NEIGHBOR AVERAGING AS PERTURBATION: RISK-AWARE DECENTRALIZED FEDERATED LEARNING VIA NOISE-PROBED SENSITIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Decentralized federated learning (DFL) enables collaborative training by repeatedly exchanging and averaging locally updated models. Under heterogeneous data, however, an update beneficial to its local client may not benefit a receiving neighbor when averaged. Existing topology-driven or learning-adaptive schemes regulate collaboration structure or influence, while the state-dependent effect of a specific averaging operation on the receiving model remains insufficiently characterized. We introduce a perturbation view of neighbor averaging, treating each candidate average as a finite parameter perturbation along the neighbor-model difference. We derive a local output-drift bound showing that the resulting effect depends jointly on the receiving model's sensitivity and the induced displacement. Motivated by this bound, we propose **NoRA** (*Noise-Probed Risk-Aware Averaging*) and the **NPAR** score, which combines noise-probed sensitivity with pairwise model displacement to estimate averaging risk. NoRA attenuates high-risk interactions with only one additional scalar per client beyond standard model exchange. NoRA preserves symmetric doubly stochastic mixing under state-dependent, time-varying weights and retains the standard $\mathcal{O}(1/\sqrt{KT})$ leading convergence rate under finite-window spectral contraction. Experiments across diverse datasets, heterogeneity settings, and network topologies validate its effectiveness. On CIFAR-10 with Dirichlet $\alpha = 0.3$, NoRA improves mean global test accuracy by 1.7% over the strongest baseline and reaches the target accuracy with 29.3% fewer communication rounds than the fastest competing method.

1 INTRODUCTION

Decentralized federated learning (DFL) removes the central coordinator and distributes training and model exchange over peer-to-peer networks (Lalitha et al., 2018; 2019; Sun et al., 2023; Gao et al., 2023). Despite improved resilience and scalability, statistical heterogeneity remains a fundamental challenge because clients optimize over different local data distributions (Hsieh et al., 2020; Li et al., 2020; Hsu et al., 2019). Consequently, an update that benefits its originating client may not be equally beneficial to another client. This issue is especially salient in DFL, where repeated neighbor averaging directly couples the evolution of locally optimized models.

Recent decentralized learning methods increasingly move beyond fixed topology-based averaging by adapting collaboration using learning-dependent signals (Li et al., 2022; Dai et al., 2022; Tastan et al., 2025; Wang et al., 2026; Yang et al., 2026). Such signals, including validation feedback, model relationships, and training dynamics, are used to adjust the structure or strength of collaboration. However, a complementary question remains: *what effect will a particular candidate average have on the receiving model at its current state?* This distinction is important because decentralized training proceeds through a sequence of concrete averaging operations, whose effects can change as the participating models evolve even for the same client pair. For empirical analysis, we measure the receiver's loss change on a held-out local set after a candidate pairwise average and use an increase as an operational indicator of *one-step adverse averaging*. This diagnostic is used only for offline evaluation and is not observed or optimized during training. To characterize this state-dependent effect before averaging, we adopt a *perturbation view of neighbor averaging*. As illustrated in Fig. 1(a), a candidate average can be viewed as a finite parameter perturbation of the receiving

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.

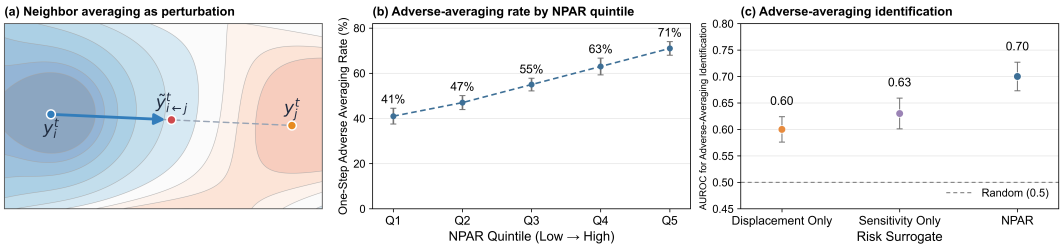


Figure 1: Motivation and empirical diagnosis of pairwise averaging risk. (a) Perturbation view of neighbor averaging from $\mathbf{y}_i^t$ toward $\mathbf{y}_j^t$. (b) One-step adverse averaging rates across NPAR quintiles. (c) AUROC of NPAR, displacement-only, and sensitivity-only surrogates for detecting adverse averaging. Error bars denote standard deviation across independent runs. See Appendix A for details.

model toward its neighbor. Our analysis establishes a local output-drift bound showing that the induced functional change depends jointly on the parameter displacement and the receiver's local sensitivity along the perturbation path. This bound motivates *Noise-Probed Pairwise Averaging Risk* (**NPAR**), an averaging-specific risk surrogate that combines pairwise model displacement with a lightweight noise-probed estimate of receiver sensitivity. Building on NPAR, we propose **NoRA** (*Noise-Probed Risk-Aware Averaging*), which attenuates high-risk neighbor interactions with only one additional scalar communicated per client. Figures 1(b) and 1(c) further examine whether higher NPAR is associated with more frequent one-step adverse averaging and whether the joint sensitivity-displacement surrogate provides stronger discrimination than either component alone.

Because NPAR depends on the current client states, NoRA induces state-dependent, time-varying mixing. We enforce symmetric doubly stochastic mixing to preserve the network average. Under a standard finite-window spectral contraction condition for time-varying decentralized networks (Nedić et al., 2017), we establish non-convex convergence to first-order stationarity with the standard $\mathcal{O}(1/\sqrt{KT})$ leading rate. Our contributions are summarized as follows:

- We introduce a perturbation view of neighbor averaging and derive NPAR, an averaging-specific surrogate that jointly captures receiver sensitivity and pairwise model displacement.
- We develop NoRA for risk-aware decentralized averaging and establish its non-convex convergence under state-dependent, time-varying mixing.
- Extensive experiments across diverse datasets, heterogeneity levels, and network topologies demonstrate improved predictive performance and communication efficiency.

2 Related Work

2.1 Decentralized Learning under Data Heterogeneity

Decentralized learning replaces centralized aggregation with peer-to-peer model exchange over a communication graph, as established in gossip-based optimization and D-PSGD (Boyd et al., 2005; Lian et al., 2017; Lalitha et al., 2018; 2019). Under heterogeneous data, divergent local objectives and optimization trajectories complicate both decentralized optimization and consensus (Hsieh et al., 2020; Li et al., 2020). Existing methods address these challenges through improved local optimization and consensus mechanisms: momentum-based methods accelerate decentralized training (Sun et al., 2023), sharpness-aware methods improve local generalization and model consistency (Shi et al., 2023), and ADMM-based approaches explicitly enforce inter-client agreement (Li et al., 2025). More broadly, heterogeneous DFL has been studied through improved local updates, regularization, communication protocols, and consensus mechanisms to stabilize optimization under non-IID data. These efforts primarily improve optimization and consensus under heterogeneous data, forming the basis of decentralized learning in non-IID settings.

2.2 Adaptive Collaboration in Decentralized Learning

Recent work further shows that decentralized collaboration need not be determined solely by the communication topology. L2C learns client-specific mixing weights from local validation objec-

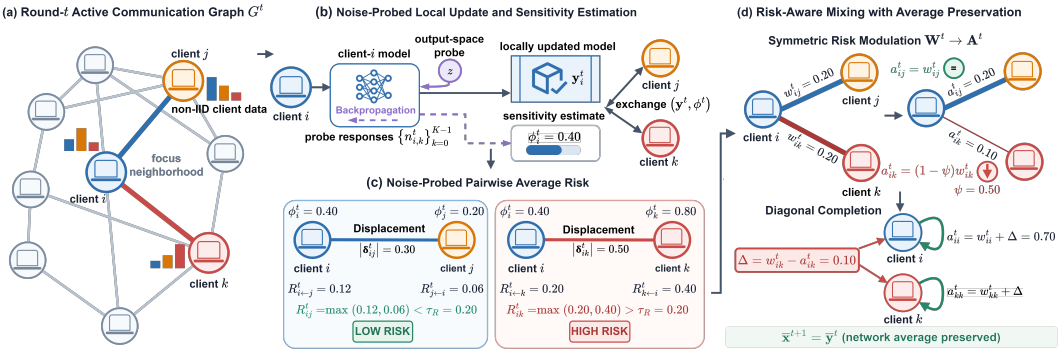


Figure 2: NoRA overview. (a) Round-t decentralized communication. (b) Noise-probed local updates estimate receiver sensitivity. (c) Sensitivity and pairwise displacement define symmetric NPAR. (d) NPAR guides risk-aware mixing while diagonal completion preserves the network average. Numerical values are illustrative.

tives, while meta-L2C extends this mechanism to unseen clients or tasks (Li et al., 2022). DisPFL enables personalized sparse collaboration (Dai et al., 2022), whereas CYCle uses reputation scores derived from gradient alignment to promote positive and fair collaboration gains (Tastan et al., 2025). More recent methods adapt neighbor influence through learned knowledge representations (Wang et al., 2026) or combine layer exchange and averaging based on loss variation and model similarity (Yang et al., 2026). Collectively, these methods use learning-dependent signals to adapt collaboration structure, strength, or granularity. Complementing these approaches, NoRA assesses the state-dependent effects of individual averaging operations using a surrogate that combines receiver sensitivity and model displacement, guiding aggregation without evaluating candidate averages on held-out data during training.

3 NoRA: Noise-Probed Risk-Aware Neighbor Averaging

3.1 Problem Formulation and Decentralized Communication

Consider m clients with heterogeneous data distributions $\{\mathcal{D}_i\}_{i=1}^m$. They collaboratively optimize a shared model parameterized by $\mathbf{x} \in \mathbb{R}^d$:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) := \frac{1}{m}\sum_{i=1}^m f_i(\mathbf{x}), \qquad f_i(\mathbf{x}) := \mathbb{E}_{\xi\sim\mathcal{D}_i} F_i(\mathbf{x};\xi), \tag{1}$$

where $\xi \sim \mathcal{D}_i$ denotes a mini-batch sampled from client i's local data distribution, $F_i(\mathbf{x};\xi)$ is the corresponding stochastic loss, f_i is the local objective of client i, and f is the global objective. As illustrated in Fig. 2(a), communication at round t is governed by an undirected graph $\mathcal{G}^t = (\mathcal{N}, \mathcal{E}^t)$, where $\mathcal{N} = \{1, \dots, m\}$ and $\mathcal{N}_i^t := j : (i,j) \in \mathcal{E}^t$ denotes the neighbor set of client i. Following standard gossip and DFL formulations (Boyd et al., 2005; Dimakis et al., 2010; Sun et al., 2023), let $\mathbf{W}^t = [w_{i,j}^t] \in [0,1]^{m\times m}$ denote a graph-supported, symmetric, doubly stochastic gossip matrix.

