BREAKING THE LEARNING-RATE DILEMMA IN CLUSTERED FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Clustered federated learning (CFL) jointly learns the client partition and the model for each cluster, where the cluster models are derived via model averaging. Existing CFL methods have shown that CFL depends on good initialization, and these methods often use small or diminishing training learning rates for stable model optimization. If the initialization is inaccurate in CFL, a client may be initially assigned to the wrong cluster; in this way, the small or diminishing learning rate and model averaging can restrict the updates that the client makes to the wrong cluster model, leaving it trapped in the wrong cluster persistently. We refer to this as the *learning-rate dilemma* in CFL. To break this learning-rate dilemma, we propose Probe-based Cluster Recovery for CFL (`PCR-CFL`), which calculates the accumulated distances between the responses of the clients to a sequence of probe models, and uses spectral clustering to estimate the number of clusters and obtain the client partition. This process allows a misassigned client to leave the wrong cluster since its responses to the same sequence of probe models are close to those of clients in its ground-truth cluster. We theoretically prove that `PCR-CFL` recovers the ground-truth cluster structure with high probability after a finite number of probes, and further establish model convergence under both constant and diminishing learning rates. Experiments on extensive datasets, together with extensive ablations, demonstrate the effectiveness and robustness of `PCR-CFL`.

1 INTRODUCTION

Federated learning enables collaborative model training among clients without exchanging their raw data. The most prominent algorithm `FedAvg` (McMahan et al., 2017) has been applied to various real-world applications (Brisimi et al., 2018; Yang et al., 2020). In practice, however, clients may have different data distributions or learning tasks; thus, a single global model may not perform well for each client (Deng et al., 2020a; Even et al., 2022; Ding & Wang, 2022; Ye et al., 2023). Clustered federated learning (CFL) addresses this heterogeneity by choosing appropriate group structures for the clients and learning the model for each group (Ghosh et al., 2020; Long et al., 2023).

Existing CFL methods typically initialize several models first, after which the clients perform a small number of local updates to refine these models. The server then averages the updated models and broadcasts the resulting cluster models for the next round. Representative CFL methods with theoretical analyses differ primarily in how they assign clients to the clusters. `IFCA` (Ghosh et al., 2020) alternates between selecting the cluster model with the smallest loss value and updating the model, but it requires the ground-truth number of clusters and well-initialized models. `SR-FCA` (Vardhan et al., 2024) avoids specifying the number of clusters by starting from singleton clusters and iteratively merging them according to model distances. However, its refinement guarantee requires the fraction of misassigned members in each initial cluster to be less than one half, and thus potentially relies on good initialization. `CFL-GP` (Kim et al., 2024) and the method in (Werner et al., 2023) group clients based on gradient information, under the assumption that the gradients or momentum from different clusters remain separable throughout training. However, under poor initialization, a model affected by an incorrect grouping may remain near a point where gradients are difficult to separate, yielding a weak clustering signal. Consequently, although existing CFL methods generally depend on good initialization, few works explain why initialization becomes critical in CFL. Here, we investigate how small training learning rates can make inaccurate initialization lead to persistent clustering failure. Stable model optimization generally requires a small or diminishing learning rate

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.

(Stich, 2019; Khaled et al., 2020; Haddadpour & Mahdavi, 2019; Li et al., 2020b). If the initialization is inaccurate in CFL, a client may initially be assigned to the wrong cluster and then perform only a few local updates on the wrong cluster model. Since the learning rate is small and its model update is averaged with those of other clients, its effect on the cluster model may remain too small to change the subsequent clustering assignment. The client can therefore be trapped in the wrong cluster persistently. We refer to this as the *learning-rate dilemma* in CFL, and provide an example in Section 4 to show that an initially incorrect partition can persist when the learning rate is small.

A basic but challenging question in CFL is therefore: *How can we break the learning-rate dilemma in CFL?* In this paper, we propose Probe-based Cluster Recovery for CFL (`PCR-CFL`) to break the learning-rate dilemma without requiring the ground-truth number of clusters or an accurate initial partition. `PCR-CFL` measures the pairwise distances between the responses of the clients to a sequence of probe models, and uses the accumulated distances to construct the affinity matrix. It then estimates the number of clusters through the eigengap and applies spectral clustering to obtain the client partition. This process allows a misassigned client to leave the wrong cluster since its responses to the same sequence of probe models are close to those of clients in its ground-truth cluster. We theoretically prove that `PCR-CFL` recovers the ground-truth cluster structure from any feasible initialization with high probability after a finite number of probes, and further establish model convergence under both constant and diminishing learning rates. Experiments on extensive datasets with different cluster structures, together with extensive ablations, demonstrate the effectiveness and robustness of `PCR-CFL` in terms of cluster recovery and model performance.

2 Related Works

Clustered federated learning (CFL) aims to address the data heterogeneity problem in federated learning (McMahan et al., 2017) by grouping clients with similar data or objectives and training a model for each group (Ghosh et al., 2020; Chen et al., 2020a; Ruan & Joe-Wong, 2022; Vahidian et al., 2023; Vardhan et al., 2024; Long et al., 2023; Kim et al., 2024; Werner et al., 2023; Guo et al., 2024; Yu et al., 2024; Carrillo et al., 2024). Existing heuristic CFL methods differ in their notion of client similarity and their assumptions on latent groups. For example, Su et al. (2022) assumed that the data on each client are generated from one of the regression models; Ruan & Joe-Wong (2022) assumed that the data on each client follow a mixture of multiple distributions, and exploited the similarity between source distributions to learn the group models; Vahidian et al. (2023) utilized the principal angles between the client data subspaces to derive well-performing models; Guo et al. (2024) proposed FedRC based on soft clustering in the presence of multiple types of distribution shifts across clients. However, these works mainly study different clustering models or optimization mechanisms heuristically, and do not theoretically establish cluster-recovery guarantees for CFL.

Representative CFL methods with theoretical analyses primarily differ in how they assign clients to clusters. `IFCA` (Ghosh et al., 2020) assigns each client to the cluster model with the smallest loss value, but its guarantees require the ground-truth number of clusters and well-initialized models. `SR-FCA` (Vardhan et al., 2024) starts from singleton clusters and merges them according to model distances. It does not require the number of clusters, but its refinement guarantee requires the fraction of misassigned members in each initial cluster to be less than one half, and thus potentially relies on good initialization. `CFL-GP` (Kim et al., 2024) and the methods in (Werner et al., 2023) group clients using gradients or momentum information and require the ground-truth number of clusters. Their analyses assume that the gradients or momentum from different clusters remain separated during training. However, with a poor initialization and a small learning rate, a model affected by a wrong group may remain near a point where gradients are hard to separate. We provide the examples in Appendix D.1 to show that, under poor initialization, limited local computation can cause `SR-FCA`'s initial model-based clustering to merge distinct tasks and violate its refinement requirements, while a small training learning rate can sustain weak gradient separation along an incorrectly mixed trajectory, weakening the clustering signal used by `CFL-GP`.

Federated learning typically requires a small or diminishing learning rate for stable model optimization (Stich, 2019; Haddadpour & Mahdavi, 2019; Khaled et al., 2020; Li et al., 2020b; Karimireddy et al., 2020; Wang et al., 2020b; Reddi et al., 2021; Wang & Joshi, 2021). In particular, Li et al. (2020b) have shown that a diminishing client learning rate is necessary for the convergence of federated learning on non-IID data, Karimireddy et al. (2020) further explained that `FedAvg` requires

smaller local learning rates to control client drift under data heterogeneity, and Khaled et al. (2020) have shown that the learning rate should become even smaller as the number of local updates grows. Therefore, a small learning rate is necessary for stable model optimization in federated learning. Additional related works are provided in Appendix A.

3 PRELIMINARIES

Consider N clients indexed by $[N] \coloneqq \{1, \dots, N\}$. Let $\mathcal{W} \subseteq \mathbb{R}^d$ be a nonempty, closed, convex model parameter space with finite positive diameter $R \coloneqq \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{W}} \|\boldsymbol{u} - \boldsymbol{v}\|$, where $d \in \mathbb{N}$ is the model dimension. Let $\mathcal{D}_i$ be the data distribution of client i, let $\xi \sim \mathcal{D}_i$ be a random local example, and let $\ell(\boldsymbol{w}; \xi)$ be its loss for model parameters $\boldsymbol{w} \in \mathcal{W}$, where $\ell(\cdot)$ is the loss function. Let $f_i(\boldsymbol{w}) \coloneqq \mathbb{E}_{\xi\sim\mathcal{D}_i}[\ell(\boldsymbol{w}; \xi)]$ be the risk of client i, and let $\boldsymbol{w}_i^\star \coloneqq \arg\min_{\boldsymbol{w}\in\mathcal{W}} f_i(\boldsymbol{w})$ be its risk minimizer. The goal of clustered federated learning is to assign clients to clusters and learn a model for each cluster. Let $K^\star$ denote the ground-truth number of clusters, and let $\mathcal{C}^\star = \{\mathcal{C}_1^\star, \dots, \mathcal{C}_{K^\star}^\star\}$ denote the ground-truth client partition, where $\bigcup_{k=1}^{K^\star} \mathcal{C}_k^\star = [N]$ and $\mathcal{C}_k^\star \cap \mathcal{C}_{k'}^\star = \varnothing, \forall k \neq k'$. Let the ground-truth cluster assignment of client i be $z_i^\star$, where $z_i^\star = k$ if $i \in \mathcal{C}_k^\star$. Let the client weights satisfy $p_i > 0$ and $\sum_{i=1}^N p_i = 1$, and denote the mass of cluster k as $P_k \coloneqq \sum_{i\in\mathcal{C}_k^\star} p_i$. For the ground-truth cluster $k \in [K^\star]$, define its weighted risk by $F_k(\boldsymbol{w}) \coloneqq \frac{1}{P_k}\sum_{i\in\mathcal{C}_k^\star} p_i f_i(\boldsymbol{w})$, and let $\boldsymbol{\theta}_k^\star \coloneqq \arg\min_{\boldsymbol{w}\in\mathcal{W}} F_k(\boldsymbol{w})$ be its optimal cluster model. Let $\nabla f_i(\boldsymbol{w}) \in \mathbb{R}^d$ denote the gradient of f_i with respect to $\boldsymbol{w}$. At round $t \in \mathbb{N}_0$, let $\xi_i^t \sim \mathcal{D}_i$ be the random local sample of client i, and let $\mathbf{g}_i(\boldsymbol{w}; \xi_i^t) \in \mathbb{R}^d$ be the stochastic gradient evaluated at $\boldsymbol{w}$. Throughout this paper, we make the following commonly used assumptions in the literature on stochastic optimization (Rakhlin et al., 2012; Johnson & Zhang, 2013; Ghadimi & Lan, 2013; Bottou et al., 2018) and theoretical analyses of CFL (Ghosh et al., 2020; Vardhan et al., 2024; Kim et al., 2024).

Assumption 3.1 (Smoothness). There exists a constant $L > 0$ such that, for every $i \in [N]$, f_i is differentiable and L-smooth on $\mathcal{W}$:

$$\|\nabla f_i(\boldsymbol{u}) - \nabla f_i(\boldsymbol{v})\| \leqslant L\|\boldsymbol{u} - \boldsymbol{v}\|, \qquad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathcal{W}. \tag{1}$$

Assumption 3.2 (Strong Convexity). There exists a constant $\mu \in (0, L]$ such that, for every $i \in [N]$, f_i is μ-strongly convex on $\mathcal{W}$:

$$f_i(\boldsymbol{v}) \geqslant f_i(\boldsymbol{u}) + \langle \nabla f_i(\boldsymbol{u}), \boldsymbol{v} - \boldsymbol{u}\rangle + \frac{\mu}{2}\|\boldsymbol{v} - \boldsymbol{u}\|^2, \qquad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathcal{W}. \tag{2}$$

Assumption 3.3 (Unbiased Stochastic Gradients with Bounded Variance). Let $\mathcal{F}_t$ be the sigma-field generated by all randomness before the samples $\{\xi_i^t\}_{i=1}^N$ are drawn at round t. There exists a constant $\sigma^2 \geqslant 0$ such that, for every $i \in [N]$, every round $t \in \mathbb{N}_0$, and every $\mathcal{F}_t$-measurable $\boldsymbol{w} \in \mathcal{W}$,

$$\mathbb{E}[\mathbf{g}_i(\boldsymbol{w}; \xi_i^t) \mid \mathcal{F}_t] = \nabla f_i(\boldsymbol{w}), \qquad \mathbb{E}[\|\mathbf{g}_i(\boldsymbol{w}; \xi_i^t) - \nabla f_i(\boldsymbol{w})\|^2 \mid \mathcal{F}_t] \leqslant \sigma^2. \tag{3}$$

4 OUR WORK

A small or diminishing training learning rate is commonly used for stable model optimization (Li et al., 2020b; Karimireddy et al., 2020; Wang et al., 2020b; Reddi et al., 2021; Wang & Joshi, 2021), but it can trap misassigned clients in incorrect clusters in CFL. We refer to this conflict as the *learning-rate dilemma*, and provide the following example for better illustration.

Example of the Learning-Rate Dilemma. Consider $N = 4$ clients and 2 ground-truth clusters $\mathcal{C}_1^\star = \{1, 2\}$ and $\mathcal{C}_2^\star = \{3, 4\}$. For any $a > 0$, let $\mathcal{W} = [-a, 2a] \times [-a, 3a]$, and $\boldsymbol{H} = \mathrm{diag}(4, 1)$. Define the risk as $f_i(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}_i^\star)^\mathsf{T}\boldsymbol{H}(\boldsymbol{w}-\boldsymbol{w}_i^\star)$, where $\boldsymbol{w}_1^\star = (a, 0)^\mathsf{T}$, $\boldsymbol{w}_2^\star = (0, 0)^\mathsf{T}$, and $\boldsymbol{w}_3^\star = \boldsymbol{w}_4^\star = (0, 2a)^\mathsf{T}$. Initialize the models such that $\|\boldsymbol{\theta}_1^0 - (a, 0)^\mathsf{T}\| < a/10$ and $\|\boldsymbol{\theta}_2^0 - (0, 4a/3)^\mathsf{T}\| < a/10$. Client 1 is initialized with $\boldsymbol{\theta}_1^0$, while clients 2, 3, and 4 are initialized with $\boldsymbol{\theta}_2^0$. Therefore, the initial client partition is $\mathcal{C}_1^0 = \{1\}$ and $\mathcal{C}_2^0 = \{2, 3, 4\}$, where client 2 is incorrectly assigned since its ground-truth cluster is $\mathcal{C}_1^\star$. At round t, let $z_i^t \in [2]$ be the cluster assignment of client i, and let $\mathcal{C}^t = \{\mathcal{C}_1^t, \mathcal{C}_2^t\}$ be the current client partition, where $\mathcal{C}_k^t \coloneqq \{i \in [4] : z_i^t = k\}$. The cluster models are updated by $\boldsymbol{\theta}_k^{t+1} = \boldsymbol{\theta}_k^t - \frac{\gamma}{|\mathcal{C}_k^t|}\sum_{j\in\mathcal{C}_k^t} \nabla f_j(\boldsymbol{\theta}_k^t)$, where the learning rate $0 < \gamma < 1/4$. Here,

$1/4 = 1/L$ is the standard monotone-descent scale for the 4-smooth risks. Consider the following assignment rules: ① The *model-distance rule* assigns client i by the distance between each cluster model and its local update $\boldsymbol{u}_i^t := \boldsymbol{\theta}_{z_i^t}^{t+1} - \alpha \nabla f_i(\boldsymbol{\theta}_{z_i^t}^{t+1})$, where $0 < \alpha < 1/4$. ② The *loss-value rule* assigns client i to the model with the smallest f_i value. ③ The *gradient-distance rule* evaluates the gradient $\mathbf{g}_i^t := \nabla f_i(\boldsymbol{\phi}_t)$ at the same model $\boldsymbol{\phi}_t$ for the clients and applies Euclidean two-means clustering to $\{\mathbf{g}_i^t\}_{i=1}^4$. Under the above initialization, each of the three rules gives $\mathcal{C}_1^t = \{1\}$ and $\mathcal{C}_2^t = \{2, 3, 4\}$ for every $t \geqslant 0$. In particular, client 2 remains incorrectly assigned, while $\boldsymbol{\theta}_1^t$ and $\boldsymbol{\theta}_2^t$ converge to $(a, 0)^\mathsf{T}$ and $(0, 4a/3)^\mathsf{T}$, respectively.

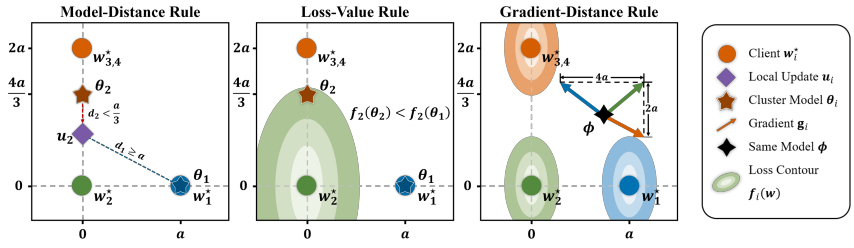


Figure 1: **Persistence of incorrect client assignments under a small learning rate.** A misassigned client updates the wrong cluster model, which preserves the assignment error in subsequent rounds.

The derivation of the above result is deferred to Appendix D.1 due to space limitations. The example can be generalized to multiple clusters and is illustrated in Figure 1. This example shows that a small learning rate does not ensure that the clustering process can correct itself: the wrong partition remains invariant, and the cluster models converge to the risk minimizers induced by the wrong partition. In this regime, a small training learning rate can lead to persistent cluster-recovery failure by keeping the iterates in a region that preserves the incorrect assignments.

4.1 A Strategy for Breaking the Learning-Rate Dilemma

We propose Probe-based Cluster Recovery for Clustered Federated Learning (`PCR-CFL`) to break the learning-rate dilemma. `PCR-CFL` uses the responses of clients to the probe models to estimate the number of clusters and assign clients to the clusters, and reuses the trained cluster models when the client partition changes. The complete procedure of `PCR-CFL` is shown in Algorithm 1.

The algorithm starts with the client models $\{\boldsymbol{w}_i^0\}_{i=1}^N \subset \mathcal{W}$, which yield the initial cluster models $\{\boldsymbol{\theta}_k^0\}_{k=1}^{K_0} \subset \mathcal{W}$ via clustering, where K_0 is the initial number of clusters. Denote the initial client partition as $\mathcal{C}^0 = \{\mathcal{C}_k^0\}_{k=1}^{K_0}$. At round t, let K_t denote the current number of clusters and let $\mathcal{C}^t = \{\mathcal{C}_k^t\}_{k=1}^{K_t}$ denote the current client partition. Denote the cluster assignment of client i by z_i^t, where $z_i^t = k$ if $i \in \mathcal{C}_k^t$. For each current cluster $k \in [K_t]$, denote its mass and aggregated stochastic gradient by $P_k^t := \sum_{i \in \mathcal{C}_k^t} p_i$ and $\mathbf{g}_k^t := \sum_{i \in \mathcal{C}_k^t} \frac{p_i}{P_k^t} \mathbf{g}_i(\boldsymbol{\theta}_k^t; \xi_i^t)$ respectively. The server then updates the cluster model by $\widetilde{\boldsymbol{\theta}}_k^{t+1} = \Pi_{\mathcal{W}}\left(\boldsymbol{\theta}_k^t - \eta_t \mathbf{g}_k^t\right), k \in [K_t]$, where $\Pi_{\mathcal{W}}$ is the Euclidean projection onto $\mathcal{W}$, and η_t is the learning rate for training the cluster models.

Let $H \in \mathbb{N}$ be the probe interval, where the r-th clustering probe is performed at round $t_r := rH$. At probe r, the server selects a current cluster model $\boldsymbol{\phi}_r := \widetilde{\boldsymbol{\theta}}_{c_r}^{t_r}$, where $c_r := 1 + (r-1) \bmod K_{t_r - 1}$. The vector $\boldsymbol{\phi}_r \in \mathcal{W}$ is the probe model sent to the clients, and $c_r \in [K_{t_r-1}]$ is the index of the selected cluster model. Let $E \in \mathbb{N}$ be the number of local probe updates, let $B \in \mathbb{N}$ be the probe minibatch size, and let $\alpha > 0$ be the probe learning rate. At probe r, client i initializes $\boldsymbol{\phi}_i^{r,0} = \boldsymbol{\phi}_r$ and performs the following temporary local updates:

$$\boldsymbol{\phi}_i^{r,e+1} := \Pi_{\mathcal{W}}\left(\boldsymbol{\phi}_i^{r,e} - \frac{\alpha}{B} \sum_{b=1}^{B} \mathbf{g}_i(\boldsymbol{\phi}_i^{r,e}; \xi_{i,b}^{r,e})\right), \quad e = 0, \ldots, E-1, \qquad \widetilde{\boldsymbol{\phi}}_i^r := \boldsymbol{\phi}_i^{r,E}. \tag{4}$$

Here $\widetilde{\boldsymbol{\phi}}_i^r$ is the adapted probe model returned by client i, and $\xi_{i,b}^{r,e} \sim \mathcal{D}_i$ is the b-th sample in client i's probe minibatch at local step e. The clients start from the same $\boldsymbol{\phi}_r$, so each client's probe response is determined by its local task rather than its current cluster assignment. Across the same sequence of probe models, clients in the same ground-truth cluster produce similar responses;

Algorithm 1 Probe-based Cluster Recovery for Clustered Federated Learning (PCR-CFL)

Input: affinity threshold h, probe interval H, local probe updates E, probe minibatch size B, and learning rates $\{\eta_t\}_{t=0}^{T-1}$ and α.
Initialize: initial number of clusters K_0, client partition $\mathcal{C}^0$, and cluster models $\{\boldsymbol{\theta}_k^0\}_{k=1}^{K_0}$.
for $t = 0, \dots, T-1$ **do**
  Update the cluster models by $\widetilde{\boldsymbol{\theta}}_k^{t+1} = \Pi_{\mathcal{W}}\left(\boldsymbol{\theta}_k^t - \eta_t \mathbf{g}_k^t\right), k \in [K_t]$.
  if $t+1 = rH$ for some $r \in \mathbb{N}$ **then**
    Select the probe model $\boldsymbol{\phi}_r \coloneqq \widetilde{\boldsymbol{\theta}}_{c_r}^{t_r}$, where $c_r \coloneqq 1 + (r-1) \bmod K_{t_r-1}$.
    Each client obtains the adapted probe model $\widetilde{\boldsymbol{\phi}}_i^r$ by Eq. (4).
    Construct the normalized Laplacian matrix $\boldsymbol{L}_r$ by Eq. (5).
    Estimate K_{t+1} by Eq. (6) and obtain the client partition $\mathcal{C}^{t+1}$ by Eq. (7).
    Reuse the old cluster models for $\mathcal{C}^{t+1}$ by Eq. (8).
  else
    Set $K_{t+1} = K_t$, $\mathcal{C}^{t+1} = \mathcal{C}^t$, and $\boldsymbol{\theta}_k^{t+1} = \widetilde{\boldsymbol{\theta}}_k^{t+1}$ for every $k \in [K_t]$.
  end if
end for
Output: number of clusters K_T, client partition $\mathcal{C}^T$, and cluster models $\{\boldsymbol{\theta}_k^T\}_{k=1}^{K_T}$.

thus, a misassigned client can be identified by comparing its responses with those of other clients. Since a probe is performed every H rounds, the average extra computation is E/H local probe updates per round. The server then averages the pairwise response distances over the completed probes: $d_{i,j}(r) \coloneqq \frac{1}{r}\sum_{u=1}^{r} \mathrm{dist}(\widetilde{\boldsymbol{\phi}}_i^u, \widetilde{\boldsymbol{\phi}}_j^u), \forall i, j \in [N]$, where $\mathrm{dist}(\cdot,\cdot)$ is the squared Euclidean distance between two adapted probe responses, *i.e.*, $\mathrm{dist}(\boldsymbol{u}, \boldsymbol{v}) \coloneqq \|\boldsymbol{u} - \boldsymbol{v}\|^2$, where a small $d_{i,j}(r)$ means the two clients adapt similarly to a sequence of models. In Appendix D.1, we establish the within-/cross-cluster gap after finitely many adaptation steps for the multi-probe procedure, which shows that the multi-probe procedure can separate task responses and accumulate stable distance estimates under stochastic sampling. To convert the accumulated distances into clustering decisions, let $h > 0$ be the affinity threshold; the server then constructs the affinity matrix $\boldsymbol{W}_r \in \mathbb{R}^{N\times N}$ with $W_{i,j}^r = \frac{1}{N}\mathbf{1}\{d_{i,j}(r) < h^2\}, \forall i \neq j$ and $W_{i,i}^r = 0, \forall i \in [N]$. Let $\boldsymbol{D}_r$ be the diagonal degree matrix with $D_{i,i}^r = \sum_{j=1}^{N} W_{i,j}^r$. The normalized Laplacian matrix is defined as:

$$\boldsymbol{L}_r = \boldsymbol{I}_N - \boldsymbol{D}_r^{-1/2}\boldsymbol{W}_r\boldsymbol{D}_r^{-1/2}, \tag{5}$$

where $\boldsymbol{I}_N$ is the $N \times N$ identity matrix and a zero entry of $\boldsymbol{D}_r^{-1/2}$ is used when $D_{i,i}^r = 0$. The eigenvalues and eigenvectors of $\boldsymbol{L}_r$ provide the cluster-number estimate and the spectral representations. For the symmetric Laplacian matrix $\boldsymbol{L}$, denote its ordered eigenvalues by $\lambda_1(\boldsymbol{L}) \leqslant \cdots \leqslant \lambda_N(\boldsymbol{L})$. The server estimates the number of clusters according to the eigengap as follows:

$$K_{t_r} = \underset{1\leqslant k<N}{\arg\max}\left\{\lambda_{k+1}(\boldsymbol{L}_r) - \lambda_k(\boldsymbol{L}_r)\right\}. \tag{6}$$

The server maps each client to a spectral representation for partitioning. Denote $\boldsymbol{Q}_r \in \mathbb{R}^{N\times K_{t_r}}$ as the matrix containing orthonormal eigenvectors associated with the K_{t_r} smallest eigenvalues of $\boldsymbol{L}_r$, where client i is represented by the normalized i-th row $\boldsymbol{q}_i^r \coloneqq (\boldsymbol{Q}_r)_{i,:}^{\intercal}/\|(\boldsymbol{Q}_r)_{i,:}\|$ (a zero row is kept as zero). The server then applies the following rule to obtain the K_{t_r} representatives: it sets $j_1^r = 1$ and selects $j_\ell^r \coloneqq \arg\max_{i\in[N]\setminus\{j_1^r,\dots,j_{\ell-1}^r\}} \min_{a\in[\ell-1]} \|\boldsymbol{q}_i^r - \boldsymbol{q}_{j_a^r}^r\|$ for $\ell = 2, \dots, K_{t_r}$. Here, $j_\ell^r \in [N]$ is the client index of the ℓ-th spectral representative. Each representative keeps its own label, and every other client is assigned to its nearest representative:

$$z_i^{t_r} \coloneqq \begin{cases} \ell, & i = j_\ell^r, \\ \underset{\ell\in[K_{t_r}]}{\arg\min}\|\boldsymbol{q}_i^r - \boldsymbol{q}_{j_\ell^r}^r\|, & i \notin \{j_1^r, \dots, j_{K_{t_r}}^r\}, \end{cases} \qquad \mathcal{C}_\ell^{t_r} \coloneqq \{i \in [N] : z_i^{t_r} = \ell\}. \tag{7}$$

This step forms a new partition from the probe distances, where a misassigned client can leave the wrong cluster when its probe responses are closer to those of clients in its ground-truth cluster. The eigengap also estimates K_{t_r}, so the correction does not require the ground-truth number of clusters.

