AVERAGING MODELS IS NOT AVERAGING SPECTRA: STEERING FEDERATED REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning averages model parameters, whereas representation diversity belongs to the features produced by the aggregated model. We study the gap between these two objects through class-conditional von Neumann entropy (VNE). Exact linear constructions reveal a two-sided phenomenon: parameter averaging can either increase or decrease VNE relative to the local encoders. This spectral ambiguity leads to a second question—how far can the aggregated representation depart from centralized training? Under explicit regularity and heterogeneity assumptions, we bound the one-round diagnostic deviation from a same-start centralized reference, including rank-deficient empirical covariances. We then turn from diagnosis to control: interpolation toward a shared class-conditional density matrix contracts a weighted spectral disagreement, motivating the class-conditional FedVNA regularizer. On CIFAR-10 with ResNet-18 and Dirichlet $\alpha = 0.1$, FedVNA raises final VNE from FedAvg's 3.0655 to 4.0782, while their best accuracies remain close (0.7464 and 0.7392). Together, the results show that federated spectra can be diagnosed and deliberately shaped, while spectral diversity and predictive performance remain distinct outcomes.

1 INTRODUCTION

Averaging models does not mean averaging their representation geometry. FedAvg (McMahan et al., 2017) combines client parameters, but a representation spectrum is observed only after the resulting model acts on data. This distinction produces a simple paradox. If one client's encoder keeps only the first feature coordinate and another keeps only the second, each representation is collapsed onto a line, yet their average spans the plane. If two encoders differ only in the sign of one coordinate, each representation is fully spread, yet their average discards that coordinate. The direction of spectral change is therefore a property of how client encoders interact, not an automatic consequence of averaging.

The question matters because the representation spectrum is where client heterogeneity leaves a visible trace. Under label skew, local and global models can undergo dimensional collapse, with features concentrating in a few directions (Shi et al., 2023), and a growing line of federated methods targets representation geometry directly through unsupervised, multimodal, probabilistic, angle-based, or drift-aware alignment objectives (Liao et al., 2024; Yi et al., 2024; Wang et al., 2024d; Chen et al., 2024a; Yi et al., 2025; Nguyen et al., 2025). Aggregation methods, meanwhile, refine what the server combines (Fanì et al., 2024; Chen et al., 2024b; Liu et al., 2024b; Shi et al., 2025). Both lines rely on a prior question that neither answers: what does parameter averaging itself do to the spectrum of the deployed encoder?

We organize the answer around *direction* and *magnitude*. For direction, an exact covariance decomposition and two opposing linear constructions show that parameter averaging admits no universal VNE trend. For magnitude, we compare full-participation local training with a centralized reference restarted from the same round initialization. Contemporary analyses connect heterogeneity and local computation to optimization drift (Wang et al., 2024c; Crawshaw & Liu, 2024; Chen & Vikalo, 2024; Li et al., 2024; Guo & Yang, 2024; Ying et al., 2025); our comparison carries this viewpoint to a class-conditional spectral diagnostic based on centered, trace-normalized covariance (Kim et al., 2023). The resulting one-round bound remains valid when empirical covariance is rank deficient, explains how local trajectory separation reaches the representation spectrum, and, applied also to

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.

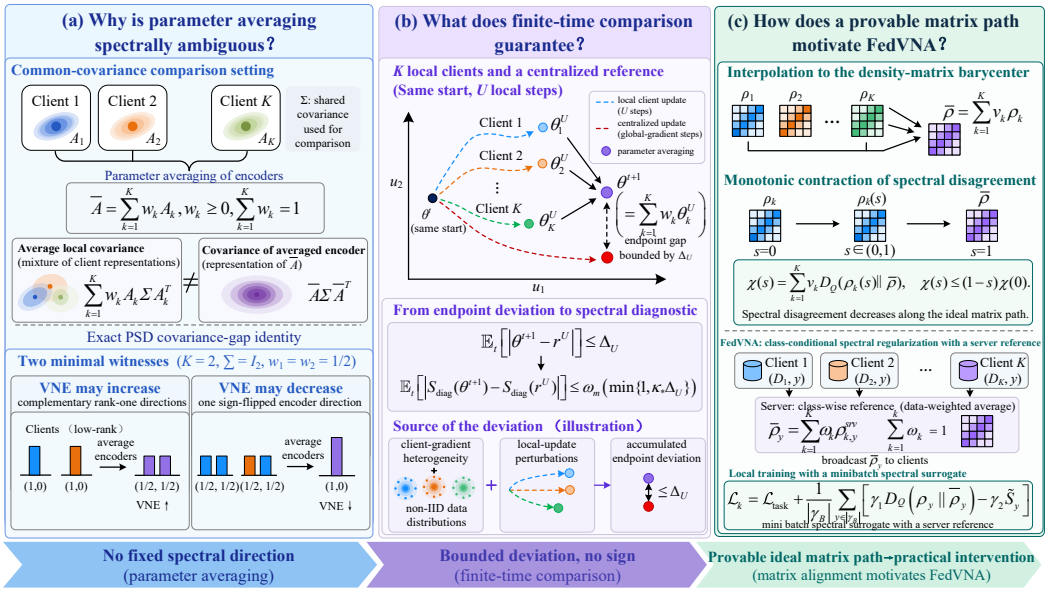

Figure 1: Overview of the theoretical and methodological progression for K participating clients. (a) Parameter averaging is distinct from averaging the clients' induced covariances and can either increase or decrease VNE. (b) A same-start, within-round comparison bounds the deviation of the aggregated model from a centralized gradient reference, and entropy continuity transfers this control to the spectral diagnostic without determining its sign. (c) Spectral disagreement contracts along an explicitly defined interpolation toward the density-matrix barycenter; FedVNA translates this shared-reference principle into a minibatch spectral surrogate.

the local models, bounds the size of the averaging-induced change whose sign the two constructions leave open.

The third question is *control*. If parameter averaging supplies no preferred spectral direction, a reliable intervention must act on representation objects themselves. Existing methods exchange prototypes, align feature distributions, or regularize covariance and uniformity (Tan et al., 2022; Shi et al., 2023; Son et al., 2024; McLaughlin & Su, 2024; Goswami et al., 2025; Liao et al., 2025a). We identify a matrix operation with a definite direction: moving class-conditioned client density matrices toward their shared reference contracts weighted spectral disagreement. FedVNA translates this shared-reference principle into a class-conditional minibatch objective that combines spectral alignment with an entropy incentive.

On CIFAR-10 with ResNet-18, FedVNA raises final VNE over FedAvg under strong label heterogeneity while accuracy stays close, and the VNE ordering of the spectral baselines changes with heterogeneity: spectral controllability and predictive utility are related but distinct outcomes.

Our contributions follow the same three stages. *Direction:* an exact covariance identity and two 2×2 witnesses show that parameter averaging can raise or lower class-conditional VNE, so any spectral effect of averaging is a property of how client encoders interact. *Magnitude:* a same-start, within-round comparison bounds how far the averaged model's diagnostic can move from a centralized reference, and hence how large the averaging-induced change can be, in terms of client heterogeneity and local updates. *Control:* interpolation toward the class-conditional density-matrix barycenter contracts weighted spectral disagreement; FedVNA turns this shared-reference principle into a minibatch regularizer that moves the CIFAR-10 spectrum while keeping accuracy close to FedAvg. Figure 1 summarizes the progression.

Scope. The two constructions are algebraic parameter endpoints, not trajectories of a particular non-IID optimizer. The finite-time comparison restarts from the same parameters each round and bounds a magnitude; it neither orders the two entropies nor establishes long-run contraction. The contraction result concerns an explicitly defined matrix path, which FedVNA approximates with a minibatch surrogate rather than follows exactly.

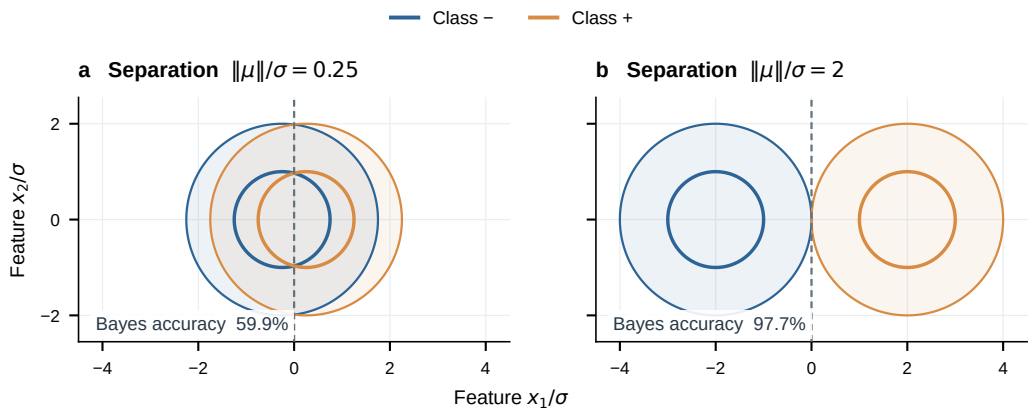

Figure 2: Equal within-class VNE does not fix classification accuracy. Both two-class Gaussian examples have centered covariance $\sigma^2 I_2$, density matrix $I_2/2$, and VNE $\log 2$. Changing only $\|\mu\|_2/\sigma$ from 0.25 to 2 changes the equal-prior Bayes accuracy from 59.9% to 97.7%. Contours are analytic density levels; no training data are plotted.

2 Class-conditional spectral diagnostic

The object of interest is the representation produced by a specified model, evaluated on the same observations throughout a comparison. Within a class, a concentrated covariance spectrum means that most feature variation lies in a few directions; a diffuse spectrum means that variation occupies more directions. Neither statement says whether the class is separated from another class. Conditioning on client and class keeps these two questions apart before the client-level results are combined.

Fix a finite set $\mathcal{A}$ of client–class pairs $a = (k, y)$, with $n_a \geq 2$ fixed diagnostic samples for each pair. Their representations form $Z_a(\theta) \in \mathbb{R}^{n_a \times m}$, whose ith row is denoted by $z_{a,i}(\theta)^\top$. For matrices, $\|\cdot\|_1$ denotes the Schatten trace norm and $\|\cdot\|_F$ the Frobenius norm. Let $H_a = I - \mathbf{1}\mathbf{1}^\top/n_a$ and $\widetilde{Z}_a = H_a Z_a$. Define

$$C_a(\theta) = \frac{\widetilde{Z}_a(\theta)^\top \widetilde{Z}_a(\theta)}{n_a - 1}, \qquad G_a(\theta) = \frac{\widetilde{Z}_a(\theta)\widetilde{Z}_a(\theta)^\top}{n_a - 1}, \qquad \rho_a(\theta) = \frac{C_a(\theta)}{\operatorname{Tr} C_a(\theta)}. \tag{1}$$

We use only pairs for which $\operatorname{Tr} C_a(\theta) > 0$ on the parameter region in question. With fixed $q_a \geq 0$ satisfying $\sum_{a \in \mathcal{A}} q_a = 1$, and with the equivalent notation $q_{k,y} := q_{(k,y)}$, our global-model diagnostic is

$$S_{\mathrm{diag}}(\theta) = \sum_{a \in \mathcal{A}} q_a S(\rho_a(\theta)), \qquad S(\rho) = -\operatorname{Tr}(\rho \log \rho), \tag{2}$$

where $0 \log 0 = 0$. We call the trace-normalized covariance ρ_a the class-conditional *density matrix*; it is positive semidefinite with unit trace, so $S(\rho_a)$ is its von Neumann entropy. C_a and G_a share their nonzero eigenvalues and trace, so their normalized entropies agree even when $n_a - 1 < m$; the appendix proof of Lemma 3 records how ridge and logarithm regularization modify this identity.

What the diagnostic captures. Class-conditioned VNE describes how broadly feature energy is distributed within each class. Class separation supplies a complementary view of how distinct those classes are. To make the distinction transparent, let $X \mid Y = \pm 1 \sim \mathcal{N}(\pm\mu, \sigma^2 I_2)$ with equal class priors. Centering within either class gives density matrix $I_2/2$ and VNE $\log 2$ for every μ, whereas the Bayes accuracy is $\Phi(\|\mu\|_2/\sigma)$, where Φ is the standard-normal CDF. Figure 2 changes only this separation while preserving the within-class spectrum. We therefore report spectral diversity and task accuracy as complementary outcomes throughout the paper.

