How Communication Shapes Convergence and Generalization in Decentralized Training: A Feature Learning Perspective

Anonymous authors
Paper under double-blind review

Abstract

As data, model, and client populations continue to grow, decentralized learning has emerged as an important paradigm for training without a central server. Despite its practical relevance, existing theory has largely focused on optimization and consensus, with generalization typically studied through algorithmic stability. It remains unclear how decentralized communication shapes feature learning in neural networks and, in turn, affects the generalization of individual client models. To fill this gap, we analyze decentralized gradient descent for two-layer convolutional neural networks from a feature learning perspective. The derived dynamics of signal and noise coefficients characterize how communication propagates predictive signals and sample-specific noise across clients. The corresponding topology-aware quantities yield early generalization bounds and reveal client-placement effects that graph spectra and feature similarity alone cannot capture. Our analysis further establishes training convergence and client-level risk guarantees at a common fitted iterate. An extension to periodic local updates identifies a sufficient local-update budget under which the early generalization guarantee retains the well-mixed scale. Experiments on synthetic and real-world datasets under IID and non-IID settings support the predicted effects of communication topology, client placement, and communication frequency.

1 Introduction

Decentralized learning enables clients to train models collaboratively through local computation and peer-to-peer communication, without relying on a central server (Tsitsiklis et al., 1986; Nedić et al., 2018; Martínez Beltrán et al., 2023). This architecture is particularly attractive when centralized coordination is impractical or creates communication bottlenecks (Assran et al., 2019; Zantedeschi et al., 2020). In such systems, locally learned information propagates across clients through successive model exchanges over a communication graph. Communication topology can therefore influence not only how quickly clients reach consensus, but also what information each local model acquires during training (Xiao & Boyd, 2004; Vogels et al., 2021; Palmieri et al., 2024).

Yet existing theory on decentralized learning largely characterizes optimization error, consensus disagreement, or generalization through algorithmic stability (Koloskova et al., 2020b; Zhu et al., 2022; Le Bars et al., 2024). These algorithm-level guarantees do not explain how communication topology shapes learning predictive features and memorizing sample-specific noise in neural networks, or why different communication structures may produce different test behavior despite comparable training losses. This motivates a fundamental question:

How does communication topology interact with client heterogeneity to shape feature learning, and what convergence and generalization guarantees follow?

In this paper, we address this question by analyzing decentralized gradient descent for a two-layer convolutional neural network under a heterogeneous signal–noise data model. Building on feature learning theory (Cao et al., 2022; Kou et al., 2023; Huang et al., 2024), we explicitly track how predictive signal and sample-specific noise components evolve in each client's model parameters

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.

during training. Our analysis shows that, over a finite training horizon, each client's signal acquisition depends jointly on cross-client feature similarity, communication topology, and the placement of heterogeneous signals on the graph.

Specifically, our main contributions are summarized as follows:

1. Topology-aware feature learning. We derive coefficient dynamics and finite-time bounds for signal learning and noise memorization in individual client models. Our analysis identifies signal exposure $S_k(T)$ and noise accumulation $Q_k(T)$ for any client K after T iterations, whose ratio $S_k(T)^2/Q_k(T)$ enters the early generalization bound. This ratio captures signal-placement effects that the graph spectrum and aggregate signal similarity alone cannot determine.

2. Convergence and generalization guarantees. We establish convergence of the empirical loss and characterize test error at a common fitted iterate, giving sufficient conditions for small test error and identifying a weak-signal regime where successful fitting coexists with constant-order test error. For periodic local updates, we derive a sufficient local-update budget that preserves the well-mixed scale of the early generalization bound. A larger spectral gap permits longer local-update intervals within this guarantee, while the fitting bound incurs no additional interval-dependent factor.

3. Empirical validation. Experiments on synthetic and real-world datasets support the predicted relationships among topology-aware signal exposure, graph mixing, and client-level performance. They further illustrate that different client placements can induce different learning behavior even on the same communication graph.

2 Related Work

Decentralized optimization and communication topology. Decentralized optimization enables network-wide learning through local computation and peer-to-peer communication, with classical work establishing convergence guarantees for decentralized gradient and stochastic methods (Nedić & Ozdaglar, 2009; Yuan et al., 2016; Lian et al., 2017). Subsequent work addressed heterogeneous data, communication constraints, and time-varying topologies (Tang et al., 2018; Koloskova et al., 2019; 2020b; Kong et al., 2021). Beyond convergence, recent work shows that the spectral gap alone may not fully characterize topology-dependent training behavior, motivating measures that jointly account for graph structure and data heterogeneity (Vogels et al., 2022; Le Bars et al., 2023; Zhu et al., 2025). While these studies primarily focus on optimization, consensus, and communication, we study how graph communication shapes the predictive signal and sample-specific noise features learned by nonlinear neural networks.

Generalization theory of decentralized learning. Algorithmic stability provides a general framework for studying the generalization of gradient-based algorithms (Hardt et al., 2016) and has been extended to decentralized learning (Richards & Rebeschini, 2020; Sun et al., 2021; Zhu et al., 2022). Recent analyses show the role of topology can vary substantially: worst-case stability may be graph-independent, while refined bounds recover graph-dependent effects (Le Bars et al., 2024; Li et al., 2025). Other works study generalization or implicit regularization under structured assumptions, including separable data and overparameterized regression (Taheri & Thrampoulidis, 2023; Wu & Sun, 2024). In contrast, we analyze generalization through feature learning dynamics, explicitly tracking how predictive signal and sample-specific noise accumulate across the communication graph.

Feature learning in neural networks. Neural-network training has been studied through both kernel-based descriptions and feature learning dynamics (Jacot et al., 2018; Chizat et al., 2019; Yang & Hu, 2021; Allen-Zhu & Li, 2023; Jelassi & Li, 2022). Signal–noise decompositions characterize benign and harmful overfitting in two-layer CNNs (Cao et al., 2022; Kou et al., 2023). In federated learning, prior work establishes convergence guarantees for overparameterized neural networks (Huang et al., 2021; Deng et al., 2022), while feature learning analyses explain how aggregation and local updates shape feature acquisition and generalization under heterogeneous data (Huang et al., 2024; Bao et al., 2024). We build on this perspective to study decentralized learning, where multi-hop communication makes signal and noise accumulation depend jointly on graph topology, cross-client feature similarity, and client placement.

3 PRELIMINARIES

Notation. We use lower-case bold letters for vectors, upper-case bold letters for matrices, and non-bold letters for scalars. For a vector $\mathbf{v}$, we denote its Euclidean norm by $\|\mathbf{v}\|_2$. For a matrix $\mathbf{A}$, $\|\mathbf{A}\|_2$ denotes its spectral norm. For any positive integer n, we write $[n] = \{1, 2, \ldots, n\}$. We use asymptotic notation $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$, and use $\widetilde{O}(\cdot)$, $\widetilde{\Omega}(\cdot)$, and $\widetilde{\Theta}(\cdot)$ to hide logarithmic factors.

Date and Model. We consider a decentralized learning problem with K clients. Following the two-patch distributional framework commonly used in feature learning, we first assign each client a signal vector $\boldsymbol{\mu}_k \in \mathbb{R}^d$ sampled from a discrete distribution $\boldsymbol{\mu}_k \sim P\big(\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \ldots, \boldsymbol{\mu}^{(C)}\big)$, where P assigns probabilities to the candidate signal vectors $\{\boldsymbol{\mu}^{(c)}\}_{c=1}^{C}$. For simplicity, we assume that these signal vectors are mutually orthogonal.

Conditional on the client signals, a sample from the local distribution $\mathcal{D}_k$ has the form. The label $y_{k,i} \in \{+1, -1\}$ is a Rademacher random variable, and the input $\mathbf{x}_{k,i} = [\mathbf{x}_{k,i}^{(1)}, \mathbf{x}_{k,i}^{(2)}] \in (\mathbb{R}^d)^2$ consists of one signal patch and one noise patch: $\mathbf{x}_{k,i}^{(1)} = y_{k,i}\boldsymbol{\mu}_k, \mathbf{x}_{k,i}^{(2)} = \boldsymbol{\xi}_{k,i}$, where $\boldsymbol{\xi}_{k,i} \sim \mathcal{N}(0, \sigma_p^2\mathbf{H})$ is independent of $y_{k,i}$ with σ_p being the strength of noise, and $\mathbf{H} = \mathbf{I}_d - \sum_{c=1}^{C} \frac{\boldsymbol{\mu}^{(c)}(\boldsymbol{\mu}^{(c)})^\top}{\|\boldsymbol{\mu}^{(c)}\|_2^2}$. Thus, the first patch contains label-dependent signal information, while the second patch represents label-independent noise lying in the orthogonal complement of the signal subspace.

Each client maintains a two-layer ReLU CNN with m filters per branch and fixed second-layer signs. For client k, the model is

$$f_k(\mathbf{W}_k, \mathbf{x}) = F_{+1}(\mathbf{W}_{k,+1}, \mathbf{x}) - F_{-1}(\mathbf{W}_{k,-1}, \mathbf{x}), \tag{1}$$

where, for $j \in \{+1, -1\}$,

$$F_j(\mathbf{W}_{k,j}, \mathbf{x}) = \frac{1}{m}\sum_{r=1}^{m}\Big[\sigma\big(\mathbf{w}_{k,j,r}^\top \mathbf{x}^{(1)}\big) + \sigma\big(\mathbf{w}_{k,j,r}^\top \mathbf{x}^{(2)}\big)\Big]. \tag{2}$$

Here m is the network width, $\mathbf{w}_{k,j,r} \in \mathbb{R}^d$ is the r-th first-layer filter associated with sign j, $\mathbf{W}_{j,k}$ is the symbol to represent the model's weights collectively, and $\sigma(z) = \max\{0, z\}$ is the ReLU activation. Moreover, each weight in the first layer is initialized from a Gaussian distribution, $\mathbf{w}_{j,r} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2\mathbf{I}_{d\times d})$ for all $r \in [m]$ and $j \in \{-1, 1\}$.

Objectives in Decentralized Learning. Let $\mathcal{S}_k = \{(\mathbf{x}_{k,i}, y_{k,i})\}_{i=1}^{n_k}$ be the local training set of client k, and let $\mathcal{S} = \{\mathcal{S}_k\}_{k=1}^{K}$ with $n = \sum_{k=1}^{K} n_k$. We distinguish between empirical logistic loss and test classification error:

$$\begin{aligned} L_{\mathcal{S}}(\mathbf{W}) &= \sum_{k=1}^{K}\frac{n_k}{n}L_{\mathcal{S}_k}(\mathbf{W}_k), \qquad \ell(z) = \log(1+\exp(-z)) \\ L_{\mathcal{S}_k}(\mathbf{W}_k) &= \frac{1}{n_k}\sum_{i=1}^{n_k}\ell\big(y_{k,i}f_k(\mathbf{W}_k, \mathbf{x}_{k,i})\big), \qquad \mathcal{R}_k(\mathbf{W}_k) = \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_k}\{yf(\mathbf{W}_k, \mathbf{x}) \le 0\} \end{aligned} \tag{3}$$

where $\mathbf{W}_k$ denotes the local model maintained by client k. The clients collectively seek to minimize the empirical objective $L_{\mathcal{S}}(\mathbf{W})$, while client-level generalization is evaluated by $\mathcal{R}_k(\mathbf{W}_k)$ on the corresponding local distribution $\mathcal{D}_k$. Unlike centralized federated learning, which enforces consensus through server aggregation, decentralized learning relies on repeated peer-to-peer communication. We next formalize this communication mechanism.

Communication and Updates. Clients communicate over an undirected graph $\mathcal{G} = ([K], E)$. Their information exchange is characterized by a mixing matrix $\mathbf{M} \in \mathbb{R}^{K\times K}$, where $\mathbf{M}_{kk'}$ specifies the weight that client k assigns to the current model of client k'.

Assumption 3.1 (Mixing matrix)**.** *The communication matrix $\mathbf{M} \in \mathbb{R}^{K\times K}$ is nonnegative, symmetric, doubly stochastic, and positive semidefinite, with $\lambda_2(\mathbf{M}) < 1$. It is supported on $\mathcal{G} = ([K], E)$: $M_{kl} = 0$ whenever $k \neq l$ and $(k, l) \notin E$.*

This assumption is standard in decentralized optimization and consensus algorithms (Xiao & Boyd, 2004; Lian et al., 2017; Koloskova et al., 2020a). The condition $\lambda_2(\mathbf{M}) < 1$ guarantees that information eventually mixes across the connected communication graph. With gossip communication specified by $\mathbf{M}$, we now combine it with local gradient descent.

At iteration t, for any $k \in [K], j \in \{-1, +1\}, r \in [m]$, decentralized gradient descent updates the local models as

$$\begin{aligned}\mathbf{w}_{k,j,r}^{(t+1)} &= \sum_{l=1}^{K} M_{kl}\mathbf{w}_{l,j,r}^{(t)} - \eta\nabla_{\mathbf{w}_{k,j,r}} L_{\mathcal{S}_k}(\mathbf{W}_k^{(t)}) \\ &= \sum_{l=1}^{K} M_{kl}\mathbf{w}_{l,j,r}^{(t)} - \frac{\eta}{n_k m}\sum_{i=1}^{n_k} \ell_{k,i}'^{(t)}\sigma'\big(\langle \mathbf{w}_{k,j,r}^{(t)}, \boldsymbol{\xi}_{k,i}\rangle\big) j y_{k,i}\boldsymbol{\xi}_{k,i} - \frac{\eta}{n_k m}\sum_{i=1}^{n_k} \ell_{k,i}'^{(t)}\sigma'\big(\langle \mathbf{w}_{k,j,r}^{(t)}, y_{k,i}\boldsymbol{\mu}_k\rangle\big) j\boldsymbol{\mu}_k,\end{aligned} \tag{4}$$

where

$$\ell_{k,i}'^{(t)} = \ell'\big(y_{k,i} f_k(\mathbf{W}_k^{(t)}, \mathbf{x}_{k,i})\big) = -\frac{\exp\{-y_{k,i} f_k(\mathbf{W}_k^{(t)}, \mathbf{x}_{k,i})\}}{1 + \exp\{-y_{k,i} f_k(\mathbf{W}_k^{(t)}, \mathbf{x}_{k,i})\}}.$$

The first term in Equation 4 propagates the features learned by neighboring clients, whereas the remaining terms incorporate the signal and noise contained in the local data. This interaction between local feature acquisition and network communication is the central object of our analysis.

4 Main Results

In this section, we present our main theoretical results. Throughout the analysis, we work under the following scaling conditions. For simplicity, we assume $\bar{n} = n_k$ and $\|\boldsymbol{\mu}_k\|_2 = \|\boldsymbol{\mu}\|_2$ for all $k \in [K]$.

Assumption 4.1. *Fix $\delta \in (0, 1)$. Suppose that:*

- *The dimension d is sufficiently large: $d = \tilde{\Omega}(\max\{n^2, \sigma_p^{-2} K^2 \bar{n}\|\mu\|_2^2\})$.*
- *The size of training sample $\bar{n}$ and width of CNNs m adhere to $\bar{n} = \Omega(\log(d/\delta), m = \Omega(\log d/\delta)$. All network models share the same initialization.*
- *The learning rate η satisfies $\eta \leq O(m \min\{\frac{1+\lambda_{min}(M)}{\|\mu\|_2^2+\sigma_p^2 d}, \frac{\bar{n}\min_k M_{kk}}{\sigma_p^2 d}\}$. The standard deviation of Gaussian initialization σ_0 is chosen such that $0 < \sigma_0 \leq \tilde{O}(1) \cdot \min\{(\sigma_p\sqrt{d})^{-1}, \|\mu\|_2^{-1}\}$.*

These conditions are standard in feature learning analyses (Cao et al., 2022; Kou et al., 2023; Huang et al., 2024). The dimension condition controls correlations among sample-specific noise vectors, while the sample-size and width conditions ensure the required concentration properties. The step-size condition keeps the gradient dynamics stable over the training horizon, and the initialization upper bound keeps the initial network responses sufficiently small.

4.1 Coefficient dynamics

Under the two-patch model, each local gradient is a linear combination of the client's signal and training-noise vectors, while communication mixes these contributions across clients, as in Equation 4. The following lemma gives the resulting decomposition of each filter into shared initialization, signal, and noise components, together with the corresponding coefficient dynamics.

Lemma 4.2 (Signal–noise decomposition)**.** *Under the same initialization and Assumptions 3.1. For every integer $t \geq 0$, $k \in [K]$, $j \in \{-1, +1\}$, and $r \in [m]$, the weights admit the decomposition*

$$\mathbf{w}_{k,j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j\sum_{k'=1}^{K} \gamma_{k,k',j,r}^{(t)} \frac{\boldsymbol{\mu}_{k'}}{\|\boldsymbol{\mu}_{k'}\|_2^2} + \sum_{k'=1}^{K}\sum_{i=1}^{n_{k'}} \rho_{k,k',j,r,i}^{(t)} \frac{\boldsymbol{\xi}_{k',i}}{\|\boldsymbol{\xi}_{k',i}\|_2^2}, \tag{5}$$

Furthermore, the coefficients satisfy

$$\gamma^{(0)}_{k,k',j,r} = \rho^{(0)}_{k,k',j,r,i} = 0, \tag{6}$$

$$\gamma^{(t+1)}_{k,k',j,r} = \sum_{l=1}^{K} M_{kl}\gamma^{(t)}_{l,k',j,r} - \frac{\eta\|\boldsymbol{\mu}_{k'}\|_2^2}{\bar{n}m}\mathbf{1}\{k'=k\}\sum_{i=1}^{n_{k'}} \ell'^{(t)}_{k',i}\sigma'\big(\langle \mathbf{w}^{(t)}_{k',j,r}, y_{k',i}\boldsymbol{\mu}_{k'}\rangle\big), \tag{7}$$

$$\psi^{(t+1)}_{k,k',j,r,i} = \sum_{l=1}^{K} M_{kl}\psi^{(t)}_{l,k',j,r,i} - \frac{\eta\|\boldsymbol{\xi}_{k',i}\|_2^2}{\bar{n}m}\ell'^{(t)}_{k',i}\sigma'\big(\langle \mathbf{w}^{(t)}_{k',j,r}, \boldsymbol{\xi}_{k',i}\rangle\big)\mathbf{1}\{y_{k',i}=j\}\mathbf{1}\{k'=k\}, \tag{8}$$

$$\phi^{(t+1)}_{k,k',j,r,i} = \sum_{l=1}^{K} M_{kl}\phi^{(t)}_{l,k',j,r,i} + \frac{\eta\|\boldsymbol{\xi}_{k',i}\|_2^2}{\bar{n}m}\ell'^{(t)}_{k',i}\sigma'\big(\langle \mathbf{w}^{(t)}_{k',j,r}, \boldsymbol{\xi}_{k',i}\rangle\big)\mathbf{1}\{y_{k',i}=-j\}\mathbf{1}\{k'=k\}. \tag{9}$$

where $\psi^{(t)}_{k,k',j,r,i} = \max\{\rho^{(t)}_{k,k',j,r,i}, 0\}$ *and* $\phi^{(t)}_{k,k',j,r,i} = \min\{\rho^{(t)}_{k,k',j,r,i}, 0\}$.

We defer the proof to Appendix A.2. In Lemma 4.2, γ and ρ track signal and training-noise contributions from source client k' to receiving client k, respectively. Equation 7–Equation 9 show that local gradients introduce new contributions at their source, while mixing propagates existing contributions through the graph. We next quantify the resulting signal and noise accumulation. Table 2 in the appendix summarizes the assumptions and scope of these two groups of results below.

4.2 Early feature learning and risk certificates

To characterize how communication topology affects feature learning, for client k and training horizon $T \geq 1$, define

$$\mathbf{A}_T = \sum_{t=0}^{T-1}\mathbf{M}^t, \chi_k = \sum_{q=1}^{K}\frac{\langle\boldsymbol{\mu}_k,\boldsymbol{\mu}_q\rangle}{\|\boldsymbol{\mu}_k\|_2^2}, S_k(T) = \sum_{q=1}^{K}[\mathbf{A}_T]_{kq}\frac{\langle\boldsymbol{\mu}_k,\boldsymbol{\mu}_q\rangle}{\|\boldsymbol{\mu}_k\|_2^2}, Q_k(T) = \sum_{q=1}^{K}[\mathbf{A}_T]_{kq}^2. \tag{10}$$

Here, χ_k measures aggregate feature similarity, $S_k(T)$ captures the cumulative predictive signal reaching client k, and $Q_k(T)$ controls the corresponding accumulation of training noise. These quantities characterize finite-time feature learning dynamics and their implications for generalization.

Lemma 4.3 (Finite-time signal and noise scales)**.** *Under Assumption 3.1 and 4.1, let* $T_0 = \left\lfloor \frac{c_T m}{\eta}\min\left\{\|\mu\|_2^{-2}, \frac{\bar{n}}{\sigma_p^2 d}\right\}\right\rfloor$, *where* $c_T > 0$ *is a sufficiently small universal constant. With probability at least* $1-\delta$ *and* $1 \leq T \leq T_0$, *at least* $\Omega(m)$ *fixed initially aligned filters satisfy* $j\left\langle \mathbf{w}^{(T)}_{k,j,r} - \mathbf{w}^{(0)}_{j,r}, \boldsymbol{\mu}_k\right\rangle = \Theta\Big(\frac{\eta\|\mu\|_2^2}{m}S_k(T)\Big)$. *and every filter satisfies* $\|\mathbf{H}\mathbf{w}^{(T)}_{k,j,r}\|_2^2 = O\Big(\sigma_0^2 d + \frac{\eta^2\sigma_p^2 d}{\bar{n}m^2}Q_k(T)\Big)$. *The training logits remain* $O(1)$ *throughout this interval.*

The proof appears in Appendix A.3. Lemma 4.3 shows that contributions from clients sharing the same signal accumulate through $S_k(T)$, whereas the near-orthogonality of training-noise vectors yields an energy bound involving $Q_k(T)$. This energy bound controls the network's response to fresh test noise. Comparing this response with the learned signal, while accounting for initialization, gives the following sufficient condition for small client-level test error.

Theorem 4.4 (Early generalization)**.** *Under the conditions of Lemma 4.3, let* $L = \log(2mnK/\delta)$. *On the same event, the following holds simultaneously for every client* $k \in [K]$ *and time* $1 \leq T \leq T_0$. *If*

$$\sigma_0\mu\sqrt{L} \leq c_{\text{init}}\frac{\eta\|\mu\|_2^2}{m}S_k(T), \qquad \sigma_0^2 d \leq \frac{\eta^2\sigma_p^2 d}{\bar{n}m^2}Q_k(T), \tag{11}$$

and

$$\frac{\bar{n}\|\mu\|_2^4}{\sigma_p^4 d}\frac{S_k(T)^2}{Q_k(T)} \geq C,$$

Then

$$\mathcal{R}_k(\mathbf{W}_k^{(T)}) \leq \exp\left[-c\frac{\bar{n}\|\mu\|_2^4}{\sigma_p^4 d}\frac{S_k(T)^2}{Q_k(T)}\right]. \tag{12}$$

Here $c_{\text{init}} > 0$ is a sufficiently small universal constant and $C > 0$ is a sufficiently large universal constant.

Theorem 4.4 separates the statistical signal-to-noise factor $\bar{n}\mu^4/(\sigma_p^4 d)$ from the communication-dependent ratio $S_k(T)^2/Q_k(T)$. The following corollary illustrates how communication topology and signal placement affect the ratio $S_k(T)^2/Q_k(T)$.

Proposition 4.5 (Spectral characterization)**.** *Under the stated signal geometry, let $\mathbf{M}$ be a fixed mixing matrix satisfying Assumption 3.1. For every client $k \in [K]$ and integer $T \geq 1$,*

$$\left| S_k(T) - \frac{T\chi_k}{K} \right| \leq \frac{1-\lambda_2^T}{1-\lambda_2} \sqrt{\chi_k(1-\chi_k/K)}. \tag{13}$$

Moreover,

$$\lim_{T\to\infty} \frac{S_k(T)^2}{Q_k(T)} = \frac{\chi_k^2}{K}. \tag{14}$$

The limit depends only on aggregate signal similarity, but the finite-time factor can distinguish signal placements with the same χ_k on the same graph. The following example makes this distinction explicit.

Signal placement on a fixed graph. The asymptotic ratio depends only on χ_k, whereas finite-time exposure also depends on placement. For the lazy four-cycle $M = \frac{1}{2}I + \frac{1}{4}\operatorname{Adj}(C_4)$ and orthogonal equal-norm signals a, b, the assignments (a, a, b, b) and (a, b, a, b) have the same $\chi_k = 2$ and $Q_1(2)$, but yield $S_1(2)^2/Q_1(2) = 49/38$ and $18/19$, respectively. Only the adjacent same-type client can contribute to node 1 by $T = 2$.

4.3 Fitting and risk at a common iterate

Moving beyond early feature learning, we establish convergence of the average empirical loss and characterize client-level test error at a common fitted iterate.

To preserve a fixed fraction of training-noise activations throughout fitting, we additionally assume

$$\sigma_0 = \widetilde{\Omega}\left(\frac{n}{\sigma_p d}\right). \tag{15}$$

This initialization condition also appears in Cao et al. (2022); Huang et al. (2024) and is compatible with the upper bound in Assumption 4.1 under our dimension condition; see Appendix A.1 for explicit logarithmic factors.