At a probe round, the number of clusters and client partition may change. Denote the overlap mass between the new cluster $\ell \in [K_{t_r}]$ and the previous cluster $k \in [K_{t_r-1}]$ by $\omega_{\ell k}^{t_r} \coloneqq \sum_{i\in\mathcal{C}_\ell^{t_r}\cap\mathcal{C}_k^{t_r-1}} p_i$.

To avoid restarting model training after repartitioning, the server reuses the old cluster models:

$$\boldsymbol{\theta}_{\ell}^{t_r} := \sum_{k=1}^{K_{t_r-1}} \frac{\omega_{\ell k}^{t_r}}{\sum_{k'=1}^{K_{t_r-1}} \omega_{\ell k'}^{t_r}} \widetilde{\boldsymbol{\theta}}_{k}^{t_r}. \tag{8}$$

At a non-probe round, the client partition is unchanged and $\boldsymbol{\theta}_k^{t+1} = \widetilde{\boldsymbol{\theta}}_k^{t+1}$. This step preserves the training information carried by clients that remain together; thus, correcting the partition does not require restarting model training, since the groups can continue training from their previous models.

4.2 Theoretical Analysis

We next study whether PCR-CFL recovers the ground-truth cluster structure and whether the cluster models converge. We first define the following separable cluster structure, which is commonly used in the literature on clustering (Kumar & Kannan, 2010; Awasthi & Sheffet, 2012; Ostrovsky et al., 2013) and CFL (Ghosh et al., 2020; Dennis et al., 2021; Vardhan et al., 2024).

Definition 4.1 (Separable Cluster Structure). We say $\mathcal{C}^\star = \{\mathcal{C}_1^\star, \dots, \mathcal{C}_{K^\star}^\star\}$ satisfies the separable cluster structure if there exist cluster centers $\{\boldsymbol{c}_k^\star\}_{k=1}^{K^\star}$, constants $\epsilon \geqslant 0$ and $\Delta > 0$ such that $\|\boldsymbol{w}_i^\star - \boldsymbol{c}_k^\star\| \leqslant \epsilon, \forall i \in \mathcal{C}_k^\star$, and $\min_{k \neq \ell} \|\boldsymbol{c}_k^\star - \boldsymbol{c}_\ell^\star\| \geqslant \Delta$. Moreover, $|\mathcal{C}_k^\star| \geqslant 2$ for every $k \in [K^\star]$.

The separable cluster structure property reveals that the risk minimizers of clients in the same cluster stay close, whereas the centers of different clusters are separated. Each cluster has at least two clients since comparing the consistency of probe distances requires at least one within-cluster client pair. Kim et al. (2024) assumed a positive average gradient gap over the queried models; we instead analyze the more natural cluster structure. Before proceeding to the theorems, we first introduce the notation. The goal of CFL is to recover the cluster structure and learn the cluster models. When $K_t = K^\star$, we define the weighted clustering error e_t^{c} and the weighted cluster-model error e_t^{m} as:

$$e_t^{\mathrm{c}} := \min_{\pi \in \mathfrak{S}_{K^\star}} \sum_{i=1}^{N} p_i \mathbf{1}\{z_i^t \neq \pi(z_i^\star)\}, \qquad e_t^{\mathrm{m}} := \sum_{k=1}^{K^\star} P_k \|\boldsymbol{\theta}_{\pi_t(k)}^t - \boldsymbol{\theta}_k^\star\|^2. \tag{9}$$

Here, $\mathfrak{S}_{K^\star}$ is the set of permutations of $[K^\star]$, P_k is the weight of ground-truth cluster k, and π_t is the permutation that attains the minimum in e_t^{c}. When $K_t \neq K^\star$, we set $e_t^{\mathrm{c}} := 1$ and $e_t^{\mathrm{m}} := R^2$. We define the exact-recovery event as $\mathcal{E}_t := \{e_t^{\mathrm{c}} = 0\}$, which holds exactly when $K_t = K^\star$ and $\mathcal{C}^t = \mathcal{C}^\star$ up to a label permutation by the definition of e_t^{c}. The first theorem answers whether PCR-CFL can leave an inaccurate initial client partition and identify the ground-truth cluster structure.

Theorem 4.2 (Cluster Structure Recovery). *Suppose Assumption 3.1, 3.2, 3.3 and Definition 4.1 hold, all the probe samples are drawn independently at each probe, and $0 < \alpha \leqslant \mu/L^2$. Denote $r_{\mathrm{p}} := (e^{-\mu\alpha E} R^2 + \alpha\sigma^2/(B\mu))^{1/2}$. Suppose $\Delta > 4(\epsilon + r_{\mathrm{p}})$, choose h such that $2(\epsilon + r_{\mathrm{p}}) < h < \Delta - 2(\epsilon + r_{\mathrm{p}})$, and denote $\kappa_h := \min\{h^2 - (2\epsilon + 2r_{\mathrm{p}})^2, (\Delta - 2\epsilon - 2r_{\mathrm{p}})^2 - h^2\}$. The following result holds for Algorithm 1:*

$$\Pr(\mathcal{E}_{t_r}) \geqslant 1 - 2N^2 e^{-2r\kappa_h^2/R^4}. \tag{10}$$

For any $\delta \in (0,1)$, we have $\Pr(\mathcal{E}_t) \geqslant 1 - \delta, \forall t \geqslant H r_\delta$, where $r_\delta := \max\left\{1, \left\lceil \frac{R^4}{2\kappa_h^2} \log\left(\frac{2N^2}{\delta}\right)\right\rceil\right\}$.

We leave the proof of Theorem 4.2 to Appendix D.3. Theorem 4.2 reveals that, once the number of probes reaches r_δ, the number of clusters and client partition at any subsequent round are correct with probability at least $1 - \delta$. The interval for h is nonempty when $\Delta > 4(\epsilon + r_{\mathrm{p}})$, where r_{p} is the probe-adaptation error radius. Theorem 4.2 gives exact cluster recovery with high probability at any selected round after sufficiently many probes, while the second theorem reveals how the clustering and cluster-model errors change when the cluster models use a constant learning rate.

Theorem 4.3 (Model Convergence with a Constant Learning Rate). *Suppose Assumption 3.1, 3.2, 3.3 and Definition 4.1 hold, all the probe samples are drawn independently at each probe, and $0 < \alpha \leqslant \mu/L^2$. Denote $r_{\mathrm{p}} := (e^{-\mu\alpha E} R^2 + \alpha\sigma^2/(B\mu))^{1/2}$. Suppose $\Delta > 4(\epsilon + r_{\mathrm{p}})$, choose h such that $2(\epsilon + r_{\mathrm{p}}) < h < \Delta - 2(\epsilon + r_{\mathrm{p}})$, and denote $\kappa_h := \min\{h^2 - (2\epsilon + 2r_{\mathrm{p}})^2, (\Delta - 2\epsilon - 2r_{\mathrm{p}})^2 - h^2\}$. Let $r(t) := \lfloor t/H \rfloor$ be the number of completed probes by round t, let $0 < \eta \leqslant \min\{1/L, \mu/L^2\}$ and set $\eta_t = \eta$. Denote $\rho := 1 - \mu\eta$. Then, for every $t \geqslant 1$, the following result holds for Algorithm 1:*

$$\mathbb{E}[e_t^{\mathrm{c}}] \leqslant \min\left\{1, 2N^2 e^{-2r(t)\kappa_h^2/R^4}\right\}, \tag{11}$$

$$\mathbb{E}[e_t^{\mathrm{m}}] \leqslant \rho^t R^2 + \frac{\eta\sigma^2}{\mu}\left(1-\rho^t\right) + 4N^2R^2\sum_{\tau=0}^{t-1}\rho^{t-1-\tau}e^{-2r(\tau)\kappa_h^2/R^4}. \tag{12}$$

Consequently, $\limsup_{t\to\infty}\mathbb{E}[e_t^{\mathrm{c}}] = 0$, *and* $\limsup_{t\to\infty}\mathbb{E}[e_t^{\mathrm{m}}] \leqslant \frac{\eta\sigma^2}{\mu}$.

We leave the proof of Theorem 4.3 to Appendix D.4 due to space limitations. Theorem 4.3 reveals that the clustering bound of $\mathbb{E}[e_t^{\mathrm{c}}]$ decreases exponentially with $r(t)$, while in the model bound of $\mathbb{E}[e_t^{\mathrm{m}}]$, $\rho^t R^2$ is the remaining initial error, the second term is caused by stochastic gradients, and the third term contains the effect of incorrect client partitions. The clustering error $\mathbb{E}[e_t^{\mathrm{c}}]$ vanishes since clients in the same ground-truth cluster give similar responses to the sequence of probe models, which allows the affinity matrix to correct an inaccurate client partition. The cluster-model error $\mathbb{E}[e_t^{\mathrm{m}}]$ also decreases, but a constant training learning rate leaves the error floor $\eta\sigma^2/\mu$.

When the gradients contain no noise, we provide Proposition D.10 in Appendix D.6 to show that a constant learning rate gives exact clustering and linear model convergence. With stochastic gradients, the following theorem uses a diminishing learning rate to remove the remaining model error.

Theorem 4.4 (Model Convergence with a Diminishing Learning Rate). *Suppose Assumption 3.1, 3.2, 3.3 and Definition 4.1 hold, all the probe samples are drawn independently at each probe, and* $0 < \alpha \leqslant \mu/L^2$. *Denote* $r_{\mathrm{p}} := (e^{-\mu\alpha E}R^2 + \alpha\sigma^2/(B\mu))^{1/2}$. *Suppose* $\Delta > 4(\epsilon + r_{\mathrm{p}})$, *choose* h *such that* $2(\epsilon+r_{\mathrm{p}}) < h < \Delta - 2(\epsilon+r_{\mathrm{p}})$, *and denote* $\kappa_h := \min\{h^2-(2\epsilon+2r_{\mathrm{p}})^2, (\Delta-2\epsilon-2r_{\mathrm{p}})^2-h^2\}$. *Let* $\eta_t = \frac{c_\eta}{t+t_0}$ *where* $\mu c_\eta > 1$ *and* $t_0 \geqslant \max\left\{c_\eta L, c_\eta L^2/\mu\right\}$. *Denote the exact recovery time as* $T_{\mathrm{exact}} := \inf\{T \in \mathbb{N}_0 : \mathcal{E}_t \text{ holds}, \forall t \geqslant T\}$. *For any* $\delta \in (0,1)$, *denote*

$$\bar{r}_\delta := \max\left\{1, \left\lceil \frac{R^4}{2\kappa_h^2}\log\left(\frac{2N^2}{\delta\left(1-e^{-2\kappa_h^2/R^4}\right)}\right)\right\rceil\right\}, \tag{13}$$

then the result $\Pr(T_{\mathrm{exact}} \leqslant H\bar{r}_\delta) \geqslant 1-\delta$ *holds for Algorithm 1, and* $\limsup_{t\to\infty}\mathbb{E}[e_t^{\mathrm{m}}] = 0$.

We leave the proof of Theorem 4.4 to Appendix D.5 due to space limitations. Theorem 4.4 reveals that the diminishing learning rate can make the expected weighted squared cluster-model error vanish despite stochastic gradients. Note that Theorem 4.4 gives a stronger conclusion than recovery at one selected round as in Theorem 4.2, where the ground-truth number of clusters and client partition remain correct after a finite round.

5 Experiments

Our experiments are organized around the following questions: ① Do the empirical cluster-recovery and model convergence results of `PCR-CFL` agree with the theoretical results? ② Does `PCR-CFL` improve model and clustering performance across different cluster structures? ③ Do small learning rates trap the originally misassigned clients for existing CFL methods, and does `PCR-CFL` break this learning-rate dilemma? ④ Are the conclusions stable when changing the number of clients, the batch size, the number of local updates, and the clustering intervals? ⑤ How sensitive is `PCR-CFL` to the affinity threshold, probe minibatch size, and number of local probe updates?

Experimental Setup. We use the following datasets: a synthetic dataset with strongly convex objectives for theory verification; Rotated MNIST and Rotated CIFAR following the standard CFL construction (Ghosh et al., 2020; Vardhan et al., 2024), which have a clear cluster structure; MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky et al., 2009) partitioned using the Dirichlet distribution with $\beta \in \{0.3, 0.5\}$ as in (Yurochkin et al., 2019; Wang et al., 2020a), which have weaker separable cluster structures; and the realistic federated learning dataset FEMNIST (Caldas et al., 2018), which has no prescribed ground-truth cluster structure. We compare with `FedAvg` (McMahan et al., 2017), `IFCA` (Ghosh et al., 2020), `CFL-GP` (Kim et al., 2024), and `SR-FCA` (Vardhan et al., 2024). For datasets with group structures, `IFCA` and `CFL-GP` receive the ground-truth number of clusters since they require a fixed number of clusters, while `SR-FCA` and `PCR-CFL` estimate the number of clusters. `PCR-CFL` selects the affinity threshold h adaptively: after probe r, we apply two-means clustering to $\{d_{i,j}(r) : i < j\}$ and set $h_r^2 = (a_r + b_r)/2$, where $a_r < b_r$ are the means of the two groups. We provide Theorem D.11 in Appendix D.7 to prove that this threshold lies in the recovery interval with high probability when the within-cluster and cross-cluster probe

distances form two separated groups. We record Score, P20, and ARI: Score is the average accuracy for classification and an R^2-style score for the synthetic problem; P20 is the mean score of the worst-performing 20% of clients; the Adjusted Rand index (ARI) (Hubert & Arabie, 1985; Steinley, 2004) evaluates the clustering performance. `FedAvg` has no ARI since it does not perform clustering. The Score and P20 of `FedAvg` are omitted on Synthetic since the score is close to zero. ARI is omitted on FEMNIST since it has no prescribed ground-truth cluster structure. All reported points are means over five seeds, and the error bars represent the standard deviation. Detailed experimental settings are deferred to Appendix B.1.

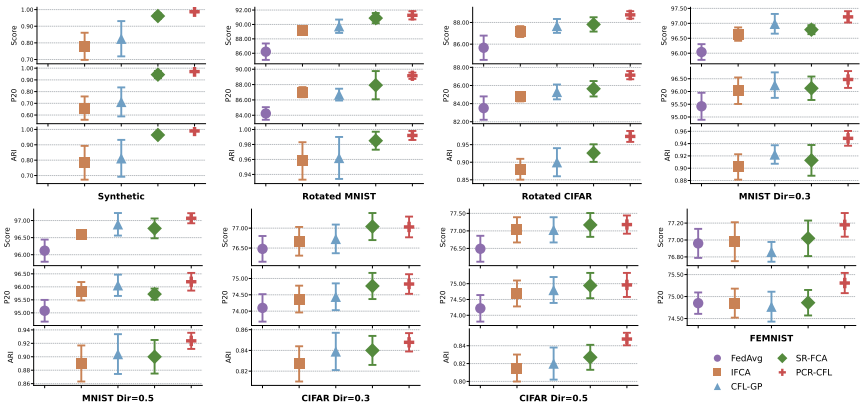


Figure 2: Comparison with baseline methods under the default configuration.

Answer to Question 1. We run `PCR-CFL` on the synthetic dataset to validate the theoretical results. The experimental settings are deferred to Appendix B.2, and the results are shown in Figure 5. Panels (a)–(d) and (i) are consistent with Theorem 4.2, where the recovery frequencies increase with the number of probes and the separation Δ. The results in (i) support finite-probe recovery. Panels (e)–(g) are consistent with Theorem 4.3, where the clustering error decreases as the number of probes grows, and the cluster-model error under a constant learning rate follows the scale of $\eta\sigma^2/\mu$. Panel (h) supports Theorem 4.4, where the cluster-model error stabilizes at an error floor under a constant learning rate and continues to decrease toward zero under a diminishing learning rate. These results empirically support the finite-probe recovery, constant-learning-rate convergence, and diminishing-learning-rate convergence in our theorems.

Answer to Question 2. Comparisons of `PCR-CFL` with baseline methods under the default configuration are shown in Figure 2, where `PCR-CFL` obtains higher average results for Score, P20, and ARI than the baselines. On Synthetic, Rotated MNIST, and Rotated CIFAR datasets, where the cluster structure is clear, `PCR-CFL` obtains both strong model performance and nearly exact clustering performance. On the datasets with Dirichlet partitioning, the separable cluster structure may hold less clearly, but `PCR-CFL` still shows higher ARI, Score, and P20 than the baselines. FEMNIST has no prescribed ground-truth cluster structure, yet `PCR-CFL` remains competitive in both Score and P20 without relying on the clustering structure. We also evaluate the effectiveness of `PCR-CFL` on the large-scale datasets EMNIST (Cohen et al., 2017) and FedC4 (Charles et al., 2023), where the results are shown in Figures 19 and 20 of Appendix B.6. These results show that `PCR-CFL` improves both clustering and model performance regardless of the cluster structures.

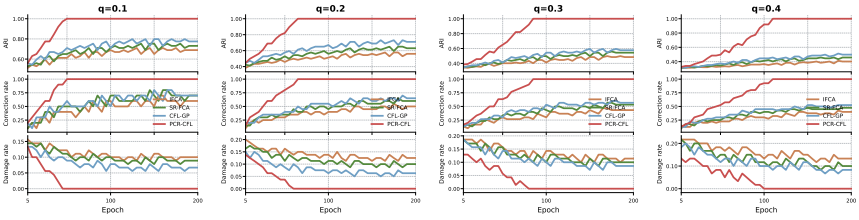
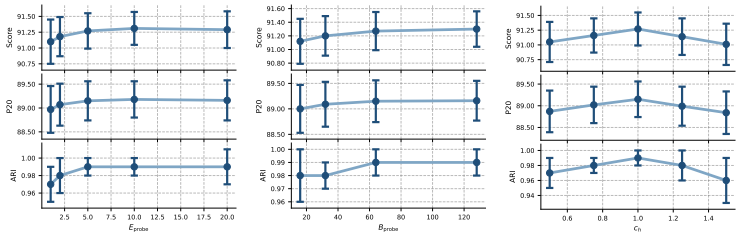


Figure 3: Recovery from incorrect initial client assignments on Synthetic at $\eta/\eta_0 = 1/2$.

Answer to Question 3. On Synthetic and Rotated MNIST, we compare the CFL methods over various η/η_0 (η_0 is the default learning rate) and initial misassignment fractions q. We report ARI,

the *correction rate* (the fraction of initially misassigned clients returned to the ground-truth clusters), and the *damage rate* (the fraction of initially correct clients that become misassigned). Figure 3 shows the Synthetic results at $\eta/\eta_0 = 1/2$. The detailed experimental settings are deferred to Appendix B.3 and results are reported in Figure 6 and 7. The results show that, across the learning rates and initial misassignment fractions, our PCR−CFL achieves a high ARI and a high correction rate while driving the damage rate toward zero. In contrast, the other CFL methods retain part of the initial assignment error and move some initially correct clients to incorrect clusters. Therefore, even with a small learning rate, PCR−CFL can effectively resolve the learning-rate dilemma.

Answer to Question 4. We compare PCR−CFL with the baseline methods across various numbers of clients, batch sizes, numbers of local updates, and clustering intervals. The experimental settings are deferred to Appendix B.4 and the results are reported in Figure 8, 9, 10, 11, 12, 13, 14, and 15. On the datasets with clear cluster structures, our PCR−CFL maintains a stable ARI, and its Score and P20 remain stable. In contrast, the baseline methods exhibit sensitivity to these factors. Moreover, when varying the number of local training updates, PCR−CFL with a smaller E_{train} achieves higher ARI and Score than the other CFL methods with a larger E_{train}. Specifically, PCR−CFL uses only $E_{\text{train}} + E_{\text{probe}}/H = 6$ local gradient updates per client per round on average, yet achieves higher mean Score and ARI than the CFL baselines with $E_{\text{train}} = 10$. The results support the conclusion that client responses to the common probe models provide a more reliable signal for cluster recovery. PCR−CFL also maintains superior model performance on the datasets with weaker cluster structures or no prescribed ground-truth cluster structure. These results further demonstrate the effectiveness and robustness of PCR−CFL under different training configurations.



(a) Probe updates (b) Probe minibatch size (c) Threshold multiplier

Figure 4: Sensitivity of PCR-CFL to probe hyperparameters on Rotated MNIST.

Answer to Question 5. We evaluate PCR−CFL across affinity thresholds, probe minibatch sizes, and numbers of probe updates, where the threshold ablation applies a scaling factor c_h to h_r. The experimental settings are deferred to Appendix B.5 and the results are reported in Figure 16, 17, and 18. The Rotated MNIST results are shown in Figure 4. Results show that PCR−CFL is robust to the number of local probe updates and the probe minibatch size. On datasets with clear and weak cluster structures, the best model and clustering results are obtained near $c_h = 1$, since this choice places the threshold at the midpoint between the empirical within-cluster and cross-cluster distance groups, making it more likely to lie inside the interval required by the theorems. On FEMNIST, which has no prescribed ground-truth cluster structure, changing c_h has little effect on Score or P20. We also ablate the distance accumulation scheme, the choice of the probe model, and the probe learning-rate schedule, where the results are shown in Table 1 in Appendix B.5. Results show that averaging the cluster models for the probe or learning rate schedule has little effect, whereas replacing accumulated distances with distances from a single probe lowers Score and P20, which shows the effectiveness of the distance accumulation scheme in PCR−CFL.

6 Conclusion

In this paper, we propose PCR−CFL to break the learning-rate dilemma in CFL. PCR−CFL calculates the accumulated distances between the responses of the clients to a sequence of probe models, and uses spectral clustering to estimate the number of clusters and obtain the client partition. This process allows a misassigned client to leave the wrong cluster since its responses to the same sequence of probe models are close to those of clients in its ground-truth cluster. We theoretically prove that PCR−CFL recovers the ground-truth cluster structure with high probability after a finite number of probes, and further establish model convergence under both constant and diminishing learning rates. Extensive experiments show the effectiveness and robustness of PCR−CFL.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the writing, assisting in the writing of proofs, checking the correctness of mathematical proofs, assisting with translation, sourcing and searching for information, and identifying relevant literature. We have not used generative AI tools for generating synthetic datasets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, or supporting qualitative and thematic data analysis or interpreting results, and cleaning and reformatting datasets is not applicable to this work. We have reviewed all AI-assisted work. Specifically, all LLM-assisted mathematical proofs were independently checked by multiple authors, and all references identified with the aid of AI tools were verified for correctness and relevance by the authors. The AI-polished portions of the manuscript were further revised and improved by several authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *Proceedings of the 9th International Conference on Learning Representations (ICLR'21)*, 2021a.

Durmus Alp Emre Acar, Yue Zhao, Ruizhao Zhu, Ramon Matas Navarro, Matthew Mattina, Paul N. Whatmough, and Venkatesh Saligrama. Debiasing model updates for improving personalized federated training. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21)*, pp. 21–31, 2021b.