3 DIRECTION: SPECTRAL AMBIGUITY OF PARAMETER AVERAGING

The mixture $\sum_k v_k \rho_k$ of client density matrices and the density matrix $\rho(\sum_k w_k \theta_k)$ of the parameter-averaged encoder are different objects. The following exact calculation exposes the distinction in a linear model.

Proposition 1 (Covariance identity for a shared linear diagnostic)**.** *Suppose every encoder acts on the same centered input distribution with covariance $\Sigma \succeq 0$ through $z = A_k x$. For $w_k \geq 0$, $\sum_k w_k = 1$, and $\bar{A} = \sum_k w_k A_k$,*

$$\sum_k w_k A_k \Sigma A_k^\top - \bar{A}\Sigma\bar{A}^\top = \sum_k w_k (A_k - \bar{A})\Sigma(A_k - \bar{A})^\top \succeq 0. \tag{3}$$

The Loewner inequality does not impose an ordering on entropy after trace normalization.

Proposition 2 (No universal entropy direction under parameter averaging)**.** *Even for two equally weighted linear encoders evaluated on the same centered class distribution with covariance I_2, the von Neumann entropy of the parameter-averaged encoder can be either higher or lower than the mean entropy of the two local encoders.*

For a nonzero linear encoder, write $\rho(A) = AA^\top / \operatorname{Tr}(AA^\top)$. For equal client weights, set $\bar{A} = (A_1 + A_2)/2$ and $\bar{S}_{\mathrm{loc}} = \{S(\rho(A_1)) + S(\rho(A_2))\}/2$. Then

$$\begin{aligned} &(A_1, A_2) = (\operatorname{diag}(1,0), \operatorname{diag}(0,1)), \quad &&(\bar{S}_{\mathrm{loc}}, S(\rho(\bar{A}))) = (0, \log 2), \\ &(A_1, A_2) = (I_2, \operatorname{diag}(1,-1)), &&(\bar{S}_{\mathrm{loc}}, S(\rho(\bar{A}))) = (\log 2, 0). \end{aligned} \tag{4}$$

These are the two encoder pairs from the introduction and Figure 1(a): complementary rank-one directions spread the averaged spectrum, and a sign flip collapses it. In both cases $S((\rho(A_1) + \rho(A_2))/2) = \log 2$: entropy concavity applies to the mixture of density matrices, not to the density matrix of averaged parameters. The full calculation appears in the appendix proof of Proposition 2.

4 MAGNITUDE: FINITE-TIME COMPARISON WITH A CENTRALIZED REFERENCE

Spectral ambiguity changes the next question from direction to deviation. We compare two updates that begin at the same parameter vector: clients take U local optimizer steps and average their endpoints, while a reference model takes U steps on the weighted global objective. Both branches restart at the beginning of each communication round, making the centralized branch a common within-round reference. We first control how a change in parameters can alter the diagnostic, then bound the separation of the two updates.

For $m \geq 2$ let $h_2(u) = -u \log u - (1-u)\log(1-u)$ and define the clipped entropy modulus

$$\omega_m(u) = \begin{cases} u \log(m-1) + h_2(u), & 0 \leq u \leq 1 - 1/m, \\ \log m, & 1 - 1/m < u \leq 1. \end{cases} \tag{5}$$

Assumption 1 (Diagnostic regularity)**.** *On a convex parameter region $\mathcal{R}_t$, the fixed diagnostic features satisfy $\|z_{a,i}(\theta)\|_2 \leq R$ and $\|z_{a,i}(\theta') - z_{a,i}(\theta)\|_2 \leq L_z \|\theta' - \theta\|_2$. For every $a \in \mathcal{A}$ and $\theta \in \mathcal{R}_t$, $\operatorname{Tr} C_a(\theta) \geq \tau_a > 0$. This lower trace bound excludes the zero covariance but permits rank deficiency.*

Lemma 3 (Entropy continuity with rank deficiency)**.** *Under Assumption 1, let*

$$L_{C,a} = \frac{4n_a}{n_a - 1} R L_z, \qquad \kappa_a = \frac{\sqrt{m}\, L_{C,a}}{\tau_a}. \tag{6}$$

Then, for $\theta, \theta' \in \mathcal{R}_t$,

$$\begin{aligned} \frac{1}{2}\|\rho_a(\theta') - \rho_a(\theta)\|_1 &\leq \min\{1, \kappa_a \|\theta' - \theta\|_2\}, \\ |S(\rho_a(\theta')) - S(\rho_a(\theta))| &\leq \omega_m(\min\{1, \kappa_a \|\theta' - \theta\|_2\}). \end{aligned} \tag{7}$$

Assumption 2 (Idealized full-participation local updates)**.** *Let $U \geq 1$ be an integer, $w_k > 0$, $\sum_k w_k = 1$, $L = \sum_k w_k L_k$, and let each $\eta_e \geq 0$ be deterministic for $e = 0, \ldots, U-1$. Every L_k has β-Lipschitz gradient on $\mathcal{R}_t$. Starting from θ^t, local and reference iterates obey*

$$\theta_k^{e+1} = \theta_k^e - \eta_e[\nabla L_k(\theta_k^e) + \xi_k^e], \qquad r^{e+1} = r^e - \eta_e \nabla L(r^e), \qquad \theta_k^0 = r^0 = \theta^t. \tag{8}$$

For every $\theta \in \mathcal{R}_t$, $\sum_k w_k \|\nabla L_k(\theta) - \nabla L(\theta)\|_2^2 \le \zeta^2$. For round t, let $(\mathcal{G}_{t,e})_{e=0}^U$ be an increasing filtration whose initial sigma-field $\mathcal{G}_{t,0}$ contains the round-start information. The variables θ^t, θ_k^e, and r^e are $\mathcal{G}_{t,e}$-measurable; ξ_k^e is $\mathcal{G}_{t,e+1}$-measurable and satisfies

$$\sum_k w_k \mathbb{E}[\|\xi_k^e\|_2^2 \mid \mathcal{G}_{t,e}] \le \nu_e^2 \qquad \textit{almost surely}, \tag{9}$$

for deterministic $\nu_e \ge 0$. All θ_k^e, r^e, and the endpoint $\sum_k w_k \theta_k^U$ lie in $\mathcal{R}_t$ almost surely.

Theorem 4 (One-round spectral diagnostic deviation)**.** *Under Assumptions 1–2, define $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{G}_{t,0}]$, $\Gamma_0 = 0$, and*

$$\Gamma_{e+1} = (1 + \beta\eta_e)\Gamma_e + \eta_e(\zeta + \nu_e), \tag{10}$$

$$\Delta_U = \sum_{e=0}^{U-1} \eta_e(\beta\Gamma_e + \nu_e). \tag{11}$$

For the actual parameter average $\theta^{t+1} = \sum_k w_k \theta_k^U$, let $\kappa_* = \max_{a \in \mathcal{A}} \kappa_a$. Then

$$\mathbb{E}_t \left\|\theta^{t+1} - r^U\right\|_2 \le \Delta_U, \qquad \mathbb{E}_t \left|S_{\mathrm{diag}}(\theta^{t+1}) - S_{\mathrm{diag}}(r^U)\right| \le \omega_m(\min\{1, \kappa_* \Delta_U\}). \tag{12}$$

Both conditional inequalities in (12) *hold almost surely.*

Proof roadmap. The proof separates two quantities that are easily conflated. First, the clientwise deviations $\theta_k^e - r^e$ obey a recursion driven directly by heterogeneity and update perturbations, producing Γ_e. Second, after averaging the client endpoints, the same-point heterogeneity terms cancel because the global objective and the server average use the same weights. Only the smoothness-amplified local spread and perturbations remain, producing Δ_U. The diagnostic regularity lemma then maps the parameter deviation to a trace-distance bound between normalized covariances, and the clipped Fannes–Audenaert inequality maps trace distance to entropy difference. This last step accommodates zero eigenvalues. All intermediate inequalities and the conditional-expectation argument are given in the appendix.

For a distinct mean local-model diagnostic, let $S_{\mathrm{local}}(\{\theta_k\}) = \sum_{(k,y) \in \mathcal{A}} q_{k,y} S(\rho_{k,y}(\theta_k))$ and $q_k = \sum_{y:(k,y) \in \mathcal{A}} q_{k,y}$. The same proof yields, almost surely,

$$\mathbb{E}_t \left|S_{\mathrm{local}}(\{\theta_k^U\}) - S_{\mathrm{diag}}(r^U)\right| \le \omega_m(\min\{1, \kappa_* c_q \Gamma_U\}), \qquad c_q = \max_{k: w_k > 0} \frac{q_k}{w_k}. \tag{13}$$

Only when $q_k = w_k$ for all k may c_q be dropped. Combining the two bounds returns to the quantity whose sign Section 3 left open.

Corollary 5 (Size of the averaging-induced spectral change)**.** *Under the assumptions of Theorem 4, almost surely,*

$$\mathbb{E}_t \left|S_{\mathrm{diag}}(\theta^{t+1}) - S_{\mathrm{local}}(\{\theta_k^U\})\right| \le \omega_m(\min\{1, \kappa_* \Delta_U\}) + \omega_m(\min\{1, \kappa_* c_q \Gamma_U\}). \tag{14}$$

Proposition 2 shows that this difference has no fixed sign; Corollary 5 shows that its size is nonetheless controlled by the same heterogeneity and local-update quantities.

Constant steps and special cases. For constant $\eta_e = \eta$, $\nu_e = 0$ and $\beta > 0$, the theorem gives

$$\Delta_U = \frac{\zeta}{\beta}\left[(1 + \beta\eta)^U - 1 - U\beta\eta\right] = \tfrac{1}{2}\beta\zeta\eta^2 U(U-1) + O(\eta^3). \tag{15}$$

For fixed U, β, and ζ, the expansion is taken as $\eta \to 0$. Thus the reference and noiseless full-participation FedAvg agree exactly at $U = 1$ under Assumption 2. If $\beta = 0$ and $\nu_e = 0$, $\Delta_U = 0$ for every U. With nonzero perturbations, the endpoint bound can remain positive even at $U = 1$. Its numerical tightness depends on trajectory-level constants, and the entropy modulus reaches its ceiling $\log m$ when $\kappa_* \Delta_U \ge 1 - 1/m$.

Interpretation. In the small-local-drift regime $\beta\eta U \ll 1$, the leading noiseless term scales as $\eta^2 U(U-1)$; for large U, the exact geometric expression above must be used.

5 CONTROL: A DIRECTIONAL MATRIX INTERVENTION

The absence of a fixed spectral direction under parameter averaging suggests intervening on a spectral object more directly. Suppose clients can express their class-conditioned covariances in a common feature space. Their density matrices then have a barycenter, and the gap between the entropy of that barycenter and the clients' mean entropy measures spectral disagreement. Unlike the entropy of a parameter-averaged encoder, this gap has a tractable behavior under an explicitly defined matrix operation: interpolation toward the shared barycenter.

For density matrices, define $D_{\mathrm{Q}}(\rho\|\sigma) = \mathrm{Tr}[\rho(\log\rho - \log\sigma)]$ when $\mathrm{supp}(\rho) \subseteq \mathrm{supp}(\sigma)$, and $+\infty$ otherwise; logarithms of singular matrices are evaluated on their supports.

Proposition 6 (Idealized class-conditional alignment)**.** *Let ρ_k be density matrices of a common dimension, $v_k > 0$ with $\sum_k v_k = 1$, and $\bar\rho = \sum_k v_k\rho_k$. For $0 \le s \le 1$, set $\rho_k(s) = (1-s)\rho_k + s\bar\rho$. The barycenter remains $\bar\rho$. With*

$$\chi(s) = S(\bar\rho) - \sum_k v_k S(\rho_k(s)) = \sum_k v_k D_{\mathrm{Q}}(\rho_k(s)\|\bar\rho), \tag{16}$$

one has $\chi(s) \le (1-s)\chi(0)$. More generally, for $0 \le s_1 < s_2 < 1$,

$$\chi(s_2) \le \frac{1-s_2}{1-s_1}\chi(s_1). \tag{17}$$

Consequently, the weighted mean entropy is nondecreasing along this matrix path, although no individual client's entropy need be.

This proposition supplies the direction missing from parameter averaging: moving each client toward the population's spectrum for the same class contracts their weighted disagreement.

6 FEDVNA: TRANSLATING MATRIX CONTROL INTO TRAINING

FedVNA operationalizes the matrix-level motivation by providing each client with a server-aggregated, class-conditioned spectral reference. Local training then balances the task loss with a spectral surrogate: one term encourages within-class feature diversity, while a reference-dependent term responds to the shared class spectrum.