Theorem 4.6 (Training convergence)**.** *Under Assumptions 3.1 and 4.1, suppose that condition 15 holds. Fix $\epsilon \in (0,1)$ and let $T_1 = \Theta(Km\bar{n}/(\eta\sigma_p^2 d))$. There exists the $T_\epsilon = \widetilde{O}(mK^2\bar{n}/(\eta\epsilon\sigma_p^2 d))$ such that, with probability at least $1-\delta$, some integer $t_* \in [T_1, T_\epsilon)$ satisfies*

$$L_S(\mathbf{W}^{(t_*)}) \leq \epsilon.$$

Theorem 4.6 guarantees small average empirical loss at a common iterate. The following theorem characterizes the client-level test error at this same iterate by comparing the learned signal with the response to fresh test noise.

Theorem 4.7 (Generalization at a fitted iterate)**.** *Under the assumptions of Theorem 4.6, with probability at least $1-\delta$, there exists a common iterate $t_* \in [T_1, T_\epsilon)$ satisfying $L_S(\mathbf{W}^{(t_*)}) \leq \epsilon$ such that the following hold simultaneously for every client $k \in [K]$.*

- ***Sufficient signal for small test error.*** *Suppose additionally that $\eta = O\left(\frac{\bar{n}m(1-\lambda_2)}{\sigma_p^2 d}\right)$. For any $\zeta \in (0,1)$, if*

$$\frac{\bar{n}\chi_k\|\mu\|_2^2}{\sigma_p^2 d} = \tilde{\Omega}\left[\left(\sigma_0\|\mu\|_2 + \sigma_0\sigma_p\sqrt{d} + \sqrt{\frac{n}{d}}\right)\sqrt{\log\frac{2}{\zeta}}\right], \tag{16}$$

then $\mathcal{R}_k(\mathbf{W}_k^{(t_*)}) \leq \zeta$.

- ***Insufficient signal and constant test error.*** *If $\bar{n}\chi_k\|\mu\|_2^4/(\sigma_p^4 d) \le c_0$, then $\mathcal{R}_k(\mathbf{W}_k^{(t_*)}) \ge c_*$, where $c_0 > 0$ is a sufficiently small universal constant and $c_* > 0$ is universal.*

Theorem 4.7 distinguishes fitting the training data from learning predictive features. Under the stated signal geometry, $\bar{n}\chi_k$ counts the training examples sharing client k's signal, whereas the noise bound involves all n training examples. For fixed sample sizes and signal strength, greater cross-client similarity therefore strengthens the useful-signal term in Equation 16 without enlarging its sample size-dependent noise term. The initialization terms remain relevant because test predictions depend on the full learned weights, not only on their increments.

Conditioning on η links communication speed to the local learning rate. A smaller spectral gap requires a smaller step size for the stated upper guarantee, allowing shared signals to propagate over the corresponding learning timescale. The constant-error conclusion reveals that the average empirical loss can be small while a client's test error remains bounded away from zero. Thus, successful fitting alone does not ensure that a sufficient predictive signal has been learned relative to the response to fresh noise.

4.4 Periodic Local Updates

We next examine how communication frequency affects signal acquisition and noise accumulation. The same propagation argument accommodates multiple local updates between communications. Fix integers $x \ge 0$ and $y \ge 1$. Each period consists of x local-only updates followed by y updates with communication, implemented by setting $\mathbf{M}_t = I$ and $\mathbf{M}_t = \mathbf{M}$, respectively:

$$\mathbf{W}_k^{(t+1)} = \sum_{l=1}^{K} (\mathbf{M}_t)_{kl} \mathbf{W}_l^{(t)} - \eta \nabla L_{\mathcal{S}_k}(\mathbf{W}_k^{(t)}). \tag{17}$$

Every iteration evaluates a full local gradient; only the last y iterations of each period communicate. Thus $x = 0, y = 1$ recovers Equation 4. The coefficient recursions retain their form (Equation 7–9), with $\mathbf{M}$ replaced by $\mathbf{M}_t$. Accordingly, the accumulated propagation matrix becomes

$$\mathbf{A}_T = \sum_{s=0}^{T-1} \mathbf{M}_{T-1} \cdots \mathbf{M}_{s+1}, \tag{18}$$

where an empty product is I. Equivalently, $\mathbf{A}_{t+1} = \mathbf{M}_t \mathbf{A}_t + I$ with $\mathbf{A}_0 = 0$. The product excludes $\mathbf{M}_s$ because the gradient generated at step s is applied locally and propagates only in subsequent iterations. We use the definitions of $S_k(T)$ and $Q_k(T)$ in Equation 10 and $\mathbf{A}_T$ above.

Theorem 4.4 relates early test error to the ratio $S_k(T)^2/Q_k(T)$. The following proposition quantifies how the communication schedule changes these quantities and gives a sufficient local-update budget for retaining the well-mixed scale χ_k^2/K.

Proposition 4.8 (Communication frequency and early generalization)**.** *Let $\mathbf{M}$ satisfy Assumption 3.1. Under their respective assumptions, the feature bounds in Lemma 4.3 and the risk bound in Theorem 4.4 remain valid for Equation 17, with the same early window T_0 and with $S_k(T), Q_k(T)$ computed from Equation 18.*

In particular, consider $y = 1$ and an integer $1 \le T \le T_0$. If

$$x + 1 \le \frac{(1 - \lambda_2) T \sqrt{\chi_k}}{2K}, \tag{19}$$

then

$$\frac{S_k(T)^2}{Q_k(T)} \ge \frac{\chi_k^2}{8K}. \tag{20}$$

Under the initialization and signal-strength conditions of Theorem 4.4, this yields

$$\mathcal{R}_k(\mathbf{W}_k^{(T)}) \le \exp\left[-c \frac{\bar{n}\|\mu\|^4}{\sigma_p^4 d} \frac{\chi_k^2}{K} \right], \tag{21}$$

where $c > 0$ is universal.

Condition Equation 19 provides a sufficient local-update budget to retain the well-mixed scale of $S_k(T)^2/Q_k(T)$. At a horizon T, a larger spectral gap permits a longer interval within this guarantee. For intervals exceeding this budget, the proposition no longer ensures this well-mixed scale.

Communication frequency plays a different role in the fitting analysis. Under the fitting assumptions, the training guarantee and the constant-test-error lower bound remain of order T_ϵ, without an additional factor depending on x or y. Hence, infrequent communication can still allow for fitting through local noise memorization, even when the early cross-client generalization guarantee is unavailable. Consistent with this distinction, our periodic experiments do not show a universal monotonic dependence of test error on the communication interval; the effect also depends on client placement and the evaluation distribution. We provide detailed proofs and experiments in Appendix B and C.

5 Experiments

Experimental setup. Our synthetic experiments follow the two-patch model in Section 3 and train the two-layer ReLU CNN using decentralized gradient descent. We examine how graph structure, client placement, and signal strength affect the measured signal and noise dynamics and own-type test error. Within each comparison, we share the initial weights and the randomness used to generate client data; we report results over five seeds. We also examine client placement on binary rotated MNIST. Appendix C contains detailed experimental settings and additional results.

5.1 Signals and Training Noise

We separately examine signal exposure $S_k(T)$ and noise accumulation $Q_k(T)$, and then relate their combined factor $\mathcal{G}_k(T) = S_k(T)^2/Q_k(T)$ to own-type generalization.

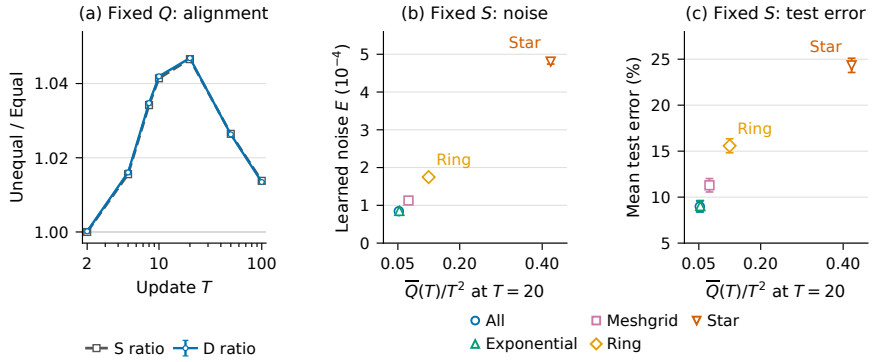

Figure 1: **Testing the two graph quantities separately.** (a) Fixed graph and Q: S and signal ratios, Unequal-block / Equal-block. (b,c) Fixed $S = T$: noise energy and test error across all five IID graphs at $T = 20$. D: good-filter signal increment; E: noise-subspace increment energy.

Fixed Q: multi-hop signal exposure. We first isolate the role of $S_k(T)$. On the same ring, we construct two signal layouts with identical type counts, graph spectrum, Q, and matched one-hop same-type-neighbor statistics, but different multi-hop exposure. Specifically, *Equal-block* (AAABBBAAABBB) consists of four equal-length blocks, whereas *Unequal-block* (AAAABBBBAABB) uses block lengths $(4, 4, 2, 2)$. This leaves the predicted signal exposure S as the main variable in Figure 1(a); the layout with larger S also exhibits a larger learned signal increment, while the learned noise energy remains essentially unchanged. This supports the role of $S_k(T)$ in capturing signal acquisition that is not explained by graph-only or one-hop statistics.

Fixed S: topology-dependent noise accumulation. We next isolate the role of $Q_k(T)$ by assigning the same signal to every client. In this setting, $S_k(T) = T$ for every topology, while $Q_k(T)$ still depends on how concentrated the cumulative communication weights are. As shown in Figure 1(b,c), topologies with larger Q exhibit greater learned noise energy, whereas the learned signal remains

similar. The corresponding early test errors follow the same qualitative trend, supporting the role of $Q_k(T)$ in finite-time noise accumulation. These differences become smaller later in training.

5.2 Connecting signal and noise scales to generalization

We use $K = 16$ clients with two orthogonal signal types on a lazy ring. For Table 1, communication occurs every step ($h = 1$), and we evaluate at $T = 32$. Table 1 compares client placements at fixed signal strength and signal strengths within each fixed layout. Here $\overline{\mathcal{G}} = K^{-1} \sum_k S_k(T)^2 / Q_k(T)$.

Communication-dependent factor. We first fix the statistical factor at 1 and rearrange the signal types on the ring. The graph and $Q_k(T)$ remain unchanged, while $\overline{\mathcal{G}}$ increases from Alternating through Random to Clustered. Own-type error decreases in the same order, from 18.09% to 14.43% to 9.38%. This comparison links the placement-dependent signal exposure to early client-level performance.

Statistical factor. We then vary the signal strength while fixing the topology, client placement, and communication schedule, so that $S_k(T)^2/Q_k(T)$ is unchanged within each layout. Table 1(b) shows that increasing the statistical factor $\bar{n}\|\mu\|^4/(\sigma_p^4 d)$ substantially reduces own-type test error for both Clustered and Alternating layouts. These results are consistent with the statistical factor in our generalization bound. Together with the communication-dependent intervention, they support the joint role of statistical signal strength and communication-dependent signal-to-noise accumulation in generalization.

(a) Communication-dependent factor

Layout	$\overline{\mathcal{G}}$	Own Error (%)
Alternating	2.350	18.09±1.04
Random	3.384±0.705	14.43±2.13
Clustered	5.009	9.38±0.84

(b) Statistical factor

Factor	Clustered	Alternating
0.1	33.93±0.77	39.01±1.12
1	9.38±0.84	18.09±1.04
10	0.018±0.008	0.239±0.051

Table 1: Two complementary interventions on early generalization. Left: the ring layouts vary $\overline{\mathcal{G}}$ while keeping the data scale fixed. Right: the statistical factor $\bar{n}\|\mu\|^4/(\sigma_p^4 d)$ is varied while the ring topology and communication schedule are fixed.

Client-level performance on real data. We examine how client placement affects own-domain performance on binary rotated MNIST. On the same ring, clustering clients with the same rotation yields a lower mean own-domain error than alternating the two rotations when communication occurs every $h = 16$ updates (0.218% versus 0.300%). The difference is much smaller when clients communicate every update ($h = 1$). This qualitative result suggests that placement has a more visible effect on client-level performance when communication is less frequent. We report full results for other evaluation distributions in Appendix C.

Additional experiments. Appendix C reports complete experimental settings, training trajectories, additional topology and placement comparisons, communication-interval studies, initialization controls, and real-data results.

6 Conclusion

We analyze decentralized gradient descent for a two-layer ReLU CNN through feature learning. Our finite-time generalization bounds identify distinct roles for signal exposure $S_k(T)$ and noise accumulation $Q_k(T)$, capturing signal-placement effects beyond graph spectra and aggregate signal similarity. At a common iterate with small average empirical loss, we give sufficient conditions for small client-level test error and identify a weak-signal regime in which test error remains constant. Extending the analysis to periodic local updates, we identify a sufficient local-update budget for preserving the well-mixed scale of the early generalization bound. Controlled synthetic experiments support these learning mechanisms, and real-data experiments illustrate placement-dependent performance on a fixed communication graph. Together, these results show that understanding the learning effects of communication requires considering not only mixing speed, but also which predictive signals reach each client and how sample-specific noise accumulates.

AI USE STATEMENT

We used AI to organize and edit the manuscript, check notation and proof dependencies, and assist with mathematical derivations.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Uuf2q9TfXGA.

Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. Stochastic gradient push for distributed deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 344–353. PMLR, 2019. URL https://proceedings.mlr.press/v97/assran19a.html.

Yajie Bao, Michael Crawshaw, and Mingrui Liu. Provable benefits of local steps in heterogeneous federated learning for neural networks: A feature learning perspective. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2857–2902. PMLR, 2024. URL https://proceedings.mlr.press/v235/bao24a.html.

Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25237–25250. Curran Associates, Inc., 2022. doi: 10.52202/068431-1830. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/a12c999be280372b157294e72a4bbc8b-Abstract-Conference.html.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1812.07956.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Local SGD optimizes overparameterized neural networks in polynomial time. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 6840–6861. PMLR, 2022. URL https://proceedings.mlr.press/v151/deng22a.html.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1225–1234. PMLR, 2016. URL https://proceedings.mlr.press/v48/hardt16.html.

Baihe Huang, Xiaoxiao Li, Zhao Song, and Xin Yang. FL-NTK: A neural tangent kernel-based framework for federated learning analysis. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4423–4434. PMLR, 2021. URL https://proceedings.mlr.press/v139/huang21c.html.

Wei Huang, Ye Shi, Zhongyi Cai, and Taiji Suzuki. Understanding convergence and generalization in federated learning through feature learning theory. In *The Twelfth International Conference on Learning Representations*, pp. 25655–25686, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6d892051f91812700f427250e95e04f5-Abstract-Conference.html.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1806.07572.

Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9965–10040. PMLR, 2022. URL https://proceedings.mlr.press/v162/jelassi22a.html.

Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3478–3487. PMLR, 2019. URL https://proceedings.mlr.press/v97/koloskova19a.html.

Anastasia Koloskova, Tao Lin, Sebastian U. Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. In *International Conference on Learning Representations*, 2020a.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U. Stich. A unified theory of decentralized SGD with changing topology and local updates. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5381–5393. PMLR, 2020b. URL https://proceedings.mlr.press/v119/koloskova20a.html.

Lingjing Kong, Tao Lin, Anastasia Koloskova, Martin Jaggi, and Sebastian U. Stich. Consensus control for decentralized deep learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5686–5696. PMLR, 2021. URL https://proceedings.mlr.press/v139/kong21a.html.

Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting in two-layer ReLU convolutional neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17615–17659. PMLR, 2023. URL https://proceedings.mlr.press/v202/kou23a.html.

Batiste Le Bars, Aurélien Bellet, Marc Tommasi, Erick Lavoie, and Anne-Marie Kermarrec. Refined convergence and topology learning for decentralized SGD with heterogeneous data. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 1672–1702. PMLR, 2023. URL https://proceedings.mlr.press/v206/le-bars23a.html.

Batiste Le Bars, Aurélien Bellet, Marc Tommasi, Kevin Scaman, and Giovanni Neglia. Improved stability and generalization guarantees of the decentralized SGD algorithm. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26215–26240. PMLR, 2024. URL https://proceedings.mlr.press/v235/le-bars24a.html.

Qinglun Li, Yingqi Liu, Miao Zhang, Xiaochun Cao, Quanjun Yin, and Li Shen. Unveiling the power of multiple gossip steps: A stability-based generalization analysis in decentralized training. In *Advances in Neural Information Processing Systems*, volume 38, pp. 82058–82077. Neural Information Processing Systems Foundation, Inc., 2025. doi: 10.52202/085713-2474. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/6ad5d39b10e37915d7dfda2893d8e924-Abstract-Conference.html. Spotlight.

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. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5330–5340. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/f75526659f31040afeb61cb7133e4e6d-Abstract.html.

Enrique Tomás Martínez Beltrán, Mario Quiles Pérez, Pedro Miguel Sánchez Sánchez, Sergio López Bernal, Gérôme Bovet, Manuel Gil Pérez, Gregorio Martínez Pérez, and Alberto Huertas Celdrán. Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges. *IEEE Communications Surveys & Tutorials*, 25(4):2983–3013, 2023. doi: 10.1109/COMST.2023.3315746.

Angelia Nedić and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, January 2009. doi: 10.1109/TAC.2008.2009515. URL https://ieeexplore.ieee.org/document/4749425.

Angelia Nedić, Alex Olshevsky, and Michael G. Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018. doi: 10.1109/JPROC.2018.2817461.

Luigi Palmieri, Chiara Boldrini, Lorenzo Valerio, Andrea Passarella, and Marco Conti. Impact of network topology on the performance of decentralized federated learning. *Computer Networks*, 253:110681, 2024. doi: 10.1016/j.comnet.2024.110681. URL https://doi.org/10.1016/j.comnet.2024.110681.

Dominic Richards and Patrick Rebeschini. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *Journal of Machine Learning Research*, 21(34):1–44, 2020. URL https://jmlr.org/papers/v21/18-638.html.

Tao Sun, Dongsheng Li, and Bao Wang. Stability and generalization of decentralized stochastic gradient descent. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9756–9764, 2021. doi: 10.1609/aaai.v35i11.17173. URL https://ojs.aaai.org/index.php/AAAI/article/view/17173.

Hossein Taheri and Christos Thrampoulidis. On generalization of decentralized learning with separable data. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 4917–4945. PMLR, 2023. URL https://proceedings.mlr.press/v206/taheri23a.html.

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. D^2: Decentralized training over decentralized data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4848–4856. PMLR, 2018. URL https://proceedings.mlr.press/v80/tang18a.html.

John N. Tsitsiklis, Dimitri P. Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, 1986. doi: 10.1109/TAC.1986.1104412.

Thijs Vogels, Lie He, Anastasiia Koloskova, Sai Praneeth Karimireddy, Tao Lin, Sebastian U. Stich, and Martin Jaggi. RelaySum for decentralized deep learning on heterogeneous data. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/ebbdfea212e3a756a1fded7b35578525-Abstract.html.

Thijs Vogels, Hadrien Hendrikx, and Martin Jaggi. Beyond spectral gap: The role of the topology in decentralized learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15039–15050. Curran Associates, Inc., 2022. doi: 10.52202/068431-1094. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/61162d94822d468ee6e92803340f2040-Abstract-Conference.html.

Tongle Wu and Ying Sun. Implicit regularization of decentralized gradient descent for sparse regression. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1df5c96e327ea0cbd32e0d8bae835994-Abstract-Conference.html.

Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.

Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737. PMLR, 2021. URL https://proceedings.mlr.press/v139/yang21c.html.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016. doi: 10.1137/130943170.

Valentina Zantedeschi, Aurélien Bellet, and Marc Tommasi. Fully decentralized joint learning of personalized models and collaboration graphs. In *Proceedings of the Twenty Third International*

Conference on Artificial Intelligence and Statistics, volume 108 of *Proceedings of Machine Learning Research*, pp. 864–874. PMLR, 2020. URL https://proceedings.mlr.press/v108/zantedeschi20a.html.

Tongtian Zhu, Fengxiang He, Lan Zhang, Zhengyang Niu, Mingli Song, and Dacheng Tao. Topology-aware generalization of decentralized SGD. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27479–27503. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhu22d.html.

Tongtian Zhu, Wenhao Li, Can Wang, and Fengxiang He. DICE: Data influence cascade in decentralized learning. In *The Thirteenth International Conference on Learning Representations*, pp. 69007–69042, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ac8fbba029dadca99d6b8c3f913d3ed6-Abstract-Conference.html.

A PROOFS FOR DECENTRALIZED GRADIENT DESCENT

The argument separates the finite-time feature growth from the additive activation control needed to fit the training data. All probabilities below concern the training data and initialization.

A.1 NOTATION AND PARAMETER CONVENTIONS

Throughout the appendix, $n = K\bar{n}$, $\mu = \|\boldsymbol{\mu}_k\|_2$, and $c, C > 0$ denote universal constants that may change between displays. We use the single concentration factor $L = \log(2mnK/\delta)$.

We write down the parameters explicitly,

$$d \geq C \max\left\{n^2 L^2, \frac{Kn\mu^2}{\sigma_p^2}\right\}, \quad \bar{n}, m \geq C \log(d/\delta), \quad \sigma_0 \leq \frac{c}{\sqrt{L}} \min\{\mu^{-1}, (\sigma_p\sqrt{d})^{-1}\},$$

and the following sufficient version of the step-size condition:

$$0 < \eta \leq c_\eta m \min\left\{\frac{1 + \lambda_{\min}(\mathbf{M})}{\mu^2 + 2\sigma_p^2 d}, \frac{\bar{n} \min_k M_{kk}}{\sigma_p^2 d}\right\}. \tag{22}$$

Here c_η is a sufficiently small universal constant, chosen so that $c_\eta \leq c_T/8$, where c_T specifies the Stage 1 horizons below. In particular, $c_\eta \leq 1/48$ gives the two inequalities used in the later descent and gate arguments:

$$\eta \leq \frac{m}{4(\mu^2 + 2\sigma_p^2 d)}, \qquad \frac{2\eta\sigma_p^2 d}{\bar{n}m} \leq \min_k M_{kk}. \tag{23}$$

A.2 SIGNAL-NOISE DECOMPOSITION

Lemma A.1 (Restatement of Lemma 4.2)**.** *Under the shared initialization and communication assumptions, for every $t \geq 0$, $k \in [K]$, $j \in \{-1, +1\}$, and $r \in [m]$, the weights admit the decomposition*

$$\mathbf{w}_{k,j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j \sum_{k'=1}^{K} \gamma_{k,k',j,r}^{(t)} \frac{\boldsymbol{\mu}_{k'}}{\|\boldsymbol{\mu}_{k'}\|_2^2} + \sum_{k'=1}^{K} \sum_{i=1}^{n_{k'}} \rho_{k,k',j,r,i}^{(t)} \frac{\boldsymbol{\xi}_{k',i}}{\|\boldsymbol{\xi}_{k',i}\|_2^2}, \tag{24}$$

The coefficients are initialized at zero. We define the positive and negative parts

$$\psi_{k,k',j,r,i}^{(t)} = \max\{\rho_{k,k',j,r,i}^{(t)}, 0\}, \qquad \phi_{k,k',j,r,i}^{(t)} = \min\{\rho_{k,k',j,r,i}^{(t)}, 0\},$$

and write

$$\Psi_{k,k',i}^{(t)} := \max_{j,r} |\rho_{k,k',j,r,i}^{(t)}| = \max_{j,r}\{\psi_{k,k',j,r,i}^{(t)}, -\phi_{k,k',j,r,i}^{(t)}\}.$$

The coefficients satisfy

$$\gamma_{k,k',j,r}^{(t+1)} = \sum_{l=1}^{K} M_{kl} \gamma_{l,k',j,r}^{(t)} - \frac{\eta\|\boldsymbol{\mu}_{k'}\|_2^2}{n_{k'}m} \mathbf{1}\{k' = k\} \sum_{i=1}^{n_{k'}} \ell_{k',i}'^{(t)} \sigma'\left(\left\langle \mathbf{w}_{k',j,r}^{(t)}, y_{k',i}\boldsymbol{\mu}_{k'} \right\rangle\right),$$

$$\psi_{k,k',j,r,i}^{(t+1)} = \sum_{l=1}^{K} M_{kl} \psi_{l,k',j,r,i}^{(t)} - \frac{\eta\|\boldsymbol{\xi}_{k',i}\|_2^2}{n_{k'}m} \ell_{k',i}'^{(t)} \sigma'\left(\left\langle \mathbf{w}_{k',j,r}^{(t)}, \boldsymbol{\xi}_{k',i} \right\rangle\right) \mathbf{1}\{y_{k',i} = j\}\mathbf{1}\{k' = k\},$$

$$\phi_{k,k',j,r,i}^{(t+1)} = \sum_{l=1}^{K} M_{kl} \phi_{l,k',j,r,i}^{(t)} + \frac{\eta\|\boldsymbol{\xi}_{k',i}\|_2^2}{n_{k'}m} \ell_{k',i}'^{(t)} \sigma'\left(\left\langle \mathbf{w}_{k',j,r}^{(t)}, \boldsymbol{\xi}_{k',i} \right\rangle\right) \mathbf{1}\{y_{k',i} = -j\}\mathbf{1}\{k' = k\}.$$

Proof. The decomposition holds at $t = 0$ by shared initialization. Suppose it holds at time t; we proceed to the next deductive step.

