FEDSPM: ROUTING-ENABLED FEDERATED LEARNING UNDER DUAL HETEROGENEITY VIA SEMIPARAMETRIC MIXTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Routing-prediction federated learning has emerged as a new paradigm that reframes inter-client heterogeneity as a resource for system-level intelligence: at inference time, the server routes each external query to the best-matched client for prediction. Existing approaches, however, typically treat each client as internally homogeneous, overlooking latent subpopulations within local data. For example, patients with the same diagnosis at one hospital may exhibit morphologically distinct disease subtypes. The coexistence of inter-client and intra-client heterogeneity, which we call *dual heterogeneity*, can impair both routing and prediction. To address this challenge, we propose `FedSPM`, a routing-enabled semiparametric mixture framework that represents each client using client-specific latent components. Each component combines a predictive distribution for classification with a feature distribution for routing. To flexibly model feature distributions while effectively sharing information across clients, `FedSPM` models their density ratios relative to a common nonparametric distribution estimated via empirical likelihood. We develop a federated expectation-maximization algorithm that optimizes a tractable surrogate and prove convergence of the exact profiled objective at the standard $\mathcal{O}(1/\sqrt{T})$ rate when the surrogate errors are properly controlled. Experiments on controlled benchmarks and real-world medical data demonstrate consistent improvements in routing and prediction under dual heterogeneity. Code is available at `https://anonymous.4open.science/r/FedSPM`.

1 INTRODUCTION

Federated learning (FL) enables multiple clients to collaboratively train a model without centralizing their raw data. In a typical FL system, a server distributes a shared model to participating clients, each client updates the model using its local data, and the server aggregates the resulting locally updated models into an improved global model. This distributed training paradigm allows knowledge to be shared across clients such as hospitals (Xu et al., 2021) and mobile devices (Li et al., 2020a), making FL particularly promising when data are sensitive, geographically dispersed, or impractical to centralize (Li et al., 2020b; Kairouz et al., 2021; Yang et al., 2019).

Traditionally, *inter-client heterogeneity*, where data distributions vary across clients, is viewed as an obstacle in FL. Since clients optimize different local objectives, their gradients may drift in conflicting directions (Karimireddy et al., 2020), thereby slowing convergence (Li et al., 2020d) and degrading global model performance (Zhao et al., 2018). In contrast, the recent *routing-prediction* FL paradigm (Wang et al., 2026) reframes inter-client heterogeneity as a useful signal of client specialization. Like personalized FL (Li et al., 2021; Arivazhagan et al., 2019), it learns specialized models for individual client domains. Beyond personalization, it further estimates how well an external query matches each client's data distribution. At inference time, the server uses these distributional match scores to route the query to the most suitable client, whose specialized model then makes the final prediction, thereby turning client-specific expertise into system-level intelligence.

However, such a routing-prediction FL paradigm typically assumes that the data within each client are drawn from a homogeneous distribution, thereby overlooking *intra-client heterogeneity*. This assumption is often violated in practice, as local data may arise from a mixture of latent compo-

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.

nents (Marfoq et al., 2021; Wu et al., 2023). For example, within one hospital, dermoscopic images may involve different lesion types, anatomical sites, and patient age groups: cases with the same diagnosis may exhibit markedly different visual patterns (Zalaudek et al., 2006), whereas visually similar cases may correspond to different diagnoses (Braga et al., 2008; Carrera et al., 2017). Since these factors and their interactions are rarely fully observed or annotated, the resulting subgroup memberships are latent, making it infeasible to fit a separate model to each predefined group. Ignoring such latent structure forces a homogeneous local model to fit a mixture of heterogeneous feature and predictive distributions, thereby degrading both routing and prediction accuracy. Taken together, intra-client and inter-client heterogeneity constitute what we call *dual heterogeneity*. This raises our key question: *how can routing and prediction be jointly improved under dual heterogeneity?*

No existing heterogeneous FL approach fully addresses this problem. Methods centered on a global model mitigate client drift through regularization (Li et al., 2020c; Acar et al., 2021), aggregation reweighting (Wang et al., 2020; Li et al., 2023), or refined optimization (Hsu et al., 2019; Reddi et al., 2021), but their shared predictor still struggles to adapt to client-specific distributions. Conventional personalized methods, including local fine-tuning (Wang et al., 2019), regularization (Dinh et al., 2020; Li et al., 2021), client clustering (Ghosh et al., 2020; Sattler et al., 2021), and representation learning (Arivazhagan et al., 2019; Oh et al., 2022), improve prediction within each client's local domain, but provide no mechanism for server-side routing of external queries. Routing-based personalization (Wang et al., 2026) enables server-side routing, yet treats each client as internally homogeneous and overlooks intra-client heterogeneity. Mixture-model-based personalization captures intra-client heterogeneity (Marfoq et al., 2021; Wu et al., 2023), but its assumption of shared component distributions across clients limits flexibility in modeling client-specific latent structures.

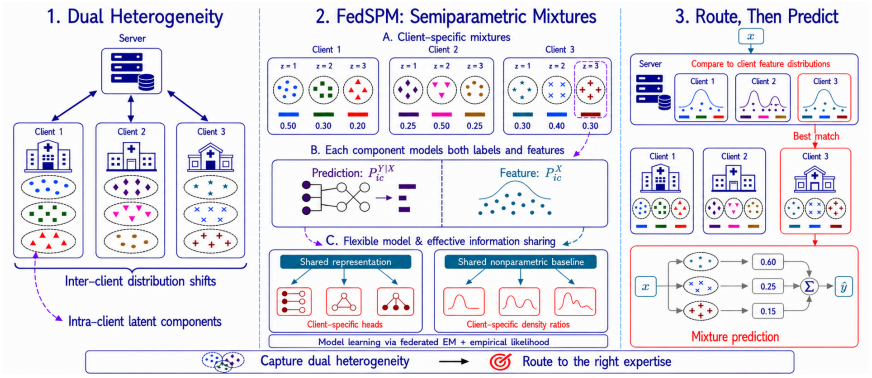

Figure 1: **Overview of the `FedSPM` framework.**

To address this challenge, we propose `FedSPM` (Fig. 1), a routing-enabled FL framework that represents each client as a mixture of latent components rather than a single homogeneous population. This mixture discovers hidden variation within each client, while allowing its components to differ across clients captures inter-client distribution shifts. Each component contains two complementary parts: a predictive distribution that relates features to labels and a feature distribution that characterizes the inputs covered by the component. To balance model flexibility with effective information sharing, the predictive distributions combine shared representations with client-specific prediction heads, while the feature distributions use a density ratio model (DRM) (Anderson, 1979) relative to a shared nonparametric baseline. We learn the model using empirical likelihood (EL) (Owen, 1990) and a federated expectation-maximization (EM) algorithm. At inference time, the server routes an external query to the most suitable client using the learned feature distributions, and the selected client combines its component-wise predictions to produce the final output. Our main contributions are summarized as follows:

- We formulate routing-prediction FL under dual heterogeneity and propose `FedSPM`, a semiparametric mixture framework that flexibly captures latent variation within clients and distribution shifts across clients while enabling effective information sharing.
- We develop a federated EM algorithm based on EL and establish convergence guarantees for the resulting nonconvex optimization under local stochastic gradient descent (SGD) with momentum.
- We evaluate `FedSPM` on controlled benchmarks and a real-world medical dataset, demonstrating consistent improvements in both routing and prediction over competitive FL baselines.

2 METHOD

Problem Formulation. Consider an FL system with m clients for K-class classification. Each client $i \in [m] := \{1, \ldots, m\}$ has a local dataset $\mathcal{D}_i := \{(x_{ij}, y_{ij})\}_{j \in [n_i]}$ with samples drawn independently from a client-specific distribution $P_i^{X,Y}$. Let $n := \sum_{i=1}^m n_i$ be the total sample size, $\rho_i := n_i/n$ be the sample fraction of client i, and $\mathcal{D} := \bigcup_{i=1}^m \mathcal{D}_i$ be the pooled dataset.

2.1 SEMIPARAMETRIC MIXTURE MODEL

To capture *intra-client heterogeneity*, we augment each client distribution $P_i^{X,Y}$ with a latent component $Z \in [C]$, yielding $P_i^{X,Y,Z}$. Specifically, we assume

$$(x_{ij}, y_{ij}) \mid (z_{ij} = c) \sim P_{ic}^{X,Y},\ z_{ij} \sim P_i^Z \text{ with } P_i^Z(\{c\}) = \pi_{ic},\ \forall c \in [C],\ j \in [n_i]. \tag{1}$$

Here, $\pi_i := (\pi_{i1}, \ldots, \pi_{iC})$ with $\pi_{ic} \geq 0$ and $\sum_c \pi_{ic} = 1$. Marginalizing over Z gives $P_i^{X,Y} = \sum_c \pi_{ic} P_{ic}^{X,Y}$. Notably, we allow both the mixing weights π_i and the component distributions $P_{ic}^{X,Y}$ to vary across clients, providing a flexible model of intra-client heterogeneity.

While allowing fully client-specific $P_{ic}^{X,Y}$ is expressive, it prevents effective information sharing across clients. We factor $P_{ic}^{X,Y}$ into its predictive distribution $P_{ic}^{Y|X}$ and feature distribution P_{ic}^X, and impose structure on both factors to share information while modeling *inter-client heterogeneity*:

Concept shift. We model the component predictive distribution of client i as

$$P_{ic}^{Y|X}(\{k\} \mid x) \propto \exp(\alpha_{ikc} + \beta_{ikc}^\top g_{\theta_c}(x)), \tag{2}$$

where $g_{\theta_c}(x)$ is a *shared* embedding for component c. The client-specific parameters $(\alpha_{ikc}, \beta_{ikc})$ allow the marginal predictive distribution $P_i^{Y|X}$ to vary across clients, hence capturing concept shift.

Covariate shift. We model the component feature distribution of client i via a DRM with respect to a shared baseline distribution G:

$$dP_{ic}^X / dG(x) = \exp\left(\gamma_{ic} + \xi_{ic}^\top h_{\nu_c}(x)\right), \tag{3}$$

where dP_{ic}^X / dG is the Radon–Nikodym derivative of P_{ic}^X relative to G, and h_{ν_c} is a *shared* basis for component c. The client-specific tilting parameters (γ_{ic}, ξ_{ic}) induce differences in the marginal feature distribution P_i^X across clients, capturing covariate shift while preserving a shared structure.

Label shift. Finally, differences in the mixing weights π_i and the component distribution $P_{ic}^{X,Y}$ jointly induce variation in the marginal label distribution P_i^Y across clients, capturing label shift.

The baseline distribution G in equation 3 remains unspecified. Restricting G to a parametric family imposes an unjustified distributional assumption, which limits model flexibility and increases the risk of misspecification. Instead, we estimate G via EL, a nonparametric likelihood framework that assigns unknown probability masses to the observed samples, subject to the DRM constraints (see App. A for related work on DRM and EL). Specifically, G is represented as a discrete distribution supported on the pooled observations, taking the form $G = \sum_{i,j} r_{ij} \delta_{x_{ij}}$ with unknown $r_{ij} \geq 0$. Crucially, *all samples across all clients jointly determine* G, serving as a key mechanism for information sharing. To ensure that G and each P_{ic}^X define valid probability distributions, the masses $\{r_{ij}\}$ must satisfy $\int G(dx) = 1$ and $\int \exp\left(\gamma_{i'c} + \xi_{i'c}^\top h_{\nu_c}(x)\right) G(dx) = 1$ for all i', c, *i.e.,*

$$\sum_{i,j} r_{ij} = 1, \quad \sum_{i,j} \exp\left(\gamma_{i'c} + \xi_{i'c}^\top h_{\nu_c}(x_{ij})\right) r_{ij} = 1, \quad \forall i' \in [m], c \in [C]. \tag{4}$$

In summary, the predictive model in equation 2, the DRM-EL feature model in equation 3, and the mixing weights π_i together define our *semiparametric mixture model*.

Remark 1 (Identifiability and interpretability)**.** Our model is intended as a flexible approximation to the client-specific joint distribution $P_i^{X,Y}$, rather than as a tool for recovering identifiable or interpretable latent structures. Its non-identifiability arises from both the neural network implementations of the representation maps g_{θ_c} and h_{ν_c}, and the mixture structure itself. For the latter, if two components on client i collapse, *i.e.,* $P_{ic_1}^{X,Y} = P_{ic_2}^{X,Y}$ for some $c_1 \neq c_2$, then redistributing mass between π_{ic_1} and π_{ic_2} leaves $P_i^{X,Y} = \sum_c \pi_{ic} P_{ic}^{X,Y}$ unchanged. Thus, the learned components should be viewed as auxiliary constructs for routing and prediction, not as recovered true subgroups.

Remark 2 (Comparison to existing work)**.** Prior methods (Marfoq et al., 2021; Wu et al., 2023) assume $P_{ic}^{X,Y} = P_{i'c}^{X,Y}$ for all i, i', c, *i.e.,* identical component distributions across clients. This assumption is often unrealistic. For instance, even data from clinically similar patient groups may differ substantially across hospitals due to site-specific acquisition protocols and imaging devices.

2.2 Profile Log-EL and EM Algorithm

We develop a practical learning procedure for the proposed semiparametric mixture model. The main difficulty is that both the latent component assignments and the nonparametric baseline G are unknown. We first profile out G through EL, and then exploit the latent mixture structure to derive an EM algorithm that avoids direct optimization of the resulting intractable objective.