Dmitrii Avdiukhin and Shiva Prasad Kasiviswanathan. Federated learning under arbitrary communication patterns. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21)*, pp. 425–435, 2021.

Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. In *Proceedings of the 15th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX'12)*, pp. 37–49, 2012. doi: 10.1007/978-3-642-32512-0_4.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Theodora S. Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch. Paschalidis, and Wei Shi. Federated learning of predictive models from federated electronic health records. *International Journal of Medical Informatics*, 112:59–67, 2018.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A benchmark for federated settings. *arXiv:1812.01097*, 2018.

José A. Carrillo, Nicolás García Trillos, Sixu Li, and Yuhua Zhu. FedCBO: Reaching group consensus in clustered federated learning through consensus-based optimization. *Journal of Machine Learning Research*, 25:214:1–214:51, 2024.

Zachary Charles, Zachary Garrett, Zhouyuan Huo, Sergei Shmulyian, and Virginia Smith. On large-cohort training for federated learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS'21)*, pp. 20461–20475, 2021.

Zachary Charles, Nicole Mitchell, Krishna Pillutla, Michael Reneer, and Zachary Garrett. Towards federated foundation models: Scalable dataset pipelines for group-structured learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS'23)*, 2023.

Cheng Chen, Ziyi Chen, Yi Zhou, and Bhavya Kailkhura. Fedcluster: Boosting the convergence of federated learning via cluster-cycling. In *Proceedings of the 2020 IEEE International Conference on Big Data (BigData'20)*, pp. 5017–5026. IEEE, 2020a.

Wei Chen, Kartikeya Bhardwaj, and Radu Marculescu. Fedmax: Mitigating activation divergence for accurate and communication-efficient federated learning. In *Proceedings of 2020 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD'20)*, pp. 348–363, 2020b.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards understanding biased client selection in federated learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS'22)*, pp. 10351–10375, 2022.

Yae Jee Cho, Pranay Sharma, Gauri Joshi, Zheng Xu, Satyen Kale, and Tong Zhang. On the convergence of federated averaging with cyclic client participation. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, pp. 5677–5721, 2023.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: extending MNIST to handwritten letters. In *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN'17)*, pp. 2921–2926, 2017.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv:2003.13461*, 2020a.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. In *Advances in Neural Information Processing Systems 33 (NeurIPS'20)*, pp. 15111–15122, 2020b.

Don Kurian Dennis, Tian Li, and Virginia Smith. Heterogeneity for the win: One-shot federated clustering. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21)*, pp. 2611–2620, 2021.

Nikita Dhawan, Nicole Mitchell, Zachary Charles, Zachary Garrett, and Gintare Karolina Dziugaite. Leveraging function space aggregation for federated learning at scale. *Transactions on Machine Learning Research*, 2024, 2024.

Shu Ding and Wei Wang. Collaborative learning by detecting collaboration partners. In *Advances in Neural Information Processing Systems 35 (NeurIPS'22)*, pp. 15629–15641, 2022.

Canh T. Dinh, Nguyen Hoang Tran, and Tuan Dung Nguyen. Personalized federated learning with Moreau envelopes. In *Advances in Neural Information Processing Systems 33 (NeurIPS'20)*, pp. 21394–21405, 2020.

Mathieu Even, Laurent Massoulié, and Kevin Scaman. On sample optimality in personalized collaborative and federated learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS'22)*, pp. 212–225, 2022.

Yann Fraboni, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. Clustered sampling: Low-variance and improved representativity for clients selection in federated learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21)*, pp. 3407–3416, 2021.

Yann Fraboni, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. A general theory for federated optimization with asynchronous and heterogeneous clients updates. *Journal of Machine Learning Research*, 24:110:1–110:43, 2023.

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

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS'20)*, pp. 19586–19597, 2020.

Xinran Gu, Kaixuan Huang, Jingzhao Zhang, and Longbo Huang. Fast federated learning in the presence of arbitrary device unavailability. In *Advances in Neural Information Processing Systems 34 (NeurIPS'21)*, pp. 12052–12064, 2021.

Yongxin Guo, Xiaoying Tang, and Tao Lin. FedRC: Tackling diverse distribution shifts challenge in federated learning by robust clustering. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, 2024.

Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv:1910.14425*, 2019.

Xinmeng Huang, Donghwan Lee, Edgar Dobriban, and Hamed Hassani. Collaborative learning of discrete distributions under heterogeneity and communication constraints. In *Advances in Neural Information Processing Systems 35 (NeurIPS'22)*, pp. 31915–31928, 2022.

Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized federated learning: An attentive collaboration approach. *arXiv:2007.03797*, 2020.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Xiaowen Jiang, Anton Rodomanov, and Sebastian U. Stich. Federated optimization with doubly regularized drift correction. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, 2024.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26 (NIPS'13)*, pp. 315–323, 2013.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, pp. 5132–5143, 2020.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS'20)*, pp. 4519–4529, 2020.

Heasung Kim, Hyeji Kim, and Gustavo de Veciana. Clustered federated learning via gradient-based partitioning. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, pp. 24137–24193, 2024.

Nikita Yurevich Kotelevskii, Maxime Vono, Alain Durmus, and Eric Moulines. Fedpop: A bayesian approach for personalised federated learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS'22)*, pp. 8687–8701, 2022.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS'10)*, pp. 299–308, 2010. doi: 10.1109/FOCS.2010.35.

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

Taehwan Lee and Sung Whan Yoon. Rethinking the flat minima searching in federated learning. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, 2024.

Bo Li, Mikkel N. Schmidt, Tommy S. Alstrøm, and Sebastian U. Stich. On the effectiveness of partial variance reduction in federated learning with heterogeneous data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'2023)*, pp. 3964–3973, 2023.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems (MLSys'20)*, pp. 429–450, 2020a.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *Proceedings of the 8th International Conference on Learning Representations (ICLR'20)*, 2020b.

Guodong Long, Ming Xie, Tao Shen, Tianyi Zhou, Xianzhi Wang, and Jing Jiang. Multi-center federated learning: clients clustering for better personalization. *World Wide Web*, 26(1):481–500, 2023.

Jie Ma, Guodong Long, Tianyi Zhou, Jing Jiang, and Chengqi Zhang. On the convergence of clustered federated learning. *arXiv:2202.06187*, 2022.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv:2002.10619*, 2020.

Othmane Marfoq, Giovanni Neglia, Richard Vidal, and Laetitia Kameni. Personalized federated learning through local memorization. In *Proceedings of the 39th International Conference on Machine Learning (ICML'22)*, pp. 15070–15092, 2022.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS'17)*, pp. 1273–1282, 2017.

Aritra Mitra, Rayana H. Jaafar, George J. Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. In *Advances in Neural Information Processing Systems 34 (NeurIPS'21)*, pp. 14606–14619, 2021.

Solmaz Niknam, Harpreet S. Dhillon, and Jeffrey H. Reed. Federated learning for wireless communications: Motivation, opportunities, and challenges. *IEEE Communications Magazine*, 58(6): 46–51, 2020.

Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *Journal of the ACM*, 59(6), 2013.

Younghyun Park, Dong-Jun Han, Do-Yeon Kim, Jun Seo, and Jaekyun Moon. Few-round learning for federated learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS'21)*, pp. 28612–28622, 2021.

Kumar Kshitij Patel, Margalit Glasgow, Ali Zindari, Lingxiao Wang, Sebastian U. Stich, Ziheng Cheng, Nirmit Joshi, and Nathan Srebro. The limits and potentials of local SGD for distributed heterogeneous learning with intermittent communication. In *The 37th Annual Conference on Learning Theory (COLT'24)*, pp. 4115–4157, 2024.

Akbar Rafiey. Decomposable submodular maximization in federated setting. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, 2024.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning (ICML'12)*, 2012.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *Proceedings of the 9th International Conference on Learning Representations (ICLR'21)*, 2021.

Yichen Ruan and Carlee Joe-Wong. Fedsoft: Soft clustered federated learning with proximal local updating. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI'22)*, pp. 241–251, 2022.

Yichen Ruan, Xiaoxi Zhang, Shu-Che Liang, and Carlee Joe-Wong. Towards flexible device participation in federated learning. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS'21)*, pp. 3403–3411, 2021.

Sumudu Samarakoon, Mehdi Bennis, Walid Saad, and Mérouane Debbah. Federated learning for ultra-reliable low-latency v2v communications. In *Proceedings of the IEEE Global Communications Conference (GLOBECOM'18)*, pp. 1–7, 2018.

Karan Singhal, Hakim Sidahmed, Zachary Garrett, Shanshan Wu, John Rush, and Sushant Prakash. Federated reconstruction: Partially local federated learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS'21)*, pp. 11220–11232, 2021.

Douglas Steinley. Properties of the hubert-arabie adjusted rand index. *Psychological Methods*, 9: 386–396, 09 2004.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *Proceedings of the 7th International Conference on Learning Representations (ICLR'19)*, 2019.

Lili Su, Jiaming Xu, and Pengkun Yang. Global convergence of federated learning for mixed regression. In *Advances in Neural Information Processing Systems 35 (NeurIPS'22)*, pp. 29889–29902, 2022.

Tiffany Tuor, Shiqiang Wang, Bong Jun Ko, Changchang Liu, and Kin K. Leung. Overcoming noisy and irrelevant data in federated learning. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR'20)*, pp. 5020–5027, 2020.

Saeed Vahidian, Mahdi Morafah, Weijia Wang, Vyacheslav Kungurtsev, Chen Chen, Mubarak Shah, and Bill Lin. Efficient distribution similarity identification in clustered federated learning via principal angles between client data subspaces. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI'23)*, pp. 10043–10052, 2023.

Harsh Vardhan, Avishek Ghosh, and Arya Mazumdar. An improved federated clustering algorithm with model-based clustering. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris S. Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *Proceedings of the 8th International Conference on Learning Representations (ICLR'20)*, 2020a.

Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified framework for the design and analysis of local-update SGD algorithms. *Journal of Machine Learning Research*, 22:213:1–213:50, 2021.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS'20)*, pp. 7611–7623, 2020b.

Jianyu Wang, Rudrajit Das, Gauri Joshi, Satyen Kale, Zheng Xu, and Tong Zhang. On the unreasonable effectiveness of federated averaging with heterogeneous data. *Transactions on Machine Learning Research*, 2024, 2024.

Shiqiang Wang and Mingyue Ji. A unified analysis of federated learning with arbitrary client participation. In *Advances in Neural Information Processing Systems 35 (NeurIPS'22)*, pp. 19124–19137, 2022.

Mariel Werner, Lie He, Michael Jordan, Martin Jaggi, and Sai Praneeth Karimireddy. Provably personalized and robust federated learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Haibo Yang, Zhuqing Liu, Xin Zhang, and Jia Liu. SAGDA: achieving $\mathcal{O}(\epsilon^{-2})$ communication complexity in federated min-max learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS'22)*, pp. 7142–7154, 2022a.

Haibo Yang, Xin Zhang, Prashant Khanduri, and Jia Liu. Anarchic federated learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML'22)*, pp. 25331–25363, 2022b.

Haibo Yang, Peiwen Qiu, Prashant Khanduri, Minghong Fang, and Jia Liu. Understanding server-assisted federated learning in the presence of incomplete client participation. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, 2024.

Liu Yang, Ben Tan, Vincent W. Zheng, Kai Chen, and Qiang Yang. Federated recommendation systems. In *Federated Learning - Privacy and Incentive*, volume 12500 of *Lecture Notes in Computer Science*, pp. 225–239. 2020.

Rui Ye, Zhenyang Ni, Fangzhao Wu, Siheng Chen, and Yanfeng Wang. Personalized federated learning with inferred collaboration graphs. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, pp. 39801–39817, 2023.

Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *arXiv:2002.04758*, 2020.

Xue Yu, Ziyi Liu, Wu Wang, and Yifan Sun. Clustered federated learning based on nonconvex pairwise fusion. *Information Sciences*, 678:120956, 2024.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan H. Greenewald, Trong Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML'19)*, pp. 7252–7261, 2019.

Hao Zhang, Chenglin Li, Wenrui Dai, Junni Zou, and Hongkai Xiong. Fedcr: Personalized federated learning based on across-client common representation with conditional mutual information regularization. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, pp. 41314–41330, 2023.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv:1806.00582*, 2018.

Wujun Zhou, Shu Ding, Zelin Li, and Wei Wang. Enhancing the performance of global model by improving the adaptability of local models in federated learning. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI'25)*, pp. 7173–7181, 2025.

Ligeng Zhu, Hongzhou Lin, Yao Lu, Yujun Lin, and Song Han. Delayed gradient averaging: Tolerate the communication latency for federated learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS'21)*, pp. 29995–30007, 2021.

A Additional Related Works

Federated Learning. `FedAvg` performs model averaging after several rounds of local stochastic gradient descent to orchestrate a global model (McMahan et al., 2017), which has achieved remarkable success in various real-world applications (Brisimi et al., 2018; Samarakoon et al., 2018; Yang et al., 2020; Niknam et al., 2020; Zhou et al., 2025). However, further studies have revealed that in the presence of data heterogeneity, `FedAvg` struggles with performance degradation since the learned global model may not generalize well on each client individually (Zhao et al., 2018; Haddadpour & Mahdavi, 2019; Li et al., 2020a; Wang et al., 2024; Patel et al., 2024).

To mitigate the impact of data heterogeneity, some studies incorporate control variates to correct client drift during local updates, which can take advantage of the similarity between clients to accelerate the convergence of federated learning and improve the performance of the learned model (Karimireddy et al., 2020; Wang et al., 2020b; Mitra et al., 2021; Yang et al., 2022a; Li et al., 2023). Another line of work adds regularization terms to the optimization objectives to regularize the difference between the global model and local models to expedite convergence and enhance the model performance (Li et al., 2020a; Dinh et al., 2020; Acar et al., 2021a; Zhang et al., 2023; Jiang et al., 2024). Mechanisms such as delayed gradient averaging, gradient accumulation, and asynchronous schemes can also improve communication efficiency by delaying the model aggregation step and handling the clients' historical gradients (Deng et al., 2020b; Avdiukhin & Kasiviswanathan, 2021; Zhu et al., 2021; Fraboni et al., 2023). Some works focus on studying the impact of partial client participation on the convergence of federated learning and propose various client sampling strategies to boost the performance of the learned model (Tuor et al., 2020; Ruan et al., 2021; Gu et al., 2021; Fraboni et al., 2021; Charles et al., 2021; Cho et al., 2022; Kotelevskii et al., 2022; Wang & Ji, 2022; Yang et al., 2022b; Cho et al., 2023; Rafiey, 2024; Yang et al., 2024). Some works utilize the model interpolation scheme and local fine-tuning techniques to mitigate the bias introduced by data heterogeneity (Yu et al., 2020; Deng et al., 2020a; Mansour et al., 2020; Acar et al., 2021b; Singhal et al., 2021; Park et al., 2021; Huang et al., 2022; Marfoq et al., 2022; Lee & Yoon, 2024). Nevertheless, these methods are based on a single global model and thus still encounter the performance degradation challenge when the data heterogeneity is severe among the clients. Besides, Deng et al. (2020a) have also illustrated that solely optimizing for the global model's performance will lead to poor generalization on local clients.

Clustered Federated Learning. To tackle the challenge of data heterogeneity, another important approach is clustered federated learning (Ghosh et al., 2020; Mansour et al., 2020; Chen et al., 2020b; Huang et al., 2020; Chen et al., 2020a; Dennis et al., 2021; Su et al., 2022; Ruan & Joe-Wong, 2022; Ma et al., 2022; Even et al., 2022; Ding & Wang, 2022; Vahidian et al., 2023; Vardhan et al., 2024; Long et al., 2023; Werner et al., 2023; Kim et al., 2024; Guo et al., 2024; Yu et al., 2024; Carrillo et al., 2024). These methods generally cluster the clients into groups according to client similarity and leverage each group for federated learning to enhance the performance.

Existing heuristic CFL methods differ in their notion of client similarity and their assumptions on the latent groups. Among these works, Su et al. (2022) contextualized federated learning in mixed regression, where the data on each client are generated from one of the regression models; Ruan & Joe-Wong (2022) assumed that the data on each client follow a mixture of multiple distributions, and proposed a novel method to exploit the similarity between source distributions to learn the group models; Long et al. (2023) proposed a general formulation of clustered FL as a bilevel optimization problem; Vahidian et al. (2023) utilized the principal angles between the client data subspaces to derive well-performing models; Guo et al. (2024) proposed FedRC based on soft clustering to maintain robust global model performance in the presence of multiple types of distribution shifts across clients. However, these works heuristically study different clustering models or optimization mechanisms, and do not theoretically establish cluster-recovery guarantees for CFL.

Representative CFL methods with theoretical analyses primarily differ in how they assign clients to clusters. `IFCA` (Ghosh et al., 2020) lets each client select the cluster model with the lowest loss value and uses that client to update the selected model. However, its recovery guarantee requires the ground-truth number of clusters and sufficiently good initial models, which are often inaccessible in practice. `SR-FCA` (Vardhan et al., 2024) starts from singleton clusters and repeatedly merges and refines them using model distances. It does not require the ground-truth number of clusters, but its refinement guarantee requires the fraction of misassigned members in each initial cluster to be less

than one half, a condition that relies on sufficiently accurate local model estimates before merging. This requirement is restrictive when each client does not have enough data to learn an accurate local model; therefore, it places a hidden requirement similar to good initialization. `CFL-GP` (Kim et al., 2024) and the methods in (Werner et al., 2023) group clients using gradients or momentum information and require the ground-truth number of clusters. Their analyses assume that the gradients or momentum from different clusters remain separated during training. However, with a poor initialization and a small learning rate, a model affected by a wrong group may remain near a point where gradients are hard to separate. We provide concrete examples in Appendix D.1 to show that, under poor initialization, limited local computation can cause `SR-FCA`'s initial model-based clustering to merge distinct tasks and violate its refinement requirements, while a small training learning rate can sustain weak gradient separation along an incorrectly mixed trajectory, weakening the clustering signal used by `CFL-GP`.

Federated learning typically requires a small or diminishing learning rate for stable model optimization (Stich, 2019; Haddadpour & Mahdavi, 2019; Khaled et al., 2020; Li et al., 2020b; Karimireddy et al., 2020; Wang et al., 2020b; Reddi et al., 2021; Wang & Joshi, 2021). In particular, Li et al. (2020b) have shown that a diminishing client learning rate is necessary for the convergence of federated learning on non-IID data, Karimireddy et al. (2020) further explained that `FedAvg` requires smaller local learning rates to control client drift under data heterogeneity, and Khaled et al. (2020) have shown that the learning rate should become even smaller as the number of local updates grows. Therefore, a small learning rate is necessary for stable model optimization in federated learning, and increasing the number of local updates is impractical since federated learning uses modest local computation (McMahan et al., 2017; Karimireddy et al., 2020; Li et al., 2020a). A basic but challenging problem in CFL is therefore: *Can a client recover after it has been assigned to a wrong cluster while the training learning rate remains small?* Werner et al. (2023) revealed that a misassigned client under a small learning rate may still be stuck even after multiple local update steps. This is the *learning-rate dilemma* in CFL. Existing CFL methods cannot guarantee recovery from this state when the number of clusters, initial cluster models, and initial client partition are unknown.

B Experimental Details

B.1 Detailed Experimental Setup

Datasets and Models. We use the following eight datasets that contain different cluster structures. The synthetic strongly convex problem has four clusters and is used for verification of the theoretical results. Each client has the objective $f_i(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_i^\star)^\top \boldsymbol{H}(\boldsymbol{w} - \boldsymbol{w}_i^\star)$, where the cluster centers form a regular simplex and client optima are perturbed within a ball around their center. Rotated MNIST and Rotated CIFAR follow the standard clustered federated learning construction as in (Ghosh et al., 2020; Vardhan et al., 2024), where clients belong to one of four groups formed by rotations of 0, 90, 180, and 270 degrees. Rotated MNIST and Rotated CIFAR have a clear cluster structure. We additionally partition MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky et al., 2009) using the Dirichlet distribution with concentration parameters $\beta \in \{0.3, 0.5\}$ to obtain heterogeneous partitions, following common federated learning configurations as in (Yurochkin et al., 2019; Wang et al., 2020a). The Dirichlet partitioning produces 10 clusters by default. For N clients, we sort the samples in each cluster by label and split them into $N/10$ disjoint shards, where each client receives one shard. Here, samples with the same label may be split across multiple shards. These four datasets have weaker separable cluster structures than the rotated datasets. We use the natural client partition of FEMNIST from LEAF (Caldas et al., 2018). FEMNIST has no prescribed ground-truth cluster structure. Thus, the FEMNIST dataset measures whether our PCR-CFL remains useful when a prescribed cluster structure is absent. We use the same CNN model as in (McMahan et al., 2017) for MNIST variants, the same ResNet-9 model as in (Vardhan et al., 2024) for Rotated CIFAR, the CNN model similar to that in (Ghosh et al., 2020) for CIFAR-10 with the Dirichlet partitioning, and the same CNN model as in (Caldas et al., 2018) for the FEMNIST dataset.

Baselines. We compare with `FedAvg` (McMahan et al., 2017), `IFCA` (Ghosh et al., 2020), `CFL-GP` (Kim et al., 2024), and `SR-FCA` (Vardhan et al., 2024). These baselines represent three different clustering signals: `IFCA` assigns clients through model loss, `CFL-GP` forms groups from gradient information, and `SR-FCA` uses parameter distances. For datasets with group structures, `IFCA` and `CFL-GP` receive the ground-truth number of clusters since they require a fixed number of clusters, while `SR-FCA` and our `PCR-CFL` estimate the number of clusters. FEMNIST has no prescribed ground-truth cluster structure; therefore, `IFCA` and `CFL-GP` use the same fixed number of clusters, while `SR-FCA` and `PCR-CFL` estimate the number of clusters.

Experimental Setup. We use $N = 100$ clients, training minibatch size $B_{\text{train}} = 64$, local training updates $E_{\text{train}} = 5$ per round, and clustering/probe interval $H = 5$ as the default configuration. For `IFCA`, `CFL-GP`, and `SR-FCA`, H is the number of communication rounds between clustering operations, while for `PCR-CFL`, it is the number of rounds between model probes. The training budget is 200 epochs for each dataset. The default probe configuration is $E_{\text{probe}} = 5$ local probe updates and probe minibatch size $B_{\text{probe}} = 64$. With full client participation, each probe adds E_{probe} local gradient updates per client every H rounds. Relative to E_{train} training updates per round, the additional local-update count is $E_{\text{probe}}/(HE_{\text{train}})$: it is 20% for the default configuration. For Rotated MNIST and MNIST with Dirichlet partitioning, we use SGD with learning rate 0.1. For Rotated CIFAR and CIFAR-10 with Dirichlet partitioning, we use SGD with learning rate 0.5, momentum 0.9, weight decay 5×10^{-4}, and a cosine annealing learning-rate schedule. For the Synthetic and FEMNIST datasets, we use SGD with learning rate 0.04. `PCR-CFL` selects the affinity threshold h adaptively: after probe r, we apply two-means clustering to the accumulated pairwise distances $\{d_{i,j}(r) : i < j\}$. Let $a_r < b_r$ be the means of the two resulting groups, which represent the smaller within-cluster distances and the larger cross-cluster distances, respectively. We set $h_r^2 = (a_r + b_r)/2$ and construct the affinity matrix using h_r. This procedure adapts to the distance scale of each dataset and avoids dataset-specific threshold tuning. We provide Theorem D.11 in Appendix D.7 to prove that this data-based threshold lies in the recovery interval with high probability when the within-cluster and cross-cluster probe distances form two sufficiently separated groups. We record the following metrics on these datasets: Score, P20, and ARI. For classification datasets, Score is the average client test accuracy, while on the synthetic problem, Score is an R^2-style score, *i.e.*, a parameter-recovery coefficient averaged across clients $1 - N\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i^\star\|^2 / \sum_j \|\boldsymbol{w}_j^\star - \overline{\boldsymbol{w}}^\star\|^2$. P20 is the mean score of the worst-performing 20% of clients, which measures lower-tail performance. We use the Adjusted Rand index (ARI) (Hubert & Arabie, 1985; Steinley, 2004) to evaluate

the quality of the clustering process. FedAvg has no ARI since it does not perform clustering. The Score and P20 of FedAvg are omitted on the synthetic dataset since the score is close to zero. FEMNIST has no prescribed ground-truth cluster structure, so ARI is omitted for all methods on this dataset. Unless otherwise stated, all reported points are means over five random seeds, and the error bars represent the standard deviation.

