FEDORS: SAMPLE-EFFICIENT PERSONALIZED FEDERATED LEARNING VIA ORTHOGONAL RESIDUAL SPARSE COMPONENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized federated learning (PFL) often combines a shared representation with client-specific prediction heads. However, these heads cannot recover predictive information absent from the shared representation, limiting performance when clients exhibit heterogeneous structures. We propose FedORS, which augments this architecture with client-specific residual components to capture information beyond the shared representation. In its linear formulation, these components are modeled as sparse and orthogonal to the shared representation subspace, enabling local corrections without increasing the dimension of the shared representation. Under structural and sparsity assumptions, we establish theoretical guarantees on convergence and sample complexity, and further characterize adaptation to previously unseen clients with limited local data. Our analysis highlights how low-dimensional shared structure and residual sparsity can support sample-efficient PFL. Experiments on synthetic and real-world datasets demonstrate improved performance across heterogeneous client distributions and effective adaptation to unseen clients.

1 INTRODUCTION

Personalized federated learning (PFL) aims to learn models adapted to individual clients while retaining the benefits of collaborative learning across distributed data. Achieving this objective is particularly challenging when data distributions vary across clients and each client has access to only limited local samples (Kairouz et al., 2021; Tan et al., 2022). A single shared model may fail to capture client-specific predictive structure, while purely local training cannot exploit information available from other clients. Representation-based PFL addresses this tension by decomposing the model into a shared representation and personalized components (Arivazhagan et al., 2019; Collins et al., 2021; Oh et al., 2022). The shared representation captures predictive structure common across clients, while the personalized components adapt it to each client's local prediction task.

Despite this advantage, representation-based personalization is fundamentally constrained by the information captured in the shared representation. Recent approaches enhance representation-based PFL by separating client-specific representation bias from client-invariant information (Zhang et al., 2023), or by adapting global generative classifiers to client-specific feature distributions (Mclaughlin & Su, 2024). However, these approaches still rely on predictive information encoded by the shared representation. When the shared representation omits predictive information needed by a client, personalization restricted to that representation cannot recover it. This limitation becomes more pronounced when clients exhibit distinct predictive structures, as a single shared representation may not preserve all information required for individual prediction. Effective personalization may therefore require modeling client-specific predictive structure beyond the shared representation.

One way to address this limitation is to model client-specific predictive structure beyond the shared representation. However, extending personalization beyond the shared representation can increase the complexity of local estimation. When the additional client-specific structure is modeled in a high-dimensional space without appropriate structural constraints, reliable estimation may require substantially more local samples (Collins et al., 2021). This creates a central tension: the personalized model should capture predictive structure not represented by the shared features, while the

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.

additional client-specific component should remain sufficiently structured to be learned from limited local data. Addressing this tension calls for a form of personalization that extends beyond the shared representation without introducing an unrestricted high-dimensional local model.

Motivated by this tension, we propose FedORS, a PFL framework that augments the shared representation with a client-specific residual component. The key idea is to decompose each client's predictive structure into a component captured through the shared representation and an additional residual learned locally. We formalize this principle in a linear setting, where the residual is modeled as sparse and orthogonal to the shared representation subspace. Orthogonality separates the residual from information already captured by the shared component, while sparsity controls the complexity of the additional local structure. This decomposition allows the personalized model to capture predictive structure beyond the shared features without introducing an unrestricted d-dimensional local model. We then generalize the residual-based formulation to nonlinear models by parameterizing the shared representation, personalized head, and residual with corresponding neural modules.

To characterize the sample efficiency of the proposed residual decomposition, we establish theoretical guarantees for learning the shared and personalized components from limited local data. Under structural and sparsity assumptions, FedORS requires $m = O(s \log d + k + \log n)$ samples per client, where the $s \log d$ term reflects the sparse complexity of the residual rather than direct dependence on the ambient dimension d. This shows that additional client-specific predictive structure can be introduced while keeping its local sample requirement controlled by sparsity. We further study adaptation to previously unseen clients. Once the shared representation has been learned from participating clients, a new client only needs to estimate its personalized head and sparse residual, requiring $m_{\text{new}} = O(s \log d + k)$ local samples. Experiments on synthetic and real-world datasets complement the theoretical analysis, demonstrating consistent improvements under heterogeneous data distributions and effective adaptation to new clients with limited local data. Our contributions are summarized as follows.

- **Residual-based personalization beyond shared representations.** We propose FedORS, which augments a collaboratively learned shared representation with a client-specific residual component to capture predictive structure not adequately represented by shared features. The residual is structured through orthogonality and sparsity to separate additional client-specific information from the shared representation while controlling its complexity. We formalize this residual decomposition in a linear setting and further generalize the resulting framework to nonlinear models with neural representations and lightweight personalized residual modules.
- **Sample-efficient learning and adaptation.** We establish theoretical guarantees for the linear formulation that characterize recovery of the shared representation and personalized components from finite local samples. The resulting sample requirement scales as $O(s \log d + k + \log n)$, with the additional client-specific structure governed by sparse complexity rather than unrestricted ambient-dimensional estimation. We further characterize adaptation to previously unseen clients, showing that once a well-learned shared representation is available, a new client can recover its personalized components using $O(s \log d + k)$ local samples.
- **Empirical evaluation.** We conduct extensive experiments on synthetic and real-world datasets under multiple heterogeneous data settings. The results demonstrate the benefit of augmenting shared representations with client-specific residual structure, validate the effectiveness of FedORS, and show strong adaptation performance on previously unseen clients with limited local samples.

2 Problem Formulation

In this section, we introduce the notation used throughout the paper and formulate the personalized federated learning problem considered in this work.

2.1 Notation

For an integer n, we write $[n] := \{1, 2, \ldots, n\}$, and for a set $S \subseteq [n]$, we denote its cardinality by $\mathrm{card}(S)$. For a vector $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}_S$ represents the subvector containing the entries indexed

by S. The support of $\mathbf{x}$ is defined as $\mathrm{supp}(\mathbf{x}) := \{i \in [n] : x_i \neq 0\}$, and $\mathbf{x}$ is s-sparse if $\mathrm{card}(\mathrm{supp}(\mathbf{x})) \leq s$. For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{A}_S$ denotes the submatrix of $\mathbf{A}$ consisting of the columns indexed by S. We use $\sigma_{\max}(\mathbf{A})$ and $\sigma_{\min}(\mathbf{A})$ to denote the largest and smallest singular values of $\mathbf{A}$, respectively. For the normalized matrix $\frac{1}{\sqrt{n}}\mathbf{A}$, we define $\bar{\sigma}_{\max}(\mathbf{A}) := \sigma_{\max}\big(\frac{1}{\sqrt{n}}\mathbf{A}\big)$ and $\bar{\sigma}_{\min}(\mathbf{A}) := \sigma_{\min}\big(\frac{1}{\sqrt{n}}\mathbf{A}\big)$. The notation $\|\cdot\|$ refers to the spectral norm for matrices or the Euclidean norm (ℓ_2 norm) for vectors, while $\|\cdot\|_1$ denotes the ℓ_1 norm.

2.2 Personalized Federated Learning Problem

We consider a personalized federated learning setting with n clients. Each client $i \in [n]$ has a local data distribution $\mathcal{D}_i$ over $\mathbb{R}^d \times \mathcal{Y}$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathcal{Y}$ denote the input and label, respectively. The distributions $\{\mathcal{D}_i\}_{i=1}^n$ may differ across clients, and each client learns a personalized model g_i. The learning objective is to minimize the average population risk

$$\min_{\{g_i\}_{i\in[n]}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{(\mathbf{x}_i,y_i)\sim\mathcal{D}_i}\left[\ell\big(g_i(\mathbf{x}_i), y_i\big)\right], \tag{2.1}$$

where ℓ denotes the task loss. We decompose each client model into shared and client-specific components. The shared component captures predictive structure learned collaboratively across clients, while the client-specific component includes a personalized head together with an additional residual that accounts for predictive structure beyond the shared representation. The shared representation is learned collaboratively, while the personalized head and residual are maintained locally by each client. We formulate this decomposition in a linear setting, and generalize it to nonlinear models.

3 Linear FedORS

We begin by formulating FedORS in a linear setting. We first introduce the linear model and its residual decomposition, followed by the corresponding federated optimization algorithm. We then establish theoretical guarantees for learning the shared and personalized components and characterize sample-efficient adaptation to new clients.

3.1 Linear formulation

Consider client $i \in [n]$ with input $\mathbf{x}_i \in \mathbb{R}^d$ and label $y_i \in \mathbb{R}$. The model consists of a shared representation matrix $\mathbf{B} \in \mathbb{R}^{d\times k}$, a personalized head $\mathbf{w}_i \in \mathbb{R}^k$, and a client-specific residual $\mathbf{s}_i \in \mathbb{R}^d$. The prediction is given by

$$\hat{y}_i = \mathbf{w}_i^\top \mathbf{B}^\top \mathbf{x}_i + \mathbf{s}_i^\top \mathbf{x}_i. \tag{3.2}$$

The first term $\mathbf{w}_i^\top \mathbf{B}^\top \mathbf{x}_i$ captures predictive structure through the shared k-dimensional representation, while the residual term $\mathbf{s}_i^\top \mathbf{x}_i$ captures additional client-specific predictive structure. We constrain $\mathbf{s}_i$ to be orthogonal to the column space of $\mathbf{B}$ so that it represents structure outside the shared subspace, and assume it to be sparse. When $\mathbf{s}_i = \mathbf{0}$, the model reduces to the standard shared-representation and personalized-head formulation.

For client i, we define the local objective as

$$f_i(\mathbf{w}_i, \mathbf{B}, \mathbf{s}_i) := \frac{1}{2}\mathbb{E}_{(\mathbf{x}_i,y_i)\sim\mathcal{D}_i}\left[\big(\mathbf{w}_i^\top \mathbf{B}^\top \mathbf{x}_i + \mathbf{s}_i^\top \mathbf{x}_i - y_i\big)^2\right]. \tag{3.3}$$

Let $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_n] \in \mathbb{R}^{k\times n}$ and $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_n] \in \mathbb{R}^{d\times n}$ collect the personalized heads and residual components, respectively. The federated learning objective is

$$\min_{\substack{\mathbf{B}\in\mathbb{R}^{d\times k}\\ \mathbf{W}\in\mathbb{R}^{k\times n},\mathbf{S}\in\mathbb{R}^{d\times n}}} F(\mathbf{W}, \mathbf{B}, \mathbf{S}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w}_i, \mathbf{B}, \mathbf{s}_i). \tag{3.4}$$

We assume that the samples of client i are generated according to the ground-truth linear model

$$y_i = \mathbf{w}_i^{*\top} \mathbf{B}^{*\top} \mathbf{x}_i + \mathbf{s}_i^{*\top} \mathbf{x}_i, \tag{3.5}$$

where $\mathbf{B}^* \in \mathbb{R}^{d\times k}$ denotes the ground-truth shared representation, $\mathbf{w}_i^* \in \mathbb{R}^k$ denotes the ground-truth personalized head, and $\mathbf{s}_i^* \in \mathbb{R}^d$ denotes the ground-truth client-specific residual.

Algorithm 1 Linear FedORS

Input: Number of clients n, number of communication rounds T, step size η

Server initializes a column-orthonormal shared representation $\mathbf{B}^0$; each client i initializes $\mathbf{w}_i^0$ and $\mathbf{s}_i^0$
for $t = 0, 1, \ldots, T-1$ **do**
  Server sends the current shared representation $\mathbf{B}^t$ to clients
  for each client i **in parallel do**
    Client i samples a fresh mini-batch of m local data samples
    Estimate the sparse residual:
      $\tilde{\mathbf{s}}_i^{t+1} \leftarrow \arg\min_{\mathbf{s}_i} \|\mathbf{s}_i\|_1 \quad \text{s.t. } \mathbf{X}_i^\top \mathbf{B}^t \mathbf{w}_i^t + \mathbf{X}_i^\top \mathbf{s}_i = \mathbf{y}_i$
    Project the residual outside the current shared subspace:
      $\mathbf{s}_i^{t+1} \leftarrow \mathbf{B}_\perp^t \mathbf{B}_\perp^{t\top} \tilde{\mathbf{s}}_i^{t+1}$
    Update the personalized head:
      $\mathbf{w}_i^{t+1} \leftarrow \arg\min_{\mathbf{w}_i} f_i(\mathbf{w}_i, \mathbf{B}^t, \mathbf{s}_i^{t+1})$
    Update the local shared representation:
      $\tilde{\mathbf{B}}_i^{t+1} \leftarrow \mathbf{B}^t - \eta \left(\mathbf{B}^t \mathbf{w}_i^{t+1} - \tilde{\mathbf{s}}_i^{t+1} - \mathbf{B}^t \mathbf{w}_i^t\right) \mathbf{w}_i^{t+1\top}.$
    Client i sends $\tilde{\mathbf{B}}_i^{t+1}$ to the server
  end for
  Server aggregates $\tilde{\mathbf{B}}^{t+1} \leftarrow \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{B}}_i^{t+1}$
  $\mathbf{B}^{t+1}, \mathbf{R}^{t+1} \leftarrow \mathrm{QR}(\tilde{\mathbf{B}}^{t+1})$
end for

3.2 Linear FedORS algorithm

We next present the optimization procedure for the linear formulation of FedORS. At each communication round, clients first estimate their personalized residuals and heads with the current shared representation fixed, and then update the shared representation using their local data. Only the shared representation is communicated to the server, while the personalized components remain local throughout training. The complete procedure is summarized in Algorithm 1.