Profile log-EL. Let $\zeta \coloneqq (\alpha, \beta, \theta, \gamma, \xi, \nu, \pi)$ be the parameters of interest. The log-EL is:

$$\ell(\zeta, G) = \sum_{i,j} \log r_{ij} + \sum_{i,j} \log \left[\sum_{c} \pi_{ic} \exp\left(\gamma_{ic} + \xi_{ic}^\top h_{\nu_c}(x_{ij})\right) P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}) \right].$$

See App. C.1 for the derivation. Since G is a nuisance parameter, we work with the profile log-EL $p\ell(\zeta) \coloneqq \sup_G \ell(\zeta, G)$, where the supremum is taken under the constraints in equation 4. By the Lagrange multiplier method, the optimal baseline weights are:

$$r_{ij}^*(\zeta) = n^{-1} \left\{ 1 + \sum_{i',c} \lambda_{i'c} \left[\exp\left(\gamma_{i'c} + \xi_{i'c}^\top h_{\nu_c}(x_{ij})\right) - 1 \right] \right\}^{-1}, \tag{5}$$

where $\{\lambda_{ic}\}$ are the solutions to

$$\sum_{i,j} \frac{\exp\left(\gamma_{i'c} + \xi_{i'c}^\top h_{\nu_c}(x_{ij})\right) - 1}{1 + \sum_{i'',c'} \lambda_{i''c'} \left[\exp\left(\gamma_{i''c'} + \xi_{i''c'}^\top h_{\nu_{c'}}(x_{ij})\right) - 1 \right]} = 0, \quad \forall i' \in [m], c \in [C]. \tag{6}$$

See App. C.2 for the derivation. Substituting equation 5 back into $\ell(\zeta, G)$ yields $p\ell(\zeta)$. However, directly maximizing $p\ell(\zeta)$ remains computationally impractical because each evaluation requires solving the nonlinear system in equation 6 for the Lagrange multipliers $\{\lambda_{ic}\}$, making end-to-end optimization with standard automatic-differentiation tools such as *PyTorch* infeasible.

EM algorithm. To address this challenge, we treat the component assignments $\{z_{ij}\}$ as missing data and employ the EM algorithm to maximize $p\ell(\zeta)$. Let $z_{ijc} \coloneqq \mathbb{1}(z_{ij} = c)$ denote the indicator that the j-th sample on the i-th client belongs to the c-th subpopulation. If $\{z_{ij}\}$ were observed, the complete-data profile log-EL would be

$$p\ell^{\mathrm{c}}(\zeta) = \sum_{i,j} \log r_{ij}^*(\zeta) + \sum_{i,j,c} z_{ijc} \left[\log P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}) + \gamma_{ic} + \xi_{ic}^\top h_{\nu_c}(x_{ij}) + \log \pi_{ic} \right].$$

See App. C.3 for the derivation. Since $\{z_{ij}\}$ are unobserved, EM alternates between estimating their posterior distributions under the current parameters and updating the parameters based on these estimates. Specifically, the two steps are:

E-step. Given $\zeta^{(t)}$, we compute the posterior responsibilities for each z_{ij} as

$$w_{ijc}^{(t)} \coloneqq \mathbb{E}\left[z_{ijc} \middle| \mathcal{D}; \zeta^{(t)} \right] \propto P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}; \zeta^{(t)}) \exp\left(\gamma_{ic}^{(t)} + (\xi_{ic}^{(t)})^\top h_{\nu_c^{(t)}}(x_{ij}) \right) \pi_{ic}^{(t)}. \tag{7}$$

The resulting profile Q-function, defined as the conditional expectation of the complete-data profile log-EL, decomposes up to an additive constant independent of ζ as

$$Q^{(t)}(\zeta) \coloneqq \mathbb{E}\left[p\ell^{\mathrm{c}}(\zeta) \middle| \mathcal{D}; \zeta^{(t)} \right] \doteq Q_1^{(t)}(\alpha, \beta, \theta) + Q_2^{(t)}(\gamma, \xi, \nu) + Q_3^{(t)}(\pi), \text{ where} \tag{8}$$

$$Q_1^{(t)}(\alpha, \beta, \theta) \coloneqq \sum_{i,j,c} w_{ijc}^{(t)} \log P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}), \quad Q_3^{(t)}(\pi) \coloneqq \sum_{i,j,c} w_{ijc}^{(t)} \log \pi_{ic}, \text{ and}$$

$$Q_2^{(t)}(\gamma, \xi, \nu) \coloneqq \sum_{i,j,c} w_{ijc}^{(t)} \left[\gamma_{ic} + \xi_{ic}^\top h_{\nu_c}(x_{ij}) \right] - \sum_{i,j} \log \left(1 + \sum_{i',c} \lambda_{i'c} \left[\exp\left(\gamma_{i'c} + \xi_{i'c}^\top h_{\nu_c}(x_{ij})\right) - 1 \right] \right).$$

M-step. We update $\zeta^{(t+1)} = \arg\max Q^{(t)}(\zeta)$. Since $Q_1^{(t)}$, $Q_2^{(t)}$, and $Q_3^{(t)}$ are separable, the three parameter blocks can be optimized independently. For convenience, we refer to (α, β, θ) as the *supervised block*, and to (γ, ξ, ν) as the *DRM block*. For the mixing weights, maximizing $Q_3^{(t)}$ under $\sum_c \pi_{ic} = 1$ yields

$$\pi_{ic}^{(t+1)} = \frac{1}{n_i} \sum_{j=1}^{n_i} w_{ijc}^{(t)}. \tag{9}$$

See App. C.4 for the EM derivation. For the supervised block, maximizing $Q_1^{(t)}$ is equivalent to minimizing a weighted cross-entropy loss, which can be directly solved via gradient-based methods: $\left(\alpha^{(t+1)}, \beta^{(t+1)}, \theta^{(t+1)}\right) = \arg\max Q_1^{(t)}(\alpha, \beta, \theta)$. For the DRM block, directly maximizing $Q_2^{(t)}$ faces the same computational bottleneck as $p\ell(\zeta)$: each evaluation still requires solving the nonlinear system for the Lagrange multipliers $\{\lambda_{ic}\}$. In contrast, the EM algorithm offers a pivotal resolution: at any critical point of $Q_2^{(t)}$, each λ_{ic} admits the analytic form $n^{-1} \sum_{j=1}^{n_i} w_{ijc}^{(t)}$ (see App. C.5), implying that every critical point of $Q_2^{(t)}$ is also that of the tractable surrogate

$$\widetilde{Q}_2^{(t)}(\gamma, \xi, \nu) \coloneqq \sum_{i,j,c} w_{ijc}^{(t)} \left[\gamma_{ic} + \xi_{ic}^\top h_{\nu_c}(x_{ij})\right] - \sum_{i,j} \log\Bigg(\sum_{i',j',c} w_{i'j'c}^{(t)} \exp\left(\gamma_{i'c} + \xi_{i'c}^\top h_{\nu_c}(x_{ij})\right)\Bigg). \tag{10}$$

We optimize $\widetilde{Q}_2^{(t)}$ instead, giving the update $\left(\gamma^{(t+1)}, \xi^{(t+1)}, \nu^{(t+1)}\right) = \arg\max \widetilde{Q}_2^{(t)}(\gamma, \xi, \nu)$.

2.3 Federated Training and Inference

Algorithm 1 in App. B presents `FedSPM`, which extends the centralized EM procedure to the client-server setting. Although the tilting parameters (γ_{ic}, ξ_{ic}) are client-specific, their full collection (γ, ξ) must be globally maintained and shared across clients due to the coupling in the second term of equation 10. Accordingly, at round t, the server broadcasts $(\theta^{(t)}, \gamma^{(t)}, \xi^{(t)}, \nu^{(t)})$ to all clients. Each client i then performs a local E-step followed by a local M-step:

Local E-step. Client i computes $\{w_{ijc}^{(t)}\}_{j\in[n_i], c\in[C]}$ via equation 7 using only its own dataset $\mathcal{D}_i$.

Local M-step. After the local E-step, client i updates the three parameter blocks separately. For $u \in \{\theta, \gamma, \xi, \nu\}$, let $u^{(t+1;i)}$ denote the full local copy of u updated by client i in round t. First, client i updates its own mixing weights π_i via equation 9. For the supervised block, client i updates $(\alpha_i^{(t+1)}, \beta_i^{(t+1)}, \theta^{(t+1;i)})$ by maximizing

$$Q_{1i}^{(t)}(\alpha_i, \beta_i, \theta) \coloneqq \sum_{j,c} w_{ijc}^{(t)} \log P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}),$$

which depends only on $\mathcal{D}_i$ and $\{w_{ijc}^{(t)}\}_{j\in[n_i], c\in[C]}$ available on client i. For the DRM block, client i updates $(\gamma^{(t+1;i)}, \xi^{(t+1;i)}, \nu^{(t+1;i)})$ by maximizing

$$\widetilde{Q}_{2i}^{(t)}(\gamma, \xi, \nu) \coloneqq \sum_{j,c} w_{ijc}^{(t)} \left[\gamma_{ic} + \xi_{ic}^\top h_{\nu_c}(x_{ij})\right] - \sum_{j} \log\Bigg(\sum_{i',c} \tau_{i'c}^{(t)} \exp\left(\gamma_{i'c} + \xi_{i'c}^\top h_{\nu_c}(x_{ij})\right)\Bigg),$$

where $\tau_{i'c}^{(t)} \coloneqq \sum_{j'=1}^{n_{i'}} w_{i'j'c}^{(t)}$ denotes the total responsibility assigned to component c on client i'. Unlike $Q_{1i}^{(t)}$, client i cannot evaluate $\widetilde{Q}_{2i}^{(t)}$ using only its local information, because $\widetilde{Q}_{2i}^{(t)}$ involves responsibilities from other clients. However, this extra information is needed only in the form of summary statistics $\{\tau_{i'c}^{(t)}\}_{i'\in[m], c\in[C]}$. Accordingly, after the local E-step, each client i transmits $\{\tau_{ic}^{(t)}\}_{c\in[C]}$ to the server, and the server broadcasts the collection $\{\tau_{ic}^{(t)}\}_{i\in[m], c\in[C]}$ to all clients. This additional information comprises only $m \times C$ scalars, which is negligible relative to model transmission and does not reveal raw data or per-sample responsibilities. Armed with these summary statistics, client i can maximize both $Q_{1i}^{(t)}$ and $\widetilde{Q}_{2i}^{(t)}$ via any gradient-based `LocalSolver`.

Routing and prediction. At inference time, given an external query x, the server first routes it to the most suitable client using maximum a posteriori estimation, with the sample fraction ρ_i as the prior and the client feature density relative to the baseline distribution G as the likelihood:

$$i^* \coloneqq \arg\max_{i\in[m]} \left\{\rho_i \cdot dP_i^X / dG(x)\right\} = \arg\max_{i\in[m]} \left\{\rho_i \sum_c \pi_{ic} \exp\left(\gamma_{ic} + \xi_{ic}^\top h_{\nu_c}(x)\right)\right\}. \tag{11}$$

Let $P_i^{Z|X}$ denote the conditional distribution of the latent component given the query on client i. The selected client i^* then makes the final prediction:

$$\widehat{y} \coloneqq \arg\max_{k\in[K]} P_{i^*}^{Y|X}(\{k\} \mid x) = \arg\max_{k\in[K]} \left\{ \sum_c P_{i^*c}^{Y|X}(\{k\} \mid x) P_{i^*}^{Z|X}(\{c\} \mid x) \right\}. \tag{12}$$

2.4 Convergence Analysis

We analyze the convergence of `FedSPM` when `LocalSolver` is instantiated as local SGD (Stich, 2019) with momentum. Since the mixing weights and the supervised block follow standard EM/GEM updates and inherit the convergence guarantees established by Marfoq et al. (2021), we focus exclusively on the DRM block $\phi \coloneqq (\gamma, \xi, \nu)$. For convenience, we cast the optimization of ϕ as a standard minimization problem using sample-averaged objectives:

$$F(\zeta) \coloneqq -\frac{1}{n} p\ell(\zeta), \quad f^{(t)}(\phi) \coloneqq -\frac{1}{n} Q_2^{(t)}(\phi), \quad \widetilde{f}^{(t)}(\phi) \coloneqq -\frac{1}{n} \widetilde{Q}_2^{(t)}(\phi), \quad \widetilde{f}_i^{(t)}(\phi) \coloneqq -\frac{1}{n_i} \widetilde{Q}_{2i}^{(t)}(\phi).$$

Assume that F is bounded below by F^*. In communication round t, each client i performs

$$\phi_{i,0}^{(t)} \coloneqq \phi^{(t)}, \qquad d_{i,-1}^{(t)} \coloneqq 0, \qquad d_{i,e}^{(t)} \coloneqq \mu d_{i,e-1}^{(t)} + \widetilde{g}_{i,e}^{(t)}, \qquad \phi_{i,e+1}^{(t)} \coloneqq \phi_{i,e}^{(t)} - \eta d_{i,e}^{(t)},$$

for $e \in \{0, \dots, E-1\}$, where $\mu \in [0, 1)$ denotes the momentum, E denotes the number of local steps, η denotes the learning rate, and $\widetilde{g}_{i,e}^{(t)}$ denotes the stochastic gradient.

Assumption 1 (Smoothness)**.** For all i and t, $\widetilde{f}_i^{(t)}$ is L-smooth.