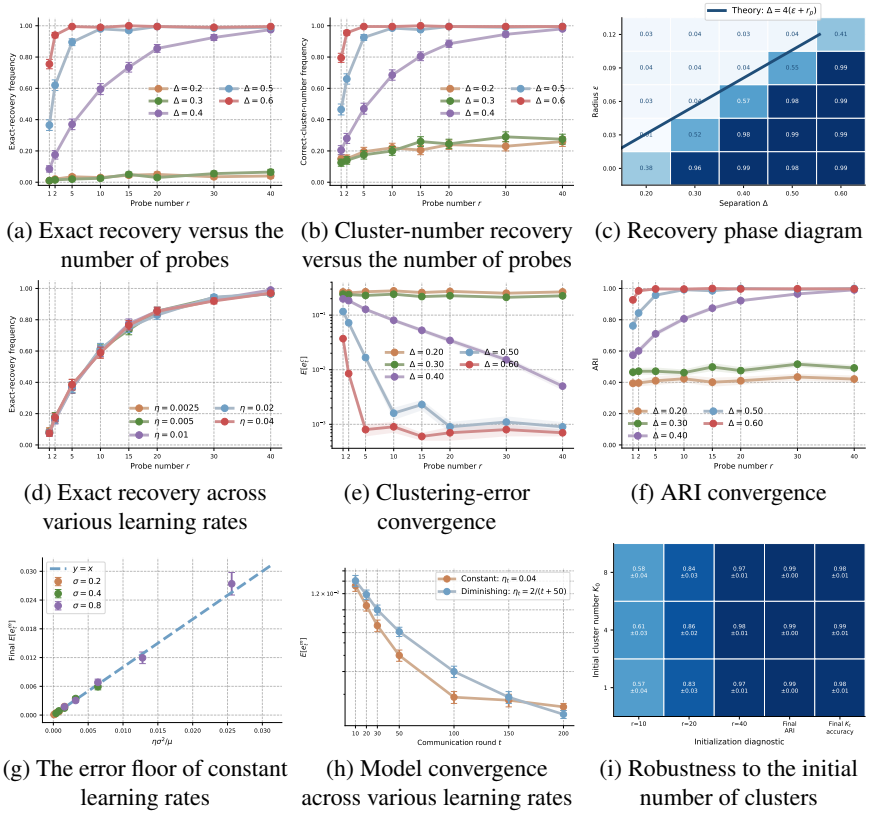
B.2 Theory Verification Results

We run PCR-CFL on the synthetic strongly convex problem to validate the theoretical results. The synthetic problem has $N = 100$ clients, four balanced ground-truth clusters, and dimension $d = 20$, with $\mu = 1$ and $L = 5$. For the default configuration, we use $\epsilon = 0.05$, $\Delta = 0.4$, $\sigma = 0.2$, $\alpha = \mu/L^2 = 0.04$, $H = 5$, $B_{\text{probe}} = 64$, and $E_{\text{probe}} = 200$, where the number of probe updates is used only in this synthetic experiment to ensure the nonvacuous separation condition in Theorem 4.2, 4.3, 4.4. Each run starts from $K_0 \in \{1, 4, 8\}$, a random client partition, and randomly initialized cluster models. For the theory verification, we use the fixed threshold at the midpoint of the squared-distance interval, *i.e.*, $h^2 = ((2\epsilon + 2r_{\text{p}})^2 + (\Delta - 2\epsilon - 2r_{\text{p}})^2)/2$. We first vary $\Delta \in \{0.20, 0.30, 0.40, 0.50, 0.60\}$, $\epsilon \in \{0, 0.03, 0.06, 0.09, 0.12\}$, and the number of completed probes $r \in \{1, 2, 5, 10, 15, 20, 30, 40\}$. We report the exact-recovery frequency $\Pr(\mathcal{E}_{t_r})$, the probability of estimating $K^\star$, ARI, and $\mathbb{E}[e_t^{\text{c}}]$ to validate finite-probe recovery and the separation condition in Theorem 4.2. We then vary the constant learning rate over $\eta \in \{0.0025, 0.005, 0.01, 0.02, 0.04\}$ and the gradient-noise level over $\sigma \in \{0, 0.2, 0.4, 0.8\}$. Recovery curves under different η test the sensitivity of recovery to the training learning rate, while the cluster-model error is compared with the constant-learning-rate upper bound in Theorem 4.3. We then compare the constant learning rate $\eta_t = 0.04$ with the diminishing learning rate $\eta_t = 2/(t + 50)$ from Theorem 4.4. We report $\mathbb{E}[e_t^{\text{m}}]$ at communication rounds $t \in \{10, 20, 30, 50, 100, 150, 200\}$ and the finite-horizon stable recovery time, defined as the first round after which $\mathcal{E}_t$ holds. All the results are shown in Figure 5, where the frequencies in the theory verification use 200 independent synthetic trials since they estimate a probability. Panels (a)–(d) and (i) are consistent with Theorem 4.2, where the exact-recovery and correct-cluster-number frequencies increase with the number of probes and the separation Δ, and the recovery region is consistent with the sufficient condition $\Delta > 4(\epsilon + r_{\text{p}})$. The nearly overlapping curves under different training learning rates support the predicted insensitivity of recovery to the training learning rate. The nearly identical results for $K_0 \in \{1, 4, 8\}$ support finite-probe recovery without knowing the ground-truth number of clusters or requiring an accurate initial client partition. Panels (e)–(g) are consistent with Theorem 4.3, where the clustering error decreases as the number of probes grows, ARI approaches one, and the final cluster-model error under a constant learning rate follows the scale of the upper bound $\eta\sigma^2/\mu$. Panel (h) supports Theorem 4.4, where the cluster-model error stabilizes at a noise-dependent error floor under a constant learning rate and continues to decrease toward zero under a diminishing learning rate. These results empirically support the finite-probe recovery, constant-learning-rate convergence, and diminishing-learning-rate convergence.

B.3 Learning-Rate-Dilemma Results

Let η_0 denote the default training learning rate for each dataset. On Synthetic and Rotated MNIST, we compare IFCA, CFL-GP, SR-FCA, and PCR-CFL over $\eta/\eta_0 \in \{1/16, 1/8, 1/4, 1/2, 1, 2\}$ and initial misassignment fractions $q \in \{0.1, 0.2, 0.3, 0.4\}$ while keeping the other clustering hyperparameters fixed. For each prescribed initial misassignment fraction q, all four methods start from the same corrupted client partition. For SR-FCA, this partition initializes the refinement stage in place of singleton initialization, so the comparison measures recovery from the prescribed assignment errors. We report ARI, the *correction rate*, defined as the fraction of initially misassigned clients returned to their ground-truth clusters, and the *damage rate*, defined as the fraction of initially correct clients that become misassigned. The complete results on Synthetic and Rotated MNIST are reported in Figure 6 and 7. The results show that, across the learning rates and initial misassignment fractions, our PCR-CFL achieves a high ARI and a high correction rate while driving the damage rate toward zero. In contrast, the other CFL methods retain part of the initial assignment error since some misassigned clients remain in incorrect clusters; they also move some initially correct clients to incorrect clusters. A possible reason for the lower recovery accuracy of CFL-GP is that its queried models are updated using the current groups. When the learning rate is small, a model affected by a wrong group changes slowly and may continue to give weak gradient separation. Gradient differences may also change direction over the queried models, so averaging them may further reduce the

gap. The gap increases as the initial misassignment fraction q becomes larger, and the gap also increases as the learning rate becomes smaller. Therefore, even with a small learning rate, PCR−CFL can recover misassigned clients without damaging the correctly assigned clients, showing that it effectively resolves the learning-rate dilemma.



(a) Exact recovery versus the number of probes

(b) Cluster-number recovery versus the number of probes

(c) Recovery phase diagram

(d) Exact recovery across various learning rates

(e) Clustering-error convergence

(f) ARI convergence

(g) The error floor of constant learning rates

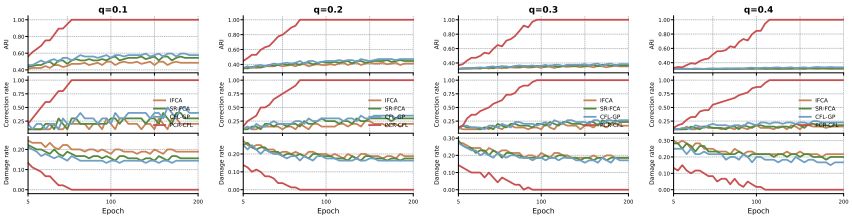
(h) Model convergence across various learning rates

(i) Robustness to the initial number of clusters

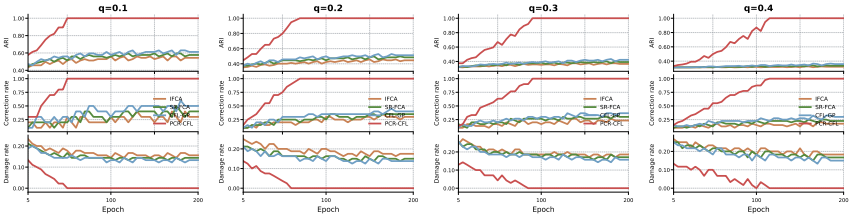
Figure 5: **Empirical verification of finite-probe recovery and cluster-model convergence on the Synthetic dataset.** The nine panels provide empirical evidence for cluster-structure recovery, recovery of the number of clusters, the separation condition, insensitivity to the learning rate, clustering-error and ARI convergence, the constant-learning-rate model-error floor, convergence under constant and diminishing learning rates, and robustness to the initial number of clusters. Curves show means, and error bars or shaded bands show standard deviations across five seeds. For recovery frequencies estimated from 200 trials, error bars show the binomial standard error.

B.4 Ablation Results

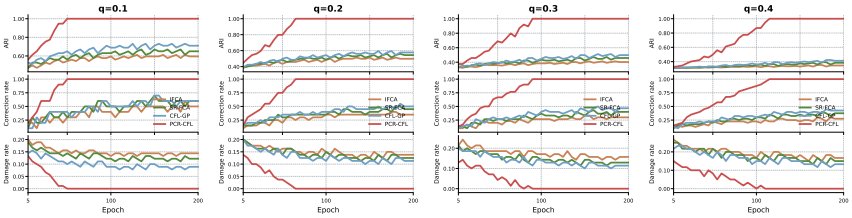
We change one factor at a time while keeping the remaining configuration fixed: the number of clients $N \in \{50, 100, 150, 200\}$, the batch size $B_{\text{train}} \in \{32, 64, 96, 128\}$, and the number of local updates $E_{\text{train}} \in \{1, 5, 10, 20\}$. We also vary the interval between clustering operations or probes; the synthetic dataset uses $H \in \{3, 5, 7, 10\}$ and the other datasets use $H \in \{1, 5, 7, 10\}$. FedAvg is included when varying N, B_{train}, and E_{train}, but is omitted from the H study since it has no clustering operation. The results are reported in Figure 8, 9, 10, 11, 12, 13, 14, and 15. The curves report the mean value over five seeds, and the shaded bands show the standard deviation. The results show that, across different numbers of clients, minibatch sizes, numbers of local updates, and clustering or probe intervals, our PCR−CFL preserves its advantage in both model and clustering performance. On the three datasets with clear cluster structures (Synthetic, Rotated MNIST, and Rotated CIFAR), our PCR−CFL maintains a stable ARI across these four factors, and its Score and P20 remain stable. In contrast, the baseline methods exhibit sensitivity to these four factors. PCR−CFL also maintains superior clustering and model performance on the datasets with weaker cluster structures. FEMNIST has no prescribed ground-truth cluster structure; PCR−CFL still obtains higher average model performance on this dataset. Moreover, when varying the number of local training updates, PCR−CFL



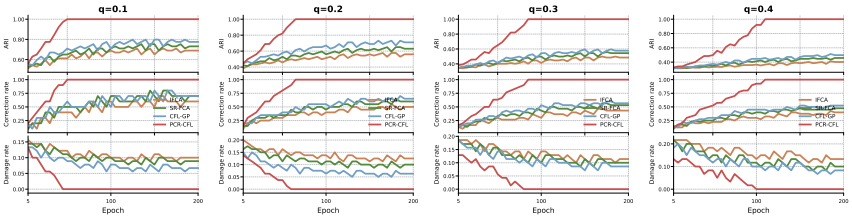
(a) Learning-rate ratio $\eta/\eta_0 = 1/16$



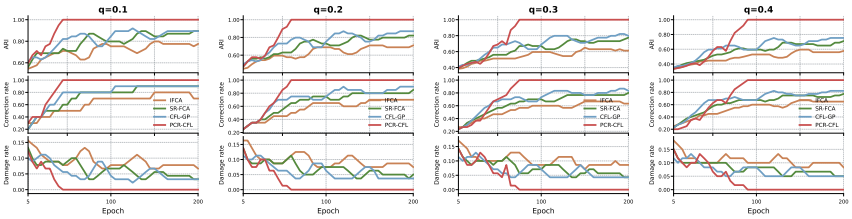
(b) Learning-rate ratio $\eta/\eta_0 = 1/8$



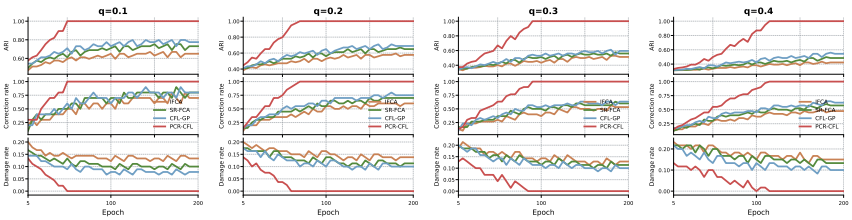
(c) Learning-rate ratio $\eta/\eta_0 = 1/4$



(d) Learning-rate ratio $\eta/\eta_0 = 1/2$

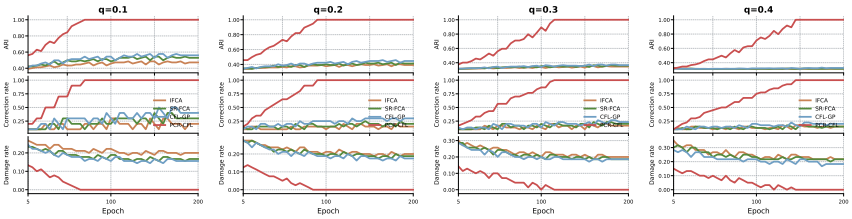


(e) Learning-rate ratio $\eta/\eta_0 = 1$

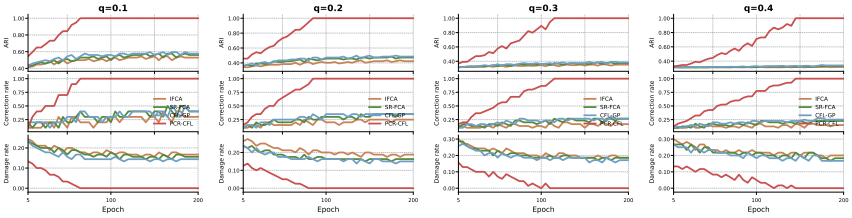


(f) Learning-rate ratio $\eta/\eta_0 = 2$

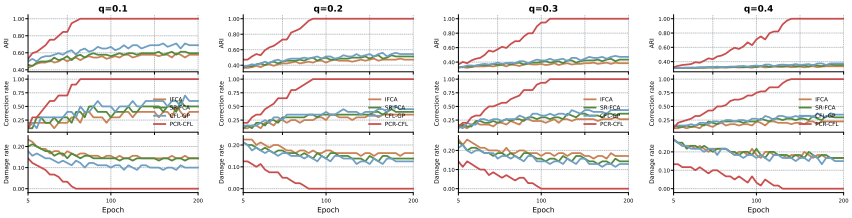
Figure 6: **Recovery from incorrect initial client assignments across learning rates on the Synthetic dataset.** Rows vary η/η_0, and columns vary the initial misassignment fraction q.



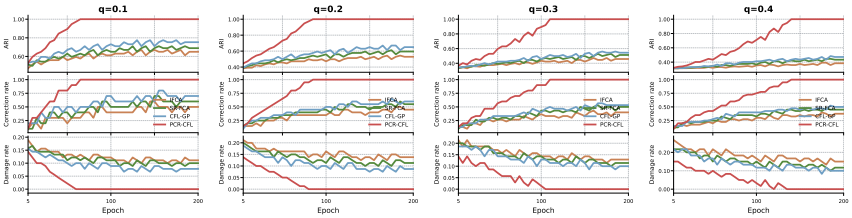
(a) Learning-rate ratio $\eta/\eta_0 = 1/16$



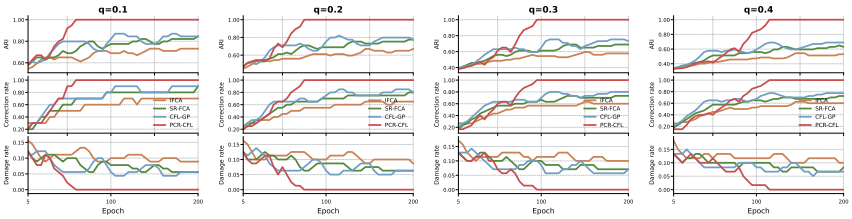
(b) Learning-rate ratio $\eta/\eta_0 = 1/8$



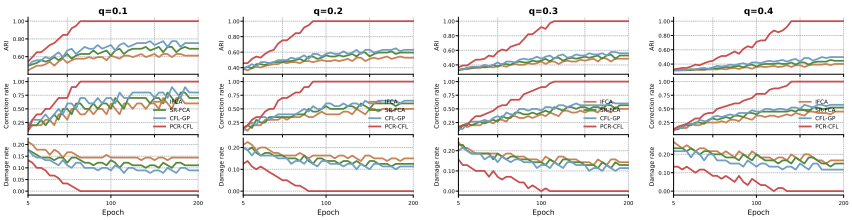
(c) Learning-rate ratio $\eta/\eta_0 = 1/4$



(d) Learning-rate ratio $\eta/\eta_0 = 1/2$



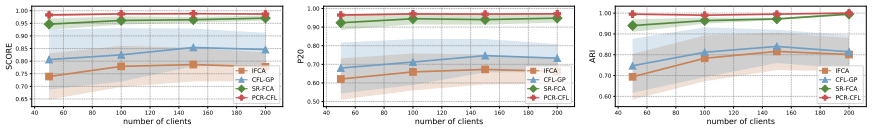
(e) Learning-rate ratio $\eta/\eta_0 = 1$



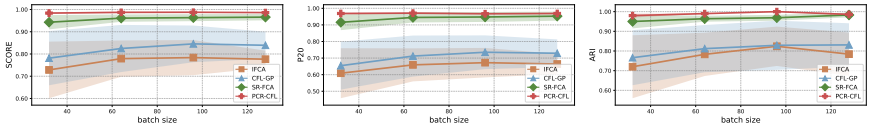
(f) Learning-rate ratio $\eta/\eta_0 = 2$

Figure 7: **Recovery from incorrect initial client assignments across learning rates on the Rotated MNIST dataset.** Rows vary η/η_0, and columns vary the initial misassignment fraction q.

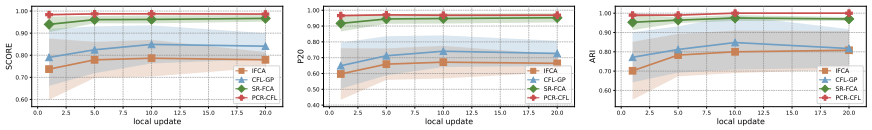
with a smaller E_{train} achieves higher ARI and Score than the other CFL methods with a larger E_{train}. The local-update panels in Figures 8 and 9 make the computation comparison more concrete. With $E_{\text{train}} = 5$, $E_{\text{probe}} = 5$, and $H = 5$, PCR−CFL performs five training updates and one additional probe update per client per communication round on average, which is 20% more than the five-update training setting. Raising a baseline from $E_{\text{train}} = 5$ to $E_{\text{train}} = 10$ doubles its training-update steps, yet its Score and ARI remain below those of PCR−CFL at $E_{\text{train}} = 5$ in these panels. Including the probe updates, PCR−CFL therefore uses 40% fewer local gradient updates per client per round than the ten-update baselines' training updates alone, while retaining higher mean Score and ARI in these panels. This supports effective recovery by PCR−CFL with a smaller per-round local-update budget. The results show that the gain of PCR−CFL does not arise from increasing the number of local updates; instead, the results support the conclusion that client responses to the common probe models provide a more reliable signal for cluster recovery. These results further demonstrate the effectiveness and robustness of PCR−CFL under different configurations.



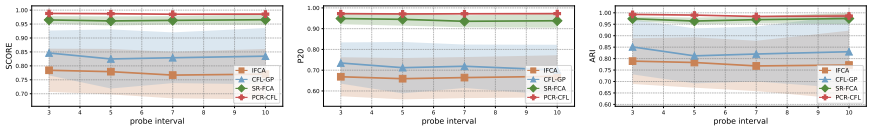
(a) Sensitivity to the number of clients



(b) Sensitivity to the batch size

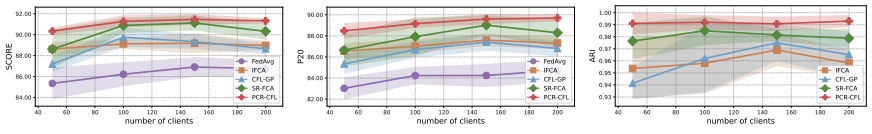


(c) Sensitivity to the number of local updates

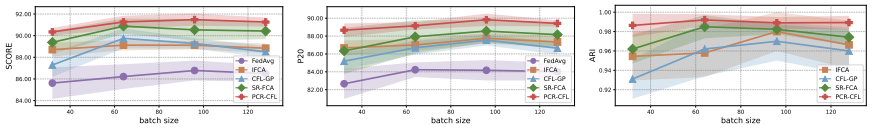


(d) Sensitivity to the probe interval

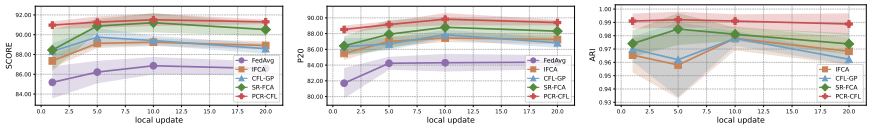
Figure 8: **Sensitivity to different hyperparameters on the Synthetic dataset.** Rows vary the number of clients, the minibatch size, the number of local training updates, and the interval between clustering operations or probes. Columns report Score, P20, and ARI.



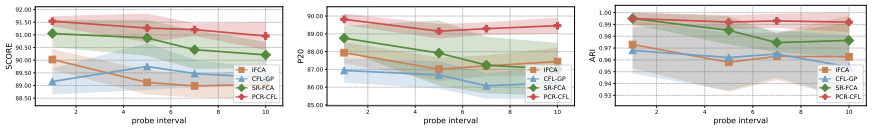
(a) Sensitivity to the number of clients



(b) Sensitivity to the batch size

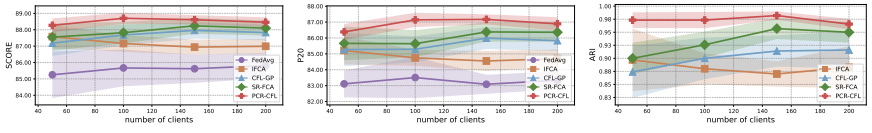


(c) Sensitivity to the number of local updates

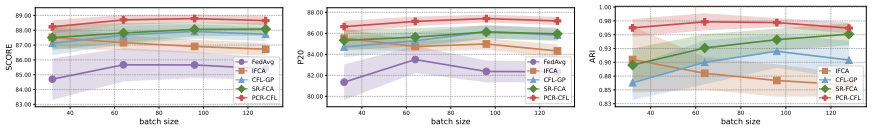


(d) Sensitivity to the probe interval

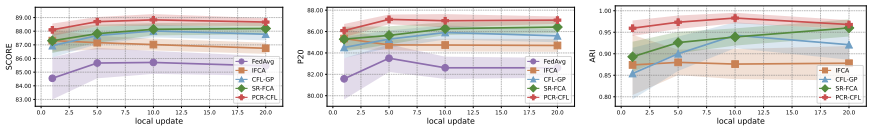
Figure 9: **Sensitivity to different hyperparameters on the Rotated MNIST dataset.** Rows vary the number of clients, the minibatch size, the number of local training updates, and the interval between clustering operations or probes. Columns report Score, P20, and ARI.



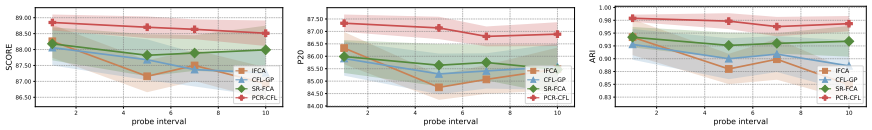
(a) Sensitivity to the number of clients



(b) Sensitivity to the batch size

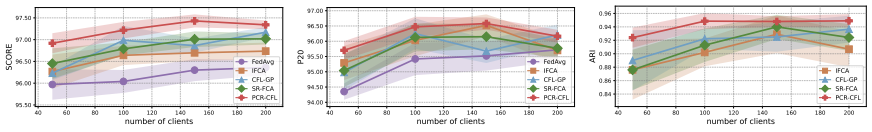


(c) Sensitivity to the number of local updates

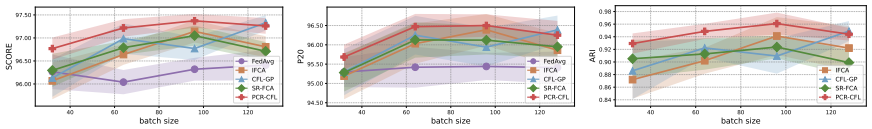


(d) Sensitivity to the probe interval

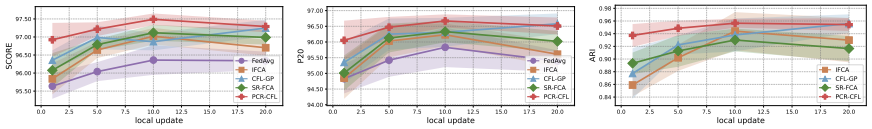
Figure 10: **Sensitivity to different hyperparameters on the Rotated CIFAR dataset.** Rows vary the number of clients, the minibatch size, the number of local training updates, and the interval between clustering operations or probes. Columns report Score, P20, and ARI.



