JOINING LATE IN DECENTRALIZED FEDERATED LEARNING: A THEORETICAL ANALYSIS AND TWO LOCAL RULES

Anonymous authors
Paper under double-blind review

ABSTRACT

In decentralized federated learning (DFL), a client arriving mid-run changes the objective the network minimizes, and classical reasoning cannot quantify that change, since without a unique minimizer the displacement one would bound is undefined. We ask what that arrival costs, and what any treatment of it can achieve, and answer both without convexity of any kind. We prove two identities, holding under no assumption, that fix its entire geometric effect. It displaces the network average by D/n, where D is the arriving client's distance from the incumbents and n the network size, and adds to the consensus error a *join shock* of order D^2/n. Carried into a post-join stationarity bound with every constant explicit, they yield four terms that separate exactly. Two transients, the inherited gap and the join shock, decay as $1/T'$. Two persistent floors are set by gradient noise and by the spectral gap. The floors involve neither the join time nor the arriving client, so in this bound joining late carries no persistent penalty, only a transient one. The decomposition also fixes what can be acted on, and we act on both terms it leaves open. The shock admits no step size, so only initialization acts on it; averaging with the neighbors' models brings it to the network's existing consensus scale, and to zero at the incumbent average. The bound-minimizing step size depends on an oracle gap no client can evaluate, yet it is computable from a purely local misfit, with n canceling, no schedule beating it in this bound class, and degradation at most $(\sqrt{\kappa}+1/\sqrt{\kappa})/2$ under κ-fold misestimation. Experiments on CIFAR-10/100 and Tiny-ImageNet test each prediction: the identities hold to machine precision, and networks with and without a late joiner reach the same late-window stationarity. The prescribed step, computed from a single local loss gap, costs a factor 2.6 to 3.3 against the grid-optimal constant on the two CIFAR-10 topologies, inside the 4.1 to 7.0 its tolerance permits.

1 INTRODUCTION

Decentralized federated learning (DFL) dispenses with the central server. Clients hold private data, take local stochastic gradient steps, and mix their models with neighbors over a communication graph (Lian et al., 2017; Koloskova et al., 2020; Boyd et al., 2006). The set of participating clients is seldom fixed in advance. In cross-device deployments devices become eligible as they come online and meet the conditions for participation (Bonawitz et al., 2019; Kairouz et al., 2021), and in cross-silo deployments new institutions are onboarded during a training campaign rather than before it (Warnat-Herresthal et al., 2021; Kairouz et al., 2021). A client that begins participating at some round $\tau_k > 0$ is therefore the ordinary case rather than a corner case.

Nothing in a standard DFL protocol treats such an arrival as a distinct event, since a client entering the mixing graph is averaged into the network as an ordinary consequence of the update rule (Lian et al., 2017; Sun et al., 2023a; De Vos et al., 2023). Whether that is adequate has not been established. What makes the arrival more than bookkeeping is that it changes the optimization problem. For $t < \tau_k$ the incumbents descend $f_{n-1} = \frac{1}{n-1}\sum_{i\neq k} f_i$. From $t = \tau_k$ the network descends $f_n = \frac{1}{n}\sum_i f_i$. Two distinct things happen at that instant. The iterates are a good solution to a problem the network has stopped solving, and, through a separate channel, the arriving model sits at some distance from the incumbent consensus and perturbs the network's geometry. Neither is what partial

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.

participation and client sampling describe (McMahan et al., 2017; Li et al., 2020; Cho et al., 2022), which randomize who contributes while holding the objective fixed, nor staleness (Lian et al., 2018; Liu et al., 2024), which delays *gradients* for a fixed objective. Here the objective itself changes, and no participation or staleness constant absorbs that (Section 6).

Quantifying that change requires something to measure it against. Under convexity there is a single minimum value, a run is certified by the gap to it, and a change of objective becomes a displacement between two minimizers that can be bounded (Karimi et al., 2016). That is the route the two existing analyses of a mid-run arrival take, and it is why both are stated under strong convexity (Ruan et al., 2021; Sun et al., 2023b). A deep network offers no such reference point. Non-convex analysis therefore certifies a weaker property that survives its absence, namely that the run approaches stationarity, measured by the average squared gradient norm along the trajectory (Ghadimi & Lan, 2013), a criterion matched by the known lower bounds (Arjevani et al., 2023) and the one against which decentralized SGD is stated (Lian et al., 2017; Koloskova et al., 2020). We adopt it, and in place of the displacement convexity would have supplied we use the one the arrival imposes, which is defined whatever the landscape (Theorem 1).

The criterion is weaker in one respect that governs everything below. It certifies that the network approaches a critical point of the new objective and says nothing about the value attained there, so it supports no claim about accuracy or generalization. That is a property of the criterion rather than of our analysis, and the unified treatment we work against reports no test error at all (Koloskova et al., 2020). Accuracy is measured throughout Section 5, labeled observational, and never derived from a bound, so the two columns need not agree and we report both where they do not.

Theorem 1 computes that displacement exactly. An arriving client whose initial model sits at distance D_k from the incumbent average moves the network average by D_k/n, and adds $\frac{n-1}{n^2}D_k^2$ to the consensus error the network carries. Both statements are identities rather than bounds and hold under no assumption whatever, not smoothness, not bounded noise, not a connected graph. An implementation can therefore check them to floating-point precision at its first join event, before any downstream number is reproduced, which is the first thing Section 5 does.

We analyze this event directly, without convexity of any kind, under the assumptions standard for DFL. They are smoothness, unbiased gradient noise of bounded variance, bounded heterogeneity, and a mixing graph that contracts at every round. Our central result (Theorem 2) bounds the criterion over the window following the join with every constant explicit, and its four terms separate exactly into two that fade as that window lengthens and two that do not. The first two are the gap inherited at the switch and the *join shock*, which we call *transients*. The other two are fixed by gradient noise and by the graph, and we call them *persistent floors*. Those floors are precisely what a network with no late joiner already pays, so in this bound *joining late carries no persistent penalty*.

We give the first such analysis under assumptions appropriate to deep networks. Our contributions are the following.

1. We use that separation to delimit what any treatment of a late join can achieve (Section 3.3). The join shock admits no step size, the step size minimizing the remaining transient depends on a quantity no client can evaluate, and the floors can be neither attributed to the join nor removed. Two terms remain open, and we characterize the optimum of each.
2. We prove exact identities for the join event and use them to eliminate part of that cost (Sections 3.1 and 4.1). Averaging the arriving model with its neighbors' brings the displacement it imposes down to the consensus scale the network already carries, and to zero at the incumbent average.
3. We derive the step size the bound prescribes and make it computable (Section 4.2). No schedule improves on a correctly chosen constant, and that constant depends on the loss the joining client incurs on the model it inherits rather than on how different its data is. Where the incumbents have plateaued the network size cancels, a branch our sparse topologies never enter (Appendix K.3).
4. We test every falsifiable prediction on three image benchmarks (Section 5). The identities hold to machine precision, the n^{-1} suppression of the shock is confirmed over a 10-fold range of network size, and a network with no late joiner reaches the same late-window stationarity as one that admits a client mid-run. Each prediction is stated with the condition, fixed in advance, under which it would have counted as refuted.

2 PROBLEM FORMULATION

2.1 DECENTRALIZED SGD WITH A LATE JOIN

Clients $[n] = \{1, \dots, n\}$ hold differentiable local objectives $f_i : \mathbb{R}^d \to \mathbb{R}$. Write $f_n := \frac{1}{n}\sum_{i=1}^{n} f_i$ and $f_{n-1} := \frac{1}{n-1}\sum_{i\neq k} f_i$ for the objectives of the full network and of the incumbents, and $f_n^* := \inf_\theta f_n(\theta)$, with f_{n-1}^* and f_k^* defined analogously. Client k joins at round τ_k, so the active set is $\mathcal{A}^t = [n] \setminus \{k\}$ for $t < \tau_k$ and $\mathcal{A}^t = [n]$ for $t \geq \tau_k$. One round of decentralized SGD is

$$\theta_i^{t+1} = \sum_{j\in\mathcal{A}^t} W_{ij}^t\big(\theta_j^t - \eta\, g_j^t\big), \qquad i \in \mathcal{A}^t, \tag{1}$$

where g_j^t is a stochastic gradient of f_j evaluated at θ_j^t and $W^t \in \mathbb{R}^{|\mathcal{A}^t|\times|\mathcal{A}^t|}$ is doubly stochastic on the active set. Write $\bar\theta^t := \frac{1}{|\mathcal{A}^t|}\sum_{i\in\mathcal{A}^t}\theta_i^t$ for the network average and $\Omega^t := \frac{1}{|\mathcal{A}^t|}\sum_{i\in\mathcal{A}^t}\left\|\theta_i^t - \bar\theta^t\right\|^2$ for the consensus error, and refer to $[0, \tau_k)$ as Phase 1 and to $[\tau_k, T]$ as Phase 2.

Client k selects an initial model $\theta_k^{\tau_k}$ and enters the graph at the *start* of round τ_k, so that $\bar\theta^{\tau_k}$ and Ω^{τ_k} are taken over all n clients before the update of Equation (1) is applied in that round. We write $\bar\theta^{\tau_k-}$ and Ω^{τ_k-} for the average and consensus error of the $n-1$ incumbents immediately before the join, and set $D_k := \left\|\theta_k^{\tau_k} - \bar\theta^{\tau_k-}\right\|$. The two pairs are taken over different index sets, which is exactly what Theorem 1 computes, so an implementation misplacing the join by one round violates it measurably (Section 5.1).

The join enters the analysis through exactly two quantities measured at τ_k, the *inherited gap* $G_{\tau_k} := f_n(\bar\theta^{\tau_k}) - f_n^* \geq 0$ and the consensus error Ω^{τ_k}, and every statement below about the cost of a late join reduces to these two. Both, with D_k, are $\mathcal{F}_{\tau_k}$-measurable, so Phase-2 statements are conditional on $\mathcal{F}_{\tau_k}$ and unconditional versions follow by the tower property.

2.2 THE CONVERGENCE CRITERION

Section 1 fixes the criterion. Stated for this setting, fix a horizon T, write $T' := T - \tau_k$ for the rounds remaining after the join, and measure

$$\mathcal{M}(\tau_k, T) \;:=\; \frac{1}{T'}\sum_{t=\tau_k}^{T-1}\mathbb{E}\left\|\nabla f_n(\bar\theta^t)\right\|^2, \tag{2}$$

the expected squared gradient norm of the full objective at the network average, averaged over the rounds following the join.

The criterion admits a direct operational reading. If $\mathcal{M}(\tau_k, T) \leq \epsilon^2$ and t is drawn uniformly from $[\tau_k, T)$, then $\bar\theta^t$ is ϵ-stationary in expectation, so Equation (2) bounds the quality of a model the run delivers under randomized stopping rather than summarizing a trajectory. It is also why the average rather than the minimum is the object of interest, since no implementable stopping rule attains the minimum. We evaluate the gradient at the network average, Ω^t carrying the dispersion about it separately, and begin the window at τ_k because f_n becomes the network's objective only then. Unlike the value of f_n it can be evaluated without knowing f_n^*. The constraint this places on what may be claimed, together with three companion conventions governing how comparisons are drawn, is stated in full in Appendix C.2.

3 CONVERGENCE ANALYSIS

The analysis rests on five assumptions, stated formally in Appendix C. (A1) Each f_i is L-smooth. (A2) f_n^*, f_{n-1}^* and f_k^* are finite. (A3) Stochastic gradients are unbiased, have variance at most σ^2, and are independent across clients within a round. (A4) Heterogeneity is bounded by ζ_n^2, uniformly in θ. (A5) Each W^t is symmetric and doubly stochastic, and the post-join sequence has a uniform spectral gap, $\lambda_n := \sup_{t\geq\tau_k}\left\|W^t - J\right\|_2 < 1$ for $J = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. This admits graphs that change at every round but excludes sporadic gossip. None requires convexity, and a Polyak–Łojasiewicz condition appears only in Appendix G.3.

3.1 THE JOIN EVENT

Before any optimization dynamics enter, the arrival of client k displaces the network average and alters the consensus error by amounts fixed entirely by the definitions of Section 2.1. We call the two statements below the *join identities*. They hold under no assumption at all, which is what allows the floating-point check of Section 5.1.

Theorem 1 (Join identities)**.** *Let the $n-1$ incumbents have average $\bar{\theta}^{\tau_k-}$ and consensus error Ω^{τ_k-}, and let client k join with initial model $\theta_k^{\tau_k}$ at distance $D_k = \left\|\theta_k^{\tau_k} - \bar{\theta}^{\tau_k-}\right\|$ from that average. Then*

$$\left\|\bar{\theta}^{\tau_k} - \bar{\theta}^{\tau_k-}\right\| = \frac{D_k}{n}, \qquad \Omega^{\tau_k} = \frac{n-1}{n}\,\Omega^{\tau_k-} + \frac{n-1}{n^2}\,D_k^2. \tag{3}$$

Proof. $\bar{\theta}^{\tau_k} = \frac{1}{n}\big((n-1)\bar{\theta}^{\tau_k-} + \theta_k^{\tau_k}\big)$ gives the first identity. Applying the parallel-axis identity to the n post-join models about $\bar{\theta}^{\tau_k-}$ equates $(n-1)\Omega^{\tau_k-} + D_k^2$ with $n\Omega^{\tau_k} + D_k^2/n$, which gives the second (Appendix D.2). □

We call $\frac{n-1}{n^2}D_k^2$ the *join shock*, the entire contribution of the arrival to the geometry of the network, quadratic in D_k and suppressed by one factor of n (Appendix D.2). The contraction of (A5) exists only from τ_k onward, since embedding the pre-join dynamics on all n clients leaves the mixing matrix with a repeated unit eigenvalue and no spectral gap (Remark 11), which makes the two-phase treatment necessary rather than convenient.

3.2 THE POST-JOIN STATIONARITY BOUND

A descent lemma for the average iterate, telescoped over $[\tau_k, T)$, combines with a consensus recursion of contraction factor $\rho = \frac{3+\lambda_n^2}{4}$ (Appendix D). That recursion runs from Ω^{τ_k}, so Theorem 1 enters the analysis at exactly this point and nowhere else.

Theorem 2 (Post-join stationarity bound)**.** *Under (A1)–(A5), for $\eta \le \frac{1-\lambda_n}{7L}$,*

$$\mathcal{M}(\tau_k, T) \;\le\; \underbrace{\frac{4G_{\tau_k}}{\eta T'}}_{\text{(i) inherited gap}} + \underbrace{\frac{16L^2\Omega^{\tau_k}}{(1-\lambda_n^2)T'}}_{\text{(ii) consensus transient}} + \underbrace{\frac{2L\eta\sigma^2}{n}}_{\text{(iii) variance floor}} + \underbrace{\frac{96\,\eta^2L^2(\sigma^2+\zeta_n^2)}{(1-\lambda_n)^2}}_{\text{(iv) topology floor}}. \tag{4}$$

Term (ii) is where the join enters, through Ω^{τ_k} and hence through Equation (3). It is not the join shock alone, that identity giving it the incumbents' own consensus error as well, so it does not vanish at $D_k = 0$. Terms (i) and (ii) are *transient*, decaying as $1/T'$. Terms (iii) and (iv) are *persistent*, depending on $\eta, \sigma^2, n, \lambda_n$ and ζ_n^2 alone, so they take the same values for a network in which no client joins late. In this bound, therefore, *joining late incurs no persistent penalty*. The claim concerns stationarity alone and does not assert that a late joiner reaches an equally good solution (Appendix G). Whether the transients are larger for a later join is separate. Appendix F answers it at the bound level, where the floors cancel exactly and explicit instances on both sides, one of them strongly convex, show that no available assumption fixes the sign of what remains.

Theorem 3 (Decomposition of the inherited gap)**.** *Under (A2), $0 \le G_{\tau_k} \le \frac{n-1}{n}\epsilon_1 + \frac{1}{n}\Delta_k^{\text{misfit}}$, where $\epsilon_1 := f_{n-1}(\bar{\theta}^{\tau_k}) - f_{n-1}^*$ is the incumbents' residual suboptimality and $\Delta_k^{\text{misfit}} := f_k(\bar{\theta}^{\tau_k}) - f_k^*$ is the* misfit*, the excess loss of the inherited model on client k's own data.*

Both quantities are evaluated at $\bar{\theta}^{\tau_k}$, which is what permits them to be combined, and the join enters Equation (4) only through Δ_k^{misfit}, with weight $1/n$. Because the misfit is a loss gap rather than a distance between distributions, client k can measure it on its own data, and the calibration of Section 4.2 rests on this. Once Phase 1 has plateaued, ϵ_1 is negligible and the inherited gap is governed by the misfit alone. Appendix D.9 gives an exact Phase-1 drift law.

3.3 LIMITS ON ANY TREATMENT OF THE JOIN

Matching the four terms against the quantities an implementation sets gives three consequences (Table 2, Appendix D.7), and together they delimit what any treatment of a late join may claim.

The join shock is unaffected by the step size. Term (ii) contains no η, and by Equation (3) it is set by Ω^{τ_k-}, which the arriving client does not control, and by D_k, which it fixes when it chooses its initial model. Initialization is therefore the only available control.

The step size that minimizes the bound is not computable. Term (i) is the only *transient* on which η acts, and balancing it against (iii) amounts to minimizing $4G_{\tau_k}/(\eta T') + 2L\eta\sigma^2/n$. That balance sets term (iv) aside, which is legitimate only in the regime Proposition 21 identifies and not at every operating point. The minimizer depends on G_{τ_k} and hence on f_n^*, which no client can evaluate, so it is exact but not directly usable.

The persistent floors are join-independent. Terms (iii) and (iv) involve none of G_{τ_k}, Ω^{τ_k}, D_k or τ_k, so they can be neither attributed to the join nor removed by any treatment of it. Nor is the inherited gap removable, since Theorem 3 identifies Δ_k^{misfit} as a property of the instance rather than an artifact of the analysis.

4 Initialization and Step-Size Rules

Equation (4) leaves two terms open to a treatment of the join, and we act on both. The *initialization rule* fixes how an arriving client initializes, the only quantity acting on the join shock, and the *step-size rule* replaces the oracle prescription of Section 3.3 with one that client can evaluate. The persistent floors are untouched by construction, and neither rule requires a server, adds a hyperparameter, alters the mixing schedule, or costs more than one exchange of models.

4.1 Removing the join shock

Corollary 4 (Warm initialization)**.** *If client k initializes as the average of models received from a neighbor set N_k, then $D_k^2 \le \frac{n-1}{|N_k|}\Omega^{\tau_k-}$ by Jensen's inequality, so the join shock obeys $\frac{n-1}{n^2}D_k^2 \le \frac{(n-1)^2}{n^2|N_k|}\Omega^{\tau_k-}$, and it therefore falls from the exogenous scale of a fresh initialization to the endogenous consensus scale. At the exact incumbent average, $D_k = 0$ and Equation* (3) *gives $\Omega^{\tau_k} = \frac{n-1}{n}\Omega^{\tau_k-} < \Omega^{\tau_k-}$, so the join strictly* decreases *network consensus error.*

The resolution is exact, since Equation (3) is an identity and D_k is the only quantity in it the arriving client sets. Nor is it what mixing already achieves, which reduces consensus error *after* τ_k whereas Ω^{τ_k} is the initial condition that recursion runs from. One round of averaging already reaches consensus scale, so waiting longer buys a constant factor while the incumbent average drifts (Appendix I.3).

Corollary 4 minimizes term (ii) alone, and trading a larger displacement for a lower term (i) is not advantageous.

Proposition 5 (The neighbor average is optimal to leading order)**.** *Let $\delta := \theta_k^{\tau_k} - \bar{\theta}^{\tau_k-}$ and let $U(\delta)$ be terms (i) and (ii) of Equation* (4) *with $f_n(\bar{\theta}^{\tau_k})$ replaced by its L-smooth quadratic majorant at $\bar{\theta}^{\tau_k-}$, a global upper bound under (A1). Then U is a strictly convex quadratic, minimized uniquely at*

$$\delta^\star = -\Big(L + \tfrac{8\eta L^2(n-1)}{1-\lambda_n^2}\Big)^{-1} n\,\nabla f_n(\bar{\theta}^{\tau_k-}), \tag{5}$$

and every other choice loses $\frac{\beta}{2}\|\delta - \delta^\star\|^2$, with $\beta > 0$ the curvature of U (Appendix I.3).

Averaging with neighbors is therefore optimal to leading order and not merely convenient. Estimating $\|\nabla f_n(\bar{\theta}^{\tau_k-})\|$ by $\sqrt{M} \approx 0.8$ at the default operating point gives $\|\delta^\star\| \approx 0.15$ against $D_k = 1.06$ for a fresh initialization, so $\delta = 0$ forgoes about 2×10^{-2} of what the neighbor average removes, a factor 10^{-2} of an effect Section 5.3 already finds macroscopically invisible. That number is the non-plateau regime, where $\|\nabla f_n\| = \Theta(1)$ and the residual is $\Theta(1)$ in n. Only at plateau, where $\nabla f_{n-1} \approx 0$ forces $\nabla f_n = \frac{1}{n}\nabla f_k$, does it acquire a further n^{-2}. Equation (5) is not a local quantity, and Appendix I.3 derives it and quantifies the cost of attaining it.

4.2 Calibrating the step size

A disruptive event might be expected to warrant a change of step-size schedule. The bound provides no support for one.

Theorem 6 (The optimal step size is a constant, and it is explicit)**.** *Let $\eta_t \in (0, \eta_c]$ with $\eta_c := \frac{1-\lambda_n}{7L}$ and $T' = T - \tau_k$, and measure stationarity by the step-size-weighted analogue of Equation (2), the standard randomized-stopping certificate. For $\sigma^2 > 0$ the part of the bound corresponding to terms (i) and (iii) of Equation (4) is at least $4\sqrt{2L\sigma^2 G_{\tau_k}/(nT')} = O\big((T')^{-1/2}\big)$, with equality* iff *$\eta_t \equiv \eta^\star = \sqrt{2nG_{\tau_k}/(L\sigma^2 T')}$. That optimum is admissible when $T' \geq 98nLG_{\tau_k}/(\sigma^2(1-\lambda_n)^2)$, and otherwise the constrained optimum is $\eta_t \equiv \eta_c$. No warmup, decay, or restart schedule beats a correctly chosen constant in the leading term.*

The proof is Cauchy–Schwarz followed by AM–GM (Appendix E). Within this bound class a schedule's profile carries no leading-order benefit, so the design problem is a constant rather than a shape. It is specific to the $\Theta(1/(\eta T'))$ transient of Equation (4), does not transfer to a Polyak–Łojasiewicz setting where a decaying schedule regains a logarithmic advantage (Appendix G.3), and concerns terms (i) and (iii) alone, term (iv) being set aside as lower order under the condition Proposition 21 states. Since $\eta^\star \propto (T')^{-1/2}$ at fixed G_{τ_k}, while G_{τ_k} stops growing once Phase 1 plateaus, *the prescribed step size increases as the join gets later*, until the cap binds. Section 5.2 finds that monotonicity on exactly the region where the precondition holds. On the capped branch a late joiner takes the largest step the spectral gap permits, leaving a residual transient $28LG_{\tau_k}/((1-\lambda_n)T')$ in which the delay penalty is transferred onto the graph. Neither branch is usable as it stands, since both depend on the unobservable f_n^* through G_{τ_k}.

Proposition 7 (Observable calibration rule)**.** *Substituting Theorem 3 into Theorem 6 gives*

$$\hat{\eta} = \min\left\{ \frac{1-\lambda_n}{7L},\ \sqrt{\frac{2\big[(n-1)\hat{\epsilon}_1 + \hat{\Delta}_k\big]}{L\sigma^2(T-\tau_k)}} \right\}, \tag{6}$$

which, once $(n-1)\hat{\epsilon}_1$ is negligible beside $\hat{\Delta}_k$, reduces to $\hat{\eta} = \sqrt{2\hat{\Delta}_k/(L\sigma^2(T-\tau_k))}$, in which n has canceled*.*

On that branch the joining client alone can compute the prescribed step size, from (L, σ^2, λ_n), the remaining horizon, and the local $\hat{\Delta}_k = f_k(\theta_{\mathrm{warm}}) - \hat{f}_k^{\mathrm{loc}}$, set by the loss client k incurs on the model it inherits rather than by how far its data lies from the rest. The local reference comes from a few local steps, and either direction of bias in it is covered below. The branch carries two qualifications. It needs $(n-1)\hat{\epsilon}_1 \leq \hat{\Delta}_k$ rather than a flat training curve, which is the stronger condition at large n, and $\hat{\Delta}_k$ is evaluated at $\theta_k^{\tau_k}$ rather than at the $\bar{\theta}^{\tau_k}$ of Theorem 3, a displacement warm initialization holds at consensus scale and Lemma 8 absorbs (Appendix E.4). The topologies of Section 5 never enter it, the cap being active throughout (Appendix K.3).

Lemma 8 (Square-root tolerance)**.** *Let $\hat{G} > 0$ estimate G_{τ_k}, let $\hat{\eta} = \sqrt{2n\hat{G}/(L\sigma^2 T')}$ (uncapped), and let $\kappa = \max\{\hat{G}/G_{\tau_k}, G_{\tau_k}/\hat{G}\}$. Then the leading term is inflated by exactly $\frac{1}{2}\big(\sqrt{\hat{G}/G_{\tau_k}} + \sqrt{G_{\tau_k}/\hat{G}}\big) \leq \frac{1}{2}\big(\sqrt{\kappa} + 1/\sqrt{\kappa}\big)$.*

A 4-fold error in $\hat{G}$ costs a factor 1.25 and a 16-fold error 2.125, which permits an estimator as coarse as $\hat{\Delta}_k$. The $\min$ with η_c makes the tolerance one-sided by absorbing over-estimates.

Algorithm 1 collects the two rules. Every client adopts the same $\hat{\eta}$ at τ_k from a tuple pinned at the join, and Section 5.1 checks the values agree bit for bit. L and σ^2 are estimated online by deliberately conservative estimators, whose cost is bounded by Lemma 8. On sparse graphs the $\min$ is usually active, so the rule reduces to the largest step the spectral gap permits, which is the content of Theorem 6 and not a failure of it (Appendix I).

5 Experiments

5.1 Experimental setup and implementation verification

We run CIFAR-10 (LeNet, from scratch), CIFAR-100 and Tiny-ImageNet (ResNet18-GN, from an ImageNet-pretrained backbone) (Krizhevsky, 2009; Le & Yang, 2015; LeCun et al., 1998; He

Algorithm 1 The two rules at a late join. The first two blocks run once, at τ_k, and the third is the round of Equation (1) with the step size pinned at the join.

