# Global Convergence of Federated Learning for Mixed Regression

Lili Su, Jiaming Xu, and Pengkun Yang

***Abstract*— This paper studies the problem of model training under Federated Learning when clients exhibit cluster structures. We contextualize this problem in mixed regression, where each client has limited local data generated from one of $k$ unknown regression models. We design an algorithm that achieves global convergence from any arbitrary initialization, and works even when local data volume is highly unbalanced – there could exist clients that contain $O(1)$ data points only. Our algorithm is intended for the scenario where the parameter server can recruit one client per cluster referred to as "anchor clients", and each anchor client possesses $\tilde{\Omega}(k)$ data points. Our algorithm first runs moment descent on this set of anchor clients to obtain coarse model estimates. Subsequently, every client alternately estimates its cluster labels and refines the model estimates based on FedAvg or FedProx. A key innovation in our analysis is a uniform estimate of the clustering errors, which we prove by bounding the Vapnik–Chervonenkis dimension of general polynomial concept classes based on the theory of algebraic geometry.**

***Index Terms*— Federated Learning, mixed regression, clustering, global convergence, empirical process.**

## I. Introduction

FEDERATED learning (FL) [1] enables a massive number of clients to collaboratively train models without disclosing raw data. Data heterogeneity includes non-IID local distributions and unbalanced data volume. For example, smartphone users have different preferences in article categories (e.g. politics, sports or entertainment) and have a wide range of reading frequencies. In fact, distribution of the local dataset sizes is often heavy-tailed [2], [3], [4].

Based on the number of models trained, existing methods to deal with data heterogeneity can be roughly classified into three categories: a common model, fully personalized models, and clustered personalized models; see Section II for detailed discussion. Using a common model to serve highly heterogeneous clients has fundamental drawbacks; recent work [5] rigorously quantified the heterogeneity level that the common model can tolerate with. Fully personalized models [6], [7] often do not come with performance guarantees as the underlying optimization problem is generally hard to solve. In this work, we focus on clustered personalized models [8], i.e., clients within the same cluster share the same underlying model and clients across clusters have relatively different underlying models. The main challenge is that the cluster identities of the clients are unknown. We target at designing algorithms that simultaneously learn the clusters and train models for each cluster.

A handful of simultaneous clustering and training algorithms have been proposed [9], [10], [11], [12], mostly of which are heuristic and lack of convergence guarantees [9], [10], [11]. Towards formal assurance, [12] studied this problem through the lens of statistical learning yet postulated a number of strong assumptions such as the initial model estimates are very close to the true ones, linear models, strong convexity, balanced and high-volume of local data. Their numerical results [12] suggested that sufficiently many random initializations would lead to at least one good realization satisfying the required closeness assumption. However, the necessary number of random initializations scales exponentially in both the input dimension and the number of clusters. Besides, in practice, it is hard to recognize and winnow out good initialization. In this work, following [12], we adopt a statistical learning setup. In particular, we contextualize our problem as the canonical mixed non-parametric regression with a known feature map, where each client has a set of local data generated from one of $k$ unknown regression models. We would like to extend our results to the general convex/non-convex loss functions in future work. Departing from standard mixed regression, in which each client keeps one data point only [13], [14], in our problem, the sizes of local datasets can vary significantly across clients. Our algorithm is intended for the scenario where the parameter server can recruit one client per cluster referred to as "anchor clients", and each anchor client possesses $\tilde{\Omega}(k)$ data points. The non-anchor clients may contain as few as two data points and we further assume that there are $\tilde{\Omega}(d)$ clients in total. A similar mixed regression

Manuscript received 10 November 2022; revised 4 September 2023; accepted 27 June 2024. Date of publication 11 July 2024; date of current version 20 August 2024. The work of Lili Su was supported in part by the NSF Faculty Early Career Development Program (CAREER) Award under Grant CCF-2340482 and in part by the Army Research Office (ARO) under Grant W911 NF-23-2-0014. The work of Jiaming Xu was supported in part by NSF under Grant CCF-1856424 and in part by the NSF CAREER Award under Grant CCF-2144593. The work of Pengkun Yang was supported in part by NSFC under Grant 12101353 and in part by Tsinghua University Initiative Scientific Research Program. An earlier version of this paper was presented at the Proceedings of the 2022 Conference on Neural Information Processing Systems. *(Corresponding author: Lili Su.)*

Lili Su is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: l.su@northeastern.edu).

Jiaming Xu is with the Fuqua School of Business, Duke University, Durham, NC 27708 USA (e-mail: jiaming.xu868@duke.edu).

Pengkun Yang is with the Center for Statistical Science, Tsinghua University, Beijing 100190, China (e-mail: yangpengkun@tsinghua.edu.cn).

Communicated by S. Jalali, Associate Editor for Machine Learning and Statistics.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2024.3425758.

Digital Object Identifier 10.1109/TIT.2024.3425758

setup with data heterogeneity has been considered in [15] in a different context of meta-learning; the focus therein is on exploring structural similarities among a large number of tasks in centralized learning. On the technical side, their analysis only works when the covariance matrices of all the clusters are identical and each client has $\Omega(\log k)$ data points. Please refer to Remark 1 for more detailed technical comparisons.

*Contributions:* The main contributions of this work are summarized as follows.

- We design a two-phase federated learning algorithm to learn clustered personalized models in the context of mixed regression problems. In Phase 1, the parameter server runs a federated moment descent on the set of anchor clients to obtain coarse model estimates based on subspace estimation. In each global iteration of Phase 2, every client alternately estimates its cluster label and refines the model estimates based on FedAvg or FedProx. The algorithm works even when local data volume is highly unbalanced – there could exist clients that contain $O(1)$ data points only.
- We prove the global convergence of our algorithm from any initialization. The proof is built upon two key ingredients: 1) We develop a novel eigengap-free bound to control the projection errors in subspace estimation; 2) To deal with the sophisticated interdependence between the two phases and across iterations, we develop a novel uniform estimate on the clustering errors, which we derive by bounding the VC dimension of general polynomial concept classes based on the theory of algebraic geometry. Our analysis reveals that the final estimation error is dominated by the uniform deviation of the clustering errors, which is largely overlooked by the previous work.

Furthermore, we empirically evaluate our algorithm using a synthetic mixed linear regression dataset and extend its applicability beyond mixed regression to general statistical learning, evidenced by testing on the MNIST dataset. Numerically, our algorithm is comparable to the Oracle algorithm and outperforms the other benchmarks. Yet, our theoretical analysis is still limited to mixed regression. Extending our theory to general statistical learning is an important future work.

## II. Related Work

FedAvg [1] is a widely adopted FL algorithm due to its simplicity and low communication cost. However, severe data heterogeneity could lead to unstable training trajectories and land in suboptimal models [16], [17], [18]. Based on the number of models trained, existing methods to deal with data heterogeneity can be roughly classified into three categories.

### A. A Common Model

To limit the negative impacts of data heterogeneity on the obtained common model, a variety of techniques based on variance reduction [16], [18], [19] and normalization [20] have been introduced. Their convergence results mostly are derived under strong technical assumptions such as bounded gradient and/or bounded Hessian dissimilarity which may not hold when the underlying truth in the data generation is taken into account [16], [18], [19]. In fact, none of them strictly outperform others in different instances of data heterogeneity [21]. Besides, the generalization errors of the common model with respect to local data are mostly overlooked except for a recent work [5], which shows that the common model can tolerate a moderate level of model heterogeneity.

### B. Fully Personalized Models

Fully personalized methods are more general than clustered federated learning approaches in the sense that they do not require the existence of cluster structures among the trained models. Nevertheless, existing work imposed stringent technical requirements to derive assured performance characterization such as convergence rates and final errors/accuracy.

Federated Multi-Task Learning (MTL) was proposed in [6] wherein different models are learned for each of the massive population of clients [6], [7]. Despite recent efforts, the convergence behaviors of Federated MTL are far from well-understood because the objective is not jointly convex in the model parameters and the model relationships [6], [22], [23]. Specifically, [6] focused on solving the subproblem of updating the model parameters only. Even in the centralized setting, convergence is only shown under rather restricted assumptions such as equal dataset sizes for different tasks (i.e. balanced local data) [23] and small number of common features [22]. Moreover, the average excess error rather than the error of individual tasks is shown to decay with the dominating term $O(1/\sqrt{n})$, where $n$ is the size of the balanced local dataset [23]. Despite recent progress [24], [25] in the centralized training, their results are mainly for linear representation learning with equal data volume of different tasks, which precludes the applicability of their results to the real-world setting wherein the distributions of the local data volume are often heavy-tailed and there might exist clients whose local data volume $n_i$ is small. Parallel to Federated MTL, model personalization is also studied under the Model-Agnostic Meta-Learning (MAML) framework [26], [27] where the global objective is modified to account for the cost of fine-tuning a global model at individual clients. Empirically, MAML-based methods are observed to fail to train models with low generalization errors [28, Appendix 1]. Theoretically, those approaches generally yield complicated non-convex objectives, making even heuristic guarantees hard to ensure; the convergence is shown to stationary points only [26], [27], [28].

### C. Clustered Personalized Models

Clustered Federated Learning (CFL) [8], [9], [10], [11], [12], [29] can be viewed as a special case of Federated MTL where tasks across clients form cluster structures. In addition to the algorithms mentioned in Section I (i.e., the algorithms that simultaneously learn clusters and models for each cluster), other attempts have been made to integrate clustering with model training. [8] hierarchically clustered the clients in a post-processing fashion. To recover the $k$ clusters, $\Omega(k)$ empirical risk minimization problems need to be solved sequentially – which is time-consuming. [29] proposed a modular algorithm that contains one-shot clustering stage, followed by $k$ individual adversary-resilient model training. Their algorithm scales poorly in the input dimension, and requires local datasets to be balanced and sufficiently large (i.e., $n \geq d^2$). Moreover, each client sequentially solves two empirical risk minimization problems. To utilize information across different clusters, [11] proposed soft clustering and provided numerical evidence on MNIST and Fashion-MNIST datasets. Soft clustering was later on formally analyzed in [7], which required that the covariate distributions $P_i(\mathbf{x})$ are the same for all clusters, and that the local gradients are uniformly bounded. These two assumptions are the key enablers for knowledge transfer across clusters. Neither of them is assumed in our work. The novelty of our work consists in providing theoretical convergence bounds even when no good initialization is available and in allowing highly unbalanced local datasets such as O(1) data points at some clients. Some key limitations of our work are: We contextualize the statistical learning problems as mixed regressions with the known feature map $\phi$. When $\phi$ is the identity mapping, our setup reduces to mixed linear regressions. We require that the parameter server can successfully recruit a set of anchor clients that cover all clusters, and that each client possesses $\tilde{\Omega}(k)$ data points. In addition, we assume that there are $\tilde{\Omega}(d)$ clients in total. We would like to extend the acquired insights of this work to the more general non-convex setting in future work.

Notably, after the posting of our work, [30] independently proposes an algorithm that also can converge from any initialization. Nevertheless, their analysis requires strong convexity, a high volume of local data (with $\Omega(\text{poly}(d))$ data points at each client), and re-sampling of fresh data at each iteration. In contrast, our algorithm works as long as there exist anchor clients with $\tilde{\Omega}(k)$ that cover all clusters. Oftentimes $k \ll d$ holds in practice. Moreover, our analysis even accommodates clients that contain $O(1)$ data points only. Thus our requirement on the local data volume is much weaker.

## III. Problem Formulation

A FL system consists of a parameter server (PS) and $M$ clients. Each client $i \in [M]$ keeps a dataset $\mathcal{D}_i = \{(x_{ij}, y_{ij})\}_{j=1}^{n_i}$ that are generated from one of $k$ unknown regression models. Let $N = \sum_{i=1}^{M} n_i$. The local datasets are highly unbalanced with varying $n_i$ across clients. If $n_i = \tilde{\Omega}(k)$, we refer to client $i$ as *anchor* client, which corresponds to an active user in practice. Anchor clients play a crucial role in our algorithm design. We consider the challenging yet practical scenario wherein a non-anchor client may have $O(1)$ data points only.

We adopt a canonical mixture model setup: For each client $i \in [M]$,

$$y_i = \phi(\boldsymbol{x}_i)\theta^*_{z_i} + \zeta_i, \tag{1}$$

where $z_i \in [k]$ is the *hidden* local cluster label, $\theta^*_1, \cdots, \theta^*_k$ are the true models of the clusters, $\phi(\boldsymbol{x}_i) \in \mathbb{R}^{n_i \times d}$ is the feature matrix with rows given by $\phi(x_{ij})$, $y_i = (y_{ij}) \in \mathbb{R}^{n_i}$ is the response vector, and $\zeta_i = (\zeta_{ij}) \in \mathbb{R}^{n_i}$ is the noise vector. Examples of the feature maps $\phi$ are polynomials (which can be highly nonlinear) and random features. When $\phi$ is the identity function, Eq.(1) reduces to the mixed linear regression model [14], [15].[1] The cluster label of client $i$ is randomly generated from one of the $k$ components from some unknown $p = (p_1, \dots, p_k)$ in probability simplex $\boldsymbol{\Delta}^{k-1}$. That is, $\mathbb{P}\{z_i = \ell\} = p_\ell$ for $\ell \in [k]$. In addition, $\|\theta^*_\ell\|_2 \leq R$ for each component. The feature covariate $\phi(x_{ij})$ is independent and sub-Gaussian with $\alpha I_d \preceq \mathbb{E}[\phi(x_{ij})\phi(x_{ij})^\top] \preceq \beta I_d$ for $\beta \geq 1$. We assume that the covariance matrix is identical within the same cluster but may vary across different clusters, i.e., $\mathbb{E}[\phi(x_{ij})\phi(x_{ij})^\top] = \Sigma_\ell$ if $z_i = \ell$. The noise $\zeta_{ij}$ is independent and sub-Gaussian with $\mathbb{E}[\zeta_{ij}] = 0$ and $\mathbb{E}[\zeta_{ij}^2] \leq \sigma^2$.

Our formulation accommodates statistical heterogeneity in feature covariates, local models, and observation noises [31]. For the identifiability of the true cluster models $\theta^*_\ell$'s, we assume a minimum proportion and a pairwise separation of clusters. Formally, let $\Delta = \min_{\ell \neq \ell'} \|\theta^*_\ell - \theta^*_{\ell'}\|_2$ and $p_{\min} = \min_{\ell \in [k]} p_\ell$. For ease of presentation, we assume the parameters $\alpha, \beta = \Theta(1)$, $\sigma/\Delta = O(1)$, and $R/\Delta = O(1)$, but our main results show explicit dependencies on these parameters. Note that even under these assumptions, we still allow $R$, $\Delta$, $\sigma$ to scale with model dimension $d$. Also, the assumption $R/\Delta = O(1)$ basically requires the radius of $\theta^*_\ell$'s is on the same scale as their pairwise separation. It rules out the extreme setting where $\theta^*_\ell$'s themselves are extremely large while their pairwise separations are tiny.

*Notations:* Let $[n] \triangleq \{1, \dots, n\}$. For two sets $A$ and $B$, let $A \ominus B$ denote the symmetric difference $(A - B) \cup (B - A)$. We use standard asymptotic notation: for two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ (or $a_n \lesssim b_n$) if $a_n \leq C b_n$ for some constant $C$ and sufficiently large $n$; $a_n = \Omega(b_n)$ (or $a_n \gtrsim b_n$) if $b_n = O(a_n)$; $a_n = \Theta(b_n)$ (or $a_n \asymp b_n$) if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$; Poly-logarithmic factors are hidden in $\tilde{\Omega}$. Given a matrix $A \in \mathbb{R}^{n \times d}$, let $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ denote its singular value decomposition, where $r = \min\{n, d\}$, $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$ are the singular values, and $u_i$ ($v_i$) are the corresponding left (right) singular vectors. We call $U = [u_1, u_2, \dots, u_k]$ as the top-$k$ left singular matrix of $A$. Let $\text{span}(U) = \text{span}\{u_1, \dots, u_k\}$ denote the $k$-dimensional subspace spanned by $\{u_1, \dots, u_k\}$.

## IV. Main Results

We propose a two-phase FL algorithm that enables clients to learn the model parameters $\theta^*_1, \dots, \theta^*_k$ and their clusters simultaneously:

*(i) Coarse estimation via FedMD.* Run the federated moment descent algorithm to obtain coarse estimates of model parameters $\theta^*_i$'s.

*(ii) Fine-tuning via iterative FedX+clustering.* In each iteration, each client first estimates its cluster label and then refines its local model estimate via either FedAvg or FedProx (which we refer to as FedX) [1], [16].

FedMD and FedX+clustering are detailed in the pseudocode in Phase 1 and Phase 2, respectively.

[1] Detailed comparison to previous work [14], [15] can be found in Remark 1.

### *A. Federated Moment Descent*

With multiple clusters and sub-Gaussian features, simple procedures such as power method will no longer provide a reasonably good coarse estimation. The reasons are two-fold: 1) With sub-Gaussian features, it is difficult to construct a matrix whose leading eigenspace approximately aligns with the space spanned by the true model parameters $(\theta_1^*, \dots, \theta_k^*)$; 2) Even this is achievable, there still remains significant ambiguity in determining the model parameters from their spanned subspace.

The key idea of the first phase of our algorithm is to leverage the existence of anchor clients. Specifically, the PS chooses a set $H$ of $m_H$ anchor clients uniformly at random. Each selected anchor client $i \in H$ maintains a sequence of estimators $\{\theta_{i,t}\}$ that approaches $\theta^*_{z_i}$, achieving $\|\theta_{i,t} - \theta^*_{z_i}\|_2 \le \epsilon\Delta$ for some small constant $\epsilon > 0$ when $t$ is sufficiently large.

At high-level, we hope to have $\theta_{i,t}$ move along a well calibrated direction $r_{i,t}$ that decreases the residual estimation error $\|\Sigma_{z_i}^{1/2}(\theta^*_{z_i} - \theta_{i,t})\|_2^2$, i.e., the variance of the residual $\langle \phi(x_{ij}), \theta^* - \theta_{i,t}\rangle$. As such, we like to choose $r_{i,t}$ to be positively correlated with $\Sigma_{z_i}(\theta^*_{z_i} - \theta_{i,t})$. However, to estimate $\Sigma_{z_i}(\theta^*_{z_i} - \theta_{i,t})$ solely based on the local data of anchor client $i$, it requires $n_i = \tilde\Omega(d)$, which is unaffordable in typical FL systems with high model dimension $d$ and limited local data. To resolve the curse of dimensionality, we decompose the estimation task at each chosen anchor client into two subtasks: we first estimate a $k$-dimensional subspace that $\Sigma_{z_i}(\theta^*_{z_i} - \theta_{i,t})$ lies in by "pooling" local datasets across sufficiently many non-anchor clients; then we project the local data of anchor client $i$ onto the estimated subspace and reduce the estimation problem from $d$-dimension to $k$-dimension.

The precise description of our Phase 1 procedure is given as below. For ease of notation, let $\varepsilon(x, y, \theta) \triangleq (y - \langle \phi(x), \theta\rangle)\phi(x)$.

In Step 9, PS estimates the subspace that the residual estimation errors $\{\Sigma_\ell(\theta^*_\ell - \theta_{i,t})\}_{\ell=1}^k$ lie in, in collaboration with clients in $\mathcal{S}_t$. In particular, for each anchor client $i \in H$, define

$$Y_{i,t} = \frac{1}{m}\sum_{i' \in \mathcal{S}_t} \varepsilon(x_{i'1}, y_{i'1}, \theta_{i,t})\varepsilon(x_{i'2}, y_{i'2}, \theta_{i,t})^\top,$$

where $(x_j, y_j)$ and $(\tilde x_j, \tilde y_j)$ are two data points on client $j$. We approximate the subspace spanned by $\{\Sigma_\ell(\theta^*_\ell - \theta_{i,t})\}_{\ell=1}^k$ via that spanned by the top-$k$ left singular vectors of $Y_{i,t}$. To compute the latter, we adopt the following multi-dimensional generalization of the power method, known as *orthogonal iteration* [32, Section 8.2.4]. In general, given a symmetric matrix $Y \in \mathbb{R}^{d\times d}$, the orthogonal iteration generates a sequence of matrices $Q_t \in \mathbb{R}^{d\times k}$ as follows: $Q_0 \in \mathbb{R}^{d\times k}$ is initialized as a random orthogonal matrix $Q_0^\top Q_0 = I$ and $YQ_t = Q_{t+1}R_{t+1}$ with QR factorization. When $t$ is large, $Q_t$ approximates the top-$k$ left singualr matrix of $Y$, provided the existence of an eigen-gap $\lambda_k > \lambda_{k+1}$. When $k = 1$, this is just the *power iteration* and we can further approximate the leading eigenvalue of $Y$ by the Raleigh quotient $Q_t^\top Y Q_t$. When $Y$ is asymmetric, by running the orthogonal iteration on $YY^\top$, we can compute the top-$k$ left singular matrix of $Y$.