At round $t \in \{0, \dots, T-1\}$, client i starts from $\mathbf{x}_{i,0}^t = \mathbf{x}_i^t$ and performs K local steps to obtain the *locally updated model* $\mathbf{y}_i^t := \mathbf{x}_{i,K}^t$. For each $(i,j) \in \mathcal{E}^t$, we isolate the receiver-side effect of neighbor j through the pairwise interpolation

$$\mathbf{y}_i^t(\lambda) := \mathbf{y}_i^t + \lambda\boldsymbol{\delta}_{i,j}^t, \qquad \boldsymbol{\delta}_{i,j}^t := \mathbf{y}_j^t - \mathbf{y}_i^t, \qquad \lambda \in [0,1]. \tag{2}$$

Here $\boldsymbol{\delta}_{i,j}^t$ denotes the pairwise parameter displacement, with λ controlling its magnitude. For a c-class classification problem, let $\mathbf{H}_i(\mathbf{x};\xi) \in \mathbb{R}^c$ denote the logits produced by client i, and define the output Jacobian $\mathbf{J}_i(\mathbf{x};\xi) := \nabla_{\mathbf{x}}\mathbf{H}_i(\mathbf{x};\xi) \in \mathbb{R}^{c\times d}$. While Eq. (2) characterizes the parameter-space displacement induced by neighbor averaging, the corresponding output change also depends on the receiver-side Jacobian sensitivity (Chen et al., 2021). We therefore estimate this sensitivity during local training.

3.2 NOISE-PROBED LOCAL UPDATE AND SENSITIVITY ESTIMATION

As illustrated in Fig. 2(b), directly evaluating the aggregate Jacobian sensitivity is costly for high-dimensional models. NoRA instead estimates it through Gaussian probing. At local step k of round t, client i draws a mini-batch $\xi_{i,k}^t \sim \mathcal{D}_i$ and computes the stochastic gradient $\mathbf{g}_{i,k}^t := \nabla_{\mathbf{x}} F_i(\mathbf{x}_{i,k}^t; \xi_{i,k}^t)$. Let $\mathbf{J}_{i,k}^t := \mathbf{J}_i(\mathbf{x}_{i,k}^t; \xi_{i,k}^t)$. NoRA draws a mean-zero isotropic Gaussian probe $\mathbf{z}_{h,i,k}^t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_c)$ and defines

$$\mathbf{n}_{i,k}^t := \rho(\mathbf{J}_{i,k}^t)^\top \mathbf{z}_{h,i,k}^t, \qquad \widetilde{\mathbf{g}}_{i,k}^t := \mathbf{g}_{i,k}^t + \mathbf{n}_{i,k}^t, \qquad \mathbf{x}_{i,k+1}^t := \mathbf{x}_{i,k}^t - \eta \widetilde{\mathbf{g}}_{i,k}^t, \tag{3}$$

where $\rho > 0$ is the probe scale and $\eta > 0$ is the learning rate. Each probe is sampled independently of the current mini-batch and preceding sampling randomness. The probe response $(\mathbf{J}_{i,k}^t)^\top \mathbf{z}_{h,i,k}^t$ can be computed by backpropagation without explicitly forming $\mathbf{J}_{i,k}^t$.

Following the anisotropic perturbation characterization in Chen et al. (2021), conditioned on $\mathbf{J}_{i,k}^t$,

$$\mathbf{n}_{i,k}^t \mid \mathbf{J}_{i,k}^t \sim \mathcal{N}\Big(\mathbf{0},\, \rho^2 (\mathbf{J}_{i,k}^t)^\top \mathbf{J}_{i,k}^t\Big). \tag{4}$$

Hence,

$$\mathbb{E}_{\mathbf{z}_h}\left[\|\mathbf{n}_{i,k}^t\|^2 \,\middle|\, \mathbf{J}_{i,k}^t\right] = \rho^2 \|\mathbf{J}_{i,k}^t\|_F^2. \tag{5}$$

The probe energy therefore provides a stochastic estimate of the aggregate Jacobian sensitivity. NoRA summarizes the K realized probe responses as

$$\phi_i^t := \left(\frac{1}{Kd\rho^2} \sum_{k=0}^{K-1} \|\mathbf{n}_{i,k}^t\|^2 \right)^{1/2}. \tag{6}$$

The normalization by d accounts for model dimension, while ρ^2 removes the dependence on the probe scale. Thus, ϕ_i^t serves as a stochastic proxy for the receiver-side Jacobian sensitivity along client i's local training trajectory.

3.3 NOISE-PROBED PAIRWISE AVERAGING RISK

As illustrated in Fig. 2(c), we next relate receiver-side sensitivity to the parameter displacement induced by neighbor averaging.

Proposition 1 (Aggregation-Induced Output Drift)**.** *For any fixed mini-batch ξ and $\lambda \in [0,1]$, suppose $\mathbf{H}_i(\cdot;\xi)$ is continuously differentiable on a neighborhood of $\{\mathbf{y}_i^t + s\boldsymbol{\delta}_{i,j}^t : 0 \le s \le \lambda\}$. Then*

$$\left\|\mathbf{H}_i(\mathbf{y}_i^t + \lambda\boldsymbol{\delta}_{i,j}^t; \xi) - \mathbf{H}_i(\mathbf{y}_i^t; \xi)\right\| \le \lambda \sup_{0\le s\le\lambda} \left\|\mathbf{J}_i(\mathbf{y}_i^t + s\boldsymbol{\delta}_{i,j}^t; \xi)\right\|_F \left\|\boldsymbol{\delta}_{i,j}^t\right\|. \tag{7}$$

See Appendix B.1 for the proof. Proposition 1 relates output drift to receiver-side Jacobian sensitivity and pairwise parameter displacement. By Eqs. (5)-(6), ϕ_i^t provides a lightweight trajectory-level proxy for the former. This motivates the directional averaging-risk proxy

$$R_{i\leftarrow j}^t := \phi_i^t \left\|\boldsymbol{\delta}_{i,j}^t\right\|. \tag{8}$$

Notably, Eq. (7) bounds the magnitude of output change without determining the sign of the receiver's loss change. Moreover, ϕ_i^t is estimated along the local training trajectory rather than the candidate averaging path. Thus, $R_{i\leftarrow j}^t$ is a theoretically motivated risk surrogate, not a direct evaluation of the pathwise bound or a guarantee of adverse averaging. For undirected communication, we define the *Noise-Probed Pairwise Averaging Risk (NPAR)*

$$R_{i,j}^t := \max\left\{R_{i\leftarrow j}^t, R_{j\leftarrow i}^t\right\} = \max\left\{\phi_i^t, \phi_j^t\right\} \left\|\mathbf{y}_i^t - \mathbf{y}_j^t\right\|. \tag{9}$$

The maximum retains the larger endpoint-specific risk and guarantees $R_{i,j}^t = R_{j,i}^t$. More generally, for any convex neighbor aggregation $\mathbf{y}_i^{\mathrm{agg}} := \sum_{j=1}^m \alpha_{i,j}\mathbf{y}_j^t$ with $\alpha_{i,j} \ge 0$ and $\sum_j \alpha_{i,j} = 1$,

$$\mathbf{y}_i^{\mathrm{agg}} - \mathbf{y}_i^t = \sum_{j\ne i} \alpha_{i,j}\boldsymbol{\delta}_{i,j}^t, \qquad \left\|\mathbf{y}_i^{\mathrm{agg}} - \mathbf{y}_i^t\right\| \le \sum_{j\ne i} \alpha_{i,j} \left\|\boldsymbol{\delta}_{i,j}^t\right\|. \tag{10}$$

Thus, pairwise displacements constitute the elementary perturbations of general neighbor aggregation, making NPAR a natural edge-level surrogate for risk-aware modulation of mixing weights.

Algorithm 1 NoRA: Noise-Probed Risk-Aware Neighbor Averaging

Input: $T, K, \eta, \rho, \tau_R, \psi$; $\{\mathbf{x}_i^0\}_{i=1}^m$.
Base mixing: $\{\mathbf{W}^t\}_{t=0}^{T-1}$.

for $t = 0$ **to** $T-1$ **do**
  for each client i **in parallel do**
    $\mathbf{x}_{i,0}^t \leftarrow \mathbf{x}_i^t$, $s_i^t \leftarrow 0$
    for $k = 0$ **to** $K-1$ **do**
      Compute $\mathbf{g}_{i,k}^t$ and $\mathbf{n}_{i,k}^t$ via Eq. (3)
      $\mathbf{x}_{i,k+1}^t \leftarrow \mathbf{x}_{i,k}^t - \eta(\mathbf{g}_{i,k}^t + \mathbf{n}_{i,k}^t)$; $s_i^t \leftarrow s_i^t + \|\mathbf{n}_{i,k}^t\|^2$
    end for
    $\mathbf{y}_i^t \leftarrow \mathbf{x}_{i,K}^t$, $\phi_i^t \leftarrow \sqrt{s_i^t/(Kd\rho^2)}$
    Send $(\mathbf{y}_i^t, \phi_i^t)$ to neighbors
  end for
  for each client i **in parallel do**
    Compute $R_{i,j}^t$ and $M_{i,j}^t$, $j \in \mathcal{N}_i^t$, via Eqs. (9)-(11)
    Construct $\{a_{i,j}^t\}_{j=1}^m$ via Eq. (12)
    $\mathbf{x}_i^{t+1} \leftarrow \sum_{j=1}^m a_{i,j}^t \mathbf{y}_j^t$
  end for
end for

3.4 Risk-Aware Mixing with Average Preservation

As illustrated in Fig. 2(d), given the symmetric NPAR scores, NoRA defines the edge-wise risk mask

$$M_{i,j}^t := \mathbb{I}\left[R_{i,j}^t \le \tau_R\right], \qquad (i,j) \in \mathcal{E}^t, \tag{11}$$

where $\tau_R > 0$ is the risk threshold. Since $R_{i,j}^t = R_{j,i}^t$, we have $M_{i,j}^t = M_{j,i}^t$. NoRA then transforms the base gossip matrix $\mathbf{W}^t$ into the risk-aware mixing matrix $\mathbf{A}^t = [a_{i,j}^t] \in \mathbb{R}^{m\times m}$. For $\psi \in [0,1)$,

$$a_{i,j}^t := w_{i,j}^t\left[(1-\psi) + \psi M_{i,j}^t\right] \quad (i \ne j), \qquad a_{i,i}^t := 1 - \sum_{j\ne i} a_{i,j}^t. \tag{12}$$

For non-neighboring pairs, $a_{i,j}^t = 0$. Thus, low-risk edges retain their base weights, whereas high-risk edges are attenuated by $1-\psi$, with the removed mass reassigned to the diagonal.