Require: join round τ_k, horizon T, network constants $n, L, \sigma^2, \lambda_n$, incumbent residual estimate $\hat{\epsilon}_1$ (zero under the plateau gate, else the gossip fallback of Appendix I.2)

Once, at round τ_k: the arriving client k

1: $\theta_k^{\tau_k} \leftarrow |N_k|^{-1} \sum_{j \in N_k} \theta_j^{\tau_k -}$ from the models its neighbors N_k send ▷ Corollary 4

2: $\hat{\Delta}_k \leftarrow f_k(\theta_k^{\tau_k}) - \hat{f}_k^{\mathrm{loc}}$, with $\hat{f}_k^{\mathrm{loc}}$ from a few local steps ▷ misfit, Theorem 3

Once, at round τ_k: every client $i \in [n]$

3: $\eta_c \leftarrow (1-\lambda_n)/(7L)$ and $\hat{\eta} \leftarrow \min\left\{\eta_c, \sqrt{2[(n-1)\hat{\epsilon}_1 + \hat{\Delta}_k]/(L\sigma^2(T-\tau_k))}\right\}$ ▷ Proposition 7

Each round $t = \tau_k, \ldots, T-1$: every client $i \in [n]$, in parallel

4: $g_i^t \leftarrow$ stochastic gradient of f_i at θ_i^t

5: $\theta_i^{t+1} \leftarrow \sum_{j \in \mathcal{A}^t} W_{ij}^t \left(\theta_j^t - \hat{\eta}\, g_j^t\right)$ ▷ Equation (1), Theorem 6

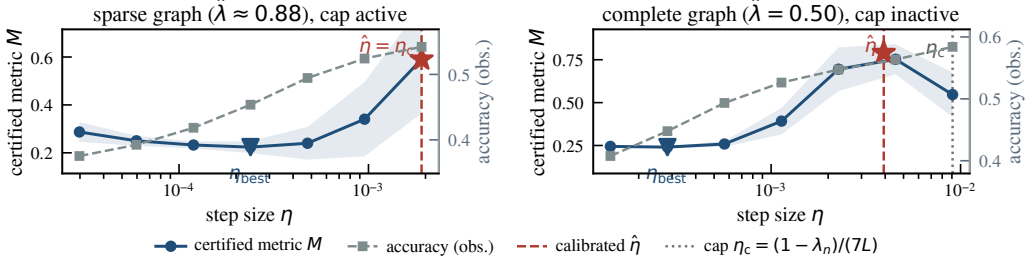


Figure 1: Certified metric M against step size on the two topology branches (CIFAR-10, band one s.d.). The interior minimum trades term (i) against terms (iii)–(iv) of Equation (4). Observational accuracy (gray, right axis) peaks at the largest step, so the two optima do not coincide.

et al., 2016; Wu & He, 2020) on a fixed random 6-regular graph with a join at $\tau_k = 0.5T$ under Dirichlet heterogeneity, λ_n from exact eigendecomposition. Every configuration uses plain SGD with momentum 0 on the same graph and horizon, and a seed fixes the join schedule, making the design paired. The ablation switches the two rules on and off independently, and with neither it is D-PSGD (Lian et al., 2017). We compare against DFedAvg (Sun et al., 2023a), Epidemic Learning (De Vos et al., 2023), and a reference without collaboration. On CIFAR-100 and Tiny-ImageNet the fixed-step rows use a per-dataset step size from a separate grid, so the ablation is read against a tuned floor, and their pretrained backbones carry a smaller $\hat{L}$, which raises η_c and is why the square-root branch of Equation (6) runs there while the cap binds on CIFAR-10 (Appendices J and J.1). Code is available online.[1]

Since Equation (3) is an identity, a systematic residual would mean the join is misplaced and nothing downstream could be trusted. Over every join at which it could be evaluated, including rounds where clients arrive together and Appendix D.3 applies, the largest relative residual is 1.3×10^{-15}, with sixteen two-client rounds excluded rather than counted as passing for want of a logged field (Appendix J.4). The step size adopted at τ_k is bit-identical across clients in every run where the rule is enabled. The check certifies the algebra and not the behavior of the quantity it governs, which the consensus trajectory supplies. A cold initialization raises Ω^{τ_k} by about 50% and a warm one lowers it to the value Corollary 4 predicts, to within 1% in every seed (Appendix K.2).

5.2 The calibrated step size behaves as the bound predicts

Shape of the curve and quality of the calibration. Equation (4) runs a transient in $1/\eta$ against floors increasing in η, so $M(\eta)$ should have an interior minimum. It does on both branches, at $\eta_c/8$ and $\eta_c/32$ (Figure 1). Lemma 8 bounds what calibration loses against that minimum by $(c+1/c)/2$, with $c = \hat{\eta}/\eta_{\mathrm{best}}$. On the sparse graph $c = 8.0$ permits 4.06 and we observe 2.64, and on the complete graph $c = 13.86$ permits 6.97 against 3.29. The calibrated step sits an order of magnitude

[1] https://anonymous.4open.science/r/DecentralizedFedLab-E067

Table 1: Ablation and baselines at the default operating point (5 seeds, paired by seed, momentum 0, ± across-seed s.d.). M is certified and lower is better, accuracy is observational and personalized, best per column in bold. The design being paired, an overlapping spread does not by itself make a difference insignificant. All rows share one Phase-1 step size and differ only from τ_k onward, so Appendix K.4 cannot be interpolated to them. *Columns are not comparable across datasets, CIFAR-10 being trained from scratch and the other two from a pretrained backbone.*

	CIFAR-10		CIFAR-100		Tiny-ImageNet	
Method	acc (obs.)	M	acc (obs.)	M	acc (obs.)	M
Init + calib.	0.541 ± 0.014	0.64 ± 0.18	0.479 ± 0.008	3.87 ± 0.59	0.487 ± 0.014	$\mathbf{0.25} \pm 0.03$
Calib. only	0.541 ± 0.014	0.63 ± 0.18	0.488 ± 0.004	4.35 ± 0.37	$\mathbf{0.505} \pm 0.003$	0.32 ± 0.04
Init only	$\mathbf{0.581} \pm 0.010$	0.56 ± 0.12	$\mathbf{0.488} \pm 0.002$	4.57 ± 0.25	0.483 ± 0.001	1.54 ± 0.07
D-PSGD	0.580 ± 0.010	0.56 ± 0.12	0.486 ± 0.003	4.53 ± 0.27	0.481 ± 0.004	1.55 ± 0.08
Epidemic Learning	0.467 ± 0.016	0.38 ± 0.12	0.426 ± 0.001	4.07 ± 0.20	0.458 ± 0.002	1.30 ± 0.10
DFedAvg	0.402 ± 0.013	$\mathbf{0.33} \pm 0.10$	0.391 ± 0.002	$\mathbf{3.76} \pm 0.11$	0.424 ± 0.003	1.16 ± 0.06
No collaboration	0.539 ± 0.007	–	0.382 ± 0.001	–	0.306 ± 0.002	–

above the empirical optimum, for reasons Appendix H locates in the derivation rather than in the estimators, yet the cost stays inside the envelope rather than growing with the ratio. On the complete graph M falls back at the largest grid point, which the bound does not predict.

Dependence on the join time and its precondition. By Theorem 6 the prescribed step grows as $(T')^{-1/2}$ once Phase 1 has plateaued. Sweeping τ_k on the complete graph, $\hat{\eta}$ increases monotonically over $\tau_k \geq 0.5T$ in every seed and falls below it, where $\hat{\Delta}_k$ is still descending. On the sparse graph $\hat{\eta} \equiv \eta_c$ throughout, the capped branch, and across topologies M follows neither the step nor accuracy (Appendices K.5 and K.6).

5.3 Component ablation and baselines

The initialization rule has no macroscopic effect, as predicted. Proposition 5 predicts that no initialization can be macroscopically visible at these sizes, and none is. Between Init only and D-PSGD, which differ in initialization alone, the paired accuracy difference is 0.001 to 0.002 across the three datasets and ΔM is 0.00, $+0.04$, -0.01 against spreads of 0.12, 0.25, 0.07. The two rules nonetheless couple through the channel Theorem 3 identifies, warm initialization lowering the measured misfit so that Equation (6) prescribes a smaller step. Tiny-ImageNet shows it, Init + calib. and Calib. only separating by 0.018 in accuracy and 0.07 in M in opposite directions. CIFAR-10 is the control, the cap binding throughout whatever the misfit.

The step-size rule acts on the square-root branch. Section 4.2 fixes that domain before any measurement. Once the cap binds, $\hat{\eta}$ is set by the graph and by $\hat{L}$ rather than by the misfit. Against Init only, where the branch runs, the certified metric improves in all 5 seeds on CIFAR-100 (-0.70) and Tiny-ImageNet (-1.29), costing 0.009 in observational accuracy on the first and gaining 0.004 on the second. The floor those improvements are measured against was selected on accuracy, and on CIFAR-100 the -0.70 does not survive against a larger constant applied from round 0. M still falls at $\eta = 0.3$, roughly 70 times η_c, so its minimizer lies outside the range Equation (4) covers while $\hat{\eta}$ is admissible by construction (Appendix K.4). CIFAR-10 lies outside the domain, a larger $\hat{L}$ putting η_c a factor of 5 below the fixed step, and the rule gives up 0.08 in M and 0.040 in accuracy there. Being off the cap is necessary for Equation (6) to express its trade and not sufficient for the rule to help (Appendix K.9).

Baselines and deferred results. Every configuration of the two rules exceeds both baselines in observational accuracy on all three datasets, while DFedAvg attains the lowest M on CIFAR-10 and CIFAR-100, at 0.179 and 0.097 in accuracy against the best row on each. On Tiny-ImageNet the two rules attain both the lowest M (0.25) and an accuracy above every baseline, while on the other two the lowest M and the highest accuracy belong to different rows, the criterion split of Appendix C.2 at its widest where the cap binds. Collaboration yields 0.042 on CIFAR-10 against 0.106 and 0.199 on the other two, ordered as the tasks are, CIFAR-10 training LeNet from scratch on 500

examples per client. A further 6 groups are deferred. No schedule beats the grid-optimal constant, the misestimation envelope holds without being stressed, and the n^{-1} suppression is confirmed from $n = 20$ to 200, while the cap-activity and heterogeneity groups locate where the step-size rule does not help. Under 20 staggered arrivals the identity still holds exactly while the effect on M reverses, and drawing the arrivals inside the plateau region removes the reversal in every seed, improving M by 3.8 across-seed standard deviations, the largest effect this paper measures (Appendices K.3 and K.7 to K.11).

6 Related Work

Decentralized SGD and consensus analysis. Decentralized SGD over a doubly stochastic mixing matrix (Lian et al., 2017; 2018), its unified treatment under time-varying topology (Koloskova et al., 2020), the gossip averaging underneath both (Boyd et al., 2006; Nedic & Ozdaglar, 2009) and the federated variants inheriting it (Sun et al., 2023a; De Vos et al., 2023) all contract consensus error from a *given* initial condition, over a client set fixed in advance. We take that recursion as given and act on its initial condition. Work acting instead on the communication schedule, by merging late (Zhu et al., 2026), by replacing outer synchronization with gossip (Hewa Koneputugodage et al., 2026), or by learning the graph (Singh et al., 2026), likewise fixes the participant set, so none of it defines a join. Neither those nor Shi et al. (2023) can be run at the floor Table 1 holds constant (Appendix J.3). One family does reach a floor. Term (iv) of Equation (4) scales with ζ_n^2, which gradient tracking removes by carrying an auxiliary estimate of the average gradient (Nedić et al., 2017; Pu & Nedić, 2021; Yuan et al., 2019), so our claim that the floors are irremovable concerns Equation (1) and not decentralized optimization in general. Trackers obey a conservation invariant $\sum_i s_i^t = \sum_i \nabla f_i^t$ an arriving client must be initialized to respect, and whether that absorbs or amplifies the join shock is open.

Client participation, availability, and asynchrony. Partial participation and client sampling, uniform (McMahan et al., 2017; Li et al., 2020) or biased toward high-loss clients (Cho et al., 2022), randomize the active subset, asynchronous aggregation (Liu et al., 2024) tolerates stale updates, and Zehtabi et al. (2025) unify sporadic computation and exchange. All fix the objective and the population in advance and parameterize the error by a participation probability or a staleness bound, neither of which absorbs a change of objective from f_{n-1} to f_n.

Newcomers, warm starting, and target shift. Ruan et al. (2021) admit mid-run arrivals but answer them with a rebooting heuristic rather than a bound, and Sun et al. (2023b) give a radius of stability for open systems. Both are server based, so neither has a consensus error to perturb, and both rest on the convex reading of Section 1 (Karimi et al., 2016), so a port would be a new method (Appendix J.3). Further out, Ash & Adams (2020) measure a generalization cost of warm starting, complementing Corollary 4 rather than competing with it, since that corollary acts on Ω^{τ_k} and asserts nothing about test error. Non-stationary optimization (Besbes et al., 2015) treats target change as drift under a variation budget rather than a discrete arrival.

Step sizes for non-convex SGD. Against the rate of Section 1, practice favors schedules over a tuned constant (Loshchilov & Hutter, 2017; Goyal et al., 2017). Learning-rate-free methods remove the unobservable constants by adapting toward a global distance (Defazio & Mishchenko, 2023), where Section 4.2 uses a local loss gap.

7 Conclusion

This paper asked what a decentralized network incurs when a client joins mid-run, and answered it without convexity. Exact identities quantify the arrival, and a post-join stationarity bound with explicit constants splits that cost into two vanishing transients and two floors a network with no late joiner already pays. The decomposition settles the design question and not only the descriptive one. Two terms remain open and each has an optimum, the neighbor average for the initialization and a local loss gap for the step size, the second only where the incumbents have plateaued. No analysis in this class can settle whether joining late costs anything that survives, the join entering none of the floors where such a cost would appear.

Reproducibility Statement

The theoretical claims are supported as follows. The five assumptions are stated formally in Appendix C, together with the scope of the mixing assumption and the probabilistic convention under which the Phase-2 statements are made. Complete proofs of the join identities, the post-join bound and the misfit decomposition are given in Appendix D, and of the step-size results in Appendix E. Every constant appearing in the bound is traced to the step of the proof that produces it in the table closing Appendix D.7, so that a reader checking one number does not have to reconstruct the whole argument. The results that are stated in the flow limit or at the level of upper bounds rather than realized errors are confined to Appendix D.9, Appendix F and Appendix G.2, and each is identified as such where it begins.

The experiments are specified as follows. Appendix J.1 gives the datasets, models, partitions, initialization and augmentation, including the two respects in which the CIFAR-10 setting differs from the other two. Appendix J.2 gives the conditions the common floor holds fixed and the argument for the momentum-free choice. Appendix J.4 gives the experiment groups, the seeds and the run counts. Appendix J.5 defines the certified metric and every derived quantity exactly as computed, and Appendix J.6 states the statistical procedure and its limits at these sample sizes. Appendix J.7 states each prediction with the condition under which it would have been refuted. Appendix I.1 expands Algorithm 1 into the form that was run, with every estimator defined and its direction of bias recorded.

One check should be performed before any other. Equation (3) and its multi-client composition in Equation (8) are identities that hold under no assumption, so a reimplementation can verify its fidelity to Theorem 1 directly, at the first join event, before reproducing any downstream number. A systematic residual there means the mixing order or the join instant is misplaced and that nothing computed afterwards can be trusted. Section 5.1 reports what the residual should look like when the implementation is correct.

Use of Generative AI

Generative AI tools were used throughout this work, in both the research and the writing. We disclose by category, following the taxonomy in the ICLR policy.

Categories requiring disclosure. We used generative AI tools to help develop the theoretical framework, to formulate mathematical claims, to supply critical ingredients for proving them, to assist in writing the proofs, to propose and refine hypotheses, to design the experimental methodology and to give feedback on it, to implement the method and the experimental harness, and to interpret results. We did not use generative AI tools to generate synthetic datasets, or for qualitative or thematic data analysis. Dataset preparation is limited to the partitioning of public benchmarks, which is covered by the implementation disclosure above.

Categories where disclosure is recommended. We additionally used generative AI tools to draft parts of this paper, to edit it for readability, to suggest its structure, to write the analysis and verification code, to search for information, and to brainstorm.

Verification. We have reviewed all AI-assisted work, and the following checks are the substance of that review rather than a summary of it. Every constant in Equation (4), the admissibility threshold, the capped-branch value, the optimal constant step size and the tolerance factor of Lemma 8 were re-derived symbolically and checked against the proofs as written. The implementation is tested against the analysis by Equation (3), which holds under no assumption and therefore produces a detectable residual if the code does not implement what the theorem describes, and Section 5.1 reports the outcome over every logged join. The step size adopted at the join is separately verified to be bit-identical across clients, which is an equality rather than a tolerance. Appendix J.7 states each prediction together with the condition under which it would have been refuted, and reports the outcomes that departed rather than only those that did not. We take responsibility for the final content of this work, including the text, the claims and the artifacts produced with the aid of generative AI.

REFERENCES

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.

Jordan T. Ash and Ryan P. Adams. On warm-starting neural network training. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3884–3894, 2020.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konečný, Stefano Mazzocchi, H. Brendan McMahan, et al. Towards federated learning at scale: System design. In *Proceedings of Machine Learning and Systems*, 2019.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006.

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

Martijn De Vos, Sadegh Farhadkhani, Rachid Guerraoui, Anne-Marie Kermarrec, Rafael Pires, and Rishi Sharma. Epidemic learning: Boosting decentralized learning with randomized communication. *Advances in Neural Information Processing Systems*, 36:36132–36164, 2023.

Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by D-adaptation. In *International Conference on Machine Learning*, pp. 7449–7479. PMLR, 2023.

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

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

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

Chamin Hewa Koneputugodage, Thalaiyasingam Ajanthan, Sameera Ramasinghe, Hadi Mohaghegh Dolatabadi, Shamane Siriwardhana, Gil Avraham, Violetta Shevchenko, Karol Pajak, James Snewin, and Alexander Long. Factored gossip DiLoCo: Reducing blocking communication in DiLoCo. In *International Conference on Machine Learning*, 2026.

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

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In *European Conference on Machine Learning*, 2016.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning*, pp. 5381–5393. PMLR, 2020.

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

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. CS231N course report, Stanford University, 2015.

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

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

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning*, pp. 3043–3052. PMLR, 2018.

Ji Liu, Juncheng Jia, Tianshi Che, Chao Huo, Jiaxiang Ren, Yang Zhou, Huaiyu Dai, and Dejing Dou. FedASMU: Efficient asynchronous federated learning with dynamic staleness-aware model update. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13900–13908, 2024.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Angelia Nedić, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187:409–457, 2021.

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

Yifan Shi, Li Shen, Kang Wei, Yan Sun, Bo Yuan, Xueqian Wang, and Dacheng Tao. Improving the model consistency of decentralized federated learning. In *International Conference on Machine Learning*, pp. 31269–31291. PMLR, 2023.

Inderjeet Singh, Eleonore Vissol-Gaudin, Andikan Otung, and Motoyoshi Sekiya. Learning to collaborate: An orchestrated-decentralized framework for peer-to-peer LLM federation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25472–25480, 2026.

Tao Sun, Dongsheng Li, and Bao Wang. Decentralized federated averaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4289–4301, 2023a.

Youbang Sun, Heshan Fernando, Tianyi Chen, and Shahin Shahrampour. On the stability analysis of open federated learning systems. In *American Control Conference*, pp. 867–872, 2023b.

Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N Ahmad Aziz, et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 594(7862):265–270, 2021.

Yuxin Wu and Kaiming He. Group normalization. *International Journal of Computer Vision*, 128 (3):742–755, 2020.

Kun Yuan, Bicheng Ying, Xiaochuan Zhao, and Ali H. Sayed. Exact diffusion for distributed optimization and learning – part I: Algorithm development. *IEEE Transactions on Signal Processing*, 67(3):708–723, 2019.

Shahryar Zehtabi, Dong-Jun Han, Rohit Parasnis, Seyyedali Hosseinalipour, and Christopher G. Brinton. Decentralized sporadic federated learning: A unified algorithmic framework with convergence guarantees. In *International Conference on Learning Representations*, 2025.

Tongtian Zhu, Tianyu Zhang, Mingze Wang, Zhanpeng Zhou, and Can Wang. On the surprising effectiveness of a single global merging in decentralized learning. In *International Conference on Learning Representations*, 2026.

Appendix: Supplementary Material

A Overview of the Supplementary Material

The supplement is organized so that it can be read starting from any individual result. Appendix B fixes notation and gives the correspondence between the theoretical symbols and the ones the experiments report. Appendix C states (A1)–(A5) formally, records the scope of the mixing assumption and the probabilistic convention, and gives in full the four conventions governing measurement and comparison, of which the main text states only one. Appendix D proves the join identities and the post-join bound with every constant traced to its origin, together with the multi-client composition that Section 5.1 actually checks and an exact Phase-1 drift law. Appendix E proves the step-size results, defines the estimators, and records the direction of bias of each.

Two sections carry material cut from the main text for length. Appendix F gives the exact bound-level penalty for joining later rather than earlier, together with explicit instances showing that no regime assumption supplies a universal sign for it, one of them strongly convex. Appendix G sets out what the stationarity criterion does and does not certify, describes basin selection as the single mechanism by which the join time could change the attained value rather than the rate, and gives the Polyak–Łojasiewicz case. Section 3.2 draws on the first for the sign of the delay penalty and on the second for what the criterion leaves out. No experiment reported here tests either, and both are written to be read as bound-level or flow-limit statements.

The remainder is apparatus. Appendix H states the limitations in full, Appendix I gives the estimators and the message schedule in enough detail to reimplement from, Appendix J gives the experimental configurations together with the condition under which each prediction would have been refuted, and Appendix K gives the additional results the main text defers, among them the network with no late joiner that Theorem 2 is stated against (Appendix K.1) and the sweep of network size that measures the suppression of the join shock (Appendix K.3).

B NOTATION

Symbol	Meaning	First used
Network and schedule		
n, k, τ_k	client count, the late client, its join round	Section 2.1
$\mathcal{A}^t$	active set, $[n] \setminus \{k\}$ before τ_k and $[n]$ after	Section 2.1
W^t, J	mixing matrix at round t; $\frac{1}{n}\mathbf{1}\mathbf{1}^\top$	Equation (1)
λ_n	uniform post-join spectral gap parameter, $\sup_{t\geq\tau_k} \left\|W^t - J\right\|_2$	Assumption 5
T, T'	horizon; post-join window length $T - \tau_k$	Section 2.2
Iterates and geometry		
θ_i^t, $\bar{\theta}^t$	client model; average over the active set	Section 2.1
Ω^t	consensus error, $\frac{1}{\lvert\mathcal{A}^t\rvert}\sum_i \left\|\theta_i^t - \bar{\theta}^t\right\|^2$	Section 2.1
$\bar{\theta}^{\tau_k-}$, Ω^{τ_k-}	incumbent average and consensus error just before the join	Section 2.1
D_k	join displacement, $\left\|\theta_k^{\tau_k} - \bar{\theta}^{\tau_k-}\right\|$	Section 2.1
$\tilde{\theta}^t$, $\widetilde{\Omega}^{\tau_k}$	average and consensus error of the earlier-joining run	Appendix F
Objectives and gaps		
f_i, f_n, f_{n-1}, f_k	local, full-network, incumbent and joiner objectives	Section 2.1
f_n^*, f_{n-1}^*, f_k^*	their infima, finite by (A2)	Assumption 2
G_{τ_k}, $\widetilde{G}_{\tau_k}$	inherited gap $f_n(\bar{\theta}^{\tau_k}) - f_n^*$; its earlier-run counterpart	Section 2.1
G_0	gap at initialization, $f_n(\bar{\theta}^0) - f_n^*$	Corollary 19
ϵ_1	Phase-1 residual, $f_{n-1}(\bar{\theta}^{\tau_k}) - f_{n-1}^*$	Theorem 3
Δ_k^{misfit}	misfit, $f_k(\bar{\theta}^{\tau_k}) - f_k^*$	Theorem 3
Constants		
L, σ^2, ζ_n^2	smoothness, gradient-noise variance, heterogeneity	(A1), (A3), (A4)
μ	Polyak–Łojasiewicz constant, used only in Appendix G.3	Assumption 6
ρ, C_Ω	consensus contraction $\frac{3+\lambda_n^2}{4}$; forcing constant 24	Lemma 14
Step sizes		
η, η_t	constant step size; a schedule	Equation (1)
η_{c}	admissibility cap $\frac{1-\lambda_n}{7L}$	Theorem 6
$\eta^\star$	bound-optimal constant $\sqrt{2nG_{\tau_k}/(L\sigma^2T')}$	Theorem 6
$\hat{\eta}$	the calibrated step actually adopted	Equation (6)
S, $\mathcal{M}_w$	schedule sum $\sum_t \eta_t$; step-weighted criterion	Appendix E.1
Flow limit, used only in Appendices D.9, F and G.2		
Φ_{n-1}^s, B_n	time-s flow map of $-\nabla f_{n-1}$; basin map of f_n	Definition 17
Measured quantities		
$\mathcal{M}(\tau_k, T)$, M	the criterion of Equation (2); its measured value	Equation (2)
$\hat{\Delta}_k$, $\hat{\epsilon}_1$	measured misfit and Phase-1 residual	Proposition 7
$\hat{L}$, $\hat{\sigma}^2$, $\hat{\lambda}$	estimated smoothness, noise variance, spectral quantity	Appendix E.4
κ	two-sided misestimation ratio of the gap	Lemma 8
c	ratio of $\hat{\eta}$ to the empirically best step size	Section 5.2

The last block requires one comment. $\mathcal{M}(\tau_k, T)$ is defined as an expectation over the trajectory, whereas M is a finite-sample estimate of it computed from full-batch gradients at the network average in float64 and averaged over $[\tau_k, T)$. Appendix J.5 gives the exact procedure. The two are written differently throughout so that a statement about the bound is never mistaken for a statement about a measurement.

A small number of objects recur often enough to be named. They are collected here so that no name has to be reconstructed from context.

Term	What it names	Defined at
join identities	the two assumption-free identities for the arrival	Theorem 1
join shock	$\frac{n-1}{n^2} D_k^2$, the arrival's entire contribution to network geometry	Section 3.1
inherited gap	G_{τ_k}, suboptimality of the average under the new objective	Section 2.1
misfit	Δ_k^{misfit}, excess loss of the inherited model on client k's own data	Theorem 3
initialization rule	neighbor-averaged initialization at the join	Section 4.1
step-size rule	the calibrated post-join step size of Equation (6)	Section 4.2
certified metric	M, the measured value of Equation (2)	Appendix J.5
capped, square-root branch	the two branches of the minimum in Equation (6)	Equation (6)

Of these, *join identities*, *join shock* and *misfit* are ours; the rest are descriptive labels for objects the analysis introduces.

C Assumptions and Measurement Conventions

C.1 Assumptions

Assumption 1 (Smoothness)**.** Each f_i is L-smooth: $\|\nabla f_i(x) - \nabla f_i(y)\| \le L \|x - y\|$ for all x, y.

Assumption 2 (Lower boundedness)**.** $f_n^*,\ f_{n-1}^*,\ f_k^* > -\infty$.

Assumption 3 (Stochastic gradients)**.** $\mathbb{E}[g_i^t \mid \theta_i^t] = \nabla f_i(\theta_i^t)$, $\mathbb{E}\,\|g_i^t - \nabla f_i(\theta_i^t)\|^2 \le \sigma^2$, and the noises $\{g_i^t - \nabla f_i(\theta_i^t)\}_i$ are independent across clients at each round.

Assumption 4 (Bounded heterogeneity)**.** $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\theta) - \nabla f_n(\theta)\|^2 \le \zeta_n^2$ for all θ.

Assumption 5 (Mixing)**.** Each W^t is symmetric and doubly stochastic on the active set, and the post-join sequence has a uniform spectral gap: with $J = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, $\lambda_n := \sup_{t \ge \tau_k} \|W^t - J\|_2 < 1$. For a fixed graph, $\lambda_n = \max_{i \ge 2} |\lambda_i(W)|$.

Convexity is assumed nowhere in Assumptions 1 to 5. A Polyak–Łojasiewicz condition is stated separately as Assumption 6 and is used only in Appendix G.3.

Remark 9 (Scope of the mixing assumption)*.* Assumption 5 permits the graph to change at every round, provided each round's matrix is by itself uniformly contracting, so that every per-round graph is connected with gap at least $1 - \lambda_n$. Schemes whose individual rounds may be disconnected, such as sporadic gossip, are *not* covered. They require joint spectral control of the products $\prod_s W^s$ over multi-round windows (Koloskova et al., 2020), under which the results below hold with $1 - \lambda_n$ replaced by the corresponding multi-round gap. We do not pursue that extension, and no experiment reported here exercises it.

Remark 10 (Probabilistic convention)*.* The quantities G_{τ_k}, Ω^{τ_k} and D_k depend on the Phase-1 trajectory and are therefore $\mathcal{F}_{\tau_k}$-measurable random variables rather than deterministic constants. Every Phase-2 statement below is made conditional on $\mathcal{F}_{\tau_k}$, so that $\mathbb{E}[\cdot]$ abbreviates $\mathbb{E}[\cdot \mid \mathcal{F}_{\tau_k}]$ and these three quantities appear on right-hand sides as known numbers. Unconditional versions follow by the tower property, with G_{τ_k} and Ω^{τ_k} replaced by their expectations. The join identities of Theorem 1 are pathwise and need no such convention.

C.2 Conventions governing measurement and comparison

The four statements below are constraints on what the analysis is permitted to assert. Only the second is stated in the main text (Section 2.2). The others bind where they are applied but are not stated there, so this subsection is the only place they appear in full. The body refers to them by what they say rather than by number.

Convention 1 (Measurement window)**.** A comparison between two training strategies must average the same metric over the *same* round window, with the same $(\eta, \sigma^2, \zeta_n^2, \lambda_n)$. Comparing a post-join window average for one strategy against a full-horizon average for another is inadmissible, since the full-horizon average is inflated by large early-round gradients and the comparison then measures the window rather than the strategy. A comparison that tunes η separately per strategy is not a same-floor comparison and must be labeled as one that is not.

This is why the criterion of Equation (2) averages from τ_k rather than from round 0, and why Theorem 24 fixes a common window and a common step size before any difference of bounds is taken. It is also the reason Table 1 holds momentum, topology, round count and join schedule fixed across configurations, and flags the two datasets on which the fixed-step configurations use a separately tuned constant.