**Phase 1** Federated Moment Descent (FedMD)

```
1  Input: $m_H, k, m, n_H, T, T_1, T_2 \in \mathbb{N}$, $\alpha, \beta, \epsilon, \Delta \in \mathbb{R}$,
      $\theta_0 \in \mathbb{R}^d$ with $\|\theta_0\|_2 \le R$
2  Output: $\hat\theta_1, \dots, \hat\theta_k$
3  PS chooses a set $H$ of $m_H$ anchor clients;
4  for each anchor client $i \in H$ do
5      $\theta_{i,0} \leftarrow \theta_0$;
6  for $t = 0, 1, \dots, T-1$ do
7      PS selects a set $\mathcal{S}_t$ of $m$ clients from
         $[M] \setminus \left(H \cup \left(\cup_{\tau=0}^{t-1}\mathcal{S}_\tau\right)\right)$;
8      PS broadcasts $\{\theta_{i,t}, i \in H\}$ to all clients $i'$ in $\mathcal{S}_t$;
         /* where $\cup_{\tau=0}^{-1}\mathcal{S}_\tau = \emptyset$ */;
9      PS calls federated-orthogonal-iteration ($\mathcal{S}_t$,
         $\{\varepsilon(x_{i'1}, y_{i'1}, \theta_{i,t}), \varepsilon(x_{i'2}, y_{i'2}, \theta_{i,t})\}_{i' \in \mathcal{S}_t}$, $k$, $T_1$)
         to output $\hat U_{i,t}$ for each anchor client $i \in H$;
         /* described in Algorithm 3 */
10     PS sends $\hat U_{i,t}$ to each anchor client $i \in H$;
11     Each anchor client $i$ runs power iteration on
         $A_{i,t}A_{i,t}^\top$ for $T_2$ steps to compute the leading
         eigenvector $\hat\beta_{i,t}$ and $\hat\sigma^2_{i,t} = \hat\beta_{i,t}^\top A_{i,t}\hat\beta_{i,t}$ with $A_{i,t}$
         defined in (2);
12     At each anchor client $i$, if $\hat\sigma_{i,t} > \epsilon\alpha\Delta/\sqrt{2}$ then
13         $\theta_{i,t+1} \leftarrow \theta_{i,t} + r_{i,t}\eta_{i,t}$ and reports $\theta_{i,t+1}$ to the
             PS, where $r_{i,t} = \hat U_{i,t}\hat\beta_{i,t}$ and
             $\eta_{i,t} = \alpha\hat\sigma_{i,t}/(2\beta^2)$ ;
14     else
15         Stop updating $\theta_{i,t}$ and let $\theta_{i,\tau} \leftarrow \theta_{i,t}$ for all
             $t+1 \le \tau \le T$ for anchor client $i$;
16 PS computes the pairwise distance $\left\|\tilde\theta_{i,T} - \tilde\theta_{i',T}\right\|_2$ for
     every pair of anchor clients $i, i' \in H$, assigns them in
     the same cluster when the pairwise distance is
     smaller than $\Delta/2$, and outputs $\hat\theta_\ell$ to be the center of
     the estimated $\ell$-th cluster for $\ell \in [k]$.
```

In our setting, the orthogonal iteration can be implemented in a distributed manner in FL systems as shown in Algorithm 3 in the Appendix A-A.

In Step 11, each anchor client $i$ estimates the residual error $\Sigma_{z_i}(\theta^*_{z_i} - \theta_{i,t})$ by projecting $\varepsilon(x_{ij}, y_{ij}, \theta_{i,t})$ onto the previously estimated subspace, that is, $\hat U_{i,t}^\top \varepsilon(x_{ij}, y_{ij}, \theta_{i,t})$. This reduces the estimation from $d$-dimension to $k$-dimension and hence $\tilde\Omega(k)$ local data points suffice. Specifically, define

$$A_{i,t} = \frac{1}{n_H}\sum_{j \in \mathcal{D}_{i,t}} \left(\hat U_{i,t}^\top \varepsilon(x_{ij}, y_{ij}, \theta_{i,t})\right)\left(\hat U_{i,t}^\top \varepsilon(\tilde x_{ij}, \tilde y_{ij}, \theta_{i,t})\right)^\top, \tag{2}$$

where $\mathcal{D}_{i,t}$ consists of $2n_H$ local data points $(x_{ij}, y_{ij})$ and $(\tilde x_{ij}, \tilde y_{ij})$ freshly drawn from $\mathcal{D}_i$ at iteration $t$. Client $i$ runs the power iteration to output $\hat\beta_{i,t}$ and $\hat\sigma^2_{i,t}$ as approximations of the leading left singular vector and singular value of $A_{i,t}$.

Then anchor client $i$ updates $\theta_{i,t+1}$ by moving along the direction of the estimated residual error $r_{i,t}$ with an appropriately chosen step size $\eta_{i,t}$.

We show that $\theta_{i,T}$ is close to $\theta^*_{z_i}$ for every anchor client $i \in H$ and the outputs $\hat{\theta}_\ell$'s are close to $\theta^*_\ell$'s up to a permutation of cluster indices.

*Theorem 1:* Let $\epsilon \in (0, \alpha/8)$ be a small but fixed constant. Suppose that

$$m \geq C\frac{\beta^8(R^4+\sigma^4)d\log^3 N}{\alpha^{12}\Delta^4 p_{\min}^2\epsilon^4}, n_H \geq C\frac{\beta^8(R^4+\sigma^4)k\log^3 N}{\alpha^{12}\Delta^4\epsilon^4}$$

and

$$T \geq \frac{16\beta^2}{\alpha^2}\log\frac{2\beta R}{\alpha\epsilon\Delta}, T_1 \geq Ck\log\frac{Nd\beta R}{\alpha\epsilon\Delta}, T_2 \geq C\log\frac{Nd\beta R}{\alpha\epsilon\Delta}, \tag{3}$$

where $C > 0$ is a constant. With probability at least $1 - 10m_H T/N^{10}$, for all initialization $\theta_0$ with $\|\theta_0\|_2 \leq R$,

$$\sup_{i\in H}\left\|\theta_{i,T} - \theta^*_{z_i}\right\|_2 \leq \epsilon\Delta. \tag{4}$$

Furthermore, suppose $H \cap \{i : z_i = \ell\} \neq \emptyset$ for all $\ell \in [k]$. Then

$$d(\hat{\theta}, \theta^*) \triangleq \min_\pi \max_{\ell\in[k]}\left\|\hat{\theta}_{\pi(\ell)} - \theta^*_\ell\right\|_2 \leq \epsilon\Delta, \tag{5}$$

where $\pi$ is permutation over $[k]$.

Note that in (5) we take a minimization over permutation $\pi$, as the cluster indices are unidentifiable. Moreover, the condition $H \cap \{i : z_i = \ell\} \neq \emptyset$ for all $\ell \in [k]$ means that there exists at least one anchor client from each cluster. This condition holds with probability at least $1 - \delta$, if we choose $m_H \geq \log(k/\delta)/p_{\min}$ anchor clients uniformly at random, following the standard coupon collector's analysis.

Phase 1 uses fresh data at every iteration. In total we need $p_{\min}^{-2}\tilde{\Omega}(d)$ clients with at least two data points and at least one anchor client (with $\tilde{\Omega}(k)$ data points) from each cluster. This requirement is relatively mild, as typical FL systems have a large number of clients with $O(1)$ data points and a few anchor clients with moderate data volume.

We defer the detailed proof of Theorem 1 to Appendix A. A key step in our proof is to show the residual estimation errors $\{\Sigma_\ell(\theta^*_\ell - \theta_{i,t})\}_{\ell=1}^k$ approximately lie in $\text{span}(\hat{U}_{i,t})$. Unfortunately, the eigengap of $Y_{i,t}$ could be small, especially when $\theta_{i,t}$ gets close to $\theta^*_{z_i}$; and hence the standard Davis-Kahan theorem [33] cannot be applied. This issue is further exacerbated by the fact that the convergence rate of the orthogonal iteration also crucially depends on the eigengaps [32]. For these reasons, $\text{span}(\hat{U}_{i,t})$ may not be close to $\text{span}\{\Sigma_\ell(\theta^*_\ell - \theta_{i,t})\}_{\ell=1}^k$ at all. To resolve this issue, we develop a novel gap-free bound to show that projection errors $\hat{U}_{i,t}^\top\Sigma_\ell(\theta^*_\ell - \theta_{i,t})$ are small for every $\ell \in [k]$ (cf. Lemma 5).

*Remark 1 (Comparison to Previous Work [14], [15]):* Our algorithm is partly inspired by [14] which focuses on the noiseless mixed linear regression, but deviates in a number of crucial aspects. First, our algorithm crucially utilizes the fact that each client chosen in $\mathcal{S}_t$ has at least two data points and hence the space of the singular vectors of $\mathbb{E}[Y_{i,t}]$ is spanned by $\{\Sigma_\ell(\theta^*_\ell - \theta_{i,t})\}_{\ell=1}^k$. In contrast, [14] relies on the sophisticated method of moments which only works under the Gaussian features and requires exponential in $k^2$ many data points. Second, our algorithm crucially exploits the existence of anchor clients and greatly simplifies the moment descent algorithm in [14].

Our algorithm also bears similarities with the meta-learning algorithm in [15], which also uses clients collectively for subspace estimation and anchor clients for estimating cluster centers. However, there are several key differences. First, [15] focuses on the centralized setting and relies on one-shot estimation, under the additional assumption that the covariance matrix of features across all clusters are identical. Instead, our moment descent algorithm is iterative, is amenable to a distributed implementation in FL systems, and allows for covariance matrices varying across clusters. Second, in the fine-tuning phase, [15] uses the centralized least squares to refine the clusters estimated with anchor clients, under the additional assumption that $\Omega(\log k)$ data points for every client. In contrast, as we will show later, we use the FedX+clustering to iteratively cluster clients and refine cluster center estimation.

*Remark 2 (Data Privacy Risk):* Compared to the standard FedAvg algorithm wherein only aggregated local updates/gradients are broadcasted by the parameter server, the major step of our two-phase algorithm that may leak additional privacy is Step 8 wherein the local model estimates of the anchor clients are broadcasted to many other non-anchor clients. However, this privacy leakage is minor and can be further mitigated by a simple privacy-preserving mechanism according to the following considerations.

First, in our algorithm, each chosen non-anchor client only receives a collection of local model estimates (without ID for anchor clients) from the parameter server, it does not know which broadcasted model corresponds to which anchor client and hence cannot directly identify each individual anchor client's local true model. Second, we only choose a very few number of anchor clients (roughly on the order of the number of clusters) and in practice these anchor clients are often specially recruited by the PS; hence they can be made less concerned about privacy leakage through some incentivizing schemes. Last but not least, we can better preserve the privacy of anchor clients by broadcasting perturbed versions of their local models to each client. Specifically, fix any anchor client $i$, each non-anchor client $i'$ receives $\theta_{i',i,t}$ and $\tilde{\theta}_{i',i,t}$ that are equal to $\theta_{i,t}$ subject to two independent noise perturbations. Then for the subspace estimation in Step 9, we can replace one $\theta_{i,t}$ by $\theta_{i',i,t}$ and the other by $\tilde{\theta}_{i',i,t}$, in the definition of $Y_{i,t}$. Crucially, $Y_{i,t}$ involves an average over $m$ non-anchor clients; hence these independent noise perturbations for different $i'$ will be averaged out. Since $m$ is large, this implies that the injected random noises can be made large without deteriorating too much the accuracy of the subspace estimation, in a similar spirit as privatizing the model averaging step in FedAvg. This gives a promising pathway to maintain anchor clients' privacy; we leave rigorously analyzing its privacy guarantee as future work.

*Remark 3 (Beyond Mixed Regression):* While we present the federated moment descent algorithm in the context of mixed regression, the algorithm can be adapted to the general statistical learning setup. Recall that for mixed regression, the high-level idea of federated moment descent is to find a descent direction $r_{i,t}$ for each selected anchor client $i$ to decrease the local residual error $\|\Sigma_{z_i}^{1/2}(\theta^*_{z_i} - \theta_{i,t})\|^2$. For general statistical learning setup, such a choice of local loss may no longer be appropriate to measure the error of the local model $\theta_{i,t}$. To address this, we define $L_i(\theta) = \frac{1}{|\mathcal{D}_i|}\sum_{j\in\mathcal{D}_i} L(\theta; x, y)$ to be the local population loss, where $L(\theta; x, y)$ is the loss with $\theta$ evaluated at a data point $(x, y)$. In this manner, a natural choice of descent direction to decrease the local population loss is the negative gradient $-\nabla L_i(\theta)$. Thus, the federated moment descent algorithm can be executed with $\varepsilon$ replaced by

$$\varepsilon(x, y, \theta) \triangleq -\nabla_\theta L(\theta; x, y).$$

All other steps remain exactly the same as before. Furthermore, as we will see, Phase II is already stated under a general loss function $L$ and can be executed verbatim. In Section V-D, we have tested this adjusted algorithm in the real data experiment on the MNIST handwritten digits database. The results are promising: Our adjusted algorithm is comparable to the Oracle algorithm and outperforms other benchmarks. Yet, our current analysis is mostly focused on mixed regression. Extending these theoretical guarantees to the broader algorithm is not straightforward, so we are setting this as a goal for future work.

### *B. FedX+clustering*

At the end of Phase 1, only the selected anchor clients in $H$ obtained coarse estimates of their local true models (characterized in (4)). In Phase 2, both anchor clients in $H$ and all the other clients (anchor or not) will participate and update their local model estimates.

Phase 2 is stated in a generic form for any loss function $L(\theta, \lambda; \mathcal{D})$, where $\theta = (\theta_1, \dots, \theta_k) \in \mathbb{R}^{dk}$ is the cluster parameters, $\lambda \in \mathbf{\Delta}^{k-1}$ represents the likelihood of the cluster identity of a client, and $\mathcal{D}$ denotes the client's dataset. This generic structure covers the idea of soft clustering [11]. Note that unlike Phase 1 where each anchor client $i$ only maintains an estimate $\theta_{i,t}$ of its own model, in Phase 2, each client $i$ maintains model estimates $\theta_{i\cdot,t} = (\theta_{i1,t}, \dots, \theta_{ik,t})$ for all clusters.

In Phase 2, the local estimation at each client has a flavor of alternating minimization: It first runs a minimization step to estimate its cluster, and then runs a FedAvg or FedProx update to refine model estimates. To allow the participation of clients with $O(1)$ data points only, at every iteration the clients are allowed to reuse all local data, including those used in the first phase. Similar alternating update is analyzed in [12] yet under the strong assumption that the update in each round is over fresh data with Gaussian distribution. Moreover, the analysis therein is restricted to the setting where the model refinement at each client is via running a single gradient step, which is barely used in practice but much simpler to analyze than FedAvg or FedProx update.

In our analysis, we consider the square loss

$$L(\theta, \lambda; \mathcal{D}_i) = \frac{1}{2n_i}\sum_{\ell=1}^{k} \lambda_\ell \left\|y_i - \phi(\boldsymbol{x}_i)\theta_\ell\right\|_2^2.$$

```
**Phase 2** FedX+clustering
Input: $\theta = (\theta_1, \dots, \theta_k)$ from the output of Phase 1,
    $\eta, T'$.
Output: $\hat\theta = (\hat\theta_1, \dots, \hat\theta_k)$
PS sets $\theta_T \leftarrow \theta$.
for $t = T+1, \dots, T+T'$ do
    PS broadcasts $\theta_{t-1}$ to all clients;
    Each client $i$ estimates the likelihood of its local
      cluster label by
          $\lambda_{i,t} \leftarrow \arg\min_{\lambda\in\mathbf{\Delta}^{k-1}} L(\theta_{t-1}, \lambda; \mathcal{D}_i);$      (6)
    Each client $i$ refines its local model based on either
      FedAvg or FedProx with $L_i(\theta) = L(\theta, \lambda_{i,t}; \mathcal{D}_i)$,
      and reports the updated local parameters
      $\theta_{i\cdot,t} = (\theta_{i1,t}, \dots, \theta_{ik,t})$.
    *FedAvg-based: it runs $s$ steps of local gradient
      descent:
          $\theta_{i\cdot,t} \leftarrow \mathcal{G}_i^s(\theta_{t-1}),$   where $\mathcal{G}_i(\theta) = \theta - \eta\nabla L_i(\theta)$
      *FedProx-based: it solves the local proximal
      optimization:
          $\theta_{i\cdot,t} \leftarrow \arg\min_\theta L_i(\theta) + \frac{1}{2\eta}\|\theta - \theta_{t-1}\|_2^2$
    PS updates the global model as $\theta_t \leftarrow \sum_{i=1}^M w_i\theta_{i\cdot,t}$,
      where $w_i = n_i/N$.
```

In this context, (6) yields a vertex of the probability simplex $\lambda_{i\ell,t} = \mathbb{1}\{\ell = z_{i,t}\}$, where

$$z_{i,t} = \arg\min_{\ell\in[k]} \|y_i - \phi(\boldsymbol{x}_i)\theta_{\ell,t-1}\|_2. \tag{7}$$

The estimate $z_{i,t}$ provides a hard clustering label. Hence, in each round, only one regression model will be updated per client.

To capture the tradeoff between communication cost and statistical accuracy using FedAvg or FedProx, we introduce the following quantities from [5]:

$$\gamma \triangleq \eta \max_{i\in[M]} \frac{1}{n_i}\left\|\phi(\boldsymbol{x}_i)\right\|_2^2, \qquad \kappa \triangleq \begin{cases} \frac{\gamma s}{1-(1-\gamma)^s} & \text{for FedAvg,}\\ 1+\gamma & \text{for FedProx.}\end{cases}$$

We choose a properly small learning rate $\eta$ such that $\gamma < 1$. Here, $\kappa \geq 1$ quantifies the stability of local updates. Notably, $\kappa \approx 1$ using a relatively small $\eta$.

For the learnability of model parameters, we assume that collectively there are sufficient data in each cluster. In particular, we assume $N_\ell \gtrsim d$, where $N_\ell = \sum_{i:z_i=\ell} n_i$ denotes the number of data points in cluster $\ell$. To further characterize the *quantity skewness* (i.e., the imbalance of data partition $n = (n_1, \dots, n_M)$ across clients), we adopt the $\chi^2$-divergence, which is defined as $\chi^2(P\|Q) = \int \frac{(dP-dQ)^2}{dQ}$ for a distribution $P$ absolutely continuous with respect to a distribution $Q$. Let $\chi^2(n)$ be the chi-squared divergence between data partition $p_n$ over the clients $p_n(i) = n_i/N$ and the uniform distribution over $[M]$. Note that when data partition is balanced (i.e., $n_i = N/M$ for all $i$), it holds that $\chi^2(n) = 0$.

We have the following theoretical guarantee of Phase 2, where $s$ is the number of local steps in FedAvg. Notably, $s$ is an algorithmic parameter for FedAvg only. To recover the results for FedProx, we only need to set $s = 1$.

*Theorem 2:* Let $c_0, C_0, c_1, C_1, c_2, C_2, c, C$ denote some constants. Suppose that $k \geq 2$, $\eta \leq c_0/(\beta s)$, and $\min_{\ell \in [k]} N_\ell \geq C_0 d$. Let $\rho = \min_\ell N_\ell / N$. If $\nu \log(e/\nu) \leq c_1 \rho\alpha/(\kappa\beta)$, then with probability $1 - C_1 k e^{-d}$, for all $t \geq T+1$ and all $\theta_T$ such that $d(\theta_T, \theta^*) \leq \epsilon\Delta$, where $\epsilon \leq \frac{1}{3}\sqrt{\alpha/\beta}$, it holds that

$$d(\theta_t, \theta^*) \leq (1 - c_2 s\eta\alpha\rho/\kappa)\, d(\theta_{t-1}, \theta^*) + C_2 s\eta\sqrt{\beta}\sigma\nu \log\frac{e}{\nu}, \tag{8}$$

where

$$\nu \triangleq \frac{1}{N}\sum_{i=1}^{M} n_i p_e(n_i) + C\sqrt{\frac{dk\log k}{M}(\chi^2(n) + 1)}, \tag{9}$$

and $p_e(n_i) = 4ke^{-cn_i\alpha^2\left(1 \wedge \frac{\Delta^2}{\sigma^2}\right)^2}$. Furthermore, if $t \geq T+1$, for each client $i$, with probability at least $1 - p_e(n_i)$,

$$\left\|\hat{\theta}_{i,t} - \theta^*_{z_i}\right\|_2 \leq \epsilon\Delta \exp\left(-c_2 s\eta\alpha\rho(t-T)/\kappa\right) + \frac{C_2\sqrt{\beta}\sigma\kappa}{c_2\alpha\rho}\nu\log\frac{e}{\nu},$$

where $\hat{\theta}_{i,t}$ is client $i$'s estimate of its own model parameter at time $t$.