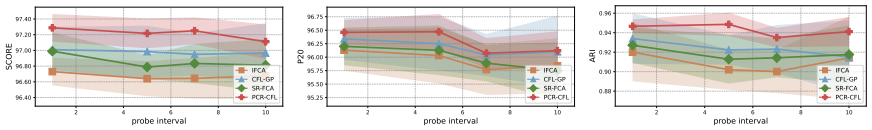
(a) Sensitivity to the number of clients



(b) Sensitivity to the batch size

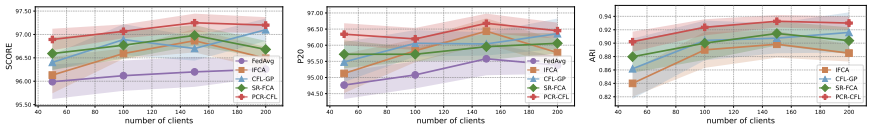


(c) Sensitivity to the number of local updates

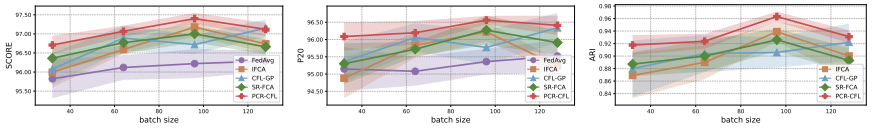


(d) Sensitivity to the probe interval

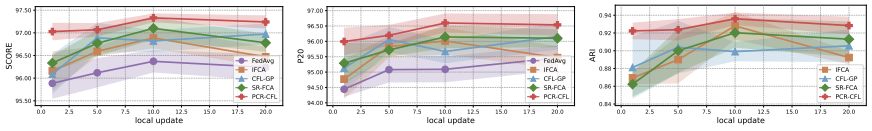
Figure 11: **Sensitivity to different hyperparameters on the MNIST ($\beta = 0.3$) dataset.** Rows vary the number of clients, the minibatch size, the number of local training updates, and the interval between clustering operations or probes. Columns report Score, P20, and ARI.



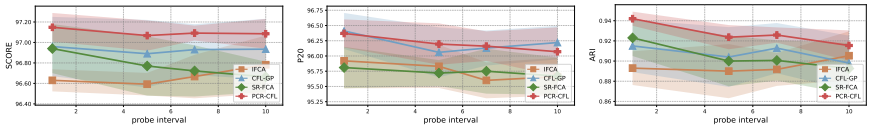
(a) Sensitivity to the number of clients



(b) Sensitivity to the batch size

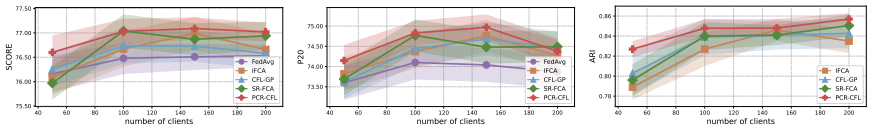


(c) Sensitivity to the number of local updates

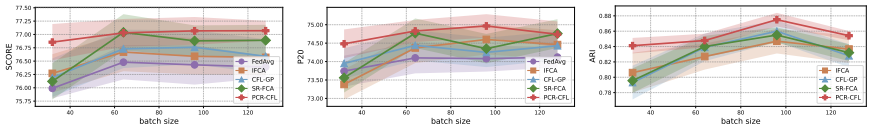


(d) Sensitivity to the probe interval

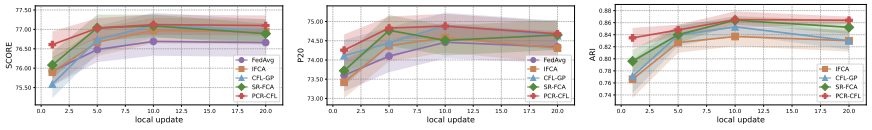
Figure 12: **Sensitivity to different hyperparameters on the MNIST ($\beta = 0.5$) dataset.** Rows vary the number of clients, the minibatch size, the number of local training updates, and the interval between clustering operations or probes. Columns report Score, P20, and ARI.



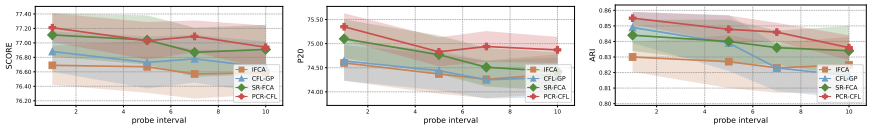
(a) Sensitivity to the number of clients



(b) Sensitivity to the batch size

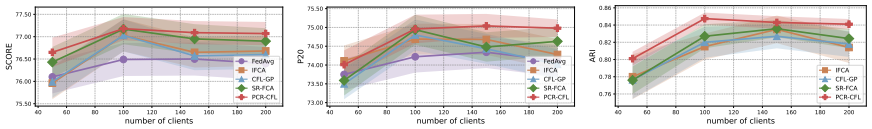


(c) Sensitivity to the number of local updates

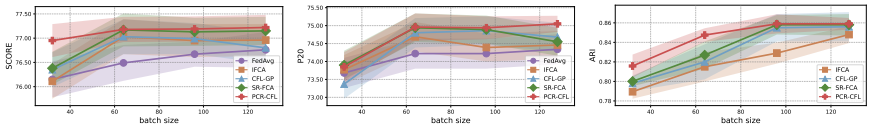


(d) Sensitivity to the probe interval

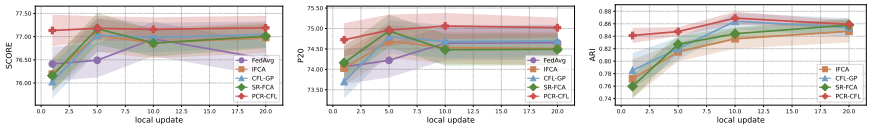
Figure 13: **Sensitivity to different hyperparameters on the CIFAR-10 ($\beta = 0.3$) dataset.** Rows vary the number of clients, the minibatch size, the number of local training updates, and the interval between clustering operations or probes. Columns report Score, P20, and ARI.



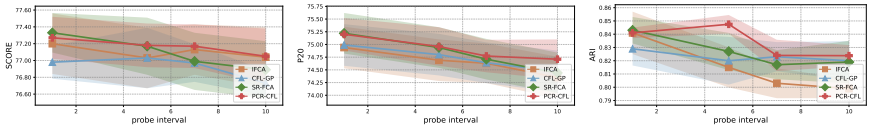
(a) Sensitivity to the number of clients



(b) Sensitivity to the batch size

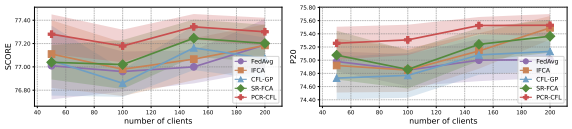


(c) Sensitivity to the number of local updates

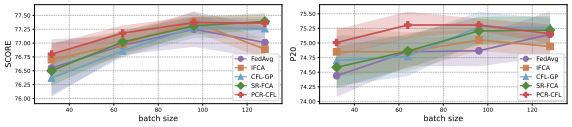


(d) Sensitivity to the probe interval

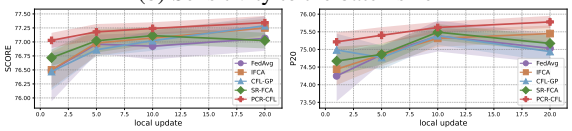
Figure 14: **Sensitivity to different hyperparameters on the CIFAR-10 ($\beta = 0.5$) dataset.** Rows vary the number of clients, the minibatch size, the number of local training updates, and the interval between clustering operations or probes. Columns report Score, P20, and ARI.



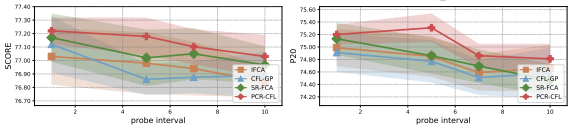
(a) Sensitivity to the number of clients



(b) Sensitivity to the batch size



(c) Sensitivity to the number of local updates



(d) Sensitivity to the probe interval

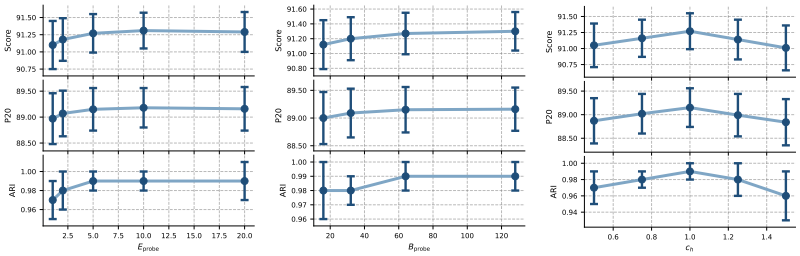
Figure 15: **Sensitivity to different hyperparameters on the FEMNIST dataset.** Rows vary the number of clients, training minibatch size, number of local training updates, and the interval between clustering operations or probes. Columns report Score and P20.

B.5 Ablation Studies of PCR-CFL Probe Hyperparameters

We evaluate the parameters in PCR−CFL during the probe process: the adaptive affinity threshold h_r, probe minibatch size B_{probe}, and number of local probe updates E_{probe}. The threshold ablation applies a scaling factor c_h to the adaptive threshold h_r, and we vary the scaling factor $c_h \in \{0.5, 0.75, 1, 1.25, 1.5\}$. We also vary $B_{\text{probe}} \in \{16, 32, 64, 128\}$ and $E_{\text{probe}} \in \{1, 2, 5, 10, 20\}$ while keeping the remaining configuration fixed. The studies use Rotated MNIST and MNIST with $\beta = 0.3$, which have clear and weaker cluster structures, respectively, and FEMNIST. FEMNIST has no prescribed ground-truth cluster structure. We report Score, P20, and ARI. The complete results are reported in Figure 16, 17, and 18. The results show that our PCR−CFL is robust to the number of local probe updates and the probe minibatch size. On Rotated MNIST and MNIST with $\beta = 0.3$, which represent clear and weak cluster structures, respectively, the best model and clustering results are obtained near $c_h = 1$, since the choice $c_h = 1$ places the threshold at the midpoint between the empirical within-cluster and cross-cluster distance groups, making it more likely to lie inside the interval required by the theorems. FEMNIST has no prescribed ground-truth cluster structure. On this dataset, changing c_h has little effect on Score or P20. These results show that PCR−CFL is robust to these probe hyperparameters.

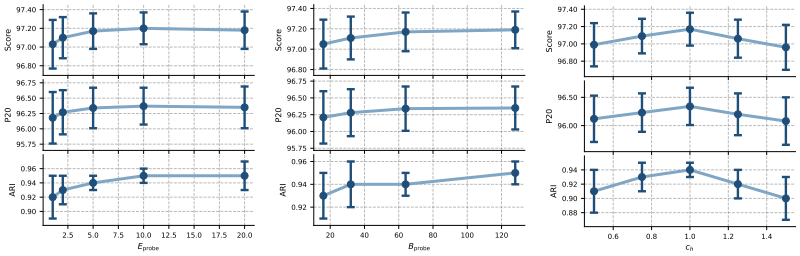
We further conduct additional comparisons on Rotated MNIST, MNIST with $\beta = 0.3$, and FEMNIST to evaluate distance accumulation, the choice of the common probe model, and the probe learning-rate schedule. Each comparison changes one component of PCR−CFL, with $E_{\text{probe}} = 5$, $B_{\text{probe}} = 64$, $H = 5$, and $c_h = 1$. The initial probe learning rate is $\alpha_0 = 0.1$ for the two MNIST datasets and $\alpha_0 = 0.04$ for FEMNIST. Temporary probe updates use minibatch SGD without momentum or weight decay, with temporary optimizer state reset at each probe. For the distance accumulation ablation, at probe r, we replace the accumulated distance by $d_{i,j}^{\text{current}}(r) = \|\widetilde{\boldsymbol{\phi}}_i^r - \widetilde{\boldsymbol{\phi}}_j^r\|^2$. The adaptive threshold is recomputed from these current distances using the same two-means rule. For the choice of the common probe model, we replace the round-robin choice of a cluster model by the unweighted mean of the updated cluster models before repartitioning: $\boldsymbol{\phi}_r^{\text{mean}} = \frac{1}{K_{t_r-1}} \sum_{k=1}^{K_{t_r-1}} \widetilde{\boldsymbol{\theta}}_k^{t_r}$. The mean is recomputed at every probe, so it accommodates changes in both the cluster models and their number. For the probe learning-rate ablation, we compare the constant rate $\alpha_r = \alpha_0$ with a cosine-annealed rate $\alpha_r = \frac{\alpha_0}{2}\left[1 + \cos\left(\pi(r-1)/R_{\text{probe}}\right)\right]$, where

$R_{\text{probe}} = \lfloor T/H \rfloor$ is the number of probes within the common T communication rounds. The rate is fixed across the local updates of each probe and decreases across probes. As shown in Table 1, using the average of the cluster models as the common probe model or annealing the probe learning rate yields similar Score and P20, and ARI to the default configuration. In contrast, replacing accumulated distances with distances from a single probe lowers Score and P20 on all three datasets and reduces ARI on both MNIST datasets, which shows that distance accumulation could reduce the influence of noise from any individual probe on the clustering decisions.



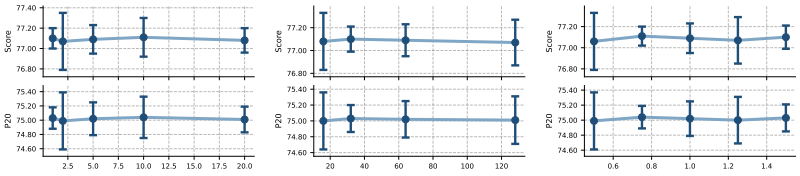
(a) Probe updates (b) Probe minibatch size (c) Threshold multiplier

Figure 16: **Sensitivity to probe hyperparameters on Rotated MNIST.** Panels vary E_{probe}, B_{probe}, and c_h, respectively; each panel reports Score, P20, and ARI.



(a) Probe updates (b) Probe minibatch size (c) Threshold multiplier

Figure 17: **Sensitivity to probe hyperparameters on MNIST ($\beta = 0.3$).** Panels vary E_{probe}, B_{probe}, and c_h, respectively; each panel reports Score, P20, and ARI.



(a) Probe updates (b) Probe minibatch size (c) Threshold multiplier

Figure 18: **Sensitivity to probe hyperparameters on FEMNIST.** Panels vary E_{probe}, B_{probe}, and c_h, respectively; each panel reports Score and P20.

B.6 Large-Scale Datasets

We evaluate the effectiveness of `PCR-CFL` on various large-scale real-world applications.

For the image classification task, we use EMNIST (Cohen et al., 2017) as in (Kim et al., 2024). EMNIST consists of handwritten digits and alphabets, resulting in 62 classes, which has natural heterogeneity stemming from characters being written by different authors. We partition the handwritten characters in EMNIST according to their author as that in (Caldas et al., 2018). We vary the number of clients as $N \in \{40, 80, 120, 160\}$. Following the construction of `CFL-GP` (Kim et al., 2024), we partition the EMNIST examples among N clients with heterogeneous label proportions drawn from the Dirichlet distribution with concentration parameter 1.0. We then divide the clients into two

Table 1: **Probe ablations on Rotated MNIST, MNIST with $\beta = 0.3$, and FEMNIST.**

Dataset	Metric	Default configuration	Non-accumulated distances	Averaged cluster models	Cosine-annealed probe learning rate
Rotated MNIST	Score	91.27 ± 0.57	87.97 ± 0.38	$\mathbf{91.51 \pm 0.43}$	91.46 ± 0.52
	P20	89.15 ± 0.41	84.52 ± 0.51	89.30 ± 0.52	$\mathbf{89.39 \pm 0.48}$
	ARI	$\mathbf{0.99 \pm 0.01}$	0.94 ± 0.03	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$
MNIST ($\beta = 0.3$)	Score	97.22 ± 0.19	93.92 ± 0.46	$\mathbf{97.37 \pm 0.15}$	97.07 ± 0.15
	P20	96.47 ± 0.33	89.75 ± 0.62	$\mathbf{96.50 \pm 0.29}$	96.24 ± 0.34
	ARI	0.95 ± 0.01	0.82 ± 0.02	$\mathbf{0.96 \pm 0.01}$	0.93 ± 0.02
FEMNIST	Score	$\mathbf{77.18 \pm 0.14}$	76.80 ± 0.11	77.10 ± 0.13	77.13 ± 0.16
	P20	$\mathbf{75.31 \pm 0.23}$	74.37 ± 0.36	74.86 ± 0.19	75.26 ± 0.19

groups: one group retains the original images, and the other rotates every image by 180 degrees. We use a similar two-layer LeNet CNN model as in (Dhawan et al., 2024) for EMNIST. We set the training minibatch size $B_{\text{train}} = 64$, local training updates $E_{\text{train}} = 5$ per round, and clustering/probe interval $H = 5$ as the default configuration. The training budget is 1000 epochs. The default probe configuration is $E_{\text{probe}} = 5$ local probe updates and probe minibatch size $B_{\text{probe}} = 64$. We follow the configurations used in (Caldas et al., 2018), with stochastic gradient descent as the local optimizer and Adam as the probe model optimizer. For local and probe learning rates, we conducted a grid search over $\{1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}\}$ and chose the best hyperparameters. Score is averaged classification accuracy across clients.

For natural language processing, we use the domain-partitioned FedC4 (Charles et al., 2023; Dhawan et al., 2024), where each domain is a client. We select $N \in \{40, 80, 120, 160\}$ clients with more than 10,000 words in the text. To construct the cluster structure, we rank ordinary vocabulary tokens by their frequency in the text of the selected N clients, and place consecutive tokens into groups of two. For a group (v_0, v_1), a client in cluster $k \in \{0, 1\}$ replaces each v_j by $v_{(j+k) \bmod 2}$. The $k = 0$ cluster uses the original tokens, while the $k = 1$ cluster uses the one-position cyclic permutation. Special tokens and any incomplete final group remain unchanged. We assign half the clients to each cluster and apply its permutation to both input tokens and next-token targets. Clients in the same cluster share a permutation but retain their distinct domains. We use the similar 1.5B-parameter decoder-only Transformer with sequence length 1024 for FedC4 as in (Dhawan et al., 2024). We set the training minibatch size $B_{\text{train}} = 8$, local training updates $E_{\text{train}} = 5$ per round, and clustering/probe interval $H = 5$ as the default configuration. The number of participation clients is 8 at each round. The training budget is 10000 epochs. The default probe configuration is $E_{\text{probe}} = 5$ local probe updates and probe minibatch size $B_{\text{probe}} = 8$. We follow the configurations used in (Caldas et al., 2018), with stochastic gradient descent as the local optimizer and Adam as the probe model optimizer. For local and probe learning rates, we conducte a grid search over $\{1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}\}$ and choose the best hyperparameters. Score is averaged next-token prediction accuracy across clients.

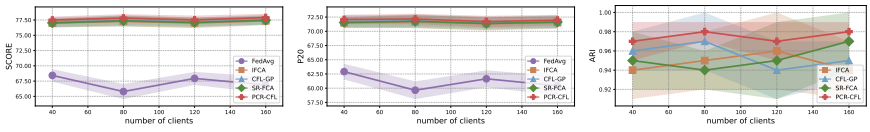


Figure 19: **Large-scale results on EMNIST.** The panels report Score, P20, and ARI, respectively.

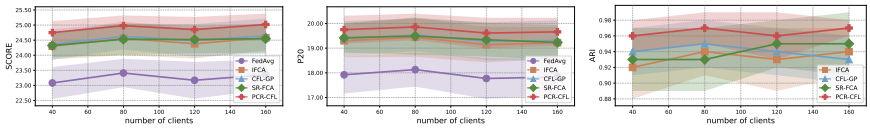


Figure 20: **Large-scale results on FedC4.** The panels report Score, P20, and ARI, respectively.

C NOTATION

Symbol	Meaning
$N, [N]$	Number of clients and client index set.
$d, \mathcal{W}, R$	Model dimension, model parameter space, and its diameter.
$f_i, \nabla f_i, \mathbf{g}_i$	Risk, risk gradient, and stochastic gradient of client i.
p_i	Client weight.
$K^\star, \mathcal{C}^\star$	Ground-truth number of clusters and ground-truth client partition.
$\boldsymbol{w}_i^\star, \boldsymbol{c}_k^\star$	Client risk minimizer, ground-truth cluster center.
ϵ, Δ	Within-cluster radius and minimum center separation.
$F_k, \boldsymbol{\theta}_k^\star, P_k$	Risk, optimal model, and client mass of ground-truth cluster k.
$K_t, \mathcal{C}^t$	Number of clusters and client partition.
$\boldsymbol{\theta}_k^t, \widetilde{\boldsymbol{\theta}}_k^{t+1}$	Current cluster model and intermediate cluster model at round t.
$H, t_r, \boldsymbol{\phi}_r$	Probe interval, round of probe r, and common probe model.
E, B	Number of local probe updates and probe minibatch size.
$\alpha, \widetilde{\boldsymbol{\phi}}_i^r$	Probe learning rate and adapted probe model of client i.
$\rho_{\mathrm{p}}, r_{\mathrm{p}}$	One-step probe contraction factor and probe-adaptation error bound.
$\mathrm{dist}(\cdot,\cdot)$	Probe-response distance.
$d_{i,j}(r)$	Accumulated probe-response distance between clients i and j.
$\boldsymbol{W}_r, \boldsymbol{D}_r, \boldsymbol{L}_r$	Affinity, degree, and normalized Laplacian matrices after probe r.
h, κ_h	Affinity threshold and distance-separation margin.
$r(t)$	Number of completed probes by round t.
$r_\delta, \bar{r}_\delta$	Probe count for recovery at one round, and probe count for permanent recovery.
η_t, ρ	Cluster-model learning rate and one-round model contraction factor.
$e_t^{\mathrm{c}}, e_t^{\mathrm{m}}$	Weighted clustering error and weighted cluster-model error.
$\mathcal{E}_t, T_{\mathrm{exact}}$	Exact-recovery event $\{e_t^{\mathrm{c}} = 0\}$ and first round after which it holds permanently.

Superscript $\star$ marks a ground-truth or optimal quantity. The index t denotes a communication round, while r denotes a clustering probe. Expectations and probabilities are written as $\mathbb{E}[\cdot]$ and $\Pr(\cdot)$, and $\mathbf{1}\{\cdot\}$ denotes an indicator.

D TECHNICAL RESULTS

D.1 ANALYSIS OF THE LEARNING-RATE-DILEMMA EXAMPLE

Proof. The set $\mathcal{W} = [-a, 2a] \times [-a, 3a]$ is convex and has finite diameter $5a$. It contains all four client minimizers and the initialization regions used below. The eigenvalues of $\boldsymbol{H}$ are 1 and 4, so every client risk is 1-strongly convex and 4-smooth. Set $\boldsymbol{c}_1^\star = (a/2, 0)^\mathsf{T}$ and $\boldsymbol{c}_2^\star = (0, 2a)^\mathsf{T}$. The client minimizers satisfy the separable cluster structure with $\epsilon = a/2$ and $\Delta = \sqrt{17}a/2 > 4\epsilon$.

Under the stated incorrect client partition, the two cluster risks are minimized at

$$\overline{\boldsymbol{\theta}}_1 = (a, 0)^\mathsf{T}, \qquad \overline{\boldsymbol{\theta}}_2 = (0, 4a/3)^\mathsf{T}. \tag{14}$$

Since all four risks have Hessian $\boldsymbol{H} = \mathrm{diag}(4, 1)$, the cluster updates satisfy

$$\boldsymbol{\theta}_k^{t+1} - \overline{\boldsymbol{\theta}}_k = (\boldsymbol{I}_2 - \gamma \boldsymbol{H})(\boldsymbol{\theta}_k^t - \overline{\boldsymbol{\theta}}_k), \qquad k \in [2]. \tag{15}$$

For $0 < \gamma < 1/4$, $\|\boldsymbol{I}_2 - \gamma \boldsymbol{H}\|_2 = 1 - \gamma < 1$. Consequently, if $\|\boldsymbol{\theta}_k^t - \overline{\boldsymbol{\theta}}_k\| < a/10$, then

$$\|\boldsymbol{\theta}_k^{t+1} - \overline{\boldsymbol{\theta}}_k\| \leqslant (1 - \gamma)\|\boldsymbol{\theta}_k^t - \overline{\boldsymbol{\theta}}_k\| < \frac{a}{10}. \tag{16}$$

Thus, both initialization regions are invariant as long as the client partition remains unchanged. They lie in the interior of $\mathcal{W}$, so no projection is active.

We next show that each assignment rule preserves the partition throughout these regions.

