Global Convergence of Federated Learning for Mixed Regression

Lili Su
Electrical and Computer Engineering
Northeastern University
l.su@northeastern.edu

Jiaming Xu
The Fuqua School of Business
Duke University
jiaming.xu868@duke.edu

Pengkun Yang
Center for Statistical Science
Tsinghua University
yangpengkun@tsinghua.edu.cn

Abstract

This paper studies the problem of model training under Federated Learning when clients exhibit cluster structure. 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 initialization, and works even when local data volume is highly unbalanced – there could exist clients that contain $O(1)$ data points only. Our algorithm first runs moment descent on a few anchor clients (each with $\tilde{\Omega}(k)$ data points) to obtain coarse model estimates. Then each 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 on the clustering errors, which we prove by bounding the VC dimension of general polynomial concept classes based on the theory of algebraic geometry.

1 Introduction

Federated learning (FL) [MMR$^+$17] enables a massive number of clients to collaboratively train models without disclosing raw data. Heterogeneity in clients renders local data non-IID and highly unbalanced. 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 [DDPM13, FAC, FHK12].

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 2 for detailed discussion. Using a common model to serve highly heterogeneous clients has fundamental drawbacks; recent work [SXY21] rigorously quantified the heterogeneity level that the common model can tolerate with. Fully personalized models [SCST17, MNB$^+$21] 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 [SMS20], 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 [MMRS20, XLS$^+$21, LLV21, GCYR20], mostly of which are heuristic and lack of convergence guarantees

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

[MMRS20, XLS+21, LLV21]. Towards formal assurance, [GCYR20] 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 [GCYR20] 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 [GCYR20], we adopt a statistical learning setup. In particular, we contextualize our problem as the canonical mixed regression, where each client has a set of local data generated from one of k unknown regression models. Departing from standard mixed regression, in which each client keeps one data point only [MN19, LL18], in our problem the sizes of local datasets can vary significantly across clients. We only make a mild assumption that there exist a few anchor clients (each with $\tilde{\Omega}(k)$ data points) and sufficiently many clients with at least two data points. Similar mixed regression setup with data heterogeneity has been considered in [KSS+20] in a different context of meta-learning; the focus there 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 a few anchor clients to obtain coarse model estimates based on subspace estimation. In each global iteration of Phase 2, each 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.

2 Related Work

FedAvg [MMR+17] 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 [LSZ+20, ZLL+18, KKM+20]. Based on the number of models trained, existing methods to deal with data heterogeneity can be roughly classified into three categories.

A Common Model: To limit the negative impacts of data heterogeneity on the obtained common model, a variety of techniques based on variance reduction [LSZ+20, LXC+19, KKM+20] and normalization [WLL+20] have been introduced. Their convergence results mostly are derived under strong technical assumptions such as bounded gradient and/or bounded Hessian dissimilarity which do not hold when the underlying truth in the data generation is taken into account [LSZ+20, LXC+19, KKM+20]. In fact, none of them strictly outperform others in different instances of data heterogeneity [LDCH21]. Besides, the generalization errors of the common model with respect to local data are mostly overlooked except for a recent work [SXY21], which shows that the common model can tolerate only a moderate level of model heterogeneity.

Fully Personalized Models: [SCST17] proposed Federated Multi-Task Learning (MTL) wherein different models are learned for each of the massive population of clients [SCST17, MNB+21]. Though conceptually attractive, the convergence behaviors of Federated MTL is far from well-understood because the objective is not jointly convex in the model parameters and the model relationships [SCST17, AEP08, AZB05]. Specifically, [SCST17] 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) [AZB05] and small number of common features [AEP08]. 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 homogeneous local dataset size [AZB05]. Despite recent progress [TJJ21, DHK$^+$20], their results are mainly for linear representation learning and for balanced local data. Parallel to Federated MTL, model personalization is also studied under the Model-Agnostic Meta-Learning (MAML) framework [FMO20, JKRK19] where the global objective is modified to account for the cost of fine-tuning a global model at individual clients. However, they focused on studying the convergence in training errors only– no characterization of the generalization error is given.

Clustered Personalized Models: Clustered Federated Learning (CFL) [SMS20, GCYR20, GHYR19, XLS$^+$21, MMRS20, LLV21] 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 1, i.e., which simultaneously learn clusters and models for each cluster, other attempts have been made to integrate clustering with model training. [SMS20] 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. [GHYR19] 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, [LLV21] proposed soft clustering. Unfortunately, the proposed algorithm is only tested on the simple MNIST and Fashion-MNIST datasets, and no theoretical justifications are given.

3 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 active user in practice. Anchor clients play a crucial rule 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. Note that the feature map ϕ can be non-linear (e.g. polynomials). The cluster label of client i is randomly generated from one of the k components from some unknown $p = (p_1, \ldots, p_k)$ in probability simplex $\boldsymbol{\Delta}^{k-1}$. That is, $\mathbb{P}\{z_i = j\} = p_j$ for $j \in [k]$. In addition, $\|\theta^*_j\|_2 \leq R$ for each component. The feature covariate $\phi(x_{ij})$ is independent and sub-Gaussian $\alpha I_d \preceq \mathbb{E}[\phi(x_{ij})\phi(x_{ij})^\top] \preceq \beta I_d$. We assume that the covariance matrix is identical within the same cluster but may vary across different clusters, i.e., $\mathbb{E}[\phi(x_{ij})\phi(x_{ij})^\top] = \Sigma_j$ if $z_i = j$. The noise ζ_{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 [KMA$^+$21]. For the identifiability of the true cluster models θ^*_j's, we assume a minimum proportion and a pairwise separation of clusters. Formally, let $\Delta = \min_{j \neq j'} \|\theta^*_j - \theta^*_{j'}\|_2$ and $p_{\min} = \min_{j \in [k]} p_j$. For ease of presentation, we assume the parameters $\alpha, \beta = \Theta(1)$, $\sigma/\Delta = O(1)$, and $R/\Delta = O(1)$, while our main results can be extended to show more explicit dependencies on these parameters with careful bookkeeping calculations. Note that even under these assumptions, we still allow R, Δ, σ to scale with model dimension d. Also, the assumption $R/\Delta = O(1)$ basically requires the radius of θ^*_j's is on the same scale as their pairwise separation. It rules out the extreme setting where θ^*_j's themselves are extremely large while their pairwise separations are tiny.

Notations: Let $[n] \triangleq \{1, \ldots, n\}$. 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 $a_n \le Cb_n$ 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 $r = \min\{n, d\}$, $\sigma_1 \ge \cdots \ge \sigma_r \ge 0$ are the singular values, and u_i (v_i) are the corresponding left (right) singular vectors. We call $U = [u_1, u_2, \ldots, u_k]$ as the top-k left singular matrix of A. Let $\mathrm{span}(U) = \mathrm{span}\{u_1, \ldots, u_k\}$ denote the k-dimensional subspace spanned by $\{u_1, \ldots, u_k\}$.

4 Main Results

We propose a two-phase FL algorithm that enables clients to learn the model parameters $\theta_1^*, \ldots, \theta_k^*$ and their clusters simultaneously:

(i) Coarse estimation via FedMD. Run the federated moment descent algorithm to obtain coarse estimates of model parameters θ_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) [MMR+17, LSZ+20].

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

4.1 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 $(\theta_1^*, \ldots, \theta_k^*)$; 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 n_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 $\{\Sigma_j(\theta_j^* - \theta_{i,t})\}_{j=1}^k$ lie in, in collaboration with clients in $\mathcal{S}_t$. In particular, for each anchor client $i \in H$, define

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

We approximate the subspace spanned by $\{\Sigma_j(\theta_j^* - \theta_{i,t})\}_{j=1}^k$ 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 [GVL13, 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 the

```
Phase 1: Federated Moment Descent (FedMD)
1  Input: $n_H, k, m, \ell, T, T_1, T_2 \in \mathbb{N}$, $\alpha, \beta, \epsilon, \Delta \in \mathbb{R}$, $\theta_0 \in \mathbb{R}^d$ with $\|\theta_0\|_2 \le R$
2  Output: $\hat{\theta}_1, \dots, \hat{\theta}_k$
3  PS chooses a set $H$ of $n_H$ anchor clients uniformly at random;
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 $i$ calls power-iteration($A_{i,t} A_{i,t}^\top$, $T_2$) to output $(\hat{\beta}_{i,t}, \hat{\sigma}_{i,t}^2)$ with $A_{i,t}$
         defined in (2);
12     if $\hat{\sigma}_{i,t} > \epsilon\Delta$ then
13         $\theta_{i,t+1} \leftarrow \theta_{i,t} + r_{i,t}\eta_{i,t}$ and reports $\theta_{i,t+1}$, 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         $\theta_{i,t+1} \leftarrow \theta_{i,t}$ and reports $\theta_{i,t+1}$;
16 PS computes the pairwise distance $\|\theta_{i,T} - \theta_{i',T}\|_2$ 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 $\hat{\theta}_j$ to be the center of the estimated $j$-th cluster for $j \in [k]$.
```

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. In our setting, the orthogonal iteration can be implemented in a distributed manner in FL systems as shown in Algorithm 3 in the Appendix D.1.

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

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

where $\mathcal{D}_{i,t}$ consists of 2ℓ 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 $\hat{\sigma}_{i,t}^2$ 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 $\hat{\theta}_j$ are close to θ_j^* up to a permutation of cluster indices.