Fix ridge and clipping levels $\epsilon_{\mathrm{r}} > 0$ and $\epsilon_{\mathrm{c}} > 0$. For a class represented by $n \ge 2$ features in the minibatch, let $W \in \mathbb{R}^{n\times m}$ be centered, $C = W^\top W/(n-1)$, and

$$\widetilde G_{\epsilon_{\mathrm{r}}} = \frac{WW^\top/(n-1) + \epsilon_{\mathrm{r}} I_n}{\mathrm{Tr}\, C + n\epsilon_{\mathrm{r}}}, \qquad \widetilde S_{\epsilon_{\mathrm{r}},\epsilon_{\mathrm{c}}} = -\sum_j \max\{\lambda_j(\widetilde G_{\epsilon_{\mathrm{r}}}), \epsilon_{\mathrm{c}}\} \log\max\{\lambda_j(\widetilde G_{\epsilon_{\mathrm{r}}}), \epsilon_{\mathrm{c}}\}. \tag{18}$$

The server averages local $m \times m$ covariance density matrices from the final local epoch into a common feature-space reference $\bar\rho_y$. With K participating clients, write $\rho_{k,y}^{\mathrm{srv}}$ for the uploaded density matrix and let w_k be the server aggregation weight; the reference is

$$\bar\rho_y = \sum_{k=1}^K w_k \rho_{k,y}^{\mathrm{srv}}, \qquad \sum_{k=1}^K w_k = 1. \tag{19}$$

If $\bar\rho_y = Q_y\,\mathrm{diag}(\lambda_{y,1},\dots,\lambda_{y,m})Q_y^\top$, define its eigenvalue-clipped matrix logarithm by

$$\Lambda_y = Q_y\,\mathrm{diag}(\log\max\{\lambda_{y,1},\epsilon_{\mathrm{c}}\},\dots,\log\max\{\lambda_{y,m},\epsilon_{\mathrm{c}}\})\,Q_y^\top. \tag{20}$$

For the valid class set $\mathcal{Y}_B$ in a minibatch, let $W_y \in \mathbb{R}^{n_y\times m}$ be the centered feature matrix and $C_y = W_y^\top W_y/(n_y-1)$. With coefficients $\gamma_1, \gamma_2 \ge 0$, FedVNA adds to the task loss

$$\frac{1}{|\mathcal{Y}_B|}\sum_{y\in\mathcal{Y}_B}\left[\gamma_1\left(-\widetilde S_{\epsilon_{\mathrm{r}},\epsilon_{\mathrm{c}},y} - \frac{\mathrm{Tr}(C_y\Lambda_y)}{\mathrm{Tr}\, C_y + n_y\epsilon_{\mathrm{r}}}\right) - \gamma_2\widetilde S_{\epsilon_{\mathrm{r}},\epsilon_{\mathrm{c}},y}\right]. \tag{21}$$

The γ_1 bracket mirrors the decomposition $D_{\mathrm{Q}}(\rho\|\bar\rho) = -S(\rho) - \mathrm{Tr}(\rho\log\bar\rho)$ of each summand of χ in Proposition 6, with the server reference (19) in the role of the barycenter for $v_k = w_k$; the γ_2 term

adds a direct entropy incentive. Because the entropy is computed on a regularized $n_y \times n_y$ Gram matrix while the cross term uses an $m \times m$ feature covariance, the bracket is a spectral *surrogate* rather than an exact relative entropy. This form makes a class-conditioned intervention feasible in minibatch training while preserving a shared feature-space reference. Algorithm 1 in Appendix D summarizes one round; the appendix also gives the communication cost and the value of ϵ_{r} and ϵ_{c}.

7 Experiments

The experiments ask four questions, each tied to one part of the analysis. Q1 and Q2 test the practical side of Proposition 6: does the shared-reference surrogate move the class-conditional spectrum, and does the movement persist across rounds and heterogeneity levels? Q3 tests the separation illustrated in Section 2: must a spectral change carry an accuracy change? Q4 asks whether the alignment term acts on the disagreement χ that Proposition 6 contracts. The local-computation study at the end of this section is a qualitative check on the role of U in Theorem 4, not a numerical test of the bound. CIFAR-10 has 50,000 training and 10,000 test images. A ResNet-18 encoder (He et al., 2016) produces $m = 512$-dimensional features. We use $K = 10$ clients, which receive label-skewed training partitions drawn with Dirichlet concentration $\alpha \in \{0.1, 0.5, 1.0\}$; smaller α represents stronger label heterogeneity. All runs use 100 communication rounds, five local epochs per round, full participation, batch size 64, and three seeds. Local SGD uses momentum 0.9 and a cosine learning-rate schedule from 10^{-2} to 10^{-4}. The main FedVNA runs use $\gamma_1 = 10^{-3}$ and $\gamma_2 = 10^{-2}$. We compare against FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020), MOON (Li et al., 2021), FedTopo (Hu et al., 2026), and the additional spectral baselines FedDecorr (Shi et al., 2023) and FedUV (Son et al., 2024).

Reporting protocol. Test accuracy is evaluated on the global test set after every round. We report the mean and sample standard deviation across three seeds; "final" denotes round 100, and "best" is selected within each run before averaging. Spectra are reported as the class-conditional VNE of centered features (Section 2) during training. Appendix D gives further details.

	Method	$\alpha = 0.1$			$\alpha = 0.5$			$\alpha = 1.0$		
		Final VNE	Best acc.	Final acc.	Final VNE	Best acc.	Final acc.	Final VNE	Best acc.	Final acc.
Optim.	FedAvg	3.0655 (−) ±.0934	.7464 ±.0382	.6946 ±.0504	2.6931 (−) ±.0276	.8688 ±.0076	.8584 ±.0096	2.8896 (−) ±.0460	.8804 ±.0017	.8705 ±.0040
Optim.	FedProx	3.0262 (−) ±.0777	.7406 ±.0443	.7077 ±.0338	2.6940 (−) ±.0221	.8689 ±.0037	.8566 ±.0068	2.8725 (−) ±.0703	.8772 ±.0007	.8689 ±.0029
Repr.	MOON	3.1975 (−) ±.1122	.7232 ±.0464	.7064 ±.0524	2.7653 (≈) ±.2182	.8655 ±.0067	.8603 ±.0072	2.5039 (−) ±.0945	.8730 ±.0051	.8685 ±.0060
Repr.	FedTopo	3.0465 (−) ±.0809	.7398 ±.0472	.6939 ±.0626	2.7077 (−) ±.0216	.8705 ±.0049	.8600 ±.0050	2.8946 (−) ±.0629	.8771 ±.0012	.8688 ±.0049
Spec.	FedDecorr	3.4844 (−) ±.1877	.7005 ±.0350	.6516 ±.0339	3.4459 (+) ±.0210	.8263 ±.0075	.8067 ±.0031	3.5689 (+) ±.0457	.8458 ±.0029	.8311 ±.0100
Spec.	FedUV	3.2655 (−) ±.1410	.7359 ±.0452	.7035 ±.0595	2.6675 (−) ±.1420	.8663 ±.0071	.8598 ±.0049	2.6484 (−) ±.0560	.8766 ±.0015	.8699 ±.0008
	FedVNA (ours)	**4.0782** ±.0457	**.7392** ±.0432	**.6951** ±.0621	**3.1473** ±.0496	**.8684** ±.0058	**.8592** ±.0075	**3.2350** ±.0716	**.8776** ±.0019	**.8713** ±.0033

Table 1: CIFAR-10/ResNet-18 summary over three seeds (mean, with the sample standard deviation in grey). Best accuracy is selected within each run; final VNE and final accuracy are measured at round 100. Baselines are grouped into optimization-side (Optim.), representation-side (Repr.), and spectral (Spec.) methods. (−)/(+): final VNE lower/higher than FedVNA under a two-sided paired t-test across seeds ($p < 0.05$); (≈): no significant difference.

Q1: Does the intervention change the spectrum? Table 1 shows that at $\alpha = 0.1$ FedVNA has the largest final VNE: 4.0782, compared with 3.0655 for FedAvg and 3.4844 for FedDecorr. Its best accuracy is 0.7392, compared with FedAvg's 0.7464, while their final accuracies are essentially tied (0.6951 and 0.6946). The spectrum therefore changes much more strongly than task performance. At $\alpha = 0.5$ and 1.0, FedDecorr has higher VNE than FedVNA (3.4459 versus 3.1473, and 3.5689 versus 3.2350), but lower accuracy (0.8263 versus 0.8684, and 0.8458 versus 0.8776). FedVNA and FedUV have closer accuracy at $\alpha = 0.5$ and 1.0, while their VNE differs. These contrasts show that the intervention reaches the representation spectrum and that spectral and predictive orderings can differ.

Q2: Does the change persist? The trajectories in Figure 3 show that the spectral ordering changes with heterogeneity: FedVNA is highest at $\alpha = 0.1$, whereas FedDecorr is highest at $\alpha = 0.5$ and 1.0. The aligned contrasts in Figure 8 (Appendix C) place these changes beside task performance,

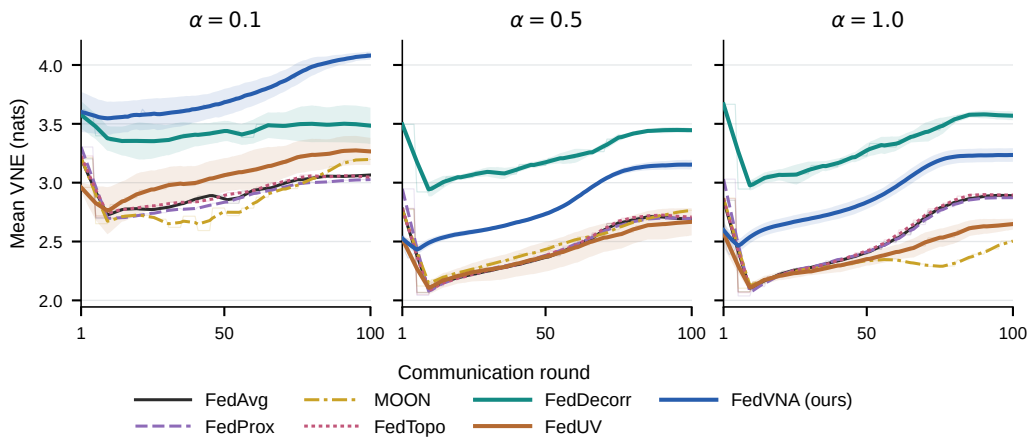

Figure 3: Mean VNE over 100 rounds at three heterogeneity levels. Dark lines are 9-round moving averages of the three-seed mean; faint lines show the raw mean. Shading shows one (smoothed) standard deviation for FedVNA, FedDecorr, and FedUV.

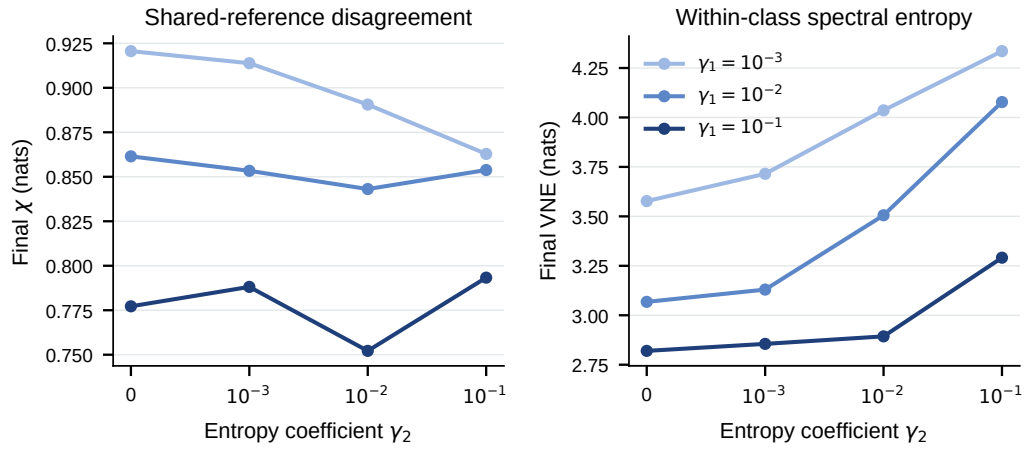

Figure 4: Final disagreement χ (left) and final VNE (right) in the coefficient sweep at $\alpha = 0.1$ (one seed), for the three nonzero alignment coefficients. The $\gamma_1 = 0$ settings are not shown because their client matrices lie in a 64-dimensional feature space, so their χ is not comparable.

where the ordering need not agree. The coefficient sweep in Appendix B, whose grid contains the main configuration, shows how the alignment and entropy terms trade off.