① At the fixed points in Eq. (14), the local models produced by the model-distance rule are

$$\overline{\boldsymbol{u}}_1 = (a, 0)^\mathsf{T}, \quad \overline{\boldsymbol{u}}_2 = (0, (1 - \alpha)4a/3)^\mathsf{T}, \quad \overline{\boldsymbol{u}}_3 = \overline{\boldsymbol{u}}_4 = (0, (4 + 2\alpha)a/3)^\mathsf{T}. \tag{17}$$

The difference between the distance to the other cluster model and the distance to the current cluster model is $5a/3$ for client 1,

$$a\sqrt{1+\frac{16}{9}(1-\alpha)^2}-\frac{4\alpha a}{3}>a \tag{18}$$

for client 2, and

$$a\sqrt{1+\frac{(4+2\alpha)^2}{9}}-\frac{2\alpha a}{3}>a \tag{19}$$

for clients 3 and 4. Both inequalities hold for $0<\alpha<1/4$.

For cluster models in the stated initialization regions, let $\boldsymbol{u}_i$ and $\overline{\boldsymbol{u}}_i$ denote the corresponding local models at the perturbed and fixed cluster models, respectively. Since $\|\boldsymbol{I}_2-\alpha\boldsymbol{H}\|_2=1-\alpha<1$,

$$\|\boldsymbol{u}_i-\overline{\boldsymbol{u}}_i\|<\frac{a}{10}. \tag{20}$$

The reverse triangle inequality then shows that each client-to-model distance changes by less than $a/5$. Hence, the difference between the two distances changes by less than $2a/5$. Since its fixed-point value is greater than a, it remains greater than $3a/5$. The model-distance rule therefore preserves the incorrect client partition.

② At the fixed points, the loss advantage of the current cluster model is $26a^2/9$ for client 1, $10a^2/9$ for client 2, and $34a^2/9$ for clients 3 and 4. In particular, the smallest advantage is

$$f_2(\overline{\boldsymbol{\theta}}_1)-f_2(\overline{\boldsymbol{\theta}}_2)=2a^2-\frac{8a^2}{9}=\frac{10a^2}{9}. \tag{21}$$

Direct calculation gives $\max_{i\in[4],k\in[2]}\|\nabla f_i(\overline{\boldsymbol{\theta}}_k)\|=2\sqrt{5}a$. For any $\boldsymbol{\theta}_k$ satisfying $\|\boldsymbol{\theta}_k-\overline{\boldsymbol{\theta}}_k\|<a/10$, the quadratic expansion gives

$$\left|f_i(\boldsymbol{\theta}_k)-f_i(\overline{\boldsymbol{\theta}}_k)\right|<2\sqrt{5}a\frac{a}{10}+2\left(\frac{a}{10}\right)^2. \tag{22}$$

Therefore, the difference between the two loss values changes by less than

$$\left(\frac{2\sqrt{5}}{5}+\frac{1}{25}\right)a^2<\frac{10a^2}{9}. \tag{23}$$

Thus, the loss rule also preserves the incorrect client partition.

③ For the gradient-distance rule, let $\boldsymbol{\phi}$ be any common reference model. Since the four risks have the same Hessian, translating all gradients by $\boldsymbol{H}\boldsymbol{\phi}$ does not change their pairwise distances. After removing this common translation, the four gradient vectors are

$$(-4a,0)^\mathsf{T},\qquad(0,0)^\mathsf{T},\qquad(0,-2a)^\mathsf{T},\qquad(0,-2a)^\mathsf{T}. \tag{24}$$

The ground-truth partition $\{\{1,2\},\{3,4\}\}$ has a two-means objective value of $8a^2$, whereas $\{\{1\},\{2,3,4\}\}$ has an objective value of $8a^2/3$. Every other nonempty two-partition has an objective value of at least $8a^2$. Consequently, Euclidean two-means uniquely returns $\{\{1\},\{2,3,4\}\}$ up to label permutation.

Starting from the two initialization regions, apply Eq. (16) and the three strict comparisons above inductively. Each rule preserves the incorrect client partition at every round. Finally, Eq. (15) implies that $\boldsymbol{\theta}_1^t\to\overline{\boldsymbol{\theta}}_1$ and $\boldsymbol{\theta}_2^t\to\overline{\boldsymbol{\theta}}_2$.

The following loss-value comparison uses $\gamma=3/4$, which exceeds the stable gradient-descent interval $0<\gamma<2/\lambda_{\max}(\boldsymbol{H})=1/2$ for this quadratic problem. We take the loss-value rule with $\gamma=3/4$, $\boldsymbol{\theta}_1^0=(41a/40,0)^\mathsf{T}$, and $\boldsymbol{\theta}_2^0=(0,4a/3)^\mathsf{T}$. As long as the partition is unchanged, the first coordinates of $\boldsymbol{\theta}_1^t/a$ for $t=0,\dots,5$ are $41/40,19/20,11/10,4/5,7/5,1/5$; all iterates remain in $\mathcal{W}$. Before the fifth update, client 2 prefers $\boldsymbol{\theta}_2^t$, since $f_2(\boldsymbol{\theta}_1^t)\geqslant 2(4a/5)^2>8a^2/9=f_2(\boldsymbol{\theta}_2^t)$ for $t=0,\dots,4$. After the fifth update, $f_2(\boldsymbol{\theta}_1^5)=2a^2/25<8a^2/9=f_2(\boldsymbol{\theta}_2^5)$, so client 2 selects cluster 1 and clients $1,3,4$ retain their assignments through this update. □

Initialization sensitivity of SR-FCA under limited local computation. The refinement analysis of SR−FCA (Vardhan et al., 2024) relies on an initial partition in which every cluster has an impurity fraction below α_{imp}, for some $0 < \alpha_{\mathrm{imp}} < \beta_{\mathrm{trim}} < 1/2$, and every true cluster has a representative initial cluster. Here, impurities are members outside the majority true cluster, and β_{trim} is the trimming parameter. Its ONE_SHOT stage constructs this partition from locally trained models, and the analysis bounds the probability of violating these requirements. With limited local computation, the quality of this constructed partition can depend on the model initialization. Consider two equally sized true clusters with population risks

$$f_A(w) = \tfrac{1}{2}(w-1)^2, \qquad f_B(w) = (w-\tfrac{1}{2})^2, \qquad w_A^\star = 1, \quad w_B^\star = \tfrac{1}{2}. \tag{25}$$

Let each client take one exact gradient step with learning rate $1/4$ from a common initialization w_0, and fix the model-distance threshold at $\lambda = 1/10$. If $w_0 = 0$, both types return $w_1 = 1/4$. Thus, all pairwise model distances vanish and the threshold graph is complete: ONE_SHOT merges the two types into one cluster, with one half of its members from each true cluster. This partition fails the strict-majority and coverage requirements for refinement. In contrast, if $w_0 = 1/2$, the returned models are $w_{A,1} = 5/8$ and $w_{B,1} = 1/2$. Their cross-cluster distance is $1/8 > \lambda$, whereas within-cluster distances are zero, so the threshold graph separates the two true clusters. In both cases, set the minimum retained cluster size no larger than the size of either true cluster. The tasks, threshold, learning rate, and local computation budget are identical; changing only the initialization changes whether ONE_SHOT produces a partition satisfying the refinement requirements. This provides a concrete mechanism for potential initialization dependence when local computation is limited.

Initialization sensitivity of gradient separation in CFL-GP. CFL−GP (Kim et al., 2024) accumulates gradients evaluated at shared cluster models. Separation between optimal parameters need not yield separation between gradients at the models visited during training. Consider two task types

$$f_A(w) = \tfrac{1}{2}w^2, \qquad f_B(w) = (w-1)^2, \qquad w_A^\star = 0, \quad w_B^\star = 1. \tag{26}$$

At the common query point $w_0 = 2$, both gradients equal 2, even though the optimal parameters are separated by 1. Suppose an incorrect cluster contains equal proportions of the two types. For as long as this mixture is retained, exact gradient descent on its average risk $F = (f_A + f_B)/2$, with effective learning rate $0 < \eta \leqslant 2/3$, gives

$$w_{t+1} = w_t - \eta(\tfrac{3}{2}w_t - 1), \qquad w_t = \tfrac{2}{3} + \tfrac{4}{3}(1 - \tfrac{3}{2}\eta)^t. \tag{27}$$

Consequently, the gradient separation at these query points satisfies

$$|\nabla f_A(w_t) - \nabla f_B(w_t)| = \tfrac{4}{3}\left[1 - (1 - \tfrac{3}{2}\eta)^t\right] \leqslant 2\eta t. \tag{28}$$

For any finite horizon $T \geqslant 1$ and tolerance $\varepsilon > 0$, choosing $\eta \leqslant \min\{2/3, \varepsilon/(2T)\}$ keeps this gap at most ε for all $t \leqslant T$ along the mixed trajectory. The same bound applies to the difference between any common convex averages of these gradient observations. At the alternative initial query point $w_0 = 0$, the gradient gap is already 2. Thus, a poor initialization combined with a small learning rate can sustain weak gradient separation over a finite training window, reducing the distinction available for clustering in the presence of sampling noise.

Finite-step probe geometry and accumulated distances. Consider the four-client quadratic example above with exact gradients and a fixed $0 < \alpha \leqslant 1/4$. Let $\boldsymbol{A}_E = \boldsymbol{I}_2 - (\boldsymbol{I}_2 - \alpha\boldsymbol{H})^E$. Each coordinate of a local update is a convex combination of the current coordinate and the corresponding client minimizer, so all iterates remain in $\mathcal{W}$ and projection is inactive. Unrolling the E updates from any common probe $\boldsymbol{\phi}_u \in \mathcal{W}$ gives

$$\widetilde{\boldsymbol{\phi}}_i^u = \boldsymbol{w}_i^\star + (\boldsymbol{I}_2 - \alpha\boldsymbol{H})^E(\boldsymbol{\phi}_u - \boldsymbol{w}_i^\star), \qquad \widetilde{\boldsymbol{\phi}}_i^u - \widetilde{\boldsymbol{\phi}}_j^u = \boldsymbol{A}_E(\boldsymbol{w}_i^\star - \boldsymbol{w}_j^\star). \tag{29}$$

Consequently, for every number of completed probes $r \geqslant 1$,

$$d_{i,j}(r) = \frac{1}{r}\sum_{u=1}^{r} \|\boldsymbol{A}_E(\boldsymbol{w}_i^\star - \boldsymbol{w}_j^\star)\|^2 = \|\boldsymbol{A}_E(\boldsymbol{w}_i^\star - \boldsymbol{w}_j^\star)\|^2. \tag{30}$$

This expression is independent of the common probe sequence and hence of the training learning rate and initial client assignments in this example. Writing $s_E = 1 - (1-4\alpha)^E$ and $t_E = 1 - (1-\alpha)^E$, the largest within-cluster and smallest cross-cluster accumulated distances are

$$\max_{i \neq j:\, z_i^\star = z_j^\star} d_{i,j}(r) = a^2 s_E^2, \qquad \min_{z_i^\star \neq z_j^\star} d_{i,j}(r) = 4a^2 t_E^2. \tag{31}$$

Thus, $s_E < 2t_E$ suffices for exact affinity recovery with any $a^2 s_E^2 < h^2 < 4a^2 t_E^2$. For example, take $\alpha = 1/16 = \mu/L^2$ and $E = 10$. Then $s_E = 1-(3/4)^{10} \approx 0.943686$ and $2t_E = 2[1-(15/16)^{10}] \approx 0.951079$, so the gap is strictly positive. Finite-step probing therefore restores the distance ordering needed for the true partition even while the training models remain trapped under the small-learning-rate dynamics above. It does not require the adapted models to reach their individual minimizers. For stochastic probes, let $m_{i,j}(r) = r^{-1} \sum_{u=1}^{r} \mathbb{E}[\|\widetilde{\boldsymbol{\phi}}_i^u - \widetilde{\boldsymbol{\phi}}_j^u\|^2 \mid \mathcal{H}_{u-1}]$. The concentration argument in Eq. (68) yields

$$\Pr\left(\max_{i<j} |d_{i,j}(r) - m_{i,j}(r)| \geqslant x\right) \leqslant N(N-1)\exp\left(-\frac{2rx^2}{R^4}\right), \qquad x > 0. \tag{32}$$

Once the conditional mean distances have a positive separation margin, accumulation reduces the probability that sampling fluctuations reverse a pairwise decision. The fixed-E deterministic expression in Eq. (30) does not change with r; the benefit of increasing r is concentration around the mean response distances. The results distinguish the benefits of the probe procedure. First, each probe provides a common starting point for all clients, so their responses are compared under the same initialization rather than under different assigned cluster models. In this quadratic example, the shared starting-point term cancels exactly in Eq. (29). Second, accumulating response distances improves statistical reliability: for a fixed positive mean separation margin, the bound in Eq. (32) decreases exponentially with the number of probes. Third, the direct calculation gives a strict within-/cross-cluster distance gap. Thus, recovery in this example requires neither adaptation to a local minimizer nor a limit in which the number of local adaptation steps tends to infinity.

D.2 Intermediate Results

The following quantities are used only in the proofs. Let

$$\rho_{\mathrm{p}}^2 := 1 - 2\mu\alpha + L^2\alpha^2. \tag{33}$$

Here, ρ_{p} is the contraction factor used to derive the E-step probe-adaptation error bound used below.

We also use the ground-truth affinity matrix $\boldsymbol{W}^\star \in \mathbb{R}^{N\times N}$, defined by

$$W_{i,j}^\star := \frac{1}{N}\mathbf{1}\{i \neq j,\ z_i^\star = z_j^\star\}. \tag{34}$$

Thus, $W_{i,j}^\star = 1/N$ exactly when two distinct clients belong to the same ground-truth cluster, and it is zero otherwise. This matrix is used only as the target of the affinity-matrix recovery proof and is not available to the algorithm.

Lemma D.1 (Separation produced by multi-step probes). *Suppose Assumption 3.1, 3.2, 3.3 and Definition 4.1 hold, and all the probe samples are drawn independently at each probe. If $0 < \alpha \leqslant \mu/L^2$, define the probe-adaptation error radius $r_{\mathrm{p}} := (e^{-\mu\alpha E}R^2 + \alpha\sigma^2/(B\mu))^{1/2}$. Then, for every client i and probe r,*

$$\mathbb{E}\left[\|\widetilde{\boldsymbol{\phi}}_i^r - \boldsymbol{w}_i^\star\|^2 \mid \mathcal{H}_{r-1}\right] \leqslant r_{\mathrm{p}}^2, \tag{35}$$

where $\mathcal{H}_{r-1}$ contains all randomness revealed before probe r. Moreover, if $\Delta > 2\epsilon + 2r_{\mathrm{p}}$, then

$$\begin{aligned}
&\max_{\substack{i,j\in[N]\\ z_i^\star = z_j^\star}} \mathbb{E}\left[\|\widetilde{\boldsymbol{\phi}}_i^r - \widetilde{\boldsymbol{\phi}}_j^r\|^2 \mid \mathcal{H}_{r-1}\right] \leqslant (2\epsilon + 2r_{\mathrm{p}})^2,\\
&\min_{\substack{i,j\in[N]\\ z_i^\star \neq z_j^\star}} \mathbb{E}\left[\|\widetilde{\boldsymbol{\phi}}_i^r - \widetilde{\boldsymbol{\phi}}_j^r\|^2 \mid \mathcal{H}_{r-1}\right] \geqslant (\Delta - 2\epsilon - 2r_{\mathrm{p}})^2.
\end{aligned} \tag{36}$$

Proof of Lemma D.1. Fix a probe r and client i. The common probe model $\boldsymbol{\phi}_r$ is selected before the local samples of probe r are drawn and is therefore $\mathcal{H}_{r-1}$-measurable. For local probe update e and sample b, define the single-sample gradient noise by

$$\varepsilon_{i,b}^{r,e} := \mathbf{g}_i(\boldsymbol{\phi}_i^{r,e}; \xi_{i,b}^{r,e}) - \nabla f_i(\boldsymbol{\phi}_i^{r,e}). \tag{37}$$

The minibatch gradient and its average noise are then

$$\overline{\mathbf{g}}_i^{r,e} := \frac{1}{B}\sum_{b=1}^{B} \mathbf{g}_i(\boldsymbol{\phi}_i^{r,e}; \xi_{i,b}^{r,e}), \qquad \varepsilon_i^{r,e} := \overline{\mathbf{g}}_i^{r,e} - \nabla f_i(\boldsymbol{\phi}_i^{r,e}) = \frac{1}{B}\sum_{b=1}^{B} \varepsilon_{i,b}^{r,e}. \tag{38}$$

Let $\mathcal{H}_{r,e}$ contain $\mathcal{H}_{r-1}$ and all probe samples revealed before local probe update e. Thus, $\boldsymbol{\phi}_i^{r,e}$ is $\mathcal{H}_{r,e}$-measurable, whereas $\{\xi_{i,b}^{r,e}\}_{b=1}^{B}$ are fresh conditionally independent samples. By the unbiasedness and variance bounds in Eq. (3),

$$\mathbb{E}[\varepsilon_{i,b}^{r,e} \mid \mathcal{H}_{r,e}] = 0, \qquad \mathbb{E}[\|\varepsilon_{i,b}^{r,e}\|^2 \mid \mathcal{H}_{r,e}] \leqslant \sigma^2. \tag{39}$$

For $b \neq b'$, conditional independence and the zero conditional means give

$$\mathbb{E}\Big[\Big\langle \varepsilon_{i,b}^{r,e}, \varepsilon_{i,b'}^{r,e} \Big\rangle \,\Big|\, \mathcal{H}_{r,e}\Big] = 0. \tag{40}$$

Consequently,

$$\mathbb{E}[\|\varepsilon_i^{r,e}\|^2 \mid \mathcal{H}_{r,e}] = \frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}[\|\varepsilon_{i,b}^{r,e}\|^2 \mid \mathcal{H}_{r,e}] + \frac{2}{B^2}\sum_{1\leqslant b<b'\leqslant B}\mathbb{E}\Big[\langle \varepsilon_{i,b}^{r,e}, \varepsilon_{i,b'}^{r,e}\rangle \,\Big|\, \mathcal{H}_{r,e}\Big] \leqslant \frac{\sigma^2}{B}. \tag{41}$$

Together with linearity of conditional expectation, this proves

$$\mathbb{E}[\varepsilon_i^{r,e} \mid \mathcal{H}_{r,e}] = 0, \qquad \mathbb{E}[\|\varepsilon_i^{r,e}\|^2 \mid \mathcal{H}_{r,e}] \leqslant \frac{\sigma^2}{B}. \tag{42}$$

We next compare one local iterate with the constrained minimizer $\boldsymbol{w}_i^\star$. Since f_i is differentiable and convex and $\mathcal{W}$ is closed and convex, first-order optimality gives

$$\langle \nabla f_i(\boldsymbol{w}_i^\star), \boldsymbol{w} - \boldsymbol{w}_i^\star\rangle \geqslant 0, \qquad \forall \boldsymbol{w} \in \mathcal{W}. \tag{43}$$

Euclidean projection satisfies $\boldsymbol{x} = \Pi_{\mathcal{W}}(\boldsymbol{y})$ if and only if $\langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{w} - \boldsymbol{x}\rangle \leqslant 0$ for every $\boldsymbol{w} \in \mathcal{W}$. Apply this characterization with $\boldsymbol{x} = \boldsymbol{w}_i^\star$ and $\boldsymbol{y} = \boldsymbol{w}_i^\star - \alpha\nabla f_i(\boldsymbol{w}_i^\star)$. Then Eq. (43) is equivalent to

$$\boldsymbol{w}_i^\star = \Pi_{\mathcal{W}}\left(\boldsymbol{w}_i^\star - \alpha\nabla f_i(\boldsymbol{w}_i^\star)\right). \tag{44}$$

For compactness, set

$$\boldsymbol{e}_i^{r,e} := \boldsymbol{\phi}_i^{r,e} - \boldsymbol{w}_i^\star, \qquad \boldsymbol{\delta}_i^{r,e} := \nabla f_i(\boldsymbol{\phi}_i^{r,e}) - \nabla f_i(\boldsymbol{w}_i^\star). \tag{45}$$

The Euclidean projection is nonexpansive. Using Eqs. (4) and (44) and $\overline{\mathbf{g}}_i^{r,e} = \nabla f_i(\boldsymbol{\phi}_i^{r,e}) + \varepsilon_i^{r,e}$ therefore gives the pointwise bound

$$\|\boldsymbol{\phi}_i^{r,e+1} - \boldsymbol{w}_i^\star\|^2 \leqslant \|\boldsymbol{e}_i^{r,e} - \alpha\boldsymbol{\delta}_i^{r,e} - \alpha\varepsilon_i^{r,e}\|^2. \tag{46}$$

Expanding the right-hand side before taking expectations gives

$$\|\boldsymbol{e}_i^{r,e} - \alpha\boldsymbol{\delta}_i^{r,e} - \alpha\varepsilon_i^{r,e}\|^2 = \|\boldsymbol{e}_i^{r,e} - \alpha\boldsymbol{\delta}_i^{r,e}\|^2 + \alpha^2\|\varepsilon_i^{r,e}\|^2 - 2\alpha\langle \boldsymbol{e}_i^{r,e} - \alpha\boldsymbol{\delta}_i^{r,e}, \varepsilon_i^{r,e}\rangle. \tag{47}$$

Both $\boldsymbol{e}_i^{r,e}$ and $\boldsymbol{\delta}_i^{r,e}$ are $\mathcal{H}_{r,e}$-measurable. Thus, the last term has conditional expectation zero by Eq. (42). Expanding the first term in Eq. (47) and using the minibatch variance bound yields

$$\mathbb{E}\Big[\|\boldsymbol{\phi}_i^{r,e+1} - \boldsymbol{w}_i^\star\|^2 \mid \mathcal{H}_{r,e}\Big] \leqslant \|\boldsymbol{e}_i^{r,e}\|^2 - 2\alpha\,\langle \boldsymbol{e}_i^{r,e}, \boldsymbol{\delta}_i^{r,e}\rangle + \alpha^2\|\boldsymbol{\delta}_i^{r,e}\|^2 + \frac{\alpha^2\sigma^2}{B}. \tag{48}$$

For completeness, apply strong convexity once with $(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{\phi}_i^{r,e}, \boldsymbol{w}_i^\star)$ and once with the arguments reversed. Adding the two inequalities cancels the function values and gives

$$\langle \nabla f_i(\boldsymbol{\phi}_i^{r,e}) - \nabla f_i(\boldsymbol{w}_i^\star), \boldsymbol{\phi}_i^{r,e} - \boldsymbol{w}_i^\star\rangle \geqslant \mu\|\boldsymbol{\phi}_i^{r,e} - \boldsymbol{w}_i^\star\|^2. \tag{49}$$

Smoothness directly bounds the same gradient difference. In the notation of Eq. (45), these two facts read

$$\langle \boldsymbol{e}_i^{r,e}, \boldsymbol{\delta}_i^{r,e}\rangle \geqslant \mu\|\boldsymbol{e}_i^{r,e}\|^2, \qquad \|\boldsymbol{\delta}_i^{r,e}\| \leqslant L\|\boldsymbol{e}_i^{r,e}\|. \tag{50}$$

Substituting Eq. (50) into Eq. (48) gives

$$\mathbb{E}\Big[\|\boldsymbol{\phi}_i^{r,e+1} - \boldsymbol{w}_i^\star\|^2 \mid \mathcal{H}_{r,e}\Big] \leqslant (1 - 2\mu\alpha + L^2\alpha^2)\|\boldsymbol{\phi}_i^{r,e} - \boldsymbol{w}_i^\star\|^2 + \frac{\alpha^2\sigma^2}{B}. \tag{51}$$

By Eq. (33), the coefficient in parentheses equals ρ_{p}^2, so

$$\mathbb{E}\Big[\|\boldsymbol{\phi}_i^{r,e+1} - \boldsymbol{w}_i^\star\|^2 \mid \mathcal{H}_{r,e}\Big] \leqslant \rho_{\mathrm{p}}^2\|\boldsymbol{\phi}_i^{r,e} - \boldsymbol{w}_i^\star\|^2 + \frac{\alpha^2\sigma^2}{B}. \tag{52}$$

To verify the claimed range of this coefficient, complete the square:

$$\rho_{\mathrm{p}}^2 = L^2\left(\alpha - \frac{\mu}{L^2}\right)^2 + 1 - \frac{\mu^2}{L^2} \geqslant 0. \tag{53}$$

Since $\alpha \leqslant \mu/L^2$,

$$\rho_{\mathrm{p}}^2 = 1 - 2\mu\alpha + L^2\alpha^2 \leqslant 1 - \mu\alpha \leqslant e^{-\mu\alpha} < 1. \tag{54}$$

We now unroll the recursion. Repeated use of the tower property shows by induction that, for any $s \in [E]$,

$$\mathbb{E}\big[\|\boldsymbol{\phi}_i^{r,s} - \boldsymbol{w}_i^\star\|^2 \mid \mathcal{H}_{r-1}\big] \leqslant \rho_{\mathrm{p}}^{2s}\|\boldsymbol{\phi}_r - \boldsymbol{w}_i^\star\|^2 + \frac{\alpha^2\sigma^2}{B}\sum_{e=0}^{s-1}\rho_{\mathrm{p}}^{2e}. \tag{55}$$

Indeed, the claim is Eq. (52) for $s = 1$. If it holds for s, condition first on $\mathcal{H}_{r,s}$, apply Eq. (52), and then condition again on $\mathcal{H}_{r-1}$; this multiplies the existing bound by ρ_{p}^2 and adds $\alpha^2\sigma^2/B$. Taking $s = E$, using $\boldsymbol{\phi}_i^{r,0} = \boldsymbol{\phi}_r$ and $\widetilde{\boldsymbol{\phi}}_i^r = \boldsymbol{\phi}_i^{r,E}$, yields

$$\begin{aligned}
\mathbb{E}\Big[\|\widetilde{\boldsymbol{\phi}}_i^r - \boldsymbol{w}_i^\star\|^2 \mid \mathcal{H}_{r-1}\Big] &\leqslant \rho_{\mathrm{p}}^{2E}\|\boldsymbol{\phi}_r - \boldsymbol{w}_i^\star\|^2 + \frac{\alpha^2\sigma^2}{B}\sum_{e=0}^{E-1}\rho_{\mathrm{p}}^{2e} \\
&\leqslant \rho_{\mathrm{p}}^{2E}R^2 + \frac{\alpha^2\sigma^2}{B}\frac{1-\rho_{\mathrm{p}}^{2E}}{1-\rho_{\mathrm{p}}^2} \\
&\leqslant e^{-\mu\alpha E}R^2 + \frac{\alpha\sigma^2}{B\mu} = r_{\mathrm{p}}^2.
\end{aligned} \tag{56}$$

The last inequality uses Eq. (54), $1 - \rho_{\mathrm{p}}^{2E} \leqslant 1$, and

$$1 - \rho_{\mathrm{p}}^2 = 2\mu\alpha - L^2\alpha^2 \geqslant \mu\alpha. \tag{57}$$

This proves Eq. (35).

