TRUSS: TRANSMISSION-REDUCED UNIFIED-STATE SCAFFOLDING UNDER NON-STATIONARY CLIENT PARTICIPATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Heterogeneous client objectives and unequal, time-varying availability create coupled sources of error in federated optimization: local drift and participation bias. Correcting both while preserving the speedup from local computation and keeping communication and memory costs low is a central challenge. We introduce *Transmission-Reduced Unified-State Scaffolding* (TRUSS), which addresses both with a single persistent, model-sized memory per client. The difference between this memory and the current server model supplies the correction applied at every local step. Each active client therefore exchanges one model-sized vector in each direction per round, matching FedAvg's per-round communication, and the server stores only the global model. For smooth non-convex objectives under independent participation whose probabilities may vary across clients and rounds, we establish an $O((NKR)^{-1/2}+R^{-1})$ stationarity bound, yielding linear speedup in the number of clients N and local steps K, with no bounded-gradient-heterogeneity assumption and no knowledge of client inclusion probabilities or importance weighting. Experiments demonstrate the effectiveness and communication efficiency of TRUSS.

1 INTRODUCTION

Federated learning trains a shared model across clients while keeping their training data local. Federated averaging (FedAvg) implements this principle by alternating several local stochastic-gradient steps with server aggregation (McMahan et al., 2017). Increasing the number of local steps between communications is a central way to reduce communication frequency (Stich, 2019), but its benefit depends on whether the resulting updates remain aligned with the global objective.

Two properties of deployed systems undermine this alignment: *objective heterogeneity* and *unequal, time-varying client availability*. Because client data distributions differ, repeated local steps pull client models toward different local solutions, a phenomenon known as client drift (Karimireddy et al., 2020). Unequal availability gives frequently participating clients disproportionate influence and can change the objective the system actually optimizes (Wang and Ji, 2024). The two effects compound: local drift strengthens the client-specific directions that unequal participation overrepresents, while that imbalance moves the shared model toward frequently available clients and increases its disagreement with the objectives of absent clients (Xiang et al., 2024; Ying et al., 2025). Correcting drift without correcting participation imbalance, or correcting the imbalance while letting drift accumulate, therefore leaves part of the same optimization problem unresolved.

Existing approaches address important parts of this problem. Variance reduction methods like SCAFFOLD and SCALLION correct local drift without a bounded gradient-heterogeneity assumption, but their standard guarantees assume uniform client sampling, and their control variates require additional server state and communication (Karimireddy et al., 2020; Huang et al., 2024). Availability-focused methods instead reuse past updates, estimate aggregation weights, or compensate for missed work, as in MIFA, FedAU, and FedAWE (Gu et al., 2021; Wang and Ji, 2024; Xiang et al., 2024). These methods do consider heterogeneous data, but they retain heterogeneity-dependent assumptions or error terms and do not correct each local step with a control variate, so longer local trajectories can increase drift and limit the benefit of additional computation between communications. FOCUS and its stochastic variant SG-FOCUS jointly correct local-update and

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.

Table 1: Comparison of representative methods.

Method	No het. bound	K speedup	Participation	Comm.	Client mem.	Server mem.
FedAvg (McMahan et al., 2017)	×	✓	Uniform partial	2	0	1
FedProx (Li et al., 2020a)	×	✓	IID sampling	2	0	1
MIFA[a] (Gu et al., 2021)	×	✓	Bounded inactivity	2	0	$N+1$
FedAWE (Xiang et al., 2024)	×	✓	Non-stationary	2	1	1
FedAU (Wang and Ji, 2024)	×	✓	Stationary	2	0	1
ProxSkip (Mishchenko et al., 2022)	✓	×	Full	2	1	1
SCAFFOLD (Karimireddy et al., 2020)	✓	✓	Uniform	4	1	2
SCALLION (Huang et al., 2024)	✓	✓	Uniform	3	1	2
SG-FOCUS (Ying et al., 2025)	✓	×	Stationary	2	1	2
FedSUM-CR[b] (You and Pu, 2025)	✓	✓	Bounded inactivity	2	2	2
TRUSS[e]	✓	✓	Non-stationary	**2**	**1**	**1**

"No het. bound" means no uniform gradient-heterogeneity bound. "K speedup" shows linear speedup of the convergence rate with local steps. "Comm." counts uncompressed model-sized vectors sent in both directions per active client per communication round. Memory counts persistent model-sized vectors, including the server model and excluding temporary workspaces, data, and scalar counters. [a]MIFA's non-convex result requires bounded inactivity and additional oracle and smoothness assumptions. [b]FedSUM rates and step-size caps depend explicitly on the worst-case participation delay.

participation bias without a heterogeneity bound (Ying et al., 2025); they assume time-invariant participation statistics and do not establish linear speedup in the number of local steps. Amplified SCAFFOLD addresses heterogeneous objectives and periodic availability, with linear speedup under a participation-balance condition over windows of rounds (Crawshaw and Liu, 2024). FedSUM-CR attains FedAvg-level communication under bounded inactivity participation without a heterogeneity bound (You and Pu, 2025); its communication-reduced variant, however, stores a second model-sized vector at every client, and its rates are stated in terms of worst-case and average participation delays. What remains open is the combination of per-step drift correction, persistently unequal and changing availability, and FedAvg-level per-round communication and server memory, using only one persistent model-sized vector per client and requiring no balanced participation window.

We introduce *Transmission-Reduced Unified-State Scaffolding* (TRUSS), which addresses both challenges through a single persistent, model-sized memory on each client. The difference between that memory and the current server model supplies the correction applied at every local step. The memory updates are constructed so that the client memories always average to the server model; the corrections consequently average to zero, as SCAFFOLD's control variates do, while no client ever receives a second vector. When a client is absent, its memory stays fixed while the server model moves, so their difference also records the net server movement during the absence. This unified state therefore serves drift correction and availability compensation at once, without estimating participation probabilities or tracking inactivity durations. Each active client receives one model and sends one model displacement per round, matching FedAvg's per-round communication, and the server retains only the global model. To the best of our knowledge, TRUSS is the first method to combine FedAvg-level per-round communication and server memory with only one persistent model-sized vector per client, together with convergence under non-stationary participation, no bounded-gradient-heterogeneity assumption, and linear speedup in the number of clients and local steps. Table 1 summarizes these distinctions, which we discuss in detail in Section 2.

Contributions.

- **Joint correction through a single client memory.** We introduce TRUSS, which uses one persistent, model-sized vector per client to correct local drift and compensate for unequal, time-varying availability. Each active client exchanges one vector in each direction per round, matching FedAvg's communication load while requiring no auxiliary server vector (Algorithm 1).
- **Convergence under non-stationary participation.** For smooth non-convex objectives, we establish convergence to stationarity at the server iterates under independent participation with unequal, time-varying probabilities. The guarantee requires no uniform gradient-heterogeneity bound and explicitly accounts for initial memory-tracking error (Theorem 1).

- **Linear speedup in clients and local steps.** We obtain an $O((NKR)^{-1/2} + R^{-1})$ stationarity bound over R communication rounds. The leading term yields linear speedup in the number of clients N and local steps K (Corollary 1).
- **Empirical accuracy and communication efficiency.** Experiments evaluate TRUSS under uniform, unequal stationary, and non-stationary availability, with ablations of data heterogeneity and availability fluctuations. TRUSS achieves comparable accuracy over communication rounds to SCAFFOLD and SCALLION at a lower communication cost (Section 7).

2 Related work

Local computation, heterogeneity, and availability. FedAvg and local SGD reduce communication through local updates (McMahan et al., 2017; Stich, 2019), which induce drift under heterogeneity (Li et al., 2020b; Karimireddy et al., 2020). FedProx uses proximal regularization (Li et al., 2020a), with later analysis removing bounded dissimilarity (Yuan and Li, 2022). SCAFFOLD corrects drift with client/server control variates (Karimireddy et al., 2020); SCALLION supports compression and merges uplink updates (Huang et al., 2024). ProxSkip/ScaffNew skips communication under full participation (Mishchenko et al., 2022); PAdaMFed combines adaptive steps, momentum, and control variates under uniform partial participation (Yan et al., 2025). Participation-aware FedAvg analysis links convergence to availability (Wang and Ji, 2022). MIFA and FedVARP reuse client updates with server storage proportional to client count in their server-memory formulations (Gu et al., 2021; Jhunjhunwala et al., 2022); FedAU estimates weights under stationary participation (Wang and Ji, 2024), whereas FedAWE combines missed-work compensation with implicit gossiping under non-stationary availability (Xiang et al., 2024). FedSWE allows zero-probability rounds under window-averaged availability (Xiang et al., 2026). These availability-focused guarantees retain bounded heterogeneity assumptions and leave local trajectories susceptible to drift.

Joint correction and the role of local steps. Amplified SCAFFOLD jointly addresses heterogeneity and periodic availability, with linear speedup under a participation-balance condition over windows (Crawshaw and Liu, 2024). FOCUS/SG-FOCUS correct both participation and local-update bias without a heterogeneity bound, but assume fixed participation probabilities (Ying et al., 2025). SG-FOCUS provides client-count speedup, but its stochastic non-convex bound lacks local-step speedup; FedSUM and FedSUM-CR also correct both effects without a heterogeneity bound, supporting participation with bounded inactivity (You and Pu, 2025). FedSUM exchanges three model-sized vectors per active round; FedSUM-CR reduces this to two by storing an additional client vector. Both retain an auxiliary server vector. TRUSS has the best footprint with two-vector communication with one persistent client vector and no auxiliary server vector. Under independent, non-stationary participation, Corollary 1 gives $O((NKR)^{-1/2} + R^{-1})$, allowing linear speedup when the leading term dominates, without participation balance across clients.

Analysis of time-varying mixing. Backward weights connect to absolute probability sequences for stochastic-matrix products (Touri and Nedić, 2014) and optimization on time-varying directed graphs (Saadatniaki et al., 2020). We combine them with heterogeneity tracking and covariance control for retained noise to establish server-model descent and stochastic averaging without requiring unbiased server updates.

3 Problem setup

In a federated learning system with N clients, we minimize $f(x) = N^{-1} \sum_{i=1}^{N} f_i(x)$, where client i has local data distribution $\mathcal{D}_i$ and local objective $f_i(x) := \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[\ell(x; \xi_i)]$, with ℓ denoting the sample loss. We have the following assumptions:

Assumption 1 (Non-stationary independent participation)**.** Let $\mathcal{S}^r \subseteq \{1, \dots, N\}$ denote the set of active clients in round r. The indicators $I_i^r = \mathbf{1}_{\{i \in \mathcal{S}^r\}}$ are independent across clients and rounds. We assume $p_i^r = \mathbb{P}(I_i^r = 1) \geq \delta > 0$ for every client and round. Also define $\delta_{\max} = \sup_{i,r} p_i^r \leq 1$.

There is no bound on the length of an inactivity period, as required in MIFA and FedSUM.

Assumption 2 (Stochastic gradients)**.** For $\xi_i \sim \mathcal{D}_i$, define the stochastic gradient $g_i(x; \xi_i) := \nabla_x \ell(x; \xi_i)$. These gradients are unbiased with bounded variance:

$$\mathbb{E}[g_i(x; \xi_i) \mid x] = \nabla f_i(x), \quad \mathbb{E}\Big[\|g_i(x; \xi_i) - \nabla f_i(x)\|^2 \mid x\Big] \leq \sigma^2, \quad \forall i \in \{1, \dots, N\},\ x \in \mathbb{R}^d.$$

Every oracle call uses a fresh sample, independently across clients and calls and independently of the participation indicators. We write $g_i(x)$ when the sample is clear.

Algorithm 1 TRUSS

1: **Notation:** x^r is the server model at round r; $y_i^{r,k}$ is client i's local model at step k of round r; m_i^r is client i's memory in round r.
2: **Initialize:** R rounds, K local steps , and local and server step sizes $\eta_\ell, \eta_g > 0$; set $h = K\eta_g\eta_\ell$. Initialize the server model x^0 and memories with $N^{-1}\sum_i m_i^0 = x^0$ (default: $m_i^0 = x^0$).
3: **for** $r = 1, \dots, R$ **do**
4: Observe the available set $\mathcal{S}^r$; send x^{r-1} to its clients.
5: **for all** $i \in \mathcal{S}^r$ in parallel **do**
6: $y_i^{r,0} \leftarrow x^{r-1}$; $d_i^{r-1} \leftarrow (m_i^{r-1} - x^{r-1})/h$ ▷ receive x^{r-1} and compute correction d_i^{r-1}
7: **for** $k = 0, \dots, K-1$ **do**
8: $y_i^{r,k+1} \leftarrow y_i^{r,k} - \eta_\ell\big(g_i(y_i^{r,k}; \xi_i^{r,k}) + d_i^{r-1}\big)$ ▷ apply d_i^{r-1} in SGD
9: **end for**
10: $\Delta y_i^r \leftarrow y_i^{r,K} - x^{r-1}$.
11: $m_i^r \leftarrow m_i^{r-1} + \eta_g \Delta y_i^r$; send Δy_i^r to the server ▷ refresh m_i^r and send Δy_i^r
12: **end for**
13: Inactive clients retain $m_i^r \leftarrow m_i^{r-1}$.
14: Server: $x^r \leftarrow x^{r-1} + \frac{\eta_g}{N}\sum_{i\in\mathcal{S}^r} \Delta y_i^r$.
15: **end for**
16: **return** x^R.

Assumption 3 (Smoothness)**.** Each local objective f_i is differentiable and β-smooth:

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le \beta \|x - y\|, \qquad \forall x, y \in \mathbb{R}^d,\ i \in \{1, \dots, N\}.$$

4 TRUSS ALGORITHM

TRUSS is presented in Algorithm 1. Each client stores one persistent memory m_i (lines 1–2). Each active round communicates only x^{r-1} downstream (line 4) and Δy_i^r upstream (line 11), one vector in each direction; the memories remain local. On the client side, d is computed from the local memory and the received server model (line 6), and corrects the local gradient in SGD (line 8) to compensate for both objective heterogeneity and irregular participation. In the following section, we show how this correction shapes the model update throughout training.

Memory and model updates. Summing the K local steps with the fixed correction d_i^{r-1} gives

$$\Delta y_i^r = y_i^{r,K} - x^{r-1} = -K\eta_\ell\left(\overline{g}_i^r + \frac{m_i^{r-1} - x^{r-1}}{h}\right), \qquad \overline{g}_i^r = \frac{1}{K}\sum_{k=0}^{K-1} g_i(y_i^{r,k}; \xi_i^{r,k}), \tag{4.1}$$

where $h = K\eta_g\eta_\ell$ is the effective step size for the memory update. Here $\xi_i^{r,k}$ is the fresh sample for local step k; $\overline{g}_i^r$ averages the raw stochastic gradients computed by client i in round r. The implemented memory update is $m_i^r = m_i^{r-1} + \eta_g \Delta y_i^r$. We can rewrite this update as:

$$m_i^{r-1} + \eta_g \Delta y_i^r = m_i^{r-1} - h\overline{g}_i^r - (m_i^{r-1} - x^{r-1}) = x^{r-1} - h\overline{g}_i^r.$$

The server uses $x^r = x^{r-1} + \frac{\eta_g}{N}\sum_{i:I_i^r=1} \Delta y_i^r$. Here the normalization is by the total number of clients N. Its increment equals the average memory increment. Hence the initialization $x^0 = N^{-1}\sum_i m_i^0$, satisfied in particular by $m_i^0 = x^0$, preserves $x^r = N^{-1}\sum_i m_i^r$ at every round. The algorithm therefore has the equivalent form

$$m_i^r = \begin{cases} x^{r-1} - h\overline{g}_i^r, & I_i^r = 1, \\ m_i^{r-1}, & I_i^r = 0, \end{cases} \qquad x^r = \frac{1}{N}\sum_i m_i^r. \tag{4.2}$$

The identity $x^r = N^{-1}\sum_i m_i^r$ is preserved by the updates; it does not require an additional communication step. Neither implemented update requires the participation probabilities.

Refreshing the memory. Suppose client i returns in round r after last participating in round $s < r$. The model displacement moves the reference from x^{s-1} to the current server model x^{r-1}. The gradient term replaces the old contribution $-h\overline{g}_i^s$ with the new contribution $-h\overline{g}_i^r$. This refreshed memory remains fixed throughout the client's next inactive period. Only the combined vector is stored, not the reference and gradient separately. This identity explains memory maintenance; the following section connects it to the convergence analysis.

5 CORRECTION WITH UNIFIED MEMORY

The memory identity (4.2) shows how information from all clients is retained in the global model throughout training, but two questions remain: how does a client-specific correction recover the population gradient, and how can unequal refresh rates be reconciled with equal client weights in the objective? We first turn heterogeneity into tracking error, then account for its propagation across rounds. The final step connects this analytical compensation to descent at the actual server iterates.

5.1 HETEROGENEITY BECOMES A TRACKING PROBLEM

To identify the global direction, we separate the client-specific gradient component and introduce an analytical shifted memory:

$$b_i(x) = \nabla f_i(x) - \nabla f(x), \qquad q_i^r = m_i^r + h b_i(x^r), \qquad \frac{1}{N}\sum_i q_i^r = x^r. \tag{5.1}$$

Neither the clients nor the server compute b_i or q_i; the shift is only a change of variables for the analysis. Averaging q_i^r gives x^r because $\sum_i b_i(x^r) = 0$. We also write $\overline{g}_i^r = \nabla f_i(x^{r-1}) + u_i^r + \zeta_i^r$, where u_i^r averages the changes in ∇f_i along the local trajectory relative to $\nabla f_i(x^{r-1})$, and ζ_i^r averages the stochastic errors around those local gradients. Their precise definitions appear in (A.7). The corrected direction used in local SGD (line 8) then becomes

$$\overline{g}_i^r + \frac{m_i^{r-1} - x^{r-1}}{h} = \nabla f(x^{r-1}) + \frac{q_i^{r-1} - x^{r-1}}{h} + u_i^r + \zeta_i^r. \tag{5.2}$$

The mismatch $q_i^{r-1} - x^{r-1}$ measures how far the stored correction is from cancelling $b_i(x^{r-1})$: it is zero exactly when $(m_i^{r-1} - x^{r-1})/h = -b_i(x^{r-1})$. Equation (5.2) therefore exposes the population gradient plus a tracking error, local drift, and noise. To see how the client mismatches enter the server update, let $\mathcal{F}_{r-1}$ denote the history before round r and define $\bar{p}_r = N^{-1}\sum_i p_i^r$. Combining the server update with (5.2) gives

$$\mathbb{E}[x^r - x^{r-1} \mid \mathcal{F}_{r-1}] = -h\bar{p}_r \nabla f(x^{r-1}) - \frac{1}{N}\sum_i p_i^r (q_i^{r-1} - x^{r-1}) - \frac{h}{N}\sum_i p_i^r \mathbb{E}[u_i^r \mid \mathcal{F}_{r-1}]. \tag{5.3}$$

The stochastic-gradient noise ζ_i^r has zero conditional mean. Under uniform participation, $p_i^r = p_r$ for all clients, the second term vanishes because $\sum_i (q_i^{r-1} - x^{r-1}) = 0$. The server update therefore follows the true population gradient direction up to local drift and zero-mean noise, simplifying the convergence analysis. In general, however, unequal participation assigns different expected weights to the client mismatches, so the second term need not vanish even when the local drift is zero. To control this residual error, we first characterize the evolution of the shifted memories and then define backward participation weights and an associated weighted trajectory that account for unequal refresh rates.

Lemma A.1 in Appendix A shows that every shifted memory refreshes toward the *same* global descent target:

$$q_i^r = (1 - I_i^r) q_i^{r-1} + I_i^r \big(x^{r-1} - h\nabla f(x^{r-1})\big) + \varepsilon_i^r, \tag{5.4}$$

$$\varepsilon_i^r = h\big(b_i(x^r) - b_i(x^{r-1})\big) - h I_i^r (u_i^r + \zeta_i^r). \tag{5.5}$$

In (5.4), the first term represents retention and the second term represents refresh toward a common target. The remainder ε_i^r accounts for the change of coordinates and the local drift and noise. Note that, when a client is absent its actual memory m_i is fixed, but its analytical memory q_i can still change through $b_i(x^r) - b_i(x^{r-1})$. Smoothness bounds the change in heterogeneity by $\|b_i(x^r) - b_i(x^{r-1})\| \le 2\beta \|x^r - x^{r-1}\|$, without requiring a uniform bound on $\|b_i(x)\|$ itself.