Notably, $\hat{\theta}_{i,t}$ is the $z_{i,t}$-th entry of $\theta_{i\cdot,t}$. Theorem 2 shows that the model estimation errors decay geometrically starting from any realization that is within a small neighborhood of $\theta^*$. The parameter $\nu$ captures the additional clustering errors injected at each iteration. It consists of two parts: the first term of (9) bounds the clustering error in expectation which diminishes exponentially in the local data size and the signal-to-noise ratio $\Delta/\sigma$; the second term bounds the uniform deviation of the clustering error across all initialization and iterations. Note that if the cluster structure were known exactly, we would get a model estimation error of $\theta^*_\ell$ scaling as $\sqrt{d/N_\ell}$. However, it turns out that this estimation error is dominated by our uniform deviation bound of the clustering error and hence is not explicitly shown in our bound (8). In comparison, the previous work [12] assumes fresh samples at each iteration by sample-splitting and good initialization independent of everything else provided a priori; hence their analysis fails to capture the influence of the uniform deviation of the clustering error.

In passing, we briefly comment on the key assumption $\nu\log(e/\nu) \lesssim \rho\alpha/(\kappa\beta)$. As aforementioned, the clustering error $\nu$ consists of two parts shown in (9): the first term decays exponentially in the local dataset size and the signal-to-noise ratio and hence is very small in most typical scenarios; the second term is on the order of $dk\log k/M$ when the quantity skewness (imbalance of data partition) is of a constant order. Finally, $\rho$ captures the imbalance of cluster sizes, which is typically of a constant order, and we can choose a small enough step size to ensure $\kappa$ is close to 1. Thus the key assumption $\nu\log(e/\nu) \lesssim \rho\alpha/(\kappa\beta)$ roughly translates to $\nu$ being a subconstant, which further means that $M$ (the number of clients) needs to be larger than $d$ (the model dimension) by polylog factors. This is often satisfied in the typical FL applications which involve a very large collection of clients.

*1) Analysis of Global Iterations:* Without loss of generality, assume the optimal permutation in (5) is identity. In this case, if $z_{i,t} = \ell$, then client $i$ will refine $\theta_{\ell,t-1}$. To prove Theorem 2, we need to analyze the global iteration of $\theta_t$. Following a similar argument to [5] with a careful examination of cluster labels, we obtain the following lemma. The proof is deferred to Appendix B-A.

*Lemma 1:* Let $\phi(\boldsymbol{x})$ be the matrix that stacks all $\phi(\boldsymbol{x}_i)$ vertically, and similarly for $y$. It holds that

$$\theta_{\ell,t} = \theta_{\ell,t-1} - \eta B\Lambda_{\ell,t}(\phi(\boldsymbol{x})\theta_{\ell,t-1} - y), \quad \ell \in [k], \tag{10}$$

where $B = \frac{1}{N}\phi(\boldsymbol{x})^\top P$, $P$ is a block diagonal matrix with $i$th block $P_i$ of size $n_i \times n_i$ given by

$$P_i = \begin{cases} \sum_{\tau=0}^{s-1}(I - \eta\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^\top/n_i)^\tau & \text{for FedAvg,} \\ [I + \eta\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^\top/n_i]^{-1} & \text{for FedProx,} \end{cases}$$

and $\Lambda_{\ell,t}$ is another block diagonal matrix with $i$th block being $\lambda_{i\ell,t}I_{n_i}$.

Lemma 1 immediately yields the evolution of estimation error. Let $\Lambda_\ell$ be the matrix with $i$th block being $\mathbb{1}\{z_i = \ell\}I_{n_i}$ representing the true client identities. Plugging model (1), the estimation error evolves as

$$\theta_{\ell,t} - \theta^*_\ell = (I - \eta K_\ell)(\theta_{\ell,t-1} - \theta^*_\ell) - \eta B\mathcal{E}_{\ell,t}(\phi(\boldsymbol{x})\theta_{\ell,t-1} - y) + \eta B\Lambda_\ell\zeta, \quad \forall \ell \in [k], \tag{11}$$

where $K_\ell = B\Lambda_\ell\phi(\boldsymbol{x})$ and $\mathcal{E}_{\ell,t} = \Lambda_{\ell,t} - \Lambda_\ell$. The estimation error is decomposed into three terms: 1) the main contribution to the decrease of estimation error; 2) the clustering error; and 3) the noisy perturbation. Let $I_\ell = \{i : z_i = \ell\}$ be the clients belonging to $\ell$th cluster, and $I_{\ell,t} = \{i : z_{i,t} = \ell\}$ be the clients with estimated label $\ell$. The indices of nonzero blocks of $\mathcal{E}_{\ell,t}$ are $I_\ell \ominus I_{\ell,t}$ indicating the clustering errors pertaining to $\ell$th cluster.

For ease of presentation, we introduce a few additional notations for the collective data over a subset of clients. Given a subset $I \subseteq [M]$ of clients, let $\phi(\boldsymbol{x}_I)$ denote the matrix that vertically stacks $\phi(\boldsymbol{x}_i)$ for $i \in I$, and we similarly use notations $y_I$ and $\zeta_I$; let $P_I$ be the matrix with diagonal blocks $P_i$ for $i \in I$. Using those notations, we have $K_\ell = \frac{1}{N}\phi(\boldsymbol{x}_{I_\ell})^\top P_{I_\ell}\phi(\boldsymbol{x}_{I_\ell})$, which differs from the usual covariance matrix by an additional matrix $P_{I_\ell}$. Therefore, the analysis of the first and third terms on the right-hand side of (11) follows from standard concentration inequalities for random matrices. In the remaining of this subsection, we focus on the second term, which is a major challenge in the analysis. The proof details are all deferred to Appendix B-B.

*Lemma 2:* There exists a universal constant $C$ such that, with probability $1 - Ce^{-d}$,

$$\|B\mathcal{E}_{\ell,t}(\phi(\boldsymbol{x})\theta_{\ell,t-1} - y)\|_2 \leq Cs(\beta d(\theta_{t-1}, \theta^*) + \sqrt{\beta}\sigma)\nu\log\frac{e}{\nu}, \quad \forall\, \ell \in [k]. \tag{12}$$

Lemma 2 aims to upper bound the error of

$$\begin{aligned} &B\mathcal{E}_{\ell,t}(\phi(\boldsymbol{x})\theta_{\ell,t-1} - y) \\ &= \frac{1}{N}\phi(\boldsymbol{x}_{S_{\ell,t}})^\top P_{S_{\ell,t}}(\phi(\boldsymbol{x}_{S_{\ell,t}})\theta_{\ell,t-1} - y_{S_{\ell,t}}), \end{aligned} \tag{13}$$

where $S_{\ell,t} = I_\ell \ominus I_{\ell,t}$. The technical difficulty arises from the involved dependency between the clustering error $S_{\ell,t}$ and the estimated parameter $\theta_{\ell,t-1}$ as estimating label $z_{i,t}$ and updating $\theta_{\ell,t-1}$ use a common set of local data.

*Proof Sketch of Lemma 2:* It follows from the definition of $z_{i,t}$ in (7) that

$$\|\phi(\boldsymbol{x}_i)\theta_{\ell,t-1} - y_i\|_2 \le \|\phi(\boldsymbol{x}_i)\theta_{z_i,t-1} - y_i\|_2, \quad \forall i \in S_{\ell,t}.$$

Then,

$$\begin{aligned} \|\phi(\boldsymbol{x}_{S_{\ell,t}})\theta_{\ell,t-1} - y_{S_{\ell,t}}\|_2^2 &= \sum_{i\in S_{\ell,t}} \|\phi(\boldsymbol{x}_i)\theta_{\ell,t-1} - y_i\|_2^2 \\ &\le \sum_{i\in S_{\ell,t}} \|\phi(\boldsymbol{x}_i)\theta_{z_i,t-1} - y_i\|_2^2 \\ &\le \sum_{i\in S_{\ell,t}} 2\left(\|\phi(\boldsymbol{x}_i)(\theta_{z_i,t-1} - \theta^*_{z_i})\|_2^2 + \|\zeta_i\|_2^2)\right) \\ &\le 2\left(d(\theta_{t-1},\theta^*)\cdot\|\phi(\boldsymbol{x}_{S_{\ell,t}})\|_2 + \|\zeta_{S_{\ell,t}}\|_2\right)^2. \end{aligned} \tag{14}$$

Hence, it suffices to upper bound $\|\phi(\boldsymbol{x}_{S_{\ell,t}})\|_2$ and $\|\zeta_{S_{\ell,t}}\|_2$ given a small estimation error $d(\theta_{t-1},\theta^*)$ from the last iteration. To this end, we show a uniform upper bound of the total clustering error $\sum_{i\in S_{\ell,t}} n_i$ by analyzing a weighted empirical process. Let

$$\begin{aligned} f^{\mathrm{I}}_{\ell,\theta}(\boldsymbol{x}_i,y_i) &\triangleq \max_{\ell'\ne\ell} \mathbb{1}\{P_{\ell\ell'}[\boldsymbol{x}_i,y_i](\theta) \ge 0\} \quad \text{for } i \in I_\ell, \\ f^{\mathrm{II}}_{\ell,\theta}(\boldsymbol{x}_i,y_i) &\triangleq \prod_{\ell'\ne\ell} \mathbb{1}\{P_{\ell'\ell}[\boldsymbol{x}_i,y_i](\theta) \ge 0\}, \quad \text{for } i \notin I_\ell. \end{aligned}$$

Using the decision rule (7), the set $S_{\ell,t}$ can be written as a function $S_\ell(\theta_{t-1})$ with

$$\mathbb{1}\{i \in S_\ell(\theta)\} = \begin{cases} f^{\mathrm{I}}_{\ell,\theta}(\boldsymbol{x}_i,y_i), & \text{if } i \in I_\ell, \\ f^{\mathrm{II}}_{\ell,\theta}(\boldsymbol{x}_i,y_i), & \text{if } i \notin I_\ell, \end{cases} \tag{15}$$

where

$$P_{\ell\ell'}[\boldsymbol{x}_i,y_i](\theta) \triangleq \|y_i - \phi(\boldsymbol{x}_i)\theta_\ell\|_2^2 - \|y_i - \phi(\boldsymbol{x}_i)\theta_{\ell'}\|_2^2.$$

Then we derive the following uniform deviation of the incorrectly clustered data points

$$\begin{aligned} &\sup_{\theta\in\mathbb{R}^{dk}} \left| \sum_{i=1}^M n_i \mathbb{1}\{i \in S_\ell(\theta)\} - \sum_{i=1}^M n_i \mathbb{P}\{i \in S_\ell(\theta)\} \right| \\ &\le CN\sqrt{\frac{dk\log k}{M}(\chi^2(n)+1)}. \end{aligned}$$

This is proved via upper bounds on the Vapnik–Chervonenkis (VC) dimensions of the binary function classes

$$\mathcal{F}^{\mathrm{I}}_\ell \triangleq \{f^{\mathrm{I}}_{\ell,\theta} : \theta \in \mathbb{R}^{dk}\}, \quad \mathcal{F}^{\mathrm{II}}_\ell \triangleq \{f^{\mathrm{II}}_{\ell,\theta} : \theta \in \mathbb{R}^{dk}\}. \tag{16}$$

Using classical results of VC dimensions, those functions are equivalently intersections of hyperplanes in ambient dimension $O(d^2)$, which yields an upper bound $O(d^2)$. However, the hyperplanes are crucially rank-restricted as the total number of parameters in $\theta$ is $dk$. We prove that the VC dimensions are at most $O(dk\log k)$ using the algebraic geometry of polynomials given by the celebrated Milnor-Thom theorem (see, e.g., [34, Theorem 6.2.1]).[2] Consequently, $\|\phi(\boldsymbol{x}_{S_{\ell,t}})\|_2, \|\zeta_{S_{\ell,t}}\|_2$ and thus (14) can be uniformly upper bounded using sub-Gaussian concentration and the union bound, concluding the proof of Lemma 2. □

### C. Global Convergence

Combining Theorem 1 and Theorem 2, we immediately deduce the global convergence from any initialization within the $\ell_2$ ball of radius $R$.

*Theorem 3:* Suppose the conditions of Theorem 1 and Theorem 2 hold. Let $\hat\theta$ be the output of our two-phase algorithm by running Phase 1 with $T \ge \frac{24\beta^2}{\alpha^2}\log\frac{2\beta R}{\alpha\Delta}$ iterations starting from any initialization $\theta_0$ with $\|\theta_0\|_2 \le R$, followed by Phase 2 with $T' \ge C\frac{\kappa}{s\eta\alpha\rho}\log\frac{\Delta}{\nu}$ iterations. Then with probability $1 - N^{-9} - Cke^{-d}$, it is true that

$$d(\hat\theta,\theta^*) \le C\frac{\sqrt{\beta}\sigma\kappa}{\alpha\rho}\nu\log\frac{e}{\nu}, \tag{17}$$

Furthermore, for each client $i$, with probability $1 - p_e(n_i)$, it holds that $\|\hat\theta_{i,T+T'} - \theta^*_{z_i}\|_2 \le C\frac{\sqrt{\beta}\sigma\kappa}{\alpha\rho}\nu\log\frac{e}{\nu}$.

To the best of our knowledge, this is the first result that proves the global convergence of clustered federated learning from any initialization. Our bound (17) reveals that the final estimation error is dominated by the clustering error captured by $\nu$, and scales linearly in $\kappa$ which characterizes the stability of local updates under FedAvg or FedProx. Moreover, Theorem 3 shows that Phase 1 converges very fast with only $\Theta(1)$ iterations and hence is relatively inexpensive in both computation and communication. Instead, the number of iterations needed for Phase 2 grows logarithmically in $\Delta/\nu$ and linearly in $\kappa/(s\eta\alpha\rho)$. Thus, by choosing $s$ relatively large while keeping $\kappa$ close to 1, FedAvg enjoys a saving of the total communication cost.

*Remark 4 (On Memory and Communication Cost):* We briefly comment on the communication and memory cost at the parameter server in Phase 1 and Phase 2 separately.

The dominating operation in Phase 1 is line 9 in the for-loop. Throughout the $T$ rounds in Phase 1, there are $T \times m_H$ calls of the function federated-orthogonal-iteration. The execution of each call will consume $\Theta(T_1 \times (dk) \times m)$ bits – the exact multiplicative constant involved depends on the chosen operating systems. Different from communication cost, the memory across iterations can be reused. The memory cost of the parameter server in Phase 1 is also dominated by the execution of line 9. The memory cost is $\Theta(dkm_H)$.

The memory cost of the parameter server in Phase 2 is $\Theta(dk)$. The communication cost of the parameter server in Phase 2 is $\Theta(dkT')$, where $T'$ is the number of global rounds in Phase 2.

## V. Experimental Results

In this section, we provide experimental results on synthetic and real data corroborating our theoretical findings.

[2] Similar applications of the Milnor-Thom theorem have been known in the literature (see e.g. [35, Theorem 2.2] and [36, Theorem 2]).

For the synthetic data experiments, we consider the mixed linear regression with $k = 3$ clusters. The true model parameter for each cluster $\theta_1^*$, $\theta_2^*$, $\theta_3^*$ are independently sampled from Gaussian distribution $\frac{2}{\sqrt{d}} * \mathcal{N}(0, I_d)$ with $d = 100$. Then we generate the local dataset $\mathcal{D}_i = \{x_{ij}, y_{ij}\}_{j=1}^{n_i}$ for each client $i$ according to the linear regression model (1), where each $x_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ and $\zeta_{ij} \overset{\text{i.i.d.}}{\sim} 0.2 * \mathcal{N}(0, 1)$.

We simulate our two-phase algorithm as follows. Phase 1 randomly selects $\lceil 3k \log k \rceil$ anchors clients and runs 5 iterations starting from a random initialization $\frac{2}{\sqrt{d}} * \mathcal{N}(0, I_d)$, followed by Phase 2 running 400 global iterations. We further adopt the following simplifications for ease of implementation. In particular, Phase 1 reuses the local data on all participating clients, and all clients including anchor clients participate in the subspace estimation subroutine in Algorithm 3. Finally, we implement all orthogonal iterations by direct singular value decomposition.

We compare the performance of our two-phase algorithm with existing FL algorithms including (1) vanilla FedAvg, (2) one-shot clustering, (3) IFCA, and (4) oracle iterative clustering. We provide a brief description of each of these algorithms below:

- Vanilla FedAvg [1]: It learns a common model, ignoring the underlying cluster structure.
- One-shot clustering [29]: This method contains three phases: First, each client estimates its underlying model based on its local data. Then, the PS clusters the locally estimated models via $k$-means. Finally, for each estimated cluster of clients, the PS runs FedAvg to obtain the model estimate for each cluster.
- IFCA [12]: This method is the same as Phase 2 of our algorithm yet requires good initialization.
- Oracle iterative clustering: This method is an ideal implementation of IFCA with the true model parameters as initialization. Clearly, the oracle iterative clustering algorithm is infeasible in practice, but we use it as a benchmark.

For each of the methods, we choose FedAvg with the number of local update steps $s = 5$. We randomly initialize our two-phase algorithm, vanilla FedAvg, and IFCA.

In the following, we consider three federated learning configurations with a total of $N = 1000$ data points but at increasing levels of data heterogeneity. The configuration for the real data experiments are described in Section V-D

### A. Balanced Local Data and Balanced Cluster Partition

In this configuration, we consider balanced local data and balanced cluster partition. Specifically, we let $M = 200$, $n_i = 50$ for $i \in [M]$, and $p_1 = p_2 = p_3 = 1/3$. That is, this configuration contains 200 clients, each with 50 data points. For each client, it belongs to one of the 3 clusters with equal probability $1/3$.

In the left panel of Fig. 1, we show the performance of our two-phase algorithm, where the second phase is based on FedAvg for different local steps $s$ or FedProx. We see that during the first 5 rounds (Phase 1), the errors quickly (exponentially with a large rate) converge to a relatively small value. Starting from iteration 6 (upon entering Phase 2), the errors further decay exponentially fast (with a smaller rate than Phase 1). These observations are consistent with our theoretical predictions. We also notice that as the number of local steps $s$ increases, FedAvg converges faster, while the final estimation errors stay almost the same. This is because the data partition is perfectly balanced, so the local updates of FedAvg are relatively stable with $\kappa \approx 1$; hence according to Theorem 2 and Theorem 3, the convergence rate increases proportionally to $s$, while the final estimation does not change.

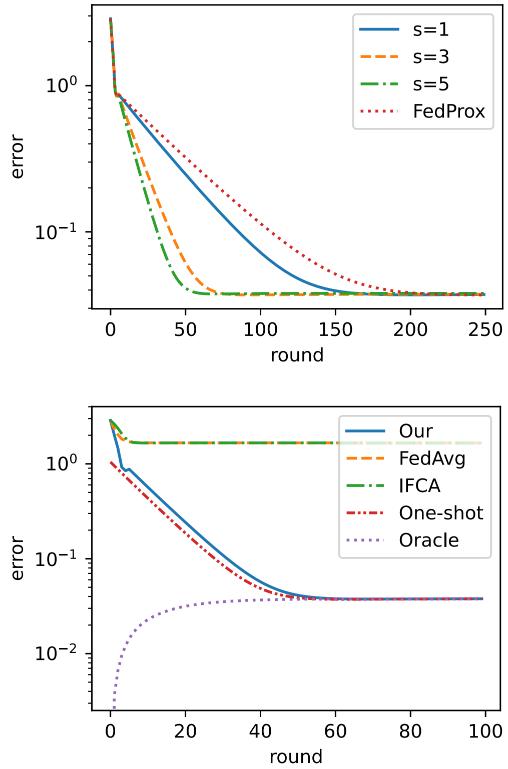

Fig. 1. Balanced local data and balanced cluster partition.

The right panel of Fig. 1 shows that our method significantly outperforms vanilla FedAvg and IFCA, and quickly converges to the same estimation error attainable by the oracle algorithm. Note that FedAvg does not converge to small errors due to lack of model personalization in the presence of model heterogeneity. The performance of IFCA is highly dependent on the quality of initialization. With a random initialization, IFCA gets stuck on an error floor. The one-shot clustering algorithm performs well in this setting. This is because the local data partition and cluster partition are perfectly balanced, so each client can well estimate its underlying model solely based on its local data and the PS can correctly cluster all the locally estimated models via $k$-means.

