

Let

$$\Phi_0^{\text{nc}} \triangleq \mathbb{E}[F(\boldsymbol{w}^{(0,0)})] - F^*.$$

If $\widehat{A}(\eta) > 0$, telescoping (102) over k yields

$$\begin{aligned} \widehat{A}(\eta)\sum_{k=0}^{K-1}\mathcal{V}_k &\le \Phi_0^{\text{nc}} + K\widehat{C}(\eta), \\ \frac{1}{T}\sum_{k=0}^{K-1}\sum_{p=1}^{P}\mathbb{E}\|\nabla F(\boldsymbol{w}^{(k,p-1)})\|^2 &\le \frac{\Phi_0^{\text{nc}}}{T\widehat{A}(\eta)} + \frac{\widehat{C}(\eta)}{P\widehat{A}(\eta)}. \end{aligned} \tag{103}$$

E.3 Admissible Step Size and Rate Extraction

We use the same step-size cap as in the PL analysis, namely $\bar{\eta}_{\text{nc}} \triangleq 1/(64LP\tau)$. For $0 < \eta \le \bar{\eta}_{\text{nc}}$, Appendix D.2 already verifies the local- and squared-drift restrictions. In addition, (73), the definitions of Γ_U and Λ_U, and $(P-1)P(2P-1)/6 \le P^3/3$ give

$$\frac{2L^2\Gamma_U}{P} \le \frac{32}{3}(L\eta\tau P)^2\left(1 + \frac{6C_{\psi,2}}{\tau^2}\right) < \frac{1}{2},$$

so (98) holds. The same range gives

$$\begin{aligned} \frac{C_\psi}{\tau} &\le \frac{1}{128}, \\ \frac{L\eta\Gamma_{g,1}}{2\tau} &= 2L\eta\tau\left(1 + \frac{6C_{\psi,2}}{\tau^2}\right) \le \frac{1}{64}\left(1 + \frac{1}{4096}\right) < \frac{1}{63}, \\ A_{\text{step}} &\ge \frac{31}{32}\eta\tau. \end{aligned}$$

Finally, $\chi_2 = 2r_P(q)$, $r_P(q)/P < 1/120$, and $\lambda_1 = 4/P$, so

$$\left(\chi_2 + \frac{P}{32}\right)\lambda_1 = \frac{8r_P(q)}{P} + \frac{1}{8} < \frac{1}{15} + \frac{1}{8}.$$

Substituting the preceding bounds into (100), and then using (103), gives

$$\widehat{A}(\eta) \geq \left(\frac{31}{32} - \frac{1}{15} - \frac{1}{8}\right)\eta\tau = \frac{373}{480}\eta\tau \geq \frac{3}{4}\eta\tau,$$

$$\frac{1}{T}\sum_{k=0}^{K-1}\sum_{p=1}^{P}\mathbb{E}\|\nabla F(\boldsymbol{w}^{(k,p-1)})\|^2 \leq \frac{4\Phi_0^{\mathrm{nc}}}{3\eta\tau T} + \frac{4\widehat{C}(\eta)}{3P\eta\tau}. \tag{104}$$

To express the residual in terms of the system parameters, recall

$$\Sigma_{\mathrm{eff}}^2 = \frac{V_\nu}{\tau^2} = \omega_{B,S}\gamma^2 + \frac{\sigma^2}{\tau S},$$
$$J_{\mathrm{CW}} \triangleq \tau\mathcal{R}_{\mathrm{CW}} = P(\tau\alpha + \sqrt{V_\nu}) + (\tau-1)H.$$

Lemma E.1 (Normalized non-convex residual)**.** *There exist universal constants $C_{\mathrm{nc}}, C'_{\mathrm{nc}} > 0$ such that, for every $0 < \eta \leq \frac{1}{64LP\tau}$,*

$$\frac{\widehat{C}(\eta)}{P\eta\tau} \leq C_{\mathrm{nc}}L\eta\left[\tau H^2 + \frac{V_\nu}{\tau} + \alpha J_{\mathrm{CW}}\right] + C'_{\mathrm{nc}}L^2\eta^2 J_{\mathrm{CW}}^2. \tag{105}$$

Proof. The step-size condition implies $C_\psi H \leq (\tau-1)H/(64P)$ and hence $C_{\mathrm{dr},2} \leq J_{\mathrm{CW}}/P$. Equations (52)–(53) and (84) then give, for a universal constant $c > 0$,

$$\alpha_{\mathrm{d}} \leq c\eta\tau P^2, \qquad \beta_{\mathrm{d}} \leq c\eta P J_{\mathrm{CW}},$$
$$\chi_2 \leq c\eta L\tau P^2, \qquad \chi_1 \leq c\eta LP J_{\mathrm{CW}}, \qquad \chi_0 \leq c\eta LP\alpha J_{\mathrm{CW}}.$$

Furthermore, the definitions of $C_{\psi,2}$, D_U, and Λ_U imply

$$D_{\psi,2} \leq cL^2\eta^2\tau^2(\tau-1)^2H^2,$$
$$D_U \leq c\eta^2\left(\frac{J_{\mathrm{CW}}}{P}\right)^2,$$
$$\lambda_0 \leq cL^2\eta^2 J_{\mathrm{CW}}^2,$$
$$\frac{C_{\mathrm{step}}}{\eta\tau} \leq cL\eta\left(\tau H^2 + \frac{V_\nu}{\tau}\right). \tag{106}$$

For the D_U estimate, recall the identity $D_U = 8\eta^2\big(\tau^2\alpha^2 + D_{\psi,2} + V_\nu\big)$. Combining this identity with $\tau\alpha + \sqrt{V_\nu} \leq J_{\mathrm{CW}}/P$ and $L\eta\tau(\tau-1)H \leq J_{\mathrm{CW}}/P$ gives the claimed bound. Likewise, substituting the definitions of $C_{\psi,2}$ and $D_{\psi,2}$ into C_{step} shows that the $D_{\psi,2}$ contribution has an additional factor $[L\eta(\tau-1)]^2$; the common step-size restriction absorbs this higher-order term and gives (106). Dividing (101) by $P\eta\tau$, the remaining terms satisfy

$$\frac{\chi_0}{P} \leq cL\eta\alpha J_{\mathrm{CW}}, \qquad \frac{8\chi_1^2}{P^2} + \frac{(\chi_2 + P/32)\lambda_0}{P} \leq cL^2\eta^2 J_{\mathrm{CW}}^2.$$

Combining these estimates proves (105). □

Horizon-dependent rate. For $T = KP$, use the constant learning rate $\eta = \eta_T^{\mathrm{nc}}$, where

$$\eta_T^{\mathrm{nc}} = \min\left\{\bar{\eta}_{\mathrm{nc}}, \frac{1}{L\tau\sqrt{T}}\right\}. \tag{107}$$

When $T \geq 4096P^2$, the second branch is active. Substituting (105) into (104) at this step size, with $V_\nu = \tau^2\Sigma_{\mathrm{eff}}^2$ and $J_{\mathrm{CW}} = \tau\mathcal{R}_{\mathrm{CW}}$, gives universal constants $C_3, C_4 > 0$ such that

$$\frac{1}{T}\sum_{k=0}^{K-1}\sum_{p=1}^{P}\mathbb{E}\|\nabla F(\boldsymbol{w}^{(k,p-1)})\|^2 \leq \frac{4L\Phi_0^{\mathrm{nc}}}{3\sqrt{T}} + \frac{C_3}{\sqrt{T}}\left[H^2 + \Sigma_{\mathrm{eff}}^2 + \alpha\mathcal{R}_{\mathrm{CW}}\right] + \frac{C_4}{T}\mathcal{R}_{\mathrm{CW}}^2, \tag{108}$$

where

$$\Sigma_{\mathrm{eff}}^2 = \omega_{B,S}\gamma^2 + \frac{\sigma^2}{\tau S}.$$

Equation (108) proves Theorem 3.2 and yields the following simplified stationarity rate.

Corollary E.1 (Simplified non-convex rate)**.** *For fixed system parameters,*

$$\frac{1}{T}\sum_{k=0}^{K-1}\sum_{p=1}^{P}\mathbb{E}\left\|\nabla F(\boldsymbol{w}^{(k,p-1)})\right\|^2 = \mathcal{O}\left(\frac{L\Phi_0^{\mathrm{nc}} + H^2 + \Sigma_{\mathrm{eff}}^2 + \alpha\mathcal{R}_{\mathrm{CW}}}{\sqrt{T}} + \frac{\mathcal{R}_{\mathrm{CW}}^2}{T}\right). \tag{109}$$

Thus, the averaged stationarity measure is $\mathcal{O}(T^{-1/2})$. With heterogeneity and noise levels held fixed, $\mathcal{R}_{\mathrm{CW}}$ is affine in P, so the period-dependent contributions are at most linear in P at order $T^{-1/2}$ and quadratic at order T^{-1}.