Theorem 1. *Let $\epsilon \in (0, 1/4)$ be a small but fixed constant. Suppose that*

$$m \ge p_{\min}^{-2} \tilde{\Omega}(d), \ \ell = \tilde{\Omega}(k), \ T = \Omega(1), \ T_1 = \Omega(k \log(Nd)), \ T_2 = \Omega(\log(Nd)). \tag{3}$$

With probability at least $1 - O(n_H T / N^{10})$, for all initialization θ_0 with $\|\theta_0\|_2 \le R$,

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

Furthermore, when $n_H \ge \log N / p_{\min}$, with probability at least $1 - O(n_H T / N^{10})$,

$$d(\hat{\theta}, \theta^*) \le \epsilon\Delta, \tag{5}$$

where

$$d(\hat{\theta}, \theta^*) \triangleq \min_{\pi} \max_{j \in [k]} \left\| \hat{\theta}_{\pi(j)} - \theta_j^* \right\|_2, \quad \textit{where } \pi \textit{ is permutation over } [k]. \tag{6}$$

Note that in (6) we take a minimization over permutation, as the cluster indices are unidentifiable.

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 $\tilde{\Omega}(1/p_{\min})$ anchor clients. 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 D. A key step in our proof is to show the residual estimation errors $\{\Sigma_j(\theta_j^* - \theta_{i,t})\}_{j=1}^k$ approximately lie in $\mathrm{span}(\hat{U}_{i,t})$. 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 [DK70] cannot be be applied. This issue is further exaggerated by the fact that the convergence rate of the orthogonal iteration also crucially depends on the eigengaps [GVL13]. For these reasons, $\mathrm{span}(\hat{U}_{i,t})$ may not be close to $\mathrm{span}\{\Sigma_j(\theta_j^* - \theta_{i,t})\}_{j=1}^k$ at all. To resolve this issue, we develop a novel gap-free bound to show that projection errors $\hat{U}_{i,t}^\top \Sigma_j(\theta_j^* - \theta_{i,t})$ are small for every $j \in [k]$ (cf. Lemma 5).

Remark 1 (Comparison to previous work [LL18, KSS$^+$20]**).** Our algorithm is partly inspired by [LL18] 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 $\{\Sigma_j(\theta_j^* - \theta_{i,t})\}_{j=1}^k$. In contrast, [LL18] 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 [LL18].

Our algorithm also bears similarities with the meta-learning algorithm in [KSS$^+$20], which also uses clients collectively for subspace estimation and anchor clients for estimating cluster centers. However, there are several key differences. First, [KSS$^+$20] focuses on the centralized setting and relies on one-shot estimation, under the additional assumption that the covariance matrix of features across all clusters are identical. Instead, our moment descent algorithm is iterative, is amenable to a distributed implementation in FL systems, and allows for covariance matrices varying across clusters. Second, in the fine-tuning phase, [KSS$^+$20] 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.