Assumption 2 (Stochastic gradient)**.** For all i, t, and e, $\widetilde{g}_{i,e}^{(t)}$ is unbiased with bounded variance σ^2, *i.e.,* $\mathbb{E}\big[\widetilde{g}_{i,e}^{(t)} \mid \phi_{i,e}^{(t)}\big] = \nabla \widetilde{f}_i^{(t)}(\phi_{i,e}^{(t)})$ and $\mathbb{E}\big[\big\|\widetilde{g}_{i,e}^{(t)} - \nabla \widetilde{f}_i^{(t)}(\phi_{i,e}^{(t)})\big\|^2 \mid \phi_{i,e}^{(t)}\big] \le \sigma^2$.

Assumption 3 (Inter-client heterogeneity)**.** There exist constants $\Gamma_0, \Gamma_1 \ge 0$ such that, for all t, $\sum_{i=1}^m \rho_i \big\|\nabla \widetilde{f}_i^{(t)}(\phi^{(t)})\big\|^2 \le \Gamma_0 + \Gamma_1 \big\|\nabla \widetilde{f}^{(t)}(\phi^{(t)})\big\|^2$.

Assumptions 1–3 are standard in federated optimization (Wang et al., 2020; Marfoq et al., 2021).

Assumption 4 (Gradient bridge)**.** There exist a constant $\Gamma_2 \ge 0$ and a non-negative sequence $\{\varepsilon^{(t)}\}_{t\ge 0}$ such that, for all t, $\mathbb{E}\big[\big\|\nabla f^{(t)}(\phi^{(t)})\big\|^2\big] \le \Gamma_2 \mathbb{E}\big[\big\|\nabla \widetilde{f}^{(t)}(\phi^{(t)})\big\|^2\big] + \varepsilon^{(t)}$.

Assumption 5 (Function-value bridge)**.** There exists a non-negative sequence $\{\delta^{(t)}\}_{t\ge 0}$ such that, for all t, $\mathbb{E}\big[f^{(t)}(\phi^{(t+1)}) - f^{(t)}(\phi^{(t)})\big] \le \mathbb{E}\big[\widetilde{f}^{(t)}(\phi^{(t+1)}) - \widetilde{f}^{(t)}(\phi^{(t)})\big] + \delta^{(t)}$.

Assumption 4 requires that, along the iterates, the gradient norm of the DRM objective $f^{(t)}$ is bounded by that of its tractable surrogate $\widetilde{f}^{(t)}$ up to an error $\varepsilon^{(t)}$, without imposing any alignment between their gradient directions, while Assumption 5 quantifies the one-step mismatch between their function-value changes. We further provide empirical support for these assumptions in Sec. 3.2.

Theorem 1. Under Assumptions 1–5, when $\eta = \Theta(1/\sqrt{T})$ and $\eta L a_E < \min\{1/3, 1/\sqrt{3 + 8\Gamma_1}\}$, `FedSPM` under local SGD with momentum satisfies:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla_\phi F(\zeta^{(t)})\right\|^2 = \mathcal{O}\left(\frac{\Gamma_2 (F(\zeta^{(0)}) - F^*)}{a_E \sqrt{T}} + \frac{\Gamma_2}{a_E} \bar{\delta}_T + \frac{\Gamma_2 L \sigma^2 s_E}{a_E \sqrt{T}} \sum_{i=1}^m \rho_i^2 + \frac{\Gamma_2 L^2 (a_E^2 \Gamma_0 + \sigma^2 s_E)}{T} + \bar{\varepsilon}_T \right),$$

where

$$\mu_{e,s} \coloneqq \frac{1-\mu^{e-s}}{1-\mu}, \; a_E \coloneqq \sum_{e=0}^{E-1} \mu_{E,e}, \; s_E \coloneqq \sum_{e=0}^{E-1} \mu_{E,e}^2, \; \bar{\delta}_T \coloneqq \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} \delta^{(t)}, \; \bar{\varepsilon}_T \coloneqq \frac{1}{T} \sum_{t=0}^{T-1} \varepsilon^{(t)}.$$

Crucially, although `FedSPM` optimizes the tractable surrogate $\widetilde{f}^{(t)}$, Theorem 1 establishes convergence toward a stationary point of the exact profiled objective F. For fixed E, if $\sum_{t=0}^{T-1} \delta^{(t)} = \mathcal{O}(1)$ and $\sum_{t=0}^{T-1} \varepsilon^{(t)} = \mathcal{O}(\sqrt{T})$, the surrogate errors do not change the standard $\mathcal{O}(1/\sqrt{T})$ convergence rate. The bound further characterizes the effect of momentum through the accumulation factors a_E and s_E. When $\mu = 0$, $a_E = s_E = E$, recovering the usual local SGD scaling. As $\mu \to 1$, $a_E = \mathcal{O}(E^2)$ and $s_E = \mathcal{O}(E^3)$. Consequently, the initial-gap term improves from order $1/(E\sqrt{T})$ to $1/(E^2\sqrt{T})$, whereas the stochastic term grows from order $1/\sqrt{T}$ to $E/\sqrt{T}$, revealing a trade-off between faster optimization and amplified stochastic noise. The proof is deferred to App. C.6.

3 EXPERIMENTS ON BENCHMARK DATASETS

We conduct experiments on three image classification benchmarks of increasing complexity: FMNIST (Xiao et al., 2017), CIFAR-10 (Krizhevsky, 2009), and CIFAR-100 (Krizhevsky, 2009), with dataset details deferred to App. D.1. To study each heterogeneity type in isolation and assess their joint effect, we construct semi-synthetic FL settings by superimposing controlled inter-client and intra-client heterogeneity onto the original data.

3.1 TRAINING AND EVALUATION SETUP

Baselines. We compare `FedSPM` with representative FL baselines from seven categories: global-model methods (FedAvg (McMahan et al., 2017) and FedProx (Li et al., 2020c)), a fine-tuning-based method (FedAvgFT (Wang et al., 2019)), a regularization-based method (Ditto (Li et al., 2021)), a cluster-based method (ClusterFL (Ghosh et al., 2020)), a representation-learning-based method (FedBABU (Oh et al., 2022)), mixture-model-based methods (FedEM (Marfoq et al., 2021) and FedGMM (Wu et al., 2023)), and a routing-based method (FedDRM (Wang et al., 2026)). Notably, only FedGMM, FedDRM, and `FedSPM` are capable of client routing by design.

Evaluation metrics. Following Wang et al. (2026), we evaluate all methods using system and average accuracies. System accuracy is measured on the pooled test set across all clients, computed via equation 11 and equation 12 for routing-capable methods, and via majority voting otherwise. Average accuracy is the mean of local test accuracies weighted by client sample fractions.

Model architecture. We instantiate the classification encoder g_{θ_c} with ResNet (He et al., 2016) and the routing encoder h_{ν_c} with a lightweight CNN (LeCun et al., 1998). For a fair comparison, all methods use the same classification encoder architecture, and routing-capable methods use the same routing encoder architecture. Notably, the model capacity of cluster- and mixture-model-based methods scales linearly with their number of components. For `FedSPM`, we report two variants to separate algorithmic gains from increased model capacity. In `FedSPM` ($1\times$), the encoders are shared across components, *i.e.,* $g_{\theta_c} \equiv g_\theta$ and $h_{\nu_c} \equiv h_\nu$ for all c. In `FedSPM` ($C\times$), each component has its own encoders g_{θ_c} and h_{ν_c}.

Dual heterogeneity settings. Since benchmark datasets do not inherently exhibit statistical heterogeneity, we explicitly construct dual heterogeneity based on the mixture representation $P_i^{X,Y} = \sum_c \pi_{ic} P_{ic}^{X,Y}$, following common practice (Tan et al., 2023; Wu et al., 2023; Wang et al., 2026). Specifically, we first partition the full dataset across 8 clients via class-wise Dirichlet partitioning (Yurochkin et al., 2019) with concentration parameter $\alpha_{\text{inter}} = 1.0$, inducing label shift and unequal client dataset sizes. Within each client, we further divide the local dataset into 2 latent components, with the mixing weights π_i sampled from a Dirichlet distribution with $\alpha_{\text{intra}} = 2.0$. To induce component-wise covariate shift in P_{ic}^X, we apply transformations at two levels: client-level transformations, which introduce inter-client heterogeneity through combinations of color shifts and spatial intensity biases (Wang et al., 2026); and component-level transformations, which introduce intra-client heterogeneity through color-channel inversion (Wu et al., 2023). In addition, for each client-component pair, we randomly generate a label permutation (Wu et al., 2023) to induce component-wise concept shift in $P_{ic}^{Y|X}$. Since the type and degree of each heterogeneity are specified by construction, the ground-truth data-generating mechanism is fully known, making these benchmarks well-suited for controlled comparisons. See App. D.2 for illustrative visualizations.

Training details. We employ local SGD with momentum as the `LocalSolver`. For fine-tuning-based methods, we additionally perform one epoch of local fine-tuning before evaluation. For cluster- and mixture-model-based methods, we set the number of components C to 3. Additional implementation and hyperparameter details are provided in App. D.3.

3.2 EXPERIMENTAL RESULTS

Main results. Tab. 1 yields four main observations. First, `FedSPM` consistently outperforms all baselines in both metrics across all datasets, regardless of model capacity. Second, routing-free personalization methods (red background) trade system accuracy for average accuracy compared with global-model methods, revealing severe model drift. The substantial recovery in system accuracy achieved by FedDRM and `FedSPM` further underscores the necessity of client routing for

Table 1: **System and average accuracies on benchmark datasets.** Results are reported as mean ± standard deviation over 3 random seeds. Superscripts (1), (2), and (3) denote the first-, second-, and third-best results, respectively. $1\times$ and $C\times$ denote the model capacity.

Method	System Accuracy			Average Accuracy		
	FMNIST	CIFAR-10	CIFAR-100	FMNIST	CIFAR-10	CIFAR-100
FedAvg ($1\times$)	39.15 ± 1.31	35.02 ± 0.32	18.17 ± 0.53	39.15 ± 1.31	35.02 ± 0.32	18.17 ± 0.53
FedProx ($1\times$)	39.14 ± 1.32	34.94 ± 0.36	18.15 ± 0.49	39.14 ± 1.32	34.94 ± 0.36	18.15 ± 0.49
FedAvgFT ($1\times$)	35.42 ± 1.90	32.71 ± 0.22	16.29 ± 0.69	50.55 ± 1.04	41.40 ± 0.91	23.51 ± 0.54
Ditto ($1\times$)	34.30 ± 2.42	33.12 ± 0.45	16.02 ± 0.69	54.87 ± 1.18	43.91 ± 0.26	25.06 ± 0.54
ClusterFL ($C\times$)	31.36 ± 2.20	28.38 ± 2.48	14.68 ± 1.59	$55.77 \pm 2.32^{(3)}$	45.33 ± 1.20	27.31 ± 1.26
FedBABU ($1\times$)	34.27 ± 1.19	33.39 ± 0.55	16.64 ± 0.25	38.36 ± 1.76	36.90 ± 0.47	19.09 ± 0.62
FedEM ($C\times$)	29.73 ± 2.68	26.71 ± 2.07	13.69 ± 0.62	38.79 ± 0.19	38.50 ± 0.42	19.98 ± 0.44
FedGMM ($C\times$)	31.98 ± 1.99	25.36 ± 0.48	11.97 ± 0.96	38.34 ± 0.58	38.86 ± 0.42	20.70 ± 0.25
FedDRM ($1\times$)	$53.68 \pm 1.01^{(3)}$	$47.47 \pm 0.09^{(3)}$	$27.15 \pm 0.37^{(3)}$	53.69 ± 1.01	$47.62 \pm 0.11^{(3)}$	$27.55 \pm 0.37^{(3)}$
FedSPM ($1\times$)	$\mathbf{58.82 \pm 0.61}^{(1)}$	$\mathbf{48.39 \pm 0.21}^{(1)}$	$\mathbf{28.33 \pm 0.37}^{(1)}$	$\mathbf{58.86 \pm 0.62}^{(1)}$	$\mathbf{48.59 \pm 0.22}^{(1)}$	$\mathbf{28.84 \pm 0.40}^{(1)}$
FedSPM ($C\times$)	$58.23 \pm 0.17^{(2)}$	$48.03 \pm 0.58^{(2)}$	$27.54 \pm 0.45^{(2)}$	$58.32 \pm 0.18^{(2)}$	$48.32 \pm 0.61^{(2)}$	$28.07 \pm 0.46^{(2)}$

system-level performance. Third, the poor performance of mixture-model-based methods indicates that personalizing only the mixing weights is insufficient to handle severe concept shift without component-level personalization. Fourth, FedSPM ($1\times$) slightly outperforms FedSPM ($C\times$), suggesting that under fixed data budgets on these relatively simple tasks, the benefit of increased model capacity does not outweigh the sample dilution from separating encoders per component.

Impact of dual heterogeneity intensity. We further evaluate the robustness of FedSPM under different levels of inter-client and intra-client heterogeneity on FMNIST. To isolate their effects, we vary the Dirichlet concentration parameter $\alpha_{\text{inter}} \in \{0.5, 1.0, 2.0\}$ while fixing $\alpha_{\text{intra}} = 2.0$, and vary $\alpha_{\text{intra}} \in \{0.5, 1.0, 2.0\}$ while fixing $\alpha_{\text{inter}} = 1.0$. As shown in Tab. 2, FedSPM consistently achieves the best performance across all settings, demonstrating its strong adaptability to dual heterogeneity.