F Relation to Cyclic Client Participation

This section shows how Cycwin-FL reduces to CyCP and compares their convergence results.

F.1 Recovering CyCP from Cycwin-FL

Let $K_c \geq 2$ denote the number of equal-size, non-overlapping client groups in CyCP, with groups $\mathcal{G}_1^{\mathrm{CyCP}}, \ldots, \mathcal{G}_{K_c}^{\mathrm{CyCP}}$ visited in a fixed cyclic order. Let N_{CyCP} denote the number of clients sampled from each group. Without loss of generality, number the clients so that each group forms a consecutive block, and set

$$B = \frac{M}{K_c}, \qquad N_w = B, \qquad S = N_{\mathrm{CyCP}}.$$

Then $d = B$ and $P = M/B = K_c$. Starting from $v^{(0,1)} = 1$, the cyclic windows satisfy

$$\mathcal{A}^{(k,p)} = \{(p-1)B + 1, \ldots, pB\} = \mathcal{G}_p^{\mathrm{CyCP}}, \qquad p \in [P].$$

Thus, each client appears exactly once per cycle, and the cyclic pattern is balanced.

Sampling S clients uniformly without replacement from each window gives the same client-sampling rule as CyCP. The selected clients perform the same local SGD steps, and the server applies the same aggregation. Therefore, given the same initial model, step size, sampled clients, and mini-batches, the two algorithms generate the same sequence of global models. When $K_c = 1$, all clients are available in every round, and the algorithm reduces to FedAvg with uniform random client sampling.

F.2 Comparison of PL Convergence Results

Under the assumptions and step-size conditions of Theorem 3.1, the parameter choices above give a convergence bound for CyCP with local SGD. For $B > 1$, the effective variance becomes

$$\Sigma_{\mathrm{eff}}^2 = \frac{M/K_c - S}{S(M/K_c - 1)}\gamma^2 + \frac{\sigma^2}{\tau S}.$$

The client-sampling term is zero when $B = S = 1$. Substituting this expression and $P = K_c$ into (13) gives

$$\mathbb{E}[F(\boldsymbol{w}^{(K,0)})] - F^* \leq \frac{\Phi_0^{\mathrm{PL}}}{T^2} + \widetilde{\mathcal{O}}\left(\frac{L\Sigma_{\mathrm{eff}}^2}{\mu^2 T} + \frac{L^2}{\mu^3 T^2}\left[P(\alpha + \Sigma_{\mathrm{eff}}) + \frac{\tau - 1}{\tau}H\right]^2\right), \tag{110}$$

where $T = KP$, and $\widetilde{\mathcal{O}}$ suppresses logarithmic factors and universal constants.

For fixed system parameters, (110) gives a $\widetilde{\mathcal{O}}(T^{-1})$ convergence rate, matching the rate in Theorem 2 of Cho et al. (2023). The two terms in Σ_{eff}^2 account for random client sampling and stochastic gradients, respectively. The client-sampling term includes the factor $\omega_{B,S}$ arising from sampling without replacement, with $\omega_{B,S} = (B - S)/(S(B - 1))$ for $B > 1$ and $\omega_{1,1} = 0$, while the stochastic-gradient term decreases with τS.

When all clients in each window participate and gradients are exact, namely $S = B$ and $\sigma = 0$, the effective variance is zero. Equation (110) then gives

$$\mathbb{E}[F(\boldsymbol{w}^{(K,0)})] - F^* \leq \frac{\Phi_0^{\mathrm{PL}}}{T^2} + \widetilde{\mathcal{O}}\left(\frac{L^2}{\mu^3 T^2}\left[P\alpha + \frac{\tau - 1}{\tau}(\alpha + \gamma)\right]^2\right).$$

Hence, the rate improves to $\widetilde{\mathcal{O}}(T^{-2})$. If additionally $\tau = 1$, the drift term depends only on $P\alpha$ and contains no γ term. This recovers the convergence order and the dependence on inter-group heterogeneity in the full-group CyCP+GD result of Theorem 1 in Cho et al. (2023).

The two analyses differ in their dependence on the cycle length. Using our notation and the same total number of communication rounds T, the leading variance term in their Theorem 2 is

$$\widetilde{\mathcal{O}}\left(\frac{LP}{\mu^2 T}\left[\omega_{B,S}\gamma^2 + \frac{\sigma^2}{\tau S}\right]\right) = \widetilde{\mathcal{O}}\left(\frac{LP\Sigma_{\text{eff}}^2}{\mu^2 T}\right).$$

Our bound in (110) instead contains $\widetilde{\mathcal{O}}(L\Sigma_{\text{eff}}^2/(\mu^2 T))$.

This difference comes from jointly bounding the random errors caused by client sampling and stochastic gradients. By Lemma B.1, the combined random error in each round has zero conditional mean. Consequently, the cross terms between these errors from different rounds vanish when taking conditional expectations. The variance accumulated over a cycle is therefore bounded by the sum of the per-round variances, avoiding an additional factor P in the leading variance term. The cycle length still affects the T^{-2} drift term and the step-size condition. The initial-error and drift terms also have different coefficients in the two bounds.

G Unbalanced Cyclic Patterns and the Cyclic-Weighted Objective

We now relax the balanced cyclic pattern condition in Assumption 4. We first derive the induced client weights and gradient bias, then carry this bias through the PL and non-convex proofs to bound optimization error relative to the equally weighted global objective. For $P \geq 2$, the analysis allows any integer $1 \leq N_w \leq B$, including unbalanced cyclic patterns.

The fundamental period is still $P = M/d$, where $d = \gcd(M, N_w)$. To describe the availability counts per cycle, write $B = q_{\text{cyc}}d + r$ with $0 \leq r < d$. A balanced cyclic pattern corresponds to $r = 0$.

G.1 Availability Counts per Cycle

Lemma G.1 (Availability counts per cycle for general cyclic-window schedules). *Let $B = q_{\text{cyc}}d + r$ with $0 \leq r < d$. Over one fundamental period, every client belongs to either q_{cyc} or $q_{\text{cyc}} + 1$ available windows. Exactly rP clients have availability count $q_{\text{cyc}} + 1$, and $(d - r)P$ clients have availability count q_{cyc}.*

Proof. The $P = M/d$ distinct window starts form one residue class modulo d and can be written as $0, d, 2d, \ldots, (P-1)d$ after a cyclic relabeling. A fixed client is available when a start position falls in the length-B circular interval ending at that client, including the client's own position. Since $B = q_{\text{cyc}}d + r$, such an interval contains either q_{cyc} or $q_{\text{cyc}} + 1$ points of the d-spaced start grid. Hence every availability count belongs to $\{q_{\text{cyc}}, q_{\text{cyc}} + 1\}$.

Let n_+ be the number of clients with availability count $q_{\text{cyc}} + 1$. Counting client–window incidences in two ways and using $M = dP$ and $B = q_{\text{cyc}}d + r$ gives

$$q_{\text{cyc}}M + n_+ = PB, \qquad n_+ = P(q_{\text{cyc}}d + r) - q_{\text{cyc}}dP = rP.$$

The remaining $M - n_+ = (d - r)P$ clients therefore have availability count q_{cyc}. The coverage counts also show that all clients are available equally often if and only if $r = 0$, or equivalently $d \mid B$, proving Lemma 3.1. □

Let c_m denote the number of available windows containing client m during one fundamental period. Normalizing these counts gives the weights $\pi_m \triangleq c_m/(PB)$, which satisfy $\sum_{m=1}^{M} \pi_m = 1$. We measure their deviation from uniform weights by the cyclic discrepancy

$$\Delta_{\text{cyc}} \triangleq \sum_{m=1}^{M} \left| \pi_m - \frac{1}{M} \right|. \tag{111}$$

The individual weight deviations are

$$\pi_m - \frac{1}{M} = \frac{d-r}{dPB}, \quad \text{if } c_m = q_{\text{cyc}} + 1,$$
$$\frac{1}{M} - \pi_m = \frac{r}{dPB}, \quad \text{if } c_m = q_{\text{cyc}}.$$

Summing these absolute deviations using the client counts in Lemma G.1 gives

$$\Delta_{\text{cyc}} = \frac{2r(d-r)}{Bd}. \tag{112}$$

In particular, $\Delta_{\text{cyc}} = 0$ if and only if the cyclic pattern is balanced.