It remains to compare two clients. For a random vector $\boldsymbol{x}$, define its conditional L_2 norm by $\|\boldsymbol{x}\|_{2,r} := (\mathbb{E}[\|\boldsymbol{x}\|^2 \mid \mathcal{H}_{r-1}])^{1/2}$. This is a norm after conditioning, so both Minkowski's inequality and its reverse form apply. We first translate Definition 4.1 into pairwise bounds. If $z_i^\star = z_j^\star = k$, then

$$\|\boldsymbol{w}_i^\star - \boldsymbol{w}_j^\star\| \leqslant \|\boldsymbol{w}_i^\star - \boldsymbol{c}_k^\star\| + \|\boldsymbol{w}_j^\star - \boldsymbol{c}_k^\star\| \leqslant 2\epsilon. \tag{58}$$

If $z_i^\star = k \neq \ell = z_j^\star$, then

$$\|\boldsymbol{w}_i^\star - \boldsymbol{w}_j^\star\| \geqslant \|\boldsymbol{c}_k^\star - \boldsymbol{c}_\ell^\star\| - \|\boldsymbol{w}_i^\star - \boldsymbol{c}_k^\star\| - \|\boldsymbol{w}_j^\star - \boldsymbol{c}_\ell^\star\| \geqslant \Delta - 2\epsilon. \tag{59}$$

For two clients in the same ground-truth cluster, insert their deterministic optimal models between the two adapted models. Minkowski's inequality, Eqs. (35) and (58) give

$$\|\widetilde{\boldsymbol{\phi}}_i^r - \widetilde{\boldsymbol{\phi}}_j^r\|_{2,r} \leqslant \|\widetilde{\boldsymbol{\phi}}_i^r - \boldsymbol{w}_i^\star\|_{2,r} + \|\boldsymbol{w}_i^\star - \boldsymbol{w}_j^\star\| + \|\boldsymbol{w}_j^\star - \widetilde{\boldsymbol{\phi}}_j^r\|_{2,r} \leqslant 2r_{\mathrm{p}} + 2\epsilon. \tag{60}$$

For clients in different ground-truth clusters, write

$$\boldsymbol{w}_i^\star - \boldsymbol{w}_j^\star = (\boldsymbol{w}_i^\star - \widetilde{\boldsymbol{\phi}}_i^r) + (\widetilde{\boldsymbol{\phi}}_i^r - \widetilde{\boldsymbol{\phi}}_j^r) + (\widetilde{\boldsymbol{\phi}}_j^r - \boldsymbol{w}_j^\star). \tag{61}$$

Apply Minkowski's inequality to Eq. (61) and rearrange. Using Eqs. (35) and (59) then gives

$$\|\widetilde{\boldsymbol{\phi}}_i^r - \widetilde{\boldsymbol{\phi}}_j^r\|_{2,r} \geqslant \|\boldsymbol{w}_i^\star - \boldsymbol{w}_j^\star\| - \|\widetilde{\boldsymbol{\phi}}_i^r - \boldsymbol{w}_i^\star\|_{2,r} - \|\widetilde{\boldsymbol{\phi}}_j^r - \boldsymbol{w}_j^\star\|_{2,r} \geqslant \Delta - 2\epsilon - 2r_{\mathrm{p}}. \tag{62}$$

The final quantity is positive by the condition in the lemma. Since both sides of Eqs. (60) and (62) are nonnegative, squaring preserves their directions. Finally, by the definition of the conditional L_2 norm, $\|\widetilde{\boldsymbol{\phi}}_i^r - \widetilde{\boldsymbol{\phi}}_j^r\|_{2,r}^2$ equals the conditional expected squared distance. Therefore, squaring Eqs. (60) and (62) proves Eq. (36). □

Lemma D.2 (Recovery of the ground-truth affinity matrix). *Suppose Assumption 3.1, 3.2, 3.3 and Definition 4.1 hold, and all the probe samples are drawn independently at each probe. Suppose* $0 < \alpha \leqslant \mu/L^2$. *Define* $r_{\mathrm{p}} := (e^{-\mu\alpha E}R^2 + \alpha\sigma^2/(B\mu))^{1/2}$. *Choose* h *such that* $2(\epsilon + r_{\mathrm{p}}) < h < \Delta - 2(\epsilon + r_{\mathrm{p}})$, *and denote* $\kappa_h := \min\{h^2 - (2\epsilon + 2r_{\mathrm{p}})^2, (\Delta - 2\epsilon - 2r_{\mathrm{p}})^2 - h^2\}$. *Then, for every* $r \in \mathbb{N}$,

$$\Pr(\boldsymbol{W}_r = \boldsymbol{W}^\star) \geqslant 1 - 2N^2 e^{-2r\kappa_h^2/R^4}. \tag{63}$$

Proof of Lemma D.2. For each $u \in \mathbb{N}$, let $\mathcal{H}_{u-1}$ contain the complete history immediately before probe u, including all training rounds since probe $u-1$. Let $\mathcal{H}_u$ additionally contain probe u and all subsequent training rounds before probe $u+1$. These sigma-fields are nested, $X^u_{i,j}$ is $\mathcal{H}_u$-measurable, and the selected probe model $\boldsymbol{\phi}_u$ is $\mathcal{H}_{u-1}$-measurable. For a fixed pair $i, j \in [N]$ and probe $u \in [r]$, define

$$X^u_{i,j} := \operatorname{dist}(\widetilde{\boldsymbol{\phi}}^u_i, \widetilde{\boldsymbol{\phi}}^u_j) = \|\widetilde{\boldsymbol{\phi}}^u_i - \widetilde{\boldsymbol{\phi}}^u_j\|^2, \qquad M^u_{i,j} := X^u_{i,j} - \mathbb{E}[X^u_{i,j} \mid \mathcal{H}_{u-1}]. \tag{64}$$

Both adapted probe models lie in $\mathcal{W}$, whose diameter is R. Consequently, $0 \leqslant X^u_{i,j} \leqslant R^2$ almost surely. Moreover, $\mathbb{E}[M^u_{i,j} \mid \mathcal{H}_{u-1}] = 0$, so $\{M^u_{i,j}\}_{u=1}^r$ is a martingale difference sequence. Conditionally on $\mathcal{H}_{u-1}$, subtracting the conditional mean only translates the range, so its length remains at most R^2.

We include the concentration calculation because the probes need not be independent across u: the common model $\boldsymbol{\phi}_u$ may depend on all earlier training and probe outcomes. Conditional Hoeffding's lemma gives, for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda M^u_{i,j}} \,\middle|\, \mathcal{H}_{u-1}\right] \leqslant e^{\lambda^2 R^4/8}. \tag{65}$$

Applying the tower property successively for $u = r, r-1, \ldots, 1$ yields

$$\mathbb{E}\left[e^{\lambda \sum_{u=1}^r M^u_{i,j}}\right] \leqslant e^{r\lambda^2 R^4/8}. \tag{66}$$

For $\lambda > 0$, Markov's inequality and Eq. (66) imply

$$\Pr\left(\sum_{u=1}^r M^u_{i,j} \geqslant rx\right) \leqslant e^{-\lambda rx + r\lambda^2 R^4/8}. \tag{67}$$

The right-hand side is minimized at $\lambda = 4x/R^4$, giving e^{-2rx^2/R^4}. Repeating the calculation for $-M^u_{i,j}$ and applying a union bound to the two tails gives, for every $x > 0$,

$$\Pr\left(\left|\frac{1}{r}\sum_{u=1}^r X^u_{i,j} - \frac{1}{r}\sum_{u=1}^r \mathbb{E}[X^u_{i,j} \mid \mathcal{H}_{u-1}]\right| \geqslant x\right) \leqslant 2e^{-2rx^2/R^4}. \tag{68}$$

If $z^\star_i = z^\star_j$, Lemma D.1 implies

$$\frac{1}{r}\sum_{u=1}^r \mathbb{E}[X^u_{i,j} \mid \mathcal{H}_{u-1}] \leqslant (2\epsilon + 2r_{\mathrm{p}})^2. \tag{69}$$

If $z^\star_i \neq z^\star_j$, the same lemma gives

$$\frac{1}{r}\sum_{u=1}^r \mathbb{E}[X^u_{i,j} \mid \mathcal{H}_{u-1}] \geqslant (\Delta - 2\epsilon - 2r_{\mathrm{p}})^2. \tag{70}$$

Set $x = \kappa_h$ in Eq. (68). On the complement of this deviation event, a within-cluster pair satisfies

$$d_{i,j}(r) = \frac{1}{r}\sum_{u=1}^r X^u_{i,j} \leqslant \frac{1}{r}\sum_{u=1}^r \mathbb{E}[X^u_{i,j} \mid \mathcal{H}_{u-1}] + \kappa_h \leqslant (2\epsilon + 2r_{\mathrm{p}})^2 + \kappa_h \leqslant h^2, \tag{71}$$

where the last inequality follows from the definition of κ_h. Similarly, a cross-cluster pair satisfies

$$d_{i,j}(r) \geqslant \frac{1}{r}\sum_{u=1}^r \mathbb{E}[X^u_{i,j} \mid \mathcal{H}_{u-1}] - \kappa_h \geqslant (\Delta - 2\epsilon - 2r_{\mathrm{p}})^2 - \kappa_h \geqslant h^2. \tag{72}$$

Thus, $d_{i,j}(r) < h^2$ holds exactly for within-cluster pairs on the simultaneous concentration event.

There are $N(N-1)/2$ unordered client pairs. Applying a union bound to Eq. (68) shows that all pairwise decisions are correct with probability at least

$$1 - N(N-1)e^{-2r\kappa_h^2/R^4} \geqslant 1 - 2N^2 e^{-2r\kappa_h^2/R^4}. \tag{73}$$

On this event, the definition of $\boldsymbol{W}_r$ in the algorithm section and Eq. (34) give $\boldsymbol{W}_r = \boldsymbol{W}^\star$, which proves the claim. □

Lemma D.2 shows why repeated probes correct the client partition without using its current labels. The empirical pairwise distances concentrate around separated conditional means, and the probability of any incorrect affinity entry decreases exponentially in the number of probes.

Lemma D.3 (Recovery by spectral clustering). *Suppose every ground-truth cluster contains at least two clients. If $\boldsymbol{W}_r = \boldsymbol{W}^\star$, then $\boldsymbol{L}_r$ has $K^\star$ zero eigenvalues and $N - K^\star$ positive eigenvalues. For cluster k, the positive eigenvalue $|\mathcal{C}_k^\star|/(|\mathcal{C}_k^\star| - 1)$ has multiplicity $|\mathcal{C}_k^\star| - 1$. The largest eigengap is uniquely attained at $K^\star$, and Eqs. (6) and (7) recover $K^\star$ and $\mathcal{C}^\star$ up to label permutation.*

Proof of Lemma D.3. When $\boldsymbol{W}_r = \boldsymbol{W}^\star$, the affinity matrix is block diagonal with one complete block for each ground-truth cluster. Write $n_k := |\mathcal{C}_k^\star|$. Within cluster k, the affinity and degree blocks are

$$\boldsymbol{W}_k^\star = \frac{1}{N}(\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\mathsf{T} - \boldsymbol{I}_{n_k}), \qquad \boldsymbol{D}_k^\star = \frac{n_k - 1}{N}\boldsymbol{I}_{n_k}. \tag{74}$$

Since $n_k \geqslant 2$, the degree block is invertible. The corresponding normalized-Laplacian block is

$$\boldsymbol{L}_k^\star = \frac{n_k}{n_k - 1}\boldsymbol{I}_{n_k} - \frac{1}{n_k - 1}\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\mathsf{T}. \tag{75}$$

The vector $\mathbf{1}_{n_k}$ is an eigenvector of Eq. (75) with eigenvalue zero. Every vector orthogonal to $\mathbf{1}_{n_k}$ is an eigenvector with eigenvalue $n_k/(n_k - 1)$, and this orthogonal subspace has dimension $n_k - 1$. Combining all blocks gives exactly $K^\star$ zero eigenvalues and $N - K^\star$ positive eigenvalues.

Each positive eigenvalue $n_k/(n_k - 1)$ lies in $(1, 2]$. Therefore, the gap between the last zero eigenvalue and the first positive eigenvalue is strictly greater than one. The difference between any two positive eigenvalues is strictly smaller than one. All gaps among the zero eigenvalues are zero. It follows that the largest consecutive eigengap is unique and occurs between eigenvalues $K^\star$ and $K^\star + 1$, so Eq. (6) returns $K_{t_r} = K^\star$. Equivalently,

$$\lambda_{K^\star+1}(\boldsymbol{L}_r) - \lambda_{K^\star}(\boldsymbol{L}_r) > 1, \qquad \lambda_{j+1}(\boldsymbol{L}_r) - \lambda_j(\boldsymbol{L}_r) < 1 \quad \big(j \in [N-1] \setminus \{K^\star\}\big). \tag{76}$$

Let $\boldsymbol{Z} \in \{0, 1\}^{N \times K^\star}$ be the membership matrix. Every orthonormal basis of the zero eigenspace has the form

$$\boldsymbol{Q}^\star = \boldsymbol{Z}\,\mathrm{diag}(n_1, \ldots, n_{K^\star})^{-1/2}\boldsymbol{O}, \tag{77}$$

where $\boldsymbol{O}$ is orthogonal. After row normalization, clients in the same ground-truth cluster k all have a representation equal to the k-th row of $\boldsymbol{O}$. Hence, two clients in the same cluster have spectral distance zero, whereas two clients in different clusters have distance $\sqrt{2}$. If $\boldsymbol{q}_i^\star$ denotes the normalized row for client i, then

$$\|\boldsymbol{q}_i^\star - \boldsymbol{q}_j^\star\| = \begin{cases} 0, & z_i^\star = z_j^\star, \\ \sqrt{2}, & z_i^\star \neq z_j^\star. \end{cases} \tag{78}$$

After representatives from $q < K^\star$ different ground-truth clusters have been selected, every client in one of those clusters has distance zero from a selected representative. Every client in an unselected cluster has distance $\sqrt{2}$ from every selected representative. The farthest-first rule must therefore select a client from an unselected cluster. By induction, the first $K^\star$ representatives contain exactly one client from each ground-truth cluster. The nearest-representative assignment in Eq. (7) then assigns every client to the representative of its own cluster. This recovers $\mathcal{C}^\star$ up to the arbitrary ordering of the representatives. □

Lemma D.4 (Cluster-model contraction under the correct partition). *Suppose Assumption 3.1, 3.2 and 3.3 hold. Let $\mathcal{E}_t$ hold and let the cluster labels be aligned with the ground-truth labels. If $0 < \eta_t \leqslant \min\{1/L, \mu/L^2\}$, then the models produced by the projected cluster-model update in Algorithm 1 satisfy*

$$\mathbb{E}\left[\sum_{k=1}^{K^\star} P_k \|\widetilde{\boldsymbol{\theta}}_k^{t+1} - \boldsymbol{\theta}_k^\star\|^2 \,\middle|\, \mathcal{F}_t\right] \leqslant (1 - \mu\eta_t)e_t^{\mathrm{m}} + \eta_t^2\sigma^2. \tag{79}$$

Proof of Lemma D.4. On $\mathcal{E}_t$, the aggregated stochastic gradient for ground-truth cluster k is

$$\mathbf{g}_k^t = \sum_{i \in \mathcal{C}_k^\star} \frac{p_i}{P_k} \mathbf{g}_i(\boldsymbol{\theta}_k^t; \xi_i^t). \tag{80}$$

It is conditionally unbiased for $\nabla F_k(\boldsymbol{\theta}_k^t)$. Jensen's inequality and Assumption 3.3 give

$$\begin{aligned} \left\| \mathbf{g}_k^t - \nabla F_k(\boldsymbol{\theta}_k^t) \right\|^2 &= \left\| \sum_{i \in \mathcal{C}_k^\star} \frac{p_i}{P_k} \left(\mathbf{g}_i(\boldsymbol{\theta}_k^t; \xi_i^t) - \nabla f_i(\boldsymbol{\theta}_k^t) \right) \right\|^2 \\ &\leqslant \sum_{i \in \mathcal{C}_k^\star} \frac{p_i}{P_k} \left\| \mathbf{g}_i(\boldsymbol{\theta}_k^t; \xi_i^t) - \nabla f_i(\boldsymbol{\theta}_k^t) \right\|^2. \end{aligned} \tag{81}$$

Taking the conditional expectation shows that

$$\mathbb{E}\left[\|\mathbf{g}_k^t - \nabla F_k(\boldsymbol{\theta}_k^t)\|^2 \,\middle|\, \mathcal{F}_t \right] \leqslant \sigma^2. \tag{82}$$

Both strong convexity and smoothness are preserved by positive weighted averages, so F_k is μ-strongly convex and L-smooth. The first-order optimality condition on the convex set $\mathcal{W}$ gives $\boldsymbol{\theta}_k^\star = \Pi_{\mathcal{W}}(\boldsymbol{\theta}_k^\star - \eta_t \nabla F_k(\boldsymbol{\theta}_k^\star))$. Nonexpansiveness of projection and conditional unbiasedness therefore yield

$$\begin{aligned} &\mathbb{E}[\|\widetilde{\boldsymbol{\theta}}_k^{t+1} - \boldsymbol{\theta}_k^\star\|^2 \mid \mathcal{F}_t] \\ &\quad \leqslant \|\boldsymbol{\theta}_k^t - \boldsymbol{\theta}_k^\star\|^2 - 2\eta_t \langle \boldsymbol{\theta}_k^t - \boldsymbol{\theta}_k^\star, \nabla F_k(\boldsymbol{\theta}_k^t) - \nabla F_k(\boldsymbol{\theta}_k^\star) \rangle \\ &\qquad + \eta_t^2 \|\nabla F_k(\boldsymbol{\theta}_k^t) - \nabla F_k(\boldsymbol{\theta}_k^\star)\|^2 + \eta_t^2 \sigma^2. \end{aligned} \tag{83}$$

Strong convexity and smoothness give

$$\langle \boldsymbol{\theta}_k^t - \boldsymbol{\theta}_k^\star, \nabla F_k(\boldsymbol{\theta}_k^t) - \nabla F_k(\boldsymbol{\theta}_k^\star) \rangle \geqslant \mu \|\boldsymbol{\theta}_k^t - \boldsymbol{\theta}_k^\star\|^2, \qquad \|\nabla F_k(\boldsymbol{\theta}_k^t) - \nabla F_k(\boldsymbol{\theta}_k^\star)\| \leqslant L \|\boldsymbol{\theta}_k^t - \boldsymbol{\theta}_k^\star\|. \tag{84}$$

Substitution into Eq. (83) yields

$$\begin{aligned} \mathbb{E}[\|\widetilde{\boldsymbol{\theta}}_k^{t+1} - \boldsymbol{\theta}_k^\star\|^2 \mid \mathcal{F}_t] &\leqslant (1 - 2\mu\eta_t + L^2\eta_t^2) \|\boldsymbol{\theta}_k^t - \boldsymbol{\theta}_k^\star\|^2 + \eta_t^2 \sigma^2 \\ &\leqslant (1 - \mu\eta_t) \|\boldsymbol{\theta}_k^t - \boldsymbol{\theta}_k^\star\|^2 + \eta_t^2 \sigma^2, \end{aligned} \tag{85}$$

where the last inequality uses $\eta_t \leqslant \mu / L^2$. Multiplying by P_k, summing over k, and using $\sum_k P_k = 1$ proves Eq. (79). □

Lemma D.5 (One-round cluster-model recursion). *Suppose Assumption 3.1, 3.2 and 3.3 hold and $0 < \eta_t \leqslant \min\{1/L, \mu/L^2\}$. Suppose $\Pr(\mathcal{E}_t^c) \leqslant q_t$ and $\Pr(\mathcal{E}_{t+1}^c) \leqslant q_{t+1}$, where $\mathcal{E}_t^c$ is the complement of $\mathcal{E}_t$. Then*

$$\mathbb{E}[e_{t+1}^{\mathrm{m}}] \leqslant (1 - \mu\eta_t) \mathbb{E}[e_t^{\mathrm{m}}] + \eta_t^2 \sigma^2 + R^2 (q_t + q_{t+1}). \tag{86}$$

Proof of Lemma D.5. Let

$$u_{t+1} := \sum_{k=1}^{K^\star} P_k \|\widetilde{\boldsymbol{\theta}}_k^{t+1} - \boldsymbol{\theta}_k^\star\|^2 \tag{87}$$

when $\mathcal{E}_t$ holds and the labels are aligned. If $\mathcal{E}_t$ does not hold, set $u_{t+1} = 0$. If both $\mathcal{E}_t$ and $\mathcal{E}_{t+1}$ hold and $t+1$ is a probe round, the overlap weights satisfy

$$\omega^{t+1}_{\pi_{t+1}(g), \pi_t(g')} = P_g \mathbf{1}\{g = g'\}, \qquad g, g' \in [K^\star]. \tag{88}$$

Thus, $\boldsymbol{\theta}^{t+1}_{\pi_{t+1}(g)} = \widetilde{\boldsymbol{\theta}}^{t+1}_{\pi_t(g)}$. The same identity follows directly from the algorithm at a non-probe round. Therefore, $e_{t+1}^{\mathrm{m}} = u_{t+1}$. If either event fails, compactness of $\mathcal{W}$ still bounds the error: every model update is projected onto $\mathcal{W}$, and every transferred model is a convex combination because

$$\frac{\omega^{t+1}_{\ell k}}{\sum_{k'} \omega^{t+1}_{\ell k'}} \geqslant 0, \qquad \sum_k \frac{\omega^{t+1}_{\ell k}}{\sum_{k'} \omega^{t+1}_{\ell k'}} = 1. \tag{89}$$

The denominator is positive because every new cluster is nonempty and $p_i > 0$. Hence, the definition of e^{m}_{t+1} in Section 4 gives $0 \leqslant e^{\mathrm{m}}_{t+1} \leqslant R^2$. Therefore,

$$e^{\mathrm{m}}_{t+1} \leqslant u_{t+1}\mathbf{1}\{\mathcal{E}_t\} + R^2\mathbf{1}\{\mathcal{E}_t^c \cup \mathcal{E}_{t+1}^c\}. \tag{90}$$

Take expectations in Eq. (90). Since $\mathcal{E}_t \in \mathcal{F}_t$, the tower property and Lemma D.4 give

$$\begin{aligned}\mathbb{E}[u_{t+1}\mathbf{1}\{\mathcal{E}_t\}] &= \mathbb{E}[\mathbf{1}\{\mathcal{E}_t\}\mathbb{E}[u_{t+1} \mid \mathcal{F}_t]] \\ &\leqslant (1-\mu\eta_t)\mathbb{E}[e^{\mathrm{m}}_t\mathbf{1}\{\mathcal{E}_t\}] + \eta_t^2\sigma^2\Pr(\mathcal{E}_t) \\ &\leqslant (1-\mu\eta_t)\mathbb{E}[e^{\mathrm{m}}_t] + \eta_t^2\sigma^2. \end{aligned} \tag{91}$$

Finally, $\Pr(\mathcal{E}_t^c \cup \mathcal{E}_{t+1}^c) \leqslant q_t + q_{t+1}$ by the union bound. Substitution into Eq. (90) proves Eq. (86). □

Lemma D.6 (Vanishing-error recursion). *Let $x_t \geqslant 0$ satisfy*

$$x_{t+1} \leqslant \left(1 - \frac{a}{t+t_0}\right)x_t + \frac{c}{(t+t_0)^2} + u_t, \tag{92}$$

where $a > 1$, $c \geqslant 0$, $t_0 \geqslant a$, and there exist $C_u \geqslant 0$ and $\beta \in (0,1)$ such that $0 \leqslant u_t \leqslant C_u\beta^t$. Then $x_t \to 0$.