Table 2: **System and average accuracies under varying dual heterogeneity intensities.**

Method	Inter-Client Heterogeneity (α_{inter})						Intra-Client Heterogeneity (α_{intra})					
	System Accuracy			Average Accuracy			System Accuracy			Average Accuracy		
	0.5	1.0	2.0	0.5	1.0	2.0	0.5	1.0	2.0	0.5	1.0	2.0
FedAvg ($1\times$)	49.85	39.15	31.37	49.85	39.15	31.37	42.15	37.59	39.15	42.15	37.59	39.15
FedProx ($1\times$)	49.88	39.14	31.34	49.88	39.14	31.34	42.10	37.58	39.14	42.10	37.58	39.14
FedAvgFT ($1\times$)	45.87	35.42	27.67	62.45	50.55	45.01	36.72	33.42	35.42	54.11	50.99	50.55
Ditto ($1\times$)	45.28	34.30	25.56	65.87	54.87	49.36	34.26	31.42	34.30	58.22	55.44	54.87
ClusterFL ($C\times$)	40.77	31.36	23.96	68.80	$55.77^{(3)}$	52.41	34.79	31.22	31.36	$61.48^{(3)}$	$57.66^{(3)}$	$55.77^{(3)}$
FedBABU ($1\times$)	42.93	34.27	27.82	50.14	38.36	31.45	37.16	35.09	34.27	40.70	39.21	38.36
FedEM ($C\times$)	37.08	29.73	25.70	49.68	38.79	31.57	33.43	28.88	29.73	41.78	39.47	38.79
FedGMM ($C\times$)	34.33	31.98	24.57	49.22	38.34	31.83	28.92	27.90	31.98	44.08	38.59	38.34
FedDRM ($1\times$)	$69.03^{(3)}$	$53.68^{(3)}$	$52.51^{(3)}$	$69.05^{(3)}$	53.69	$52.51^{(3)}$	$60.25^{(3)}$	$55.52^{(3)}$	$53.68^{(3)}$	60.28	55.53	53.69
FedSPM ($1\times$)	$\mathbf{74.20}^{(1)}$	$\mathbf{58.82}^{(1)}$	$55.68^{(2)}$	$\mathbf{74.24}^{(1)}$	$\mathbf{58.86}^{(1)}$	$55.70^{(2)}$	$\mathbf{62.83}^{(1)}$	$\mathbf{63.10}^{(1)}$	$\mathbf{58.82}^{(1)}$	$\mathbf{62.89}^{(1)}$	$\mathbf{63.20}^{(1)}$	$\mathbf{58.86}^{(1)}$
FedSPM ($C\times$)	$73.15^{(2)}$	$58.23^{(2)}$	$\mathbf{56.57}^{(1)}$	$73.21^{(2)}$	$58.32^{(2)}$	$\mathbf{56.65}^{(1)}$	$62.48^{(2)}$	$61.33^{(2)}$	$58.23^{(2)}$	$62.58^{(2)}$	$61.40^{(2)}$	$58.32^{(2)}$

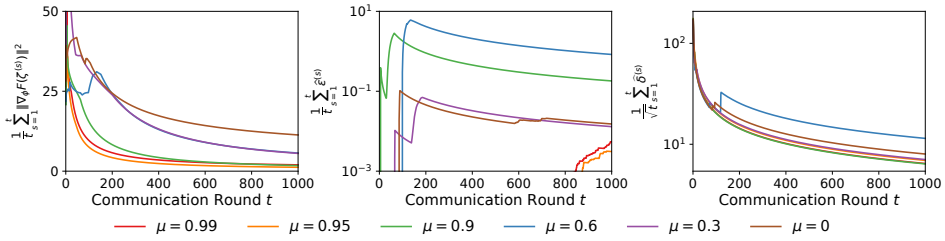

Figure 2: **Impact of momentum μ.** From left to right, the panels show the running average of $\|\nabla_\phi F(\zeta^{(t)})\|^2$ and the normalized cumulative sums of $\widehat{\varepsilon}^{(t)}$ and $\widehat{\delta}^{(t)}$.

Impact of μ. We study the impact of momentum μ under a constant learning rate on FMNIST. To align the empirical evaluation with our theory, we track the running average of $\|\nabla_\phi F(\zeta^{(t)})\|^2$, together with the normalized cumulative sums of the estimated bridge errors $\widehat{\delta}^{(t)}$ and $\widehat{\varepsilon}^{(t)}$. See detailed estimation procedure in App. D.4. Since computing these quantities requires evaluating the profiled objective F and the DRM objective $f^{(t)}$, both of which depend on the Lagrange multipliers $\{\lambda_{ic}\}$, we numerically solve equation 6 at each communication round. Fig. 2 yields three observations. First, the averaged gradient norm consistently decreases for all tested values of μ,

suggesting that FedSPM empirically converges toward a stationary point of F. Second, as μ increases, the convergence first improves and then deteriorates, with the best performance achieved at $\mu = 0.95$, consistent with the theoretical trade-off. Third, the normalized cumulative bridge errors remain small or decrease during training, providing empirical support for the asymptotic behaviors $\bar{\delta}_T = o(1)$ and $\bar{\varepsilon}_T = o(1)$, as well as for the reasonableness of Assumptions 4 and 5.

Impact of C. We study the impact of the number of components C on FMNIST. To largely isolate the effect of varying C from model capacity changes, we employ a shared encoder across all components so that increasing C only adds lightweight component-specific heads. As shown in Fig. 3, increasing C from 1 to 2 yields the largest gain, supporting the benefit of mixture modeling under dual heterogeneity. Further increasing C brings smaller but consistent improvements, suggesting that mild over-specification can provide a more flexible approximation to the predictive distribution, consistent with theoretical results on over-specified mixture-of-experts models (Ho et al., 2022; Nguyen et al., 2023).

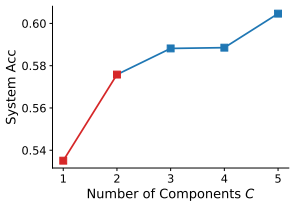

Figure 3: **Impact of C.**

4 Experiment on Real Medical Dataset

We further conduct a case study using Fed-ISIC2019 (Ogier du Terrail et al., 2022), a real-world 8-class dermoscopic image dataset collected from 6 clinical centers, each serving as one client. Unlike the controlled benchmarks, its dual heterogeneity arises naturally from differences in imaging protocols, class proportions, and pathological features. This setting therefore serves two purposes: assessing the necessity of jointly modeling dual heterogeneity in real-world FL and evaluating the effectiveness of our semiparametric model against parametric alternatives. For routing-capable methods, we additionally report *routing accuracy*, defined as the fraction of test samples routed to their source clinical center. See App. D.5 for other experimental details.

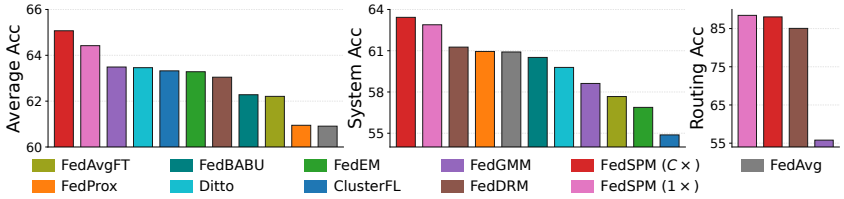

Figure 4: **Average, system, and routing accuracies on Fed-ISIC2019.**

Fig. 4 yields four main observations. First, FedSPM achieves the best performance across all three metrics, showing its practical feasibility in real-world medical scenarios. Second, in terms of average accuracy, mixture-model-based methods generally outperform those without explicit modeling of intra-client heterogeneity, validating the necessity of jointly modeling dual heterogeneity in real-world medical data. Third, FedSPM substantially improves routing accuracy over FedGMM, highlighting the limited flexibility of parametric GMMs and the benefit of our semiparametric model. Finally, in contrast to the benchmark results, FedSPM ($C\times$) outperforms FedSPM ($1\times$), suggesting that on this medically complex dataset, the benefit of increased model capacity in capturing diverse pathological features outweighs the sample dilution effect.

5 Conclusion

This paper presents FedSPM, a semiparametric mixture framework that extends the routing-prediction FL paradigm from inter-client heterogeneity to dual heterogeneity. By combining DRM-based EL with mixture modeling, FedSPM leverages latent intra-client structure to improve both routing and prediction while balancing model flexibility with effective information sharing. A federated EM procedure further enables practical optimization in the client-server setting. Empirical results on controlled benchmarks and real-world medical data show that FedSPM outperforms existing approaches in both routing and prediction performance under complex dual heterogeneity. Overall, this work paves the way for expertise-aware FL systems that transform dual heterogeneity from an obstacle into a source of system-level intelligence.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and polishing, provide feedback on mathematical arguments and experimental design, and help interpret experimental results. All AI-assisted suggestions were critically reviewed and verified by the authors, who take full responsibility for the final content of this work.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

J. A. Anderson. Multivariate logistic compounds. *Biometrika*, 66(1):17–26, 1979.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

J. C. T. Braga, A. Scope, I. Klaz, P. Mecca, P. Spencer, and A. A. Marghoob. Melanoma mimicking seborrheic keratosis: an error of perception precluding correct dermoscopic diagnosis. *Journal of the American Academy of Dermatology*, 58(5):875–880, 2008.

Song Cai, Jiahua Chen, and James V. Zidek. Hypothesis testing in the presence of multiple samples under density ratio models. *Statistica Sinica*, 27(2):761–783, 2017.

C. Carrera, S. Segura, P. Aguilera, M. Scalvenzi, C. Longo, A. Barreiro, P. Broganelli, S. Cavicchini, A. Llambrich, P. Zaballos, L. Thomas, J. Malvehy, S. Puig, and I. Zalaudek. Dermoscopic clues for diagnosing melanomas that resemble seborrheic keratosis. *JAMA Dermatology*, 153(6):544–551, 2017.

Jiahua Chen and Yukun Liu. Quantile and quantile-function estimations under density ratio model. *The Annals of Statistics*, 41(3):1669–1692, 2013.

Canh T. Dinh, Nguyen H. Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems*, 2020.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In *Advances in Neural Information Processing Systems*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Nhat Ho, Chiao-Yu Yang, and Michael I. Jordan. Convergence rates for gaussian mixtures of experts. *Journal of Machine Learning Research*, 23(323):1–81, 2022.

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

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2): 1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 2020.

Richard Kay and Sarah Little. Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika*, 74(3):495–501, 1987.

Amor Keziou and Samuela Leoni-Aubin. On empirical likelihood for semiparametric two-sample density ratio models. *Journal of Statistical Planning and Inference*, 138(4):915–928, 2008.

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

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.

Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020a.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020b.

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*, 2020c.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, 2021.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020d.

Zexi Li, Tao Lin, Xinyi Shang, and Chao Wu. Revisiting weighted aggregation in federated learning with neural networks. In *International Conference on Machine Learning*, 2023.

Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, and Richard Vidal. Federated multi-task learning under a mixture of distributions. In *Advances in Neural Information Processing Systems*, 2021.

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

Huy Nguyen, TrungTin Nguyen, and Nhat Ho. Demystifying Softmax gating function in Gaussian mixture of experts. In *Advances in Neural Information Processing Systems*, 2023.

Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, Boris Muzellec, Constantin Philippenko, Santiago Silva, Maria Teleńczuk, Shadi Albarqouni, Salman Avestimehr, Aurélien Bellet, Aymeric Dieuleveut, Martin Jaggi, Sai Praneeth Karimireddy, Marco Lorenzi, Giovanni Neglia, Marc Tommasi, and Mathieu Andreux. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. In *Advances in Neural Information Processing Systems*, 2022.

Jaehoon Oh, SangMook Kim, and Se-Young Yun. Fedbabu: Toward enhanced representation for federated image classification. In *International Conference on Learning Representations*, 2022.

A. B. Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120, 1990.

J. Qin. Empirical likelihood in biased sample problems. *The Annals of Statistics*, 21(3):1182–1196, 1993.

J. Qin. Empirical likelihood ratio based confidence intervals for mixture proportions. *The Annals of Statistics*, 27(4):1368–1384, 1999.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multi-task optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8):3710–3722, 2021.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

Yue Tan, Chen Chen, Weiming Zhuang, Xin Dong, Lingjuan Lyu, and Guodong Long. Is heterogeneity notorious? taming heterogeneity to handle test-time shift in federated learning. In *Advances in Neural Information Processing Systems*, 2023.

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*, 2020.

Kangkang Wang, Rajiv Mathews, Chloé Kiddon, Hubert Eichner, Françoise Beaufays, and Daniel Ramage. Federated Evaluation of On-device Personalization. *arXiv preprint arXiv:1910.10252*, 2019.

Zijian Wang, Xiaofei Zhang, Xin Zhang, Yukun Liu, and Qiong Zhang. Beyond aggregation: Guiding clients in heterogeneous federated learning. In *International Conference on Learning Representations*, 2026.

Yue Wu, Shuaicheng Zhang, Wenchao Yu, Yanchi Liu, Quanquan Gu, Dawei Zhou, Haifeng Chen, and Wei Cheng. Personalized federated learning under mixture of distributions. In *International Conference on Machine Learning*, 2023.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5(1):1–19, 2021.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2):1–19, 2019.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, 2019.