4.2 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 $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^{dk}$ 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 [LLV21]. 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 $\theta_{i\cdot,t} = (\theta_{i1,t}, \ldots, \theta_{ik,t})$ for all clusters.

```
Phase 2: FedX+clustering
Input: $\theta = (\theta_1, \ldots, \theta_k)$ from the output of Phase 1, $\eta, T'$.
Output: $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$
PS sets $\theta_T \leftarrow \theta$.
for $t = T+1, \ldots, T+T'$ do
    PS broadcasts $\theta_{t-1}$ to all clients;
    Each client $i$ estimates the likelihood of its local cluster label by
        $$\lambda_{i,t} \leftarrow \arg\min_{\lambda \in \mathbf{\Delta}^{k-1}} L(\theta_{t-1}, \lambda; \mathcal{D}_i); \qquad (7)$$
    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
      $\theta_{i\cdot,t} = (\theta_{i1,t}, \ldots, \theta_{ik,t})$.
        • FedAvg-based: it runs $s$ steps of local gradient descent:
            $$\theta_{i\cdot,t} \leftarrow \mathcal{G}_i^s(\theta_{t-1}), \quad \text{where } \mathcal{G}_i(\theta) = \theta - \eta \nabla L_i(\theta)$$
        • 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 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 [GCYR20] 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

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

In this context, (7) yields a vertex of the probability simplex $\lambda_{ij,t} = \mathbb{1}\{j = z_{i,t}\}$, where

$$z_{i,t} = \arg\min_{j \in [k]} \|y_i - \phi(\boldsymbol{x}_i)\theta_{j,t-1}\|_2. \qquad (8)$$

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 [SXY21]:

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

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

For the learnability of model parameters, we assume that collectively there are sufficient data in each cluster. In particular, we assume $N_j \gtrsim d$, where $N_j = \sum_{i:z_i=j} n_i$ denotes the number of data points in cluster j. To further characterize the *quantity skewness* (i.e., the imbalance of data partition $n = (n_1, \ldots, n_M)$ across clients), we adopt the χ^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. *Suppose that $k \geq 2$, $\eta \lesssim 1/s$, and $N_j \gtrsim d$. Let $\rho = \min_j N_j/N$. If $\nu \log(e/\nu) \lesssim \rho/\kappa$, then with probability $1 - Cke^{-d}$, for all $t \geq T+1$ and all θ_T such that $d(\theta_T, \theta^*) \leq \epsilon\Delta$, where ϵ is some constant, it holds that*

$$d(\theta_t, \theta^*) \leq (1 - C_1 s\eta\rho/\kappa)\, d(\theta_{t-1}, \theta^*) + C_2 s\eta\sigma\nu \log\frac{e}{\nu}, \tag{9}$$

where

$$\nu \triangleq \frac{1}{N}\sum_{i=1}^{M} n_i p_e(n_i) + C\sqrt{\frac{dk\log k}{M}(\chi^2(n)+1)}, \tag{10}$$

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

$$\left\|\hat{\theta}_{i,t} - \theta^*_{z_i}\right\|_2 \lesssim \Delta \cdot e^{-C_1 s\eta\rho(t-T)/\kappa} + \frac{\sigma\kappa}{\rho}\nu\log\frac{e}{\nu},$$

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 θ^*. The parameter ν captures the additional clustering errors injected at each iteration. It consists of two parts: the first term of (10) bounds the clustering error in expectation which diminishes exponentially in the local data size and the signal-to-noise ratio Δ/σ; 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 an model estimation error of θ_j^* scaling as $\sqrt{d/N_j}$. 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 (9). In comparison, the previous work [GCYR20] assumes fresh samples at each iteration by sample-splitting and good initialization independent of everything else provided a prior; 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 $\nu\log(e/\nu) \lesssim \rho/\kappa$. As aforementioned, the clustering error ν consists of two parts shown in (10): 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, ρ captures imbalance of cluster sizes, which is typically of a constant order, and we can choose a small enough step size to ensure κ is close to 1. Thus the key assumption $\nu\log(e/\nu) \lesssim \rho/\kappa$ roughly translates to ν 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.

4.2.1 Analysis of global iterations

Without loss of generality, assume the optimal permutation in (6) is identity. In this case, if $z_{i,t} = j$, then client i will refine $\theta_{j,t-1}$. To prove Theorem 2, we need to analyze the global iteration of θ_t. Following a similar argument to [SXY21] with a careful examination of cluster labels, we obtain the following lemma. The proof is deferred to Appendix E.1.

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

$$\theta_{j,t} = \theta_{j,t-1} - \eta B \Lambda_{j,t}(\phi(\boldsymbol{x})\theta_{j,t-1} - y), \quad j \in [k], \tag{11}$$

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

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

and $\Lambda_{j,t}$ is another block diagonal matrix with ith block being $\lambda_{ij,t} I_{n_i}$.

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

$$\theta_{j,t} - \theta_j^* = (I - \eta K_j)(\theta_{j,t-1} - \theta_j^*) - \eta B\mathcal{E}_{j,t}(\phi(\boldsymbol{x})\theta_{j,t-1} - y) + \eta B\Lambda_j\zeta, \quad \forall j \in [k], \tag{12}$$

where $K_j = B\Lambda_j\phi(\boldsymbol{x})$ and $\mathcal{E}_{j,t} = \Lambda_{j,t} - \Lambda_j$. 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 $I_j = \{i : z_i = j\}$ be the clients belonging to jth cluster, and $I_{j,t} = \{i : z_{i,t} = j\}$ be the clients with estimated label j. The indices of nonzero blocks of $\mathcal{E}_{j,t}$ are $I_j \ominus I_{j,t}$ indicating the clustering errors pertaining to jth cluster.

For the 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 ζ_I; let P_I be the matrix with diagonal blocks P_i for $i \in I$. Using those notations, we have $K_j = \frac{1}{N}\phi(\boldsymbol{x}_{I_j})^\top P_{I_j}\phi(\boldsymbol{x}_{I_j})$, which differs from the usual covariance matrix by an additional matrix P_{I_j}. Therefore, the analysis of the first and third terms on the right-hand side of (12) 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 E.2.

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

$$\|B\mathcal{E}_{j,t}(\phi(\boldsymbol{x})\theta_{j,t-1} - y)\|_2 \lesssim s(d(\theta_{t-1}, \theta^*) + \sigma)\nu \log\frac{e}{\nu}, \qquad \forall\, j \in [k]. \tag{13}$$

Lemma 2 aims to upper bound the error of

$$B\mathcal{E}_{j,t}(\phi(\boldsymbol{x})\theta_{j,t-1} - y) = \frac{1}{N}\phi(\boldsymbol{x}_{S_{j,t}})^\top P_{S_{j,t}}(\phi(\boldsymbol{x}_{S_{j,t}})\theta_{j,t-1} - y_{S_{j,t}}), \tag{14}$$

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

Proof Sketch of Lemma 2. It follows from the definition of $z_{i,t}$ in (8) that

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

Then,

$$\begin{aligned} \|\phi(\boldsymbol{x}_{S_{j,t}})\theta_{j,t-1} - y_{S_{j,t}}\|_2^2 &= \sum_{i \in S_{j,t}} \|\phi(\boldsymbol{x}_i)\theta_{j,t-1} - y_i\|_2^2 \le \sum_{i \in S_{j,t}} \|\phi(\boldsymbol{x}_i)\theta_{z_i,t-1} - y_i\|_2^2 \\ &\le \sum_{i \in S_{j,t}} 2\left(\|\phi(\boldsymbol{x}_i)(\theta_{z_i,t-1} - \theta_{z_i}^*)\|_2^2 + \|\zeta_i\|_2^2)\right) \\ &\le 2\left(d(\theta_{t-1}, \theta^*) \cdot \|\phi(\boldsymbol{x}_{S_{j,t}})\|_2 + \|\zeta_{S_{j,t}}\|_2\right)^2. \end{aligned} \tag{15}$$

Hence, it suffices to upper bound $\|\phi(\boldsymbol{x}_{S_{j,t}})\|_2$ and $\|\zeta_{S_{j,t}}\|_2$ 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 $\sum_{i \in S_{j,t}} n_i$ by analyzing a weighted empirical process. Using the decision rule (8), the set $S_{j,t}$ can be written as a function $S_j(\theta_{t-1})$ with

$$\mathbb{1}\{i \in S_j(\theta)\} = \begin{cases} \max\limits_{\ell \neq j} \mathbb{1}\{P_{j\ell}[\boldsymbol{x}_i, y_i](\theta) \geq 0\} \triangleq f^{\mathrm{I}}_{j,\theta}(\boldsymbol{x}_i, y_i), & i \in I_j, \\ \prod\limits_{\ell \neq j} \mathbb{1}\{P_{\ell j}[\boldsymbol{x}_i, y_i](\theta) \geq 0\} \triangleq f^{\mathrm{II}}_{j,\theta}(\boldsymbol{x}_i, y_i), & i \notin I_j, \end{cases} \tag{16}$$

where

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

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

$$\sup_{\theta \in \mathbb{R}^{dk}} \left| \sum_{i=1}^{M} n_i \mathbb{1}\{i \in S_j(\theta)\} - \sum_{i=1}^{M} n_i \mathbb{P}\{i \in S_j(\theta)\} \right| \leq CN \sqrt{\frac{dk \log k}{M} (\chi^2(n) + 1)}\,.$$

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

$$\mathcal{F}^{\mathrm{I}}_j \triangleq \{f^{\mathrm{I}}_{j,\theta} : \theta \in \mathbb{R}^{dk}\}, \quad \mathcal{F}^{\mathrm{II}}_j \triangleq \{f^{\mathrm{II}}_{j,\theta} : \theta \in \mathbb{R}^{dk}\}. \tag{17}$$

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 θ 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., [Mat13, Theorem 6.2.1]).[1] Consequently, $\|\phi(\boldsymbol{x}_{S_{j,t}})\|_2, \|\zeta_{S_{j,t}}\|_2$ and thus (15) can be uniformly upper bounded using sub-Gaussian concentration and the union bound, concluding the proof of Lemma 2.

□

4.3 Global convergence