Client update. At the beginning of communication round t, client i samples a mini-batch $\{\mathbf{x}_i^{t,j}, y_i^{t,j}\}_{j=1}^m \sim \mathcal{D}_i$, and let $\mathbf{X}_i \in \mathbb{R}^{d\times m}$ and $\mathbf{y}_i \in \mathbb{R}^m$ denote the corresponding input matrix and label vector. Given the current shared representation $\mathbf{B}^t$ and personalized head $\mathbf{w}_i^t$, the client first estimates the residual through the ℓ_1-minimization problem

$$\tilde{\mathbf{s}}_i^{t+1} = \arg\min_{\mathbf{s}_i} \|\mathbf{s}_i\|_1 \quad \text{s.t.} \quad \mathbf{X}_i^\top \mathbf{B}^t \mathbf{w}_i^t + \mathbf{X}_i^\top \mathbf{s}_i = \mathbf{y}_i. \tag{3.6}$$

The ℓ_1 objective promotes a sparse residual correction that explains predictive structure not captured by the current shared component. The recovered vector is then projected onto the orthogonal complement of the current shared representation:

$$\mathbf{s}_i^{t+1} = \mathbf{B}_\perp^t \mathbf{B}_\perp^{t\top} \tilde{\mathbf{s}}_i^{t+1}, \tag{3.7}$$

where $\mathbf{B}_\perp^t$ denotes an orthonormal basis for the orthogonal complement of the column space of $\mathbf{B}^t$. This projection removes the component aligned with the current shared subspace, so that $\mathbf{s}_i^{t+1}$ captures residual predictive structure outside the representation modeled by $\mathbf{B}^t$.

Given the updated residual $\mathbf{s}_i^{t+1}$, client i next estimates its personalized head by minimizing

$$f_i(\mathbf{w}_i, \mathbf{B}^t, \mathbf{s}_i^{t+1}) := \frac{1}{2m} \sum_{j=1}^m \left(\mathbf{w}_i^\top \mathbf{B}^{t\top} \mathbf{x}_i^{t,j} + \mathbf{s}_i^{t+1\top} \mathbf{x}_i^{t,j} - y_i^{t,j}\right)^2, \tag{3.8}$$

which gives $\mathbf{w}_i^{t+1} = \arg\min_{\mathbf{w}_i} f_i(\mathbf{w}_i, \mathbf{B}^t, \mathbf{s}_i^{t+1})$. After updating the personalized components, client i updates the shared representation as

$$\tilde{\mathbf{B}}_i^{t+1} = \mathbf{B}^t - \eta \left(\mathbf{B}^t \mathbf{w}_i^{t+1} - \tilde{\mathbf{s}}_i^{t+1} - \mathbf{B}^t \mathbf{w}_i^t\right) \mathbf{w}_i^{t+1\top}. \tag{3.9}$$

Client i then sends $\tilde{\mathbf{B}}_i^{t+1}$ to the server.

Server update. After receiving the local representation updates, the server averages them as $\tilde{\mathbf{B}}^{t+1} = \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{B}}_i^{t+1}$. The server then applies QR factorization $\mathbf{B}^{t+1}, \mathbf{R}^{t+1} = \mathrm{QR}(\tilde{\mathbf{B}}^{t+1})$, and uses the resulting column-orthonormal matrix $\mathbf{B}^{t+1}$ as the new shared representation.

3.3 THEORETICAL ANALYSIS

We now provide a theoretical analysis of linear FedORS. Our analysis characterizes the local sample requirement for recovering the shared representation and personalized components and shows how residual sparsity affects the complexity of client-specific estimation. We first introduce the subspace distance and the structural assumptions used in the analysis.

Definition 1 (Subspace distance) *The distance between the column spaces of two column-orthonormal matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d\times k}$ is defined as $\operatorname{dist}(\mathbf{U}, \mathbf{V}) := \left\|\mathbf{U}^\top \mathbf{V}_\perp\right\|$, where $\mathbf{V}_\perp \in \mathbb{R}^{d\times(d-k)}$ is an orthonormal basis for the orthogonal complement of the column space of $\mathbf{V}$.*

Definition 1 measures the discrepancy between two representation subspaces through their principal angles. The singular values of $\mathbf{U}^\top \mathbf{V}_\perp$ correspond to the sines of these angles. This distance is equivalent, up to constant factors, to the projection distance $\|\mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top\|$ and the optimal rotation distance between the two subspaces (Chen et al., 2021). We then make the following standard assumption on the local data samples (Collins et al., 2021; Deng et al., 2020).

Assumption 1 (Sub-gaussian) *The samples $\mathbf{x}_i \in \mathbb{R}^d$ are i.i.d. with mean $\mathbf{0}$, covariance $\mathbf{I}_d$, and are sub-gaussian, i.e., $\mathbb{E}[e^{\mathbf{v}^\top \mathbf{x}_i}] \le e^{\|\mathbf{v}\|^2/2}$ for all $\mathbf{v} \in \mathbb{R}^d$.*

Assumption 1 enables concentration of the finite-sample quantities used in the recovery analysis. Over all clients considered in the analysis, we denote the maximum norms of the ground-truth personalized heads and residuals by $r_w := \max_i \|\mathbf{w}_i^*\|$ and $r_s := \max_i \|\mathbf{s}_i^*\|$, respectively. Without loss of generality, we take the ground-truth representation $\mathbf{B}^*$ to have orthonormal columns. We next introduce the structural assumptions on the client-specific residuals.

Assumption 2 (Orthogonality) *The ground-truth residual matrix $\mathbf{S}^*$ lies in the orthogonal complement of the column space of $\mathbf{B}^*$, i.e., $\mathbf{S}^* = \mathbf{B}_\perp^* \mathbf{B}_\perp^{*\top} \mathbf{S}^*$.*

Assumption 3 (Residual sparsity) *For any client $i \in [n]$ and any round $t = 0, \dots, T-1$, the residual vector $\mathbf{s}_i^* + \mathbf{B}^* \mathbf{w}_i^* - \mathbf{B}^t \mathbf{w}_i^t$ is s-sparse.*

Assumption 4 (Incoherence) *For any round $t = 0, \dots, T-1$, the ground-truth residual matrix $\mathbf{S}^*$ satisfies $\|\mathbf{B}^{t\top} \mathbf{S}^*\| \le \mu_{bs}$ and $\|\mathbf{W}^* \mathbf{S}^{*\top}\| \le \mu_{ws}$, where μ_{bs} and μ_{ws} are non-negative constants.*

Assumption 2 separates the ground-truth residual from the shared representation subspace. Assumption 3 ensures that the residual satisfies the sparsity condition required for recovery through the ℓ_1-minimization step in Eq. (3.6). Assumption 4 controls the interaction of the residual components with the representation and personalized heads.

We now state the main result, which characterizes the local sample requirement for recovering the shared representation and personalized components.

Theorem 1 *Let Assumptions 1–4 hold. Suppose each client uses m fresh local samples in each round, where $m \ge C(s \log d + k + \log n)$ for some constant C. Define $\beta := \big(1 - \operatorname{dist}^2(\mathbf{B}^0, \mathbf{B}^*)\big)\bar{\sigma}_{\min}^2(\mathbf{W}^*) - \frac{2\mu_{ws}}{n}$, and $\varepsilon_0 := \frac{\mu_{bs} r_s}{\sqrt{n}\beta} + \frac{\mu_{ws}}{n\beta}$. Suppose that $\beta > 0$ and $\varepsilon_0 \le \operatorname{dist}(\mathbf{B}^0, \mathbf{B}^*)$. Then, for any step size $0 < \eta \le \frac{1}{2(r_w^2 + r_s^2)}$, we have*

$$\operatorname{dist}(\mathbf{B}^T, \mathbf{B}^*) \le (1 - \beta\eta)^T \operatorname{dist}(\mathbf{B}^0, \mathbf{B}^*) + \varepsilon_0 \tag{3.10}$$

with probability at least $1 - 2T \exp\big(-\tilde{C} \min\{s \log d + \log n,\ k + \log n\}\big)$, where $\tilde{C} > 0$ is a constant.

Sample efficiency. Theorem 1 shows that the learned representation contracts toward an ε_0-neighborhood of the ground-truth subspace with a per-round local sample requirement of $m = O(s \log d + k + \log n)$. The $s \log d$ term arises from recovering the sparse client-specific residual, the k term corresponds to estimating the personalized head within the shared representation, and the $\log n$ term provides simultaneous control across the n participating clients. Thus, the additional client-specific structure introduced by the residual is governed by the sparse complexity

$s \log d$ rather than unrestricted estimation in the ambient dimension d. For comparison, the corresponding sample requirement of FedRep is $O(\frac{k^2 d}{n} + k^3 \log n)$ (Collins et al., 2021). When the client-specific residual is sufficiently sparse relative to the ambient dimension d, the $s \log d$ term in FedORS can be substantially smaller than the d-dependent term in FedRep, leading to a lower local sample requirement.

Building on Theorem 1, we next establish conditions for ε-accurate recovery of both the shared representation and personalized components.

Corollary 1 *Under the setting of Theorem 1, let* $\Lambda := \big(1 - \mathrm{dist}^2(\mathbf{B}^0, \mathbf{B}^*)\big)\bar{\sigma}^2_{\min}(\mathbf{W}^*)$. *For any target accuracy* $0 < \varepsilon \le 1$, *suppose the number of communication rounds satisfies* $T \ge \frac{2}{\eta\Lambda} \log \frac{2\lceil r_w + r_s \rceil}{\varepsilon}$, *and* $\mu_{bs} \le \frac{\sqrt{n}\Lambda\varepsilon}{8 r_s \lceil r_w + r_s \rceil}, \mu_{ws} \le \frac{n\Lambda\varepsilon}{8\lceil r_w + r_s \rceil}$. *Then, the learned shared representation satisfies* $\mathrm{dist}(\mathbf{B}^T, \mathbf{B}^*) \le \varepsilon$, *with the same probability as in Theorem 1. Furthermore, for every client* $i \in [n]$, *the personalized components satisfy*

$$\big\|\mathbf{B}^T \mathbf{w}_i^{T+1} - \mathbf{B}^* \mathbf{w}_i^*\big\| \le \varepsilon, \qquad \big\|\mathbf{s}_i^{T+1} - \mathbf{s}_i^*\big\| \le \varepsilon. \tag{3.11}$$

Corollary 1 translates the representation contraction in Theorem 1 into ε-accurate recovery of both the shared and personalized components. The allowable incoherence levels μ_{bs} and μ_{ws} scale with $\sqrt{n}$ and n, respectively, so the corresponding conditions become less restrictive as the number of participating clients increases. Theorem 1 and Corollary 1 establish that FedORS can recover both shared and client-specific predictive structure with a local sample requirement determined by the low-dimensional head and sparse residual structure.

3.4 Adaptation to new clients

We next consider adaptation to previously unseen clients after federated training. Let $\hat{\mathbf{B}}$ denote the shared representation learned from the participating clients and provided to the new clients. During adaptation, $\hat{\mathbf{B}}$ is kept fixed, while the personalized heads and residuals are estimated from local data. Given $\hat{\mathbf{B}}$ and an initial personalized head $\mathbf{w}_i^0$, a new client i estimates its sparse residual and projects it onto the orthogonal complement of the learned shared subspace:

$$\tilde{\mathbf{s}}_i = \arg\min_{\mathbf{s}_i} \|\mathbf{s}_i\|_1 \quad \text{s.t.} \quad \mathbf{X}_i^\top \hat{\mathbf{B}} \mathbf{w}_i^0 + \mathbf{X}_i^\top \mathbf{s}_i = \mathbf{y}_i, \tag{3.12}$$

$$\mathbf{s}_i = \hat{\mathbf{B}}_\perp \hat{\mathbf{B}}_\perp^\top \tilde{\mathbf{s}}_i. \tag{3.13}$$

It then estimates the personalized head while keeping the shared representation $\hat{\mathbf{B}}$ and the updated residual $\mathbf{s}_i$ fixed:

$$\mathbf{w}_i = \arg\min_{\mathbf{w}_i} f_i(\mathbf{w}_i, \hat{\mathbf{B}}, \mathbf{s}_i). \tag{3.14}$$

We consider new clients under the same linear model and residual structural conditions as the participating clients, with the orthogonality and sparsity conditions extended to their local adaptation. The following result characterizes the local sample requirement for recovering these personalized components.