Proposition 2 (Preservation of Mixing Structure)**.** *If $\mathbf{W}^t$ is nonnegative, symmetric, and doubly stochastic, then $\mathbf{A}^t$ is also nonnegative, symmetric, and doubly stochastic. Moreover,*

$$a_{i,j}^t \ge (1-\psi) w_{i,j}^t, \qquad i \ne j. \tag{13}$$

See Appendix B.2 for the proof. Each client finally aggregates its locally updated model with those received from its neighbors as $\mathbf{x}_i^{t+1} := \sum_{j=1}^m a_{i,j}^t \mathbf{y}_j^t$. Since $\mathbf{A}^t$ is doubly stochastic, the network average is preserved: $\overline{\mathbf{x}}^{t+1} := \frac{1}{m}\sum_i \mathbf{x}_i^{t+1} = \frac{1}{m}\sum_j \mathbf{y}_j^t =: \overline{\mathbf{y}}^t$. Thus, risk-aware mixing attenuates high-risk edges while preserving the network average. Moreover, $\psi < 1$ retains every positive base edge. Algorithm 1 summarizes the complete NoRA procedure.

4 Convergence Analysis

We establish the non-convex convergence of NoRA under standard assumptions for decentralized stochastic optimization (Lian et al., 2017; Sun et al., 2023; Li et al., 2025). We further assume $f^\star := \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > -\infty$.

Assumption 1 (Lipschitz smoothness). Each local objective $f_i : \mathbb{R}^d \to \mathbb{R}$ is differentiable and L-smooth, i.e., $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \le L\|\mathbf{x}-\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $i \in \mathcal{N}$.

Assumption 2 (Stochastic gradients and heterogeneity). For any $i \in \mathcal{N}$, round t, and local step k, the stochastic gradient $\mathbf{g}_{i,k}^t$ is unbiased with bounded variance: $\mathbb{E}_{\xi_{i,k}^t}[\mathbf{g}_{i,k}^t] = \nabla f_i(\mathbf{x}_{i,k}^t)$ and $\mathbb{E}_{\xi_{i,k}^t}\left\|\mathbf{g}_{i,k}^t - \nabla f_i(\mathbf{x}_{i,k}^t)\right\|^2 \le \sigma_l^2$. Moreover, the gradient heterogeneity is bounded as $\frac{1}{m}\sum_{i=1}^m \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \le \sigma_g^2$ for all $\mathbf{x} \in \mathbb{R}^d$.

Assumption 3 (Bounded Jacobian second moment). For any $i \in \mathcal{N}$, round t, and local step k, $\mathbb{E}_{\xi_{i,k}^t}\left[\|\mathbf{J}_{i,k}^t\|_F^2\right] \le G_H^2$.

Under Assumptions 2-3 and Eq. (5), the perturbed gradient remains unbiased and its error has bounded second moment

$$\sigma_\rho^2 := \sigma_l^2 + \rho^2 G_H^2. \tag{14}$$

The derivation is given in Appendix C.1. We next characterize the time-varying mixing induced by NoRA.

Assumption 4 (Joint spectral contraction). Let $\mathbf{A}^t$ be the risk-aware mixing matrix generated by Eq. (12), $\mathbf{P} := m^{-1}\mathbf{1}\mathbf{1}^\top$, and $\mathbf{A}_{t:s} := \mathbf{A}^t\mathbf{A}^{t-1}\cdots\mathbf{A}^s$ for $t \ge s$. There exist an integer $B \ge 1$ and $\beta \in (0,1)$ such that

$$\|\mathbf{A}_{t:t-B+1} - \mathbf{P}\|_2 \le \beta, \qquad \forall\, t \ge B-1. \tag{15}$$

Assumption 4 is the standard finite-window spectral contraction condition for time-varying decentralized networks (Nedić et al., 2017). It implies

$$\mathcal{C}_{A,T} := \max_{1\le t\le T}\sum_{s=0}^{t-1}\|\mathbf{A}_{t-1:s} - \mathbf{P}\|_2 \le \mathcal{C}_B := \frac{B}{1-\beta}. \tag{16}$$

Here, $\mathcal{C}_{A,T}$ measures the cumulative disagreement contraction of the realized mixing sequence, while $\mathcal{C}_B$ provides a uniform upper bound. See Appendix C.2 for the derivation.

Theorem 1 (Non-convex convergence of NoRA). Suppose Assumptions 1-4 hold and $\mathbf{x}_i^0 = \mathbf{x}^0$ for all $i \in \mathcal{N}$. Let $\overline{\mathbf{x}}^t := m^{-1}\sum_{i=1}^m \mathbf{x}_i^t$. If $\eta LK \le [64(1+\mathcal{C}_B)]^{-1}$, then the iterates generated by NoRA satisfy

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^t)\right\|^2 \le \frac{8\big(f(\overline{\mathbf{x}}^0) - f^\star\big)}{\eta KT} + 16L\eta\sigma_\rho^2 + 1280L^2\eta^2K(1+\mathcal{C}_B^2)(\sigma_\rho^2 + K\sigma_g^2). \tag{17}$$

The proof is provided in Appendix C.5.

Corollary 1 (Convergence rate). Under the conditions of Theorem 1, let $\rho > 0$ be fixed and choose $\eta = \Theta(1/(L\sqrt{KT}))$ sufficiently small to satisfy the step-size condition. Then

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^t)\right\|^2 = \mathcal{O}\left(\frac{L\big(f(\overline{\mathbf{x}}^0) - f^\star\big) + \sigma_\rho^2}{\sqrt{KT}} + \frac{(1+\mathcal{C}_B^2)(\sigma_\rho^2 + K\sigma_g^2)}{T}\right). \tag{18}$$

Corollary 1 recovers the standard $\mathcal{O}(1/\sqrt{KT})$ leading rate, while the additional term captures the effects of data heterogeneity and network mixing. NoRA affects the bound through the noise-induced second-moment term σ_ρ^2 and the mixing factor $\mathcal{C}_B$; the risk parameters (τ_R, ψ) influence the latter through the resulting adaptive mixing matrices $\mathbf{A}^t$.

5 Experiments

We evaluate NoRA in terms of predictive performance, convergence and communication efficiency, robustness to network topology, and component ablation and hyperparameter sensitivity.

5.1 Experimental settings

Datasets and Data Partition. We conduct experiments on MNIST (LeCun et al., 2010), CIFAR-10, and CIFAR-100 (Krizhevsky et al., 2009) with $m = 100$ clients. We consider IID data and three non-IID settings: Dirichlet partitions with concentration parameters $\alpha \in \{0.3, 0.6\}$ (Hsu et al., 2019) and a Pathological partition (Zhang et al., 2020), where smaller α induces stronger statistical heterogeneity. For the Pathological partition, training samples are sorted by label, divided into 200 shards, and each client is randomly assigned two shards, yielding severe label skew.

Communication Configurations. Following established DFL experimental protocols (Shi et al., 2023; Li et al., 2025), we use a sparse, time-varying communication topology with a target neighbor budget of 10 per client. At each round, an undirected random graph is generated. We set

Table 1: Test accuracy (%) on MNIST, CIFAR-10, and CIFAR-100 under different data partitions. The highest mean accuracy in each row is highlighted in bold.

Dataset	Heter. Set.	Test Accuracy (%)						
		D-PSGD	DFedAvg	DisPFL	DFedSAM	DFedADMM	DecFus	**NoRA**
MNIST	IID	95.39±0.06	97.66±0.21	97.86±0.10	97.92±0.08	97.56±0.43	97.73±0.03	**98.19**±0.10
	0.6	95.11±0.10	97.48±0.06	97.64±0.08	97.87±0.05	97.58±0.32	97.50±0.05	**98.06**±0.11
	0.3	94.86±0.17	97.31±0.27	97.54±0.13	97.73±0.05	97.51±0.33	97.36±0.08	**98.03**±0.13
	Patho	94.15±0.11	95.93±0.13	95.14±0.14	97.05±0.19	97.18±0.24	94.65±0.40	**97.29**±0.48
CIFAR-10	IID	68.40±0.75	78.06±0.11	73.41±1.04	79.05±0.49	78.02±0.39	77.01±0.48	**79.26**±0.56
	0.6	49.03±1.15	76.26±0.51	73.07±0.99	75.58±0.24	76.32±0.55	75.65±0.33	**77.87**±0.54
	0.3	43.94±0.61	74.44±0.95	71.79±0.89	74.29±0.23	73.78±0.25	73.83±0.44	**76.14**±0.22
	Patho	46.01±0.39	67.00±0.92	54.86±0.86	60.22±0.97	54.96±0.62	56.21±2.70	**69.40**±0.96
CIFAR-100	IID	22.66±0.55	33.86±0.45	34.79±0.43	31.70±2.54	34.98±0.61	31.20±0.61	**36.89**±0.96
	0.6	21.27±0.19	35.13±0.60	36.28±0.86	32.29±3.05	35.92±1.10	31.49±1.59	**37.20**±0.48
	0.3	20.88±0.15	35.16±0.59	36.29±0.58	32.82±2.05	35.38±2.05	30.34±0.39	**36.61**±0.44
	Patho	5.27±0.70	12.06±0.62	7.58±2.30	10.08±0.82	10.71±0.51	11.42±2.31	**12.34**±0.08

the connectivity window to $B = 5$ and require the union of every five consecutive graphs to be connected, allowing individual graphs to be disconnected while ensuring information propagation across successive rounds. Base aggregation uses symmetric Metropolis weights with one gossip step per round. For each random seed, all methods use the same graph sequence.