According to Equation 4, the local update is

$$\begin{aligned}
\mathbf{w}_{k,j,r}^{(t+1)} &= \sum_{l=1}^{K} M_{kl}\mathbf{w}_{l,j,r}^{(t)} - \eta\nabla_{\mathbf{w}_{k,j,r}} L_{S_k}(\mathbf{W}_k^{(t)}) \\
&= \sum_{l=1}^{K} M_{kl}\mathbf{w}_{l,j,r}^{(t)} - \frac{\eta}{n_k m}\sum_{i=1}^{n_k} \ell_{k,i}'^{(t)}\sigma'(\langle \mathbf{w}_{k,j,r}^{(t)}, \boldsymbol{\xi}_{k,i}\rangle) j y_{k,i}\boldsymbol{\xi}_{k,i} - \frac{\eta}{n_k m}\sum_{i=1}^{n_k} \ell_{k,i}'^{(t)}\sigma'(\langle \mathbf{w}_{k,j,r}^{(t)}, y_{k,i}\cdot\boldsymbol{\mu}_k\rangle) j\boldsymbol{\mu}_k \\
&= \sum_{l=1}^{K} M_{kl}(\mathbf{w}_{j,r}^{(0)} + \sum_{k'=1}^{K} j\gamma_{l,k',j,r}^{(t)}\|\boldsymbol{\mu}_{k'}\|_2^{-2}\boldsymbol{\mu}_{k'} + \sum_{i=1}^{n_{k'}} \rho_{l,k',j,r,i}^{(t)}\|\boldsymbol{\xi}_{k',i}\|_2^{-2}\boldsymbol{\xi}_{k',i}) \\
&\quad - \frac{\eta}{n_k m}\sum_{i=1}^{n_k} \ell_{k,i}'^{(t)}\sigma'(\langle \mathbf{w}_{k,j,r}^{(t)}, \boldsymbol{\xi}_{k,i}\rangle) j y_{k,i}\boldsymbol{\xi}_{k,i} - \frac{\eta}{n_k m}\sum_{i=1}^{n_k} \ell_{k,i}'^{(t)}\sigma'(\langle \mathbf{w}_{k,j,r}^{(t)}, y_{k,i}\cdot\boldsymbol{\mu}_k\rangle) j\boldsymbol{\mu}_k \\
&= \mathbf{w}_{j,r}^{(0)} + \sum_{k'=1}^{K}(j\gamma_{k,k',j,r}^{(t+1)}\|\boldsymbol{\mu}_{k'}\|_2^{-2}\boldsymbol{\mu}_{k'} + \sum_{i=1}^{n_{k'}} \rho_{k,k',j,r,i}^{(t+1)}\|\boldsymbol{\xi}_{k',i}\|_2^{-2}\boldsymbol{\xi}_{k',i}),
\end{aligned} \tag{25}$$

where the final equality takes the property of M and the coefficients can be expressed as

$$\gamma_{k,k',j,r}^{(t+1)} = \sum_{l=1}^{K} M_{kl}\gamma_{l,k',j,r}^{(t)} - \frac{\eta\|\boldsymbol{\mu}_{k'}\|_2^2}{n_{k'}m}\mathbf{1}\{k'=k\}\sum_{i=1}^{n_{k'}} \ell_{k',i}'^{(t)}\sigma'\Big(\Big\langle \mathbf{w}_{k',j,r}^{(t)}, y_{k',i}\boldsymbol{\mu}_{k'}\Big\rangle\Big),$$

$$\psi_{k,k',j,r,i}^{(t+1)} = \sum_{l=1}^{K} M_{kl}\psi_{l,k',j,r,i}^{(t)} - \frac{\eta\|\boldsymbol{\xi}_{k',i}\|_2^2}{n_{k'}m}\ell_{k',i}'^{(t)}\sigma'\Big(\Big\langle \mathbf{w}_{k',j,r}^{(t)}, \boldsymbol{\xi}_{k',i}\Big\rangle\Big)\mathbf{1}\{y_{k',i}=j\}\mathbf{1}\{k'=k\},$$

$$\phi_{k,k',j,r,i}^{(t+1)} = \sum_{l=1}^{K} M_{kl}\phi_{l,k',j,r,i}^{(t)} + \frac{\eta\|\boldsymbol{\xi}_{k',i}\|_2^2}{n_{k'}m}\ell_{k',i}'^{(t)}\sigma'\Big(\Big\langle \mathbf{w}_{k',j,r}^{(t)}, \boldsymbol{\xi}_{k',i}\Big\rangle\Big)\mathbf{1}\{y_{k',i}=-j\}\mathbf{1}\{k'=k\}.$$

□

A.3 Finite-time Feature Learning

We use three bounds repeatedly in this proof. β_μ and β_ξ control the initial signal and noise projections, and κ controls the interaction between two distinct training noise vectors. Their values are

$$\beta_\mu = C\sigma_0\mu\sqrt{L}, \quad \beta_\xi = C\sigma_0\sigma_p\sqrt{dL}, \quad \kappa = C\sqrt{L/d}.$$

We choose the initialization upper bound small enough that $\beta_\mu + \beta_\xi \le 1/16$. The filter sets below select an initial positive margin on a signal or a noise vector,

$$\mathcal{R}_{k,j}^{\mu} = \{r : j\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k\rangle \ge \sigma_0\mu/4\}, \quad \mathcal{R}_{k,i}^{\xi} = \{r : \langle \mathbf{w}_{y_{k,i},r}^{(0)}, \boldsymbol{\xi}_{k,i}\rangle \ge \sigma_0\|\boldsymbol{\xi}_{k,i}\|_2/4\}.$$

Lemma A.2 (Noise geometry)**.** *Under the dimension assumption, with probability at least* $1-\delta/4$*,*

$$\begin{aligned}
\frac{\sigma_p^2 d}{4} \le \|\boldsymbol{\xi}_{k,i}\|_2^2 &\le 2\sigma_p^2 d, \\
\frac{|\langle \boldsymbol{\xi}_{k,i}, \boldsymbol{\xi}_{k',i'}\rangle|}{\|\boldsymbol{\xi}_{k',i'}\|_2^2} &\le \kappa, \quad (k,i) \ne (k',i').
\end{aligned}$$

The Gram matrix of the normalized noise vectors has all eigenvalues in $[3/4, 5/4]$.

Proof. Let $r_H = \operatorname{rank}(\mathbf{H})$. Since $\mathbf{H}$ is the orthogonal projector onto the complement of the signal subspace, the data model gives

$$\frac{d}{2} \le r_H \le d, \qquad \boldsymbol{\xi}_{k,i} = \sigma_p \mathbf{H}\mathbf{g}_{k,i}, \quad \mathbf{g}_{k,i} \sim \mathcal{N}(0, \mathbf{I}_d).$$

Thus, after an orthogonal change of coordinates,

$$Z = \frac{\|\boldsymbol{\xi}_{k,i}\|_2^2}{\sigma_p^2} \sim \chi_{r_H}^2.$$

Because $r_H \geq d/2$ and $r_H \leq d$,

$$\{Z < d/4\} \subseteq \{Z < r_H/2\}, \qquad \{Z > 2d\} \subseteq \{Z > 2r_H\},$$

and

$$\mathbb{P}\{Z \leq r/2\} \leq e^{-r/16}, \qquad \mathbb{P}\{Z \geq 2r\} \leq e^{-r/8}.$$

Consequently, for each (k, i),

$$\mathbb{P}\left\{\|\boldsymbol{\xi}_{k,i}\|_2^2 \notin \left[\frac{\sigma_p^2 d}{4}, 2\sigma_p^2 d\right]\right\} \leq e^{-d/32} + e^{-d/16}.$$

There are $n = K\bar{n}$ noise vectors. The dimension and sample-size conditions imply $d \geq CL$, so, after increasing the universal constant in the dimension assumption, a union bound gives an event $\mathcal{E}_{\text{norm}}$ with

$$\mathbb{P}(\mathcal{E}_{\text{norm}}^c) \leq \delta/8$$

on which all the displayed norm bounds hold simultaneously.

We next control the correlations. Fix two distinct observations $a = (k, i)$ and $b = (k', i')$. Conditional on $\boldsymbol{\xi}_b$, independence of the Gaussian noise patches and $\mathbf{H}\boldsymbol{\xi}_b = \boldsymbol{\xi}_b$ give

$$\frac{\langle \boldsymbol{\xi}_a, \boldsymbol{\xi}_b \rangle}{\sigma_p \|\boldsymbol{\xi}_b\|_2} \sim \mathcal{N}(0, 1).$$

On $\mathcal{E}_{\text{norm}}$, $\|\boldsymbol{\xi}_b\|_2 \geq \sigma_p\sqrt{d}/2$. Therefore conditioning only on $\boldsymbol{\xi}_b$ gives

$$\begin{aligned}\mathbb{P}\left\{\frac{|\langle \boldsymbol{\xi}_a, \boldsymbol{\xi}_b \rangle|}{\|\boldsymbol{\xi}_b\|_2^2} > \kappa \,\middle|\, \boldsymbol{\xi}_b\right\} &= \mathbb{P}\left\{|Z| > \frac{\kappa \|\boldsymbol{\xi}_b\|_2}{\sigma_p}\right\} \\ &\leq 2\exp\left(-\frac{\kappa^2 d}{8}\right), \qquad Z \sim \mathcal{N}(0, 1).\end{aligned}$$

Equivalently, after averaging over $\boldsymbol{\xi}_b$,

$$\mathbb{P}\left(\left\{\frac{|\langle \boldsymbol{\xi}_a, \boldsymbol{\xi}_b \rangle|}{\|\boldsymbol{\xi}_b\|_2^2} > \kappa\right\} \cap \left\{\|\boldsymbol{\xi}_b\|_2 \geq \frac{\sigma_p\sqrt{d}}{2}\right\}\right) \leq 2\exp\left(-\frac{\kappa^2 d}{8}\right).$$

Recall that $\kappa = C_\kappa\sqrt{L/d}$, where C_κ is a sufficiently large universal constant. Hence the last probability is at most $2e^{-C_\kappa^2 L/8}$. A union bound over fewer than n^2 ordered pairs gives

$$\mathbb{P}(\mathcal{E}_{\text{cross}}^c \cap \mathcal{E}_{\text{norm}}) \leq \delta/8,$$

where $\mathcal{E}_{\text{cross}}$ denotes the event that the following bound holds simultaneously for every ordered pair of distinct samples,

$$\frac{|\langle \boldsymbol{\xi}_{k,i}, \boldsymbol{\xi}_{k',i'} \rangle|}{\|\boldsymbol{\xi}_{k',i'}\|_2^2} \leq \kappa \qquad ((k, i) \neq (k', i')). \tag{26}$$

The two events therefore hold simultaneously with probability at least $1 - \delta/4$.

It remains to pass from the asymmetric bound in Equation 26 to normalized inner products. Since the ordered-pair union bound also contains the reversed pair, Equation 26 gives both

$$|\langle \boldsymbol{\xi}_a, \boldsymbol{\xi}_b \rangle| \leq \kappa \|\boldsymbol{\xi}_b\|_2^2, \qquad |\langle \boldsymbol{\xi}_a, \boldsymbol{\xi}_b \rangle| \leq \kappa \|\boldsymbol{\xi}_a\|_2^2.$$

Multiplying these inequalities and taking a square root yields

$$\frac{|\langle \boldsymbol{\xi}_a, \boldsymbol{\xi}_b \rangle|}{\|\boldsymbol{\xi}_a\|_2 \|\boldsymbol{\xi}_b\|_2} \leq \kappa.$$

Let $\widetilde{\mathbf{G}}$ be the Gram matrix of the normalized vectors $\boldsymbol{\xi}_a/\|\boldsymbol{\xi}_a\|_2$. Its diagonal entries are one, and every off-diagonal entry has absolute value at most κ. The dimension condition, with the constant chosen above, also gives $n\kappa \leq 1/64$. Thus the sum of the absolute off-diagonal entries in every row is at most $(n-1)\kappa \leq 1/64$. Because $\widetilde{\mathbf{G}}$ is symmetric, Gershgorin's theorem then places every eigenvalue in$[1 - \frac{1}{64}, 1 + \frac{1}{64}] \subseteq [\frac{3}{4}, \frac{5}{4}]$. □

Lemma A.3 (Gaussian initialization and sample counts)**.** *Under the dimension, sample-size, width, and initialization upper bounds in Assumption 4.1, with probability at least $1-\delta/4$, simultaneously for every indicated index,*

$$|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k \rangle| \le \beta_\mu, \qquad |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_{k',i} \rangle| \le \beta_\xi,$$
$$\|\mathbf{H}\mathbf{w}_{j,r}^{(0)}\|_2^2 \le C\sigma_0^2 d, \qquad \frac{\bar{n}}{3} \le \sum_{i=1}^{\bar{n}} \mathbf{1}\{y_{k,i} = j\} \le \frac{2\bar{n}}{3},$$
$$|\mathcal{R}_{k,j}^{\mu}| \ge m/8, \qquad |\mathcal{R}_{k,i}^{\xi}| \ge m/8.$$

Proof. For each signal vector and each filter, Gaussian initialization gives

$$\frac{\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k \rangle}{\sigma_0 \mu} \sim \mathcal{N}(0,1).$$

For a fixed noise vector, conditioning on the training data gives the same argument with $\sigma_0 \|\boldsymbol{\xi}_{k',i}\|_2$. The bound $\mathbb{P}\{|Z| \ge u\} \le 2e^{-u^2/2}$, followed by a union bound over the indicated projections, therefore gives

$$|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k \rangle| \le C\sigma_0 \mu \sqrt{L} = \beta_\mu,$$
$$|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_{k',i} \rangle| \le C\sigma_0 \|\boldsymbol{\xi}_{k',i}\|_2 \sqrt{L} \le \beta_\xi.$$

The last inequality uses the noise-norm bound in Lemma A.2.

Also, $\|\mathbf{H}\mathbf{w}_{j,r}^{(0)}\|_2^2/\sigma_0^2$ has a chi-square distribution. Its upper-tail bound gives the stated norm estimate. The noise-norm failure and these Gaussian failures together are at most $\delta/8$. Each label count is binomial with parameters $(\bar{n}, 1/2)$, so its probability of falling outside $[\bar{n}/3, 2\bar{n}/3]$ is at most $e^{-c\bar{n}}$, after adjusting c.

Finally, for a signal-filter set, every filter belongs to the set with probability

$$\mathbb{P}\left\{ \frac{j\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k \rangle}{\sigma_0 \mu} \ge \frac{1}{4} \right\} = \mathbb{P}\{Z \ge 1/4\} > \frac{1}{3}, \qquad Z \sim \mathcal{N}(0,1).$$

The indicators are independent across r. Thus a Chernoff bound gives $\mathbb{P}\{|\mathcal{R}_{k,j}^{\mu}| < m/8\} \le e^{-cm}$. Conditioning on a nonzero training noise vector gives exactly the same argument for $\mathcal{R}_{k,i}^{\xi}$. □

A.4 Proof of Lemma 4.3

The short horizon is

$$T_0 = \left\lfloor \frac{c_T m}{\eta} \min\left\{ \mu^{-2}, \frac{\bar{n}}{\sigma_p^2 d} \right\} \right\rfloor, \tag{27}$$

where $c_T > 0$ is a sufficiently small universal constant. We will prove the early training and risk certificate in this horizon.

Proof. According to the decomposition of coefficients, we first consider the signal learning by eq. (7):

$$\gamma_{k,k',j,r}^{(t+1)} = \sum_{l=1}^{K} M_{kl} \gamma_{l,k',j,r}^{(t)} - \frac{\eta}{n_k m} \sum_{i=1}^{n_k} \ell_{k,i}'^{(t)} \sigma'(\langle \mathbf{w}_{k,j,r}^{(t)}, y_{k,i} \boldsymbol{\mu}_k \rangle) \|\boldsymbol{\mu}_k\|_2^2 \mathbb{1}(k'=k)$$
$$\overset{(a)}{\le} \sum_{l=1}^{K} M_{kl} \gamma_{l,k',j,r}^{(t)} + \frac{\eta}{m} \|\boldsymbol{\mu}_k\|_2^2,$$

where (a) holds due to $0 \leq -\ell' \leq 1$. Then, taking a telescoping sum over $t = 1, ..., T$ when $k' = k$ yields:

$$\begin{aligned}
\gamma_{k,k,j,r}^{(T+1)} &\leq \sum_{l_1=1}^{K} M_{kl_1} \gamma_{l_1,k,j,r}^{(T)} + \frac{\eta}{m} \|\boldsymbol{\mu}_k\|_2^2 \\
&\leq \sum_{l_1=1}^{K} M_{kl_1} \left(\sum_{l_2=1}^{K} M_{l_1 l_2} \gamma_{l_2,k,j,r}^{(T-1)} + \frac{\eta}{m} \|\boldsymbol{\mu}_k\|_2^2 \mathbb{1}(l_1 = k) \right) + \frac{\eta}{m} \|\boldsymbol{\mu}_k\|_2^2 \\
&= \sum_{l_1,l_2=1}^{K} M_{kl_1} M_{l_1 l_2} \gamma_{l_2,k,j,r}^{(T-1)} + \frac{\eta}{m} \|\boldsymbol{\mu}_k\|_2^2 \left(M_{kk} + 1\right) \\
&\vdots \\
&\leq \underbrace{\sum_{l_1,\ldots,l_T=1}^{K} M_{kl_1} \cdots M_{l_{T-1} l_T} \gamma_{l_T,k,j,r}^{(0)}}_{\text{Vanishes since } \gamma^{(0)}=0} + \frac{\eta}{m} \|\boldsymbol{\mu}_k\|_2^2 \sum_{t=0}^{T} [M^t]_{kk}.
\end{aligned}$$

Since $\gamma_{l_T,k,j,r}^{(0)} = 0$, the first term vanishes, leaving:

$$\gamma_{k,k,j,r}^{(T+1)} \leq \frac{\eta}{m} \|\boldsymbol{\mu}_k\|_2^2 \sum_{t=0}^{T} [M^t]_{kk}, \tag{28}$$

Following the same process, we also have that

$$\Psi_{k,k',i}^{(T)} \leq \frac{\eta \|\boldsymbol{\xi}_{k',i}\|_2^2}{\bar{n}m} \sum_{s=0}^{T-1} [\mathbf{M}^s]_{kk'}. \tag{29}$$

To derive a matching bound on the coefficients, we also need to bound the stage-1 output. Take the inner product of Equation 5 with $\boldsymbol{\mu}_q$. The noise terms vanish by orthogonality, while $0 \leq \langle \boldsymbol{\mu}_{k'}, \boldsymbol{\mu}_q \rangle / \mu^2 \leq 1$. Consequently,

$$\begin{aligned}
|\langle \mathbf{w}_{k,j,r}^{(T)}, \boldsymbol{\mu}_q \rangle| &\leq \beta_\mu + \sum_{k'=1}^{K} \gamma_{k,k',j,r}^{(T)} \\
&\leq \beta_\mu + \frac{\eta \mu^2}{m} \sum_{k'=1}^{K} [\mathbf{A}_T]_{kk'} = \beta_\mu + \frac{\eta \mu^2 T}{m}.
\end{aligned}$$

For $\boldsymbol{\xi}_{q,i}$, the signal terms vanish. Its own noise coefficient appears with multiplier one, while every other coefficient has multiplier at most κ in absolute value by Lemma A.2. Thus

$$\begin{aligned}
|\langle \mathbf{w}_{k,j,r}^{(T)}, \boldsymbol{\xi}_{q,i} \rangle| &\leq \beta_\xi + \Psi_{k,q,i}^{(T)} + \kappa \sum_{(k',i') \neq (q,i)} \Psi_{k,k',i'}^{(T)} \\
&\leq \beta_\xi + \frac{2\eta\sigma_p^2 d}{\bar{n}m} [\mathbf{A}_T]_{kq} + \frac{2\kappa\eta\sigma_p^2 dT}{m} \\
&\leq \beta_\xi + \frac{2\eta\sigma_p^2 dT}{\bar{n}m} + \frac{2\kappa\eta\sigma_p^2 dT}{m}.
\end{aligned}$$

If we take T_0 into the inequalities above, then each branch output is in $[0, 2]$. The difference of the two branch outputs therefore satisfies

$$|f_k(\mathbf{W}_k^{(T)}, \mathbf{x}_{k,i})| \leq 2, \qquad C_1 \leq -\ell_{k,i}'^{(T)} \leq 1, \qquad C_1 = (1 + e^2)^{-1}.$$

For $r \in \mathcal{R}_{k',j}^{\mu}$, the exact decomposition yields

$$j \langle \mathbf{w}_{k',j,r}^{(t)}, \boldsymbol{\mu}_{k'} \rangle = j \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_{k'} \rangle + \sum_{l=1}^{K} \gamma_{k',l,j,r}^{(t)} \frac{\langle \boldsymbol{\mu}_{k'}, \boldsymbol{\mu}_l \rangle}{\mu^2} > 0.$$

where the initial term is strictly positive because $\sigma_0 > 0$, and every learned term is nonnegative. Combining this with $-\ell_{k,i}'^{(t)} \geq C_1$, we have:

$$\begin{aligned}\gamma_{k,k,j,r}^{(t+1)} &= \sum_{l=1}^{K} \mathbf{M}_{kl}\gamma_{l,k,j,r}^{(t)} - \frac{\eta}{n_k m} \cdot \sum_{i=1}^{n_k} \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{k,j,r}^{(t)}, y_i \boldsymbol{\mu}_k \rangle) \cdot \|\boldsymbol{\mu}_k\|_2^2 \\ &\geq \sum_{l=1}^{K} \mathbf{M}_{kl}\gamma_{l,k,j,r}^{(t)} + \frac{\eta}{3m} C_1 \cdot \|\boldsymbol{\mu}_k\|_2^2.\end{aligned}$$

Similar to the upper bound, we obtain

$$\gamma_{k,k,j,r}^{(T+1)} \geq \frac{\eta C_1}{3m} \|\boldsymbol{\mu}_k\|_2^2 \sum_{t=0}^{T} [\mathbf{M}^t]_{kk}.$$

Following the same bounding technique and telescoping sum over the gradient injection at node k', the upper bound is governed by the transition probability from k' to k:

$$\frac{\eta C_1}{3m} \|\boldsymbol{\mu}_{k'}\|_2^2 \sum_{t=0}^{T-1} [\mathbf{M}^t]_{kk'} \leq \gamma_{k,k',j,r}^{(T)} \leq \frac{\eta}{m} \|\boldsymbol{\mu}_{k'}\|_2^2 \sum_{t=0}^{T-1} [\mathbf{M}^t]_{kk'}. \tag{30}$$

Taking the signal inner product in Equation 5 gives

$$j\langle \mathbf{w}_{k,j,r}^{(T)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k \rangle = \sum_{k'=1}^{K} \gamma_{k,k',j,r}^{(T)} \frac{\langle \boldsymbol{\mu}_k, \boldsymbol{\mu}_{k'} \rangle}{\mu^2}.$$

Under the stated signal geometry, a positive similarity implies $\boldsymbol{\mu}_{k'} = \boldsymbol{\mu}_k$, and hence $\mathcal{R}_{k',j}^{\mu} = \mathcal{R}_{k,j}^{\mu}$. Thus, for every $r \in \mathcal{R}_{k,j}^{\mu}$, multiplying Equation 30 by the corresponding similarities and summing proves

$$\frac{\eta C_1 \|\boldsymbol{\mu}_k\|_2^2}{3m} S_k(T) \leq j\langle \mathbf{w}_{k,j,r}^{(T)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k \rangle \leq \frac{\eta \|\boldsymbol{\mu}_k\|_2^2}{m} S_k(T). \tag{31}$$

The initialization event supplies at least $m/8$ such fixed filters for each (k, j), uniformly over the entire window.

Applying $\mathbf{H}$ to Equation 5 removes the learned signal component and leaves

$$\mathbf{H}(\mathbf{w}_{k,j,r}^{(T)} - \mathbf{w}_{j,r}^{(0)}) = \sum_{k',i} \frac{\rho_{k,k',j,r,i}^{(T)}}{\|\boldsymbol{\xi}_{k',i}\|_2} \frac{\boldsymbol{\xi}_{k',i}}{\|\boldsymbol{\xi}_{k',i}\|_2}.$$

Using $|\rho| \leq \Psi$ and Lemma A.2, we obtain

$$\begin{aligned}\|\mathbf{H}(\mathbf{w}_{k,j,r}^{(T)} - \mathbf{w}_{j,r}^{(0)})\|_2^2 &\leq \frac{5}{4} \sum_{k',i} \frac{(\rho_{k,k',j,r,i}^{(T)})^2}{\|\boldsymbol{\xi}_{k',i}\|_2^2} \leq \frac{5\eta^2}{4\bar{n}^2 m^2} \sum_{k'=1}^{K} [\mathbf{A}_T]_{kk'}^2 \sum_{i=1}^{\bar{n}} \|\boldsymbol{\xi}_{k',i}\|_2^2 \\ &\leq \frac{5\eta^2 \sigma_p^2 d}{2\bar{n}m^2} \sum_{k'=1}^{K} [\mathbf{A}_T]_{kk'}^2 = \frac{5\eta^2 \sigma_p^2 d}{2\bar{n}m^2} Q_k(T),\end{aligned}$$

where the squared norm is a quadratic form in the Gram matrix of the normalized noise vectors. By theorem A.2, it is at most $5/4$ times the squared norm of its coefficient vector.

Adding initialization by $\|\mathbf{a} + \mathbf{b}\|_2^2 \leq 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2$ gives

$$\|\mathbf{H}\mathbf{w}_{k,j,r}^{(T)}\|_2^2 \leq C \left[\sigma_0^2 d + \frac{\eta^2 \sigma_p^2 d}{\bar{n}m^2} Q_k(T) \right]. \tag{32}$$

In fact, this energy upper bound holds for every $T \geq 0$, since its coefficient upper bound is unconditional. Only the signal lower bound and the bounded-logit conclusion use $T \leq T_0$. □

A.5 EARLY GENERALIZATION CERTIFICATE

We first turn a lower bound on the learned signal and an upper bound on the noise projection into a test-error bound. We then substitute the finite-time coefficient estimates to obtain Theorem 4.4.