Theorem 2 *Suppose the learned shared representation* $\hat{\mathbf{B}}$ *satisfies* $\mathrm{dist}(\hat{\mathbf{B}}, \mathbf{B}^*) \le \frac{\varepsilon}{r_w + r_s}$. *For a new client i, if the local sample size satisfies* $m_{\mathrm{new}} \ge C(s \log d + k)$ *for some constant* $C > 0$, *then with high probability,*

$$\left\|\hat{\mathbf{B}} \mathbf{w}_i - \mathbf{B}^* \mathbf{w}_i^*\right\| \le \varepsilon, \qquad \|\mathbf{s}_i - \mathbf{s}_i^*\| \le \varepsilon. \tag{3.15}$$

Sample-efficient adaptation. For a new client, the shared representation learned from the participating clients can be directly reused, so adaptation reduces to estimating only the personalized head and sparse residual from local data. This avoids relearning the shared structure and reduces adaptation to a single-client recovery problem. Accordingly, Theorem 2 shows that the personalized components can be recovered with $m_{\mathrm{new}} = O(s \log d + k)$ local samples. The k term corresponds to estimating the personalized head within the learned shared representation, while the $s \log d$ term arises from recovering the sparse residual. With the shared representation fixed, the recovery condition only needs to hold for the new client, rather than uniformly across all participating clients as in federated representation learning. Consequently, the adaptation sample requirement does not include the additional $\log n$ term.

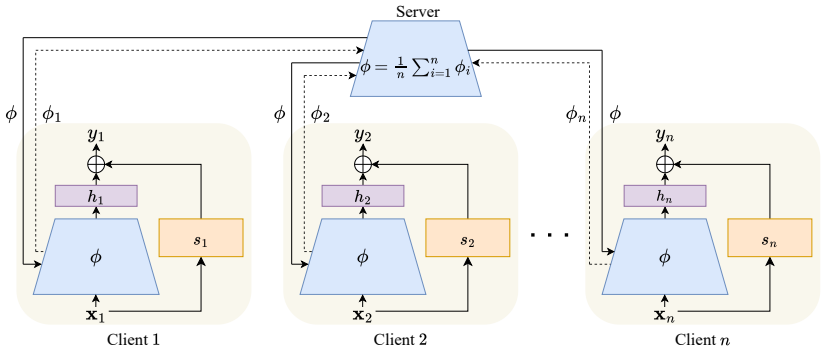

Figure 1: General FedORS framework.

4 Generalizing FedORS to Nonlinear Models

We next generalize the FedORS formulation to nonlinear models. The linear shared representation $\mathbf{B}^\top \mathbf{x}$ is generalized to a nonlinear representation $\phi(\mathbf{x})$, with the personalized head implemented by h_i, while the linear residual $\mathbf{s}_i^\top \mathbf{x}$ is generalized to a personalized residual module $s_i(\mathbf{x})$. The resulting prediction retains the additive structure of FedORS, combining the shared-representation branch $h_i(\phi(\mathbf{x}))$ with the client-specific residual $s_i(\mathbf{x})$.

4.1 General FedORS formulation

As illustrated in Figure 1, general FedORS consists of three components for each client i: a shared representation $\phi : \mathbb{R}^d \to \mathbb{R}^k$, a personalized head $h_i : \mathbb{R}^k \to \mathbb{R}^C$, and a personalized residual $s_i : \mathbb{R}^d \to \mathbb{R}^C$, where C denotes the output dimension. For classification tasks, $h_i(\phi(\mathbf{x}_i))$ and $s_i(\mathbf{x}_i)$ correspond to predictive logits. The prediction for $\mathbf{x}_i$ is given by $h_i\big(\phi(\mathbf{x}_i)\big)+s_i(\mathbf{x}_i)$. The term $h_i(\phi(\mathbf{x}_i))$ performs prediction based on the shared representation, while $s_i(\mathbf{x}_i)$ captures additional client-specific predictive structure.

Given the local data distribution $\mathcal{D}_i$, the local objective of client i is

$$f_i(h_i, \phi, s_i) := \mathbb{E}_{(\mathbf{x}_i, y_i)\sim\mathcal{D}_i}\left[\ell\left(h_i\big(\phi(\mathbf{x}_i)\big) + s_i(\mathbf{x}_i), y_i\right)\right], \tag{4.16}$$

where ℓ denotes the task loss. The corresponding federated objective is

$$\min_{\phi, \{h_i, s_i\}_{i\in[n]}} \frac{1}{n}\sum_{i=1}^{n}\left[f_i(h_i, \phi, s_i) + \lambda\|s_i\|_1\right], \tag{4.17}$$

where $\lambda \geq 0$ controls the sparsity regularization applied to the personalized residual, and $\|s_i\|_1$ denotes the ℓ_1 norm of the trainable parameters of the residual module. Under this decomposition, the shared representation ϕ is learned collaboratively and communicated between the clients and server, whereas h_i and s_i are optimized locally and remain local to client i.

4.2 Optimization

We optimize the general FedORS objective through sequential local updates of the personalized residual, personalized head, and shared representation. In each communication round, each client first updates its residual, followed by the personalized head and its local copy of the shared representation. The local modules are optimized using gradient-based methods, after which the updated shared representation is sent to the server for aggregation. For simplicity, we use the same symbols s_i, h_i, and ϕ to denote the corresponding modules and their trainable parameters when the meaning is clear from context. The complete training procedure is summarized in Algorithm 2 in Appendix B.

Client update. At the beginning of communication round t, client i receives the current shared representation ϕ^{t-1} from the server. It first updates the personalized residual for τ_s local steps. For $q = 1, \ldots, \tau_s$,

$$s_i^{t,q} = s_i^{t,q-1} - \eta\,\mathcal{G}_s\left(f_i(h_i^{t,0}, \phi^{t-1}, s_i^{t,q-1}) + \lambda\|s_i^{t,q-1}\|_1\right), \tag{4.18}$$

where $\mathcal{G}_s(\cdot)$ denotes the optimization direction with respect to the residual parameters, such as that produced by SGD or Adam. After updating the residual, client i optimizes its personalized head for τ_h steps while keeping the shared representation and residual fixed. For $q = 1, \ldots, \tau_h$,

$$h_i^{t,q} = h_i^{t,q-1} - \eta\, \mathcal{G}_h \left(f_i(h_i^{t,q-1}, \phi^{t-1}, s_i^{t,\tau_s}) \right). \tag{4.19}$$

Finally, client i updates its local copy of the shared representation while fixing the updated personalized components. For $q = 1, \ldots, \tau_\phi$,

$$\phi_i^{t,q} = \phi_i^{t,q-1} - \eta\, \mathcal{G}_\phi \left(f_i(h_i^{t,\tau_h}, \phi_i^{t,q-1}, s_i^{t,\tau_s}) \right). \tag{4.20}$$

After the local updates, h_i^{t,τ_h} and s_i^{t,τ_s} are retained by client i, while only the updated representation ϕ_i^{t,τ_ϕ} is uploaded to the server.

Server update. The server aggregates the representations received from the clients by $\phi^t = \frac{1}{n}\sum_{i=1}^n \phi_i^{t,\tau_\phi}$. The resulting shared representation ϕ^t is then broadcast to the clients for the next communication round. Through this procedure, the shared representation is learned collaboratively across clients, while the personalized head and residual are optimized locally for each client's prediction task.

5 Experiments

In this section, we evaluate FedORS on synthetic and real-world datasets. The synthetic experiments focus on the linear setting and examine residual-based personalization and sample-efficient learning under limited local data. For real-world datasets, we report the main results under label-per-client heterogeneity, with additional experiments on Dirichlet heterogeneity and new-client adaptation provided in Appendix C.

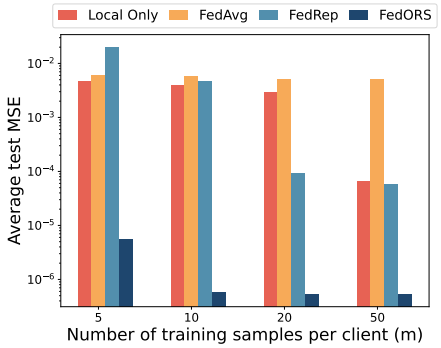

Figure 2: Test MSE under different numbers of training samples per client

5.1 Synthetic Data

For synthetic experiments, we focus on the linear setting described in Section 3. Data samples are generated as $\mathbf{x}_i^j \sim \mathcal{N}(0, \mathbf{I}_d)$, with labels y_i^j generated by a fixed linear transform of $\mathbf{x}_i^j$ with independent zero-mean Gaussian noise of standard deviation 10^{-3}. We set the data dimension $d = 50$ and the number of clients $n = 50$, and vary the number of training samples per client $m \in \{5, 10, 20, 50\}$. We compare FedORS with three baselines: Local Only, FedAvg (McMahan et al., 2017), and FedRep (Collins et al., 2021). Local Only denotes that each client solves a linear regression problem using only its local data, without collaboration with other clients.

Figure 2 illustrates the average test MSE under different numbers of training samples per client. We observe that as the number of training samples increases, FedORS consistently achieves the lowest test MSE among all compared methods. Its improvement over FedRep indicates that the residual component can capture client-specific predictive structure not represented by the shared representation and personalized head alone, while the improvement over Local Only and FedAvg shows the

benefit of combining collaboratively learned shared structure with client-specific residual modeling. More importantly, FedORS maintains strong performance with few local samples, providing empirical support for the sample-efficient learning characterized by our theoretical analysis.

5.2 REAL DATA

Datasets and heterogeneity scenarios. We conduct experiments on three benchmark datasets: CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), and Tiny-ImageNet (Chrabaszcz et al., 2017). In the label-per-client (LPC) setting, each client is assigned data from a fixed number of classes. We use LPC values $\{2, 4\}$ for CIFAR-10, $\{2, 10\}$ for CIFAR-100, and $\{5, 20\}$ for Tiny-ImageNet. We also evaluate FedORS under Dirichlet heterogeneity, with results reported in Appendix C.

Table 1: Mean test accuracy (%) $\pm$ standard deviation under LPC heterogeneity over three runs.

Datasets	CIFAR-10		CIFAR-100		Tiny-ImageNet	
LPC	2	4	2	10	5	20
Local Only	86.87±0.10	75.11±0.33	90.98±0.20	62.21±0.30	62.19±0.13	38.29±0.07
FedAvg	50.45±0.32	51.98±0.33	44.66±0.10	28.88±0.53	19.97±0.51	12.10±0.32
FedPer	88.77±0.19	80.22±0.19	80.97±0.98	66.22±0.22	59.41±0.68	43.65±0.33
APFL	89.01±0.59	80.81±0.24	88.40±0.09	67.14±0.11	65.41±0.14	44.65±0.14
Ditto	86.84±0.08	76.13±0.10	86.70±0.17	66.80±0.18	63.38±0.08	38.72±0.08
FedRep	88.97±0.17	80.66±0.21	83.61±0.10	66.80±0.18	61.78±0.10	44.95±0.21
FedBABU	88.93±0.20	80.73±0.19	89.52±0.02	68.26±0.30	63.92±0.12	45.20±0.06
FedDBE	86.43±0.55	77.68±0.71	76.40±1.20	54.83±0.33	63.31±0.31	43.54±0.55
pFedFDA	86.97±0.02	79.39±0.08	85.33±0.17	60.59±0.45	58.96±0.03	43.83±0.12
PFedCS	89.35±0.09	80.69±0.15	81.32±0.11	66.24±0.25	60.31±0.20	44.22±0.50
FedRIR	88.34±0.09	80.49±0.26	78.81±0.19	63.61±0.17	60.49±0.33	41.16±0.11
FedORS	**90.05±0.22**	**81.70±0.41**	**91.46±0.10**	**68.95±0.19**	**67.48±0.11**	**46.07±0.13**

Baselines and implementation. We compare FedORS with FedAvg (McMahan et al., 2017) and a range of PFL baselines: FedPer (Arivazhagan et al., 2019), APFL (Deng et al., 2020), Ditto (Li et al., 2021b), FedRep (Collins et al., 2021), FedBABU (Oh et al., 2022), FedDBE (Zhang et al., 2023), pFedFDA (Mclaughlin & Su, 2024), PFedCS (Wu et al., 2025), and FedRIR (Huang et al., 2025). All results are averaged over three runs. Detailed implementation settings are provided in Appendix C.

Performance under LPC heterogeneity.

Table 1 reports the main results under LPC heterogeneity. FedORS achieves the best performance across all evaluated datasets and LPC settings. FedAvg, which uses a fully shared model, yields lower accuracy, indicating the limitation of relying on a single model under heterogeneous data distributions. Local Only is competitive when each client contains fewer labels, while its gap to representation-based methods generally widens as the LPC value increases across all three datasets. This trend indicates a greater benefit from collaborative learning as the local prediction task involves a broader set of classes. Among representation-based methods, FedORS further improves performance by introducing a client-specific residual in addition to the shared representation and personalized head, supporting the benefit of modeling additional client-specific structure beyond the standard representation-head architecture.