Convention 2 (What a stationarity bound certifies)**.** A bound on $\mathcal{M}(\tau_k, T) = \frac{1}{T'} \sum_t \mathbb{E} \left\| \nabla f_n(\bar{\theta}^t) \right\|^2$ certifies approach to stationarity of f_n and nothing else. It carries no information about the value of f_n attained at or near the stationary point, and therefore none about accuracy. No claim of the form "this configuration attains better accuracy" may be derived from such a bound.

This is the reason accuracy is labeled observational everywhere in Section 5, and the reason the step size that minimizes $\mathcal{M}$ and the step size that maximizes accuracy are allowed to differ, as they do in Figure 1. Appendix G.1 develops the point.

Convention 3 (Comparing bounds)**.** If U_A and U_B upper bound the errors of two strategies, then $U_A < U_B$ does not imply $\mathrm{err}_A < \mathrm{err}_B$, because bounds differ in slack. Orderings of upper bounds are directional evidence at the bound level and are never reported as orderings of true errors. A sign or an "if and only if" statement about a difference of bounds is admissible only when that difference is computed with exact constants, never from unsigned $O(\cdot)$ expressions.

The exact-constants condition is what makes Theorem 24 admissible and what makes any informal ordering of $O(\cdot)$ expressions inadmissible. It binds nowhere else in the paper, since Section 5 compares measured quantities rather than bounds.

Convention 4 (Transient and persistent)**.** Fix η. A term in a convergence bound is *persistent* if it does not vanish as $T \to \infty$, and *transient* otherwise. A term of order $(T')^{-\alpha}$ with $\alpha > 0$ is transient however slowly it decays. The qualifier "slow" may be attached to such a term. The label "persistent" may not.

This is the definition used in Section 3.2 and it is what justifies the reading of Equation (4) given there. Terms (i) and (ii) decay as $1/T'$, which is far slower than the geometric decay a strongly convex or Polyak–Łojasiewicz objective would give (Theorem 30), and calling them persistent on that ground would be a category error.

The conventions are restrictive rather than decorative, and the clearest way to see it is to take a claim they forbid. Suppose one asserted that delaying a client's arrival improves final accuracy. Such a claim breaks three of the four at once: it compares unequal windows, it reads a stationarity bound as an accuracy statement, and it treats an ordering of bounds as an ordering of outcomes. What can be said in its place is what Theorem 24 and the two-sided instances of Proposition 26 give, which is that nothing available in this setting fixes the sign.

D Proofs: The Join Event and the Post-Join Bound

D.1 Spectral collapse before the join

Remark 11 (The spectral gap is created by the join). Embed the pre-join dynamics in $\mathbb{R}^n$ by writing $\hat{W}^t = \mathrm{blockdiag}(W^t, 1)$ for $t < \tau_k$, with the trailing block acting on the not-yet-participating client k. The doubly stochastic block and that trailing block each contribute an eigenvalue 1, so $\lambda_2(\hat{W}^t) = 1$ and the system on all n clients has no spectral gap before τ_k. The contraction supplied by (A5) therefore exists only from the join onward.

Proof. $\hat{W}^t$ is block diagonal, so its spectrum is the union of the spectra of its blocks. The 1×1 block contributes the eigenvalue 1. The block W^t is doubly stochastic, so $W^t \mathbf{1} = \mathbf{1}$ and its leading eigenvalue is also 1. The eigenvalue 1 of $\hat{W}^t$ therefore has multiplicity at least two, whence $\lambda_2(\hat{W}^t) = 1$ and $\left\| \hat{W}^t - \frac{1}{n} \mathbf{1}\mathbf{1}^\top \right\|_2 = 1$. □

The consequence is structural rather than technical. Any attempt to run a single consensus recursion across the join must do so with contraction factor 1 on $[0, \tau_k)$, which is no contraction at all, so

the analysis has to restart at τ_k from whatever consensus error the join leaves behind. That restart is what makes Ω^{τ_k} an initial condition rather than a term in a running average, and it is the reason Theorem 1 enters Theorem 2 at exactly one place.

A second, weaker observation points the same way. Writing $a_i^t = t - \tau_i$ for the number of rounds client i has been mixing by round t, the difference $a_i^t - a_k^t = \tau_k - \tau_i$ is constant in t for every incumbent i. The late client therefore retains the smallest mixing age throughout, and the difference does not diminish however long the run continues. This does not by itself imply a penalty, since Theorem 2 shows the cost of lateness decays, but it does explain why no amount of post-join mixing can undo the asymmetry at the level of the schedule.

D.2 Proof of Theorem 1

Theorem 1 (restated). *Let the $n-1$ incumbents have average $\bar\theta^{\tau_k-}$ and consensus error Ω^{τ_k-}, and let client k join with initial model $\theta_k^{\tau_k}$ at distance $D_k = \big\|\theta_k^{\tau_k} - \bar\theta^{\tau_k-}\big\|$ from that average. Then $\big\|\bar\theta^{\tau_k} - \bar\theta^{\tau_k-}\big\| = D_k/n$ and $\Omega^{\tau_k} = \frac{n-1}{n}\Omega^{\tau_k-} + \frac{n-1}{n^2}D_k^2$.*

Proof. Both statements are algebra and use no assumption. For the first,

$$\bar\theta^{\tau_k} = \frac{1}{n}\Big(\sum_{i\neq k}\theta_i^{\tau_k} + \theta_k^{\tau_k}\Big) = \frac{1}{n}\big((n-1)\bar\theta^{\tau_k-} + \theta_k^{\tau_k}\big) = \bar\theta^{\tau_k-} + \frac{1}{n}\big(\theta_k^{\tau_k} - \bar\theta^{\tau_k-}\big),$$

so $\big\|\bar\theta^{\tau_k} - \bar\theta^{\tau_k-}\big\| = \frac{1}{n}\big\|\theta_k^{\tau_k} - \bar\theta^{\tau_k-}\big\| = D_k/n$.

For the second, recall the parallel-axis identity. For points $x_1,\dots,x_m$ with mean $\bar x$ and any $c\in\mathbb{R}^d$,

$$\sum_{i=1}^m \|x_i - c\|^2 = \sum_{i=1}^m \|x_i - \bar x\|^2 + m\,\|\bar x - c\|^2, \tag{7}$$

which follows by expanding $\|x_i - c\|^2 = \|x_i - \bar x\|^2 + 2\langle x_i - \bar x, \bar x - c\rangle + \|\bar x - c\|^2$ and noting that the cross terms sum to zero. Apply Equation (7) to the $m = n$ post-join models $\{\theta_i^{\tau_k}\}_{i\in[n]}$, whose mean is $\bar\theta^{\tau_k}$, with $c = \bar\theta^{\tau_k-}$. The left side splits over the incumbents and the joiner,

$$\sum_{i\in[n]}\big\|\theta_i^{\tau_k} - \bar\theta^{\tau_k-}\big\|^2 = \underbrace{\sum_{i\neq k}\big\|\theta_i^{\tau_k} - \bar\theta^{\tau_k-}\big\|^2}_{=(n-1)\Omega^{\tau_k-}} + \underbrace{\big\|\theta_k^{\tau_k} - \bar\theta^{\tau_k-}\big\|^2}_{=D_k^2},$$

the first sum being $(n-1)\Omega^{\tau_k-}$ by the definition of Ω^{τ_k-} as the consensus error of the $n-1$ incumbents about their own average. The right side is $n\,\Omega^{\tau_k} + n\,\big\|\bar\theta^{\tau_k} - \bar\theta^{\tau_k-}\big\|^2 = n\,\Omega^{\tau_k} + D_k^2/n$ by the first identity. Equating and solving,

$$n\,\Omega^{\tau_k} = (n-1)\Omega^{\tau_k-} + D_k^2 - \frac{D_k^2}{n} = (n-1)\Omega^{\tau_k-} + \frac{n-1}{n}D_k^2,$$

which is the claim after division by n. □

Two checks. First, for $n = 2$ with a single incumbent, $\Omega^{\tau_k-} = 0$ and two models at distance D_k have their average midway between them, so $\Omega^{\tau_k} = \frac{1}{2}\big(2\cdot(D_k/2)^2\big) = D_k^2/4$, matching $\frac{n-1}{n^2}D_k^2 = \frac{1}{4}D_k^2$. Second, setting $D_k = 0$ gives $\Omega^{\tau_k} = \frac{n-1}{n}\Omega^{\tau_k-} < \Omega^{\tau_k-}$, the strict decrease recorded in Corollary 4, which is simply the statement that adding a point at the mean of a set reduces its mean squared deviation.

The two identities suppress the arrival by different powers of n, and it is worth separating them because the main text turns on the second. The network average moves by D_k/n, so its squared displacement is $\Theta(n^{-2})$ at fixed D_k. The consensus error instead picks up $\frac{n-1}{n^2}D_k^2$, whose numerator grows with n, so at fixed D_k that term is $\Theta(n^{-1})$; measured against the incumbents' own consensus error it is exactly $D_k^2/(n\Omega^{\tau_k-})$, again $\Theta(n^{-1})$. Doubling n therefore halves the join shock rather than quartering it, and any claim in this paper about an arrival becoming invisible at large n rests on the weaker of the two rates. Appendix K.3 sweeps n over a 10-fold range and finds the second rate, with a fitted exponent of -1.085 against the -2 the first would give.

D.3 MULTI-CLIENT JOINS

The runs of Section 5 contain rounds at which more than one client arrives, so the identity that is actually checked there is the following composition rather than Equation (3) itself.

Proposition 12 (Simultaneous joins)**.** *Let n_{pre} incumbents have average $\bar\theta^{\tau_k-}$ and consensus error Ω^{τ_k-}, and let a set $\mathcal{J}$ of clients join at round τ_k, so that $n = n_{\mathrm{pre}} + |\mathcal{J}|$. Write $\delta_j = \theta_j^{\tau_k} - \bar\theta^{\tau_k-}$ and $D_j = \|\delta_j\|$ for $j \in \mathcal{J}$. The symbol $\mathcal{J}$ is distinct from the averaging matrix $J = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ of Assumption 5. Then, exactly,*

$$\bar\theta^{\tau_k} = \bar\theta^{\tau_k-} + \frac{1}{n}\sum_{j\in\mathcal{J}}\delta_j, \qquad \Omega^{\tau_k} = \frac{1}{n}\Big[n_{\mathrm{pre}}\,\Omega^{\tau_k-} + \sum_{j\in\mathcal{J}} D_j^2 - \frac{1}{n}\Big\|\sum_{j\in\mathcal{J}}\delta_j\Big\|^2\Big]. \tag{8}$$

Proof. The first identity is the definition of the average. For the second, apply Equation (7) to the n post-join models with $c = \bar\theta^{\tau_k-}$. The left side is $n_{\mathrm{pre}}\Omega^{\tau_k-} + \sum_{j\in\mathcal{J}} D_j^2$ and the right side is $n\,\Omega^{\tau_k} + n\,\big\|\bar\theta^{\tau_k} - \bar\theta^{\tau_k-}\big\|^2 = n\,\Omega^{\tau_k} + \frac{1}{n}\Big\|\sum_{j\in\mathcal{J}}\delta_j\Big\|^2$. □

For $|\mathcal{J}| = 1$ the last term is D_k^2/n and Equation (8) reduces to Equation (3). The negative sign runs opposite to the intuition one might bring to it. The subtracted term is largest when the displacements *align*, so arrivals that agree on a direction inject the *least* consensus error, $|\mathcal{J}|$ clients at a common point $\delta_j \equiv \delta$ injecting $\frac{|\mathcal{J}|(n-|\mathcal{J}|)}{n^2}\,\|\delta\|^2$, which is $O(n^{-1})$ relative to $\|\delta\|^2$ for small $|\mathcal{J}|$. Arrivals whose displacements cancel make that term vanish and inject the *most*, namely $\frac{1}{n}\sum_{j\in\mathcal{J}} D_j^2$. The mechanism is that alignment drags the post-join average toward the newcomers, whereas cancellation leaves the average where it was and leaves every newcomer at its full distance from it. What cancellation does minimize is the displacement of the average itself, which the first identity of Equation (8) sends to zero. We do not analyze the bound-level composition of several joins, and Theorem 2 is applied only with Ω^{τ_k} read off Equation (8) as an initial condition.

D.4 DESCENT OF THE AVERAGE ITERATE

The lemma is stated for a general step-size sequence, since Appendix E needs that form. Setting $\eta_t \equiv \eta$ recovers the constant-step case used in Theorem 2.

Lemma 13 (Non-convex descent)**.** *Under Assumptions 1 to 3 and the double stochasticity of Assumption 5, for $t \ge \tau_k$ and $\eta_t \le \frac{1}{4L}$,*

$$\mathbb{E} f_n(\bar\theta^{t+1}) \le \mathbb{E} f_n(\bar\theta^t) - \frac{\eta_t}{4}\mathbb{E}\,\big\|\nabla f_n(\bar\theta^t)\big\|^2 + \eta_t L^2 \mathbb{E}[\Omega^t] + \frac{L\eta_t^2\sigma^2}{2n}. \tag{9}$$

Proof. Step 1, the average obeys a plain SGD recursion. Each W^t is doubly stochastic on the active set, which for $t \ge \tau_k$ is all of $[n]$, so $\frac{1}{n}\sum_i\sum_j W_{ij}^t x_j = \frac{1}{n}\sum_j x_j$ for any $\{x_j\}$. Applying this to Equation (1) gives $\bar\theta^{t+1} = \bar\theta^t - \eta_t \bar g^t$ with $\bar g^t = \frac{1}{n}\sum_i g_i^t$. Write $\bar\nabla^t := \frac{1}{n}\sum_i \nabla f_i(\theta_i^t) = \mathbb{E}[\bar g^t \,|\, \mathcal{F}_t]$.

Step 2, deviation control. f_n is an average of L-smooth functions and is therefore L-smooth. By the triangle inequality, Assumption 1, and Cauchy–Schwarz in the form $\frac{1}{n}\sum_i a_i \le (\frac{1}{n}\sum_i a_i^2)^{1/2}$,

$$\big\|\bar\nabla^t - \nabla f_n(\bar\theta^t)\big\| \le \frac{1}{n}\sum_i \big\|\nabla f_i(\theta_i^t) - \nabla f_i(\bar\theta^t)\big\| \le \frac{L}{n}\sum_i \big\|\theta_i^t - \bar\theta^t\big\| \le L\sqrt{\Omega^t}. \tag{10}$$

Step 3, smoothness and the cross term. L-smoothness of f_n gives $f_n(\bar\theta^{t+1}) \le f_n(\bar\theta^t) - \eta_t\,\langle \nabla f_n(\bar\theta^t), \bar g^t\rangle + \frac{L\eta_t^2}{2}\,\|\bar g^t\|^2$. Take $\mathbb{E}[\cdot\,|\,\mathcal{F}_t]$, which replaces $\bar g^t$ by $\bar\nabla^t$ in the inner product. Writing $e^t := \bar\nabla^t - \nabla f_n(\bar\theta^t)$ and using Young's inequality $\langle a, b\rangle \le \frac{1}{2}\,\|a\|^2 + \frac{1}{2}\,\|b\|^2$ together with Equation (10),

$$-\eta_t\,\langle \nabla f_n(\bar\theta^t), \bar\nabla^t\rangle = -\eta_t\,\big\|\nabla f_n(\bar\theta^t)\big\|^2 - \eta_t\,\langle \nabla f_n(\bar\theta^t), e^t\rangle \le -\frac{\eta_t}{2}\,\big\|\nabla f_n(\bar\theta^t)\big\|^2 + \frac{\eta_t L^2}{2}\Omega^t.$$

Step 4, second moment. The noises are conditionally unbiased and independent across clients by Assumption 3, so their average has second moment at most σ^2/n and

$$\mathbb{E}\left\|\bar{g}^t\right\|^2 = \left\|\bar{\nabla}^t\right\|^2 + \mathbb{E}\left\|\bar{g}^t - \bar{\nabla}^t\right\|^2 \le \left\|\bar{\nabla}^t\right\|^2 + \frac{\sigma^2}{n} \le 2\left\|\nabla f_n(\bar{\theta}^t)\right\|^2 + 2L^2\Omega^t + \frac{\sigma^2}{n},$$

the last step by $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ and Equation (10).

Step 5, collect. Combining Steps 3 and 4,

$$\mathbb{E} f_n(\bar{\theta}^{t+1}) \le f_n(\bar{\theta}^t) - \eta_t\Big(\frac{1}{2} - L\eta_t\Big)\left\|\nabla f_n(\bar{\theta}^t)\right\|^2 + \Big(\frac{\eta_t L^2}{2} + L^3\eta_t^2\Big)\Omega^t + \frac{L\eta_t^2\sigma^2}{2n}.$$

The step-size restriction enters here and only here. For $\eta_t \le \frac{1}{4L}$ we have $\frac{1}{2} - L\eta_t \ge \frac{1}{4}$ and $\frac{\eta_t L^2}{2} + L^3\eta_t^2 = \eta_t L^2(\frac{1}{2} + L\eta_t) \le \frac{3}{4}\eta_t L^2 \le \eta_t L^2$. Taking a further expectation over $\mathcal{F}_t$ gives Equation (9). Assumption 2 is not used in the lemma itself and is required only when the telescope of Appendix D.7 is truncated at f_n^*. □

D.5 The consensus recursion

Lemma 14 (Consensus recursion with explicit constants)**.** *Under Assumptions 1 and 3 to 5, for a constant step size $\eta \le \frac{1-\lambda_n}{7L}$ and all $t \ge \tau_k$,*

$$\mathbb{E}[\Omega^t] \;\le\; \rho^{\,t-\tau_k}\,\Omega^{\tau_k} + \frac{C_\Omega\,\eta^2(\sigma^2+\zeta_n^2)}{(1-\lambda_n)^2}, \qquad \rho := \frac{3+\lambda_n^2}{4} < 1, \qquad C_\Omega = 24. \tag{11}$$

Proof. Stack the n active models as the rows of $\Theta^t \in \mathbb{R}^{n\times d}$, put $J = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and $\hat{\Theta}^t = (I-J)\Theta^t$, so that $\Omega^t = \frac{1}{n}\left\|\hat{\Theta}^t\right\|_F^2$. Let G^t stack the stochastic gradients, ∇^t their conditional means, and $\Xi^t = G^t - \nabla^t$, with hats again denoting left multiplication by $I-J$. If $\lambda_n = 0$ then $W^t = J$, the deviation vanishes after a single step and Equation (11) is immediate, so assume $\lambda_n > 0$.

Step 1, the deviation recursion. From $\Theta^{t+1} = W^t(\Theta^t - \eta G^t)$ and $JW^t = W^tJ = J$, which holds because W^t is symmetric and doubly stochastic,

$$\hat{\Theta}^{t+1} = (I-J)W^t(\Theta^t - \eta G^t) = (W^t - J)(\Theta^t - \eta G^t) = (W^t-J)\hat{\Theta}^t - \eta\,(W^t-J)\hat{G}^t,$$

using $(W^t - J)J = 0$. Assumption 5 gives $\|(W^t-J)X\|_F \le \lambda_n\|X\|_F$ for every X.

Step 2, split off the noise before Young's inequality. Conditionally on $\mathcal{F}_t$ the noise $\hat{\Xi}^t$ has zero mean, so the cross term vanishes and

$$\begin{aligned}\mathbb{E}\left\|\hat{\Theta}^{t+1}\right\|_F^2 &= \mathbb{E}\left\|(W^t-J)(\hat{\Theta}^t - \eta\hat{\nabla}^t)\right\|_F^2 + \eta^2\,\mathbb{E}\left\|(W^t-J)\hat{\Xi}^t\right\|_F^2 \\ &\le \lambda_n^2\,\mathbb{E}\left\|\hat{\Theta}^t - \eta\hat{\nabla}^t\right\|_F^2 + \eta^2\lambda_n^2\,\mathbb{E}\left\|\hat{\Xi}^t\right\|_F^2,\end{aligned}$$

with $\mathbb{E}\left\|\hat{\Xi}^t\right\|_F^2 \le \mathbb{E}\,\|\Xi^t\|_F^2 \le n\sigma^2$ by Assumption 3 and because $I-J$ is an orthogonal projection. Performing the split at this point matters, since the Young parameter of Step 3 would otherwise multiply the noise term by $1+\beta^{-1}$ and cost a factor $(1-\lambda_n^2)^{-1}$ in C_Ω.

Step 3, Young's inequality on the deterministic part. With $\beta = \frac{1-\lambda_n^2}{2\lambda_n^2} > 0$ we get $(1+\beta)\lambda_n^2 = \frac{1+\lambda_n^2}{2}$ and $(1+\beta^{-1})\lambda_n^2 = \frac{\lambda_n^2(1+\lambda_n^2)}{1-\lambda_n^2} \le \frac{2}{1-\lambda_n^2}$, so

$$\lambda_n^2\left\|\hat{\Theta}^t - \eta\hat{\nabla}^t\right\|_F^2 \le \frac{1+\lambda_n^2}{2}\left\|\hat{\Theta}^t\right\|_F^2 + \frac{2\eta^2}{1-\lambda_n^2}\left\|\hat{\nabla}^t\right\|_F^2.$$

Step 4, the gradient-deviation bound. Row i of $\hat{\nabla}^t$ is $\nabla f_i(\theta_i^t) - \bar{\nabla}^t$. Insert the exact identity

$$\nabla f_i(\theta_i^t) - \bar{\nabla}^t = \big[\nabla f_i(\theta_i^t) - \nabla f_i(\bar{\theta}^t)\big] + \big[\nabla f_i(\bar{\theta}^t) - \nabla f_n(\bar{\theta}^t)\big] + \big[\nabla f_n(\bar{\theta}^t) - \bar{\nabla}^t\big]$$

and use $\|a+b+c\|^2 \le 3(\|a\|^2+\|b\|^2+\|c\|^2)$. Averaged over i, the first bracket contributes at most $L^2\Omega^t$ by Assumption 1, the second at most ζ_n^2 by Assumption 4 evaluated at the common point $\bar{\theta}^t$, and the third at most $L^2\Omega^t$ by Equation (10). Hence

$$\frac{1}{n}\left\|\hat{\nabla}^t\right\|_F^2 \le 3\big(L^2\Omega^t+\zeta_n^2+L^2\Omega^t\big) = 6L^2\Omega^t+3\zeta_n^2. \tag{12}$$

Step 5, contraction. Dividing the recursion by n and collecting Steps 2 to 4, and using $\lambda_n^2 \le 1$ on the noise term,

$$\mathbb{E}[\Omega^{t+1}] \le \Big(\frac{1+\lambda_n^2}{2}+\frac{12\,\eta^2L^2}{1-\lambda_n^2}\Big)\mathbb{E}[\Omega^t]+\frac{6\,\eta^2\zeta_n^2}{1-\lambda_n^2}+\eta^2\sigma^2.$$

The step-size condition is what closes the contraction. For $\eta \le \frac{1-\lambda_n}{7L}$,

$$\frac{12\,\eta^2L^2}{1-\lambda_n^2} \le \frac{12\,(1-\lambda_n)^2}{49\,(1-\lambda_n^2)} = \frac{12\,(1-\lambda_n)}{49\,(1+\lambda_n)} \le \frac{1-\lambda_n^2}{4} \iff 48 \le 49\,(1+\lambda_n)^2,$$

which holds for every $\lambda_n \ge 0$ and is tight only in the limit $\lambda_n \to 0$. The contraction factor is therefore at most $\frac{1+\lambda_n^2}{2}+\frac{1-\lambda_n^2}{4} = \frac{3+\lambda_n^2}{4} = \rho$, and $1-\rho = \frac{1-\lambda_n^2}{4}$.

Step 6, unroll. A recursion $x_{t+1} \le \rho x_t + F$ with $\rho < 1$ gives $x_t \le \rho^{t-\tau_k}x_{\tau_k} + \frac{F}{1-\rho}$. Here $F = \frac{6\eta^2\zeta_n^2}{1-\lambda_n^2}+\eta^2\sigma^2$ and $\frac{1}{1-\rho} = \frac{4}{1-\lambda_n^2}$, so the steady-state part is

$$\frac{24\,\eta^2\zeta_n^2}{(1-\lambda_n^2)^2}+\frac{4\,\eta^2\sigma^2}{1-\lambda_n^2} \le \frac{24\,\eta^2\zeta_n^2}{(1-\lambda_n)^2}+\frac{24\,\eta^2\sigma^2}{(1-\lambda_n)^2} = \frac{24\,\eta^2(\sigma^2+\zeta_n^2)}{(1-\lambda_n)^2},$$

using $(1-\lambda_n^2)^2 \ge (1-\lambda_n)^2$ for the first fraction and $\frac{4}{1-\lambda_n^2} \le \frac{4}{1-\lambda_n} \le \frac{24}{(1-\lambda_n)^2}$ for the second. This is Equation (11) with $C_\Omega = 24$. □

Remark 15 (Why no gradient-norm forcing appears). The forcing term of Equation (11) contains no $\mathbb{E}\left\|\nabla f_n(\bar{\theta}^t)\right\|^2$ contribution, which is what allows the consensus bound to be substituted into the descent telescope without absorbing a term back into the left side. The reason is that Assumption 4 is assumed uniformly in θ, so the middle bracket of Step 4 is controlled at the common point $\bar{\theta}^t$ and no gradient of f_n survives. Under the weaker relaxed-growth condition $\frac{1}{n}\sum_i \|\nabla f_i(\theta)\|^2 \le \hat{\zeta}^2+B^2\|\nabla f_n(\theta)\|^2$ often used in the decentralized literature, an additional $B^2\mathbb{E}\left\|\nabla f_n(\bar{\theta}^t)\right\|^2$ forcing term does appear and has to be absorbed, tightening the admissible step size by a factor depending on B. Assumption 4 is the case $B=0$.

Remark 16 (Time-varying step sizes). If η is replaced by a schedule $\{\eta_t\}$ with $\eta_t \le \eta_{\max} \le \frac{1-\lambda_n}{7L}$, the proof goes through unchanged. The contraction factor of Step 5 depends on η_t only through the inequality that is verified at $\eta_{\max}$, and the forcing is increasing in η_t^2, so Equation (11) holds with η replaced by $\eta_{\max}$.

D.6 Summing from the join shock

Summing Equation (11) over the post-join window and using $\sum_{j\ge0}\rho^j = \frac{1}{1-\rho} = \frac{4}{1-\lambda_n^2}$,

$$\sum_{t=\tau_k}^{T-1}\mathbb{E}[\Omega^t] \;\le\; \frac{4\,\Omega^{\tau_k}}{1-\lambda_n^2}+\frac{24\,\eta^2(\sigma^2+\zeta_n^2)\,(T-\tau_k)}{(1-\lambda_n)^2}. \tag{13}$$

This is the one place at which the join enters the dynamics. The first term is the join shock of Theorem 1, appearing as the initial condition of a recursion that is otherwise the standard one for decentralized SGD, and the second is the steady state that the same recursion would reach in a network where no client joined late. The separation between the two is exactly the separation between the transient and persistent terms of Equation (4), and it is what makes the cost of a late join computable rather than entangled with the cost of decentralization.

D.7 Proof of Theorem 2

Theorem 2 (restated). *Under (A1)–(A5), for $\eta \le \frac{1-\lambda_n}{7L}$, the criterion $\mathcal{M}(\tau_k, T)$ of Equation (2) is bounded by the four terms of Equation (4).*

Proof. The condition $\eta \le \frac{1-\lambda_n}{7L}$ implies $\eta \le \frac{1}{7L} \le \frac{1}{4L}$, so both Lemma 13 and Lemma 14 apply. Rearranging Equation (9) at constant η and summing over $t = \tau_k, \dots, T-1$, the function values telescope and $f_n(\bar{\theta}^T) \ge f_n^*$ by Assumption 2, leaving

$$\frac{\eta}{4} \sum_{t=\tau_k}^{T-1} \mathbb{E}\left\|\nabla f_n(\bar{\theta}^t)\right\|^2 \le \left(f_n(\bar{\theta}^{\tau_k}) - f_n^*\right) + \eta L^2 \sum_{t=\tau_k}^{T-1} \mathbb{E}[\Omega^t] + \frac{L\eta^2\sigma^2(T-\tau_k)}{2n}.$$

The first bracket is G_{τ_k} by definition. Divide by $\frac{\eta(T-\tau_k)}{4}$ and substitute Equation (13). The initial-condition part of that sum contributes $\frac{4L^2}{T-\tau_k} \cdot \frac{4\Omega^{\tau_k}}{1-\lambda_n^2} = \frac{16L^2\Omega^{\tau_k}}{(1-\lambda_n^2)(T-\tau_k)}$, which is term (ii). The steady-state part contributes $4L^2 \cdot \frac{24\eta^2(\sigma^2+\zeta_n^2)}{(1-\lambda_n)^2} = \frac{96\eta^2L^2(\sigma^2+\zeta_n^2)}{(1-\lambda_n)^2}$, which is term (iv). The noise term contributes $\frac{4}{\eta(T-\tau_k)} \cdot \frac{L\eta^2\sigma^2(T-\tau_k)}{2n} = \frac{2L\eta\sigma^2}{n}$, which is term (iii), and the inherited gap contributes $\frac{4G_{\tau_k}}{\eta(T-\tau_k)}$, which is term (i). □

Every constant in Equation (4) is traceable to one place, which is worth recording because a bound with named constants admits exactly that check.