Lemma A.4 (Test error from the actual coefficients)**.** *Condition on the initial geometry in Lemma A.3. For a receiving client k, a test distribution $\mathcal{D}_q$, and $t \geq 0$, let $s_{kq}(t)$ be a lower bound on the signed signal response and let $v_k(t)$ bound every filter's noise projection:*

$$s_{kq}(t) = \min_{y\in\{\pm 1\}} \frac{1}{m} \sum_{r\in\mathcal{R}^{\mu}_{q,y}} \sum_{l=1}^{K} \gamma^{(t)}_{k,l,y,r} \frac{\langle \boldsymbol{\mu}_q, \boldsymbol{\mu}_l \rangle}{\mu^2} - \beta_\mu, \tag{33}$$

$$v_k(t) = \max_{j,r} \|\mathbf{H}\mathbf{w}^{(t)}_{k,j,r}\|_2. \tag{34}$$

Then, simultaneously for every k, q and every integer $t \geq 0$ with $v_k(t) > 0$,

$$\mathcal{R}_q(\mathbf{W}^{(t)}_k) \leq \exp\left[-\frac{1}{2}\left(\frac{s_{kq}(t)}{\sigma_p v_k(t)} - \frac{1}{\sqrt{2\pi}}\right)^2_+\right]. \tag{35}$$

If $v_k(t) = 0$, the risk is zero when $s_{kq}(t) > 0$, and is at most one otherwise. This coefficient-based conclusion does not require the small-step or mixing-time conditions: any positive step size is allowed, provided the stated decomposition and its coefficient signs hold. Its numerical usefulness depends on the resulting margin and noise bounds.

Proof. Fix the training data and initialization on the stated event, and consider a fresh test point with label y and signal $y\boldsymbol{\mu}_q$.

Step 1: separate the two signal branches. The decomposition therefore gives

$$\langle \mathbf{w}^{(t)}_{k,y,r}, y\boldsymbol{\mu}_q \rangle = y\langle \mathbf{w}^{(0)}_{y,r}, \boldsymbol{\mu}_q \rangle + \sum_{l=1}^{K} \gamma^{(t)}_{k,l,y,r} \frac{\langle \boldsymbol{\mu}_q, \boldsymbol{\mu}_l \rangle}{\mu^2},$$

$$\langle \mathbf{w}^{(t)}_{k,-y,r}, y\boldsymbol{\mu}_q \rangle = y\langle \mathbf{w}^{(0)}_{-y,r}, \boldsymbol{\mu}_q \rangle - \sum_{l=1}^{K} \gamma^{(t)}_{k,l,-y,r} \frac{\langle \boldsymbol{\mu}_q, \boldsymbol{\mu}_l \rangle}{\mu^2}.$$

Every term in the coefficient sums is nonnegative. For $r \in \mathcal{R}^{\mu}_{q,y}$, the first initial inner product is also positive, so the correct-branch ReLU output is at least its coefficient sum. The opposite-branch preactivation is at most β_μ, so its ReLU output is as well. Keeping the initially aligned filters in the correct branch and all filters in the opposite branch yields

$$\frac{1}{m}\sum_{r=1}^{m}\left[\sigma\Big(\langle \mathbf{w}^{(t)}_{k,y,r}, y\boldsymbol{\mu}_q \rangle\Big) - \sigma\Big(\langle \mathbf{w}^{(t)}_{k,-y,r}, y\boldsymbol{\mu}_q \rangle\Big)\right] \geq s_{kq}(t).$$

Step 2: control the opposite branch on fresh noise. Write the fresh noise as $\boldsymbol{\xi} = \sigma_p \mathbf{H}\mathbf{g}$, where $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_d)$ is independent of training. Its opposite-branch response is

$$N_{-y}(\mathbf{g}) = \frac{1}{m}\sum_{r=1}^{m} \sigma\Big(\sigma_p \langle \mathbf{H}\mathbf{w}^{(t)}_{k,-y,r}, \mathbf{g} \rangle\Big).$$

A centered Gaussian of standard deviation a has expected positive part $a/\sqrt{2\pi}$. Consequently,

$$\mathbb{E}N_{-y}(\mathbf{g}) = \frac{\sigma_p}{m\sqrt{2\pi}} \sum_{r=1}^{m} \|\mathbf{H}\mathbf{w}^{(t)}_{k,-y,r}\|_2 \leq \frac{\sigma_p v_k(t)}{\sqrt{2\pi}}.$$

Since ReLU is 1-Lipschitz, Cauchy–Schwarz also gives

$$\begin{aligned} |N_{-y}(\mathbf{g}) - N_{-y}(\mathbf{g}')| &\leq \frac{\sigma_p}{m}\sum_{r=1}^{m} \left|\langle \mathbf{H}\mathbf{w}^{(t)}_{k,-y,r}, \mathbf{g} - \mathbf{g}' \rangle\right| \\ &\leq \sigma_p v_k(t) \|\mathbf{g} - \mathbf{g}'\|_2. \end{aligned}$$

Thus Gaussian concentration implies, for $u \geq 0$ and $v_k(t) > 0$,

$$\mathbb{P}\left\{ N_{-y}(\mathbf{g}) \geq \frac{\sigma_p v_k(t)}{\sqrt{2\pi}} + u \right\} \leq \exp\left[-\frac{u^2}{2\sigma_p^2 v_k(t)^2} \right]. \tag{36}$$

Step 3: combine signal and noise. The correct-branch noise response is nonnegative, so the signed test output satisfies

$$y f(\mathbf{W}_k^{(t)}, \mathbf{x}) \geq s_{kq}(t) - N_{-y}(\mathbf{g}).$$

If $s_{kq}(t) > \sigma_p v_k(t)/\sqrt{2\pi}$, an error requires the event in Equation 36 with $u = s_{kq}(t) - \sigma_p v_k(t)/\sqrt{2\pi}$. Otherwise the right-hand side of Equation 35 is one. The same bound holds for both labels, so averaging over y proves Equation 35.

If $v_k(t) = 0$, all noise responses vanish and the asserted zero-noise conclusion follows from the signal bound. This argument applies to every fixed trajectory and time on the same geometry. □

A.6 Proof of Theorem 4.4

Proof of the upper bound in Theorem 4.4. This follows from the geometry and Lemma 4.3. For every $1 \leq T \leq T_0$, the coefficient estimates of Equation 33 and give

$$s_{kk}(T) \geq c_s \frac{\eta \mu^2}{m} S_k(T) - \beta_\mu, \tag{37}$$

$$v_k(T)^2 \leq C_v \left[\sigma_0^2 d + \frac{\eta^2 \sigma_p^2 d}{\bar{n} m^2} Q_k(T) \right], \tag{38}$$

where $c_s > 0$ is sufficiently small and $C_v > 0$ sufficiently large.

To see the two inequalities, each initially aligned filter contributes at least $c\eta\mu^2 S_k(T)/m$ to the sum in Equation 33, and at least a fixed positive fraction of m such filters exist. Averaging over these filters gives the first term in Equation 37. The second inequality is precisely the noise-energy estimate, maximized over the filters.

Substitution into Equation 35 now gives the explicit bound

$$\mathcal{R}_k(\mathbf{W}_k^{(T)}) \leq \exp\left[-\frac{1}{2} \left(\frac{c_s \frac{\eta\mu^2}{m} S_k(T) - \beta_\mu}{\sigma_p \sqrt{C_v \left[\sigma_0^2 d + \frac{\eta^2 \sigma_p^2 d}{\bar{n} m^2} Q_k(T) \right]}} - \frac{1}{\sqrt{2\pi}} \right)_+^2 \right]. \tag{39}$$

For the pure topology bound, assume

$$\beta_\mu \leq \frac{c_s}{2} \frac{\eta \mu^2}{m} S_k(T), \tag{40}$$

$$\sigma_0^2 d \leq \frac{\eta^2 \sigma_p^2 d}{\bar{n} m^2} Q_k(T). \tag{41}$$

The first condition follows from Equation 11 by choosing c_{init} sufficiently small relative to the universal constant in β_μ.

These conditions simplify Equation 37–Equation 38 to

$$s_{kk}(T) \geq \frac{c_s}{2} \frac{\eta \mu^2}{m} S_k(T) > 0, \qquad v_k(T)^2 \leq \frac{2 C_v \eta^2 \sigma_p^2 d}{\bar{n} m^2} Q_k(T).$$

If $v_k(T) = 0$, the risk is zero by the preceding lemma. Otherwise, dividing these estimates shows that,

$$\left(\frac{s_{kk}(T)}{\sigma_p v_k(T)} \right)^2 \geq \frac{(c_s^2/4)(\eta^2 \mu^4 / m^2) S_k(T)^2}{2 C_v \sigma_p^2 (\eta^2 \sigma_p^2 d / (\bar{n} m^2)) Q_k(T)}$$
$$= \frac{c_s^2}{8 C_v} \frac{\bar{n} \mu^4}{\sigma_p^4 d} \frac{S_k(T)^2}{Q_k(T)}.$$

If $\bar{n}\mu^4 S_k(T)^2/(\sigma_p^4 d Q_k(T))$ is larger than a sufficiently large universal constant, this also proves

$$\frac{s_{kk}(T)}{\sigma_p v_k(T)} \geq \frac{2}{\sqrt{2\pi}}. \tag{42}$$

The inequality $\frac{1}{2}(x - 1/\sqrt{2\pi})^2 \geq x^2/8$ for $x \geq 2/\sqrt{2\pi}$ then gives

$$\mathcal{R}_k(\mathbf{W}_k^{(T)}) \leq \exp\left[-c\frac{\bar{n}\mu^4}{\sigma_p^4 d}\frac{S_k(T)^2}{Q_k(T)}\right]. \tag{43}$$

□

A.7 Proof of the spectral characterization

Proof of Proposition 4.5. Set $\mathbf{J} = \mathbf{1}\mathbf{1}^\top/K$. The row of $T\mathbf{J}$ assigns weight T/K to every source client, whereas $\mathbf{A}_T - T\mathbf{J}$ records the deviation from these equal weights.

Let

$$\mathbf{c}_k = \left(\frac{\langle \boldsymbol{\mu}_k, \boldsymbol{\mu}_1\rangle}{\mu^2}, \ldots, \frac{\langle \boldsymbol{\mu}_k, \boldsymbol{\mu}_K\rangle}{\mu^2}\right)^\top.$$

Under the stated signal geometry, $\mathbf{c}_k$ indicates the clients sharing client k's signal. Hence

$$\mathbf{1}^\top \mathbf{c}_k = \|\mathbf{c}_k\|_2^2 = \chi_k, \qquad \left\|\mathbf{c}_k - \frac{\chi_k}{K}\mathbf{1}\right\|_2^2 = \chi_k\left(1 - \frac{\chi_k}{K}\right).$$

To bound that deviation, write

$$\mathbf{B}_T = \mathbf{A}_T - T\mathbf{J} = \sum_{s=0}^{T-1}(\mathbf{M}^s - \mathbf{J}).$$

Since $\mathbf{M}$ is symmetric, positive semidefinite, and doubly stochastic, it has an orthonormal eigenbasis with $\mathbf{u}_1 = \mathbf{1}/\sqrt{K}$ and $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_K \geq 0$. Consequently,

$$\mathbf{B}_T = \sum_{l=2}^{K}\left(\sum_{s=0}^{T-1}\lambda_l^s\right)\mathbf{u}_l\mathbf{u}_l^\top = \sum_{l=2}^{K}\frac{1-\lambda_l^T}{1-\lambda_l}\mathbf{u}_l\mathbf{u}_l^\top.$$

Every residual eigenvector is orthogonal to $\mathbf{1}$, and every geometric sum is at most the one for λ_2. Thus

$$\|\mathbf{B}_T\|_2 \leq \sum_{s=0}^{T-1}\lambda_2^s = \frac{1-\lambda_2^T}{1-\lambda_2}, \qquad \mathbf{B}_T\mathbf{1} = 0. \tag{44}$$

Since $S_k(T) = \mathbf{e}_k^\top \mathbf{A}_T \mathbf{c}_k$,

$$\begin{aligned} S_k(T) - \frac{T\chi_k}{K} &= \mathbf{e}_k^\top \mathbf{B}_T \mathbf{c}_k \\ &= \mathbf{e}_k^\top \mathbf{B}_T\left(\mathbf{c}_k - \frac{\chi_k}{K}\mathbf{1}\right). \end{aligned}$$

The last equality uses $\mathbf{B}_T\mathbf{1} = 0$. Taking absolute values and using $\|\mathbf{e}_k\|_2 = 1$ gives

$$\begin{aligned} \left|S_k(T) - \frac{T\chi_k}{K}\right| &\leq \|\mathbf{B}_T\|_2 \left\|\mathbf{c}_k - \frac{\chi_k}{K}\mathbf{1}\right\|_2 \\ &\leq \frac{1-\lambda_2^T}{1-\lambda_2}\sqrt{\chi_k(1-\chi_k/K)}, \end{aligned}$$

which is Equation 13.

The stationary row $T\mathbf{1}^\top/K$ is orthogonal to the residual row because $\mathbf{B}_T\mathbf{1} = 0$. In particular,

$$\begin{aligned} Q_k(T) = \|\mathbf{e}_k^\top \mathbf{A}_T\|_2^2 &= \left\|\frac{T}{K}\mathbf{1}^\top + \mathbf{e}_k^\top \mathbf{B}_T\right\|_2^2 \\ &= \frac{T^2}{K} + \frac{2T}{K}\mathbf{e}_k^\top \mathbf{B}_T\mathbf{1} + \|\mathbf{e}_k^\top \mathbf{B}_T\|_2^2. \end{aligned}$$

The cross term vanishes, leaving the exact identity

$$Q_k(T) = \frac{T^2}{K} + \|\mathbf{e}_k^\top \mathbf{B}_T\|_2^2. \tag{45}$$

Using the orthogonality of the eigenvectors gives the equivalent spectral expression

$$Q_k(T) = \frac{T^2}{K} + \sum_{l=2}^{K} (\mathbf{u}_l)_k^2 \left(\frac{1-\lambda_l^T}{1-\lambda_l}\right)^2.$$

Since $\|\mathbf{B}_T\|_2 \le 1/(1-\lambda_2)$ uniformly in T, the signal residual divided by T and the noise residual divided by T^2 both vanish. Hence

$$\frac{S_k(T)}{T} \longrightarrow \frac{\chi_k}{K}, \qquad \frac{Q_k(T)}{T^2} \longrightarrow \frac{1}{K}.$$

Since the second limit is positive,

$$\frac{S_k(T)^2}{Q_k(T)} = \frac{(S_k(T)/T)^2}{Q_k(T)/T^2} \longrightarrow \frac{(\chi_k/K)^2}{1/K} = \frac{\chi_k^2}{K},$$

proving Equation 14.

Evaluate the four-cycle example. Here $\mathbf{A}_2 = \mathbf{I} + \mathbf{M}$, so the receiving row at node 1 is

$$\mathbf{e}_1^\top \mathbf{A}_2 = \left(\frac{3}{2}, \frac{1}{4}, 0, \frac{1}{4}\right), Q_1(2) = \left(\frac{3}{2}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2 = \frac{19}{8}.$$

For placement (a, a, b, b), the matching signal is at nodes 1 and 2, giving $S_1(2) = 3/2 + 1/4 = 7/4$. For placement (a, b, a, b) it is at nodes 1 and 3, giving $S_1(2) = 3/2 + 0 = 3/2$. Therefore the respective ratios are

$$\frac{(7/4)^2}{19/8} = \frac{49}{38}, \qquad \frac{(3/2)^2}{19/8} = \frac{18}{19},$$

as claimed. □

A.8 Coefficient bound and Training Convergence

Fix a target accuracy $\epsilon \in (0, 1)$ and put

$$\alpha_\epsilon = 512 \log \frac{2K}{\epsilon}, \qquad a_\epsilon = 16\big(\log(8/\epsilon) + 1\big).$$

The first is the coefficient upper bound established in Lemma A.7; the second sets the amplitude of the reference solution. The fitting horizon is

$$T_\epsilon = \left\lceil \frac{32 m a_\epsilon^2 K n}{\eta \epsilon \sigma_p^2 d} \right\rceil. \tag{46}$$

The common conditions at the start of the appendix still hold. The fitting argument additionally requires

$$d \ge C n^2 L^2 \alpha_\epsilon^2, \tag{47}$$

$$\sigma_0 \ge \frac{C n \sqrt{L}\, \alpha_\epsilon}{\sigma_p d}. \tag{48}$$

For the fitted signal lower bound, use

$$T_1 = \left\lfloor \frac{c_T K \bar{n} m}{\eta \sigma_p^2 d} \right\rfloor, \tag{49}$$

$$\eta \le \frac{c_\eta \bar{n} m (1-\lambda_2)}{\sigma_p^2 d}, \qquad c_\eta \le c_T/8. \tag{50}$$

Up to universal constants and floor effects,

$$T_0 = \Theta\left(\frac{\bar{n}m}{\eta\sigma_p^2 d}\right), \qquad T_1 = \Theta\left(\frac{K\bar{n}m}{\eta\sigma_p^2 d}\right) = \Theta(KT_0).$$

The coefficient bound used in early training horizon controls the logits only on the short window, while the fitting argument needs K more iterations for each source coefficient to be spread over the network and reach its useful scale. The additional stepsize bound ensures enough mixing during this window: using $\lfloor x \rfloor \geq x - 1$ gives

$$T_1(1-\lambda_2) \geq \frac{c_T K\bar{n}m(1-\lambda_2)}{\eta\sigma_p^2 d} - (1-\lambda_2) \geq 8K - (1-\lambda_2) \geq 4K.$$

Lemma A.5 (Noise inner products)**.** *Under the assumption of Lemmas A.3–A.2, the decomposition gives*

$$\left|\left\langle \mathbf{w}_{k,j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_{q,i} \right\rangle - \rho_{k,q,j,r,i}^{(t)}\right| \leq \kappa \sum_{(l,h)\neq(q,i)} \Psi_{k,l,h}^{(t)}. \tag{51}$$

In particular, if all noise coefficients have magnitude at most A, the right-hand side is at most $n\kappa A$. For the local sample $q = k$, the own-sample coefficient equals $\psi_{k,k,j,r,i}^{(t)}$ when $y_{k,i} = j$, and $\phi_{k,k,j,r,i}^{(t)}$ when $y_{k,i} = -j$.

Proof. Fix the receiver k and the training noise vector $\boldsymbol{\xi}_{q,i}$. Taking the inner product of the full weight decomposition with this vector gives

$$\begin{aligned}\left\langle \mathbf{w}_{k,j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_{q,i} \right\rangle &= j\sum_{l=1}^{K} \gamma_{k,l,j,r}^{(t)} \frac{\langle \boldsymbol{\mu}_l, \boldsymbol{\xi}_{q,i}\rangle}{\|\boldsymbol{\mu}_l\|_2^2} + \sum_{l=1}^{K}\sum_{h=1}^{\bar{n}} \rho_{k,l,j,r,h}^{(t)} \frac{\langle \boldsymbol{\xi}_{l,h}, \boldsymbol{\xi}_{q,i}\rangle}{\|\boldsymbol{\xi}_{l,h}\|_2^2} \\ &= \rho_{k,q,j,r,i}^{(t)} + \sum_{(l,h)\neq(q,i)} \rho_{k,l,j,r,h}^{(t)} \frac{\langle \boldsymbol{\xi}_{l,h}, \boldsymbol{\xi}_{q,i}\rangle}{\|\boldsymbol{\xi}_{l,h}\|_2^2}.\end{aligned}$$

The signal sum vanishes by signal–noise orthogonality. In the noise sum, the term $(l,h) = (q,i)$ is exactly the own-sample coefficient. All remaining terms involve distinct noise vectors. Thus

$$\left|\left\langle \mathbf{w}_{k,j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_{q,i} \right\rangle - \rho_{k,q,j,r,i}^{(t)}\right| \leq \sum_{(l,h)\neq(q,i)} |\rho_{k,l,j,r,h}^{(t)}| \frac{|\langle \boldsymbol{\xi}_{l,h}, \boldsymbol{\xi}_{q,i}\rangle|}{\|\boldsymbol{\xi}_{l,h}\|_2^2} \leq \kappa \sum_{(l,h)\neq(q,i)} \Psi_{k,l,h}^{(t)}.$$

where the first inequality is the triangle inequality and the second uses Lemma A.2 and the definition of Ψ. □

Lemma A.6 (Signal inner products)**.** *For every $t \geq 0$, $k \in [K]$, $j \in \{-1, +1\}$, and $r \in [m]$,*

$$\left\langle \mathbf{w}_{k,j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k \right\rangle = j\sum_{q=1}^{K} \gamma_{k,q,j,r}^{(t)} \frac{\langle \boldsymbol{\mu}_q, \boldsymbol{\mu}_k\rangle}{\|\boldsymbol{\mu}_q\|_2^2}. \tag{52}$$

Proof. Substituting the decomposition gives

$$\left\langle \mathbf{w}_{k,j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k \right\rangle = j\sum_{q=1}^{K} \gamma_{k,q,j,r}^{(t)} \frac{\langle \boldsymbol{\mu}_q, \boldsymbol{\mu}_k\rangle}{\|\boldsymbol{\mu}_q\|_2^2} + \sum_{q=1}^{K}\sum_{i=1}^{\bar{n}} \rho_{k,q,j,r,i}^{(t)} \frac{\langle \boldsymbol{\xi}_{q,i}, \boldsymbol{\mu}_k\rangle}{\|\boldsymbol{\xi}_{q,i}\|_2^2}.$$

Every term in the second sum is zero. This proves Equation 52; in particular, multiplying by j leaves a sum of nonnegative signal coefficients and nonnegative signal similarities. □

Lemma A.7. *Under the assumptions of Lemmas A.2–A.3, for all $0 \leq t \leq T_\epsilon$,*

$$0 \leq \psi_{k,q,j,r,i}^{(t)} \leq \Psi_{k,q,i}^{(t)} \leq \alpha_\epsilon, \qquad -\alpha_\epsilon \leq -\Psi_{k,q,i}^{(t)} \leq \phi_{k,q,j,r,i}^{(t)} \leq 0, \tag{53}$$

$$0 \leq \gamma_{k,q,j,r}^{(t)} \leq 4\bar{n}\frac{\mu^2}{\sigma_p^2 d}\alpha_\epsilon. \tag{54}$$

for all $k \in [K]$, $r \in [m]$, $j \in \{\pm 1\}$, and $i \in [\bar{n}]$. Consequently, the aggregated learned signal satisfies

$$0 \le j \left\langle \mathbf{w}_{k,j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k \right\rangle \le 4\bar{n}\chi_k \frac{\mu^2}{\sigma_p^2 d}\alpha_\epsilon. \tag{55}$$

Moreover, for every $r \in \mathcal{R}_{q,i}^{\xi}$,

$$\left\langle \mathbf{w}_{k,y_{q,i},r}^{(t)}, \boldsymbol{\xi}_{q,i} \right\rangle \ge \frac{\sigma_0 \|\boldsymbol{\xi}_{q,i}\|_2}{8} > 0, \tag{56}$$

$$\psi_{k,q,y_{q,i},r,i}^{(t)} = \Psi_{k,q,i}^{(t)}, \tag{57}$$

where $\alpha_\epsilon = O(\log(2K/\epsilon))$.

Proof. To simplify this proof, we define

$$h_{\max} = \frac{2\eta\sigma_p^2 d}{\bar{n}m}, \qquad H_t = \max_{0\le s\le t} \max_{k,q,i} \Psi_{k,q,i}^{(s)}.$$

The first bounds any sample's gradient increment to a noise coefficient, since $\|\boldsymbol{\xi}_{q,i}\|_2^2 \le 2\sigma_p^2 d$ and $0 < -\ell' \le 1$. The second is the historical maximum. The step-size assumption gives $h_{\max} \le \min_q M_{qq} \le 1$. The dimension and initialization conditions imply

$$n\kappa \le \frac{1}{64}, \qquad 2\beta_\mu + 2\beta_\xi < 1, \qquad n\kappa\alpha_\epsilon \le \frac{\sigma_0 \min_{q,i} \|\boldsymbol{\xi}_{q,i}\|_2}{8}. \tag{58}$$

Assume the induction claims hold through time t. The coefficient recursions and $0 \le \sigma' \le 1$ give

$$\Psi_{k,q,i}^{(t+1)} \le \sum_{l=1}^{K} M_{kl}\Psi_{l,q,i}^{(t)} + \mathbf{1}\{k=q\}\frac{\eta\|\boldsymbol{\xi}_{q,i}\|_2^2}{\bar{n}m}(-\ell_{q,i}'^{(t)}).$$

To see explicitly why this also controls the opposite branch, take $j = -y_{q,i}$. If its source noise gate is closed, then

$$-\phi_{k,q,j,r,i}^{(t+1)} = \sum_l M_{kl}(-\phi_{l,q,j,r,i}^{(t)}) \le \sum_l M_{kl}\Psi_{l,q,i}^{(t)}.$$

If its source gate is open, then

$$\begin{aligned} -\phi_{k,q,j,r,i}^{(t+1)} &= \sum_l M_{kl}(-\phi_{l,q,j,r,i}^{(t)}) + \mathbf{1}\{k=q\}\frac{\eta\|\boldsymbol{\xi}_{q,i}\|_2^2}{\bar{n}m}(-\ell_{q,i}'^{(t)}) \\ &\le \sum_l M_{kl}\Psi_{l,q,i}^{(t)} + \mathbf{1}\{k=q\}\frac{\eta\|\boldsymbol{\xi}_{q,i}\|_2^2}{\bar{n}m}(-\ell_{q,i}'^{(t)}). \end{aligned}$$

Thus neither case can exceed the displayed upper bound. ψ is similar. For every fixed $r \in \mathcal{R}_{q,i}^{\xi}$, its matching coefficient attains this upper bound: the source activation is one, and its coefficient at every receiver equals Ψ at time t. Thus

$$\begin{aligned} \Psi_{k,q,i}^{(t+1)} &= \sum_{l=1}^{K} M_{kl}\Psi_{l,q,i}^{(t)} + \mathbf{1}\{k=q\}\frac{\eta\|\boldsymbol{\xi}_{q,i}\|_2^2}{\bar{n}m}(-\ell_{q,i}'^{(t)}) \\ &= \psi_{k,q,y_{q,i},r,i}^{(t+1)}. \end{aligned} \tag{59}$$

In particular, this equality follows from a fixed set of good filters; it does not interchange a maximum and a sum.