6 CONCLUSION

We propose FedORS, which augments shared representations with client-specific residual components to capture predictive structure beyond shared features. We develop a linear formulation with orthogonal sparse residuals, establish guarantees for sample-efficient learning and new-client adaptation, and generalize the framework to nonlinear models. Experiments on synthetic and real-world datasets demonstrate consistent improvements under heterogeneous data settings and effective adaptation to new clients with limited local data.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing. We reviewed all AI-assisted content and revised it as necessary. We take responsibility for the final content of this work, including any text produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or sensitive data. All experiments are conducted using publicly available benchmark datasets. We are not aware of any specific ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

We provide the assumptions, optimization procedures, algorithms, and complete proofs for the theoretical results in the main paper and appendix. The experimental section describes the experimental setup and implementation details, with additional information provided in the appendix to facilitate reproduction.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

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

Brian Bullins, Elad Hazan, Adam Kalai, and Roi Livni. Generalize across tasks: Efficient algorithms for linear representation learning. In *Algorithmic Learning Theory*, pp. 235–246. PMLR, 2019.

Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus. Mathematique*, 346(9-10):589–592, 2008.

Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

Gary Cheng, Karan Chadha, and John Duchi. Fine-tuning is fine in federated learning. *arXiv preprint arXiv:2108.07313*, 3, 2021.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pp. 2089–2099. PMLR, 2021.

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

Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

Simon Foucart and Holger Rauhut. An invitation to compressive sensing. In *A Mathematical Introduction to Compressive Sensing*, pp. 1–39. Springer, 2013.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. FedDC: Federated learning with non-iid data via local drift decoupling and correction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10102–10111. IEEE Computer Society, 2022.

Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Yongqiang Huang, Zerui Shao, Ziyuan Yang, Zexin Lu, and Yi Zhang. FedRIR: rethinking information representation in federated learning. In *Proceedings of the ACM on Web Conference 2025*, pp. 807–816, 2025.

Ellango Jothimurugesan, Kevin Hsieh, Jianyu Wang, Gauri Joshi, and Phillip B Gibbons. Federated learning under distributed concept drift. In *International Conference on Artificial Intelligence and Statistics*, pp. 5834–5853. PMLR, 2023.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 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 Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Royson Lee, Minyoung Kim, Da Li, Xinchi Qiu, Timothy Hospedales, Ferenc Huszár, and Nicholas Lane. Fedl2p: Federated learning to personalize. *Advances in Neural Information Processing Systems*, 36:14818–14836, 2023.

Mohan Li, Dario Fenoglio, Martin Gjoreski, and Marc Langheinrich. Federated learning with profile mapping under distribution shifts and drifts. In *International Conference on Learning Representations*, 2026.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10708–10717. IEEE, 2021a.

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

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021b.

Yichen Li, Xiuying Wang, Wenchao Xu, Haozhao Wang, Yining Qi, Jiahua Dong, and Ruixuan Li. Feature distillation is the better choice for model-heterogeneous federated learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Junkang Liu, Yuanyuan Liu, Fanhua Shang, Hongying Liu, Jin Liu, and Wei Feng. Improving generalization in federated learning with highly heterogeneous data via momentum-based stochastic controlled weight averaging. In *International Conference on Machine Learning*, 2025.

Connor Mclaughlin and Lili Su. Personalized federated learning via feature distribution adaptation. *Advances in Neural Information Processing Systems*, 37:77038–77059, 2024.

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

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

Jonathan Scott, Hossein Zakerinia, and Christoph H Lampert. PeFLL: Personalized federated learning by learning to learn. *arXiv preprint arXiv:2306.05515*, 2023.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12):9587–9603, 2022.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pp. 10434–10443. PMLR, 2021.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

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.

Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Communication-efficient federated learning via knowledge distillation. *Nature Communications*, 13(1):2032, 2022.

Siyuan Wu, Yongzhe Jia, Bowen Liu, Haolong Xiang, Xiaolong Xu, and Wanchun Dou. Pfedcs: A personalized federated learning method for enhancing collaboration among similar classifiers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21572–21580, 2025.

Jianqiao Zhang, Caifeng Shan, and Jungong Han. Fedgmkd: An efficient prototype federated learning framework through knowledge distillation and discrepancy-aware aggregation. *Advances in Neural Information Processing Systems*, 37:118326–118356, 2024.

Jianqing Zhang, Yang Hua, Jian Cao, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Eliminating domain bias for federated learning in representation space. *Advances in Neural Information Processing Systems*, 36:14204–14227, 2023.

Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning*, pp. 26311–26329. PMLR, 2022.

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

Federated Learning. Federated learning was first formalized in the FedAvg (McMahan et al., 2017) framework, which learns a single global model by averaging locally updated parameters from participating clients. Data heterogeneity is one of the essential challenges in federated optimization (Kairouz et al., 2021). Prior works address this issue through several mechanisms. FedProx (Li et al., 2020) and FedDyn (Acar et al., 2021) regularize local objectives to constrain the discrepancy between local and global optimization, while MOON (Li et al., 2021a) further introduces model-level contrastive regularization to align local and global representations. SCAFFOLD (Karimireddy et al., 2020) and FedDC (Gao et al., 2022) correct local updates to better approximate the global optimization direction. Other approaches use auxiliary information, such as public data for correcting local updates (Zhao et al., 2018) and label-distribution-aware logit calibration in FedLC (Zhang et al., 2022). More recently, federated optimization has been further improved through adaptive client grouping (Jothimurugesan et al., 2023), momentum-based stochastic weight averaging (Liu et al., 2025), and distribution-guided aggregation with client-specific model assignment (Li et al., 2026).

Personalized Federated Learning. Personalized federated learning introduces personalized components into federated settings. Personalization can be implemented through diverse approaches, such as local fine-tuning (Cheng et al., 2021; Wang et al., 2019), local and global model interpolation (Deng et al., 2020; Hanzely & Richtárik, 2020), meta-learning (Fallah et al., 2020; Lee et al., 2023; Scott et al., 2023), knowledge distillation (Li et al., 2025; Wu et al., 2022; Zhang et al., 2024). A growing line of work on representation learning (Bullins et al., 2019; Du et al., 2021; Tripuraneni et al., 2021) provided theoretical foundations for representation-based personalization. FedRep (Collins et al., 2021) introduces a decoupled representation-head framework, where clients collaboratively learn a shared representation while optimizing personalized heads locally, achieving linear convergence and strong empirical results under data heterogeneity. More recently, FedDBE (Zhang et al., 2023) and FedRIR (Huang et al., 2025) separate client-invariant and client-specific representations to improve personalization under heterogeneous data, while pFedFDA (Mclaughlin & Su, 2024) adapts global generative classifiers to local feature distributions. While these methods demonstrate the effectiveness of the representation-head paradigm, their personalization remains tied to information encoded by the shared representation. FedORS reduces this dependency by explicitly modeling client-specific information beyond the shared representation.

B General FedORS algorithm

The optimization procedure for general FedORS is summarized in Algorithm 2.

Algorithm 2 General FedORS

Input: Number of clients n, number of communication rounds T, learning rate η, local update steps τ_s, τ_h, and τ_ϕ, regularization coefficient λ

Server initializes ϕ^0; each client i initializes h_i^0 and s_i^0
for $t = 1, 2, \ldots, T$ **do**
  Server sends ϕ^{t-1} to clients
  for each client i **in parallel do**
    $s_i^{t,0} \leftarrow s_i^{t-1}$, $h_i^{t,0} \leftarrow h_i^{t-1}$, $\phi_i^{t,0} \leftarrow \phi^{t-1}$
    for $q = 1, 2, \ldots, \tau_s$ **do**
      $s_i^{t,q} \leftarrow s_i^{t,q-1} - \eta\, \mathcal{G}_s \left(f_i(h_i^{t,0}, \phi^{t-1}, s_i^{t,q-1}) + \lambda \| s_i^{t,q-1} \|_1 \right)$
      // Update personalized residual
    end for
    for $q = 1, 2, \ldots, \tau_h$ **do**
      $h_i^{t,q} \leftarrow h_i^{t,q-1} - \eta\, \mathcal{G}_h \left(f_i(h_i^{t,q-1}, \phi^{t-1}, s_i^{t,\tau_s}) \right)$ // Update personalized head
    end for
    for $q = 1, 2, \ldots, \tau_\phi$ **do**
      $\phi_i^{t,q} \leftarrow \phi_i^{t,q-1} - \eta\, \mathcal{G}_\phi \left(f_i(h_i^{t,\tau_h}, \phi_i^{t,q-1}, s_i^{t,\tau_s}) \right)$ // Update local representation
    end for
    Client i sends ϕ_i^{t,τ_ϕ} to the server
  end for
  Server updates $\phi^t \leftarrow \frac{1}{n} \sum_{i=1}^n \phi_i^{t,\tau_\phi}$ // Aggregate shared representations
end for

C Experimental details and additional results

C.1 Additional synthetic experiment

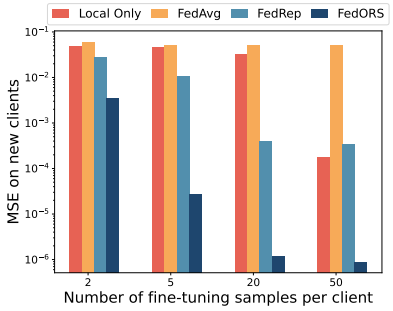

Figure 3: Test MSE on new clients under different local sample sizes

Adaptation to new clients. We evaluate adaptation to previously unseen clients with limited local data. All methods are first trained on $n = 50$ participating clients with $m = 20$ training samples per client. For FedAvg and FedRep, the learned representation is kept fixed and only the personalized head is optimized for each new client. For FedORS, the shared representation is likewise fixed, while both the personalized head and residual are optimized locally. Local Only continues to fit a linear regression model using only the new client's local data. Figure 3 reports the average test MSE of the adapted models. FedORS achieves the lowest test MSE across the evaluated local sample sizes, with a clear advantage when only a small number of samples are available. These results support the sample-efficient adaptation characterized by our theoretical analysis, showing that new clients

can effectively adapt the learned shared representation by estimating their personalized components from limited local data.

C.2 Orthogonal residuals in nonlinear models

Table 2: Mean test accuracy (%) $\pm$ standard deviation on CIFAR-10 with a nonlinear MLP.

Method	LPC = 2	Dir. $\alpha = 0.1$
Local Only	72.37 ± 0.19	75.86 ± 0.24
FedAvg	29.14 ± 0.44	29.29 ± 0.30
FedRep	70.20 ± 0.37	75.99 ± 0.03
FedORS	$\mathbf{76.61 \pm 0.11}$	$\mathbf{81.30 \pm 0.07}$

We conduct an experiment on CIFAR-10 using a simple MLP under LPC $= 2$ and Dirichlet $\alpha = 0.1$, and compare FedORS with Local Only, FedAvg, and FedRep. To impose orthogonality, each client uses a local reference mini-batch to compute the input Jacobian of the current shared representation and constructs the corresponding local representation subspace. The first layer of the residual module is then projected onto the orthogonal complement of this subspace after each residual update. This projection is applied before updating the personalized head and local shared representation, following the same residual–head–representation optimization order used in FedORS. As shown in Table 2, FedORS achieves the best performance under both heterogeneous data partitions. These results demonstrate that the orthogonal residual construction can be implemented in nonlinear neural networks while maintaining strong personalized prediction performance. This further supports the applicability of orthogonal residual modeling beyond the linear formulation.

C.3 Additional real-data experiments

C.3.1 Implementation details

We run all real-data experiments on NVIDIA L40 GPUs. We simulate 50 clients in each communication round, and split each client's local dataset into 75% for training and 25% for testing. We consider two heterogeneity settings. In the label-per-client (LPC) setting, each client is assigned data from a fixed number of classes. We use LPC values $\{2, 4\}$ for CIFAR-10, $\{2, 10\}$ for CIFAR-100, and $\{5, 20\}$ for Tiny-ImageNet. In the Dirichlet setting, client class proportions are sampled from a Dirichlet distribution with parameter α, and samples are allocated according to these proportions. We use $\alpha \in \{0.1, 0.5\}$ for CIFAR-10 and $\alpha \in \{0.05, 0.1\}$ for CIFAR-100 and Tiny-ImageNet.

The batch size is set to 10, the number of local epochs is 1 by default, and we use Adam with a learning rate of 10^{-4}. For the LPC setting, we run 25 communication rounds on CIFAR-10 and on CIFAR-100 with LPC $= 2$. For the Dirichlet setting on CIFAR-10, we run 10 communication rounds for $\alpha = 0.1$ and 50 communication rounds for $\alpha = 0.5$. For all remaining settings on CIFAR-100 and all settings on Tiny-ImageNet, we run 50 communication rounds. For the ℓ_1 regularization coefficient, we set $\lambda = 10^{-5}$ for CIFAR-10 with LPC $= 2$, $\alpha = 0.1$, and $\alpha = 0.5$, and set $\lambda = 5 \times 10^{-5}$ for CIFAR-10 with LPC $= 4$ and all settings on CIFAR-100 and Tiny-ImageNet. Each experiment is repeated three times, and we report the mean test accuracy with standard deviation.