Constant	Appears in	Origin
4	term (i)	division by $\eta(T-\tau_k)/4$, from $\frac{1}{2} - L\eta \ge \frac{1}{4}$ in Step 5 of Lemma 13
16	term (ii)	$4L^2$ from the same division, times $\frac{4}{1-\lambda_n^2} = \frac{1}{1-\rho}$
$\rho = \frac{3+\lambda_n^2}{4}$	term (ii)	$\frac{1+\lambda_n^2}{2} + \frac{1-\lambda_n^2}{4}$, Step 5 of Lemma 14
2	term (iii)	$\frac{4}{\eta} \cdot \frac{L\eta^2}{2n}$, from the second moment σ^2/n
96	term (iv)	$4L^2 \cdot C_\Omega$ with $C_\Omega = 24$ from Step 6 of Lemma 14
7	the cap	the smallest integer for which $48 \le 49(1+\lambda_n)^2$ holds at $\lambda_n = 0$

The constants are not claimed to be optimal. The threshold 7 in particular is the smallest integer that closes Step 5 of Lemma 14 at $\lambda_n = 0$, and the margin 49 against 48 leaves visible slack at every $\lambda_n > 0$, which is consistent with the perturbation study of Appendix K.8 finding no instability when the cap's premise is violated mildly.

The same bound read as a specification gives Table 2, which is what Section 3.3 argues in prose. Only the first two rows depend on the join, and they are the two Section 4 acts on.

Table 2: The terms of Equation (4), what determines each, and what can reduce it. Only the first two rows depend on the join.

Term	Determined by	Reduced by	Type
(i) inherited gap	$\epsilon_1,\ \Delta_k^{\text{misfit}},\ \eta$	the step size	transient
(ii) join shock	$D_k^2,\ \Omega^{\tau_k -}$	the initialization, and nothing else	transient
(iii) variance floor	η, σ^2, n	the step size; join-independent	persistent
(iv) topology floor	$\eta, \lambda_n, \zeta_n^2$	the step size and the graph; join-independent	persistent

D.8 Proof of Theorem 3

Theorem 3 (restated). *Under (A2), $0 \le G_{\tau_k} \le \frac{n-1}{n}\epsilon_1 + \frac{1}{n}\Delta_k^{\text{misfit}}$, where $\epsilon_1 = f_{n-1}(\bar{\theta}^{\tau_k}) - f_{n-1}^*$ and $\Delta_k^{\text{misfit}} = f_k(\bar{\theta}^{\tau_k}) - f_k^*$.*

Proof. The lower bound $G_{\tau_k} \geq 0$ is the definition of f_n^* as an infimum. For the upper bound, the identity $f_n = \frac{n-1}{n} f_{n-1} + \frac{1}{n} f_k$ evaluated at $\bar{\theta}^{\tau_k}$ gives

$$f_n(\bar{\theta}^{\tau_k}) = \frac{n-1}{n}\big(f_{n-1}^* + \epsilon_1\big) + \frac{1}{n}\big(f_k^* + \Delta_k^{\text{misfit}}\big).$$

For the infima, every θ satisfies $f_n(\theta) \geq \frac{n-1}{n} f_{n-1}^* + \frac{1}{n} f_k^*$, and taking the infimum over θ on the left preserves the inequality, so

$$f_n^* \;\geq\; \frac{n-1}{n} f_{n-1}^* + \frac{1}{n} f_k^*. \tag{14}$$

This is the only step that is not immediate. It is an inequality rather than an identity because the infimum of a sum dominates the sum of the infima, with equality only when f_{n-1} and f_k share a minimizer. Subtracting Equation (14) from the display above gives the claim, and Assumption 2 is what makes all three infima finite so that the subtraction is defined. □

The slack in Equation (14) is exactly the extent to which the incumbents and the joining client disagree about where the optimum lies, so the decomposition is loose in precisely the regime it is used to describe. We make no tightness claim. What the decomposition delivers is not a sharp estimate of G_{τ_k} but an observable upper bound on it, and Lemma 8 bounds the cost of the resulting looseness.

D.9 Phase-1 drift in the flow limit

The results in this subsection are stated in the deterministic flow limit and are the only ones in the paper that leave the discrete stochastic model. Nothing in Section 3, Section 4 or Section 5 depends on them. They are included because they answer, exactly, a question the decomposition of Theorem 3 raises but does not settle, namely what the incumbents' own optimization does to the objective that only becomes the network's at τ_k.

Definition 17 (Flow limit)**.** Statements below take $\sigma \to 0$ and $\eta \to 0$ with $s = \eta t$ held fixed, and write Φ_{n-1}^s for the time-s flow map of $-\nabla f_{n-1}$. In this limit the incumbents are at exact consensus, so $\Omega^{\tau_k -} = 0$, and a warm join in the sense of Corollary 4 gives $\theta_k^{\tau_k} = \bar{\theta}^{\tau_k -}$ and hence $D_k = 0$. By Equation (3) the post-join average then coincides with the Phase-1 switch point, $\bar{\theta}^{\tau_k} = \Phi_{n-1}^{\tau_k}(\bar{\theta}^0)$. For a join with $D_k \neq 0$ every statement holds with the switch point replaced by $\Phi_{n-1}^{\tau_k}(\bar{\theta}^0) + \frac{1}{n}(\theta_k^{\tau_k} - \Phi_{n-1}^{\tau_k}(\bar{\theta}^0))$, again by Equation (3), and we do not restate the case each time.

Lemma 18 (Exact Phase-1 drift law)**.** *Along* $\frac{d\theta^s}{ds} = -\nabla f_{n-1}(\theta^s)$*,*

$$\frac{d}{ds} f_n(\theta^s) = -\frac{n-1}{n} \left\|\nabla f_{n-1}(\theta^s)\right\|^2 - \frac{1}{n} \left\langle \nabla f_k(\theta^s), \nabla f_{n-1}(\theta^s) \right\rangle. \tag{15}$$

Consequently f_n can increase at θ^s only where $\langle \nabla f_k, \nabla f_{n-1} \rangle < -(n-1)\left\|\nabla f_{n-1}\right\|^2$, which forces $\nabla f_{n-1}(\theta^s) \neq 0$ and, by Cauchy–Schwarz, the strict inequality $\left\|\nabla f_k(\theta^s)\right\| > (n-1)\left\|\nabla f_{n-1}(\theta^s)\right\|$.

Proof. The chain rule gives $\frac{d}{ds} f_n(\theta^s) = -\left\langle \nabla f_n(\theta^s), \nabla f_{n-1}(\theta^s) \right\rangle$, and substituting $\nabla f_n = \frac{n-1}{n}\nabla f_{n-1} + \frac{1}{n}\nabla f_k$ yields Equation (15). The right side is positive exactly when $\langle \nabla f_k, \nabla f_{n-1} \rangle < -(n-1)\left\|\nabla f_{n-1}\right\|^2$. If $\nabla f_{n-1} = 0$ the left side of that inequality is 0 and the right side is 0, a contradiction, so $\nabla f_{n-1} \neq 0$. Cauchy–Schwarz then gives $(n-1)\left\|\nabla f_{n-1}\right\|^2 < -\left\langle \nabla f_k, \nabla f_{n-1} \right\rangle \leq \left\|\nabla f_k\right\| \left\|\nabla f_{n-1}\right\|$, and dividing by $\left\|\nabla f_{n-1}\right\| > 0$ gives the stated bound. □

Corollary 19 (Compatible clients)**.** *Under Definition 17, if $\left\langle \nabla f_k(\theta), \nabla f_{n-1}(\theta) \right\rangle \geq 0$ along the Phase-1 orbit, then f_n is non-increasing during Phase 1 and $G_{\tau_k} \leq G_0 := f_n(\bar{\theta}^0) - f_n^*$ for every τ_k.*

It follows that the incumbents' optimization degrades the objective it is not yet minimizing only at points where the joining client's gradient opposes theirs and exceeds it in norm by the factor $n-1$. That is a quantitative description of an atypical client rather than a generic one, and it grows more demanding as the network grows. The converse question, whether descending f_{n-1} can descend f_n faster than descending f_n itself would, has an equally short answer.

Lemma 20 (Amplified alignment)**.** *At any point θ, the instantaneous decrease of f_n along the f_{n-1}-flow exceeds that along the f_n-flow if and only if*

$$\langle \nabla f_n(\theta), \nabla f_{n-1}(\theta) - \nabla f_k(\theta) \rangle > 0. \tag{16}$$

Proof. The two rates of decrease are $\langle \nabla f_n, \nabla f_{n-1} \rangle$ and $\|\nabla f_n\|^2$. Their difference is $\langle \nabla f_n, \nabla f_{n-1} - \nabla f_n \rangle$, and $\nabla f_{n-1} - \nabla f_n = \frac{1}{n}(\nabla f_{n-1} - \nabla f_k)$ follows from $\nabla f_n = \frac{n-1}{n}\nabla f_{n-1} + \frac{1}{n}\nabla f_k$. Positivity of the difference is therefore equivalent to Equation (16). □

Equation (16) holds when the incumbent direction is at least as well aligned with descent of f_n as ∇f_n is, and longer. It is not an exotic condition, and Appendix F.1 exhibits instances on both sides of it, including one in which f_n is strongly convex. That is the mechanism behind the absence of a universal sign in Proposition 26, and it is the reason the typical direction of the delay penalty is recorded there as a heuristic rather than as a theorem.

E Proofs: The Calibrated Step Size

E.1 The weighted criterion and the optimality of a constant step size

A schedule makes the unweighted average of Equation (2) the wrong object, because the descent lemma produces one factor of η_t per round on the left of the telescope. The standard remedy is the step-size-weighted average, which is also the quantity certified by randomized stopping with round t drawn with probability η_t/S. Fix a schedule $\{\eta_t\}_{t=\tau_k}^{T-1}$ with $\eta_t \in (0, \eta_c]$, put

$$S := \sum_{t=\tau_k}^{T-1} \eta_t, \qquad \mathcal{M}_w := \frac{1}{S} \sum_{t=\tau_k}^{T-1} \eta_t \, \mathbb{E} \left\| \nabla f_n(\bar{\theta}^t) \right\|^2,$$

and note that $\mathcal{M}_w = \mathcal{M}(\tau_k, T)$ whenever the schedule is constant.

Summing Equation (9) over the window, telescoping, and dividing by $S/4$ gives the schedule-level analogue of Theorem 2,

$$\mathcal{M}_w \;\le\; \underbrace{\frac{4G_{\tau_k}}{S}}_{\text{(i)}} + \underbrace{\frac{4L^2}{S} \sum_{t=\tau_k}^{T-1} \eta_t \, \mathbb{E}[\Omega^t]}_{\text{(ii)+(iv)}} + \underbrace{\frac{2L\sigma^2}{nS} \sum_{t=\tau_k}^{T-1} \eta_t^2}_{\text{(iii)}}, \tag{17}$$

where Lemma 13 applies round by round because $\eta_t \le \eta_c \le \frac{1}{4L}$, and Remark 16 controls the middle group. Write

$$\Psi(\{\eta_t\}) := \frac{4G_{\tau_k}}{S} + \frac{2L\sigma^2}{nS} \sum_{t=\tau_k}^{T-1} \eta_t^2 \tag{18}$$

for the part of Equation (17) corresponding to terms (i) and (iii) of Equation (4). This is the quantity Theorem 6 concerns, and the middle group is treated separately in Appendix E.2.

Theorem 6 (restated). *For $\sigma^2 > 0$, $\Psi(\{\eta_t\}) \ge 4\sqrt{2L\sigma^2 G_{\tau_k}/(nT')}$ over all schedules with $\eta_t > 0$, with equality if and only if $\eta_t \equiv \eta^\star = \sqrt{2nG_{\tau_k}/(L\sigma^2 T')}$. The optimum is admissible when $T' \ge 98nLG_{\tau_k}/(\sigma^2(1-\lambda_n)^2)$, and otherwise the optimum over $\eta_t \in (0, \eta_c]$ is $\eta_t \equiv \eta_c$.*

Proof. Step 1, Cauchy–Schwarz removes the shape. By Cauchy–Schwarz applied to $\sum_t \eta_t \cdot 1$,

$$S^2 = \Big(\sum_{t=\tau_k}^{T-1} \eta_t \Big)^2 \le T' \sum_{t=\tau_k}^{T-1} \eta_t^2, \qquad \text{so} \qquad \sum_{t=\tau_k}^{T-1} \eta_t^2 \;\ge\; \frac{S^2}{T'}, \tag{19}$$

with equality if and only if η_t is the same for every t in the window. Substituting into Equation (18),

$$\Psi(\{\eta_t\}) \;\ge\; \psi(S) := \frac{4G_{\tau_k}}{S} + \frac{2L\sigma^2}{nT'} \, S, \tag{20}$$

so every schedule is bounded below by a function of its sum alone. This is the whole of the argument that the profile of a schedule carries no leading-order benefit, since ψ cannot distinguish a warmup from a decay from a restart that share a sum.

Step 2, AM–GM fixes the sum. For $A, B > 0$ the function $A/S + BS$ is minimized over $S > 0$ at $S = \sqrt{A/B}$ with value $2\sqrt{AB}$, by AM–GM. With $A = 4G_{\tau_k}$ and $B = \frac{2L\sigma^2}{nT'}$,

$$\psi(S) \;\geq\; 2\sqrt{AB} = 2\sqrt{\frac{8L\sigma^2 G_{\tau_k}}{nT'}} = 4\sqrt{\frac{2L\sigma^2 G_{\tau_k}}{nT'}},$$

with equality if and only if $S = S^\star := \sqrt{\frac{4G_{\tau_k}\, nT'}{2L\sigma^2}} = \sqrt{\frac{2nG_{\tau_k} T'}{L\sigma^2}}$.

Step 3, the two equalities together. Equality in $\Psi \geq 4\sqrt{\cdot}$ needs equality in both Equation (19) and Step 2. The first forces $\eta_t \equiv \eta$ for some η and the second then forces $\eta T' = S^\star$, that is $\eta = S^\star/T' = \sqrt{2nG_{\tau_k}/(L\sigma^2 T')} = \eta^\star$. Conversely the constant schedule $\eta^\star$ attains both, so the characterization is exact.

Step 4, admissibility. $\eta^\star \leq \eta_{\mathrm{c}} = \frac{1-\lambda_n}{7L}$ holds if and only if $\frac{2nG_{\tau_k}}{L\sigma^2 T'} \leq \frac{(1-\lambda_n)^2}{49L^2}$, that is $T' \geq \frac{98\, nLG_{\tau_k}}{\sigma^2(1-\lambda_n)^2}$.

Step 5, the constrained optimum. Suppose $\eta^\star > \eta_{\mathrm{c}}$, so the unconstrained optimum is inadmissible. On $(0, S^\star]$ the envelope ψ is strictly decreasing, since $\psi'(S) = -4G_{\tau_k}/S^2 + \frac{2L\sigma^2}{nT'} < 0$ exactly for $S < S^\star$. Any admissible schedule has $S \leq \eta_{\mathrm{c}} T' < \eta^\star T' = S^\star$, so $\Psi \geq \psi(S) \geq \psi(\eta_{\mathrm{c}} T')$, and the second inequality is an equality only at $S = \eta_{\mathrm{c}} T'$, which under $\eta_t \leq \eta_{\mathrm{c}}$ forces $\eta_t \equiv \eta_{\mathrm{c}}$. That schedule is constant, so Equation (19) is also an equality and the lower bound is attained. The constrained optimum is therefore again a constant, and its value is

$$\psi(\eta_{\mathrm{c}} T') = \frac{4G_{\tau_k}}{\eta_{\mathrm{c}} T'} + \frac{2L\eta_{\mathrm{c}}\sigma^2}{n} = \frac{28LG_{\tau_k}}{(1-\lambda_n)(T-\tau_k)} + \frac{2(1-\lambda_n)\sigma^2}{7n}. \tag{21}$$

□

Two readings should be distinguished. What Step 1 establishes is that no schedule shape helps, which is a statement about this bound rather than about the algorithm, and it is exactly as strong as Equation (17) is tight. What Step 5 establishes is that a very late joiner takes the largest step the spectral gap permits whatever its misfit, so on that branch $\hat{\eta}$ is determined by the graph rather than by the join. The first term of Equation (21) is the residual transient quoted in Section 4.2, and it is not removable by any choice of step size.

E.2 The non-leading terms and two regimes outside the present scope

Theorem 6 concerns Ψ alone, so the middle group of Equation (17) has to be checked separately at $\eta^\star$ before the leading-order conclusion is safe.

Proposition 21 (The omitted terms are lower order)**.** *At the constant schedule $\eta_t \equiv \eta^\star$, with $\eta^\star \leq \eta_{\mathrm{c}}$, terms (ii) and (iv) of Equation* (4) *equal*

$$\frac{16L^2\Omega^{\tau_k}}{(1-\lambda_n^2)(T-\tau_k)} \qquad \textit{and} \qquad \frac{192\, nLG_{\tau_k}(\sigma^2+\zeta_n^2)}{\sigma^2(1-\lambda_n)^2(T-\tau_k)}$$

respectively. Both are $O(1/T')$ and are therefore dominated by the $\Theta((T')^{-1/2})$ leading term as $T' \to \infty$ at fixed n and λ_n, so $\eta^\star$ minimizes Equation (4) *to leading order and is not claimed to be its exact minimizer. The limit is taken at fixed n and the dominance is not uniform in it. The ratio of term (iv) to the leading term is $\Theta\big(n^{3/2}\sqrt{LG_{\tau_k}}(\sigma^2+\zeta_n^2)\sigma^{-3}(1-\lambda_n)^{-2}(T')^{-1/2}\big)$, so term (iv) is genuinely lower order only when $T' \gg n^3 LG_{\tau_k}(\sigma^2+\zeta_n^2)^2\sigma^{-6}(1-\lambda_n)^{-4}$.*

Proof. Term (ii) does not contain η, so it is unchanged. For term (iv), substitute $(\eta^\star)^2 = \frac{2nG_{\tau_k}}{L\sigma^2 T'}$ into $\frac{96\eta^2 L^2(\sigma^2+\zeta_n^2)}{(1-\lambda_n)^2}$ and simplify. □

That threshold is not met at the operating points of Section 5. At $n = 100$ and $\lambda_n = 0.88$ it reads $T' \gg 5 \times 10^9$ against an actual $T' = 100$, and evaluating the two floors directly at the adopted

step gives term (iv) larger than term (iii) by more than three orders of magnitude. In that regime the balance Theorem 6 strikes is between a transient and the *smaller* of the two floors, which is one of the two reasons the prescribed step lands above the empirical minimizer of M (Appendix H). Balancing term (i) against term (iv) instead gives the cube-root law of Remark 22, and the honest reading of Theorem 6 is that it is exact for Ψ and approximate for Equation (4).

The consequence is that the stationarity rate of a network with a late join is $O\big(\sqrt{L\sigma^2 G_{\tau_k}/(nT')}\big) + O(1/T')$, in which the join appears in the leading term only through G_{τ_k} and in the lower-order term only through Ω^{τ_k} and G_{τ_k}. Neither carries a persistent contribution, which is the statement of Section 3.2 read at the optimized step size rather than at a fixed one.

Remark 22 (Near-noiseless training). If $\sigma^2 = 0$ the variance floor vanishes and the balance is between terms (i) and (iv). Minimizing $\frac{4G_{\tau_k}}{\eta T'} + C\eta^2$ with $C = \frac{96L^2\zeta_n^2}{(1-\lambda_n)^2}$ gives $\eta = \big(\frac{G_{\tau_k}(1-\lambda_n)^2}{48\,T'L^2\zeta_n^2}\big)^{1/3}$ and a rate $\Theta\big((T')^{-2/3}\big)$, a cube-root law rather than a square root. The crossover is governed by which of the two floors dominates at the candidate step size, and every configuration reported in Section 5 uses mini-batch gradients with a measured $\hat{\sigma}^2$ bounded away from zero, so this branch is never taken and we report no experiment on it.

Remark 23 (Unknown horizon). Theorem 6 presumes T is known, which every run reported here satisfies. When it is not, the horizon-free schedule $\eta_t = \eta_0(t - \tau_k + 1)^{-1/2}$ with $\eta_0 = \sqrt{2nG_{\tau_k}/(L\sigma^2)}$ and $\eta_0 \le \eta_c$ satisfies, for $T' \ge 4$,

$$\Psi \;\le\; \frac{2 + \log T'}{2} \cdot 4\sqrt{\frac{2L\sigma^2 G_{\tau_k}}{nT'}},$$

so not knowing the horizon costs a logarithmic factor and nothing more. The proof is Equation (18) with $S \ge \eta_0(2\sqrt{T'} - 2) \ge \eta_0\sqrt{T'}$ and $\sum_t \eta_t^2 \le \eta_0^2(1 + \log T')$, both by comparison of the sums with integrals. This does not contradict Theorem 6, whose optimality is asserted among schedules that may use T', and the logarithm is the penalty for not using it. If $\eta_0 > \eta_c$ the cap binds on an initial segment, on which the schedule is the constant η_c of Equation (21).

E.3 Proof of Lemma 8

Lemma 8 (restated). *Let $\hat{G} > 0$ estimate G_{τ_k}, let $\hat{\eta} = \sqrt{2n\hat{G}/(L\sigma^2 T')}$ be the uncapped calibrated step, and let $\kappa = \max\{\hat{G}/G_{\tau_k}, G_{\tau_k}/\hat{G}\}$. Then the leading term is inflated by exactly $\frac{1}{2}\big(\sqrt{\hat{G}/G_{\tau_k}} + \sqrt{G_{\tau_k}/\hat{G}}\big) \le \frac{1}{2}\big(\sqrt{\kappa} + 1/\sqrt{\kappa}\big)$.*

Proof. Write $\phi(\eta) = A/\eta + B\eta$ for the constant-schedule value of Equation (18), with $A = \frac{4G_{\tau_k}}{T'}$ and $B = \frac{2L\sigma^2}{n}$, so that $\eta^\star = \sqrt{A/B}$ and $\phi(\eta^\star) = 2\sqrt{AB}$. Since $\hat{\eta}$ and $\eta^\star$ differ only through the square root of the gap estimate, $\hat{\eta} = c\,\eta^\star$ with $c = \sqrt{\hat{G}/G_{\tau_k}}$. Then

$$\phi(c\eta^\star) = \frac{A}{c\eta^\star} + Bc\eta^\star = \frac{1}{c}\sqrt{AB} + c\sqrt{AB} = \phi(\eta^\star) \cdot \frac{c + c^{-1}}{2}.$$

The map $c \mapsto c + c^{-1}$ is invariant under $c \mapsto 1/c$, so $c + c^{-1} = \sqrt{\kappa} + 1/\sqrt{\kappa}$ and the displayed inequality is in fact an equality. It is stated as an inequality in Section 4.2 because κ is what an implementation can bound, whereas the direction of the error is usually unknown. □

The inflation factor is what justifies a coarse estimator, and it is worth tabulating because Section 5.2 reads observed ratios against it.

Misestimation κ	1	2	4	9	16	100
Inflation $\frac{1}{2}(\sqrt{\kappa} + 1/\sqrt{\kappa})$	1.000	1.061	1.250	1.667	2.125	5.050

The same arithmetic applies to any multiplicative error entering under the square root, so an error of κ in the product $L\sigma^2$ costs the same factor. The one quantity it does not protect is the cap η_c itself, which is a hypothesis of Lemma 13 and Lemma 14 rather than a term in the trade-off. Underestimating L inflates η_c and can violate the premise of both lemmas outright, which is why $\hat{L}$ is deliberately biased upward in Appendix E.4.

E.4 Estimators and the cancellation of n

Proposition 7 (restated). *Substituting Theorem 3 into Theorem 6 gives the calibration rule Equation* (6)*, which reduces to* $\hat{\eta} = \sqrt{2\hat{\Delta}_k/(L\sigma^2(T-\tau_k))}$ *once* $\hat{\epsilon}_1 \approx 0$*, with* n *canceled.*

Proof. Multiplying Theorem 3 by n gives $nG_{\tau_k} \le (n-1)\epsilon_1 + \Delta_k^{\mathrm{misfit}}$, and $\eta^\star$ depends on G_{τ_k} only through the product nG_{τ_k} under the square root, since $\eta^\star = \sqrt{2nG_{\tau_k}/(L\sigma^2 T')}$. Replacing nG_{τ_k} by its upper bound and each population quantity by its estimate gives the second argument of the minimum in Equation (6), and the minimum with η_c enforces admissibility. Setting $\hat{\epsilon}_1 = 0$ leaves $\sqrt{2\hat{\Delta}_k/(L\sigma^2(T-\tau_k))}$, in which n does not appear. □

The cancellation is a consequence of the weight $1/n$ that Theorem 3 attaches to the misfit meeting the factor n that Theorem 6 attaches to the gap, and it holds only on the plateau branch. It is what allows the arriving client to compute the network's step size alone. The same substitution turns the admissibility threshold of Theorem 6 from $T' \ge \frac{98nLG_{\tau_k}}{\sigma^2(1-\lambda_n)^2}$ into $T' \ge \frac{98L[(n-1)\hat{\epsilon}_1+\hat{\Delta}_k]}{\hat{\sigma}^2(1-\hat{\lambda}_n)^2}$, which is free of n on the plateau branch and not before it, the term $(n-1)\hat{\epsilon}_1$ being linear in n exactly as in Equation (6).

The estimators and their biases are as follows. Each is deliberately one-sided, and the direction is chosen so that the resulting error is one that Lemma 8 bounds rather than one that violates a hypothesis.

Quantity	Estimator	Bias	Consequence
$\hat{\Delta}_k$	$f_k(\theta_k^{\tau_k}) - \hat{f}_k^{\mathrm{loc}}$, with $\hat{f}_k^{\mathrm{loc}}$ from a few local steps	under	$\hat{\eta}$ small
$\hat{\epsilon}_1$	0 under the plateau gate, else a gossip fallback	under	$\hat{\eta}$ small
$\hat{L}$	network maximum of local finite-difference probes, scaled up	over	η_c small
$\hat{\sigma}^2$	network maximum of the sample variance of B_σ independent batches	over	$\hat{\eta}$ small
$\hat{\lambda}_n$	exact eigendecomposition of the fixed mixing matrix	none	η_c exact

Three of these require further comment. First, the substitution of Theorem 3 is itself a source of bias in the opposite direction to the estimators, since $nG_{\tau_k} \le (n-1)\epsilon_1 + \Delta_k^{\mathrm{misfit}}$ discards the slack of Equation (14) and therefore biases $\hat{\eta}$ *upward*. Where the incumbents have settled near stationarity of f_{n-1} and f_n obeys a quadratic growth condition, $\nabla f_n = \Theta(1/n)$ gives $G_{\tau_k} = \Theta(n^{-2})$ while the bound supplies only $\Theta(n^{-1})$, so the relaxation inflates $\hat{\eta}$ by $\Theta(\sqrt{n})$. That is a heuristic rather than a consequence of (A1)–(A5), which supply no such growth condition, but it is the second of the two reasons recorded in Appendix H. A further displacement enters because $\hat{\Delta}_k$ is evaluated at $\theta_k^{\tau_k}$ whereas $\Delta_k^{\mathrm{misfit}}$ is defined at $\bar{\theta}^{\tau_k}$, the two differing by $\frac{n-1}{n}D_k$, which Corollary 4 holds at consensus scale and which Lemma 8 absorbs along with the rest. Second, since $\hat{f}_k^{\mathrm{loc}} \ge f_k^*$, the measured misfit under-estimates $\Delta_k^{\mathrm{misfit}}$, whereas taking 0 as a lower bound for f_k^* under non-negative losses would over-estimate it. Both errors are covered by Lemma 8, in opposite directions, and we use the first. Third, the plateau gate replaces $(n-1)\epsilon_1$ by zero, which is an under-estimate of the numerator. If the gate fires only when its intent is met, namely when $(n-1)\epsilon_1 \le \Delta_k^{\mathrm{misfit}}$, then the numerator is under-estimated by at most a factor 2, so $\kappa \le 2$ and Lemma 8 bounds the cost at 1.061. The gate fires rarely in practice (Appendix H), and where it does not the fallback path retains the n-dependent form of Equation (6).