At the source client q, the own-sample and cross-sample decomposition in Lemma A.5 gives, for every $r \in \mathcal{R}_{q,i}^{\xi}$,

$$\langle \mathbf{w}_{q,y_{q,i},r}^{(t)}, \boldsymbol{\xi}_{q,i} \rangle \ge \Psi_{q,q,i}^{(t)} - \beta_\xi - n\kappa H_t.$$

There are at least $m/8$ such filters. On the opposite branch, the own-sample coefficient is nonpositive, so every filter satisfies

$$\langle \mathbf{w}_{q,-y_{q,i},r}^{(t)}, \boldsymbol{\xi}_{q,i} \rangle \le \phi_{q,q,-y_{q,i},r,i}^{(t)} + \beta_\xi + n\kappa H_t \le \beta_\xi + n\kappa H_t.$$

For the signal patch $y_{q,i}\boldsymbol{\mu}_q$, Equation 52 shows that learning increases the matching-branch preactivation and decreases the opposite-branch preactivation. Since ReLU is nondecreasing, their signed difference is at least its initialized value, which is at least $-2\beta_\mu$. Adding the signal and noise contributions to the network output yields

$$\begin{aligned} y_{q,i} f_q(\mathbf{W}_q^{(t)}, \mathbf{x}_{q,i}) &\geq \tfrac{1}{8}(\Psi_{q,q,i}^{(t)} - \beta_\xi - n\kappa H_t)_+ - \beta_\xi - n\kappa H_t - 2\beta_\mu \\ &\geq \tfrac{1}{8}\Psi_{q,q,i}^{(t)} - 2n\kappa H_t - (2\beta_\mu + 2\beta_\xi). \end{aligned} \tag{60}$$

The second line uses $(u)_+ \geq u$ and $1 + 1/8 \leq 2$.

For a receiver $k \neq q$, the next coefficient is a convex combination and cannot exceed H_t. At the source,

$$\begin{aligned} \Psi_{q,q,i}^{(t+1)} &\leq M_{qq}\Psi_{q,q,i}^{(t)} + (1 - M_{qq})H_t + h_{\max}(-\ell_{q,i}'^{(t)}) \\ &= H_t - M_{qq}(H_t - \Psi_{q,q,i}^{(t)}) + h_{\max}(-\ell_{q,i}'^{(t)}). \end{aligned}$$

If $\Psi_{q,q,i}^{(t+1)} > H_t$, then $H_t - \Psi_{q,q,i}^{(t)} < h_{\max}/M_{qq} \leq 1$. Thus a source creating a new record must already have a coefficient within one unit of the previous record. Substituting $\Psi_{q,q,i}^{(t)} > H_t - 1$ into Equation 60 gives

$$y_{q,i} f_q(\mathbf{W}_q^{(t)}, \mathbf{x}_{q,i}) \geq \left(\frac{1}{8} - 2n\kappa\right) H_t - \frac{1}{8} - 1 \geq \frac{H_t}{16} - \frac{9}{8}.$$

The bounds in Equation 58 justify both inequalities. The logistic derivative therefore satisfies, at any source that creates a new record,

$$-\ell_{q,i}'^{(t)} = \frac{1}{1 + \exp\{y_{q,i} f_q(\mathbf{W}_q^{(t)}, \mathbf{x}_{q,i})\}} \leq \exp(9/8 - H_t/16).$$

If there is no new record, $H_{t+1} = H_t$. Otherwise choose a source attaining H_{t+1} and apply the last bound. In either case,

$$0 \leq H_{t+1} - H_t \leq h_{\max} e^{9/8 - H_t/16}, \qquad H_{t+1} - H_t \leq h_{\max} \leq 1.$$

To sum this nonlinear recurrence, exponentiate the historical maximum. The inequality $e^x - 1 \leq xe^x$ gives

$$\begin{aligned} e^{H_{t+1}/16} - e^{H_t/16} &= e^{H_t/16}\left(e^{(H_{t+1} - H_t)/16} - 1\right) \\ &\leq e^{H_t/16}\frac{H_{t+1} - H_t}{16} e^{(H_{t+1} - H_t)/16} \\ &\leq \frac{e^{19/16}}{16} h_{\max}. \end{aligned}$$

The last line uses the exponential increment bound and $H_{t+1} - H_t \leq 1$. Summing from 0 to $t - 1$, with $H_0 = 0$, gives

$$H_t \leq 16 \log\left(1 + \frac{e^{19/16}}{16} h_{\max} t\right). \tag{61}$$

The same calculation bounds H_{t+1} during the induction. To check that it stays below α_ϵ throughout the required horizon, use $n = K\bar{n}$ in the definition of T_ϵ:

$$h_{\max} T_\epsilon \leq \frac{2\eta\sigma_p^2 d}{\bar{n}m} T_\epsilon \leq 64 a_\epsilon^2 K^2/\epsilon + 1.$$

Here the final 1 accounts for rounding T_ϵ upwards and uses $h_{\max} \leq 1$. The elementary bounds $a_\epsilon \leq 128 \log(2K/\epsilon)$, $K^2/\epsilon \leq (2K/\epsilon)^2$, and $[\log(2K/\epsilon)]^2 \leq (2K/\epsilon)^2$ imply

$$\log\left(1 + \frac{e^{19/16}}{16} h_{\max} T_\epsilon\right) \leq \log\left(2 + 2^{20}(2K/\epsilon)^4\right) \leq 26 \log(2K/\epsilon).$$

In the last inequality, $2K/\epsilon \geq 2$ lets us absorb the numerical constants into $22 \log(2K/\epsilon)$ in addition to the displayed fourth power. Hence Equation 61 is at most $416 \log(2K/\epsilon) \leq \alpha_\epsilon = 512 \log(2K/\epsilon)$.

At time $t+1$, we have already proved equality in Equation 59 using the gates at time t. For every $r \in \mathcal{R}_{q,i}^{\xi}$ and every receiver k, the own-sample coefficient at $t+1$ is therefore nonnegative.

The initialization margin and the cross-noise bound now give

$$\left\langle \mathbf{w}_{k,y_{q,i},r}^{(t+1)}, \boldsymbol{\xi}_{q,i} \right\rangle \geq \frac{\sigma_0 \|\boldsymbol{\xi}_{q,i}\|_2}{4} - n\kappa H_{t+1} \geq \frac{\sigma_0 \|\boldsymbol{\xi}_{q,i}\|_2}{8} > 0.$$

The final inequality is exactly the last condition in Equation 58. This proves gate persistence at $t+1$ and closes the induction. In particular, $|\rho| \leq \Psi \leq \alpha_\epsilon$ proves both the positive and negative coefficient bounds in Equation 53.

Unrolling Equation 59 from the zero initial coefficients gives the exact identity

$$\Psi_{k,q,i}^{(t)} = \frac{\eta \|\boldsymbol{\xi}_{q,i}\|_2^2}{\bar{n}m} \sum_{s=0}^{t-1} [\mathbf{M}^{t-1-s}]_{kq} (-\ell_{q,i}'^{(s)}).$$

Unrolling the signal recursion uses the same loss derivatives and mixing weights. Dropping its activation factors, which belong to $[0,1]$, therefore gives

$$\begin{aligned} \gamma_{k,q,j,r}^{(t)} &\leq \frac{\eta \|\boldsymbol{\mu}_q\|_2^2}{\bar{n}m} \sum_{i=1}^{\bar{n}} \sum_{s=0}^{t-1} [\mathbf{M}^{t-1-s}]_{kq} (-\ell_{q,i}'^{(s)}) \\ &= \sum_{i=1}^{\bar{n}} \frac{\|\boldsymbol{\mu}_q\|_2^2}{\|\boldsymbol{\xi}_{q,i}\|_2^2} \Psi_{k,q,i}^{(t)} \leq 4\bar{n} \frac{\mu^2}{\sigma_p^2 d} \alpha_\epsilon. \end{aligned}$$

The last inequality uses $\|\boldsymbol{\mu}_q\|_2^2 = \mu^2$, $\|\boldsymbol{\xi}_{q,i}\|_2^2 \geq \sigma_p^2 d/4$, and the noise bound just proved. Finally, $\langle \boldsymbol{\mu}_q, \boldsymbol{\mu}_k \rangle / \|\boldsymbol{\mu}_q\|_2^2$ equals one for clients sharing client k's signal and zero otherwise. Its sum over q is χ_k. Applying Equation 52 and summing the coefficient bound with these weights gives Equation 55. □

A.9 Convergence Analysis

Definition A.8. *For a given precision $\epsilon > 0$, we define the reference optimal network parameters $\mathbf{W}^*$ globally. For each filter $r \in [m]$ and $j \in \{\pm 1\}$, the target weight is defined as*

$$\mathbf{w}_{j,r}^* = \mathbf{w}_{j,r}^{(0)} + a_\epsilon \sum_{l=1}^{K} \left(j \cdot \chi_l \cdot \frac{\bar{n}\boldsymbol{\mu}_l}{\sigma_p^2 d} + \sum_{i=1}^{n_l} \mathbb{1}(j = y_{l,i}) \frac{\boldsymbol{\xi}_{l,i}}{\|\boldsymbol{\xi}_{l,i}\|_2^2} \right) \tag{62}$$

Lemma A.9. *Under Assumption 4.1, assume the mixing matrix $\mathbf{M}$ is doubly stochastic. Let $\lambda_2 \in [0,1)$ denote the second largest eigenvalue of $\mathbf{M}$. Let $\mathbf{W}^*$ be the target weight defined in Equation 62. At time step T_1, the Frobenius distance between the local model on any client $k \in [K]$ and the optimal reference model is bounded by*

$$\|\mathbf{W}_k^{(T_1)} - \mathbf{W}^*\|_F \leq O\left(\frac{\eta\, \sigma_p \sqrt{Kd}}{(1-\lambda_2)\sqrt{m}} \right) + \tilde{O}\left(\frac{K\sqrt{m\bar{n}}}{\sigma_p \sqrt{d}} \right). \tag{63}$$

The first terms explicitly capture the topological consensus errors (initialization disagreement and dynamic gradient tracking error) unique to decentralized communication, while the third term corresponds to the centralized distance induced by feature learning and noise memorization.

Proof of Lemma A.9. To bound the distance $\|\mathbf{W}_k^{(T_1)} - \mathbf{W}^*\|_F$, we introduce the virtual global average model $\bar{\mathbf{W}}^{(t)} = \frac{1}{K} \sum_{l=1}^{K} \mathbf{W}_l^{(t)}$. By applying the triangle inequality, we decouple the network topology effect from the optimization trajectory

$$\|\mathbf{W}_k^{(T_1)} - \mathbf{W}^*\|_F \leq \underbrace{\left\| \mathbf{W}_k^{(T_1)} - \bar{\mathbf{W}}^{(T_1)} \right\|_F}_{\text{Term I: Dynamic Consensus Error}} + \underbrace{\left\| \bar{\mathbf{W}}^{(T_1)} - \mathbf{W}^* \right\|_F}_{\text{Term II: Global Feature Learning Distance}}.$$

The single-step local update is given by

$$\mathbf{W}_k^{(t+1)} = \sum_{l=1}^{K} \mathbf{M}_{kl}\mathbf{W}_l^{(t)} - \eta\nabla L_k(\mathbf{W}_k^{(t)}).$$

Based on the updating rule and the definition of the virtual global average model, we have

$$\begin{aligned}\bar{\mathbf{W}}^{(t+1)} &= \frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{K}\mathbf{M}_{kl}\mathbf{W}_l^{(t)} - \frac{\eta}{K}\sum_{k=1}^{K}\nabla L_k(\mathbf{W}_k^{(t)}),\\ &= \bar{\mathbf{W}}^{(t)} - \frac{\eta}{K}\sum_{k=1}^{K}\nabla L_k(\mathbf{W}_k^{(t)}).\end{aligned}$$

By unrolling the D-SGD update rule for both the local weights and the global average weights from $t = 0$ to T_1, we have

$$\mathbf{W}_k^{(T_1)} = \sum_{l=1}^{K}[\mathbf{M}^{T_1}]_{kl}\mathbf{W}_l^{(0)} - \eta\sum_{\tau=0}^{T_1-1}\sum_{l=1}^{K}[\mathbf{M}^{T_1-1-\tau}]_{kl}\nabla L_l(\mathbf{W}_l^{(\tau)}).$$

$$\bar{\mathbf{W}}^{(T_1)} = \frac{1}{K}\sum_{l=1}^{K}\mathbf{W}_l^{(0)} - \eta\sum_{\tau=0}^{T_1-1}\frac{1}{K}\sum_{l=1}^{K}\nabla L_l(\mathbf{W}_l^{(\tau)}).$$

Subtracting these identities yields

$$\begin{aligned}\mathbf{W}_k^{(T_1)} - \overline{\mathbf{W}}^{(T_1)} = &\sum_{l=1}^{K}\left([\mathbf{M}^{T_1}]_{kl} - \frac{1}{K}\right)\mathbf{W}_l^{(0)}\\ &- \eta\sum_{\tau=0}^{T_1-1}\sum_{l=1}^{K}\left([\mathbf{M}^{T_1-1-\tau}]_{kl} - \frac{1}{K}\right)\nabla L_{\mathcal{S}_l}(\mathbf{W}_l^{(\tau)}).\end{aligned}$$

All clients share the same initialization. Consequently,

$$\sum_{l=1}^{K}\left([\mathbf{M}^{T_1}]_{kl} - \frac{1}{K}\right)\mathbf{W}_l^{(0)} = \left(\sum_{l=1}^{K}[\mathbf{M}^{T_1}]_{kl} - 1\right)\overline{\mathbf{W}}^{(0)} = \mathbf{0}.$$

Thus the initial-disagreement term vanishes exactly.

We next bound the local gradients explicitly. For each sample $(\mathbf{x}_{l,i}, y_{l,i})$ and filter (j, r),

$$\begin{aligned}&\nabla_{\mathbf{w}_{l,j,r}}\ell\big(y_{l,i}f_l(\mathbf{W}_l^{(\tau)}, \mathbf{x}_{l,i})\big)\\ &\quad= \frac{jy_{l,i}\ell_{l,i}'^{(\tau)}}{m}\Big[\sigma'\big(\langle\mathbf{w}_{l,j,r}^{(\tau)}, y_{l,i}\boldsymbol{\mu}_l\rangle\big)y_{l,i}\boldsymbol{\mu}_l + \sigma'\big(\langle\mathbf{w}_{l,j,r}^{(\tau)}, \boldsymbol{\xi}_{l,i}\rangle\big)\boldsymbol{\xi}_{l,i}\Big].\end{aligned}$$

Using $\langle\boldsymbol{\mu}_l, \boldsymbol{\xi}_{l,i}\rangle = 0$, $|\ell_{l,i}'^{(\tau)}| \le 1$, and $0 \le \sigma' \le 1$, and summing over the $2m$ filters, we obtain

$$\begin{aligned}\Big\|\nabla_{\mathbf{W}_l}\ell\big(y_{l,i}f_l(\mathbf{W}_l^{(\tau)}, \mathbf{x}_{l,i})\big)\Big\|_F^2 &\le \frac{2}{m}\left(\mu^2 + \|\boldsymbol{\xi}_{l,i}\|_2^2\right)\\ &\le \frac{6\sigma_p^2 d}{m}.\end{aligned}$$

Since the full-batch gradient is the average of these sample gradients, convexity of the squared norm gives

$$\begin{aligned}\|\nabla L_{\mathcal{S}_l}(\mathbf{W}_l^{(\tau)})\|_F^2 &\le \frac{1}{\bar{n}}\sum_{i=1}^{\bar{n}}\Big\|\nabla_{\mathbf{W}_l}\ell\big(y_{l,i}f_l(\mathbf{W}_l^{(\tau)}, \mathbf{x}_{l,i})\big)\Big\|_F^2\\ &\le \frac{6\sigma_p^2 d}{m}.\end{aligned}$$

Let $\mathbf{J} = \mathbf{1}\mathbf{1}^\top/K$. For every integer $h \geq 0$, symmetry and the spectral assumption imply

$$\left(\sum_{l=1}^{K}\left|[\mathbf{M}^h]_{kl} - \frac{1}{K}\right|^2\right)^{1/2} = \|\mathbf{e}_k^\top(\mathbf{M}^h - \mathbf{J})\|_2 \leq \lambda_2^h,$$

where the right-hand side is interpreted as 1 when $h = 0$.

Applying the triangle inequality and Cauchy–Schwarz therefore gives

$$\begin{aligned}
\|\mathbf{W}_k^{(T_1)} - \overline{\mathbf{W}}^{(T_1)}\|_F &\leq \eta \sum_{\tau=0}^{T_1-1}\left\|\sum_{l=1}^{K}\left([\mathbf{M}^{T_1-1-\tau}]_{kl} - \frac{1}{K}\right)\nabla L_{\mathcal{S}_l}(\mathbf{W}_l^{(\tau)})\right\|_F \\
&\leq \eta \sum_{\tau=0}^{T_1-1}\left(\sum_{l=1}^{K}\left|[\mathbf{M}^{T_1-1-\tau}]_{kl} - \frac{1}{K}\right|^2\right)^{1/2}\left(\sum_{l=1}^{K}\|\nabla L_{\mathcal{S}_l}(\mathbf{W}_l^{(\tau)})\|_F^2\right)^{1/2} \\
&\leq \eta\sqrt{\frac{6K\sigma_p^2 d}{m}}\sum_{\tau=0}^{T_1-1}\lambda_2^{T_1-1-\tau} = \eta\sqrt{\frac{6K\sigma_p^2 d}{m}}\frac{1-\lambda_2^{T_1}}{1-\lambda_2} \\
&\leq \frac{\eta}{1-\lambda_2}\sqrt{\frac{6K\sigma_p^2 d}{m}}.
\end{aligned}$$

We further decompose

$$\|\overline{\mathbf{W}}^{(T_1)} - \mathbf{W}^*\|_F \leq \|\overline{\mathbf{W}}^{(T_1)} - \overline{\mathbf{W}}^{(0)}\|_F + \|\overline{\mathbf{W}}^{(0)} - \mathbf{W}^*\|_F.$$

Define the averaged coefficients by

$$\bar{\gamma}_{q,j,r}^{(t)} = \frac{1}{K}\sum_{k=1}^{K}\gamma_{k,q,j,r}^{(t)}, \qquad \bar{\rho}_{q,j,r,i}^{(t)} = \frac{1}{K}\sum_{k=1}^{K}\rho_{k,q,j,r,i}^{(t)} = \bar{\psi}_{q,j,r,i}^{(t)} + \bar{\phi}_{q,j,r,i}^{(t)},$$

where $\bar{\psi}$ and $\bar{\phi}$ are defined by the same receiver averages. Averaging the signal–noise decomposition gives

$$\overline{\mathbf{w}}_{j,r}^{(T_1)} - \mathbf{w}_{j,r}^{(0)} = j\sum_{q=1}^{K}\bar{\gamma}_{q,j,r}^{(T_1)}\frac{\boldsymbol{\mu}_q}{\mu^2} + \sum_{q=1}^{K}\sum_{i=1}^{\bar{n}}\bar{\rho}_{q,j,r,i}^{(T_1)}\frac{\boldsymbol{\xi}_{q,i}}{\|\boldsymbol{\xi}_{q,i}\|_2^2}.$$

Column stochasticity of $\mathbf{M}$ implies that the mixing operation preserves each receiver average. Averaging the coefficient recursions and using $0 \leq -\ell' \leq 1$ and $0 \leq \sigma' \leq 1$ yields

$$0 \leq \bar{\gamma}_{q,j,r}^{(t+1)} - \bar{\gamma}_{q,j,r}^{(t)} \leq \frac{\eta\mu^2}{Km},$$

and

$$\left|\bar{\rho}_{q,j,r,i}^{(t+1)} - \bar{\rho}_{q,j,r,i}^{(t)}\right| \leq \frac{\eta\|\boldsymbol{\xi}_{q,i}\|_2^2}{K\bar{n}m}.$$

All coefficients start from zero. Since $T_1 \leq c_T K\bar{n}m/(\eta\sigma_p^2 d)$, telescoping gives

$$\begin{aligned}
0 \leq \bar{\gamma}_{q,j,r}^{(T_1)} &\leq \frac{\eta T_1\mu^2}{Km} \leq \frac{c_T\bar{n}\mu^2}{\sigma_p^2 d}, \\
|\bar{\rho}_{q,j,r,i}^{(T_1)}| &\leq \frac{\eta T_1\|\boldsymbol{\xi}_{q,i}\|_2^2}{K\bar{n}m} \leq \frac{c_T\|\boldsymbol{\xi}_{q,i}\|_2^2}{\sigma_p^2 d}.
\end{aligned}$$

For the signal contribution, the triangle inequality gives

$$\sum_{j,r}\left\|j\sum_{q=1}^{K}\bar{\gamma}_{q,j,r}^{(T_1)}\frac{\boldsymbol{\mu}_q}{\mu^2}\right\|_2^2 \leq 2m\left(\frac{c_T K\bar{n}\mu}{\sigma_p^2 d}\right)^2 = \frac{2c_T^2 mn^2\mu^2}{\sigma_p^4 d^2} \leq \frac{2c_T^2 mn}{K\sigma_p^2 d},$$

where the last step uses $\sigma_p^2 d \geq Kn\mu^2$.

For the noise contribution, Lemma A.2 gives the uniform bound

$$\left\| \sum_{q,i} c_{q,i} \frac{\boldsymbol{\xi}_{q,i}}{\|\boldsymbol{\xi}_{q,i}\|_2^2} \right\|_2^2 \le \frac{5}{4} \sum_{q,i} \frac{c_{q,i}^2}{\|\boldsymbol{\xi}_{q,i}\|_2^2}$$

for every coefficient vector $(c_{q,i})$, including coefficients selected by training.