5.2 BACKWARD PARTICIPATION WEIGHTS

We fix a terminal round R and construct an analytical weighted average $z^r = N^{-1}\sum_i w_i^r q_i^r$. We choose the weights backward from $w_i^R = 1$, so that the terminal average agrees with the actual model: $z^R = x^R$. The recursion in Lemma A.2 is

$$w_i^R = 1, \qquad \gamma_r = \frac{1}{N}\sum_i p_i^r w_i^r, \qquad w_i^{r-1} = (1 - p_i^r) w_i^r + \gamma_r. \tag{5.6}$$

The two contributions to w_i^{r-1} in (5.6) correspond to two terms in (5.4). The coefficient $(1 - p_i^r) w_i^r$ comes from the retention term $(1 - I_i^r) q_i^{r-1}$. The coefficient γ_r comes from the server-model term

$I_j^r x^{r-1}$ within each client's refresh term $I_j^r(x^{r-1} - h\nabla f(x^{r-1}))$, summed over clients j. Taking conditional expectations, weighting by w_j^r, and averaging these server-model contributions gives

$$\frac{1}{N}\sum_j p_j^r w_j^r x^{r-1} = \gamma_r x^{r-1} = \frac{1}{N}\sum_i \gamma_r q_i^{r-1}.$$

Thus every previous memory receives the same additional coefficient γ_r through the server mean, even when the refresh probabilities differ. Choosing w_i^{r-1} as the sum of these two coefficients makes the previous-memory contributions recombine into z^{r-1}:

$$\frac{1}{N}\sum_i (1-p_i^r) w_i^r q_i^{r-1} + \gamma_r x^{r-1} = \frac{1}{N}\sum_i w_i^{r-1} q_i^{r-1} = z^{r-1}.$$

This recombination is the participation cancellation. The full conditional increment, including the remainders, is

$$\mathbb{E}[z^r - z^{r-1} \mid \mathcal{F}_{r-1}] = -h\gamma_r \nabla f(x^{r-1}) + \bar{\varepsilon}_r, \qquad \bar{\varepsilon}_r = \frac{1}{N}\sum_i w_i^r \mathbb{E}[\varepsilon_i^r \mid \mathcal{F}_{r-1}]. \tag{5.7}$$

Lemma A.3 in Appendix A gives the complete calculation. Lemma A.2 also shows that $N^{-1}\sum_i w_i^r = 1$ and $\delta \le \gamma_r \le \delta_{\max}$. The leading direction in (5.7) is therefore the population gradient scaled by a positive scalar. Unequal availability still affects the effective step and the remainder bounds, but no separate participation-weighted tracking term remains in this identity. The weights and z, like the shifted memories q, are analysis-only quantities: they expose the compensation implemented by memory retention, rather than adding a probability-dependent correction to the algorithm. Both w and z depend on the chosen horizon R. If all clients have the same probability in each round, even if that probability changes over time, then $w_i^r = 1$ and $z^r = x^r$ throughout.

5.3 Descent of the server model

The leading direction in (5.7) is $\nabla f(x^{r-1})$, not $\nabla f(z^{r-1})$: the analytical trajectory z cannot be treated as ordinary gradient descent. Instead, we use its participation cancellation to prove progress at the actual model x. Let $f_{\inf} := \inf_{x\in\mathbb{R}^d} f(x)$. We define the descent potential

$$\Phi_r = \mathbb{E}[f(x^r) - f_{\inf} + \langle \nabla f(x^r), z^r - x^r\rangle]\,. \tag{5.8}$$

This is the expected first-order approximation of $f(z^r) - f_{\inf}$ around x^r. More importantly, the added inner product replaces the biased increment of x by the increment of z in the descent calculation. Using smoothness we get

$$\Phi_r - \Phi_{r-1} \le \mathbb{E}\langle \nabla f(x^{r-1}), z^r - z^{r-1}\rangle + \frac{\beta}{2}\mathbb{E}\left\|x^r - x^{r-1}\right\|^2 + \mathbb{E}\langle \nabla f(x^r) - \nabla f(x^{r-1}), z^r - x^r\rangle. \tag{5.9}$$

Substituting (5.7) into the first inner product produces the negative term $-h\gamma_r \mathbb{E}\left\|\nabla f(x^{r-1})\right\|^2$, plus remainder terms. This is the benefit of introducing z: the descent potential removes the participation-weighted mismatch from the leading term and exposes the global gradient at the actual server model x. The remaining terms still depend on participation and must be controlled. The second term in (5.9) is proportional to the squared server model update. The last term can be bounded using smoothness and Cauchy–Schwarz by

$$\left|\mathbb{E}\langle \nabla f(x^r) - \nabla f(x^{r-1}), z^r - x^r\rangle\right| \le \beta\sqrt{\mathbb{E}\left\|x^r - x^{r-1}\right\|^2}\sqrt{\mathbb{E}\left\|z^r - x^r\right\|^2}. \tag{5.10}$$

The server-model update, local-drift, and tracking estimates, combined in Lemma A.10, bound these two contributions with factors of h^2 multiplying gradient, tracking, and noise quantities. The leading term proportional to h now has the favorable sign for descent, so a sufficiently small step lets it dominate the added gradient terms. To control the memory errors, the proof adds a scaled tracking energy to Φ_r and uses the contraction due to memory refresh (Lemma A.11). Stored stochastic noise is bounded separately (Lemmas A.5–A.7) to preserve the variance reduction from averaging clients, giving the factor $1/N$ in its contribution.

Summing the resulting descent bound over rounds controls the average squared global gradient at the server iterates (Theorem A.1 in Appendix A). Since $z^R = x^R$, the added inner product also vanishes at the terminal round. Thus z enables a convergence guarantee for the actual model x without requiring the trajectory gap to vanish first.

6 CONVERGENCE GUARANTEE

Define the participation-imbalance parameters $\theta := \delta_{\max}/\delta - 1$ and $\Lambda := 3 + 6\sqrt{\theta} + 3\sqrt{\theta/\delta_{\max}}$, together with the initial objective gap and tracking energy

$$F_0 = \mathbb{E}[f(x^0) - f_{\inf}], \quad t_i^0 = \frac{x^0 - q_i^0}{h} = \frac{x^0 - m_i^0}{h} - b_i(x^0), \quad \mathcal{E}_0 = \frac{1}{N}\sum_i \mathbb{E}\left\|t_i^0\right\|^2,$$

where q_i^0 and b_i are defined in (5.1). In particular, $\mathcal{E}_0$ need not vanish: these quantities summarize the initialization rather than impose a uniform heterogeneity bound. We measure stationarity by $\overline{G}_R = R^{-1}\sum_{r=0}^{R-1}\mathbb{E}\left\|\nabla f(x^r)\right\|^2$.

Theorem 1 (TRUSS convergence). *Under Assumptions 1–3, run Algorithm 1 with step sizes satisfying $\eta_g \geq \delta$ and $0 < h = K\eta_g\eta_\ell \leq \frac{\delta^2}{200\beta\Lambda}$. Then, for every $R \geq 1$,*

$$\overline{G}_R \leq \frac{3F_0}{\delta h R} + \frac{2\beta h}{\delta R}\Big(\theta + \frac{16\Lambda}{\delta}\Big)\mathcal{E}_0 + \frac{4\beta h\Lambda\delta_{\max}\sigma^2}{N\delta K}\Big(1 + \frac{45\beta h\delta_{\max}N}{\Lambda\delta\eta_g^2}\Big). \tag{6.1}$$

The first two terms in (6.1) capture the effects of the initial objective gap and the initial memory-tracking error, respectively. The last term accounts for stochastic gradient noise, both directly and through its effect on the local trajectories. These contributions scale as h and h^2, respectively. For fixed problem parameters and finite $F_0, \mathcal{E}_0$, taking $h = \Theta(R^{-1/2})$ within the cap yields $\overline{G}_R = O(R^{-1/2})$ without requiring $\mathcal{E}_0 = 0$, as shown in the following corollary.

Corollary 1 (Convergence rate in communication rounds). *Under Assumptions 1–3 and the initialization conditions of Theorem 1, fix $R \geq 1$ and $K \geq 1$. Choose constant step sizes within the run as $\eta_g = \delta\sqrt{N}$ and $\eta_\ell = \min\{(\beta\delta\sqrt{\delta KR})^{-1}, \delta/(200\beta\Lambda K\sqrt{N})\}$. Then*

$$\overline{G}_R \leq \frac{3\beta F_0}{\sqrt{\delta NKR}} + \frac{8\Lambda\delta_{\max}\sigma^2}{\delta^{3/2}\sqrt{NKR}} + \frac{600\beta\Lambda F_0}{\delta^3 R} + \frac{\delta\theta + 16\Lambda}{100\Lambda}\frac{\mathcal{E}_0}{R}. \tag{6.2}$$

The objective-gap term has $1/\sqrt{\delta NKR}$ scaling, while participation imbalance enters the leading stochastic term through $\Lambda\delta_{\max}/\delta$. The proofs of Theorem 1 and Corollary 1 are in Appendix A.

Remark (Uniform participation). For $p_i^r = \delta$, we have $\delta_{\max} = \delta$, $\theta = 0$, and $\Lambda = 3$. Corollary 1 gives $\overline{G}_R = O((\delta NKR)^{-1/2} + R^{-1})$ for fixed δ. The leading term yields linear speedup in N and K when dominant, with δN clients active in expectation. TRUSS matches SCAFFOLD's leading stochastic scaling while halving its original uncompressed communication.

Remark (Trajectory closeness). Lemmas A.3 and A.13 bound the averaged squared gaps from the server model as $\frac{1}{NR}\sum_{r=0}^{R-1}\sum_i \mathbb{E}\left\|q_i^r - x^r\right\|^2 \lesssim h^2(\frac{\mathcal{E}_0}{\delta R} + \frac{\overline{G}_R}{\delta} + \frac{\sigma^2}{K})$ and $\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\|z^r - x^r\right\|^2 \leq \frac{\theta}{NR}\sum_{r=0}^{R-1}\sum_i \mathbb{E}\left\|q_i^r - x^r\right\|^2$. Here $\lesssim$ hides numerical constants. $h_R = \Theta(R^{-1/2})$ yields $O(R^{-1})$ averaged squared gaps decay.

7 EXPERIMENTS

We evaluate TRUSS through experiments on a real distributed cluster with 10 clients and one server on the National Research Platform (NRP) (Weitzel et al., 2025). The deployment spans 11 nodes across different U.S. states. Each client is allocated one GPU, two CPU cores, and 5 GiB of memory.

We train ResNet-20 on CIFAR-10, ResNet-56 on CIFAR-100, and a convolutional network on SVHN. We use a Dirichlet distribution to create disjoint non-i.i.d. client partitions (Hsu et al., 2019). The degree of data heterogeneity is controlled by the concentration parameter α: the smaller α is, the more likely each client holds examples from only one class. Unless stated otherwise we set $\alpha = 0.01$ (see Appendix B.2 for partition visualizations and details). We run each experiment 3 times and report the mean and one-standard-deviation over random seeds. Each experiment uses a fine-tuned local and global learning rate per algorithm.

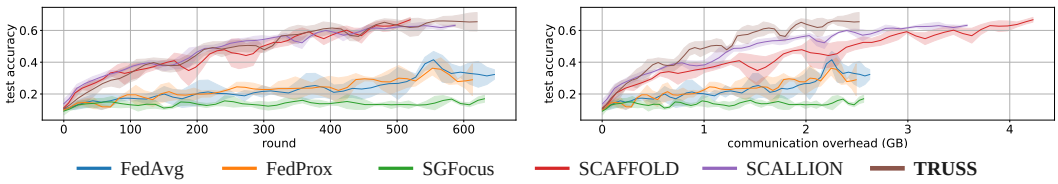
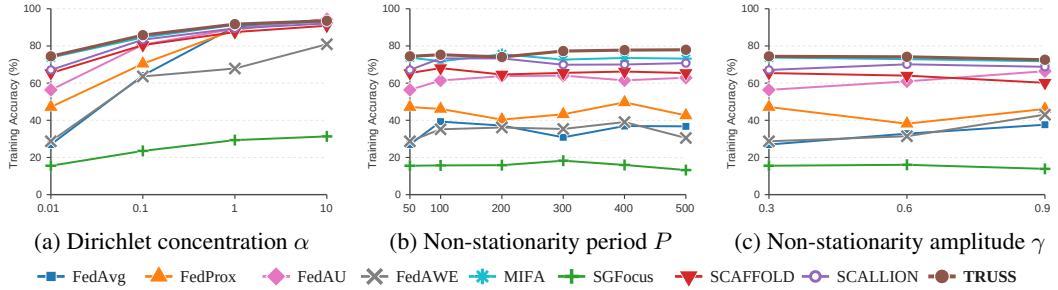


Figure 1: CIFAR-10/ResNet-20 test accuracy vs. rounds (left) and communication GB (right).



(a) Dirichlet concentration α (b) Non-stationarity period P (c) Non-stationarity amplitude γ

Figure 2: CIFAR-10/ResNet-20 training accuracy under exponential availability, varying (a) Dirichlet concentration α (log scale), (b) period P (rounds), and (c) amplitude γ.

7.1 Communication Efficiency

We compare TRUSS with FedAvg, FedProx, SGFocus, SCAFFOLD, and SCALLION on CIFAR-10 with ResNet-20, reporting test accuracy against communication rounds and cumulative communication cost (Figure 1). Over rounds, TRUSS matches SCAFFOLD and SCALLION, methods that correct client drift via variance reduction, and both outperform FedAvg, FedProx, and SGFocus. Measured by communicated GB, however, TRUSS reaches the same accuracy at a lower communication cost: each active client exchanges only a model and a displacement, while the heterogeneity correction is reconstructed locally from memory rather than sent as a separate control vector. Thus TRUSS delivers the optimization benefit of full heterogeneity correction with substantially less communication.

7.2 Non-stationary client availability

We next study training under diverse client availability regimes. We evaluate five participation scenarios: **Uniform**, where all clients share a constant participation probability $p_i^t = \delta$; **Stationary**, where baseline rates p_i vary across clients but remain constant over rounds ($f_i(t) = 1$); and three non-stationary patterns where client i participates in round t independently with time-varying probability $p_i^t = p_i\, f_i(t)$, governed by baseline rate p_i, fluctuation amplitude $\gamma \in [0, 1]$, period P, and client-specific phase ϕ_i: **Staircase** (f_i alternates between $1+\gamma$ and $1-\gamma$ over successive half-periods), **Sine** ($f_i(t) = 1 + \gamma \sin(2\pi t/P + \phi_i)$), and **Exponential** ($f_i$ follows an asymmetric decay and recovery within each period shifted by phase ϕ_i). A complete mathematical formulation and visualizations of the participation scenarios under $P = 50$ and $\gamma = 0.3$ (Figure 3) are provided in Appendix B.1.

Table 2 compares TRUSS with FedAvg, FedProx, FedAU, FedAWE, MIFA, SGFocus, SCAFFOLD, and SCALLION across these patterns and three dataset–model pairs. TRUSS achieves the highest or second-highest mean training and test accuracy in all 15 dataset–participation settings. We also observe that MIFA's strong performance is consistent with its reuse of each client's latest accumulated local gradient, which keeps inactive clients represented in server updates (Gu et al., 2021). In the server-memory formulation listed in Table 1, this requires N model-sized update vectors in addition to the global model, so server storage grows with both the client count and model size. TRUSS achieves consistently competitive accuracy while retaining only the global model at the server and one persistent model-sized vector per client.

7.3 Ablation study

Finally, in Figure 2, we ablate the sources of imbalance studied in this paper: data heterogeneity and non-stationary availability. Holding the training protocol fixed, we vary the Dirichlet concentration α to change the severity of non-i.i.d. partitions, and we vary the non-stationarity amplitude γ and

Table 2: Training and test accuracy (%) across client availability patterns (mean ± standard deviation). Each algorithm is run for a communication budget of 1000 × model-size. The metric for accuracy of the last 25 rounds is reported. Bold and underlined mark the best and second-best.