F THE DELAY PENALTY AT A COMMON WINDOW AND STEP SIZE

Whether a client that joins later is worse off than one that joins earlier is the question the title of the paper raises, and this section gives the most that the analysis supports. Two limits on the answer should be stated before it. Everything below compares *upper bounds* rather than realized errors, which is admissible only because the difference is computed with exact constants, and everything below concerns stationarity rather than attained value. Nothing in Section 5 tests any of it, since the experiments reported there vary the join time within the late-join model rather than against a counterfactual in which the same client never joined late. We include the section because the negative half of its answer, that no regime supplies a universal sign, is what keeps the rest of the paper from claiming one.

Fix a network of n clients and compare two runs over a common window $[\tau_k, T]$. The first joins client k at round τ_k. The second joins the same client at $\tilde{\tau}_k < \tau_k$, so that by round τ_k it has already been training on f_n for $\tau_k - \tilde{\tau}_k$ rounds, and the special case $\tilde{\tau}_k = 0$ is the run in which no client joins late at all. Write $\tilde{\theta}^t$ for the second run's average, $\widetilde{G}_{\tau_k} = f_n(\tilde{\theta}^{\tau_k}) - f_n^*$ for its gap at τ_k, and $\widetilde{\Omega}^{\tau_k}$ for its consensus error there. The comparison is admissible only under the measurement window convention of Appendix C.2, so both runs are averaged over the same window with the same $(\eta, \sigma^2, \zeta_n^2, \lambda_n)$. Most informal comparisons in the literature average one run over $[\tau_k, T]$ and the other over $[0, T]$, which measures the window rather than the strategy, and per-run tuning of η moves the floors and breaks the cancellation the next result depends on.

Theorem 24 (Exact bound-level delay penalty at a common step size)**.** *Compare two join times over the common window $[\tau_k, T]$ at a common step size $\eta \le \eta_c$. The floor terms (iii) and (iv) of Equation* (4) *cancel* exactly*, and the difference of bounds equals*

$$\frac{4\big(G_{\tau_k} - \widetilde{G}_{\tau_k}\big)}{\eta(T - \tau_k)} \;+\; \frac{16L^2\big(\Omega^{\tau_k} - \widetilde{\Omega}^{\tau_k}\big)}{(1 - \lambda_n^2)(T - \tau_k)}. \tag{22}$$

Consequently the delayed run has the smaller bound if and only if $\big(G_{\tau_k} - \widetilde{G}_{\tau_k}\big) + \frac{4\eta L^2}{1-\lambda_n^2}\big(\Omega^{\tau_k} - \widetilde{\Omega}^{\tau_k}\big) \le 0$.

Proof. Theorem 2 applies to either run over $[\tau_k, T]$. Its proof uses only that all n clients are active from the first round of the window onward, so that Lemma 13 and Lemma 14 hold there, and it treats the consensus error at the first round of the window as an initial condition. Both hold for the second run, whose join is earlier. Writing the two bounds out,

$$U_{\text{late}} = \frac{4G_{\tau_k}}{\eta(T - \tau_k)} + \frac{16L^2\Omega^{\tau_k}}{(1 - \lambda_n^2)(T - \tau_k)} + \frac{2L\eta\sigma^2}{n} + \frac{96\eta^2L^2(\sigma^2 + \zeta_n^2)}{(1 - \lambda_n)^2},$$
$$U_{\text{early}} = \frac{4\widetilde{G}_{\tau_k}}{\eta(T - \tau_k)} + \frac{16L^2\widetilde{\Omega}^{\tau_k}}{(1 - \lambda_n^2)(T - \tau_k)} + \frac{2L\eta\sigma^2}{n} + \frac{96\eta^2L^2(\sigma^2 + \zeta_n^2)}{(1 - \lambda_n)^2},$$

the third and fourth terms of each are built from $\eta, L, \sigma^2, n, \lambda_n$ and ζ_n^2 alone. None of those depends on the join time, on G_{τ_k}, on Ω^{τ_k} or on D_k, so the two pairs are numerically identical and subtract to zero. What remains is Equation (22). The stated equivalence follows on multiplying by $\frac{\eta(T-\tau_k)}{4} > 0$. □

The cancellation is the entire reason a sign statement is permitted at all. Under the bound-comparison convention of Appendix C.2 a sign claim requires exact constants, and here the constants that would otherwise carry unquantified slack are the same on both sides and vanish. The comparison is decided by the warm-start-gap difference and the join-shock difference and by nothing else. It remains a comparison of guarantees, and Appendix G.1 explains why that does not settle which run ends up better.

Corollary 25 (The join-shock difference is signed under a scale condition)**.** *By Equation* (3)*, $\Omega^{\tau_k} - \widetilde{\Omega}^{\tau_k} \ge \frac{n-1}{n^2}D_k^2 - \widetilde{\Omega}^{\tau_k}$, so the second term of Equation* (22) *is non-negative whenever*

$$D_k^2 \;\ge\; \frac{n^2}{n-1}\,\widetilde{\Omega}^{\tau_k}. \tag{23}$$

Equation (23) is a sufficient condition and not a consequence of (A1)–(A5), and it is a loose one, since it is obtained by discarding the non-negative term $\frac{n-1}{n}\Omega^{\tau_k-}$ and therefore carries a factor $\frac{n^2}{n-1} \approx n$ in front of $\widetilde{\Omega}^{\tau_k}$. It is loose enough to be uninformative at the scales we run. No counterfactual early-join run is performed here, so $\widetilde{\Omega}^{\tau_k}$ is not measured and we read it at the incumbent consensus scale Ω^{τ_k-} that Appendices K.2 and K.3 do report. At the default operating point, $n = 100$ with $\Omega^{\tau_k-} \approx 0.0127$ and a cold $D_k^2 \approx 0.64$, the right side of Equation (23) is 1.29 and the condition *fails* by a factor of two. In the $n = 100$ arm of the size sweep, where $\Omega^{\tau_k-} = 0.0092$ and $D_k^2 = 1.12$, it holds by 21%. Neither is the order-of-magnitude margin a scale argument would lead one to expect, and the reason is the discarded term rather than the scales.

Retaining that term gives the condition that actually governs. If the two runs carry comparable incumbent dispersion, $\Omega^{\tau_k-} \geq \widetilde{\Omega}^{\tau_k}$, then $\Omega^{\tau_k} - \widetilde{\Omega}^{\tau_k} \geq \frac{n-1}{n^2} D_k^2 - \frac{1}{n}\widetilde{\Omega}^{\tau_k}$, so the second term of Equation (22) is non-negative whenever

$$D_k^2 \;\geq\; \frac{n}{n-1}\,\widetilde{\Omega}^{\tau_k}. \tag{24}$$

The threshold is smaller by a factor of n, evaluating to 0.0128 and 0.0093 on the two configurations above, which the measured D_k^2 exceeds by 50-fold and 121-fold. The extra hypothesis is what buys that, and it is an assumption about the two runs rather than a consequence of the identity, so we state it rather than absorb it. Under a cold initialization, where D_k^2 sits at the scale of a fresh random model while $\widetilde{\Omega}^{\tau_k}$ sits at consensus scale, bounded by $\rho^{\tau_k - \tilde{\tau}_k}\widetilde{\Omega}^{\tilde{\tau}_k} + \frac{24\eta^2(\sigma^2+\zeta_n^2)}{(1-\lambda_n)^2}$ by Lemma 14 applied to the earlier run, Equation (24) holds by orders of magnitude. This is a scale argument, and we identify it as such. Under the warm initialization of Corollary 4, D_k^2 is itself driven to consensus scale and the asymmetry disappears up to that scale, which is the tractable half of the penalty and the half the paper acts on. The other half is the gap difference, and it is not tractable at all.

F.1 Neither ordering is universal

The instances below live in the deterministic flow limit of Definition 17, so they concern G_{τ_k} and $\widetilde{G}_{\tau_k}$ directly rather than the bounds built from them. They are constructions rather than observations, and their role is to rule something out.

Proposition 26 (No universal ordering of the inherited gaps)**.** *There exist instances satisfying Assumptions 1 to 5, including one in which f_n is* strongly convex *and every objective-side assumption holds globally and unmodified, realizing $G_{\tau_k} > \widetilde{G}_{\tau_k}$, and instances realizing $G_{\tau_k} < \widetilde{G}_{\tau_k}$. No regime assumption available in this setting therefore supplies a universal sign for Equation* (22)*.*

Before the construction it is worth naming the mechanism, since a counterexample that only exhibits a failure is less informative than one that explains it. The channel is Lemma 20. When $\langle \nabla f_n, \nabla f_{n-1} - \nabla f_k \rangle > 0$ the incumbent direction is at least as well aligned with descent of f_n as ∇f_n itself and longer, so descending the objective the network is not yet minimizing descends the one it will minimize *faster*, for as long as the condition persists. No curvature assumption closes that channel, which is why the strongly convex instance below behaves the same way as the non-convex one.

Proof. Instances (i) and (ii). Take $n = 2$ and, for $b > 0$,

$$f_1(x) = \tfrac{x^2}{2} + bx, \qquad f_k(x) = \tfrac{x^2}{2} - bx, \qquad \text{so} \qquad f_n(x) = \tfrac{x^2}{2}.$$

Then f_n is 1-strongly convex and hence Polyak–Łojasiewicz with $\mu = 1$, each f_i is 1-smooth and bounded below by $-b^2/2$, and $\frac{1}{2}\sum_i \|\nabla f_i(x) - \nabla f_n(x)\|^2 = b^2$ for *every* x, so Assumption 4 holds globally with $\zeta_n^2 = b^2$ and no modification of the objectives is needed. Here $f_{n-1} = f_1$ is minimized at $-b$, and the misfit $\Delta_k^{\text{misfit}}(x) = f_k(x) - f_k^* = \frac{(x-b)^2}{2}$ is strictly positive near $x = 0$, so neither instance is degenerate in the misfit.

For (i), start both runs at $x_0 > 0$. The flows are $\tilde{\theta}(s) = x_0 e^{-s}$ for the run with no delay and $x(s) = (x_0 + b)e^{-s} - b$ for Phase 1 of the delayed run. Setting $\bar{\tau} := \log \frac{x_0+b}{b} > 0$, on $0 < s < \bar{\tau}$ we have $x(s) - \tilde{\theta}(s) = b(e^{-s} - 1) < 0$ and $x(s) > 0$, hence $0 < x(s) < \tilde{\theta}(s)$ and

$$G_{\tau_k} = \tfrac{x(\tau_k)^2}{2} \;<\; \tfrac{\tilde{\theta}(\tau_k)^2}{2} = \widetilde{G}_{\tau_k} \qquad \text{for every } \tau_k \in (0, \bar{\tau}).$$

The mechanism is Equation (16), which here reads $\langle \nabla f_n, \nabla f_{n-1} - \nabla f_k \rangle = x \cdot 2b > 0$ for $x > 0$.

For (ii), start both runs at $-b$, the minimizer of f_{n-1}. Phase 1 is then stationary and $x(s) \equiv -b$, so $G_{\tau_k} = f_n(-b) - f_n^* = \frac{b^2}{2}$ for every τ_k, while $\tilde{\theta}(s) = -be^{-s}$ gives $\widetilde{G}_{\tau_k} = \frac{b^2}{2}e^{-2\tau_k} < \frac{b^2}{2} = G_{\tau_k}$ strictly for every $\tau_k > 0$.

As a check on Lemma 18, along Phase 1 of this instance $\frac{d}{ds} f_n = -x(x+b)$, so f_n increases exactly on $-b < x < 0$, which is after the orbit has crossed the minimizer of f_n. The drift law predicts increase exactly when $\langle \nabla f_k, \nabla f_1 \rangle < -\|\nabla f_1\|^2$, that is when $(x-b)(x+b) < -(x+b)^2$, which rearranges to $2x(x+b) < 0$ and hence to $-b < x < 0$.

Instance (iii). Take smooth, lower-bounded functions with bounded gradients agreeing on $[-2, 2]$ with $f_i(\theta) = \theta^2/2$ for $i \neq k$ and with $f_k \equiv c$. The modification outside $[-2, 2]$ is required for Assumption 4 to hold globally, since the unmodified pair has a gradient gap growing linearly in $|\theta|$, and it is harmless because every trajectory below stays in $[0, 1]$ where the displayed formulas are exact. On that region $\Delta_k^{\text{misfit}} \equiv 0$ and $\nabla f_{n-1} = \frac{n}{n-1}\nabla f_n$, so Phase 1 is f_n-descent along f_n's own flow lines at a speed multiplied by $\frac{n}{n-1}$, which is Equation (16) in the extreme case $\nabla f_k = 0$. Starting both runs at $\theta_0 \in (0, 1]$ gives $\theta(s) = \theta_0 e^{-s}$ and $\tilde{\theta}(s) = \theta_0 e^{-\frac{n-1}{n}s}$, whence

$$G_{\tau_k} = \tfrac{n-1}{2n}\theta_0^2 e^{-2\tau_k} \;<\; \tfrac{n-1}{2n}\theta_0^2 e^{-2\frac{n-1}{n}\tau_k} = \widetilde{G}_{\tau_k} \qquad \text{strictly, for every } \tau_k > 0. \qquad \square$$

The three instances separate cleanly. Instance (ii) is the saturating case, in which the incumbent flow settles at a stationary point of f_{n-1} whose f_n-value the earlier run strictly undercuts, and it is the behavior one expects of a client whose data the incumbents do not represent. Instances (i) and (iii) run the other way through amplified alignment, (iii) degenerately with zero misfit and (i) with genuine misfit, genuine heterogeneity and strong convexity. In (i) the advantage is not permanent, but it outlasts the crossing $\bar{\tau} = \log(1 + x_0/b)$ at which the proof stops. Since $\frac{1}{2}(\tilde{\theta}(s)^2 - x(s)^2) = \frac{1}{2}b(1 - e^{-s})\big[(2x_0 + b)e^{-s} - b\big]$, the delayed run keeps the strictly smaller gap on the whole of $\tau_k < \log(1 + 2x_0/b)$, and only past that point does the ordering reverse toward the behavior of (ii). We state the proof on $(0, \bar{\tau})$ because that interval suffices for the claim. Curvature governs the rate of forgetting and never the sign of the comparison, which is the content of the strongly convex instance and the reason the next statement is not a theorem.

Heuristic 27 (The typical direction, and it is not a theorem). For a client with substantial misfit joining after a long Phase 1, and absent sustained amplified alignment in the sense of Equation (16), the expected direction is $\widetilde{G}_{\tau_k} \leq G_{\tau_k}$, so that Equation (22) is non-negative and grows with τ_k as the earlier run's gap keeps shrinking while the delayed run's gap saturates near the f_n-value of a stationary point of f_{n-1}.

One half of its hypothesis is not a hypothesis at all in the regime Section 4.2 operates in. At exact incumbent stationarity $\nabla f_{n-1}(\theta) = 0$ forces $\nabla f_n(\theta) = \frac{1}{n}\nabla f_k(\theta)$, so

$$\langle \nabla f_n(\theta), \nabla f_{n-1}(\theta) - \nabla f_k(\theta) \rangle = \big\langle \tfrac{1}{n}\nabla f_k(\theta), -\nabla f_k(\theta) \big\rangle = -\tfrac{1}{n}\|\nabla f_k(\theta)\|^2 \;\leq\; 0,$$

and amplified alignment in the sense of Equation (16) is precluded outright, strictly so wherever $\nabla f_k(\theta) \neq 0$. Writing $a = \nabla f_{n-1}(\theta)$ and $b = \nabla f_k(\theta)$ the exact form is $\frac{n-1}{n}\|a\|^2 - \frac{n-2}{n}\langle a, b \rangle - \frac{1}{n}\|b\|^2$, which stays non-positive on a neighborhood $\|a\| = O(\|b\|/n)$ of stationarity, so the preclusion degrades gracefully rather than holding only at a single point. What this does not supply is the other half, namely that the incumbents reach such a neighborhood, and Appendix H records that our own runs do not certify it.

We attach no genericity claim of any kind to Heuristic 27, use it in no theorem, and test it in no experiment. Proposition 26 shows its negation occurs, through a channel that is transient in instance (i) and degenerate in instance (iii) but real in every curvature regime. Its testable form is a same-window, common-step-size measurement of $\mathcal{M}$ for two join times with everything else held fixed, which is a study we have not run and which would in any case measure realized errors rather than the bounds Theorem 24 compares.

The conclusion of this section is narrower than its title may suggest. Joining late carries no persistent cost in this bound. Its cost is a transient whose magnitude is what Section 4 acts on and whose sign, at the bound level and in general, is a property of the instance rather than of the lateness.

G What the Stationarity Guarantee Certifies

G.1 Stationarity does not certify function value

Theorem 2 certifies that the network spends its post-join window near stationarity of f_n on average. A stationary point of a non-convex f_n may be a saddle or a local minimum of arbitrary value, so two runs carrying identical stationarity guarantees may attain very different values, and a run with the worse stationarity bound may sit in the better basin. Every comparative statement in Appendix F is therefore a statement about guarantees rather than about outcomes. This is not a hedge added for safety. It is the logical content of the bound, and it is what the measurement convention of Appendix C.2 records.

Three consequences run through the paper. Accuracy is reported as an observational quantity everywhere in Section 5, and the theory undertakes to predict none of it. The step size that minimizes the certified metric and the step size that maximizes accuracy need not coincide, and Figure 1 shows that they do not, which confirms the semantics rather than exposing a defect in the rule. Moreover, the split in Table 1, where a baseline attains the lowest $\mathcal{M}$ on two datasets while losing 0.10 to 0.18 in accuracy, is arbitrated by exactly this distinction rather than by a preference between the two columns.

The failure runs in both directions. If $U_A < U_B$ for two upper bounds, the realized errors may be ordered either way, since the bounds differ in slack. Even the exact difference of Equation (22) compares bounds, because G_{τ_k} and $\widetilde{G}_{\tau_k}$ enter as descriptions of two different trajectories and identical machinery reduces but does not remove the asymmetry in slack.

G.2 Basin selection

Rate analysis answers how fast the system approaches stationarity. In a non-convex landscape there is a second question with no convex analogue, namely which stationary point it approaches, and the join time participates in the answer. The result below is the only place in the paper where that participation is described. It is qualitative, it holds under dynamical hypotheses that go well beyond (A1)–(A5), no experiment reported here tests it, and no claim elsewhere in the paper depends on it. We state it because Section 3.2 asserts that a late join carries no persistent penalty in Equation (4), and it is therefore necessary to state what that assertion leaves out.

Gradient flows of merely smooth functions need not converge to a point, basin boundaries need not be negligible, and an orbit may meet a boundary non-transversally, so the hypotheses have to be explicit.

(D1) f_n is real-analytic, or satisfies a Łojasiewicz inequality at each critical point, with bounded sublevel sets, so that every gradient-flow trajectory converges to a single critical point.

(D2) f_n is a Morse function, so that critical points are isolated, the basin map B_n sending a point to the limit of the f_n-flow started there is well defined everywhere, and the basin boundaries form a closed measure-zero set off which B_n is locally constant.

(D3) The Phase-1 orbit $s \mapsto \Phi^s_{n-1}(\bar{\theta}^0)$ meets the basin-boundary set at isolated parameter values, transversally, each contact passing between the interiors of two distinct basins. This is needed only for the clean jump picture.

Proposition 28 (Path dependence of the limit)**.** *In the flow limit of Definition 17 and under (D1) and (D2), the limit attained by a run that joins at τ_k is $\theta^\infty(\tau_k) = B_n(\Phi^{\tau_k}_{n-1}(\bar{\theta}^0))$, the f_n-basin containing the point the incumbent flow has reached by τ_k. As a function of τ_k this map is locally constant on every interval whose orbit segment avoids the basin boundary, so it can change only where the orbit meets that boundary. Under (D3) it is piecewise constant with jumps exactly at the crossings. It is in general distinct from the limit $B_n(\bar{\theta}^0)$ of a run with no delay.*

Proof. (D1) makes B_n well defined. (D2) makes it locally constant off a closed measure-zero set, so its composition with the continuous orbit $s \mapsto \Phi^s_{n-1}(\bar{\theta}^0)$ is locally constant wherever the orbit avoids that set, and a change therefore requires boundary contact. Without further hypotheses the orbit could run along the boundary or accumulate contacts, which (D3) excludes by making every

contact an isolated transverse crossing between basin interiors, so the map is piecewise constant and each crossing changes the basin. □

Proposition 28 establishes that the join time is a genuine degree of freedom over the attained solution. It does not establish how large the effect is or in which direction it runs, and it cannot, since that depends on the landscape of f_n and (A1)–(A5) say nothing about the landscape. It is the only mechanism compatible with Equation (4) by which a delayed run could reach a better limit than an early one, the instances of Proposition 26 improving only the transient gap, and it is equally a mechanism for a worse one. We do not propose a way to control it. The hypotheses are also real restrictions rather than technical conveniences, since overparameterized networks typically have non-isolated critical manifolds and violate (D2) outright. The conclusion should survive at the coarser level of critical components and critical values, and we do not formalize that here.

G.3 The Polyak–Łojasiewicz case

Assumption 6 (Polyak–Łojasiewicz; used only here)**.** $\|\nabla f_n(\theta)\|^2 \geq 2\mu(f_n(\theta) - f_n^*)$ for all θ, some $\mu > 0$.

Overparameterized networks near interpolation often satisfy Assumption 6 approximately (Karimi et al., 2016), which makes it the natural condition under which to ask how much of the preceding picture is an artifact of assuming nothing beyond smoothness.

Proposition 29 (PL removes the value consequences of basin selection)**.** *Under Assumption 6 every stationary point of f_n is a global minimizer, so all basins carry the value f_n^* and Proposition 28 has no consequence for the attained value, only for the attained parameter. The converse fails, so failure of Assumption 6 is necessary but not sufficient for the join time to affect the attained value.*

Proof. If $\nabla f_n(\theta) = 0$ then Assumption 6 gives $0 \geq 2\mu(f_n(\theta) - f_n^*)$ and hence $f_n(\theta) = f_n^*$, which is the first claim. The minimizer set may still be a positive-dimensional manifold, so parameter-space path dependence survives. For the converse, $f(x) = \log(1 + x^2)$ is 2-smooth and bounded below, has $x = 0$ as its unique critical point and therefore a single basin, yet is not PL, since $\|f'(x)\|^2 = \frac{4x^2}{(1+x^2)^2} \to 0$ while $f(x) - f^* = \log(1 + x^2) \to \infty$ as $|x| \to \infty$. Sufficiency would additionally require at least two attracting critical points with distinct values. □

Theorem 30 (Post-join bound under PL)**.** *Under (A1)–(A5) and Assumption 6, for a constant $\eta \leq \eta_c$ and $\rho = \frac{3+\lambda_n^2}{4}$,*

$$\mathbb{E}\big[f_n(\bar{\theta}^T) - f_n^*\big] \leq \Big(1 - \frac{\eta\mu}{2}\Big)^{T-\tau_k} G_{\tau_k} + \frac{L\eta\sigma^2}{n\mu} + \frac{48\,\eta^2 L^2(\sigma^2 + \zeta_n^2)}{\mu(1-\lambda_n)^2} + \eta L^2 \Omega^{\tau_k}\,(T - \tau_k)\,\bar{\rho}^{\,T-\tau_k-1}, \tag{25}$$

with $\bar{\rho} = \max\{\rho, 1 - \frac{\eta\mu}{2}\} < 1$ and $T > \tau_k$.

Proof. Subtract f_n^* from both sides of Equation (9) and apply Assumption 6 to the gradient term, which gives $-\frac{\eta}{4}\mathbb{E}\left\|\nabla f_n(\bar{\theta}^t)\right\|^2 \leq -\frac{\eta\mu}{2} G^t$ with $G^t := \mathbb{E}[f_n(\bar{\theta}^t) - f_n^*]$, so

$$G^{t+1} \leq \Big(1 - \frac{\eta\mu}{2}\Big) G^t + \eta L^2 \mathbb{E}[\Omega^t] + \frac{L\eta^2\sigma^2}{2n}.$$

The contraction factor is in $(0, 1)$ because $\mu \leq L$ and $\eta \leq \frac{1}{4L}$ give $\frac{\eta\mu}{2} \leq \frac{1}{8}$. Unroll over $[\tau_k, T]$. The constant noise forcing sums against the geometric weights to at most $\frac{2}{\eta\mu} \cdot \frac{L\eta^2\sigma^2}{2n} = \frac{L\eta\sigma^2}{n\mu}$. Split $\mathbb{E}[\Omega^t]$ by Equation (11). Its steady-state part sums to $\frac{2}{\eta\mu} \cdot \eta L^2 \cdot \frac{24\eta^2(\sigma^2+\zeta_n^2)}{(1-\lambda_n)^2} = \frac{48\eta^2 L^2(\sigma^2+\zeta_n^2)}{\mu(1-\lambda_n)^2}$, and its initial-condition part is the convolution $\eta L^2 \Omega^{\tau_k} \sum_{j=0}^{T-\tau_k-1} (1 - \frac{\eta\mu}{2})^{T-\tau_k-1-j} \rho^j \leq \eta L^2 \Omega^{\tau_k} (T - \tau_k) \bar{\rho}^{\,T-\tau_k-1}$, using $\sum_{j=0}^{m-1} a^{m-1-j} b^j \leq m \max\{a, b\}^{m-1}$. □

Under PL the inherited gap and the join shock are both forgotten *geometrically* in function value, rather than at the rate $1/(T - \tau_k)$ of Equation (4), so the qualitative picture familiar from the strongly convex case is restored for the value while remaining silent about the parameter reached. The truly

non-convex regime is characterized jointly by polynomial forgetting and by the possibility that basin selection carries value consequences, and Proposition 29 shows that these two are not independent.

The step-size conclusion changes with it, which matters because Section 4.2 states that it does. The exact constant-optimality of Theorem 6 rests on the transient having the form $4G_{\tau_k}/S$, so that Cauchy–Schwarz can eliminate the shape of a schedule and leave a function of its sum alone. Under PL the transient is instead at most $e^{-\mu S/2}G_{\tau_k}$, which still depends on the schedule only through S, whereas the noise accumulation $\sum_t \eta_t^2 \prod_{s>t}(1-\frac{\eta_s\mu}{2})$ now discounts early rounds and does not discount late ones. Noise injected early is forgotten and noise injected late is not, so front-loading the step size is strictly favored and the argument of Theorem 6 does not transfer. Minimizing over constants the exponential surrogate of the first two terms of Equation (25), namely $e^{-\eta\mu(T-\tau_k)/2}G_{\tau_k} + \frac{L\eta\sigma^2}{n\mu}$, which upper bounds them since $1-x \le e^{-x}$, gives $\eta^\star_{\mathrm{PL}} = \frac{2}{\mu(T-\tau_k)} \log \frac{n\mu^2(T-\tau_k)G_{\tau_k}}{2L\sigma^2} = \Theta\big(\frac{\log(T-\tau_k)}{\mu(T-\tau_k)}\big)$ whenever that logarithm is positive, with value $\frac{2L\sigma^2}{n\mu^2(T-\tau_k)}\big(1+\log\frac{n\mu^2(T-\tau_k)G_{\tau_k}}{2L\sigma^2}\big)$. Setting the derivative of the discrete expression itself to zero gives a condition in $(T-\tau_k-1)$-th powers with no such closed form, which is why the surrogate is used. The best constant therefore pays a logarithmic factor over $1/(T-\tau_k)$, and the standard weighted-averaging argument for decaying step sizes removes it (Koloskova et al., 2020). That is the sense in which decaying schedules regain their advantage under PL, and it is a logarithmic advantage rather than a change of rate.

Two consequences follow. First, in the near-interpolation regime, where PL is approximately valid, the value-level late-join question collapses, since the gap is forgotten geometrically and basin selection has no value consequence. Second, the constant-step optimality of Theorem 6 is a property of the non-convex transient rather than a general fact about schedules, so a reader who transports it to a PL setting will be transporting it out of the regime where it was derived.

H LIMITATIONS

The list is complete rather than representative, and where a limitation has a measured cost that cost is stated beside it.