Q3: How does task performance respond? Table 1 and Figure 8 give the same qualitative answer under both best and final accuracy: substantial spectral changes can coexist with small, absent, or negative task changes. Figure 7 in the appendix makes the time dependence explicit through accuracy trajectories and each run's best-to-final gap.

Q4: What does the alignment term control? Proposition 6 is a statement about χ, not about VNE. We evaluate χ in Equation (16), with $v_k = w_k$ and averaged over classes, on the class-conditional density matrices that clients send to the server, across the coefficient sweep of Appendix B. Figure 4 separates the two terms of the surrogate. At every γ_2, a larger alignment coefficient gives a lower final χ, from 0.86–0.92 at $\gamma_1 = 10^{-3}$ to 0.75–0.79 at $\gamma_1 = 10^{-1}$. A larger entropy coefficient mainly raises VNE and changes χ by at most 0.06. The alignment term therefore moves the quantity that Proposition 6 contracts, while the entropy term moves the spectrum; the sweep uses one seed at $\alpha = 0.1$. Appendix C.2 reports χ along the main runs.

Sensitivity to local computation. Theorem 4 lets local computation enter through U. Varying the number of local epochs in FedAvg (Appendix C.4) changes final VNE, but the response depends on heterogeneity and communication budget and need not be monotone.

8 Related work

Optimization and aggregation objects. FedAvg, FedProx, SCAFFOLD, and FedNova established complementary responses to heterogeneous local progress (McMahan et al., 2017; Li et al., 2020; Karimireddy et al., 2020; Wang et al., 2020). Recent analyses refine the roles of heterogeneity, participation, local computation, robustness, missing data, and client selection (Wang et al., 2024c; Crawshaw & Liu, 2024; Chen & Vikalo, 2024; Li et al., 2024; Guo & Yang, 2024; Chen et al., 2024c; Zhang et al., 2024; Ying et al., 2025). A parallel line changes the server's aggregation object—from parameters to sufficient statistics, model-space solutions, partial networks, layers, and calibrated updates (Fanì et al., 2024; Chen et al., 2024b; Shi et al., 2025; Liu et al., 2024b; Wang et al., 2024a; Huang et al., 2024; Liu et al., 2024a; Wang et al., 2024b; Zheng et al., 2025; Gao et al., 2024; Wu et al., 2025; Du et al., 2026; Zhang et al., 2025). We ask what spectrum appears after the aggregated encoder acts on data.

Federated representation interventions. Representation objectives compare models, exchange prototypes, decorrelate features, encourage uniformity, or summarize topology (Li et al., 2021; Tan et al., 2022; Shi et al., 2023; Son et al., 2024; Hu et al., 2026). The scope now includes unsupervised, multimodal, probabilistic, frequency-domain, hybrid, aligned, and modular representations (Liao et al., 2024; Yi et al., 2024; Wang et al., 2024d; McLaughlin & Su, 2024; Chen et al., 2024d; Yi et al., 2025; Liu et al., 2025; Goswami et al., 2025; Hao et al., 2025; Liao et al., 2025a; Nguyen et al., 2025). Foundation-model federation makes the aggregation object still more explicit through prompts, low-rank factors, experts, and projections (He et al., 2025; Liao et al., 2025b; Shu et al., 2025; Wang et al., 2025a;b; Zhu et al., 2025; Guan et al., 2025; Farhat et al., 2025; Zhuo et al., 2025; Meng et al., 2026). We isolate the trace-normalized within-class covariance spectrum and connect it to a class-conditioned intervention.

Spectral diagnostics of representations. VNE, effective rank, and similarity-spectrum entropy summarize distinct forms of representation diversity (Kim et al., 2023; Garrido et al., 2023; Friedman & Dieng, 2023), while neural collapse emphasizes class-mean and classifier geometry (Papyan et al., 2020). Our centered within-class diagnostic separates spectral spread from class separation and provides one object shared by the finite-time comparison and matrix intervention.

9 Discussion and conclusion

Parameter averaging has no fixed spectral direction, yet the size of the change it induces is controlled by client heterogeneity and local trajectory separation. A desired direction can instead be introduced by acting on the spectral object itself: interpolation toward a shared class density matrix contracts weighted disagreement, and FedVNA translates this shared-reference principle into minibatch training.

The experiments separate controllability from predictive utility: under strong label heterogeneity FedVNA raises VNE over FedAvg with nearly identical final accuracy, and FedDecorr reaches higher VNE at weaker heterogeneity at a larger accuracy cost. Spectral diversity describes feature geometry; it does not replace an accuracy evaluation.

The evidence is limited to CIFAR-10, ResNet-18, label skew, and ten clients. The bound can reach the entropy ceiling when its constants are large; the minibatch surrogate only approximates the ideal matrix operation; and full-covariance communication is expensive. Future work should determine when controlled spectra improve separation and prediction.

AI USE STATEMENT

Generative AI tools were used for literature search, language polishing, and checking the correctness of mathematical proofs. The authors reviewed the AI-assisted work and take responsibility for the final content.

REFERENCES

Koenraad MR Audenaert. A sharp continuity estimate for the von neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127–8136, 2007.

Huancheng Chen and Haris Vikalo. Heterogeneity-guided client sampling: Towards fast and efficient non-iid federated learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7886b9bafe76c52fd568db10ff9772df-Abstract-Conference.html.

Junbao Chen, Jingfeng Xue, Yong Wang, Zhenyan Liu, and Lu Huang. Classifier clustering and feature alignment for federated learning under distributed concept drift. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/942e820be4aa112509b3a281ff398851-Abstract-Conference.html.

Minghui Chen, Meirui Jiang, Xin Zhang, Qi Dou, Zehua Wang, and Xiaoxiao Li. Local superior soups: A catalyst for model merging in cross-silo federated learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/24f7b98aef14fcd68acf3c941af1b59e-Abstract-Conference.html.

Xiaohong Chen, Canran Xiao, and Yongmei Liu. Confusion-resistant federated learning via diffusion-based data harmonization on non-iid data. In *Advances in Neural Information Processing Systems*, volume 37, 2024c. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f896c593ae0cce758205646e1325406b-Abstract-Conference.html.

Xinyue Chen, Yazhou Ren, Jie Xu, Fangfei Lin, Xiaorong Pu, and Yang Yang. Bridging gaps: Federated multi-view clustering in heterogeneous hybrid views. In *Advances in Neural Information Processing Systems*, volume 37, 2024d. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/412fb8623bf8b6d56fb6285ea295447e-Abstract-Conference.html.

Michael Crawshaw and Mingrui Liu. Federated learning under periodic client participation and heterogeneous data: A new communication-efficient algorithm and analysis. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0f98645119923217a245735c2c4d23f4-Abstract-Conference.html.

Haizhou Du, Yiran Xiang, Yiwen Cai, Xiufeng Liu, Zonghan Wu, Huan Huo, and Guodong Long. Fedfree: Breaking knowledge-sharing barriers through layer-wise alignment in heterogeneous federated learning. In *Advances in Neural Information Processing Systems*, volume 38, 2026. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7644353d580a9e027e0069d6480d971b-Abstract-Conference.html.

Eros Fanì, Raffaello Camoriano, Barbara Caputo, and Marco Ciccone. Accelerating heterogeneous federated learning with closed-form classifiers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13029–13048. PMLR, 2024. URL https://proceedings.mlr.press/v235/fani-24a.html.

Yehya Farhat, Hamza ElMokhtar Shili, Fangshuo Liao, Chen Dun, Mirian Hipolito Garcia, Guoqing Zheng, Ahmed Awadallah, Robert Sim, Dimitrios Dimitriadis, and Anastasios Kyrillidis. Learning to specialize: Joint gating-expert training for adaptive moes in decentralized settings. In *Advances in Neural Information Processing Systems*, volume 38,

2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/03bb44eceb94537ee2e2f00a8dca60b1-Abstract-Conference.html.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

Changyu Gao, Andrew Lowy, Xingyu Zhou, and Stephen Wright. Private heterogeneous federated learning without a trusted server revisited: Error-optimal and communication-efficient algorithms for convex losses. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 14763–14789. PMLR, 2024.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/garrido23a.html.

Dipam Goswami, Simone Magistri, Kai Wang, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. Covariances for free: Exploiting mean distributions for training-free federated learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5df817c5dd95293ebf6d1583303a8c73-Abstract-Conference.html.

Kaiqi Guan, Wenke Huang, Xianda Guo, Yueyang Yuan, Bin Yang, and Mang Ye. Rethinking fair federated learning from parameter and client view. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1b1205172b8a57ad504002a081b37859-Abstract-Conference.html.

Zhishuai Guo and Tianbao Yang. Communication-efficient federated group distributionally robust optimization. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/29021b06afa4c648ee438584f7ef3e7e-Abstract-Conference.html.

Chenhe Hao, Weiying Xie, Daixun Li, Haonan Qin, Hangyu Ye, Leyuan Fang, and Yunsong Li. Fedcs: Coreset selection for federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15434–15443, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Hao_FedCS_Coreset_Selection_for_Federated_Learning_CVPR_2025_paper.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Run He, Kai Tong, Di Fang, Han Sun, Ziqian Zeng, Haoran Li, Tianyi Chen, and Huiping Zhuang. Afl: A single-round analytic approach for federated learning with pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4988–4998, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/He_AFL_A_Single-Round_Analytic_Approach_for_Federated_Learning_with_Pre-trained_CVPR_2025_paper.html.

Ke Hu, Liyao Xiang, Peng Tang, and Weidong Qiu. Fedtopo: Topology-informed representation alignment in federated learning under non-iid conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21849–21857, 2026.

Wenke Huang, Zekun Shi, Mang Ye, He Li, and Bo Du. Self-driven entropy aggregation for byzantine-robust heterogeneous federated learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20096–20110. PMLR, 2024.

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

Jaeill Kim, Suhyun Kang, Duhun Hwang, Jungwook Shin, and Wonjong Rhee. Vne: An effective method for improving deep representation by manipulating eigenvalue distribution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3799–3810, 2023.

Hanmin Li, Kirill Acharya, and Peter Richtárik. The power of extrapolation in federated learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e0b6f389739496e363a89155c9448a8a-Abstract-Conference.html.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10713–10722, 2021.

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.

Guiqiu Liao, Matjaz Jogan, Eric Eaton, and Daniel A. Hashimoto. Forla: Federated object-centric representation learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/923fd1695eaee59610c2bc033a9e2354-Paper-Conference.pdf.

Xinting Liao, Weiming Liu, Chaochao Chen, Pengyang Zhou, Fengyuan Yu, Huabin Zhu, Binhui Yao, Tao Wang, Xiaolin Zheng, and Yanchao Tan. Rethinking the representation in federated unsupervised learning with non-iid data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22841–22850, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liao_Rethinking_the_Representation_in_Federated_Unsupervised_Learning_with_Non-IID_Data_CVPR_2024_paper.html.

Yangxu Liao, Wenke Huang, Guancheng Wan, Jian Liang, Bin Yang, and Mang Ye. Splitting with importance-aware updating for heterogeneous federated learning with large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=ny0m8YEUzH.

Jiahao Liu, Yipeng Zhou, Di Wu, Miao Hu, Mohsen Guizani, and Quan Z. Sheng. Fedlmt: Tackling system heterogeneity of federated learning via low-rank model training with theoretical guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32509–32551. PMLR, 2024a.

Qiqi Liu, Jiaqiang Li, Yuchen Liu, Yaochu Jin, Lingjuan Lyu, Xiaohu Wu, and Han Yu. Personalized federated learning under local supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4069–4079, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_Personalized_Federated_Learning_under_Local_Supervision_ICCV_2025_paper.html.

Xiang Liu, Liangxi Liu, Feiyang Ye, Yunheng Shen, Xia Li, Linshan Jiang, and Jialin Li. Fedlpa: One-shot federated learning with layer-wise posterior aggregation. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9482f45fdd89aba9130bb04c44f788a9-Abstract-Conference.html.

Connor J. McLaughlin and Lili Su. Personalized federated learning via feature distribution adaptation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8ce6c5450ccddbe6adee4b3749893587-Abstract-Conference.html.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

Chuiyang Meng, Ming Tang, and Vincent W. S. Wong. Florg: Federated fine-tuning with low-rank gram matrices and procrustes alignment. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kntrZOm2AQ.