For all datasets and heterogeneity settings, we adopt ResNet-8 (He et al., 2016) and treat its final linear layer as the personalized head, with the remaining layers serving as the shared representation. FedORS further adds a lightweight personalized residual network with fewer layers and smaller hidden dimensions. Table 3 reports the average client MACs per local epoch under the CIFAR-10 setting with LPC $= 2$. In this setting, the residual requires only 12.05% of the MACs of ResNet-8.

Table 3: Average client MACs per local epoch on CIFAR-10 with LPC=2.

Module	Avg. client MACs per local epoch	Relative MACs to ResNet-8
ResNet-8	1.3079×10^{10}	-
Residual network	1.5760×10^{9}	12.05%

C.3.2 BASELINES

We compare FedORS with a range of personalized federated learning methods, along with FedAvg (McMahan et al., 2017), which represents the standard setting where clients collaboratively learn a single global model. Among the personalized methods, FedPer (Arivazhagan et al., 2019) learns a shared representation and personalized heads, where clients update both simultaneously but only the shared representation is aggregated globally. FedRep (Collins et al., 2021) decouples the training by first updating the personalized heads and then updating the shared representation locally, aggregating only the shared parameters. FedBABU (Oh et al., 2022) trains a shared representation globally and subsequently fine-tunes the personalized heads locally. FedDBE (Zhang et al., 2023) mitigates domain bias by separating client-invariant and client-specific representations and aligning the invariant ones across clients. FedRIR (Huang et al., 2025) further separates global and personalized representations through masked local feature learning and information distillation. pFedFDA (Mclaughlin & Su, 2024) trains the shared representation and personalized heads based on the global feature distribution. PFedCS (Wu et al., 2025) enhances personalization by promoting collaboration among clients with similar classifiers. APFL (Deng et al., 2020) interpolates between global and local models, while Ditto (Li et al., 2021b) applies global regularization to keep each client's model close to the global model.

C.3.3 ADDITIONAL RESULTS

Performance under Dirichlet heterogeneity. We further evaluate FedORS under Dirichlet heterogeneity. Table 4 reports the test accuracy on CIFAR-10, CIFAR-100, and Tiny-ImageNet under different Dirichlet parameters. Across these settings, FedORS achieves the best or comparable performance among the compared methods, demonstrating its effectiveness across different levels of data heterogeneity.

Table 4: Mean test accuracy (%) $\pm$ standard deviation under Dirichlet heterogeneity with parameter α over three runs. FedORS consistently achieves competitive performance across all settings.

Datasets	CIFAR-10		CIFAR-100		Tiny-ImageNet	
α	0.1	0.5	0.05	0.1	0.05	0.1
Local Only	87.90±0.03	63.34±0.12	63.65±0.41	50.86±0.02	45.85±0.13	35.98±0.18
FedAvg	43.85±5.93	45.63±0.54	18.12±0.31	24.65±0.13	10.30±0.09	9.31±0.40
FedPer	86.30±0.20	**75.24±0.17**	65.13±0.22	55.17±0.14	49.93±0.25	40.93±0.04
APFL	87.54±0.21	74.86±0.33	66.10±0.25	55.89±0.06	52.49±0.10	41.78±0.04
Ditto	85.14±0.10	72.95±0.27	60.53±0.16	45.99±0.19	46.87±0.11	38.68±0.16
FedRep	86.71±0.16	75.19±0.03	65.74±0.29	55.28±0.22	50.75±0.09	41.18±0.15
FedBABU	87.97±0.02	73.47±0.33	66.20±0.10	55.53±0.22	51.05±0.11	39.19±0.27
FedDBE	77.62±1.22	68.89±0.01	63.74±0.59	50.30±0.39	48.22±0.03	38.01±0.07
pFedFDA	85.99±0.22	70.05±0.14	59.10±0.26	51.89±0.45	47.23±0.20	39.21±0.25
PFedCS	86.68±0.19	75.18±0.09	63.06±0.13	53.61±0.12	47.89±0.16	40.36±0.17
FedRIR	85.20±0.44	73.46±0.02	59.92±0.40	49.92±0.01	46.78±0.35	37.34±0.10
FedORS	**88.41±0.16**	74.92±0.07	**68.76±0.30**	**57.84±0.02**	**52.56±0.23**	**42.73±0.10**

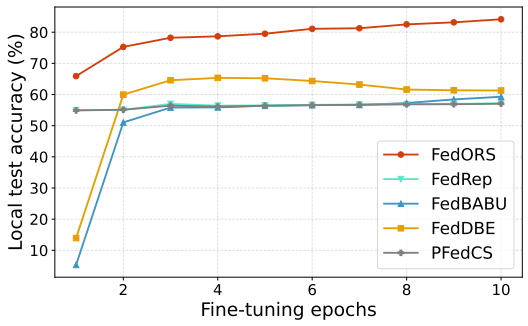

Figure 4: Test accuracy after fine-tuning on new clients

Adaptation to new clients. We further evaluate the ability of FedORS to adapt to previously unseen clients on CIFAR-10. We first conduct federated training with 50 clients, each containing data from two classes, and then introduce 10 new clients following the same heterogeneous setting with disjoint data samples. For new-client adaptation, we keep the shared representation learned during federated training fixed. For representation-based baselines, each new client updates only its personalized head using its local data. For FedORS, the shared representation is fixed, while both the personalized head and residual are optimized locally. As shown in Figure 4, FedORS achieves higher test accuracy than the compared methods throughout the adaptation process. These results show that the shared representation learned during federated training can be effectively reused by new clients, while the personalized residual provides additional capacity for adapting the model to their local data.

D PROOFS

Before we provide the detailed proof, we first recall the standard notion of the Restricted Isometry Property (RIP), which plays a central role in sparse recovery.

Definition 2 (Restricted Isometry Property) *A matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ satisfies the restricted isometry property if there exists a constant $\delta_s \geq 0$ such that, for all s-sparse vectors $\mathbf{x} \in \mathbb{R}^n$,*

$$(1-\delta_s)\,\|\mathbf{x}\|^2 \leq \|\mathbf{A}\mathbf{x}\|^2 \leq (1+\delta_s)\,\|\mathbf{x}\|^2. \tag{D.21}$$

Equivalently, the restricted isometry constant is defined as

$$\delta_s := \max_{S\subseteq[n],\,\mathrm{card}(S)\leq s} \|\mathbf{A}_S^\top \mathbf{A}_S - \mathbf{I}_S\|, \tag{D.22}$$

where $\mathbf{A}_S$ denotes the submatrix of $\mathbf{A}$ consisting of the columns indexed by S.

D.1 KEY LEMMAS

In this section, we present the auxiliary lemmas necessary for our main convergence analysis. Lemma 1 and Lemma 2 jointly guarantee the exact recovery of the personalized sparse residual, while Lemma 3 derives the closed-form solution for the personalized head.

Lemma 1 *Under Assumption 1, for each client i, let $\mathbf{X}_i \in \mathbb{R}^{d\times m}$ be the local data matrix and define $\tilde{\mathbf{X}}_i := \frac{1}{\sqrt{m}}\mathbf{X}_i$. Then for any fixed vector $\mathbf{v} \in \mathbb{R}^d$ and any $t \in (0,1)$,*

$$\mathbb{P}\Big(\big|\|\tilde{\mathbf{X}}_i^\top \mathbf{v}\|^2 - \|\mathbf{v}\|^2\big| \geq t\|\mathbf{v}\|^2\Big) \leq 2\exp\Big(-\frac{1}{32}t^2 m\Big). \tag{D.23}$$

Proof of Lemma 1. Without loss of generality, we omit the subscript of the data matrix and assume that $\|\mathbf{v}\| = 1$, then we have

$$\|\tilde{\mathbf{X}}^\top \mathbf{v}\|^2 - \|\mathbf{v}\|^2 = \frac{1}{m}\|\mathbf{X}^\top \mathbf{v}\|^2 - \|\mathbf{v}\|^2 = \frac{1}{m}\sum_{j=1}^{m}\big((\mathbf{x}_j^\top \mathbf{v})^2 - \|\mathbf{v}\|^2\big). \tag{D.24}$$

Define the random variable $Z_j = \big(\mathbf{x}_j^\top \mathbf{v}\big)^2 - \|\mathbf{v}\|^2$, where $j \in [m]$. By Assumption 1, we have $\mathbb{E}[\mathbf{x}_j\mathbf{x}_j^\top] = \mathbf{I}_d$, thus $\mathbb{E}Z_j = 0$. Further, since $\mathbf{x}_j^\top \mathbf{v}$ is sub-gaussian, Z_j is sub-exponential such that

$$\mathbb{P}\left(|Z_j| \geq r\right) \leq 2\exp\Big(-\frac{r}{2}\Big) \quad \text{for all} \quad r > 0. \tag{D.25}$$

The variables Z_j are mutually independent due to the independence of each $\mathbf{x}_j$. Therefore, from equation D.24 we have a sum of m mean-zero independent sub-exponential random variables, then we can apply Bernstein inequality for sub-exponential random variables (Corollary 7.32 in (Foucart & Rauhut, 2013)), leading to

$$\mathbb{P}\Big(\Big|\frac{1}{m}\sum_{j=1}^{m} Z_j\Big| \geq t\Big) = \mathbb{P}\Big(\Big|\sum_{j=1}^{m} Z_j\Big| \geq tm\Big) \leq 2\exp\Big(-\frac{mt^2}{32+4t}\Big). \tag{D.26}$$

By taking $t \in (0,1)$ and combining with equation D.24, we have

$$\mathbb{P}\Big(\big|\|\tilde{\mathbf{X}}^\top \mathbf{v}\|^2 - \|\mathbf{v}\|^2\big| \geq t\|\mathbf{v}\|^2\Big) \leq 2\exp\Big(-\frac{1}{32}t^2 m\Big). \tag{D.27}$$

This concludes the proof. □

Lemma 2 *Under Assumption 1 and 3, each client i exactly recovers $\tilde{\mathbf{s}}_i^{t+1} := \mathbf{B}^*\mathbf{w}_i^* + \mathbf{s}_i^* - \mathbf{B}^t\mathbf{w}_i^t$, by solving convex program*

$$\min \|\mathbf{s}_i\|_1, \quad \text{s.t. } \mathbf{X}_i^\top \mathbf{s}_i = \mathbf{X}_i^\top \tilde{\mathbf{s}}_i^{t+1}, \tag{D.28}$$

with probability at least $1-\exp(-C_1(s\log d + \log n))$, as long as $m \geq C_2(s\log d + \log n)$ for some constant C_1, C_2.

Proof of Lemma 2. Without loss of generality, we omit the subscript of the data matrix and rewrite equation D.28 as

$$\min \|\mathbf{s}_i\|_1, \quad \text{s.t. } \tilde{\mathbf{X}}^\top \mathbf{s}_i = \tilde{\mathbf{X}}^\top \tilde{\mathbf{s}}_i^{t+1}, \tag{D.29}$$

where $\tilde{\mathbf{X}} := \frac{1}{\sqrt{m}}\mathbf{X}_i$. With this transformation, we can establish a similar line of proof that closely parallels the argument in Theorem 9.9 of Foucart & Rauhut (2013). We establish the restricted isometry property of $\tilde{\mathbf{X}}$ based on equation D.27. According to Proposition C.3 in (Foucart & Rauhut, 2013), for $\rho \in (0, 1/2)$, there exists a finite subset U of the unit ball $B_S := \{\mathbf{a} \in \mathbb{R}^d, \operatorname{supp} \mathbf{a} \subset S, \|\mathbf{a}\| \le 1\}$ satisfying

$$\operatorname{card}(U) \le \left(1 + \frac{2}{\rho}\right)^{2s} \text{ and } \min_{\mathbf{u}\in U} \|\mathbf{a} - \mathbf{u}\| \le \rho \quad \text{for all } \ \mathbf{a} \in B_S. \tag{D.30}$$

Here, $S \subseteq [d]$ and $\operatorname{card}(S) = 2s$. The concentration inequality equation D.27 gives, for $t \in (0, 1)$,

$$\begin{aligned}
\mathbb{P}\big(\big|\|\tilde{\mathbf{X}}^\top \mathbf{u}\|^2 - \|\mathbf{u}\|^2\big| \ge t\|\mathbf{u}\|^2 \quad \text{for some } \mathbf{u} \in U\big) &\le \sum_{\mathbf{u}\in U} \mathbb{P}\big(\big|\|\tilde{\mathbf{X}}^\top \mathbf{u}\|^2 - \|\mathbf{u}\|^2\big| \ge t\|\mathbf{u}\|^2\big) \\
&\le 2 \operatorname{card}(U) \exp\Big(-\frac{1}{32}t^2 m\Big) &\text{(D.31)} \\
&\le 2\left(1 + \frac{2}{\rho}\right)^{2s} \exp\Big(-\frac{1}{32}t^2 m\Big). &\text{(D.32)}
\end{aligned}$$

Now we assume that the realization of the matrix $\tilde{\mathbf{X}}$ satisfies

$$\big|\|\tilde{\mathbf{X}}^\top \mathbf{u}\|^2 - \|\mathbf{u}\|^2\big| < t\|\mathbf{u}\|^2 \quad \text{for all } \mathbf{u} \in U, \tag{D.33}$$

which holds with probability at least

$$1 - 2\left(1 + \frac{2}{\rho}\right)^{2s} \exp(-\frac{1}{32}t^2 m). \tag{D.34}$$