I. Zalaudek, S. Grinschgl, G. Argenziano, A. A. Marghoob, A. Blum, E. Richtig, I. H. Wolf, R. Fink-Puches, H. Kerl, H. P. Soyer, and R. Hofmann-Wellenhof. Age-related prevalence of dermoscopy patterns in acquired melanocytic naevi. *British Journal of Dermatology*, 154(2):299–304, 2006.

Biao Zhang. Quantile estimation under a two-sample semi-parametric model. *Bernoulli*, 6(3):491–511, 2000.

Biao Zhang. Assessing goodness-of-fit of generalized logit models based on case-control data. *Journal of Multivariate Analysis*, 82(1):17–38, 2002.

Qiong Zhang, Qinglong Tian, and Pengfei Li. Neyman-pearson multiclass classification under label noise via empirical likelihood. *arXiv preprint arXiv:2603.21623*, 2026.

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

A Related Work on DRM and EL

The DRM, first introduced by Anderson (1979), provides a statistical framework for modeling multiple related populations by formulating their densities as ratios with respect to a shared baseline distribution. It is highly flexible and subsumes commonly used parametric families such as the binomial, gamma, and normal distributions (Kay & Little, 1987). As a semiparametric model, the DRM does not impose parametric assumptions on the baseline distribution, which can instead be handled nonparametrically via EL (Owen, 1990). This integration gives rise to the DRM-based EL approach, which has garnered substantial attention over the past few decades (Qin, 1993; 1999; Zhang, 2000; 2002). Moreover, Keziou & Leoni-Aubin (2008) formally established the equivalence between the maximum DRM-based EL estimators and the corresponding dual EL estimators. Since the dual EL admits an analytical form and is computationally tractable, it significantly facilitates the implementation and application of DRM-based EL methods (Chen & Liu, 2013; Cai et al., 2017; Zhang et al., 2026).

Recently, Wang et al. (2026) first introduced DRM-based EL into FL and established a routing-prediction FL paradigm, thereby enabling explicit client routing together with personalized representation learning. However, this paradigm treats each client as internally homogeneous. Since intra-client heterogeneity often arises in real-world FL scenarios, ignoring such latent structure degrades both routing and prediction performance, motivating our semiparametric mixture extension.

B The Federated EM Algorithm

Algorithm 1: `FedSPM`

for $t = 0, 1, \ldots, T-1$ **do**
  Server broadcasts $(\theta^{(t)}, \gamma^{(t)}, \xi^{(t)}, \nu^{(t)})$ to all clients
  for *client* $i \in [m]$ *in parallel* **do**
    Compute $\{w_{ijc}^{(t)}\}_{j\in[n_i],c\in[C]}$ via equation 7
    Set $\tau_{ic}^{(t)} \leftarrow \sum_{j=1}^{n_i} w_{ijc}^{(t)}$ for all $c \in [C]$
    Send $\{\tau_{ic}^{(t)}\}_{c\in[C]}$ to the server
  end
  Server broadcasts $\{\tau_{ic}^{(t)}\}_{i\in[m],c\in[C]}$ to all clients
  for *client* $i \in [m]$ *in parallel* **do**
    Update $\{\pi_{ic}^{(t+1)}\}_{c\in[C]}$ via equation 9
    $(\alpha_i^{(t+1)}, \beta_i^{(t+1)}, \theta^{(t+1;i)}) \leftarrow \texttt{LocalSolver}(\alpha_i^{(t)}, \beta_i^{(t)}, \theta^{(t)})$
    $(\gamma^{(t+1;i)}, \xi^{(t+1;i)}, \nu^{(t+1;i)}) \leftarrow \texttt{LocalSolver}(\gamma^{(t)}, \xi^{(t)}, \nu^{(t)})$
    Client i sends $(\theta^{(t+1;i)}, \gamma^{(t+1;i)}, \xi^{(t+1;i)}, \nu^{(t+1;i)})$ to the server
  end
  Server updates
  $(\theta^{(t+1)}, \gamma^{(t+1)}, \xi^{(t+1)}, \nu^{(t+1)}) \leftarrow \sum_{i=1}^{m} \rho_i(\theta^{(t+1;i)}, \gamma^{(t+1;i)}, \xi^{(t+1;i)}, \nu^{(t+1;i)})$
end

C Mathematical Details

C.1 Derivation of the Log-EL

The observed data contain only (x_{ij}, y_{ij}), while the component assignment z_{ij} is latent. Therefore, the observed-data log-EL is obtained from the augmented distribution $P_i^{X,Y,Z}$ by marginalizing over

all possible values of Z:

$$\begin{aligned}
\ell(\zeta, G) &\coloneqq \sum_{i,j} \log \sum_{c} \underbrace{P_i^{X,Y,Z}(dx_{ij}, \{y_{ij}\}, \{c\})}_{\text{complete-data contribution for } Z=c} \\
&= \sum_{i,j} \log \sum_{c} \pi_{ic} P_{ic}^{X,Y}(dx_{ij}, \{y_{ij}\}) \\
&= \sum_{i,j} \log \sum_{c} \pi_{ic} P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}) P_{ic}^{X}(dx_{ij}) \\
&= \sum_{i,j} \log \sum_{c} \pi_{ic} P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}) \frac{dP_{ic}^{X}}{dG}(x_{ij}) G(dx_{ij}) \\
&= \sum_{i,j} \log \sum_{c} \pi_{ic} P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}) \frac{dP_{ic}^{X}}{dG}(x_{ij}) r_{ij} \\
&= \sum_{i,j} \log r_{ij} + \sum_{i,j} \log \left[\sum_{c} \pi_{ic} \exp\left(\gamma_{ic} + \xi_{ic}^{\top} h_{\nu_c}(x_{ij})\right) P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}) \right].
\end{aligned}$$

C.2 Derivation of the Optimal Baseline Weights

For fixed ζ, maximizing ℓ over G reduces to maximizing $\sum_{i,j} \log r_{ij}$ subject to the constraints in equation 4. The corresponding Lagrangian is:

$$\mathcal{L} \coloneqq \sum_{i,j} \log r_{ij} - n\kappa \left[\sum_{i,j} r_{ij} - 1 \right] - n \sum_{i',c} \lambda_{i'c} \sum_{i,j} \left[\exp\left(\gamma_{i'c} + \xi_{i'c}^{\top} h_{\nu_c}(x_{ij})\right) - 1 \right] r_{ij}.$$

Setting $\partial \mathcal{L} / \partial r_{ij} = 0$ gives

$$r_{ij}^{-1} - n\kappa - n \sum_{i',c} \lambda_{i'c} \left[\exp\left(\gamma_{i'c} + \xi_{i'c}^{\top} h_{\nu_c}(x_{ij})\right) - 1 \right] = 0.$$

Multiplying the above equation by r_{ij}, summing over (i, j), and applying the constraints in equation 4 gives $\kappa = 1$, which yields

$$r_{ij}^{*}(\zeta) = \frac{1}{n} \left\{ 1 + \sum_{i',c} \lambda_{i'c} \left[\exp\left(\gamma_{i'c} + \xi_{i'c}^{\top} h_{\nu_c}(x_{ij})\right) - 1 \right] \right\}^{-1}.$$

Substituting $r_{ij}^{*}(\zeta)$ back into equation 4 gives the following nonlinear system for $\{\lambda_{ic}\}$:

$$\sum_{i,j} \frac{\exp\left(\gamma_{i'c} + \xi_{i'c}^{\top} h_{\nu_c}(x_{ij})\right) - 1}{1 + \sum_{i'',c'} \lambda_{i''c'} \left[\exp\left(\gamma_{i''c'} + \xi_{i''c'}^{\top} h_{\nu_{c'}}(x_{ij})\right) - 1 \right]} = 0, \quad \forall i' \in [m],\ c \in [C].$$

C.3 Derivation of the Complete-Data Profile Log-EL

Unlike the observed-data log-EL, the complete-data log-EL treats the latent component indicators $\{z_{ijc}\}$ as observed. Thus, it uses the corresponding complete-data contribution from $P_i^{X,Y,Z}$ di-

rectly, without marginalizing over Z:

$$\begin{aligned}
\ell^{\mathrm{c}}(\zeta, G) &\coloneqq \sum_{i,j,c} z_{ijc} \log P_i^{X,Y,Z}(dx_{ij}, \{y_{ij}\}, \{c\}) \\
&= \sum_{i,j,c} z_{ijc} \log\Big[\pi_{ic} P_{ic}^{X,Y}(dx_{ij}, \{y_{ij}\})\Big] \\
&= \sum_{i,j,c} z_{ijc} \log\Big[\pi_{ic} P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}) P_{ic}^{X}(dx_{ij})\Big] \\
&= \sum_{i,j,c} z_{ijc} \log\Big[\pi_{ic} P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}) \frac{dP_{ic}^{X}}{dG}(x_{ij}) G(dx_{ij})\Big] \\
&= \sum_{i,j,c} z_{ijc} \log\Big[\pi_{ic} \frac{dP_{ic}^{X}}{dG}(x_{ij}) r_{ij} P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij})\Big] \\
&= \sum_{i,j} \log r_{ij} + \sum_{i,j,c} z_{ijc} \Big[\log P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}) + \gamma_{ic} + \xi_{ic}^{\top} h_{\nu_c}(x_{ij}) + \log \pi_{ic}\Big].
\end{aligned}$$

Since maximizing ℓ^{c} over G for fixed ζ also reduces to maximizing $\sum_{i,j} \log r_{ij}$ subject to the constraints in equation 4, ℓ^{c} shares the same optimal baseline weights as ℓ. Thus, substituting $r_{ij}^*(\zeta)$ back into $\ell^{\mathrm{c}}(\zeta, G)$ yields the complete-data profile log-EL:

$$\begin{aligned}
p\ell^{\mathrm{c}}(\zeta) &\coloneqq \sup_{G} \ell^{\mathrm{c}}(\zeta, G) \\
&= \sum_{i,j} \log r_{ij}^*(\zeta) + \sum_{i,j,c} z_{ijc} \Big[\log P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}) + \gamma_{ic} + \xi_{ic}^{\top} h_{\nu_c}(x_{ij}) + \log \pi_{ic}\Big].
\end{aligned}$$

C.4 EM Algorithm Details

The EM algorithm is an iterative procedure that starts from an initial estimate $\zeta^{(0)}$. At iteration t, given the current parameter estimate $\zeta^{(t)}$, the algorithm alternates between the following two steps.

E-step. The latent component assignments $\{z_{ij}\}$ are estimated by their posterior responsibilities, which serve as soft guesses computed from the observed data under the current parameter estimate:

$$w_{ijc}^{(t)} \coloneqq \mathbb{E}\left[z_{ijc} \middle| \mathcal{D}; \zeta^{(t)}\right] = P_i^{Z|X,Y}(\{c\} \mid x_{ij}, y_{ij}; \zeta^{(t)}).$$

The second equality follows because z_{ijc} is the indicator of the event $\{Z = c\}$. Applying Bayes' rule with prior $\pi_{ic}^{(t)}$ and likelihood $P_{ic}^{X,Y}(dx_{ij}, \{y_{ij}\}; \zeta^{(t)})$ gives

$$w_{ijc}^{(t)} = \frac{\pi_{ic}^{(t)} P_{ic}^{X,Y}(dx_{ij}, \{y_{ij}\}; \zeta^{(t)})}{\sum_{c'} \pi_{ic'}^{(t)} P_{ic'}^{X,Y}(dx_{ij}, \{y_{ij}\}; \zeta^{(t)})}.$$

Using the factorization of $P_{ic}^{X,Y}$ into $P_{ic}^{Y|X}$ and P_{ic}^{X} gives

$$w_{ijc}^{(t)} = \frac{\pi_{ic}^{(t)} P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}; \zeta^{(t)}) P_{ic}^{X}(dx_{ij}; \zeta^{(t)})}{\sum_{c'} \pi_{ic'}^{(t)} P_{ic'}^{Y|X}(\{y_{ij}\} \mid x_{ij}; \zeta^{(t)}) P_{ic'}^{X}(dx_{ij}; \zeta^{(t)})}.$$

Under the EL representation of G, $P_{ic}^{X}(dx_{ij}) = dP_{ic}^{X}/dG(x_{ij}) r_{ij}$. Therefore,

$$w_{ijc}^{(t)} = \frac{\pi_{ic}^{(t)} P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}; \zeta^{(t)}) \frac{dP_{ic}^{X}}{dG}(x_{ij}; \zeta^{(t)}) r_{ij}}{\sum_{c'} \pi_{ic'}^{(t)} P_{ic'}^{Y|X}(\{y_{ij}\} \mid x_{ij}; \zeta^{(t)}) \frac{dP_{ic'}^{X}}{dG}(x_{ij}; \zeta^{(t)}) r_{ij}}.$$

Finally, substituting the DRM density ratio and canceling the common factor r_{ij} yields

$$w_{ijc}^{(t)} = \frac{P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}; \zeta^{(t)}) \exp\left(\gamma_{ic}^{(t)} + (\xi_{ic}^{(t)})^{\top} h_{\nu_c^{(t)}}(x_{ij})\right) \pi_{ic}^{(t)}}{\sum_{c'} P_{ic'}^{Y|X}(\{y_{ij}\} \mid x_{ij}; \zeta^{(t)}) \exp\left(\gamma_{ic'}^{(t)} + (\xi_{ic'}^{(t)})^{\top} h_{\nu_{c'}^{(t)}}(x_{ij})\right) \pi_{ic'}^{(t)}}.$$