### B. Unbalanced Local Data and Balanced Cluster Partition

In this configuration, we consider unbalanced local data but balanced cluster partition. Specifically, we let $M = 920$, $n_i = 10$ for $i = 1, \cdots, 900$, and $n_i = 50$ for $i = 901, \cdots, 920$. That is, this configuration contains 920 clients, with each of the first 900 clients keeps 10 data points, and each of the remaining

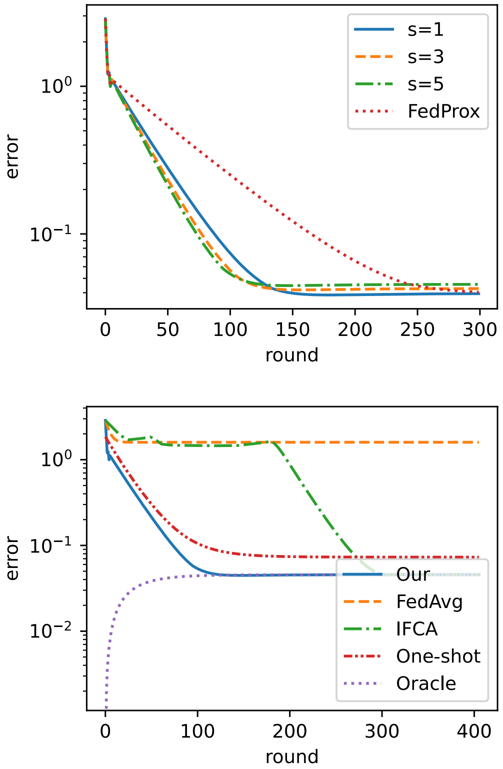

Fig. 2. Unbalanced local data and balanced cluster partition.

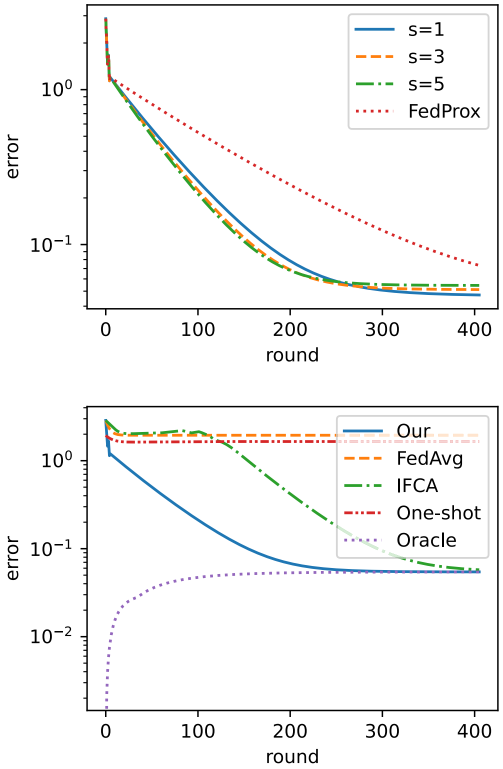

Fig. 3. Unbalanced local data and unbalanced cluster partition.

20 clients keeps 50 data points. For each client, it belongs to one of the 3 clusters with equal probability $1/3$.

The left panel of Fig. 2 stays almost the same as that of Fig. 1. The only noticeable difference is that in this setting with unbalanced local data, as $s$ increases, the convergence rate of FedAvg only slightly improves, while the final estimation error also gets slightly inflated. This is because, with unbalanced local data, the local updates of FedAvg for data-scarce clients become unstable, leading to a larger value of $\kappa$.

The right panel of Fig. 2 shows that our method still significantly outperforms vanilla FedAvg and IFCA, and quickly converges to the same estimation error attainable by the oracle algorithm. Although this time IFCA eventually also converges to the oracle estimation error, it still gets stuck on an error floor for a long time. The one-shot clustering algorithm no longer performs as well as before. This is because here the local data partition is unbalanced, so data-scarce clients cannot well estimate their underlying models solely based on their local data and the PS is likely to incorrectly cluster them. Since in one-shot clustering, the clustering is done only once and fixed throughout the remaining process, these clustering errors cannot be corrected.

### C. Unbalanced Local Data and Unbalanced Cluster Partition

In this configuration, we consider unbalanced local data and unbalanced cluster partition. Specifically, we let $M = 920$, $n_i = 10$ for $i = 1, \cdots, 900$, and $n_i = 50$ for $i = 901, \cdots, 920$. That is, this configuration contains 920 clients, with each of the first 900 clients keeps 10 data points, and each of the remaining 20 clients keeps 50 data points. For each client, it belongs to one of the 3 clusters with probability $p_1 = 0.2$, $p_2 = 0.3$, $p_3 = 0.5$.

The left panel of Fig. 3 stays almost the same as that of Fig. 2, except that convergence rates get smaller. This is consistent with our theoretical prediction in Theorem 2, which shows that the convergence rate is proportional to $\rho$ (roughly the same as $p_{\min}$).

The right panel of Fig. 3 shows that our method significantly outperforms vanilla FedAvg, IFCA, and one-shot, and quickly converges to the same estimation error attainable by the oracle algorithm. Note that one-shot clustering performs poorly in this case, because with unbalanced local data partition and unbalanced cluster partition, the one-shot clustering suffers from a large amount of errors in the initial clustering based on locally estimated models.

### D. Experiments on MNIST Real Dataset: Beyond Mixed Linear Regression

In Sections V-A -V-C, we focus on mixed linear regression on synthetic data. In this section, we consider the multi-class classification problem on MNIST dataset. We implement our algorithm adapted to general statistical learning, as described in Remark 3.

*Setup:* We create configurations with $k = 2$ and $k = 4$ clusters following [12]. For the former, to generate two clusters, the dataset of each client is either original or rotated by 90 degrees; for the latter, to generate four clusters, the dataset of each client is rotated by some degree in $\{0, 90, 180, 270\}$. The training dataset volume of MNIST is 60,000. The test dataset is copied for each cluster with the corresponding rotation transformation. The system contains $k \in \{2, 4\}$

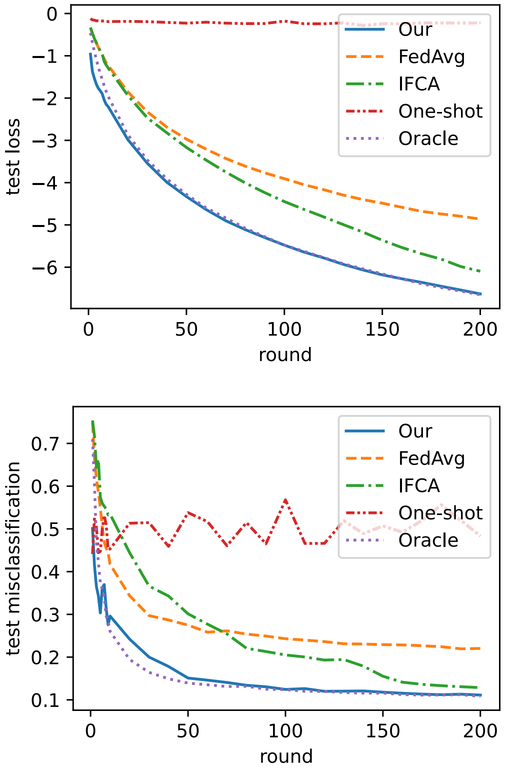

Fig. 4. Experiments with the MNIST dataset with two clusters (i.e., $k = 2$).

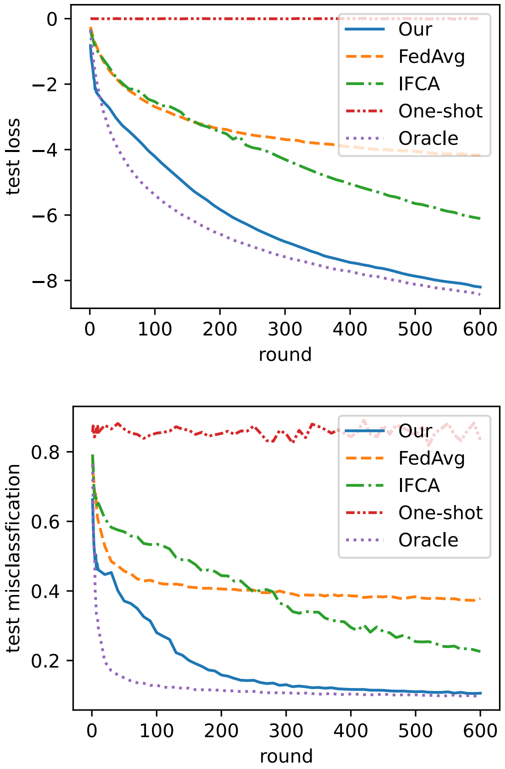

Fig. 5. Experiments with the MNIST dataset with four clusters (i.e., $k = 4$).

anchor clients – one for each cluster. Each anchor client holds $n_H = 20$ images. We consider a large client population. There are 25,000 non-anchor clients, where each contributes $n_i = 2$ images. To cope with this large client population, we downsample each image to a size of 14-by-14 pixels. For the image classification task, we adopt the logistic regression with input size 196 and output size 10.

For $k = 2$ and $k = 4$, we run Phase 1 on the anchor clients for 5 and 10 rounds, respectively. In Phase 1, the parameter server samples 1,000 clients for subspace estimation. In Phase 2, the parameter server samples 100 clients in each round. We use the FedAvg-based algorithm with $s = 5$ local updates per round and learning rate $\eta = 0.1$. For each cluster, we find the lowest misclassification error among the $k$ trained models, and then plot the average error among all clusters. We use the same benchmark comparison as in Fig. 1, 2, and 3. It is worth noting that the ground-truth model parameters do not exist for real datasets, so our previous Oracle iterative clustering algorithm cannot be directly applied. To resolve this, we modify the Oracle iterative clustering so that each client is provided with its true cluster label and only updates the corresponding model parameters using the FedAvg update rule.

We plot the trajectories of the evolution of the loss and the misclassification rate over test datasets in Fig. 4 and Fig. 5, where the loss is measured by the negative log-likelihood. Consistent with Fig. 1, 2, and 3, our algorithm is comparable to the Oracle algorithm and significantly outperforms other benchmarks. Overall, our algorithm converges fast in both Phases 1 and 2. In particular, FedAvg lacks model personalization; IFCA converges slowly due to the poor quality of random initialization; the effectiveness of one-shot averaging is compromised by the limited amount of data available at the non-anchor clients. As can be seen from the test misclassification errors in Fig. 4 and Fig. 5, during Phase 1 the errors first quickly decay and then either fluctuate around 3.5 or saturate at 4.3, respectively. Entering Phase 2, the test misclassification errors decay further.

# Appendix A
# Analysis of Phase 1

In this section, we present the analysis of our federated moment descent algorithm as described in Phase 1.

## A. Subspace Estimation via Federated Orthogonal Iteration

Recall that Phase 1 estimates the subspace that the residual estimation errors $\{\Sigma_\ell(\theta_\ell^* - \theta_{i,t})\}_{\ell=1}^k$ lie in via the federated-orthogonal iteration. We show that $\mathbb{E}[Y_{i,t}]$ is of rank at most $k$ and the eigenspace corresponding to the non-zero eigenvalues is spanned by $\{\Sigma_\ell(\theta_\ell^* - \theta_{i,t})\}_{\ell=1}^k$. Specifically, we first prove that $Y_{i,t}$ is close to $\mathbb{E}[Y_{i,t}]$ in the operator norm and then further deduce that $\{\Sigma_\ell(\theta_\ell^* - \theta_{i,t})\}_{\ell=1}^k$ approximately lie in the subspace spanned by the top-$k$ left singular vectors of $Y_{i,t}$.

Let $U_{i,t} \in \mathbb{R}^{d\times k}$ denote the top-$k$ left singular matrix of $Y_{i,t}$. To approximately compute $U_{i,t}$ in the FL systems, we adopt the following federated-orthogonal iteration algorithm. Suppose that $Y$ admits a decomposition over distributed clients, that is, $Y = \frac{1}{\sum_{i\in\mathcal{S}} n_i}\sum_{i\in\mathcal{S}}\sum_{j\in[n_i]} a_{ij} b_{ij}^\top$, where $\mathcal{S}$ is a set of clients, and $\{(a_{ij}, b_{ij})\}_{j=1}^{n_i}$ are computable based on the local dataset $\mathcal{D}_i$. Algorithm 3 approximates the top-$k$ left singular matrix of $Y$. It can be easily verified that Algorithm 3 effectively runs the orthogonal iteration on $YY^\top$.

**Algorithm 3** Federated Orthogonal Iteration

```
Input: A set $\mathcal{S}$ of clients $i$ with $\{(a_{ij}, b_{ij})\}_{i\in\mathcal{S}, j\in[n_i]}$, $k \in \mathbb{N}$, and even $T \in \mathbb{N}$
Output: $Q_T \in \mathbb{R}^{d\times k}$

PS initializes $Q_0 \in \mathbb{R}^{d\times k}$ as a random orthogonal matrix $Q_0^\top Q_0 = \mathbf{I}$.
for $t = 0, 1, \ldots, T-1$ do
    PS broadcasts $Q_t$ to all clients in $\mathcal{S}$.
    if $t$ is even then
        Each client $i \in \mathcal{S}$ computes an update $Q_{i,t} = \frac{1}{n_i}\sum_{j=1}^{n_i} b_{ij}a_{ij}^\top Q_t$ and transmits it back to the PS.
        PS updates $Q_{t+1} = \sum_{i\in S} w_i Q_{i,t}$, where $w_i = n_i / \sum_{i\in S} n_i$.
    else
        Each client $i \in \mathcal{S}$ computes an update $Q_{i,t} = \frac{1}{n_i}\sum_{j=1}^{n_i} a_{ij}b_{ij}^\top Q_t$ and transmits it back to the PS.
        PS applies the QR decomposition to obtain $Q_{t+1}$:
            $$\sum_{i\in\mathcal{S}} w_i Q_{i,t} = Q_{t+1}R_{t+1}.$$
    end if
end for
```

Recall that Phase 1 is called for each anchor client $i \in H$ and each global iteration $t$, and that $\hat{U}_{i,t}$ is the output of federated-orthogonal iteration in Step 9 of Phase 1, which approximates $U_{i,t}$. Based on the above discussion, we can show that the residual estimation errors $\{\Sigma_\ell(\theta_\ell^* - \theta_{i,t})\}_{\ell=1}^k$ approximately lie in the subspace spanned by the $k$ columns of $\hat{U}_{i,t}$.

*Proposition 1 (Subspace Estimation):* Suppose $T_1 \geq Ck \log \frac{Nd\beta^4 R}{\alpha^5\epsilon^2\Delta^2}$ and we condition on $\theta_{i,t}$ such that $\left\|\theta_{i,t}\right\|_2 \leq (1 + 2\beta/\alpha) R$. Then with probability at least $1 - 5N^{-10}$,

$$\left\|\left(\hat{U}_{i,t}\hat{U}_{i,t}^\top - I\right)\Sigma_\ell\left(\theta_\ell^* - \theta_{i,t}\right)\right\|_2^2$$
$$\leq \max\left\{C\left(\delta_{i,t}^2 + \sigma^2\right)\xi_1/p_\ell,\ \alpha^4\epsilon^2\Delta^2/(512\beta^2)\right\},\ \forall \ell \in [k],$$

where $\delta_{i,t} = \max_{\ell\in[k]} \|\theta_\ell^* - \theta_{i,t}\|_2$, $\xi_1 = \sqrt{\frac{d}{m}\log N} + \frac{d}{m}\log^3 N$, and $C$ is a constant.

We postpone the detailed proof to Appendix A-D. One key challenge in the analysis is that the eigengap of $\mathbb{E}\left[Y_{i,t}\right]$ could be small, especially when $\theta_{i,t}$ is close to $\theta_{z_i}^*$; and hence the standard Davis-Kahan theorem cannot be applied. This issue is further exacerbated by the fact that the convergence rate of the orthogonal iteration also crucially depends on the eigengaps. To resolve this issue, one key innovation of our analysis is to develop a gap-free bound to show that the projection errors $\hat{U}_{i,t}^\top \Sigma_\ell(\theta_\ell^* - \theta_{i,t})$ are small for every $\ell \in [k]$ (cf. Lemma 5).

## B. Moment Descent on Anchor Clients

Recall that in Step 11 of Phase 1, each anchor client $i \in H$ runs the power iteration to output $\hat{\beta}_{i,t}$ and $\hat{\sigma}_{i,t}^2$ as approximations of the leading left singular vector and singular value of $A_{i,t}$, respectively.

Then anchor client $i$ updates a new estimate $\theta_{i,t+1}$ by moving along the direction of the estimated residual error $r_{i,t}$ with an appropriately and adaptively chosen step size $\eta_{i,t}$. The following result shows that $\hat{\sigma}_{i,t}^2$ closely approximates the squared residual error $\|\Sigma_{z_i}(\theta_{z_i}^* - \theta_{i,t})\|_2^2$.

*Proposition 2:* Let $C$ denote a large constant. Fix an anchor client $i$, let $T_2 \geq C\log\frac{Nd\beta^4 R}{\alpha^5\epsilon^2\Delta^2}$, and condition on $\theta_{i,t}$ such that $\left\|\theta_{i,t}\right\|_2 \leq (1 + 2\beta/\alpha) R$. Then with probability at least $1 - 10N^{-10}$,

$$\left|\hat{\beta}_{i,t}^\top \hat{U}_{i,t}^\top \Sigma_{z_i}\left(\theta_{z_i}^* - \theta_{i,t}\right)\right|^2$$
$$\geq \left\|\Sigma_{z_i}\left(\theta_{z_i}^* - \theta_{i,t}\right)\right\|_2^2$$
$$- \max\left\{C\left(\delta_t^2 + \sigma^2\right)\left(\xi_1/p_{z_i} + \xi_2\right), \alpha^4\epsilon^2\Delta^2/(256\beta^2)\right\}. \tag{18}$$

and

$$\left|\left\|\Sigma_{z_i}(\theta_{z_i}^* - \theta_{i,t})\right\|_2^2 - \hat{\sigma}_{i,t}^2\right|$$
$$\leq \max\left\{C\left(\delta_t^2 + \sigma^2\right)\left(\xi_1/p_{z_i} + \xi_2\right), \alpha^4\epsilon^2\Delta^2/(128\beta^2)\right\}. \tag{19}$$

We postpone the proof to Appendix A-E. This proposition is the key to show that the descent direction $r_{i,t}$ is approximately parallel to the residual error $\Sigma_{z_i}(\theta_{z_i}^* - \theta_{i,t})$, so that the residual error decreases geometrically until it reaches a plateau. See the proof of Claim 1 for details.

## C. Proof of Theorem 1

Now, we are ready to prove our main theorem on the performance guarantee of Phase 1. Fix any anchor client $i$ and omit the subscript $i$ for simplicity. We further assume it belongs to cluster $\ell$, i.e., $z_i = \ell$. Let $\mathcal{E}_t$ denote the following event:

$$\mathcal{E}_t = \{\left\|\theta_t\right\|_2 \leq (1 + 2\beta/\alpha)R\} \cap \{\text{(18) and (19) hold}\}. \tag{20}$$

Let $\mathcal{E} = \cap_{t=1}^T \mathcal{E}_t$. We first prove the following claim.

*Claim 1:* Suppose $\mathcal{E}_t$ holds and

$$m \geq c\frac{\beta^8(R^4 + \sigma^4)d\log^3 N}{\alpha^{12}\Delta^4 p_{\min}^2\epsilon^4}, n_H \geq c\frac{\beta^8(R^4 + \sigma^4)k\log^3 N}{\alpha^{12}\Delta^4\epsilon^4} \tag{21}$$

for some sufficiently large constant $c$. Then

$$\left|\|\Sigma_\ell(\theta_\ell^* - \theta_t)\|_2^2 - \hat{\sigma}_t^2\right| \leq \alpha^2\epsilon^2\Delta^2/128. \tag{22}$$

Moreover,

- If $\hat{\sigma}_t \leq \alpha\epsilon\Delta/\sqrt{2}$, then
$$\|\Sigma_\ell(\theta_\ell^* - \theta_t)\|_2 \leq \alpha\epsilon\Delta. \tag{23}$$
- If $\hat{\sigma}_t > \alpha\epsilon\Delta/\sqrt{2}$, then
$$\left\|\Sigma_\ell\left(\theta_\ell^* - \theta_{t+1}\right)\right\|_2 \leq \left(1 - \frac{\alpha^2}{8\beta^2}\right)\left\|\Sigma_\ell\left(\theta_\ell^* - \theta_t\right)\right\|_2^2. \tag{24}$$

*Proof of Claim 1:* Since event $\mathcal{E}_t$ holds and $\|\theta_\ell^*\|_2 \le R$, it follows that $\delta_t \le 2(1+\beta/\alpha)R$ and hence

$$\begin{aligned}&\left|\|\Sigma_\ell(\theta_\ell^*-\theta_t)\|_2^2-\hat{\sigma}_t^2\right|\\&\le \max\left\{C\left(\delta_t^2+\sigma^2\right)(\xi_1/p_\ell+\xi_2),\alpha^4\epsilon^2\Delta^2/(128\beta^2)\right\}\\&\le \alpha^2\epsilon^2\Delta^2/128,\end{aligned}$$

where the last inequality follows from (21). This proves (22). Next, we divide the analysis into the following two cases.