Next, we will show that equation D.33 implies $\big|\|\tilde{\mathbf{X}}^\top \mathbf{a}\|^2 - \|\mathbf{a}\|^2\big| \le \delta$ for all $\mathbf{a} \in B_S$ and $\delta \in (0, 1)$. Let $\mathbf{H} := \tilde{\mathbf{X}}_S \tilde{\mathbf{X}}_S^\top - \mathbf{I}_d$, and equation D.33 means that

$$|\mathbf{u}_S^\top \mathbf{H} \mathbf{u}_S| < t \quad \text{for all } \mathbf{u} \in U. \tag{D.35}$$

Consider a vector $\mathbf{a} \in B_S$ and choose a vector $\mathbf{u} \in U$ satisfying $\|\mathbf{a} - \mathbf{u}\| \le \rho \le 1/2$, then we have

$$\begin{aligned}
|\mathbf{a}_S^\top \mathbf{H} \mathbf{a}_S| &= |\mathbf{u}_S^\top \mathbf{H} \mathbf{u}_S + (\mathbf{a}_S^\top + \mathbf{u}_S^\top)\mathbf{H}(\mathbf{a}_S - \mathbf{u}_S)| &\text{(D.36)} \\
&\le |\mathbf{u}_S^\top \mathbf{H} \mathbf{u}_S| + |(\mathbf{a}_S^\top + \mathbf{u}_S^\top)\mathbf{H}(\mathbf{a}_S - \mathbf{u}_S)| &\text{(D.37)} \\
&< t + \|\mathbf{H}\|\|\mathbf{a}_S + \mathbf{u}_S\|\|\mathbf{a}_S - \mathbf{u}_S\| &\text{(D.38)} \\
&\le t + 2\rho\|\mathbf{H}\|. &\text{(D.39)}
\end{aligned}$$

Taking the maximum over all $\mathbf{a} \in B_S$, we have

$$\|\mathbf{H}\| < t + 2\rho\|\mathbf{H}\|, \quad \text{i.e.} \quad \|\mathbf{H}\| < \frac{t}{1 - 2\rho}. \tag{D.40}$$

Choose $t := (1 - 2\rho)\delta$, so that $\|\mathbf{H}\| < \delta$. By equation D.34, we conclude that

$$\mathbb{P}\left(\|\mathbf{H}\| \ge \delta\right) \le 2\left(1 + \frac{2}{\rho}\right)^{2s} \exp\Big(-\frac{1}{32}(1 - 2\rho)^2\delta^2 m\Big). \tag{D.41}$$

Recall the definition that $\delta_{2s} = \operatorname{supp}_{S\subset[d],\operatorname{card}(S)=2s} \|\mathbf{H}\|$. We use equation D.41 and take the union bound over all subsets $S \subset [d]$ of cardinality $2s$,

$$\begin{aligned}
\mathbb{P}\left(\delta_{2s} \ge \delta\right) &\le \sum_{S\subseteq[d],\operatorname{card}(S)=2s} \mathbb{P}\left(\|\mathbf{H}\| \ge \delta\right) &\text{(D.42)} \\
&\le 2\binom{d}{2s}\left(1 + \frac{2}{\rho}\right)^{2s} \exp\Big(-\frac{1}{32}(1 - 2\rho)^2\delta^2 m\Big) &\text{(D.43)} \\
&\le 2\left(\frac{ed}{2s}\right)^{2s}\left(1 + \frac{2}{\rho}\right)^{2s} \exp\Big(-\frac{1}{32}(1 - 2\rho)^2\delta^2 m\Big). &\text{(D.44)}
\end{aligned}$$

Choosing $\rho = 2/(e^{7/2}-1)$ and $\delta = \frac{1}{3}$, we have $\delta_{2s} < \frac{1}{3}$ with probability at least $1 - e^{-C_2(s\log(d)+\log n)}$ if

$$m \geq C_2(s\log(d)+\log n), \tag{D.45}$$

where C_2 is some constant. By taking a union bound over $i \in [n]$, $\delta_{2s} < \frac{1}{3}$ holds for all clients with probability at least $1 - e^{-C_1(s\log(d)+\log n)}$, where C_1 is some constant.

Given that $\tilde{\mathbf{X}}$ satisfies the restricted isometry property with $\delta_{2s} < \frac{1}{3}$, the ℓ_1-minimization problem equation D.29 uniquely recovers all s-sparse vectors in $\mathbb{R}^d$, which is a standard conclusion in sparse recovery (Candes et al., 2006; Candes, 2008; Foucart & Rauhut, 2013). Thus we complete the proof. □

Lemma 3 *On the event of exact residual recovery in Lemma 2, the update of the personalized heads satisfies*

$$\mathbf{W}^{t+1} = \mathbf{B}^{t\top}(\mathbf{B}^*\mathbf{W}^* + \mathbf{S}^*), \tag{D.46}$$

with probability at least $1-\exp(-C_1(k+\log n))$*, as long as* $m \geq C_2(k+\log n)$ *for some constant* C_1, C_2.

Proof of Lemma 3. Combining equation 3.6, equation 3.7 and Lemma 2, we obtain the update for the sparse residual, such that $\mathbf{s}_i^{t+1} = \mathbf{B}_\perp^t \mathbf{B}_\perp^{t\top}(\mathbf{B}^*\mathbf{w}_i^* + \mathbf{s}_i^*)$. Further, combining equation 3.8, we obtain the update for the head $\mathbf{w}_i^{t+1}$ as follows:

$$\begin{aligned}\mathbf{w}_i^{t+1} &= \arg\min_{\mathbf{w}_i} \frac{1}{2m}\left\|\mathbf{X}_i^\top(\mathbf{B}^t\mathbf{w}_i + \mathbf{s}_i^{t+1}) - \mathbf{X}_i^\top(\mathbf{B}^*\mathbf{w}_i^* + \mathbf{s}_i^*)\right\|^2 \\ &= \arg\min_{\mathbf{w}_i} \frac{1}{2}\left\|\frac{1}{\sqrt{m}}\mathbf{X}_i^\top\mathbf{B}^t\mathbf{w}_i - \frac{1}{\sqrt{m}}\mathbf{X}_i^\top\mathbf{B}^t\mathbf{B}^{t\top}(\mathbf{B}^*\mathbf{w}_i^* + \mathbf{s}_i^*)\right\|^2.\end{aligned} \tag{D.47}$$

Treating equation D.47 as a standard linear regression problem, we obtain the closed form of solution:

$$\mathbf{w}_i^{t+1} = \Big(\frac{1}{m}\mathbf{B}^{t\top}\mathbf{X}_i\mathbf{X}_i^\top\mathbf{B}^t\Big)^{-1}\Big(\frac{1}{m}\mathbf{B}^{t\top}\mathbf{X}_i\mathbf{X}_i^\top\mathbf{B}^t\Big)\mathbf{B}^{t\top}(\mathbf{B}^*\mathbf{w}_i^* + \mathbf{s}_i^*) = \mathbf{B}^{t\top}(\mathbf{B}^*\mathbf{w}_i^* + \mathbf{s}_i^*), \tag{D.48}$$

for which we require the term $\frac{1}{m}\mathbf{B}^{t\top}\mathbf{X}_i\mathbf{X}_i^\top\mathbf{B}^t$ to be invertible. Define $\mathbf{G}_i := \frac{1}{m}\mathbf{B}^{t\top}\mathbf{X}_i\mathbf{X}_i^\top\mathbf{B}^t$ and consider that

$$\mathbf{G}_i = \sum_{j=1}^m \frac{1}{\sqrt{m}}\mathbf{B}^{t\top}\mathbf{x}_i^j\Big(\frac{1}{\sqrt{m}}\mathbf{B}^{t\top}\mathbf{x}_i^j\Big)^\top. \tag{D.49}$$

Let $\mathbf{v}_i^j := \frac{1}{\sqrt{m}}\mathbf{B}^{t\top}\mathbf{x}_i^j$ for all $i \in [n], j \in [m]$. Define $\mathbf{V}_i := [\mathbf{v}_i^1, \ldots, \mathbf{v}_i^m] \in \mathbb{R}^{k\times m}$. It follows that $\mathbf{G}_i = \mathbf{V}_i\mathbf{V}_i^\top$. Since $\mathbf{x}_i^j$ is isotropic sub-Gaussian and $\mathbf{B}^t$ has orthonormal columns, $\mathbf{B}^{t\top}\mathbf{x}_i^j$ is an isotropic sub-Gaussian vector in $\mathbb{R}^k$. Hence, $\mathbf{v}_i^j$ is a scaled sub-Gaussian vector with covariance $\mathbf{I}_k/m$. Using Theorem 4.6.1 in (Vershynin, 2018), we have

$$\sigma_{\min}(\mathbf{G}_i) = \sigma_{\min}^2(\mathbf{V}_i) \geq \Big(1 - C\Big(\sqrt{\frac{k}{m}} + \frac{t}{\sqrt{m}}\Big)\Big)^2 \geq 1 - C'\Big(\sqrt{\frac{k}{m}} + \frac{t}{\sqrt{m}}\Big), \tag{D.50}$$

with probability at least $1-e^{-t^2}$ for $m \geq k, t \geq 0$ and some constant C, C'. Let $\delta_k = C'\Big(\sqrt{\frac{k}{m}} + \frac{t}{\sqrt{m}}\Big)$ to obtain

$$\sigma_{\min}(\mathbf{G}_i) \geq 1 - \delta_k, \tag{D.51}$$

with probability at least $1 - e^{-(\delta_k\sqrt{m}/C' - \sqrt{k})^2}$ for $m > k$. Choose t such that $\delta_k = \frac{12C'}{\sqrt{m}}(\sqrt{k} + \sqrt{\log n})$, then equation D.51 holds with probability at least

$$1 - \exp\Big(-\big(12(\sqrt{k}+\sqrt{\log n}) - \sqrt{k}\big)^2\Big) \geq 1 - \exp(-121k - 144\log n). \tag{D.52}$$

Here we need $m > C_2(k+\log n)$ to enforce $\delta_k < 1$, for some constant C_2. Finally, taking a union bound over $i \in [n]$, we have, with probability at least

$$1 - n\exp(-121k - 144\log n) \geq 1 - \exp(-121k - 143\log n) \geq 1 - \exp(-C_1(k+\log n)), \tag{D.53}$$

for each client i, $\mathbf{G}_i$ is invertible, for some constant C_1. Thus we complete the proof. □

D.2 Proof of Main Theorem

With the closed-form updates for $\mathbf{S}^{t+1}$ and $\mathbf{W}^{t+1}$, we now proceed to the proof of Theorem 1.

Theorem 1 *Let Assumptions 1–4 hold. Suppose each client uses m fresh local samples in each round, where $m \geq C(s\log d + k + \log n)$ for some constant C. Define $\beta := \big(1 - \mathrm{dist}^2(\mathbf{B}^0, \mathbf{B}^*)\big)\bar{\sigma}^2_{\min}(\mathbf{W}^*) - \frac{2\mu_{ws}}{n}$, and $\varepsilon_0 := \frac{\mu_{bs} r_s}{\sqrt{n}\beta} + \frac{\mu_{ws}}{n\beta}$. Suppose that $\beta > 0$ and $\varepsilon_0 \leq \mathrm{dist}(\mathbf{B}^0, \mathbf{B}^*)$. Then, for any step size $0 < \eta \leq \frac{1}{2(r_w^2 + r_s^2)}$, we have*

$$\mathrm{dist}(\mathbf{B}^T, \mathbf{B}^*) \leq (1 - \beta\eta)^T \,\mathrm{dist}(\mathbf{B}^0, \mathbf{B}^*) + \varepsilon_0 \tag{D.54}$$

with probability at least $1 - 2T\exp\big(-\tilde{C}\min\{s\log d + \log n,\, k + \log n\}\big)$, where $\tilde{C} > 0$ is a constant.