Duong Nguyen, Nghia Hoang, Thanh Trung Huynh, Quoc Viet Hung Nguyen, and Phi Le Nguyen. Learning reconfigurable representations for multimodal federated learning with missing data. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/29978617eb250d16969e3681121ad6d7-Abstract-Conference.html.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Changlong Shi, He Zhao, Bingjie Zhang, Mingyuan Zhou, Dandan Guo, and Yi Chang. Fedawa: Adaptive optimization of aggregation weights in federated learning using client vectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30651–30660, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Shi_FedAWA_Adaptive_Optimization_of_Aggregation_Weights_in_Federated_Learning_Using_Client_CVPR_2025_paper.html.

Yujun Shi, Jian Liang, Wenqing Zhang, Vincent Tan, and Song Bai. Towards understanding and mitigating dimensional collapse in heterogeneous federated learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=EXnIyMVTL8s.

Yao Shu, Wenyang Hu, See-Kiong Ng, Bryan Kian Hsiang Low, and Fei Yu. Ferret: Federated full-parameter tuning at scale for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=mzPArjGqrs.

Ha Min Son, Moon-Hyun Kim, Tai-Myoung Chung, Chao Huang, and Xin Liu. Feduv: Uniformity and variance for heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5863–5872, June 2024.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8432–8440, 2022.

Haolin Wang, Xuefeng Liu, Jianwei Niu, Wenkai Guo, and Shaojie Tang. Why go full? elevating federated learning through partial network updates. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b4baac5d3f7508a4eb2b65376470a5a2-Abstract-Conference.html.

Haozhao Wang, Shengyu Wang, Jiaming Li, Hao Ren, Xingshuo Han, Wenchao Xu, Shangwei Guo, Tianwei Zhang, and Ruixuan Li. Bsemifl: Semi-supervised federated learning via a bayesian approach. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=fmlol78Qqf.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

Jiaqi Wang, Chenxu Zhao, Lingjuan Lyu, Quanzeng You, Mengdi Huai, and Fenglong Ma. Bridging model heterogeneity in federated learning via uncertainty-based asymmetrical reciprocity learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52290–52308. PMLR, 2024b.

Jiayi Wang, Shiqiang Wang, Rong-Rong Chen, and Mingyue Ji. A new theoretical perspective on data heterogeneity in federated optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51650–51700. PMLR, 2024c. URL https://proceedings.mlr.press/v235/wang24bu.html.

Lei Wang, Jieming Bian, Letian Zhang, Chen Chen, and Jie Xu. Taming cross-domain representation variance in federated prototype learning with heterogeneous data domains. In *Advances in Neural Information Processing Systems*, volume 37, 2024d. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a11e42a37c6bc926d6dc57e0cca0e825-Abstract-Conference.html.

Lei Wang, Jieming Bian, Letian Zhang, and Jie Xu. Adaptive lora experts allocation and selection for federated fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/6df1b2b45e64d402588746f79b68b82c-Abstract-Conference.html.

Siyuan Wu, Yongzhe Jia, Haolong Xiang, Xiaolong Xu, Xuyun Zhang, Lianyong Qi, and Wanchun Dou. A fair federated learning method for handling client participation probability inconsistencies in heterogeneous environments. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/dcc0ac74ac8b95dc1939804acce0317d-Abstract-Conference.html.

Liping Yi, Han Yu, Chao Ren, Gang Wang, Xiaoguang Liu, and Xiaoxiao Li. Federated model heterogeneous matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7a9ee756fb71d401bcb5230db02c11db-Abstract-Conference.html.

Liping Yi, Han Yu, Gang Wang, Xiaoguang Liu, and Xiaoxiao Li. Federated representation angle learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yi_Federated_Representation_Angle_Learning_ICCV_2025_paper.html.

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

Hao Zhang, Chenglin Li, Nuowen Kan, Ziyang Zheng, Wenrui Dai, Junni Zou, and Hongkai Xiong. Improving generalization in federated learning with model-data mutual information regularization: A posterior inference approach. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f6ec12322f9f9b25d07afcdd25604005-Abstract-Conference.html.

Letian Zhang, Bo Chen, Jieming Bian, Lei Wang, and Jie Xu. Fedel: Federated elastic learning for heterogeneous devices. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3327377ee37b8c87c43a7d6a4ea3d279-Abstract-Conference.html.

Hao Zheng, Zhigang Hu, Liu Yang, Meiguang Zheng, Aikun Xu, and Boyu Wang. Fedcalm: Conflict-aware layer-wise mitigation for selective aggregation in deeper personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15444–15453, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zheng_FedCALM_Conflict-aware_Layer-wise_Mitigation_for_Selective_Aggregation_in_Deeper_Personalized_CVPR_2025_paper.html.

Ran Zhu, Mingkun Yang, Shiqiang Wang, Jie Yang, and Qing Wang. Flick: Empowering federated learning with commonsense knowledge. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/02080e8e63c0db49c6f8854e2f19a272-Abstract-Conference.html.

Wei Zhuo, Zhaohuan Zhan, and Han Yu. Personalized subgraph federated learning with differentiable auxiliary projections. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2360da01c2ed6592bb691326424de184-Paper-Conference.pdf.

A PROOFS

A.1 PROOF OF PROPOSITION 1

Set $B_k = A_k - \bar{A}$. By the definition of $\bar{A}$,

$$\sum_k w_k B_k = \sum_k w_k A_k - \bar{A} = 0. \tag{22}$$

Expanding each local covariance around $\bar{A}$ gives

$$\begin{aligned}\sum_k w_k A_k \Sigma A_k^\top &= \sum_k w_k (\bar{A} + B_k)\Sigma(\bar{A} + B_k)^\top \\ &= \bar{A}\Sigma\bar{A}^\top + \left(\sum_k w_k B_k\right)\Sigma\bar{A}^\top + \bar{A}\Sigma\left(\sum_k w_k B_k\right)^\top + \sum_k w_k B_k \Sigma B_k^\top. \end{aligned} \tag{23}$$

The two cross terms in Equation (23) vanish by Equation (22). Rearranging yields the identity in Equation (3).

It remains to verify the Loewner statement. For every vector u,

$$u^\top \left(\sum_k w_k B_k \Sigma B_k^\top\right) u = \sum_k w_k \left\|\Sigma^{1/2} B_k^\top u\right\|_2^2 \geq 0, \tag{24}$$

because $w_k \geq 0$ and $\Sigma \succeq 0$. Hence the residual is positive semidefinite. The calculation is unchanged when Σ is a fixed centered empirical input covariance. The claim stops before trace normalization: dividing the two sides by their generally different traces need not preserve an entropy ordering. Indeed, $I_2 \succeq \mathrm{diag}(1, 0)$, while their normalized entropies are $\log 2$ and 0, respectively. In the opposite direction, $\mathrm{diag}(2, 1) \succeq I_2$, while $h_2(1/3) < \log 2$. Thus a larger unnormalized covariance can have either larger or smaller entropy after trace normalization.

A.2 PROOF OF PROPOSITION 2

For a centered input with covariance I_2, the feature covariance of the linear encoder A is $AA^\top$. Its density matrix is therefore

$$\rho(A) = \frac{AA^\top}{\mathrm{Tr}(AA^\top)}, \tag{25}$$

whenever $A \neq 0$.

An averaging operation that increases entropy. Take

$$A_1 = \mathrm{diag}(1, 0), \qquad A_2 = \mathrm{diag}(0, 1). \tag{26}$$

Equation (25) gives $\rho(A_1) = \mathrm{diag}(1, 0)$ and $\rho(A_2) = \mathrm{diag}(0, 1)$. Both spectra are $(1, 0)$, so both local entropies are zero. With equal weights, $\bar{A} = (A_1 + A_2)/2 = I_2/2$, and hence

$$\rho(\bar{A}) = \frac{I_2/4}{\mathrm{Tr}(I_2/4)} = \frac{I_2}{2}. \tag{27}$$

Its spectrum is $(1/2, 1/2)$ and its entropy is $\log 2$. Thus $S(\rho(\bar{A}))$ is strictly larger than the mean local entropy.

An averaging operation that decreases entropy. Now take

$$A_1 = I_2, \qquad A_2 = \mathrm{diag}(1, -1). \tag{28}$$

For both encoders, $A_k A_k^\top = I_2$, so $\rho(A_1) = \rho(A_2) = I_2/2$ and each local entropy equals $\log 2$. However, $\bar{A} = (A_1 + A_2)/2 = \mathrm{diag}(1, 0)$, which gives $\rho(\bar{A}) = \mathrm{diag}(1, 0)$ and entropy zero. Thus the averaged-model entropy is strictly smaller than the mean local entropy.

All matrices used above have positive covariance trace. Moreover, the mixture $(\rho(A_1) + \rho(A_2))/2$ of density matrices equals $I_2/2$ in both constructions. The two opposite outcomes therefore concern the density matrix of the parameter-averaged encoder, not a failure of entropy concavity for mixtures of density matrices.

A.3 Proof of Lemma 3

Fix a pair a and abbreviate $n = n_a$. We divide the argument into the covariance perturbation, trace normalization, and entropy continuity steps.

Step 1: covariance perturbation. For $\bar{z}(\theta) = n^{-1}\sum_{i=1}^{n} z_{a,i}(\theta)$, the centered sample covariance can be written as

$$C_a(\theta) = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} z_{a,i}(\theta)z_{a,i}(\theta)^\top - \bar{z}(\theta)\bar{z}(\theta)^\top\right]. \tag{29}$$

For arbitrary vectors u, v,

$$uu^\top - vv^\top = (u-v)u^\top + v(u-v)^\top,$$
$$\left\|ab^\top\right\|_F = \|a\|_2 \|b\|_2 . \tag{30}$$

The triangle inequality therefore gives

$$\left\|uu^\top - vv^\top\right\|_F \le (\|u\|_2 + \|v\|_2)\,\|u-v\|_2 . \tag{31}$$

Assumption 1 and Equation (31) imply, for every i,

$$\left\|z_{a,i}(\theta')z_{a,i}(\theta')^\top - z_{a,i}(\theta)z_{a,i}(\theta)^\top\right\|_F \le 2RL_z\,\|\theta'-\theta\|_2 . \tag{32}$$

The mean feature obeys $\|\bar{z}(\theta)\|_2 \le R$ and $\|\bar{z}(\theta') - \bar{z}(\theta)\|_2 \le L_z\,\|\theta'-\theta\|_2$ by the triangle inequality and the per-sample assumptions. Applying Equation (31) once more to the mean outer product gives the same upper bound as in Equation (32). Inserting both estimates into Equation (29) yields

$$\|C_a(\theta') - C_a(\theta)\|_F \le \frac{4n}{n-1}RL_z\,\|\theta'-\theta\|_2 = L_{C,a}\,\|\theta'-\theta\|_2 . \tag{33}$$

Step 2: trace normalization. Let $C' = C_a(\theta')$, $C = C_a(\theta)$, $t' = \operatorname{Tr} C'$, and $t = \operatorname{Tr} C$. Both matrices are positive semidefinite, and Assumption 1 gives $t, t' \ge \tau_a$. Write

$$\frac{C'}{t'} - \frac{C}{t} = \frac{C'-C}{t'} + C\left(\frac{1}{t'} - \frac{1}{t}\right). \tag{34}$$

Because $C \succeq 0$, $\|C\|_1 = \operatorname{Tr} C = t$. Moreover, trace duality with the identity gives $|t'-t| = |\operatorname{Tr}(C'-C)| \le \|C'-C\|_1$. Taking trace norms in Equation (34) therefore gives

$$\begin{aligned}\|\rho_a(\theta') - \rho_a(\theta)\|_1 &\le \frac{\|C'-C\|_1}{t'} + \frac{\|C\|_1\,|t'-t|}{tt'} \\ &\le \frac{2\,\|C'-C\|_1}{t'}.\end{aligned} \tag{35}$$

For an $m \times m$ matrix, $\|M\|_1 \le \sqrt{\operatorname{rank}(M)}\,\|M\|_F \le \sqrt{m}\,\|M\|_F$. Combining this fact with Equations (33) and (35) proves

$$\frac{1}{2}\,\|\rho_a(\theta') - \rho_a(\theta)\|_1 \le \frac{\sqrt{m}L_{C,a}}{\tau_a}\,\|\theta'-\theta\|_2 = \kappa_a\,\|\theta'-\theta\|_2 . \tag{36}$$

The left-hand side is a trace distance between density matrices and is always at most one, which gives the stated clipped bound.

