Federated Representation Learning in the Under-Parameterized Regime

Renpu Liu [1] **Cong Shen** [2] **Jing Yang** [1]

Abstract

Federated representation learning (FRL) is a popular personalized federated learning (FL) framework where clients work together to train a common representation while retaining their personalized heads. Existing studies, however, largely focus on the over-parameterized regime. In this paper, we make the initial efforts to investigate FRL in the under-parameterized regime, where the FL model is insufficient to express the variations in all ground-truth models. We propose a novel FRL algorithm `FLUTE`, and theoretically characterize its sample complexity and convergence rate for linear models in the under-parameterized regime. To the best of our knowledge, this is the first FRL algorithm with provable performance guarantees in this regime. `FLUTE` features a data-independent random initialization and a carefully designed objective function that aids the distillation of subspace spanned by the global optimal representation from the misaligned local representations. On the technical side, we bridge low-rank matrix approximation techniques with the FL analysis, which may be of broad interest. We also extend `FLUTE` beyond linear representations. Experimental results demonstrate that `FLUTE` outperforms state-of-the-art FRL solutions in both synthetic and real-world tasks.

1. Introduction

In the development of machine learning (ML), the role of representation learning has become increasingly essential. It transforms raw data into meaningful features, reveals hidden patterns and insights in data, and facilitates efficient learning of various ML tasks such as meta-learning (Tripuraneni et al., 2021), multi-task learning (Wang et al., 2016a), and few-shot learning (Du et al., 2020).

Recently, representation learning has been introduced to the federated learning (FL) framework to cope with the heterogeneous local datasets at participating clients (Liang et al., 2020). In the FL setting, it often assumes that all clients share a common representation, which works in conjunction with personalized local heads to realize personalized prediction while harnessing the collective training power (Arivazhagan et al., 2019; Collins et al., 2021; Zhong et al., 2022; Shen et al., 2023).

Existing theoretical analysis of representation learning usually assumes the adopted model is over-parameterized to almost fit the ground-truth model (Tripuraneni et al., 2021; Wang et al., 2016a). While this may be valid for expressive models like Deep Neural Networks (He et al., 2016; Liu et al., 2017) or Large Language Models (OpenAI, 2023; Touvron et al., 2023), it may be too restrictive for FL on resource-constrained devices, as adopting over-parameterized models in such a framework faces several significant challenges, as elaborated below.

- **Computation limitation.** In FL, edge devices like smartphones and Internet of Things (IoT) devices often have limited memory and lack computational power, which are not capable of either storing or training over-parameterized models (Wang et al., 2019; He et al., 2020; Kairouz et al., 2021)[1].
- **Communication overhead.** In FL, the clients need to communicate updated model information with the server frequently. It thus becomes prohibitive to transmit a huge number of model updates for devices operating with limited communication energy and bandwidth.
- **Privacy concern.** Existing works show that excessively expressive models may "memorize" relevant information from local datasets, increasing the model's susceptibility to reconstruction attacks (Hitaj et al., 2017; Melis et al., 2019; Wang et al., 2018; Li et al., 2020) or membership inference (Tan et al., 2022).

[1]Department of Electrical Engineering, The Pennsylvania State University, University Park, PA, USA [2]Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, Virginia, USA. Correspondence to: Jing Yang <yangjing@psu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]For example, two of the widely adopted neural network models suitable for IoT or embedded devices, MobileNetV3 (Howard et al., 2019) and EfficientNet-B0 (Tan & Le, 2019), only have a few million parameters and, as an example, typically process at most a few GFLOPS in a Raspberry Pi 4 (Ju et al., 2023).

Motivated by those concerns, in this work, we focus on federated representation learning (FRL) in the *under-parameterized* regime, where the parameterized model class is not rich enough to realize the ground-truth models across all clients. This is arguably a more realistic setting for edge devices supporting FL. Meanwhile, due to the inherent limitation of the expressiveness of the under-parameterized models, the algorithm design and theoretical guarantees in the over-parameterized regime do not naturally translate to this setting. We summarize our main contributions as follows.

- **Algorithm design.** A major challenge for FRL in the under-parameterized regime is the fact that the locally optimal representation may not be globally optimal. As a result, simply averaging the local representations may not converge to the global optimal solution. To cope with this challenge, we propose `FLUTE`, a novel FRL framework tailored for the under-parameterized setting. To the best of our knowledge, this is the first FRL framework that focuses on the under-parameterized regime. Our algorithm design features two primary innovations. First, we develop a new regularization term that generalizes the existing formulations in a non-trivial way. In particular, this new regularization term is designed to provably enhance the performance of FRL in the under-parameterized setting. Second, our algorithm contains a new and critical step of server-side updating by simultaneously optimizing both the representation layer and all local head layers. This represents a significant departure from existing approaches in FRL, particularly in over-parameterized settings where local heads are optimized solely on the client side. By leveraging information across these local heads, our approach could learn the ground-truth model more effectively.
- **Theoretical guarantees.** In terms of theoretical performance, we specialize `FLUTE` to the linear setting and analyze the sample complexity required for `FLUTE` to recover a near-optimal model, as well as characterizing its convergence rate. `FLUTE` achieves a sample complexity that scales in $\tilde{\mathcal{O}}\left(\frac{\max\{d,M\}}{M\epsilon^2}\right)$ for recovering an ϵ-optimal model, where d is the dimension of the input data and M is the number of clients. This result indicates a linear sample complexity speedup in terms of M in the high dimensional setting (i.e., $d \geq M$) compared with its single-agent counterpart (Hsu et al., 2012). Besides, it outperforms the sample complexity in the noiseless over-parameterized FRL setting (Collins et al., 2021) in terms of both M and d. Moreover, we show that `FLUTE` converges to the optimal model exponentially fast when the number of samples is sufficiently large.
- **Technical contributions.** In the under-parameterized regime, we must analyze the convergence of both the representation and personalized heads toward their optimal estimations. This is in sharp contrast to the over-parameterized regime, where we only need to study the convergence of the representation column space to the ground truth (Collins et al., 2021; Zhong et al., 2022). Towards this end, we adopt a low-rank matrix approximation framework (Chen et al., 2023) of the ground-truth model. However, in contrast to conventional low-rank matrix approximation, in FRL, the global model is not accessible a priori but must be learned from distributed local datasets. Thus, the technical analysis needs to bound the unavoidable gradient discrepancy in the under-parameterized regime, as well as ensure that neither gradient discrepancy nor noise-induced errors accumulate over iterations. To address these technical challenges, we first provide new concentration results to ensure that the norm of the gradient discrepancy can be bounded when local datasets are sufficiently large. We then develop iteration-dependent upper bounds for sample complexity, which guarantee that the improvement in the estimation, i.e., the 'distance' between our estimated model and the optimal low-rank model, can mitigate potential disturbances caused by gradient discrepancy and noise.
- **Empirical evaluation.**[2] We conduct a series of experiments utilizing both synthetic datasets for linear `FLUTE` and real-world datasets, specifically CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), for general `FLUTE`. The empirical results demonstrate the advantages of `FLUTE`, as evidenced by its superior performance over baselines, particularly in the scenarios where the level of under-parameterization is significant.

2. Related Work

Representation learning. Representation learning focuses on acquiring a representation across diverse tasks to effectively extract feature information (LeCun et al., 2015; Tripuraneni et al., 2021; Wang et al., 2016a; Finn et al., 2017). In the linear multi-task learning setting, Du et al. (2020) characterize the optimal solution of the empirical risk minimization (ERM) problem, demonstrating that the gap between the solution and the ground-truth representation is upper bounded by $\mathcal{O}\big(\sqrt{\frac{M+d}{MN}}\big)$, where d is the dimension of data, M is the number of clients and N is the number of samples per task. Tripuraneni et al. (2021) give an upper bound $\mathcal{O}\big(\sqrt{\frac{d}{MN}}\big)$ using the Method-of-Moment estimator. Thekumparampil et al. (2021) also show the $\mathcal{O}\big(\sqrt{\frac{d}{MN}}\big)$ upper bound in their work. Duchi et al. (2022) consider data-dependent noise and show that the sample complexity required to recover the shared subspace of the linear mod-

[2]Main experiments can be reproduced with the code provided under the following link: https://github.com/RenpuLiu/flute

els scales in $O\big(\log^3(Nd)\sqrt{\frac{d}{MN}}\big)$. These works, however, only focus on the over-parameterized regime in a centralized setting.

Federated representation learning. Recently, representation learning has been introduced to FL

(Arivazhagan et al., 2019; Liang et al., 2020; Collins et al., 2021; Yu et al., 2020). Liang et al. (2020) propose an FRL framework named Fed-LG, where the distinct representations are stored locally and the common prediction head is forwarded to the server for aggregation. In contrast, Arivazhagan et al. (2019) propose FedPer, where a common representation is shared among clients, with personalized local heads kept at the client side. A similar setting is adopted by FedRep (Collins et al., 2021), where exponential convergence to the optimal representation in the linear setting is proved. These works focus on the over-parameterized regime, while the under-parameterized regime has largely been overlooked.

Low-rank matrix factorization. Under-parameterized representation learning problem considered in this work is closely related to low-rank matrix factorization, where the objective is to find two low-rank matrices whose product is closest to a given matrix $\mathbf{\Phi}$. Pitaval et al. (2015) prove the global convergence of gradient search with infinitesimal step size for this problem. Ge et al. (2017) demonstrate that no spurious minima exists in such a problem and all saddle points are strict. Based on a revised robust strict saddle property, Zhu et al. (2021) show that the local search method such as gradient descent leads to a linear convergence rate with good initialization with a regularity condition on $\mathbf{\Phi}$. Chen et al. (2023) extend the analysis in Zhu et al. (2021) to general $\mathbf{\Phi}$, and show that with a moderate random initialization, the gradient descent method will converge globally at a linear rate. In the over-parameterized regime, Ye & Du (2021) proves that the gradient descent method will converge to a global minimum at a polynomial rate with random initialization. We note, however, that these works assume the perfect knowledge of $\mathbf{\Phi}$, which is different from the data-based *representation learning* problem considered in this work.

3. Problem Formulation

Notations. We use $\text{diag}(x_1, \cdots, x_d)$ to denote a d-dimension diagonal matrix with diagonal entries $x_1, \cdots, x_d$. $\langle x, y\rangle$ denotes the inner product of x and y, and $\|x\|$ denotes the Euclidean norm of vector x. We use $f \circ \psi$ to denote the composition of functions $f : \mathbb{R}^k \to \mathbb{R}^m$ and $\psi : \mathbb{R}^d \to \mathbb{R}^k$, i.e., $(f \circ \psi)(x) = f(\psi(x))$. $\mathbf{I}_d$ represents a $d \times d$ identify matrix, and $\mathbf{0}$ is a d-dimensional all-zero vector.

FL with common representation. We consider an FL system consisting of M clients and one server. Client i has a local dataset $\mathcal{D}_i$ that consists of n_i training samples (x, y) where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^m$. For simplicity, we assume $n_i = N$ for all client $i \in [M]$. For $(x_{i,j}, y_{i,j}) \in \mathcal{D}_i$, we assume $y_{i,j} = g_i(x_{i,j}) + \xi_{i,j}$, where $x_{i,j}$ is randomly drawn according to a sub-Gaussian distribution P_X with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_d$, $g_i : \mathbb{R}^d \to \mathbb{R}^m$ is a deterministic function, and $\xi_{i,j} \in \mathbb{R}^m$ is an independent and identically distributed (IID) centered sub-Gaussian noise vector with covariance matrix $\sigma^2\mathbf{I}_d$.

Federated representation learning (FRL) aims at learning both a common representation that suits all clients and an individual head that only fits client i. An FL framework adopting this principle was proposed by Arivazhagan et al. (2019), and we follow the same framework in this paper. More specifically, we assume that the local model of client i can be decomposed into two parts: a common representation $\psi_{\mathbf{B}} : \mathbb{R}^d \to \mathbb{R}^k$ shared by all clients and a local head $f_{w_i} : \mathbb{R}^k \to \mathbb{R}^m$, where $\mathbf{B}$ and w_i are the parameters of the corresponding functions. Then, the ERM problem considered in this FRL framework can be formulated as:

$$\min_{\mathbf{B}, \{w_i\}} \frac{1}{M} \sum_{i \in [M]} \frac{1}{N} \sum_{(x,y) \in \mathcal{D}_i} \ell((f_{w_i} \circ \psi_{\mathbf{B}})(x), y). \quad (1)$$

This formulation leverages the common representation while accommodating data heterogeneity among clients, facilitating efficient personalized model training (Arivazhagan et al., 2019; Collins et al., 2021).

In this work, we focus on the under-parameterized setting in FRL, which is formally defined as follows.

Definition 3.1 (Under-Parameterization in FRL)**.** Given a common representation class Ψ and a collection of local head classes $\{\mathcal{F}_i\}_{i=1}^M$, an FRL problem is under-parameterized if there does not exist a representation $\psi \in \Psi$, and a collection of functions $f_1 \times f_2 \ldots \times f_M \in \mathcal{F}_1 \times \mathcal{F}_2 \ldots \times \mathcal{F}_M$ such that $f_i \circ \psi = g_i$ for all $i \in [M]$.

The over-parameterization in FRL can be defined in a symmetric form. This definition aligns with the over-parameterized frameworks in matrix approximation, as detailed in Jiang et al. (2022); Ye & Du (2021), where over-parameterization is characterized by the rank of the representation being no-less than that of the ground-truth model. It also encompasses the definition in central statistical learning (Belkin et al., 2019; Oneto et al., 2023), where over-parameterization is defined as the predictor's function class being sufficiently rich to approximate the global minimum.

While various algorithms have been developed and analyzed in the over-parameterized setting (Arivazhagan et al., 2019; Liang et al., 2020; Collins et al., 2021), to the best of our knowledge, under-parameterized FRL has not been studied in the literature before. This is, however, arguably a more practical setting in large-scale FRL supported by a massive

number of resources-scarce IoT devices, as such IoT devices usually cannot support the storage, computation, and communication of models parameterized by a large number of parameters, while the task heterogeneity across massive devices imposes significant challenges on the model class to reconstruct M different local models perfectly[3].

Low-dimensional linear representation. We first focus on the *linear* setting in which all local models g_i are linear, i.e., $y_{i,j} = \phi_i^\top x_{i,j} + \xi_{i,j}$ for $(x_{i,j}, y_{i,j}) \in \mathcal{D}_i$. Denote $\mathbf{\Phi} := [\phi_1, \cdots, \phi_M] \in \mathbb{R}^{d\times M}$ and assume its rank is r. Then, similar to the works of Collins et al. (2021); Arivazhagan et al. (2019), we consider a linear prediction model where $(f_{w_i} \circ \psi_{\mathbf{B}})(x)$ can be expressed as $x^\top \mathbf{B} w_i$. Here, $\mathbf{B} \in \mathbb{R}^{d\times k}$ is the common linear representation shared across clients, and $w_i \in \mathbb{R}^k$ is the local head maintained by client i. We denote $\mathbf{W} = [w_1, \cdots, w_M]$. Then, if we further consider the ℓ_2 loss function, the ERM problem becomes

$$\min_{\mathbf{B},\mathbf{W}} \frac{1}{M} \sum_{i\in[M]} \frac{1}{N} \sum_{(x,y)\in\mathcal{D}_i} \|x^\top \mathbf{B} w_i - y\|^2. \tag{2}$$

We note that the existing literature usually assumes that $r \le k$, which falls in the over-parameterized regime (Zhu et al., 2021). The over-parameterized assumption implies the existence of a pair of $\mathbf{B}$ and $\mathbf{W}$ that can accurately recover the ground-truth model $\mathbf{\Phi}$, i.e., $\mathbf{BW} = \mathbf{\Phi}$. Thus, the learning goal in the over-parameterized regime is to identify such a pair using available training data (Du et al., 2020; Tripuraneni et al., 2021; Collins et al., 2021; Shen et al., 2023).

In contrast to the existing works, in the *under-parameterized* regime given in Definition 3.1, we have $r > k$, i.e., there does not exist matrices $\mathbf{B} \in \mathbb{R}^{d\times k}$ and $\mathbf{W} \in \mathbb{R}^{k\times M}$ such that $\mathbf{BW} = \mathbf{\Phi}$. Our objective is to learn a common representation and local heads $(\mathbf{B}, \mathbf{W})$ in the federated learning framework such that $\|\mathbf{BW} - \mathbf{\Phi}\|_F^2$ reaches its minimum, although $\mathbf{\Phi}$ is not explicitly given but embedded in local datasets.

4. The FLUTE Algorithm

In this section, we present the FLUTE algorithm for the linear model. We will first highlight the unique challenges the under-parameterized setting brings, and then introduce our algorithm design.

[3]Continuing the previous example of MobileNet, which can be adapted for object detection for autonomous driving (Chen et al., 2021), it is known that a single model may not capture very detailed or complex features of the complete environment, including pedestrians, cyclists, and various road signs (Chen et al., 2022).

4.1. Challenges

In order to understand the fundamental differences between the over- and under-parameterized regimes, we first assume $\mathbf{\Phi}$ is known beforehand, and consider solving the following optimization problem:

$$(\mathbf{B}^*, \mathbf{W}^*) = \mathop{\arg\min}_{\mathbf{B}\in\mathbb{R}^{d\times k}, \mathbf{W}\in\mathbb{R}^{k\times M}} \|\mathbf{BW} - \mathbf{\Phi}\|_F^2. \tag{3}$$

Denote the singular value decomposition (SVD) of $\mathbf{\Phi}$ as $\mathbf{U\Lambda V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are two unitary matrices, and $\mathbf{\Lambda}$ is a diagonal matrix. When $k \ge r$, i.e., the model is over-parameterized, $\mathbf{B}^*$ and $\mathbf{W}^*$ can be explicitly constructed from the SVD of $\mathbf{\Phi}$, i.e., any $(\mathbf{B}, \mathbf{W})$ satisfying $\mathbf{BW} = \mathbf{U\Lambda V}^\top$ is an optimizer to Equation (3). When $k < r$, i.e., in the under-parameterized regime, we can no longer recover the full matrix $\mathbf{\Phi}$ with $\mathbf{B}^*$ and $\mathbf{W}^*$. Instead, existing result (Golub & Van Loan, 2013) states that we can only determine that the solution must satisfy $\mathbf{B}^*\mathbf{W}^* = \mathbf{U}_k\mathbf{\Lambda}_k\mathbf{V}_k^\top$, where $\mathbf{\Lambda}_k$ is a $k \times k$ diagonal matrix with the k largest singular values of $\mathbf{\Phi}$ as the diagonal entries.

Compared with the over-parameterized setting, learning $\mathbf{B}^*$ and $\mathbf{W}^*$ from decentralized datasets is more challenging in the under-parameterized setting. Let $\mathbf{B}_i^\diamond$ be the locally optimized representation at client i, i.e., $(\mathbf{B}_i^\diamond, w_i^\diamond) = \arg\min \|\mathbf{B}_i w_i - \phi_i\|^2$. Then, in the over-parameterized setting, $\mathbf{B}_i^\diamond$ will always stay in the same column space as $\mathbf{B}^*$, i.e., $\text{span}(\mathbf{B}_i^\diamond) \subseteq \text{span}(\mathbf{B}^*)$, $\forall i \in [M]$. However, for the under-parameterized setting, it is possible that $\text{span}(\mathbf{B}_i^\diamond) \not\subset \text{span}(\mathbf{B}^*)$, $\exists i \in [M]$. How to aggregate the locally obtained $\mathbf{B}_i^\diamond$ to correctly span the column space of $\mathbf{B}^*$ thus becomes a unique challenge in the under-parameterized setting and requires novel techniques different from those in the existing over-parameterized literature.

Example 1. Consider a scenario that $\mathbf{\Phi} \in \mathbb{R}^{d\times M}$ with $M < d$. We assume $\mathbf{\Phi} = \mathbf{U}\,\text{diag}(\lambda_1, \cdots, \lambda_M)$, where $\mathbf{U} := [u_1, \ldots, u_M]$ is a unitary matrix and $\lambda_1 > \lambda_2 > \cdots > \lambda_M > 0$. Assume $k = 1$. Then, we have $\mathbf{B}^*\mathbf{W}^* = u_1\lambda_1$. Assume each client i can perfectly recover its local model $\phi_i = u_i\lambda_i$ with $\mathbf{B}_i = u_i\lambda_i/w_i$. Then, depending on the value of w_i's, the aggregated representation $\mathbf{B} := \frac{1}{M}\sum_i \mathbf{B}_i$ may exhibit different properties. For example, if $w_i = \lambda_i$, we have $\mathbf{B} = \frac{1}{M}\sum_i u_i$, which deviates significantly from the column space of $\mathbf{B}^*$. On the other hand, if $w_i = \sqrt{\lambda_i/M}$, then $\mathbf{B}_i = u_i\sqrt{M\lambda_i}$, while $\mathbf{B} = \sum_i u_i\sqrt{\lambda_i/M}$. Thus, u_1 will have a heavier weight in the aggregated representation, which will eventually help recover the column space of $\mathbf{B}^*$. Intuitively, to accurately recover the column space of $\mathbf{B}^*$, in the under-parameterized setting, it requires a more sophisticated algorithm design not just to estimate the column space of $\mathbf{\Phi}$, but also *distill the most significant components* of it from distributed datasets in each aggregation.