Proof of Theorem 1. From Lemma 2 and the update rules of equation 3.6 and equation 3.7, we have

$$\mathbf{S}^{t+1} = \mathbf{B}_\perp^t \mathbf{B}_\perp^{t\top} \tilde{\mathbf{S}}^{t+1}, \tag{D.55}$$

where $\tilde{\mathbf{S}}^{t+1} = \mathbf{B}^*\mathbf{W}^* + \mathbf{S}^* - \mathbf{B}^t\mathbf{W}^t$. From Lemma 3, we have

$$\mathbf{W}^{t+1} = \mathbf{B}^{t\top}(\mathbf{B}^*\mathbf{W}^* + \mathbf{S}^*). \tag{D.56}$$

Further, by the update of the representation, we have the update for representation as follows:

$$\tilde{\mathbf{B}}^{t+1} = \mathbf{B}^t - \frac{\eta}{n}(\mathbf{B}^t\mathbf{W}^{t+1} - \tilde{\mathbf{S}}^{t+1} - \mathbf{B}^t\mathbf{W}^t)\mathbf{W}^{t+1\top} \tag{D.57}$$

$$= \mathbf{B}^t - \frac{\eta}{n}\big(\mathbf{B}^t\mathbf{W}^{t+1} - (\mathbf{B}^*\mathbf{W}^* + \mathbf{S}^*)\big)\mathbf{W}^{t+1\top}, \tag{D.58}$$

$$\mathbf{B}^{t+1}, \mathbf{R}^{t+1} = \mathrm{QR}(\tilde{\mathbf{B}}^{t+1}). \tag{D.59}$$

Left-multiplying $\tilde{\mathbf{B}}^{t+1}$ by $\mathbf{B}_\perp^{*\top}$ and writing the QR decomposition of $\tilde{\mathbf{B}}^{t+1}$ as $\tilde{\mathbf{B}}^{t+1} = \mathbf{B}^{t+1}\mathbf{R}^{t+1}$, we have

$$\mathbf{B}_\perp^{*\top}\mathbf{B}^{t+1} = \big(\mathbf{B}_\perp^{*\top}\mathbf{B}^t - \frac{\eta}{n}\mathbf{B}_\perp^{*\top}(\mathbf{B}^t\mathbf{W}^{t+1} - (\mathbf{B}^*\mathbf{W}^* + \mathbf{S}^*))\mathbf{W}^{t+1\top}\big)(\mathbf{R}^{t+1})^{-1}. \tag{D.60}$$

Hence,

$$\begin{aligned}\mathrm{dist}\left(\mathbf{B}^{t+1}, \mathbf{B}^*\right) &= \left\|\left(\mathbf{B}_\perp^{*\top}\mathbf{B}^t - \frac{\eta}{n}\mathbf{B}_\perp^{*\top}(\mathbf{B}^t\mathbf{W}^{t+1} - (\mathbf{B}^*\mathbf{W}^* + \mathbf{S}^*))\mathbf{W}^{t+1\top}\right)(\mathbf{R}^{t+1})^{-1}\right\| \\ &\leq \left(\left\|\mathbf{B}_\perp^{*\top}\mathbf{B}^t(\mathbf{I}_k - \frac{\eta}{n}\mathbf{W}^{t+1}\mathbf{W}^{t+1\top})\right\| + \frac{\eta}{n}\left\|\mathbf{B}_\perp^{*\top}\mathbf{S}^*\mathbf{W}^{t+1\top}\right\|\right)\left\|(\mathbf{R}^{t+1})^{-1}\right\|.\end{aligned} \tag{D.61}$$

Considering the first term in equation D.61, we have

$$\begin{aligned}\lambda_{\max}\Big(\frac{1}{n}\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}\Big) &= \frac{1}{n}\left\|\mathbf{B}^{t\top}(\mathbf{B}^*\mathbf{W}^* + \mathbf{S}^*)\right\|^2 \\ &\leq \frac{2}{n}\left\|\mathbf{B}^{t\top}\mathbf{B}^*\mathbf{W}^*\right\|^2 + \frac{2}{n}\left\|\mathbf{B}^{t\top}\mathbf{S}^*\right\|^2 \\ &\leq \frac{2}{n}\left\|\mathbf{W}^*\right\|^2 + \frac{2}{n}\left\|\mathbf{S}^*\right\|^2 \qquad \text{(D.62)} \\ &\leq 2(r_w^2 + r_s^2). \qquad \text{(D.63)}\end{aligned}$$

Thus, let $\eta \leq \frac{1}{2(r_w^2 + r_s^2)}$ and we have

$$\begin{aligned}\left\|\mathbf{I}_k - \frac{\eta}{n}\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}\right\| &\leq 1 - \frac{\eta}{n}\lambda_{\min}(\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}) \qquad \text{(D.64)} \\ &= 1 - \frac{\eta}{n}\lambda_{\min}(\mathbf{B}^{t\top}(\mathbf{B}^*\mathbf{W}^* + \mathbf{S}^*)(\mathbf{B}^*\mathbf{W}^* + \mathbf{S}^*)^\top\mathbf{B}^t) \qquad \text{(D.65)} \\ &\leq 1 - \frac{\eta}{n}\sigma^2_{\min}(\mathbf{W}^{*\top}\mathbf{B}^{*\top}\mathbf{B}^t) + \frac{2\eta}{n}\sigma_{\max}(\mathbf{B}^{t\top}\mathbf{B}^*\mathbf{W}^*\mathbf{S}^{*\top}\mathbf{B}^t) \\ &\quad - \frac{\eta}{n}\sigma^2_{\min}(\mathbf{B}^{t\top}\mathbf{S}^*) \qquad \text{(D.66)} \\ &\leq 1 - \frac{\eta}{n}\sigma^2_{\min}(\mathbf{W}^{*\top}\mathbf{B}^{*\top}\mathbf{B}^t) + \frac{2\eta\mu_{ws}}{n} - \frac{\eta}{n}\sigma^2_{\min}(\mathbf{B}^{t\top}\mathbf{S}^*) \qquad \text{(D.67)} \\ &\leq 1 - \eta\bar{\sigma}^2_{\min}(\mathbf{W}^*)\,\sigma^2_{\min}(\mathbf{B}^{*\top}\mathbf{B}^t) + \frac{2\eta\mu_{ws}}{n}, \qquad \text{(D.68)}\end{aligned}$$

where equation D.66 follows by applying Weyl's inequality.

We next show by induction that the iterates remain within the initial neighborhood of $\mathbf{B}^*$, i.e., $\operatorname{dist}(\mathbf{B}^t, \mathbf{B}^*) \leq \operatorname{dist}(\mathbf{B}^0, \mathbf{B}^*)$ for all $t = 0, \ldots, T$. The claim trivially holds for $t = 0$. Suppose that it holds at iteration t for some $t = 0, \ldots, T-1$, then

$$\sigma_{\min}^2(\mathbf{B}^{*\top}\mathbf{B}^t) \geq 1 - \left\|\mathbf{B}_\perp^{*\top}\mathbf{B}^t\right\|^2 = 1 - \operatorname{dist}^2(\mathbf{B}^t, \mathbf{B}^*) \geq 1 - \operatorname{dist}^2(\mathbf{B}^0, \mathbf{B}^*) =: c_0. \tag{D.69}$$

Next, we consider the second term in equation D.61,

$$\begin{aligned}
\frac{\eta}{n}\left\|\mathbf{B}_\perp^{*\top}\mathbf{S}^*\mathbf{W}^{t+1\top}\right\| &= \frac{\eta}{n}\left\|\mathbf{B}_\perp^{*\top}\mathbf{S}^*(\mathbf{S}^{*\top} + \mathbf{W}^{*\top}\mathbf{B}^{*\top})\mathbf{B}^t\right\| \\
&\leq \frac{\eta}{n}\left\|\mathbf{B}_\perp^{*\top}\mathbf{S}^*\mathbf{S}^{*\top}\mathbf{B}^t\right\| + \frac{\eta}{n}\left\|\mathbf{B}_\perp^{*\top}\mathbf{S}^*\mathbf{W}^{*\top}\mathbf{B}^{*\top}\mathbf{B}^t\right\| \\
&\leq \frac{\eta\mu_{bs}r_s}{\sqrt{n}} + \frac{\eta\mu_{ws}}{n},
\end{aligned} \tag{D.70}$$

where equation D.70 holds based on Assumption 4 and Cauchy-Schwarz inequality. To bound $\|(\mathbf{R}^{t+1})^{-1}\|$, define $\mathbf{Q} := \mathbf{B}^t\mathbf{W}^{t+1} - (\mathbf{B}^*\mathbf{W}^* + \mathbf{S}^*)$ and note that

$$\mathbf{B}^{t\top}\mathbf{Q} = \mathbf{B}^{t\top}\left(\mathbf{B}^t\mathbf{B}^{t\top}(\mathbf{B}^*\mathbf{W}^* + \mathbf{S}^*) - (\mathbf{B}^*\mathbf{W}^* + \mathbf{S}^*)\right) = -\mathbf{B}^{t\top}\mathbf{B}_\perp^t\mathbf{B}_\perp^{t\top}(\mathbf{B}^*\mathbf{W}^* + \mathbf{S}^*) = \mathbf{0}, \tag{D.71}$$

then we have

$$\begin{aligned}
\mathbf{R}^{t+1\top}\mathbf{R}^{t+1} &= \tilde{\mathbf{B}}^{t+1\top}\tilde{\mathbf{B}}^{t+1} \\
&= \mathbf{B}^{t\top}\mathbf{B}^t - \frac{\eta}{n}(\mathbf{B}^{t\top}\mathbf{Q}\mathbf{W}^{t+1\top} + \mathbf{W}^{t+1}\mathbf{Q}^\top\mathbf{B}^t) + \frac{\eta^2}{n^2}\mathbf{W}^{t+1}\mathbf{Q}^\top\mathbf{Q}\mathbf{W}^{t+1\top} \\
&= \mathbf{B}^{t\top}\mathbf{B}^t + \frac{\eta^2}{n^2}\mathbf{W}^{t+1}\mathbf{Q}^\top\mathbf{Q}\mathbf{W}^{t+1\top}.
\end{aligned} \tag{D.72}$$

By Weyl's inequality, we have

$$\sigma_{\min}^2(\mathbf{R}^{t+1}) \geq 1 + \frac{\eta^2}{n^2}\lambda_{\min}(\mathbf{W}^{t+1}\mathbf{Q}^\top\mathbf{Q}\mathbf{W}^{t+1\top}) \geq 1, \tag{D.73}$$

where equation D.73 holds since $\mathbf{W}^{t+1}\mathbf{Q}^\top\mathbf{Q}\mathbf{W}^{t+1\top}$ is positive semi-definite. Finally, combining equation D.68, equation D.69, equation D.70, equation D.73, we have

$$\operatorname{dist}\left(\mathbf{B}^{t+1}, \mathbf{B}^*\right) \leq (1 - \beta\eta)\operatorname{dist}\left(\mathbf{B}^t, \mathbf{B}^*\right) + \frac{\eta\mu_{bs}r_s}{\sqrt{n}} + \frac{\eta\mu_{ws}}{n}, \tag{D.74}$$

with probability at least $1 - e^{-C_1(s\log d + \log n)} - e^{-C_2(k+\log n)} \geq 1 - 2e^{-\tilde{C}\min\{s\log d + \log n,\, k + \log n\}}$, where $\beta = c_0\bar{\sigma}_{\min}^2(\mathbf{W}^*) - 2\mu_{ws}/n$ and $C_1, C_2, \tilde{C}$ are some constants. Moreover, since $\eta \leq \frac{1}{2(r_w^2 + r_s^2)}$, $\beta \leq \bar{\sigma}_{\min}^2(\mathbf{W}^*) \leq r_w^2$, we have $0 < \beta\eta \leq \frac{1}{2}$. Using the induction hypothesis and the condition $\frac{\mu_{bs}r_s}{\sqrt{n}\beta} + \frac{\mu_{ws}}{n\beta} \leq \operatorname{dist}(\mathbf{B}^0, \mathbf{B}^*)$, we obtain

$$\begin{aligned}
\operatorname{dist}\left(\mathbf{B}^{t+1}, \mathbf{B}^*\right) &\leq (1 - \beta\eta)\operatorname{dist}\left(\mathbf{B}^0, \mathbf{B}^*\right) + \frac{\eta\mu_{bs}r_s}{\sqrt{n}} + \frac{\eta\mu_{ws}}{n} && \text{(D.75)} \\
&\leq (1 - \beta\eta)\operatorname{dist}\left(\mathbf{B}^0, \mathbf{B}^*\right) + \eta\beta\operatorname{dist}\left(\mathbf{B}^0, \mathbf{B}^*\right) && \text{(D.76)} \\
&= \operatorname{dist}\left(\mathbf{B}^0, \mathbf{B}^*\right). && \text{(D.77)}
\end{aligned}$$

Therefore, the induction is complete, and $\operatorname{dist}(\mathbf{B}^t, \mathbf{B}^*) \leq \operatorname{dist}(\mathbf{B}^0, \mathbf{B}^*)$ holds for all $t = 0, \ldots, T$. Since equation D.74 holds at each round $t = 0, \ldots, T-1$, recursively applying it for T iterations and taking a union bound over these rounds gives

$$\operatorname{dist}\left(\mathbf{B}^T, \mathbf{B}^*\right) \leq (1 - \beta\eta)^T\operatorname{dist}\left(\mathbf{B}^0, \mathbf{B}^*\right) + \frac{\mu_{bs}r_s}{\sqrt{n}\beta} + \frac{\mu_{ws}}{n\beta}. \tag{D.78}$$

with probability at least $1 - 2Te^{-\tilde{C}\min\{s\log d + \log n,\ k + \log n\}}$. This concludes the proof. □

We now restate Corollary 1 and provide its detailed proof.