*Case 1:* $\hat{\sigma}_t \le \alpha\epsilon\Delta/\sqrt{2}$. In this case, (23) directly follows from (22), as

$$\|\Sigma_\ell(\theta_\ell^*-\theta_t)\|_2^2 \le \hat{\sigma}_t^2+\alpha^2\epsilon^2\Delta^2/128 \le \alpha^2\epsilon^2\Delta^2.$$

*Case 2:* $\hat{\sigma}_t > \alpha\epsilon\Delta/\sqrt{2}$. In this case,

$$\begin{aligned}\|\Sigma_\ell(\theta_\ell^*-\theta_t)\|_2^2 &\ge \hat{\sigma}_t^2-\alpha^2\epsilon^2\Delta^2/128 > \alpha^2\epsilon^2\Delta^2/4\\&\ge \frac{256C\beta^2}{\alpha^2}(\delta_t^2+\sigma^2)(\xi_1/p_\ell+\xi_2),\end{aligned}$$

where the first inequality follows from (18), the second inequality follows because $\hat{\sigma}_t > \alpha\epsilon\Delta/\sqrt{2}$, and the last inequality holds due to (21). We further deduce from (18) that

$$\left|\hat{\beta}_t^\top\hat{U}_t^\top\Sigma_\ell\left(\theta_\ell^*-\theta_t\right)\right|^2 \ge \left(1-\frac{\alpha^2}{64\beta^2}\right)\|\Sigma_\ell\left(\theta_\ell^*-\theta_t\right)\|_2^2 \quad (25)$$

and from (22) that

$$\begin{aligned}\left(1-\frac{\alpha^2}{32\beta^2}\right)\|\Sigma_\ell\left(\theta_\ell^*-\theta_t\right)\|_2^2 &\le \hat{\sigma}_t^2\\&\le\left(1+\frac{\alpha^2}{32\beta^2}\right)\|\Sigma_\ell\left(\theta_\ell^*-\theta_t\right)\|_2^2.\end{aligned} \quad (26)$$

Now we show $\theta_{t+1}$ gets closer to $\theta_\ell^*$. Note that $\theta_{t+1}=\theta_t+\eta_t r_t$. It follows that

$$\begin{aligned}&\left(\theta_\ell^*-\theta_{t+1}\right)^\top\Sigma_\ell^2\left(\theta_\ell^*-\theta_{t+1}\right)\\&=\left(\theta_\ell^*-\theta_t\right)^\top\Sigma_\ell^2\left(\theta_\ell^*-\theta_t\right)-2\eta_t\left(\theta_\ell^*-\theta_t\right)^\top\Sigma_\ell^2 r_t+\eta_t^2 r_t^\top\Sigma_\ell^2 r_t.\end{aligned}$$

We decompose

$$\Sigma_\ell\left(\theta_\ell^*-\theta_t\right)=a_t r_t+b_t r_t^\perp,$$

for some unit vector $r_t^\perp$ that is perpendicular to $r_t$. Recalling $r_t=\hat{U}_t\hat{\beta}_t$, we have $\|r_t\|_2=1$ and it follows from (25) that

$$a_t=\langle r_t,\Sigma_\ell\left(\theta_\ell^*-\theta_t\right)\rangle^2 \ge \left(1-\frac{\alpha^2}{64\beta^2}\right)\|\Sigma_\ell\left(\theta_\ell^*-\theta_t\right)\|_2^2.$$

Since $a_t^2+b_t^2=\|\Sigma_\ell(\theta_\ell^*-\theta_t)\|_2^2$, we have $|b_t| \le \frac{\alpha}{8\beta}\|\Sigma_\ell(\theta_\ell^*-\theta_t)\|_2$. Hence,

$$\begin{aligned}&\left(\theta_\ell^*-\theta_t\right)^\top\Sigma_\ell^2 r_t=\left(a_t r_t+b_t r_t^\perp\right)^\top\Sigma_\ell r_t\\&\ge a_t\alpha-|b_t|\beta\\&\ge\sqrt{1-\frac{\alpha^2}{64\beta^2}}\alpha\|\Sigma_\ell(\theta_\ell^*-\theta_t)\|_2-\frac{\alpha}{8}\|\Sigma_\ell(\theta_\ell^*-\theta_t)\|_2\\&\ge\frac{\alpha}{2}\|\Sigma_\ell(\theta_\ell^*-\theta_t)\|_2,\end{aligned}$$

where $\lambda_{\min}(\Sigma_\ell)\ge\alpha$ and $\beta\ge\|\Sigma_\ell\|_2$. It follows that

$$\begin{aligned}&\left(\theta_\ell^*-\theta_{t+1}\right)^\top\Sigma_\ell^2\left(\theta_\ell^*-\theta_{t+1}\right)\\&\le\left(\theta_\ell^*-\theta_t\right)^\top\Sigma_\ell^2\left(\theta_\ell^*-\theta_t\right)-\eta_t\alpha\|\Sigma_\ell\left(\theta_\ell^*-\theta_t\right)\|_2+\eta_t^2\|\Sigma_\ell\|_2^2.\end{aligned}$$

Recall the choice of step size $\eta_t=\alpha\hat{\sigma}_t/(2\beta^2)$. We get that

$$\begin{aligned}&\left(\theta_\ell^*-\theta_{t+1}\right)^\top\Sigma_\ell^2\left(\theta_\ell^*-\theta_{t+1}\right)\\&\le\left(\theta_\ell^*-\theta_t\right)^\top\Sigma_\ell^2\left(\theta_\ell^*-\theta_t\right)-\frac{\alpha^2}{2\beta^2}\|\Sigma_\ell\left(\theta_\ell^*-\theta_t\right)\|_2\hat{\sigma}_t+\frac{\alpha^2\hat{\sigma}_t^2}{4\beta^2}\\&\le\left(1-\frac{\alpha^2}{8\beta^2}\right)\|\Sigma_\ell\left(\theta_i^*-\theta_t\right)\|_2^2,\end{aligned}$$

where the last inequality holds due to (26). Therefore,

$$\|\Sigma_\ell\left(\theta_\ell^*-\theta_{t+1}\right)\|_2^2\le\left(1-\frac{\alpha^2}{8\beta^2}\right)\|\Sigma_\ell\left(\theta_\ell^*-\theta_t\right)\|_2^2,$$

This proves (24). □

Second, we prove $\|\theta_T-\theta_\ell^*\|_2\le\epsilon\Delta$ assuming event $\mathcal{E}$ holds. Define

$$t^*=\min\{\inf\{t\ge 0:\hat{\sigma}_t\le\alpha\epsilon\Delta/\sqrt{2}\},\,T\}.$$

By the stopping criterion of our algorithm, it suffices to prove $\|\theta_{t^*}-\theta_\ell^*\|_2\le\epsilon\Delta$. We divide the analysis into the following two cases depending on the value of $t^*$.

- $t^*<T$. In this case, by definition, $\hat{\sigma}_{t^*}\le\alpha\epsilon\Delta/\sqrt{2}$. By applying (23) in Claim 1, we obtain
$$\|\Sigma_\ell(\theta_\ell^*-\theta_{t^*})\|_2\le\alpha\epsilon\Delta.$$
- $t^*=T$. In this case, $\hat{\sigma}_t>\alpha\epsilon\Delta/\sqrt{2}$ for all $t=1,\ldots,T-1$. By applying (24) in Claim 1, we obtain
$$\begin{aligned}\|\Sigma_\ell\left(\theta_\ell^*-\theta_{t^*}\right)\|_2&\le\left(1-\frac{\alpha^2}{8\beta^2}\right)^{T/2}\|\Sigma_\ell(\theta_\ell^*-\theta_0)\|_2\\&\le 2\exp\left(-T\alpha^2/(16\beta^2)\right)\beta R\le\alpha\epsilon\Delta,\end{aligned}$$
where the last inequality holds by choosing $T=\frac{16\beta^2}{\alpha^2}\log\frac{2\beta R}{\epsilon\alpha\Delta}$.

In both cases, we obtain that

$$\|\theta_\ell^*-\theta_{t^*}\|_2\le\frac{1}{\alpha}\|\Sigma_\ell(\theta_\ell^*-\theta_{t^*})\|_2\le\frac{\alpha\epsilon\Delta}{\alpha}=\epsilon\Delta.$$

Third, we prove that $\mathbb{P}\left\{\mathcal{E}_{t+1}\mid\cap_{\tau=1}^t\mathcal{E}_\tau\right\}\ge 1-10N^{-10}$. We divide the analysis into the following two cases assuming $\cap_{\tau=1}^t\mathcal{E}_\tau$ holds.

- $\hat{\sigma}_\tau\le\alpha\epsilon\Delta/\sqrt{2}$ for some $\tau\in[t]$. In this case, by the stopping criterion of our algorithm, $\theta_{t+1}=\theta_\tau$. Thus, $\|\theta_{t+1}\|_2=\|\theta_\tau\|_2\le(1+2\beta/\alpha)R$.
- $\hat{\sigma}_\tau>\alpha\epsilon\Delta/\sqrt{2}$ for all $\tau\in[t]$. In this case, by applying (24) in Claim 1, we obtain that
$$\begin{aligned}\alpha\|\theta_\ell^*-\theta_{t+1}\|_2&\le\|\Sigma_\ell(\theta_\ell^*-\theta_{t+1})\|_2\\&\le\|\Sigma_\ell(\theta_\ell^*-\theta_0)\|_2\le 2\beta R.\end{aligned}$$

In both cases, we have $\|\theta_{t+1}\|_2\le(1+2\beta/\alpha)R$. Thus, applying Proposition 2 yields $\mathbb{P}\left\{\mathcal{E}_{t+1}\mid\cap_{\tau=1}^t\mathcal{E}_\tau\right\}\ge 1-10N^{-10}$. It follows that

$$\begin{aligned}\mathbb{P}\left\{\cap_{t=1}^T\mathcal{E}_t\right\}&=\mathbb{P}\left\{\mathcal{E}_1\right\}\times\cdots\times\mathbb{P}\left\{\mathcal{E}_T\mid\mathcal{E}_1,\ldots,\mathcal{E}_{T-1}\right\}\\&\ge\left(1-10N^{-10}\right)^T\ge 1-10TN^{-10}.\end{aligned}$$

Then we apply a union bound over all anchor client $i\in H$ and conclude that with probability at least $1-10Tm_HN^{-10}$, $\|\theta_{i,T}-\theta_{z_i}^*\|_2\le\epsilon\Delta$ for all $i\in H$. This proves (4).

Finally, we prove (5). Recall that $H \cap \{i : z_i = \ell\} \neq \emptyset$ for all $\ell \in [k]$. Moreover, as long as $\epsilon < 1/4$, we have for two anchor clients $i, i' \in H$

$$\begin{aligned} \|\theta_{i,T} - \theta_{i',T}\|_2 &\le \left\|\theta_{i,T} - \theta^*_{z_i}\right\|_2 + \left\|\theta_{i',T} - \theta^*_{z_{i'}}\right\|_2 \\ &\le 2\epsilon\Delta, \quad \text{if } z_i = z_{i'}, \\ \|\theta_{i,T} - \theta_{i',T}\|_2 &\ge \Delta - \left\|\theta_{i,T} - \theta^*_{z_i}\right\|_2 - \left\|\theta_{i',T} - \theta^*_{z_{i'}}\right\|_2 \\ &\ge (1-2\epsilon)\Delta, \quad \text{if } z_i \neq z_{i'}. \end{aligned}$$

Thus, by assigning anchor clients $i, i' \in H$ in the same cluster when $\|\theta_{i,T} - \theta_{i',T}\|_2 \le \Delta/2$ we can recover the $k$ clusters of the clients users. In particular, let $\hat{z}_i$ denote the estimated cluster label of anchor client $i \in H$. Then there exists a permutation $\pi : [k] \to [k]$ such that $\pi(\hat{z}_i) = z_i$ for all $i \in H$. Let $\hat{\theta}_\ell$ denote the center of the recovered cluster $\ell$, that is

$$\hat{\theta}_\ell = \sum_{i \in H} \theta_{i,T} \mathbb{1}\{\hat{z}_i = \ell\} / \sum_{i \in H} \mathbb{1}\{\hat{z}_i = \ell\}.$$

Then we have $\|\hat{\theta}_{\pi(\ell)} - \theta^*_\ell\|_2 \le \epsilon\Delta$ for all $\ell \in [k]$. This finishes the proof of (5).

### *D. Proof of Proposition 1*

In the following analysis, we fix an anchor client $i \in H$ and omit the subscript $i$ for ease of presentation. Crucially, since $\mathcal{S}_t$ and $\mathcal{D}_t$ are freshly drawn, all the global data and local data used in iteration $t+1$ are independent of $\theta_t$. Hence, we condition on $\theta_t$ and $\mathcal{S}_t$ in the following analysis. Note that

$$\begin{aligned} \mathbb{E}[Y_t] &= \frac{1}{m} \sum_{i' \in \mathcal{S}_t} \mathbb{E}_{z_{i'}} \left[ \Sigma_{z_{i'}} \left(\theta^*_{z_{i'}} - \theta_t\right) \left(\theta^*_{z_{i'}} - \theta_t\right)^\top \Sigma_{z_{i'}} \right] \\ &= \sum_{\ell=1}^{k} p_\ell \Sigma_\ell \left(\theta^*_\ell - \theta_t\right) \left(\theta^*_\ell - \theta_t\right)^\top \Sigma_\ell, \end{aligned}$$

where $p_\ell$ is the probability that a client belongs to the $\ell$-th cluster.

Let $\hat{U}_t \in \mathbb{R}^{d \times k}$ denote the left singular matrix of $Y_t$. We aim to show that the collection of $\Sigma_\ell(\theta^*_\ell - \theta_t)$ for $\ell \in [k]$ approximately lie in the space spanned by the $k$ columns of $\hat{U}_t$. As such, we first show that $Y_t$ is close to $\mathbb{E}[Y_t]$ in the operator norm.

*Lemma 3:* With probability at least $1 - 3N^{-10}$,

$$\|Y_t - \mathbb{E}[Y_t]\|_2 \le C\left(\delta_t^2 + \sigma^2\right)\xi_1,$$

where $\delta_t = \max_{\ell \in [k]} \|\theta^*_\ell - \theta_t\|_2$ and $\xi_1 = \sqrt{\frac{d}{m}\log N} + \frac{d}{m}\log^3 N$, and $C > 0$ is some constant.

*Proof:* Let $\varepsilon_i = (y_{i1} - \langle\phi(x_{i1}), \theta_t\rangle)\phi(x_{i1})$ and $\tilde{\varepsilon}_i = (y_{i2} - \langle\phi(x_{i2}), \theta_t\rangle)\phi(x_{i2})$. Note that

$$Y_t - \mathbb{E}[Y_t] = \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i \tilde{\varepsilon}_i^\top - \mathbb{E}\left[\varepsilon_i \tilde{\varepsilon}_i^\top\right].$$

Let $a_i = \varepsilon_i / \sqrt{\delta_t^2 + \sigma^2}$ and $b_i = \tilde{\varepsilon}_i / \sqrt{\delta_t^2 + \sigma^2}$. We will apply a truncated version of the Matrix Bernstein's inequality given in Lemma 13. As such, we first check the conditions in Lemma 13 are all satisfied. Note that

$$\begin{aligned} \mathbb{E}\left[\|\varepsilon_i\|_2^2\right] &= \mathbb{E}\left[\left\|\left(\langle\phi(x_{i1}), \theta^*_{z_i} - \theta_t\rangle + \zeta_i\right)\phi(x_{i1})\right\|_2^2\right] \\ &= \mathbb{E}\left[\left\|\langle\phi(x_{i1}), \theta^*_{z_i} - \theta_t\rangle\phi(x_{i1})\right\|_2^2\right] \\ &\quad + \mathbb{E}\left[\|\zeta_{i1}\phi(x_{i1})\|_2^2\right]. \end{aligned}$$

By the sub-Gaussianity of $\phi(x_{i1})$, we have

$$\mathbb{E}\left[\|\zeta_i\phi(x_{i1})\|_2^2\right] \le \sigma^2 \mathbb{E}\left[\|\phi(x_{i1})\|_2^2\right] = C_1\sigma^2 d$$

and further by Cauchy-Schwarz inequality,

$$\begin{aligned} &\mathbb{E}\left[\left\|\langle\phi(x_{i1}), \theta^*_{z_i} - \theta_t\rangle\phi(x_{i1})\right\|_2^2\right] \\ &\le \sqrt{\mathbb{E}\left[\langle\phi(x_{i1}), \theta^*_{z_i} - \theta_t\rangle^4\right]} \sqrt{\mathbb{E}\left[\|\phi(x_{i1})\|_2^4\right]} \le C_1\delta_t^2 d, \end{aligned}$$

where $C_1$ is a constant only depending on the sub-Gaussian norm of $\phi(x_{i1})$. Combining the last three displayed equations gives that $\mathbb{E}\left[\|a_i\|_2^2\right] \le C_1 d$. The same upper bound also holds for $\mathbb{E}\left[\|b_i\|_2^2\right]$.

Moreover, $\left\|\mathbb{E}\left[a_i a_i^\top\right]\right\|_2 = \sup_{u \in \mathcal{S}^{d-1}} \mathbb{E}\left[\langle a_i, u\rangle^2\right]$. Note that for any $u \in \mathcal{S}^{d-1}$,

$$\begin{aligned} \mathbb{E}\left[\langle a_i, u\rangle^2\right] &= \frac{1}{\delta_t^2 + \sigma^2} \mathbb{E}\left[r_i^2 \langle\phi(x_{i1}), u\rangle^2\right] \\ &\le \frac{1}{\delta_t^2 + \sigma^2} \sqrt{\mathbb{E}\left[r_i^4\right]} \sqrt{\langle\phi(x_{i1}), u\rangle^4} \le C_2, \end{aligned}$$

where $r_i = y_{i1} - \langle\phi(x_{i1}), \theta_t\rangle$ and $C_2$ is some constant only depending on the sub-Gaussian norm of $\phi(x_{i1})$ and $\zeta_{i1}$. Combining the last two displayed equations gives that $\left\|\mathbb{E}\left[a_i a_i^\top\right]\right\|_2 \le C_2$. The same upper bound also holds for $\left\|\mathbb{E}\left[b_i b_i^\top\right]\right\|_2$. Finally, by the sub-Gaussian property of $\phi(x_{i1})$ and $\epsilon$-net argument, we have

$$\mathbb{P}\left\{\|\phi(x_{i1})\|_2 \ge s_1\right\} \le \exp\left(2d - cs_1^2\right)$$

and

$$\mathbb{P}\left\{\frac{|r_i|}{\sqrt{\delta_t^2 + \sigma^2}} \ge s_2\right\} \le \exp\left(-cs_2^2\right),$$

where $c > 0$ is some constant only depending on the sub-Gaussian norm of $\phi(x_{i1})$ and $\zeta_{i1}$.

Choosing $s_1 = \sqrt{Cs}d^{1/4}$ and $s_2 = \sqrt{s}/(C^{1/2}d^{1/4})$ for a constant $C = 4/c$, we get that for all $s \ge \sqrt{d}$,

$$\begin{aligned} \mathbb{P}\left\{\|a_i\|_2 \ge s\right\} &\le \mathbb{P}\left\{\|\phi(x_{i1})\|_2 \ge s_1\right\} + \mathbb{P}\left\{\frac{|r_i|}{\sqrt{\delta_t^2 + \sigma^2}} \ge s_2\right\} \\ &\le \exp\left(2d - 4\, s\sqrt{d}\right) + \exp\left(-\frac{c^2 s}{4\sqrt{d}}\right) \\ &\le 2\exp\left(-\frac{c^2 s}{4\sqrt{d}}\right). \end{aligned}$$

The same bound holds for $\mathbb{P}\left\{\|b_i\|_2 \ge s\right\}$. Applying the truncated version of the Matrix Bernstein's inequality given in Lemma 13 yields the desired result. □

The following result shows the geometric convergence of orthogonal iteration. Let $Y = U\Lambda U^\top$ denote the eigenvalue decomposition of $Y$ with $|\lambda_1| \ge |\lambda_2| \ge \cdots |\lambda_d|$ and the corresponding eigenvectors $u_i$'s. Define $U_1 = [u_1, \ldots, u_k]$ and $U_2 = [u_{k+1}, \ldots, u_d]$. Let $Q_t \in \mathbb{R}^{d \times k}$ denote the output of the *orthogonal iteration* with $Q_0$ initialized as a random orthogonal iteration $Q_0^\top Q_0 = I$ and $YQ_t = Q_{t+1}R_{t+1}$.

*Lemma 4 [32, Theorem 8.2.2]:* Assume $|\lambda_k| > |\lambda_{k+1}|$ and $\cos(\gamma) = \sigma_{\min}(U_1^\top Q_0)$ for $\gamma \in [0, \pi/2]$. Then

$$\left\|Q_t Q_t^\top - U_1 U_1^\top\right\|_2 \le \tan(\gamma) \left|\frac{\lambda_{k+1}}{\lambda_k}\right|^t, \quad \forall t.$$

Finally, we need a gap-free bound that controls the projection errors.