Baselines. We compare NoRA with six representative decentralized learning methods spanning four methodological families: (1) *standard decentralized optimization*: D-PSGD (Lian et al., 2017) and DFedAvg (Sun et al., 2023); (2) *personalized decentralized learning*: DisPFL (Dai et al., 2022); (3) *heterogeneity-aware local optimization*: DFedSAM (Shi et al., 2023); and (4) *consensus and adaptive fusion*: DFedADMM (Li et al., 2025) and DecFus (Yang et al., 2026). Detailed configurations are provided in Appendix D.1.

Evaluation Metrics. We evaluate predictive performance using *global test accuracy*. Communication efficiency is assessed using two metrics: *Rounds-to-Target*, the number of communication rounds required to reach a predefined target accuracy, and *Relative Communication Cost* (Rel. Comm.), the corresponding cumulative communication volume normalized to DFedAvg.

Implementation Details. All methods are implemented in PyTorch and trained on NVIDIA GeForce RTX 4090 GPUs for $T = 500$ communication rounds. Following Shi et al. (2023), all methods except D-PSGD perform five local epochs per round, whereas D-PSGD performs one mini-batch update. We use SGD with batch size 128, weight decay 5×10^{-4}, and an initial learning rate of 0.065 decayed by 0.9995 after each round. We adopt ResNet-18 with GroupNorm for CIFAR-100, while the MNIST and CIFAR-10 architectures follow the configurations in Li et al. (2025). For NoRA, we set the noise intensity to $\rho = 0.1$, the risk threshold to $\tau_R = 0.08$, and the target modulation strength to $\psi = 0.25$. Results are averaged over three independent runs unless noted otherwise.

5.2 Overall Performance

Table 1 compares NoRA with six baselines across three datasets and four data partitions. NoRA achieves the highest mean test accuracy in all 12 settings, with notable gains on CIFAR-10 under data heterogeneity. Under Dirichlet $\alpha = 0.6$ and 0.3, NoRA outperforms DFedADMM and DFedAvg by 1.55% and 1.70%, respectively. Under the Pathological partition, it achieves 69.40% accuracy, exceeding DFedAvg by 2.40%, compared with a 0.21% advantage over DFedSAM under IID data. On CIFAR-100, NoRA also improves upon DFedADMM by 1.91% under IID data. These results suggest that incorporating receiver sensitivity and model displacement into neighbor averaging benefits predictive performance, particularly under heterogeneous data.

5.3 Convergence and Communication Efficiency

Convergence Speed. Fig. 3 shows that NoRA exhibits relatively smooth accuracy curves with sustained improvements under non-IID data, particularly in the Pathological setting. As quantified in

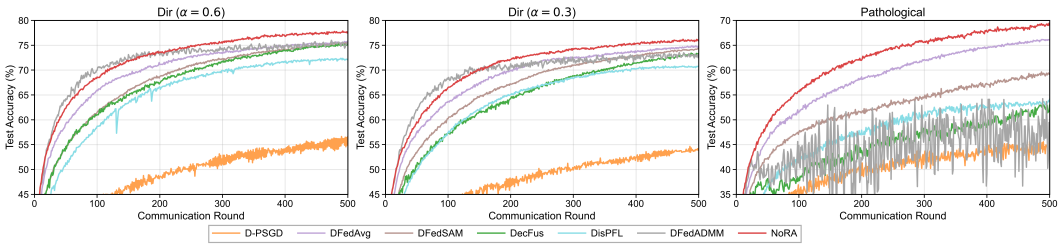


Figure 3: Convergence behavior on CIFAR-10 under three non-IID data partitions.

Table 2: Communication efficiency on CIFAR-10. >500 denotes an unreached target within 500 rounds, with the corresponding cost marked as N/R. The best results are highlighted in bold.

Method	IID@78%		Dir 0.6@76%		Dir 0.3@74%		Patho.@54%	
	Rounds	Rel. Comm.↓	Rounds	Rel. Comm.↓	Rounds	Rel. Comm.↓	Rounds	Rel. Comm.↓
D-PSGD	>500	N/R	>500	N/R	>500	N/R	>500	N/R
DFedAvg	421	1.00	416	1.00	431	1.00	116	1.00
DisPFL	>500	N/R	>500	N/R	>500	N/R	358	1.53
DFedSAM	396	0.89	>500	N/R	446	1.04	240	2.00
DFedADMM	323	0.75	409	0.94	420	0.90	375	1.68
DecFus	>500	N/R	>500	N/R	495	1.22	386	2.11
NoRA	**287**	**0.68**	**304**	**0.68**	**297**	**0.66**	**89**	**0.81**

Table 2, NoRA reaches the target accuracy with the fewest communication rounds across all four settings. Notably, it reduces the required rounds by 25.7% and 29.3% over DFedADMM, the fastest competing method, under Dirichlet $\alpha = 0.6$ and 0.3, respectively. The observed convergence behavior is consistent with NoRA's state-dependent risk modulation, which adapts neighbor influence as local models evolve.

Communication Efficiency. Table 2 further shows that NoRA achieves the lowest cumulative communication cost across all settings, requiring only 0.66-0.81 times that of DFedAvg. These savings indicate that the reduction in rounds to target outweighs the additional scalar communication required for risk estimation.

5.4 Robustness to Network Topology

Communication topology plays a critical role in information propagation and model aggregation in DFL. We evaluate five representative topologies: Ring, Grid, Exponential (Exp), Fully connected (Full), and Random, schematically illustrated in Appendix Fig. 6.

As shown in Fig. 4, most methods benefit from denser connectivity on MNIST under Dirichlet $\alpha = 0.3$, highlighting the influence of network structure on predictive performance. NoRA achieves the highest test accuracy across all five topologies, outperforming DecFus by 0.41% even on the sparsest Ring topology. Its advantage persists across the remaining topologies, reaching 98.16% on Random. These results suggest that state-dependent risk modulation helps NoRA exploit available neighbor interactions across diverse connectivity conditions.

5.5 Ablation and Sensitivity Analysis

Ablation Study. Table 3 examines NoRA's key components and aggregation signals. *Noise only* injects noise during local optimization while retaining fixed gossip weights. *Risk only* employs NPAR-guided aggregation without injecting noise into local optimization. *Random mask* applies random edge attenuation, matching NoRA's modulation strength and number of attenuated edges. *Displacement only* and *Sensitivity only* replace NPAR with model distance and receiver sensitivity, respectively.

NoRA achieves the highest mean accuracy of 69.40%. Compared with *Risk only* and *Noise only*, NoRA improves accuracy by 0.61% and 0.78%, respectively, supporting the complementary contributions of noise injection and risk modulation. NoRA also outperforms *Displacement only* and

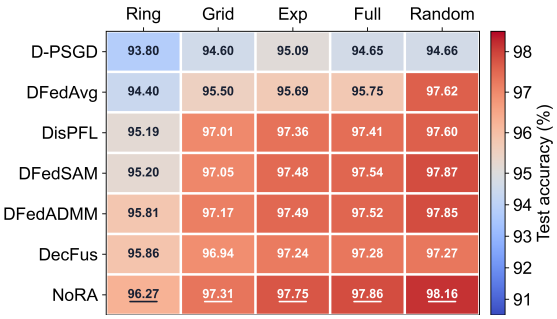


Figure 4: Test accuracy (%) across five network topologies on MNIST with Dirichlet $\alpha = 0.3$.

Table 3: Component ablation on CIFAR-10 under the Pathological partition.

Method	Noise Update	Aggregation Signal	Acc. (%)
Noise only	✓	Fixed	68.62±0.78
Risk only	✗	NPAR	68.79±0.47
Random mask	✓	Random	68.73±0.51
Displacement only	✓	Distance	68.82±0.53
Sensitivity only	✓	Sensitivity	68.96±0.66
NoRA	✓	NPAR	**69.40±0.96**

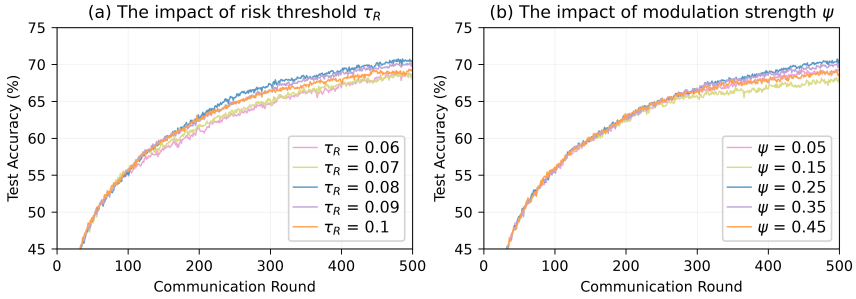


Figure 5: Sensitivity analysis of key hyperparameters: risk threshold τ_R, and modulation strength ψ

Sensitivity only by 0.58% and 0.44%, respectively, indicating the benefit of jointly incorporating sensitivity and model displacement. Its 0.67% advantage over *Random mask* further supports the effectiveness of NPAR-guided attenuation over random edge attenuation.

Risk Threshold τ_R. The threshold τ_R in Eq. (11) determines which neighbor interactions undergo risk-based attenuation. As shown in Fig. 5(a), we vary τ_R from 0.06 to 0.10. Excessively small thresholds attenuate more interactions, potentially limiting beneficial collaboration, whereas larger thresholds reduce the selectivity of risk modulation. We adopt $\tau_R = 0.08$, which yields the highest final test accuracy among the evaluated thresholds.