Combining Theorem 1 and Theorem 2, we immediately deduce the global convergence from any initialization within the ℓ_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 $T = \Theta(1)$ iterations starting from any initialization θ_0 with $\|\theta_0\|_2 \leq R$, followed by Phase 2 with $T' = \Theta(\frac{\kappa}{s\eta\rho} \log \frac{\Delta}{\nu})$ iterations. Then with probability $1 - N^{-9} - cke^{-d}$, it is true that*

$$d(\hat{\theta}, \theta^*) \lesssim \frac{\sigma\kappa}{\rho} \nu \log \frac{e}{\nu}, \tag{18}$$

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

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 (18) reveals that the final estimation error is dominated by the clustering error captured by ν, and scales linearly in κ 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 Δ/ν and linearly in $\kappa/(s\eta\rho)$. Thus, by choosing s relatively large while keeping κ close to 1, FedAvg enjoys a saving of the total communication cost.

Acknowledgement

J. Xu is supported in part by the NSF Grant CCF-1856424 and an NSF CAREER award CCF-2144593. P. Yang is supported in part by the NSFC Grant 12101353 and Tsinghua University Initiative Scientific Research Program. The authors would like to thank Philippe Rigollet for pointing out the related literature on computing VC dimensions using Milnor-Thom theorem.

[1] Similar applications of the Milnor-Thom theorem have been known in the literature (see e.g. [GJ93, Theorem 2.2] and [VR02, Theorem 2]).

References

[AEP08] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine learning*, 73(3):243–272, 2008.

[AZB05] Rie Kubota Ando, Tong Zhang, and Peter Bartlett. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(11), 2005.

[DDPM13] Simon Dooms, Toon De Pessemier, and Luc Martens. Movietweetings: a movie rating dataset collected from twitter. In *Workshop on Crowdsourcing and human computation for recommender systems, CrowdRec at RecSys*, volume 2013, page 43, 2013.

[DHK$^+$20] Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

[DK70] C. Davis and W.M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[FAC] Frequency of facebook use in the united states as of 3rd quarter 2020. `https://www.statista.com/statistics/199266/frequency-of-use-among-facebook-users-in-the-united-states/`. Accessed: 2022-05-01.

[FHK12] Andrey Feuerverger, Yu He, and Shashi Khatri. Statistical significance of the netflix challenge. *Statistical Science*, 27(2):202–231, 2012.

[FMO20] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

[GCYR20] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.

[GHYR19] Avishek Ghosh, Justin Hong, Dong Yin, and Kannan Ramchandran. Robust federated learning in a heterogeneous environment. *arXiv preprint arXiv:1906.06629*, 2019.

[GJ93] Paul Goldberg and Mark Jerrum. Bounding the vapnik-chervonenkis dimension of concept classes parameterized by real numbers. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 361–369, 1993.

[GVL13] Gene H Golub and Charles F Van Loan. *Matrix computations*. Johns Hopkins University Press, 4 edition, 2013.

[JKRK19] Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

[KKM$^+$20] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

[KMA$^+$21] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth

Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[KSS+20] Weihao Kong, Raghav Somani, Zhao Song, Sham Kakade, and Sewoong Oh. Meta-learning for mixed linear regression. In *International Conference on Machine Learning*, pages 5394–5404. PMLR, 2020.

[LDCH21] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. *arXiv preprint arXiv:2102.02079*, 2021.

[LL18] Yuanzhi Li and Yingyu Liang. Learning mixtures of linear regressions with nearly optimal complexity. In *Conference On Learning Theory*, pages 1125–1144. PMLR, 2018.

[LLV21] Chengxi Li, Gang Li, and Pramod K Varshney. Federated learning with soft clustering. *IEEE Internet of Things Journal*, 2021.

[LSZ+20] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

[LXC+19] Liping Li, Wei Xu, Tianyi Chen, Georgios B Giannakis, and Qing Ling. Rsa: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1544–1551, 2019.

[Mat13] Jiri Matousek. *Lectures on discrete geometry*, volume 212. Springer Science & Business Media, 2013.

[MMR+17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[MMRS20] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

[MN19] Peter McCullagh and John A Nelder. *Generalized linear models*. Routledge, 2019.

[MNB+21] Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, and Richard Vidal. Federated multi-task learning under a mixture of distributions. *Advances in Neural Information Processing Systems*, 34, 2021.

[RV+13] Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[SCST17] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

[SMS20] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

[SXY21] Lili Su, Jiaming Xu, and Pengkun Yang. A non-parametric view of FedAvg and FedProx: Beyond stationary points. *arXiv preprint arXiv:2106.15216*, 2021.

[Sza91] Stanislaw J Szarek. Condition numbers of random matrices. *Journal of Complexity*, 7(2):131–149, 1991.

[TJJ21] Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.

[Tro15] Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[VR02] M Vidyasagar and J Maurice Rojas. An improved bound on the vc-dimension of neural networks with polynomial activation functions. *Math ArXiV preprint math. OC/0112208, submitted for publication*, 2002.

[WLL$^+$20] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

[XLS$^+$21] Ming Xie, Guodong Long, Tao Shen, Tianyi Zhou, Xianzhi Wang, Jing Jiang, and Chengqi Zhang. Multi-center federated learning. *arXiv preprint arXiv:2108.08647*, 2021.

[ZLL$^+$18] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Appendix A
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix B
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Limitations

The limitations of our work are two-fold: 1) Our setup assumes exact cluster structure, where each client belongs to a unique cluster. An interesting generalization is to consider more relaxed cluster structure, where each client may belong to a mixture of multiple clusters. 2) Our study focuses on the mixed regression model. The Phase 1 crucially relies on the regression setup. While the Phase 2 is described under the general risk minimization setup, our analysis has been restricted to mixed regression. An interesting yet challenging future direction is to extend our algorithms and analysis to more general clustered Federated Learning setting with general risk functions.

B Broad Impact

Federated Learning has gained tremendous popularity over the past few years and has been widely deployed in products such as Apple's Siri and Google's Gboard. It opens up a world of new opportunities for training machine learning models without compromising data privacy. Our paper significantly advances the theory and algorithms for Federated Learning, providing important design insights and key enabling technologies for efficient and personalized model training under FL. Our study is highly interdisciplinary, bringing together ideas from statistics, optimization, and distributed computation. We are unaware of any potential negative societal impacts of our work.

C Experimental results

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

We consider the mixed linear regression with $k = 3$ clusters. The true model parameter for each cluster θ_1^*, θ_2^*, θ_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.

(1) The vanilla FedAvg ignores the underlying cluster structure and learns a common model.

(2) In the one-shot clustering [GHYR19], first each client estimates its underlying model based on its local data, then the PS clusters the locally estimated models via k-means, and finally within each estimated cluster, we run FedAvg to obtain the model estimate for each cluster.

(3) The IFCA algorithm [GCYR20] is the same as Phase 2 of our algorithm.

(4) Oracle iterative clustering algorithm is an ideal implementation of IFCA initialized with the true model parameters. 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 settings with a total of $N = 1000$ data points but at increasing levels of data heterogeneity.

C.1 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 $1/3$.

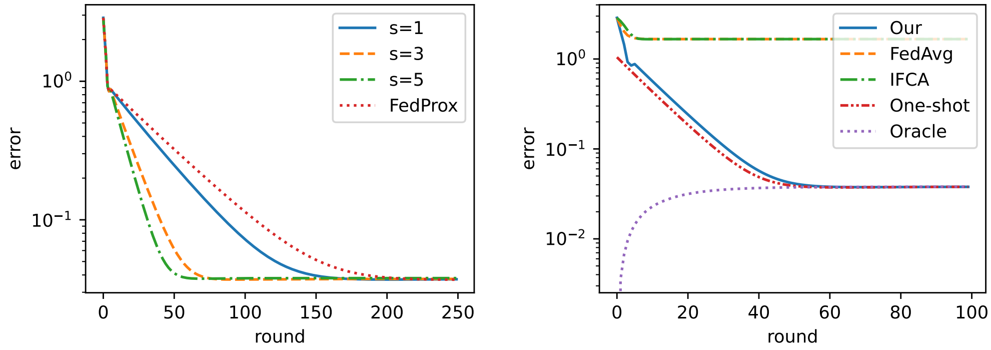

Figure 1: Balanced local data and balanced cluster partition.

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.

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.

C.2 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 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 κ.

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.

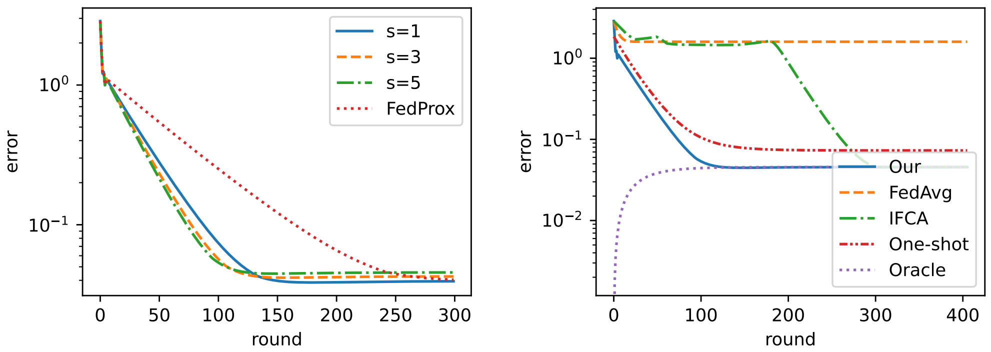

Figure 2: Unbalanced local data and balanced cluster partition.

C.3 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 $p_1 = 0.2$, $p_2 = 0.3$, $p_3 = 0.5$.

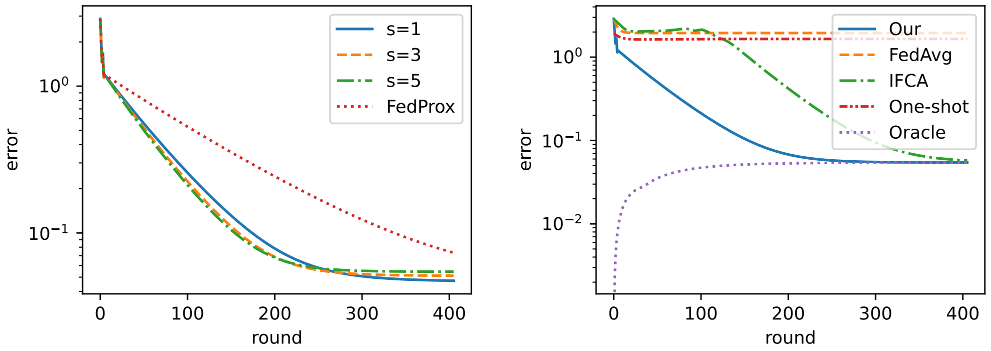

Figure 3: Unbalanced local data and unbalanced cluster partition.