*Lemma 5 (Gap-Free Bound on Projection Errors):* Suppose $M \in \mathbb{R}^{d\times d}$ satisfies that

$$\left\|M - \sum_{i=1}^k x_i x_i^\top\right\|_2 \le \epsilon,$$

where $x_i \in \mathbb{R}^d$ for $1 \le i \le k$. Let $Q_t \in \mathbb{R}^{d\times k}$ be the output of the orthogonal iteration running over $MM^\top$ with $Q_0$ initialized as a random orthogonal matrix $Q_0^\top Q_0 = I$. Assume that $\|x_i\|_2 \le H$ for all $1 \le i \le k$. There exists a universal constant $C > 0$ such that for any $\epsilon > 0$ and $t \ge Ck \log \frac{dNH}{\epsilon}$, we have with probability at least $1 - 2N^{-10}$ (over the randomness of $Q_0$),

$$\left\|Q_t Q_t^\top x_i - x_i\right\|_2 \le 3\sqrt{\epsilon}, \quad \forall 1 \le i \le k.$$

*Proof:* Let $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_d \ge 0$ denote the singular values of $M$. Then by assumption on $M$ and Weyl's inequality, $\sigma_{k+1} \le \epsilon$. We divide the analysis into two cases depending on the value of $\sigma_1$. Let $\delta > 0$ be some parameter to be tuned later.

*Case 1:* $\sigma_1 \le (1+\delta)^k \epsilon$. In this case, by Weyl's inequality,

$$\begin{aligned}\|x_i\|_2^2 &\le \left\|\sum_{i=1}^k x_i x_i^\top\right\|_2 \le \|M\|_2 + \left\|M - \sum_{i=1}^k x_i x_i^\top\right\|_2 \\ &\le \epsilon \left(1 + (1+\delta)^k\right).\end{aligned}$$

Thus,

$$\left\|Q_t Q_t^\top x_i - x_i\right\|_2 \le \|x_i\|_2 \le \sqrt{\epsilon\left(1 + (1+\delta)^k\right)}$$

*Case 2:* $\sigma_1 > (1+\delta)^k \epsilon$. Then by the pigeonhole principle there must exist $1 \le p \le k$ such that $\sigma_p/\sigma_{p+1} > 1+\delta$. Choose

$$q = \max\left\{p : \sigma_p/\sigma_{p+1} > 1+\delta\right\}.$$

It follows that $\sigma_{q+1} \le (1+\delta)^{k-q}\epsilon \le (1+\delta)^k \epsilon$. Let $U_q = [u_1, \ldots, u_q]$, where $u_i$'s are the left singular vectors of $M$ corresponding to $\sigma_i$. Given the subspace $\text{span}\{u_1, \ldots, u_q\}$, denote the unique orthogonal decomposition of $x_i$ by $x_i = \Pi_W(x_i) + e$, where $\Pi_W(x_i) = U_q U_q^\top x_i$ and $e^\top u_j = 0$ for all $j \in [q]$. Let $u = e/\|e\|_2 \in S^{d-1}$. Then,

$$\begin{aligned}\left\|U_q U_q^\top x_i - x_i\right\|_2^2 &= u^\top x_i x_i^\top u \le u^\top \left(\sum_{i=1}^k x_i x_i^\top\right) u \\ &= u^\top \left(\sum_{i=1}^k x_i x_i^\top - M\right) u + u^\top M u.\end{aligned}$$

Note that

$$u^\top \left(\sum_{i=1}^k x_i x_i^\top - M\right) u \le \left\|\sum_{i=1}^k x_i x_i^\top - M\right\|_2 \le \epsilon.$$

Moreover,

$$\begin{aligned}u^\top M u &= \sum_j \sigma_j u^\top u_j v_j^\top u \\ &= \sum_{j\ge q+1} \sigma_j u^\top u_j v_j^\top u \\ &\le \sigma_{q+1} \sum_{j \ge q+1} \left|u^\top u_j\right| \left|v_j^\top u\right| \\ &\le \sigma_{q+1} \sqrt{\sum_{j\ge q+1} \left|u^\top u_j\right|^2 \sum_{j\ge q+1} \left|v_j^\top u\right|^2} \\ &\le \sigma_{q+1} \le (1+\delta)^k \epsilon.\end{aligned}$$

Combining the last three displayed equations gives that

$$\left\|U_q U_q^\top x_i - x_i\right\|_2^2 \le \epsilon\left(1 + (1+\delta)^k\right).$$

Let $\hat{Q}_t$ be the submatrix of $Q_t$ formed by the first $q$ columns. Since $\sigma_q > \sigma_{q+1}$, the space spanned by $\hat{Q}_t$ is the same space spanned by $Q_t$ if the orthogonal iteration were run with $k$ replaced by $q$. Thus, applying Lemma 4 with $k$ replaced by $q$ gives that

$$\left\|\hat{Q}_t \hat{Q}_t^\top - U_q U_q^\top\right\|_2 \le \tan(\gamma)(1+\delta)^{-t},$$

where $\cos(\gamma) = \sigma_{\min}(U_q^\top \hat{Q}_0)$ and $\hat{Q}_0$ is the submatrix of $Q_0$ formed by its first $q$ columns. Applying Lemma 14, we get $\tan(\gamma) \le cN^{10} d \log N$ with probability at least $1 - 2N^{-10}$ for some constant $c > 0$. Therefore, when $t \ge (C/\delta) \log \frac{NdH}{\epsilon}$ for some sufficiently large constant $C > 0$, we have

$$\left\|\hat{Q}_t \hat{Q}_t^\top - U_q U_q^\top\right\|_2 \le \epsilon/H.$$

Therefore, by triangle's inequality,

$$\begin{aligned}\left\|Q_t Q_t^\top x_i - x_i\right\|_2 &\le \left\|\hat{Q}_t \hat{Q}_t^\top x_i - x_i\right\|_2 \\ &\le \left\|U_q U_q^\top x_i - x_i\right\|_2 \\ &\quad + \left\|\left(\hat{Q}_t \hat{Q}_t^\top - U_q U_q^\top\right) x_i\right\|_2 \\ &\le \sqrt{\epsilon\left(1 + (1+\delta)^k\right)} + \epsilon.\end{aligned}$$

Finally, choosing $\delta = 1/k$ and noting that $(1+\delta)^k \le e$, we get the desired conclusions. □

Applying Lemma 3 and Lemma 5 and invoking the assumption that $T_1 \ge Ck \log \frac{Nd\beta^4 R}{\alpha^5 \epsilon^2 \Delta^2}$, we obtain that conditional on $\theta_t$ with $\|\theta_t\|_2 \le (1 + 2\beta/\alpha) R$, with probability at least $1 - 5N^{-10}$,

$$\begin{aligned}&\left\|\left(\hat{U}_t \hat{U}_t^\top - I\right) \sqrt{p_\ell} \Sigma_\ell \left(\theta_\ell^* - \theta_t\right)\right\|_2^2 \\ &\le \max\left\{C\left(\delta_t^2 + \sigma^2\right)\xi_1,\ p_\ell \alpha^4 \epsilon^2 \Delta^2/(512\beta^2)\right\}, \quad \forall \ell \in [k].\end{aligned}$$

This finishes the proof of Proposition 1.

### *E. Proof of Proposition 2*

Similar to the proof of Proposition 1, for ease of exposition, we fix an anchor client $i$ and omit the subscript $i$ for simplicity. We further assume client $i$ belongs to cluster $\ell$, i.e., $z_i = \ell$. Note that crucially, the global data points on clients $\mathcal{S}_t$

are independent of the local data points on $\mathcal{D}_t$. Thus, in the following analysis, we further condition on $\hat{U}_t$. Then

$$\mathbb{E}\left[A_t\right] = \hat{U}_t^\top \Sigma_\ell \left(\theta_\ell^* - \theta_t\right) \left(\theta_\ell^* - \theta_t\right)^\top \Sigma_\ell \hat{U}_t.$$

*Lemma 6:* With probability at least $1 - 3N^{-10}$,

$$\left\| A_t - \mathbb{E}\left[A_t\right] \right\|_2 \leq C \left( \left\| \theta_\ell^* - \theta_t \right\|_2^2 + \sigma^2 \right) \xi_2,$$

where $\xi_2 = \sqrt{\frac{k}{n_H} \log N} + \frac{k}{n_H} \log^3 N$ and $C > 0$ is a constant.

*Proof:* Note that

$$A_t - \mathbb{E}\left[A_t\right] = \frac{1}{n_H} \sum_{j \in \mathcal{D}_t} \hat{U}_t^\top \left( \varepsilon_j \tilde{\varepsilon}_j^\top - \mathbb{E}\left[ \varepsilon_j \tilde{\varepsilon}_j^\top \right] \right) \hat{U}_t,$$

where $\varepsilon_j = (y_j - \langle \phi(x_j), \theta_t \rangle) \phi(x_j)$ and $\tilde{\varepsilon}_j = (\tilde{y}_j - \langle \phi(\tilde{x}_j), \theta_t \rangle) \phi(\tilde{x}_j)$. Let $a_j = \hat{U}_t^\top \varepsilon_j / \sqrt{\|\theta_\ell^* - \theta_t\|_2^2 + \sigma^2}$ and $b_j = \hat{U}_t^\top \tilde{\varepsilon}_j / \sqrt{\|\theta_\ell^* - \theta_t\|_2^2 + \sigma^2}$. The rest of the proof follows analogously as that of Lemma 3. □

Applying Lemma 6 and Lemma 5, when $T_2 \geq C \log \frac{N d \beta^3 R}{\alpha^4 \epsilon^2 \Delta^2}$, we have with probability at least $1 - 5N^{-10}$,

$$\begin{aligned} \left| \hat{\beta}_t^\top \hat{U}_t^\top \Sigma_\ell \left( \theta_\ell^* - \theta_t \right) \right|^2 &\geq \left\| \hat{U}_t^\top \Sigma_\ell \left( \theta_\ell^* - \theta_t \right) \right\|_2^2 \\ &\quad - \max\left\{ C \left( \|\theta_\ell^* - \theta_t\|_2^2 + \sigma^2 \right) \xi_2, \alpha^4 \epsilon^2 \Delta^2 / (512 \beta^2) \right\}. \end{aligned}$$

Applying Proposition 1, when $T_1 \geq C k \log \frac{N d \beta^4 R}{\alpha^5 \epsilon^2 \Delta^2}$, we have with probability at least $1 - 5N^{-10}$,

$$\begin{aligned} \left\| \hat{U}_t^\top \Sigma_\ell \left( \theta_\ell^* - \theta_t \right) \right\|_2^2 &\geq \left\| \Sigma_\ell \left( \theta_\ell^* - \theta_t \right) \right\|_2^2 \\ &\quad - \max\left\{ C \left( \delta_t^2 + \sigma^2 \right) \xi_1 / p_\ell, \alpha^4 \epsilon^2 \Delta^2 / (512 \beta^2) \right\}. \end{aligned}$$

Let $\mathcal{E}_t$ denote the event such that the above two displayed equations hold simultaneously. Then $\mathbb{P}\left\{ \mathcal{E}_t \right\} \geq 1 - 10N^{-10}$. In the following, we assume event $\mathcal{E}_t$ holds.

Combining the last two displayed equations yields that

$$\begin{aligned} \left| \hat{\beta}_t^\top \hat{U}_t^\top \Sigma_\ell \left( \theta_\ell^* - \theta_t \right) \right|^2 &\geq \left\| \Sigma_\ell \left( \theta_\ell^* - \theta_t \right) \right\|_2^2 \\ &\quad - \max\left\{ 2C \left( \delta_t^2 + \sigma^2 \right) \left( \xi_1 / p_\ell + \xi_2 \right), \alpha^4 \epsilon^2 \Delta^2 / (256 \beta^2) \right\}. \end{aligned} \tag{27}$$

This proves (18). Moreover, since

$$\hat{\sigma}_t^2 \triangleq \hat{\beta}_t^\top A_t \hat{\beta}_t = \hat{\beta}_t^\top \mathbb{E}\left[A_t\right] \hat{\beta}_t + \hat{\beta}_t^\top \left( A_t - \mathbb{E}\left[A_t\right] \right) \hat{\beta}_t,$$

it follows that

$$\left| \hat{\sigma}_t^2 - \left| \hat{\beta}_t^\top \hat{U}_t^\top \Sigma_\ell \left( \theta_\ell^* - \theta_t \right) \right|^2 \right| \leq C \left( \delta_t^2 + \sigma^2 \right) \xi_2.$$

Combining the last two displayed equations yields that

$$\begin{aligned} &\left| \hat{\sigma}_t^2 - \left\| \Sigma_\ell \left( \theta_\ell^* - \theta_t \right) \right\|_2^2 \right| \\ &\leq \max\left\{ 3C \left( \delta_t^2 + \sigma^2 \right) \left( \xi_1 / p_\ell + \xi_2 \right), \alpha^4 \epsilon^2 \Delta^2 / (128 \beta^2) \right\} \end{aligned}$$

This proves (19).

## Appendix B
## Analysis of Phase 2

Throughout the proof in this section, we assume without loss of generality that the optimal permutation in (5) is identity.

### A. Derivation of Global Iteration

*Proof of Lemma 1:* We first prove the result for FedAvg. By definition, we have

$$\nabla_\ell L_i(\theta) = \frac{\lambda_{i\ell,t}}{n_i} \phi(\boldsymbol{x}_i)^\top (\phi(\boldsymbol{x}_i) \theta_\ell - y_i),$$

where $\lambda_{i\ell,t} = \mathbb{1}\{\ell = z_{i,t}\}$ and $\nabla_\ell$ denotes the gradient with respect to $\theta_\ell$. Then the one-step local gradient descent at client $i$ is

$$[\mathcal{G}_i(\theta)]_\ell = \begin{cases} \theta_\ell, & \ell \neq z_{i,t}, \\ g_i(\theta_\ell) \triangleq \theta_\ell - \eta_i \phi(\boldsymbol{x}_i)^\top (\phi(\boldsymbol{x}_i) \theta_\ell - y_i), & \ell = z_{i,t}, \end{cases}$$

where $\eta_i = \eta / n_i$. Iterating $s$ steps yields that [5]

$$\begin{aligned} g_i^s(\theta_\ell) &= (I - \eta_i \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_i))^s \theta_\ell \\ &\quad + \sum_{\tau=0}^{s-1} (I - \eta_i \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_i))^\tau \eta_i \phi(\boldsymbol{x}_i)^\top y_i \\ &\overset{(a)}{=} \theta_\ell - \sum_{\tau=0}^{s-1} (I - \eta_i \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_i))^\tau \eta_i \phi(\boldsymbol{x}_i)^\top (\phi(\boldsymbol{x}_i) \theta_\ell - y_i) \\ &\overset{(b)}{=} \theta_\ell - \eta_i \phi(\boldsymbol{x}_i)^\top P_i (\phi(\boldsymbol{x}_i) \theta_\ell - y_i), \end{aligned}$$

where $(a)$ used $I - (I - X)^s = \sum_{\tau=0}^s (I - X)^\tau X$, and $(b)$ used $(I - X^\top X)^\tau X^\top = X^\top (I - X X^\top)^\tau$ and the definition of $P_i$. Then,

$$\begin{aligned} \theta_{i\ell,t} &= [\mathcal{G}_i^s(\theta_{t-1})]_\ell = \lambda_{i\ell,t} g_i^s(\theta_{\ell,t-1}) + (1 - \lambda_{i\ell,t}) \theta_{\ell,t-1} \\ &= \theta_{\ell,t-1} - \eta_i \lambda_{i\ell,t} \phi(\boldsymbol{x}_i)^\top P_i (\phi(\boldsymbol{x}_i) \theta_{\ell,t-1} - y_i). \end{aligned}$$

We obtain the global iteration:

$$\begin{aligned} \theta_{\ell,t} &= \sum_{i=1}^M \frac{n_i}{N} \theta_{i\ell,t} \\ &= \theta_{\ell,t-1} - \frac{\eta}{N} \sum_{i=1}^M \lambda_{i\ell,t} \phi(\boldsymbol{x}_i)^\top P_i (\phi(\boldsymbol{x}_i) \theta_{\ell,t-1} - y_i), \end{aligned}$$

which is (10) using matrix notations.

The proof for FedProx is similar. The first-order condition for the local proximal optimization is

$$\eta_i \lambda_{i\ell,t} \phi(\boldsymbol{x}_i)^\top (\phi(\boldsymbol{x}_i) \theta_{i\ell,t} - y_i) + (\theta_{i\ell,t} - \theta_{\ell,t-1}) = 0, \ \ell \in [k].$$

Therefore, if $\ell \neq z_{i,t}$, then $\theta_{i\ell,t} = \theta_{\ell,t-1}$; if $\ell = z_{i,t}$, then

$$\begin{aligned} \theta_{i\ell,t} &= (I + \eta_i \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_i))^{-1} (\theta_{\ell,t-1} + \eta_i \phi(\boldsymbol{x}_i)^\top y_i) \\ &\overset{(a)}{=} \theta_{\ell,t-1} \\ &\quad - \eta_i (I + \eta_i \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_i))^{-1} \phi(\boldsymbol{x}_i)^\top (\phi(\boldsymbol{x}_i) \theta_{\ell,t-1} - y_i) \\ &\overset{(b)}{=} \theta_{\ell,t-1} - \eta_i \phi(\boldsymbol{x}_i)^\top P_i (\phi(\boldsymbol{x}_i) \theta_{\ell,t-1} - y_i), \end{aligned}$$

where $(a)$ used $I - (I + X)^{-1} = (I + X)^{-1} X$, and $(b)$ used $(I + X^\top X)^{-1} X^\top = X^\top (I + X X^\top)^{-1}$ and the definition of $P_i$. The remaining steps are the same as those in FedAvg. □

### *B. Convergence Analysis of Phase 2*

We analyze the three terms on the right-hand side of (11) separately. The first term of (11) is the main term due to the decreasing of estimation error, and the last term is the stochastic variation due to the observation noise $\zeta$. We have the following lemmas on the eigenvalues of $K_\ell$ and the concentration of the observation noise.