- **Scope of the mixing assumption.** Schemes whose individual rounds may be disconnected, such as sporadic gossip, are not covered (Remark 9). The results hold for them with $1-\lambda_n$ replaced by a multi-round gap, which we have not carried through. Every topology we run is connected at every round, so no reported result lies outside the assumption.
- **The network-size sweep runs under a modified protocol.** Isolating n required a fixed per-client shard and a Dirichlet allocation resampled at each size, so the total data grows with n and absolute values are not comparable across the sweep (Appendix K.3). Every claim drawn from it is a paired comparison within one n or a ratio. The degree is held at 6 while n grows, so $\hat{\lambda}_n$ rises across the sweep and is reported at each size; a topology whose spectral gap is fixed in n would remove that, and we did not run one.
- **The prescribed step lands an order of magnitude above the empirical minimizer of M, and the cause is in the analysis rather than in the estimators.** It sits 8 to 14 times above it (Section 5.2). Lemma 8 bounds the cost at $(c+1/c)/2$ and the observed cost stays inside the envelope on both branches, so the overshoot costs a constant rather than a rate. Two features of the derivation account for its direction. Theorem 6 balances term (i) against term (iii) and treats term (iv) as lower order, which Proposition 21 shows requires $T' \gg n^3(1-\lambda_n)^{-4}$ and which fails by seven orders of magnitude at our operating point, where term (iv) exceeds term (iii) by more than a factor of 10^3. Separately, substituting Theorem 3 for nG_{τ_k} discards the slack of Equation (14), which under a quadratic growth condition at incumbent stationarity inflates $\hat{\eta}$ by $\Theta(\sqrt{n})$, or a factor of 10 at $n=100$ (Appendix E.4). Both push in the observed direction and the second matches the observed magnitude. The estimators push the other way. Since $\hat{L}$ and $\hat{\sigma}^2$ enter Equation (6) only in denominators, accumulating them as maxima makes $\hat{\eta}$ *smaller* and therefore closer to the empirical minimizer, so replacing them with windowed estimators would widen the gap rather than close it. They remain a defect for a different reason, recorded next.

- **The estimators of L and σ^2 are order statistics over clients.** Both are accumulated as network maxima, so $\hat{L}$ grows with the number of clients drawn rather than with any property of the objective, which Appendix K.3 measures directly as $\hat{L}$ rising from 5.70 at $n = 20$ to 8.31 at $n = 200$. A windowed or quantile estimator would remove that dependence. We have not implemented one, and the previous item records why doing so would not by itself improve the calibration.
- **The plateau gate never fires in our runs.** The gate that would set $\hat{\epsilon}_1 = 0$ and reduce Equation (6) to its n-free form requires agreement across all clients and is too strict at these network sizes, so every run takes the fallback path through $\hat{\epsilon}_1$. The fallback is Equation (6) exactly as stated, so no run is conducted outside the rule the analysis prescribes, and the plateau precondition of Theorem 6 is assessed instead through the stability of $\hat{\Delta}_k$, which Section 5.2 finds turning over at $\tau_k = 0.5T$, the same point at which the predicted monotonicity begins. The cancellation of n in Proposition 7 is not exercised. What Appendix K.3 adds is why, and the reason is the cap rather than the gate. On the sparse graph the minimum in Equation (6) selects η_c at every size from $n = 20$ to $n = 200$, so $\hat{G}_n$ never enters the adopted step and the branch on which the cancellation has content is never taken. Forcing the gate changes $\hat{G}_n$ and leaves $\hat{\eta}$ identical to the last digit. The n-dependence that does appear factors into $(1 - \hat{\lambda}_n)$, which moves because the degree is held fixed while n grows, and $\hat{L}$, which grows because it is a maximum accumulated over more clients; both lie outside Proposition 7 and the second is the estimator defect recorded above. The claim is untested by us, for a located reason, on a branch our default topology does not enter.
- **The joinless control does not resolve the size of a persistent penalty.** Appendix K.1 finds the late-window certified metric of a network with no late joiner, one joining at $0.3T$ and one joining at $0.6T$ agreeing to 20 to 80 times inside the across-seed spread, which meets the criterion fixed in advance. Withholding the client for the whole run, the largest perturbation a single client's participation history admits, also stays inside that spread. At $n = 100$ the arriving client enters f_n with weight $1/n$ and the criterion is quadratic in the gradient, so the design establishes the absence of a detectable persistent penalty and cannot bound its size. A design that could would have to perturb a constant fraction of the network, and we did not run one.
- **Uneven baseline coverage.** Methods requiring momentum fall outside the momentum-free floor and are not compared against (Appendix J.3). The heterogeneity sweep carries no baselines, EMNIST was dropped, and the staggered multi-join scenario covers only CIFAR-10 and CIFAR-100. Table 1 is accordingly an ablation at a fixed floor rather than a survey of the field, which is what Section 5 claims for it and no more.
- **The ablation isolates the rule and not the step size it selects.** Every row of Table 1 shares one Phase-1 step size and switches at τ_k, so a calibrated row differs from its fixed-step counterpart by a network-wide change of step size as well as by the misfit that sets it, on Tiny-ImageNet from 0.03 to 0.00254. No result here asserts that difference. Theorem 6 claims optimality within the bound class and Proposition 7 local computability, and P6, P8 and P9, which test them, are met. The value alone does not account for the gap either, since a constant at 0.003 from round 0 attains $M = 2.89$ on Tiny-ImageNet against 1.54 at the selected constant (Appendix K.4). A fixed step decayed at τ_k by the same factor and without the rule would separate the two, and we did not run one.
- **The step-size rule under staggered arrivals is claimed only where the precondition holds.** When 20 clients arrive at rounds drawn from $(0.25T, T]$, the effect of the rule on the certified metric reverses relative to the single controlled join and it costs 0.048 in accuracy (Appendix K.11). Drawing the same 20 arrivals from $(0.5T, 0.8T]$, so that each falls inside the plateau region, removes the reversal in every seed. The boundary is the precondition of Theorem 6 and not the composition of the measurement window, which we vary independently and which does not remove the reversal. Neither manipulation is single-factor, since the second draw compresses the arrivals as well as moving them and the fixed-membership window of the first is 18 rounds. We claim the attribution and not an exclusion, and the region in which the rule is effective is narrower than the range of conditions a deployment would present.
- **A perturbation in the dangerous direction did not destabilize training.** Under-estimating λ_n inflates η_c and violates the premise of Lemma 13 and Lemma 14, which Lemma 8 does

Algorithm 2 The two rules as implemented. Lines marked *est.* are engineering estimators rather than hypotheses of any result stated in this paper, and their bias directions are those of Appendix E.4.

Require: horizon T, conservatism constant c_L, batch count B_σ, refresh period K, plateau window W_p and tolerance tol_p

Every round, every incumbent i (background)

1: every K rounds, refresh $\hat{L}_i \leftarrow c_L \max_{\text{probes}} \|\nabla f_i(\theta + \delta u; \xi) - \nabla f_i(\theta; \xi)\| / \delta$ ▷ est., same batch ξ both sides

2: every K rounds, refresh $\hat{\sigma}_i^2 \leftarrow \frac{1}{B_\sigma - 1} \sum_{b=1}^{B_\sigma} \|g_b - \bar{g}\|^2$ from B_σ independent batches at θ_i^t ▷ est.

3: update the loss EMA and set the local plateau bit $p_i \leftarrow 1$ if the relative change over W_p rounds stays below tol_p ▷ est.

4: flood $\hat{L} \leftarrow \max_i \hat{L}_i$, $\hat{\sigma}^2 \leftarrow \max_i \hat{\sigma}_i^2$ and $P \leftarrow \bigwedge_i p_i$ ▷ network maxima, upward bias

Round $\tau_k - 1$, the incumbents

5: each neighbor $j \in N_k$ sends client k a fresh snapshot of $\theta_j^{\tau_k -}$ ▷ one exchange, no extra rounds

Round τ_k, the arriving client k

6: $\theta_k^{\tau_k} \leftarrow |N_k|^{-1} \sum_{j \in N_k} \theta_j^{\tau_k -}$ ▷ Corollary 4; any locally fitted model is discarded

7: $\hat{\Delta}_k \leftarrow f_k(\theta_k^{\tau_k}) - \hat{f}_k^{\mathrm{loc}}$, with $\hat{f}_k^{\mathrm{loc}}$ from a short local fit on a copy ▷ misfit, Theorem 3

8: $\hat{\epsilon}_1 \leftarrow 0$ if $P = 1$, else the fallback of Appendix I.2

9: broadcast the tuple $(\tau_k,\ T,\ \hat{\Delta}_k,\ \hat{\epsilon}_1,\ P,\ \hat{L},\ \hat{\sigma}^2,\ \hat{\lambda}_n)$

Round τ_k, every client $i \in [n]$ independently

10: $\hat{\eta} \leftarrow \min\left\{(1 - \hat{\lambda}_n)/(7\hat{L}),\ \sqrt{2[(n-1)\hat{\epsilon}_1 + \hat{\Delta}_k]/(\hat{L}\hat{\sigma}^2(T - \tau_k))}\right\}$ ▷ Equation (6), from the pinned tuple alone

Each round $t = \tau_k, \ldots, T - 1$, every client, in parallel

11: $\varphi_i \leftarrow \theta_i^t - \hat{\eta}\, g_i^t$, exchange φ, then $\theta_i^{t+1} \leftarrow \sum_j W_{ij}^t \varphi_j$ ▷ Equation (1), in this order

not cover, yet training did not diverge at our operating point. It cost 0.023 in accuracy, and the other perturbation that leaves the admissible range cost 0.022 (Appendix K.8). The absence of divergence is evidence of slack in the constant $7L$, consistent with the margin recorded in Appendix D.7. We report both and do not extrapolate from either.

- **Outcomes departed from what the bound leads one to expect, in two distinct ways.** Four of them fall outside every criterion we registered and are listed as such in Appendix J.7, namely the certified metric falling back at the largest grid point on the complete graph, $\hat{\eta}$ decreasing for joins before $0.5T$ where the plateau precondition fails (Section 5.2), the reversal of the step-size rule under staggered arrivals drawn across the whole run, and withholding a client for a whole run lowering rather than raising the late-window metric (Appendix K.1). Three further departures are registered criteria that were not met, P10b, P12 and P13, of which P13 is the absence of an interior minimum in the fixed-step grid up to $\eta = 0.3$, outside the range Theorem 2 covers (Appendix K.4). Table 3 and Table 4 state, for every prediction we test, the condition fixed in advance under which it would have counted as a refutation.
- **Two theoretical sections are scoped rather than settled.** Basin selection holds under (D1)–(D3), which overparameterized networks need not satisfy (Appendix G.2), and the delay penalty of Appendix F is a bound-level result with no experimental counterpart here. No claim in Section 3, Section 4 or Section 5 rests on either.

I Implementation Details

I.1 Full procedure with estimator definitions

Algorithm 1 states the two rules at the level the main text needs. Algorithm 2 gives the form that was run, in which the constants of Equation (6) are replaced by the estimators of Appendix E.4 and the mechanism that makes $\hat{\eta}$ identical across clients is explicit.

Two implementation points carry the correctness properties checked in Section 5.1.

The step size is computed rather than communicated. Line 10 is a deterministic function of the tuple broadcast at line 9, so every client evaluates the same expression on the same inputs and the results agree bit for bit without any agreement procedure. This is why the check of Section 5.1 is an equality rather than a tolerance. The tuple must reach the whole network before τ_k takes effect, so

an implementation sets $\tau_k = t_1 + d +$ slack with d at least the graph diameter, where t_1 is the round at which client k announces itself.

That schedule does not by itself reconcile the two rules, and the tension should be stated plainly. Warm initialization is strictly one-hop, since line 6 uses only the models client k's own neighbors send it. The step-size rule is not, since line 10 has every client in $[n]$ evaluate the same expression on a tuple that contains $\hat{\Delta}_k$, and $\hat{\Delta}_k$ does not exist until line 7 has run at round τ_k. On a graph of diameter $d > 1$ those two rounds cannot coincide. An implementation must choose between three options, and we record which we take. It may pause the gradient steps for d rounds while the tuple floods, which changes the horizon. It may have client k compute $\hat{\Delta}_k$ on neighbor snapshots taken at round $\tau_k - d - 1$ and accept that the warm model is stale by the incumbent drift over d rounds, which Appendix I.3 bounds at $O(d\eta(\|\nabla f_{n-1}\| + \sigma))$. Or it may let client k adopt $\hat{\eta}$ at τ_k and the incumbents adopt it at $\tau_k + d$, which breaks the bit-identical agreement of Section 5.1 during the intervening rounds. Our runs sidestep the choice, computing the tuple centrally at τ_k, which is a research convenience of the same kind as obtaining $\hat{\lambda}_n$ by exact eigendecomposition (Appendix J.5). The sparse graph we run has diameter 3, so the window at issue is short there, whereas the ring of Appendix K.6 has diameter 50 and the choice would be material.

The round is adapt-then-combine and the order matters. Each client computes its local half-step φ_i first, then all clients exchange, then each mixes. A per-round snapshot is cloned once per sender and shared by all receivers, which removes any dependence on the order in which a client processes incoming models and gives the synchronous semantics Equation (1) assumes. Mixing before the local step, or letting a receiver mix models that already reflect the current round, breaks the identity of Theorem 1 at the join round and is detectable as a systematic residual.

I.2 The plateau gate and its fallback

The gate exists because $(n-1)\hat{\epsilon}_1$ is the one non-local term in Equation (6). Each incumbent maintains an exponential moving average of its local training loss and sets its bit when the relative change over W_p rounds falls below tol_p, and the network conjunction P gates the substitution $\hat{\epsilon}_1 \leftarrow 0$. When $P = 1$ the calibration reduces to the n-free form of Proposition 7. Appendix E.4 bounds the cost of that substitution at a factor 1.061 whenever the gate fires only where its intent holds.

The conjunction over 100 clients is strict enough that $P = 1$ never occurred in our runs, so every reported run took the fallback path, in which $\hat{\epsilon}_1$ is estimated from the gossip average of incumbent losses relative to its running minimum. That path evaluates Equation (6) in the form the analysis states, with n present. A less strict gate, for instance a quorum rather than a conjunction, would be the obvious repair and we did not implement one.

I.3 Warm initialization variants

Corollary 4 bounds D_k^2 by $\frac{n-1}{|N_k|}\Omega^{\tau_k-}$ for any neighbor set, so averaging over the whole network should be indistinguishable from averaging over immediate neighbors once both are at consensus scale. Both were implemented. Averaging the models of all active clients and averaging the models of the $|N_k| = 6$ graph neighbors agree to within 0.0001 in final accuracy and in the certified metric, which is the predicted indistinguishability rather than a coincidence.

Waiting for several rounds before joining is a natural variant, and it yields little. A second round of neighbor averaging reduces D_k^2 by a further constant factor at best, since D_k^2 is already at consensus scale after the first, while the incumbent average moves by $O(m\eta(\|\nabla f_{n-1}\| + \sigma))$ over m rounds of waiting. The reduction is second order and the drift is first order, so a single round is optimal, and we use one.

Balancing the two transients. Proposition 5 is the minimizer of the majorant U, and the derivation is three lines. Writing $\delta = \theta_k^{\tau_k} - \bar{\theta}^{\tau_k-}$ and $g_\star = \nabla f_n(\bar{\theta}^{\tau_k-})$, Theorem 1 gives $\bar{\theta}^{\tau_k} = \bar{\theta}^{\tau_k-} + \delta/n$ and $\Omega^{\tau_k} = \frac{n-1}{n}\Omega^{\tau_k-} + \frac{n-1}{n^2}\|\delta\|^2$, so terms (i) and (ii) of Equation (4) become, after majorizing f_n at $\bar{\theta}^{\tau_k-}$,

$$U(\delta) = U(0) + \frac{4}{\eta T' n}\langle g_\star, \delta\rangle + \frac{\beta}{2}\|\delta\|^2, \qquad \frac{\beta}{2} = \frac{2L}{\eta T' n^2} + \frac{16L^2(n-1)}{(1-\lambda_n^2)T' n^2}.$$

Setting the gradient to zero gives Equation (5), in which T' and n^2 cancel, so the optimal initialization does not depend on the remaining horizon. The quadratic is exact in δ, which is why the loss at any other choice is the identity $\frac{\beta}{2}\left\|\delta-\delta^\star\right\|^2$ rather than a bound; what is a bound is U itself, which exceeds the two terms it majorizes by at most $4L\left\|\delta\right\|^2/(\eta T' n^2)$.

Three qualifications belong with it. First, $g_\star$ is the network gradient at the incumbent average and is not available to client k. Under the plateau condition it reduces to $\frac{1}{n}\nabla f_k(\bar{\theta}^{\tau_k-})$ and Equation (5) becomes one local gradient step of size $(L+8\eta L^2(n-1)/(1-\lambda_n^2))^{-1}$, which at the default operating point is 1.93×10^{-3}, or $0.89\,\hat{\eta}$. That is the same substitution Proposition 7 makes and it carries the same exposure. The plateau gate never fires in our runs (Appendix H), and the measured $\|\nabla f_n\|$ at the end of a run is of order $\sqrt{M}$ rather than zero. The two estimates of $\|\delta^\star\|$ differ accordingly, 9×10^{-3} under the plateau form against 1.5×10^{-1} from $\sqrt{M}$, and Section 4.1 quotes the second. Both leave Corollary 4 optimal to leading order; only the size of the residual moves, from 10^{-4} to 10^{-2} of the shock removed.

Second, no experiment here tests Proposition 5. The quantity it predicts is a correction to an effect already below the resolution of every macroscopic measurement we make, so no criterion in Table 3 or Table 4 bears on it, and we register none.

Third, the step of Equation (5) is one gradient evaluation at the inherited model and not a period of local fitting. Appendix I.4 discards the locally fitted model because it sits at the scale of a fresh initialization in D_k; a single step of size $\alpha_n\approx\hat{\eta}$ moves D_k by less than the residual one round of neighbor averaging already leaves, which is the sense in which the two are different objects. It does lower $\hat{\Delta}_k$ and therefore the step Equation (6) prescribes, so the coupling Section 5.3 reports would be slightly stronger under it; we did not run that configuration.

I.4 The cold-join case

If a deployment forbids client k from receiving any model before it contributes one, the initialization rule is unavailable and D_k sits at the scale of a fresh model rather than at consensus scale. Equation (3) still holds, so the shock is still exactly $\frac{n-1}{n^2}D_k^2$, but nothing removes it and the network must rely on the $\rho^{t-\tau_k}$ decay of Lemma 14 to forget it. The decay rate is set by the spectral gap, so accelerating gossip to enlarge $1-\lambda_n$ temporarily is the natural direction. We did not implement it and make no claim about it. The configuration we do run under this name is the cold-initialization ablation of Table 1, which measures the shock rather than mitigating it.

I.5 Violation diagnostics

The initialization rule requires that a locally fitted model be discarded rather than blended into $\theta_k^{\tau_k}$. Initializing instead from a locally fitted model moves D_k from consensus scale to misfit scale, which Equation (3) converts into a proportionally larger Ω^{τ_k}, and the join identities track that exactly. At $n=100$ the consequence is invisible in every macroscopic quantity, the certified metric moving from 0.591 to 0.605 and accuracy by 0.001, both inside the seed spread.

The violation is real and is captured where the theory says it acts, in D_k and hence in Ω^{τ_k}, and it is at the same time too small at these network sizes to reach anything a practitioner would monitor. The suppression by n in Equation (3) reconciles the two, and is what makes the initialization rule invisible in Table 1.

J Experimental Details

J.1 Datasets, models, partitions, and initialization

Three settings are used. CIFAR-10 with LeNet at $n=100$ clients (Krizhevsky, 2009; LeCun et al., 1998), CIFAR-100 with ResNet18-GN at $n=50$, and Tiny-ImageNet with ResNet18-GN at $n=50$ (Le & Yang, 2015; He et al., 2016; Wu & He, 2020). Every run is 200 rounds with one local epoch per round, batch size 64, plain SGD with momentum 0, and a constant step size. The certified metric is evaluated every 5 rounds and in the neighborhood of each join.

Clients are partitioned by a Dirichlet allocation with concentration $\alpha = 0.4$ unless stated, drawn once per dataset and reused across every run so that the partition is not a source of variation between configurations. Each client's shard is then split $80/20$ into a local training set and a local validation set, and accuracy is reported as the mean over clients of accuracy on their own validation split. That convention makes accuracy comparable across configurations at a fixed α and *not* comparable across α, since a more concentrated partition aligns each client's validation labels with its training labels and inflates the number. Appendix K.10 depends on this and notes it explicitly.

Normalization and initialization differ by dataset and both differences are deliberate. LeNet is trained from a random initialization. ResNet18-GN replaces every batch-normalization layer with a group-normalization layer, using $\gcd(32, C)$ groups for C channels, because averaging models across clients destroys batch-normalization running statistics. Its convolutional and fully-connected weights are then initialized from the ImageNet-pretrained ResNet18 checkpoint wherever the shapes match, while the normalization layers are left freshly initialized, since batch-normalization statistics have no group-normalization counterpart and the affine parameters do not carry the same meaning. The final layer is replaced to match the class count.

One consequence of that checkpoint has to be declared. Tiny-ImageNet is built from ImageNet, so a backbone pretrained on ImageNet has seen the source of its evaluation images and the Tiny-ImageNet accuracies are contaminated in the ordinary transfer-learning sense. They are reported as observational quantities that the theory does not undertake to predict, and every comparison drawn from them here is between configurations that share the same checkpoint, so the contamination is common to all of them and cancels within a column. It does not cancel against any externally reported Tiny-ImageNet number, and none is drawn.

A second consequence concerns comparability. The absolute accuracies for CIFAR-100 and Tiny-ImageNet in Table 1 are obtained from a partially pretrained backbone and are not comparable with results trained from scratch, whereas the CIFAR-10 column is trained from scratch and is not comparable with them either. Table 1 already carries that warning for other reasons, and this is one more. What the pretraining does not affect is any comparison drawn *within* a column, since every configuration in a column shares the same initialization, nor any result in Section 5.2 or Appendix K, all of which run on CIFAR-10 with LeNet. It does bear on one mechanism the main text reports. A pretrained backbone yields a smaller measured $\hat{L}$, 4.83 on CIFAR-100 and 1.99 on Tiny-ImageNet against roughly 10 on CIFAR-10, which raises η_{c} and is part of why the square-root branch of Equation (6) runs on those two datasets and the cap binds on CIFAR-10.

CIFAR-100 and Tiny-ImageNet additionally use standard training-time augmentation, random crop with padding 4 and horizontal flip, applied to the training loader only. Evaluation and the full-batch gradients that define the certified metric are computed on unaugmented data. CIFAR-10 uses no augmentation. Together with the per-dataset step size of Appendix K.4, these are the three respects in which the CIFAR-10 column and the other two differ, and they are why the caption of Table 1 forbids reading across columns.

J.2 The same-floor protocol

Every configuration in Table 1 runs at a floor held fixed by construction. Momentum is 0 everywhere, the graph is a fixed symmetric random 6-regular graph with lazy Metropolis weights, the round count is identical, and the join schedule is drawn once per seed and shared by every configuration at that seed, which makes the design paired. The step size is common in the same sense. It is 0.01 on CIFAR-10 and CIFAR-100 and 0.03 on Tiny-ImageNet, used by every configuration throughout Phase 1 and retained through Phase 2 by the fixed-step rows, so that a calibrated row differs from its fixed-step counterpart only in the step adopted at τ_k. On CIFAR-100 and Tiny-ImageNet that constant is the one selected in Appendix K.4; on CIFAR-10 it is a fixed default and was not selected by a grid, which is why Section 5.1 claims a tuned floor for the other two datasets only. Models are exchanged as per-round snapshots, cloned once per sender and shared by all receivers, so that a round has the semantics of Equation (1) rather than depending on the order in which a client processes arrivals. Connectivity is verified for the Phase-1 incumbent set and for the set after every join event, since a late joiner can be a cut vertex and Assumption 5 would otherwise fail silently.

The momentum-free choice is the one that needs an argument rather than a statement. Momentum with coefficient β_{m}, written so here because β already names the curvature of Proposition 5, multi-

plies the effective step size by roughly $(1 - \beta_{\mathrm{m}})^{-1}$, so at $\beta_{\mathrm{m}} = 0.9$ two configurations nominally sharing η are running an order of magnitude apart. Theorem 24 requires a common η for the floor terms to cancel, and Equation (6) prescribes a step size for the update of Equation (1), which has no momentum term. A comparison in which some configurations carry momentum is therefore not a comparison at a common floor, whatever its table reports. The cost of this choice is that methods tuned with momentum cannot be run here, which is Appendix J.3.

J.3 Admissibility of methods as baselines

Table 1 compares configurations that differ only in the two rules, at the floor of Appendix J.2. A method is admissible as a baseline here only if it can be run at that floor and certifies the same quantity, and three groups cannot. Zhu et al. (2026), Hewa Koneputugodage et al. (2026) and Singh et al. (2026) act on when, how often, or with whom clients communicate, so running any of them would change λ_n, the round count or the graph, which are the quantities the floor holds constant; they are orthogonal to the two rules rather than competing with them, since those act at a single round and leave the schedule untouched. Ruan et al. (2021) and Sun et al. (2023b) assume a central aggregator, so their setting has no consensus error to perturb and no quantity corresponding to Ω^{τ_k}; a port to Equation (1) would be a new method rather than a reimplementation, and Corollary 4 uses only the models a joining client's own neighbors send it whereas a server-side rule may condition on the whole population. Shi et al. (2023) is tuned against a different optimizer floor. Its reported configuration uses local momentum, which Appendix J.2 sets to zero, and a perturbation radius of 0.25 paired with an initial step size of 0.1, 3 to 10 times ours, so the radius and the step size are no longer in the ratio the method was tuned at. Run under our floor it degrades to chance on all three datasets and we report no number for it in either direction, not having isolated which of momentum or the ratio is responsible.

J.4 Experiment groups and run counts

The results in this paper come from 507 completed runs. Section 5 reports no counts at all, by design, so they are given here in full.

Group	What it varies	Runs	Reported in
Main, CIFAR-10	7 configurations single join, 6 staggered, 5 seeds	65	Table 1, Appendix K.11
Main, CIFAR-100	7 configurations single join, 6 staggered, 5 seeds	65	Table 1, Appendix K.11
Main, Tiny-ImageNet	7 configurations, 5 seeds	35	Table 1
Step-size selection	4 constants, 3 seeds, two datasets	24	Appendix K.4
Step-size grid	8 points, 3 seeds, two graphs	48	Figure 1, Appendix K.4
Join-time sweep	8 join times, 3 seeds, two graphs	48	Appendix K.5
Topology	4 graphs, 3 seeds	12	Appendix K.6
Schedules	3 schedules, 3 seeds, two graphs	18	Appendix K.7
Perturbation	κ, c_L, $\hat{\lambda}_n$, warm variants	42	Appendix K.8, Appendix I
Cap probe	graph, join time, c_L	30	Appendix K.9
Heterogeneity	$\alpha \in \{0.1, 1.0\}$, 5 seeds	20	Appendix K.10
No late joiner	4 participation histories, 5 seeds	20	Appendix K.1
Step-size grid ext.	2 constants, 3 seeds, two datasets	12	Appendix K.4
Network size	4 sizes, 4 configurations, 3 seeds	48	Appendix K.3
Staggered attribution	4 configurations, 5 seeds, one arrival window	20	Appendix K.11

The topology row counts 4 graphs. The five-row table of Appendix K.6 adds the default 6-regular configuration, which is reported again there rather than rerun.

Seeds are 42 to 46 for the five-seed groups and 42 to 44 for the three-seed ones. Wall-clock time for the groups above is 289 hours on a Slurm-managed cluster of two nodes, one with an Intel Core i9-13900K, 64 GB of memory and one NVIDIA GeForce RTX 4090, the other with an Intel Xeon Silver 4309Y, 128 GB of memory and two NVIDIA A40s. Groups share no state, so the assignment of groups to nodes does not affect any result.

Of the 507 runs, 497 contain a join event; the remaining 10 are the two joinless arms of Appendix K.1, which by construction have none. Those 497 runs record 1624 join events, and the

identity was verified at 1608 of them: 1581 single-client joins, at which the largest relative residual is 1.27×10^{-15}, and 27 same-round two-client joins, at which it is 5.63×10^{-16}. The remaining 16 are same-round joins in the CIFAR-10 runs, whose logs do not carry $\|\sum_{j \in J} \delta_j\|$; the multi-client form of Equation (8) requires that quantity and it cannot be recovered from the individual displacements, so these events are excluded rather than counted as passing. No round in any run produced more than two simultaneous arrivals, so the composition of Proposition 12 is exercised at $|J| = 2$ and no further. The agreement of $\hat{\eta}$ across clients is defined only where the step-size rule is enabled, which is 277 runs, and the deduplicated count of adopted values is 1 in every one of them. Appendix K.3 reports the identity residual at each network size.