G.2 Cyclic Bias at a Common Model

The availability counts per cycle determine the client weights in the cycle average of window gradients evaluated at a common model. For a fixed model $\boldsymbol{w}$,

$$\frac{1}{P}\sum_{p=1}^{P}\frac{1}{B}\sum_{m\in\mathcal{A}^{(k,p)}}\nabla F_m(\boldsymbol{w}) = \sum_{m=1}^{M}\pi_m \nabla F_m(\boldsymbol{w}).$$

Thus, this cycle average equals the gradient of the cyclic-weighted objective

$$F_\pi(\boldsymbol{w}) \triangleq \sum_{m=1}^{M}\pi_m F_m(\boldsymbol{w}), \qquad \nabla F_\pi(\boldsymbol{w}) = \sum_{m=1}^{M}\pi_m \nabla F_m(\boldsymbol{w}).$$

Together with (112), this establishes Lemma 4.1. For any $\mathcal{H}_k$-measurable model $\boldsymbol{w}$, conditional uniformity gives

$$\frac{1}{P}\sum_{p=1}^{P}\mathbb{E}_k\left[\frac{1}{S}\sum_{m\in\mathcal{S}^{(k,p)}}\nabla F_m(\boldsymbol{w})\right] = \frac{1}{P}\sum_{p=1}^{P}\frac{1}{B}\sum_{m\in\mathcal{A}^{(k,p)}}\nabla F_m(\boldsymbol{w}) = \nabla F_\pi(\boldsymbol{w}).$$

Define

$$\boldsymbol{b}_{\text{cyc}}(\boldsymbol{w}) \triangleq \sum_{m=1}^{M}\left(\pi_m - \frac{1}{M}\right)\nabla F_m(\boldsymbol{w}). \tag{113}$$

Because the coefficients sum to zero,

$$\boldsymbol{b}_{\text{cyc}}(\boldsymbol{w}) = \sum_{m=1}^{M}\left(\pi_m - \frac{1}{M}\right)\left(\nabla F_m(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\right).$$

Since $N_w \le B$ implies $d \le B$, every client is covered at least once. Assumption 3 therefore gives, for every client,

$$\|\nabla F_m(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\| \le \alpha + \gamma.$$

Therefore,

$$\|\boldsymbol{b}_{\text{cyc}}(\boldsymbol{w})\| \le (\alpha+\gamma)\Delta_{\text{cyc}} \triangleq B_{\text{cyc}}. \tag{114}$$

The exact cancellation identity (34) becomes

$$\sum_{p=1}^{P}\boldsymbol{\epsilon}^{(k,p)}(\boldsymbol{w}) = P\boldsymbol{b}_{\text{cyc}}(\boldsymbol{w}).$$

For the actual iterates, add and subtract the cycle anchor in every summand:

$$\boldsymbol{\epsilon}_\Sigma^{(k)} = P\boldsymbol{b}_{\text{cyc}}(\boldsymbol{w}^{(k,0)}) + \sum_{p=1}^{P}\left[\boldsymbol{\epsilon}^{(k,p)}(\boldsymbol{w}^{(k,p-1)}) - \boldsymbol{\epsilon}^{(k,p)}(\boldsymbol{w}^{(k,0)})\right].$$

Applying (114) and the same $2L$-Lipschitz bound as in (35) gives

$$\|\boldsymbol{\epsilon}_\Sigma^{(k)}\| \le PB_{\text{cyc}} + 2L\sum_{p=1}^{P}\|\boldsymbol{w}^{(k,p-1)} - \boldsymbol{w}^{(k,0)}\|. \tag{115}$$

Thus the cycle mismatch separates into cyclic bias and trajectory drift. The fixed-anchor term $P\boldsymbol{b}_{\text{cyc}}(\boldsymbol{w}^{(k,0)})$ need not vanish under an unbalanced cyclic pattern, while smoothness controls the remaining drift as before.

G.3 Modified First and Second Moments

The trajectory bounds (42)–(57) remain valid because the per-round bound $\|\boldsymbol{\epsilon}^{(k,p)}\| \le \alpha$ continues to hold. The only change in the first moment is the nonzero fixed-anchor term in (115). Specifically, (39) becomes

$$-\eta\tau \left\langle \nabla F(\boldsymbol{w}^{(k,0)}), \mathbb{E}_k[\boldsymbol{\epsilon}_\Sigma^{(k)}] \right\rangle \le \eta\tau P B_{\mathrm{cyc}} G_k + 2\eta\tau L G_k \sum_{p=1}^{P} a_{k,p}.$$

Define

$$B_{\mathrm{align}}^{\mathrm{imb}}(\eta) \triangleq B_{\mathrm{align}}(\eta) + \eta\tau P B_{\mathrm{cyc}}.$$

Then (60) is replaced by

$$\left\langle \nabla F(\boldsymbol{w}^{(k,0)}), \mathbb{E}_k[\boldsymbol{\delta}_k] \right\rangle \le -A_{\mathrm{align}}(\eta) G_k^2 + B_{\mathrm{align}}^{\mathrm{imb}}(\eta) G_k. \tag{116}$$

For the second moment, (115) gives

$$\begin{aligned}
\|\boldsymbol{\epsilon}_\Sigma^{(k)}\|^2 &\le 2P^2 B_{\mathrm{cyc}}^2 + 8L^2 \left(\sum_{p=1}^{P} \|\boldsymbol{w}^{(k,p-1)} - \boldsymbol{w}^{(k,0)}\| \right)^2 \\
&\le 2P^2 B_{\mathrm{cyc}}^2 + 8L^2 P \sum_{p=1}^{P} \|\boldsymbol{w}^{(k,p-1)} - \boldsymbol{w}^{(k,0)}\|^2.
\end{aligned}$$

Define

$$C_{\delta,0}^{\mathrm{imb}} \triangleq C_{\delta,0} + 8\eta^2\tau^2 P^2 B_{\mathrm{cyc}}^2.$$

Then the analogue of (64) is

$$\mathbb{E}_k\|\boldsymbol{\delta}_k\|^2 \le 8\eta^2 P^2(\tau^2 + 6C_{\psi,2}) G_k^2 + 8\eta^2 P L^2 (5\tau^2 + 6C_{\psi,2}) \mathsf{U}_k + C_{\delta,0}^{\mathrm{imb}}.$$

The partial-cycle squared-drift recursion (65)–(76) is unchanged because it uses only the individual bound $\|\boldsymbol{\epsilon}^{(k,p)}\| \le \alpha$. Define

$$\mathcal{P}_{\mathrm{imb}}(\eta) \triangleq 8\eta^2 P^2(\tau^2 + 6C_{\psi,2}) + \frac{4\eta^2 P^2 L^2 (5\tau^2 + 6C_{\psi,2})(P-1)(2P-1)}{3} C_U, \tag{117}$$

$$\mathcal{Q}_{\mathrm{imb}}(\eta) \triangleq \frac{4\eta^2 P^2 L^2 (5\tau^2 + 6C_{\psi,2})(P-1)(2P-1)}{3} D_U + C_{\delta,0}^{\mathrm{imb}}. \tag{118}$$

Then

$$\mathbb{E}_k\|\boldsymbol{\delta}_k\|^2 \le \mathcal{P}_{\mathrm{imb}}(\eta) G_k^2 + \mathcal{Q}_{\mathrm{imb}}(\eta). \tag{119}$$

G.4 PL Convergence to a Neighborhood Depending on the Cyclic Pattern

Assume that F satisfies Assumption 5, as in Theorem 4.1, and define

$$D_{\mathrm{PL}}^{\mathrm{imb}}(\eta) \triangleq A_{\mathrm{align}}(\eta) - \frac{L}{2}\mathcal{P}_{\mathrm{imb}}(\eta).$$

Under the same explicit step-size range $0 < \eta \le \bar{\eta}_{\mathrm{PL}} = 1/(64LP\tau)$, the following calculation verifies that $D_{\mathrm{PL}}^{\mathrm{imb}}(\eta)$ remains positive. From the step-size restriction and the definition of $C_{\psi,2}$,

$$\frac{6C_{\psi,2}}{\tau^2} = 2L^2\eta^2(\tau-1)(2\tau-1) \le 4(L\eta\tau)^2 \le \frac{1}{4096},$$

where $P \ge 2$ is used in the last inequality. In particular, $6C_{\psi,2} \le \tau^2$.