Participation	Algorithm	CIFAR-10 (ResNet-20)		CIFAR-100 (ResNet-56)		SVHN (CNN)	
		Train	Test	Train	Test	Train	Test
Uniform	**TRUSS**	**76.8** ± 0.9	**75.8** ± 0.3	**52.5** ± 0.4	**49.2** ± 0.2	84.7 ± 1.3	84.3 ± 0.8
	FedAvg	35.3 ± 2.2	34.5 ± 0.6	25.6 ± 0.7	23.4 ± 0.8	55.5 ± 4.6	54.9 ± 3.9
	FedProx	33.2 ± 3.4	33.8 ± 3.1	30.8 ± 2.6	27.6 ± 3.1	61.9 ± 0.5	61.3 ± 1.7
	FedAU	42.8 ± 2.1	42.5 ± 1.6	40.0 ± 1.2	36.4 ± 0.7	82.9 ± 1.6	81.0 ± 2.1
	FedAWE	19.0 ± 0.0	19.4 ± 0.1	4.50 ± 0.73	4.36 ± 0.63	36.1 ± 3.2	35.7 ± 4.5
	MIFA[a]	67.0 ± 2.6	65.1 ± 2.2	46.1 ± 1.0	40.1 ± 0.9	**85.8** ± 1.2	**85.7** ± 0.9
	SGFocus	13.5 ± 0.9	13.8 ± 0.4	2.28 ± 0.31	2.40 ± 0.47	9.15 ± 2.18	8.88 ± 2.18
	SCAFFOLD	57.5 ± 6.8	59.2 ± 5.1	36.9 ± 0.3	35.8 ± 1.9	79.7 ± 0.8	78.3 ± 0.2
	SCALLION	67.2 ± 1.8	67.1 ± 2.1	41.2 ± 2.2	40.3 ± 1.6	84.4 ± 0.4	82.6 ± 0.1
Stationary	**TRUSS**	**76.1** ± 0.7	**75.4** ± 1.0	**53.4** ± 0.1	**51.5** ± 0.0	85.5 ± 1.0	83.7 ± 0.0
	FedAvg	37.6 ± 1.3	38.0 ± 1.4	24.4 ± 3.6	21.9 ± 3.5	50.6 ± 3.9	49.8 ± 1.6
	FedProx	51.2 ± 7.8	50.4 ± 7.0	31.0 ± 3.3	28.1 ± 3.3	59.6 ± 9.0	59.5 ± 7.7
	FedAU	58.2 ± 5.2	57.2 ± 4.0	49.1 ± 0.0	43.9 ± 1.0	84.5 ± 0.3	83.5 ± 0.5
	FedAWE	19.1 ± 4.5	20.0 ± 4.9	16.4 ± 2.2	15.4 ± 2.1	58.5 ± 0.9	57.4 ± 0.2
	MIFA[a]	73.4 ± 1.0	71.4 ± 0.8	45.4 ± 0.0	40.9 ± 1.0	**88.3** ± 0.9	**87.1** ± 1.0
	SGFocus	17.8 ± 0.2	18.0 ± 0.1	2.51 ± 0.52	2.39 ± 0.46	17.8 ± 2.4	17.8 ± 1.8
	SCAFFOLD	64.0 ± 0.5	63.3 ± 1.0	32.6 ± 4.1	32.2 ± 4.3	78.8 ± 3.5	78.7 ± 2.0
	SCALLION	73.6 ± 0.4	72.7 ± 0.9	41.1 ± 1.7	41.0 ± 1.7	83.6 ± 0.6	82.3 ± 0.2
Staircase (Non-stationary)	**TRUSS**	74.4 ± 3.1	74.0 ± 3.1	**51.1** ± 0.8	**49.2** ± 0.3	84.8 ± 0.9	84.5 ± 0.1
	FedAvg	41.3 ± 4.0	42.3 ± 4.6	31.8 ± 5.5	28.8 ± 5.1	51.3 ± 0.7	50.8 ± 2.3
	FedProx	38.5 ± 0.2	37.5 ± 0.6	35.6 ± 0.6	32.1 ± 0.4	46.4 ± 0.4	46.3 ± 1.9
	FedAU	54.2 ± 7.1	53.8 ± 6.9	47.7 ± 0.8	42.4 ± 0.2	84.0 ± 0.4	82.6 ± 0.2
	FedAWE	36.5 ± 2.6	38.1 ± 2.8	19.3 ± 0.7	18.3 ± 0.3	59.1 ± 3.7	58.8 ± 2.8
	MIFA[a]	**77.7** ± 1.6	**75.8** ± 0.8	49.4 ± 1.0	43.1 ± 1.1	**86.9** ± 1.1	**86.7** ± 0.6
	SGFocus	12.4 ± 3.3	12.2 ± 3.6	2.68 ± 0.22	2.63 ± 0.24	13.8 ± 5.5	13.6 ± 6.0
	SCAFFOLD	62.2 ± 2.3	62.5 ± 2.3	33.7 ± 0.2	33.6 ± 1.0	77.8 ± 2.4	76.8 ± 1.1
	SCALLION	63.8 ± 3.4	64.2 ± 3.3	43.9 ± 3.3	42.8 ± 2.9	78.7 ± 0.6	77.8 ± 1.2
Sine (Non-stationary)	**TRUSS**	**76.2** ± 3.6	**75.0** ± 3.9	**50.3** ± 0.5	**48.5** ± 0.1	84.6 ± 2.0	83.0 ± 1.5
	FedAvg	25.5 ± 0.4	25.5 ± 0.1	22.8 ± 5.2	21.8 ± 5.3	45.6 ± 4.1	45.1 ± 4.9
	FedProx	41.8 ± 6.6	41.9 ± 6.3	34.8 ± 0.8	31.7 ± 0.6	46.8 ± 2.7	46.4 ± 3.9
	FedAU	52.4 ± 2.4	52.2 ± 2.4	44.6 ± 0.8	40.9 ± 0.5	83.9 ± 0.0	83.0 ± 0.2
	FedAWE	32.1 ± 8.9	32.3 ± 8.9	14.6 ± 3.2	14.4 ± 3.5	74.6 ± 1.1	73.2 ± 1.1
	MIFA[a]	73.4 ± 0.8	71.4 ± 1.3	45.2 ± 1.1	40.5 ± 0.7	**87.7** ± 0.4	**86.2** ± 0.0
	SGFocus	12.6 ± 0.9	12.9 ± 0.8	2.86 ± 0.22	3.06 ± 0.25	14.1 ± 5.2	14.0 ± 5.6
	SCAFFOLD	61.8 ± 4.7	60.2 ± 5.7	27.7 ± 4.8	27.0 ± 6.2	73.4 ± 6.0	72.1 ± 3.1
	SCALLION	67.8 ± 0.9	68.1 ± 0.6	40.1 ± 0.0	38.7 ± 0.1	81.8 ± 3.2	81.2 ± 1.8
Exponential (Non-stationary)	**TRUSS**	**74.4** ± 2.9	**73.6** ± 2.9	**51.5** ± 1.6	**49.4** ± 0.9	86.8 ± 0.1	85.3 ± 0.0
	FedAvg	27.0 ± 5.2	27.6 ± 5.7	27.5 ± 7.6	25.4 ± 7.5	53.3 ± 16.5	53.1 ± 18.4
	FedProx	47.2 ± 3.2	46.3 ± 2.9	31.6 ± 9.2	28.6 ± 8.8	49.5 ± 7.2	49.3 ± 7.2
	FedAU	56.4 ± 3.7	55.3 ± 3.3	46.6 ± 0.5	42.1 ± 1.8	84.4 ± 0.7	82.9 ± 1.5
	FedAWE	28.7 ± 4.8	28.8 ± 5.1	16.2 ± 7.4	15.4 ± 7.2	51.8 ± 9.1	50.9 ± 11.5
	MIFA[a]	73.8 ± 0.9	72.5 ± 0.4	46.3 ± 0.8	41.8 ± 0.3	**88.4** ± 0.7	**87.3** ± 0.8
	SGFocus	15.6 ± 1.4	16.8 ± 1.9	2.70 ± 0.32	2.64 ± 0.25	14.8 ± 0.9	14.6 ± 1.2
	SCAFFOLD	65.4 ± 3.5	65.3 ± 3.1	33.9 ± 2.0	33.1 ± 1.7	80.0 ± 1.9	79.7 ± 1.5
	SCALLION	67.1 ± 6.2	67.5 ± 5.4	42.0 ± 1.8	41.4 ± 1.9	83.4 ± 0.4	81.7 ± 0.2

[a]MIFA requires N model-sized update vectors in addition to the global model at the server (Gu et al., 2021).

period P in Section 7.2. Figure 2 illustrates the training accuracy of all algorithms across these sweeps, with complete numerical values reported in Appendix C (Tables 5–7). As shown in Figure 2(a), TRUSS demonstrates robust convergence under severe label skew ($\alpha = 0.01$), sustaining over 74% training accuracy where baseline methods degrade significantly. Across varying oscillation periods P (Figure 2(b)) and fluctuation amplitudes γ (Figure 2(c)), TRUSS maintains stable and competitive training performance under fluctuating temporal participation.

8 Conclusion

We showed that a single persistent state per client suffices to correct client drift and to compensate for unequal, time-varying availability at once. In TRUSS, the gap between that memory and the current server model supplies the local correction and records whatever movement a client missed, so an active client exchanges one model-sized vector in each direction and the server stores only the global model. For smooth non-convex objectives under independent, non-stationary participation, this yields an $O((NKR)^{-1/2}+R^{-1})$ stationarity bound with no gradient-heterogeneity assumption. Relaxing the independence and $\delta > 0$ requirements to correlated or vanishing availability, and combining the single uplink vector with compression or secure aggregation, are natural next steps.

Reproducibility Statement

The theoretical framework, underlying assumptions, and complete proofs for Theorem 1 and Corollary 1 are provided self-contained in Appendix A. The algorithmic procedure for TRUSS is formally defined in Algorithm 1, with explicit initialization conditions and step-size schedules. Details regarding dataset partitioning, network architectures, benchmark baselines, simulation setups, and hyperparameter selections are provided in Section 7 and Appendix C. Code and configuration scripts to reproduce all empirical benchmarks and ablation sweeps are available at https://anonymous.4open.science/r/Truss/README.md.

AI Use Statement

In this work, generative AI tools were used for assisting in manuscript editing to improve clarity and readability, formatting numerical tables in LaTeX, and generating Python plotting scripts for vector visualization. Generative AI tools were not used for formulating mathematical claims, deriving proofs, establishing conceptual frameworks, or fabricating experimental data. All theoretical derivations, algorithmic designs, and empirical findings were conducted, verified, and validated by the authors. The authors take full responsibility for the entire content of this work, including all text, mathematical arguments, and artifacts.

References

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. https://proceedings.neurips.cc/paper_files/paper/2024/hash/0f98645119923217a245735c2c4d23f4-Abstract-Conference.html.

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, pages 12052–12064, 2021. https://proceedings.neurips.cc/paper/2021/hash/64be20f6dd1dd46adf110cf871e3ed35-Abstract.html.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019. https://arxiv.org/abs/1909.06335.

Xinmeng Huang, Ping Li, and Xiaoyun Li. Stochastic controlled averaging for federated learning with communication compression. In *International Conference on Learning Representations*, 2024. https://openreview.net/forum?id=jj5ZjZsWJe.

Divyansh Jhunjhunwala, Pranay Sharma, Aushim Nagarkatti, and Gauri Joshi. FedVARP: Tackling the variance due to partial client participation in federated learning. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180, pages 906–916, 2022. https://proceedings.mlr.press/v180/jhunjhunwala22a.html.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 5132–5143, 2020. https://proceedings.mlr.press/v119/karimireddy20a.html.

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, pages 429–450, 2020a. https://proceedings.mlsys.org/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html.

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. https://openreview.net/forum?id=HJxNAnVtDS.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 1273–1282, 2017. https://proceedings.mlr.press/v54/mcmahan17a.html.

Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. ProxSkip: Yes! Local gradient steps provably lead to communication acceleration! Finally! In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 15750–15769, 2022. https://proceedings.mlr.press/v162/mishchenko22b.html.

Fakhteh Saadatniaki, Ran Xin, and Usman A. Khan. Decentralized optimization over time-varying directed graphs with row and column-stochastic matrices. *IEEE Transactions on Automatic Control*, 65(11):4769–4780, 2020. https://arxiv.org/abs/1810.07393.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019. https://openreview.net/forum?id=S1g2JnRcFX.

Behrouz Touri and Angelia Nedić. Product of random stochastic matrices. *IEEE Transactions on Automatic Control*, 59(2):437–448, 2014. https://doi.org/10.1109/TAC.2013.2283750.

Shiqiang Wang and Mingyue Ji. A unified analysis of federated learning with arbitrary client participation. In *Advances in Neural Information Processing Systems*, volume 35, pages 19124–19137, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/hash/79ba1b827d3fc58e129d1cbfc8ff69f2-Abstract-Conference.html.

Shiqiang Wang and Mingyue Ji. A lightweight method for tackling unknown participation statistics in federated averaging. In *International Conference on Learning Representations*, 2024. https://openreview.net/forum?id=ZKEuFKfCKA.

Derek Weitzel, Ashton Graves, Sam Albin, Huijun Zhu, Frank Wuerthwein, Mahidhar Tatineni, Dmitry Mishin, Elham Khoda, Mohammad Sada, Larry Smarr, Thomas DeFanti, and John Graham. The National Research Platform: Stretched, multi-tenant, scientific Kubernetes cluster. In *Practice and Experience in Advanced Research Computing 2025: The Power of Collaboration*, article 69. ACM, 2025. https://doi.org/10.1145/3708035.3736060.

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, pages 104281–104328, 2024. https://proceedings.neurips.cc/paper_files/paper/2024/hash/bcaebb606df0c8c714756660e4a3bd7c-Abstract-Conference.html.

Ming Xiang, Stratis Ioannidis, Edmund Yeh, Carlee Joe-Wong, and Lili Su. Resilience beyond stationary client unavailability: Unlocking efficient and unbiased federated learning. *Journal of Machine Learning Research*, 27(173):1–73, 2026. https://jmlr.org/papers/v27/25-1020.html.

Wenjing Yan, Kai Zhang, Xiaolu Wang, and Xuanyu Cao. Problem-parameter-free federated learning. In *International Conference on Learning Representations*, 2025. https://openreview.net/forum?id=ZuazHmXTns.

Bicheng Ying, Zhe Li, and Haibo Yang. Exact and linear convergence for federated learning under arbitrary client participation is attainable. In *Advances in Neural Information Processing Systems*, volume 38, 2025. https://proceedings.neurips.cc/paper_files/paper/2025/hash/395409679270591fd2a70abc694cf5a1-Abstract-Conference.html.

Rong You and Shi Pu. FedSUM family: Efficient federated learning methods under arbitrary client participation. *arXiv preprint arXiv:2512.18275*, 2025. https://arxiv.org/abs/2512.18275.

Xiaotong Yuan and Ping Li. On convergence of FedProx: Local dissimilarity invariant bounds, non-smoothness and beyond. In *Advances in Neural Information Processing Systems*, volume 35, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/hash/45ecdd6cf1f507d378a3442ed89e580b-Abstract-Conference.html.

A CONVERGENCE ANALYSIS

We give a self-contained proof. The argument proceeds from a common refresh target, to a backward-weighted trajectory, and then to a descent inequality coupled with memory tracking. All expectations below include both participation and oracle randomness.

A.1 SETUP AND EQUIVALENT UPDATES

Objective and sampling. There are N clients and

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x), \qquad f_i(x) = \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[\ell(x;\xi_i)], \qquad f_{\inf} = \inf_{x\in\mathbb{R}^d} f(x) > -\infty.$$

Each f_i has a β-Lipschitz gradient, with $\beta > 0$. The active set is $\mathcal{S}^r = \{i : I_i^r = 1\}$, where the Bernoulli indicators are independent across clients and rounds and satisfy

$$\mathbb{P}(I_i^r = 1) = p_i^r, \qquad 0 < \delta \le p_i^r \le \delta_{\max} \le 1.$$

The probabilities are deterministic and may vary with i and r. Each oracle call draws a fresh sample independently across clients and calls and independently of participation. With $g_i(y;\xi) = \nabla_y \ell(y;\xi)$, the oracle satisfies

$$\mathbb{E}[g_i(y;\xi) \mid y] = \nabla f_i(y), \qquad \mathbb{E}[\|g_i(y;\xi) - \nabla f_i(y)\|^2 \mid y] \le \sigma^2.$$

Thus these bounds also hold conditional on the history before a fresh draw, since the query is measurable with respect to that history. No bound on $\|\nabla f_i(x) - \nabla f(x)\|$ is assumed.

Steps and initialization. Define

$$\theta = \frac{\delta_{\max}}{\delta} - 1, \qquad \Lambda = 3 + 6\sqrt{\theta} + 3\sqrt{\frac{\theta}{\delta_{\max}}}.$$

Fix $K \ge 1$, a horizon $R \ge 1$, and positive constant step sizes with

$$h = K\eta_g\eta_\ell, \qquad \eta_g \ge \delta, \qquad 0 < h \le \frac{\delta^2}{200\beta\Lambda}. \tag{A.1}$$

Since $\Lambda \ge 3$, this also gives $\beta h \le \delta^2/600$. We will also use

$$\Lambda \ge 3 + 6\sqrt{\theta} \ge 3\sqrt{1+\theta}, \qquad \Lambda^2 \ge \frac{9\delta_{\max}}{\delta}.$$

Initialize with $N^{-1}\sum_i m_i^0 = x^0$, independently of future randomness, and finite initial energies defined below. Let $\mathcal{F}_{r-1}$ contain all randomness revealed before round r.

Local and server updates. For analysis, generate a local path for every client, independently of the current indicators conditional on $\mathcal{F}_{r-1}$:

$$\begin{aligned} y_i^{r,0} &= x^{r-1}, \qquad d_i^{r-1} = \frac{m_i^{r-1} - x^{r-1}}{h}, \\ y_i^{r,k+1} &= y_i^{r,k} - \eta_\ell\big(g_i(y_i^{r,k};\xi_i^{r,k}) + d_i^{r-1}\big), \quad 0 \le k < K, \\ \overline{g}_i^r &= \frac{1}{K}\sum_{k=0}^{K-1} g_i(y_i^{r,k};\xi_i^{r,k}), \qquad \Delta y_i^r = y_i^{r,K} - x^{r-1}. \end{aligned} \tag{A.2}$$

Only active clients compute these paths in the algorithm. The implemented updates are

$$m_i^r = m_i^{r-1} + \eta_g I_i^r \Delta y_i^r, \qquad x^r = x^{r-1} + \frac{\eta_g}{N}\sum_i I_i^r \Delta y_i^r. \tag{A.3}$$

Their increments satisfy

$$\frac{1}{N}\sum_i (m_i^r - m_i^{r-1}) = \frac{\eta_g}{N}\sum_i I_i^r \Delta y_i^r = x^r - x^{r-1}.$$

Induction from the initialization gives $N^{-1}\sum_i m_i^r = x^r$. Summing the local steps in (A.2) also gives

$$\Delta y_i^r = -K\eta_\ell(\overline{g}_i^r + d_i^{r-1}), \qquad m_i^{r-1} + \eta_g \Delta y_i^r = m_i^{r-1} - h\overline{g}_i^r - (m_i^{r-1} - x^{r-1}).$$

Consequently, the equivalent memory refresh is

$$m_i^r = (1 - I_i^r)m_i^{r-1} + I_i^r(x^{r-1} - h\overline{g}_i^r), \qquad x^r = \frac{1}{N}\sum_i m_i^r. \tag{A.4}$$

A.2 Shifted memories and a common refresh target

Define the shifted memories by

$$b_i(x) = \nabla f_i(x) - \nabla f(x), \qquad q_i^r = m_i^r + hb_i(x^r), \qquad \frac{1}{N}\sum_i q_i^r = x^r.$$

We measure their mismatch with the actual model in gradient units:

$$t_i^r = \frac{x^r - q_i^r}{h} = \frac{x^r - m_i^r}{h} - b_i(x^r), \qquad \mathcal{E}_r = \frac{1}{N}\sum_i \mathbb{E}\left\|t_i^r\right\|^2. \tag{A.5}$$

The memory average and $\sum_i b_i(x) = 0$ give

$$\sum_i t_i^r = 0. \tag{A.6}$$

With $x = x^{r-1}$, decompose

$$\begin{aligned} \overline{g}_i^r &= \nabla f_i(x) + u_i^r + \zeta_i^r, \\ u_i^r &= \frac{1}{K}\sum_{k=0}^{K-1}\big(\nabla f_i(y_i^{r,k}) - \nabla f_i(x)\big), \\ \zeta_i^r &= \frac{1}{K}\sum_{k=0}^{K-1}\big(g_i(y_i^{r,k};\xi_i^{r,k}) - \nabla f_i(y_i^{r,k})\big). \end{aligned} \tag{A.7}$$

The corresponding error quantities are

$$U_r = \frac{1}{N}\sum_i \mathbb{E}\left\|u_i^r\right\|^2, \qquad D_r = \mathbb{E}\left\|x^r - x^{r-1}\right\|^2, \qquad G_r = \mathbb{E}\left\|\nabla f(x^r)\right\|^2.$$

$$F_r = \mathbb{E}[f(x^r) - f_{\inf}].$$

We also write $\overline{G}_R = R^{-1}\sum_{r=0}^{R-1} G_r$. At initialization,

$$F_0 = \mathbb{E}[f(x^0) - f_{\inf}], \qquad \mathcal{E}_0 = \frac{1}{N}\sum_i \mathbb{E}\left\|\frac{x^0 - m_i^0}{h} - b_i(x^0)\right\|^2.$$

The refresh remainder will be denoted by

$$\varepsilon_i^r = h\big(b_i(x^r) - b_i(x^{r-1})\big) - hI_i^r(u_i^r + \zeta_i^r).$$

Lemma A.1 (A common refresh target)**.** *For every client and round, the shifted memory satisfies*

$$q_i^r = (1 - I_i^r)q_i^{r-1} + I_i^r\big(x^{r-1} - h\nabla f(x^{r-1})\big) + \varepsilon_i^r. \tag{A.8}$$

Proof. Write $x = x^{r-1}$. First, the corrected local direction is

$$\begin{aligned} \overline{g}_i^r + \frac{m_i^{r-1} - x}{h} &= \nabla f_i(x) + u_i^r + \zeta_i^r - t_i^{r-1} - b_i(x) \\ &= \nabla f(x) - t_i^{r-1} + u_i^r + \zeta_i^r. \end{aligned} \tag{A.9}$$

Next, substitute (A.4) into $q_i^r = m_i^r + hb_i(x^r)$:

$$\begin{aligned} q_i^r &= (1 - I_i^r)(q_i^{r-1} - hb_i(x)) + I_i^r(x - h\overline{g}_i^r) + hb_i(x^r) \\ &= (1 - I_i^r)q_i^{r-1} + I_i^r\big(x - h(\overline{g}_i^r - b_i(x))\big) + h(b_i(x^r) - b_i(x)) \\ &= (1 - I_i^r)q_i^{r-1} + I_i^r(x - h\nabla f(x)) - hI_i^r(u_i^r + \zeta_i^r) + h(b_i(x^r) - b_i(x)). \end{aligned}$$

The last equality uses $\overline{g}_i^r - b_i(x) = \nabla f(x) + u_i^r + \zeta_i^r$ and proves the claim. Finally, the average objective is β-smooth, so

$$\begin{aligned} \|b_i(x^r) - b_i(x)\| &\le \|\nabla f_i(x^r) - \nabla f_i(x)\| + \|\nabla f(x^r) - \nabla f(x)\| \\ &\le 2\beta\|x^r - x\|. \end{aligned}$$

□

Thus all shifted memories refresh toward the same global-gradient target. Imperfect tracking, local drift, and oracle noise remain; none is discarded or assumed to vanish.