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 ρ (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 Analysis of Phase 1

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

D.1 Subspace estimation via federated orthogonal iteration

Recall that Phase 1 aims to estimate the subspace that the residual estimation errors $\{\Sigma_j(\theta_j^* - \theta_{i,t})\}_{j=1}^k$ lie in via the federated-orthogonal iteration. We can show that $\mathbb{E}\left[Y_{i,t}\right]$ is of rank at most k

and the eigenspace corresponding to the non-zero eigenvalues is spanned by $\{\Sigma_j(\theta_j^* - \theta_{i,t})\}_{j=1}^k$. Specifically, we first prove that $Y_{i,t}$ is close to $\mathbb{E}\left[Y_{i,t}\right]$ in the operator norm and then further deduce that $\{\Sigma_j(\theta_j^* - \theta_{i,t})\}_{j=1}^k$ 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 $i \in S$ 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 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 $\{\Sigma_j(\theta_j^* - \theta_{i,t})\}_{j=1}^k$ approximately lie in the subspace spanned by the k columns of $\hat{U}_{i,t}$.

Proposition 1 (Subspace estimation)**.** *If $T_1 \geq Ck\log(Nd)$ for some sufficiently large constant C, then with probability at least $1 - N^{-10}$,*

$$\left\|\left(\hat{U}_{i,t}\hat{U}_{i,t}^\top - I\right)\Sigma_j\left(\theta_j^* - \theta_{i,t}\right)\right\|_2^2 \leq O\left(\left(\delta_{i,t}^2 + \sigma^2\right)\xi_1/p_j\right), \quad \forall j \in [k],$$

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

We postpone the detailed proof to Appendix D.4. 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 exaggerated 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 $\hat{U}_{i,t}^\top\Sigma_j(\theta_j^* - \theta_{i,t})$ are small for every $j \in [k]$ (cf. Lemma 5).

D.2 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 $\|\Sigma_j(\theta_{z_i}^* - \theta_{i,t})\|_2^2$. Moreover, the residual error decreases geometrically until reaching a plateau.

Proposition 2. *Fix an anchor client i and suppose $T_2 \gtrsim \log(Nd)$. There exists a constant $C > 0$ and an event $\mathcal{E}_{i,t}$ with $\mathbb{P}\left\{\mathcal{E}_{i,t}\right\} \geq 1 - O(N^{-10})$ such that on event $\mathcal{E}_{i,t}$*

$$\left|\left\|\Sigma_j(\theta_{z_i}^* - \theta_{i,t})\right\|_2^2 - \hat{\sigma}_{i,t}^2\right| \leq C(\delta_{i,t}^2 + \sigma^2)(\xi_1/p_{z_i} + \xi_2), \tag{19}$$

where $\xi_2 = \sqrt{\frac{k}{\ell}\log N} + \frac{k}{\ell}\log^3 N$. Furthermore, if

$$\left\|\Sigma_{z_i}\left(\theta_{z_i}^* - \theta_{i,t}\right)\right\|_2^2 \geq C(\xi_1/p_{z_i} + \xi_2), \tag{20}$$

then

$$\left\|\Sigma_{z_i}\left(\theta_{z_i}^* - \theta_{i,t+1}\right)\right\|_2^2 \leq \left(1 - \frac{\alpha^2}{8\beta^2}\right)\left\|\Sigma_{z_i}\left(\theta_{z_i}^* - \theta_{i,t}\right)\right\|_2^2, \tag{21}$$

We postpone the proof to Appendix D.5. The key ingredient in the proof of geometric decay (21) is to show that the descent direction $r_{i,t}$ is approximately parallel to the residual error $\Sigma_{z_i}(\theta_{z_i}^* - \theta_{i,t})$ under condition (20)

D.3 Proof of Theorem 1

Now, we are ready to prove our main theorem on the performance guarantee of Phase 1. Let $\mathcal{E}_{i,t}$ denote the event under which the statement of Proposition 2 holds. Let $\mathcal{E} = \cup_{i \in H} \cup_{t=1}^{T} \mathcal{E}_{i,t}$. By the union bound, $\mathbb{P}\left\{\mathcal{E}\right\} \geq 1 - O(n_H T/N^{10})$. In the following, we assume event $\mathcal{E}$ holds.

We first prove (4). Fix any anchor client i and omit the subscript i for simplicity. We further assume it belongs to cluster j, i.e., $z_i = j$. Define

$$t^* = \min\{\inf\{t \geq 0 : \hat{\sigma}_t \leq \epsilon\Delta\}, T\}.$$

By definition,

$$\hat{\sigma}_t > \epsilon\Delta, \quad \forall\, 0 \leq t \leq t^* - 1. \tag{22}$$

Moreover, by the update rule of our algorithm, $\theta_T = \theta_{t^*}$. Thus, it suffices to bound $\|\theta_{t^*} - \theta_j^*\|_2$.

We claim that for all $0 \leq t \leq t^*$,

$$\left\|\Sigma_j\left(\theta_j^* - \theta_t\right)\right\|_2^2 \leq \left(1 - \frac{\alpha^2}{8\beta^2}\right)^t \left\|\Sigma_j\left(\theta_j^* - \theta_0\right)\right\|_2^2 \tag{23}$$

$$\|\theta_t\|_2 \leq (1 + 2\beta/\alpha)R. \tag{24}$$

If $t^* = T$, then by (23) we immediately get that

$$\left\|\Sigma_j\left(\theta_j^* - \theta_{t^*}\right)\right\|_2 \leq \left(1 - \frac{\alpha^2}{8\beta^2}\right)^{T/2} \left\|\Sigma_j \theta_j^*\right\|_2 \leq \exp\left(-T\alpha^2/(16\beta^2)\right)\beta R \leq \epsilon\Delta,$$

where the last inequality holds by choosing $T = \frac{16\beta^2}{\alpha^2}\log\frac{\beta R}{\epsilon\Delta}$.

If $t^* < T$, then by (24) and $\|\theta_j^*\|_2 \leq R$ for all $j \in [k]$, it follows that $\delta_t \leq 2(1 + \beta/\alpha)R$. Therefore, by (19)

$$\left|\left\|\Sigma_j(\theta_j^* - \theta_t)\right\|_2^2 - \hat{\sigma}_t^2\right| \leq C(\delta_t^2 + \sigma^2)(\xi_1/p_j + \xi_2) \leq \epsilon^2\Delta^2/2, \tag{25}$$

where the last inequality holds by choosing $m = \tilde{\Omega}(d/p_{\min}^2)$ and $\ell = \tilde{\Omega}(k)$ and invoking the standing assumption that $R = O(\Delta)$ and $\sigma = O(\Delta)$. It immediately follows that

$$\left\|\Sigma_j(\theta_j^* - \theta_{t^*})\right\|_2^2 \leq \hat{\sigma}_{t^*}^2 + \epsilon^2\Delta^2/2 \leq \frac{3}{2}\epsilon^2\Delta^2,$$

where the last inequality holds by the stopping rule of our algorithm so that $\hat{\sigma}_{t^*} \leq \epsilon\Delta$.

In both cases, we get that

$$\left\|\theta_j^* - \theta_{t^*}\right\|_2 \leq \frac{1}{\alpha} \left\|\Sigma_j(\theta_j^* - \theta_{t^*})\right\|_2 \leq \frac{2\epsilon\Delta}{\alpha} = \epsilon'\Delta$$

for $\epsilon' = 2\epsilon/\alpha$. This proves (4).

Now, it remains to prove the claim (23)–(24) by induction. The base case $t = 0$ trivially holds as $\left\|\theta_0\right\|_2 \leq R$. Now, suppose the induction hypothesis holds for an arbitrary t where $0 \leq t \leq t^* - 1$, we prove it also holds for $t + 1$. In view of (25),

$$\left\|\Sigma_j(\theta_j^* - \theta_t)\right\|_2^2 \geq \hat{\sigma}_t^2 - \epsilon^2\Delta^2/2 > \epsilon^2\Delta^2/2 \geq C(\delta_t^2 + \sigma^2)(\xi_1/p_j + \xi_2),$$

where the second inequality holds due to (22). Therefore, the condition (20) is satisfied. Hence, by applying Proposition 2, we get that

$$\left\|\Sigma_j\left(\theta_j^* - \theta_{t+1}\right)\right\|_2^2 \leq \left(1 - \frac{\alpha^2}{8\beta^2}\right) \left\|\Sigma_j\left(\theta_j^* - \theta_t\right)\right\|_2^2 \leq \left(1 - \frac{\alpha^2}{8\beta^2}\right)^{t+1} \left\|\Sigma_j\left(\theta_j^* - \theta_0\right)\right\|_2^2.$$

where the second inequality holds by the induction hypothesis (23). Hence

$$\alpha^2 \left\|\theta_j^* - \theta_{t+1}\right\|_2^2 \leq \left\|\Sigma_j\left(\theta_j^* - \theta_{t+1}\right)\right\|_2^2 \leq \left\|\Sigma_j\left(\theta_j^* - \theta_0\right)\right\|_2^2 \leq 4\beta^2 R^2.$$

It follows that

$$\|\theta_j^* - \theta_{t+1}\|_2 \leq \frac{2\beta}{\alpha} R$$

and hence $\|\theta_{t+1}\|_2 \leq (1 + 2\beta/\alpha)R$. This completes the induction proof.

Finally, we prove (5). Note that by standard coupon collector's problem, we deduce that if $n_H \geq \log(k/\delta)/p_{\min}$, then with probability at least $1 - \delta$, $H \cap \{i : z_i = j\} \neq \emptyset$ for all $j \in [k]$. To see this, note that

$$\begin{aligned}\mathbb{P}\left\{H \cap \{i : z_i = j\} \neq \emptyset, \forall j \in [k]\right\} &\geq 1 - \sum_{j \in [k]} \mathbb{P}\left\{H \cap \{i : z_i = j\} = \emptyset\right\} \\ &\geq 1 - k\left(1 - p_{\min}\right)^{n_H} \\ &\geq 1 - k\exp\left(-p_{\min} n_H\right) \geq 1 - \delta.\end{aligned}$$

Therefore, as long as $\epsilon' < 1/4$, we have for two anchor clients $i, i' \in H$

$$\|\theta_{i,T} - \theta_{i',T}\|_2 \leq \left\|\theta_{i,T} - \theta_{z_i}^*\right\|_2 + \left\|\theta_{i',T} - \theta_{z_{i'}}^*\right\|_2 \leq 2\epsilon'\Delta, \quad \text{if } z_i = z_{i'}$$

$$\|\theta_{i,T} - \theta_{i',T}\|_2 \geq \Delta - \left\|\theta_{i,T} - \theta_{z_i}^*\right\|_2 - \left\|\theta_{i',T} - \theta_{z_{i'}}^*\right\|_2 \geq (1 - 2\epsilon')\Delta, \quad \text{if } z_i \neq z_{i'}.$$

Thus, by assigning anchor clients $i, i' \in H$ in the same cluster when $\left\|\theta_{i,T} - \theta_{i',T}\right\|_2 \leq \Delta/2$ we can exactly 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 $\hat{\theta}_j$ denote the center of the recovered cluster j, that is

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

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

D.4 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 from θ_t. Hence, we condition on θ_t and $\mathcal{S}_t$ in the following analysis. Note that

$$\mathbb{E}\left[Y_t\right] = \frac{1}{m} \sum_{i' \mathcal{S}_t} \mathbb{E}_{z_{i'}}\left[\Sigma_{z_{i'}}\left(\theta_{z_{i'}}^* - \theta_t\right)\left(\theta_{z_{i'}}^* - \theta_t\right)\Sigma_{z_{i'}}\right] = \sum_{j=1}^{k} p_j \Sigma_j\left(\theta_j^* - \theta_t\right)\left(\theta_j^* - \theta_t\right)^\top \Sigma_j,$$

where p_j is the probability that a client belongs to the j-th cluster.