4.2. A New Loss Function

Motivated by the observation in Example 1, instead of considering the original problem in (2), we introduce two new regularization terms and consider the following ERM problem:

$$\min_{\mathbf{B},\{w_i\}_{i=1}^M} \frac{1}{M}\sum_{i\in[M]}\frac{1}{N}\sum_{(x,y)\in\mathcal{D}_i}\|x^\top \mathbf{B}w_i - y\|^2 \qquad (4)$$
$$\underbrace{-\gamma_1\|\mathbf{BW}\|_F^2}_{\text{(I)}} + \underbrace{\gamma_2(\|\mathbf{B}^\top\mathbf{B}\|_F^2 + \|\mathbf{W}\mathbf{W}^\top\|_F^2)}_{\text{(II)}}.$$

In Equation (4), we introduce the regularization term (I) into the loss function, with the purpose of preserving the top-k significant components of $\mathbf{BW}$. By preserving the significant components in $\mathbf{B}$, the term (I) mitigates local over-fitting induced during local updates. However, minimizing term (I) alone would result in a uniform enlargement of all k singular values of $\mathbf{BW}$. To address this, we further incorporate the regularization term (II). This term is specifically formulated to promote the k most significant components and suppress the less significant ones. By doing so, it aids the server in accurately distilling the correct subspace spanned by the optimal representation. We note that when $\gamma_1 = 2\lambda_2$, (I) and (II) together recover the conventional penalty term $\|\mathbf{B}^\top\mathbf{B} - \mathbf{W}\mathbf{W}^\top\|_F^2$, which has been previously adopted for low-rank matrix approximation (Chen et al., 2023; Zhu et al., 2021; Wang et al., 2016b) and multi-task learning (Tripuraneni et al., 2021).

4.3. FLUTE for Linear Model

In order to solve the optimization problem given in (4), we introduce an algorithm named FLUTE (**F**erated **L**earning in **U**nder-parame**T**erized R**E**gime), which is compactly described in Algorithm 1. Specifically, for each epoch, the algorithm consists of three major steps, namely, *server broadcast*, *client update*, and *server update*.

Server broadcast. At the beginning of epoch t, the server broadcasts the representation $\mathbf{B}^{t-1}$ to all clients, and w_i^{t-1} (i.e., the i-th column of $\mathbf{W}^{t-1}$) to each individual client i.

Client update. Denoting the local loss function as $L_i = \frac{1}{N}\sum_{(x,y)\in\mathcal{D}_i}\|x^\top\mathbf{B}w_i - y\|^2$, the client calculates the gradient of L_i with respect to w_i^{t-1} and $\mathbf{B}_i^{t-1}$, respectively, and uploads them to the server.

Server update. After receiving $\nabla_{\mathbf{w}_i^{t-1}}L_i$ and $\nabla_{\mathbf{B}^{t-1}}L_i$ from all clients, the server first aggregates them to update the global representation and local heads as follows:

$$\bar{\mathbf{B}}^t = \mathbf{B}^{t-1} - \eta_l\sum_{i\in[M]}\nabla_{\mathbf{B}^{t-1}}L_i\Big(w_i^{t-1},\mathbf{B}^{t-1}\Big),$$
$$w_i^t = w_i^{t-1} - \eta_l\nabla_{w_i^{t-1}}L_i\Big(w_i^{t-1},\mathbf{B}^{t-1}\Big), \forall i\in[M], \qquad (5)$$

Algorithm 1 FLUTE Linear

Input: Learning rates η_l and η_r, regularization parameter λ, communication round T, constant α
Initialization: All entries of $\mathbf{B}^0$ and $\mathbf{W}^0$ are independently sampled form $\mathcal{N}(0,\alpha^2)$.
for $t\in[T]$ **do**
  Server sends $\mathbf{B}^{t-1}$ and w_i^{t-1} to client i, $\forall i\in[M]$.
  for client $i\in[M]$ in parallel **do**
    Calculates $\nabla_{w_i^{t-1}}L_i\Big(w_i^{t-1},\mathbf{B}^{t-1}\Big)$ and $\nabla_{\mathbf{B}^{t-1}}L_i\Big(w_i^{t-1},\mathbf{B}^{t-1}\Big)$.
    Sends gradients to the server.
  end for
  Server updates according to Equations (5) to (6).
end for

after which it constructs matrix $\bar{\mathbf{W}}^t$ by setting $\bar{\mathbf{W}}^t := [w_1^t,\cdots,w_M^t]$. It then performs another step of gradient descent with respect to the regularization term in (4) to refine the global representation and local heads and obtain $\mathbf{B}^t$ and $\mathbf{W}^t$:

$$\mathbf{B}^t = \bar{\mathbf{B}}^t + \gamma_1\eta_r\nabla_{\mathbf{B}^{t-1}}\|\mathbf{B}^{t-1}\mathbf{W}^{t-1}\|_F^2 \qquad (6)$$
$$-\gamma_2\eta_r\nabla_{\mathbf{B}^{t-1}}(\|(\mathbf{B}^{t-1})^\top\mathbf{B}^{t-1}\|_F^2 + \|\mathbf{W}^{t-1}(\mathbf{W}^{t-1})^\top\|_F^2),$$
$$\mathbf{W}^t = \bar{\mathbf{W}}^t + \gamma_1\eta_r\nabla_{\mathbf{W}^{t-1}}\|\mathbf{B}^{t-1}\mathbf{W}^{t-1}\|_F^2$$
$$-\gamma_2\eta_r\nabla_{\mathbf{W}^{t-1}}(\|(\mathbf{B}^{t-1})^\top\mathbf{B}^{t-1}\|_F^2 + \|\mathbf{W}^{t-1}(\mathbf{W}^{t-1})^\top\|_F^2).$$

The procedure repeats until some stop criterion is satisfied.

Remark 4.1. When α is small, the initialization of $\mathbf{B}^0$ and $\mathbf{W}^0$ would ensure that the largest singular value of $\mathbf{B}^0(\mathbf{W}^0)^T$ is sufficiently small with high probability. As we will show in the next section, such initialization guarantees that FLUTE converges to the global minimum.

The major differences between FLUTE and existing FRL algorithms such as FedRep (Collins et al., 2021), FedRod (Chen & Chao, 2021), and FedCP (Zhang et al., 2023a) lie in the server-side model updating. While these existing algorithms typically involve transmitting only the shared representation layers of local models to the server, with local heads being optimized and utilized exclusively at the client side, FLUTE requires clients to transmit both the shared representation layers and the local heads to the server. The increased communication cost is fundamentally necessary due to the unique nature of FRL in the under-parameterized regime, as it allows for server-side optimization, not just aggregation, of the entire model. Furthermore, FLUTE introduces additional data-free penalty terms to the server-side updates. These terms are designed to guide the shared representation to converge toward the global minimum by leveraging the information in the local heads. This approach represents a significant paradigm shift in federated learning, aiming to enhance the overall global performance

of the FRL model.

5. Theoretical Guarantees

Before introducing our main theorem, we denote $\underline{d} = \min\{d, M\}$ and $\overline{d} = \max\{d, M\}$. We also denote $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{\underline{d}}$ as the ordered singular values of $\mathbf{\Phi}$ with $\Delta := 2(\lambda_k - \lambda_{k+1})$. Denote $E = \sum_i \lambda_i^2$. We assume $\Delta > 0$ throughout the analysis.

5.1. Main Results

Theorem 5.1 (Sample complexity)**.** *Set $\gamma_1 = \frac{1}{4}$ and $\gamma_2 = \frac{1}{8}$ in Equation (4). Let $0 < \alpha \lesssim \frac{1}{10d}$, and $\eta := \eta_l = \eta_r \lesssim \frac{\Delta^2}{228{\lambda_1}^3}$. Then, for any $\epsilon > 0$, under Algorithm 1, there exists positive constants c and c' such that when the number of samples per client satisfies*

$$N \geq c\frac{\lambda_1^4 k(\overline{d} + \log\frac{1}{\delta} + \log\log\frac{1}{\epsilon})(\sqrt{k(\lambda_1)^2 + E} + \sqrt{k}\sigma)^2}{M\eta^2\Delta^6\epsilon^2}$$

and $t \geq \frac{\log(\epsilon\sqrt{M}\eta\Delta^2/c'\lambda_1^2\sqrt{k})}{\log(1-\eta\Delta/16)}$, with probability at least $1-\delta$,

$$\frac{1}{M}\sum_{i\in[M]} \|\mathbf{B}^t w_i^t - \mathbf{B}^* w_i^*\| \leq \epsilon. \tag{7}$$

Remark 5.2. Theorem 5.1 indicates that the per-client sample complexity scales in $\tilde{\mathcal{O}}\left(\frac{\max\{d,M\}}{M\epsilon^2}\right)$. Compared with the single-client setting, which is essentially a noisy linear regression problem with sample complexity $\mathcal{O}(\frac{d}{\epsilon^2})$ (Hsu et al., 2012), FLUTE achieves a linear speedup in terms of M in the high dimensional setting (i.e., $d > M$). When $d < M$, the sample complexity of FLUTE becomes independent with M, which is due to the fact that each client requires a minimum number of samples to have the local optimization problem non-ill-conditioned. Compared with the sample complexity $\mathcal{O}\left(\frac{d}{M} + \log(M)\right)$ of FRL in the noiseless over-parameterized setting (Collins et al., 2021), FLUTE achieves more favorable dependency on M.

Remark 5.3. We note that the dependency on Δ and λ_1, especially Δ, is unique for the under-parameterized FRL. For the special case when $\lambda_{k+1}^* = 0$, the problem we consider essentially falls into the over-parameterized regime, and FLUTE can still be applied. Theorem 5.1 shows that the sample complexity scales in $\mathcal{O}(\frac{\max\{d,M\}}{M\epsilon^2}(\frac{\lambda_1^*}{\Delta})^{10})$. We note that under the assumption that $\mathbf{B}^*$ consists of orthonormal columns, the SOTA sample complexity in the over-parameterized regime scales in $\mathcal{O}(\frac{\max\{d,M\}}{M\epsilon^2}\kappa^4)$ (Tripuraneni et al., 2021), where $\kappa = \sigma_1((\mathbf{W}^*)^\top\mathbf{W}^*)/\sigma_r((\mathbf{W}^*)^\top\mathbf{W}^*)$. Under the same assumption on $\mathbf{B}^*$, we have $\lambda_1^* = \sqrt{\sigma_1((\mathbf{W}^*)^\top\mathbf{W}^*)}$, $\Delta = \sqrt{\sigma_k((\mathbf{W}^*)^\top\mathbf{W}^*)}$, and our sample complexity then becomes $\mathcal{O}(\frac{\max\{d,M\}}{M\epsilon^2}\kappa^5)$. The additional order of κ in the bound is due to an initial state-dependent quantity bounded by $\frac{\lambda_1^*}{\Delta}$. The detailed analysis can be found in Appendix A.2.3.

Remark 5.4. The sample complexity in Theorem 5.1 requires that the size of *each local dataset* be sufficiently large. This is in stark contrast to the sample complexity result in existing works (Collins et al., 2021), which imposes a requirement on the *total number of samples in the system* instead of on each individual client/task. We need the size for each local dataset to be sufficiently large to ensure that every ϕ_i can be locally estimated with a small error so that the top k components of the ground truth $\mathbf{\Phi}$ can be correctly recovered.

Theorem 5.5 (Convergence rate)**.** *Set γ_1, γ_2 and η as in Theorem 5.1. Denote $\kappa_T = (1-\frac{\eta\Delta}{16})^T$. Then, for a constant $T_{\mathcal{R}}$ (defined in Equation (16) in Appendix A) and any $T > T_{\mathcal{R}}$, there exist positive constants c_1 and c_2 such that when $N \geq c_1\frac{(d-\log\delta+\log T)(\sqrt{k(\lambda_1)^2+E}+\sqrt{k}\sigma)^2}{\kappa_T^2\Delta^2}$, for all $T_{\mathcal{R}} < t \leq T$, with probability at least $1-\delta$, we have*

$$\frac{1}{M}\sum_{i\in[M]} \|\mathbf{B}^t w_i^t - \mathbf{B}^* w_i^*\| \leq \frac{c_2\lambda_1^2\sqrt{k}}{\sqrt{M}\eta\Delta^2}\left(1-\frac{\eta\Delta}{16}\right)^t. \tag{8}$$

Remark 5.6. Theorem 5.5 shows that when the number of samples per client N is sufficiently large, FLUTE converges *exponentially* fast. We note that the required number of samples grows exponentially in the total number of iterations. Such an exponential increase in the required number of samples is essential to guarantee that the 'noise' level, which is the gradient estimation error, decays at least as fast as the decay rate of the representation estimation error, which is exponential. Similar phenomenon has been observed in the literature (Mitra et al., 2021; Zhang et al., 2023b). In our problem, there are essentially two parts of 'noise' in the learning process. One is the sub-exponential label noise $\xi_{i,j}$, and the other is the gradient discrepancy arising from the *under-parameterized* nature. This discrepancy persists even when $\mathbf{B}^t$ and $\mathbf{W}^t$ are nearly optimal, leading to an unavoidable gap between $\mathbf{B}^t\mathbf{W}^t$ and $\mathbf{\Phi}$. This gap behaves similarly to the sub-Gaussian noise in the convergence analysis, as elaborated in Section 5.2. Therefore, an exponential increase in the number of samples is required to cope with both parts of the noise and ensure the one-step improvement of the estimation error as iteration grows.

Remark 5.7. We also note that both the sample complexity in Theorem 5.1 and the convergence rate in Theorem 5.5 are influenced by Δ, the gap between λ_k and λ_{k+1}. A smaller Δ signifies a growing challenge in correctly identifying the top-k principal components of $\mathbf{\Phi}$, leading to increased sample complexity and slower convergence. This is due to the challenge of accurately distinguishing and recovering the k-th and $(k+1)$-th significant components from the dataset when Δ is small. Note that in order to successfully distin-

guish σ_k and σ_{k+1}, we need to estimate them to be $\Delta/2$-accurate, i.e., $|\hat{\sigma}_k - \sigma_k| \leq \Delta/2$ and $|\hat{\sigma}_{k+1} - \sigma_{k+1}| \leq \Delta/2$. Hence, the required number of samples per client would grow significantly when Δ is small, and this is arguably inevitable.

5.2. Proof Sketch

In this subsection, we outline the major challenges and main steps in the proof of Theorem 5.5 while deferring the complete analysis to Appendix A. Theorem 5.1 can be proved once Theorem 5.5 is established.

Challenges of the analysis. The analytical frameworks proposed by Collins et al. (2021) and Zhong et al. (2022) for over-parameterized learning scenarios, as well as by Chen et al. (2023) for low-rank matrix approximation, cannot handle the unique challenges that arise in the under-parameterized FRL framework, as elaborated below.

The first major challenge we encounter is to bound the gradient discrepancy on the update of $\mathbf{B}^t$, denoted as $(\mathbf{B}^t\mathbf{W}^t - \mathbf{\Phi})(\mathbf{W}^t)^\top - \sum_{i\in[M]} \frac{\mathbf{X}_i\mathbf{X}_i^\top}{N}\big(\mathbf{B}^t w_i^t - \phi_i\big)(w_i^t)^\top$. Such difficulty is absent in the analyses in Collins et al. (2021) and Zhong et al. (2022) because, in the over-parameterized regime and with a fixed number of samples per client per iteration, the error caused by the gradient discrepancy decays at a rate comparable to that of the representation estimation error. Therefore, the gradient discrepancy will gradually converge to zero. However, for the under-parameterized setting, even with the optimal $(\mathbf{B}^t, \mathbf{W}^t)$, i.e., when $\mathbf{B}^t\mathbf{W}^t = \mathbf{B}^*\mathbf{W}^*$, gradient discrepancy can still be non-zero, as the optimal representation cannot recover all local models, i.e., $\mathbf{B}^*\mathbf{W}^* \neq \mathbf{\Phi}$. Instead, it only decreases when the number of samples N increases. This phenomenon indicates that an increase in the number of samples is essential to ensure one-step improvements of the estimated representations toward the ground-truth representation as the iteration progresses.

Another main challenge is ensuring that neither the gradient discrepancy nor noise-induced errors accumulate over iterations. This is critical as error accumulation can lead to significant deviation from the optimal solution, resulting in poor convergence and degraded model performance. To achieve this, we need to ensure the improvement of the estimation can dominate the effect of potential disturbances.

To tackle these new challenges, we first prove two concentration lemmas (Lemma A.13 and Lemma A.14 in Appendix A.2.5) to ensure that the norm of the gradient discrepancy can be bounded when local datasets are sufficiently large. Next, to address the second challenge of avoiding the accumulation of gradient discrepancy and noise-induced errors over iterations, we develop iteration-dependent upper bounds for sample complexity (Lemma A.10 and Lemma A.11 in Appendix A.2.3). These bounds guarantee that the improvement in estimation, i.e., the 'distance' improvement between our estimated model and the optimal low-rank model, can mitigate potential disturbances caused by gradient discrepancy and noise. We establish this by introducing a novel approach to derive an accuracy-dependent upper bound for the per-client sample complexity, ensuring the error caused by the gradient discrepancy decays as fast as the increase of the signal-to-noise ratio (SNR), formally introduced in Appendix A.

Main steps of the proof. First, we transform the asymmetric matrix factorization problem into a symmetric problem by appropriately padding $\mathbf{0}$ columns or rows to $\mathbf{B}^t$ and $\mathbf{W}^t$ and constructing the updating matrices $\mathbf{\Theta}^t$ (see Appendix A). Our goal is then to prove that $\mathbf{\Theta}^t(\mathbf{\Theta}^t)^\top$ converges. We first show that, with a small random initialization, $\mathbf{\Theta}^t$ will enter a region containing the optima with high probability. Then, utilizing Lemma A.13 and Lemma A.14, we demonstrate that when $\mathbf{\Theta}^t$ enters the region $\mathcal{R}$, it will remain in this region with high probability despite gradient discrepancy and noise. Finally, utilizing Lemma A.10 and Lemma A.11, we show that when N is sufficiently large, $\mathbf{\Theta}^t(\mathbf{\Theta}^t)^\top$ converges at a linear rate with high probability under the influence of gradient discrepancy and noise, provided that the initialization condition satisfies $\mathbf{\Theta}^0 \in \mathcal{R}$.

6. General `FLUTE`

In this section, we extend `FLUTE` designed for linear models to more general settings. Specifically, we use $\psi_{\mathbf{B}}$ to denote the representation, and assume linear local heads $f_i(z) = \mathbf{H}_i^\top z + b_i$, where $\mathbf{H}_i \in \mathbb{R}^{k\times m}, b_i \in \mathbb{R}^m$. This is motivated by the neural network architecture where all layers before the last layer are abstracted as the representation layer, and the last layer is linear. Then, the objective function becomes

$$\min_{\mathbf{B},\{\mathbf{H}_i\},\{b_i\}} \frac{1}{M}\sum_{i\in[M]} \frac{1}{N}\sum_{(x,y)\in\mathcal{D}_i} \ell\big(\mathbf{H}_i^\top \psi_{\mathbf{B}}(x) + b_i, y\big) + \lambda R(\{\mathbf{H}_i\}, \mathbf{B}), \quad (9)$$

where $R(\{\mathbf{H}_i\}, \mathbf{B})$ is the regularization term to encourage the alignment of local models with the global optimum structure.

The general `FLUTE` algorithm for solving problem (9) is provided in Algorithm 2 in Appendix B. Given the nonlinearity of $\psi_{\mathbf{B}}$, the penalty introduced in linear `FLUTE` is not directly applicable to the general problem. We thus formulate and design new penalty terms, following the same principles that motivated the design in the linear setting. This is to mitigate the local over-fitting induced by local updates and to encourage a structure benefit to global optimization. As a concrete example, we present a design of the penalty term for the classification problem with CNN as a

prediction model in Section 7.2.