A.3 Backward weights: exposing the global-gradient direction

Fix a horizon $R \ge 1$. Starting from $w_i^R = 1$, define backward

$$\gamma_r = \frac{1}{N}\sum_i p_i^r w_i^r, \qquad w_i^{r-1} = \gamma_r + (1 - p_i^r)w_i^r, \qquad r = R, \dots, 1. \tag{A.10}$$

Let $v \in \mathbb{R}^{N\times d}$ be a client matrix whose ith row is $v_i^\top$, with $v_i \in \mathbb{R}^d$. Define the averaging discrepancy

$$L_r(v) = \frac{1}{N}\sum_i (1 - w_i^r)v_i = \frac{1}{N}\sum_i v_i - \frac{1}{N}\sum_i w_i^r v_i.$$

Thus L_r is the difference between uniform and backward-weighted averaging. Left multiplication by $C = I - \mathbf{1}\mathbf{1}^\top/N$ subtracts the average row:

$$(Cv)_{i,:} = v_i^\top - \frac{1}{N}\sum_j v_j^\top.$$

We call v zero-sum when $\sum_i v_i = 0$. Throughout the appendix, client-indexed collections are stacked as rows using this convention. The next lemma controls the weights. The weighted-trajectory lemma then uses their recursion to cancel unequal memory retention.

Lemma A.2 (Backward-weight bounds)**.** *The weights satisfy the following bounds, with w^r indexed by $0 \le r \le R$ and γ_r by $1 \le r \le R$:*

$$\frac{1}{N}\sum_i w_i^r = 1, \qquad 0 < w_i^r \le \frac{\delta_{\max}}{\delta}, \qquad \delta \le \gamma_r \le \delta_{\max}. \tag{A.11}$$

Their squared deviation from uniform weights satisfies

$$\frac{1}{N}\sum_i (1 - w_i^r)^2 \le \theta. \tag{A.12}$$

Proof. Positivity and unit average. At the terminal round, $w_i^R = 1$, so the weights are positive and their average is one. Suppose these properties hold at round r. Because $\delta \le p_i^r \le \delta_{\max}$,

$$\delta = \frac{\delta}{N}\sum_i w_i^r \le \gamma_r = \frac{1}{N}\sum_i p_i^r w_i^r \le \frac{\delta_{\max}}{N}\sum_i w_i^r = \delta_{\max}.$$

The backward update therefore gives $w_i^{r-1} = \gamma_r + (1 - p_i^r)w_i^r \ge \gamma_r > 0$. Averaging that update gives

$$\frac{1}{N}\sum_i w_i^{r-1} = \gamma_r + \frac{1}{N}\sum_i w_i^r - \frac{1}{N}\sum_i p_i^r w_i^r = \gamma_r + 1 - \gamma_r = 1.$$

Backward induction establishes positivity and unit average at every round, as well as the claimed bounds on γ_r.

A uniform bound on each weight. The terminal weights satisfy $w_i^R = 1 \le \delta_{\max}/\delta$. If $w_i^r \le \delta_{\max}/\delta$, the bound $\gamma_r \le \delta_{\max}$ and $p_i^r \ge \delta$ give

$$w_i^{r-1} \le \delta_{\max} + (1-\delta)\frac{\delta_{\max}}{\delta} = \frac{\delta_{\max}}{\delta}.$$

Induction proves the upper bound at every round and completes (A.11).

The squared deviation from uniform weights. Expanding the square and using unit average yields

$$\frac{1}{N}\sum_i (1-w_i^r)^2 = 1 - \frac{2}{N}\sum_i w_i^r + \frac{1}{N}\sum_i (w_i^r)^2 = \frac{1}{N}\sum_i (w_i^r)^2 - 1.$$

By the uniform weight bound, $(w_i^r)^2 \le (\delta_{\max}/\delta) w_i^r$. Consequently,

$$\frac{1}{N}\sum_i (1-w_i^r)^2 \le \frac{\delta_{\max}}{\delta}\frac{1}{N}\sum_i w_i^r - 1 = \frac{\delta_{\max}}{\delta} - 1 = \theta,$$

which proves (A.12). □

Define the analytical trajectory

$$z^r = \frac{1}{N}\sum_i w_i^r q_i^r, \qquad 0 \le r \le R.$$

Both the weights and z depend on the fixed horizon R; neither is computed by the algorithm. Define the conditional weighted remainder

$$\bar{\varepsilon}_r = -\frac{h}{N}\sum_i p_i^r w_i^r \mathbb{E}[u_i^r \mid \mathcal{F}_{r-1}] + \frac{h}{N}\sum_i w_i^r \mathbb{E}[b_i(x^r) - b_i(x^{r-1}) \mid \mathcal{F}_{r-1}]. \tag{A.13}$$

Lemma A.3 (Weighted trajectory dynamics)**.** *The trajectory satisfies $z^R = x^R$ and*

$$\mathbb{E}[z^r - z^{r-1} \mid \mathcal{F}_{r-1}] = -h\gamma_r \nabla f(x^{r-1}) + \bar{\varepsilon}_r. \tag{A.14}$$

Moreover,

$$z^r - x^r = hL_r(t^r), \qquad \mathbb{E}\|z^r - x^r\|^2 \le \theta h^2 \mathcal{E}_r. \tag{A.15}$$

Proof. Identify the conditional remainder. Unbiased fresh sampling and the tower property give $\mathbb{E}[\zeta_i^r \mid \mathcal{F}_{r-1}, I^r] = 0$; the second moment is proved explicitly in Lemma A.4 below. Independence of the virtual paths and the current indicators gives

$$\mathbb{E}[I_i^r u_i^r \mid \mathcal{F}_{r-1}] = p_i^r \mathbb{E}[u_i^r \mid \mathcal{F}_{r-1}], \qquad \mathbb{E}[I_i^r \zeta_i^r \mid \mathcal{F}_{r-1}] = 0.$$

Thus $N^{-1}\sum_i w_i^r \mathbb{E}[\varepsilon_i^r \mid \mathcal{F}_{r-1}] = \bar{\varepsilon}_r$.

Cancel the retained-memory terms. Taking the weighted conditional expectation of (A.8) gives

$$\begin{aligned}
\mathbb{E}[z^r \mid \mathcal{F}_{r-1}] &= \frac{1}{N}\sum_i (1-p_i^r) w_i^r q_i^{r-1} + \gamma_r\big(x^{r-1} - h\nabla f(x^{r-1})\big) + \bar{\varepsilon}_r \\
&= \frac{1}{N}\sum_i (w_i^{r-1} - \gamma_r) q_i^{r-1} + \gamma_r x^{r-1} - h\gamma_r \nabla f(x^{r-1}) + \bar{\varepsilon}_r \\
&= z^{r-1} - \gamma_r x^{r-1} + \gamma_r x^{r-1} - h\gamma_r \nabla f(x^{r-1}) + \bar{\varepsilon}_r.
\end{aligned}$$

The second equality uses $(1-p_i^r) w_i^r = w_i^{r-1} - \gamma_r$ and $N^{-1}\sum_i q_i^{r-1} = x^{r-1}$. This cancels the unequal retention of the memories exactly and proves (A.14).

Compare the trajectories. Since $q_i^r - x^r = -h t_i^r$ and $\sum_i t_i^r = 0$,

$$z^r - x^r = -\frac{h}{N}\sum_i w_i^r t_i^r = hL_r(t^r).$$

To bound this gap, expand $L_r(t^r)$ and apply Cauchy–Schwarz to the sum over clients. For every realization,

$$\begin{aligned}\|z^r - x^r\|^2 &= h^2 \left\| \frac{1}{N} \sum_i (1 - w_i^r) t_i^r \right\|^2 \\ &\le h^2 \left(\frac{1}{N} \sum_i (1 - w_i^r)^2 \right) \left(\frac{1}{N} \sum_i \|t_i^r\|^2 \right) \\ &\le \theta h^2 \frac{1}{N} \sum_i \|t_i^r\|^2.\end{aligned}$$

The first inequality separates the weight deviations from the client tracking errors. The second uses (A.12), which bounds the first factor by θ. Taking expectations and recalling the definition of $\mathcal{E}_r$ in (A.5) gives

$$\mathbb{E}\|z^r - x^r\|^2 \le \theta h^2 \frac{1}{N} \sum_i \mathbb{E}\|t_i^r\|^2 = \theta h^2 \mathcal{E}_r.$$

Finally, $w_i^R = 1$ and the memory-average identity give $z^R = x^R$. If $p_i^r = p^r$ for every client, backward induction gives $\gamma_r = p^r$ and $w_i^{r-1} = p^r + (1 - p^r) = 1$. Hence $z^r = x^r$ at every round in this case. □

The leading direction is the global gradient at x^{r-1}, not at z^{r-1}. Therefore (A.14) is not a standalone gradient-descent recursion for z. The next step bounds tracking and uses this identity in a joint descent argument.

A.4 Oracle noise and weighted projections

The gap bound (A.15) reduces the comparison of z and x to memory tracking. However, bounding all tracking error by one sum of squares would hide the variance reduction from averaging clients. We therefore separate retained oracle noise from the remaining tracking error, then combine the latter's recursion with the linearized objective.

Using the centering operation C, split the tracking matrix t^r, whose ith row is $(t_i^r)^\top$, into retained oracle noise n^r and a remaining component k^r:

$$n^0 = 0, \qquad n^r = C\big((1 - I_i^r) n_i^{r-1} + I_i^r \zeta_i^r\big)_{i=1}^N, \qquad k_i^r = t_i^r - n_i^r, \tag{A.16}$$

The energy of the remaining tracking error and the noise contribution to the server direction are

$$A_r = \frac{1}{N} \sum_i \mathbb{E}\,\|k_i^r\|^2\,.$$

$$B_{\text{noise}}^r = \frac{1}{N} \sum_i I_i^r (n_i^{r-1} - \zeta_i^r). \tag{A.17}$$

The virtual ζ_i^r for inactive clients are used only on the proof space; neither auxiliary matrix changes the algorithm.

Lemma A.4 (Fresh oracle noise)**.** *For every client i and round $r \ge 1$,*

$$\mathbb{E}[\zeta_i^r \mid \mathcal{F}_{r-1}] = 0, \qquad \mathbb{E}[\|\zeta_i^r\|^2 \mid \mathcal{F}_{r-1}] \le \frac{\sigma^2}{K}. \tag{A.18}$$

Proof. Let $e_i^{r,k} = g_i(y_i^{r,k}; \xi_i^{r,k}) - \nabla f_i(y_i^{r,k})$ be the error of local call k. Conditional on the history before that call, $y_i^{r,k}$ is fixed and the fresh draw has

$$\mathbb{E}[e_i^{r,k} \mid \text{history before call } (i,r,k)] = 0, \qquad \mathbb{E}[\|e_i^{r,k}\|^2 \mid \text{history before call } (i,r,k)] \le \sigma^2.$$

For $j < k$, $e_i^{r,j}$ is already known before call k. Applying the tower property therefore gives

$$\mathbb{E}[\langle e_i^{r,j}, e_i^{r,k} \rangle \mid \mathcal{F}_{r-1}] = 0, \qquad \mathbb{E}[e_i^{r,k} \mid \mathcal{F}_{r-1}] = 0.$$

Since $\zeta_i^r = K^{-1}\sum_{k=0}^{K-1} e_i^{r,k}$,

$$\begin{aligned}\mathbb{E}[\|\zeta_i^r\|^2 \mid \mathcal{F}_{r-1}] &= \frac{1}{K^2}\sum_{k=0}^{K-1}\mathbb{E}[\|e_i^{r,k}\|^2 \mid \mathcal{F}_{r-1}] \\ &\quad + \frac{2}{K^2}\sum_{0\le j<k<K}\mathbb{E}[\langle e_i^{r,j}, e_i^{r,k}\rangle \mid \mathcal{F}_{r-1}] \\ &\le \frac{K\sigma^2}{K^2} = \frac{\sigma^2}{K}.\end{aligned}$$

Averaging the zero conditional means gives the first claim. Both claims remain true after conditioning also on I^r, because the virtual local paths are independent of current participation. Moreover, distinct clients use conditionally independent fresh draws, so $\mathbb{E}[\langle \zeta_i^r, \zeta_j^r\rangle \mid \mathcal{F}_{r-1}, I^r] = 0$ for $i \neq j$. □

Lemma A.5 (Retained-noise covariance)**.** *For every deterministic coefficient vector $\alpha \in \mathbb{R}^N$ and $r \ge 0$,*

$$\mathbb{E}\left\|\sum_i \alpha_i n_i^r\right\|^2 \le \frac{\sigma^2}{K}\|C\alpha\|^2. \tag{A.19}$$

In particular,

$$\mathbb{E}\|n_i^r\|^2 \le \frac{\sigma^2}{K}\left(1-\frac{1}{N}\right), \qquad \frac{1}{N}\sum_i \mathbb{E}\|n_i^r\|^2 \le \frac{\sigma^2}{K}. \tag{A.20}$$

Proof. Separate retained and fresh noise. We prove (A.19) for every deterministic coefficient vector $\alpha \in \mathbb{R}^N$. Its entries are scalar coefficients, not the client vectors forming the rows of $v \in \mathbb{R}^{N\times d}$. It is immediate at $r = 0$ because $n^0 = 0$. Assume it holds at round $r-1$, and fix α. Since C is symmetric, the centering in (A.16) can be applied to the coefficients instead of the noise matrix:

$$\sum_i \alpha_i n_i^r = \underbrace{\sum_i (1-I_i^r)(C\alpha)_i n_i^{r-1}}_{\text{retained noise}} + \underbrace{\sum_i I_i^r (C\alpha)_i \zeta_i^r}_{\text{fresh noise}}.$$

Here $(C\alpha)_i = \alpha_i - N^{-1}\sum_j \alpha_j$. We bound the two sums separately; we do not assume that the stored noises n_i^{r-1} are independent.

Bound the retained noise using the induction hypothesis. Condition on the current indicators $I^r = (I_1^r, \dots, I_N^r)$. The coefficients $(1-I_i^r)(C\alpha)_i$ are then fixed. Moreover, I^r is independent of n^{r-1}, so conditioning on I^r does not change the distribution of the previous noise matrix. We may therefore apply the induction hypothesis with these coefficients:

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_i (1-I_i^r)(C\alpha)_i n_i^{r-1}\right\|^2 \,\middle|\, I^r\right] \\ &\le \frac{\sigma^2}{K}\left\|C\big((1-I_i^r)(C\alpha)_i\big)_{i=1}^N\right\|^2 \\ &\le \frac{\sigma^2}{K}\sum_i (1-I_i^r)(C\alpha)_i^2.\end{aligned}$$

The last inequality uses that centering cannot increase the Euclidean norm and that $(1-I_i^r)^2 = 1-I_i^r$. Averaging over participation, with $\mathbb{E}[1-I_i^r] = 1-p_i^r$, gives

$$\mathbb{E}\left\|\sum_i (1-I_i^r)(C\alpha)_i n_i^{r-1}\right\|^2 \le \frac{\sigma^2}{K}\sum_i (1-p_i^r)(C\alpha)_i^2.$$

Bound the fresh noise and remove the cross term. Given the history $\mathcal{F}_{r-1}$ and the current indicators I^r, the retained-noise sum is fixed. Fresh independent oracle sampling and the virtual-path

construction give

$$\mathbb{E}[\zeta_i^r \mid \mathcal{F}_{r-1}, I^r] = 0, \qquad \mathbb{E}[\|\zeta_i^r\|^2 \mid \mathcal{F}_{r-1}, I^r] \le \frac{\sigma^2}{K}.$$

The fresh errors from distinct clients have zero conditional cross moments. Thus, when the squared norm of the fresh-noise sum is expanded, only its diagonal terms remain:

$$\begin{aligned}\mathbb{E}\left\|\sum_i I_i^r (C\alpha)_i \zeta_i^r\right\|^2 &= \sum_i (C\alpha)_i^2 \,\mathbb{E}[I_i^r \|\zeta_i^r\|^2] \\ &\le \frac{\sigma^2}{K}\sum_i p_i^r (C\alpha)_i^2.\end{aligned}$$

Also, the conditional mean of the fresh-noise sum is zero. Its inner product with the fixed retained-noise sum therefore has conditional expectation zero, and hence

$$\mathbb{E}\left\langle \sum_i (1 - I_i^r)(C\alpha)_i n_i^{r-1}, \sum_i I_i^r (C\alpha)_i \zeta_i^r \right\rangle = 0.$$

Combine the bounds. Expand the squared norm of the retained- and fresh-noise sums. Their cross term has zero expectation, so the two second-moment bounds above give

$$\begin{aligned}\mathbb{E}\left\|\sum_i \alpha_i n_i^r\right\|^2 &\le \frac{\sigma^2}{K}\sum_i \big((1 - p_i^r) + p_i^r\big)(C\alpha)_i^2 \\ &= \frac{\sigma^2}{K}\|C\alpha\|^2.\end{aligned}$$

This completes the induction. In particular, retention and refresh contribute complementary factors $1 - p_i^r$ and p_i^r; the bound does not accumulate a new noise term at every round.

Obtain the noise-matrix energy bound. For each client i, choose α with entry i equal to one and all other entries zero. Then $\sum_j \alpha_j n_j^r = n_i^r$ and $\|C\alpha\|^2 = 1 - 1/N$. Applying (A.19) yields

$$\mathbb{E}\|n_i^r\|^2 \le \frac{\sigma^2}{K}\left(1 - \frac{1}{N}\right) \le \frac{\sigma^2}{K}.$$

Averaging this inequality over i proves (A.20). □

Lemma A.6 (Server-noise variance)**.** *For every $r \ge 1$,*

$$\mathbb{E}\|B_{\text{noise}}^r\|^2 \le \frac{3\delta_{\max}\sigma^2}{NK}. \tag{A.21}$$

Proof. Separate the three contributions. Recall $B_{\text{noise}}^r = N^{-1}\sum_i I_i^r (n_i^{r-1} - \zeta_i^r)$. Add and subtract $N^{-1}\sum_i p_i^r n_i^{r-1}$ to obtain

$$B_{\text{noise}}^r = \underbrace{\frac{1}{N}\sum_i p_i^r n_i^{r-1}}_{\text{mean retained noise}} + \underbrace{\frac{1}{N}\sum_i (I_i^r - p_i^r) n_i^{r-1}}_{\text{participation fluctuation}} - \underbrace{\frac{1}{N}\sum_i I_i^r \zeta_i^r}_{\text{fresh noise}}.$$

The first term is the expected retained-noise contribution given the past. The second is its deviation caused by the random active set. The third comes from the new stochastic gradients in this round.

Remove the cross terms. Given $\mathcal{F}_{r-1}$, the first term is fixed and the second has mean zero because $\mathbb{E}[I_i^r - p_i^r \mid \mathcal{F}_{r-1}] = 0$. Their inner product therefore has expectation zero. Next condition on both $\mathcal{F}_{r-1}$ and I^r: the first two terms are fixed, whereas the third has conditional mean zero by

Lemma A.4. Its inner product with either of the first two terms also has expectation zero. Thus the second moment of the sum is exactly the sum of the three second moments:

$$\mathbb{E}\|B^r_{\text{noise}}\|^2 = \mathbb{E}\left\|\frac{1}{N}\sum_i p_i^r n_i^{r-1}\right\|^2 + \mathbb{E}\left\|\frac{1}{N}\sum_i (I_i^r - p_i^r) n_i^{r-1}\right\|^2 + \mathbb{E}\left\|\frac{1}{N}\sum_i I_i^r \zeta_i^r\right\|^2.$$

This uses zero expected inner products, not independence of the three contributions.