For the first term of $\mathcal{P}_{\mathrm{imb}}$, the corresponding smoothness penalty therefore satisfies

$$\frac{L}{2} 8\eta^2 P^2(\tau^2 + 6C_{\psi,2}) \le \frac{1}{8}\eta\tau P.$$

For the second term, use $C_U = 16\eta^2(\tau^2 + 6C_{\psi,2})$, $(P-1)(2P-1) \le 2P^2$, and again $6C_{\psi,2} \le \tau^2$ to obtain

$$\frac{L}{2}\frac{4\eta^2P^2L^2(5\tau^2+6C_{\psi,2})(P-1)(2P-1)}{3}C_U$$
$$\le 256(L\eta\tau P)^3\,\eta\tau P \le \frac{1}{1024}\eta\tau P.$$

Adding the two contributions gives

$$\frac{L}{2}\mathcal{P}_{\mathrm{imb}}(\eta) < \frac{3}{16}\eta\tau P.$$

Combining this estimate with the bound $A_{\mathrm{align}}(\eta) \ge 15\eta\tau P/16$ from Appendix D.2 yields

$$D_{\mathrm{PL}}^{\mathrm{imb}}(\eta) \ge \frac{3}{4}\eta\tau P > 0.$$

Define

$$C_{\mathrm{PL}}^{\mathrm{imb}}(\eta) \triangleq \frac{B_{\mathrm{align}}^{\mathrm{imb}}(\eta)^2}{2D_{\mathrm{PL}}^{\mathrm{imb}}(\eta)} + \frac{L}{2}\mathcal{Q}_{\mathrm{imb}}(\eta). \tag{120}$$

Applying Young's inequality and the PL condition, as in Appendix D.1, gives

$$\Phi_{k+1}^{\mathrm{PL}} \le \big(1-\mu D_{\mathrm{PL}}^{\mathrm{imb}}(\eta)\big)\Phi_k^{\mathrm{PL}} + C_{\mathrm{PL}}^{\mathrm{imb}}(\eta). \tag{121}$$

Moreover, $D_{\mathrm{PL}}^{\mathrm{imb}}(\eta) \le A_{\mathrm{align}}(\eta) \le \eta\tau P$. The smoothness–PL argument in Appendix D.2 gives $\mu \le L$, so the same step-size range ensures $0 < \mu D_{\mathrm{PL}}^{\mathrm{imb}}(\eta) \le L\eta\tau P \le 1/64$.

Unrolling (121) therefore gives the fixed-step bound

$$\Phi_K^{\mathrm{PL}} \le \big(1-\mu D_{\mathrm{PL}}^{\mathrm{imb}}(\eta)\big)^K\Phi_0^{\mathrm{PL}} + \frac{C_{\mathrm{PL}}^{\mathrm{imb}}(\eta)}{\mu D_{\mathrm{PL}}^{\mathrm{imb}}(\eta)}. \tag{122}$$

The residual in (122) combines stochastic noise, trajectory drift, and cyclic bias. The following lemma separates these contributions and identifies the term that remains as the step size decreases.

Lemma G.2 (PL residual bound under an unbalanced cyclic pattern)**.** *Let $J_{\mathrm{CW}} \triangleq \tau\mathcal{R}_{\mathrm{CW}}$. There exist universal constants $\widetilde{C}_{\mathrm{PL},1}, \widetilde{C}_{\mathrm{PL},2}, \widetilde{C}_{\mathrm{PL},3} > 0$ such that, for every $0 < \eta \le \frac{1}{64LP\tau}$,*

$$\frac{C_{\mathrm{PL}}^{\mathrm{imb}}(\eta)}{\mu D_{\mathrm{PL}}^{\mathrm{imb}}(\eta)} \le \widetilde{C}_{\mathrm{PL},1}\frac{L\eta V_\nu}{\mu\tau} + \widetilde{C}_{\mathrm{PL},2}\frac{L^2\eta^2J_{\mathrm{CW}}^2}{\mu} + \widetilde{C}_{\mathrm{PL},3}\frac{B_{\mathrm{cyc}}^2}{\mu}. \tag{123}$$

Proof. The coefficient bounds from Lemma D.1 remain valid because the trajectory bounds do not require a balanced cyclic pattern. Together with the preceding positivity estimate, they give, for a universal constant $c > 0$,

$$B_{\mathrm{align}}(\eta) \le cL\eta^2\tau PJ_{\mathrm{CW}}, \qquad D_{\mathrm{PL}}^{\mathrm{imb}}(\eta) \ge \frac{3}{4}\eta\tau P.$$

Using these two bounds together with $B_{\mathrm{align}}^{\mathrm{imb}}(\eta) = B_{\mathrm{align}}(\eta) + \eta\tau PB_{\mathrm{cyc}}$, we obtain

$$\frac{B_{\mathrm{align}}^{\mathrm{imb}}(\eta)^2}{2\mu[D_{\mathrm{PL}}^{\mathrm{imb}}(\eta)]^2} \le c\frac{L^2\eta^2J_{\mathrm{CW}}^2}{\mu} + c\frac{B_{\mathrm{cyc}}^2}{\mu}.$$

It remains to control the second-moment remainder. The definition of $C_{\delta,0}^{\mathrm{imb}}$ and the definition in (118), together with the same estimates for D_U used in Lemma D.1, yield

$$\mathcal{Q}_{\mathrm{imb}}(\eta) \le c\eta^2PV_\nu + cL^2\eta^4\tau^2P^2J_{\mathrm{CW}}^2 + c\eta^2\tau^2P^2B_{\mathrm{cyc}}^2.$$

Multiplying by $L/[2\mu D_{\mathrm{PL}}^{\mathrm{imb}}(\eta)]$, using the lower bound above, and applying $L\eta\tau P \le 1/64$, we obtain

$$\frac{L\mathcal{Q}_{\mathrm{imb}}(\eta)}{2\mu D_{\mathrm{PL}}^{\mathrm{imb}}(\eta)} \le c\frac{L\eta V_\nu}{\mu\tau} + c\frac{L^2\eta^2J_{\mathrm{CW}}^2}{\mu} + c\frac{B_{\mathrm{cyc}}^2}{\mu}.$$

Combining these bounds with the definition in (120) proves (123). □

Now choose the horizon-dependent PL step size in Appendix D.3 and assume $T \geq 256(L/\mu)P\log T$. Since $V_\nu = \tau^2\Sigma_{\text{eff}}^2$, substituting (123) into (122) and using the same contraction estimate as in (86) gives universal constants $\widetilde{C}_1, \widetilde{C}_2, \widetilde{C}_3 > 0$ such that

$$\mathbb{E}[F(\boldsymbol{w}^{(K,0)})] - F^* \leq \frac{\Phi_0^{\text{PL}}}{T^2} + \widetilde{C}_1 \frac{L\log T}{\mu^2 T}\Sigma_{\text{eff}}^2 + \widetilde{C}_2 \frac{L^2(\log T)^2}{\mu^3 T^2}\mathcal{R}_{\text{CW}}^2 + \widetilde{C}_3 \frac{B_{\text{cyc}}^2}{\mu}. \tag{124}$$

Thus, the finite-horizon PL rate has the same form as the bound under a balanced cyclic pattern, plus a cyclic-bias term. Using $B_{\text{cyc}} = (\alpha+\gamma)\Delta_{\text{cyc}}$, the additional term is $\mathcal{O}\big((\alpha+\gamma)^2\Delta_{\text{cyc}}^2/\mu\big)$.

Equation (122) retains the stochastic and local-update residuals for a fixed learning rate. With the horizon-dependent choice, (124) separates the decaying terms from the persistent cyclic-bias contribution and proves Theorem 4.1.

Suppressing the logarithmic factors in (124) gives the following simplified PL bound.

Corollary G.1 (Simplified PL rate under an unbalanced cyclic pattern)**.** *For fixed system parameters,*

$$\mathbb{E}[F(\boldsymbol{w}^{(K,0)})] - F^* = \widetilde{\mathcal{O}}\left(\frac{\Sigma_{\text{eff}}^2}{T} + \frac{\Phi_0^{\text{PL}} + \mathcal{R}_{\text{CW}}^2}{T^2}\right) + \mathcal{O}\left(\frac{(\alpha+\gamma)^2\Delta_{\text{cyc}}^2}{\mu}\right). \tag{125}$$

G.5 General Non-Convex Stationarity with Unbalanced Cyclic Patterns

The non-convex per-step descent and the partial-cycle squared-drift recursion are unchanged. Only the cycle-level window-gradient coupling changes. The fixed-anchor term now satisfies

$$\sum_{p=1}^{P} \boldsymbol{\epsilon}^{(k,p)}(\boldsymbol{w}^{(k,0)}) = P\boldsymbol{b}_{\text{cyc}}(\boldsymbol{w}^{(k,0)}),$$

which motivates the definition

$$\chi_1^{\text{imb}} \triangleq \chi_1 + PB_{\text{cyc}}.$$

Repeating the anchor decomposition preceding (97) leaves an additional fixed-anchor contribution bounded by $PB_{\text{cyc}}\mathbb{E}[G_k]$. The coupling bound becomes

$$|\mathcal{C}_{\epsilon,k}| \leq \chi_2\mathbb{E}[G_k^2] + \chi_1^{\text{imb}}\mathbb{E}[G_k] + \chi_0.$$

The relation between G_k^2 and $\mathcal{V}_k$ in Appendix E.2 is also unchanged because it only uses the partial-cycle squared-drift bound. Consequently, the descent coefficient remains

$$\widehat{A}(\eta) = A_{\text{step}} - \eta\tau\left(\chi_2 + \frac{P}{32}\right)\lambda_1,$$

while the residual becomes

$$\widehat{C}_{\text{imb}}(\eta) = PC_{\text{step}} + \eta\tau\left[\left(\chi_2 + \frac{P}{32}\right)\lambda_0 + \frac{8(\chi_1^{\text{imb}})^2}{P} + \chi_0\right].$$