7. Experimental Results

7.1. Synthetic Datasets

We generate a synthetic dataset as follows. First, we randomly generate ϕ_i according to a d-dimensional standard Gaussian distribution. For each ϕ_i, we then randomly generate N pairs of (x, y), where x is sampled from a standard Gaussian distribution, ξ is sampled from a centered Gaussian distribution with variance σ^2, and $y = \phi_i^\top x + \xi$.

In Figure 1, we compare FLUTE with FedRep (Collins et al., 2021). We measure the quality of the learned representation $\mathbf{B}^t$ and $\mathbf{W}^t$ over the metric $\frac{1}{M}\sum_{i\in[M]} \|\mathbf{B}^t w_i^t - \phi_i\|$. We emphasize that FedRep requires empirical covariance estimated from the local datasets to be transmitted to the server for the initialization. Thus, it begins with a good estimate of the subspace spanned by $\mathbf{B}^*$. In contrast, FLUTE commences with a random initialization of both the representation and the heads. As a result, FedRep converges to a relatively small error within the few initial epochs, while FLUTE needs to go through more epochs to obtain a good estimate of the representation. However, as the learning progresses over more epochs, FLUTE eventually outperforms FedRep. To validate this hypothesis, we introduce FedRep(RI) in our experiments, which has the same initialization as FLUTE but is otherwise identical to FedRep. We see from Figure 1 that when FedRep is randomly initialized, FLUTE outperforms FedRep(RI) in much fewer iterations.

We also observe that the performance gain of FLUTE is more pronounced in highly under-parameterized scenarios, i.e., where k is relatively small. As k increases, the gap between the convergence rates of FLUTE and FedRep narrows. These results demonstrate that FLUTE achieves better performance in the under-parameterized regime. In the additional experimental results included in Appendix C, we also observe that when the number of participating clients M increases, the average error of the model learned from FLUTE decreases, which is consistent with Theorem 5.5.

7.2. Real World Datasets

Datasets and models. We now evaluate the performance of general FLUTE on multi-class classification tasks with real-world datasets CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009). For all experiments, we adopt a convolutional neural network (CNN) with two convolution layers, two fully connected layers with ReLU activation, and a final fully connected layer with a softmax activation function. A detailed description of the CNN structure is deferred to Appendix C of the Appendix.

Algorithms for comparison. We compare FLUTE with several baseline algorithms, including FedAvg (McMahan et al., 2017), Fed-LG (Liang et al., 2020), FedPer (Arivazhagan et al., 2019), FedRep (Collins et al., 2021), FedRod (Chen & Chao, 2021) and FedCP (Zhang et al., 2023a). Fed-LG is designed to learn a common head shared across clients while allowing for localized representations, while FedPer and FedRep both assume shared representation and personalized local heads. FedRod extends the model considered in FedRep by adding another head layer into the local model, and FedCP further equips a conditional policy network into the local model. We also consider variants of FLUTE and FedRep, denoted as FLUTE* and FedRep*, respectively, under which we vary the number of updates of the local heads in each communication round, as elaborated later.

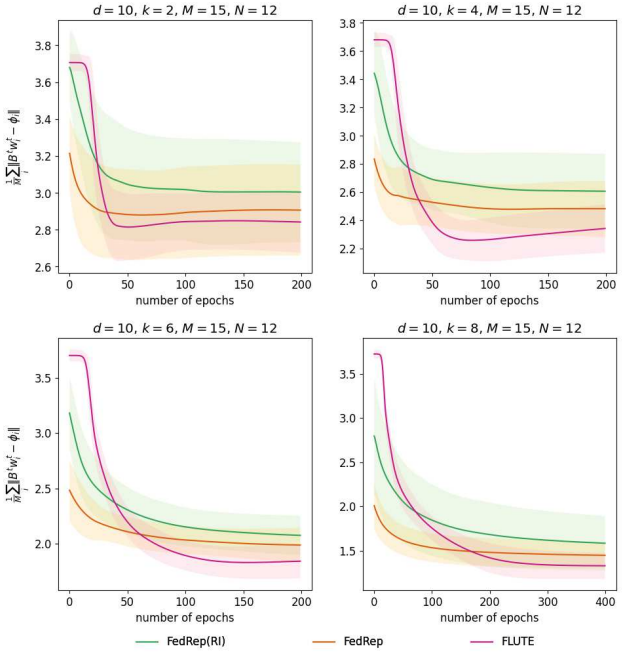

Figure 1. Experimental results with synthetic datasets.

Loss function and penalty. For algorithms other than FLUTE and FLUTE*, the local loss function is chosen as $\mathcal{L}_i = \frac{1}{N}\sum_{(x,y)\in\mathcal{D}_i} \mathcal{L}_{\text{CE}}(\mathbf{H}_i^\top \psi_{\mathbf{B}}(x) + b_i, y)$, where $\mathcal{L}_{\text{CE}}$ is the cross entropy loss. The local loss function for FLUTE and FLUTE* are specialized as

$$\mathcal{L}_i(\mathbf{B}, b, \mathbf{H}) = \frac{1}{N}\sum_{(x,y)\in\mathcal{D}_i} \mathcal{L}_{\text{CE}}\big(\mathbf{H}_i^\top \psi_{\mathbf{B}}(x) + b_i, y\big) + \lambda_1\|\psi_{\mathbf{B}}(x)\|^2 + \lambda_2\|\mathbf{H}_i\|_F^2 + \lambda_3\mathcal{NC}_i(\mathbf{H}_i), \tag{10}$$

where $y \in \mathbb{R}^m$ is a one-hot vector whose k-th entry is 1 if the corresponding x belongs to class k, λ_1, λ_2 and λ_3 are non-negative regularization parameters. $\mathcal{NC}_i(\mathbf{H}_i)$ is motivated by Papyan et al. (2020) and set as $\left\| \frac{\mathbf{H}_i^\top \mathbf{H}_i}{\|\mathbf{H}_i^\top \mathbf{H}_i\|_F} - \frac{1}{\sqrt{m-1}}\mathbf{u}_i\mathbf{u}_i^\top \odot \left(\mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top\right)\right\|_F$, where $\mathbf{u}_i$ is an m-dimensional one-hot vector whose c-th entry

Table 1. Average test accuracy on CIFAR-10 and CIFAR-100.

Dataset	CIFAR-10				CIFAR-100			
Partition	50×2	50×5	100×2	100×5	100×5	100×10	100×20	100×40
FedAvg	34.460±1.083	47.217±0.395	41.584±0.433	51.876±0.675	20.212±0.574	31.533±0.519	34.659±0.482	32.902±0.195
FedAvg-FT	83.996±0.948	71.465±0.701	84.688±0.437	70.884±0.697	78.342±0.574	66.660±0.370	54.464±0.178	44.858±0.119
Fed-LG	82.724±2.137	61.820±0.409	83.019±0.431	62.957±0.895	72.526±0.692	53.526±0.151	34.445±0.375	22.702±0.315
FedPer	85.173±1.082	74.015±0.724	86.168±0.703	73.666±0.281	76.001±0.454	67.100±0.229	56.066±0.389	44.689±0.411
FedRep	86.133±0.775	71.737±0.296	86.685±0.766	73.808±0.561	78.621±0.159	68.530±0.255	56.360±0.245	43.061±0.476
FedRep*	87.320±1.485	75.766±0.220	87.177±0.489	75.296±0.505	78.892±0.410	68.630±0.705	56.654±0.609	42.025±0.404
FedRoD	79.476±2.966	68.728±1.750	83.296±1.545	72.116±0.788	74.299±0.338	66.462±0.284	57.280±0.105	48.120±0.186
FedCP	85.361±1.605	71.603±0.885	84.798±0.489	71.344±0.587	74.266±0.559	66.426±0.372	57.067±0.483	43.638±0.415
FLUTE	87.012±0.453	76.478±0.484	86.128±1.007	76.918±0.712	77.750±0.615	70.598±0.282	59.243±0.334	48.169±0.597
FLUTE*	**87.713±1.365**	**76.543±0.921**	**88.567±0.457**	**78.255±0.688**	**79.560±0.627**	**70.844±0.419**	**59.714±0.448**	**48.170±0.440**

is 1 if $c \in \mathcal{C}_i$, and $\odot$ denotes the Hadamard product. We specialize the regularization term optimized on the server side as $R(\{\mathbf{H}_i\}) = \sum_i \mathcal{NC}_i(\mathbf{H}_i)$. Note that for general FLUTE specified to a classification problem, we penalize $\|\psi_{\mathbf{B}}(x)\|$ instead of directly penalizing the parameter $\mathbf{B}$. Since $\|\psi_{\mathbf{B}}(x)\|$ depends on data, the regularization term is optimized partially on the client side and partially on the server side.

Compared with the objective function in Equation (4) for the linear case, the term $\lambda_3 \mathcal{NC}_i(\mathbf{H}_i)$ replaces term (I) and $\lambda_1 \|\psi_{\mathbf{B}}(x)\|_2^2 + \lambda_2 \|\mathbf{H}_i\|_F^2$ replaces term (II). The primary goal of introducing $\lambda_3 \mathcal{NC}_i(\mathbf{H}_i)$ is to mitigate local over-fitting that occurs during local updates in the training process. As elaborated in Appendix B.3, such a regularization term promotes a beneficial structure for the global model, facilitating efficient learning performance. This term shares the same motivation as the term (I) in the linear scenario, which focuses on distilling significant components from the model to mitigate local over-fitting effects. For term (II), we only replace $\|\mathbf{B}^\top \mathbf{B}\|_F^2$ with $\|\psi_{\mathbf{B}}(x)\|^2$, since the representation is not linear in general.

Implementation and evaluation. We use m to denote the number of classes assigned to each client. For CIFAR-10 dataset, we consider four (N, m) pairs: $(50, 2)$, $(50, 5)$, $(100, 2)$ and $(100, 5)$; For CIFAR-100 dataset, we consider four (N, m) pairs: $(100, 5)$, $(100, 10)$, $(100, 20)$ and $(100, 40)$.

For experiments conducted on the CIFAR-10 dataset, all algorithms are executed over 100 communication rounds. For LG-Fed, FedPer, FedRoD, FedCP and FLUTE, each client performs one round of local updates in each communication round. FedRep performs one epoch of local head update and an additional epoch for the local representation update. Compared with FedRep, FedRep* processes 10 epochs to update its local heads and one epoch to update its representation. For comparison, FLUTE* also runs 11 rounds of local updates, updating both representation and local head in the first round, followed by 10 rounds of only updating the local head.

The experiments on the CIFAR-100 dataset also use 100 communication rounds. The number of local updates for LG-Fed, FedPer, FedRoD, FedCP and FLUTE are set to 5. FedRep is configured to update the local representation and head for 5 epochs each, while FedRep* allocates 5 epochs for updating the local representation and 10 for updating the local head. FLUTE* runs 15 epochs of local updates, where the initial 5 epochs update both the representation and local head while the subsequent 10 epochs solely update the local head.

Averaged performance. The results are reported in Table 1. It is evident that FLUTE and FLUTE* consistently outperform other baseline algorithms in all experiments conducted on CIFAR-10 and CIFAR-100 datasets. This superior performance is attributed to the tailored design that encourages the locally learned models to move towards a global optimal solution rather than a local optimum. We also observe that the gain of FLUTE and FLUTE* becomes more prominent with larger N and m. Intuitively, larger N and m implies more severe under-parameterization for the given CNN model, and our algorithms exhibit more advantage for such cases.

8. Conclusion

To the best of our knowledge, this paper represents the first effort in the study of federated representation learning in the under-parameterized regime, which is of great practical importance. We have proposed a novel FRL algorithm FLUTE that was inspired by asymmetric low-rank matrix approximation. FLUTE incorporates a novel regularization term in the loss function and solves the corresponding ERM problem in a federated manner. We proved the convergence of FLUTE and established the per-client sample complexity that is comparable to the over-parameterized result but with very different proof techniques. We also extended FLUTE to general (non-linear) settings which are of practical interest. FLUTE demonstrated superior performance over existing FRL solutions in both synthetic and real-world tasks, highlighting its advantages for efficient learning in the under-parameterized regime.

Acknowledgements

The work of R. Liu and J. Yang was supported in part by the U.S. National Science Foundation under grants CNS-1956276, CNS-2003131 and CNS-2114542. The work of C. Shen was supported in part by the U.S. National Science Foundation under grants ECCS-2332060, CPS-2313110, ECCS-2143559, and ECCS-2033671.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Arivazhagan, M. G., Aggarwal, V., Singh, A. K., and Choudhary, S. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849—-15854, July 2019.

Chen, H., Chen, X., Elmasri, M., and Sun, Q. Fast global convergence of gradient descent for low-rank matrix approximation. *arXiv preprint arXiv:2305.19206*, 2023.

Chen, H.-Y. and Chao, W.-L. On bridging generic and personalized federated learning for image classification. *arXiv preprint arXiv:2107.00778*, 2021.

Chen, L., Lin, S., Lu, X., Cao, D., Wu, H., Guo, C., Liu, C., and Wang, F.-Y. Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 22(6):3234–3246, 2021.

Chen, Y., Dai, X., Chen, D., Liu, M., Dong, X., Yuan, L., and Liu, Z. Mobile-Former: Bridging MobileNet and transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5270–5279, 2022.

Collins, L., Hassani, H., Mokhtari, A., and Shakkottai, S. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pp. 2089–2099. PMLR, 2021.

Davidson, K. R. and Szarek, S. J. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001.

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

Duchi, J., Feldman, V., Hu, L., and Talwar, K. Subspace recovery from heterogeneous data with non-isotropic noise, 2022.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Ge, R., Jin, C., and Zheng, Y. No spurious local minima in nonconvex low rank problems: A unified geometric analysis, 2017.

Golub, G. H. and Van Loan, C. F. *Matrix computations*. JHU Press, 2013.

He, C., Annavaram, M., and Avestimehr, S. Group knowledge transfer: Federated learning of large CNNs at the edge. *Advances in Neural Information Processing Systems*, 33:14068–14080, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Hitaj, B., Ateniese, G., and Perez-Cruz, F. Deep models under the gan: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 603–618, 2017.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324, 2019.

Hsu, D., Kakade, S. M., and Zhang, T. Random design analysis of ridge regression. In *Conference on Learning Theory*, pp. 9–1. JMLR Workshop and Conference Proceedings, 2012.

Hwang, S.-G. Cauchy's interlace theorem for eigenvalues of hermitian matrices. *The American Mathematical Monthly*, 111(2):157–159, 2004.

Jiang, L., Chen, Y., and Ding, L. Algorithmic regularization in model-free overparametrized asymmetric matrix factorization. *arXiv preprint arXiv:2203.02839*, 2022.

Ju, R.-Y., Lin, T.-Y., Jian, J.-H., and Chiang, J.-S. Efficient convolutional neural networks on Raspberry Pi for image classification. *Journal of Real-Time Image Processing*, 20(2):21, 2023.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

Liang, P. P., Liu, T., Ziyin, L., Allen, N. B., Auerbach, R. P., Brent, D., Salakhutdinov, R., and Morency, L.-P. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pp. 1273–1282. PMLR, 2017.

Melis, L., Song, C., De Cristofaro, E., and Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 691–706, 2019.

Mitra, A., Jaafar, R., Pappas, G. J., and Hassani, H. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34:14606–14619, 2021.

Oneto, L., Ridella, S., and Anguita, D. Do we really need a new theory to understand over-parameterization? *Neurocomputing*, 543:126227, 2023.

OpenAI. Gpt-4 technical report, 2023.

Papyan, V., Han, X. Y., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, Sept. 2020.

Pitaval, R.-A., Dai, W., and Tirkkonen, O. Convergence of gradient descent for low-rank matrix approximation. *IEEE Transactions on Information Theory*, 61(8):4451–4457, 2015.

Rudelson, M. and Vershynin, R. The littlewood-offord problem and invertibility of random matrices, 2008.

Shen, Z., Ye, J., Kang, A., Hassani, H., and Shokri, R. Share your representation only: Guaranteed improvement of the privacy-utility tradeoff in federated learning, 2023.

Tan, J., Mason, B., Javadi, H., and Baraniuk, R. Parameters or privacy: A provable tradeoff between overparameterization and membership inference. *Advances in Neural Information Processing Systems*, 35:17488–17500, 2022.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114. PMLR, 2019.

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. Sample efficient linear meta-learning by alternating minimization, 2021.

Tirer, T. and Bruna, J. Extended unconstrained features model for exploring deep neural collapse, 2022.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023.

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

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wang, J., Kolar, M., and Srebro, N. Distributed multi-task learning with shared representation. *arXiv preprint arXiv:1603.02185*, 2016a.

Wang, L., Zhang, X., and Gu, Q. A unified computational and statistical framework for nonconvex low-rank matrix estimation, 2016b.

Wang, S., Tuor, T., Salonidis, T., Leung, K. K., Makaya, C., He, T., and Chan, K. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6): 1205–1221, 2019.

Wang, Z., Song, M., Zhang, Z., Song, Y., Wang, Q., and Qi, H. Beyond inferring class representatives: User-level privacy leakage from federated learning, 2018.

Ye, T. and Du, S. S. Global convergence of gradient descent for asymmetric low-rank matrix factorization, 2021.

Yu, T., Bagdasaryan, E., and Shmatikov, V. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*, 2020.

Zhang, J., Hua, Y., Wang, H., Song, T., Xue, Z., Ma, R., and Guan, H. Fedcp: Separating feature information for personalized federated learning via conditional policy. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3249–3261, 2023a.

Zhang, T. T. C. K., Toso, L. F., Anderson, J., and Matni, N. Meta-learning operators to optimality from multi-task non-iid data, 2023b.

Zhong, A., He, H., Ren, Z., Li, N., and Li, Q. Feddar: Federated domain-aware representation learning. *arXiv preprint arXiv:2209.04007*, 2022.

Zhu, Z., Li, Q., Tang, G., and Wakin, M. B. The global optimization geometry of low-rank matrix optimization, 2021.

Notations. Throughout this paper, bold capital letters (e.g., $\mathbf{X}$) denote matrices, and calligraphic capital letters (e.g., $\mathcal{C}$) denote sets. We use $\mathrm{tr}(\mathbf{X})$ to denote the trace of matrix $\mathbf{X}$, $\sigma_{\min}(\mathbf{X})$ and $\sigma_{\max}(\mathbf{X})$ to denote the minimum and maximum singular values of $\mathbf{X}$, respectively, and $\mathrm{diag}(x_1, \cdots, x_d)$ to denote a d-dimensional diagonal matrix with diagonal entries $x_1, \cdots, x_d$. $|\mathcal{C}|$ denotes the cardinality of set $\mathcal{C}$, and $\{X_i\}_{i\in[N]}$ denotes the set $\{X_1, \cdots, X_N\}$. We use $\langle x, y\rangle$ to denote the inner product of x and y, and $\|x\|$ to denote the Euclidean norm of vector x. We use $f \circ \psi$ to denote the composition of functions $f : \mathbb{R}^k \to \mathbb{R}^m$ and $\psi : \mathbb{R}^d \to \mathbb{R}^k$, i.e., $(f \circ \psi)(x) = f(\psi(x))$. $a \lesssim b$ indicates $a \leq Cb$ for a positive constant C. $\mathbf{I}_d$ represents a $d \times d$ identity matrix, and $\mathbf{0}$ is a d-dimensional all-zero vector.

Denote $\overline{d} := \max\{d, M\}$, $\underline{d} := \min\{d, M\}$, and $\mathbf{\Phi}_\star \in \mathbb{R}^{\overline{d}\times\overline{d}}$ as the matrix constructed from $\mathbf{\Phi} \in \mathbb{R}^{d\times M}$ by padding all-zero columns or rows. Define its SVD as $\mathbf{\Phi}_\star = \mathbf{U}_\star\mathbf{\Lambda}_\star\mathbf{V}_\star^\top$. Denote $\tilde{\mathbf{\Lambda}} = \mathrm{diag}(2\mathbf{\Lambda}_\star, -2\mathbf{\Lambda}_\star)$ and let $\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_{2\overline{d}}^*$ be the eigenvalues of $\tilde{\mathbf{\Lambda}}$, with $\Delta = \lambda_k^* - \lambda_{k+1}^*$. Note that the definition of Δ is consistent with the definition in Section 5. For clarity of presentation, we use σ_ξ to denote the standard deviation of the noise ξ instead of σ that is used in the main paper.