Bound each contribution. For the mean retained noise, apply (A.19) at round $r-1$ with $\alpha_i = p_i^r/N$. Since centering cannot increase the norm,

$$\mathbb{E}\left\|\frac{1}{N}\sum_i p_i^r n_i^{r-1}\right\|^2 \le \frac{\sigma^2}{K}\|C\alpha\|^2 \le \frac{\sigma^2}{KN^2}\sum_i (p_i^r)^2 \le \frac{\delta_{\max}\sigma^2}{NK}.$$

The last inequality uses $(p_i^r)^2 \le p_i^r \le \delta_{\max}$.

For the participation fluctuation, condition on $\mathcal{F}_{r-1}$. The centered indicators $I_i^r - p_i^r$ are independent across clients, with variance $p_i^r(1-p_i^r)$, so all terms involving distinct clients vanish when the square is expanded. Taking total expectation and using the per-client bound (A.20) gives

$$\mathbb{E}\left\|\frac{1}{N}\sum_i (I_i^r - p_i^r) n_i^{r-1}\right\|^2 = \frac{1}{N^2}\sum_i p_i^r(1-p_i^r)\mathbb{E}\|n_i^{r-1}\|^2 \le \frac{\sigma^2}{KN^2}\sum_i p_i^r(1-p_i^r) \le \frac{\delta_{\max}\sigma^2}{NK}.$$

For the fresh noise, the conditional zero cross moments and variance bound in Lemma A.4 give

$$\mathbb{E}\left\|\frac{1}{N}\sum_i I_i^r \zeta_i^r\right\|^2 = \frac{1}{N^2}\sum_i \mathbb{E}[I_i^r\|\zeta_i^r\|^2] \le \frac{\sigma^2}{KN^2}\sum_i p_i^r \le \frac{\delta_{\max}\sigma^2}{NK}.$$

Adding these three bounds in the second-moment decomposition above proves

$$\mathbb{E}\|B^r_{\text{noise}}\|^2 \le \frac{3\delta_{\max}\sigma^2}{NK},$$

which is (A.21). □

Lemma A.7 (Weighted projections). *For $0 \le r \le R$,*

$$\mathbb{E}\|L_r(n^r)\|^2 \le \frac{\theta\sigma^2}{NK}, \tag{A.22}$$

$$\mathbb{E}\|L_r(k^r)\|^2 \le \theta A_r. \tag{A.23}$$

Proof. For the noise projection, choose the deterministic coefficients $\alpha_i = (1-w_i^r)/N$. The weights have unit average, so $\sum_i \alpha_i = 0$ and hence $C\alpha = \alpha$. Therefore

$$\mathbb{E}\|L_r(n^r)\|^2 = \mathbb{E}\left\|\sum_i \alpha_i n_i^r\right\|^2 \le \frac{\sigma^2}{K}\|C\alpha\|^2 = \frac{\sigma^2}{KN^2}\sum_i (1-w_i^r)^2 \le \frac{\theta\sigma^2}{NK}.$$

Here the first inequality is (A.19), and the last is (A.12). This proves (A.22).

For the remaining tracking error, use Cauchy–Schwarz directly:

$$\mathbb{E}\|L_r(k^r)\|^2 = \mathbb{E}\left\|\frac{1}{N}\sum_i (1-w_i^r)k_i^r\right\|^2 \le \left(\frac{1}{N}\sum_i (1-w_i^r)^2\right)\left(\frac{1}{N}\sum_i \mathbb{E}\|k_i^r\|^2\right) \le \theta A_r.$$

The final inequality uses (A.12) and the definition of A_r, proving (A.23). □

A.5 Local drift and memory tracking

Lemma A.8 (Local trajectory drift)**.** *For every round $r \ge 1$, the bias caused by the local trajectories obeys*

$$U_r \le \frac{18(\beta h)^2}{\eta_g^2}\left(G_{r-1} + A_{r-1} + \frac{\sigma^2}{K}\right). \tag{A.24}$$

Proof. Fix a round r and write $x = x^{r-1}$. We first bound how far a local iterate can move from x, then use smoothness to turn this distance bound into a bound on u_i^r.

Separate the fixed direction, fresh noise, and local drift. Write the fresh oracle error as

$$e_i^{r,j} = g_i(y_i^{r,j};\xi_i^{r,j}) - \nabla f_i(y_i^{r,j}).$$

The definition of t_i^{r-1} gives the exact decomposition

$$g_i(y_i^{r,j};\xi_i^{r,j}) + \frac{m_i^{r-1}-x}{h} = \nabla f(x) - t_i^{r-1} + \big(\nabla f_i(y_i^{r,j}) - \nabla f_i(x)\big) + e_i^{r,j}.$$

The first two terms are fixed given the history before the round. The gradient difference measures the local drift away from x. Since $y_i^{r,0} = x$, summing the first k local updates yields

$$y_i^{r,k} - x = -\eta_\ell\left(k(\nabla f(x) - t_i^{r-1}) + \sum_{j=0}^{k-1} e_i^{r,j}\right) - \eta_\ell \sum_{j=0}^{k-1}\big(\nabla f_i(y_i^{r,j}) - \nabla f_i(x)\big).$$

The centering identity $\sum_i t_i^{r-1} = 0$ also gives

$$\frac{1}{N}\sum_i \mathbb{E}\|\nabla f(x) - t_i^{r-1}\|^2 = G_{r-1} + \mathcal{E}_{r-1} - 2\mathbb{E}\left\langle \nabla f(x), \frac{1}{N}\sum_i t_i^{r-1}\right\rangle = G_{r-1} + \mathcal{E}_{r-1}.$$

Thus averaging the fixed-direction energy introduces tracking error, but no separate gradient-heterogeneity term.

Bound the two accumulated sums. We use the same fresh-oracle martingale-difference property as in Lemma A.4. Since $\nabla f(x) - t_i^{r-1}$ is measurable before the round, and $e_i^{r,j}$ is measurable before any later local draw,

$$\mathbb{E}\langle \nabla f(x) - t_i^{r-1}, e_i^{r,j}\rangle = 0, \qquad \mathbb{E}\langle e_i^{r,j}, e_i^{r,\ell}\rangle = 0 \quad (j < \ell).$$

Expanding the square therefore gives

$$\mathbb{E}\left\|k(\nabla f(x) - t_i^{r-1}) + \sum_{j=0}^{k-1} e_i^{r,j}\right\|^2 = k^2\mathbb{E}\|\nabla f(x) - t_i^{r-1}\|^2 + \sum_{j=0}^{k-1}\mathbb{E}\|e_i^{r,j}\|^2 \le k^2\mathbb{E}\|\nabla f(x) - t_i^{r-1}\|^2 + k\sigma^2.$$

For the drift sum, Cauchy–Schwarz and smoothness give, for $k \geq 1$,

$$\mathbb{E}\left\|\sum_{j=0}^{k-1}\big(\nabla f_i(y_i^{r,j}) - \nabla f_i(x)\big)\right\|^2 \leq k\sum_{j=0}^{k-1}\mathbb{E}\|\nabla f_i(y_i^{r,j}) - \nabla f_i(x)\|^2$$
$$\leq k\beta^2\sum_{j=0}^{k-1}\mathbb{E}\|y_i^{r,j} - x\|^2.$$

No orthogonality is needed for this second sum; it is controlled through the distances of the local iterates from x.

Obtain a bound valid for every local step. Set $Z_i = \max_{0\leq k\leq K}\mathbb{E}\|y_i^{r,k} - x\|^2$. Applying $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ to the unrolled update, and then using the noise and drift bounds above, yields

$$\mathbb{E}\|y_i^{r,k} - x\|^2 \leq 2\eta_\ell^2\big(k^2\mathbb{E}\|\nabla f(x) - t_i^{r-1}\|^2 + k\sigma^2\big) + 2\eta_\ell^2 k^2\beta^2 Z_i$$
$$\leq 2(K\eta_\ell)^2\left(\mathbb{E}\|\nabla f(x) - t_i^{r-1}\|^2 + \frac{\sigma^2}{K}\right) + 2(\beta K\eta_\ell)^2 Z_i.$$

The right-hand side does not depend on k, so it also bounds Z_i. Move its last term to the left. Since

$$\beta K\eta_\ell = \frac{\beta h}{\eta_g} \leq \frac{\delta}{200\Lambda} \leq \frac{1}{600}, \qquad 1 - 2(\beta K\eta_\ell)^2 \geq \frac{1}{3},$$

we obtain

$$Z_i \leq 6(K\eta_\ell)^2\left(\mathbb{E}\|\nabla f(x) - t_i^{r-1}\|^2 + \frac{\sigma^2}{K}\right).$$

In particular, averaging over clients and over $k = 0, \ldots, K-1$, and using the fixed-direction energy identity and $K\eta_\ell = h/\eta_g$, gives

$$\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\|y_i^{r,k} - x\|^2 \leq \frac{6h^2}{\eta_g^2}\left(G_{r-1} + \mathcal{E}_{r-1} + \frac{\sigma^2}{K}\right).$$

Convert local displacement into gradient drift. The definition of u_i^r, Jensen's inequality, and smoothness imply

$$\mathbb{E}\|u_i^r\|^2 \leq \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f_i(y_i^{r,k}) - \nabla f_i(x)\|^2 \leq \frac{\beta^2}{K}\sum_{k=0}^{K-1}\mathbb{E}\|y_i^{r,k} - x\|^2.$$

Also, $t_i^{r-1} = k_i^{r-1} + n_i^{r-1}$ and (A.20) give

$$\mathcal{E}_{r-1} \leq 2A_{r-1} + \frac{2}{N}\sum_i \mathbb{E}\|n_i^{r-1}\|^2 \leq 2A_{r-1} + \frac{2\sigma^2}{K}.$$

Combining these estimates shows explicitly how the stated constant arises:

$$U_r \leq \frac{6(\beta h)^2}{\eta_g^2}\left(G_{r-1} + \mathcal{E}_{r-1} + \frac{\sigma^2}{K}\right)$$
$$\leq \frac{6(\beta h)^2}{\eta_g^2}\left(G_{r-1} + 2A_{r-1} + \frac{3\sigma^2}{K}\right)$$
$$\leq \frac{18(\beta h)^2}{\eta_g^2}\left(G_{r-1} + A_{r-1} + \frac{\sigma^2}{K}\right).$$

This is (A.24). □

Lemma A.9 (Server motion and memory tracking)**.** *For every round $r \geq 1$,*

$$D_r \leq 6h^2\left(G_{r-1} + A_{r-1} + U_r + \frac{\delta_{\max}\sigma^2}{NK}\right), \tag{A.25}$$

$$A_r \leq \left(1 - \frac{\delta}{3}\right)A_{r-1} + 4G_{r-1} + 4U_r + \frac{72(\beta h)^2\delta_{\max}\sigma^2}{\delta NK}, \tag{A.26}$$

$$A_r \leq 4\left(G_{r-1} + A_{r-1} + U_r + \frac{\delta_{\max}\sigma^2}{NK}\right). \tag{A.27}$$

Proof. Fix r and write $x = x^{r-1}$. We first bound the server displacement. We then derive the update for k^r and show that refresh contracts its previous energy up to controlled gradient, local-drift, and noise terms.

Split the server direction into tracking and raw noise. The memory update and (A.9) give

$$m_i^r - m_i^{r-1} = hI_i^r\big(t_i^{r-1} - \nabla f(x) - u_i^r - \zeta_i^r\big).$$

Averaging over clients gives the exact server step

$$x^r - x = \frac{h}{N}\sum_i I_i^r\big(t_i^{r-1} - \nabla f(x) - u_i^r - \zeta_i^r\big). \tag{A.28}$$

Substitute $t_i^{r-1} = k_i^{r-1} + n_i^{r-1}$ to separate the retained oracle noise:

$$\frac{x^r - x}{h} = \frac{1}{N}\sum_i I_i^r\big(k_i^{r-1} - \nabla f(x) - u_i^r\big) + B_{\text{noise}}^r.$$

Here B_{noise}^r is the quantity already defined in (A.17). For any client vectors v_i, Cauchy–Schwarz gives the pathwise bound

$$\left\|\frac{1}{N}\sum_i I_i^r v_i\right\|^2 \le \frac{|\mathcal{S}^r|}{N^2}\sum_i I_i^r\|v_i\|^2 \le \frac{1}{N}\sum_i \|v_i\|^2.$$

Apply this with $v_i = k_i^{r-1} - \nabla f(x) - u_i^r$, and use $\|a+b+c\|^2 \le 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$, to obtain

$$\mathbb{E}\left\|\frac{1}{N}\sum_i I_i^r\big(k_i^{r-1} - \nabla f(x) - u_i^r\big)\right\|^2 \le 3(A_{r-1} + G_{r-1} + U_r).$$

Lemma A.6 gives $\mathbb{E}\|B_{\text{noise}}^r\|^2 \le 3\delta_{\max}\sigma^2/(NK)$. The two parts may be correlated, so we use the two-term square bound:

$$\begin{aligned} D_r &\le 2h^2\mathbb{E}\left\|\frac{1}{N}\sum_i I_i^r\big(k_i^{r-1} - \nabla f(x) - u_i^r\big)\right\|^2 + 2h^2\mathbb{E}\|B_{\text{noise}}^r\|^2 \\ &\le 6h^2\left(G_{r-1} + A_{r-1} + U_r + \frac{\delta_{\max}\sigma^2}{NK}\right). \end{aligned}$$

This proves (A.25).

Derive the exact tracking update. Set $s_i^r = b_i(x) - b_i(x^r)$. Substituting the memory increment into the definition of t_i^r gives, client by client,

$$t_i^r = (1 - I_i^r)t_i^{r-1} + I_i^r(\nabla f(x) + u_i^r + \zeta_i^r) + \frac{x^r - x}{h} + s_i^r.$$

The average of the first two terms is $-(x^r - x)/h$:

$$\begin{aligned} &\frac{1}{N}\sum_i\big((1 - I_i^r)t_i^{r-1} + I_i^r(\nabla f(x) + u_i^r + \zeta_i^r)\big) \\ &\quad = \frac{1}{N}\sum_i I_i^r(\nabla f(x) - t_i^{r-1} + u_i^r + \zeta_i^r) = -\frac{x^r - x}{h}. \end{aligned}$$

Here we used $\sum_i t_i^{r-1} = 0$ and (A.28). Consequently, adding the common displacement is exactly the centering operation C, and

$$t^r = C\big((1 - I_i^r)t_i^{r-1} + I_i^r(\nabla f(x) + u_i^r + \zeta_i^r)\big)_{i=1}^N + s^r.$$

Subtract the noise recursion (A.16). Its retained and fresh noise terms cancel, leaving

$$k^r = C\big((1 - I_i^r)k_i^{r-1} + I_i^r(\nabla f(x) + u_i^r)\big)_{i=1}^N + s^r. \tag{A.29}$$

Thus raw oracle noise enters this remaining tracking component only through the local trajectories and the coordinate change s^r.

Bound the energy before the coordinate change. Because I_i^r is either zero or one, the retained and refreshed terms in (A.29) are never simultaneously nonzero. Participation independence and the virtual-path construction therefore give

$$\mathbb{E}\left\|(1-I_i^r)k_i^{r-1}+I_i^r\left(\nabla f(x)+u_i^r\right)\right\|^2=(1-p_i^r)\mathbb{E}\|k_i^{r-1}\|^2+p_i^r\mathbb{E}\|\nabla f(x)+u_i^r\|^2.$$

Using $p_i^r \geq \delta$, $p_i^r \leq 1$, and the two-term square bound, we obtain

$$\frac{1}{N}\sum_i \mathbb{E}\left\|(1-I_i^r)k_i^{r-1}+I_i^r\left(\nabla f(x)+u_i^r\right)\right\|^2 \leq (1-\delta)A_{r-1}+2G_{r-1}+2U_r.$$

Centering can only decrease this energy, since for any client matrix v,

$$\frac{1}{N}\sum_i \|(Cv)_i\|^2=\frac{1}{N}\sum_i\|v_i\|^2-\left\|\frac{1}{N}\sum_i v_i\right\|^2 \leq \frac{1}{N}\sum_i\|v_i\|^2.$$

Include the coordinate change. Apply Young's inequality in the form

$$\|a+b\|^2 \leq (1+\delta/2)\|a\|^2+(1+2/\delta)\|b\|^2$$

to the centered term and s^r in (A.29). Smoothness gives $N^{-1}\sum_i \mathbb{E}\|s_i^r\|^2 \leq 4\beta^2 D_r$. Hence the pre-coordinate-change energy bound yields

$$\begin{aligned}A_r &\leq (1+\delta/2)(1-\delta)A_{r-1}+2(1+\delta/2)(G_{r-1}+U_r)\\ &\quad +4\beta^2(1+2/\delta)D_r\\ &\leq (1-\delta/2)A_{r-1}+3G_{r-1}+3U_r+\frac{12\beta^2}{\delta}D_r.\end{aligned}$$

For the last line, we used $0<\delta\leq 1$ to bound $(1+\delta/2)(1-\delta)\leq 1-\delta/2$, $2(1+\delta/2)\leq 3$, and $1+2/\delta \leq 3/\delta$. Substituting the server-motion bound (A.25) gives

$$\begin{aligned}A_r &\leq \left(1-\frac{\delta}{2}\right)A_{r-1}+3G_{r-1}+3U_r\\ &\quad +\frac{72(\beta h)^2}{\delta}\left(G_{r-1}+A_{r-1}+U_r+\frac{\delta_{\max}\sigma^2}{NK}\right).\end{aligned} \tag{A.30}$$

Collect the coefficients. The step-size condition implies

$$\frac{72(\beta h)^2}{\delta}\leq\frac{72\delta^3}{600^2}\leq\frac{\delta}{6}\leq 1.$$

The coefficients of A_{r-1}, G_{r-1}, and U_r in (A.30) consequently satisfy

$$1-\frac{\delta}{2}+\frac{72(\beta h)^2}{\delta}\leq 1-\frac{\delta}{3},\qquad 3+\frac{72(\beta h)^2}{\delta}\leq 4.$$

Keeping the noise coefficient $72(\beta h)^2/\delta$ unchanged proves (A.26). Bounding all four coefficients by four also proves (A.27). □

What the tracking bounds imply. The noise lemma gives $\mathcal{E}_r \leq 2A_r+2\sigma^2/K$. Hence (A.15) controls the trajectory separation through the tracking recursion above. In particular, Lemma A.13 derives the explicit consequence

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\|z^r-x^r\|^2 \leq \theta h^2\left(\frac{12\mathcal{E}_0}{\delta R}+\frac{60}{\delta}\overline{G}_R+\frac{3\sigma^2}{K}\right).$$

This is a coupled estimate: its right-hand side still contains the stationarity measure $\overline{G}_R$. We close the argument below by adding tracking energy to descent, rather than assuming that the gap already vanishes.

A.6 Analysis of the descent potential

We now use the dynamics of z without treating it as an independent optimization process. The descent potential is based on the first-order approximation of $f(z^r) - f_{\inf}$ around the actual model x^r. This keeps the gradient evaluated at the same point as in (A.14).

For the fixed horizon R, recall the descent potential

$$\Phi_r = F_r + \mathbb{E}\langle \nabla f(x^r), z^r - x^r \rangle = F_r + h\mathbb{E}\langle \nabla f(x^r), L_r(t^r) \rangle. \tag{A.31}$$

Its dependence on R is suppressed, as for w^r and z^r.

Lemma A.10 (One-round bound for the descent potential)**.** *For every* $1 \le r \le R$,

$$\begin{aligned}\Phi_r - \Phi_{r-1} \le &-\frac{7\gamma_r h}{8} G_{r-1} + \frac{4h\delta_{\max}^2}{\delta} U_r \\ &+ \beta h^2 \Lambda \left(G_{r-1} + A_{r-1} + U_r + \frac{\delta_{\max}\sigma^2}{NK} \right).\end{aligned} \tag{A.32}$$

Proof. We first derive a descent inequality with four error terms. We then bound each term separately, keeping the retained oracle noise apart from the remaining tracking error. Fix r and write $x = x^{r-1}$. As in Lemma A.9, let $s_i^r = b_i(x) - b_i(x^r)$ denote the coordinate change. We write $\mu_i^r = \mathbb{E}[u_i^r \mid \mathcal{F}_{r-1}]$ for the conditional local bias.