Thus the one-cycle recursion is

$$\mathbb{E}[F(\boldsymbol{w}^{(k+1,0)})] \leq \mathbb{E}[F(\boldsymbol{w}^{(k,0)})] - \widehat{A}(\eta)\mathcal{V}_k + \widehat{C}_{\text{imb}}(\eta).$$

Whenever $\widehat{A}(\eta) > 0$, telescoping over $k = 0, \ldots, K-1$ and using the lower bound on the objective gives the finite-step guarantee

$$\frac{1}{T}\sum_{k=0}^{K-1}\sum_{p=1}^{P}\mathbb{E}\|\nabla F(\boldsymbol{w}^{(k,p-1)})\|^2 \leq \frac{\Phi_0^{\text{nc}}}{T\widehat{A}(\eta)} + \frac{\widehat{C}_{\text{imb}}(\eta)}{P\widehat{A}(\eta)}. \tag{126}$$

This is the counterpart of (103) under an unbalanced cyclic pattern. To isolate the additional cyclic-bias contribution, recall the residual expression $\widehat{C}(\eta)$ defined in (101). The elementary inequality

$$(\chi_1 + PB_{\text{cyc}})^2 \leq 2\chi_1^2 + 2P^2B_{\text{cyc}}^2$$

then gives

$$\widehat{C}_{\mathrm{imb}}(\eta) \le 2\widehat{C}(\eta) + 16\eta\tau P B_{\mathrm{cyc}}^2.$$

For comparison, denote the right-hand side of (103), evaluated at the current system parameters, by

$$\mathcal{R}_{\mathrm{bal}}(T,\eta) \triangleq \frac{\Phi_0^{\mathrm{nc}}}{T\widehat{A}(\eta)} + \frac{\widehat{C}(\eta)}{P\widehat{A}(\eta)}.$$

The non-convex step-size range in Appendix E.3 ensures $\widehat{A}(\eta) \ge (3/4)\eta\tau$. Substituting this lower bound into (126) gives

$$\frac{1}{T}\sum_{k=0}^{K-1}\sum_{p=1}^{P}\mathbb{E}\|\nabla F(\boldsymbol{w}^{(k,p-1)})\|^2 \le 2\mathcal{R}_{\mathrm{bal}}(T,\eta) + \frac{64}{3}B_{\mathrm{cyc}}^2. \tag{127}$$

Therefore, an unbalanced cyclic pattern adds a term of order B_{cyc}^2 to the guarantee under a balanced cyclic pattern. Under the step-size choice in Appendix E.3, $\mathcal{R}_{\mathrm{bal}}(T,\eta_T^{\mathrm{nc}})$ vanishes with T, while the cyclic-bias term in the bound is independent of T. This proves Theorem 4.2.

Applying the prescribed step size to (127) gives the following simplified non-convex bound.

Corollary G.2 (Simplified non-convex rate under an unbalanced cyclic pattern)**.** *For fixed system parameters,*

$$\frac{1}{T}\sum_{k=0}^{K-1}\sum_{p=1}^{P}\mathbb{E}\left\|\nabla F(\boldsymbol{w}^{(k,p-1)})\right\|^2 = \mathcal{O}(T^{-1/2}) + \mathcal{O}\big((\alpha+\gamma)^2\Delta_{\mathrm{cyc}}^2\big). \tag{128}$$

H Additional Experimental Details and Results

This section provides the experimental configurations and supplementary results for Section 5. The studies examine a balanced cyclic pattern, within-window sampling, and the interaction between an unbalanced cyclic pattern and client heterogeneity.

H.1 Evaluation protocol and common training procedure

For each replicate, Cycwin-FL and the method used for comparison share the same data partition and model initialization. Cycwin-FL follows Algorithm 1; all compared methods use constant learning rates. Image-classification metrics are evaluated every five rounds and the PL objective every fifty rounds. At each evaluation round, the plotted line shows the mean across replicates, and the shaded band denotes one standard deviation. The CIFAR-100 studies use 10 Monte Carlo replicates. The CIFAR-10, MNIST, and EuroSAT studies use 20 Monte Carlo replicates. The PL experiment uses 1000 Monte Carlo replicates.

The experiments examine the qualitative effects of sampling, model drift, and cyclic weighting using empirically chosen learning rates and horizons. They do not directly test the asymptotic rates under the sufficient step-size and horizon conditions in Sections 3 and 4. Test accuracy measures predictive performance, whereas the non-convex theory bounds the average squared gradient norm.

For the benchmark comparisons, window-size ablations, sampling-ratio studies, and heterogeneity studies, the client partition is regenerated in each replicate and paired across the compared schedules. All image experiments use 50 clients, mini-batches of size 64, SGD without momentum, and a server test batch size of 256. Unless a study specifies a controlled partition, each client receives data according to a class-proportion vector drawn from a symmetric Dirichlet distribution with concentration α_{Dir}.

H.2 Data sets, models, and optimization

CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) each contain 50,000 training and 10,000 test RGB images of size 32×32, divided into 10 and 100 classes, respectively. MNIST (LeCun et al., 1998) contains 60,000 training and 10,000 test grayscale digit images of size 28×28. EuroSAT (Helber et al., 2019) contains 27,000 RGB satellite patches of size 64×64 from 10 land-use and land-cover

classes; a seed-42 random split assigns 80% to training and 20% to testing. Together, these data sets cover natural-image and remote-sensing classification. All experiments use synthetic cyclic-window schedules, with EuroSAT providing the remote-sensing benchmark.

CIFAR training uses random crops with four-pixel padding and horizontal flips, followed by data-set-specific normalization; its test images are only normalized. MNIST images are normalized without augmentation. EuroSAT training uses a random resized crop with scale in $[0.8, 1]$, horizontal and vertical flips, and normalization; its test images are normalized. CIFAR and EuroSAT use a ResNet-18 adapted to the input resolution (He et al., 2016), while MNIST uses a two-convolutional-layer network. Table 2 gives the remaining data-set-specific settings. CIFAR-100 uses $\tau = 20$ local steps and 1,000 global rounds. The other image benchmarks use $\tau = 15$ and 500 rounds.

Table 2: Common setup for the image-classification experiments.

Data set	Model	Samples/client	τ	Rounds	Replicates
CIFAR-10	ResNet-18	1,000	15	500	20
CIFAR-100	ResNet-18	1,000	20	1,000	10
EuroSAT	ResNet-18	432	15	500	20
MNIST	Two-layer CNN	1,000	15	500	20

All four tasks use a learning rate of 0.005. For the Cycwin-FL–FedAvg comparison, Cycwin-FL advances the window by $N_w = 1$ and selects the complete window, while FedAvg samples the same number of clients uniformly without replacement from the full population. The main comparison uses $B = S \in \{1, 5, 10\}$; the supplementary window-size ablation uses $B = S \in \{1, 3, 5, 7, 9\}$ with $N_w = 1$. Thus, every availability cycle has a balanced cyclic pattern in these comparisons.

H.3 Controlled PL objective

To compare participation mechanisms on an objective with known F^* and PL constants, we construct a 20-dimensional finite-sum quadratic. For client m and local sample j,

$$f_{m,j}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w} - \boldsymbol{a}_m)^\top \boldsymbol{H}(\boldsymbol{w} - \boldsymbol{a}_m) + \boldsymbol{e}_{m,j}^\top(\boldsymbol{w} - \boldsymbol{a}_m). \tag{129}$$

Here, $\boldsymbol{H} \in \mathbb{R}^{20\times 20}$ is generated by a random orthogonal basis and geometrically spaced eigenvalues from 1 to 20, and $\boldsymbol{w}^*$ denotes the global minimizer. The 50 client centers satisfy $\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{a}_m = \boldsymbol{w}^*$ and have a root-mean-square distance of 2 from $\boldsymbol{w}^*$. Each client has 128 finite-sum samples; its perturbations obey $\frac{1}{128}\sum_{j=1}^{128} \boldsymbol{e}_{m,j} = \boldsymbol{0}$ and have a root-mean-square norm of 1. Therefore,

$$F(\boldsymbol{w}) - F^* = \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}^*)^\top \boldsymbol{H}(\boldsymbol{w} - \boldsymbol{w}^*), \tag{130}$$

$$\nabla F(\boldsymbol{w}) = \boldsymbol{H}(\boldsymbol{w} - \boldsymbol{w}^*), \tag{131}$$

so F is 1-strongly convex, is L-smooth with $L = 20$, and satisfies the PL inequality with $\mu = 1$.