Then, the resulting profile Q-function is the conditional expectation of the complete-data profile log-EL with respect to $\{z_{ij}\}$. Up to an additive constant independent of ζ, it decomposes as

$$\begin{aligned} Q^{(t)}(\zeta) &\coloneqq \mathbb{E}\left[p\ell^{\mathrm{c}}(\zeta)\middle|\mathcal{D};\zeta^{(t)}\right] \\ &\doteq Q_1^{(t)}(\alpha,\beta,\theta) + Q_2^{(t)}(\gamma,\xi,\nu) + Q_3^{(t)}(\pi), \end{aligned}$$

where

$$\begin{aligned} Q_1^{(t)}(\alpha,\beta,\theta) &\coloneqq \sum_{i,j,c} w_{ijc}^{(t)} \log P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij}), \\ Q_2^{(t)}(\gamma,\xi,\nu) &\coloneqq \sum_{i,j,c} w_{ijc}^{(t)} \left[\gamma_{ic} + \xi_{ic}^\top h_{\nu_c}(x_{ij})\right] \\ &\quad - \sum_{i,j} \log\left(1 + \sum_{i',c} \lambda_{i'c}\left[\exp\left(\gamma_{i'c} + \xi_{i'c}^\top h_{\nu_c}(x_{ij})\right) - 1\right]\right), \\ Q_3^{(t)}(\pi) &\coloneqq \sum_{i,j,c} w_{ijc}^{(t)} \log \pi_{ic}. \end{aligned}$$

M-step. Instead of directly maximizing the profile log-EL, the M-step maximizes $Q^{(t)}$ with respect to ζ, *i.e.,*

$$\zeta^{(t+1)} = \arg\max Q^{(t)}(\zeta).$$

Here we focus on the analytic update of the mixing weights π, which solves

$$\begin{aligned} \max_{\pi} \quad & \sum_{ijc} w_{ijc}^{(t)} \log \pi_{ic} \\ \text{s.t.} \quad & \sum_{c} \pi_{ic} = 1, \qquad i \in [m], \\ & \pi_{ic} \geq 0, \qquad i \in [m], \quad c \in [C]. \end{aligned}$$

Using the method of Lagrange multipliers, the Lagrangian is

$$\mathcal{L}_\pi \coloneqq \sum_{ijc} w_{ijc}^{(t)} \log \pi_{ic} + \sum_i \iota_i \left(\sum_c \pi_{ic} - 1\right).$$

At the optimum, we require

$$\frac{\partial \mathcal{L}_\pi}{\partial \pi_{ic}} = 0, \quad \frac{\partial \mathcal{L}_\pi}{\partial \iota_i} = 0, \quad i \in [m], \quad c \in [C].$$

Solving these equations gives

$$\pi_{ic}^{(t+1)} = \frac{1}{n_i} \sum_{j=1}^{n_i} w_{ijc}^{(t)}.$$

C.5 Derivation of the Analytical Form of the Lagrange Multipliers

We derive the analytical form of $\{\lambda_{ic}\}$ from the first-order optimality condition of $Q_2^{(t)}$. For each $i_0 \in [m]$ and $c_0 \in [C]$, taking the derivative with respect to γ_{i_0,c_0} gives

$$\frac{\partial Q_2^{(t)}}{\partial \gamma_{i_0,c_0}} = \sum_{j=1}^{n_{i_0}} w_{i_0,j,c_0}^{(t)} - \lambda_{i_0,c_0} \sum_{i,j} \frac{\exp\left(\gamma_{i_0,c_0} + \xi_{i_0,c_0}^\top h_{\nu_{c_0}}(x_{ij})\right)}{1 + \sum_{i',c} \lambda_{i'c}\left[\exp\left(\gamma_{i'c} + \xi_{i'c}^\top h_{\nu_c}(x_{ij})\right) - 1\right]},$$

where the term involving $\partial\lambda_{i_0,c_0}/\partial\gamma_{i_0,c_0}$ vanishes due to equation 6. By equation 4 and equation 5, we further have

$$\sum_{i,j} \frac{\exp\left(\gamma_{i_0,c_0} + \xi_{i_0,c_0}^\top h_{\nu_{c_0}}(x_{ij})\right)}{1 + \sum_{i',c} \lambda_{i'c}\left[\exp\left(\gamma_{i'c} + \xi_{i'c}^\top h_{\nu_c}(x_{ij})\right) - 1\right]} = n \sum_{i,j} \exp\left(\gamma_{i_0,c_0} + \xi_{i_0,c_0}^\top h_{\nu_{c_0}}(x_{ij})\right) r_{ij}^* = n.$$

Therefore,

$$\frac{\partial Q_2^{(t)}}{\partial \gamma_{i_0,c_0}} = \sum_{j=1}^{n_{i_0}} w_{i_0,j,c_0}^{(t)} - n\lambda_{i_0,c_0}.$$

Setting the derivative to zero yields

$$\lambda_{i_0,c_0} = \frac{1}{n} \sum_{j=1}^{n_{i_0}} w_{i_0,j,c_0}^{(t)}.$$

C.6 Convergence Proofs

The proof proceeds in four steps. First, we unroll the local momentum recursion and rewrite the aggregated update as a single descent step driven by a weighted effective gradient. The weights are the momentum accumulation factors $\{\mu_{E,e}\}_{e=0}^{E-1}$, which make the dependence on μ explicit through a_E and s_E. Second, we decompose this effective gradient into the averaged exact gradient, the local-update drift, and the stochastic noise. Lemmas 5 and 6 bound the stochastic noise and drift terms, respectively. Third, Lemma 7 establishes a one-round descent estimate for the tractable surrogate $\widetilde{f}^{(t)}$. Finally, we combine Assumptions 4 and 5 with the EM bridge arguments in Lemmas 1 and 2 to transfer the descent to the profiled log-EL and telescope over the outer rounds.

C.6.1 Additional Notations

For the convergence proof, define the stochastic gradient noise as

$$\omega_{i,e}^{(t)} \coloneqq \widetilde{g}_{i,e}^{(t)} - \nabla \widetilde{f}_i^{(t)}(\phi_{i,e}^{(t)}).$$

Let $\mathcal{F}^{(-1)}$ be the trivial σ-algebra. For $t \geq 0$, define

$$\mathcal{F}^{(t)} \coloneqq \sigma\left(\omega_{i,e}^{(s)} : 0 \leq s \leq t,\ i \in [m],\ 0 \leq e \leq E-1\right).$$

For each client i and local step e, define

$$\mathcal{F}_{i,-1}^{(t)} \coloneqq \mathcal{F}^{(t-1)}, \quad \mathcal{F}_{i,e}^{(t)} \coloneqq \mathcal{F}^{(t-1)} \vee \sigma\left(\omega_{i,0}^{(t)}, \ldots, \omega_{i,e}^{(t)}\right).$$

C.6.2 Key Lemmas

Lemma 1 (EM function-value bridge)**.** At communication round t, for any ζ, we have

$$F(\zeta) - F(\zeta^{(t)}) \leq f^{(t)}(\phi) - f^{(t)}(\phi^{(t)}).$$

Proof. We first prove the evidence lower bound (ELBO) for the profiled log-EL:

$$p\ell(\zeta) = \sum_{i,j} \log r_{ij}^*(\zeta) + \sum_{i,j} \log\left[\sum_c \pi_{ic} \exp\left(\gamma_{ic} + \xi_{ic}^\top h_{\nu_c}(x_{ij})\right) P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij})\right].$$

For each sample (i,j), using $\sum_c w_{ijc}^{(t)} = 1$, Jensen's inequality gives

$$\begin{aligned} p\ell(\zeta) &= \sum_{i,j} \log r_{ij}^*(\zeta) + \sum_{i,j} \log\left[\sum_c w_{ijc}^{(t)} \frac{\pi_{ic} \exp\left(\gamma_{ic} + \xi_{ic}^\top h_{\nu_c}(x_{ij})\right) P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij})}{w_{ijc}^{(t)}}\right] \\ &\geq \sum_{i,j} \log r_{ij}^*(\zeta) + \sum_{i,j,c} w_{ijc}^{(t)} \log \frac{\pi_{ic} \exp\left(\gamma_{ic} + \xi_{ic}^\top h_{\nu_c}(x_{ij})\right) P_{ic}^{Y|X}(\{y_{ij}\} \mid x_{ij})}{w_{ijc}^{(t)}} \\ &= Q^{(t)}(\zeta) - \sum_{i,j,c} w_{ijc}^{(t)} \log w_{ijc}^{(t)}. \end{aligned}$$

By equation 7, when $\zeta = \zeta^{(t)}$, the Jensen inequality above becomes tight. Hence,

$$p\ell(\zeta^{(t)}) = Q^{(t)}(\zeta^{(t)}) - \sum_{i,j,c} w_{ijc}^{(t)} \log w_{ijc}^{(t)}.$$

Subtracting this equality from the previous lower bound gives

$$p\ell(\zeta) - p\ell(\zeta^{(t)}) \geq Q^{(t)}(\zeta) - Q^{(t)}(\zeta^{(t)}).$$

By the monotone ascent property of the standard EM/GEM updates for $Q_1^{(t)}$ and $Q_3^{(t)}$, we have

$$Q^{(t)}(\zeta) - Q^{(t)}(\zeta^{(t)}) \geq Q_2^{(t)}(\phi) - Q_2^{(t)}(\phi^{(t)}).$$

Therefore, since $F = -p\ell/n$ and $f^{(t)} = -Q_2^{(t)}/n$, we obtain

$$F(\zeta) - F(\zeta^{(t)}) \leq f^{(t)}(\phi) - f^{(t)}(\phi^{(t)}).$$

□

Lemma 2 (EM gradient bridge)**.** At communication round t, we have

$$\nabla_\phi F(\zeta^{(t)}) = \nabla f^{(t)}(\phi^{(t)}).$$

Proof. By the ELBO for the profiled log-EL established in the proof of Lemma 1, for any ζ,

$$p\ell(\zeta) - p\ell(\zeta^{(t)}) \geq Q^{(t)}(\zeta) - Q^{(t)}(\zeta^{(t)}).$$

Equivalently,

$$p\ell(\zeta) - Q^{(t)}(\zeta) \geq p\ell(\zeta^{(t)}) - Q^{(t)}(\zeta^{(t)}).$$

Therefore, $p\ell(\zeta) - Q^{(t)}(\zeta)$ attains a local minimum at $\zeta^{(t)}$. Since the involved functions are differentiable, its ϕ-gradient vanishes at $\zeta^{(t)}$, namely

$$\nabla_\phi p\ell(\zeta^{(t)}) = \nabla_\phi Q^{(t)}(\zeta^{(t)}).$$

By equation 8, only $Q_2^{(t)}$ depends on ϕ. Thus,

$$\nabla_\phi Q^{(t)}(\zeta^{(t)}) = \nabla Q_2^{(t)}(\phi^{(t)}).$$

Since $F = -p\ell/n$ and $f^{(t)} = -Q_2^{(t)}/n$, we obtain

$$\nabla_\phi F(\zeta^{(t)}) = -\frac{1}{n}\nabla_\phi p\ell(\zeta^{(t)}) = -\frac{1}{n}\nabla Q_2^{(t)}(\phi^{(t)}) = \nabla f^{(t)}(\phi^{(t)}).$$

□

Lemma 3 (Momentum unrolling identity)**.** At communication round t, we have

$$\phi^{(t+1)} = \phi^{(t)} - \eta\widetilde{g}^{(t)},$$

where

$$\widetilde{g}^{(t)} \coloneqq \sum_{i=1}^{m} \rho_i \sum_{e=0}^{E-1} \mu_{E,e}\widetilde{g}_{i,e}^{(t)}.$$

Proof. By unrolling the momentum recursion, for every $e = 0, \ldots, E-1$,

$$d_{i,e}^{(t)} = \mu d_{i,e-1}^{(t)} + \widetilde{g}_{i,e}^{(t)} = \sum_{s=0}^{e} \mu^{e-s}\widetilde{g}_{i,s}^{(t)}.$$

Therefore,

$$\begin{aligned}
\phi_{i,E}^{(t)} - \phi^{(t)} &= -\eta \sum_{e=0}^{E-1} d_{i,e}^{(t)} \\
&= -\eta \sum_{e=0}^{E-1} \sum_{s=0}^{e} \mu^{e-s}\widetilde{g}_{i,s}^{(t)} \\
&= -\eta \sum_{s=0}^{E-1} \left(\sum_{e=s}^{E-1} \mu^{e-s} \right) \widetilde{g}_{i,s}^{(t)} \\
&= -\eta \sum_{s=0}^{E-1} \mu_{E,s}\widetilde{g}_{i,s}^{(t)}.
\end{aligned}$$

Using the aggregation rule and $\sum_{i=1}^{m} \rho_i = 1$ gives

$$\begin{aligned}\phi^{(t+1)} - \phi^{(t)} &= \sum_{i=1}^{m} \rho_i \left(\phi_{i,E}^{(t)} - \phi^{(t)}\right) \\ &= -\eta \sum_{i=1}^{m} \rho_i \sum_{e=0}^{E-1} \mu_{E,e} \widetilde{g}_{i,e}^{(t)} \\ &= -\eta \widetilde{g}^{(t)}.\end{aligned}$$