Expand the descent potential. Smoothness gives the pathwise Taylor bound

$$\left| f(z^r) - f(x^r) - \langle \nabla f(x^r), z^r - x^r \rangle \right| \le \frac{\beta}{2} \|z^r - x^r\|^2.$$

Taking expectations shows how Φ_r relates to the objective at z^r:

$$|\mathbb{E}[f(z^r) - f_{\inf}] - \Phi_r| \le \frac{\beta}{2} \mathbb{E}\|z^r - x^r\|^2. \tag{A.33}$$

For the one-round change, expand the added linear term around the old gradient $\nabla f(x)$:

$$\begin{aligned}(\Phi_r - F_r) - (\Phi_{r-1} - F_{r-1}) = &\mathbb{E}\langle \nabla f(x), (z^r - x^r) - (z^{r-1} - x) \rangle \\ &+ \mathbb{E}\langle \nabla f(x^r) - \nabla f(x), z^r - x^r \rangle.\end{aligned}$$

On the other hand, $F_r - F_{r-1} \le \mathbb{E}\langle \nabla f(x), x^r - x \rangle + \beta D_r/2$. Adding these two relations cancels the explicit server displacement inside the first inner product:

$$\begin{aligned}&\Phi_r - \Phi_{r-1} \\ &\le \mathbb{E}\langle \nabla f(x), z^r - z^{r-1} \rangle + \frac{\beta}{2} D_r + \mathbb{E}\langle \nabla f(x^r) - \nabla f(x), z^r - x^r \rangle.\end{aligned}$$

Since $\nabla f(x)$ is measurable before the round, the tower property and (A.14) give

$$\begin{aligned}\mathbb{E}\langle \nabla f(x), z^r - z^{r-1} \rangle &= \mathbb{E}\langle \nabla f(x), \mathbb{E}[z^r - z^{r-1} \mid \mathcal{F}_{r-1}] \rangle \\ &= -h\gamma_r G_{r-1} + \mathbb{E}\langle \nabla f(x), \bar{\varepsilon}_r \rangle.\end{aligned}$$

The coordinate-change remainder has the useful sign identity

$$\frac{1}{N} \sum_i w_i^r \big(b_i(x^r) - b_i(x) \big) = -\frac{1}{N} \sum_i w_i^r s_i^r = L_r(s^r),$$

where the last equality uses $\sum_i s_i^r = 0$. Together with $z^r - x^r = hL_r(t^r)$, this gives

$$\begin{aligned}&\Phi_r - \Phi_{r-1} \\ &\le -\gamma_r h G_{r-1} - h\mathbb{E}\left\langle \nabla f(x), \frac{1}{N} \sum_i p_i^r w_i^r \mu_i^r \right\rangle \\ &\quad + h\mathbb{E}\langle \nabla f(x), L_r(s^r) \rangle + \frac{\beta}{2} D_r + h\mathbb{E}\langle \nabla f(x^r) - \nabla f(x), L_r(t^r) \rangle.\end{aligned} \tag{A.34}$$

We now bound the local bias, the coordinate change, and the two parts of the last term obtained by writing $t^r = k^r + n^r$.

Bound the local-bias term. The weights satisfy $p_i^r w_i^r \geq 0$, $N^{-1}\sum_i p_i^r w_i^r = \gamma_r$, and $p_i^r w_i^r \leq \delta_{\max}^2/\delta$. Weighted Jensen therefore gives

$$\left\| \frac{1}{N}\sum_i p_i^r w_i^r \mu_i^r \right\|^2 = \gamma_r^2 \left\| \sum_i \frac{p_i^r w_i^r}{N\gamma_r} \mu_i^r \right\|^2 \leq \frac{\gamma_r}{N}\sum_i p_i^r w_i^r \|\mu_i^r\|^2.$$

Conditional Jensen also gives $N^{-1}\sum_i \mathbb{E}\|\mu_i^r\|^2 \leq U_r$. Hence

$$\mathbb{E}\left\| \frac{1}{N}\sum_i p_i^r w_i^r \mu_i^r \right\|^2 \leq \frac{\gamma_r \delta_{\max}^2}{\delta} U_r.$$

Apply Young's inequality in the form

$$|\langle a, b\rangle| \leq \frac{\gamma_r}{16}\|a\|^2 + \frac{4}{\gamma_r}\|b\|^2.$$

Using the preceding second-moment bound cancels the factor γ_r in the second term and yields

$$h\left|\mathbb{E}\left\langle \nabla f(x), \frac{1}{N}\sum_i p_i^r w_i^r \mu_i^r \right\rangle\right| \leq \frac{\gamma_r h}{16} G_{r-1} + \frac{4h\delta_{\max}^2}{\delta} U_r. \tag{A.35}$$

Bound the coordinate-change term. Cauchy–Schwarz over clients, followed by the weight bound (A.12), gives

$$\mathbb{E}\|L_r(s^r)\|^2 \leq \theta\,\frac{1}{N}\sum_i \mathbb{E}\|s_i^r\|^2 \leq 4\theta\beta^2 D_r.$$

The second inequality uses $\|b_i(x) - b_i(x^r)\| \leq 2\beta\|x^r - x\|$. Applying the same Young inequality used for the local-bias term, and then $\gamma_r \geq \delta$, gives

$$h|\mathbb{E}\langle \nabla f(x), L_r(s^r)\rangle| \leq \frac{\gamma_r h}{16} G_{r-1} + \frac{16\beta^2 h\theta}{\delta} D_r. \tag{A.36}$$

Bound the change of gradient against tracking and noise. Before separating the two components, smoothness gives the general bound

$$\begin{aligned} &|\mathbb{E}\langle \nabla f(x^r) - \nabla f(x), z^r - x^r\rangle| \\ &\quad\leq \sqrt{\mathbb{E}\|\nabla f(x^r) - \nabla f(x)\|^2}\sqrt{\mathbb{E}\|z^r - x^r\|^2} \\ &\quad\leq \beta\sqrt{D_r}\sqrt{\mathbb{E}\|z^r - x^r\|^2}. \end{aligned}$$

To retain the sharper noise scale, use $z^r - x^r = hL_r(k^r) + hL_r(n^r)$ and bound the two inner products separately. Smoothness and Cauchy–Schwarz in expectation give

$$\begin{aligned} &h|\mathbb{E}\langle \nabla f(x^r) - \nabla f(x), L_r(k^r)\rangle| \\ &\quad\leq \beta h\sqrt{D_r}\sqrt{\mathbb{E}\|L_r(k^r)\|^2} \leq \beta h\sqrt{D_r \theta A_r}. \end{aligned}$$

The second inequality uses (A.23). Applying (A.25) and (A.27) to D_r and A_r, respectively, bounds the last expression by

$$\sqrt{24}\,\beta h^2\sqrt{\theta}\left(G_{r-1} + A_{r-1} + U_r + \frac{\delta_{\max}\sigma^2}{NK}\right).$$

Since $\sqrt{24} \leq 5$, we have

$$\begin{aligned} &h|\mathbb{E}\langle \nabla f(x^r) - \nabla f(x), L_r(k^r)\rangle| \\ &\quad\leq 5\beta h^2\sqrt{\theta}\left(G_{r-1} + A_{r-1} + U_r + \frac{\delta_{\max}\sigma^2}{NK}\right). \end{aligned} \tag{A.37}$$

For the noise component, use the sharper projection bound (A.22):

$$\begin{aligned} &h|\mathbb{E}\langle \nabla f(x^r) - \nabla f(x), L_r(n^r)\rangle| \\ &\quad \le \beta h \sqrt{D_r}\sqrt{\mathbb{E}\|L_r(n^r)\|^2} \le \beta h \sqrt{D_r \frac{\theta\sigma^2}{NK}} \\ &\quad \le \sqrt{6}\,\beta h^2 \sqrt{\frac{\theta}{\delta_{\max}}}\sqrt{\left(G_{r-1} + A_{r-1} + U_r + \frac{\delta_{\max}\sigma^2}{NK}\right)\frac{\delta_{\max}\sigma^2}{NK}}. \end{aligned}$$

The last line uses (A.25). Because all four terms in the parenthesis are nonnegative,

$$\sqrt{\left(G_{r-1} + A_{r-1} + U_r + \frac{\delta_{\max}\sigma^2}{NK}\right)\frac{\delta_{\max}\sigma^2}{NK}} \le G_{r-1} + A_{r-1} + U_r + \frac{\delta_{\max}\sigma^2}{NK}.$$

Together with $\sqrt{6} \le 3$, this gives

$$\begin{aligned} &h|\mathbb{E}\langle \nabla f(x^r) - \nabla f(x), L_r(n^r)\rangle| \\ &\quad \le 3\beta h^2 \sqrt{\frac{\theta}{\delta_{\max}}}\left(G_{r-1} + A_{r-1} + U_r + \frac{\delta_{\max}\sigma^2}{NK}\right). \end{aligned} \tag{A.38}$$

These applications of Cauchy–Schwarz allow the tracking and noise components to be correlated with the server displacement. Separating k^r and n^r preserves the factor $1/N$ in the noise bound.

Collect the descent and error coefficients. The remaining coordinate-change contribution in (A.36) satisfies, by (A.25),

$$\frac{16\beta^2 h\theta}{\delta} D_r \le \frac{96\beta^2 h^3 \theta}{\delta}\left(G_{r-1} + A_{r-1} + U_r + \frac{\delta_{\max}\sigma^2}{NK}\right).$$

Since $\beta h\Lambda \le \delta^2/200$ and $\Lambda \ge 6\sqrt{\theta}$,

$$\frac{96\beta h\sqrt{\theta}}{\delta} \le \frac{96\delta\sqrt{\theta}}{200\Lambda} \le \frac{2\delta}{25} \le 1.$$

Consequently,

$$\frac{16\beta^2 h\theta}{\delta} D_r \le \beta h^2\sqrt{\theta}\left(G_{r-1} + A_{r-1} + U_r + \frac{\delta_{\max}\sigma^2}{NK}\right).$$

When $\theta = 0$, both sides are zero. The smoothness term also satisfies

$$\frac{\beta}{2} D_r \le 3\beta h^2\left(G_{r-1} + A_{r-1} + U_r + \frac{\delta_{\max}\sigma^2}{NK}\right).$$

Substitute these estimates and the local-bias, coordinate-change, and gradient-change bounds into (A.34). The two Young inequalities consume $\gamma_r h G_{r-1}/16$ each, giving

$$-\gamma_r h + \frac{\gamma_r h}{16} + \frac{\gamma_r h}{16} = -\frac{7\gamma_r h}{8}.$$

The remaining terms share the factor

$$\beta h^2\left(G_{r-1} + A_{r-1} + U_r + \frac{\delta_{\max}\sigma^2}{NK}\right).$$

Their coefficients sum to

$$\underbrace{3}_{\text{smoothness}} + \underbrace{\sqrt{\theta}}_{\text{coordinate change}} + \underbrace{5\sqrt{\theta}}_{\text{tracking}} + \underbrace{3\sqrt{\theta/\delta_{\max}}}_{\text{noise}} = \Lambda.$$

The local-bias remainder is $4h\delta_{\max}^2 U_r/\delta$. Combining these terms gives exactly (A.32). □

A.7 CLOSING THE COUPLED DESCENT BOUND

Add tracking energy to the descent potential:

$$\Psi_r^{(R)} = \Phi_r + \frac{16\beta h^2 \Lambda}{\delta} A_r. \tag{A.39}$$

Lemma A.11 (Coupled-potential descent)**.** *For every* $1 \le r \le R$*,*

$$\Psi_r^{(R)} - \Psi_{r-1}^{(R)} \le -\frac{\delta h}{2} G_{r-1} + \frac{90\delta_{\max}^2 \beta^2 h^3 \sigma^2}{\delta \eta_g^2 K} + \frac{2\beta h^2 \Lambda \delta_{\max} \sigma^2}{NK}. \tag{A.40}$$

Proof. The tracking term in $\Psi_r^{(R)}$ supplies a negative multiple of A_{r-1} that pays for the tracking error in the bound for the descent potential. We check each coefficient explicitly.

Add the two one-round inequalities. Subtracting A_{r-1} from (A.26) gives

$$A_r - A_{r-1} \le -\frac{\delta}{3} A_{r-1} + 4G_{r-1} + 4U_r + \frac{72(\beta h)^2}{\delta} \frac{\delta_{\max} \sigma^2}{NK}.$$

Multiply this inequality by $16\beta h^2 \Lambda/\delta$ and add (A.32). Dividing the result by $h > 0$ gives

$$\begin{aligned} \frac{\Psi_r^{(R)} - \Psi_{r-1}^{(R)}}{h} \le & -\frac{7\gamma_r}{8} G_{r-1} + \frac{4\delta_{\max}^2}{\delta} U_r \\ & + \beta h \Lambda \left(G_{r-1} + A_{r-1} + U_r + \frac{\delta_{\max} \sigma^2}{NK} \right) \\ & + \frac{16\beta h \Lambda}{\delta} \left(-\frac{\delta}{3} A_{r-1} + 4G_{r-1} + 4U_r + \frac{72(\beta h)^2}{\delta} \frac{\delta_{\max} \sigma^2}{NK} \right). \end{aligned}$$

Collecting the coefficient of each nonnegative quantity yields

$$\begin{aligned} \frac{\Psi_r^{(R)} - \Psi_{r-1}^{(R)}}{h} \le & \left[-\frac{7\gamma_r}{8} + \beta h \Lambda \left(1 + \frac{64}{\delta} \right) \right] G_{r-1} - \frac{13\beta h \Lambda}{3} A_{r-1} \\ & + \left[\frac{4\delta_{\max}^2}{\delta} + \beta h \Lambda \left(1 + \frac{64}{\delta} \right) \right] U_r \\ & + \beta h \Lambda \left(1 + \frac{1152(\beta h)^2}{\delta^2} \right) \frac{\delta_{\max} \sigma^2}{NK}. \end{aligned}$$

In particular, the tracking coefficient is $\beta h \Lambda - 16\beta h \Lambda/3 = -13\beta h \Lambda/3$.

Bound the positive coefficients involving Λ. The step-size condition (A.1) gives $\beta h \Lambda \le \delta^2/200$ directly. Hence

$$\beta h \Lambda \left(1 + \frac{64}{\delta} \right) \le \frac{\delta^2 + 64\delta}{200} \le \frac{13\delta}{40} \le \delta \le \frac{\delta_{\max}^2}{\delta}.$$

Thus the coefficient multiplying U_r in the combined one-round inequality is at most $5\delta_{\max}^2/\delta$.

Substitute the local-drift bound. Lemma A.8 now gives

$$\frac{5\delta_{\max}^2}{\delta} U_r \le \frac{90\delta_{\max}^2 (\beta h)^2}{\delta \eta_g^2} \left(G_{r-1} + A_{r-1} + \frac{\sigma^2}{K} \right).$$

Substitution into the combined one-round inequality adds $90\delta_{\max}^2 (\beta h)^2/(\delta \eta_g^2)$ to each of its gradient and tracking coefficients. It also contributes this same coefficient times σ^2/K to the noise terms. We check these three contributions separately below.

Absorb tracking error and retain gradient descent. Using $\eta_g \ge \delta$, the step-size cap, and $\Lambda^2 \ge 9\delta_{\max}/\delta$ gives

$$\frac{90\delta_{\max}^2 (\beta h)^2}{\delta \eta_g^2} \le \beta h \Lambda \frac{90\delta_{\max}^2}{200\delta \Lambda^2} \le \frac{\delta_{\max}}{20} \beta h \Lambda.$$

Therefore,

$$-\frac{13\beta h\Lambda}{3}+\frac{90\delta_{\max}^2(\beta h)^2}{\delta\eta_g^2}\le\beta h\Lambda\left(-\frac{13}{3}+\frac{\delta_{\max}}{20}\right)\le 0.$$

We may consequently discard the entire A_{r-1} term. For the gradient coefficient, use $\eta_g\ge\delta$, $\delta_{\max}\le 1$, and $\beta h\le\delta^2/600$ to obtain

$$\frac{90\delta_{\max}^2(\beta h)^2}{\delta\eta_g^2}\le\frac{90(\beta h)^2}{\delta^3}\le\frac{90\delta}{600^2}\le\frac{\delta}{4000}.$$

Together with the bound $\beta h\Lambda(1+64/\delta)\le 13\delta/40$, this gives

$$\beta h\Lambda\left(1+\frac{64}{\delta}\right)+\frac{90\delta_{\max}^2(\beta h)^2}{\delta\eta_g^2}\le\left(\frac{13}{40}+\frac{1}{4000}\right)\delta\le\frac{3\delta}{8}.$$

Because $\gamma_r\ge\delta$, the full gradient coefficient is at most $-7\gamma_r/8+3\delta/8\le-\delta/2$.

Restore the scale h and bound the noise terms. Multiplying the noise term introduced by the local-drift bound by h gives exactly

$$h\,\frac{90\delta_{\max}^2(\beta h)^2}{\delta\eta_g^2}\,\frac{\sigma^2}{K}=\frac{90\delta_{\max}^2\beta^2h^3\sigma^2}{\delta\eta_g^2K}.$$

For the remaining noise term, the step-size cap implies

$$\frac{1152(\beta h)^2}{\delta^2}\le\frac{1152\delta^2}{600^2}\le 1.$$

Multiplying the remaining noise term in the combined one-round inequality by h, we obtain

$$\beta h^2\Lambda\left(1+\frac{1152(\beta h)^2}{\delta^2}\right)\frac{\delta_{\max}\sigma^2}{NK}\le\frac{2\beta h^2\Lambda\delta_{\max}\sigma^2}{NK}.$$

Together with the gradient coefficient bound $-\delta/2$ and the nonpositive tracking coefficient, these are precisely the three terms in (A.40). □

Lemma A.12 (Endpoint bounds)**.** *For the fixed horizon R,*

$$\Psi_0^{(R)}\le\frac{3}{2}F_0+\beta h^2\left(\theta+\frac{16\Lambda}{\delta}\right)\mathcal{E}_0, \tag{A.41}$$

$$\Psi_R^{(R)}=F_R+\frac{16\beta h^2\Lambda}{\delta}A_R\ge 0. \tag{A.42}$$

Proof. Control the initial potential. Since $n^0=0$, we have $k^0=t^0$ and $A_0=\mathcal{E}_0$. To control the added linear term, first apply smoothness at the trial point $x-\nabla f(x)/\beta$:

$$f_{\inf}\le f\left(x-\frac{\nabla f(x)}{\beta}\right)\le f(x)-\frac{1}{2\beta}\|\nabla f(x)\|^2.$$

Thus $G_0\le 2\beta F_0$. The weight bound also gives $\mathbb{E}\|L_0(t^0)\|^2\le\theta\mathcal{E}_0$. Cauchy–Schwarz in expectation, followed by Young's inequality, yields

$$\begin{aligned}|\Phi_0-F_0|&=h\left|\mathbb{E}\langle\nabla f(x^0),L_0(t^0)\rangle\right|\\&\le h\sqrt{G_0}\sqrt{\mathbb{E}\|L_0(t^0)\|^2}\le h\sqrt{2\beta F_0\theta\mathcal{E}_0}\\&\le\frac{1}{2}F_0+\beta h^2\theta\mathcal{E}_0.\end{aligned}$$

Substituting this into the definition of the potential gives

$$\begin{aligned}\Psi_0^{(R)}&\le F_0+|\Phi_0-F_0|+\frac{16\beta h^2\Lambda}{\delta}\mathcal{E}_0\\&\le\frac{3}{2}F_0+\beta h^2\left(\theta+\frac{16\Lambda}{\delta}\right)\mathcal{E}_0,\end{aligned}$$

which is (A.41).