The centers are first constructed as a smooth, centered cyclic collection with relative jitter 0.15. They are then randomly permuted over positions in the cyclic order, so nearby positions are not artificially assigned similar objectives. The common initialization satisfies $\|\boldsymbol{w}^{(0,0)} - \boldsymbol{w}^*\| = 5$. We use stochastic mini-batches of size 16, $\tau = 5$, constant learning rate 10^{-4}, 5,000 rounds, and 1000 Monte Carlo replicates. Cycwin-FL uses $N_w = 1$ and $B = S \in \{1, 5, 10\}$; the comparator samples S clients uniformly from all 50 clients. Full-window participation eliminates client-sampling variance, but the mini-batch perturbations retain stochastic-gradient noise, so this experiment does not represent the deterministic full-window case with $\sigma = 0$.

H.4 Sampling-ratio and unbalanced cyclic pattern studies

The sampling-ratio experiment uses the CIFAR-10 base setup with $\alpha_{\mathrm{Dir}} = 1$, fixes $B = N_w = 5$, and varies $S \in \{5, 4, 3, 2, 1\}$, corresponding to $\varrho = S/B \in \{1, 0.8, 0.6, 0.4, 0.2\}$. This varies within-window sampling under a balanced cyclic pattern.

The unbalanced cyclic pattern experiment uses a controlled correlation between cyclic weights and client heterogeneity to make the effects of an unbalanced cyclic pattern observable within 500 rounds. It uses CIFAR-10, $M = 50$, full-window participation $S = B$, $B \in \{5, 3\}$, and $N_w \in \{1, 2, 3\}$. Both values of B use the same optimization setup: $\tau = 15$, learning rate 0.005, batch size 64, 500 rounds, and 20 replicates. For $N_w = 2$, the cyclic pattern is unbalanced, with $d = 2$ and $\Delta_{\text{cyc}} = 1/5$ for $B = 5$ and $1/3$ for $B = 3$; the $N_w \in \{1, 3\}$ controls have balanced cyclic patterns.

Examples from classes 0–4 are assigned to the 25 clients that are available more often when $N_w = 2$, while the 25 less frequently available clients receive examples from classes 5–9. Each client holds 1,000 examples, and the two groups are equally large, so the population is class-balanced even though the cyclic weighting favors classes 0–4. This assignment correlates client heterogeneity with cyclic weights in (15).

Table 3: Configurations used for the main-paper experiments.

Study	Participation	Optimization
CIFAR-10 comparison	Cycwin-FL $N_w = 1$, $B = S \in \{1, 5, 10\}$; matched FedAvg S	ResNet-18, $\tau = 15$, $\eta = 0.005$, 500 rounds
PL quadratic	Cycwin-FL $N_w = 1$, $B = S \in \{1, 5, 10\}$; matched FedAvg S	$\mu = 1$, $L = 20$, $\tau = 5$, $\eta = 10^{-4}$, 5,000 rounds
Sampling ratio	Cycwin-FL $B = N_w = 5$, $\varrho \in \{1, .8, .6, .4, .2\}$	CIFAR-10 base setup
Cyclic pattern, $B = 5$	Cycwin-FL $N_w \in \{1, 2, 3\}$; $N_w = 2$ gives an unbalanced cyclic pattern	$\tau = 15$, $\eta = 0.005$, 500 rounds
Cyclic pattern, $B = 3$	Cycwin-FL $N_w \in \{1, 2, 3\}$; $N_w = 2$ gives an unbalanced cyclic pattern	$\tau = 15$, $\eta = 0.005$, 500 rounds

H.5 Supplementary experimental results

The supplementary experiments examine sampling and heterogeneity effects across data sets, Dirichlet concentrations, and window sizes. We first compare Cycwin-FL with FedAvg, then vary the window size, within-window sampling ratio, and inter- or intra-window heterogeneity. Accuracy and loss are reported separately to show both predictive performance and optimization behavior.

H.5.1 Cycwin-FL versus FedAvg across data sets and heterogeneity levels

We compare Cycwin-FL with FedAvg at matched participation budgets on CIFAR-10, CIFAR-100, EuroSAT, and MNIST, using $\alpha_{\text{Dir}} \in \{0.5, 1, 2\}$. Smaller concentrations produce more class-concentrated local data and hence stronger client heterogeneity. Across these settings, Cycwin-FL remains close to FedAvg at the same S, while run-to-run variability generally decreases as the full window grows. This behavior is consistent with the equal client weighting induced by a balanced cyclic pattern, even when the participating clients in individual rounds have highly non-IID data. See Figures 6–9. The CIFAR-10 accuracy curve for $\alpha_{\text{Dir}} = 1$ in this Cycwin-FL–FedAvg comparison also appears in Figure 3.

H.5.2 Window-size ablations

The window-size suite uses $N_w = 1$, full-window participation $S = B$, and $B \in \{1, 3, 5, 7, 9\}$ while retaining the data-set-specific settings in Table 2. It evaluates the same four data sets and three Dirichlet concentrations. Larger windows average more local directions and typically improve accuracy while reducing run-to-run variability. Here $P = M$ for every B, and $\Sigma^2_{\text{eff}} = \sigma^2/(\tau B)$: the explicit variance reduction concerns mini-batch noise, since client-sampling variance is zero. Changing B can also change the window heterogeneity, so the observed improvements need not arise from variance reduction alone. Larger windows require more client computation at the same round budget. See Figures 10–13.

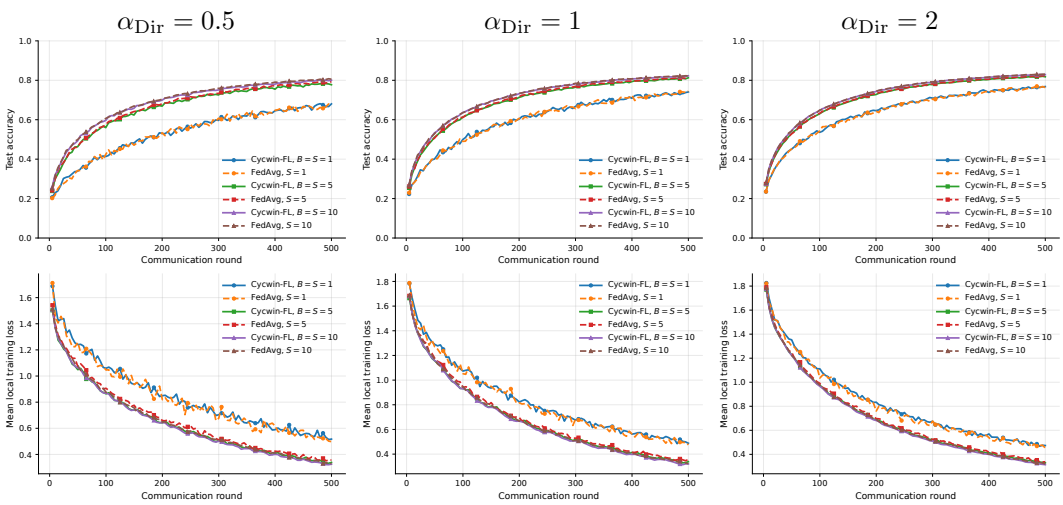


Figure 6: Cycwin-FL versus FedAvg on CIFAR-10: test accuracy (top) and training loss (bottom).

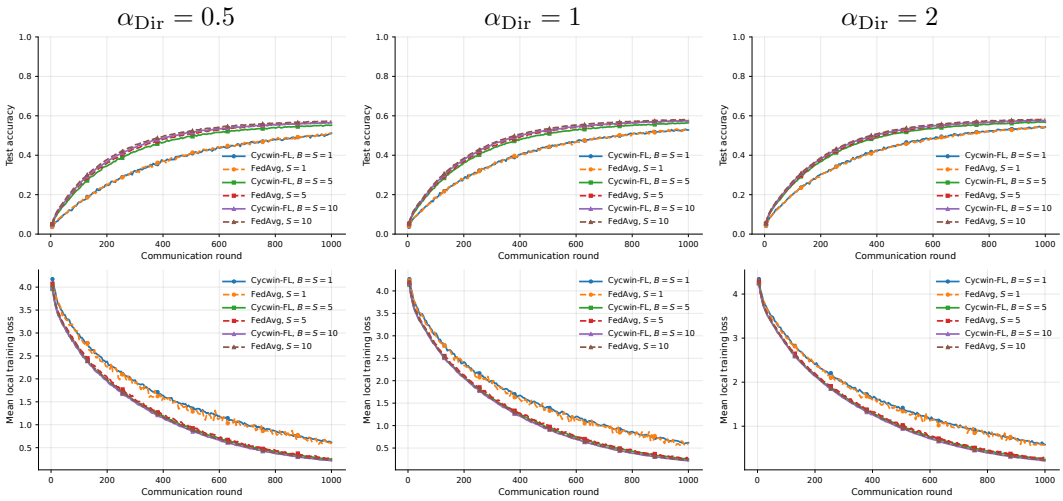


Figure 7: Cycwin-FL versus FedAvg on CIFAR-100: test accuracy (top) and training loss (bottom).