Modulation Strength ψ. The parameter ψ in Eq. (12) controls the attenuation of high-risk neighbor interactions. As shown in Fig. 5(b), varying ψ from 0.05 to 0.45 yields similar early convergence trajectories, with performance differences emerging later in training. We set $\psi = 0.25$, achieving the highest final accuracy while exhibiting limited sensitivity to moderate variations in modulation strength.

Additional sensitivity analyses of noise intensity ρ, local iteration steps K, and the number of clients m are provided in Appendix D.2.

6 Conclusion

We investigated the state-dependent effects of neighbor averaging in DFL under data heterogeneity. By interpreting each averaging operation as a parameter perturbation, we characterized its potential impact through receiver sensitivity and pairwise model displacement. Building on this formulation, we developed NoRA, which employs the noise-probed NPAR surrogate to adapt aggregation to evolving client states. We established non-convex convergence guarantees under state-dependent, time-varying mixing, retaining the standard leading convergence rate. Experiments across diverse datasets, data partitions, and network topologies demonstrated improvements in test accuracy and communication efficiency. Ablation results further supported the benefits of NPAR-guided attenuation over random edge attenuation. As a surrogate, NPAR does not explicitly account for perturbation direction or sensitivity along the entire averaging path. Future work will explore refined risk estimation and extend NoRA to asynchronous and directed networks.

AI USE STATEMENT

We used generative AI tools to assist with manuscript organization, academic writing, experimental methodology feedback, and results interpretation. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This study uses publicly available benchmark datasets. NoRA does not require clients to share raw training data. However, protecting against information leakage from exchanged model parameters is beyond its scope.

REPRODUCIBILITY STATEMENT

We provide the algorithmic details of NoRA and its theoretical analysis in the main text, with complete proofs provided in the appendix. Experimental settings and implementation details, including data partitioning, model architectures, communication configurations, baseline implementations, and hyperparameters, are documented in the main text and appendix to facilitate reproducibility.

REFERENCES

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Gossip algorithms: Design, analysis and applications. In *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, volume 3, pp. 1653–1664. IEEE, 2005.

Pengfei Chen, Guangyong Chen, Junjie Ye, Pheng-Ann Heng, et al. Noise against noise: stochastic label noise helps combat inherent label noise. In *International Conference on Learning Representations*, 2021.

Rong Dai, Li Shen, Fengxiang He, Xinmei Tian, and Dacheng Tao. Dispfl: Towards communication-efficient personalized federated learning via decentralized sparse training. In *International Conference on Machine Learning*, pp. 4587–4604. PMLR, 2022.

Alexandros G Dimakis, Soummya Kar, José MF Moura, Michael G Rabbat, and Anna Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010.

Hongchang Gao, My T Thai, and Jie Wu. When decentralized optimization meets federated learning. *IEEE network*, 37(5):233–239, 2023.

Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pp. 4387–4398. PMLR, 2020.

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

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

Anusha Lalitha, Shubhanshu Shekhar, Tara Javidi, and Farinaz Koushanfar. Fully decentralized federated learning. In *Third workshop on bayesian deep learning (NeurIPS)*, volume 12, 2018.

Anusha Lalitha, Osman Cihan Kilinc, Tara Javidi, and Farinaz Koushanfar. Peer-to-peer federated learning on graphs. *arXiv preprint arXiv:1901.11173*, 2019.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Qinglun Li, Li Shen, Guanghao Li, Quanjun Yin, and Dacheng Tao. Dfedadmm: Dual constraint controlled model inconsistency for decentralize federated learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4803 – 4815, 2025.

Shuangtong Li, Tianyi Zhou, Xinmei Tian, and Dacheng Tao. Learning to collaborate in decentralized learning of personalized models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9756–9765. IEEE, 2022.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30, 2017.

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

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

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

Nurbek Tastan, Samuel Horváth, and Karthik Nandakumar. CYCle: Choosing your collaborators wisely to enhance collaborative fairness in decentralized learning. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=ygqNiLQqfH.

Shuai Wang, Youliang Tian, Axin Xiang, Jinbo Xiong, and Jianfeng Ma. pdfl: Knowledge-aware adaptive aggregation for personalized decentralized federated learning. *Knowledge-Based Systems*, pp. 116749, 2026.

Li Yang, Jialong Sun, Chuhai Cai, Xinyang Liu, Yichen Li, Bowen Peng, Jialong Li, and Bo Liu. Decfus: Decentralized layer-wise fusion with dynamic exploration and exploitation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=QNYVYUa3aa.

Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. *arXiv preprint arXiv:2012.08565*, 2020.

A Offline Diagnosis of Pairwise Averaging Risk

Diagnostic Setup. We conduct offline diagnosis on CIFAR-10 under the Pathological partition with 100 clients, each assigned data from two classes. We follow the main experimental setup, including the model architecture, mini-batch size, and five local epochs per round, but train for 400 communication rounds with an initial learning rate of 0.1, decayed by 0.998 per round. To exclude the effects of noise injection and unequal client degrees, diagnostic trajectories are generated using standard local updates and Metropolis gossip on a static, connected, undirected 10-regular graph. Candidate averages are evaluated at rounds 50, 100, 200, and 400 using a held-out set comprising 10% of each client's local data. Results are averaged over three independent runs.

Risk Diagnosis. At each diagnostic round, we evaluate both receiving directions of all 500 undirected edges and assign one label per edge, yielding 2,000 edge-level samples across four rounds. For each edge $\{i, j\}$, we measure the held-out loss changes at both endpoints under pairwise averaging with Metropolis weight w_{ij}. An edge is labeled adverse if either endpoint exhibits a loss increase. Client sensitivity ϕ_i^t is estimated through independent Gaussian probing without injecting probe noise into local updates. We compare symmetric NPAR with displacement-only and sensitivity-only surrogates, defined by pairwise model distance and $\max\{\phi_i^t, \phi_j^t\}$, respectively.

Within each run, samples from all diagnostic rounds are pooled and grouped into five NPAR quintiles to compute the adverse averaging rates in Fig. 1(b). Figure 1(c) compares the three surrogates using AUROC on the same samples. Results are reported as mean and standard deviation across three independent runs.

B Proofs of Propositions

B.1 Proof of Proposition 1

Proof. Let $\mathbf{h}(s) := \mathbf{H}_i(\mathbf{y}_i^t + s\boldsymbol{\delta}_{i,j}^t; \xi)$ for $s \in [0, \lambda]$. By the fundamental theorem of calculus,

$$\mathbf{H}_i(\mathbf{y}_i^t + \lambda\boldsymbol{\delta}_{i,j}^t; \xi) - \mathbf{H}_i(\mathbf{y}_i^t; \xi) = \int_0^\lambda \mathbf{J}_i(\mathbf{y}_i^t + s\boldsymbol{\delta}_{i,j}^t; \xi)\boldsymbol{\delta}_{i,j}^t \, \mathrm{d}s. \tag{19}$$

Taking norms and using $\|\mathbf{J}\mathbf{v}\| \le \|\mathbf{J}\|_2\|\mathbf{v}\| \le \|\mathbf{J}\|_F\|\mathbf{v}\|$,

$$\begin{aligned}\left\|\mathbf{H}_i(\mathbf{y}_i^t + \lambda\boldsymbol{\delta}_{i,j}^t; \xi) - \mathbf{H}_i(\mathbf{y}_i^t; \xi)\right\| &\le \int_0^\lambda \left\|\mathbf{J}_i(\mathbf{y}_i^t + s\boldsymbol{\delta}_{i,j}^t; \xi)\right\|_F \left\|\boldsymbol{\delta}_{i,j}^t\right\| \mathrm{d}s \\ &\le \lambda \sup_{0\le s\le\lambda} \left\|\mathbf{J}_i(\mathbf{y}_i^t + s\boldsymbol{\delta}_{i,j}^t; \xi)\right\|_F \left\|\boldsymbol{\delta}_{i,j}^t\right\|,\end{aligned}$$

which proves Eq. (7). □

B.2 Proof of Proposition 2

Proof. For $i \ne j$, Eq. (12) and $M_{i,j}^t \in \{0, 1\}$ imply

$$(1 - \psi)w_{i,j}^t \le a_{i,j}^t \le w_{i,j}^t.$$

Thus, all off-diagonal entries of $\mathbf{A}^t$ are nonnegative, while

$$a_{i,i}^t = 1 - \sum_{j\ne i} a_{i,j}^t \ge 1 - \sum_{j\ne i} w_{i,j}^t = w_{i,i}^t \ge 0,$$

where row stochasticity of $\mathbf{W}^t$ is used in the equality. Hence $\mathbf{A}^t$ is nonnegative.

Since $R_{i,j}^t = R_{j,i}^t$, we have $M_{i,j}^t = M_{j,i}^t$. Together with $w_{i,j}^t = w_{j,i}^t$, this yields $a_{i,j}^t = a_{j,i}^t$, so $\mathbf{A}^t$ is symmetric. By the diagonal completion in Eq. (12), $\sum_{j=1}^m a_{i,j}^t = 1$ for every i, hence $\mathbf{A}^t\mathbf{1} = \mathbf{1}$. Symmetry further gives $\mathbf{1}^\top\mathbf{A}^t = \mathbf{1}^\top$, establishing double stochasticity. The lower bound above is exactly Eq. (13). □

C Convergence Proofs and Auxiliary Results

For the proofs below, let $\mathcal{F}_{t,k}$ denote the history before local step (t,k) and define $\mathbb{E}_{t,k}[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t,k}]$. An unqualified $\mathbb{E}$ denotes total expectation. Stochastic mini-batches and Gaussian probes are independently sampled across clients and local steps, with each probe independent of its corresponding mini-batch.