Control the terminal potential. The terminal weights are $w_i^R = 1$, so $L_R(v) = 0$ for every client matrix v. Hence $z^R = x^R$, $\Phi_R = F_R$, and

$$\Psi_R^{(R)} = F_R + \frac{16\beta h^2 \Lambda}{\delta} A_R \geq 0,$$

because $f(x^R) \geq f_{\inf}$ and A_R is an average of squared norms. This proves (A.42). Only this terminal nonnegativity is needed for telescoping; no sign condition on the intermediate linear terms is required. □

A.8 Stationarity and step-size consequences

Theorem A.1 (Finite-time stationarity). *Under the setup of Appendix A.1, for every $R \geq 1$,*

$$\begin{aligned}\overline{G}_R \leq \frac{3F_0}{\delta h R} + \frac{2\beta h}{\delta R}\left(\theta + \frac{16\Lambda}{\delta}\right)\mathcal{E}_0 \\ + \frac{4\beta h \Lambda \delta_{\max}\sigma^2}{N\delta K}\left(1 + \frac{45\beta h \delta_{\max} N}{\Lambda \delta \eta_g^2}\right).\end{aligned} \tag{A.43}$$

Proof. Telescope the coupled potential. Rearrange (A.40) and sum over $r = 1, \ldots, R$:

$$\begin{aligned}\frac{\delta h}{2}\sum_{r=1}^{R} G_{r-1} &\leq \sum_{r=1}^{R}(\Psi_{r-1}^{(R)} - \Psi_r^{(R)}) \\ &\quad + R\left(\frac{90\delta_{\max}^2\beta^2 h^3\sigma^2}{\delta\eta_g^2 K} + \frac{2\beta h^2\Lambda\delta_{\max}\sigma^2}{NK}\right) \\ &= \Psi_0^{(R)} - \Psi_R^{(R)} + R\left(\frac{90\delta_{\max}^2\beta^2 h^3\sigma^2}{\delta\eta_g^2 K} + \frac{2\beta h^2\Lambda\delta_{\max}\sigma^2}{NK}\right).\end{aligned}$$

Lemma A.12 bounds $\Psi_0^{(R)}$ and gives $\Psi_R^{(R)} \geq 0$. Substituting these bounds yields

$$\begin{aligned}\frac{\delta h}{2}\sum_{r=0}^{R-1} G_r \leq \frac{3}{2}F_0 + \beta h^2\left(\theta + \frac{16\Lambda}{\delta}\right)\mathcal{E}_0 \\ + R\left(\frac{90\delta_{\max}^2\beta^2 h^3\sigma^2}{\delta\eta_g^2 K} + \frac{2\beta h^2\Lambda\delta_{\max}\sigma^2}{NK}\right).\end{aligned}$$

Divide by the descent coefficient. Divide by $\delta h R/2$ and use $\overline{G}_R = R^{-1}\sum_{r=0}^{R-1} G_r$ to obtain

$$\begin{aligned}\overline{G}_R \leq \frac{3F_0}{\delta h R} + \frac{2\beta h}{\delta R}\left(\theta + \frac{16\Lambda}{\delta}\right)\mathcal{E}_0 \\ + \frac{180\beta^2 h^2\delta_{\max}^2\sigma^2}{\delta^2\eta_g^2 K} + \frac{4\beta h\Lambda\delta_{\max}\sigma^2}{N\delta K}.\end{aligned}$$

The higher-order noise coefficient factors as

$$\frac{180\beta^2 h^2\delta_{\max}^2\sigma^2}{\delta^2\eta_g^2 K} = \frac{4\beta h\Lambda\delta_{\max}\sigma^2}{N\delta K}\,\frac{45\beta h\delta_{\max} N}{\Lambda\delta\eta_g^2}.$$

This proves (A.43). Since $G_r = \mathbb{E}\|\nabla f(x^r)\|^2$, this is stationarity of the actual server iterates; no separate convergence claim for z is needed. □

Corollary A.1 (Convergence rate in communication rounds). *Under the assumptions and initialization conditions of Appendix A.1, choose the step sizes of Corollary 1. Then the bound (6.2) holds for every $R \geq 1$.*

Proof. Check the step sizes and control stochastic local drift. The chosen steps satisfy $\eta_g = \delta\sqrt{N} \geq \delta$ and

$$h = K\eta_g\eta_\ell = \min\left\{\frac{1}{\beta}\sqrt{\frac{NK}{\delta R}}, \frac{\delta^2}{200\beta\Lambda}\right\}.$$

Thus (A.1) holds. Using the step-size cap and $\Lambda^2 \geq 9\delta_{\max}/\delta$, the extra noise factor satisfies

$$\frac{45\beta h\delta_{\max}N}{\Lambda\delta\eta_g^2} = \frac{45\beta h\delta_{\max}}{\Lambda\delta^3} \leq \frac{45\delta_{\max}}{200\Lambda^2\delta} \leq \frac{1}{40} \leq 1.$$

Consequently, Theorem A.1 gives

$$\overline{G}_R \leq \frac{3F_0}{\delta hR} + \frac{2\beta h}{\delta R}\left(\theta + \frac{16\Lambda}{\delta}\right)\mathcal{E}_0 + \frac{8\beta h\Lambda\delta_{\max}\sigma^2}{N\delta K}.$$

Express each contribution in terms of the horizon. The reciprocal of the chosen step obeys

$$\frac{1}{h} = \max\left\{\beta\sqrt{\frac{\delta R}{NK}}, \frac{200\beta\Lambda}{\delta^2}\right\} \leq \beta\sqrt{\frac{\delta R}{NK}} + \frac{200\beta\Lambda}{\delta^2}.$$

It follows that

$$\frac{3F_0}{\delta hR} \leq \frac{3\beta F_0}{\sqrt{\delta NKR}} + \frac{600\beta\Lambda F_0}{\delta^3 R}.$$

For the noise term, use $\beta h \leq \sqrt{NK/(\delta R)}$ to obtain

$$\frac{8\beta h\Lambda\delta_{\max}\sigma^2}{N\delta K} \leq \frac{8\Lambda\delta_{\max}\sigma^2}{\delta^{3/2}\sqrt{NKR}}.$$

For the initial tracking term, use the cap $\beta h \leq \delta^2/(200\Lambda)$:

$$\frac{2\beta h}{\delta R}\left(\theta + \frac{16\Lambda}{\delta}\right)\mathcal{E}_0 \leq \frac{\delta\theta + 16\Lambda}{100\Lambda}\frac{\mathcal{E}_0}{R}.$$

Combining these three estimates proves (6.2) and Corollary 1. Substituting $\delta_{\max} = \delta$, $\theta = 0$, and $\Lambda = 3$ gives the rate stated in the uniform-participation remark. □

A.9 Averaged tracking and trajectory separation

For completeness, we derive the averaged separation bound from the tracking recursion. This is an additional consequence; the stationarity proof above uses the sharper noise projections instead. Write $\overline{\mathcal{E}}_R = R^{-1}\sum_{r=0}^{R-1}\mathcal{E}_r$.

Lemma A.13 (Averaged tracking)**.** *For every horizon $R \geq 1$,*

$$\overline{\mathcal{E}}_R \leq \frac{12\mathcal{E}_0}{\delta R} + \frac{60}{\delta}\overline{G}_R + \frac{3\sigma^2}{K}. \tag{A.44}$$

Proof. Eliminate local drift from the tracking recursion. Lemma A.8 bounds U_r by $18(\beta h)^2(G_{r-1} + A_{r-1} + \sigma^2/K)/\eta_g^2$. Substituting it into (A.26) gives

$$A_r \leq \left(1 - \frac{\delta}{3} + \frac{72(\beta h)^2}{\eta_g^2}\right)A_{r-1} + \left(4 + \frac{72(\beta h)^2}{\eta_g^2}\right)G_{r-1} + \frac{72(\beta h)^2\sigma^2}{\eta_g^2 K} + \frac{72(\beta h)^2\delta_{\max}\sigma^2}{\delta NK}.$$

The step conditions imply

$$\frac{72(\beta h)^2}{\eta_g^2} \leq \frac{72\delta^2}{600^2} \leq \frac{\delta}{6} \leq 1.$$

Bounding the tracking and gradient coefficients separately yields

$$A_r \leq \left(1 - \frac{\delta}{6}\right)A_{r-1} + 5G_{r-1} + \frac{72(\beta h)^2\sigma^2}{\eta_g^2 K} + \frac{72(\beta h)^2\delta_{\max}\sigma^2}{\delta NK}.$$

Sum the tracking contraction. Rearranging the preceding inequality and summing gives

$$\frac{\delta}{6}\sum_{r=1}^{R} A_{r-1} \le A_0 - A_R + 5\sum_{r=1}^{R} G_{r-1} + R\left(\frac{72(\beta h)^2\sigma^2}{\eta_g^2 K} + \frac{72(\beta h)^2\delta_{\max}\sigma^2}{\delta N K}\right).$$

Use $A_R \ge 0$, $A_0 = \mathcal{E}_0$, and divide by $\delta R/6$:

$$\frac{1}{R}\sum_{r=0}^{R-1} A_r \le \frac{6\mathcal{E}_0}{\delta R} + \frac{30}{\delta}\overline{G}_R + \frac{432(\beta h)^2\sigma^2}{\delta\eta_g^2 K} + \frac{432(\beta h)^2\delta_{\max}\sigma^2}{\delta^2 N K}.$$

Restore the retained-noise energy. The decomposition $t^r = k^r + n^r$ and Lemma A.5 imply $\mathcal{E}_r \le 2A_r + \frac{2\sigma^2}{K}$. Therefore

$$\overline{\mathcal{E}}_R \le \frac{12\mathcal{E}_0}{\delta R} + \frac{60}{\delta}\overline{G}_R + \left(2 + \frac{864(\beta h)^2}{\delta\eta_g^2} + \frac{864(\beta h)^2\delta_{\max}}{\delta^2 N}\right)\frac{\sigma^2}{K}.$$

The coefficient of $\frac{\sigma^2}{K}$ is at most three, since

$$\frac{864(\beta h)^2}{\delta\eta_g^2} + \frac{864(\beta h)^2\delta_{\max}}{\delta^2 N} \le 864(\beta h)^2\left(\frac{1}{\delta^3} + \frac{1}{\delta^2}\right) \le \frac{864}{600^2}(\delta + \delta^2) < 1.$$

This proves (A.44) without assuming that $\mathcal{E}_r$ vanishes. □

Trajectory comparison. Since $\mathcal{E}_r = (Nh^2)^{-1}\sum_i \mathbb{E}\|q_i^r - x^r\|^2$, the preceding lemma directly bounds the shifted-memory gap:

$$\frac{1}{NR}\sum_{r=0}^{R-1}\sum_i \mathbb{E}\|q_i^r - x^r\|^2 = h^2\overline{\mathcal{E}}_R \le h^2\left(\frac{12\mathcal{E}_0}{\delta R} + \frac{60}{\delta}\overline{G}_R + \frac{3\sigma^2}{K}\right). \tag{A.45}$$

The pointwise bound (A.15) also gives

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\|z^r - x^r\|^2 \le \theta h^2\frac{1}{R}\sum_{r=0}^{R-1}\mathcal{E}_r \le \theta h^2\left(\frac{12\mathcal{E}_0}{\delta R} + \frac{60}{\delta}\overline{G}_R + \frac{3\sigma^2}{K}\right). \tag{A.46}$$

Meaning of a vanishing average gap. For fixed problem parameters and fixed finite $F_0, \mathcal{E}_0$, choose $h_R = \min\{\delta^2/(200\beta\Lambda), \kappa/\sqrt{R}\}$ for any fixed $\kappa > 0$, hold this step constant within the R-round run, and keep η_g fixed with $\eta_g \ge \delta$. In (A.43), the objective, initialization, leading-noise, and higher-order-noise terms are respectively $O(R^{-1/2})$, $O(R^{-3/2})$, $O(R^{-1/2})$, and $O(R^{-1})$. Hence $\overline{G}_R = O(R^{-1/2})$. Applying (A.45) and (A.15) gives

$$\frac{1}{NR}\sum_{r=0}^{R-1}\sum_i \mathbb{E}\left\|q_i^r - x^r\right\|^2 \le h_R^2\left(\frac{12\mathcal{E}_0}{\delta R} + \frac{60}{\delta}\overline{G}_R + \frac{3\sigma^2}{K}\right) = O(R^{-1}),$$

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\|z^r - x^r\right\|^2 \le \frac{\theta}{NR}\sum_{r=0}^{R-1}\sum_i \mathbb{E}\left\|q_i^r - x^r\right\|^2 = O(R^{-1}).$$

This comparison does not require $\mathcal{E}_0 = 0$; for example, $m_i^0 = x^0$ gives $\mathcal{E}_0 = N^{-1}\sum_i \left\|b_i(x^0)\right\|^2$, independently of h_R for a fixed deterministic x^0. The asymptotic statement compares horizon-dependent runs and their horizon-dependent analytical trajectories. It does not assert pointwise decay for a single noisy run with fixed h, and it is a consequence of the convergence theorem rather than a premise of its proof.

B EXPERIMENTAL SETUP

B.1 CLIENT PARTICIPATION SCENARIOS

We evaluate five client participation scenarios spanning stationary and non-stationary availability models. In each communication round $t \geq 1$, each client $i \in \{1, \ldots, N\}$ participates independently with probability

$$p_i^t = p_i \, f_i(t), \tag{B.1}$$

where $p_i \in (0, 1]$ represents the client's baseline participation rate and $f_i(t) \geq 0$ is a temporal modulation function satisfying $\frac{1}{P}\sum_{t=1}^{P} f_i(t) \approx 1$ over an oscillation period P. The non-stationary patterns are parameterized by non-stationarity period P, fluctuation amplitude $\gamma \in [0, 1]$, and client-specific initial phase $\phi_i \in [0, 2\pi)$.

1. Uniform Participation. All clients share an identical, time-invariant participation probability across all rounds:

$$p_i^t = \delta \quad \forall i \in \{1, \ldots, N\}, \ \forall t \geq 1, \tag{B.2}$$

corresponding to $f_i(t) = 1$ and $p_i = \delta$ (we set $\delta = 0.3$ across all datasets).

2. Stationary (Heterogeneous) Participation. Baseline participation probabilities p_i vary across clients but remain constant across rounds:

$$f_i(t) = 1 \implies p_i^t = p_i \quad \forall t \geq 1, \tag{B.3}$$

where p_i are sampled to reflect persistently unequal device availability while maintaining an average active client fraction of $\frac{1}{N}\sum_{i=1}^{N} p_i = \delta$.

3. Staircase Participation (Non-Stationary). Availability modulates in discrete steps, alternating between a high-availability half-period $(1 + \gamma)$ and a low-availability half-period $(1 - \gamma)$:

$$f_i(t) = \begin{cases} 1 + \gamma, & \text{if } \theta_i(t) \in [0, \pi), \\ 1 - \gamma, & \text{if } \theta_i(t) \in [\pi, 2\pi), \end{cases} \tag{B.4}$$

where $\theta_i(t) = \left(\frac{2\pi t}{P} + \phi_i\right) \bmod 2\pi$ is the client-specific phase angle.

4. Sine Participation (Non-Stationary). Availability follows a smooth sinusoidal wave around the baseline rate:

$$f_i(t) = 1 + \gamma \sin\left(\frac{2\pi t}{P} + \phi_i\right), \tag{B.5}$$

where client-specific phase offsets ϕ_i stagger peak availability across clients.

5. Exponential Participation (Non-Stationary). Availability follows an asymmetric, non-linear trajectory within each period P, featuring a sharp exponential decay followed by an exponential recovery:

$$f_i(t) = \begin{cases} f_{\text{decay}}\left(\frac{\theta_i(t)}{\pi}\right), & \text{if } \theta_i(t) \in [0, \pi), \\ f_{\text{rise}}\left(\frac{\theta_i(t) - \pi}{\pi}\right), & \text{if } \theta_i(t) \in [\pi, 2\pi), \end{cases} \tag{B.6}$$

where $\theta_i(t) = \left(\frac{2\pi t}{P} + \phi_i\right) \bmod 2\pi$ denotes the phase angle for client i at round t. The decay and rise profile functions over normalized local parameter $s \in [0, 1)$ are defined by:

$$f_{\text{decay}}(s) = 1 + \gamma\big(2w(s) - 1\big), \tag{B.7}$$

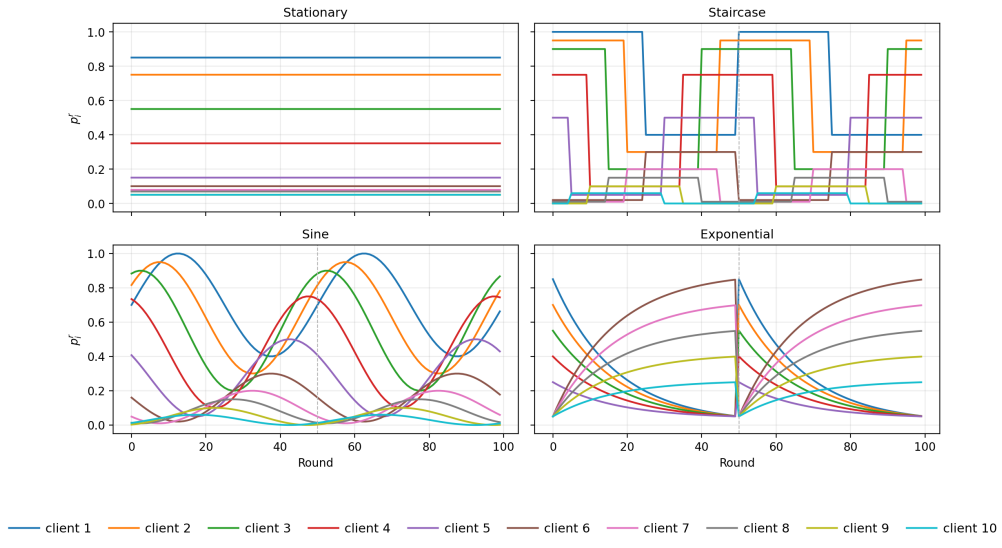
$$f_{\text{rise}}(s) = 1 - \gamma\big(2w(s) - 1\big), \tag{B.8}$$

with the normalized exponential kernel

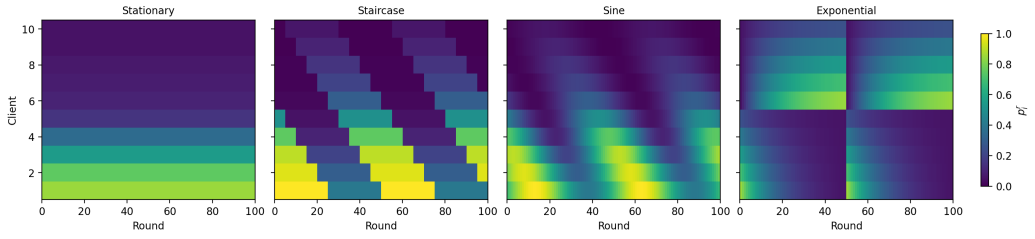
$$w(s) = \frac{e^{-3s} - e^{-3}}{1 - e^{-3}}, \quad s \in [0, 1). \tag{B.9}$$

Here $w(0) = 1$ and $\lim_{s \to 1} w(s) = 0$. Consequently, $f_{\text{decay}}(s)$ decreases exponentially from $1 + \gamma$ down to $1 - \gamma$ over the first half-period, while $f_{\text{rise}}(s)$ increases exponentially from $1 - \gamma$ back to $1 + \gamma$ over the second half-period. The client-specific initial phase ϕ_i staggers these decay–recovery cycles across devices to model asynchronous, non-stationary participation dynamics.

Figure 3 illustrates these participation scenarios under our default benchmark configuration ($P = 50$ rounds and fluctuation amplitude $\gamma = 0.3$) across 10 clients. Specifically, Figure 3(a) shows the individual participation probability trajectories p_i^t over communication rounds for the four heterogeneous settings (Stationary, Staircase, Sine, and Exponential), while Figure 3(b) depicts the corresponding spatio-temporal availability heatmaps across clients and rounds.