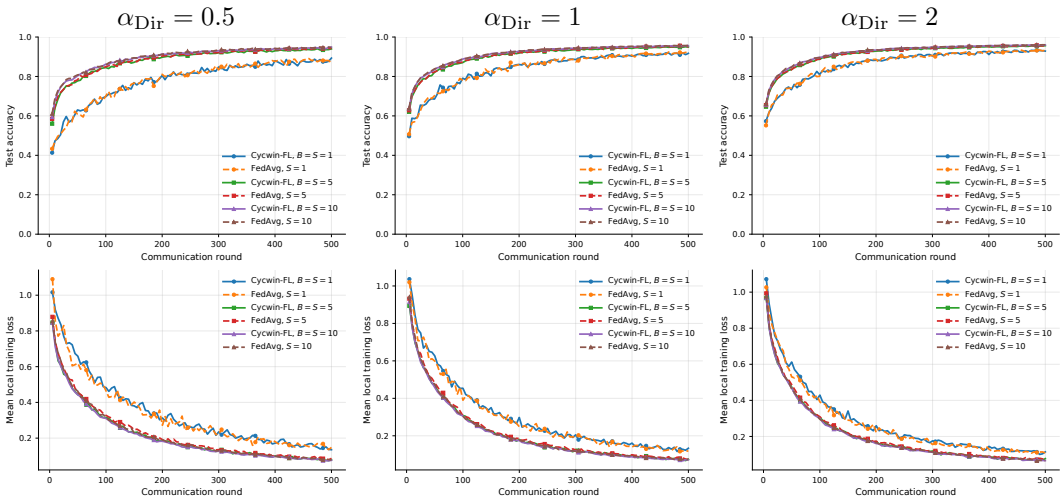


Figure 8: Cycwin-FL versus FedAvg on EuroSAT: test accuracy (top) and training loss (bottom).

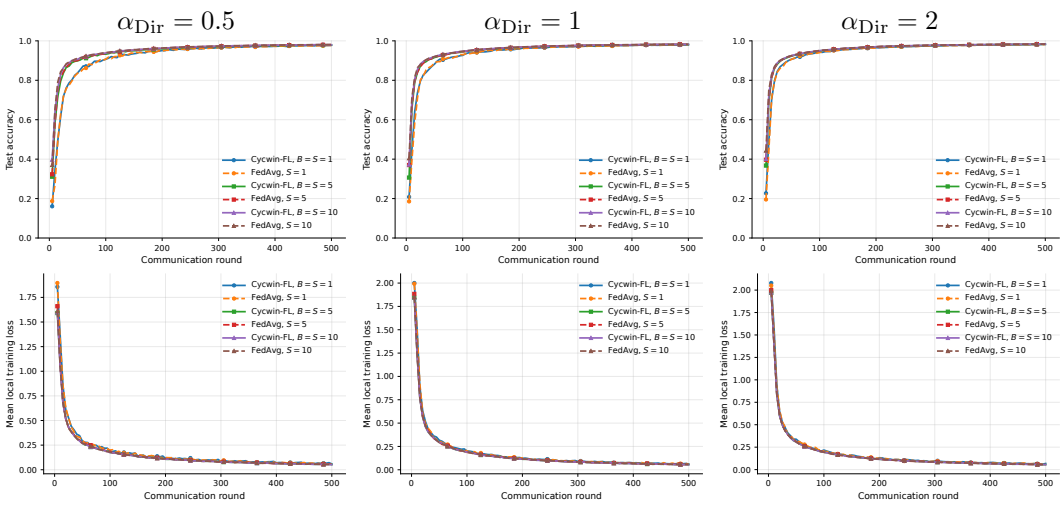


Figure 9: Cycwin-FL versus FedAvg on MNIST: test accuracy (top) and training loss (bottom).

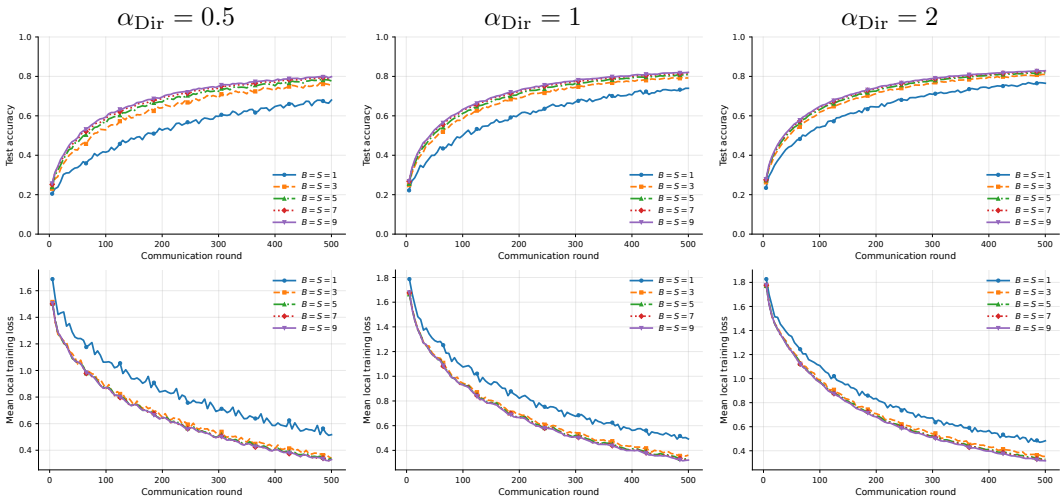


Figure 10: Window-size ablations on CIFAR-10: test accuracy (top) and training loss (bottom).

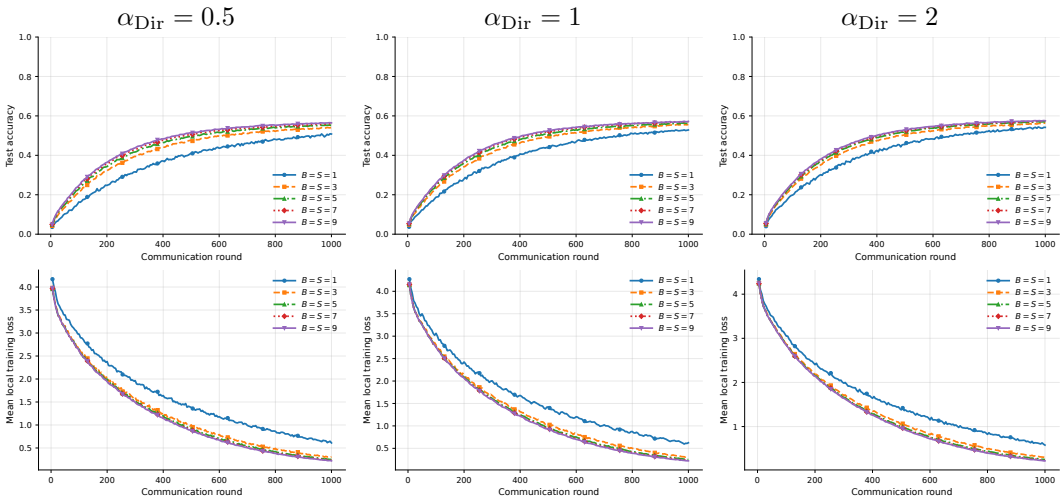


Figure 11: Window-size ablations on CIFAR-100: test accuracy (top) and training loss (bottom).

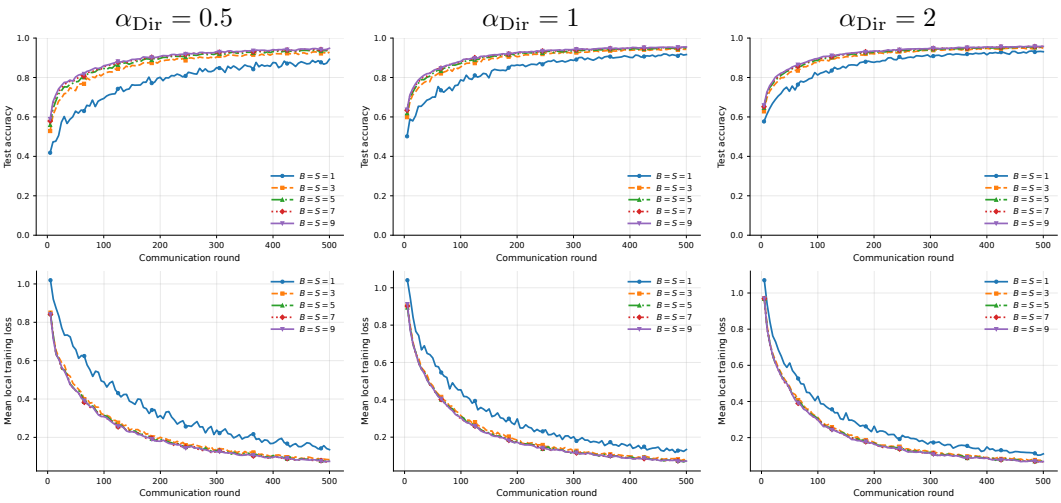


Figure 12: Window-size ablations on EuroSAT: test accuracy (top) and training loss (bottom).