A. Analysis of the FLUTE Linear Algorithm

A.1. Preliminaries

We start with the updating rule of $\mathbf{B}^t$ and $\mathbf{W}^t$ in Algorithm 1.

For $\mathbf{B}^t$, from FLUTE we have the following updating rule:

$$\begin{aligned}\mathbf{B}^{t+1} &= \mathbf{B}^t - \frac{\eta}{N}\sum_{i\in[M]}\sum_{j\in[N]} x_{i,j}\big(x_{i,j}^\top \mathbf{B}^t w_i^t - y_{i,j}\big)(w_i^t)^\top - \frac{\eta}{2}\mathbf{B}^t\big((\mathbf{B}^t)^\top\mathbf{B}^t - \mathbf{W}^t(\mathbf{W}^t)^\top\big)\\ &= \mathbf{B}^t - \frac{\eta}{N}\sum_{i\in[M]} \mathbf{X}_i\mathbf{X}_i^\top\big(\mathbf{B}^t w_i^t - \phi_i\big)(w_i^t)^\top - \frac{\eta}{2}\mathbf{B}^t\big((\mathbf{B}^t)^\top\mathbf{B}^t - \mathbf{W}^t(\mathbf{W}^t)^\top\big).\end{aligned}$$

Since data points $\{x_{i,j}\}$ are sampled from a standard Gaussian distribution, for large N, it holds that $\mathbf{X}_i\mathbf{X}_i^\top/N \approx \mathbf{I}$. Then, we introduce the following definition:

$$\mathbf{Q}^{t+1} \coloneqq \eta\sum_{i\in[M]}\big(\mathbf{B}^t w_i^t - \phi_i\big)(w_i^t)^\top - \eta\sum_{i\in[M]}\frac{\mathbf{X}_i\mathbf{X}_i^\top}{N}\big(\mathbf{B}^t w_i^t - \phi_i\big)(w_i^t)^\top + \eta\sum_{i\in[M]}\frac{\mathbf{X}_i\mathbf{E}_i(w_i^t)^\top}{N}. \tag{11}$$

With this definition, the updating rule of $\mathbf{B}^t$ can be rewritten as

$$\mathbf{B}^{t+1} = \mathbf{B}^t - \eta\big(\mathbf{B}^t\mathbf{W}^t - \mathbf{\Phi}\big)(\mathbf{W}^t)^\top - \frac{\eta}{2}\mathbf{B}^t\big((\mathbf{B}^t)^\top\mathbf{B}^t - \mathbf{W}^t(\mathbf{W}^t)^\top\big) + \mathbf{Q}^{t+1}.$$

Now we consider the updating rule of $\mathbf{W}$. Observe that each of its columns satisfies

$$\begin{aligned}w_i^{t+1} &= w_i^t - \frac{\eta}{N}\sum_{j\in[N]}(\mathbf{B}^t)^\top x_{i,j}\big((x_{i,j})^\top\mathbf{B}^t w_i^t - y_{i,j}\big)\\ &= w_i^t - \frac{\eta}{N}(\mathbf{B}^t)^\top\mathbf{X}_i\mathbf{X}_i^\top\big(\mathbf{B}^t w_i^t - \phi_i\big).\end{aligned}$$

We define $\tilde{\mathbf{Q}}^{t+1} \coloneqq [\tilde{q}_1^{t+1}, \cdots, \tilde{q}_M^{t+1}]$, where each of its columns is given by

$$\tilde{q}_i^{t+1} \coloneqq \eta(\mathbf{B}^t)^\top\big(\mathbf{B}^t w_i^t - \phi_i\big) - \frac{\eta}{N}(\mathbf{B}^t)^\top\mathbf{X}_i\mathbf{X}_i^\top\big(\mathbf{B}^t w_i^t - \phi_i\big) + \frac{\eta}{N}(\mathbf{B}^t)^\top\mathbf{X}_i\mathbf{E}_i. \tag{12}$$

Then, $\mathbf{W}^t$ is updated according to

$$\mathbf{W}^{t+1} = \mathbf{W}^t - \eta(\mathbf{B}^t)^\top(\mathbf{B}^t\mathbf{W}^t - \mathbf{\Phi}) + \frac{\eta}{2}\big((\mathbf{B}^t)^\top\mathbf{B}^t - \mathbf{W}^t(\mathbf{W}^t)^\top\big)\mathbf{W}^t + \tilde{\mathbf{Q}}^{t+1}.$$

Recall the SVD of $\mathbf{\Phi}$ is denoted as $\mathbf{\Phi} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$. Further denote $\tilde{\mathbf{B}}^t = \mathbf{U}^\top\mathbf{B}^t$ and $\tilde{\mathbf{W}}^t = \mathbf{W}^t\mathbf{V}$. Then, we have

$$\tilde{\mathbf{B}}^{t+1} = \tilde{\mathbf{B}}^t - \eta\big(\tilde{\mathbf{B}}^t\tilde{\mathbf{W}}^t - \mathbf{\Lambda}\big)\big(\tilde{\mathbf{W}}^t\big)^\top - \frac{\eta}{2}\tilde{\mathbf{B}}^t\big((\tilde{\mathbf{B}}^t)^\top\tilde{\mathbf{B}}^t - \tilde{\mathbf{W}}^{t-1}(\tilde{\mathbf{W}}^t)^\top\big) + \mathbf{U}^\top\mathbf{Q}^{t+1},$$

$$\tilde{\mathbf{W}}^{t+1} = \tilde{\mathbf{W}}^t - \eta(\tilde{\mathbf{B}}^t)^\top\big(\tilde{\mathbf{B}}^t\tilde{\mathbf{W}}^t - \mathbf{\Lambda}\big) - \frac{\eta}{2}\big((\tilde{\mathbf{B}}^t)^\top\tilde{\mathbf{B}}^t - \tilde{\mathbf{W}}^t(\tilde{\mathbf{W}}^t)^\top\big)\tilde{\mathbf{W}}^t + \tilde{\mathbf{Q}}^{t+1}\mathbf{V}.$$

Similar to the definition of $\mathbf{\Phi}_\star$, we construct $\mathbf{B}_\star^t \in \mathbb{R}^{\bar{d}\times k}$ and $\mathbf{W}_\star^t \in \mathbb{R}^{k\times\bar{d}}$ by padding all-zero columns or rows to $\tilde{\mathbf{B}}^t$ and $\tilde{\mathbf{W}}$, respectively. Similarly, we obtain $\mathbf{Q}_\star^t \in \mathbb{R}^{\bar{d}\times k}$ and $\tilde{\mathbf{Q}}_\star^t \in \mathbb{R}^{k\times\bar{d}}$ by padding all-zero columns or rows to $\mathbf{Q}^t$ and $\tilde{\mathbf{Q}}^t$, respectively. Then, we define $\mathbf{\Theta}^t$ and $\mathbf{R}^t$ as

$$\mathbf{\Theta}^t = \begin{bmatrix} \frac{(\mathbf{B}_\star^t)^\top + \mathbf{W}_\star^t}{\sqrt{2}} & \frac{(\mathbf{B}_\star^t)^\top - \mathbf{W}_\star^t}{\sqrt{2}} \end{bmatrix}^\top,$$
$$\mathbf{R}^t = \begin{bmatrix} \frac{(\mathbf{Q}_\star^t)^\top\mathbf{U}_\star + \tilde{\mathbf{Q}}_\star^t\mathbf{V}_\star}{\sqrt{2}} & \frac{(\mathbf{Q}_\star^t)^\top\mathbf{U}_\star - \tilde{\mathbf{Q}}_\star^t\mathbf{V}_\star}{\sqrt{2}} \end{bmatrix}^\top.$$

Then, the updating rule of $\mathbf{\Theta}^t$ can be described as

$$\mathbf{\Theta}^{t+1} = \mathbf{\Theta}^t + \frac{\eta}{2}\tilde{\mathbf{\Lambda}}\mathbf{\Theta}^t - \frac{\eta}{2}\mathbf{\Theta}^t(\mathbf{\Theta}^t)^\top\mathbf{\Theta}^t + \mathbf{R}^{t+1}. \tag{13}$$

Let $\mathbf{\Theta}^t = [(\mathbf{\Theta}_k^t)^\top(\mathbf{\Theta}_{\text{res}}^t)^\top]^\top$ and $\mathbf{R}^t = [(\mathbf{R}_k^t)^\top \quad (\mathbf{R}_{2\bar{d}-k}^t)^\top]^\top$ where $\mathbf{\Theta}_k^t \in \mathbb{R}^{k\times k}$, $\mathbf{\Theta}_{\text{res}}^t \in \mathbb{R}^{(2\bar{d}-k)\times k}$, $\mathbf{R}_k^t \in \mathbb{R}^{k\times k}$ and $\mathbf{R}_{2\bar{d}-k}^t \in \mathbb{R}^{(2\bar{d}-k)\times k}$. Then, we decompose the updating rule of $\mathbf{\Theta}^t$ as

$$\mathbf{\Theta}_k^t = \mathbf{\Theta}_k^{t-1} + \frac{\eta}{2}\tilde{\mathbf{\Lambda}}_k\mathbf{\Theta}_k^{t-1} - \frac{\eta}{2}\mathbf{\Theta}_k^{t-1}(\mathbf{\Theta}^{t-1})^\top\mathbf{\Theta}^{t-1} + \mathbf{R}_k^t, \tag{14}$$
$$\mathbf{\Theta}_{\text{res}}^t = \mathbf{\Theta}_{\text{res}}^{t-1} + \frac{\eta}{2}\tilde{\mathbf{\Lambda}}_{\text{res}}\mathbf{\Theta}_{\text{res}}^{t-1} - \frac{\eta}{2}\mathbf{\Theta}_{\text{res}}^{t-1}(\mathbf{\Theta}^{t-1})^\top\mathbf{\Theta}^{t-1} + \mathbf{R}_{2d-k}^t. \tag{15}$$

A.2. Proof of Theorem 5.5

First, we restate Theorem 5.5 as follows.

Theorem A.1 (Restatement of Theorem 5.5)**.** *Set λ and η as in Theorem 5.1. Then for constant $T_{\mathcal{R}}$ and any $T > T_{\mathcal{R}}$, there exist positive constants c_1 and c_2 such that when the number of samples per client satisfies $N \geq c_1\frac{(\bar{d}-\log\delta+\log T)(k\sqrt{(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)^2}{\kappa_T^2\Delta^2}$, for all $T_{\mathcal{R}} < t \leq T$ we have*

$$\frac{1}{M}\sum_{i\in[M]}\|\mathbf{B}^t w_i^t - \mathbf{B}^* w_i^*\| \leq \frac{c_2\kappa\sqrt{k}}{\sqrt{M}}\Big(1-\frac{\eta\Delta}{16}\Big)^t,$$

with probability at least $1-\delta$, where $\kappa_T = (1-\frac{\eta\Delta}{16})^T$.

Overview of the proof. The proof of Theorem 5.5 consists of three main steps.

- **Step 1:** We show that with a small random initialization, $\mathbf{\Theta}^t$ will enter a region containing the optima with high probability (see Appendix A.2.1).
- **Step 2:** We show that once $\mathbf{\Theta}^t$ enters this region, it will stay in it with high probability (see Appendix A.2.2).
- **Step 3:** We show that when N is sufficiently large, with high probability it holds that $\|\mathbf{\Theta}^t(\mathbf{\Theta}^t)^\top - \text{diag}(\tilde{\mathbf{\Lambda}}_k, \mathbf{0})\|$ converges to 0 at a linear rate when the initialization satisfies $\mathbf{\Theta}^0 \in \mathcal{R}$ (see Appendix A.2.3).

We then put pieces together and prove Theorem 5.5 in Appendix A.2.4. We introduce some auxiliary lemmas in Appendix A.2.5.

A.2.1. Step 1: Entering a Region with Small Random Initialization

We first introduce the following definitions, adapted from the proof in Chen et al. (2023). Recall that $\Delta = \lambda_k^* - \lambda_{k+1}^*$. We define

$$\mathcal{R} = \Big\{\mathbf{\Theta}^t = \begin{bmatrix}\mathbf{\Theta}_k^t \\ \mathbf{\Theta}_{\text{res}}^t\end{bmatrix} \in \mathbb{R}^{2d\times k}\Big|\sigma_1^2(\mathbf{\Theta}^t) \leq 2\lambda_1^*, \quad \sigma_1^2(\mathbf{\Theta}_{\text{res}}^t) \leq \lambda_k^* - \Delta/2, \quad \sigma_k^2(\mathbf{\Theta}_k^t) \geq \Delta/4\Big\},$$

$$\mathcal{R}_s = \left\{ \mathbf{\Theta}^t = \begin{bmatrix} \mathbf{\Theta}_k^t \\ \mathbf{\Theta}_{\text{res}}^t \end{bmatrix} \in \mathbb{R}^{2d \times k} \middle| \sigma_1^2(\mathbf{\Theta}^t) \leq 2\lambda_1^*, \quad \sigma_1^2(\mathbf{\Theta}_{\text{res}}^t) \leq \lambda_k^* - \Delta/2 \right\}.$$

Then, we establish the following proposition.

Proposition A.2. *Assume $\eta \leq \frac{1}{6\lambda_1^*}$ and all entries of $\mathbf{B}^0$ and $\mathbf{W}^0$ are independently sampled from $\mathcal{N}(0, \alpha^2)$ with a sufficiently small α. Then, if*

$$\sqrt{N} \geq \max\left\{ \frac{\sqrt{d - \log\delta}}{\sqrt{c_1}}, \frac{3456\sqrt{d - \log\delta}\sqrt{\lambda_1^*}(\sqrt{k(\lambda_1^*)^2 + E} + \sqrt{k}\sigma_\xi)}{\min\{\sigma_1(\mathbf{\Theta}_{res}^0), \sigma_1(\mathbf{\Theta}_{res}^0)\}\Delta\sqrt{c_1}} \right\},$$

with probability at least $1 - ct\delta$ for some constant $c > 0$, $\mathbf{\Theta}^t$ will enter region $\mathcal{R}$ for some $t \in [T_{\mathcal{R}}]$, where

$$T_{\mathcal{R}} = \frac{\log(\Delta/(4\sigma_k^2(\mathbf{\Theta}_k^0)))}{2\log(1 + \frac{\eta}{2}(\lambda_k^* - \Delta/2))}. \tag{16}$$

The proof of Proposition A.2 relies on Lemma A.4 and Lemma A.5, which will be introduced shortly. Before that, we state the following claim introduced in Chen et al. (2023):

Claim A.3. *$\sigma_1^2(\mathbf{\Theta}^0) \leq \lambda_1^*$, $\sigma_1^2(\mathbf{\Theta}_{res}^0) \leq \lambda_k^* - \Delta/2$, $\sigma_k^2(\mathbf{\Theta}_k^0) \leq \Delta/4$ and $\sigma_1^2(\mathbf{\Theta}_{res}^0) \leq c_1 \cdot \sigma_k(\mathbf{\Theta}_k^0)^{1+\kappa}$, where $c_1 = \frac{\Delta^{1-\kappa/2}}{2^{3-\kappa}}$ and $\kappa = \frac{\log(1+\frac{\eta}{2}\lambda_{k+1}^* + \frac{\eta}{8}\Delta)}{\log(1+\frac{\eta}{2}(\lambda_k^* - \Delta/2))} < 1$.*

The following lemma shows that with a small random initialization, Claim A.3 holds with high probability.

Lemma A.4. *Assume all entries of $\mathbf{B}^0$ and $\mathbf{W}^0$ are independently sampled from $\mathcal{N}(0, \alpha^2)$. Then, for any $\delta \in [0, 1]$, if α is sufficiently small, Claim A.3 holds with probability at least $1 - \delta$.*

Proof of Lemma A.4. Using Lemma A.19, we have $\sigma_1(\mathbf{\Theta}^0) \leq \sqrt{\lambda_1^*}$ and $\sigma_1(\mathbf{\Theta}_{\text{res}}^0) \leq \sqrt{\lambda_k^* - \Delta/2}$ hold with probability at least $1 - 2\exp(-(\frac{1}{\alpha^2}\sqrt{\lambda_k^* - \Delta/2} - 2\sqrt{d})^2/2)$, and $\sigma_k(\mathbf{\Theta}_k^0) \leq \sqrt{\Delta}/2$ holds with probability at least $1 - 2\exp(-(\sqrt{\Delta}/(2\alpha^2) - 2\sqrt{d})^2/2)$. Then for α small enough such that

$$\alpha \leq \min\left\{ \frac{\sqrt{\Delta}}{4\sqrt{d} + 2\sqrt{2\log(2/\delta')}}, \frac{\sqrt{\lambda_1^* - \Delta/2}}{2\sqrt{d} + \sqrt{2\log(2/\delta')}} \right\},$$

$\sigma_1(\mathbf{\Theta}^0) \leq \sqrt{\lambda_1^*}$, $\sigma_1(\mathbf{\Theta}_{\text{res}}^0) \leq \sqrt{\lambda_k^* - \Delta/2}$ and $\sigma_k(\mathbf{\Theta}_k^0) \leq \sqrt{\Delta}/2$ hold with probability at least $1 - 2\delta'$ for any $\delta' \in [0, 1]$.

From Rudelson & Vershynin (2008), there exists a constant K that only depends on δ' such that with probability at least $1 - \delta'$, we have $\sigma_k(\mathbf{\Theta}_k^0) \geq \alpha^2 K\sqrt{k}$.

Thus, when α is sufficiently small such that $\alpha^{4-2(1+\kappa)} \leq \frac{c_1(K\sqrt{k})^{1+\kappa}}{8d + 4\log(2/\delta')}$, with probability at least $1 - \delta'$, we have

$$c_1\sigma_k(\mathbf{\Theta}_k^0)^{1+\kappa} \geq c_1(\alpha^2 K\sqrt{k})^{1+\kappa} \geq \alpha^4(2\sqrt{d} + \sqrt{2\log(2/\delta')})^2.$$

Note that from Lemma A.19, with probability at least $1 - \delta'$ we have

$$\sigma_1^2(\mathbf{\Theta}_{\text{res}}^0) \leq \alpha^4(2\sqrt{d} + \sqrt{2\log(2/\delta')})^2.$$

Then we conclude that with probability at least $1 - 4\delta'$, we have $\sigma_1(\mathbf{\Theta}^0) \leq \sqrt{\lambda_1^*}$, $\sigma_1(\mathbf{\Theta}_{\text{res}}^0) \leq \sqrt{\lambda_1^* - \Delta/2}$, $\sigma_k(\mathbf{\Theta}_k^0) \leq \sqrt{\Delta}/2$ and $\sigma_1^2(\mathbf{\Theta}_{\text{res}}^0) \leq c_1 \cdot \sigma_k(\mathbf{\Theta}_k^0)^{1+\kappa}$. Finally the lemma follows by setting $\delta = 4\delta'$. □

Next, we introduce the following lemma, which shows that when Claim A.3 holds, $\mathbf{\Theta}^t$ will enter the region $\mathcal{R}$ in a short time period.

Lemma A.5. *Assume $\eta \leq \frac{1}{6\lambda_1^*}$ and Claim A.3 holds. Then, if*

$$\sqrt{N} \geq \max\left\{ \frac{\sqrt{d - \log\delta}}{\sqrt{c_1}}, \frac{3456\sqrt{d - \log\delta}\sqrt{\lambda_1^*}(\sqrt{k(\lambda_1^*)^2 + E} + \sqrt{k}\sigma_\xi)}{\min\{\sigma_1(\mathbf{\Theta}_{res}^0), \sigma_1(\mathbf{\Theta}_{res}^0)\}\Delta\sqrt{c_1}} \right\}, \tag{17}$$

with probability at least $1 - ct\delta$, we have $\sigma_k(\mathbf{\Theta}_k^t) \geq \sqrt{\Delta}/2$ for some $t \in [0, \frac{\log(\Delta/4\sigma_k^2(\mathbf{\Theta}_k^0))}{2\log(1+\frac{\eta}{2}(\lambda_k^* - \Delta/2))}]$.