J.5 The certified metric and derived quantities

The certified metric M estimates $\mathcal{M}(\tau_k, T)$ of Equation (2). At each evaluation round the network average model is formed, the full-batch gradient of f_n at that model is accumulated in float64 over every client's local training set, and its squared norm is recorded. M is the mean of those values over the evaluation rounds in $[\tau_k, T)$, so it is a sampled estimate of the round average and not the round average itself. Where a run contains several join events, M is taken over the window beginning at the first, which is the convention Appendix K.11 discusses.

Four auxiliary quantities appear in the text. The consensus error Ω^t is accumulated in float64 every round. The spectral quantity $\hat{\lambda}_n$ is obtained by exact eigendecomposition of the mixing matrix on the active induced subgraph, recomputed at every membership change, which is a research convenience rather than something a deployment could do. The measured misfit $\hat{\Delta}_k$ and the adopted step size $\hat{\eta}$ are read from the join event. Cap activity is the number of seeds in which the minimum of Equation (6) selected η_c.

J.6 Statistical procedure

Configurations at a common seed share a join schedule, so comparisons between them are paired and are treated as such. Across-seed standard deviations are reported everywhere. Pairwise comparisons use the paired t-test with the Wilcoxon signed rank test alongside, given the seed counts. With 5 paired samples the smallest attainable Wilcoxon p is 0.0625, so no comparison in this paper can reach $p < 0.05$ by that test, and we report the sign pattern instead. Where Section 5.3 states that the certified metric improves in all 5 seeds, that is the statement being made, and it is the strongest available at this sample size. Section 5.2 likewise requires monotonicity to hold within each seed rather than only in the seed mean.

One group produces an effect that is large against seed noise, and the nulls reported elsewhere are read against it. Under staggered arrivals drawn from the plateau region the step-size rule improves the certified metric by 3.8 across-seed standard deviations, in 5 seeds of 5 (Appendix K.11). The same instrumentation, seed counts and paired design yield differences between 0.03 and 0.88 standard deviations wherever this paper reports an absence, so those are statements about the effects and not about what the design can resolve. Appendix K.1 is the one place where the resolution itself is the binding constraint, and this is stated there.

J.7 Falsification criteria

Section 5 states its predictions in prose and does not tag them. They are collected here in the form in which they were testable: for each, the result it follows from, the condition fixed in advance under which we would have counted it refuted, and the outcome. In total, 15 criteria are stated, 11 of them met, 3 departures and 1 untested. Of those met, one holds only inside the region where its precondition holds and one is met without being stressed, and each criterion not met is discussed where it arises as well as here. The departures the criteria do not cover are listed after them.

Two conventions govern the table. A criterion is admitted only if some run we performed could have produced it, so predictions about quantities we never varied are excluded and listed separately after the table. Moreover, the seed counts of Appendix J.6 do not support significance testing, since with 5 paired samples no comparison here can reach $p < 0.05$ by a Wilcoxon signed rank test. Each

Table 3: Falsification criteria, refutation conditions fixed in advance, and outcomes; continued in Table 4. In total, 15 criteria, 11 met, 3 departures and 1 untested. Verdicts use three labels and three qualifiers. *Pass* means the criterion was met, *departure* that it was not, and *untested* that the branch it concerns was never entered, so the criterion could not bear on it. The qualifier *not stressed* marks a criterion met without being approached, *sub-plateau departure* one met inside its stated precondition while the quantity behaves unexpectedly outside it, and *content holds* one not met because the effect it tracks is below the detection threshold throughout, so the statement it was registered for is not contradicted. Δ denotes the paired difference at matched seeds. Datasets are listed in the order CIFAR-10, CIFAR-100, Tiny-ImageNet.

	Prediction (source)	Counted refuted if	Observed	Verdict
P1	The join event satisfies Equation (3) exactly, and Equation (8) at simultaneous arrivals	any relative residual above float64 accumulation noise at any logged join; alarm threshold 10^{-2}	1.27×10^{-15} over 1581 single-client joins and 5.63×10^{-16} over 27 two-client rounds, in 497 runs; 16 two-client rounds excluded for want of the logged field	pass
P2	Every client adopts the same $\hat{\eta}$ at a join (Algorithm 1)	more than one distinct adopted value at any join	one deduplicated value in each of the 277 runs with the rule enabled	pass
P3	Warm initialization leaves $\Omega^{\tau_k} = \frac{n-1}{n}\Omega^{\tau_k-}$ (Corollary 4)	measured Ω^{τ_k} departing from the prediction by more than 1% in any seed	within 0.8% in all 5 seeds; cold initialization instead raises Ω^{τ_k} by 47 to 53%	pass
P4	The initialization rule is macroscopically invisible in the certified metric at these n (the factor $\frac{n-1}{n^2}$ in Equation (3))	$\lvert\Delta M\rvert$ against D-PSGD exceeding the across-seed spread on any dataset	0.00, $+0.04$, -0.01 against spreads of 0.12, 0.25, 0.07	pass
P5	The same, in observational accuracy	$\lvert\Delta\text{acc}\rvert$ against D-PSGD exceeding the across-seed spread on any dataset	$+0.001$, $+0.002$, $+0.002$ against spreads of 0.010, 0.003, 0.004	pass
P6	Calibration lies inside the square-root envelope of Lemma 8	$M(\hat{\eta})/M(\eta_{\text{best}}) > (c + 1/c)/2$ on either graph	2.64 against 4.06 at $c = 8.0$, and 3.29 against 6.97 at $c = 13.86$	pass
P7	κ-fold misestimation of the gap costs at most $(\sqrt{\kappa} + 1/\sqrt{\kappa})/2$ (Lemma 8)	any ratio $M_\kappa/M_{\kappa=1}$ above the envelope, over $\kappa \in \{\frac{1}{16}, \frac{1}{4}, 1, 4, 16\}$	every ratio at or below 1.00, against envelopes up to 2.13; the envelope is satisfied without being approached	pass, not stressed
P8	$\hat{\eta}$ increases with τ_k once Phase 1 has plateaued (Theorem 6, square-root branch)	failure in any single seed inside the plateau region $\tau_k \geq 0.5T$ on the off-cap graph	strictly increasing per seed, 0.00393 to 0.00720; below $0.5T$ it decreases, where $\hat{\Delta}_k$ is still descending and the precondition fails	pass; sub-plateau departure

criterion is therefore stated as a sign pattern, a per-seed monotonicity, or a comparison against the across-seed standard deviation, and never as a p-value.

Limitations of the criteria. Two quantities that might be expected here are absent, and their absence is a limit on the evidence rather than an omission from the tables. First, observational accuracy is not predicted by any result in the paper, so no criterion in the table is stated on accuracy except P5, where the theory predicts the *absence* of an effect and an absence is falsifiable. Second, the capped branch of Theorem 6 asserts no monotonicity in τ_k, so the sparse-graph column of Appendix K.5, on which $\hat{\eta} \equiv \eta_c$ throughout, is outside P8 rather than a further instance of it.

Sensitivity to the thresholds. Every criterion above that predicts the absence of an effect is stated against the across-seed spread, and that is the loosest choice the seed counts support rather than a significance level. With 3 to 5 seeds the spread itself carries a relative error of 35 to 50%, so a

Table 4: Falsification criteria, continued from Table 3. Neither the three departures nor the untested criterion contradicts a claim stated elsewhere in this paper. P10b and P12 depart because the quantity whose change they predict is below the detection threshold at every point of the sweep, which is the suppression by n that Equation (3) asserts, P13 departs outside the range of η on which Equation (4) holds, and P14 concerns a branch the rule never takes here.

	Prediction (source)	Counted refuted if	Observed	Verdict
P9	No schedule improves on the best constant step size (Theorem 6)	any tested schedule attaining M below the grid-optimal constant on either graph	none does; closest is cosine at 0.338 against 0.224 and at 0.515 against 0.240	pass
P10a	The persistent floors do not depend on τ_k and are those a network with no late joiner pays (Theorem 2)	any pairwise $\lvert\Delta M_{\mathrm{tail}}\rvert$ among the joinless network and joins at $0.3T$ and $0.6T$ exceeding the pooled across-seed s.d.	0.0040, 0.0054 and 0.0014 against a pooled s.d. of 0.1098, five seeds	pass
P10b	The join transient decays window by window (Theorem 2, terms (i)–(ii))	$g(w)$ not decreasing over successive 20-round windows, or $\lvert g\rvert$ outside the spread at the last	not monotone; but $\lvert g\rvert/\mathrm{s.d.} \le 0.07$ in every window including the first, so nothing above noise is present to decay	departure; content holds
P11	The join shock is $\Theta(n^{-1})$ as deployed, not $\Theta(n^{-2})$ (Equation (3), Appendix D.2)	slope of $\log E_n$ on $\log n$ outside $[-1.3, -0.7]$ while R_n varies by less than a factor of two	slope -1.085 with R_n varying by 1.26 over n from 20 to 200, and nE_n flat to 1.20	pass
P12	The initialization rule becomes macroscopically visible as n falls (the factor $\frac{n-1}{n^2}$ in Equation (3))	$\lvert\Delta M\rvert/\mathrm{s.d.}$ not rising monotonically as n falls, or not exceeding 1 at $n = 20$	0.13, 0.16, 0.10, 0.03 at $n = 20, 50, 100, 200$, while the consensus error at the join is multiplied by 7.94 at $n = 20$	departure
P13	The fixed-step grid has an interior minimum in M (Equation (4))	M monotonically decreasing until accuracy collapse or divergence	monotone to $\eta = 0.3$ on both datasets, at roughly 70 and 30 times η_{c} and so outside the range Equation (4) covers	departure
P14	$\hat{\eta}$ is free of n on the plateau branch (Proposition 7)	$\hat{\eta}$ varying across n by more than the across-seed s.d. with the gate forced	0.00263 against a spread of 0.00057; the cap is active at every n, so $\hat{G}_n$ never enters $\hat{\eta}$	untested

threshold an order of magnitude tighter would not be resolvable, and Appendix J.6 explains why no criterion here can be a significance test at all. We therefore report the margin and not only the verdict. Taking the largest ratio each criterion attains, P10a holds at any threshold above 0.05, P4 above 0.16, P6 above 0.65, P5 above 0.67, and P3 and P7 above 0.80. The claim of Theorem 2 that the persistent terms are join-independent is thus the one that survives the tightest reading, by a factor of three against the next; of the two initialization criteria, P4 survives a 6-fold tightening of its threshold and P5 only a 1.5-fold one. P9 is not of this form and is reported as a margin instead, the closest schedule attaining 0.338 against 0.224.

Outcomes outside the criteria. Four outcomes departed from what the analysis leads one to expect and are covered by no criterion above, so they are distinct from the three criteria the table records as departures. They are listed rather than absorbed, and none of the four is explained by a mechanism we have isolated.

1. **The certified metric falls back at the largest grid point on the complete graph**, from 0.753 at $\eta_{\mathrm{c}}/2$ to 0.547 at η_{c}, which no term of Equation (4) predicts, since every term is monotone in η over that range (Appendix K.4).

2. **The calibrated step decreases for joins before $0.5T$.** This is where Appendix K.5 finds the plateau precondition failing, so it lies outside what Theorem 6 asserts and is not a refutation of P8, but it is a departure from the shape the curve would otherwise be expected to have.
3. **Under staggered arrivals drawn across the whole run the effect of the step-size rule on the certified metric reverses sign**, from -0.70 under a single controlled join to $+0.673$ on CIFAR-100, and the accuracy cost rises to 0.048. Appendix K.11 attributes both to the plateau precondition rather than to the measurement window, since drawing the arrivals inside the plateau region restores the improvement in every seed. The departure is that the region within which the rule is effective is narrower than the range of conditions a deployment would present.
4. **Withholding a client entirely lowered the certified metric.** A network of 99 clients that never admits the 100th attains $M_{\text{tail}} = 0.561$, against 0.603 for the network containing it from round 0, measured against the same f_n in both cases (Appendix K.1). The difference is inside one across-seed standard deviation and we draw no conclusion from its sign. What it establishes is that at $n = 100$ the criterion does not resolve single-client participation at all, which is the resolution limit P10a has to be read against.

K Additional Results

Unless stated otherwise these results use CIFAR-10 with LeNet at $n = 100$, 3 seeds, and the floor of Appendix J.2. The exceptions are Appendix K.1, Appendix K.2, Appendix K.10 and Appendix K.11, which run at 5 seeds and state this where they begin, and Appendix K.3, which varies n by construction and states its own protocol. Two mixing configurations recur and are named throughout by what they are. The *sparse* graph is the random 6-regular graph of Section 5.1, with $\hat{\lambda}_n = 0.88$ on average over seeds, on which the cap of Equation (6) is active. The *complete* graph has $\hat{\lambda}_n = 0.50$ and is off the cap. Mechanism groups are reported on both because the two exercise different branches of Theorem 6, and a result that appears only on one is labeled as such.

The seed counts carry a consequence that is easier to state once here than to rediscover at each table. A configuration appearing both in a three-seed group below and in Table 1 is a three-seed estimate of a five-seed quantity, so the two entries are different estimates of the same number rather than the same number. The default sparse configuration is the case a reader is most likely to check. It reads $M = 0.591$ in Appendix K.4 and in Appendix K.5, against $M = 0.64$ in Table 1 and 0.636 before rounding in Appendix K.10, and its paired difference against the fixed-step configuration is correspondingly $+0.043$ at 3 seeds (Appendix K.9) and $+0.077$ at 5 (Appendix K.10). Comparisons drawn within a group are paired at matched seeds and are the ones the text makes. Comparisons drawn between a three-seed group and Table 1 are not paired, and we draw none.

K.1 The persistent floors against a network with no late joiner

Theorem 2 asserts that terms (iii) and (iv) contain neither τ_k nor the arriving client, so that they take the same values for a network in which nobody joins late. Every other configuration in this paper contains a join. This section measures the network the claim is stated against. All configurations use 5 seeds and the floor of Appendix J.2 throughout. All 4 configurations share one partition, one graph, one designated client k and one fixed step size; the last because a calibrated arm would leave the joinless configuration at the Phase-1 step while the others switched at τ_k, and a 5-fold difference in η enters terms (iii) and (iv) directly and would confound exactly the comparison being drawn.

Write M_{tail} for the certified metric restricted to $[0.9T, T)$, evaluated for every configuration against the same f_n over all $n = 100$ clients, including the configuration in which client k never participates.

	Configuration	M_{tail}	acc (obs.)
A0	no late joiner, all 100 clients from round 0	0.6034 ± 0.1058	0.5804
A1	client k joins at $\tau_k = 0.3T$	0.6074 ± 0.1116	0.5811
A2	client k joins at $\tau_k = 0.6T$	0.6089 ± 0.1119	0.5807
A3	client k never joins, f_n measured regardless	0.5609 ± 0.1238	0.5813

The 3 configurations the claim concerns agree. Their pooled across-seed standard deviation is 0.1098 and the three pairwise differences, computed before rounding and so not exactly repro-

ducible from the table, are 0.0040, 0.0054 and 0.0014, between 20 and 80 times smaller, so the criterion fixed in advance (P10a, Table 4) is met.

What that permits must be stated exactly. Theorem 2 bounds $\mathcal{M}(\tau_k, T)$ from above, and a bound is not a prediction of the measured value, so the agreement establishes the empirical content of the claim, that no persistent penalty is detectable, and not that the floors themselves have been measured. Appendix G draws the same distinction for the criterion.

The fourth row fixes the resolution of the test, which is why we report it. Withholding client k for the whole run is the largest perturbation a single client's participation history can produce, and it moves M_{tail} by 0.042, itself inside one standard deviation. At $n = 100$ the arriving client enters f_n with weight $1/n$ and the criterion is quadratic in the gradient, so no single-client effect of any kind is resolvable at this network size; a design able to measure the *size* of a persistent penalty rather than its absence would have to perturb a constant fraction of the network. We read the agreement of A0, A1 and A2 accordingly, and Appendix H records it.

The transient is not resolvable either. Writing $g(w) = M_w(\text{A1}) - M_w(\text{A0})$ for successive 20-round windows,

w	$[60, 80)$	$[80, 100)$	$[100, 120)$	$[120, 140)$	$[140, 160)$	$[160, 180)$	$[180, 200)$
$g(w)$	$+0.0101$	$+0.0015$	$+0.0023$	$+0.0011$	$+0.0048$	-0.0085	$+0.0040$
$\lvert g(w)\rvert$/s.d.	0.06	0.01	0.02	0.01	0.04	0.07	0.04

the registered monotonicity (P10b) fails, since g rises again over the last two windows. It fails because there is nothing to decay. The magnitude is already 0.06 standard deviations in the window immediately following the join, so the sequence measures seed noise from its first entry onward, and a quantity below the detection threshold at the start cannot be observed to decrease. The empirical content of P10b, that the transient is invisible at the end of the measurement window, holds in every window rather than only at the end. This is the same $1/n$ suppression that makes the initialization rule macroscopically invisible in Section 5.3, seen through a different quantity; Appendix K.3 measures its size directly.

K.2 The join shock in the consensus trajectory

Table 1 reports that the initialization rule leaves accuracy and the certified metric unchanged, which is what the factor $\frac{n-1}{n^2}$ in Equation (3) predicts for a macroscopic quantity at these network sizes. The consensus error is not a macroscopic quantity but the one the identity governs directly, and the effect of the rule is clearly visible there.

Three readings of Figure 2, in the order the analysis produces them. First, the shock is real and large in this quantity. Cold initialization raises Ω^{τ_k} by 47 to 53% over its pre-join value in every seed, and inverting Equation (3) gives an implied displacement of $D_k = 0.79$ to 0.83 across the 5 seeds, which is the scale of a freshly initialized model rather than of the incumbent spread.

Second, warm initialization does not merely avoid the shock. Corollary 4 states that a join at the incumbent average leaves $\Omega^{\tau_k} = \frac{n-1}{n}\Omega^{\tau_k -}$, a strict decrease, and the measured values match that prediction to within 0.8% in every seed.

Seed	42	43	44	45	46
$\Omega^{\tau_k -}$ (round $\tau_k - 1$)	0.01301	0.01278	0.01269	0.01274	0.01238
warm, measured Ω^{τ_k}	0.01292	0.01258	0.01262	0.01258	0.01217
warm, predicted $\frac{n-1}{n}\Omega^{\tau_k -}$	0.01288	0.01265	0.01257	0.01261	0.01226
cold, measured Ω^{τ_k}	0.01908	0.01917	0.01941	0.01893	0.01871
cold, implied D_k	0.79	0.81	0.83	0.80	0.81

Third, the excess decays. Measured against the warm configuration round by round it falls from 52% at the join to 20, 9.0, 4.3, 2.2, 1.1, 0.5 and 0.2% over the following rounds, so it is below the seed spread within about 8 rounds. That is a per-round factor near 0.47, comfortably faster than the $\rho = (3 + \lambda_n^2)/4 = 0.94$ that Lemma 14 requires at $\hat{\lambda}_n = 0.88$. The lemma bounds the contraction

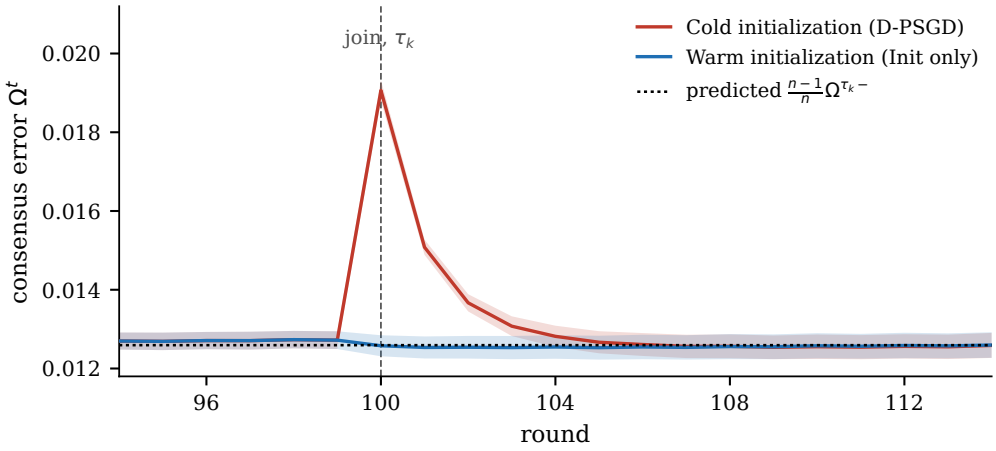


Figure 2: Consensus error around a join at $\tau_k = 0.5T$ (CIFAR-10, $n = 100$, band one standard deviation over 5 seeds). The two configurations differ in initialization alone and are identical before τ_k, since the design is paired. Cold initialization injects the join shock of Equation (3); warm initialization removes it and leaves the strict decrease Corollary 4 predicts.

rather than predicting it, and a displacement concentrated on a single client is not aligned with the slowest mixing mode, so decay faster than ρ is expected rather than anomalous.

Taken together this is the positive form of the initialization result, complementing the identity check of Section 5.1, which certifies the algebra but not that the quantity it governs behaves as described. What it does not establish, being a single network size, is the n^{-1} scaling, which Appendix K.3 measures.

K.3 Network-size sweep

Appendix D.2 separates two rates. The network average moves by D_k/n, so its squared displacement is $\Theta(n^{-2})$, while the join shock added to the consensus error is $\Theta(n^{-1})$ at fixed D_k, and only the second enters Equation (4). This section measures the second.

Protocol and its consequences for comparability. CIFAR-10 with LeNet, $n \in \{20, 50, 100, 200\}$, $\tau_k = 0.5T$, 3 seeds, on the sparse graph, with the initialization arms paired at matched seeds. Two departures from Appendix J.1 are needed to isolate n and are declared here rather than left to be inferred. Each client holds a fixed shard of 250 examples, 200 for training and 50 for validation, drawn from a Dirichlet allocation with $\alpha = 0.4$ resampled at each n; the total data therefore grows with n and equals CIFAR-10 exactly at $n = 200$. With batch size 64 and a dropped final partial batch, every n runs exactly three minibatches per round, which holds the per-round work fixed. Under the protocol of Appendix J.1, at a fixed total dataset, one local epoch would be 10 times longer at $n = 20$ than at $n = 200$ and would dominate everything measured here. The cost of the change is that absolute values are not comparable across n, so every claim below is either a paired comparison within one n or a ratio.

The shock at the join instant.

n	Ω^{τ_k-}	D_k	Ω^{τ_k} (cold)	E_n	$R_n = D_k^2/\Omega^{\tau_k-}$	identity residual
20	0.0079	1.075	0.0624	6.937	147.1	3.5×10^{-16}
50	0.0085	1.074	0.0309	2.653	136.4	3.5×10^{-16}
100	0.0092	1.060	0.0202	1.198	122.0	3.5×10^{-16}
200	0.0097	1.066	0.0153	0.577	117.0	3.5×10^{-16}

Here $E_n := \Omega^{\tau_k}/\Omega^{\tau_k-} - 1$ is the relative jump under cold initialization. Equation (3) gives $E_n = -\frac{1}{n} + \frac{n-1}{n^2}R_n$ identically, so E_n is fixed once D_k and Ω^{τ_k-} are measured and no fit to it can by

itself be evidence about anything; the residual column confirms the algebra at the order Section 5.1 reports and nothing more. The empirical content is whether R_n is stable in n, and it is. It varies by a factor of 1.26 over a 10-fold range, and nE_n is correspondingly flat to a factor of 1.20, at 138.7, 132.7, 119.8 and 115.4. The regression of $\log E_n$ on $\log n$ gives a slope of -1.085, inside the interval $[-1.3, -0.7]$ fixed in advance, and the precondition attached to that interval, that R_n vary by less than a factor of two, is met. This is P11 of Table 4.

The residual departure from -1 has a measured cause rather than being error. D_k is constant across the sweep, from 1.060 to 1.075, as a fresh initialization at a fixed scale should be. What moves is Ω^{τ_k-}, which rises from 0.0079 to 0.0097 because $\hat{\lambda}_n$ rises from 0.815 to 0.882. The degree is held at 6 while n grows, so the graph mixes less well and the incumbents' own consensus error is larger. The same quantity accounts for part of the step-size behavior below, and a topology whose spectral gap does not move with n would remove it.

The two rates are separated by the data and not by the algebra alone. Normalizing the $n = 100$ measurement forward, an n^{-1} law predicts $E_{20} = 5.99$ and an n^{-2} law predicts 29.9, against a measured 6.94. The two laws agree at $n = 100$, which before this sweep was the only size we had run.

Quantities the shock does not affect. The table below reports the paired difference between warm and cold initialization, per n, against the across-seed standard deviation at that n.

n	ΔM	$\lvert\Delta M\rvert$/s.d.	Δacc	$\lvert\Delta\text{acc}\rvert$/s.d.
20	-0.0661	0.13	-0.0030	0.88
50	$+0.0366$	0.16	$+0.0005$	0.06
100	$+0.0046$	0.10	$+0.0009$	0.17
200	$+0.0039$	0.03	$+0.0010$	0.11

The registered criterion (P12) was that this ratio rise monotonically as n falls and exceed one at $n = 20$. It does neither, and the criterion is not met. The shock at $n = 20$ is nonetheless real and large, since cold initialization multiplies the consensus error by 7.94 at the join, against 1.58 at $n = 200$. It does not reach the certified metric at any size we run.

The two statements are reconciled by the route the shock has to travel. Equation (3) governs the dispersion of the clients about their average, whereas Equation (2) is evaluated at that average, and the excess dispersion is removed by mixing within about 8 rounds at these spectral gaps (Appendix K.2). A window average over the 100 rounds following the join therefore carries what remains of an eight-round transient, and the factor $\frac{n-1}{n^2}$ suppresses it again. Both act at every network size we run, which is why the crossing P12 anticipates does not occur at $n = 20$.

The reading is therefore the one P4 and P5 state, extended and given a mechanism. The initialization rule is an intervention on the consensus layer, exact where the identity acts, whose effect on the average iterate is suppressed both by n and by the 8-round half-life of the excess it removes (Appendix K.2). In the certified metric the invisibility holds at every size in the sweep, the largest normalized difference being 0.16 at $n = 50$. In accuracy it holds at $n = 50$, 100 and 200, at 0.06, 0.17 and 0.11, while at $n = 20$ the paired difference reaches 0.88 across-seed standard deviations, the largest normalized effect in this group and short of the registered threshold rather than comfortably inside it. Two things keep us from reading that as the crossing P12 anticipates. It is a cost and not a gain, since warm initialization lowers the certified metric and the accuracy together at $n = 20$ whereas removing the shock would favor the warm arm on both, so it is Convention 2 rather than the shock becoming visible. In addition, the accuracy column is not monotone in n, so at 3 seeds it is one point and not a trend. We report it and draw no conclusion from it.

The step size across network sizes.

n	$\hat{G}_n$ (fallback)	$\hat{G}_n$ (gate forced)	$\hat{\eta}$ (fallback)	$\hat{\eta}$ (forced)	$\hat{L}$	$\hat{\sigma}^2$	$\hat{\lambda}_n$	cap
20	2.272	2.167	0.00469	0.00469	5.70	21.5	0.815	active
50	2.107	1.765	0.00324	0.00324	6.68	23.0	0.854	active
100	2.816	1.863	0.00217	0.00217	8.22	24.7	0.879	active
200	2.990	1.853	0.00205	0.00205	8.31	26.2	0.882	active

Proposition 7 states that once Phase 1 has plateaued the prescribed step size is free of n. The criterion registered for it (P14) was that $\hat{\eta}$ vary across n by less than the across-seed standard deviation with the plateau gate forced. It varies by 0.00263 against a spread of 0.00057, but the criterion cannot be evaluated here. The table indicates why, and the reason is not Proposition 7.

The cap is active at every network size, so $\hat{\eta} = \eta_c = (1-\hat{\lambda}_n)/(7\hat{L})$ throughout and $\hat{G}_n$ never enters the adopted step at all. That is why the fallback and forced-gate columns are identical to the last digit while their $\hat{G}_n$ columns are not, since both are absorbed by the minimum in Equation (6). The square-root branch, which is the only branch on which the cancellation of n has content, is never taken here. The claim is therefore untested rather than refuted, and the sparse graph cannot test it.