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

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

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

where $\delta_t=\max_{j\in[k]}\|\theta_j^*-\theta_t\|_2$ *and* $\xi_1=\sqrt{\frac{d}{m}\log N}+\frac{d}{m}\log^3 N$.

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

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

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

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

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

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

and further by Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\left\|\langle\phi(x_{i1}),\theta_{z_i}^*-\theta_t\rangle\phi(x_{i1})\right\|_2^2\right]\leq\sqrt{\mathbb{E}\left[\langle\phi(x_{i1}),\theta_{z_i}^*-\theta_t\rangle^4\right]}\sqrt{\mathbb{E}\left[\left\|\phi(x_{i1})\right\|_2^4\right]}=O\left(\delta_t^2 d\right).$$

Combining the last three displayed equation gives that $\mathbb{E}\left[\left\|a_i\right\|_2^2\right]\leq O\left(d\right)$. The same upper bound also holds for $\mathbb{E}\left[\left\|b_i\right\|_2^2\right]$.

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

$$\mathbb{E}\left[\langle a_i,u\rangle^2\right]=\frac{1}{\delta_t^2+\sigma^2}\mathbb{E}\left[r_i^2\langle\phi(x_{i1}),u\rangle^2\right]\leq\frac{1}{\delta_t^2+\sigma^2}\sqrt{\mathbb{E}\left[r_i^4\right]}\sqrt{\langle\phi(x_{i1}),u\rangle^4}\leq O\left(1\right),$$

where $r_i=y_{i1}-\langle\phi(x_{i1}),\theta_t\rangle$. Combining the last two displayed equations gives that $\left\|\mathbb{E}\left[a_ia_i^\top\right]\right\|_2=O\left(1\right)$. The same upper bound also holds for $\left\|\mathbb{E}\left[b_ib_i^\top\right]\right\|_2$. Finally, by the sub-Gaussian property of $\phi(x_{i1})$, we have

$$\mathbb{P}\left\{\left\|\phi(x_{i1})\right\|_2\geq s_1\right\}\leq\exp\left(O(d)-\Omega(s_1^2)\right)$$

and

$$\mathbb{P}\left\{\frac{|r_i|}{\sqrt{\delta_t^2+\sigma^2}}\geq s_2\right\}\leq\exp\left(-\Omega\left(s_2^2\right)\right).$$

Choosing $s_1=C\sqrt{s}d^{1/4}$ and $s_2=\sqrt{s}/(Cd^{1/4})$ for a sufficiently large constant C, we get that for all $s\geq\sqrt{d}$,

$$\begin{aligned}\mathbb{P}\left\{\left\|a_i\right\|_2\geq s\right\}&\leq\mathbb{P}\left\{\left\|\phi(x_{i1})\right\|_2\geq s_1\right\}+\mathbb{P}\left\{\frac{|r_i|}{\sqrt{\delta_t^2+\sigma^2}}\geq s_2\right\}\\&\leq\exp\left(O(d)-\Omega(Cs\sqrt{d})\right)+\exp\left(-\Omega\left(\frac{s}{\sqrt{d}}\right)\right)\\&\leq\exp\left(-\Omega\left(\frac{s}{\sqrt{d}}\right)\right).\end{aligned}$$

The same bound holds for $\mathbb{P}\left\{\left\|b_i\right\|_2\geq s\right\}$. 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 $|\lambda_1| \geq |\lambda_2| \geq \cdots |\lambda_d|$ and the corresponding eigenvectors u_i's. Define $U_1 = [u_1, \dots, u_k]$ and $U_2 = [u_{k+1}, \dots, u_d]$.

Lemma 4. *[GVL13, 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*

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

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*

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

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

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

Proof. Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d \geq 0$ denote the singular values of M. Then by assumption on M and Weyl's inequality, $\sigma_{k+1} \leq \epsilon$. We divide the analysis into two cases depending on the value of σ_1. Let $\delta > 0$ be some parameter to be tuned later.

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

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

Thus,

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

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

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

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

$$\left\|U_\ell U_\ell^\top x_i - x_i\right\|_2^2 = u^\top x_i x_i^\top u \leq u^\top \left(\sum_{i=1}^k x_i x_i^\top\right) u = u^\top \left(\sum_{i=1}^k x_i x_i^\top - M\right) u + u^\top M u.$$

Note that

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

Moreover,

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

Combining the last three displayed equations gives that

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

Let $\hat{Q}_t$ be the submatrix of Q_t formed by the first ℓ columns. Since $\sigma_\ell > \sigma_{\ell+1}$, 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 ℓ. Thus, applying Lemma 4 with k replaced by ℓ gives that

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

where $\cos(\gamma) = \sigma_{\min}(U_\ell^\top \hat{Q}_0)$ and $\hat{Q}_0$ is the submatrix of Q_0 formed by its first ℓ columns. Applying Lemma 14, we get that $\tanh(\gamma) = O(N^{10}d)$ with probability at least $1 - O(N^{-10})$. Therefore, when $t \ge (C/\delta)\log\frac{NdH}{\epsilon}$, we have

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

Therefore, by triangle's inequality,

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

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

Applying Lemma 3 and Lemma 5 and invoking the assumption that $T_1 \gtrsim k\log(Nd)$, we have with probability at least $1 - O(1/N)$,

$$\left\|\left(\hat{U}_t\hat{U}_t^\top - I\right)\sqrt{p_j}\Sigma_j\left(\theta_j^* - \theta_t\right)\right\|_2^2 \le O\left(\left(\delta_t^2 + \sigma^2\right)\xi_1\right),$$

or equivalently,

$$\left\|\hat{U}_t^\top\Sigma_j\left(\theta_j^* - \theta_t\right)\right\|_2^2 \ge \left\|\Sigma_j\left(\theta_j^* - \theta_t\right)\right\|_2^2 - O\left(\left(\delta_t^2 + \sigma^2\right)\xi_1/p_j\right). \tag{26}$$

D.5 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 j, i.e., $z_i = j$. Note that crucially, the global data points on clients $\mathcal{S}_t$ are independent from the local data points on $\mathcal{D}_t$. Thus, in the following analysis, we further condition on $\hat{U}_t$. Then

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

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

$$\left\|A_t - \mathbb{E}\left[A_t\right]\right\|_2 \le O\left(\left(\left\|\theta_j^* - \theta_t\right\|_2^2 + \sigma^2\right)\xi_2\right),$$

where $\xi_2 = \sqrt{\frac{k}{\ell}\log N} + \frac{k}{\ell}\log^3 N$.

Proof. Note that

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

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

Applying Lemma 6 and Lemma 5, when $T_2 \gtrsim \log(Nd)$, we have with probability at least $1 - O(N^{-10})$

$$\left|\hat{\beta}_t^\top \hat{U}_t^\top \Sigma_j\left(\theta_j^* - \theta_t\right)\right|^2 \geq \left\|\hat{U}_t^\top \Sigma_j\left(\theta_j^* - \theta_t\right)\right\|_2^2 - O\left(\left(\left\|\theta_j^* - \theta_t\right\|_2^2 + \sigma^2\right)\xi_2\right).$$

Applying Proposition 1, we have with probability at least $1 - O(N^{-10})$

$$\left\|\hat{U}_t^\top \Sigma_j\left(\theta_j^* - \theta_t\right)\right\|_2^2 \geq \left\|\Sigma_j\left(\theta_j^* - \theta_t\right)\right\|_2^2 - O\left(\left(\delta_t^2 + \sigma^2\right)\xi_1/p_j\right).$$

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

Combining the last two displayed equations yields that

$$\left\|\Sigma_j\left(\theta_j^* - \theta_t\right)\right\|_2^2 - O\left(\left(\delta_t^2 + \sigma^2\right)\left(\xi_1/p_j + \xi_2\right)\right) \leq \left|\hat{\beta}_t^\top \hat{U}_t^\top \Sigma_j\left(\theta_j^* - \theta_t\right)\right|^2 \leq \left\|\Sigma_j\left(\theta_j^* - \theta_t\right)\right\|_2^2.$$

Moreover, since

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

it follows that

$$\left|\hat{\sigma}_t^2 - \left|\hat{\beta}_t^\top \hat{U}_t^\top \Sigma_j\left(\theta_j^* - \theta_t\right)\right|^2\right| \leq O\left(\left(\delta_t^2 + \sigma^2\right)\xi_2\right).$$

Combining the last two displayed equations yields that

$$\left|\hat{\sigma}_t^2 - \left\|\Sigma_j\left(\theta_j^* - \theta_t\right)\right\|_2^2\right| \leq O\left(\left(\delta_t^2 + \sigma^2\right)\left(\xi_1/p_j + \xi_2\right)\right).$$

This proves (19).

Under condition (20), we have

$$\left|\hat{\beta}_t^\top \hat{U}_t^\top \Sigma_j\left(\theta_j^* - \theta_t\right)\right|^2 \geq \left(1 - \frac{\alpha^2}{64\beta^2}\right)\left\|\Sigma_j\left(\theta_j^* - \theta_t\right)\right\|_2^2 \tag{27}$$

and

$$\left(1 - \frac{\alpha^2}{32\beta^2}\right)\left\|\Sigma_j\left(\theta_j^* - \theta_t\right)\right\|_2^2 \leq \hat{\sigma}_t^2 \leq \left(1 + \frac{\alpha^2}{32\beta^2}\right)\left\|\Sigma_j\left(\theta_j^* - \theta_t\right)\right\|_2^2 \tag{28}$$

Now we show that θ_t converges to θ_j^*. Note that

$$\left(\theta_j^* - \theta_{t+1}\right)^\top \Sigma_j^2\left(\theta_j^* - \theta_{t+1}\right) = \left(\theta_j^* - \theta_t\right)^\top \Sigma_j^2\left(\theta_j^* - \theta_t\right) - 2\eta_t\left(\theta_j^* - \theta_t\right)^\top \Sigma_j^2 r_t + \eta_t^2 r_t^\top \Sigma_j^2 r_t.$$

In view of (27), and recalling $r_t = \hat{U}_t \hat{\beta}_t$, we have $\left\|r_t\right\|_2 = 1$ and under condition (20)