Consequently, for each (j, r),

$$\left\| \sum_{q,i} \bar{\rho}_{q,j,r,i}^{(T_1)} \frac{\boldsymbol{\xi}_{q,i}}{\|\boldsymbol{\xi}_{q,i}\|_2^2} \right\|_2^2 \le \frac{5}{4} \sum_{q,i} \frac{|\bar{\rho}_{q,j,r,i}^{(T_1)}|^2}{\|\boldsymbol{\xi}_{q,i}\|_2^2} \le \frac{5c_T^2}{4\sigma_p^4 d^2} \sum_{q,i} \|\boldsymbol{\xi}_{q,i}\|_2^2 \le \frac{5c_T^2 n}{2\sigma_p^2 d}.$$

Summing over the $2m$ filters gives

$$\sum_{j,r} \left\| \sum_{q,i} \bar{\rho}_{q,j,r,i}^{(T_1)} \frac{\boldsymbol{\xi}_{q,i}}{\|\boldsymbol{\xi}_{q,i}\|_2^2} \right\|_2^2 \le \frac{5c_T^2 mn}{\sigma_p^2 d}.$$

All noise vectors are orthogonal to all signal vectors. Hence the squared signal and noise contributions add exactly, so

$$\|\overline{\mathbf{W}}^{(T_1)} - \overline{\mathbf{W}}^{(0)}\|_F^2 \le \left(5 + \frac{2}{K}\right) \frac{c_T^2 mn}{\sigma_p^2 d} \le \frac{7c_T^2 mn}{\sigma_p^2 d}.$$

We now bound the reference-solution distance. Recall from Equation 62 that

$$\mathbf{w}_{j,r}^* - \mathbf{w}_{j,r}^{(0)} = \frac{a_\epsilon j\bar{n}}{\sigma_p^2 d} \sum_{q=1}^{K} \chi_q \boldsymbol{\mu}_q + a_\epsilon \sum_{q,i} \mathbf{1}\{y_{q,i} = j\} \frac{\boldsymbol{\xi}_{q,i}}{\|\boldsymbol{\xi}_{q,i}\|_2^2}.$$

Since $1 \le \chi_q \le K$,

$$\left\| \sum_{q=1}^{K} \chi_q \boldsymbol{\mu}_q \right\|_2 \le \mu \sum_{q=1}^{K} \chi_q \le \mu K^2.$$

Thus the squared signal contribution, summed over all filters, is bounded by

$$\sum_{j,r} \left\| \frac{a_\epsilon j\bar{n}}{\sigma_p^2 d} \sum_{q} \chi_q \boldsymbol{\mu}_q \right\|_2^2 \le \frac{2ma_\epsilon^2 \bar{n}^2 \mu^2 K^4}{\sigma_p^4 d^2} \le \frac{2ma_\epsilon^2 Kn}{\sigma_p^2 d}.$$

For the noise contribution, applying the same Gram bound gives

$$\sum_{j,r} \left\| a_\epsilon \sum_{q,i} \mathbf{1}\{y_{q,i} = j\} \frac{\boldsymbol{\xi}_{q,i}}{\|\boldsymbol{\xi}_{q,i}\|_2^2} \right\|_2^2 \le \frac{5a_\epsilon^2}{4} \sum_{j,r} \sum_{q,i} \frac{\mathbf{1}\{y_{q,i} = j\}}{\|\boldsymbol{\xi}_{q,i}\|_2^2} \le \frac{5ma_\epsilon^2 n}{\sigma_p^2 d}.$$

Here the equality uses $\sum_{j \in \{\pm 1\}} \mathbf{1}\{y_{q,i} = j\} = 1$. By signal–noise orthogonality,

$$\|\mathbf{W}^* - \overline{\mathbf{W}}^{(0)}\|_F^2 \le \frac{ma_\epsilon^2(2Kn + 5n)}{\sigma_p^2 d} \le \frac{8ma_\epsilon^2 Kn}{\sigma_p^2 d}.$$

Therefore,

$$\text{Term II} \le \frac{\sqrt{7}\, c_T \sqrt{mn}}{\sigma_p \sqrt{d}} + \frac{\sqrt{8}\, a_\epsilon \sqrt{mKn}}{\sigma_p \sqrt{d}}.$$

Combining the estimates for Terms I and II yields

$$\begin{aligned} \|\mathbf{W}_k^{(T_1)} - \mathbf{W}^*\|_F &\le \frac{\eta}{1 - \lambda_2} \sqrt{\frac{6K\sigma_p^2 d}{m}} + \frac{\sqrt{7}\, c_T \sqrt{mn}}{\sigma_p \sqrt{d}} + \frac{\sqrt{8}\, a_\epsilon \sqrt{mKn}}{\sigma_p \sqrt{d}} \\ &= O\left(\frac{\eta \sqrt{K}\, \sigma_p \sqrt{d}}{(1 - \lambda_2)\sqrt{m}} + \frac{a_\epsilon \sqrt{mKn}}{\sigma_p \sqrt{d}} \right), \end{aligned}$$

where we used $c_T \le 1$, $a_\epsilon \ge 1$, and $K \ge 1$. This proves A.9. Since $n = K\bar{n}$ and $a_\epsilon = O(\log(8/\epsilon))$, the second term can equivalently be written as $\widetilde{O}(K\sqrt{m\bar{n}}/(\sigma_p\sqrt{d}))$. □

Lemma A.10. *For every $t \geq 0$,*

$$\max_{k\in[K]} \|\mathbf{W}_k^{(t)} - \overline{\mathbf{W}}^{(t)}\|_F \leq \mathcal{O}\left(\frac{\eta\sqrt{Kd}\sigma_p}{(1-\lambda_2)m}\right). \tag{64}$$

The same bound holds pathwise for the uniform mini-batch updates.

Lemma A.11. *For all $k \in [K], i \in [n_k]$, and $t \in [0, T_\epsilon]$, we have the following under Assumption 4.1,*

$$y_{k,i}\langle \nabla f_k(W_k^{(t)}, \boldsymbol{x}_{k,i}), \mathbf{W}^*\rangle \geq \log(8/\epsilon). \tag{65}$$

Proof. Fix (k, i) and expand the directional derivative:

$$y_{k,i}\langle \nabla f_k(\mathbf{W}_k^{(t)}, \mathbf{x}_{k,i}), \mathbf{W}^*\rangle = \frac{1}{m}\sum_{j,r} \sigma'\Big(\langle \mathbf{w}_{k,j,r}^{(t)}, y_{k,i}\boldsymbol{\mu}_k\rangle\Big)\,\langle \boldsymbol{\mu}_k, j\mathbf{w}_{j,r}^*\rangle + \frac{1}{m}\sum_{j,r} \sigma'\Big(\langle \mathbf{w}_{k,j,r}^{(t)}, \boldsymbol{\xi}_{k,i}\rangle\Big)\,\langle y_{k,i}\boldsymbol{\xi}_{k,i}, j\mathbf{w}_{j,r}^*\rangle.$$

We now substitute Equation 62 into the two signed projections in this display. Orthogonality of the signal and noise subspaces gives

$$\langle \boldsymbol{\mu}_k, j\mathbf{w}_{j,r}^*\rangle = j\langle \boldsymbol{\mu}_k, \mathbf{w}_{j,r}^{(0)}\rangle + \frac{a_\epsilon \bar{n}}{\sigma_p^2 d}\sum_q \chi_q \langle \boldsymbol{\mu}_k, \boldsymbol{\mu}_q\rangle,$$

$$\langle y_{k,i}\boldsymbol{\xi}_{k,i}, j\mathbf{w}_{j,r}^*\rangle = y_{k,i}j\langle \boldsymbol{\xi}_{k,i}, \mathbf{w}_{j,r}^{(0)}\rangle + a_\epsilon y_{k,i} j \sum_{q,h:y_{q,h}=j} \frac{\langle \boldsymbol{\xi}_{k,i}, \boldsymbol{\xi}_{q,h}\rangle}{\|\boldsymbol{\xi}_{q,h}\|_2^2}.$$

The signal increment is nonnegative because the signal directions are either identical or orthogonal:

$$\frac{a_\epsilon \bar{n}}{\sigma_p^2 d}\sum_{q=1}^K \chi_q \langle \boldsymbol{\mu}_k, \boldsymbol{\mu}_q\rangle \geq 0.$$

It can therefore be dropped in a lower bound. For the noise increment, the term $(q, h) = (k, i)$ appears exactly when $j = y_{k,i}$, and its normalized inner product is 1. Every other term has absolute value at most κ by Lemma A.2. Separating the self-term gives

$$a_\epsilon y_{k,i} j \sum_{q,h:y_{q,h}=j} \frac{\langle \boldsymbol{\xi}_{k,i}, \boldsymbol{\xi}_{q,h}\rangle}{\|\boldsymbol{\xi}_{q,h}\|_2^2} \geq \begin{cases} a_\epsilon(1-2n\kappa), & j = y_{k,i}, \\ -2a_\epsilon n\kappa, & j = -y_{k,i}. \end{cases}$$

Here $2n\kappa$ uniformly bounds the sum of cross-example errors in either branch.

By Equation 56, at least $m/8$ matching filters remain active. Each contributes at least $a_\epsilon(1-2n\kappa)$. Any remaining matching filters contribute nonnegatively. In the opposite branch, at most m filters are active, each costing at most $2a_\epsilon n\kappa$.

After the factor $1/m$ in the directional derivative, the total noise increment is consequently at least

$$a_\epsilon\{\tfrac{1}{8}(1-2n\kappa) - 2n\kappa\} \geq \frac{a_\epsilon}{16},$$

where $n\kappa \leq 1/64$. Finally, each initialized signal projection has absolute value at most β_μ, and each initialized noise projection at most β_ξ. There are $2m$ filters and all gate values lie in $[0, 1]$, so their total contribution is at least $-2\beta_\mu - 2\beta_\xi > -1$. Combining the three contributions gives

$$y_{k,i}\langle \nabla f_k(\mathbf{W}_k^{(t)}, \mathbf{x}_{k,i}), \mathbf{W}^*\rangle \geq \frac{a_\epsilon}{16} - 1 = \log(8/\epsilon).$$

□

Lemma A.12 (One-step local-distance potential bound)**.** *For full-batch DGD, define*

$$V_t = \frac{1}{K}\sum_{k=1}^K \|\mathbf{W}_k^{(t)} - \mathbf{W}^*\|_F^2. \tag{66}$$

If $0 < \eta \leq m(1 + \lambda_{\min}(\mathbf{M}))/[4(\mu^2 + 2\sigma_p^2 d)]$, then, for every $0 \leq t < T_\epsilon$,

$$V_{t+1} \leq V_t - \eta L_{\mathcal{S}}^{(t)} + \frac{\eta\epsilon}{4}. \tag{67}$$

Proof. Write $\mathbf{G}_k^{(t)} = \nabla L_{\mathcal{S}_k}(\mathbf{W}_k^{(t)})$ in this proof. The ReLU network is positively homogeneous of degree one, so $\langle \nabla f_k(\mathbf{W}_k, \mathbf{x}), \mathbf{W}_k \rangle = f_k(\mathbf{W}_k, \mathbf{x})$. Convexity of the scalar logistic loss and Lemma A.11 imply

$$\begin{aligned}\langle \mathbf{G}_k^{(t)}, \mathbf{W}_k^{(t)} - \mathbf{W}^* \rangle &\geq L_{\mathcal{S}_k}(\mathbf{W}_k^{(t)}) - \frac{1}{\bar{n}} \sum_i \ell\Big(y_{k,i} \langle \nabla f_k(\mathbf{W}_k^{(t)}, \mathbf{x}_{k,i}), \mathbf{W}^* \rangle \Big) \\ &\geq L_{\mathcal{S}_k}(\mathbf{W}_k^{(t)}) - \frac{\epsilon}{8}.\end{aligned}$$

The final inequality uses $\ell(z) \leq e^{-z}$. For the scalar logistic loss,

$$|\ell'(z)|^2 \leq |\ell'(z)| = \frac{e^{-z}}{1 + e^{-z}} \leq \log(1 + e^{-z}) = \ell(z).$$

Retaining this loss factor in the single-example gradient estimate, and using Jensen's inequality and the orthogonality of the two patches, yields

$$\begin{aligned}\|\mathbf{G}_k^{(t)}\|_F^2 &\leq \frac{1}{\bar{n}} \sum_i \frac{2(\mu^2 + \|\boldsymbol{\xi}_{k,i}\|_2^2)}{m} |\ell'^{(t)}_{k,i}|^2 \\ &\leq \frac{2(\mu^2 + 2\sigma_p^2 d)}{m} L_{\mathcal{S}_k}(\mathbf{W}_k^{(t)}).\end{aligned} \tag{68}$$

We next exploit mixing directly within V_t. Let $\mathbf{M} = \mathbf{U} \operatorname{diag}(\lambda_1, \ldots, \lambda_K) \mathbf{U}^\top$, where $\mathbf{U}$ is orthogonal and $0 \leq \lambda_a \leq 1$. For each eigenmode a, denote, only in this calculation,

$$\mathbf{E}_a = \sum_k U_{ka} (\mathbf{W}_k^{(t)} - \mathbf{W}^*), \qquad \mathbf{G}_a = \sum_k U_{ka} \mathbf{G}_k^{(t)}.$$

Row stochasticity ensures that the transformed error after one update is $\lambda_a \mathbf{E}_a - \eta \mathbf{G}_a$. Completing the square gives, for every $\lambda \in [0, 1]$,

$$\begin{aligned}\|\lambda \mathbf{E} - \eta \mathbf{G}\|_F^2 &= \|\mathbf{E}\|_F^2 - 2\eta \langle \mathbf{E}, \mathbf{G} \rangle + \eta^2 \|\mathbf{G}\|_F^2 \\ &\quad - (1 - \lambda^2) \left\| \mathbf{E} - \frac{\eta}{1 + \lambda} \mathbf{G} \right\|_F^2 + \eta^2 \frac{1 - \lambda}{1 + \lambda} \|\mathbf{G}\|_F^2 \\ &\leq \|\mathbf{E}\|_F^2 - 2\eta \langle \mathbf{E}, \mathbf{G} \rangle + \frac{2\eta^2}{1 + \lambda} \|\mathbf{G}\|_F^2.\end{aligned}$$

This also holds at $\lambda = 1$; no division by a spectral gap occurs. Summing over eigenmodes and using the orthogonality of $\mathbf{U}$ yields

$$\begin{aligned}V_{t+1} &\leq V_t - \frac{2\eta}{K} \sum_k \langle \mathbf{G}_k^{(t)}, \mathbf{W}_k^{(t)} - \mathbf{W}^* \rangle + \frac{2\eta^2}{K(1 + \lambda_{\min}(\mathbf{M}))} \sum_k \|\mathbf{G}_k^{(t)}\|_F^2 \\ &\leq V_t - \left(2\eta - \frac{4\eta^2(\mu^2 + 2\sigma_p^2 d)}{m(1 + \lambda_{\min}(\mathbf{M}))} \right) L_{\mathcal{S}}^{(t)} + \frac{\eta \epsilon}{4}.\end{aligned} \tag{69}$$

The stated step-size bound proves Equation 67. □

Proof of Theorem 4.6. The first term of the base step-size bound implies the sufficient step size of Lemma A.12. Sum Equation 67 from $t = 0$ to $T - 1$ and discard the nonnegative final potential. Shared initialization and A.9 give $V_0 \leq 8 m a_\epsilon^2 K n / (\sigma_p^2 d)$. Thus, for every $1 \leq T \leq T_\epsilon$,

$$\frac{1}{T} \sum_{t=0}^{T-1} L_{\mathcal{S}}^{(t)} \leq \frac{8 m a_\epsilon^2 K n}{\eta T \sigma_p^2 d} + \frac{\epsilon}{4}. \tag{70}$$

With

$$T_\epsilon = \left\lceil \frac{32 m a_\epsilon^2 K n}{\eta \epsilon \sigma_p^2 d} \right\rceil,$$

the right-hand side is at most $\epsilon/2$. Moreover, $T_1 \le T_\epsilon/8$. Since all training losses are nonnegative,

$$\frac{1}{T_\epsilon - T_1} \sum_{t=T_1}^{T_\epsilon - 1} L_{\mathcal{S}}^{(t)} \le \frac{T_\epsilon}{T_\epsilon - T_1} \frac{\epsilon}{2} \le \frac{4\epsilon}{7} < \epsilon.$$

There is therefore one common integer $t_* \in [T_1, T_\epsilon)$ such that $L_{\mathcal{S}}^{(t_*)} \le \epsilon$. The probability statement follows from the common geometry event. This conclusion concerns the average of the local training losses at the same iteration; it does not assert that each local loss is at most ϵ.

□

A.10 Generalization analysis

Lemma A.13 (Bounded logits during Stage 2)**.** *Let $T_1 = \lfloor c_T nm/(\eta \sigma_p^2 d) \rfloor$, where $c_T > 0$ is a sufficiently small universal constant. Under the parameter assumptions and Equation 50, $T_1 \ge 1$, and every $0 \le t \le T_1$ satisfies*

$$|f_k(\mathbf{W}_k^{(t)}, \mathbf{x}_{k,i})| \le 2, \qquad C_1 \le -\ell_{k,i}'^{(t)} \le 1, \qquad C_1 = (1+e^2)^{-1}.$$

Lemma A.14 (Signal retained after T_1)**.** *Assume the long-window condition Equation 50 additionally. For every $T_1 \le t \le T_\epsilon$, every client k, and every $r \in \mathcal{R}_{k,j}^{\mu}$,*

$$j \langle \mathbf{w}_{k,j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_k \rangle \ge c\bar{n}\chi_k \frac{\mu^2}{\sigma_p^2 d}. \tag{71}$$

Also, for every filter,

$$\|\mathbf{H}\mathbf{w}_{k,j,r}^{(t)}\|_2^2 \le C\left(\sigma_0^2 d + \frac{n\alpha_\epsilon^2}{\sigma_p^2 d}\right). \tag{72}$$

The noise bound Equation 72 holds for every $0 \le t \le T_\epsilon$.

Proof of the upper bound in Theorem 4.7. For each label y, the signal patch has signed output at least

$$\frac{1}{m}\sum_r \left[\sigma(\langle \mathbf{w}_{k,y,r}^{(t)}, y\boldsymbol{\mu}_k\rangle) - \sigma(\langle \mathbf{w}_{k,-y,r}^{(t)}, y\boldsymbol{\mu}_k\rangle)\right] \ge c\bar{n}\chi_k \frac{\mu^2}{\sigma_p^2 d} - \beta_\mu. \tag{73}$$

On the opposite branch,

$$\langle \mathbf{w}_{k,-y,r}^{(t)}, y\boldsymbol{\mu}_k\rangle = \langle \mathbf{w}_{-y,r}^{(0)}, y\boldsymbol{\mu}_k\rangle - \sum_q \gamma_{k,q,-y,r}^{(t)} \frac{\langle \boldsymbol{\mu}_q, \boldsymbol{\mu}_k\rangle}{\mu^2} \le \beta_\mu.$$

Its averaged ReLU output is therefore at most β_μ, proving Equation 73.

Next we bound how much test noise can subtract from this margin. Write the independent test noise as $\boldsymbol{\xi} = \sigma_p \mathbf{H}\mathbf{z}$ with $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. The nonnegative opposite-branch noise response

$$g_{-y}(\mathbf{z}) = \frac{1}{m}\sum_r \sigma(\sigma_p \langle \mathbf{H}\mathbf{w}_{k,-y,r}^{(t)}, \mathbf{z}\rangle)$$

has a directly computable mean: a centered Gaussian with standard deviation s has expected ReLU output $s/\sqrt{2\pi}$, so

$$\mathbb{E} g_{-y}(\mathbf{z}) = \frac{\sigma_p}{m\sqrt{2\pi}} \sum_r \|\mathbf{H}\mathbf{w}_{k,-y,r}^{(t)}\|_2 \le \frac{\sigma_p}{\sqrt{2\pi}} \max_r \|\mathbf{H}\mathbf{w}_{k,-y,r}^{(t)}\|_2.$$

Because ReLU is 1-Lipschitz, for any $\mathbf{z}, \mathbf{z}'$,

$$\begin{aligned} |g_{-y}(\mathbf{z}) - g_{-y}(\mathbf{z}')| &\le \frac{\sigma_p}{m}\sum_r |\langle \mathbf{H}\mathbf{w}_{k,-y,r}^{(t)}, \mathbf{z} - \mathbf{z}'\rangle| \\ &\le \sigma_p \max_r \|\mathbf{H}\mathbf{w}_{k,-y,r}^{(t)}\|_2 \, \|\mathbf{z} - \mathbf{z}'\|_2. \end{aligned}$$

Following 72, Gaussian concentration gives with probability at least $1-\zeta$ over the test noise,

$$g_{-y}(\mathbf{z}) \le \sigma_p \max_r \|\mathbf{H}\mathbf{w}_{k,-y,r}^{(t)}\|_2 \left(\frac{1}{\sqrt{2\pi}} + \sqrt{2\log(1/\zeta)}\right)$$
$$\le C\left(\sigma_0\sigma_p\sqrt{d} + \sqrt{\frac{n}{d}}\,\alpha_\epsilon\right)\sqrt{\log\frac{2}{\zeta}}.$$

The last line uses $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. The factor σ_p cancels the σ_p^{-1} in the learned-noise norm, which explains the term $\sqrt{n/d}\,\alpha_\epsilon$. All filters share the same test noise; this argument requires no independence among their responses.

The matching-branch noise response is nonnegative. Dropping it and using Equation 73, we obtain on the preceding event

$$yf(\mathbf{W}_k^{(t)}, (y\boldsymbol{\mu}_k, \boldsymbol{\xi})) \ge c\bar{n}\chi_k\frac{\mu^2}{\sigma_p^2 d} - \beta_\mu - C\left(\sigma_0\sigma_p\sqrt{d} + \sqrt{\frac{n}{d}}\,\alpha_\epsilon\right)\sqrt{\log\frac{2}{\zeta}}.$$

Since $\beta_\mu = C\sigma_0\mu\sqrt{L}$, the explicit sufficient condition

$$\bar{n}\chi_k\frac{\mu^2}{\sigma_p^2 d} \ge C\left[\sigma_0\mu\sqrt{L} + \left(\sigma_0\sigma_p\sqrt{d} + \sqrt{\frac{n}{d}}\,\alpha_\epsilon\right)\sqrt{\log\frac{2}{\zeta}}\right] \tag{74}$$

with a sufficiently large constant makes the displayed margin positive. The failure probability is at most ζ for each label, hence also after averaging over the two labels: $\mathcal{R}_k(\mathbf{W}_k^{(t)}) \le \zeta$. All the training-dependent bounds used above hold uniformly in $t \in [T_1, T_\epsilon]$.

Applying the deterministic Gaussian estimate to each such parameter matrix therefore bounds its population risk. In particular, it holds at the common fitted iterate supplied by Theorem 4.6. □

Initialization geometry for the lower bound. All learned noise vectors lie in the span of the n training noises. We will show that the initial odd noise vector has a substantial component outside this span, which training cannot cancel. Let $\mathcal{V} = \operatorname{span}\{\boldsymbol{\xi}_{q,i} : q \in [K], i \in [\bar{n}]\}$. The initial odd noise vector and initial odd signal response are, respectively,

$$\mathbf{v}^{(0)} = \frac{1}{m}\sum_r \mathbf{H}(\mathbf{w}_{+1,r}^{(0)} - \mathbf{w}_{-1,r}^{(0)}), \qquad \Delta_k^{(0)} = \frac{1}{m}\sum_r \langle \mathbf{w}_{+1,r}^{(0)} - \mathbf{w}_{-1,r}^{(0)}, \boldsymbol{\mu}_k\rangle.$$

With additional failure probability at most $\delta/2$,

$$\|\mathbf{v}^{(0)}\|_2 \ge c\sigma_0\sqrt{d/m}, \qquad \|\operatorname{Proj}_{\mathcal{V}}\mathbf{v}^{(0)}\|_2 \le \tfrac{1}{2}\|\mathbf{v}^{(0)}\|_2, \qquad \max_k |\Delta_k^{(0)}| \le C\sigma_0\mu\sqrt{L/m}. \tag{75}$$

For completeness, conditional on the training data, independence of the initial filters gives

$$\mathbf{v}^{(0)} \sim \mathcal{N}(\mathbf{0}, 2\sigma_0^2\mathbf{H}/m), \qquad \Delta_k^{(0)} \sim \mathcal{N}(0, 2\sigma_0^2\mu^2/m).$$

The two branches contribute the factor 2 in these variances, and averaging m independent filters contributes the factor $1/m$. Since $\mathcal{V} \subseteq \operatorname{range}(\mathbf{H})$,

$$\frac{m\|\mathbf{v}^{(0)}\|_2^2}{2\sigma_0^2} \sim \chi^2_{\operatorname{rank}(\mathbf{H})}, \qquad \frac{m\|\operatorname{Proj}_{\mathcal{V}}\mathbf{v}^{(0)}\|_2^2}{2\sigma_0^2} \sim \chi^2_{\dim\mathcal{V}}.$$

Assuming $\operatorname{rank}(\mathbf{H}) \ge d/2$, whereas $\dim\mathcal{V} \le n$. Applying chi-square tails at a constant multiple of L gives, except on events of probability Ce^{-cL},

$$\|\mathbf{v}^{(0)}\|_2^2 \ge \frac{\sigma_0^2 d}{2m}, \qquad \|\operatorname{Proj}_{\mathcal{V}}\mathbf{v}^{(0)}\|_2^2 \le \frac{C\sigma_0^2(n+L)}{m} \le \frac{1}{4}\|\mathbf{v}^{(0)}\|_2^2.$$

The last inequality follows by choosing the constant in $d \ge C(n+L)$ sufficiently large. Gaussian tails for $\Delta_k^{(0)}$, followed by a union bound over k, give its bound in Equation 75. The constants can be chosen so that these events have total failure probability at most $\delta/2$.

To see why this prevents cancellation, for every $\mathbf{z} \in \mathcal{V}$,

$$\begin{aligned} |\langle \mathbf{v}^{(0)}, \mathbf{z} \rangle| &= |\langle \mathrm{Proj}_{\mathcal{V}} \mathbf{v}^{(0)}, \mathbf{z} \rangle| \\ &\leq \tfrac{1}{2} \|\mathbf{v}^{(0)}\|_2 \|\mathbf{z}\|_2, \\ \|\mathbf{v}^{(0)} + \mathbf{z}\|_2^2 &= \|\mathbf{v}^{(0)}\|_2^2 + \|\mathbf{z}\|_2^2 + 2\langle \mathbf{v}^{(0)}, \mathbf{z} \rangle \\ &\geq \|\mathbf{v}^{(0)}\|_2^2 + \|\mathbf{z}\|_2^2 - \|\mathbf{v}^{(0)}\|_2 \|\mathbf{z}\|_2. \end{aligned}$$

Using $ab \leq (a^2 + b^2)/2$ therefore gives

$$\|\mathbf{v}^{(0)} + \mathbf{z}\|_2^2 \geq \tfrac{1}{2} \left(\|\mathbf{v}^{(0)}\|_2^2 + \|\mathbf{z}\|_2^2 \right), \tag{76}$$

simultaneously for every $\mathbf{z} \in \mathcal{V}$. In particular, Equation 76 also holds when $\mathbf{z}$ is the learned noise vector, which depends on the initialization.