The n-dependence that does appear factors exactly into two quantities outside Proposition 7. From $n = 20$ to $n = 200$ the adopted step falls by a factor of 2.29, and $(1 - \hat{\lambda}_n)$ contributes 1.57 and $\hat{L}$ contributes 1.46, and $1.57 \times 1.46 = 2.29$. The first is the fixed-degree confound already identified above, since holding the degree at 6 while n grows lowers the spectral gap. The second is a property of our estimator rather than of the objective. $\hat{L}$ is accumulated as a maximum over clients, so it is an order statistic that grows with the number of clients drawn, which is the estimator defect that Appendix H records and Appendix I quantifies. A windowed estimator would test both readings, and would also be the only way to move this configuration off the cap; Appendix K.9 shows that halving the conservatism constant does not.

K.4 Full step-size grids

Two grids answer different questions and are kept apart. The first is the mechanism grid behind Figure 1, sweeping $\eta = \eta_c \cdot 2^{-j}$ for $j = 0, \ldots, 6$ and adding the calibrated configuration as a reference point.

	Sparse graph, $\hat{\lambda}_n = 0.88$			Complete graph, $\hat{\lambda}_n = 0.50$		
η/η_c	η	M	acc (obs.)	η	M	acc (obs.)
$1/64$	0.000030	0.287	0.375	0.000142	0.245	0.408
$1/32$	0.000060	0.250	0.392	0.000284	**0.240**	0.448
$1/16$	0.000119	0.233	0.418	0.000567	0.259	0.493
$1/8$	0.000238	**0.224**	0.454	0.001134	0.392	0.526
$1/4$	0.000477	0.240	0.505	0.002269	0.695	0.547
$1/2$	0.000954	0.340	0.544	0.004537	0.753	0.564
1	0.001907	0.591	0.578	0.009074	0.547	**0.584**
calibrated $\hat{\eta}$	0.001907	0.591	0.578	0.003931	0.791	0.559

The arithmetic behind Section 5.2 reads off this table. On the sparse graph $\hat{\eta}$ coincides with η_c exactly, which is the capped branch, and the ratio to the grid minimizer is $c = 8.0$, so Lemma 8 permits an inflation of $(c + 1/c)/2 = 4.06$ against a measured 2.64. On the complete graph $c = 13.86$ permits 6.97 against a measured 3.29. Both ratios are computed from the unrounded step sizes and metric values rather than from the rounded entries above. The sparse row at $\eta/\eta_c = 1$ and the calibrated row are the same run reported twice, since $\hat{\eta} = \eta_c$ there, and are not a duplicated entry. Two features of the table matter beyond those two numbers. Accuracy increases monotonically in η across the whole grid on both graphs while M has an interior minimum, so the two criteria disagree about the best step size by three grid points on the sparse graph and five on the complete one, a factor of 8 and 32 in the step size, which is Convention 2 made visible. That monotonicity also places the calibrated step on the complete graph, at $\eta = 0.003931$, between the $\eta_c/4$ and $\eta_c/2$ rows, and its accuracy of 0.559 sits where log-interpolation between them puts it. And M falls back at $\eta = \eta_c$ on the complete graph, from 0.753 at $\eta_c/2$ to 0.547, which Equation (4) does not predict, since every term of the bound is monotone in η on this range. We have no explanation for it and record it as a departure.

The second grid is the per-dataset step-size selection that gives the ablation of Table 1 a tuned floor, extended upward to locate the constant that minimizes the certified metric. In total, 6 constants were run for the fixed-step configurations, at 3 seeds, under the augmented setting of Appendix J.1.

	CIFAR-100		Tiny-ImageNet	
η	acc (obs.)	M	acc (obs.)	M
0.001	0.317	14.57	0.182	3.56
0.003	0.430	10.52	0.341	2.89
0.010	**0.487**	4.53	0.472	1.78
0.030	0.482	2.34	**0.483**	1.54
0.100	0.475	0.65	0.430	0.76
0.300	0.404	**0.26**	0.352	**0.30**

Selection was made on mean final accuracy, giving 0.01 on CIFAR-100 and 0.03 on Tiny-ImageNet. The extension moves neither, so Table 1 is unaffected by it. That choice is the one least favorable to the step-size rule on accuracy and the one most favorable to it on the certified metric, and the two criteria disagree on CIFAR-100, where 0.03 attains $M = 2.34$ against 4.53 at the selected 0.01. The improvement in M that Section 5.3 reports on CIFAR-100, from 4.57 to 3.87, is measured against an accuracy-tuned floor and does not survive against a larger constant. The extension locates that constant.

The extension ran under a stopping rule fixed in advance: stop when M rises at two consecutive points, when accuracy falls below the no-collaboration reference of Table 1, or on divergence. None of the three fired. M falls monotonically to the largest constant tried on both datasets, so no interior minimum is exhibited and the criterion registered as P13 in Table 4 is not met.

Where that minimum should have been is what the failure locates. Theorem 2 holds for $\eta \leq \eta_{\mathrm{c}} = (1 - \lambda_n)/(7L)$, which at the measured $\hat{L}$ of 4.83 and 1.99 (Appendix J.1) and $\hat{\lambda}_n \approx 0.85$ at $n = 50$ (Appendix K.3) is about 0.0043 on CIFAR-100 and 0.0105 on Tiny-ImageNet. The constants that decide anything here lie above it. The accuracy-selected floors exceed it by factors of 2.3 and 2.9 and the M-minimizing $\eta = 0.3$ by about 70 and 30, while the two smallest constants on CIFAR-100 and the three smallest on Tiny-ImageNet are admissible and are also the worst on M. The grid therefore locates a minimum of M outside the range in which Equation (4) asserts anything, which is also why the interior minimum the bound does predict appears on the CIFAR-10 grid above and not here, since that sweep is anchored to η_{c} and reaches it, at $\eta_{\mathrm{c}}/8$ and $\eta_{\mathrm{c}}/32$. The calibrated $\hat{\eta}$, at 0.00253 and 0.00254, is admissible on both datasets by construction, since Equation (6) takes a minimum with η_{c}; the constants that beat it on M are not.

Two further facts belong with the constant that minimizes M. It costs 0.083 in accuracy on CIFAR-100 and 0.131 on Tiny-ImageNet against the accuracy-selected floors, which is Convention 2 at its widest. The two criteria are optimized by constants two orders of magnitude apart, and the trade-off between them is 8 to 13 accuracy points. Moreover, it is a constant applied from round 0, whereas every row of Table 1 shares one Phase-1 step size and differs only from τ_k onward, which is the reading the caption of Table 1 forbids; it is therefore not a floor those rows can be paired against and we draw no paired comparison to it. As an unpaired observation, the certified metric Table 1 reports for the two rules on Tiny-ImageNet, 0.25, is below every constant in this grid including 0.30 at $\eta = 0.3$, while on CIFAR-100 the 3.87 of Table 1 is not.

K.5 Full join-time sweeps

Both branches are swept, with $\tau_k \in \{0.2, \ldots, 0.9\}T$ at 3 seeds (Figure 3). The mechanism traces are included because they are what explains the sub-plateau region, and the per-seed curves are given because the prediction is about individual seeds rather than about their mean.

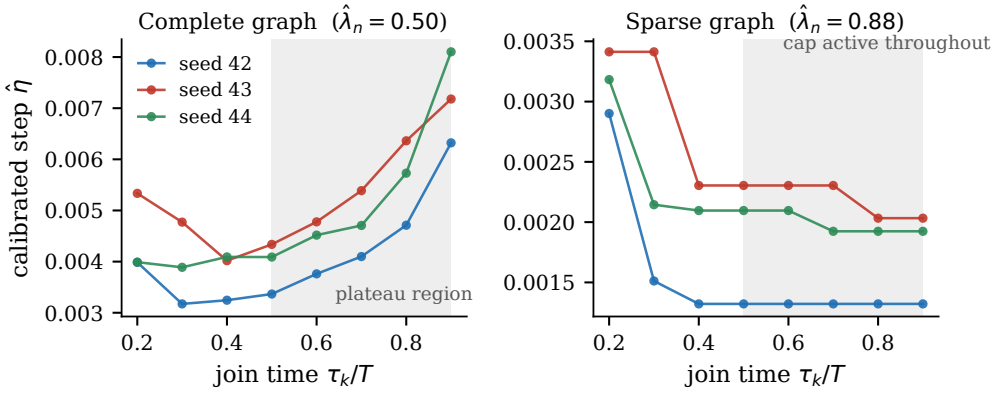


Figure 3: Calibrated step size against join time, one line per seed. Shading marks the region $\tau_k \geq 0.5T$ on which Appendix K.5 finds the plateau precondition holding. On the complete graph $\hat{\eta}$ rises there in every seed and falls below it. On the sparse graph $\hat{\eta}$ equals η_c throughout, and the decline down the range is η_c itself moving as $\hat{L}$ ratchets upward.

	Sparse graph (capped)			Complete graph (off cap)		
τ_k/T	$\hat{\eta}$	$\hat{\Delta}_k$	M	$\hat{\eta}$	$\hat{\Delta}_k$	M
0.2	0.00317	1.62	1.368	0.00444	1.67	1.078
0.3	0.00236	1.51	1.012	0.00394	1.53	1.010
0.4	0.00191	1.41	0.739	0.00379	1.42	0.895
0.5	0.00191	1.33	0.591	0.00393	1.39	0.791
0.6	0.00191	1.36	0.484	0.00435	1.37	0.715
0.7	0.00185	1.35	0.399	0.00473	1.26	0.646
0.8	0.00176	1.30	0.328	0.00560	1.25	0.593
0.9	0.00176	1.24	0.281	0.00720	1.23	0.545

On the complete graph $\hat{\eta}$ increases strictly over $\tau_k \geq 0.5T$, from 0.00393 to 0.00720, and does so in every seed rather than only in the mean. Below $0.5T$ it decreases, and the traces say why. The measured misfit is still falling there, from 1.67 at $0.2T$ to 1.39 at $0.5T$, so the network has not plateaued and the precondition of Theorem 6 does not hold. The max-accumulating estimators of L and σ^2 are also still rising over the same region, and the two together outweigh the $(T')^{-1/2}$ factor. The turning point coincides with the point at which $\hat{\Delta}_k$ flattens, which is what makes this sweep a measurement of where the precondition begins rather than a partial confirmation of the prediction.

On the sparse graph $\hat{\eta}$ equals η_c at every join time, so the square-root branch never runs and no monotonicity in τ_k is expressible. The slow decline from 0.00317 to 0.00176 down the column is η_c itself moving, since $\hat{L}$ ratchets upward over a longer Phase 1 and $\eta_c = (1 - \hat{\lambda}_n)/(7\hat{L})$ falls with it. That is the capped branch of Theorem 6 behaving as stated and not a failure of the monotonicity prediction, which the capped branch does not make.

K.6 Topology and the spectral gap

Five topologies at a fixed join time $\tau_k = 0.5T$, with $\hat{\lambda}_n$ from exact eigendecomposition of the active mixing matrix.

One row needs reading with care, and it is the row the section turns on. The ring is the only topology here whose graph changes in kind at the join. Withholding one vertex from a 100-node cycle leaves a 99-node path, so Phase 1 runs on a path and Phase 2 on a cycle, whereas every other topology loses a vertex and remains a graph of the same kind. Both are connected, and what the arrival changes is the vertex connectivity rather than the connectedness, closing the path into a cycle. The two mix differently enough to matter. Under lazy Metropolis weights the second eigenvalue is 0.9998321808 for the path and 0.9993422428 for the cycle, and the column reports the latter, since (A5) and Equation (6) both take λ_n over the post-join rounds and $\hat{\eta}$ is computed from it. Even that

value leaves $1 - \hat{\lambda}_n = 6.6 \times 10^{-4}$, smaller by a factor of about 90 than on the next topology in the table, which is why this row sits at what is effectively no spectral gap.

Topology	$\hat{\lambda}_n$	$\hat{\eta}$	M	acc (obs.)
complete	0.500	0.00393	0.791	0.559
random, 10-regular	0.806	0.00352	0.797	0.553
random, 6-regular	0.882	0.00191	0.591	**0.578**
random, 4-regular	0.940	0.00088	0.315	0.522
ring	0.99934	0.000011	1.026	0.388

The calibrated step falls monotonically as $\hat{\lambda}_n$ rises, which is $\eta_c = (1 - \hat{\lambda}_n)/(7\hat{L})$ collapsing. Observational accuracy does not follow it down. It is highest on the 6-regular graph, 2 to 3 points above the two best-connected topologies and nearly 6 above the 4-regular one, and falls away only below that. Two effects run against each other here and we have isolated neither. A larger $\hat{\lambda}_n$ shrinks the step, which costs accuracy; it also mixes less, which raises an accuracy measured against a validation split drawn from each client's own label distribution (Appendix K.10). Appendix K.4 exhibits the second at a matched step, where at $\eta = 0.0019$ the sparse graph reaches 0.578 while the complete graph interpolates to about 0.542. Below the 6-regular graph the collapsing step dominates and accuracy falls steeply. We report the shape and claim no account of it. The certified metric follows neither, and the ring is the reason it does not. Its step size is smaller than the complete graph's by between 277 and 434 across seeds, 350 as the mean of the per-seed ratios, yet its M is the largest of any topology in the table. The bound accounts for it through the transient rather than through the floor. The ring is on the cap, and at $\eta = \eta_c = \frac{1-\lambda_n}{7L}$ term (iv) evaluates to $\frac{96}{49}(\sigma^2 + \zeta_n^2)$, in which the factor $(1 - \lambda_n)^{-2}$ is cancelled exactly by η_c^2, so the topology floor takes the same value on the ring as on every other capped topology in the table and cannot be what separates them. Term (i) at the cap is instead $\frac{28LG\tau_k}{(1-\lambda_n)T'}$, which diverges as $\lambda_n \to 1$, and $1 - \hat{\lambda}_n$ is 6.6×10^{-4} here against 0.118 on the 6-regular graph. We do not convert that into a ratio between the two rows, since $\hat{L}$ and $\hat{G}$ enter the same term and are not held fixed across topologies. The mechanism the bound describes is therefore that a spectral gap this small forces a step so short that the inherited gap is not worked off within the window, and the ring's accuracy of 0.388, the lowest in the table, is consistent with a network that has barely moved from the model it started the window with. The bound is loose by orders of magnitude in this regime and we claim the direction rather than the value.

K.7 Schedule comparison

Four post-join schedules, anchored against the grid-optimal constant of Appendix K.4 rather than against the calibrated constant, since Theorem 6 compares schedules to the best constant and not to any particular one.

	Sparse graph		Complete graph	
Schedule	M	acc (obs.)	M	acc (obs.)
constant at $\hat{\eta}$	0.591	0.578	0.791	0.559
square-root decay	0.591	0.578	0.657	0.575
warmup	0.571	0.577	0.767	0.558
cosine	0.338	0.399	0.515	0.417
grid-optimal constant	**0.224**	0.454	**0.240**	0.448

No schedule reaches the grid-optimal constant on either graph, the closest being cosine at 0.338 against 0.224 and at 0.515 against 0.240. That is the comparison Theorem 6 makes, and the prediction is confirmed. The improvement cosine and square-root decay show over the *calibrated* constant is a different statement and not a counterexample, since $\hat{\eta}$ sits well above the grid minimizer on both graphs and any schedule that spends part of the window at a smaller step recovers some of that gap. Cosine incurs a substantial accuracy cost for this, 0.399 against 0.578 on the sparse graph. The decay is implemented as a rescaling that starts at η_c and decays toward $\hat{\eta}$, and on the sparse graph $\hat{\eta} = \eta_c$, so the two configurations run the same step size at every round and their sparse entries coincide.

K.8 MISESTIMATION AND ESTIMATOR PERTURBATION

The gap estimate was perturbed multiplicatively by κ and the resulting M compared against the unperturbed run. Lemma 8 bounds the ratio by $(\sqrt{\kappa} + 1/\sqrt{\kappa})/2$.

	Sparse graph		Complete graph		
κ	$\hat{\eta}$	$M/M_{\kappa=1}$	$\hat{\eta}$	$M/M_{\kappa=1}$	envelope
$1/16$	0.00092	0.56	0.00098	0.44	2.13
$1/4$	0.00183	0.97	0.00197	0.79	1.25
1	0.00191	1.00	0.00393	1.00	1.00
4	0.00191	1.00	0.00751	0.78	1.25
16	0.00191	1.00	0.00907	0.69	2.13

Every measured ratio is inside the envelope, and every ratio is also below 1, meaning that misestimating the gap in either direction improved the certified metric rather than degrading it. That follows from the grid of Appendix K.4, where $\hat{\eta}$ already sits above the minimizer of M, so perturbations that move the step away from $\hat{\eta}$ move it toward a better point on one side and onto the anomalous fall-back on the other. The envelope is therefore satisfied without being stressed, and this group confirms that Lemma 8 is not violated rather than confirming that it is tight.

On the sparse graph the entries for $\kappa \in \{1, 4, 16\}$ are numerically identical. That is the one-sided protection of the minimum in Equation (6) working as designed, since over-estimates of the gap are absorbed by the cap while under-estimates pass through, and it is why $\kappa < 1$ is the direction that moves anything there.

The estimator perturbations act on the constants rather than on the gap and are not protected by Lemma 8, since η_c is a hypothesis of Lemma 13 and Lemma 14 rather than a term in the trade-off.

Perturbation	Direction	$\hat{\eta}$	acc (obs.)	Outcome
$c_L = 1$ (halves $\hat{L}$)	less conservative	0.00382	0.556	stable
$c_L = 4$ (doubles $\hat{L}$)	more conservative	0.00095	0.552	stable, constant cost
$\hat{\lambda}_n = 0.99$	safe	0.00016	0.433	stable, $M = 0.229$, large accuracy cost
$\hat{\lambda}_n = 0.5$	*unsafe*	0.00366	0.555	did not destabilize

The last row is the informative one and it did not go the way the theory would license us to expect. Setting $\hat{\lambda}_n = 0.5$ on a graph whose true value is 0.88 inflates η_c and violates the premise of both lemmas outright, and training did not diverge. It was not free either. Accuracy falls from 0.578 at the unperturbed step to 0.555, a cost of 0.023, and $c_L = 1$ costs 0.022, so the two rows that push $\hat{\eta}$ above η_c are also the two that lose accuracy. We read the absence of divergence as slack in the constant $7L$, consistent with the margin $48 \leq 49(1 + \lambda_n)^2$ recorded in Appendix D.7, which is loosest at small λ_n. We report both facts and draw no conclusion about how far the premise can be violated.

K.9 CAP-ACTIVITY DIAGNOSTICS

Three controls were varied to locate where the step-size rule carries an accuracy cost and where it does not. Cap activity is the number of seeds out of three in which the minimum in Equation (6) was active.

Graph	τ_k	$\hat{\eta}$	cap active	Δacc	ΔM
sparse	$0.5T$	0.00191	3/3	-0.041	$+0.043$
sparse, $c_L = 1$	$0.5T$	0.00381	3/3	-0.027	$+0.237$
complete	$0.5T$	0.00393	0/3	-0.027	$+0.266$
sparse	$0.9T$	0.00176	3/3	-0.012	-0.273
complete	$0.9T$	0.00720	1/3	$+0.000$	$+0.009$

Differences are calibrated minus fixed-step at the same graph and join time. Two readings follow, and only one of them is a limitation of the rule.

The accuracy cost is largest at the default operating point, which is the first row and the configuration Section 5.3 reports, where it is -0.041. Three groups estimate that quantity independently and agree: -0.040 at 5 seeds in Table 1, -0.040 in the $\alpha = 0.4$ row of Appendix K.10, and -0.041 here at three.

It shrinks under either control this group varies. Holding the join time at $0.5T$ and widening η_c, by halving the conservatism constant or by moving to the complete graph, takes the cost from -0.041 to -0.027; holding the graph and moving the join later takes it from -0.041 to -0.012 on the sparse graph and from -0.027 to $+0.000$ on the complete one. The cost therefore tracks how far below the fixed step the prescribed one lands, and on the sparse graph the cap is what decides that, so it is a cap effect in that sense even though it is not cap saturation. One point does not fit. The sparse graph at $0.9T$ has the smallest $\hat{\eta}$ in the group and only -0.012, and there the post-join window is a fifth of the run, so the smaller step has correspondingly less of it to act over. The mechanism is the criterion mismatch Appendix K.4 exhibits directly, where accuracy increases monotonically in η while M has an interior minimum, so a rule targeting M gives up accuracy in proportion to how far below the accuracy optimum it lands. Convention 2 states that the two criteria need not agree; this group measures how much they disagree at our operating points.

The certified metric moves the other way in most of this group. ΔM is positive in four of the five rows, and the largest of them, $+0.266$, is on the complete graph at $0.5T$, where the cap is inactive and the square-root branch is the one running. That row coincides with the anomaly Appendix K.4 records on the same graph, where M rises from $\eta_c/32$ to $\eta_c/2$ and falls back at η_c, and the calibrated $\hat{\eta} = 0.00393$ sits inside that rise while the fixed step 0.01 sits at the fall-back. We record the coincidence and do not offer it as an explanation, having established no mechanism for the anomaly itself.

What is a limitation is cap saturation in the certified metric. It is structural on the sparse graph rather than a matter of estimator tuning. Halving the conservatism constant doubles η_c and still leaves all 3 seeds capped, so no plausible retuning of $\hat{L}$ moves that configuration onto the square-root branch. On the capped branch $\hat{\eta}$ is determined by the graph and by $\hat{L}$ rather than by the misfit, and Equation (6) no longer expresses the trade it was derived to express, which is why ΔM is positive at $0.5T$ on the sparse graph and the rule yields no improvement there.

K.10 Heterogeneity sweep

Dirichlet concentration $\alpha \in \{0.1, 0.4, 1.0\}$, 5 seeds, with the 0.4 row taken from the default configuration.

α	Calib. acc / M	Fixed acc / M	Δacc	ΔM
0.1	0.694 / 2.568	0.708 / 2.406	-0.014	$+0.162$
0.4	0.541 / 0.636	0.581 / 0.558	-0.040	$+0.077$
1.0	0.470 / 0.183	0.505 / 0.240	-0.034	-0.058

The accuracy cost does not change direction across an order of magnitude in α, ranging from 0.014 to 0.040 with no monotone trend in the concentration. The certified metric reverses, from $+0.162$ at $\alpha = 0.1$ to -0.058 at $\alpha = 1.0$, so the sign of ΔM on the capped branch is not itself stable in α. All three rows are on the sparse graph, where the cap is active, so this sweep measures the capped branch and is uninformative about the square-root branch. Two caveats. First, this group carries no external comparison, since the baselines were not run at these concentrations. Second, accuracy is higher at $\alpha = 0.1$ for every configuration, which is an artifact of evaluating each client against a validation split drawn from its own label distribution, so absolute accuracies are not comparable across rows of this table.

K.11 Staggered multi-client joins

The scenario replaces the single controlled join with 20 clients arriving at rounds drawn uniformly from $(0.25T, T]$, on CIFAR-10 and CIFAR-100 at 5 seeds. The no-collaboration reference of Table 1 is not run here, so the external comparison is against the two collaborative baselines alone. It is the scenario closest to a deployment, and it is where a conclusion of Section 5 reverses. This section reports the reversal and then attributes it.

	CIFAR-10		CIFAR-100	
Method	acc (obs.)	M	acc (obs.)	M
Init + calib.	0.520	0.990	0.4392	5.822
Calib. only	0.518	0.795	0.4393	6.684
Init only	**0.581**	0.627	**0.4870**	5.149
D-PSGD	0.574	0.577	0.4782	5.171
Epidemic Learning	0.462	0.418	0.4230	4.540
DFedAvg	0.400	**0.370**	0.3940	**4.180**

Two things hold. The join identities hold exactly, including at the rounds where two clients arrive together, which is what Proposition 12 covers and Section 5.1 reports. Furthermore, every configuration of the two rules still exceeds both external baselines in observational accuracy, by 0.016 at the narrowest against Epidemic Learning.

What reverses is the step-size rule. Against the initialization-only configuration it improved M by 0.70 on CIFAR-100 under a single controlled join; here it degrades it by 0.673 on CIFAR-100 and by 0.363 on CIFAR-10, and it costs 0.048 and 0.061 in observational accuracy on the two, against 0.009 and 0.040 under the controlled join. Both criteria move against the rule, on both datasets, and by more than the single controlled join would suggest.

Attributing the reversal. Two features of the scenario are candidate explanations. The first arrival falls at round 55 on average, which is $0.28T$, inside the region where Appendix K.5 finds the plateau precondition failing and the measured misfit still descending; $\hat{\Delta}_k$ is correspondingly larger here than under the controlled join, 1.33 against 0.76. Separately, the window over which M is measured begins at the first arrival and therefore contains 19 further arrivals, each injecting its own shock, whereas Theorem 2 bounds a window whose membership is fixed after τ_k. The two are separated by varying them independently: the arrival window, by drawing the 20 rounds from $(0.5T, 0.8T]$ instead, which places every arrival inside the plateau region and moves the first from round 55 to round 102; and the measurement window, by starting M at the last arrival instead of the first, which is a re-analysis of the same runs. Entries are ΔM between Init + calib. and Init only, negative where the rule improves the certified metric.

Arrivals drawn from	M from first arrival	M from last arrival
$(0.25T, T]$, first at round 55	$+0.673$ $(1/5$ seeds negative$)$	$+3.733$ $(0/5)$
$(0.5T, 0.8T]$, first at round 102	$\mathbf{-0.879}$ (**5/5**, 3.8 s.d.)	-0.001 $(3/5)$
single controlled join (Table 1)	-0.696 $(5/5)$	–

Moving the arrivals into the plateau region removes the reversal outright. The rule improves M by 0.879, in every seed and by 3.8 across-seed standard deviations, which is a larger improvement than the single controlled join it was derived for. Restricting the measurement window instead does not remove it. The precondition of Theorem 6, and not the composition of the measurement window, is therefore what governs whether the rule helps under staggered arrivals, and it governs it in the same direction as in the single-join setting of Section 5.3. This is the largest effect reported anywhere in this paper, and Appendix J.6 reads the null results elsewhere against it.

Three limits belong with that attribution. The cell at $+3.733$ is not a clean refutation of the second explanation. With the last arrival of the original schedule at round 182, the fixed-membership window is 18 rounds and carries about 4 evaluation points, immediately after a join, so the second explanation is unsupported on the data available rather than excluded, and its clean test is blocked by the geometry of the arrival window. The second draw is not a more realistic scenario than the first but a precondition-respecting one, and it changes two things at once, since arrivals compressed into $0.3T$ rather than $0.75T$ are also denser; that confound runs against the second explanation, which the compression should aggravate, and only partly with the first. Finally, the cell at -0.001 indicates that the gain is concentrated in the arrivals and the rounds following them rather than in the fixed-membership tail, which is what absorbing a join transient by the step size would produce.

Two further interpretations. The arms behind the second draw separate the rules more sharply than Table 1 does.

	arrivals in $(0.25T, T]$		arrivals in $(0.5T, 0.8T]$	
Method	acc (obs.)	M	acc (obs.)	M
Init + calib.	0.4392	5.822	0.4823	**4.112**
Calib. only	0.4393	6.684	0.4674	5.030
Init only	**0.4870**	5.149	**0.4899**	4.991
D-PSGD	0.4782	5.171	0.4820	5.042

First, the initialization rule is visible here, and it is the only place in this paper where it is. Between Init + calib. and Calib. only, which differ in initialization alone, M separates by 0.862 under the first draw and 0.918 under the second, against 0.07 for the same contrast on Tiny-ImageNet in Table 1. The channel is the one Theorem 3 identifies and Section 5.3 describes. Warm initialization lowers the measured misfit and Equation (6) prescribes a smaller step, $\hat{\eta} = 0.00270$ against 0.00462 under the second draw. What is new is the amplification, since the channel fires once per arrival and there are 20 of them. The rule remains macroscopically inert on its own, as Appendix K.3 measures across network size; it is not inert when composed with the calibration under repeated arrivals.

Second, the accuracy cost of the calibration is governed by the same precondition as the certified metric. It is 0.048 under the first draw and 0.008 under the second, alongside ΔM of $+0.673$ and -0.879. Two quantities that Convention 2 holds apart move together across the precondition boundary, which is stronger evidence for the attribution than either alone, and it also places the widest instance of the criterion split in this paper at the point where the precondition fails.

The cap is inactive in all 10 calibrated runs of this group and the plateau gate takes the fallback path throughout ($\hat{\epsilon}_1 = 0.036$), so the square-root branch of Theorem 6 is the one running here, in contrast to the sparse-graph configurations of Appendix K.3 where the cap binds at every network size.