Proof of Lemma A.5. With Claim A.3 holds, we have $\sigma_1^2(\mathbf{\Theta}^0) \leq 2\lambda_1^*$, $\sigma_1^2(\mathbf{\Theta}_{\text{res}}^0) \leq \lambda_k^* - \Delta/2$. Then, based on Lemma A.17, for N satisfying inequality (17), we have

$$\sigma_1(\mathbf{\Theta}_{\text{res}}^t) \leq \Big(1 + \frac{\eta}{2}\lambda_{k+1}^* + \frac{\eta}{8}\Delta\Big)^t \sigma_1(\mathbf{\Theta}_{\text{res}}^0),$$

holds with probability at least $1 - ct\delta$. Combining with Claim A.3, we obtain

$$\Big(1 + \frac{\eta}{2}\lambda_{k+1}^* + \frac{\eta}{8}\Delta\Big)^{T_{\mathcal{R}}} \sigma_1^2(\mathbf{\Theta}_{\text{res}}^0) = \Big(\frac{\Delta}{4\sigma_k^2(\mathbf{\Theta}_k^0)}\Big)^{\kappa/2} \sigma_1^2(\mathbf{\Theta}_{\text{res}}^0) \leq \frac{\Delta}{8\sqrt{\lambda_1^*}}\sigma_k(\mathbf{\Theta}_k^0).$$

Then, for all $t \leq T_{\mathcal{R}}$, we have

$$\Big(1 + \frac{\eta}{2}\lambda_{k+1}^* + \frac{\eta}{8}\Delta\Big)^{2t} \sigma_1^2(\mathbf{\Theta}_{\text{res}}^0) \leq \Big(1 + \frac{\eta}{2}\lambda_{k+1}^* + \frac{\eta}{8}\Delta\Big)^{t+T_{\mathcal{R}}} \sigma_1^2(\mathbf{\Theta}_{\text{res}}^0) \leq \Big(1 + \frac{\eta}{2}\lambda_k^* - \frac{\eta}{4}\Delta\Big)^t \frac{\Delta}{8\sqrt{\lambda_1^*}}\sigma_k(\mathbf{\Theta}_k^0).$$

Let $T' = \min\{t > 0 | \sigma_k^2(\mathbf{\Theta}_k^t) \geq \Delta/4\}$. We then aim to prove that

$$\sigma_k(\mathbf{\Theta}_k^t) \geq \Big(1 + \frac{\eta}{2}\lambda_k^* - \frac{\eta}{4}\Delta\Big)^t \sigma_k(\mathbf{\Theta}_k^0), \quad \forall t \leq \min\{T_{\mathcal{R}}, T'\}. \tag{18}$$

We prove it by induction.

Assume Equation (18) holds for some $\tau \leq t$, where $t \leq \min\{T_{\mathcal{R}}, T'\}$. Then we have

$$\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau) \leq \Big(1 + \frac{\eta}{2}\lambda_{k+1}^* + \frac{\eta}{8}\Delta\Big)^{2\tau} \sigma_1^2(\mathbf{\Theta}_{\text{res}}^0) \leq \frac{\Delta}{8\sqrt{\lambda_1^*}}\Big(1 + \frac{\eta}{2}\lambda_{k+1}^* - \frac{\eta}{4}\Delta\Big)^\tau \sigma_k(\mathbf{\Theta}_k^0) \leq \frac{\Delta}{8\sqrt{\lambda_1^*}}\sigma_k(\mathbf{\Theta}_k^\tau).$$

We consider the next time step $\tau + 1$. Note that $\sigma_k(\mathbf{\Theta}_k^{\tau+1})$ can be lower bounded as

$$\begin{aligned}
\sigma_k(\mathbf{\Theta}_k^{\tau+1}) &\geq \sigma_k(\mathbf{\Theta}_k^\tau + \frac{\eta}{2}\tilde{\mathbf{\Lambda}}_k\mathbf{\Theta}_k^\tau - \frac{\eta}{2}\mathbf{\Theta}_k^\tau(\mathbf{\Theta}^\tau)^\top\mathbf{\Theta}^\tau) - \sigma_1(\mathbf{R}_k^{\tau+1}) \\
&\geq \sigma_k(\mathbf{\Theta}_k^\tau + \frac{\eta}{2}\tilde{\mathbf{\Lambda}}_k\mathbf{\Theta}_k^\tau - \frac{\eta}{2}\mathbf{\Theta}_k^\tau(\mathbf{\Theta}_k^\tau)^\top\mathbf{\Theta}_k^\tau) - \frac{\eta}{2}\sigma_1(\mathbf{\Theta}_k^\tau(\mathbf{\Theta}_{\text{res}}^\tau)^\top\mathbf{\Theta}_{\text{res}}^\tau) - \sigma_1(\mathbf{R}^{\tau+1}).
\end{aligned}$$

Applying Lemma D.4 in Jiang et al. (2022) gives

$$\begin{aligned}
&\sigma_k(\mathbf{\Theta}_k^\tau + \frac{\eta}{2}\tilde{\mathbf{\Lambda}}_k\mathbf{\Theta}_k^\tau - \frac{\eta}{2}\mathbf{\Theta}_k^\tau(\mathbf{\Theta}_k^\tau)^\top\mathbf{\Theta}_k^\tau) - \frac{\eta}{2}\sigma_1(\mathbf{\Theta}_k^\tau(\mathbf{\Theta}_{\text{res}}^\tau)^\top\mathbf{\Theta}_{\text{res}}^\tau) \\
&\geq \big(1 - \frac{(\eta)^2}{2}\sigma_1(\tilde{\mathbf{\Lambda}}_k(\mathbf{\Theta}_k^\tau)^\top\mathbf{\Theta}_k^\tau)\big)(1 + \frac{\eta}{2}\lambda_k^*)\sigma_k(\mathbf{\Theta}_k^\tau)\big(1 - \frac{\eta}{2}\sigma_k^2(\mathbf{\Theta}_k^\tau)\big) - \frac{\eta}{2}\sigma_1(\mathbf{\Theta}_k^\tau(\mathbf{\Theta}_{\text{res}}^\tau)^\top\mathbf{\Theta}_{\text{res}}^\tau) \\
&\geq \Big(1 - \frac{(\eta)^2}{2}\lambda_1^{*2}\Big)\Big(1 + \frac{\eta\lambda_k^*}{2}\Big)\Big(1 - \frac{\eta\Delta}{4}\Big)\sigma_k(\mathbf{\Theta}_k^\tau) - \frac{\eta\Delta}{8}\sigma_k(\mathbf{\Theta}_k^\tau).
\end{aligned}$$

Then, for $\eta \leq \frac{\Delta^2}{18\lambda_1^{*3}}$, we have

$$\sigma_k(\mathbf{\Theta}_k^\tau + \frac{\eta}{2}\tilde{\mathbf{\Lambda}}_k\mathbf{\Theta}_k^\tau - \frac{\eta}{2}\mathbf{\Theta}_k^\tau(\mathbf{\Theta}_k^\tau)^\top\mathbf{\Theta}_k^\tau) - \frac{\eta}{2}\sigma_1(\mathbf{\Theta}_k^\tau(\mathbf{\Theta}_{\text{res}}^\tau)^\top\mathbf{\Theta}_{\text{res}}^\tau) \geq \Big(1 + \frac{\eta\lambda_k^*}{2} - \eta\Delta\frac{71}{288}\Big)\sigma_k(\mathbf{\Theta}_k^\tau). \tag{19}$$

According to Lemma A.13 and Lemma A.14, if

$$\sqrt{N} \geq \max\Big\{\frac{\sqrt{d - \log\delta}}{\sqrt{c_1}}, \frac{3456\sqrt{\lambda_1^*}(\sqrt{k(\lambda_1^*)^2 + E} + \sqrt{k}\sigma_\xi)\sqrt{d - \log\delta}}{\sigma_k(\mathbf{\Theta}_k^0)\Delta\sqrt{c_1}}\Big\},$$

then, with probability at least $1 - 2\delta$, we have $\sigma_1(\mathbf{R}^{\tau+1}) \leq \frac{1}{288}\eta\Delta\sigma_k(\mathbf{\Theta}_k^0)$.

Combining with Equation (19) gives

$$\sigma_k(\mathbf{\Theta}_k^{\tau+1}) \geq \Big(1 + \frac{\eta\lambda_k^*}{2} - \eta\Delta\frac{71}{288}\Big)\sigma_k(\mathbf{\Theta}_k^\tau) - \frac{1}{288}\eta\Delta\sigma_k(\mathbf{\Theta}_k^0)$$

$$\geq \Big(1+\frac{\eta\lambda_k^*}{2}-\eta\Delta\frac{71}{288}\Big)\Big(1+\frac{\eta}{2}\lambda_k^*-\frac{\eta}{4}\Delta\Big)^{\tau}\sigma_k(\mathbf{\Theta}_k^0)-\frac{1}{288}\eta\Delta\sigma_k(\mathbf{\Theta}_k^0)$$
$$\geq \Big(1+\frac{\eta\lambda_k^*}{2}-\eta\Delta\frac{71}{288}\Big)\Big(1+\frac{\eta}{2}\lambda_k^*-\frac{\eta}{4}\Delta\Big)^{\tau}\sigma_k(\mathbf{\Theta}_k^0)-\frac{1}{288}\Big(1+\frac{\eta}{2}\lambda_k^*-\frac{\eta}{4}\Delta\Big)^{\tau}\eta\Delta\sigma_k(\mathbf{\Theta}_k^0)$$
$$= \Big(1+\frac{\eta}{2}\lambda_k^*-\frac{\eta}{4}\Delta\Big)^{\tau+1}\sigma_k(\mathbf{\Theta}_k^0).$$

Then, we conclude that with probability at least $1-ct\delta$ for some constant c, we have

$$\sigma_k(\mathbf{\Theta}_k^t)\geq\Big(1+\eta\lambda_k^*/2-\eta\Delta/4\Big)^t\sigma_k(\mathbf{\Theta}_k^0).$$

Here we claim $T_{\mathcal{R}}\leq T'$ always holds, since if $T_{\mathcal{R}}>T'$, we must have

$$\sigma_k(\mathbf{\Theta}_k^{T_{\mathcal{R}}})\geq\Big(1+\eta\lambda_k^*/2-\eta\Delta/4\Big)^{T_{\mathcal{R}}}\sigma_k(\mathbf{\Theta}_k^0)\geq\sqrt{\Delta}/2,$$

which contradicts the definition of T'. The proof is thus complete. □

A.2.2. Step 2: Trapped in the Absorbing Region

We start by introducing Lemma A.6, Lemma A.7 and Lemma A.8.

Lemma A.6. *Assume $\eta\leq\frac{2}{5\lambda_1^*}$ and $\mathbf{\Theta}^0\in\mathcal{R}_s$. Then, if $\sqrt{N}\geq\frac{12\sqrt{2}\sqrt{\bar{d}k-k\log\delta}}{\sqrt{c_1}}$ for constant c_1, with probability at least $1-2t\delta$, we have $\sigma_1(\mathbf{\Theta}^\tau)\leq\sqrt{2\lambda_1^*}$ hold for all $\tau\leq t$.*

Proof of Lemma A.6. Assume that $\mathbf{\Theta}^{\tau-1}\in\mathcal{R}_s$. Then, utilizing Equation (13), we have

$$\sigma_1(\mathbf{\Theta}^\tau)\leq\sigma_1(\mathbf{\Theta}^{\tau-1})\big(1+\frac{\eta}{2}\lambda_1^*-\frac{\eta}{2}\sigma_1^2(\mathbf{\Theta}^{\tau-1})\big)+\sigma_1(\mathbf{R}^\tau).$$

Note that $\sigma_1(\mathbf{\Theta}^{\tau-1})\big(1+\frac{\eta}{2}\lambda_1^*-\frac{\eta}{2}\sigma_1^2(\mathbf{\Theta}^{\tau-1})\big)$ reaches its maximum at $\sigma_1(\mathbf{\Theta}^{\tau-1})=\sqrt{\frac{2+\eta\lambda_1^*}{3\eta}}$. For $\eta\leq\frac{2}{5\lambda_1^*}$, we have $\sqrt{\frac{2+\eta\lambda_1^*}{3\eta}}\geq\sqrt{2\lambda_1^*}$. Thus, $\sigma_1(\mathbf{\Theta}^{\tau-1})\big(1+\frac{\eta}{2}\lambda_1^*-\frac{\eta}{2}\sigma_1^2(\mathbf{\Theta}^{\tau-1})\big)$ is monotonically increasing for $\sigma_1(\mathbf{\Theta}^{\tau-1})\in[0,\sqrt{2\lambda_1^*}]$. Then,

$$\sigma_1(\mathbf{\Theta}^t\tau)\leq\sqrt{2\lambda_1^*}\big(1-\frac{\eta}{2}\lambda_1^*\big)+\sigma_1(\mathbf{R}^\tau).$$

We prove $\sigma_1(\mathbf{\Theta}^\tau)\leq\sqrt{2\lambda_1^*}$ by induction. First, since $\mathbf{\Theta}^0\in\mathcal{R}_s$, we have $\sigma_1(\mathbf{\Theta}^0)\leq\sqrt{2\lambda_1^*}$. Then, assume $\sigma_1(\mathbf{\Theta}^\tau)\leq\sqrt{2\lambda_1^*}$ holds for time step $0\leq\tau<t$. According to Lemma A.13 and Lemma A.14, if $\sigma_1(\mathbf{\Theta}^\tau)\leq\sqrt{2\lambda_1^*}$, and $\sqrt{N}\geq\frac{12\sqrt{2}\sqrt{\bar{d}k-k\log\delta}}{\sqrt{c_1}}$, with probability at least $1-2\delta$, we have $\sigma_1(\mathbf{R}^{\tau+1})\leq\frac{\sqrt{2}}{2}\eta(\lambda_1^*)^{\frac{3}{2}}$. Thus,

$$\sigma_1(\mathbf{\Theta}^{\tau+1})\leq\sqrt{2\lambda_1^*}\big(1-\frac{\eta}{2}\lambda_1^*\big)+\sigma_1(\mathbf{R}^{\tau+1})\leq\sqrt{2\lambda_1^*}.$$

Then, by induction, with probability at least $1-2t\delta$, we have $\sigma_1(\mathbf{\Theta}^\tau)\leq\sqrt{2\lambda_1^*}$ for all $\tau\leq t$. Then the proof is complete. □

Lemma A.7. *Assume $\eta\leq\frac{1}{6\lambda_1^*}$ and $\mathbf{\Theta}^0\in\mathcal{R}_s$. Then, if*

$$\sqrt{N}\geq\max\left\{\frac{\sqrt{\bar{d}-\log\delta}}{\sqrt{c_1}},\frac{48\sqrt{\lambda_1^*}(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)\sqrt{\bar{d}-\log\delta}}{\Delta\sqrt{c_1(\lambda_k^*-\Delta/2)}}\right\}$$

for constant c_1, with probability at least $1-2t\delta$, we have $\sigma_1(\mathbf{\Theta}_{res}^\tau)\leq\sqrt{\lambda_k^*-\Delta/2}$ holds for all $\tau\leq t$.

Proof of Lemma A.7. We prove it by induction. Note that $\sigma_1(\mathbf{\Theta}^0_{\text{res}}) \leq \sqrt{\lambda^*_k - \Delta/2}$ since $\mathbf{\Theta}^0 \in \mathcal{R}_s$. Assume $\sigma_1(\mathbf{\Theta}^{\tau-1}_{\text{res}}) \leq \sqrt{\lambda^*_k - \Delta/2}$ holds for some τ. We aim to show that the inequality holds for $\sigma_1(\mathbf{\Theta}^{\tau}_{\text{res}})$ as well.

Based on Lemma A.15, for $\eta \leq 1/6\lambda^*_1$ and $\sigma_1(\mathbf{\Theta}^\tau) \leq \sqrt{2\lambda^*_1}$, we have

$$\begin{aligned}\sigma_1(\mathbf{\Theta}^{\tau}_{\text{res}}) &\leq \Big(1 + \frac{\eta}{2}\big(\lambda^*_{k+1} - \sigma_1^2(\mathbf{\Theta}^{\tau-1}_{\text{res}}) - \sigma_k^2(\mathbf{\Theta}^{\tau-1}_{k})\big)\Big)\sigma_1(\mathbf{\Theta}^{\tau-1}_{\text{res}}) + \sigma_1(\mathbf{R}^{\tau}_{2\bar{d}-k}) \\ &\leq \Big(1 + \frac{\eta}{2}\big(\lambda^*_{k+1} - \sigma_1^2(\mathbf{\Theta}^{\tau-1}_{\text{res}})\big)\Big)\sigma_1(\mathbf{\Theta}^{\tau-1}_{\text{res}}) + \sigma_1(\mathbf{R}^{\tau}_{2\bar{d}-k}).\end{aligned}$$

Note that when $\sigma_1(\mathbf{\Theta}^{\tau-1}_{\text{res}}) \geq 0$, $\Big(1 + \frac{\eta}{2}\big(\lambda^*_{k+1} - \sigma_1^2(\mathbf{\Theta}^{\tau-1}_{\text{res}})\big)\Big)\sigma_1(\mathbf{\Theta}^{\tau-1}_{\text{res}})$ is maximized at $\sigma_1(\mathbf{\Theta}^{\tau-1}_{\text{res}}) = \sqrt{\frac{2+\eta\lambda^*_{k+1}}{3\eta}}$. Since we assume $\eta \leq \frac{1}{6\lambda^*_1} \leq \frac{2}{3\lambda^*_k + \lambda^*_{k+1}}$, it holds that $\sqrt{\frac{2+\eta\lambda^*_{k+1}}{3\eta}} \geq \sqrt{\lambda^*_k - \Delta/2}$. Then, $\Big(1 + \frac{\eta}{2}\big(\lambda^*_{k+1} - \sigma_1^2(\mathbf{\Theta}^{\tau-1}_{\text{res}})\big)\Big)\sigma_1(\mathbf{\Theta}^{\tau-1}_{\text{res}})$ is monotonically increasing for $0 \leq \sigma_1(\mathbf{\Theta}^{\tau-1}_{\text{res}}) \leq \sqrt{\lambda^*_k - \Delta/2}$. We thus have

$$\sigma_1(\mathbf{\Theta}^{\tau}_{\text{res}}) \leq \sqrt{\lambda^*_k - \Delta/2}\big(1 - \frac{\eta\Delta}{4}\big) + \sigma_1(\mathbf{R}^\tau). \tag{20}$$

According to Lemma A.13 and Lemma A.14, if $\sqrt{N} \geq \max\big\{\frac{\sqrt{\bar{d} - \log\delta}}{\sqrt{c_1}}, \frac{48\sqrt{\lambda^*_1}(\sqrt{k(\lambda^*_1)^+ E} + \sqrt{k}\sigma_\xi)\sqrt{\bar{d} - \log\delta}}{\Delta\sqrt{c_1(\lambda^*_k - \Delta/2)}}\big\}$, with probability at least $1 - 2\delta$, we have

$$\sigma_1(\mathbf{R}^\tau) \leq \frac{\Delta\eta\sqrt{\lambda^*_k - \Delta/2}}{4}. \tag{21}$$

Then by combining (20) and (21), we have $\sigma_1(\mathbf{\Theta}^{\tau}_{\text{res}}) \leq \sqrt{\lambda^*_k - \Delta/2}$. The proof is thus complete. □

Lemma A.8. *Assume $\eta \leq \frac{\Delta^2}{32\lambda^{*3}_1}$, $\sigma_k(\mathbf{\Theta}^0_k) \geq \sqrt{\Delta}/2$ and $\mathbf{\Theta}^\tau \in \mathcal{R}_s$ for all $0 < \tau \leq t$. Then, if*

$$\sqrt{N} \geq \max\left\{\frac{\sqrt{\bar{d} - \log\delta}}{\sqrt{c_1}}, \quad \frac{6144\sqrt{\lambda^*_1}(\sqrt{k(\lambda^*_1)^2 + E} + \sqrt{k}\sigma_\xi)\sqrt{\bar{d} - \log\delta}}{\Delta\sqrt{c_1}}\right\},$$

for some contact c_1, with probability at least $1 - 2t\delta$, we have $\sigma_k(\mathbf{\Theta}^\tau_k) \geq \sqrt{\Delta}/2$ hold for all $0 < \tau \leq t$.

Proof of Lemma A.8. We prove it by induction. First note that $\sigma_k(\mathbf{\Theta}^0_k) \geq \sqrt{\Delta}/2$ under the assumption of Lemma A.8. Assume that $\sigma_k(\mathbf{\Theta}^\tau_k) \geq \sqrt{\Delta}/2$ holds for some t. We then show that $\sigma_k(\mathbf{\Theta}^{\tau+1}_k) \geq \sqrt{\Delta}/2$ holds as well.