$$\left\langle r_t, \Sigma_j\left(\theta_j^* - \theta_t\right)\right\rangle^2 \geq \left(1 - \frac{\alpha^2}{64\beta^2}\right)\left\|\Sigma_j\left(\theta_j^* - \theta_t\right)\right\|_2^2.$$

We decompose

$$\Sigma_j\left(\theta_j^* - \theta_t\right) = a_t r_t + b_t r_t^\perp,$$

for some unit vector $r_t^\perp$ that is perpendicular to r_t. Since $a_t^2 + b_t^2 = \|\Sigma_j(\theta_j^* - \theta_t)\|_2^2$, we have $|b_t| \leq \frac{\alpha}{8\beta}\|\Sigma_j(\theta_j^* - \theta_t)\|_2$. Hence,

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

where $\lambda_{\min}(\Sigma_j) \geq \alpha$ and $\beta \geq \max_{j\in[k]} \|\Sigma_j\|_2$. It follows that

$$\left(\theta_j^* - \theta_{t+1}\right)^\top \Sigma_j^2 \left(\theta_j^* - \theta_{t+1}\right) \leq \left(\theta_j^* - \theta_t\right)^\top \Sigma_j^2 \left(\theta_j^* - \theta_t\right) - \eta_t \alpha \left\|\Sigma_j \left(\theta_j^* - \theta_t\right)\right\|_2 + \eta_t^2 \left\|\Sigma_j\right\|_2^2 .$$

Recall the choice of step size $\eta_t = \alpha\hat{\sigma}_t/(2\beta^2)$. In view of (28), we get that

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

Therefore,

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

This proves (21).

E Analysis of Phase 2

Throughout the proof in this section, we assume without loss of generality that the optimal permutation in (6) is identity.

E.1 Derivation of global iteration

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

$$\nabla_j L_i(\theta) = \frac{\lambda_{ij,t}}{n_i} \phi(\boldsymbol{x}_i)^\top (\phi(\boldsymbol{x}_i)\theta_j - y_i),$$

where $\lambda_{ij,t} = \mathbb{1}\{j = z_{i,t}\}$ and ∇_j denotes the gradient with respect to θ_j. Then the one-step local gradient descent at client i is

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

where $\eta_i = \eta/n_i$. Iterating s steps yields that [SXY21]

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

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

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

We obtain the global iteration:

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

which is (11) using matrix notations.

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

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

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

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

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

E.2 Convergence analysis of Phase 2

We analyze the three terms on the right-hand side of (12) separately. The first term of (12) is the main term due to the decreasing of estimation error, and the last term is the stochastic variation due to the observation noise ζ. We have the following lemmas on the eigenvalues of K_j and the concentration of the observation noise.

Lemma 7. *There exists constants c and C such that, with probability* $1 - 2ke^{-d}$,

$$c\alpha \frac{sN_j}{\kappa N} \le \lambda_{\min}(K_j) \le \lambda_{\max}(K_j) \le C\beta \frac{sN_j}{N}, \quad \forall j \in [k].$$

Proof. Since $\phi(\boldsymbol{x}_{I_j})$ of size $N_j \times d$ consists of independent and sub-Gaussian rows, by a covering argument [Ver18, Theorem 4.6.1], with probability $1 - 2e^{-d}$,

$$\alpha N_j - C(\sqrt{dN_j} \vee d) \le \sigma^2_{\min}(\phi(\boldsymbol{x}_{I_j})) \le \sigma^2_{\max}(\phi(\boldsymbol{x}_{I_j})) \le \beta N_j + C(\sqrt{dN_j} \vee d),$$

where $\sigma_{\max}$ and $\sigma_{\min}$ denote the largest and smallest singular values, respectively, and C is an absolute constant. By definition, $K_j = \frac{1}{N}\phi(\boldsymbol{x}_{I_j})^\top P_{I_j} \phi(\boldsymbol{x}_{I_j})$, where P_{I_j} is a symmetric matrix. It is shown in [SXY21, Lemma 3] that

$$s/\kappa \le \lambda_{\min}(P_{I_j}) \le \lambda_{\max}(P_{I_j}) \le s.$$

The conclusion follows from the condition $N_j \gtrsim d$ and a union bound over $j \in [k]$. □

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

$$\|B\Lambda_j\zeta\|_2^2 \le C\frac{\sigma^2 sd}{N}\|K_j\|_2, \quad \forall\, j \in [k].$$

Proof. Note that

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

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

$$\mathbb{E}\left[\|B\Lambda_j\zeta\|_2^2\right] = \mathbb{E}\left[\langle \Lambda_j B^\top B \Lambda_j, \zeta\zeta^\top\rangle\right] \le \sigma^2 \mathsf{Tr}\left(\Lambda_j B^\top B \Lambda_j\right) = \sigma^2 \mathsf{Tr}\left(B\Lambda_j^2 B^\top\right).$$

Recall that

$$B\Lambda_j^2 B^\top = \frac{1}{N^2}\phi(\boldsymbol{x}_{I_j})^\top P_{I_j}^2 \phi(\boldsymbol{x}_{I_j}) \overset{(a)}{\preceq} \frac{s}{N^2}\phi(\boldsymbol{x}_{I_j})^\top P_{I_j}\phi(\boldsymbol{x}_{I_j}) = \frac{s}{N}K_j, \tag{29}$$

where (a) holds because $\|P_{I_j}\|_2 \le s$. Therefore,

$$\mathbb{E}\left[\|B\Lambda_j\zeta\|_2^2\right] = \mathbb{E}\left[\langle \Lambda_j B^\top B \Lambda_j, \zeta\zeta^\top\rangle\right] \le \frac{\sigma^2 sd}{N}\left\|K_j\right\|_2.$$

Next, using Hanson-Wright's inequality [RV$^+$13], we get

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

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

$$\|\Lambda_j B^\top B \Lambda_j\|_2 = \|B\Lambda_j^2 B^\top\|_2 \le \frac{s}{N}\|K_j\|_2,$$

$$\|\Lambda_j B^\top B \Lambda_j\|_{\mathrm{F}} = \|B\Lambda_j^2 B^\top\|_{\mathrm{F}} \le s\|K_j\|_{\mathrm{F}} \le \frac{s\sqrt{d}}{N}\left\|K_j\right\|_2.$$

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

$$\langle \Lambda_j B^\top B \Lambda_j, \zeta\zeta^\top \rangle \le \mathbb{E}\left[\langle \Lambda_j B^\top B \Lambda_j, \zeta\zeta^\top \rangle\right] + \delta \le (C+1)\,\sigma^2 \frac{sd}{N}\|K_j\|_2.$$

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

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

Proof of Theorem 2. We prove the result conditioning on the high probability events in Lemmas 2, 7, and 8 that happen with probability at least $1 - Cke^{-d}$. We obtain from Lemma 7 that

$$\|I - \eta K_j\|_2 \le 1 - C\eta s\rho/\kappa.$$

Combining Lemmas 7 and 8 yields

$$\|B\Lambda_j\zeta\|_2 \lesssim s\sigma\sqrt{\frac{d}{N}}.$$

Plugging the above upper bounds and Lemma 2 into (12), we get

$$\|\theta_{j,t} - \theta_j^*\|_2 \le \left(1 - C\eta s\left(\frac{\rho}{\kappa} - \nu\log\frac{e}{\nu}\right)\right) d(\theta_{t-1}, \theta^*) + C\eta s\sigma\left(\sqrt{\frac{d}{N}} + \nu\log\frac{e}{\nu}\right), \quad \forall j \in [k].$$

Since $\nu\log(e/\nu) \lesssim \rho/\kappa$ and $\nu \gtrsim \sqrt{d/N}$, we conclude (9).

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 (9) that, for $t \ge T + 1$,

$$\|\hat{\theta}_{i,t} - \theta_{z_i}^*\|_2 \le d(\theta_t, \theta^*) \le (1 - C_1 s\eta\rho/\kappa)^{t-T} d(\theta_T, \theta^*) + \frac{C_2}{C_1}\frac{\sigma\kappa}{\rho}\nu\log\frac{e}{\nu}.$$

The proof is completed. □

E.2.1 Proof of Lemma 2

This subsection is devoted to the proof of Lemma 2 using the following road map:

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

Specifically, a small estimation error $d(\theta_t, \theta^*)$ implies an upper bound on the total number of incorrectly clustered data points $\sum_{i \in S_{j,t}} n_i$; then we upper bound $\|\phi(\boldsymbol{x}_{S_j^t})\|_2$ and $\|\zeta_{S_j^t}\|_2$ using sub-Gaussian concentration and the union bound; finally we conclude the result from (15).

We first upper bound $\sum_{i \in S_{j,t}} n_i$. Using (8), the set $S_{j,t} = I_j \ominus I_{j,t}$ is equivalently the union of

$$I_j - I_{j,t} = \left\{ i \in I_j : \|y_i - \phi(\boldsymbol{x}_i)\theta_{j,t-1}\|_2 \ge \min_{\ell \ne j} \|y_i - \phi(\boldsymbol{x}_i)\theta_{\ell,t-1}\|_2 \right\},$$

$$I_{j,t} - I_j = \left\{ i \notin I_j : \|y_i - \phi(\boldsymbol{x}_i)\theta_{j,t-1}\|_2 \le \min_{\ell \ne j} \|y_i - \phi(\boldsymbol{x}_i)\theta_{\ell,t-1}\|_2 \right\}.$$

Therefore, $S_{j,t} = S_j(\theta_{t-1})$, where S_j is defined in (16). The next lemma upper bounds the VC dimensions of the binary function classes specified in (17).

Lemma 9. *For $k \geq 2$, the VC dimensions of $\mathcal{F}_j^{\mathrm{I}}$ and $\mathcal{F}_j^{\mathrm{II}}$ are at most $O(dk \log k)$.*

Proof. We focus on the proof for $\mathcal{F}_j^{\mathrm{I}}$ for a fixed $j \in [k]$, and the proof for $\mathcal{F}_j^{\mathrm{II}}$ 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 $\mathcal{F}_j^{\mathrm{I}}$ shatters m points denoted by $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m)$. Define binary functions

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

It is necessary that the number of sign patterns of $g_1(\theta), \dots, g_m(\theta)$ is 2^m. Note that every $P_{\ell,j}[\boldsymbol{x}_i, y_i]$ is a (dk)-variate quadratic function. By the Milnor-Thom theorem (see, e.g., [Mat13, Theorem 6.2.1]), if $m(k-1) \geq dk \geq 2$, the number of sign patterns of $m(k-1)$ binary functions $q_{1,\ell}, \dots, q_{m,\ell}$ for $\ell \neq j$ is at most $(\frac{100m(k-1)}{dk})^{dk}$. Since each g_i is the maximum of $q_{i,\ell}$ over $\ell \neq j$, the number of sign patterns of $g_1, \dots, g_m$ is upper bounded by $(\frac{100m(k-1)}{dk})^{dk}$. Consequently, we obtain $2^m \leq (\frac{100m(k-1)}{dk})^{dk}$, and hence $m \lesssim dk \log k$. If instead, $m(k-1) < dk$, then the conclusion $m \lesssim dk \log k$ 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 $G_j(\theta) = \sum_{i=1}^{M} n_i \mathbb{1}\{i \in S_j(\theta)\}$. Since the local data $(\boldsymbol{x}_i, y_i)$ on different clients are independent, for a fixed θ, the events $\{i \in S_j(\theta)\}$ as functions of $(\boldsymbol{x}_i, y_i)$ are mutually independent. Using the binary function classes in (17), we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\theta} |G_j(\theta) - \mathbb{E}[G_j(\theta)]|\right] \\