□

Lemma 4 (Gradient decomposition)**.** At communication round t, we have

$$\widetilde{g}^{(t)} = a_E \nabla \widetilde{f}^{(t)}(\phi^{(t)}) + D^{(t)} + \Xi^{(t)},$$

where

$$D^{(t)} \coloneqq \sum_{i=1}^{m} \rho_i \sum_{e=0}^{E-1} \mu_{E,e} \Big[\nabla \widetilde{f}_i^{(t)}(\phi_{i,e}^{(t)}) - \nabla \widetilde{f}_i^{(t)}(\phi^{(t)})\Big],$$

and

$$\Xi^{(t)} \coloneqq \sum_{i=1}^{m} \rho_i \sum_{e=0}^{E-1} \mu_{E,e} \omega_{i,e}^{(t)}.$$

Proof. By definition,

$$\begin{aligned}\widetilde{g}^{(t)} &= \sum_{i=1}^{m} \rho_i \sum_{e=0}^{E-1} \mu_{E,e} \widetilde{g}_{i,e}^{(t)} \\ &= \sum_{i=1}^{m} \rho_i \sum_{e=0}^{E-1} \mu_{E,e} \nabla \widetilde{f}_i^{(t)}(\phi_{i,e}^{(t)}) + \Xi^{(t)}.\end{aligned}$$

Adding and subtracting $\nabla \widetilde{f}_i^{(t)}(\phi^{(t)})$ gives

$$\begin{aligned}\widetilde{g}^{(t)} &= \sum_{i=1}^{m} \rho_i \sum_{e=0}^{E-1} \mu_{E,e} \nabla \widetilde{f}_i^{(t)}(\phi^{(t)}) + D^{(t)} + \Xi^{(t)} \\ &= \left(\sum_{e=0}^{E-1} \mu_{E,e}\right) \sum_{i=1}^{m} \rho_i \nabla \widetilde{f}_i^{(t)}(\phi^{(t)}) + D^{(t)} + \Xi^{(t)} \\ &= a_E \nabla \widetilde{f}^{(t)}(\phi^{(t)}) + D^{(t)} + \Xi^{(t)}.\end{aligned}$$

□

Lemma 5 (Stochastic noise bound)**.** Under Assumption 2, at communication round t, we have

$$\mathbb{E}\left[\left\|\Xi^{(t)}\right\|^2 \,\middle|\, \mathcal{F}^{(t-1)}\right] \le \sigma^2 s_E \sum_{i=1}^{m} \rho_i^2.$$

Proof. For each client i, $\{\omega_{i,e}^{(t)}\}_{e=0}^{E-1}$ is a martingale difference sequence with

$$\mathbb{E}\left[\omega_{i,e}^{(t)} \,\middle|\, \mathcal{F}_{i,e-1}^{(t)}\right] = 0, \qquad \mathbb{E}\left[\left\|\omega_{i,e}^{(t)}\right\|^2 \,\middle|\, \mathcal{F}_{i,e-1}^{(t)}\right] \le \sigma^2.$$

Thus the within-client cross terms vanish, and

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{e=0}^{E-1} \mu_{E,e} \omega_{i,e}^{(t)}\right\|^2 \,\middle|\, \mathcal{F}^{(t-1)}\right] &= \sum_{e=0}^{E-1} \mu_{E,e}^2 \mathbb{E}\left[\left\|\omega_{i,e}^{(t)}\right\|^2 \,\middle|\, \mathcal{F}^{(t-1)}\right] \\ &\le \sigma^2 s_E.\end{aligned}$$

Since the accumulated noises are conditionally independent across clients given $\mathcal{F}^{(t-1)}$, the cross-client terms also vanish. Hence,

$$\mathbb{E}\left[\left\|\Xi^{(t)}\right\|^2 \,\middle|\, \mathcal{F}^{(t-1)}\right] = \sum_{i=1}^{m} \rho_i^2 \mathbb{E}\left[\left\|\sum_{e=0}^{E-1} \mu_{E,e}\omega_{i,e}^{(t)}\right\|^2 \,\middle|\, \mathcal{F}^{(t-1)}\right] \leq \sigma^2 s_E \sum_{i=1}^{m} \rho_i^2.$$

□

Lemma 6 (Gradient drift bound)**.** Under Assumptions 1–3, at communication round t, if $\eta L a_E < 1/\sqrt{3}$, we have

$$\mathbb{E}\left[\left\|D^{(t)}\right\|^2 \,\middle|\, \mathcal{F}^{(t-1)}\right] \leq \frac{2a_E^2 L^2 \eta^2}{1 - 3a_E^2 L^2 \eta^2}\left[a_E^2\left(\Gamma_0 + \Gamma_1\left\|\nabla \widetilde{f}^{(t)}(\phi^{(t)})\right\|^2\right) + s_E\sigma^2\right].$$

Proof. By Jensen's inequality and L-smoothness,

$$\left\|D^{(t)}\right\|^2 \leq a_E \sum_{i=1}^{m}\rho_i \sum_{e=0}^{E-1}\mu_{E,e}\left\|\nabla \widetilde{f}_i^{(t)}(\phi_{i,e}^{(t)}) - \nabla \widetilde{f}_i^{(t)}(\phi^{(t)})\right\|^2 \leq a_E L^2 \sum_{i=1}^{m}\rho_i \sum_{e=0}^{E-1}\mu_{E,e}\left\|\phi_{i,e}^{(t)} - \phi^{(t)}\right\|^2.$$

Set

$$M_e^{(t)} \coloneqq \sum_{i=1}^{m}\rho_i \mathbb{E}\left[\left\|\phi_{i,e}^{(t)} - \phi^{(t)}\right\|^2 \,\middle|\, \mathcal{F}^{(t-1)}\right], \qquad M_{\max}^{(t)} \coloneqq \max_{0\leq e\leq E-1} M_e^{(t)}.$$

Then

$$\mathbb{E}\left[\left\|D^{(t)}\right\|^2 \,\middle|\, \mathcal{F}^{(t-1)}\right] \leq a_E^2 L^2 M_{\max}^{(t)}.$$

It remains to bound $M_{\max}^{(t)}$. For $e \geq 1$,

$$\phi_{i,e}^{(t)} - \phi^{(t)} = -\eta \sum_{s=0}^{e-1}\mu_{e,s}\widetilde{g}_{i,s}^{(t)}.$$

Using

$$\widetilde{g}_{i,s}^{(t)} = \nabla \widetilde{f}_i^{(t)}(\phi^{(t)}) + \left[\nabla \widetilde{f}_i^{(t)}(\phi_{i,s}^{(t)}) - \nabla \widetilde{f}_i^{(t)}(\phi^{(t)})\right] + \omega_{i,s}^{(t)},$$

martingale cancellation, Young's inequality with equal weights, and L-smoothness imply

$$M_e^{(t)} \leq 2\eta^2\left(\sum_{s=0}^{e-1}\mu_{e,s}\right)^2 \sum_{i=1}^{m}\rho_i\left\|\nabla \widetilde{f}_i^{(t)}(\phi^{(t)})\right\|^2 + 3\eta^2 L^2\left(\sum_{s=0}^{e-1}\mu_{e,s}\right)\sum_{s=0}^{e-1}\mu_{e,s}M_s^{(t)} + 2\eta^2\sigma^2\sum_{s=0}^{e-1}\mu_{e,s}^2.$$

By Assumption 3,

$$\sum_{i=1}^{m}\rho_i\left\|\nabla \widetilde{f}_i^{(t)}(\phi^{(t)})\right\|^2 \leq \Gamma_0 + \Gamma_1\left\|\nabla \widetilde{f}^{(t)}(\phi^{(t)})\right\|^2.$$

Since

$$\sum_{s=0}^{e-1}\mu_{e,s} \leq a_E, \qquad \sum_{s=0}^{e-1}\mu_{e,s}^2 \leq s_E, \qquad M_s^{(t)} \leq M_{\max}^{(t)},$$

we get

$$M_e^{(t)} \leq 2\eta^2 a_E^2\left(\Gamma_0 + \Gamma_1\left\|\nabla \widetilde{f}^{(t)}(\phi^{(t)})\right\|^2\right) + 2\eta^2 s_E\sigma^2 + 3\eta^2 L^2 a_E^2 M_{\max}^{(t)}.$$

Taking the maximum over e gives

$$M_{\max}^{(t)} \le \frac{2\eta^2\left[a_E^2\left(\Gamma_0+\Gamma_1\left\|\nabla\widetilde{f}^{(t)}(\phi^{(t)})\right\|^2\right)+s_E\sigma^2\right]}{1-3a_E^2L^2\eta^2}.$$

Substituting this into the bound for $D^{(t)}$ proves the result. □

Lemma 7 (One-round descent)**.** Under Assumptions 1–3, at communication round t, if $\eta L a_E < \min\{1/3, 1/\sqrt{3+8\Gamma_1}\}$, we have

$$\mathbb{E}\Big[\widetilde{f}^{(t)}(\phi^{(t+1)}) - \widetilde{f}^{(t)}(\phi^{(t)}) \;\Big|\; \mathcal{F}^{(t-1)}\Big] \le -\frac{a_E\eta}{4}\left\|\nabla\widetilde{f}^{(t)}(\phi^{(t)})\right\|^2 + s_E L\eta^2\sigma^2\sum_{i=1}^m\rho_i^2 + 3a_E L^2\eta^3\left(a_E^2\Gamma_0+s_E\sigma^2\right).$$

Proof. By Lemma 3 and L-smoothness,

$$\widetilde{f}^{(t)}(\phi^{(t+1)}) - \widetilde{f}^{(t)}(\phi^{(t)}) \le -\eta\left\langle\nabla\widetilde{f}^{(t)}(\phi^{(t)}), \widetilde{g}^{(t)}\right\rangle + \frac{L\eta^2}{2}\left\|\widetilde{g}^{(t)}\right\|^2.$$

Using Lemma 4, expanding the square, and taking conditional expectation, the terms involving $\langle\nabla\widetilde{f}^{(t)}(\phi^{(t)}), \Xi^{(t)}\rangle$ vanish. Since

$$2\langle D^{(t)}, \Xi^{(t)}\rangle \le \left\|D^{(t)}\right\|^2 + \left\|\Xi^{(t)}\right\|^2,$$

we obtain

$$\begin{aligned}\mathbb{E}\left[\widetilde{f}^{(t)}(\phi^{(t+1)}) - \widetilde{f}^{(t)}(\phi^{(t)}) \,\middle|\, \mathcal{F}^{(t-1)}\right] &\le -a_E\eta\left\|\nabla\widetilde{f}^{(t)}(\phi^{(t)})\right\|^2 + \frac{La_E^2\eta^2}{2}\left\|\nabla\widetilde{f}^{(t)}(\phi^{(t)})\right\|^2 \\ &\quad + \left(-\eta+La_E\eta^2\right)\mathbb{E}\left[\left\langle\nabla\widetilde{f}^{(t)}(\phi^{(t)}), D^{(t)}\right\rangle \,\middle|\, \mathcal{F}^{(t-1)}\right] \\ &\quad + L\eta^2\mathbb{E}\left[\left\|D^{(t)}\right\|^2 \,\middle|\, \mathcal{F}^{(t-1)}\right] + L\eta^2\mathbb{E}\left[\left\|\Xi^{(t)}\right\|^2 \,\middle|\, \mathcal{F}^{(t-1)}\right].\end{aligned}$$

Since $a_E L\eta < 1/3$,

$$-\eta + La_E\eta^2 = -\eta(1-a_E L\eta) \le 0.$$

Using

$$-\left\langle\nabla\widetilde{f}^{(t)}(\phi^{(t)}), D^{(t)}\right\rangle \le \frac{a_E}{2}\left\|\nabla\widetilde{f}^{(t)}(\phi^{(t)})\right\|^2 + \frac{1}{2a_E}\left\|D^{(t)}\right\|^2,$$

we have

$$\begin{aligned}\mathbb{E}\left[\widetilde{f}^{(t)}(\phi^{(t+1)}) - \widetilde{f}^{(t)}(\phi^{(t)}) \,\middle|\, \mathcal{F}^{(t-1)}\right] &\le -\frac{a_E\eta}{2}\left\|\nabla\widetilde{f}^{(t)}(\phi^{(t)})\right\|^2 + \left(\frac{\eta}{2a_E}+\frac{L\eta^2}{2}\right)\mathbb{E}\left[\left\|D^{(t)}\right\|^2 \,\middle|\, \mathcal{F}^{(t-1)}\right] \\ &\quad + L\eta^2\mathbb{E}\left[\left\|\Xi^{(t)}\right\|^2 \,\middle|\, \mathcal{F}^{(t-1)}\right].\end{aligned}$$

By Lemmas 5 and 6,

$$\begin{aligned}&\left(\frac{\eta}{2a_E}+\frac{L\eta^2}{2}\right)\mathbb{E}\left[\left\|D^{(t)}\right\|^2 \,\middle|\, \mathcal{F}^{(t-1)}\right] \\ &\quad\le \frac{a_E L^2\eta^3(1+a_E L\eta)}{1-3a_E^2L^2\eta^2}\left[a_E^2\left(\Gamma_0+\Gamma_1\left\|\nabla\widetilde{f}^{(t)}(\phi^{(t)})\right\|^2\right)+s_E\sigma^2\right],\end{aligned}$$

and

$$L\eta^2\mathbb{E}\left[\left\|\Xi^{(t)}\right\|^2 \,\middle|\, \mathcal{F}^{(t-1)}\right] \le s_E L\eta^2\sigma^2\sum_{i=1}^m\rho_i^2.$$