Lemma A.15 (Odd-part test-error bound)**.** *For any fixed parameter matrix* $\mathbf{W}$ *and client* k*, define*

$$\Delta_k = \frac{1}{m} \sum_r \langle \mathbf{w}_{+1,r} - \mathbf{w}_{-1,r}, \boldsymbol{\mu}_k \rangle, \qquad \mathbf{v} = \frac{1}{m} \sum_r \mathbf{H}(\mathbf{w}_{+1,r} - \mathbf{w}_{-1,r}).$$

Let Φ *be the standard Gaussian cumulative distribution function. If* $\mathbf{v} \neq \mathbf{0}$*, then*

$$\mathcal{R}_k(\mathbf{W}) \geq \frac{1}{2} \Phi \left(-\frac{(\Delta_k)_+}{\sigma_p \|\mathbf{v}\|_2} \right). \tag{77}$$

Proof. For this proof, write $F(\mathbf{z}) = m^{-1} \sum_r [\sigma(\langle \mathbf{w}_{+1,r}, \mathbf{z} \rangle) - \sigma(\langle \mathbf{w}_{-1,r}, \mathbf{z} \rangle)]$ and $s_+ = F(\boldsymbol{\mu}_k)$, $s_- = F(-\boldsymbol{\mu}_k)$. The identity $\sigma(a) - \sigma(-a) = a$ gives $s_+ - s_- = \Delta_k$ and $F(\boldsymbol{\xi}) - F(-\boldsymbol{\xi}) = \langle \mathbf{v}, \boldsymbol{\xi} \rangle$. Balanced labels and independence of the test noise give

$$\mathcal{R}_k(\mathbf{W}) = \tfrac{1}{2} \mathbb{P}\{F(\boldsymbol{\xi}) \leq -s_+\} + \tfrac{1}{2} \mathbb{P}\{F(\boldsymbol{\xi}) \geq -s_-\}.$$

The weak inequalities agree with the definition of risk, which counts a zero margin as an error.

If $\Delta_k < 0$, these two events cover the sample space and the error is at least $1/2$, which implies the claim. If $\Delta_k \geq 0$, put $I = [-s_+, -s_-]$, an interval of length Δ_k. If both $F(\boldsymbol{\xi})$ and $F(-\boldsymbol{\xi})$ belong to I, their difference has magnitude at most Δ_k. Consequently,

$$\{|\langle \mathbf{v}, \boldsymbol{\xi} \rangle| > \Delta_k\} \subseteq \{F(\boldsymbol{\xi}) \notin I\} \cup \{F(-\boldsymbol{\xi}) \notin I\}.$$

The two events on the right have equal probability, because $\boldsymbol{\xi}$ and $-\boldsymbol{\xi}$ have the same distribution. A union bound and $\langle \mathbf{v}, \boldsymbol{\xi} \rangle \sim \mathcal{N}(0, \sigma_p^2 \|\mathbf{v}\|_2^2)$ thus imply

$$2\mathbb{P}\{F(\boldsymbol{\xi}) \notin I\} \geq \mathbb{P}\{|\langle \mathbf{v}, \boldsymbol{\xi} \rangle| > \Delta_k\} = 2\Phi \left(-\frac{\Delta_k}{\sigma_p \|\mathbf{v}\|_2} \right).$$

Finally, $F(\boldsymbol{\xi}) \notin I$ is the disjoint union of $F(\boldsymbol{\xi}) < -s_+$ and $F(\boldsymbol{\xi}) > -s_-$. Therefore

$$\mathcal{R}_k(\mathbf{W}) \geq \tfrac{1}{2} \mathbb{P}\{F(\boldsymbol{\xi}) \notin I\} \geq \tfrac{1}{2} \Phi \left(-\frac{\Delta_k}{\sigma_p \|\mathbf{v}\|_2} \right),$$

which proves Equation 77 when $\Delta_k \geq 0$ and completes the proof. □

Proof of the lower bound in Theorem 4.7. Work on the intersection of Equation 75 and the events of Lemmas A.3–A.2, which has probability at least $1 - \delta$. Fix any client and any $0 \leq t \leq T_\epsilon$. Let $\Delta_k^{(t)}$ and $\mathbf{v}_k^{(t)}$ be the quantities in Lemma A.15 for $\mathbf{W}_k^{(t)}$. The decomposition gives

$$\begin{aligned} \mathbf{v}_k^{(t)} &= \mathbf{v}^{(0)} + \mathbf{z}_k^{(t)}, \\ \mathbf{z}_k^{(t)} &= \sum_{q,i} \left[\frac{1}{m} \sum_r (\rho_{k,q,+1,r,i}^{(t)} - \rho_{k,q,-1,r,i}^{(t)}) \right] \frac{\boldsymbol{\xi}_{q,i}}{\|\boldsymbol{\xi}_{q,i}\|_2^2} \in \mathcal{V}. \end{aligned}$$

We first lower-bound the norm of this learned vector. For an example with $y_{q,i} = +1$, the coefficient signs are $\rho_{k,q,+1,r,i}^{(t)} \geq 0$ and $\rho_{k,q,-1,r,i}^{(t)} \leq 0$, so their difference is the sum of their magnitudes. For

$y_{q,i} = -1$, the signs reverse and the difference is the negative of that sum. Thus averaging over filters cannot cancel this sample's contribution. Moreover, the matching branch has at least $m/8$ good filters, each attaining $\Psi^{(t)}_{k,q,i}$ by Equation 57. It follows that

$$\left| \frac{1}{m} \sum_r \big(\rho^{(t)}_{k,q,+1,r,i} - \rho^{(t)}_{k,q,-1,r,i}\big) \right| \geq \tfrac{1}{8} \Psi^{(t)}_{k,q,i}.$$

Applying the Gram lower bound to the expansion of $\mathbf{z}^{(t)}_k$, and then using $\|\boldsymbol{\xi}_{q,i}\|_2^2 \leq 2\sigma_p^2 d$, gives

$$\begin{aligned} \|\mathbf{z}^{(t)}_k\|_2^2 &\geq \frac{3}{4} \sum_{q,i} \frac{\Big[m^{-1} \sum_r \big(\rho^{(t)}_{k,q,+1,r,i} - \rho^{(t)}_{k,q,-1,r,i}\big)\Big]^2}{\|\boldsymbol{\xi}_{q,i}\|_2^2} \\ &\geq \frac{3}{512\sigma_p^2 d} \sum_{q,i} (\Psi^{(t)}_{k,q,i})^2. \end{aligned}$$

Taking square roots proves

$$\|\mathbf{z}^{(t)}_k\|_2 \geq \frac{c}{\sigma_p \sqrt{d}} \left(\sum_{q,i} (\Psi^{(t)}_{k,q,i})^2 \right)^{1/2}. \tag{78}$$

Next we upper-bound the odd signal response. The signal recursion uses a gate bounded by one, whereas the good noise filters remain active. Comparing their injections, as in Lemma A.7, therefore gives

$$\gamma^{(t)}_{k,q,j,r} \leq \sum_i \frac{\mu^2}{\|\boldsymbol{\xi}_{q,i}\|_2^2} \Psi^{(t)}_{k,q,i}.$$

Both branches contribute positively to the learned odd signal,

$$\Delta^{(t)}_k = \Delta^{(0)}_k + \sum_q \frac{\langle \boldsymbol{\mu}_k, \boldsymbol{\mu}_q \rangle}{\mu^2} \frac{1}{m} \sum_r \big(\gamma^{(t)}_{k,q,+1,r} + \gamma^{(t)}_{k,q,-1,r}\big).$$

Using the preceding coefficient bound for each of the two branches, and $\|\boldsymbol{\xi}_{q,i}\|_2^{-2} \leq 4/(\sigma_p^2 d)$, yields

$$\begin{aligned} (\Delta^{(t)}_k)_+ &\leq |\Delta^{(0)}_k| + \frac{8\mu^2}{\sigma_p^2 d} \sum_{q,i} \frac{\langle \boldsymbol{\mu}_k, \boldsymbol{\mu}_q \rangle}{\mu^2} \Psi^{(t)}_{k,q,i} \\ &\leq |\Delta^{(0)}_k| + \frac{8\mu^2 \sqrt{\bar{n}\chi_k}}{\sigma_p^2 d} \left(\sum_{q,i} (\Psi^{(t)}_{k,q,i})^2 \right)^{1/2}. \end{aligned}$$

For the Cauchy–Schwarz step, the squared weights sum to

$$\sum_{q=1}^{K} \sum_{i=1}^{\bar{n}} \left(\frac{\langle \boldsymbol{\mu}_k, \boldsymbol{\mu}_q \rangle}{\mu^2} \right)^2 = \bar{n}\chi_k,$$

because the original signal similarities are zero or one. Thus no additional factor of K enters the signal estimate.

We now divide the signal estimate by the odd noise scale. Equation Equation 76 bounds that denominator from below in two useful ways:

$$\begin{aligned} \sigma_p \|\mathbf{v}^{(t)}_k\|_2 &\geq \frac{\sigma_p}{\sqrt{2}} \|\mathbf{v}^{(0)}\|_2 \geq c\sigma_p \sigma_0 \sqrt{d/m}, \\ \sigma_p \|\mathbf{v}^{(t)}_k\|_2 &\geq \frac{\sigma_p}{\sqrt{2}} \|\mathbf{z}^{(t)}_k\|_2 \geq \frac{c}{\sqrt{d}} \left(\sum_{q,i} (\Psi^{(t)}_{k,q,i})^2 \right)^{1/2}. \end{aligned}$$

Use the first bound for the initialization term in the numerator:

$$\frac{|\Delta_k^{(0)}|}{\sigma_p\|\mathbf{v}_k^{(t)}\|_2} \le \frac{C\sigma_0\mu\sqrt{L/m}}{c\sigma_p\sigma_0\sqrt{d/m}} \le C\sqrt{\frac{\mu^2 L}{\sigma_p^2 d}}.$$

Use the second bound for the learned signal term. If $\sum_{q,i}(\Psi_{k,q,i}^{(t)})^2 > 0$, its square root cancels and gives

$$\frac{\frac{8\mu^2\sqrt{\bar{n}\chi_k}}{\sigma_p^2 d}\left(\sum_{q,i}(\Psi_{k,q,i}^{(t)})^2\right)^{1/2}}{\sigma_p\|\mathbf{v}_k^{(t)}\|_2} \le C\frac{\mu^2\sqrt{\bar{n}\chi_k}}{\sigma_p^2\sqrt{d}}.$$

If the sum is zero, the learned signal term is zero as well, so the same bound holds. Adding these two estimates proves

$$\frac{(\Delta_k^{(t)})_+}{\sigma_p\|\mathbf{v}_k^{(t)}\|_2} \le C\sqrt{\frac{\mu^2 L}{\sigma_p^2 d}} + C\sqrt{\frac{\bar{n}\chi_k\mu^4}{\sigma_p^4 d}}. \tag{79}$$

In particular $\mathbf{v}_k^{(t)} \neq \mathbf{0}$ by Equation 76. Under the condition,

$$\frac{\mu^2 L}{\sigma_p^2 d} \le \frac{\sqrt{c_0}L}{\sqrt{\bar{n}\chi_k d}},$$

and the dimension condition, together with $\bar{n}\chi_k \ge 1$, gives $L/\sqrt{\bar{n}\chi_k d} \le C$. Thus the first term on the right of Equation 79 is at most $Cc_0^{1/4}$, while the second is at most $C\sqrt{c_0}$. Choosing c_0 sufficiently small makes the right-hand side of Equation 79 at most one. Lemma A.15 then proves the lower bound simultaneously for every client satisfying the lower signal condition and every $0 \le t \le T_\epsilon$.

In particular, it holds at the same fitted iterate as the upper bound. This completes the proof of Theorem 4.7. □

Table 2: **Theorem map.** All rows share Assumptions 3.1 and 4.1; only additional conditions are listed. Fitting means that the average-client empirical loss satisfies $L_{\mathcal{S}}^{(t_*)} \le \epsilon$ at a common $t_* \in [T_1, T_\epsilon)$. Here S, Q abbreviate $S_k(T), Q_k(T)$. The last column highlights graph and initialization dependence beyond the common assumptions.

Result	Additional conditions	Time	Fitting	Metric	Key quantities
Early feature learning and risk					
Lemma 4.3	None	$1 \le T \le T_0$	Not asserted	Signal projection; noise energy	S, Q, σ_0
Theorem 4.4	Initialization absorption Equation 11 and the theorem's signal-strength condition	$1 \le T \le T_0$	Not asserted	$\mathcal{R}_k(\mathbf{W}_k^{(T)})$	S^2/Q; σ_0 in Equation 11
Fitting and risk at a common iterate					
Theorem 4.6	Initialization lower bound Equation 15	Some $t_* \in [T_1, T_\epsilon)$	Yes	$L_{\mathcal{S}}^{(t_*)}$	σ_0 in Equation 15
Theorem 4.7 (upper bound)	Initialization lower bound Equation 15, step-size condition Equation 50, and signal condition Equation 16	Same t_*	Yes	$\mathcal{R}_k(\mathbf{W}_k^{(t_*)}) \le \zeta$	χ_k, σ_0; λ_2 through Equation 50
Theorem 4.7 (lower bound)	Initialization lower bound Equation 15 and $\bar{n}\chi_k\mu^4 \le c_0\sigma_p^4 d$	Same t_*	Yes	$\mathcal{R}_k(\mathbf{W}_k^{(t_*)}) \ge c_*$	χ_k; σ_0 in Equation 15

B PERIODIC LOCAL UPDATES

Proof of Proposition 4.8. The only change in the coefficient formulas is the sequence of mixing matrices. We first identify the propagation of each injection, and then bound the accumulated propagation over periodic blocks.

For $0 \le s < T$, put

$$\mathbf{P}_{T,s} := \mathbf{M}_{T-1}\mathbf{M}_{T-2}\cdots\mathbf{M}_{s+1}, \qquad \mathbf{P}_{T,T-1} = \mathbf{I}.$$

The coefficient induction in Section A.2 applies with $\mathbf{M}$ replaced by $\mathbf{M}_t$. For a fixed source q, a coefficient vector with zero initialization and nonnegative source injection $b_q^{(s)}$ satisfies

$$\mathbf{z}_{:,q}^{(s+1)} = \mathbf{M}_s \mathbf{z}_{:,q}^{(s)} + \mathbf{e}_q b_q^{(s)}.$$

The injection at time s is added after multiplication by $\mathbf{M}_s$. It is therefore first propagated by $\mathbf{M}_{s+1}$, which explains why $\mathbf{M}_s$ does not appear in $\mathbf{P}_{T,s}$. For example,

$$\mathbf{z}_{:,q}^{(2)} = \mathbf{M}_1 \mathbf{e}_q b_q^{(0)} + \mathbf{e}_q b_q^{(1)}.$$

Iterating the same identity up to time T gives the exact form

$$z_{kq}^{(T)} = \sum_{s=0}^{T-1} [\mathbf{P}_{T,s}]_{kq} b_q^{(s)}, \qquad \mathbf{A}_T = \sum_{s=0}^{T-1} \mathbf{P}_{T,s}. \tag{80}$$

Every product $\mathbf{P}_{T,s}$ is entrywise nonnegative and doubly stochastic. Consequently, the unconditional signal and noise bounds used in the proof of Lemma 4.3 become

$$0 \le \gamma_{k,q,j,r}^{(T)} \le \frac{\eta\mu^2}{m}[\mathbf{A}_T]_{kq},$$
$$|\rho_{k,q,j,r,i}^{(T)}| \le \frac{\eta\|\boldsymbol{\xi}_{q,i}\|_2^2}{\bar{n}m}[\mathbf{A}_T]_{kq}.$$

For every intermediate time $u \le T$, nonnegativity and the row sums give

$$\sum_q [\mathbf{A}_u]_{kq} = u, \qquad 0 \le [\mathbf{A}_u]_{kq} \le u \le T.$$

These are exactly the matrix properties used to control the training preactivations in Lemma 4.3. Substituting $\mathbf{A}_u$ into that argument therefore gives the same bounds on the schedule-independent window $1 \le T \le T_0$.

In particular, the fixed aligned signal filters stay active, their logistic factors have the same lower bound, and unrolling their lower injections gives

$$\gamma_{k,q,j,r}^{(T)} \ge \frac{\eta C_1 \mu^2}{3m}[\mathbf{A}_T]_{kq} \qquad \text{for } r \in \mathcal{R}_{q,j}^{\mu}.$$

The noise-energy calculation is equally unchanged. More explicitly, the normalized-noise Gram bound and the coefficient upper bound give

$$\begin{aligned}\|\mathbf{H}(\mathbf{w}_{k,j,r}^{(T)} - \mathbf{w}_{j,r}^{(0)})\|_2^2 &\le \frac{5}{4}\sum_{q,i}\frac{(\rho_{k,q,j,r,i}^{(T)})^2}{\|\boldsymbol{\xi}_{q,i}\|_2^2} \\ &\le \frac{5\eta^2}{4\bar{n}^2m^2}\sum_{q,i}\|\boldsymbol{\xi}_{q,i}\|_2^2[\mathbf{A}_T]_{kq}^2 \\ &\le \frac{C\eta^2\sigma_p^2 d}{\bar{n}m^2}\sum_q [\mathbf{A}_T]_{kq}^2.\end{aligned}$$

There are $\bar{n}$ examples for each source q, and $Q_k(T) = \sum_q [\mathbf{A}_T]_{kq}^2$. Adding the initialization thus yields

$$\|\mathbf{H}\mathbf{w}_{k,j,r}^{(T)}\|_2^2 \le C\left[\sigma_0^2 d + \frac{\eta^2\sigma_p^2 d}{\bar{n}m^2}Q_k(T)\right].$$

This proves the transfer asserted in Lemma 4.3. The proof of Theorem 4.4 uses only these signal and noise bounds, so its certificate transfers as well.

We still need to prove the communication-budget statement. The schedule has period $x + y$, with y occurrences of $\mathbf{M}$ in every period. A product of length $a = T - 1 - s$ can be divided, starting from its first factor, into $\lfloor a/(x+y) \rfloor$ complete periods and a remainder.

Each complete period contains exactly y occurrences of $\mathbf{M}$, regardless of its starting phase. Thus the product contains at least $y\lfloor a/(x+y) \rfloor$ occurrences. Every other factor is $\mathbf{I}$, so the product is a power of $\mathbf{M}$ and

$$\|\mathbf{P}_{T,s} - \mathbf{J}\|_2 \le \lambda_2^{y\lfloor (T-1-s)/(x+y) \rfloor}.$$

As s runs from 0 to $T-1$, its lag $a = T - 1 - s$ runs through $0, \dots, T-1$. Grouping these lags into blocks of $x + y$ gives

$$\begin{aligned}
\|\mathbf{A}_T - T\mathbf{J}\|_2 &\le \sum_{a=0}^{T-1} \lambda_2^{y\lfloor a/(x+y) \rfloor} \\
&\le \sum_{b=0}^{\infty} \sum_{a=b(x+y)}^{(b+1)(x+y)-1} \lambda_2^{by} \\
&= (x+y) \sum_{b=0}^{\infty} \lambda_2^{by}.
\end{aligned}$$

Since $\lambda_2 < 1$ and $y \ge 1$, the geometric series converges and yields the residual bound

$$\|\mathbf{A}_T - T\mathbf{J}\|_2 \le \frac{x+y}{1-\lambda_2^y}. \tag{81}$$

To turn this matrix estimate into bounds on $S_k(T)$ and $Q_k(T)$, use the fact that the residual annihilates $\mathbf{1}$:

$$S_k(T) - \frac{T\chi_k}{K} = \sum_q [\mathbf{A}_T - T\mathbf{J}]_{kq} \left(\frac{\langle \boldsymbol{\mu}_k, \boldsymbol{\mu}_q \rangle}{\mu^2} - \frac{\chi_k}{K} \right),$$

$$\sum_q \left(\frac{\langle \boldsymbol{\mu}_k, \boldsymbol{\mu}_q \rangle}{\mu^2} - \frac{\chi_k}{K} \right)^2 = \chi_k(1 - \chi_k/K).$$

Cauchy–Schwarz proves the bound on $S_k(T)$ below. For $Q_k(T)$, the row $\mathbf{e}_k^\top T\mathbf{J} = (T/K)\mathbf{1}^\top$ is orthogonal to the corresponding residual row, because the residual row sums to zero. Squaring their sum therefore gives

$$\left| S_k(T) - \frac{T\chi_k}{K} \right| \le \frac{x+y}{1-\lambda_2^y} \sqrt{\chi_k(1-\chi_k/K)}, \tag{82}$$

$$Q_k(T) = \frac{T^2}{K} + \|\mathbf{e}_k^\top (\mathbf{A}_T - T\mathbf{J})\|_2^2 \le \frac{T^2}{K} + \left(\frac{x+y}{1-\lambda_2^y} \right)^2. \tag{83}$$

For $y = 1$, condition Equation 19 says that the residual bound is at most $T\sqrt{\chi_k}/(2K)$. Substitution into Equation 82 bounds its error term by $T\chi_k\sqrt{1-\chi_k/K}/(2K) \le T\chi_k/(2K)$.

Likewise, substituting into Equation 83 and using $\chi_k \le K$ gives

$$S_k(T) \ge \frac{T\chi_k}{2K}, \qquad Q_k(T) \le \frac{T^2}{K} + \frac{T^2\chi_k}{4K^2} \le \frac{2T^2}{K}.$$

Therefore

$$\frac{S_k(T)^2}{Q_k(T)} \ge \frac{\chi_k^2}{8K},$$

which is Equation 20. Substitution into Theorem 4.4, with the same initialization and signal-strength conditions, gives Equation 21. □

C EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

C.1 EXPERIMENTAL SETTINGS AND MEASURED QUANTITIES

Synthetic Settings. The synthetic experiments follow the two-patch model in Section 3. Each client trains a two-layer ReLU CNN using full-batch gradients on its local data. For a fixed communication matrix, the update is

$$W^{(t+1)} = MW^{(t)} - \eta \nabla L(W^{(t)}).$$

In the periodic experiments, the mixing matrix at step t is M_t, with $M_t = I$ at local-only steps. Complete DGD uses $M = \mathbf{1}\mathbf{1}^\top/K$ in this update; FedAvg, where reported, is labeled separately because it aggregates after the local update. All clients share the same initialization. Within each seed, graph and placement comparisons use paired client data, initialization, and test samples.

Table 3: Synthetic experimental configurations. C is the number of signal types, $\bar{n}$ the number of examples per client, and m the number of filters per branch. All configurations use $\sigma_p = 1$ and five independent seeds.

Study	K	C	$\bar{n}$	d	m	η	$T_{\max}$
Matched layouts	12	2	100	1000	50	0.1	400
One-type graph comparison	20	1	100	1000	50	0.1	400
Block layouts	20	2	100	1000	50	0.1	400
Communication and strength sweep	16	2	32	4096	64	0.001549724	8192

The first three studies set $\|\boldsymbol{\mu}_c\|_2 = 1$. The matched-ring and block-layout studies rearrange two signal types on a fixed ring, whereas the one-type study assigns all clients the same signal type and compares communication graphs, including a ring and a star. The matched-ring study uses $\sigma_0 = 0.01$; the primary one-type and block-layout comparisons use isotropic initialization with $\sigma_0 = 0.001$. The final study uses a lazy ring with self-weight $1/2$ and weight $1/4$ on each neighbor, and varies client placement, communication interval, and signal strength. It uses $\sigma_0 = 0.0001$; we report signal strengths with the corresponding results. Its $h = 1$ runs supply both comparisons in Table 1.

Topology-aware signal and noise exposure. We compute $S_k(T)$ and $Q_k(T)$ using the mixing matrices applied in each experiment. For a fixed matrix, $A_T = \sum_{s=0}^{T-1} M^s$. For a periodic schedule, we instead use $A_0 = 0$ and $A_{t+1} = M_t A_t + I$. Both constructions satisfy $A_T \mathbf{1} = T\mathbf{1}$. Writing

$$p_{kq}(T) = \frac{[A_T]_{kq}}{T}, \qquad a_{kq} = \frac{\langle \boldsymbol{\mu}_k, \boldsymbol{\mu}_q \rangle}{\|\boldsymbol{\mu}_k\|_2^2},$$

gives

$$\mathcal{G}_k(T) = \frac{S_k(T)^2}{Q_k(T)} = \frac{\left(\sum_q p_{kq}(T) a_{kq}\right)^2}{\sum_q p_{kq}(T)^2}.$$

In the equal-norm, orthogonal-type synthetic experiments, $a_{kq} = \mathbf{1}\{c(q) = c(k)\}$. Hence

$$\mathcal{G}_k(T) = \underbrace{\frac{1}{\sum_q p_{kq}(T)^2}}_{\text{effective number of sources}} \underbrace{\left(\sum_{q:c(q)=c(k)} p_{kq}(T)\right)^2}_{\text{own-type exposure squared}}.$$

The two factors distinguish the spread of source weights from their alignment with client k's signal. The spectral gap alone does not specify this alignment. When all clients share one signal, $S_k(T) = T$ and $\mathcal{G}_k(T) = T^2/Q_k(T)$. We compute $\overline{\mathcal{G}} = K^{-1} \sum_k \mathcal{G}_k$ by first forming each client's ratio and then averaging across clients.

Measured feature increments. Let H project onto the orthogonal complement of the signal subspace, and write $\Delta \mathbf{w}_{k,j,r}^{(T)} = \mathbf{w}_{k,j,r}^{(T)} - \mathbf{w}_{j,r}^{(0)}$. For each client, the signal diagnostic D averages $j\langle \Delta \mathbf{w}_{k,j,r}^{(T)}, \boldsymbol{\mu}_{c(k)} \rangle$ over filters satisfying

$$j\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_{c(k)} \rangle \geq 0.25\sigma_0 \|\boldsymbol{\mu}_{c(k)}\|_2,$$

and then averages across clients. The learned-noise diagnostic is

$$E_{\mathrm{inc}}(T) = \mathrm{mean}_{k,j,r} \left\| H \Delta \mathbf{w}_{k,j,r}^{(T)} \right\|_2^2.$$

These diagnostics measure aggregate changes in the weights. The coefficient bounds in the analysis track contributions from individual training sources, while test predictions use the full weights, including their initialization.