Since for all $\tau \leq t$ it holds that $\mathbf{\Theta}^\tau \in \mathcal{R}_s$ and N satisfies the condition described in Lemma A.8, based on Lemmas A.13 and A.14, with probability at least $1 - 2t\delta$, it holds that $\sigma_1(\mathbf{R}^\tau_k) \leq \frac{\eta\Delta^2}{512\sqrt{\lambda^*_1}}$ for all $0 \leq \tau \leq t$. From the intermediate result of Lemma A.16, for $\eta \leq \frac{\Delta^2}{16\lambda^{*3}_1}$, we have

$$\sigma_k^2(\mathbf{\Theta}^{\tau+1}_k) \geq \Big(1 + \eta\big(\lambda^*_k - \sigma_1^2(\mathbf{\Theta}^\tau_{\text{res}}) - \sigma_k^2(\mathbf{\Theta}^\tau_k)\big)\Big)\sigma_k^2(\mathbf{\Theta}^\tau_k) - \eta^2\lambda^{*3}_1 - 4\sqrt{\lambda^*_1}\sigma_1(\mathbf{R}^\tau_k).$$

Combining with the fact that

$$\sigma_k^2(\mathbf{\Theta}^{\tau+1}_k) \geq \Big(1 + \eta\big(\lambda^*_k - \sigma_1^2(\mathbf{\Theta}^\tau_{\text{res}}) - \sigma_k^2(\mathbf{\Theta}^\tau_k)\big)\Big)\sigma_k^2(\mathbf{\Theta}^\tau_k) - \eta^2\lambda^{*3}_1 - \frac{\eta\Delta^2}{128},$$

we have

$$\begin{aligned}\sigma_k^2(\mathbf{\Theta}^{\tau+1}_k) &\geq \Big(1 + \eta\big(\Delta/2 - \sigma_k^2(\mathbf{\Theta}^\tau_k)\big)\Big)\sigma_k^2(\mathbf{\Theta}^\tau_k) - \eta^2\lambda^{*3}_1 - \frac{\eta\Delta^2}{128} \\ &\geq (1 + \eta(\Delta/2 - \Delta/4))\Delta/4 - \eta^2\lambda^{*3}_1 - \frac{\eta\Delta^2}{128} \\ &\geq \frac{\Delta}{4} + \eta\frac{\Delta^2}{16} - \eta\frac{\Delta^2}{32} - \frac{\eta\Delta^2}{128} \\ &\geq \frac{\Delta}{4}.\end{aligned}$$

The proof is thus complete. □

Combining Lemmas A.6 and A.7, we conclude that for N sufficiently large, $\mathcal{R}_s$ is an absorbing region with high probability, i.e., starting from $\mathbf{\Theta}^0 \in \mathcal{R}_s$, the subsequent $\mathbf{\Theta}^t$ will stay in $\mathcal{R}_s$ for all $t > 0$ with high probability, which is summarized in the following proposition.

Proposition A.9. *Assume $\mathbf{\Theta}^0 \in \mathcal{R}$. If $\sqrt{N} \geq c\frac{\sqrt{\lambda_1^*}(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma)\sqrt{\bar{d}-\log\delta}}{\Delta^{\frac{3}{2}}}$ for some constant c, then, with probability at least $1-t\delta$, we have $\mathbf{\Theta}^\tau \in \mathcal{R}$ for all $0 \leq \tau \leq t$.*

A.2.3. Step 3: Local Convergence of $\mathbf{\Theta}^t(\mathbf{\Theta}^t)^\top$

We next show that when N is sufficiently large, with high probability, $\|\mathbf{\Theta}^t(\mathbf{\Theta}^t)^\top - \text{diag}(\tilde{\mathbf{\Lambda}}_k, \mathbf{0})\|$ converges to 0 exponentially fast when $\mathbf{\Theta}^0 \in \mathcal{R}$.

Firstly, we establish the following lemma that lower bounds the number of samples needed for the inverse SNR to converge exponentially fast with high probability.

Lemma A.10. *Denote $\sigma_{ref}^{t+1} = \sqrt{\lambda_1^*}\Big(1-\frac{\eta\Delta}{16}\Big)^{t+1}$. Assume $\eta \leq \Delta^2/(36\lambda_1^{*3})$, $\mathbf{\Theta}^\tau \in \mathcal{R}$ for all $0 \leq \tau \leq t$, and*

$$\sqrt{N} \geq c \cdot \max\left\{\sqrt{\bar{d}-\log\delta}, \frac{\sqrt{\bar{d}-\log\delta}\sqrt{\lambda_1^*}(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\sigma_{ref}^{t+1}\Delta}, \frac{\sqrt{\bar{d}-\log\delta}\lambda_1^*(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma)}{\Delta^2}\right\}$$

for some constant c. Then, with probability at least $1-(t+1)\delta$, we have

$$\sigma_1(\mathbf{\Theta}_{res}^{t+1}) \leq \frac{2\sqrt{2}\lambda_1^*}{\sqrt{\Delta}}\Big(1-\frac{\eta\Delta}{16}\Big)^{t+1}.$$

Proof of Lemma A.10. Based on Lemma A.20, we have $\sigma_1(\mathbf{R}_{2\bar{d}-k}^{\tau+1}) \leq \sigma_1(\mathbf{R}^{\tau+1})$ and $\sigma_1(\mathbf{R}_k^{\tau+1}) \leq \sigma_1(\mathbf{R}^{\tau+1})$. Then, it follows that

$$\sigma_1(\mathbf{R}_{2\bar{d}-k}^{\tau+1}) \leq \sqrt{\sigma_1^2((\mathbf{Q}_\star^{\tau+1})^\top\mathbf{U}_\star)+\sigma_1^2(\tilde{\mathbf{Q}}_\star^{\tau+1}\mathbf{V}_\star)} \quad \text{and} \quad \sigma_1(\mathbf{R}_k^{\tau+1}) \leq \sqrt{\sigma_1^2((\mathbf{Q}_\star^{\tau+1})^\top\mathbf{U}_\star)+\sigma_1^2(\tilde{\mathbf{Q}}_\star^{\tau+1}\mathbf{V}_\star)}.$$

Substitute the σ in Lemmas A.13 and A.14 by $\sigma_{\text{ref}}^{t+1}$. Then, if

$$\sqrt{N} \geq \max\left\{\frac{\sqrt{\bar{d}-\log\delta}}{\sqrt{c_1}}, \frac{192\sqrt{\bar{d}-\log\delta}\sqrt{\lambda_1^*}(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\sigma_{\text{ref}}^{t+1}\Delta\sqrt{c_1}}, \frac{6144\sqrt{\bar{d}-\log\delta}\lambda_1^*(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\Delta^2\sqrt{c_1}}\right\},$$

with probability at least $1-2\delta$, we have

$$\sigma_1(\mathbf{R}_{2d-k}^\tau) \leq \frac{\sigma_{\text{ref}}^{t+1}\eta\Delta}{16} \quad \text{and} \quad \sigma_1(\mathbf{R}_k^\tau) \leq \frac{\eta\Delta^2}{512\sqrt{\lambda_1^*}}. \tag{22}$$

We prove the lemma by considering two cases. In the first case, we assume $\sigma_1(\mathbf{\Theta}_{\text{res}}^\tau) \geq \sigma_{\text{ref}}^{t+1}$ for all $0 \leq \tau \leq t$. In the second case, we assume there exists at least one time step in $[0,t]$ such that $\sigma_1(\mathbf{\Theta}_{\text{res}}^\tau) < \sigma_{\text{ref}}^{t+1}$, and we denote the last time step satisfying this condition as t'.

We start from the first case. Combining Equation (22) with Lemma A.15 gives

$$\begin{aligned}\sigma_1(\mathbf{\Theta}_{\text{res}}^{\tau+1}) &\leq \Big(1+\frac{\eta}{2}\big(\lambda_{k+1}^*-\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)-\sigma_k^2(\mathbf{\Theta}_k^\tau)\big)\Big)\sigma_1(\mathbf{\Theta}_{\text{res}}^\tau)+\frac{\sigma_1(\mathbf{\Theta}_{\text{res}}^\tau)\eta\Delta}{16} \\ &= \Big(1+\frac{\eta}{2}\big(\lambda_{k+1}^*+\Delta/8-\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)-\sigma_k^2(\mathbf{\Theta}_k^\tau)\big)\Big)\sigma_1(\mathbf{\Theta}_{\text{res}}^\tau),\end{aligned} \tag{23}$$

where in Equation (23) we use the assumption that $\sigma_{\text{ref}}^{t+1} \leq \sigma_1(\mathbf{\Theta}_{\text{res}}^\tau)$.

Then, using the fact that $\sigma_1^2(\mathbf{\Theta}^\tau) \leq 2\lambda_1^*$ and $\eta \leq \frac{\Delta}{16\lambda_1^{*2}}$ we obtain

$$\sigma_1^2(\mathbf{\Theta}_{\text{res}}^{\tau+1}) \leq \Big(1+\eta\big(\lambda_{k+1}^*+\Delta/8-\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)-\sigma_k^2(\mathbf{\Theta}_k^\tau)\big)+4\eta^2\lambda_1^{*2}\Big)\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau) \tag{24}$$

$$\leq \Big(1+\eta\big(\lambda_{k+1}^*+\Delta/8-\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)-\sigma_k^2(\mathbf{\Theta}_k^\tau)\big)+\frac{\eta\Delta}{4}\Big)\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)$$
$$\leq \Big(1-\eta\Delta/8+\eta\big(\lambda_k^*-\Delta/2-\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)-\sigma_k^2(\mathbf{\Theta}_k^\tau)\big)\Big)\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau), \tag{25}$$

where Equation (24) holds since $\big(\lambda_{k+1}^*+\Delta/8-\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)-\sigma_k^2(\mathbf{\Theta}_k^\tau)\big)^2\leq 16\lambda_1^{*2}$.

Next, combining Lemma A.16 and Equation (22) leads to

$$\begin{aligned}
\sigma_k^2(\mathbf{\Theta}_k^{\tau+1}) &\geq \sigma_k^2(\tilde{\mathbf{\Theta}}_k^{\tau+1})-4\sqrt{\lambda_1^*}\sigma_1(\mathbf{R}_k^{\tau+1})\\
&\geq \sigma_k^2(\tilde{\mathbf{\Theta}}_k^{\tau+1})-\frac{\eta\Delta^2}{128}\\
&\geq \sigma_k^2(\tilde{\mathbf{\Theta}}_k^{\tau+1})-\frac{\eta\Delta}{8}\sigma_k^2(\mathbf{\Theta}_k^\tau) \qquad (26)\\
&\geq \Big(1+\eta\big(\lambda_k^*-\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)-\sigma_k^2(\mathbf{\Theta}_k^\tau)\big)-\frac{\eta\Delta}{4}\Big)\sigma_k^2(\mathbf{\Theta}_k^\tau)-\frac{\eta\Delta}{8}\sigma_k^2(\mathbf{\Theta}_k^\tau)\\
&= \Big(1+\eta\Delta/8+\eta\big(\lambda_k^*-\Delta/2-\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)-\sigma_k^2(\mathbf{\Theta}_k^\tau)\big)\Big)\sigma_k^2(\mathbf{\Theta}_k^\tau), \qquad (27)
\end{aligned}$$

where in Equation (26) we use the fact $\sigma_k(\mathbf{\Theta}_k^\tau)\geq\frac{\Delta}{4}$.

Then, combining Equation (25) with Equation (27) we have

$$\begin{aligned}
\frac{\sigma_1^2(\mathbf{\Theta}_{\text{res}}^{\tau+1})}{\sigma_k^2(\mathbf{\Theta}_k^{\tau+1})} &\leq \frac{\Big(1-\eta\Delta/8+\eta\big(\lambda_k^*-\Delta/2-\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)-\sigma_k^2(\mathbf{\Theta}_k^\tau)\big)\Big)\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)}{\Big(1+\eta\Delta/8+\eta\big(\lambda_k^*-\Delta/2-\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)-\sigma_k^2(\mathbf{\Theta}_k^\tau)\big)\Big)\sigma_k^2(\mathbf{\Theta}_k^\tau)}\\
&\leq \frac{3/2-\eta\Delta/8}{3/2+\eta\Delta/8}\cdot\frac{\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)}{\sigma_k^2(\mathbf{\Theta}_k^\tau)} \qquad (28)\\
&\leq \Big(1-\frac{\eta\Delta}{6}\Big)\frac{\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)}{\sigma_k^2(\mathbf{\Theta}_k^\tau)}\\
&\leq \Big(1-\frac{\eta\Delta}{16}\Big)^2\frac{\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)}{\sigma_k^2(\mathbf{\Theta}_k^\tau)}, \qquad (29)
\end{aligned}$$

where Equation (28) holds when $-1/2\leq\eta\big(\lambda_k^*-\Delta/2-\sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau)-\sigma_k^2(\mathbf{\Theta}_k^\tau)\big)\leq 1/2$, which is valid when $\eta\leq\Delta^2/(36\lambda_1^{*3})$, and Equation (29) holds since $(1-\eta\Delta/6)\leq(1-\eta\Delta/16)$ is valid for positive η.

Then, with probability at least $1-2(t+1)\delta$, we have

$$\sigma_1^2(\mathbf{\Theta}_{\text{res}}^{t+1})\leq\Big(1-\frac{\eta\Delta}{16}\Big)^{2(t+1)}\frac{\sigma_1^2(\mathbf{\Theta}_{\text{res}}^0)}{\sigma_k^2(\mathbf{\Theta}_k^0)}\sigma_k^2(\mathbf{\Theta}_k^{t+1})\leq\frac{8\lambda_1^{*2}}{\Delta}\Big(1-\frac{\eta\Delta}{16}\Big)^{2(t+1)}.$$

For the second case, note at time step t' we have $\sigma_1(\mathbf{\Theta}_{\text{res}}^{t'})<\sigma_{\text{ref}}^{t+1}$, and for all $t'<\tau\leq t$ we have $\sigma_1(\mathbf{\Theta}_{\text{res}}^\tau)\geq\sigma_{\text{ref}}^{t+1}$. Similar to the previous analysis, we show that with probability at least $1-2(t+1-t')\delta$, we have

$$\begin{aligned}
\sigma_1^2(\mathbf{\Theta}_{\text{res}}^{t+1}) &\leq \Big(1-\frac{\eta\Delta}{6}\Big)^{t+1-t'}\frac{\big(\sigma_{\text{ref}}^{t+1}\big)^2}{\sigma_k^2(\mathbf{\Theta}_k^{t'})}\sigma_k^2(\mathbf{\Theta}_k^{t+1})\\
&\leq \frac{8\lambda_1^*}{\Delta}\big(\sigma_{\text{ref}}^{t+1}\big)^2\Big(1-\frac{\eta\Delta}{16}\Big)^{2(t+1-t')}\\
&\leq \frac{8\lambda_1^{*2}}{\Delta}\Big(1-\frac{\eta\Delta}{16}\Big)^{2(2t+2-t')}\\
&\leq \frac{8\lambda_1^{*2}}{\Delta}\Big(1-\frac{\eta\Delta}{16}\Big)^{2(t+1)}.
\end{aligned}$$

The proof is complete by combining the two cases. □

The following lemma characterizes the number of samples needed for $\mathbf{\Theta}_k$ to converge to $\tilde{\mathbf{\Lambda}}_k$, which is based on the convergence of the inverse SNR.

Lemma A.11. *Assume $\eta \leq \Delta^2/(36\lambda_1^{*3})$, $\mathbf{\Theta}^t \in \mathcal{R}$ for all $0 \leq \tau \leq t$ and N satisfies $\sqrt{N} \geq c \cdot \max\left\{\sqrt{d-\log\delta}, \frac{\sqrt{d-\log\delta}\lambda_1^*(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\sigma_D^{t+1}\Delta}\right\}$ for some constant c. Then, with probability at least $1-(t+1)\delta$, we have*

$$\sigma_1(\mathbf{D}^{t+1}) \leq \frac{200\lambda_1^{*2}}{\eta\Delta^2}\Big(1-\frac{\eta\Delta}{16}\Big)^{t+1},$$

where $\sigma_D^{t+1} = \min\Big\{3\lambda_1^*, (1-\frac{\eta\Delta}{16})^{t+1}\frac{\lambda_1^{*2}}{\eta\Delta^2}\Big\}$.

Proof of Lemma A.11. We denote $\mathbf{D}^\tau = \mathbf{\Theta}_k^\tau(\mathbf{\Theta}_k^\tau)^\top - \tilde{\mathbf{\Lambda}}_k$. For $\tilde{\mathbf{\Theta}}_k^\tau$ defined in Lemma A.16, we have

$$\mathbf{D}^\tau = \tilde{\mathbf{\Theta}}_k^\tau(\tilde{\mathbf{\Theta}}_k^\tau)^\top - \tilde{\mathbf{\Lambda}}_k + \tilde{\mathbf{\Theta}}_k^\tau(\mathbf{R}_k^\tau)^\top + \mathbf{R}_k^\tau(\tilde{\mathbf{\Theta}}_k^\tau)^T + \mathbf{R}_k^\tau(\mathbf{R}_k^\tau)^\top. \tag{30}$$

Let $\sigma_D^{t+1} = \min\Big\{3\lambda_1^*, (1-\frac{\eta\Delta}{16})^{t+1}\frac{\lambda_1^{*2}}{\eta\Delta^2}\Big\}$. Then, if

$$\sqrt{N} \geq \max\left\{\frac{\sqrt{d-\log\delta}}{\sqrt{c_1}}, \frac{1152\sqrt{d-\log\delta}\lambda_1^*(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\sigma_D^{t+1}\Delta\sqrt{c_1}}\right\}, \tag{31}$$

we have $\|\mathbf{R}_k^\tau\| \leq \frac{\sigma_D^{t+1}\eta\Delta}{96\sqrt{\lambda_1^*}}$. It follows that

$$\|\tilde{\mathbf{\Theta}}_k^\tau(\mathbf{R}_k^\tau)^\top\| \leq \frac{\sigma_D^{t+1}\eta\Delta}{48},$$

and

$$\|\mathbf{R}_k^\tau(\mathbf{R}_k^\tau)^\top\| \leq \Big(\frac{\sigma_D^{t+1}\eta\Delta}{96\sqrt{\lambda_1^*}}\Big)^2 \leq \frac{\sigma_D^{t+1}\eta\Delta}{48}, \tag{32}$$

where Equation (32) holds since $\frac{\sigma_D^{t+1}\eta\Delta}{96\sqrt{\lambda_1^*}} \leq 1$. Then, with probability at least $1-2\delta$, we have

$$\sigma_1\Big(\tilde{\mathbf{\Theta}}_k^\tau(\mathbf{R}_k^\tau)^\top + \mathbf{R}_k^\tau(\tilde{\mathbf{\Theta}}_k^\tau)^\top + \mathbf{R}_k^\tau(\mathbf{R}_k^\tau)^\top\Big) \leq \frac{\eta\Delta}{16}\sigma_D^{t+1}.$$

We prove the lemma by considering two cases: In the first case, we assume $\sigma_1(\mathbf{D}^\tau) \geq \sigma_D^{t+1}$ for all $0 \leq \tau \leq t$; In the second case, we assume there is at least one time step in $[0,t]$ such that $\sigma_1(\mathbf{D}^\tau) < \sigma_D^{t+1}$, and we denote the latest time step satisfies this condition as t'.

We start from the first case. From Section A.3 in Chen et al. (2023), we have

$$\begin{aligned}
\sigma_1(\mathbf{D}^\tau) &\leq (1-\frac{\eta\Delta}{8})\sigma_1(\mathbf{D}^{\tau-1}) + \sigma_1^2(\mathbf{\Theta}_{\text{res}}^{\tau-1}) + \frac{\eta\Delta}{16}\sigma_D^{t+1} \\
&\leq (1-\frac{\eta\Delta}{8})\sigma_1(\mathbf{D}^{\tau-1}) + (1-\frac{\eta\Delta}{16})^{2(\tau-1)}\frac{8\lambda_1^{*2}}{\Delta} + \frac{\eta\Delta}{16}\sigma_1(\mathbf{D}^{\tau-1}) \\
&\leq (1-\frac{\eta\Delta}{16})\sigma_1(\mathbf{D}^{\tau-1}) + (1-\frac{\eta\Delta}{16})^{2(\tau-1)}\frac{8\lambda_1^{*2}}{\Delta}.
\end{aligned}$$

Then, for N satisfying Equation (31), with probability at least $1-2(t+1)\delta$, we have

$$\frac{\sigma_1(\mathbf{D}^{t+1})}{(1-\frac{\eta\Delta}{16})^{t+1}} \leq \sigma_1(\mathbf{D}^0) + \sum_{i=0}^{t}\Big(1-\eta\Delta/16\Big)^i\frac{8\lambda_1^{*2}}{(1-\eta\Delta/16)\Delta}$$

$$\leq \sigma_1(\mathbf{D}^0) + \frac{130\lambda_1^{*2}}{\eta\Delta^2}$$
$$\leq \frac{200\lambda_1^{*2}}{\eta\Delta^2}, \tag{33}$$

where Equation (33) follows from the fact that $\lambda_1^*/\eta\Delta^2 \geq 1$ and $\sigma_1(\mathbf{D}^0) \leq 3\lambda_1^* \leq 3\lambda_1^{*2}/\eta\Delta^2$. Therefore, we conclude that $\sigma_1(\mathbf{D}^{t+1}) \leq (1-\eta\Delta/16)^{t+1} \cdot \frac{200\lambda_1^{*2}}{\eta\Delta^2}$.