Step 3: entropy continuity. Set

$$d_a = \frac{1}{2}\,\|\rho_a(\theta') - \rho_a(\theta)\|_1 . \tag{37}$$

When $d_a \le 1 - 1/m$, the sharp Fannes–Audenaert inequality (Audenaert, 2007) gives

$$|S(\rho_a(\theta')) - S(\rho_a(\theta))| \le d_a \log(m-1) + h_2(d_a). \tag{38}$$

For $d_a > 1 - 1/m$, the trivial range $0 \le S(\rho) \le \log m$ gives the upper bound $\log m$. These two cases are exactly $\omega_m(d_a)$. On $0 < u < 1 - 1/m$, its nonconstant branch satisfies

$$\omega_m'(u) = \log\frac{(m-1)(1-u)}{u} \ge 0, \qquad \omega_m''(u) = -\frac{1}{u(1-u)} < 0. \tag{39}$$

At $u = 1 - 1/m$, the branch equals $\log m$ and its left derivative is zero, so joining it to the constant branch preserves concavity and monotonicity. Since ω_m is nondecreasing, substituting the bound from Step 2 proves Equation (7). No positive lower eigenvalue of C_a was used, so the argument includes rank-deficient empirical covariances.

Related spectral identities. If $\widetilde{Z}_a = P\Sigma Q^\top$ is a singular-value decomposition, then $C_a = Q\Sigma^2 Q^\top/(n_a-1)$ and $G_a = P\Sigma^2 P^\top/(n_a-1)$. Thus they have the same positive eigenvalues and trace. Also, for any density matrix ρ, $\log(1+x) \le x$ gives

$$S(\rho) - S_{\epsilon_{\log}}(\rho) = \sum_{\lambda_j>0} \lambda_j \log(1+\epsilon_{\log}/\lambda_j) \le \operatorname{rank}(\rho)\epsilon_{\log} \le m\epsilon_{\log}. \tag{40}$$

This last calculation concerns placing $\epsilon_{\log}$ inside the logarithm; adding $\epsilon_{\mathrm{r}} I$ before trace normalization defines a different density matrix and changes both its spectrum and rank.

A.4 Proof of Theorem 4

Set $d_k^e = \theta_k^e - r^e$ and define

$$D_e = \sum_k w_k \mathbb{E}_t \left\| d_k^e \right\|_2 . \tag{41}$$

We first bound the spread of the individual local trajectories and then exploit cancellation when their endpoints are averaged.

Step 1: deviation of each local trajectory. Subtracting the reference update from the client update and adding and subtracting $\nabla L_k(r^e)$ gives

$$\begin{aligned} d_k^{e+1} = d_k^e &- \eta_e[\nabla L_k(\theta_k^e) - \nabla L_k(r^e)] \\ &- \eta_e[\nabla L_k(r^e) - \nabla L(r^e)] - \eta_e \xi_k^e. \end{aligned} \tag{42}$$

The β-Lipschitz gradient assumption implies

$$\begin{aligned} \left\| d_k^e - \eta_e[\nabla L_k(\theta_k^e) - \nabla L_k(r^e)] \right\|_2 &\le \left\| d_k^e \right\|_2 + \eta_e \left\| \nabla L_k(\theta_k^e) - \nabla L_k(r^e) \right\|_2 \\ &\le (1+\beta\eta_e) \left\| d_k^e \right\|_2 . \end{aligned} \tag{43}$$

The heterogeneity assumption and weighted Cauchy–Schwarz give

$$\begin{aligned} &\sum_k w_k \left\| \nabla L_k(r^e) - \nabla L(r^e) \right\|_2 \\ &\quad \le \left(\sum_k w_k \right)^{1/2} \left(\sum_k w_k \left\| \nabla L_k(r^e) - \nabla L(r^e) \right\|_2^2 \right)^{1/2} \le \zeta, \end{aligned} \tag{44}$$

because $\sum_k w_k = 1$. Similarly, conditional Jensen and weighted Cauchy–Schwarz imply

$$\begin{aligned} \sum_k w_k \mathbb{E}[\|\xi_k^e\|_2 \mid \mathcal{G}_{t,e}] &\le \sum_k w_k \sqrt{\mathbb{E}[\|\xi_k^e\|_2^2 \mid \mathcal{G}_{t,e}]} \\ &\le \sqrt{\sum_k w_k \mathbb{E}[\|\xi_k^e\|_2^2 \mid \mathcal{G}_{t,e}]} \le \nu_e. \end{aligned} \tag{45}$$

Taking norms in Equation (42) and conditioning first on $\mathcal{G}_{t,e}$ yields

$$\begin{aligned} \mathbb{E}[\left\| d_k^{e+1} \right\|_2 \mid \mathcal{G}_{t,e}] &\le (1+\beta\eta_e) \left\| d_k^e \right\|_2 \\ &\quad + \eta_e \left\| \nabla L_k(r^e) - \nabla L(r^e) \right\|_2 + \eta_e \mathbb{E}[\|\xi_k^e\|_2 \mid \mathcal{G}_{t,e}]. \end{aligned} \tag{46}$$

Multiply by w_k, sum over k, and condition on $\mathcal{G}_{t,0}$. Because $\mathcal{G}_{t,0} \subseteq \mathcal{G}_{t,e}$, the tower property $\mathbb{E}_t X = \mathbb{E}_t[\mathbb{E}[X \mid \mathcal{G}_{t,e}]]$, together with Equations (44) and (45), yields

$$D_{e+1} \le (1+\beta\eta_e) D_e + \eta_e(\zeta + \nu_e). \tag{47}$$

Since $D_0 = 0 = \Gamma_0$, induction on e and the definition of Γ_e show that $D_e \le \Gamma_e$ for every $e \le U$.

Step 2: cancellation at the averaged endpoint. Introduce the algebraic average $\bar{\theta}^e = \sum_k w_k \theta_k^e$ and $\bar{d}^e = \bar{\theta}^e - r^e = \sum_k w_k d_k^e$. This definition does not insert aggregation between local steps; it only averages the hypothetical local iterates at the same index e. Averaging the local updates gives

$$\begin{aligned}\bar{d}^{e+1} &= \bar{d}^e - \eta_e \sum_k w_k [\nabla L_k(\theta_k^e) - \nabla L(r^e) + \xi_k^e] \\ &= \bar{d}^e - \eta_e \sum_k w_k [\nabla L_k(\theta_k^e) - \nabla L_k(r^e)] - \eta_e \sum_k w_k \xi_k^e. \end{aligned} \tag{48}$$

The second equality uses the exact cancellation

$$\sum_k w_k [\nabla L_k(r^e) - \nabla L(r^e)] = 0, \tag{49}$$

which holds because both the global objective and parameter aggregation use the same weights w_k. Taking norms in Equation (48), using smoothness, and then the triangle inequality gives

$$\begin{aligned}\left\|\bar{d}^{e+1}\right\|_2 &\le \left\|\bar{d}^e\right\|_2 + \eta_e \beta \sum_k w_k \left\|d_k^e\right\|_2 + \eta_e \left\|\sum_k w_k \xi_k^e\right\|_2 \\ &\le \left\|\bar{d}^e\right\|_2 + \eta_e \beta \sum_k w_k \left\|d_k^e\right\|_2 + \eta_e \sum_k w_k \left\|\xi_k^e\right\|_2 . \end{aligned} \tag{50}$$

Condition first on $\mathcal{G}_{t,e}$ and then on $\mathcal{G}_{t,0}$. The tower property and Equation (45) give

$$\mathbb{E}_t \left\|\bar{d}^{e+1}\right\|_2 \le \mathbb{E}_t \left\|\bar{d}^e\right\|_2 + \eta_e(\beta D_e + \nu_e) \le \mathbb{E}_t \left\|\bar{d}^e\right\|_2 + \eta_e(\beta \Gamma_e + \nu_e). \tag{51}$$

Because $\bar{d}^0 = 0$ and $\bar{d}^U = \theta^{t+1} - r^U$, summing this inequality from $e = 0$ to $U - 1$ proves

$$\mathbb{E}_t \left\|\theta^{t+1} - r^U\right\|_2 \le \sum_{e=0}^{U-1} \eta_e(\beta \Gamma_e + \nu_e) = \Delta_U. \tag{52}$$

Notice why ζ is absent as a direct summand in Δ_U: the same-point heterogeneity terms cancel in Equation (49). Heterogeneity still affects Δ_U indirectly through the local-spread bound Γ_e.

Step 3: transfer from parameters to the diagnostic. For any two parameter vectors in $\mathcal{R}_t$, Lemma 3 and $\kappa_a \le \kappa_*$ imply

$$\begin{aligned}|S_{\mathrm{diag}}(\theta) - S_{\mathrm{diag}}(\theta')| &\le \sum_{a \in \mathcal{A}} q_a \omega_m(\min\{1, \kappa_a \left\|\theta - \theta'\right\|_2\}) \\ &\le \omega_m(\min\{1, \kappa_* \left\|\theta - \theta'\right\|_2\}) . \end{aligned} \tag{53}$$

The second line uses monotonicity of ω_m and $\sum_a q_a = 1$. The function $u \mapsto \omega_m(u)$ is concave and nondecreasing on $[0, 1]$: Equation (39) proves this on the nonconstant branch, its left derivative is zero at $1 - 1/m$, and the remaining branch is constant. Consequently, $x \mapsto \omega_m(\min\{1, \kappa_* x\})$ is concave and nondecreasing for $x \ge 0$. Denote this function by $g(x)$. Applying Equation (53) to the random endpoints and then using conditional Jensen gives the complete chain

$$\begin{aligned}\mathbb{E}_t \left|S_{\mathrm{diag}}(\theta^{t+1}) - S_{\mathrm{diag}}(r^U)\right| &\le \mathbb{E}_t g(\left\|\theta^{t+1} - r^U\right\|_2) \\ &\le g(\mathbb{E}_t \left\|\theta^{t+1} - r^U\right\|_2) \le g(\Delta_U). \end{aligned} \tag{54}$$

The middle inequality is conditional Jensen for concave g, and the last uses its monotonicity and Equation (52). This proves Equation (12). Since only an absolute difference was bounded, no ordering of the two entropies follows.

Step 4: the mean local-model diagnostic. Recall $q_k = \sum_y q_{k,y}$, $c_q = \max_{k:w_k>0} q_k/w_k$, and the concave nondecreasing function g from Step 3. The triangle inequality and pairwise continuity give

$$\begin{aligned}
&\mathbb{E}_t \left| S_{\mathrm{local}}(\{\theta_k^U\}) - S_{\mathrm{diag}}(r^U) \right| \\
&\quad \le \sum_{k,y} q_{k,y} \mathbb{E}_t g(\left\|\theta_k^U - r^U\right\|_2) \\
&\quad \le \sum_k q_k g(\mathbb{E}_t \left\|\theta_k^U - r^U\right\|_2) \\
&\quad \le g\left(\sum_k q_k \mathbb{E}_t \left\|\theta_k^U - r^U\right\|_2 \right).
\end{aligned} \tag{55}$$

The second inequality is conditional Jensen for each client; the third is Jensen with weights q_k, which sum to one. By the definition of c_q,

$$\sum_k q_k \mathbb{E}_t \left\|\theta_k^U - r^U\right\|_2 \le c_q \sum_k w_k \mathbb{E}_t \left\|\theta_k^U - r^U\right\|_2 = c_q D_U \le c_q \Gamma_U. \tag{56}$$

Substitution into Equation (55) proves Equation (13). When $q_k = w_k$ for every client, $c_q = 1$; otherwise it cannot in general be omitted.

Step 5: constant-step specialization. If $\eta_e = \eta$, $\nu_e = 0$, and $\beta > 0$, solving the recursion for Γ_e gives

$$\Gamma_e = \frac{\zeta}{\beta}[(1+\beta\eta)^e - 1]. \tag{57}$$

Therefore

$$\begin{aligned}
\Delta_U &= \sum_{e=0}^{U-1} \eta\beta\Gamma_e = \eta\zeta \sum_{e=0}^{U-1} [(1+\beta\eta)^e - 1] \\
&= \frac{\zeta}{\beta}[(1+\beta\eta)^U - 1 - U\beta\eta].
\end{aligned} \tag{58}$$

For fixed U, expanding $(1+\beta\eta)^U$ as $\eta \to 0$ gives the expression in Equation (15). If $U = 1$, the sum defining Δ_U contains only $\beta\Gamma_0 = 0$, so the noiseless endpoints agree. If $\beta = 0$ and $\nu_e = 0$, every summand in Δ_U is again zero even though the individual local iterates may separate.