Client risk and uncertainty. The primary test metric evaluates each client model on fresh examples from its own type or domain, matching the client-level risk in the theory. The reported mean own error first averages across clients within a seed, then across seeds. Supplementary tables also report mixture error, obtained by evaluating each local model on a common type or domain mixture and averaging across clients.

For the five-seed synthetic comparisons, tables report sample standard deviations across seeds. Paired contrasts are formed within each seed; their pointwise 95% Student-t intervals are $\bar{d} \pm t_{4,0.975} s_d / \sqrt{5}$. Clients, filters, and saved checkpoints are measurements within a run, rather than additional independent seeds.

C.2 Full trajectories of the signal and noise controls

Figure 2(a)–(c) extends the matched-layout comparison. In panel (a), the measured signal-increment ratio closely follows the graph-predicted S ratio: both rise to about 1.05 near $T = 20$ and return close to 1 by $T = 400$. Panel (b) provides the fixed-Q control. The Q ratio is 1 throughout, and the learned noise-increment ratio remains close to 1. Panel (c) shows that the normalized signal gain $D/[(\eta\mu^2/m)\bar{S}]$ is nearly the same for the two layouts, although it gradually decreases from about 0.250 to 0.226. At the saved checkpoints through $T = 400$, the bounded-logit fraction and survival of initially good gates remain at 1, and the smallest logged derivative exceeds $(1 + e^2)^{-1}$.

Panels (d)–(f) extend the one-type topology comparison. Since all clients share the same signal type, $S_k(T) = T$ for every graph, and the learned signal trajectories nearly overlap in panel (d). The graphs separate more clearly in learned noise energy in panel (e), with the star higher during early training. Panel (f) shows a corresponding early difference in own-type error, which narrows as training proceeds; the error curves are close by $T = 400$. Thus the full trajectories support the early signal and noise comparisons while showing that their effect on error diminishes later.

C.3 Client placement across training on a fixed ring

We complement the main-text comparison with a separate study on a 20-client ring with two equally represented signal types. We permute complete client datasets into blocks of length 1, 2, 5, or 10; the Random condition uses a saved permutation for each seed. We fix the ring, data scale, and initialization across these layouts. Hence $Q_k(T)$ is unchanged, while placement changes the useful signal exposure $S_k(T)$. Local and Complete DGD use different mixing matrices and are included as reference conditions. Table 4 reports the graph factor and own-type error at three checkpoints.

Within the fixed-ring comparison, longer blocks have larger $\overline{\mathcal{G}}$ and lower own-type error at each checkpoint in 4. At $T = 20$, moving from block length 1 to 10 raises $\overline{\mathcal{G}}$ from 2.159 to 5.444 and reduces own-type error from 29.82% to 20.77%. This ordering persists at $T = 100$ and $T = 400$, but the error gap between these two layouts narrows from 9.05 to 4.59 and then 1.11 percentage points. The Random layout falls between the ordered block layouts. Thus placement has its most visible effect at the earlier recorded checkpoints, while the ring layouts give more similar own-type errors later in training.

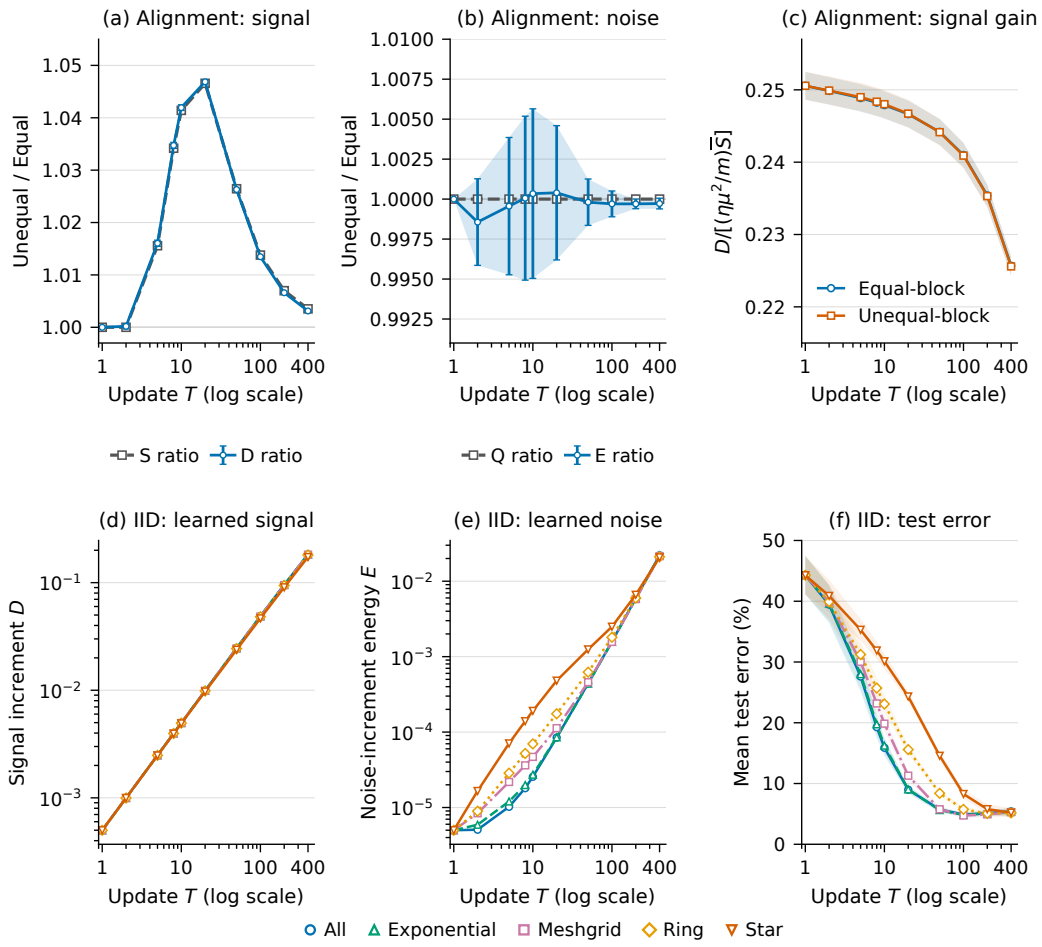

Figure 2: Full saved trajectories for the matched-layout and one-type topology comparisons. On the matched ring, Equal-block uses A/B block lengths $(3, 3, 3, 3)$ and Unequal-block uses $(4, 4, 2, 2)$; plotted ratios are Unequal-block / Equal-block. Graph quantities are computed directly from the mixing matrices, whereas signal and noise curves measure learned weight increments. Results use five paired seeds with pointwise 95% intervals.

C.4 Statistical signal strength over training

We extend the statistical-factor comparison in Table 1 from $T = 32$ to $T = 8192$. The experiments use the same $K = 16$ lazy ring with communication at every step ($h = 1$). Within each layout, we vary only the signal norm, keeping all other data and training parameters fixed. Thus $S_k(T)^2/Q_k(T)$ is unchanged across signal strengths at each fixed checkpoint.

The nominal factor values in the implementation are $\bar{n}\|\mu\|^4/[\sigma_p^4(d-2)]$, where $d-2$ is the dimension of the noise subspace. The statistical factor defined in the main text equals $(d-2)/d$ times these values, a difference of less than 0.05% at $d = 4096$.

At $T = 32$, stronger signals substantially reduce own-type error within both layouts, even though training losses remain near 0.69. The ordering across signal strengths persists at $T = 8192$, while the differences between Clustered and Alternating shrink substantially. At the largest factor, both layouts make no errors on the final test pools. These results extend the main-text comparison: the effect of statistical signal strength remains visible later in training, whereas placement differences are most pronounced at the early checkpoint.

Table 4: Graph factor and own-type error across training in the 20-client study. $\overline{\mathcal{G}}$ averages client-specific $S_k(T)^2/Q_k(T)$. Errors are in percent; reported uncertainties are sample SD over five paired seeds. The ring layouts share $Q_k(T)$ at each checkpoint.

Step	Configuration	$\overline{\mathcal{G}}$	Own error (%)
20	Local	1.000	31.74 ± 0.21
20	Ring, block 1	2.159	29.82 ± 0.42
20	Ring, block 2	2.317	29.23 ± 0.30
20	Ring, block 5	3.499	25.34 ± 0.16
20	Ring, block 10	5.444	20.77 ± 0.22
20	Ring, random	2.989 ± 0.263	27.17 ± 0.87
20	Complete DGD	5.263	24.01 ± 0.84
100	Local	1.000	29.99 ± 0.16
100	Ring, block 1	4.259	20.37 ± 0.63
100	Ring, block 2	4.323	20.22 ± 0.57
100	Ring, block 5	4.784	19.21 ± 0.55
100	Ring, block 10	6.560	15.78 ± 0.48
100	Ring, random	4.735 ± 0.223	19.43 ± 0.95
100	Complete DGD	5.091	19.40 ± 0.94
400	Local	1.000	29.81 ± 0.11
400	Ring, block 1	4.955	18.80 ± 0.75
400	Ring, block 2	4.974	18.77 ± 0.75
400	Ring, block 5	5.105	18.44 ± 0.75
400	Ring, block 10	5.590	17.69 ± 0.76
400	Ring, random	5.088 ± 0.061	18.59 ± 0.91
400	Complete DGD	5.024	19.17 ± 0.88

Table 5: Training loss and own-type error at early and final checkpoints for the three nominal factor levels shown in the main text. Results are the mean $\pm$ sample SD over five paired seeds. Random is evaluated only at factor 1.

		$T = 32$		$T = 8192$	
Factor	Layout	Train loss	Own error (%)	Train loss	Own error (%)
0.1	Clustered	0.68889 ± 0.00003	33.929 ± 0.773	0.4461 ± 0.0030	19.738 ± 0.813
0.1	Alternating	0.68896 ± 0.00004	39.012 ± 1.121	0.4461 ± 0.0031	19.853 ± 0.791
1	Clustered	0.68834 ± 0.00004	9.383 ± 0.840	0.4006 ± 0.0025	1.830 ± 0.360
1	Random	0.68848 ± 0.00004	14.432 ± 2.126	0.4006 ± 0.0025	1.887 ± 0.346
1	Alternating	0.68858 ± 0.00004	18.087 ± 1.036	0.4006 ± 0.0026	1.883 ± 0.363
10	Clustered	0.68658 ± 0.00004	0.018 ± 0.008	0.2973 ± 0.0014	0.000 ± 0.000
10	Alternating	0.68737 ± 0.00005	0.239 ± 0.051	0.2974 ± 0.0014	0.000 ± 0.000

Table 6: Early graph quantities and network responses at $T = 64$. Signal margins and noise RMS are reported in units of 10^{-3}. Measured quantities are mean $\pm$ sample SD over five paired seeds; graph quantities are deterministic for these layouts.

Layout	h	$\overline{S}$	$\overline{Q}$	$\overline{\mathcal{G}}$	Signal	Noise	Own error (%)
Clustered	1	42.100	336.760	5.315	2.683 ± 0.085	1.436 ± 0.020	2.698 ± 0.332
Clustered	4	51.970	614.371	4.496	3.283 ± 0.093	1.567 ± 0.021	1.924 ± 0.280
Clustered	16	57.574	1068.092	3.164	3.585 ± 0.102	1.738 ± 0.021	1.970 ± 0.270
Clustered	64	60.062	1552.375	2.354	3.691 ± 0.107	1.892 ± 0.021	2.306 ± 0.255
Alternating	1	32.500	336.760	3.137	2.061 ± 0.070	1.439 ± 0.019	6.389 ± 0.457
Alternating	4	32.500	614.371	1.719	2.027 ± 0.062	1.575 ± 0.018	8.329 ± 0.464
Alternating	16	32.500	1068.092	0.989	1.981 ± 0.062	1.747 ± 0.017	10.743 ± 0.492
Alternating	64	32.500	1552.375	0.680	1.928 ± 0.076	1.900 ± 0.017	12.801 ± 0.473

C.5 Communication intervals and generalization over time

We vary the communication interval in the 16-client lazy-ring setup, fixing the nominal statistical factor at 1. For both Clustered and Alternating layouts, we consider $h \in \{1, 4, 16, 64\}$. Each period consists of $h - 1$ local-only updates followed by one DGD update with communication, with gradients evaluated at the local parameters before the update. This corresponds to $x = h - 1$ and $y = 1$ in the periodic-update formulation. Across conditions, we fix the learning rate and number of local updates, and pair the underlying client datasets and initialization within each of five seeds. Local training provides a communication-free reference.

Early signal and noise responses. Table 6 reports graph quantities, network-output measurements, and own-type error at $T = 64$. For periodic communication, the cumulative mixing weights satisfy $A_0 = 0$ and $A_{t+1} = M_t A_t + I$, where $M_t = M$ when $(t + 1)$ is divisible by h and $M_t = I$ otherwise. We compute $S_k(T)$ and $Q_k(T)$ from these weights and report their client averages, together with $\overline{\mathcal{G}} = K^{-1} \sum_k S_k(T)^2 / Q_k(T)$. The signal measurement is the mean own-type signal-only margin. The noise measurement is the RMS network output over clients and held-out noise patches, computed within each seed. Both output measurements include initialization.

At each communication interval, the two layouts have identical $Q_k(T)$ and similar mean noise responses. Clustered has greater mean signal exposure, larger signal margins, and lower mean own-type error. Thus, the placement-dependent signal differences observed under communication at every step also appear under less frequent communication.

Changing the interval reveals a complementary effect. For Alternating, $S_k(T) = 32.5$ remains fixed while $Q_k(T)$ increases with h, accompanied by larger noise responses and higher own-type error. The paired error increase from $h = 1$ to $h = 64$ is 6.412 percentage points, with a 95% confidence interval of $[6.262, 6.563]$. For Clustered, both signal exposure and noise concentration increase with h, and own-type error varies nonmonotonically. These measurements illustrate early signal and noise responses to the communication schedule; the periodic-update proposition's sufficient communication-budget condition is not satisfied at this checkpoint.

Final performance. Table 7 reports training loss and own-type error after $T = 8192$ local updates.

Larger communication intervals yield lower mean final training loss in both layouts, while their effects on own-type error remain placement-dependent. For Alternating, mean error increases with h. For Clustered, the dependence is nonmonotone, with the lowest mean error among the tested intervals occurring at $h = 64$. The error differences between layouts are smaller than at $T = 64$, but remain more pronounced at longer communication intervals.

Local training reaches an average client training loss of 0.1 in all five seeds, with the first crossing occurring at 4740 ± 72 updates. None of the ring runs reaches this threshold by $T = 8192$. Despite its lower training loss, Local has substantially higher final own-type error than every communicating

Table 7: Training loss and own-type error at $T = 8192$ for different communication intervals, with nominal statistical factor 1. Results are mean $\pm$ sample SD over five paired seeds. Local training performs no communication.

Method / layout	h	Train loss	Own error (%)
Ring, Alternating	1	0.401 ± 0.003	1.883 ± 0.363
Ring, Alternating	4	0.398 ± 0.003	1.958 ± 0.398
Ring, Alternating	16	0.389 ± 0.003	2.119 ± 0.409
Ring, Alternating	64	0.356 ± 0.003	2.438 ± 0.366
Ring, Clustered	1	0.401 ± 0.002	1.830 ± 0.360
Ring, Clustered	4	0.398 ± 0.002	1.875 ± 0.319
Ring, Clustered	16	0.387 ± 0.002	1.866 ± 0.321
Ring, Clustered	64	0.351 ± 0.003	1.460 ± 0.270
Local	–	0.058 ± 0.001	10.110 ± 0.223

condition in the table. This comparison illustrates that fitting the local training data more closely need not improve prediction on fresh samples from the same client distribution.

C.6 Initialization sensitivity of the topology comparison

The graph quantities describe signal exposure and noise accumulation, while predictions also depend on the initial weights $w_{j,r}$. We examine this dependence in a 20-client, one-type comparison between Complete DGD and Ring. We vary the initialization standard deviations σ_p in the signal and noise subspaces separately, with $\sigma_{\text{signal}}, \sigma_{\text{noise}} \in \{0.01, 0.001\}$. These changes leave $S_k(T)$ and $Q_k(T)$ unchanged for each graph. Equal scales give isotropic initialization; unequal scales isolate the effects of the two subspaces.

Table 8 reports $\Delta_{\text{err}} = \text{Err}_{\text{Ring}} - \text{Err}_{\text{Complete}}$, so positive values indicate lower own-type error for Complete DGD.

Table 8: Initialization sensitivity of the topology comparison. Each entry gives Δ_{err} in percentage points, followed by its paired 95% confidence interval over five seeds.

σ_{signal}	σ_{noise}	$T = 20$	$T = 100$	$T = 400$
0.01	0.01	$+0.329$ $[+0.171, +0.487]$	$+0.777$ $[+0.434, +1.121]$	-0.031 $[-0.078, +0.016]$
0.01	0.001	$+6.425$ $[+5.810, +7.041]$	$+0.878$ $[+0.444, +1.312]$	-0.313 $[-0.364, -0.263]$
0.001	0.01	$+0.320$ $[+0.088, +0.553]$	$+0.821$ $[+0.574, +1.068]$	-0.010 $[-0.089, +0.068]$
0.001	0.001	$+6.620$ $[+6.263, +6.977]$	$+0.922$ $[+0.520, +1.323]$	-0.308 $[-0.361, -0.254]$

At $T = 20$, the Complete DGD advantage is much larger when the noise-subspace initialization scale is smaller, under either signal-subspace scale. The error gap is about 6.5 percentage points for $\sigma_{\text{noise}} = 0.001$, compared with about 0.3 for $\sigma_{\text{noise}} = 0.01$. Changing the signal-subspace scale has a comparatively small effect on this early gap.

At $T = 100$, the gaps are below one percentage point in all four settings. By $T = 400$, the smaller-noise settings instead favor Ring by about 0.31 percentage points, while the intervals for the larger-noise settings include zero. Thus initialization affects both the magnitude and the evolution of the observed topology differences.

C.7 Additional signal and noise diagnostics

Network-output trajectories. We complement the weight-increment measurements with network-output diagnostics in the 16-client synthetic study. For signal type c, we compute the signal-only margin

$$m_{k,c}(t) = \min_{y \in \{-1,1\}} y f\left(W_k^{(t)}; [y\mu_c, 0]\right).$$

The own-type and other-type curves average these margins over matching and nonmatching clients, respectively. The noise response is the within-seed root mean square

$$N_{\text{out}}(t) = \left(\text{mean}_{k,b} \, f\left(W_k^{(t)}; [0, \xi_b] \right)^2 \right)^{1/2}.$$

The noise patches reuse a fixed test noise pool independent of training data. Both diagnostics evaluate the full current network and include initialization.

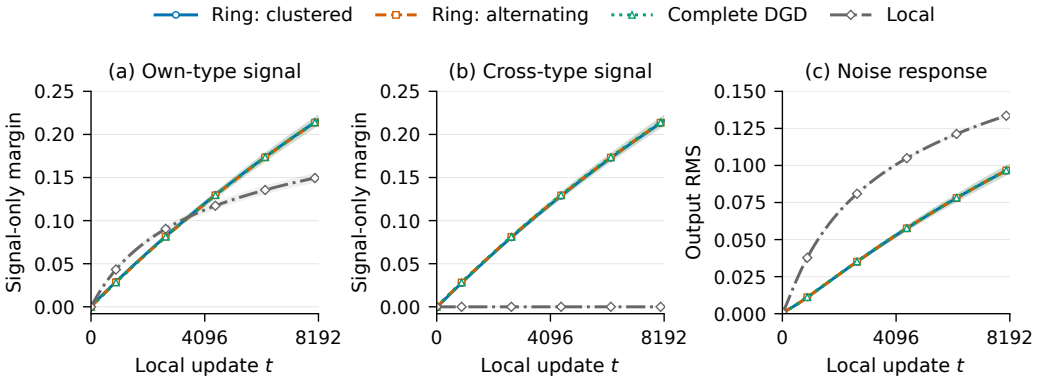

Figure 3: Network-output trajectories at nominal statistical factor 1 and $h = 1$. Panels show own-type signal margins, other-type signal margins, and responses to held-out noise patches.

At later checkpoints, the communicating methods have larger own-type signal margins and smaller noise responses than Local. Their other-type margins also increase, while the Local response to the other signal type remains close to zero. The communicating curves are close on this full-horizon scale. These output measurements complement Local's higher final own-type error in Table 7, despite its lower training loss.

Extension to five signal types. We also examine Clustered and Alternating placements on a fixed 20-client ring with five signal types. For this study, we measure signal increments on initially good filters, averaged separately over matching and other-type clients, together with the RMS norm of the learned noise-subspace weight increments. These measurements subtract the initial weights.

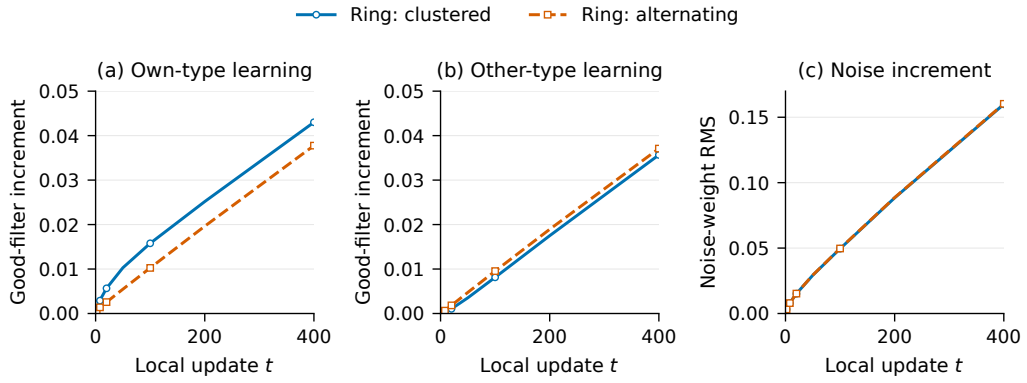

Figure 4: Weight-increment trajectories in the five-type placement study, with $K = 20$, $\bar{n} = 100$, $d = 1000$, $m = 50$, $\sigma_0 = 0.01$, and unit signal norms. Panels show own-type signal increments, other-type signal increments, and the RMS norm of noise-subspace increments.

Figure 4(a) shows larger own-type signal increments for Clustered, while panel (b) shows slightly larger other-type increments for Alternating. The noise-increment trajectories nearly overlap in

panel (c). Since the graph is fixed, both placements have the same $Q_k(T)$, while their signal exposures differ. This extends the placement comparison beyond two signal types: the observed differences remain concentrated in signal acquisition, with similar learned noise accumulation.

C.8 Client placement on real data

The synthetic experiments relate client placement on a fixed graph to signal exposure and own-type prediction. We further examine whether placement also affects own-domain performance when client heterogeneity arises from image domains. Rotated MNIST tests the interaction between placement and communication frequency, while Digits-5 extends the comparison to five domains.

Rotated MNIST. We use binary classification of digits 0 and 1, with rotations of 0° and 60° defining two domains. The 16 clients form a lazy ring, and each receives a disjoint, class-balanced shard of 250 examples. We train a CNN without batch normalization or dropout, using a learning rate of 0.01, a batch size of 50, and 512 local updates. We selected these settings using separate validation runs. We pair client datasets and minibatch sequences across conditions within each of five seeds. Both communication intervals use the same number of local updates.

Table 9: Rotated-MNIST results after 512 local updates. Values are mean $\pm$ sample SD over five paired seeds. Errors are in percent; Comm. counts synchronous mixing events.

Layout	h	Comm.	Train loss	Own error (%)	Mixture error (%)
Clustered	1	512	0.0214 ± 0.0057	0.283 ± 0.031	0.320 ± 0.090
Alternating	1	512	0.0225 ± 0.0067	0.302 ± 0.032	0.305 ± 0.032
Clustered	16	32	0.0173 ± 0.0035	0.218 ± 0.029	7.214 ± 2.241
Alternating	16	32	0.0225 ± 0.0080	0.300 ± 0.042	0.303 ± 0.044

On the same ring, Clustered has lower mean own-domain error than Alternating, with a larger gap when communication occurs every 16 updates. The paired error difference is -0.082 percentage points at $h = 16$, with a 95% interval of $[-0.140, -0.023]$; at $h = 1$, the difference is smaller and its interval includes zero. This provides a real-data counterpart to the synthetic placement comparisons: the arrangement of client domains matters, and its observed effect depends on the communication interval.

Digits-5. We also compare placements on a 20-client ring using Digits-5. Each client has 100 training examples, and training uses full local gradients with one mixing event per update for 100 updates. We report three seeds for each placement. We consider two model configurations: a CNN with batch normalization and a CNN without batch normalization using Kaiming initialization. The latter changes both normalization and initialization.

Table 10: Digits-5 error after 100 updates on a fixed ring. Values are mean $\pm$ sample SD (%) over seeds 1–3. Mixture error evaluates every local model on the balanced five-domain test set.

	With BN		No BN + Kaiming	
Placement	Own error	Mixture error	Own error	Mixture error
Clustered	30.03 ± 1.82	37.50 ± 1.50	43.81 ± 1.49	50.83 ± 1.54
Random	32.67 ± 1.37	35.23 ± 0.83	46.92 ± 0.88	48.55 ± 0.09
Alternating	33.55 ± 0.79	34.57 ± 1.06	46.75 ± 0.70	46.52 ± 1.25

Clustered has the lowest mean own-domain error in both model configurations. This extends the observed placement effect from two domains to five: grouping clients from the same domain is associated with lower own-domain error under the fixed ring.