C.1 Effective gradient moments

Define $\boldsymbol{\varepsilon}_{i,k}^t := \mathbf{g}_{i,k}^t - \nabla f_i(\mathbf{x}_{i,k}^t)$ and $\mathbf{e}_{i,k}^t := \widetilde{\mathbf{g}}_{i,k}^t - \nabla f_i(\mathbf{x}_{i,k}^t) = \boldsymbol{\varepsilon}_{i,k}^t + \mathbf{n}_{i,k}^t$. Although $\boldsymbol{\varepsilon}_{i,k}^t$ and $\mathbf{J}_{i,k}^t$ are computed from the same mini-batch, conditioning on $\mathcal{F}_{t,k}$ and the current mini-batch makes both quantities fixed. Since $\mathbf{n}_{i,k}^t = \rho(\mathbf{J}_{i,k}^t)^\top \mathbf{z}_{h,i,k}^t$ and the Gaussian probe is zero mean,

$$\mathbb{E}\left[\left\langle \boldsymbol{\varepsilon}_{i,k}^t, \mathbf{n}_{i,k}^t \right\rangle \middle| \mathcal{F}_{t,k}, \xi_{i,k}^t\right] = 0. \tag{20}$$

Therefore, by Eq. (5), the tower property, and Assumptions 2-3,

$$\mathbb{E}_{t,k}\|\mathbf{e}_{i,k}^t\|^2 = \mathbb{E}_{t,k}\|\boldsymbol{\varepsilon}_{i,k}^t\|^2 + \rho^2 \mathbb{E}_{t,k}\|\mathbf{J}_{i,k}^t\|_F^2 \le \sigma_l^2 + \rho^2 G_H^2 = \sigma_\rho^2. \tag{21}$$

Moreover, $\mathbb{E}_{t,k}[\mathbf{e}_{i,k}^t] = \mathbf{0}$. Thus, $\widetilde{\mathbf{g}}_{i,k}^t$ is conditionally unbiased, with conditional mean-squared estimation error bounded by σ_ρ^2.

C.2 Contraction of Adaptive Mixing Products

By Proposition 2, every $\mathbf{A}^t$ is nonnegative, symmetric, and doubly stochastic. Hence $\|\mathbf{A}^t\|_1 = \|\mathbf{A}^t\|_\infty = 1$ and $\|\mathbf{A}^t - \mathbf{P}\|_2 \le 1$. Moreover, $\mathbf{A}^t\mathbf{P} = \mathbf{P}\mathbf{A}^t = \mathbf{P}$ implies

$$(\mathbf{C} - \mathbf{P})(\mathbf{D} - \mathbf{P}) = \mathbf{C}\mathbf{D} - \mathbf{P} \tag{22}$$

for any compatible products $\mathbf{C}$ and $\mathbf{D}$ of doubly stochastic matrices. Consider a product of length $\ell = qB + r$, where $q = \lfloor \ell/B \rfloor$ and $0 \le r < B$. Partitioning it into q complete B-step blocks and one remainder block, and repeatedly applying Eq. (22) together with $\|\mathbf{C}\mathbf{D}\|_2 \le \|\mathbf{C}\|_2\|\mathbf{D}\|_2$, gives

$$\|\mathbf{A}_{t:t-\ell+1} - \mathbf{P}\|_2 \le \beta^{\lfloor \ell/B \rfloor}. \tag{23}$$

Therefore, $\sum_{s=0}^{t-1} \|\mathbf{A}_{t-1:s} - \mathbf{P}\|_2 \le B \sum_{q \ge 0} \beta^q = B/(1-\beta)$, which proves Eq. (16). Importantly, this bound holds pathwise and therefore does not require independence between the adaptive matrices $\mathbf{A}^t$ and the local updates.

The joint contraction in Assumption 4 follows from standard connectivity and positive-weight conditions for time-varying decentralized networks (Nedić et al., 2017). In particular, Proposition 2 preserves the graph support, while Eq. (13) ensures $a_{i,j}^t \ge (1-\psi) w_{i,j}^t$ on active edges and diagonal completion ensures $a_{i,i}^t \ge w_{i,i}^t$.

Corollary 2 (Per-round spectral-gap specialization). Suppose every $\mathbf{W}^t$ satisfies $1 - \max\{|\lambda_2(\mathbf{W}^t)|, |\lambda_m(\mathbf{W}^t)|\} \ge \gamma > 0$. For any fixed $\psi < 1$, $\|\mathbf{A}^t - \mathbf{P}\|_2 \le 1 - (1-\psi)\gamma$, and $\mathcal{C}_B$ in the preceding bounds can be replaced by $\mathcal{C}_\gamma := 1/[(1-\psi)\gamma]$.

Proof. For every $\mathbf{v} \perp \mathbf{1}$, the Laplacian quadratic-form identity and Eq. (13) yield

$$\mathbf{v}^\top(\mathbf{I} - \mathbf{A}^t)\mathbf{v} = \sum_{i<j} a_{i,j}^t (v_i - v_j)^2 \ge (1-\psi)\mathbf{v}^\top(\mathbf{I} - \mathbf{W}^t)\mathbf{v} \ge (1-\psi)\gamma\|\mathbf{v}\|^2.$$

Hence, $\lambda_2(\mathbf{A}^t) \le 1 - (1-\psi)\gamma$. Moreover, Eq. (12) gives

$$\mathbf{A}^t - \mathbf{W}^t = \sum_{i<j} (w_{i,j}^t - a_{i,j}^t)(\mathbf{q}_i - \mathbf{q}_j)(\mathbf{q}_i - \mathbf{q}_j)^\top \succeq \mathbf{0},$$

where $\mathbf{q}_i$ denotes the ith standard basis vector. Thus $\lambda_m(\mathbf{A}^t) \ge \lambda_m(\mathbf{W}^t)$ and therefore $-\lambda_m(\mathbf{A}^t) \le 1 - \gamma \le 1 - (1-\psi)\gamma$. By symmetry, $\|\mathbf{A}^t - \mathbf{P}\|_2 \le 1 - (1-\psi)\gamma$. Applying submultiplicativity to the ordered products and summing the resulting geometric series gives $\mathcal{C}_\gamma = 1/[(1-\psi)\gamma]$. □

C.3 Deviation During Local Updates

Let $\overline{\mathbf{x}}^t := m^{-1}\sum_{i=1}^m \mathbf{x}_i^t$, stack the client models row-wise as $\mathbf{X}^t := [(\mathbf{x}_1^t)^\top; \dots; (\mathbf{x}_m^t)^\top]$, and define the consensus error

$$\mathcal{E}_t := \frac{1}{m}\mathbb{E}\|\mathbf{X}^t - \mathbf{P}\mathbf{X}^t\|_F^2 = \frac{1}{m}\sum_{i=1}^m \mathbb{E}\|\mathbf{x}_i^t - \overline{\mathbf{x}}^t\|^2. \tag{24}$$

Lemma 1 (Local-update deviation). Under Assumptions 1-3, if $\eta LK \le 1/4$, then

$$\mathcal{D}_t := \frac{1}{m}\sum_{i=1}^m\sum_{k=0}^{K-1} \mathbb{E}\|\mathbf{x}_{i,k}^t - \mathbf{x}_i^t\|^2 \le 2\eta^2K^2\sigma_\rho^2 + 16\eta^2K^3\left(\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^t)\|^2 + L^2\mathcal{E}_t + \sigma_g^2\right). \tag{25}$$

Proof. Let $\mathbf{e}_{i,r}^t := \widetilde{\mathbf{g}}_{i,r}^t - \nabla f_i(\mathbf{x}_{i,r}^t)$. By Eq. (14), $\mathbb{E}_{t,r}[\mathbf{e}_{i,r}^t] = \mathbf{0}$ and $\mathbb{E}_{t,r}\|\mathbf{e}_{i,r}^t\|^2 \le \sigma_\rho^2$. For $p < r$, $\mathbf{e}_{i,p}^t$ is $\mathcal{F}_{t,r}$-measurable, and hence

$$\mathbb{E}\langle \mathbf{e}_{i,p}^t, \mathbf{e}_{i,r}^t\rangle = 0. \tag{26}$$

Therefore,

$$\mathbb{E}\left\|\sum_{r=0}^{k-1}\mathbf{e}_{i,r}^t\right\|^2 = \sum_{r=0}^{k-1}\mathbb{E}\|\mathbf{e}_{i,r}^t\|^2 \le k\sigma_\rho^2. \tag{27}$$

From (3),

$$\mathbf{x}_{i,k}^t - \mathbf{x}_i^t = -\eta\sum_{r=0}^{k-1}\left(\nabla f_i(\mathbf{x}_{i,r}^t) + \mathbf{e}_{i,r}^t\right).$$

Using $\|\mathbf{a}+\mathbf{b}\|^2 \le 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$, Eq. (27), and $\|\sum_{r<k}\mathbf{v}_r\|^2 \le k\sum_{r<k}\|\mathbf{v}_r\|^2$, we obtain

$$\mathbb{E}\|\mathbf{x}_{i,k}^t - \mathbf{x}_i^t\|^2 \le 2\eta^2k\sigma_\rho^2 + 2\eta^2k\sum_{r<k}\mathbb{E}\|\nabla f_i(\mathbf{x}_{i,r}^t)\|^2. \tag{28}$$

For the gradient term,

$$\begin{aligned}\nabla f_i(\mathbf{x}_{i,r}^t) = {}& [\nabla f_i(\mathbf{x}_{i,r}^t) - \nabla f_i(\mathbf{x}_i^t)] \\ &+ [\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\overline{\mathbf{x}}^t)] \\ &+ [\nabla f_i(\overline{\mathbf{x}}^t) - \nabla f(\overline{\mathbf{x}}^t)] + \nabla f(\overline{\mathbf{x}}^t).\end{aligned}$$

By $\|\sum_{j=1}^4 \mathbf{v}_j\|^2 \le 4\sum_{j=1}^4\|\mathbf{v}_j\|^2$, L-smoothness, and Assumption 2,

$$\frac{1}{m}\sum_i \mathbb{E}\|\nabla f_i(\mathbf{x}_{i,r}^t)\|^2 \le 4L^2\frac{1}{m}\sum_i\mathbb{E}\|\mathbf{x}_{i,r}^t - \mathbf{x}_i^t\|^2 + 4L^2\mathcal{E}_t + 4\sigma_g^2 + 4\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^t)\|^2. \tag{29}$$

Summing Eq. (28) over clients and local steps, and using $\sum_{k=0}^{K-1}k \le K^2/2$ and $\sum_k k\sum_{r<k}b_r \le K^2\sum_r b_r$, yields

$$\mathcal{D}_t \le \eta^2K^2\sigma_\rho^2 + 8\eta^2K^2L^2\mathcal{D}_t + 8\eta^2K^3\left(\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^t)\|^2 + L^2\mathcal{E}_t + \sigma_g^2\right). \tag{30}$$