A.5 Proof of Corollary 5

Both S_{diag} and S_{local} are defined on the same pairs $\mathcal{A}$ with the same weights q_a. For every realization, the triangle inequality gives

$$\left|S_{\mathrm{diag}}(\theta^{t+1}) - S_{\mathrm{local}}(\{\theta_k^U\})\right| \le \left|S_{\mathrm{diag}}(\theta^{t+1}) - S_{\mathrm{diag}}(r^U)\right| + \left|S_{\mathrm{diag}}(r^U) - S_{\mathrm{local}}(\{\theta_k^U\})\right|. \tag{59}$$

Taking $\mathbb{E}_t$ and using linearity of conditional expectation, the first term is bounded by Equation (12) and the second by Equation (13), both almost surely. This proves Equation (14). As in Theorem 4, only an absolute difference is bounded, so the corollary is compatible with either sign in Proposition 2.

A.6 Proof of Proposition 6

We first address support, because the matrices need not be full rank. If a vector u lies in the kernel of $\bar{\rho}$, then

$$0 = u^\top \bar{\rho} u = \sum_k v_k u^\top \rho_k u. \tag{60}$$

Every summand is nonnegative and every v_k is positive, so $u^\top \rho_k u = 0$ for every k. Since $u^\top \rho_k u = \left\|\rho_k^{1/2} u\right\|_2^2$, this implies $\rho_k^{1/2} u = 0$ and hence $\rho_k u = 0$. Therefore $\mathrm{supp}(\rho_k) \subseteq \mathrm{supp}(\bar{\rho})$, and the same inclusion holds for $\rho_k(s)$. All relative entropies below are therefore finite on the support of $\bar{\rho}$.

Step 1: preservation of the barycenter. Using $\sum_k v_k = 1$ and the definition of $\bar{\rho}$,

$$\begin{aligned}\sum_k v_k \rho_k(s) &= (1-s)\sum_k v_k \rho_k + s\sum_k v_k \bar{\rho} \\ &= (1-s)\bar{\rho} + s\bar{\rho} = \bar{\rho}.\end{aligned} \tag{61}$$

Step 2: relative-entropy representation of the gap. Expanding the definition of quantum relative entropy and using Equation (61),

$$\begin{aligned}\sum_k v_k D_{\mathrm{Q}}(\rho_k(s)\|\bar{\rho}) &= \sum_k v_k \operatorname{Tr}[\rho_k(s)\log\rho_k(s)] - \operatorname{Tr}\left[\left(\sum_k v_k \rho_k(s)\right)\log\bar{\rho}\right] \\ &= -\sum_k v_k S(\rho_k(s)) + S(\bar{\rho}) = \chi(s).\end{aligned} \tag{62}$$

This proves the identity in Equation (16).

Step 3: contraction from the initial matrices. Quantum relative entropy is convex in its first argument. Since $\rho_k(s) = (1-s)\rho_k + s\bar{\rho}$,

$$\begin{aligned}D_{\mathrm{Q}}(\rho_k(s)\|\bar{\rho}) &\le (1-s)D_{\mathrm{Q}}(\rho_k\|\bar{\rho}) + sD_{\mathrm{Q}}(\bar{\rho}\|\bar{\rho}) \\ &= (1-s)D_{\mathrm{Q}}(\rho_k\|\bar{\rho}).\end{aligned} \tag{63}$$

Multiplying by v_k, summing over k, and applying Equation (62) at s and at 0 gives $\chi(s) \le (1-s)\chi(0)$.

Step 4: contraction between arbitrary path points. Let $0 \le s_1 < s_2 < 1$ and define $a = (1-s_2)/(1-s_1) \in (0,1)$. Direct substitution shows that

$$\rho_k(s_2) = a\rho_k(s_1) + (1-a)\bar{\rho}. \tag{64}$$

Convexity in the first argument now gives

$$D_{\mathrm{Q}}(\rho_k(s_2)\|\bar{\rho}) \le aD_{\mathrm{Q}}(\rho_k(s_1)\|\bar{\rho}). \tag{65}$$

After weighting and summing, this becomes

$$\chi(s_2) \le \frac{1-s_2}{1-s_1}\chi(s_1), \tag{66}$$

which is Equation (17). Finally, Klein's inequality gives $D_{\mathrm{Q}}(\rho\|\sigma) \ge 0$, so Equation (62) implies $\chi(s) \ge 0$. Because $(1-s_2)/(1-s_1) \le 1$, the displayed contraction therefore also gives $\chi(s_2) \le \chi(s_1)$. At the excluded endpoint $s_2 = 1$, every $\rho_k(1) = \bar{\rho}$, and hence $\chi(1) = 0$. Since $S(\bar{\rho})$ is constant, Equation (62) now shows that the weighted mean entropy $\sum_k v_k S(\rho_k(s))$ is nondecreasing along the path.

This conclusion cannot be strengthened to every client. For example, let $v_1 = v_2 = 1/2$, $\rho_1 = I_2/2$, and $\rho_2 = \operatorname{diag}(1,0)$. Then $\bar{\rho} = \operatorname{diag}(3/4, 1/4)$ and

$$\rho_1(s) = \operatorname{diag}(1/2 + s/4, 1/2 - s/4). \tag{67}$$

For $s > 0$, this spectrum moves away from $(1/2, 1/2)$, so strict Schur concavity of entropy gives $S(\rho_1(s)) < S(\rho_1(0))$ even though the weighted mean entropy is nondecreasing.

B COEFFICIENT SENSITIVITY

The sweep at $\alpha = 0.1$ varies the alignment and entropy coefficients over $\{0, 10^{-3}, 10^{-2}, 10^{-1}\}$ with the dataset, architecture, and local-epoch count of the main experiments; its grid contains the main configuration $(\gamma_1, \gamma_2) = (10^{-3}, 10^{-2})$. With $\gamma_2 = 10^{-1}$, final VNE is 4.34, 4.08, and 3.29 for $\gamma_1 = 10^{-3}$, 10^{-2}, and 10^{-1}. The right panel reports the corresponding best test accuracy.

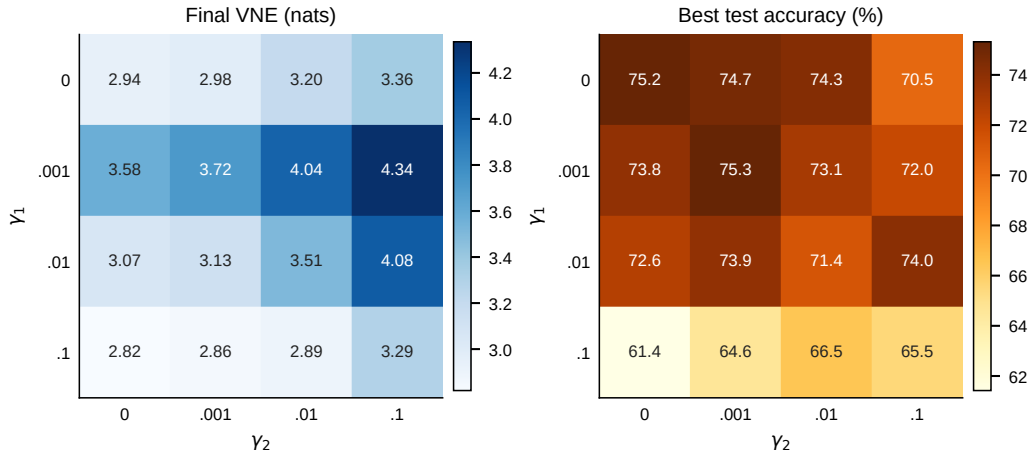

Figure 5: Sensitivity to γ_1 and γ_2 at $\alpha = 0.1$: final VNE (left) and best test accuracy (right). Each cell is a single run (seed 42).

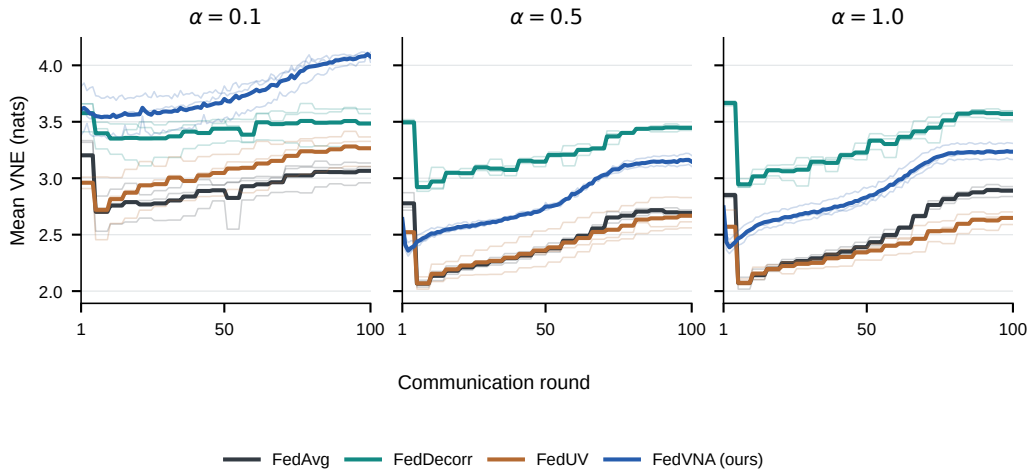

Figure 6: Seed-level VNE trajectories over 100 communication rounds for FedAvg, FedDecorr, FedUV, and FedVNA at three Dirichlet concentrations. Pale lines are the three individual training runs per method and setting; dark lines are their means.

C Additional diagnostics

C.1 Run-to-run and roundwise behavior

The mean curves in Figure 3 summarize three training seeds. Figure 6 exposes the individual trajectories for the four focal methods. Figure 7 adds the corresponding task-side view: an accuracy course under strong label skew and the within-run gap between best and final accuracy.

C.2 Disagreement along training

Figure 9 reports χ for the main FedVNA runs every ten rounds. From round 10 to round 100, the three-seed mean decreases from 0.98 to 0.89 at $\alpha = 0.1$ and increases from 0.59 to 0.67 at $\alpha = 0.5$ and from 0.36 to 0.49 at $\alpha = 1.0$; all three seeds move in the same direction at each α. Along a training run both the client features and the shared reference change, so this trajectory is not the matrix path of Proposition 6. The coefficient sweep in Figure 4 compares settings at the same round and isolates the alignment term.

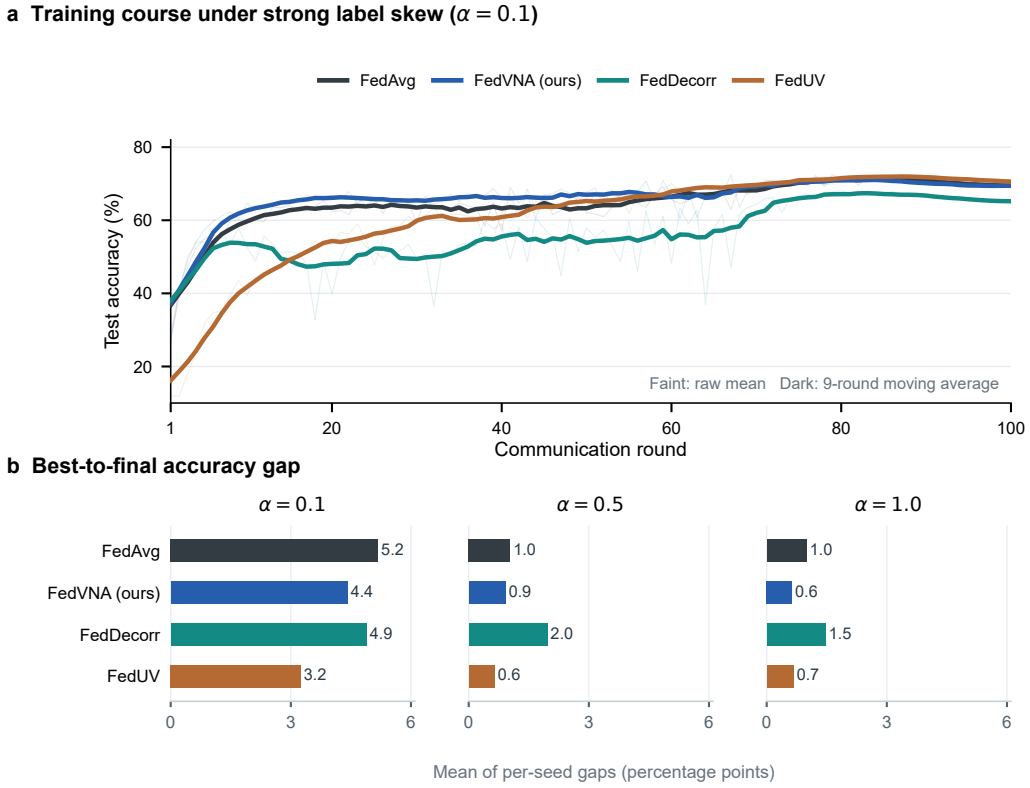

Figure 7: Accuracy during training and its endpoint dependence. The upper panel shows three-seed mean test accuracy at $\alpha = 0.1$: faint lines are unsmoothed values and dark lines are nine-round moving averages for display only. The lower panels show the mean across three seeds of each run's best-minus-final test accuracy, in percentage points, for $\alpha \in \{0.1, 0.5, 1.0\}$.