*Lemma 7:* Suppose that $\min_{\ell\in[k]} N_\ell \geq C_0 d$ for a sufficiently large constant $C_0$. There exists constants $c$ and $C$ such that, with probability $1-2ke^{-d}$,

$$c\alpha\frac{sN_\ell}{\kappa N} \leq \lambda_{\min}(K_\ell) \leq \lambda_{\max}(K_\ell) \leq C\beta\frac{sN_\ell}{N}, \quad \forall \ell \in [k].$$

*Proof:* Since $\phi(\boldsymbol{x}_{I_\ell})$ of size $N_\ell \times d$ consists of independent and sub-Gaussian rows, by a covering argument [37, Theorem 4.6.1], with probability $1-2e^{-d}$,

$$\begin{aligned}\alpha N_\ell - C(\sqrt{dN_\ell}\vee d) &\leq \sigma^2_{\min}(\phi(\boldsymbol{x}_{I_\ell})) \leq \sigma^2_{\max}(\phi(\boldsymbol{x}_{I_\ell}))\\ &\leq \beta N_\ell + C(\sqrt{dN_\ell}\vee d),\end{aligned}$$

where $\sigma_{\max}$ and $\sigma_{\min}$ denote the largest and smallest singular values, respectively, and $C$ is an absolute constant. By definition, $K_\ell = \frac{1}{N}\phi(\boldsymbol{x}_{I_\ell})^\top P_{I_\ell}\phi(\boldsymbol{x}_{I_\ell})$, where $P_{I_\ell}$ is a symmetric matrix. It is shown in [5, Lemma 3] that

$$s/\kappa \leq \lambda_{\min}(P_{I_\ell}) \leq \lambda_{\max}(P_{I_\ell}) \leq s.$$

The conclusion follows from the condition $N_\ell \geq C_0 d$ and a union bound over $\ell \in [k]$. □

*Lemma 8:* Given the input features $\phi(\boldsymbol{x})$, there exists a constant $C$ such that with probability at least $1 - k\exp(-d)$,

$$\|B\Lambda_\ell\zeta\|_2^2 \leq C\frac{\sigma^2 sd}{N}\|K_\ell\|_2, \quad \forall\ \ell \in [k].$$

*Proof:* Note that

$$\|B\Lambda_\ell\zeta\|_2^2 = \zeta^\top\Lambda_\ell B^\top B\Lambda_\ell\zeta = \langle \Lambda_\ell B^\top B\Lambda_\ell, \zeta\zeta^\top\rangle.$$

Since $\mathbb{E}\left[\zeta\zeta^\top\right] \preceq \sigma^2 I$, it follows that

$$\begin{aligned}\mathbb{E}\left[\|B\Lambda_\ell\zeta\|_2^2\right] &= \mathbb{E}\left[\langle\Lambda_\ell B^\top B\Lambda_\ell, \zeta\zeta^\top\rangle\right]\\ &\leq \sigma^2\mathsf{Tr}\left(\Lambda_\ell B^\top B\Lambda_\ell\right) = \sigma^2\mathsf{Tr}\left(B\Lambda_\ell^2 B^\top\right).\end{aligned}$$

Recall that

$$\begin{aligned}B\Lambda_\ell^2 B^\top &= \frac{1}{N^2}\phi(\boldsymbol{x}_{I_\ell})^\top P_{I_\ell}^2\phi(\boldsymbol{x}_{I_\ell})\\ &\overset{(a)}{\preceq} \frac{s}{N^2}\phi(\boldsymbol{x}_{I_\ell})^\top P_{I_\ell}\phi(\boldsymbol{x}_{I_\ell}) = \frac{s}{N}K_\ell, \end{aligned} \tag{28}$$

where $(a)$ holds because $\|P_{I_\ell}\|_2 \leq s$. Therefore,

$$\mathbb{E}\left[\|B\Lambda_\ell\zeta\|_2^2\right] = \mathbb{E}\left[\langle\Lambda_\ell B^\top B\Lambda_\ell, \zeta\zeta^\top\rangle\right] \leq \frac{\sigma^2 sd}{N}\|K_\ell\|_2.$$

Next, using Hanson-Wright's inequality [38], we get

$$\begin{aligned}&\mathbb{P}\left\{\langle\Lambda_\ell B^\top B\Lambda_\ell, \zeta\zeta^\top\rangle - \mathbb{E}\left[\langle\Lambda_\ell B^\top B\Lambda_\ell, \zeta\zeta^\top\rangle\right] \geq \delta\right\}\\ &\leq \exp\left(-c_1\min\left\{\frac{\delta}{\sigma^2\|\Lambda_\ell B^\top B\Lambda_\ell\|_2}, \frac{\delta^2}{\sigma^4\|\Lambda_\ell B^\top B\Lambda_\ell\|_{\mathrm{F}}^2}\right\}\right),\end{aligned}$$

where $c_1 > 0$ is a universal constant. Note that

$$\|\Lambda_\ell B^\top B\Lambda_\ell\|_2 = \|B\Lambda_\ell^2 B^\top\|_2 \leq \frac{s}{N}\|K_\ell\|_2,$$

$$\|\Lambda_\ell B^\top B\Lambda_\ell\|_{\mathrm{F}} = \|B\Lambda_\ell^2 B^\top\|_{\mathrm{F}} \leq s\|K_\ell\|_{\mathrm{F}} \leq \frac{s\sqrt{d}}{N}\|K_\ell\|_2.$$

Therefore, by choosing $\delta = C\frac{\sigma^2 sd}{N}\|K_\ell\|_2$ for a sufficiently large constant $C$, we get that with probability at least $1 - \exp(-d)$,

$$\begin{aligned}\langle\Lambda_\ell B^\top B\Lambda_\ell, \zeta\zeta^\top\rangle &\leq \mathbb{E}\left[\langle\Lambda_\ell B^\top B\Lambda_\ell, \zeta\zeta^\top\rangle\right] + \delta\\ &\leq (C+1)\,\sigma^2\frac{sd}{N}\|K_\ell\|_2.\end{aligned}$$

The conclusion follows from a union bound over all $\ell \in [k]$. □

Combining Lemmas 2, 7, and 8, next we prove Theorem 2.

*Proof of Theorem 2:* We prove the result by conditioning on the high probability events in Lemmas 2, 7, and 8 that happen with probability at least $1 - Cke^{-d}$. In view of Lemma 7 and the assumption that $\beta s\eta \leq c_0$, we obtain that

$$\|I - \eta K_\ell\|_2 \leq 1 - c\alpha\eta s\rho/\kappa.$$

Combining Lemmas 7 and 8 yields

$$\|B\Lambda_\ell\zeta\|_2 \leq Cs\sigma\sqrt{\frac{\beta d}{N}}.$$

Plugging the above upper bounds and Lemma 2 into (11), we get

$$\begin{aligned}\|\theta_{\ell,t} - \theta_\ell^*\|_2 \leq &\left(1 - \eta s\left(c\frac{\alpha\rho}{\kappa} - C\beta\nu\log\frac{e}{\nu}\right)\right)d(\theta_{t-1}, \theta^*)\\ &+ C\eta s\sigma\sqrt{\beta}\left(\sqrt{\frac{d}{N}} + \nu\log\frac{e}{\nu}\right), \quad \forall\ell \in [k].\end{aligned}$$

Since $\nu\log\frac{e}{\nu} \leq \frac{c\rho\alpha}{2C\kappa\beta}$ and $\nu \gtrsim \sqrt{d/N}$, we conclude (8).

Let $\hat{\theta}_{i,t} = \theta_{z_{i,t},t}$ be client $i$'s estimate of its own model parameter. If client $i$ is clustered correctly such that $z_{i,t} = z_i$, where the success probability $\mathbb{P}\{z_{i,t} = z_i\}$ is shown in Lemma 10 (which can be found in Appendix 2), it follows from (8) that, for $t \geq T+1$,

$$\begin{aligned}&\|\hat{\theta}_{i,t} - \theta_{z_i}^*\|_2 \leq d(\theta_t, \theta^*)\\ &\leq (1 - c_2 s\eta\rho\alpha/\kappa)^{t-T}\,d(\theta_T, \theta^*) + \frac{C_2}{c_2}\frac{\sigma\kappa\sqrt{\beta}}{\rho\alpha}\nu\log\frac{e}{\nu}.\end{aligned}$$

The proof is completed. □

*1) Proof of Lemma 2:* This subsection is devoted to the proof of Lemma 2 using the following road map:

$$\begin{aligned}&d(\theta_t, \theta^*)\downarrow \Longrightarrow \sum_{i:i\in S_{\ell,t}} n_i \downarrow \Longrightarrow \|\phi(\boldsymbol{x}_{S_{\ell,t}})\|_2, \|\zeta_{S_{\ell,t}}\|_2\\ &\downarrow \Longrightarrow \|B\mathcal{E}_{\ell,t}(\phi(\boldsymbol{x})\theta_{\ell,t-1} - y)\|_2 \downarrow.\end{aligned}$$

Specifically, a small estimation error $d(\theta_t, \theta^*)$ implies an upper bound on the total number of incorrectly clustered data points $\sum_{i\in S_{\ell,t}} n_i$; then we upper bound $\|\phi(\boldsymbol{x}_{S_\ell^t})\|_2$ and $\|\zeta_{S_\ell^t}\|_2$ using sub-Gaussian concentration and the union bound; finally we conclude the result from (14).

We first upper bound $\sum_{i\in S_{\ell,t}} n_i$. Using (7), the set $S_{\ell,t} = I_\ell \ominus I_{\ell,t}$ is equivalently the union of

$$\begin{aligned}&I_\ell - I_{\ell,t} =\\ &\left\{i \in I_\ell : \|y_i - \phi(\boldsymbol{x}_i)\theta_{\ell,t-1}\|_2 \geq \min_{\ell'\neq\ell}\|y_i - \phi(\boldsymbol{x}_i)\theta_{\ell',t-1}\|_2\right\},\end{aligned}$$

$$I_{\ell,t} - I_\ell = \left\{ i \notin I_\ell : \|y_i - \phi(\boldsymbol{x}_i)\theta_{\ell,t-1}\|_2 \le \min_{\ell' \neq \ell} \|y_i - \phi(\boldsymbol{x}_i)\theta_{\ell',t-1}\|_2 \right\}.$$

Therefore, $S_{\ell,t} = S_\ell(\theta_{t-1})$, where $S_\ell$ is defined in (15). The next lemma upper bounds the VC dimensions of the binary function classes specified in (16).

*Lemma 9:* For $k \ge 2$, the VC dimensions of $\mathcal{F}_\ell^{\mathrm{I}}$ and $\mathcal{F}_\ell^{\mathrm{II}}$ are at most $Cdk \log k$ for a constant $C > 0$.

*Proof:* We focus on the proof for $\mathcal{F}_\ell^{\mathrm{I}}$ for a fixed $\ell \in [k]$, and the proof for $\mathcal{F}_\ell^{\mathrm{II}}$ is similar. We count the number of faces in the arrangement of geometric objects, which is also known as the number of sign patterns. Specifically, here we define the sign patterns of binary functions $g_1(\theta), \dots, g_m(\theta)$ as the set

$$\left\{ (g_1(\theta), \dots, g_m(\theta)) : \theta \in \mathbb{R}^{dk} \right\}.$$

Suppose $\mathcal{F}_\ell^{\mathrm{I}}$ shatters $m$ points denoted by $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m)$. Define binary functions

$$q_{i,\ell'}(\theta) \triangleq \mathbb{1}\{P_{\ell\ell'}[\boldsymbol{x}_i, y_i](\theta) \ge 0\}, \quad g_i(\theta) \triangleq \max_{\ell' \neq \ell} q_{i,\ell'}(\theta).$$

It is necessary that the number of sign patterns of $g_1(\theta), \dots, g_m(\theta)$ is $2^m$. Note that every $P_{\ell',\ell}[\boldsymbol{x}_i, y_i]$ is a $(dk)$-variate quadratic function. By the Milnor-Thom theorem (see, e.g., [34, Theorem 6.2.1]), if $m(k-1) \ge dk \ge 2$, the number of sign patterns of $m(k-1)$ binary functions $q_{1,\ell'}, \dots, q_{m,\ell'}$ for $\ell' \neq \ell$ is at most $(\frac{100m(k-1)}{dk})^{dk}$. Since each $g_i$ is the maximum of $q_{i,\ell'}$ over $\ell' \neq \ell$, the number of sign patterns of $g_1, \dots, g_m$ is upper bounded by $(\frac{100m(k-1)}{dk})^{dk}$. Consequently, we obtain $2^m \le (\frac{100m(k-1)}{dk})^{dk}$, and hence $m \le Cdk \log k$. If instead, $m(k-1) < dk$, then the conclusion $m \le Cdk \log k$ trivially holds. □

Next, we show the uniform deviation of the incorrectly clustered data points. Due to the quantity skew, we consider a weighted empirical process $G_\ell(\theta) = \sum_{i=1}^M n_i \mathbb{1}\{i \in S_\ell(\theta)\}$. Since the local data $(\boldsymbol{x}_i, y_i)$ on different clients are independent, for a fixed $\theta$, the events $\{i \in S_\ell(\theta)\}$ as functions of $(\boldsymbol{x}_i, y_i)$ are mutually independent. Using the binary function classes in (16), we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_\theta |G_\ell(\theta) - \mathbb{E}[G_\ell(\theta)]|\right] \\
&\le \mathbb{E}\left[\sup_{f \in \mathcal{F}_\ell^{\mathrm{I}}} \left| \sum_{i \in I_\ell} n_i (f(\boldsymbol{x}_i, y_i) - \mathbb{E}[f(\boldsymbol{x}_i, y_i)]) \right|\right] \\
&\quad + \mathbb{E}\left[\sup_{f \in \mathcal{F}_\ell^{\mathrm{II}}} \left| \sum_{i \notin I_\ell} n_i (f(\boldsymbol{x}_i, y_i) - \mathbb{E}[f(\boldsymbol{x}_i, y_i)]) \right|\right] \\
&\le C\sqrt{dk \log k \sum_{i \in I_\ell} n_i^2} + C\sqrt{dk \log k \sum_{i \notin I_\ell} n_i^2} \\
&\le 2C\sqrt{dk \log k \sum_{i=1}^M n_i^2},
\end{aligned} \tag{29}$$

where the second inequality follows from the uniform deviation of weighted empirical processes in Lemma 12 and the upper bound of VC dimensions in Lemma 9. Finally, we use the McDiarmid's inequality to establish a high-probability tail bound. Note that we can write

$$\sup_\theta |G_\ell(\theta) - \mathbb{E}[G_\ell(\theta)]| \triangleq h(Z_1, \dots, Z_M)$$

as a function $h$ of $Z_i = (\boldsymbol{x}_i, y_i)$ with bounded differences: for any $i, z_i, z_i'$,

$$|h(z_1, \dots, z_i, \dots, z_M) - h(z_1, \dots, z_i', \dots, z_M)| \le n_i.$$

By McDiarmid's inequality, we have

$$\begin{aligned}
&\mathbb{P}\left\{h(Z_1, \dots, Z_M) - \mathbb{E}\left[h(Z_1, \dots, Z_M)\right] \ge t\right\} \\
&\le \exp\left(-\frac{2t^2}{\sum_{i=1}^M n_i^2}\right).
\end{aligned} \tag{30}$$

Therefore, combining (29) and (30), and by a union bound, with probability at least $1 - k^{-2dk}$,

$$\begin{aligned}
&\sup_\theta |G_\ell(\theta) - \mathbb{E}[G_\ell(\theta)]| \\
&\le (2C+1)\sqrt{dk \log k \sum_{i=1}^M n_i^2} \\
&= (2C+1)N\sqrt{\frac{dk \log k}{M}}(\chi^2(n)+1), \quad \forall \ell \in [k].
\end{aligned} \tag{31}$$

*Lemma 10:* Suppose $\epsilon \le \frac{\sqrt{\alpha/\beta}}{3}\Delta$. Then,

$$\sup_{\theta : d(\theta, \theta^*) \le \epsilon} \mathbb{P}[i \in S_\ell(\theta)] \le 4k \exp\left(-cn_i\alpha^2\left(1 \wedge \frac{\Delta^2}{\sigma^2}\right)^2\right),$$

for all $\ell \in [k]$, where $c$ is an absolute constant.

*Proof:* For $i \in I_\ell$, it follows from (15) and the union bound that

$$\begin{aligned}
&\mathbb{P}\{i \in S_\ell(\theta)\} \\
&\le \sum_{\ell' \neq \ell} \mathbb{P}\{\|y_i - \phi(\boldsymbol{x}_i)\theta_\ell\|_2 \ge \|y_i - \phi(\boldsymbol{x}_i)\theta_{\ell'}\|_2\} \\
&= \sum_{\ell' \neq \ell} \mathbb{P}\{\|\phi(\boldsymbol{x}_i)(\theta_\ell^* - \theta_\ell) + \zeta_i\|_2 \ge \|\phi(\boldsymbol{x}_i)(\theta_\ell^* - \theta_{\ell'}) + \zeta_i\|_2\}.
\end{aligned} \tag{32}$$

For any $u \in \mathbb{R}^d$, the $n_i$-dimensional random vector $\phi(\boldsymbol{x}_i)u + \zeta_i$ has independent and $(\|u\|_2^2 + \sigma^2)$-sub-Gaussian coordinates. Applying Bernstein inequality yields that

$$\begin{aligned}
&\mathbb{P}\left\{\left|\frac{1}{n_i}\|\phi(\boldsymbol{x}_i)u + \zeta_i\|_2^2 - \left(\mathbb{E}[\zeta_{i1}^2] + \|u\|_{\Sigma_i}^2\right)\right| \ge (\|u\|_2^2 + \sigma^2)t\right\} \\
&\le 2\exp\left(-cn_i(t \wedge t^2)\right),
\end{aligned} \tag{33}$$

where $\Sigma_i = \mathbb{E}[\phi(x_{i1})\phi(x_{i1})^\top]$. Let $u_1 \triangleq \theta_\ell^* - \theta_\ell$ and $u_2 \triangleq \theta_\ell^* - \theta_{\ell'}$. By assumptions that $\|\theta_{\ell'} - \theta_{\ell'}^*\|_2 \le \epsilon$ for all $\ell' \in [k]$ and $\|\theta_\ell^* - \theta_{\ell'}^*\|_2 \ge \Delta$ for $\ell' \neq \ell$, we have $\|u_1\|_2 \le \epsilon$, $\|u_2\|_2 \ge \Delta - \epsilon$. Applying the condition $\epsilon \le \frac{1}{3\sqrt{\beta/\alpha}}\Delta \le \frac{1}{3}\Delta$, we get

$$\|u_2\|_{\Sigma_i}^2 - \|u_1\|_{\Sigma_i}^2 \ge \alpha(\Delta - \epsilon)^2 - \beta\epsilon^2 \ge \alpha\Delta^2/3. \tag{34}$$

Therefore, let $m = \mathbb{E}[\zeta_{i1}^2] + (1-p)\|u_1\|_{\Sigma_i}^2 + p\|u_2\|_{\Sigma_i}^2$ with $p = \frac{\|u_1\|_2^2 + \sigma^2}{\|u_1\|_2^2 + \|u_2\|_2^2 + 2\sigma^2}$, and we obtain from (33) that

$$\mathbb{P}\{\|\phi(\boldsymbol{x}_i)(\theta_\ell^* - \theta_\ell) + \zeta_i\|_2 \ge \|\phi(\boldsymbol{x}_i)(\theta_\ell^* - \theta_{\ell'}) + \zeta_i\|_2\}$$

$$
\begin{aligned}
&\le \mathbb{P}\left\{\frac{1}{n_i}\|\phi(\boldsymbol{x}_i)u_1+\zeta_i\|_2^2 \ge m\right\} \\
&\quad + \mathbb{P}\left\{\frac{1}{n_i}\|\phi(\boldsymbol{x}_i)u_2+\zeta_i\|_2^2 \le m\right\} \\
&\le 4\exp\left(-cn_i(t\wedge t^2)\right),
\end{aligned}
$$

where $t=\frac{\|u_2\|_{\Sigma_i}^2-\|u_1\|_{\Sigma_i}^2}{\|u_1\|_2^2+\|u_2\|_2^2+2\sigma^2}\ge c_0\alpha(1\wedge\frac{\Delta^2}{\sigma^2})$ for a constant $c_0>0$ using the lower bound of seperation in (34). We conclude the proof for $i\in I_\ell$ from (32). Similarly, for $i\in I_{\ell'}$ with $\ell'\ne\ell$, we have

$$
\begin{aligned}
&\mathbb{P}\left\{i\in S_\ell(\theta)\right\} \\
&\le \mathbb{P}\left\{\|\phi(\boldsymbol{x}_i)(\theta_{\ell'}^*-\theta_{\ell'})+\zeta_i\|_2 \ge \|\phi(\boldsymbol{x}_i)(\theta_{\ell'}^*-\theta_\ell)+\zeta_i\|_2\right\}.
\end{aligned}
$$

The conclusion follows from a similar argument. □

Let $N_I\triangleq\sum_{i\in I}n_i$ denote the total number of data in a subset of clients $I\subseteq[M]$. It follows from (31) and Lemma 10 that, with probability $1-k^{-dk}$,

$$
N_{S_{\ell,t}}=\sum_{i\in S_{\ell,t}}n_i\le\nu N, \tag{35}
$$

where $\nu$ is defined in (9). Conditioning on total number of incorrectly clustered data points $N_I$, the next lemma upper bounds $\|\phi(\boldsymbol{x}_I)\|_2$ and $\|\zeta_I\|_2$.