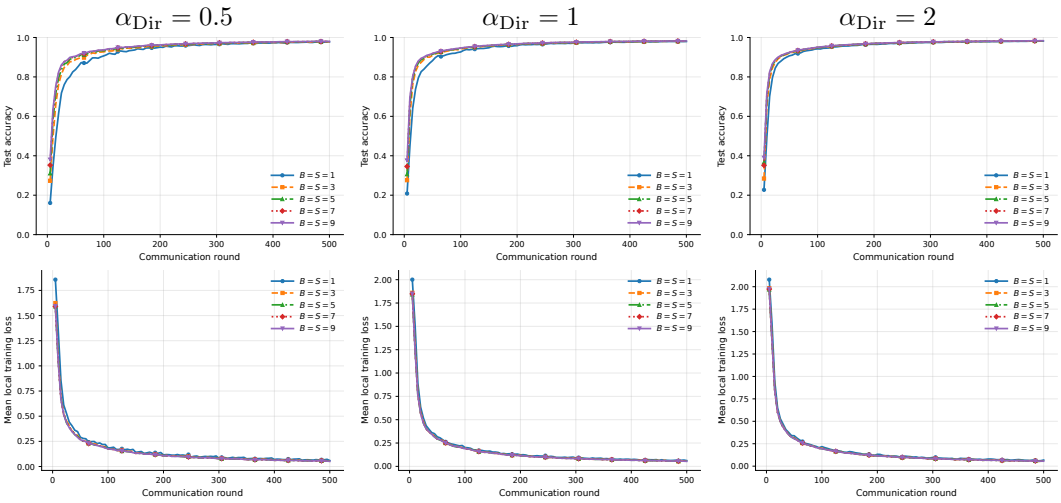


Figure 13: Window-size ablations on MNIST: test accuracy (top) and training loss (bottom).

H.5.3 Within-window subsampling

The sampling-ratio suite uses CIFAR-10, $B = N_w = 5$, $\tau = 15$, learning rate 0.005, and 500 rounds. It varies $S \in \{5, 4, 3, 2, 1\}$ at each $\alpha_{\text{Dir}} \in \{0.5, 1, 2\}$. Across all three concentrations, smaller sampling ratios slow learning and increase variability. This trend is consistent with the finite-population factor $\omega_{B,S}$ in Σ^2_{eff}: as S decreases, the server observes a noisier estimate of the visible-window direction and averages fewer local stochastic updates. See Figure 14. The CIFAR-10 accuracy curve for $\alpha_{\text{Dir}} = 1$ in this subsampling study also appears in Figure 5.

H.5.4 Inter- and intra-window heterogeneity

The final two studies use controlled data partitions to examine the two heterogeneity scales in Assumption 3. Both studies use CIFAR-10 with 50 clients, $B = N_w = S = 5$, 1,000 samples per client, $\tau = 15$, learning rate 0.005, 500 rounds, and 20 replicates. Since $N_w = B$, each cycle contains ten disjoint windows.

For the inter-window study, each window draws one class-proportion prototype from a symmetric Dirichlet distribution with concentration $\alpha_{\text{proto}} \in \{0.2, 0.5, 1, 2, 5\}$. All five clients in that window receive the same prototype. Thus, within-window client distributions are matched, while smaller

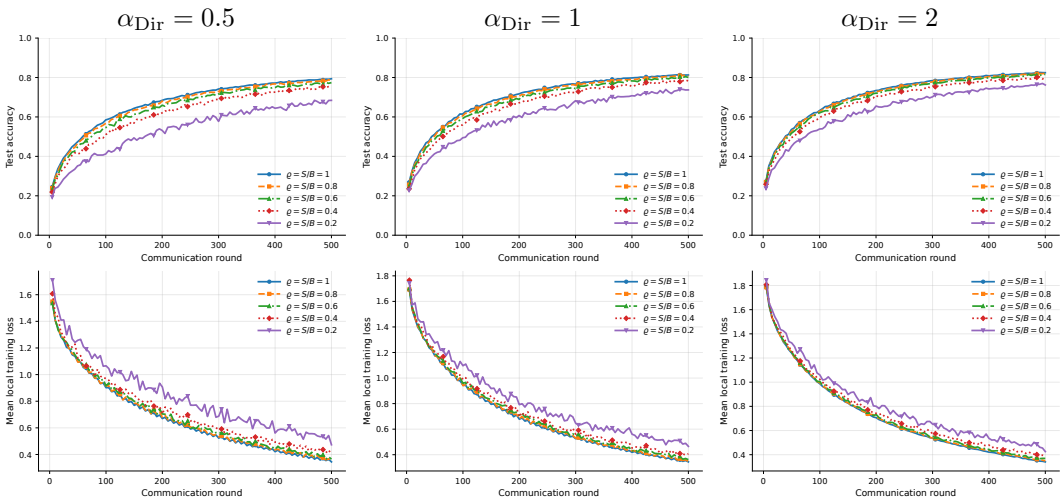


Figure 14: CIFAR-10 subsampling: test accuracy (top) and training loss (bottom).

α_{proto} creates more concentrated and more distinct window objectives. Raising α_{proto} from 0.2 to 5 increases final accuracy from 0.5372 ± 0.0459 to 0.8207 ± 0.0091, consistent with sensitivity to window-to-global heterogeneity. See Figure 15.

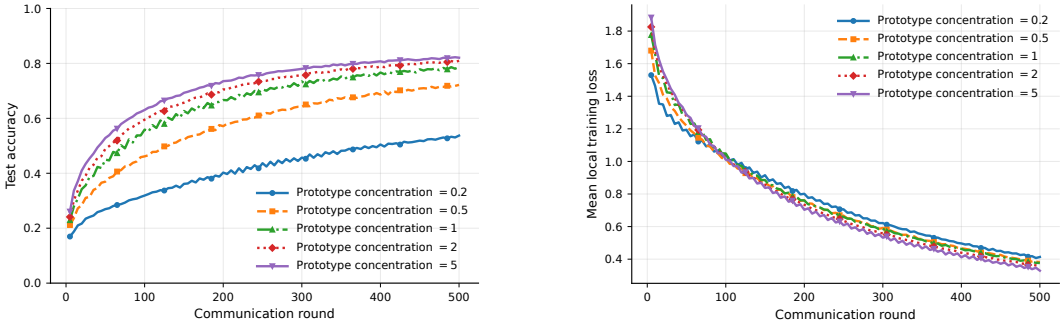


Figure 15: Inter-window heterogeneity on CIFAR-10: test accuracy (left) and training loss (right).

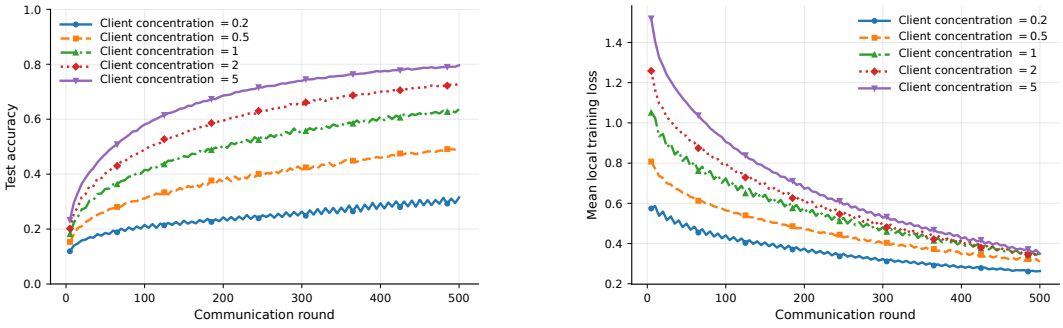


Figure 16: Intra-window heterogeneity on CIFAR-10: test accuracy (left) and training loss (right).

For the intra-window study, every client distribution is centered in expectation at the same uniform class prototype. Each client independently draws its proportions from $\mathrm{Dirichlet}((\beta/10)\mathbf{1})$ with $\beta \in \{0.2, 0.5, 1, 2, 5\}$, where $\mathbf{1} \in \mathbb{R}^{10}$ is the all-ones vector. Smaller β creates sharper, more distinct client distributions inside the same window, whereas larger β concentrates them near the common uniform prototype. Final accuracy rises from 0.3151 ± 0.0641 at $\beta = 0.2$ to 0.7963 ± 0.0091 at $\beta = 5$. The trend is consistent with sensitivity to stronger client-level heterogeneity. See Figure 16.

Both studies use $S = B$, so client-sampling variance is absent, while intra-window heterogeneity can still affect local drift through H in $\mathcal{R}_{\mathrm{CW}}$. The partition parameters control data distributions, not the gradient bounds α and γ directly; finite client samples and varying window averages can affect both scales.