(a) Per-client participation probability trajectories p_i^t over communication rounds



(b) Client participation probability heatmaps across clients (vertical axis) and rounds (horizontal axis)

Figure 3: Visualizations of client participation probabilities over communication rounds for the benchmark configurations ($P = 50, \gamma = 0.3$) across 10 clients under Stationary, Staircase, Sine, and Exponential participation patterns. Top: per-client availability trajectories p_i^t. Bottom: corresponding availability heatmaps across clients and rounds.

B.2 Data Heterogeneity and Dirichlet Partitioning

To evaluate federated optimization under realistic non-i.i.d. data distributions, we synthesize label heterogeneity using a symmetric Dirichlet distribution $\mathrm{Dir}(\alpha)$ following standard federated benchmark protocols (Hsu et al., 2019). For each dataset with C classes and $N = 10$ clients:

1. For each class $c \in \{1, \ldots, C\}$, we draw a client proportion vector $\mathbf{q}_c = (q_{c,1}, \ldots, q_{c,N}) \sim \mathrm{Dir}(\alpha \mathbf{1}_N)$, where $\sum_{i=1}^{N} q_{c,i} = 1$.

2. We then assign a fraction $q_{c,i}$ of the training samples of class c to client i, ensuring disjoint partitions across all clients.

The concentration parameter $\alpha > 0$ controls the severity of label skew:

- **Near-i.i.d. regime** ($\alpha = 10$)**:** Each client receives an approximately uniform proportion of samples from all C classes, minimizing local client drift.

- **Moderate heterogeneity** ($\alpha = 1.0$ **and** $\alpha = 0.1$)**:** Client class proportions begin to diverge substantially, producing distinct local loss landscapes.

- **Severe non-i.i.d. regime** ($\alpha = 0.01$)**:** The Dirichlet distribution places almost all mass on one or two clients per class. As a result, each client holds samples from only a small subset of classes (often only 1 or 2 dominant classes for CIFAR-10), inducing severe gradient divergence and extreme client drift.

Our default experimental benchmarks on CIFAR-10, CIFAR-100, and SVHN use $\alpha = 0.01$ to rigorously test algorithm resilience under acute heterogeneity.

Figure 4 illustrates the client-wise sample allocations across classes for our benchmark configurations.

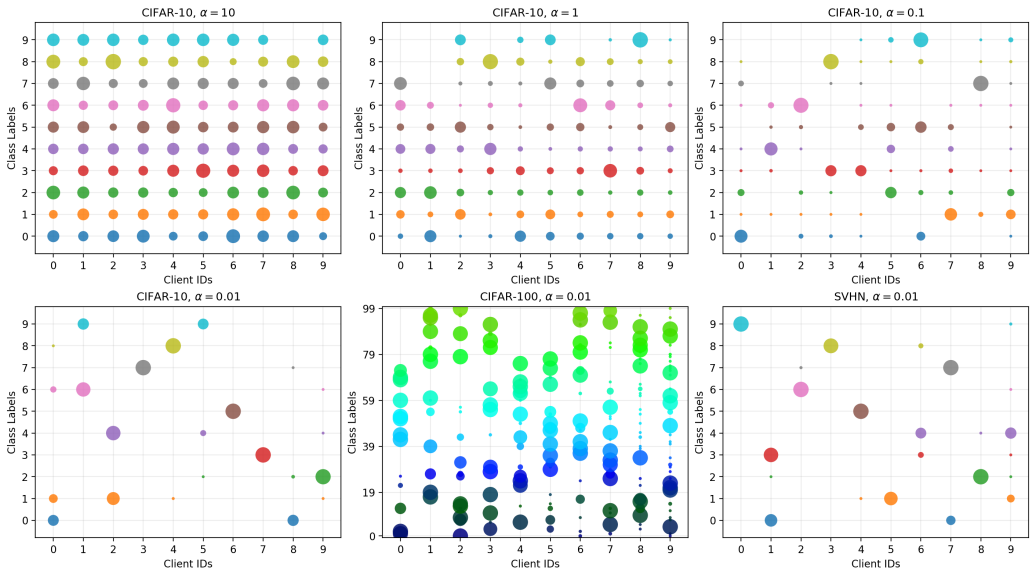


Figure 4: Sample distributions across classes (horizontal axis) for each client $i \in \{0, \ldots, 9\}$ (vertical axis) under Dirichlet partitions $\mathrm{Dir}(\alpha)$. Bubble sizes correspond to the relative number of training samples of each class held by the respective client, illustrating the transition from near-uniform class coverage ($\alpha = 10$) to extreme label concentration ($\alpha = 0.01$).

B.3 TUNED HYPERPARAMETERS

For each algorithm and experimental setting, the local client learning rate $\eta_\ell \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.25\}$ and server learning rate $\eta_g \in \{0.1, 0.25, 0.5, 1.0, 2.5, 5.0\}$ were selected by grid search to maximize average test accuracy over the final 25 rounds. Table 3 reports the tuned (η_ℓ, η_g) hyperparameter configurations across all experimental scenarios and benchmark datasets corresponding to the main results in Table 2.

Table 3: Tuned client learning rate η_ℓ and server learning rate η_g across datasets and client participation scenarios corresponding to Table 2.

Dataset	**Algorithm**	**Uniform**		**Stationary**		**Staircase**		**Sine**		**Exponential**	
		η_ℓ	η_g	η_ℓ	η_g	η_ℓ	η_g	η_ℓ	η_g	η_ℓ	η_g
CIFAR-10 (ResNet-20)	**TRUSS**	0.05	2.5	0.25	1.0	0.1	2.5	0.05	2.5	0.05	2.5
	FedAvg	0.005	1	0.01	1	0.005	1	0.05	1	0.05	1
	FedProx	0.05	1	0.1	1	0.1	1	0.01	1	0.05	1
	FedAU	0.05	0.5	0.05	0.5	0.05	0.5	0.05	0.5	0.05	0.5
	FedAWE	0.005	0.5	0.1	0.5	0.05	0.5	0.1	0.5	0.1	0.5
	MIFA	0.01	1	0.1	1	0.05	1	0.1	1	0.1	1
	SGFocus	0.05	1	0.005	1	0.005	1	0.01	1	0.005	1
	SCAFFOLD	0.05	0.5	0.1	0.5	0.1	0.5	0.1	0.5	0.05	0.5
	SCALLION	0.05	0.5	0.1	0.5	0.05	0.5	0.1	0.5	0.05	0.5
CIFAR-100 (ResNet-56)	**TRUSS**	0.1	2.5	0.25	1.0	0.25	1.0	0.25	1.0	0.25	1.0
	FedAvg	0.01	1	0.01	1	0.05	1	0.1	1	0.05	1
	FedProx	0.1	1	0.01	1	0.1	1	0.1	1	0.1	1
	FedAU	0.05	0.5	0.05	0.5	0.1	0.5	0.05	0.5	0.05	0.5
	FedAWE	0.1	0.5	0.1	0.5	0.1	0.5	0.05	0.5	0.1	0.5
	MIFA	0.01	1	0.05	1	0.05	1	0.01	1	0.05	1
	SGFocus	0.005	1	0.005	1	0.005	1	0.005	1	0.005	1
	SCAFFOLD	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5
	SCALLION	0.1	0.5	0.1	0.5	0.1	0.5	0.05	0.5	0.05	0.5
SVHN (CNN)	**TRUSS**	0.01	1.0	0.01	1.0	0.01	1.0	0.01	1.0	0.01	1.0
	FedAvg	0.001	1	0.001	1	0.001	1	0.001	1	0.001	1
	FedProx	0.001	1	0.001	1	0.001	1	0.001	1	0.001	1
	FedAU	0.01	0.25	0.01	0.5	0.01	0.25	0.01	0.25	0.01	0.25
	FedAWE	0.05	0.25	0.05	0.25	0.05	0.25	0.05	0.25	0.05	0.25
	MIFA	0.005	1	0.01	1	0.005	1	0.01	1	0.01	1
	SGFocus	0.005	1	0.001	1	0.001	1	0.001	1	0.005	1
	SCAFFOLD	0.01	0.25	0.01	0.25	0.01	0.25	0.005	0.5	0.01	0.25
	SCALLION	0.01	0.25	0.01	0.25	0.01	0.25	0.01	0.25	0.01	0.25

B.4 Benchmark Settings and Hardware

Table 4 details the default experimental settings across CIFAR-10, CIFAR-100, and SVHN. All experiments are executed on a distributed cluster on the National Research Platform (NRP) (Weitzel et al., 2025), deployed across 11 nodes spanning distinct U.S. states. Each client is allocated one GPU, two CPU cores, and 5 GiB of memory.

Table 4: Default experimental settings for CIFAR-10, CIFAR-100, and SVHN benchmarks.

Dataset-Specific Settings	
Architecture	ResNet-20 (CIFAR-10), ResNet-56 (CIFAR-100), SimpleCNN* (SVHN)
Train-eval subsample	100% / full (CIFAR-10), 10% (CIFAR-100), 25% (SVHN) of the training set
Shared Benchmark and Federated Settings	
Number of clients	10 (plus 1 server)
Participation patterns	Uniform (p=0.3); stationary, staircase, sine, and exponential (heterogeneous per-client rates; period P=50 rounds for non-stationary patterns)
Data distribution	Non-IID via Dirichlet label partition with α=0.01
Local steps	50
Communication budget	1000 $\times$ model-size
Algorithms	TRUSS, FedAvg, FedProx (μ=0.1), FedAU, FedAWE, MIFA, SGFocus, SCAFFOLD, HC-Scaffold, Scallion
Shared Optimization and Evaluation Parameters	
Optimizer	SGD with momentum
Batch size	64 per client
Momentum	0.9 (Nesterov)
Learning rate	Selected by grid search over local and (where applicable) global learning rates
Learning rate schedule	None (constant after selection)
Weight decay	10^{-4}
Learning rate warm-up	None
Loss function	Cross entropy
Accuracy objective	Top-1 accuracy (evaluated on full test set)
Repetitions	3 seeds
Reported metric	Mean and standard deviation

*SimpleCNN consists of two 3×3 convolutional blocks with max-pooling, followed by fully connected layers to 10 classes.

C Additional Experiments

This section provides numerical results for the ablation experiments described in Section 7.3. All runs evaluate ResNet-20 on CIFAR-10 under the non-stationary exponential availability model (Section 7.2). We investigate empirical sensitivity across three axes:

- **Data heterogeneity (α):** Dirichlet concentration $\alpha \in \{10, 1, 0.1, 0.01\}$ with fixed $P = 50, \gamma = 0.3$ (Table 5).
- **Fluctuation amplitude (γ):** Participation amplitude $\gamma \in \{0.3, 0.6, 0.9\}$ with fixed $\alpha = 0.01, P = 50$ (Table 6).
- **Oscillation period (P):** Participation cycle period $P \in \{50, 100, 200, 300, 400, 500\}$ with fixed $\alpha = 0.01, \gamma = 0.3$ (Table 7).

Table 5: Training and test accuracy (%) of ResNet-20 on CIFAR-10 under exponential client availability across Dirichlet concentration parameters $\alpha \in \{10, 1, 0.1, 0.01\}$ ($P = 50, \gamma = 0.3$). Mean $\pm$ std across seeds; bold and underlined values indicate highest and second-highest accuracy, respectively.

Algorithm	$\alpha = 10$		$\alpha = 1$		$\alpha = 0.1$		$\alpha = 0.01$	
	Train	Test	Train	Test	Train	Test	Train	Test
TRUSS	93.6 ± 0.2	$\underline{89.1} \pm 0.3$	$\mathbf{91.7} \pm 0.3$	$\mathbf{87.6} \pm 0.1$	$\mathbf{85.8} \pm 1.0$	$\mathbf{83.9} \pm 0.9$	$\mathbf{74.4} \pm 2.9$	$\mathbf{73.6} \pm 2.9$
FedAvg	93.8 ± 0.2	88.4 ± 0.2	90.4 ± 0.2	86.4 ± 0.2	64.1 ± 10.5	63.3 ± 10.2	27.0 ± 5.2	27.6 ± 5.7
FedProx	$\mathbf{94.5} \pm 0.0$	88.7 ± 0.1	88.7 ± 1.5	84.3 ± 1.3	70.4 ± 2.5	68.8 ± 2.5	47.2 ± 3.2	46.3 ± 2.9
FedAU	$\underline{94.4} \pm 0.1$	$\mathbf{89.3} \pm 0.1$	89.3 ± 0.4	86.4 ± 0.1	80.7 ± 1.6	78.6 ± 1.7	56.4 ± 3.7	55.3 ± 3.3
FedAWE	80.9 ± 3.0	78.9 ± 3.1	67.9 ± 1.2	67.5 ± 1.2	63.6 ± 0.8	63.1 ± 1.4	28.7 ± 4.8	28.8 ± 5.1
MIFA	92.8 ± 0.5	88.5 ± 0.5	$\underline{91.4} \pm 0.2$	86.4 ± 0.2	$\underline{84.6} \pm 0.4$	$\underline{82.0} \pm 0.5$	$\underline{73.8} \pm 0.9$	$\underline{72.5} \pm 0.4$
SGFocus	31.3 ± 1.6	32.3 ± 1.2	29.3 ± 1.4	29.2 ± 1.2	23.5 ± 1.6	25.1 ± 1.6	15.6 ± 1.4	16.8 ± 1.9
SCAFFOLD	90.9 ± 0.1	87.4 ± 0.3	87.5 ± 0.2	85.2 ± 0.3	80.5 ± 0.9	79.1 ± 0.6	65.4 ± 3.5	65.3 ± 3.1
SCALLION	92.1 ± 0.0	88.6 ± 0.1	89.5 ± 0.2	$\underline{86.7} \pm 0.0$	83.4 ± 0.2	80.9 ± 0.2	67.1 ± 6.2	67.5 ± 5.4

Table 6: Training and test accuracy (%) of ResNet-20 on CIFAR-10 under exponential client availability across non-stationarity amplitudes $\gamma \in \{0.3, 0.6, 0.9\}$ ($\alpha = 0.01, P = 50$). Mean $\pm$ std across seeds; bold and underlined values indicate highest and second-highest accuracy, respectively.

Algorithm	$\gamma = 0.3$		$\gamma = 0.6$		$\gamma = 0.9$	
	Train	Test	Train	Test	Train	Test
TRUSS	$\mathbf{74.4} \pm 2.9$	$\mathbf{73.6} \pm 2.9$	$\mathbf{74.2} \pm 1.6$	$\mathbf{73.1} \pm 1.3$	$\mathbf{72.6} \pm 1.4$	$\mathbf{72.3} \pm 1.0$
FedAvg	27.0 ± 5.2	27.6 ± 5.7	32.8 ± 4.6	32.8 ± 4.3	37.6 ± 5.7	37.4 ± 6.2
FedProx	47.2 ± 3.2	46.3 ± 2.9	38.2 ± 9.9	37.8 ± 9.8	46.1 ± 0.3	45.7 ± 0.7
FedAU	56.4 ± 3.7	55.3 ± 3.3	61.0 ± 1.0	59.5 ± 0.4	66.4 ± 0.3	64.6 ± 0.2
FedAWE	28.7 ± 4.8	28.8 ± 5.1	31.4 ± 9.1	33.2 ± 9.6	43.0 ± 3.0	45.1 ± 3.2
MIFA	$\underline{73.8} \pm 0.9$	$\underline{72.5} \pm 0.4$	$\underline{72.9} \pm 3.2$	$\underline{70.8} \pm 3.0$	$\underline{71.7} \pm 0.0$	$\underline{69.6} \pm 1.0$
SGFocus	15.6 ± 1.4	16.8 ± 1.9	16.1 ± 6.1	16.1 ± 6.1	13.9 ± 3.9	14.5 ± 4.4
SCAFFOLD	65.4 ± 3.5	65.3 ± 3.1	64.0 ± 2.0	64.3 ± 1.3	60.2 ± 1.1	61.6 ± 0.8
SCALLION	67.1 ± 6.2	67.5 ± 5.4	70.1 ± 3.0	69.5 ± 3.0	68.8 ± 2.9	68.9 ± 2.3

Table 7: Training and test accuracy (%) of ResNet-20 on CIFAR-10 under exponential client availability across non-stationarity periods $P \in \{50, 100, 200, 300, 400, 500\}$ ($\alpha = 0.01, \gamma = 0.3$). Mean $\pm$ std across seeds; bold and underlined values indicate highest and second-highest accuracy, respectively.

Algorithm	$P = 50$		$P = 100$		$P = 200$	
	Train	Test	Train	Test	Train	Test
TRUSS	$\mathbf{74.4} \pm 2.9$	$\mathbf{73.6} \pm 2.9$	$\mathbf{75.3} \pm 0.4$	$\mathbf{74.5} \pm 0.6$	$\underline{74.1} \pm 0.6$	$\mathbf{73.1} \pm 0.6$
FedAvg	27.0 ± 5.2	27.6 ± 5.7	39.3 ± 3.3	39.2 ± 2.9	37.2 ± 5.5	37.0 ± 4.2
FedProx	47.2 ± 3.2	46.3 ± 2.9	46.1 ± 2.4	45.5 ± 2.7	40.4 ± 3.2	40.1 ± 2.8
FedAU	56.4 ± 3.7	55.3 ± 3.3	61.4 ± 2.5	60.4 ± 2.2	63.8 ± 0.6	62.4 ± 0.5
FedAWE	28.7 ± 4.8	28.8 ± 5.1	35.2 ± 9.9	35.7 ± 11.0	36.1 ± 4.0	36.0 ± 4.0
MIFA	$\underline{73.8} \pm 0.9$	$\underline{72.5} \pm 0.4$	71.8 ± 1.6	69.8 ± 1.4	$\mathbf{75.5} \pm 1.6$	$\underline{73.1} \pm 1.7$
SGFocus	15.6 ± 1.4	16.8 ± 1.9	15.7 ± 0.1	16.7 ± 0.4	15.9 ± 0.4	16.7 ± 0.9
SCAFFOLD	65.4 ± 3.5	65.3 ± 3.1	68.1 ± 0.3	68.2 ± 0.6	64.6 ± 0.9	65.2 ± 0.9
SCALLION	67.1 ± 6.2	67.5 ± 5.4	$\underline{73.3} \pm 0.2$	$\underline{72.1} \pm 0.1$	73.3 ± 1.6	72.3 ± 1.5
Algorithm	$P = 300$		$P = 400$		$P = 500$	
	Train	Test	Train	Test	Train	Test
TRUSS	$\mathbf{77.3} \pm 1.3$	$\mathbf{75.7} \pm 1.0$	$\mathbf{77.8} \pm 1.1$	$\mathbf{76.5} \pm 1.2$	$\mathbf{77.9} \pm 0.3$	$\mathbf{76.4} \pm 0.7$
FedAvg	30.8 ± 1.2	31.6 ± 0.6	36.9 ± 5.5	37.5 ± 5.7	36.7 ± 1.0	38.3 ± 0.2
FedProx	43.2 ± 2.7	42.1 ± 2.2	49.7 ± 7.4	48.7 ± 6.3	42.6 ± 5.2	42.3 ± 5.4
FedAU	64.0 ± 2.7	62.8 ± 3.2	61.5 ± 1.5	60.2 ± 1.5	62.9 ± 1.8	62.1 ± 1.5
FedAWE	35.3 ± 11.2	36.8 ± 12.1	39.1 ± 2.4	39.4 ± 2.0	30.5 ± 4.2	31.2 ± 4.9
MIFA	$\underline{72.7} \pm 0.8$	$\underline{70.3} \pm 0.8$	$\underline{73.6} \pm 1.1$	$\underline{71.4} \pm 0.3$	$\underline{73.2} \pm 2.1$	$\underline{71.6} \pm 2.5$
SGFocus	18.3 ± 0.5	18.4 ± 0.5	16.0 ± 3.8	16.3 ± 3.8	13.2 ± 3.1	13.3 ± 3.4
SCAFFOLD	65.5 ± 1.3	65.4 ± 1.8	66.2 ± 2.5	66.4 ± 2.6	65.4 ± 2.7	66.2 ± 2.2
SCALLION	69.9 ± 4.6	69.3 ± 5.3	70.0 ± 2.4	69.2 ± 2.0	70.8 ± 1.4	70.4 ± 0.7