*Lemma 11:* With probability $1-4e^{-d}$, there exists a constant $C>0$ such that

$$
\sup_{N_I\le\nu N}\frac{1}{N}\|\phi(\boldsymbol{x}_I)\|_2^2\le\beta C\nu\log\frac{e}{\nu}, \tag{36}
$$

$$
\sup_{N_I\le\nu N}\frac{1}{N}\|\zeta_I\|_2^2\le C\sigma^2\nu\log\frac{e}{\nu}. \tag{37}
$$

*Proof:* Since $\phi(x_{ij})$ are independent and sub-Gaussian random vectors in $\mathbb{R}^d$, for a fixed $I\subseteq[M]$, with probability at least $1-2e^{-t}$,

$$
\|\phi(\boldsymbol{x}_I)\|_2^2\le\beta N_I+C'\left(\sqrt{(d+t)N_I}+(d+t)\right),
$$

for some absolute constant $C'>0$. There are at most $\binom{N}{\nu N}\le\exp(N\nu\log(e/\nu))$ many different $I$ with $N_I\le N'$. Hence, applying the union bound yields that, with probability at least $1-2e^{-d}$,

$$
\sup_{N_I\le\nu N}\|\phi(\boldsymbol{x}_I)\|_2^2\lesssim\beta N\nu\log\frac{e}{\nu},
$$

where we used $\nu\gtrsim\frac{d}{N}$. Since $\zeta_{ij}$ are independent and sub-Gaussian with $\mathbb{E}[\zeta_{ij}^2]\le\sigma^2$, the inequality in (37) follows from a similar argument. □

Conditioning on the high probability events of (35), (36) and (37), we obtain

$$
\|\phi(\boldsymbol{x}_{S_{\ell,t}})\|_2\le C\sqrt{\beta N\nu\log\frac{e}{\nu}},\quad \|\zeta_{S_{\ell,t}}\|_2\le C\sigma\sqrt{N\nu\log\frac{e}{\nu}}.
$$

Since $\|P_{S_{\ell,t}}\|_2\le s$, we conclude from (13) and (14) that

$$
\begin{aligned}
&\|B\mathcal{E}_{\ell,t}(\phi(\boldsymbol{x})\theta_{\ell,t-1}-y)\|_2 \\
&\le\frac{1}{N}\|\phi(\boldsymbol{x}_{S_{\ell,t}})\|_2\|P_{S_{\ell,t}}\|_2\|\phi(\boldsymbol{x}_{S_{\ell,t}})\theta_{\ell,t-1}-y_{S_{\ell,t}}\|_2 \\
&\le\frac{s}{N}\left(d(\theta_{t-1},\theta^*)\|\phi(\boldsymbol{x}_{S_{\ell,t}})\|_2^2+\|\phi(\boldsymbol{x}_{S_{\ell,t}})\|_2\|\zeta_{S_{\ell,t}}\|_2\right) \\
&\le C^2s(\beta d(\theta_{t-1},\theta^*)+\sigma\sqrt{\beta})\nu\log\frac{e}{\nu}.
\end{aligned}
$$

*2) Auxiliary Lemma:*

*Lemma 12:* Consider a weighted empirical process $G_n(f)=\sum_{i=1}^n\lambda_i f(X_i)$ for binary functions $f\in\mathcal{F}$, where $X_i$'s are independent and the VC dimension of $\mathcal{F}$ is at most $d$. Then

$$
\mathbb{E}\left[\sup_{f\in\mathcal{F}}|G_n(f)-\mathbb{E}G_n(f)|\right]\lesssim\sqrt{d\sum_{i=1}^n\lambda_i^2}.
$$

*Proof:* Since $X_i$'s are independent, by symmetrization,

$$
\mathbb{E}\left[\sup_{f\in\mathcal{F}}|G_n(f)-\mathbb{E}G_n(f)|\right]\le 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n\epsilon_i\lambda_i f(X_i)\right|\right],
$$

where $\epsilon_i$ are i.i.d. Rademacher random variables. Next, by conditioning on $X_i$'s, we aim to apply Dudley's integral. Since $\epsilon_i$ are independent and 1-sub-Gaussian, for any $f,g\in\mathcal{F}$, the increment $\sum_i\epsilon_i\lambda_i f(X_i)-\sum_i\epsilon_i\lambda_i g(X_i)$ is also sub-Gaussian with a variance parameter

$$
\sum_{i=1}^n\lambda_i^2(f-g)(X_i)^2=\left(\sum_{i=1}^n\lambda_i^2\right)\|f-g\|_{L^2(\mu_n)}^2,
$$

where $\mu_n$ denotes the weighted empirical measure $\frac{1}{\sum_i\lambda_i^2}\sum_i\lambda_i^2\delta_{X_i}$. Apply Dudley's integral (see, e.g., [37, Theorem 8.1.3]) conditioning on $X_i$'s, we get that

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n\epsilon_i\lambda_i f(X_i)\right|\right] \\
&\lesssim\sqrt{\sum_{i=1}^n\lambda_i^2}\times\mathbb{E}\left[\int_0^1\sqrt{\log\mathcal{N}(\mathcal{F},L^2(\mu_n),\epsilon)}d\epsilon\right],
\end{aligned}
$$

where $\mathcal{N}(\mathcal{F},L^2(\mu_n),\epsilon)$ denotes the $\epsilon$-covering number of $\mathcal{F}$ under $L^2(\mu_n)$. Finally, we can bound the covering number by the VC dimension of $\mathcal{F}$ as (see, e.g., [37, Theorem 8.3.18])

$$
\log\mathcal{N}(\mathcal{F},L^2(\mu_n),\epsilon)\lesssim d\log\frac{2}{\epsilon}.
$$

The conclusion follows. □

## Appendix C
## Truncated Matrix Bernstein Inequality

*Lemma 13:* Let $\{a_i:i\in[N]\}$ and $\{b_i:i\in[N]\}$ denote two independent sequences of independent random vectors in $\mathbb{R}^d$. Suppose that $\mathbb{E}\left[\|a_i\|_2^2\right]\le C_1d$, $\mathbb{E}\left[\|b_i\|_2^2\right]\le C_1d$, $\left\|\mathbb{E}\left[a_ia_i^\top\right]\right\|_2\le C_1$, $\left\|\mathbb{E}\left[b_ib_i^\top\right]\right\|_2\le C_1$, and

$$
\mathbb{P}\left\{\|a_i\|_2\ge t\right\},\mathbb{P}\left\{\|b_i\|_2\ge t\right\}\le\exp\left(-C_2t/\sqrt{d}\right),\ \forall t\ge\sqrt{d}
$$

for some universal constants $C_1,C_2>0$. Let

$$
Y=\sum_{i=1}^N\left(a_ib_i^\top-\mathbb{E}\left[a_ib_i^\top\right]\right).
$$

Then there exists a constant $C>0$ (only depending on $C_1,C_2$) such that with probability at least $1-3\delta$,

$$
\|Y\|_2\le C\left(\sqrt{Nd\log\frac{1}{\delta}}+d\log^3(N/\delta)\right).
$$

*Proof:* Given $\tau$ to be specified later, define event $\mathcal{E}_i = \{\|a_i b_i^\top\|_2 \le \tau\}$. It follows that

$$\begin{aligned} Y = &\sum_{i=1}^N \left(a_i b_i^\top \mathbb{1}\{\mathcal{E}_i\} - \mathbb{E}\left[a_i b_i^\top \mathbb{1}\{\mathcal{E}_i\}\right]\right) \\ &+ \sum_{i=1}^N a_i b_i^\top \mathbb{1}\{\mathcal{E}_i^c\} - \sum_{i=1}^N \mathbb{E}\left[a_i b_i^\top \mathbb{1}\{\mathcal{E}_i^c\}\right] \end{aligned}$$

and hence

$$\begin{aligned} \|Y\|_2 \le &\left\| \sum_{i=1}^N \left(a_i b_i^\top \mathbb{1}\{\mathcal{E}_i\} - \mathbb{E}\left[a_i b_i^\top \mathbb{1}\{\mathcal{E}_i\}\right]\right) \right\|_2 \\ &+ \left\| \sum_{i=1}^N a_i b_i^\top \mathbb{1}\{\mathcal{E}_i^c\} \right\|_2 + \left\| \sum_{i=1}^N \mathbb{E}\left[a_i b_i^\top \mathbb{1}\{\mathcal{E}_i^c\}\right] \right\|_2 . \end{aligned} \tag{38}$$

In the sequel, we bound each term in the RHS separately.

To bound the first term, we will use the matrix Bernstein inequality. Let $Y_i = a_i b_i^\top \mathbb{1}\{\mathcal{E}_i\} - \mathbb{E}\left[a_i b_i^\top \mathbb{1}\{\mathcal{E}_i\}\right]$. Then $\mathbb{E}[Y_i] = 0$ and

$$\|Y_i\|_2 \le \left\|a_i b_i^\top \mathbb{1}\{\mathcal{E}_i\}\right\|_2 + \left\|\mathbb{E}\left[a_i b_i^\top \mathbb{1}\{\mathcal{E}_i\}\right]\right\|_2 \le 2\tau.$$

Moreover,

$$\begin{aligned} \sum_{i=1}^N \mathbb{E}\left[Y_i Y_i^\top\right] &= \sum_{i=1}^N \mathbb{E}\Big[\left(a_i b_i^\top \mathbb{1}\{\mathcal{E}_i\} - \mathbb{E}\left[a_i b_i^\top \mathbb{1}\{\mathcal{E}_i\}\right]\right) \\ &\qquad \times \left(a_i b_i^\top \mathbb{1}\{\mathcal{E}_i\} - \mathbb{E}\left[a_i b_i^\top \mathbb{1}\{\mathcal{E}_i\}\right]\right)^\top\Big] \\ &= \sum_{i=1}^N \Big(\mathbb{E}\left[a_i a_i^\top \|b_i\|_2^2 \mathbb{1}\{\mathcal{E}_i\}\right] \\ &\qquad - \mathbb{E}\left[a_i b_i^\top \mathbb{1}\{\mathcal{E}_i\}\right] \mathbb{E}\left[a_i b_i^\top \mathbb{1}\{\mathcal{E}_i\}\right]^\top\Big). \end{aligned}$$

Therefore,

$$\begin{aligned} \sum_{i=1}^N \mathbb{E}\left[Y_i Y_i^\top\right] &\preceq \sum_{i=1}^N \mathbb{E}\left[a_i a_i^\top \|b_i\|_2^2 \mathbb{1}\{\mathcal{E}_i\}\right] \\ &\preceq \sum_{i=1}^N \mathbb{E}\left[a_i a_i^\top \|b_i\|_2^2\right] \\ &= \sum_{i=1}^N \mathbb{E}\left[\|b_i\|_2^2\right] \mathbb{E}\left[a_i a_i^\top\right] \preceq C_1^2 N d \mathbf{I}. \end{aligned}$$

Moreover, $Y_i Y_i^\top \succeq 0$. Hence, $\left\|\sum_{i=1}^N \mathbb{E}\left[Y_i Y_i^\top\right]\right\|_2 \le C_1^2 \ Nd$. Similarly, we can show that $\left\|\sum_{i=1}^N \mathbb{E}\left[Y_i^\top Y_i\right]\right\|_2 \le C_1^2 \ Nd$. Applying the matrix Bernstein inequality [39], we get that with probability at least $1-\delta$,

$$\left\| \sum_{i=1}^N Y_i \right\|_2 \le C_3 \left( \sqrt{Nd \log \frac{1}{\delta}} + \tau \log \frac{1}{\delta} \right), \tag{39}$$

where $C_3 > 0$ is a constant only depending on $C_1$. Next, we bound the second term in (38). Note that on the event $\cap_{i=1}^N \mathcal{E}_i$, $\left\|\sum_{i=1}^N a_i b_i^\top \mathbb{1}\{\mathcal{E}_i^c\}\right\|_2 = 0$. Note that

$$\begin{aligned} \mathbb{P}\{\mathcal{E}_i^c\} &= \mathbb{P}\left\{\|a_i b_i^\top\|_2 > \tau\right\} \\ &\le \mathbb{P}\left\{\|a_i\|_2 \ge \sqrt{\tau}\right\} + \mathbb{P}\left\{\|b_i\|_2 \ge \sqrt{\tau}\right\} \\ &\le 2 \ e^{-C_2 \sqrt{\tau/d}}. \end{aligned}$$

Hence by choosing $\tau = C_2^{-2} d \log^2 \frac{N}{\delta}$ for some sufficiently large constant $C$, we get that $\mathbb{P}\{\mathcal{E}_i^c\} \le 2\delta/N$. Thus by union bound,

$$\mathbb{P}\left\{\cap_{i=1}^N \mathcal{E}_i\right\} \ge 1 - \sum_{i=1}^N \mathbb{P}\{\mathcal{E}_i^c\} \ge 1 - 2\delta. \tag{40}$$

Finally, we bond the third term in (38). Note that

$$\begin{aligned} \left\| \sum_{i=1}^N \mathbb{E}\left[a_i b_i^\top \mathbb{1}\{\mathcal{E}_i^c\}\right] \right\|_2 &\le \sum_{i=1}^N \left\|\mathbb{E}\left[a_i b_i^\top \mathbb{1}\{\mathcal{E}_i^c\}\right]\right\|_2 \\ &\le \sum_{i=1}^N \mathbb{E}\left[\left\|a_i b_i^\top \mathbb{1}\{\mathcal{E}_i^c\}\right\|_2\right]. \end{aligned}$$

Moreover,

$$\begin{aligned} &\mathbb{E}\left[\left\|a_i b_i^\top \mathbb{1}\{\mathcal{E}_i^c\}\right\|_2\right] \\ &= \int_0^\infty \mathbb{P}\left\{\left\|a_i b_i^\top \mathbb{1}\{\mathcal{E}_i^c\}\right\|_2 \ge t\right\} \mathrm{d}t \\ &= \int_0^\tau \mathbb{P}\left\{\left\|a_i b_i^\top\right\|_2 \ge \tau\right\} \mathrm{d}t + \int_\tau^\infty \mathbb{P}\left\{\left\|a_i b_i^\top\right\|_2 \ge t\right\} \mathrm{d}t \\ &\le \tau \frac{\delta}{N} + \int_\tau^\infty \mathbb{P}\left\{\left\|a_i b_i^\top\right\|_2 \ge t\right\} \mathrm{d}t \end{aligned}$$

By assumption, for $t \ge \tau = C_2^{-2} d \log^2 \frac{N}{\delta}$,

$$\begin{aligned} \mathbb{P}\left\{\left\|a_i b_i^\top\right\|_2 \ge t\right\} &\le \mathbb{P}\left\{\|a_i\|_2 \ge \sqrt{t}\right\} + \mathbb{P}\left\{\|b_i\|_2 \ge \sqrt{t}\right\} \\ &\le 2 \ e^{-C_2 \sqrt{t/d}}. \end{aligned}$$

It follows that

$$\begin{aligned} \int_\tau^\infty \mathbb{P}\left\{\left\|a_i b_i^\top\right\|_2 \ge t\right\} \mathrm{d}t &\le 2 \int_\tau^\infty e^{-C_2 \sqrt{t/d}} \mathrm{d}t \\ &= 4d \left(\sqrt{\tau/d} + 1/C_2\right) e^{-C_2 \sqrt{\tau/d}}, \end{aligned}$$

where the equality holds by the identity that $\int_\tau^\infty e^{-\alpha\sqrt{t}} \mathrm{d}t = \frac{2}{\alpha^2}(\sqrt{\tau}\alpha + 1) e^{-\alpha\sqrt{\tau}}$. Therefore,

$$\begin{aligned} \mathbb{E}\left[\left\|a_i b_i^\top \mathbb{1}\{\mathcal{E}_i^c\}\right\|_2\right] &\le \tau \frac{\delta}{N} + 4d \left(\sqrt{\tau/d} + 1/C_2\right) e^{-C_2 \sqrt{\tau/d}} \\ &\le \frac{6 \ d\delta}{N C_2^2} \log^2(N/\delta). \end{aligned} \tag{41}$$

Plugging (39), (40), and (41) into (38) yields the desired conclusion. □

## Appendix D
## Bound on the Largest Principal Angle Between Random Subspaces

Let $U \in \mathbb{R}^{d \times \ell}$ denote an orthogonal matrix and $Q \in \mathbb{R}^{d \times \ell}$ denote a random orthogonal matrix chosen uniformly at random, where $\ell \le d$.

*Lemma 14:* With probability at least $1 - 2\epsilon$,

$$\sigma_{\min}(U^\top Q) \ge c \frac{\epsilon}{\sqrt{\ell}(\sqrt{d} + \log(1/\epsilon))}$$

for a constant $c > 0$.

*Proof:* Since $Q \in \mathbb{R}^{d\times \ell}$ is a random orthogonal matrix, to prove the claim, without loss of generality, we can assume $U = [e_1, e_2, \ldots, e_\ell]$, where $e_i$'s are the standard basis vectors in $\mathbb{R}^d$. Let $A \in \mathbb{R}^{d\times \ell}$ denote a random Gaussian matrix with i.i.d. $\mathcal{N}(0, 1)$ entries and write $A = \begin{bmatrix} X \\ Y \end{bmatrix}$, where $X \in \mathbb{R}^{\ell\times\ell}$ and $Y \in \mathbb{R}^{(d-\ell)\times\ell}$. Then $U^\top Q$ has the same distribution as $X(A^\top A)^{-1/2}$. It follows that $\sigma_{\min}(U^\top Q)$ has the same distribution as $\sigma_{\min}(X(A^\top A)^{-1/2})$. Note that

$$\sigma_{\min}\left(X(A^\top A)^{-1/2}\right) \geq \sigma_{\min}(X)\sigma_{\min}\left((A^\top A)^{-1/2}\right) = \frac{\sigma_{\min}(X)}{\sigma_{\max}(A)}.$$

In view of [40, Corollary 5.35], $\sigma_{\max}(A) \leq C\sqrt{d}+C\log(1/\epsilon)$ with probability at least $1 - \epsilon$. Moreover, in view of [41, Theorem 1.2], $\sigma_{\min}(X) \geq c\epsilon/\sqrt{\ell}$ with probability at least $1 - \epsilon$. The desired conclusion readily follows. □

## Acknowledgment

The authors would like to thank Philippe Rigollet for pointing out the related literature on computing VC dimensions using the Milnor-Thom theorem.

## References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[2] S. Dooms, T. De Pessemier, and L. Martens, "Movietweetings: A movie rating dataset collected from Twitter," in *Proc. Workshop Crowdsourcing Human Comput. Recommender Syst.*, 2013, p. 43.

[3] *Frequency of Facebook Use in the United States As of 3rd Quarter 2020*. Accessed: May 1, 2022. [Online]. Available: https://www.statista.com/statistics/199266/frequency-of-use-among-facebook-users-in-the-united-states/

[4] A. Feuerverger, Y. He, and S. Khatri, "Statistical significance of the Netflix challenge," *Stat. Sci.*, vol. 27, no. 2, pp. 202–231, May 2012.

[5] L. Su, J. Xu, and P. Yang, "A non-parametric view of FedAvg and FedProx: Beyond stationary points," 2021, *arXiv:2106.15216*.

[6] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[7] O. Marfoq, G. Neglia, A. Bellet, L. Kameni, and R. Vidal, "Federated multi-task learning under a mixture of distributions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–14.

[8] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3710–3722, Aug. 2020.

[9] Y. Mansour, M. Mohri, J. Ro, and A. Theertha Suresh, "Three approaches for personalization with applications to federated learning," 2020, *arXiv:2002.10619*.

[10] G. Long et al., "Multi-center federated learning: Clients clustering for better personalization," 2021, *arXiv:2108.08647*.

[11] C. Li, G. Li, and P. K. Varshney, "Federated learning with soft clustering," *IEEE Internet Things J.*, vol. 9, no. 10, pp. 7773–7782, May 2022.

[12] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19586–19597.

[13] P. McCullagh and J. A. Nelder, *Generalized Linear Models*. Evanston, IL, USA: Routledge, 2019.

[14] Y. Li and Y. Liang, "Learning mixtures of linear regressions with nearly optimal complexity," in *Proc. Conf. Learn. Theory*, 2018, pp. 1125–1144.

[15] W. Kong, R. Somani, Z. Song, S. Kakade, and S. Oh, "Meta-learning for mixed linear regression," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5394–5404.

[16] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, vol. 2, 2020, pp. 429–450.

[17] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," 2018, *arXiv:1806.00582*.

[18] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5132–5143.

[19] L. Li, W. Xu, T. Chen, G. B. Giannakis, and Q. Ling, "RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 1544–1551.

[20] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7611–7623.

[21] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-IID data silos: An experimental study," 2021, *arXiv:2102.02079*.

[22] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, 2008.

[23] R. K. Ando, T. Zhang, and P. Bartlett, "A framework for learning predictive structures from multiple tasks and unlabeled data," *J. Mach. Learn. Res.*, vol. 6, no. 11, 2005.

[24] N. Tripuraneni, C. Jin, and M. Jordan, "Provable meta-learning of linear representations," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10434–10443.

[25] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei, "Few-shot learning via learning the representation, provably," 2020, *arXiv:2002.09434*.

[26] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 3557–3568.

[27] Y. Jiang, J. Konecný, K. Rush, and S. Kannan, "Improving federated learning personalization via model agnostic meta learning," 2019, *arXiv:1909.12488*.

[28] S. Caldas et al., "LEAF: A benchmark for federated settings," 2018, *arXiv:1812.01097*.

[29] A. Ghosh, J. Hong, D. Yin, and K. Ramchandran, "Robust federated learning in a heterogeneous environment," 2019, *arXiv:1906.06629*.

[30] A. Ghosh and A. Mazumdar, "An improved algorithm for clustered federated learning," 2022, *arXiv:2210.11538*.

[31] P. Kairouz et al., "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, nos. 1–2, pp. 1–210, 2021.

[32] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., Baltimore, MD, USA: Johns Hopkins Univ. Press, 2013.

[33] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. III," *SIAM J. Numer. Anal.*, vol. 7, no. 1, pp. 1–46, Mar. 1970.

[34] J. Matousek, *Lectures on Discrete Geometry*, vol. 212. Berlin, Germany: Springer, 2013.

[35] P. Goldberg and M. Jerrum, "Bounding the vapnik-chervonenkis dimension of concept classes parameterized by real numbers," in *Proc. 6th Annu. Conf. Comput. Learn. Theory*, 1993, pp. 361–369.

[36] J. Maurice Rojas and M. Vidyasagar, "An improved bound on the VC-dimension of neural networks with polynomial activation functions," 2001, *arXiv:OC/0112208*.

[37] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*, vol. 47. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[38] M. Rudelson and R. Vershynin, "Hanson-wright inequality and sub-Gaussian concentration," *Electron. Commun. Probab.*, vol. 18, pp. 1–9, Jan. 2013.

[39] J. A. Tropp, "An introduction to matrix concentration inequalities," 2015, *arXiv:1501.01571*.

[40] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," 2010, *arXiv:1011.3027*.

[41] S. J. Szarek, "Condition numbers of random matrices," *J. Complex.*, vol. 7, no. 2, pp. 131–149, Jun. 1991.