&\leq \mathbb{E}\left[\sup_{f \in \mathcal{F}_j^{\mathrm{I}}} \left|\sum_{i \in I_j} n_i (f(\boldsymbol{x}_i, y_i) - \mathbb{E}[f(\boldsymbol{x}_i, y_i)])\right|\right] + \mathbb{E}\left[\sup_{f \in \mathcal{F}_j^{\mathrm{II}}} \left|\sum_{i \notin I_j} n_i (f(\boldsymbol{x}_i, y_i) - \mathbb{E}[f(\boldsymbol{x}_i, y_i)])\right|\right] \\
&\lesssim \sqrt{dk \log k \sum_{i \in \mathcal{I}_j} n_i^2} + \sqrt{dk \log k \sum_{i \notin \mathcal{I}_j} n_i^2} \\
&\leq \sqrt{2dk \log k \sum_{i=1}^{M} n_i^2},
\end{aligned} \tag{30}$$

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

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

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

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

By McDiarmid's inequality, we have

$$\mathbb{P}\left\{h(Z_1, \dots, Z_M) - \mathbb{E}\left[h(Z_1, \dots, Z_M)\right] \geq t\right\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{M} n_i^2}\right). \tag{31}$$

Therefore, combining (30) and (31), and by a union bound, with probablity at least $1 - k^{-dk}$,

$$\sup_{\theta} |G_j(\theta) - \mathbb{E}[G_j(\theta)]| \lesssim \sqrt{dk \log k \sum_{i=1}^{M} n_i^2} = N \sqrt{\frac{dk \log k}{M} (\chi^2(n) + 1)}, \quad \forall j \in [k]. \tag{32}$$

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

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

where c is an absolute constant.

Proof. For $i \in I_j$, it follows from (16) and the union bound that

$$\begin{aligned}\mathbb{P}\{i \in S_j(\theta)\} &\le \sum_{\ell\ne j}\mathbb{P}\{\|y_i - \phi(\boldsymbol{x}_i)\theta_j\|_2 \ge \|y_i - \phi(\boldsymbol{x}_i)\theta_\ell\|_2\}\\ &= \sum_{\ell\ne j}\mathbb{P}\left\{\|\phi(\boldsymbol{x}_i)(\theta_j^* - \theta_j) + \zeta_i\|_2 \ge \|\phi(\boldsymbol{x}_i)(\theta_j^* - \theta_\ell) + \zeta_i\|_2\right\}. \end{aligned} \tag{33}$$

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

$$\mathbb{P}\left\{\left|\frac{1}{n_i}\|\phi(\boldsymbol{x}_i)u + \zeta_i\|_2^2 - \left(\mathbb{E}[\zeta_{i1}^2] + \|u\|_{\Sigma_i}^2\right)\right| \ge (\|u\|_2^2 + \sigma^2)t\right\} \le 2\exp\left(-cn_i(t \wedge t^2)\right), \tag{34}$$

where $\Sigma_i = \mathbb{E}[\phi(x_{i1})\phi(x_{i1})^\top]$. Let $u_1 \triangleq \theta_j^* - \theta_j$ and $u_2 \triangleq \theta_j^* - \theta_\ell$. By assumptions that $\|\theta_\ell - \theta_\ell^*\|_2 \le \epsilon$ for all $\ell \in [k]$ and $\|\theta_j^* - \theta_\ell^*\|_2 \ge \Delta$ for $\ell \ne j$, 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

$$\|u_2\|_{\Sigma_i}^2 - \|u_1\|_{\Sigma_i}^2 \ge \alpha(\Delta - \epsilon)^2 - \beta\epsilon^2 \ge \alpha\Delta^2/3. \tag{35}$$

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 (34) that

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

where $t = \frac{\|u_2\|_{\Sigma_i}^2 - \|u_1\|_{\Sigma_i}^2}{\|u_1\|_2^2+\|u_2\|_2^2+2\sigma^2} \gtrsim \alpha(1 \wedge \frac{\Delta^2}{\sigma^2})$ using the lower bound of seperation in (35). We conclude the proof for $i \in I_j$ from (33). Similarly, for $i \in I_\ell$ with $\ell \ne j$, we have

$$\mathbb{P}\{i \in S_j(\theta)\} \le \mathbb{P}\{\|\phi(\boldsymbol{x}_i)(\theta_\ell^* - \theta_\ell) + \zeta_i\|_2 \ge \|\phi(\boldsymbol{x}_i)(\theta_\ell^* - \theta_j) + \zeta_i\|_2\}.$$

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 (32) and Lemma 10 that, with probability $1 - k^{-dk}$,

$$N_{S_{j,t}} = \sum_{i\in S_{j,t}} n_i \le \nu N, \tag{36}$$

where ν is defined in (10). 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 $1 - 4e^{-d}$,*

$$\sup_{N_I\le\nu N}\frac{1}{N}\|\phi(\boldsymbol{x}_I)\|_2^2 \lesssim \nu\log\frac{e}{\nu}, \qquad \sup_{N_I\le\nu N}\frac{1}{N}\|\zeta_I\|_2^2 \lesssim \sigma^2\nu\log\frac{e}{\nu}. \tag{37}$$

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 $1 - 2e^{-d}$,

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

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

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

$$\|\phi(\boldsymbol{x}_{S_{j,t}})\|_2 \lesssim \sqrt{N\nu \log \frac{e}{\nu}}, \qquad \|\zeta_{S_{j,t}}\|_2 \lesssim \sigma\sqrt{N\nu \log \frac{e}{\nu}}.$$

Since $\|P_{S_{j,t}}\|_2 \le s$, we conclude from (14) and (15) that

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

E.2.2 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 ϵ_i are i.i.d. Rademacher random variables. Next, by conditioning on X_i's, we aim to apply Dudley's integral. Since ϵ_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 μ_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., [Ver18, Theorem 8.1.3]) conditioning on X_i's, we get that

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

where $\mathcal{N}(\mathcal{F}, L^2(\mu_n), \epsilon)$ denotes the ϵ-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., [Ver18, Theorem 8.3.18])

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

The conclusion follows. □

F 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* $\mathbb{E}\left[\|a_i\|_2^2\right] = O(d)$, $\mathbb{E}\left[\|b_i\|_2^2\right] = O(d)$, $\left\|\mathbb{E}\left[a_i a_i^\top\right]\right\|_2 = O(1)$, $\left\|\mathbb{E}\left[b_i b_i^\top\right]\right\|_2 = O(1)$*, and*

$$\mathbb{P}\left\{\|a_i\|_2 \geq t\right\}, \mathbb{P}\left\{\|b_i\|_2 \geq t\right\} \leq \exp\left(-\Omega\left(t/\sqrt{d}\right)\right), \quad \forall t \geq \sqrt{d}.$$

Let

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

Then there exists a univeral constant $C > 0$ *such that with probability at least* $1 - 3\delta$,

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

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

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

and hence

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

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

To bound the first term, we will use 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 $\mathbb{E}\left[Y_i\right] = 0$ and

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

Moreover,

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

Therefore,

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

Moreover, $Y_i Y_i^\top \succeq 0$. Hence, $\left\|\sum_{i=1}^N \mathbb{E}\left[Y_i Y_i^\top\right]\right\|_2 = O(Nd)$. Similarly, we can show that $\left\|\sum_{i=1}^N \mathbb{E}\left[Y_i^\top Y_i\right]\right\|_2 = O(Nd)$. Applying matrix Bernstein inequality [Tro15], we get that with probability at least $1 - \delta$,

$$\left\|\sum_{i=1}^N Y_i\right\|_2 \lesssim \sqrt{Nd\log\frac{1}{\delta}} + \tau\log\frac{1}{\delta}. \tag{39}$$

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

$$\mathbb{P}\left\{\mathcal{E}_i^c\right\} = \mathbb{P}\left\{\left\|a_i b_i^\top\right\|_2 > \tau\right\} \le \mathbb{P}\left\{\|a_i\|_2 \ge \sqrt{\tau}\right\} + \mathbb{P}\left\{\|b_i\|_2 \ge \sqrt{\tau}\right\} \le 2e^{-\Omega(\sqrt{\tau/d})}.$$

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

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

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

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

Moreover,

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

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

$$\mathbb{P}\left\{\left\|a_i b_i^\top\right\|_2 \ge t\right\} \le \mathbb{P}\left\{\|a_i\|_2 \ge \sqrt{t}\right\} + \mathbb{P}\left\{\|b_i\|_2 \ge \sqrt{t}\right\} \le 2e^{-C'\sqrt{t/d}}$$

for some universal constant $C' > 0$. It follows that

$$\int_\tau^\infty \mathbb{P}\left\{\left\|a_i b_i^\top\right\|_2 \ge t\right\} \mathrm{d}t \le 2\int_\tau^\infty e^{-C'\sqrt{t/d}} \mathrm{d}t = 4d\left(\sqrt{\tau/d} + 1/C'\right) e^{-C'\sqrt{\tau/d}},$$

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,

$$\mathbb{E}\left[\left\|a_i b_i^\top \mathbb{1}\{\mathcal{E}_i^c\}\right\|_2\right] \le \tau \frac{\delta}{N} + 4d\left(\sqrt{\tau/d} + 1/C'\right) e^{-C'\sqrt{\tau/d}} = O\left(\frac{d}{N} \log^2(N/\delta)\right). \tag{41}$$

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

G 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* $1 - O(\epsilon)$,

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

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

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

In view of [Ver10, Corollary 5.35], $\sigma_{\max}(A) \lesssim \sqrt{d} + \log(1/\epsilon)$ with probability at least $1 - \epsilon$. Moreover, in view of [Sza91, Theorem 1.2], $\sigma_{\min}(X) \ge \epsilon/\sqrt{\ell}$ with probability at least $1 - O(\epsilon)$. The desired conclusion readily follows. □