Proof of Lemma D.6. Repeated substitution gives

$$x_t \leqslant x_0\prod_{q=0}^{t-1}\left(1-\frac{a}{q+t_0}\right) + \sum_{s=0}^{t-1}\left(\frac{c}{(s+t_0)^2} + u_s\right)\prod_{q=s+1}^{t-1}\left(1-\frac{a}{q+t_0}\right). \tag{93}$$

Since $t_0 \geqslant a$, every factor in the products belongs to $[0,1]$. Using $1-x \leqslant e^{-x}$ and comparing the harmonic sum with an integral,

$$\prod_{q=s+1}^{t-1}\left(1-\frac{a}{q+t_0}\right) \leqslant e^{-a\sum_{q=s+1}^{t-1}1/(q+t_0)} \leqslant \left(\frac{s+t_0+1}{t+t_0}\right)^a. \tag{94}$$

The same argument with the product starting at $q = 0$ bounds the initial term by a constant multiple of $(t+t_0)^{-a}$.

For the deterministic quadratic term, Eq. (94) gives

$$\sum_{s=0}^{t-1}\frac{c}{(s+t_0)^2}\prod_{q=s+1}^{t-1}\left(1-\frac{a}{q+t_0}\right) \leqslant \frac{c}{(t+t_0)^a}\sum_{s=0}^{t-1}\frac{(s+t_0+1)^a}{(s+t_0)^2} \leqslant \frac{C_c}{t+t_0}, \tag{95}$$

for a finite constant C_c depending only on a, c, t_0. The last step follows from the integral bound $\sum_{s=0}^{t-1}(s+t_0)^{a-2} = \mathcal{O}((t+t_0)^{a-1})$, which is valid because $a > 1$.

For the exponentially decreasing term,

$$\sum_{s=0}^{t-1}u_s\prod_{q=s+1}^{t-1}\left(1-\frac{a}{q+t_0}\right) \leqslant \frac{C_u}{(t+t_0)^a}\sum_{s=0}^{t-1}\beta^s(s+t_0+1)^a \leqslant \frac{C_\beta}{(t+t_0)^a}, \tag{96}$$

where $C_\beta < \infty$ because a geometrically weighted polynomial series converges. Each term on the right-hand side of Eq. (93) now tends to zero. This proves $x_t \to 0$ and also shows that the two forcing terms contribute $\mathcal{O}(1/t)$ and $\mathcal{O}(t^{-a})$, respectively. □

D.3 Proof of Theorem 4.2

Theorem D.7 (Cluster Structure Recovery). *Suppose Assumption 3.1, 3.2, 3.3 and Definition 4.1 hold, all the probe samples are drawn independently at each probe, and $0 < \alpha \leqslant \mu/L^2$. Denote $r_{\mathrm{p}} := (e^{-\mu\alpha E}R^2 + \alpha\sigma^2/(B\mu))^{1/2}$. Suppose $\Delta > 4(\epsilon + r_{\mathrm{p}})$, choose h such that $2(\epsilon + r_{\mathrm{p}}) < h < \Delta - 2(\epsilon + r_{\mathrm{p}})$, and denote $\kappa_h := \min\{h^2 - (2\epsilon + 2r_{\mathrm{p}})^2, (\Delta - 2\epsilon - 2r_{\mathrm{p}})^2 - h^2\}$. Then the following result holds for Algorithm 1:*

$$\Pr(\mathcal{E}_{t_r}) \geqslant 1 - 2N^2e^{-2r\kappa_h^2/R^4}. \tag{97}$$

For any $\delta \in (0,1)$, we have $\Pr(\mathcal{E}_t) \geqslant 1-\delta, \forall t \geqslant Hr_\delta$, where $r_\delta := \max\left\{1, \left\lceil \frac{R^4}{2\kappa_h^2}\log\left(\frac{2N^2}{\delta}\right)\right\rceil\right\}$.

Proof of Theorem 4.2. By Lemmas D.1 and D.2,

$$\Pr(\boldsymbol{W}_r = \boldsymbol{W}^\star) \geqslant 1 - 2N^2 e^{-2r\kappa_h^2/R^4} \tag{98}$$

for every $r \in \mathbb{N}$. On the event $\{\boldsymbol{W}_r = \boldsymbol{W}^\star\}$, Lemma D.3 shows that the eigengap estimator returns $K^\star$ and the spectral assignment returns $\mathcal{C}^\star$ up to label permutation. In set notation,

$$\{\boldsymbol{W}_r = \boldsymbol{W}^\star\} \subseteq \mathcal{E}_{t_r}. \tag{99}$$

Taking probabilities and using Eq. (63) therefore gives

$$\Pr(\mathcal{E}_{t_r}) \geqslant 1 - 2N^2 e^{-2r\kappa_h^2/R^4}. \tag{100}$$

This argument uses the common probe responses and does not use the initial number of clusters, client partition, or cluster-model values.

By the definition of r_δ in Theorem 4.2, every $r \geqslant r_\delta$ satisfies

$$\frac{2r\kappa_h^2}{R^4} \geqslant \log\left(\frac{2N^2}{\delta}\right). \tag{101}$$

Exponentiating the negative of both sides and multiplying by $2N^2$ gives

$$2N^2 e^{-2r\kappa_h^2/R^4} \leqslant \delta. \tag{102}$$

The client partition does not change between two consecutive probe rounds. For every $t \geqslant Hr_\delta$, its most recent probe index is $r(t) = \lfloor t/H \rfloor \geqslant r_\delta$. Applying the preceding bound at that probe proves $\Pr(\mathcal{E}_t) \geqslant 1 - \delta$. □

D.4 Proof of Theorem 4.3

Theorem D.8 (Cluster and Model Convergence with a Constant Learning Rate). *Suppose Assumption 3.1, 3.2, 3.3 and Definition 4.1 hold, all the probe samples are drawn independently at each probe, and $0 < \alpha \leqslant \mu/L^2$. Denote $r_{\mathrm{p}} := (e^{-\mu\alpha E}R^2 + \alpha\sigma^2/(B\mu))^{1/2}$. Suppose $\Delta > 4(\epsilon + r_{\mathrm{p}})$, choose h such that $2(\epsilon + r_{\mathrm{p}}) < h < \Delta - 2(\epsilon + r_{\mathrm{p}})$, and denote $\kappa_h := \min\{h^2 - (2\epsilon + 2r_{\mathrm{p}})^2, (\Delta - 2\epsilon - 2r_{\mathrm{p}})^2 - h^2\}$. Let $r(t) := \lfloor t/H \rfloor$ be the number of completed probes by round t, let $0 < \eta \leqslant \min\{1/L, \mu/L^2\}$ and set $\eta_t = \eta$. Denote $\rho := 1 - \mu\eta$. Then, for every $t \geqslant 1$, the following result holds for Algorithm 1:*

$$\mathbb{E}[e_t^{\mathrm{c}}] \leqslant \min\left\{1, 2N^2 e^{-2r(t)\kappa_h^2/R^4}\right\}, \tag{103}$$

$$\mathbb{E}[e_t^{\mathrm{m}}] \leqslant \rho^t R^2 + \frac{\eta\sigma^2}{\mu}\left(1 - \rho^t\right) + 4N^2R^2 \sum_{\tau=0}^{t-1} \rho^{t-1-\tau} e^{-2r(\tau)\kappa_h^2/R^4}. \tag{104}$$

Consequently, $\limsup_{t\to\infty} \mathbb{E}[e_t^{\mathrm{c}}] = 0$, *and* $\limsup_{t\to\infty} \mathbb{E}[e_t^{\mathrm{m}}] \leqslant \frac{\eta\sigma^2}{\mu}$.

Proof of Theorem 4.3. The most recent probe by round t has index $r(t)$. The client partition is unchanged after that probe. The trivial probability bound before the first probe and Lemmas D.2 and D.3 thereafter give

$$\Pr(\mathcal{E}_t^c) \leqslant \min\left\{1, 2N^2 e^{-2r(t)\kappa_h^2/R^4}\right\}. \tag{105}$$

Since $0 \leqslant e_t^{\mathrm{c}} \leqslant 1$ and $e_t^{\mathrm{c}} = 0$ on $\mathcal{E}_t$,

$$\mathbb{E}[e_t^{\mathrm{c}}] \leqslant \Pr(\mathcal{E}_t^c) \leqslant \min\left\{1, 2N^2 e^{-2r(t)\kappa_h^2/R^4}\right\}, \tag{106}$$

which proves Eq. (103).

Apply Lemma D.5 with the preceding failure-probability bounds at rounds t and $t+1$, and set $\eta_t = \eta$. This gives

$$\mathbb{E}[e_{t+1}^{\mathrm{m}}] \leqslant \rho\mathbb{E}[e_t^{\mathrm{m}}] + \eta^2\sigma^2 + 4N^2R^2 e^{-2r(t)\kappa_h^2/R^4}, \tag{107}$$

where we use $r(t+1) \geqslant r(t)$ and $\min\{1, x\} \leqslant x$. Repeated substitution and $e_0^{\mathrm{m}} \leqslant R^2$ give

$$\mathbb{E}[e_t^{\mathrm{m}}] \leqslant \rho^t R^2 + \eta^2\sigma^2 \sum_{\tau=0}^{t-1} \rho^{t-1-\tau} + 4N^2R^2 \sum_{\tau=0}^{t-1} \rho^{t-1-\tau} e^{-2r(\tau)\kappa_h^2/R^4}. \tag{108}$$

Evaluating the geometric sum in the second term gives Eq. (104).

For any fixed integer $s < t$, the last sum in Eq. (108) satisfies

$$\sum_{\tau=0}^{t-1} \rho^{t-1-\tau} e^{-2r(\tau)\kappa_h^2/R^4} \leqslant \rho^{t-s} \sum_{\tau=0}^{s-1} \rho^{s-1-\tau} + \frac{e^{-2r(s)\kappa_h^2/R^4}}{1-\rho}. \tag{109}$$

First let $t \to \infty$ and then $s \to \infty$; the two terms on the right vanish in this order. Finally,

$$\eta^2\sigma^2 \frac{1}{1-\rho} = \frac{\eta\sigma^2}{\mu}, \tag{110}$$

which proves the claimed limits. □

D.5 Proof of Theorem 4.4

Theorem D.9 (Cluster and Model Convergence with a Diminishing Learning Rate). *Suppose Assumption 3.1, 3.2, 3.3 and Definition 4.1 hold, all the probe samples are drawn independently at each probe, and* $0 < \alpha \leqslant \mu/L^2$. *Denote* $r_{\mathrm{p}} := (e^{-\mu\alpha E}R^2 + \alpha\sigma^2/(B\mu))^{1/2}$. *Suppose* $\Delta > 4(\epsilon + r_{\mathrm{p}})$, *choose* h *such that* $2(\epsilon + r_{\mathrm{p}}) < h < \Delta - 2(\epsilon + r_{\mathrm{p}})$, *and denote* $\kappa_h := \min\{h^2 - (2\epsilon + 2r_{\mathrm{p}})^2, (\Delta - 2\epsilon - 2r_{\mathrm{p}})^2 - h^2\}$. *Let* $\eta_t = \frac{c_\eta}{t+t_0}$ *where* $\mu c_\eta > 1$ *and* $t_0 \geqslant \max\{c_\eta L, c_\eta L^2/\mu\}$. *Denote the exact recovery time as* $T_{\mathrm{exact}} := \inf\{T \in \mathbb{N}_0 : \mathcal{E}_t \text{ holds}, \forall t \geqslant T\}$. *For any* $\delta \in (0,1)$, *denote*

$$\bar{r}_\delta := \max\left\{1, \left\lceil \frac{R^4}{2\kappa_h^2} \log\left(\frac{2N^2}{\delta\left(1 - e^{-2\kappa_h^2/R^4}\right)}\right)\right\rceil\right\}. \tag{111}$$

Then the result $\Pr(T_{\mathrm{exact}} \leqslant H\bar{r}_\delta) \geqslant 1 - \delta$ *holds for Algorithm 1, and* $\limsup_{t\to\infty} \mathbb{E}[e_t^{\mathrm{m}}] = 0$.

Proof of Theorem 4.4. Let $r(t) := \lfloor t/H \rfloor$ be the number of completed probes by round t, and let

$$c_h := \frac{2\kappa_h^2}{R^4}. \tag{112}$$

By Lemmas D.2 and D.3, the failure probability at probe r is at most $2N^2e^{-c_h r}$ for every $r \in \mathbb{N}$. Hence,

$$\sum_{r=1}^{\infty} \Pr(\mathcal{E}_{t_r}^c) \leqslant 2N^2 \sum_{r=1}^{\infty} e^{-c_h r} < \infty. \tag{113}$$

The first Borel–Cantelli lemma, which does not require independence of the events, implies

$$\Pr(\mathcal{E}_{t_r}^c \text{ infinitely often}) = 0. \tag{114}$$

Thus, with probability one, there is a finite random index r_{last} such that every probe with $r > r_{\mathrm{last}}$ is correct. The probe at $r_{\mathrm{last}} + 1$ establishes the ground-truth client partition, and the algorithm does not change that partition before the next probe. Since every later probe is also correct, one may take $T_{\mathrm{exact}} \leqslant H(r_{\mathrm{last}} + 1) < \infty$ on this probability-one event. If no probe fails, set $r_{\mathrm{last}} = 0$. This proves both almost-sure statements.

For a deterministic tail bound, the union bound and the geometric series give

$$\Pr\left(\bigcup_{r=q}^{\infty} \mathcal{E}_{t_r}^c\right) \leqslant \frac{2N^2e^{-c_h q}}{1 - e^{-c_h}}, \qquad q \in \mathbb{N}. \tag{115}$$

By the definition of $\bar{r}_\delta$,

$$c_h\bar{r}_\delta \geqslant \log\left(\frac{2N^2}{\delta(1 - e^{-c_h})}\right). \tag{116}$$

Substitution into Eq. (115) with $q = \bar{r}_\delta$ makes its right-hand side at most δ. If no probe fails from $\bar{r}_\delta$ onward, the client partition is correct for every $t \geqslant H\bar{r}_\delta$. This proves the high-probability bound for T_{exact}.

For every round t, the latest completed probe has index $r(t)$. The trivial bound before the first probe and the probe-recovery bound thereafter give

$$\Pr(\mathcal{E}_t^c) \leqslant \min\{1, 2N^2 e^{-c_h r(t)}\}. \tag{117}$$

It remains to prove model convergence in expectation. Apply Lemma D.5 with Eq. (117) at rounds t and $t+1$:

$$\mathbb{E}[e_{t+1}^{\mathrm{m}}] \leqslant \left(1 - \frac{\mu c_\eta}{t+t_0}\right)\mathbb{E}[e_t^{\mathrm{m}}] + \frac{c_\eta^2\sigma^2}{(t+t_0)^2} + R^2\left(\min\{1, 2N^2 e^{-c_h r(t)}\} + \min\{1, 2N^2 e^{-c_h r(t+1)}\}\right). \tag{118}$$

For $t \geqslant H$, $r(t) \geqslant t/H - 1$ and $r(t+1) \geqslant r(t)$. Hence, the last line of Eq. (118) satisfies

$$R^2\left(\min\{1, 2N^2 e^{-c_h r(t)}\} + \min\{1, 2N^2 e^{-c_h r(t+1)}\}\right) \leqslant 4N^2R^2 e^{c_h} e^{-c_h t/H}. \tag{119}$$

Moreover, $\mu c_\eta > 1$. The condition on t_0 gives

$$\eta_t = \frac{c_\eta}{t+t_0} \leqslant \min\left\{\frac{1}{L}, \frac{\mu}{L^2}\right\} \tag{120}$$

for every $t \in \mathbb{N}_0$, and also $t_0 \geqslant c_\eta L \geqslant \mu c_\eta$ because $\mu \leqslant L$. Apply Lemma D.6 to the recursion from $t = H$ onward with

$$x_t = \mathbb{E}[e_t^{\mathrm{m}}], \qquad a = \mu c_\eta, \qquad c = c_\eta^2\sigma^2, \qquad u_t = 4N^2R^2 e^{c_h} e^{-c_h t/H}. \tag{121}$$

It follows that $\mathbb{E}[e_t^{\mathrm{m}}] \to 0$. □

D.6 Deterministic Recovery

Proposition D.10 (Model convergence without gradient noise). *Suppose Assumption 3.1, 3.2, 3.3 and Definition 4.1 hold, all the probe samples are drawn independently at each probe, $\sigma = 0$, and $0 < \alpha \leqslant \mu/L^2$. Denote $r_{\mathrm{p}} := e^{-\mu\alpha E/2}R$. Suppose $\Delta > 4(\epsilon + r_{\mathrm{p}})$ and choose h such that $2(\epsilon + r_{\mathrm{p}}) < h < \Delta - 2(\epsilon + r_{\mathrm{p}})$. Then $\mathcal{E}_t$ holds for every $t \geqslant H$, and if $0 < \eta \leqslant \min\{1/L, \mu/L^2\}$ is constant, then the following result holds for Algorithm 1:*

$$e_t^{\mathrm{m}} \leqslant (1 - \mu\eta)^{t-H} R^2, \qquad t \geqslant H. \tag{122}$$

Proof of Proposition D.10. When $\sigma = 0$, Eq. (42) gives $\mathbb{E}[\|\varepsilon_i^{r,e}\|^2 \mid \mathcal{H}_{r,e}] = 0$, and hence $\varepsilon_i^{r,e} = 0$ almost surely. Therefore,

$$\|\widetilde{\boldsymbol{\phi}}_i^r - \boldsymbol{w}_i^\star\|^2 \leqslant \rho_{\mathrm{p}}^{2E}\|\boldsymbol{\phi}_r - \boldsymbol{w}_i^\star\|^2 \leqslant e^{-\mu\alpha E}R^2 = r_{\mathrm{p}}^2. \tag{123}$$

Consequently, at every probe $r \in \mathbb{N}$,

$$\begin{aligned} z_i^\star = z_j^\star &\Longrightarrow \|\widetilde{\boldsymbol{\phi}}_i^r - \widetilde{\boldsymbol{\phi}}_j^r\| \leqslant 2\epsilon + 2r_{\mathrm{p}} < h, \\ z_i^\star \neq z_j^\star &\Longrightarrow \|\widetilde{\boldsymbol{\phi}}_i^r - \widetilde{\boldsymbol{\phi}}_j^r\| \geqslant \Delta - 2\epsilon - 2r_{\mathrm{p}} > h. \end{aligned} \tag{124}$$

Thus, every term accumulated in $d_{i,j}(r)$ is on the correct side of h^2, so $\boldsymbol{W}_r = \boldsymbol{W}^\star$ almost surely. Lemma D.3 implies that the first probe, performed at round H, recovers $K^\star$ and $\mathcal{C}^\star$. Every later probe does the same, and the partition remains unchanged between probes. Thus, $\mathcal{E}_t$ holds for all $t \geqslant H$.

At round H, every transferred model is feasible, so $e_H^{\mathrm{m}} \leqslant R^2$. With the correct client partition and $\sigma = 0$, Lemma D.4 gives

$$e_{t+1}^{\mathrm{m}} \leqslant (1 - \mu\eta)e_t^{\mathrm{m}}. \tag{125}$$

Hence, for every $t \geqslant H$,

$$e_t^{\mathrm{m}} \leqslant (1 - \mu\eta)^{t-H} e_H^{\mathrm{m}} \leqslant (1 - \mu\eta)^{t-H} R^2, \tag{126}$$

which proves Eq. (122). □

D.7 Adaptive Threshold Selection

We now give a sufficient condition under which the two-means rule used in the experiments selects a threshold in the theoretical recovery interval from the observed probe distances.

Theorem D.11 (Adaptive affinity threshold)**.** *Suppose Assumption 3.1, 3.2, 3.3, and Definition 4.1 hold, all the probe samples are drawn independently at each probe, $K^\star \geqslant 2$ and $0 < \alpha \leqslant \mu/L^2$. Write $r_{\mathrm{p}} \coloneqq (e^{-\mu\alpha E}R^2 + \alpha\sigma^2/(B\mu))^{1/2}$. At probe r, let $m_{i,j}(r) \coloneqq r^{-1}\sum_{u=1}^{r}\mathbb{E}[\mathrm{dist}(\widetilde{\boldsymbol{\phi}}_i^u, \widetilde{\boldsymbol{\phi}}_j^u) \mid \mathcal{H}_{u-1}]$. Suppose there are constants $m_{\mathrm{in}} < m_{\mathrm{out}}$ and $\omega \geqslant 0$ such that, almost surely, $|m_{i,j}(r) - m_{\mathrm{in}}| \leqslant \omega$ for every within-cluster pair and $|m_{i,j}(r) - m_{\mathrm{out}}| \leqslant \omega$ for every cross-cluster pair. For $\delta \in (0,1)$, set $s_r \coloneqq \omega + R^2\sqrt{\log(2N^2/\delta)/(2r)}$. If*

$$m_{\mathrm{out}} - m_{\mathrm{in}} > (2+\sqrt{N(N-1)})s_r, \qquad (2\epsilon+2r_{\mathrm{p}})^2 + s_r < \frac{m_{\mathrm{in}}+m_{\mathrm{out}}}{2} < (\Delta-2\epsilon-2r_{\mathrm{p}})^2 - s_r, \tag{127}$$

then exact one-dimensional two-means clustering of $\{d_{i,j}(r) : i < j\}$ yields centers $a_r < b_r$ whose midpoint $h_r^2 = (a_r + b_r)/2$ satisfies $2(\epsilon + r_{\mathrm{p}}) < h_r < \Delta - 2(\epsilon + r_{\mathrm{p}})$ with probability at least $1-\delta$. On the same event, thresholding at h_r^2 gives $\boldsymbol{W}_r = \boldsymbol{W}^\star$, and the spectral step recovers $K^\star$ and $\mathcal{C}^\star$ up to label permutation.

Proof. For any client pair, Eq. (68) bounds the deviation of $d_{i,j}(r)$ from $m_{i,j}(r)$. A union bound over at most $N(N-1)/2$ pairs shows that, with probability at least $1-\delta$, all deviations are smaller than $R^2\sqrt{\log(2N^2/\delta)/(2r)}$. On this event every within-cluster distance belongs to $[m_{\mathrm{in}} - s_r, m_{\mathrm{in}} + s_r]$ and every cross-cluster distance belongs to $[m_{\mathrm{out}} - s_r, m_{\mathrm{out}} + s_r]$.

There are $M \coloneqq N(N-1)/2$ distances. Their true two-group split has one-dimensional two-means objective at most Ms_r^2, since choosing m_{in} and m_{out} as centers already attains this bound. Any different two-group split puts at least one within-cluster distance and one cross-cluster distance in the same group. For two such values, the sum of squared deviations from any common center is at least $(m_{\mathrm{out}} - m_{\mathrm{in}} - 2s_r)^2/2$. The first inequality in Eq. (127) makes this strictly larger than Ms_r^2. Thus, the exact two-means solution separates the two types of pairwise distances. Its centers satisfy $|a_r - m_{\mathrm{in}}| \leqslant s_r$ and $|b_r - m_{\mathrm{out}}| \leqslant s_r$. The second inequality in Eq. (127) therefore places h_r^2 strictly between $(2\epsilon + 2r_{\mathrm{p}})^2$ and $(\Delta - 2\epsilon - 2r_{\mathrm{p}})^2$. The first inequality also gives $m_{\mathrm{out}} - m_{\mathrm{in}} > 4s_r$, so $m_{\mathrm{in}} + s_r < h_r^2 < m_{\mathrm{out}} - s_r$. All within-cluster distances therefore lie below h_r^2 and all cross-cluster distances lie above it, giving $\boldsymbol{W}_r = \boldsymbol{W}^\star$. Lemma D.3 then gives the stated recovery. □

Theorem D.11 shows that, under the conditional-mean concentration and separation conditions, including Eq. (127), the data-driven threshold h_r lies in the theoretical recovery interval $2(\epsilon + r_{\mathrm{p}}) < h_r < \Delta - 2(\epsilon + r_{\mathrm{p}})$ with probability at least $1 - \delta$. On the same event, it produces exactly the target affinity matrix $\boldsymbol{W}^\star$ and recovers the true number of clusters and the client partition. Thus, the adaptive rule supplies a threshold satisfying the recovery requirement from the observed probe distances. The additional conditions in Theorem D.11 describe the spread of the within-cluster and cross-cluster response distances. They are compatible with cluster centers whose pairwise distances are approximately equal, small within-cluster perturbations, and sufficiently accurate probe adaptation. For example, in the noiseless limit with identical client optima within each cluster and cluster centers at the vertices of a regular simplex, exact probe responses give $m_{\mathrm{in}} = 0$, $m_{\mathrm{out}} = \Delta^2$, and $\omega = 0$. For sufficiently small r_{p} and sufficiently many probes, Eq. (127) then holds. More generally, if both inequalities in Eq. (127) hold strictly with $s_r = \omega$, they also hold for sufficiently large r, since the sampling term in s_r vanishes. Thus, repeated probes reduce sampling uncertainty.