The drift contribution to the gradient coefficient equals

$$a_E\eta\frac{\Gamma_1(a_EL\eta)^2(1+a_EL\eta)}{1-3a_E^2L^2\eta^2}\left\|\nabla\widetilde{f}^{(t)}(\phi^{(t)})\right\|^2.$$

The learning rate condition implies

$$\frac{\Gamma_1(a_EL\eta)^2(1+a_EL\eta)}{1-3a_E^2L^2\eta^2}\le\frac{1}{4},\qquad\frac{1+a_EL\eta}{1-3a_E^2L^2\eta^2}\le 3.$$

Combining the above inequalities proves the lemma. □

C.6.3 Proof of Theorem 1

Proof. By Lemma 7 and the tower property,

$$\mathbb{E}\left[\widetilde{f}^{(t)}(\phi^{(t+1)})-\widetilde{f}^{(t)}(\phi^{(t)})\right]\le-\frac{a_E\eta}{4}\mathbb{E}\left[\left\|\nabla\widetilde{f}^{(t)}(\phi^{(t)})\right\|^2\right]+s_EL\eta^2\sigma^2\sum_{i=1}^m\rho_i^2+3a_EL^2\eta^3\left(a_E^2\Gamma_0+s_E\sigma^2\right).$$

Combining the preceding inequality with Lemma 1 and Assumption 5, we obtain

$$\frac{a_E\eta}{4}\mathbb{E}\left[\left\|\nabla\widetilde{f}^{(t)}(\phi^{(t)})\right\|^2\right]\le\mathbb{E}\left[F(\zeta^{(t)})-F(\zeta^{(t+1)})\right]+\delta^{(t)}+s_EL\eta^2\sigma^2\sum_{i=1}^m\rho_i^2+3a_EL^2\eta^3\left(a_E^2\Gamma_0+s_E\sigma^2\right).$$

Summing over $t=0,\dots,T-1$ and using $F(\zeta^{(T)})\ge F^*$,

$$\frac{a_E\eta}{4}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{f}^{(t)}(\phi^{(t)})\right\|^2\right]\le F(\zeta^{(0)})-F^*+\sum_{t=0}^{T-1}\delta^{(t)}+Ts_EL\eta^2\sigma^2\sum_{i=1}^m\rho_i^2+3Ta_EL^2\eta^3\left(a_E^2\Gamma_0+s_E\sigma^2\right).$$

Thus,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\widetilde{f}^{(t)}(\phi^{(t)})\right\|^2\right]\le\frac{4\left(F(\zeta^{(0)})-F^*+\sum_{t=0}^{T-1}\delta^{(t)}\right)}{a_E\eta T}+\frac{4s_EL\eta\sigma^2}{a_E}\sum_{i=1}^m\rho_i^2+12L^2\eta^2\left(a_E^2\Gamma_0+s_E\sigma^2\right).$$

By Lemma 2 and Assumption 4,

$$\mathbb{E}\left[\left\|\nabla_\phi F(\zeta^{(t)})\right\|^2\right]\le\Gamma_2\,\mathbb{E}\left[\left\|\nabla\widetilde{f}^{(t)}(\phi^{(t)})\right\|^2\right]+\varepsilon^{(t)}.$$

Averaging over t and substituting the preceding bound gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla_\phi F(\zeta^{(t)})\right\|^2\right]\le\frac{4\Gamma_2\left(F(\zeta^{(0)})-F^*+\sum_{t=0}^{T-1}\delta^{(t)}\right)}{a_E\eta T}+\frac{4\Gamma_2s_EL\eta\sigma^2}{a_E}\sum_{i=1}^m\rho_i^2+12\Gamma_2L^2\eta^2\left(a_E^2\Gamma_0+s_E\sigma^2\right)+\frac{1}{T}\sum_{t=0}^{T-1}\varepsilon^{(t)}.$$

Taking $\eta = \Theta(1/\sqrt{T})$ yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla_\phi F(\zeta^{(t)})\right\|^2 = \mathcal{O}\Bigg(\frac{\Gamma_2(F(\zeta^{(0)})-F^*)}{a_E\sqrt{T}} + \frac{\Gamma_2}{a_E}\bar{\delta}_T + \frac{\Gamma_2 L\sigma^2 s_E}{a_E\sqrt{T}}\sum_{i=1}^{m}\rho_i^2 + \frac{\Gamma_2 L^2(a_E^2\Gamma_0+\sigma^2 s_E)}{T} + \bar{\varepsilon}_T\Bigg),$$

which completes the proof. □

D Experimental Details

D.1 Benchmark Dataset Details

We conduct experiments on FMNIST (Xiao et al., 2017), CIFAR-10 (Krizhevsky, 2009), and CIFAR-100 (Krizhevsky, 2009). FMNIST contains 70,000 28×28 grayscale images from 10 fashion classes. CIFAR-10 contains 60,000 32×32 RGB images from 10 object classes. CIFAR-100 contains 60,000 32×32 RGB images from 100 object classes, grouped into 20 superclasses. We construct three tasks of increasing complexity: (a) 10-class classification on FMNIST, (b) 10-class classification on CIFAR-10, and (c) 20-class classification using the CIFAR-100 superclasses.

D.2 Visualization of Dual Heterogeneity

We provide visualizations of the dual heterogeneity settings in our benchmark experiments.

To induce covariate shift, images are transformed at both the client and component levels. At the component level, component 0 retains the original image, while component 1 applies grayscale inversion for FMNIST and green-channel inversion for CIFAR-10/100. At the client level, the 8 clients correspond to the 2^3 binary combinations of three orthogonal dimensions: color shift (red- vs. blue-dominant), vertical spatial bias (top- vs. bottom-brightened), and horizontal spatial bias (left- vs. right-brightened). As shown in Fig. 5, these hierarchical transformations yield a visually conspicuous covariate shift across clients.

Fig. 6 illustrates the impact of α_{inter} on client label distributions. A smaller α_{inter} produces increasingly skewed class proportions and uneven dataset sizes across clients, corresponding to intensified inter-client heterogeneity. Similarly, Fig. 7 visualizes how α_{intra} modulates the intra-client mixing weights. While a small α_{intra} concentrates local data into a single dominant component, a larger α_{intra} yields more balanced mixtures, indicating stronger intra-client heterogeneity.

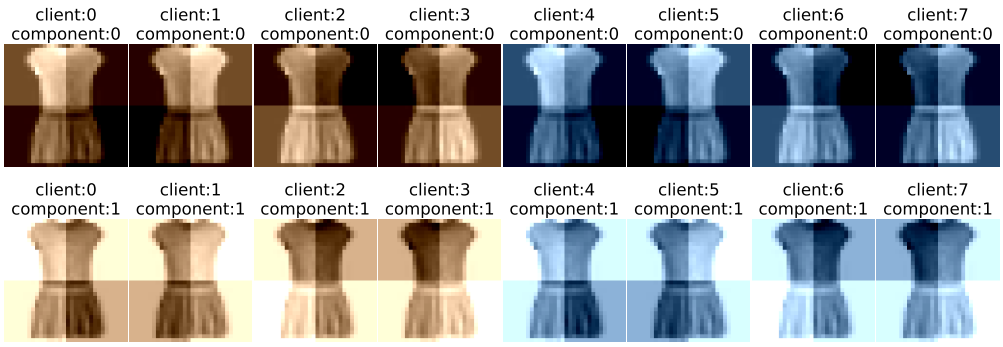

Figure 5: Visualization of an FMNIST sample under client- and component-level transformations.

D.3 Training Details on Benchmark Datasets

We employ local SGD with momentum, a batch size of 128, and an initial learning rate of 0.01 with cosine annealing. Each client performs 10 local steps per round with momentum 0.9 for image classification, and 15 local steps with momentum 0.95 for client routing. The encoder output

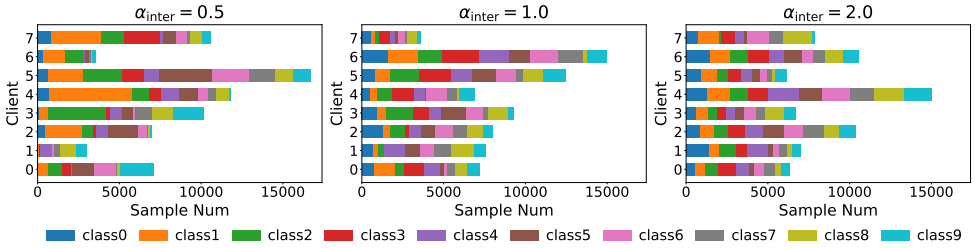

Figure 6: Visualization of client label distributions under varying α_{inter}.

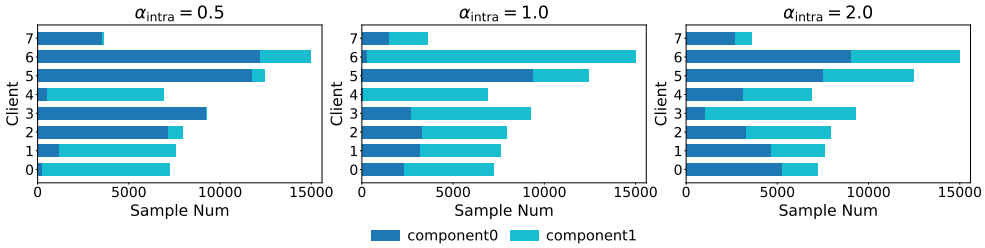

Figure 7: Visualization of client mixing weights under varying α_{intra}.

dimensions and total communication rounds are set to 32 and 200 for FMNIST, and 64 and 400 for both CIFAR-10 and CIFAR-100. For fine-tuning-based methods, we additionally perform one epoch of local fine-tuning before evaluation. For cluster- and mixture-model-based methods, we set the number of components C to 3.

D.4 Estimation of the Bridge Errors

This section details the empirical estimation of the bridge errors $\varepsilon^{(t)}$ and $\delta^{(t)}$ reported in Fig. 2. Computing these quantities requires evaluating the DRM objective $f^{(t)}$ and its gradient $\nabla f^{(t)}$, both of which depend on the Lagrange multipliers $\{\lambda_{ic}\}$. Therefore, at each communication round, we numerically solve the nonlinear system in equation 6 for $\{\lambda_{ic}\}$ using a Jacobian-based root solver, followed by a trust-region least-squares refinement when necessary. The resulting multipliers are then used to compute $\{f^{(t)}(\phi^{(t)})\}_{t=0}^{T-1}$ and $\{\nabla f^{(t)}(\phi^{(t)})\}_{t=0}^{T-1}$ along the training trajectory.

Estimation of the gradient bridge error $\varepsilon^{(t)}$. Assumption 4 gives

$$\mathbb{E}\left[\left\|\nabla f^{(t)}(\phi^{(t)})\right\|^2\right] \le \Gamma_2 \mathbb{E}\left[\left\|\nabla \widetilde{f}^{(t)}(\phi^{(t)})\right\|^2\right] + \varepsilon^{(t)}.$$

Motivated by this inequality, we first estimate the coefficient Γ_2 from the ratios between the gradient norms of the DRM objective and its tractable surrogate. To prevent numerical instability when the surrogate gradient norm vanishes in the denominator, our evaluation is restricted to the valid set:

$$\mathcal{T}_{\text{valid}} := \left\{ t : \left\|\nabla \widetilde{f}^{(t)}(\phi^{(t)})\right\|^2 > 10^{-9} \right\}.$$

We then estimate Γ_2 by the 95% quantile of the valid ratios to ensure robustness against outliers:

$$\widehat{\Gamma}_2 := \text{Quantile}_{0.95}\left\{ \frac{\left\|\nabla f^{(t)}(\phi^{(t)})\right\|^2}{\left\|\nabla \widetilde{f}^{(t)}(\phi^{(t)})\right\|^2} : t \in \mathcal{T}_{\text{valid}} \right\}.$$

Finally, given $\widehat{\Gamma}_2$, $\widehat{\varepsilon}^{(t)}$ is computed as the remaining positive gap:

$$\widehat{\varepsilon}^{(t)} := \max\left(\left\|\nabla f^{(t)}(\phi^{(t)})\right\|^2 - \widehat{\Gamma}_2 \left\|\nabla \widetilde{f}^{(t)}(\phi^{(t)})\right\|^2, 0 \right).$$

Estimation of the function-value bridge error $\delta^{(t)}$. Assumption 5 gives

$$\mathbb{E}\left[f^{(t)}(\phi^{(t+1)}) - f^{(t)}(\phi^{(t)})\right] \leq \mathbb{E}\left[\widetilde{f}^{(t)}(\phi^{(t+1)}) - \widetilde{f}^{(t)}(\phi^{(t)})\right] + \delta^{(t)}.$$

Accordingly, we estimate $\delta^{(t)}$ as

$$\widehat{\delta}^{(t)} := \max\left(\left[f^{(t)}(\phi^{(t+1)}) - f^{(t)}(\phi^{(t)})\right] - \left[\widetilde{f}^{(t)}(\phi^{(t+1)}) - \widetilde{f}^{(t)}(\phi^{(t)})\right], 0\right).$$

D.5 Real Medical Dataset and Training Details

Fed-ISIC2019 (Ogier du Terrail et al., 2022) comprises 23,247 RGB dermoscopic images collected from 6 clinical centers for an 8-class skin-lesion classification task, naturally forming a 6-client FL system with one client per clinical center.

The training setup largely follows that of the benchmark experiments, with several adjustments. Specifically, the encoder output dimension is set to 256 for image classification and 128 for client routing. The number of communication rounds is set to 400, and the batch size is set to 32.