For the second case, note at time step t' we have $\sigma_1(\mathbf{D}^{t'}) < \sigma_D^{t+1}$, and for all $t' < \tau \leq t$ we have $\sigma_1(\mathbf{D}^{\tau}) \geq \sigma_D^{t+1}$. Similar to the previous analysis, we show that with probability at least $1 - 2(t+1-t')\exp(-c_2(d+k))$, it has

$$\frac{\sigma_1(\mathbf{D}^{t+1})}{(1-\frac{\eta\Delta}{16})^{t+1-t'}} \leq \sigma_1(\mathbf{D}^{t'}) + \sum_{i=0}^{t+1-t'} \Big(1-\eta\Delta/16\Big)^i \frac{8\lambda_1^{*2}}{(1-\eta\Delta/16)\Delta},$$
$$\leq \frac{\lambda_1^{*2}}{\eta\Delta^2}(1-\frac{\eta\Delta}{16})^{t+1} + \frac{130\lambda_1^{*2}}{\eta\Delta^2}$$
$$\leq \frac{200\lambda_1^{*2}}{\eta\Delta^2}.$$

The proof is complete by combining the two cases. □

Then, we aim to show the local convergence property of $\mathbf{\Theta}^t$ stated in the following proposition.

Proposition A.12. *Assume $\mathbf{\Theta}^0 \in \mathcal{R}^0$, $\eta \leq \Delta^2/(36\lambda_1^{*3})$ and N satisfies*

$$\sqrt{N} \geq \max\left\{\frac{\sqrt{d-\log\delta}}{\sqrt{c_1}}, \frac{1152\sqrt{d-\log\delta}(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\kappa^t\Delta\sqrt{c_1}}, \frac{6144\sqrt{d-\log\delta}\lambda_1^*(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\Delta^2\sqrt{c_1}}\right\}, \tag{34}$$

for constant c_1 and $\kappa^t = \big(1-\frac{\eta\Delta}{16}\big)^{t+1}$. Define $\kappa = \frac{\lambda_1^{*2}}{\eta\Delta^2}$. Then, with probability at least $1-ct\delta$, we have

$$\|\mathbf{\Theta}^t(\mathbf{\Theta}^t)^\top - \mathrm{diag}(\tilde{\mathbf{\Lambda}}_k, \mathbf{0})\|_F \leq 400\kappa\sqrt{k}\Big(1-\frac{\eta\Delta}{16}\Big)^t.$$

Proof. First, by combining Lemmas A.6 to A.8, we conclude that if N satisfies (34), $\mathbf{\Theta}^\tau \in \mathcal{R}^\tau$ holds for all $\tau \leq t$ with probability at least $1-ct\delta$.

Then, based on Lemma A.10, if N satisfies Equation (34), with probability at least $1-2t\delta$, we have

$$\sigma_1(\mathbf{\Theta}_{\text{res}}^t) \leq \frac{2\sqrt{2}\lambda_1^*}{\sqrt{\Delta}}\Big(1-\frac{\eta\Delta}{16}\Big)^t. \tag{35}$$

Define $\mathbf{D}^t = \mathbf{\Theta}_k^t(\mathbf{\Theta}_k^t)^\top - \tilde{\mathbf{\Lambda}}_k$. Based on Lemma A.11, we have

$$\sqrt{N} \geq \frac{1152\sqrt{d-\log\delta}(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\kappa^t\Delta\sqrt{c_1}} \geq \frac{1152\sqrt{d-\log\delta}\lambda_1^*(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\sigma_D^{t+1}\Delta\sqrt{c_1}}.$$

Under the same conditions in Equation (34), it holds that

$$\sigma_1(\mathbf{D}^t) \leq \frac{200\lambda_1^{*2}}{\eta\Delta^2}\Big(1-\frac{\eta\Delta}{16}\Big)^t. \tag{36}$$

By combining Equation (35) and Equation (36), we have

$$\|\mathbf{\Theta}^t(\mathbf{\Theta}^t)^\top - \mathrm{diag}(\tilde{\mathbf{\Lambda}}_k, \mathbf{0})\|_F \leq \sqrt{k}\|\mathbf{D}^t\| + 2\sqrt{k\lambda_1^*}\|\mathbf{\Theta}_{\text{res}}^t\| \tag{37}$$

$$\leq 400 \max\Big\{\frac{\sqrt{k}\lambda_1^{*\frac{3}{2}}}{\sqrt{\Delta}}, \frac{\sqrt{k}\lambda_1^{*2}}{\eta\Delta^2}\Big\}\Big(1-\frac{\eta\Delta}{16}\Big)^t.$$

Note that the randomness in $\{\mathbf{\Theta}^t\}_t$ comes from $\{\mathbf{R}^t\}_t$. If N satisfies Equation (34), we have Equation (35) Equation (36) and the event $\mathbf{\Theta}^\tau \in \mathcal{R}^\tau, \forall 0 < \tau \leq t$ holds with probability at least $1 - ct\delta$. Noting that $\max\Big\{\frac{\sqrt{k}\lambda_1^{*\frac{3}{2}}}{\sqrt{\Delta}}, \frac{\sqrt{k}\lambda_1^{*2}}{\eta\Delta^2}\Big\} = \frac{\sqrt{k}\lambda_1^{*2}}{\eta\Delta^2}$, the proof is complete. □

A.2.4. Putting All Together

Combining Propositions A.2, A.9 and A.12, it is straightforward to show that if $t > T_{\mathcal{R}}$, N satisfies $N \geq c_2\frac{(\bar{d}-\log\delta+\log T)(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)^2}{\kappa_T^2\Delta^2}$ for some constant c_2, and $\mathbf{\Theta}^0$ satisfies the small random initialization condition stated in Proposition A.2, then it holds that $\|\mathbf{\Theta}^t(\mathbf{\Theta}^t)^\top - \text{diag}(\tilde{\mathbf{\Lambda}}_k, \mathbf{0})\|_F \leq c_4\kappa\sqrt{k}\Big(1-\frac{\eta\Delta}{16}\Big)^t$ for all t satisfies $T_{\mathcal{R}} < t < T$ with probability at least $1-\delta$, where $\kappa_T = (1-\frac{\eta\Delta}{16})^T$ and constant $c_4 = 400\Big(1-\frac{\eta\Delta}{16}\Big)^{-T_{\mathcal{R}}}$.

Applying Lemma A.18, with probability at least $1-\delta$, we have

$$\|\tilde{\mathbf{B}}^t\tilde{\mathbf{W}}^t - \text{diag}(\mathbf{\Lambda}_k, \mathbf{0})\|_F \leq c_4\kappa\sqrt{k}\Big(1-\frac{\eta\Delta}{16}\Big)^t.$$

Note that $\|\mathbf{B}^t\mathbf{W}^t - \mathbf{B}^*\mathbf{W}^*\|_F = \|\tilde{\mathbf{B}}^t\tilde{\mathbf{W}}^t - \text{diag}(\mathbf{\Lambda}_k, \mathbf{0})\|_F$. Combing with the fact that $\sum_{i\in[M]}\|\mathbf{B}^t w_i^t - \mathbf{B}^* w_i^*\|^2 = \|\mathbf{B}^t\mathbf{W}^t - \mathbf{B}^*\mathbf{W}^*\|_F^2$, we have

$$\sum_{i\in[M]}\|\mathbf{B}^t w_i^t - \mathbf{B}^* w_i^*\|^2 \leq c_4^2\kappa^2 k\Big(1-\frac{\eta\Delta}{16}\Big)^{2t}.$$

Then, applying the Cauchy-Schwarz inequality gives

$$\Big(\sum_{i\in[M]}\|\mathbf{B}^t w_i^t - \mathbf{B}^* w_i^*\|\Big)^2 \leq c_4^2 M\kappa^2 k\Big(1-\frac{\eta\Delta}{16}\Big)^{2t},$$

which immediately implies that

$$\frac{1}{M}\sum_{i\in[M]}\|\mathbf{B}^t w_i^t - \mathbf{B}^* w_i^*\| \leq c_4\kappa\sqrt{\frac{k}{M}}\Big(1-\frac{\eta\Delta}{16}\Big)^t.$$

A.2.5. Auxiliary Lemmas

Lemma A.13 (Concentration of $\|\mathbf{U}_\star^\top\mathbf{Q}_\star^{\tau+1}\|$)**.** *For any $T \geq 0$, assume $\mathbf{\Theta}^\tau \in \mathcal{R}_s$ holds for all $0 < \tau \leq t$. Then, we have the following results for any $0 \leq \tau \leq t$ and $c_2 \geq 0$ with probability at least $1-2\delta$:*

- *If $\sqrt{N} \geq \max\Big\{\frac{\sqrt{\bar{d}-\log\delta}}{\sqrt{c_1}}, \frac{192\sqrt{\bar{d}-\log\delta}\sqrt{\lambda_1^*}(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\sigma\Delta\sqrt{c_1}}\Big\}$, then it holds that*

$$\|\mathbf{U}_\star^\top\mathbf{Q}_\star^{\tau+1}\| \leq \frac{\sigma\eta\Delta}{16\sqrt{2}}. \tag{38}$$

- *If $\sqrt{N} \geq \max\Big\{\frac{\sqrt{\bar{d}-\log\delta}}{\sqrt{c_1}}, \frac{6144\sqrt{\bar{d}-\log\delta}\lambda_1^*(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\Delta^2\sqrt{c_1}}\Big\}$, then it holds that*

$$\|\mathbf{U}_\star^\top\mathbf{Q}_\star^{\tau+1}\| \leq \frac{\eta\Delta^2}{512\sqrt{2}\sqrt{\lambda_1^*}}. \tag{39}$$

Proof of Lemma A.13. Recall that $\mathbf{Q}_\star^{\tau+1}$ is defined as

$$\mathbf{Q}_\star^{\tau+1} = \eta\sum_{i\in[M]}\big(\mathbf{B}_\star^\tau w_i^\tau - \phi_{\star i}\big)(w_i^\tau)^\top - \eta\sum_{i\in[M]}\frac{\mathbf{X}_{\star i}\mathbf{X}_{\star i}^\top}{N}\big(\mathbf{B}_\star^\tau w_i^\tau - \phi_{\star i}\big)(w_i^\tau)^\top + \eta\sum_{i\in[M]}\frac{\mathbf{X}_{\star i}\mathbf{E}_i(w_i^\tau)^\top}{N}, \tag{40}$$

where $\phi_{\star i}$ and $\mathbf{X}_{\star i}$ are padded versions of ϕ_i and $\mathbf{X}_i$, respectively. To upper bound the norm of $\mathbf{Q}_\star^{\tau+1}$, we decompose it into two parts:

$$\frac{1}{\eta}\|\mathbf{Q}_\star^{\tau+1}\| \leq \underbrace{\left\|\sum_{i\in[M]}\big(\mathbf{B}_\star^\tau w_i^\tau - \phi_{\star i}\big)(w_i^\tau)^\top - \sum_{i\in[M]}\frac{\mathbf{X}_{\star i}\mathbf{X}_{\star i}^\top}{N}\big(\mathbf{B}_\star^\tau w_i^\tau - \phi_{\star i}\big)(w_i^\tau)^\top\right\|}_{\mathcal{A}_1^{\tau+1}} + \underbrace{\left\|\sum_{i\in[M]}\frac{\mathbf{X}_{\star i}\mathbf{E}_i(w_i^\tau)^\top}{N}\right\|}_{\mathcal{A}_2^{\tau+1}}.$$

For $\mathcal{A}_1^{\tau+1}$,by applying Lemma 5.4 in Vershynin (2010), there exists a $\frac{1}{4}$-net $\mathcal{N}_k$ on the unit sphere S^{k-1} and a $\frac{1}{4}$-net $\mathcal{N}_d$ on the unit sphere S^{d-1} such that

$$\mathcal{A}_1^{\tau+1} \leq 2\max_{u\in\mathcal{N}_d, v\in\mathcal{N}_k}\left|\sum_{i\in[M]}\frac{1}{N}\sum_{j\in[N]}u^\top\big(\mathbf{B}^\tau w_i^\tau - \phi_i\big)(w_i^\tau)^\top v - \sum_{i\in[M]}\frac{1}{N}\sum_{j\in[N]}u^\top x_{i,j}x_{i,j}^\top\big(\mathbf{B}^\tau w_i^\tau - \phi_i\big)(w_i^\tau)^\top v\right|.$$

Denote $c_i^\tau = \|\mathbf{B}^\tau w_i^\tau - \phi_i\|$ and $c_w^\tau = \max_i\{\|w_i^\tau\|\}$. Observe that $u^\top x_{i,j}x_{i,j}^\top\big(\mathbf{B}^\tau w_i^\tau - \phi_i\big)(w_i^\tau)^\top v - u^\top(\mathbf{B}^\tau w_i^\tau - \phi_i)(w_i^\tau)^\top v$ is a sub-exponential random variable with sub-exponential norm $c'c_i^\tau c_w^\tau$ for some constant c', where c' depends on the distribution of x. Then, based on the tail bound for sub-exponential random variables, there exists a constant $c_2' > 0$ such that for any $s \geq 0$,

$$\begin{aligned}&\mathbb{P}\left\{\frac{1}{N}\left(\sum_{i\in[M]}\sum_{j\in[N]}u^\top\big(\mathbf{B}^\tau w_i^\tau - \phi_i\big)(w_i^\tau)^\top v - \sum_{i\in[M]}\sum_{j\in[N]}u^\top x_{i,j}x_{i,j}^\top\big(\mathbf{B}^\tau w_i^\tau - \phi_i\big)(w_i^\tau)^\top v\right) \geq s\right\}\\ &\leq \exp\left(-Nc_2'\min\left(\frac{s^2}{\sum_{i\in[M]}(c_i^\tau c_w^\tau)^2}, \frac{s}{\max_i\{c_i^\tau c_w^\tau\}}\right)\right).\end{aligned}$$

Taking the union bound over all $u \in \mathcal{N}_d$ and $v \in \mathcal{N}_k$, with probability at least $1 - 9^{d+k}\exp\Big(-Nc_2'\min\Big\{\frac{s^2}{\sum_{i\in[M]}(c_i^\tau c_w^\tau)^2}, \frac{s}{\max_i\{c^\tau c_w^\tau\}}\Big\}\Big)$, we have

$$\left\|\sum_{i\in[M]}\big(\mathbf{B}^\tau w_i^\tau - \phi_i\big)(w_i^\tau)^\top - \sum_{i\in[M]}\frac{\mathbf{X}_i\mathbf{X}_i^\top}{N}\big(\mathbf{B}^\tau w_i^\tau - \phi_i\big)(w_i^\tau)^\top\right\| \leq 2s.$$

Since $\sigma_1^2(\mathbf{\Theta}^\tau) \leq 2\lambda_1^*$, we have

$$\|(\tilde{\mathbf{B}}^\tau)^\top\tilde{\mathbf{B}}^\tau + \tilde{\mathbf{W}}^\tau(\tilde{\mathbf{W}}^\tau)^\top\| = \sigma_1^2(\mathbf{\Theta}^\tau) \leq 2\lambda_1^*.$$

Note that $(\tilde{\mathbf{B}}^\tau)^\top\tilde{\mathbf{B}}^\tau$ and $\tilde{\mathbf{W}}^\tau(\tilde{\mathbf{W}}^\tau)^\top$ are PSD matrices. It follows that

$$\|\tilde{\mathbf{B}}^\tau\| \leq \sqrt{2\lambda_1^*} \quad \text{and} \quad \|\tilde{\mathbf{W}}^\tau\| \leq \sqrt{2\lambda_1^*},$$

which implies that $\|\mathbf{B}^\tau\| \leq \sqrt{2\lambda_1^*}$ and $\|\mathbf{W}^\tau\| \leq \sqrt{2\lambda_1^*}$. Since $c_i^\tau = \|\mathbf{B}^\tau w_i^\tau - \phi_i\|$ and $c_w^\tau = \max_i\{\|w_i^\tau\|\}$, we have $\sum_{i\in[M]}(c_i^\tau c_w^\tau)^2 \leq 2\lambda_1^*\|\mathbf{B}^\tau\mathbf{W}^\tau - \mathbf{\Phi}\|_F^2 \leq 4\lambda_1^*(k(\lambda_1^*)^2 + E)$ and $\max_i\{c_i^\tau c_w^\tau\} \leq 3\sqrt{2}(\lambda_1^*)^{\frac{3}{2}}$, where $E = \sum_i(\lambda_i)^2$. Let $s = \sqrt{18\lambda_1^*(k(\lambda_1^*)^2 + E)}\cdot\sqrt{\log(1/\delta)/d + 6}\cdot\sqrt{d}/\sqrt{Nc_2'}$. Then, if N is sufficiently large such that $(\sqrt{\log(1/\delta)/d + 2})\sqrt{d}/\sqrt{Nc_2'} \leq 1$, we have

$$\frac{s}{3\sqrt{2}(\lambda_1^*)^{\frac{3}{2}}} \geq \frac{s}{\sqrt{18\lambda_1^*(k(\lambda_1^*)^2 + E)}} \geq \frac{s^2}{18\lambda_1^*(k(\lambda_1^*)^2 + E)}.$$

Then, with probability at least $1 - \delta$, we have

$$\left\|\sum_{i\in[M]}\big(\mathbf{B}^\tau w_i^\tau - \phi_i\big)(w_i^\tau)^\top - \sum_{i\in[M]}\frac{\mathbf{X}_i\mathbf{X}_i^\top}{N}\big(\mathbf{B}^\tau w_i^\tau - \phi_i\big)(w_i^\tau)^\top\right\| \leq 2\sqrt{18\lambda_1^*(k(\lambda_1^*)^2 + E)}\cdot\sqrt{\log(1/\delta)/d + 6}\cdot\sqrt{\frac{d}{Nc}}.$$

Therefore, with probability at least $1-\delta$, we have

$$\mathcal{A}_1^{\tau+1} \leq \sqrt{\lambda_1^*\big(k(\lambda_1^*)^2+E\big)}\frac{6\sqrt{2}\sqrt{d-\log\delta}}{\sqrt{Nc_2}},$$

where $c_2 = \frac{c_2'}{6}$.