Since $\eta LK \le 1/4$, $8\eta^2L^2K^2 \le 1/2$. Moving the corresponding term to the left gives Eq. (25). □

C.4 Consensus Error Under Adaptive Mixing

Lemma 2 (Consensus error under adaptive mixing). Suppose Assumptions 1-4 hold, $\mathbf{x}_i^0 = \mathbf{x}^0$ for all $i \in \mathcal{N}$, and $\eta LK\mathcal{C}_B \le 1/8$. Then

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{E}_t \le 32\eta^2K\mathcal{C}_B^2\left[\sigma_\rho^2 + K\sigma_g^2 + \frac{K}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^t)\|^2\right]. \tag{31}$$

Proof. Let $\mathbf{Y}^t$ stack the locally updated models and define $\mathbf{U}^t := \mathbf{Y}^t - \mathbf{X}^t$. The algorithmic recursion is

$$\mathbf{X}^{t+1} = \mathbf{A}^t\mathbf{Y}^t = \mathbf{A}^t(\mathbf{X}^t + \mathbf{U}^t). \tag{32}$$

Since $\mathbf{A}^t\mathbf{P} = \mathbf{P}\mathbf{A}^t = \mathbf{P}$, recursive substitution gives

$$\mathbf{X}^t = \mathbf{A}_{t-1:0}\mathbf{X}^0 + \sum_{s=0}^{t-1} \mathbf{A}_{t-1:s}\mathbf{U}^s. \tag{33}$$

Common initialization implies $(\mathbf{I} - \mathbf{P})\mathbf{X}^0 = \mathbf{0}$ and therefore

$$(\mathbf{I} - \mathbf{P})\mathbf{X}^t = \sum_{s=0}^{t-1} (\mathbf{A}_{t-1:s} - \mathbf{P})\mathbf{U}^s. \tag{34}$$

Define $\bar{h}_{t,s} := \beta^{\lfloor (t-s)/B \rfloor}$. By Eq. (23), $\|\mathbf{A}_{t-1:s} - \mathbf{P}\|_2 \le \bar{h}_{t,s}$. Weighted Cauchy-Schwarz then gives

$$\frac{1}{m}\mathbb{E}\|(\mathbf{I} - \mathbf{P})\mathbf{X}^t\|_F^2 \le \mathcal{C}_B \sum_{s<t} \bar{h}_{t,s} \frac{1}{m}\mathbb{E}\|\mathbf{U}^s\|_F^2.$$

Applying Eq. (28) at $k = K$ and Eq. (29) gives

$$\begin{aligned}\frac{1}{m}\mathbb{E}\|\mathbf{U}^s\|_F^2 &\le 2\eta^2 K\sigma_\rho^2 + 2\eta^2 K \sum_{r=0}^{K-1} \frac{1}{m}\sum_i \mathbb{E}\|\nabla f_i(\mathbf{x}_{i,r}^s)\|^2 \\ &\le 2\eta^2 K\sigma_\rho^2 + 8\eta^2 K L^2 \mathcal{D}_s + 8\eta^2 K^2 \big(\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^s)\|^2 + L^2\mathcal{E}_s + \sigma_g^2\big),\end{aligned} \tag{35}$$

Substituting Lemma 1 and using $\eta LK \le 1/4$ yields

$$\frac{1}{m}\mathbb{E}\|\mathbf{U}^s\|_F^2 \le 16\eta^2 K\big[\sigma_\rho^2 + K\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^s)\|^2 + KL^2\mathcal{E}_s + K\sigma_g^2\big]. \tag{36}$$

Summing over t and exchanging the order of summation gives

$$\sum_{t=0}^{T-1} \mathcal{E}_t \le \mathcal{C}_B \sum_{s=0}^{T-1} \left(\frac{1}{m}\mathbb{E}\|\mathbf{U}^s\|_F^2\right) \sum_{t=s+1}^{T} \bar{h}_{t,s}.$$

For fixed s, set $\ell = t - s$. Using (23),

$$\sum_{t=s+1}^{T} \bar{h}_{t,s} = \sum_{\ell=1}^{T-s} \beta^{\lfloor \ell/B \rfloor} \le B\sum_{q=0}^{\infty} \beta^q = \mathcal{C}_B. \tag{37}$$

Hence $\sum_t \mathcal{E}_t \le \mathcal{C}_B^2 \sum_s m^{-1}\mathbb{E}\|\mathbf{U}^s\|_F^2$. Substituting Eq. (36) and using $\eta LK\mathcal{C}_B \le 1/8$ allows the resulting $\sum_t \mathcal{E}_t$ term to be absorbed into the left-hand side, which proves Eq. (31). □

C.5 Proof of Theorem 1

Proof. Since $\mathcal{C}_B \ge 1$, the condition $\eta LK \le [64(1 + \mathcal{C}_B)]^{-1}$ in Theorem 1 implies all step-size requirements used in Lemmas 1 and 2, as well as the intermediate descent bounds below.

Because $\mathbf{A}^t$ is doubly stochastic, averaging the local updates over all clients gives

$$\overline{\mathbf{x}}^{t+1} = \overline{\mathbf{x}}^t - \frac{\eta}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1} \widetilde{\mathbf{g}}_{i,k}^t. \tag{38}$$

Let $\mathbf{g}^t := \nabla f(\overline{\mathbf{x}}^t)$ and define

$$\mathbf{H}_t := \frac{1}{m}\sum_{i,k} \nabla f_i(\mathbf{x}_{i,k}^t) = K\mathbf{g}^t + \mathbf{R}_t,$$

$$\mathbf{R}_t := \frac{1}{m}\sum_{i,k} [\nabla f_i(\mathbf{x}_{i,k}^t) - \nabla f_i(\overline{\mathbf{x}}^t)].$$

The equality follows from $m^{-1}\sum_i \nabla f_i(\overline{\mathbf{x}}^t) = \nabla f(\overline{\mathbf{x}}^t)$. Further define $\overline{\mathbf{e}}_t := m^{-1}\sum_{i,k} \mathbf{e}_{i,k}^t$. When expanding its squared norm, cross terms across local steps of the same client vanish by Eq. (26), while cross-client terms vanish by independence and conditional zero mean. Therefore,

$$\mathbb{E}\|\overline{\mathbf{e}}_t\|^2 = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{k=0}^{K-1} \mathbb{E}\|\mathbf{e}_{i,k}^t\|^2 \le \frac{K\sigma_\rho^2}{m} \le K\sigma_\rho^2. \tag{39}$$

We next bound the local-drift remainder $\mathbf{R}_t$. By its definition,

$$\begin{aligned}\mathbb{E}\|\mathbf{R}_t\|^2 &= \mathbb{E}\left\|\sum_{k=0}^{K-1}\frac{1}{m}\sum_i[\nabla f_i(\mathbf{x}_{i,k}^t) - \nabla f_i(\overline{\mathbf{x}}^t)]\right\|^2 \\ &\le K\sum_{k=0}^{K-1}\frac{1}{m}\sum_i \mathbb{E}\|\nabla f_i(\mathbf{x}_{i,k}^t) - \nabla f_i(\overline{\mathbf{x}}^t)\|^2 \\ &\le L^2K\sum_{k=0}^{K-1}\frac{1}{m}\sum_i \mathbb{E}\|\mathbf{x}_{i,k}^t - \overline{\mathbf{x}}^t\|^2 \\ &\le 2L^2K(\mathcal{D}_t + K\mathcal{E}_t). \end{aligned} \tag{40}$$

The first inequality follows from Cauchy-Schwarz over the local steps and Jensen's inequality over clients, the second uses L-smoothness, and the last follows from $\|\mathbf{x}_{i,k}^t - \overline{\mathbf{x}}^t\|^2 \le 2\|\mathbf{x}_{i,k}^t - \mathbf{x}_i^t\|^2 + 2\|\mathbf{x}_i^t - \overline{\mathbf{x}}^t\|^2$.

Applying Young's inequality $2|\langle \mathbf{a}, \mathbf{b}\rangle| \le \tau\|\mathbf{a}\|^2 + \tau^{-1}\|\mathbf{b}\|^2$, with $\tau = K/2$ yields

$$|\langle \mathbf{g}^t, \mathbf{R}_t\rangle| \le \frac{K}{4}\|\mathbf{g}^t\|^2 + \frac{1}{K}\|\mathbf{R}_t\|^2. \tag{41}$$

By L-smoothness of f and Eq. (38),

$$\begin{aligned}\mathbb{E}f(\overline{\mathbf{x}}^{t+1}) &\le \mathbb{E}f(\overline{\mathbf{x}}^t) - \eta\mathbb{E}\langle \mathbf{g}^t, \mathbf{H}_t + \overline{\mathbf{e}}_t\rangle + \frac{L\eta^2}{2}\mathbb{E}\|\mathbf{H}_t + \overline{\mathbf{e}}_t\|^2 \\ &\le \mathbb{E}f(\overline{\mathbf{x}}^t) - \eta K\mathbb{E}\|\mathbf{g}^t\|^2 - \eta\mathbb{E}\langle \mathbf{g}^t, \mathbf{R}_t\rangle \\ &\quad + L\eta^2\mathbb{E}\|K\mathbf{g}^t + \mathbf{R}_t\|^2 + L\eta^2\mathbb{E}\|\overline{\mathbf{e}}_t\|^2 \\ &\le \mathbb{E}f(\overline{\mathbf{x}}^t) - \left(\frac{3\eta K}{4} - 2L\eta^2K^2\right)\mathbb{E}\|\mathbf{g}^t\|^2 \\ &\quad + \left(\frac{\eta}{K} + 2L\eta^2\right)\mathbb{E}\|\mathbf{R}_t\|^2 + L\eta^2K\sigma_\rho^2. \end{aligned} \tag{42}$$

Here, $\mathbb{E}\langle \mathbf{g}^t, \overline{\mathbf{e}}_t\rangle = 0$ by the conditional zero-mean property, Eq. (41) controls the remaining inner product, and $\|K\mathbf{g}^t + \mathbf{R}_t\|^2 \le 2K^2\|\mathbf{g}^t\|^2 + 2\|\mathbf{R}_t\|^2$.