Corollary 1 *Under the setting of Theorem 1, let*

$$\Lambda := \big(1 - \mathrm{dist}^2(\mathbf{B}^0, \mathbf{B}^*)\big)\bar{\sigma}^2_{\min}(\mathbf{W}^*). \tag{D.79}$$

For any target accuracy $0 < \varepsilon \leq 1$, suppose the number of communication rounds satisfies

$$T \geq \frac{2}{\eta\Lambda} \log \frac{2\lceil r_w + r_s \rceil}{\varepsilon}, \tag{D.80}$$

and

$$\mu_{bs} \leq \frac{\sqrt{n}\Lambda\varepsilon}{8r_s\lceil r_w + r_s \rceil}, \qquad \mu_{ws} \leq \frac{n\Lambda\varepsilon}{8\lceil r_w + r_s \rceil}. \tag{D.81}$$

Then, with the same probability as in Theorem 1, the learned shared representation satisfies

$$\mathrm{dist}(\mathbf{B}^T, \mathbf{B}^*) \leq \varepsilon. \tag{D.82}$$

Furthermore, for every client $i \in [n]$, the personalized components satisfy

$$\left\|\mathbf{B}^T\mathbf{w}_i^{T+1} - \mathbf{B}^*\mathbf{w}_i^*\right\| \leq \varepsilon, \qquad \left\|\mathbf{s}_i^{T+1} - \mathbf{s}_i^*\right\| \leq \varepsilon. \tag{D.83}$$

Proof of Corollary 1.

From Theorem 1,

$$\mathrm{dist}\left(\mathbf{B}^T, \mathbf{B}^*\right) \leq (1 - \beta\eta)^T \,\mathrm{dist}\left(\mathbf{B}^0, \mathbf{B}^*\right) + \varepsilon_0,$$

where

$$\varepsilon_0 = \frac{\mu_{bs} r_s}{\sqrt{n}\beta} + \frac{\mu_{ws}}{n\beta},$$

and

$$\beta = \Lambda - \frac{2\mu_{ws}}{n}.$$

We first establish a lower bound on β. From the assumed bound on μ_{ws},

$$\frac{2\mu_{ws}}{n} \leq \frac{\Lambda\varepsilon}{4\lceil r_w + r_s \rceil}.$$

Since $0 < \varepsilon \leq 1$ and $\lceil r_w + r_s \rceil \geq 1$,

$$\frac{2\mu_{ws}}{n} \leq \frac{\Lambda}{4}.$$

Therefore,

$$\beta = \Lambda - \frac{2\mu_{ws}}{n} \geq \frac{3\Lambda}{4} \geq \frac{\Lambda}{2}.$$

We now bound the contraction term. Since $0 < \beta\eta < 1$, using $1 - x \leq e^{-x}$,

$$(1 - \beta\eta)^T \,\mathrm{dist}\left(\mathbf{B}^0, \mathbf{B}^*\right) \leq e^{-\beta\eta T} \,\mathrm{dist}\left(\mathbf{B}^0, \mathbf{B}^*\right) \leq e^{-\eta\Lambda T/2},$$

where we used $\mathrm{dist}\left(\mathbf{B}^0, \mathbf{B}^*\right) \leq 1$ and $\beta \geq \Lambda/2$. Based on

$$T \geq \frac{2}{\eta\Lambda} \log \frac{2\lceil r_w + r_s \rceil}{\varepsilon},$$

we obtain

$$e^{-\eta\Lambda T/2} \leq \frac{\varepsilon}{2\lceil r_w + r_s \rceil}.$$

Therefore,

$$(1 - \beta\eta)^T \,\mathrm{dist}\left(\mathbf{B}^0, \mathbf{B}^*\right) \leq \frac{\varepsilon}{2\lceil r_w + r_s \rceil}. \tag{D.84}$$

Next, we bound the neighborhood term ε_0. Using the condition on μ_{bs}, μ_{ws} and $\beta \geq \Lambda/2$, we have

$$\frac{\mu_{bs} r_s}{\sqrt{n}\beta} \leq \frac{\Lambda\varepsilon}{8\beta\lceil r_w + r_s \rceil} \leq \frac{\varepsilon}{4\lceil r_w + r_s \rceil},$$
$$\frac{\mu_{ws}}{n\beta} \leq \frac{\Lambda\varepsilon}{8\beta\lceil r_w + r_s \rceil} \leq \frac{\varepsilon}{4\lceil r_w + r_s \rceil}.$$

Thus,

$$\varepsilon_0 = \frac{\mu_{bs} r_s}{\sqrt{n}\beta} + \frac{\mu_{ws}}{n\beta} \leq \frac{\varepsilon}{2\lceil r_w + r_s \rceil}. \tag{D.85}$$

Combining D.84 and D.85 with Theorem 1 gives

$$\operatorname{dist}\left(\mathbf{B}^T, \mathbf{B}^*\right) \leq \frac{\varepsilon}{2\lceil r_w + r_s \rceil} + \frac{\varepsilon}{2\lceil r_w + r_s \rceil} \leq \frac{\varepsilon}{\lceil r_w + r_s \rceil} \leq \varepsilon. \tag{D.86}$$

The personalized head $\mathbf{w}_i$ and sparse residual $\mathbf{s}_i$ converge as follows:

$$\begin{aligned}
\left\|\mathbf{s}_i^{T+1} - \mathbf{s}_i^*\right\| &= \left\|(\mathbf{B}^T)_\perp (\mathbf{B}^T)_\perp^\top (\mathbf{B}^* \mathbf{w}_i^* + \mathbf{s}_i^* - \mathbf{B}^T \mathbf{w}_i^T) - \mathbf{s}_i^*\right\| \\
&= \left\|(\mathbf{B}^T)_\perp (\mathbf{B}^T)_\perp^\top \mathbf{B}^* \mathbf{w}_i^* - \mathbf{B}^T (\mathbf{B}^T)^\top \mathbf{s}_i^*\right\| \\
&\leq \left\|(\mathbf{B}^T)_\perp (\mathbf{B}^T)_\perp^\top \mathbf{B}^* \mathbf{w}_i^*\right\| + \left\|\mathbf{B}^T (\mathbf{B}^T)^\top \mathbf{B}_\perp^* \mathbf{B}_\perp^{*\top} \mathbf{s}_i^*\right\| \\
&\leq \frac{\varepsilon}{\lceil r_w + r_s \rceil} r_w + \frac{\varepsilon}{\lceil r_w + r_s \rceil} r_s && \text{(D.87)} \\
&\leq \varepsilon, && \text{(D.88)} \\
\left\|\mathbf{B}^T \mathbf{w}_i^{T+1} - \mathbf{B}^* \mathbf{w}_i^*\right\| &= \left\|\mathbf{B}^T (\mathbf{B}^T)^\top (\mathbf{B}^* \mathbf{w}_i^* + \mathbf{s}_i^*) - \mathbf{B}^* \mathbf{w}_i^*\right\| \\
&= \left\|\mathbf{B}^T (\mathbf{B}^T)^\top \mathbf{s}_i^* - (\mathbf{B}^T)_\perp (\mathbf{B}^T)_\perp^\top \mathbf{B}^* \mathbf{w}_i^*\right\| \\
&\leq \left\|\mathbf{B}^T (\mathbf{B}^T)^\top \mathbf{B}_\perp^* \mathbf{B}_\perp^{*\top} \mathbf{s}_i^*\right\| + \left\|(\mathbf{B}^T)_\perp (\mathbf{B}^T)_\perp^\top \mathbf{B}^* \mathbf{w}_i^*\right\| \\
&\leq \frac{\varepsilon}{\lceil r_w + r_s \rceil} r_s + \frac{\varepsilon}{\lceil r_w + r_s \rceil} r_w && \text{(D.89)} \\
&\leq \varepsilon, && \text{(D.90)}
\end{aligned}$$

where equation D.87 and equation D.89 hold by applying the Cauchy-Schwarz inequality and $\operatorname{dist}\left(\mathbf{B}^T, \mathbf{B}^*\right) \leq \frac{\varepsilon}{\lceil r_w + r_s \rceil}$. This completes the proof. □

Theorem 2 *Suppose the learned shared representation $\hat{\mathbf{B}}$ satisfies $\operatorname{dist}(\hat{\mathbf{B}}, \mathbf{B}^*) \leq \frac{\varepsilon}{r_w + r_s}$. For a new client i, if the local sample size satisfies $m_{\text{new}} \geq C(s \log d + k)$ for some constant $C > 0$, then with high probability,*

$$\left\|\hat{\mathbf{B}} \mathbf{w}_i - \mathbf{B}^* \mathbf{w}_i^*\right\| \leq \varepsilon, \qquad \|\mathbf{s}_i - \mathbf{s}_i^*\| \leq \varepsilon. \tag{D.91}$$

Proof of Theorem 2.

For the new client, we suppose that the residual sparsity condition in Assumption 3 also holds for its local fitting step, i.e., $\mathbf{B}^* \mathbf{w}_i^* + \mathbf{s}_i^* - \hat{\mathbf{B}} \mathbf{w}_i^0$ is s-sparse. Therefore, by the single-client sparse-recovery argument used in Lemma 2, the local sparse recovery problem exactly recovers

$$\tilde{\mathbf{s}}_i = \mathbf{B}^* \mathbf{w}_i^* + \mathbf{s}_i^* - \hat{\mathbf{B}} \mathbf{w}_i^0$$

with high probability as long as $m_{\text{new}} \geq C_1 s \log d$ for some constant $C_1 > 0$. After projecting $\tilde{\mathbf{s}}_i$ onto the orthogonal complement of $\hat{\mathbf{B}}$, we obtain

$$\mathbf{s}_i = \hat{\mathbf{B}}_\perp \hat{\mathbf{B}}_\perp^\top (\mathbf{B}^* \mathbf{w}_i^* + \mathbf{s}_i^* - \hat{\mathbf{B}} \mathbf{w}_i^0) = \hat{\mathbf{B}}_\perp \hat{\mathbf{B}}_\perp^\top (\mathbf{B}^* \mathbf{w}_i^* + \mathbf{s}_i^*), \tag{D.92}$$

Given $\mathbf{s}_i$, by the corresponding single-client concentration argument used in Lemma 3, with high probability, the personalized head is obtained by

$$\mathbf{w}_i = \hat{\mathbf{B}}^\top (\mathbf{B}^* \mathbf{w}_i^* + \mathbf{s}_i^*).$$

as long as $m_{\text{new}} \geq C_2 k$ for some constant $C_2 > 0$. Combining the sample requirements for the sparse residual and the personalized head gives $m_{\text{new}} \geq C(s \log d + k)$ for some constant $C > 0$.

Then by the condition $\mathrm{dist}(\hat{\mathbf{B}}, \mathbf{B}^*) \leq \frac{\varepsilon}{r_w + r_s}$, we have

$$\begin{aligned}
\|\mathbf{s}_i - \mathbf{s}_i^*\| &= \left\|\hat{\mathbf{B}}_\perp \hat{\mathbf{B}}_\perp^\top (\mathbf{B}^* \mathbf{w}_i^* + \mathbf{s}_i^*) - \mathbf{s}_i^*\right\| \\
&= \left\|\hat{\mathbf{B}}_\perp \hat{\mathbf{B}}_\perp^\top \mathbf{B}^* \mathbf{w}_i^* - \hat{\mathbf{B}}\hat{\mathbf{B}}^\top \mathbf{s}_i^*\right\| && \text{(D.93)} \\
&\leq \left\|\hat{\mathbf{B}}_\perp \hat{\mathbf{B}}_\perp^\top \mathbf{B}^* \mathbf{w}_i^*\right\| + \left\|\hat{\mathbf{B}}\hat{\mathbf{B}}^\top \mathbf{s}_i^*\right\| \\
&\leq \frac{\varepsilon}{r_w + r_s} r_w + \frac{\varepsilon}{r_w + r_s} r_s && \text{(D.94)} \\
&\leq \varepsilon. && \text{(D.95)}
\end{aligned}$$

As for the convergence on the heads, we have:

$$\begin{aligned}
\left\|\hat{\mathbf{B}}\mathbf{w}_i - \mathbf{B}^* \mathbf{w}_i^*\right\| &= \left\|\hat{\mathbf{B}}\hat{\mathbf{B}}^\top (\mathbf{B}^* \mathbf{w}_i^* + \mathbf{s}_i^*) - \mathbf{B}^* \mathbf{w}_i^*\right\| \\
&= \left\|-\hat{\mathbf{B}}_\perp \hat{\mathbf{B}}_\perp^\top \mathbf{B}^* \mathbf{w}_i^* + \hat{\mathbf{B}}\hat{\mathbf{B}}^\top \mathbf{s}_i^*\right\| \\
&\leq \left\|\hat{\mathbf{B}}_\perp \hat{\mathbf{B}}_\perp^\top \mathbf{B}^* \mathbf{w}_i^*\right\| + \left\|\hat{\mathbf{B}}\hat{\mathbf{B}}^\top \mathbf{s}_i^*\right\| \\
&\leq \frac{\varepsilon}{r_w + r_s} r_w + \frac{\varepsilon}{r_w + r_s} r_s && \text{(D.96)} \\
&\leq \varepsilon. && \text{(D.97)}
\end{aligned}$$

Thus we complete the proof. □