C.3 Measurement sensitivity

VNE depends on how the spectrum is estimated. Figure 10 compares three alternative estimators with the full-dimensional Gram estimate on the same FedAvg models across 11 Dirichlet concentrations. Pairing isolates the effect of the estimator on each model; PCA-64 also reduces the feature dimension from 512 to 64.

C.4 Sensitivity to local computation

Figure 11 examines the qualitative role of local computation in FedAvg. Here E_{ep} denotes local epochs, whereas U in Theorem 4 denotes individual optimizer updates; an epoch generally contains multiple updates. Relative to one local epoch, increasing E_{ep} usually changes final VNE, and the change depends on both heterogeneity and the communication budget. The reversal between $E_{\text{ep}} = 10$ and $E_{\text{ep}} = 20$ at $\alpha = 0.1$ shows that the response can itself change direction. Comparisons are made within each fixed round budget; increasing local epochs also increases total computation.

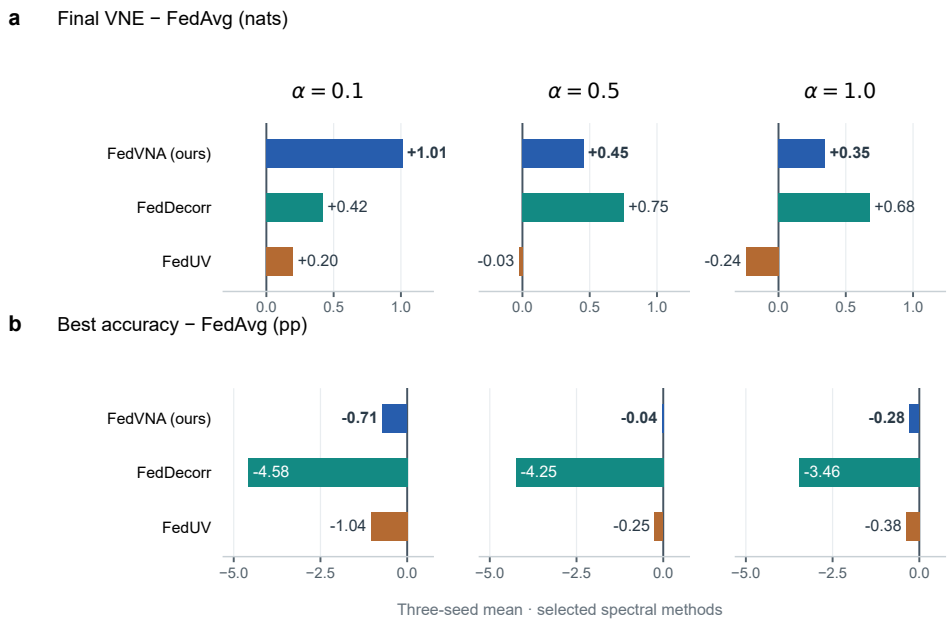

Figure 8: Three-seed mean changes from FedAvg for final VNE (nats) and best test accuracy (percentage points). Table 1 reports the absolute results and seed variation.

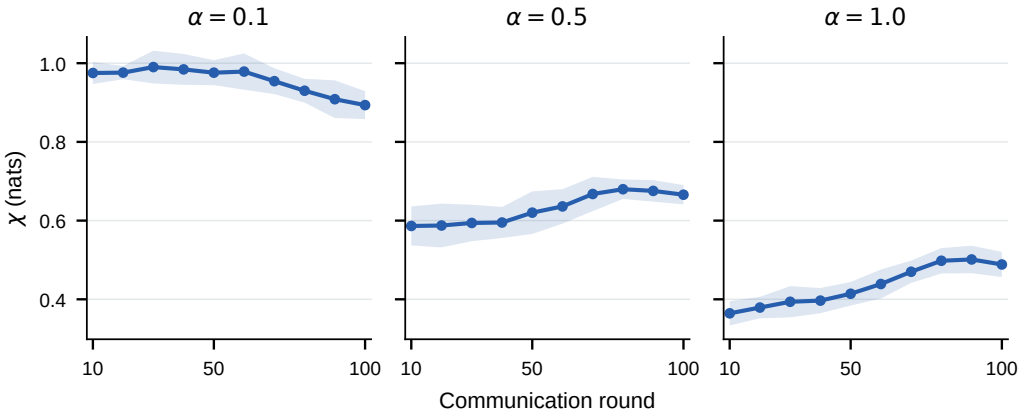

Figure 9: Disagreement χ of the main FedVNA runs every ten rounds. Lines are three-seed means; shading shows one standard deviation.

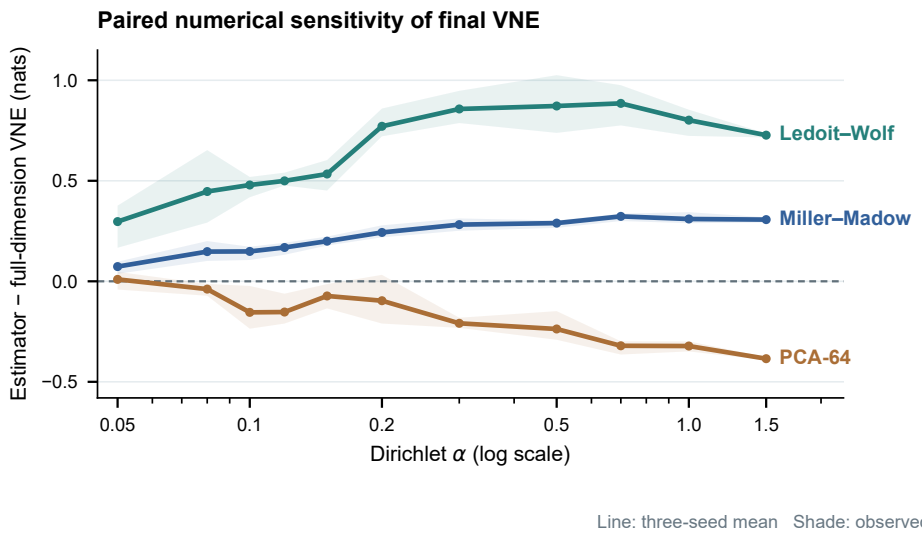

Figure 10: Sensitivity of final VNE to its numerical estimator across 11 Dirichlet concentrations. For each of three seeds, the plotted quantity is the alternative estimate minus the full-dimensional Gram estimate on the same FedAvg model. Lines are means of paired differences; translucent bands span the observed seed range. PCA-64 uses a 64-dimensional feature space whereas the reference and the other alternatives use 512 dimensions.

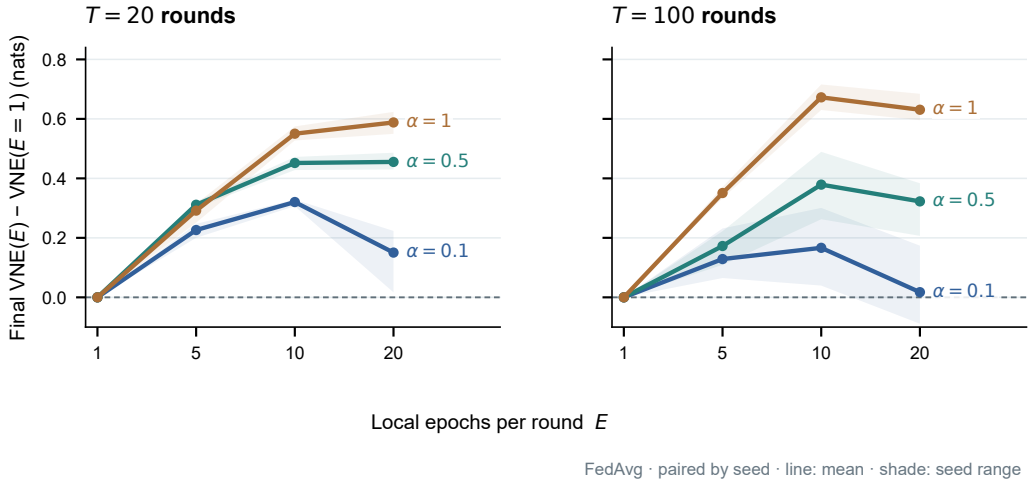

Figure 11: Sensitivity of FedAvg's final VNE to local epochs. Within each fixed round budget ($T = 20$ or 100), each point is the three-seed mean of the paired final-VNE difference from $E_{\text{ep}} = 1$ at the same Dirichlet concentration; translucent bands span the observed seed range. The two panels have a common vertical scale. The $E_{\text{ep}} = 20$ response is lower than $E_{\text{ep}} = 10$ at $\alpha = 0.1$ in both budgets.

D METHOD AND EVALUATION DETAILS

D.1 FEDVNA

Algorithm 1 summarizes one communication round.

Algorithm 1 FedVNA: one communication round.

The server reference is held fixed during the round; spectral terms use minibatch features only.

```
# f(x, theta) -> (z, logits): m-dim features and class logits
# rho_bar[c]: server reference for class c (None before the first aggregation)

def client_update(theta, D_k, rho_bar):
    for epoch in range(E):
        for x, y in loader(D_k):
            z, logits = f(x, theta)
            loss = ce_loss(logits, y)
            if rho_bar is not None:
                valid = classes_with_two_samples(y)
                for c in valid:
                    W = center(z[y == c])                    # n_c x m
                    S = batch_entropy(W, eps_r, eps_c)       # Eq. 18
                    Cov = W.T @ W / (len(W) - 1)
                    cross = trace(Cov @ log_clip(rho_bar[c], eps_c)) \
                            / (trace(Cov) + len(W) * eps_r)
                    loss += (g1 * (-S - cross) - g2 * S) / len(valid)
            theta = sgd_step(theta, loss)
    # class density matrices from final-local-epoch features
    rho = [density(final_epoch_features(c)) if n_k[c] >= 2 else eye(m) / m
           for c in range(n_classes)]
    return theta, rho

# server: one round (clients run in parallel)
outs = [client_update(theta_global, D[k], rho_bar) for k in range(K)]
theta_global = sum(w[k] * outs[k][0] for k in range(K))
rho_bar = [sum(w[k] * outs[k][1][c] for k in range(K))    # Eq. 19
           for c in range(n_classes)]
```

The server reference $\bar{\rho}_y$ in Equation (19) is the data-weighted average of the clients' class-conditional density matrices, formed from final-local-epoch features in the common m-dimensional feature space (Figure 1(c)). A client with fewer than two samples of class y contributes the maximally mixed density matrix I_m/m, so every $\bar{\rho}_y$ averages all K clients with weights w_k; these weights play the role of v_k in Proposition 6 and in the disagreement χ of Section 7. It is broadcast with the model and held fixed during the next round, so gradients pass through the minibatch spectral terms but not through the aggregation. Each client sends one $m \times m$ matrix per class; with ten classes and $m = 512$ this is $10 \times 512^2 \times 4$ bytes, or 10 MiB, per client and round in each direction. On our hardware, the median time per communication round, averaged over three seeds, was 91–105 s for FedVNA and about 25 s for FedAvg across the three heterogeneity levels.

A class enters the minibatch loss (21) when it has at least two samples and a nonzero regularized Gram trace; the terms are averaged over these classes. The entropy term uses the eigenvalues of the $n_y \times n_y$ matrix in Equation (18), which avoids an $m \times m$ eigendecomposition per class, and the cross term uses the same centered features with the clipped logarithm Λ_y. We set $\epsilon_{\mathrm{r}} = \epsilon_{\mathrm{c}} = 10^{-6}$; the two symbols distinguish ridge regularization from eigenvalue clipping. No reference exists before the first aggregation, so the first round uses the task loss alone.

D.2 EVALUATION PROTOCOL

Accuracy is evaluated after every communication round on the complete CIFAR-10 test set, which gives both the final and the within-run best accuracy in Table 1. Class-conditional VNE is computed from centered features on the client training partitions; by the identity in Section 2, its nonzero spectrum can be obtained from the Gram matrix. Table entries are the arithmetic mean and sample standard deviation over seeds 42, 43, and 44; communication rounds within one run are trajectory points, not replicates.