If $\eta LK \le 1/8$, then

$$\frac{3\eta K}{4} - 2L\eta^2K^2 \ge \frac{\eta K}{2}, \qquad \frac{\eta}{K} + 2L\eta^2 \le \frac{2\eta}{K}.$$

Substituting Eq. (40) into Eq. (42) therefore gives

$$\mathbb{E}f(\overline{\mathbf{x}}^{t+1}) \le \mathbb{E}f(\overline{\mathbf{x}}^t) - \frac{\eta K}{2}\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^t)\|^2 + L\eta^2K\sigma_\rho^2 + 4L^2\eta\mathcal{D}_t + 4L^2\eta K\mathcal{E}_t. \tag{43}$$

We next substitute the local-update deviation bound from Lemma 1. Using Eq. (25),

$$4L^2\eta\mathcal{D}_t \le 8L^2\eta^3K^2\sigma_\rho^2 + 64L^2\eta^3K^3\left(\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^t)\|^2 + L^2\mathcal{E}_t + \sigma_g^2\right).$$

Under $\eta LK \le 1/16$, the resulting gradient term can be absorbed into the descent term in Eq. (43). Using the same step-size condition to simplify the remaining coefficients yields

$$\mathbb{E}f(\overline{\mathbf{x}}^{t+1}) \le \mathbb{E}f(\overline{\mathbf{x}}^t) - \frac{\eta K}{4}\mathbb{E}\|\mathbf{g}^t\|^2 + 2L\eta^2K\sigma_\rho^2 + 64\eta^3L^2K^3\sigma_g^2 + 5\eta L^2K\mathcal{E}_t. \tag{44}$$

Summing Eq. (44) over $t = 0, \dots, T-1$, using $f(\overline{\mathbf{x}}^T) \ge f^\star$, and dividing by $\eta KT/4$ gives

$$\begin{aligned} G_T &\le \frac{4(f(\overline{\mathbf{x}}^0) - f^\star)}{\eta KT} + 8L\eta\sigma_\rho^2 + 256\eta^2L^2K^2\sigma_g^2 + 20L^2E_T, \\ G_T &:= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^t)\|^2, \qquad E_T := \frac{1}{T}\sum_{t=0}^{T-1}\mathcal{E}_t. \end{aligned} \tag{45}$$

By Lemma 2, $E_T \le 32\eta^2K\mathcal{C}_B^2(\sigma_\rho^2 + K\sigma_g^2 + KG_T)$. Substituting this bound into (45), the coefficient of G_T on the right-hand side becomes $640\eta^2L^2K^2\mathcal{C}_B^2$. The step-size condition in Theorem 1 implies

$$640\eta^2L^2K^2\mathcal{C}_B^2 \le \frac{640}{64^2} < \frac{1}{4}.$$

Hence this term can be moved to the left-hand side. Using $(1 - 1/4)^{-1} \le 2$ and collecting the remaining terms gives Eq. (17). □

C.6 Proof of Corollary 1

Proof. Let $\eta = c/(L\sqrt{KT})$ for a sufficiently small constant $c > 0$ satisfying the condition in Theorem 1. For fixed c, this holds whenever $T = \Omega(K(1+\mathcal{C}_B)^2)$.

With this choice

$$\frac{1}{\eta KT} = \frac{L}{c\sqrt{KT}}, \qquad L\eta = \frac{c}{\sqrt{KT}}, \qquad L^2\eta^2 K = \frac{c^2}{T},$$

Substituting these relations into Eq. (17) yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^t)\|^2 \le \frac{8L\big(f(\overline{\mathbf{x}}^0) - f^\star\big)}{c\sqrt{KT}} + \frac{16c\,\sigma_\rho^2}{\sqrt{KT}} + \frac{1280c^2(1+\mathcal{C}_B^2)(\sigma_\rho^2 + K\sigma_g^2)}{T}.$$

Therefore, Eq. (18) follows. □

D Additional Experimental Details and Results

D.1 Baseline-Specific Configurations

We follow the original method designs and recommended hyperparameters where applicable, with necessary adaptations to our common experimental setup. The exact configurations are summarized as follows.

- **D-PSGD** (Lian et al., 2017). Each client performs one stochastic-gradient step on a randomly sampled local mini-batch per communication round, followed by neighbor mixing. This corresponds to one iteration of the original D-PSGD algorithm.
- **DFedAvg** (Sun et al., 2023). Each client performs five local training epochs before neighbor aggregation.
- **DisPFL** (Dai et al., 2022). We set the target sparsity to 0.5, retaining 50% of model parameters. The pruning rate is initialized at 0.5 and subsequently decays according to the original cosine-annealing schedule.
- **DFedSAM** (Shi et al., 2023). We set the SAM perturbation radius to $\rho = 0.1$. Each client performs five local epochs followed by one gossip step per communication round ($Q = 1$).
- **DFedADMM** (Li et al., 2025). We set the ADMM penalty parameter to $\lambda = 0.1$, controlling the proximal penalty and associated auxiliary and dual updates.
- **DecFus** (Yang et al., 2026). We set the cutoff bounds to $\tau_u = 0.75$ and $\tau_l = 0.25$, the stability threshold to $\epsilon_s = 0.35$, and the sigmoid parameters to $\alpha = 20.0$ and $\beta = 0.4$. We use $\kappa = 0.5$ for the transition duration, a sliding window of $w = 20$ rounds for the neighborhood loss variance, and the first 100 rounds as the early-training reference interval. During layer exchange, the least-similar neighbor is selected with probability $\gamma = 0.4$.

D.2 Sensitivity to Noise Intensity, Local Computation, and Network Size

We further examine the sensitivity of NoRA to noise intensity, local update steps, and the number of clients on CIFAR-10 under Dirichlet $\alpha = 0.3$.

Noise intensity ρ. The parameter ρ controls the magnitude of noise injected during local optimization. Figure 7a compares six values of ρ ranging from 0.01 to 0.3. Smaller values yield faster initial convergence, whereas excessive noise substantially slows training and reduces final test accuracy. Among the evaluated settings, $\rho = 0.1$ achieves the highest final accuracy, providing a favorable balance between convergence speed and predictive performance.

Local Iteration Steps K. Following Shi et al. (2023), we fix the mini-batch size and vary the number of local training epochs to examine the effect of K on convergence across communication rounds. As shown in Fig. 7b, increasing K accelerates empirical convergence, with diminishing gains at larger values. This trend is consistent with Corollary 1, which characterizes the trade-off

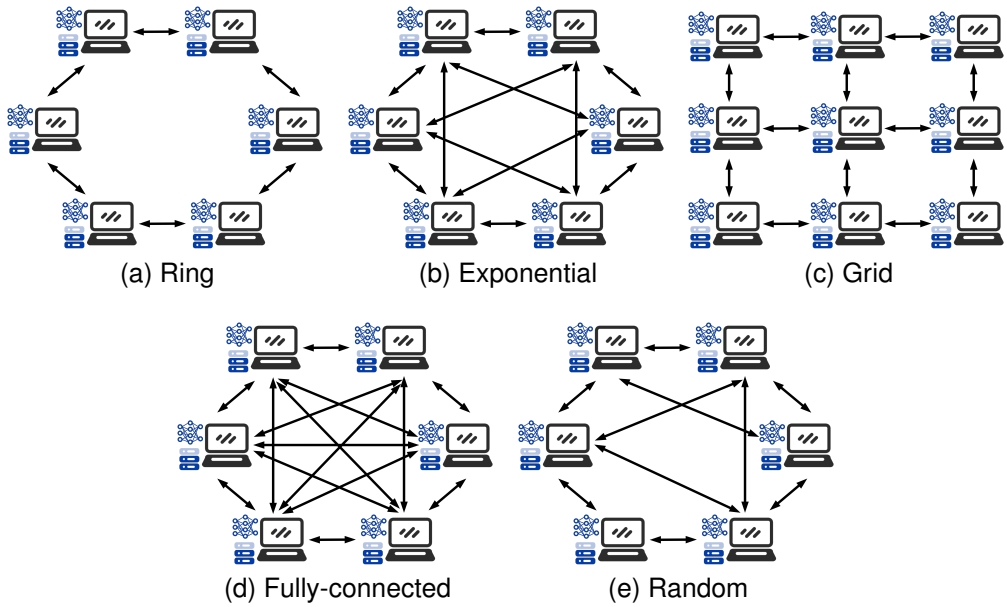


Figure 6: Schematic illustrations of the five communication topologies considered in our experiments.

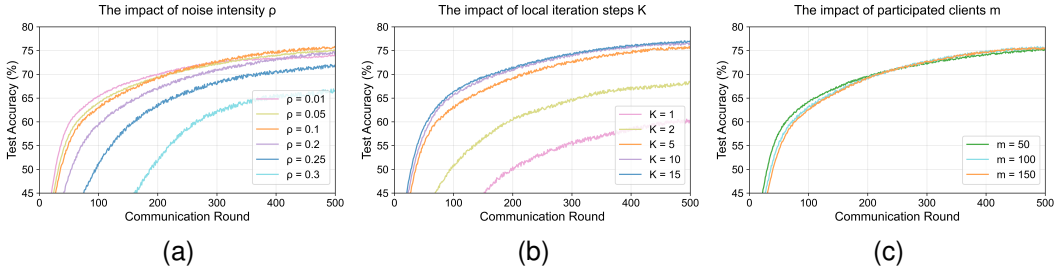


Figure 7: Sensitivity to noise intensity ρ, local iterations K and participating clients m. Here, K denotes the number of local training epochs.

between the leading $\mathcal{O}(1/\sqrt{KT})$ term and the K-dependent heterogeneity and mixing contribution $\mathcal{O}((1+\mathcal{C}_B^2)K\sigma_g^2/T)$. We adopt five local epochs to balance communication efficiency and local computation.

Number of Clients m. Fig. 7c evaluates different network sizes, with $m \in \{50, 100, 150\}$. The three configurations exhibit similar convergence trajectories and comparable final test accuracy, suggesting that NoRA maintains stable predictive performance as the number of clients increases. These observations provide empirical evidence of its scalability within the evaluated network sizes.