Next, we consider $\mathcal{A}_2^{\tau+1}$. Similar to the above analysis, note that $u^\top x_{i,j}\xi_{i,j}(w_i^\tau)^\top v$ is a centered sub-exponential random variable with sub-exponential norm $c''\sigma_\xi\|w_i^\tau\|$ for some constant c''. Based on the tail bound for sub-exponential random variables, there exists a constant $c_3' > 0$ such that for any $s \geq 0$,

$$\mathbb{P}\left\{\sum_{i\in[M]} u^\top \frac{\mathbf{X}_{\star i}\mathbf{E}_{\star i}(w_i^\tau)^\top}{N} v \geq s\right\} \leq \exp\left(-Nc_3'\min\left(\frac{s^2}{\sigma_\xi^2\|\mathbf{W}^\tau\|_F^2}, \frac{s}{\sigma_\xi\sqrt{2\lambda_1^*}}\right)\right).$$

Combining with the fact $\|\mathbf{W}^\tau\|_F^2 \leq 2k\lambda_1^*$ and taking the union bound over all $u \in \mathcal{N}_d$ and $v \in \mathcal{N}_k$, we have that inequality $\left\|\sum_{i\in[M]} \frac{\mathbf{X}_{\star i}\mathbf{E}_{\star i}^\tau(w_i^\tau)^\top}{N}\right\| \leq 2s$ holds with probability at least $1-9^{d+k}\exp\Big(-Nc\min\Big\{\frac{s^2}{2\sigma_\xi^2 k\lambda_1^*}, \frac{s}{\sigma_\xi\sqrt{2\lambda_1^*}}\Big\}\Big)$. Then, let $s = \sqrt{2\sigma^2 k\lambda_1^*}\cdot\sqrt{\log(1/\delta)/d+6}\cdot\sqrt{d}/\sqrt{Nc}$. If N is sufficiently large such that $(\sqrt{\log(1/\delta)/d+2})\sqrt{d}/\sqrt{Nc_3'} \leq 1$ we have $\min\{\frac{s^2}{2\sigma_\xi^2 k\lambda_1^*}, \frac{s}{\sigma_\xi\sqrt{2\lambda_1^*}}\} = \frac{s^2}{2\sigma_\xi^2 k\lambda_1^*}$. Therefore, with a probability at least $1-\delta$, we have

$$\mathcal{A}_2^{\tau+1} \leq 2\sqrt{2\sigma_\xi^2 k\lambda_1^*}\cdot\sqrt{\log(1/\delta)/d+6}\sqrt{\frac{d}{Nc_3'}} \leq (\lambda_1^*)^{\frac{1}{2}}\sigma_\xi\frac{6\sqrt{2}\sqrt{dk-k\log\delta}}{\sqrt{Nc_3}},$$

where $c_3 = \frac{c_3'}{24}$. Combining the upper bounds of $\mathcal{A}_1^{\tau+1}$ and $\mathcal{A}_2^{\tau+1}$, we conclude that following inequality holds with probability at least $1-\delta$:

$$\begin{aligned}
\|\mathbf{U}_\star^\top\mathbf{Q}_\star^{\tau+1}\| \leq \|\mathbf{Q}_\star^{\tau+1}\| &\leq \eta(\mathcal{A}_1^{\tau+1}+\mathcal{A}_2^{\tau+1}) \\
&\leq \eta(\lambda_1^*)^{\frac{1}{2}}\sqrt{k(\lambda_1^*)^2+E}\frac{6\sqrt{2}\sqrt{d-\log(\delta/2)}}{\sqrt{Nc_2}} + \eta(\lambda_1^*)^{\frac{1}{2}}\sigma_\xi\frac{6\sqrt{2}\sqrt{dk-k\log(\delta/2)}}{\sqrt{Nc_3}} \\
&\leq \eta\sqrt{\lambda_1^*}\bigg(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi\bigg)\frac{6\sqrt{2}\sqrt{d-\log\delta}}{\sqrt{Nc_1}},
\end{aligned}$$

where $c_1 = \frac{1}{2}\min\{c_2, c_3\}$. Thus, for any $\sigma \geq 0$, if $\sqrt{N} \geq \frac{192\sqrt{d-\log\delta}\sqrt{\lambda_1^*}(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\sigma\Delta\sqrt{c_1}}$, with probability at least $1-\delta$ it holds

$$\|\mathbf{U}_\star^\top\mathbf{Q}_\star^{\tau+1}\| \leq \frac{\sigma\eta\Delta}{16\sqrt{2}}.$$

Similarly, if $\sqrt{N} \geq \frac{6144\sqrt{d-\log\delta}\lambda_1^*(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\Delta^2\sqrt{c_1}}$, with probability at least $1-\delta$,

$$\|\mathbf{U}_\star^\top\mathbf{Q}_\star^{\tau+1}\| \leq \frac{\eta\Delta^2}{512\sqrt{2}\sqrt{\lambda_1^*}}.$$

□

Lemma A.14 (Concentration of $\|\tilde{\mathbf{Q}}_\star^{\tau+1}\mathbf{V}_\star\|$)**.** *For any $t \geq 0$, assume $\mathbf{\Theta}^\tau \in \mathcal{R}_s$ holds for all $0 < \tau \leq t$. Then, we have the following results for any $0 \leq \tau \leq t$ and $c_2 \geq 0$ with probability at least $1-2\delta$:*

- *If $\sqrt{N} \geq \max\big\{\frac{\sqrt{\bar{d}-\log\delta}}{\sqrt{c_1}}, \frac{192\sqrt{\bar{d}-\log\delta}\sqrt{\lambda_1^*}(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\sigma\Delta\sqrt{c_1}}\big\}$, then it holds that*

$$\|\tilde{\mathbf{Q}}_\star^{\tau+1}\mathbf{V}_\star\| \leq \frac{\sigma\eta\Delta}{16\sqrt{2}}.$$

- *If* $\sqrt{N} \geq \max\big\{\frac{\sqrt{\bar{d}-\log\delta}}{\sqrt{c_1}}, \frac{6144\sqrt{\bar{d}-\log\delta}\lambda_1^*(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\Delta^2\sqrt{c_1}}\big\}$*, then it holds that*

$$\|\tilde{\mathbf{Q}}_\star^{\tau+1}\mathbf{V}_\star\| \leq \frac{\eta\Delta^2}{512\sqrt{2}\sqrt{\lambda_1^*}}.$$

Proof of Lemma A.14. This proof resembles the proof of Lemma A.13. According to Lemma 5.4 in Vershynin (2010), there exists a $\frac{1}{4}$-net $\mathcal{N}_k$ on the unit sphere S^{k-1} and a $\frac{1}{4}$-net $\mathcal{N}_M$ on the unit sphere S^{M-1} so that

$$\begin{aligned}
\frac{1}{\eta}\|\tilde{\mathbf{Q}}_\star^{\tau+1}\| &\leq 2\max_{u\in\mathcal{N}_M, v\in\mathcal{N}_k}\left|\sum_{i\in[M]} v^\top\tilde{q}_i^{\tau+1}u_i + v^\top\frac{1}{N}\sum_{i\in[M]}(\mathbf{B}^\tau)^\top\mathbf{X}_i\mathbf{E}_i u_i\right| \\
&\leq \underbrace{2\max_{u\in\mathcal{N}_M, v\in\mathcal{N}_k}\left|\frac{1}{N}\left(\sum_{i\in[M]}\sum_{j\in[N]} v^\top(\mathbf{B}^\tau)^\top\big(\mathbf{B}^\tau w_i^\tau-\phi_i\big)u_i - \sum_{i\in[M]}\sum_{j\in[N]} v^\top(\mathbf{B}^\tau)^\top x_{i,j}x_{i,j}^\top\big(\mathbf{B}^\tau w_i^\tau-\phi_i\big)u_i\right)\right|}_{\mathcal{A}_3^{\tau+1}} \\
&\quad + \underbrace{2\max_{u\in\mathcal{N}_M, v\in\mathcal{N}_k}\left|v^\top\frac{1}{N}\sum_{i\in[M]}(\mathbf{B}^\tau)^\top\mathbf{X}_i\mathbf{E}_i u_i\right|}_{\mathcal{A}_4^{\tau+1}}.
\end{aligned}$$

Let $c_B^\tau = \|\mathbf{B}^\tau\|$ and recall that $c_i^\tau = \|\mathbf{B}^\tau w_i^\tau - \phi_i\|$. Based on the tail bound for sub-exponential random variables, there exists a constant $c>0$ such that for any $s\geq 0$,

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{N}\left(\sum_{i\in[M]}\sum_{j\in[N]} v^\top(\mathbf{B}^\tau)^\top\big(\mathbf{B}^\tau w_i^\tau-\phi_i\big)u_i - \sum_{i\in[M]}\sum_{j\in[N]} v^\top(\mathbf{B}^\tau)^\top x_{i,j}x_{i,j}^\top\big(\mathbf{B}^\tau w_i^\tau-\phi_i\big)u_i\right)\geq s\right\} \\
&\leq \exp\left(-Nc\min\left(\frac{s^2}{\sum_{i\in[M]}(c_i^\tau c_B^\tau)^2}, \frac{s}{\max_i\{c_i^\tau c_B^\tau\}}\right)\right).
\end{aligned}$$

Taking the union bound over all $u\in\mathcal{N}_M$ and $v\in\mathcal{N}_k$, with probability at least $1-9^{M+k}\exp\Big(-Nc\min\Big\{\frac{s^2}{\sum_{i\in[M]}(c_i^\tau c_B^\tau)^2}, \frac{s}{\max_i\{c^\tau c_B^\tau\}}\Big\}\Big)$, we have $\mathcal{A}_3^{\tau+1}\leq 2s$. Let $s = (\lambda_1^*)^{\frac{1}{2}}\sqrt{k(\lambda_1^*)^2+E}\cdot\sqrt{\log(1/\delta)/\bar{d}+6}\cdot\sqrt{\bar{d}}/\sqrt{Nc}$. If N is sufficiently large such that $(\sqrt{\log(1/\delta)/d+2})\sqrt{\bar{d}}/\sqrt{Nc}\leq 1$, there exits constant c_2 such that with probability at least $1-\delta$, we have

$$\mathcal{A}_3^{\tau+1}\leq \eta(\lambda_1^*)^{\frac{1}{2}}\sqrt{k(\lambda_1^*)^2+E}\frac{6\sqrt{2}\sqrt{\bar{d}-\log\delta}}{\sqrt{Nc_2}}.$$

For term $\mathcal{A}_4^{\tau+1}$, from the tail bound for sub-exponential random variables, there exists a constant $c>0$ such that for any $s\geq 0$,

$$\begin{aligned}
\mathbb{P}\left\{\sum_{i\in[M]} v^\top\frac{(\mathbf{B}^\tau)^\top\mathbf{X}_i\mathbf{E}_i}{N}u_i\geq s\right\} &\leq \exp\left(-Nc\min\left(\frac{s^2}{\sigma_\xi^2\sum_{i\in[M]}\|u_i\mathbf{B}^\tau\|^2}, \frac{s}{\sigma_\xi\sqrt{2\lambda_1^*}}\right)\right) \\
&\leq \exp\left(-Nc\min\left(\frac{s^2}{2\sigma_\xi^2\lambda_1^*}, \frac{s}{\sigma_\xi\sqrt{2\lambda_1^*}}\right)\right).
\end{aligned}$$

Let $s = \sqrt{2\sigma_\xi^2 k\lambda_1^*}\cdot\sqrt{\log(1/\delta)/\bar{d}+6}\cdot\sqrt{\bar{d}}/\sqrt{Nc}$. If N is sufficiently large such that $(\sqrt{\log(1/\delta)/\bar{d}+2})\sqrt{\bar{d}}/\sqrt{Nc}\leq 1$, there exists constant c_3 such that with a probability at least $1-\delta$, we have

$$\mathcal{A}_4^{\tau+1}\leq (\lambda_1^*)^{\frac{1}{2}}\sigma_\xi\frac{6\sqrt{2}\sqrt{\bar{d}k-k\log\delta}}{\sqrt{Nc_3}}.$$

Combining the upper bounds of $\mathcal{A}_3^{\tau+1}$ and $\mathcal{A}_4^{\tau+1}$, we conclude that following inequality holds with probability at least $1-\delta$:

$$\begin{aligned}\|\tilde{\mathbf{Q}}_\star^{\tau+1}\mathbf{V}_\star\| \le \|\tilde{\mathbf{Q}}_\star^{\tau+1}\| &\le \eta(\mathcal{A}_3^{\tau+1}+\mathcal{A}_4^{\tau+1}) \\ &\le \eta(\lambda_1^*)^{\frac{1}{2}}\sqrt{k(\lambda_1^*)^2+E}\frac{6\sqrt{2}\sqrt{\bar{d}-\log(\delta/2)}}{\sqrt{Nc_2}} + \eta(\lambda_1^*)^{\frac{1}{2}}\sigma_\xi\frac{6\sqrt{2}\sqrt{\bar{d}k-k\log(\delta/2)}}{\sqrt{Nc_3}} \\ &\le \eta\sqrt{\lambda_1^*}\Big(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi\Big)\frac{6\sqrt{2}\sqrt{\bar{d}-\log\delta}}{\sqrt{Nc_1}},\end{aligned}$$

where $c_1 = \frac{1}{2}\min\{c_2, c_3\}$. Thus, for any $\sigma \ge 0$, if $\sqrt{N} \ge \frac{192\sqrt{\bar{d}-\log\delta}\sqrt{\lambda_1^*}(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\sigma\Delta\sqrt{c_1}}$, with probability at least $1-\delta$ it holds that

$$\|\tilde{\mathbf{Q}}_\star^{\tau+1}\mathbf{V}_\star\| \le \frac{\sigma\eta\Delta}{16\sqrt{2}}.$$

Similarly, if $\sqrt{N} \ge \frac{6144\sqrt{\bar{d}-\log\delta}\lambda_1^*(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)}{\Delta^2\sqrt{c_1}}$, with probability at least $1-\delta$,

$$\|\tilde{\mathbf{Q}}_\star^{\tau+1}\mathbf{V}_\star\| \le \frac{\eta\Delta^2}{512\sqrt{2}\sqrt{\lambda_1^*}}.$$

□

Lemma A.15. *Suppose $\eta \le 1/6\lambda_1^*$ and $\sigma_1(\mathbf{\Theta}^\tau) \le \sqrt{2\lambda_1^*}$. Then, it holds that*

$$\sigma_1(\mathbf{\Theta}_{res}^{\tau+1}) \le \Big(1+\frac{\eta}{2}\big(\lambda_{k+1}^* - \sigma_1^2(\mathbf{\Theta}_{res}^\tau) - \sigma_k^2(\mathbf{\Theta}_k^\tau)\big)\Big)\sigma_1(\mathbf{\Theta}_{res}^\tau) + \sigma_1(\mathbf{R}_{2\bar{d}-k}^\tau).$$

Proof of Lemma A.15. According to Equation (15), we can rewrite $\mathbf{\Theta}_{\text{res}}^{\tau+1}$ as

$$\mathbf{\Theta}_{\text{res}}^{\tau+1} = \frac{1}{2}\mathbf{\Theta}_{\text{res}}^\tau - \frac{\eta}{2}\mathbf{\Theta}_{\text{res}}^\tau(\mathbf{\Theta}_{\text{res}}^\tau)^\top\mathbf{\Theta}_{\text{res}}^{\tau-1} + \Big(\frac{1}{4}\mathbf{I}_{2\bar{d}-k} + \frac{\eta}{2}\tilde{\mathbf{\Lambda}}_{\text{res}}\Big)\mathbf{\Theta}_{\text{res}}^\tau + \mathbf{\Theta}_{\text{res}}^\tau\Big(\frac{1}{4}\mathbf{I}_k - \frac{\eta}{2}(\mathbf{\Theta}_k^\tau)^\top\mathbf{\Theta}_k^\tau\Big) + \mathbf{R}_{2\bar{d}-k}^{\tau+1}. \tag{41}$$

From Lemma A.5 in Chen et al. (2023) we have the following inequalities:

$$\sigma_1\Big(\frac{1}{2}\mathbf{\Theta}_{\text{res}}^\tau - \frac{\eta}{2}\mathbf{\Theta}_{\text{res}}^\tau(\mathbf{\Theta}_{\text{res}}^\tau)^\top\mathbf{\Theta}_{\text{res}}^\tau\Big) \le \frac{1}{2}\sigma_1(\mathbf{\Theta}_{\text{res}}^\tau) - \frac{\eta}{2}\sigma_1^3(\mathbf{\Theta}_{\text{res}}^\tau), \tag{42}$$

$$\sigma_1\Big(\Big(\frac{1}{4}\mathbf{I}_{2\bar{d}-k} + \frac{\eta}{2}\mathbf{\Lambda}_{\text{res}}\Big)\mathbf{\Theta}_{\text{res}}^\tau\Big) \le (\frac{1}{4} + \frac{\eta\lambda_{k+1}^*}{2})\sigma_1(\mathbf{\Theta}_{\text{res}}^\tau), \tag{43}$$

$$\sigma_1\Big(\mathbf{\Theta}_{\text{res}}^\tau\Big(\frac{1}{4}\mathbf{I}_k - \frac{\eta}{2}(\mathbf{\Theta}_k^\tau)^\top\mathbf{\Theta}_k^t\Big)\Big) \le \sigma_1(\mathbf{\Theta}_{\text{res}}^\tau)(\frac{1}{4} - \frac{\eta}{2}\sigma_k^2(\mathbf{\Theta}_k^\tau)). \tag{44}$$

Substituting Equations (42) to (44) into Equation (41) proves the lemma. □

Lemma A.16. *Suppose $\sigma_1(\mathbf{\Theta}^t) \le \sqrt{2\lambda_1^*}$ and $\eta \le \frac{\Delta^2}{16{\lambda_1^*}^3}$. Then, it holds that*

$$\sigma_k^2(\mathbf{\Theta}_k^{\tau+1}) \ge \Big(1+\eta\big(\lambda_k^* - \sigma_1^2(\mathbf{\Theta}_{res}^\tau) - \sigma_k^2(\mathbf{\Theta}_k^\tau)\big) - \frac{\eta\Delta}{4}\Big)\sigma_k^2(\mathbf{\Theta}_k^\tau) - 4\sqrt{\lambda_1^*}\sigma_1(\mathbf{R}_k^\tau).$$

Proof of Lemma A.16. Denote $\tilde{\mathbf{\Theta}}_k^{\tau+1} = \mathbf{\Theta}_k^\tau + \frac{\eta}{2}\tilde{\mathbf{\Lambda}}_k\mathbf{\Theta}_k^\tau - \frac{\eta}{2}\mathbf{\Theta}_k^\tau(\mathbf{\Theta}^\tau)^\top\mathbf{\Theta}^\tau$. Based on Lemma A.6 and Lemma 2.3 in Chen et al. (2023), we have

$$\begin{aligned}\sigma_k^2(\tilde{\mathbf{\Theta}}_k^{\tau+1}) &\ge \Big(1+\eta\big(\lambda_k^* - \sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau) - \sigma_k^2(\mathbf{\Theta}_k^\tau)\big)\Big)\sigma_k^2(\mathbf{\Theta}_k^\tau) - \eta^2{\lambda_1^*}^3 \\ &\ge \Big(1+\eta\big(\lambda_k^* - \sigma_1^2(\mathbf{\Theta}_{\text{res}}^\tau) - \sigma_k^2(\mathbf{\Theta}_k^\tau)\big) - \frac{\eta\Delta}{4}\Big)\sigma_k^2(\mathbf{\Theta}_k^\tau).\end{aligned} \tag{45}$$

Combining with $\eta \leq \frac{1}{16\lambda_1^*}$ gives

$$\begin{aligned}\sigma_1(\tilde{\mathbf{\Theta}}_k^{\tau+1}) &\leq \sigma_1(\mathbf{\Theta}_k^\tau) + \sigma_1\Big(\frac{\eta}{2}\tilde{\mathbf{\Lambda}}_k\mathbf{\Theta}_k^\tau\Big) + \sigma_1\Big(\frac{\eta}{2}\mathbf{\Theta}_k^\tau(\mathbf{\Theta}^\tau)^\top\mathbf{\Theta}^\tau\Big)\\ &\leq \sqrt{2\lambda_1^*} + \frac{1}{32}\sqrt{\lambda_1^*} + \frac{\sqrt{2}}{16}\sqrt{\lambda_1^*}\\ &\leq 2\sqrt{\lambda_1^*}.\end{aligned}$$

Thus, $\sigma_k(\tilde{\mathbf{\Theta}}_k^{\tau+1}) \leq \sigma_1(\tilde{\mathbf{\Theta}}_k^{\tau+1}) \leq 2\sqrt{\lambda_1^*}$. Combining with the fact that $\sigma_k(\mathbf{\Theta}_k^{\tau+1}) \geq \sigma_k(\tilde{\mathbf{\Theta}}_k^{\tau+1}) - \sigma_1(\mathbf{R}_k^{\tau+1})$, we have

$$\begin{aligned}\sigma_k^2(\mathbf{\Theta}_k^{\tau+1}) &\geq \Big(\sigma_k(\tilde{\mathbf{\Theta}}_k^{\tau+1}) - \sigma_1(\mathbf{R}_k^{\tau+1})\Big)^2\\ &\geq \sigma_k^2(\tilde{\mathbf{\Theta}}_k^{\tau+1}) - 2\sigma_k(\tilde{\mathbf{\Theta}}_k^{\tau+1})\cdot\sigma_1(\mathbf{R}_k^{\tau+1})\\ &\geq \sigma_k^2(\tilde{\mathbf{\Theta}}_k^{\tau}) - 4\sqrt{\lambda_1^*}\sigma_1(\mathbf{R}_k^{\tau}).\end{aligned} \tag{46}$$

The lemma thus follows by substituting Equation (45) into Equation (46). □

Lemma A.17. *Assume $\eta \leq \frac{1}{6\lambda_1^*}$ and $\sigma_1(\mathbf{\Theta}^0) \leq \sqrt{2\lambda_1^*}$ hold. Then, if*

$$\sqrt{N} \geq \max\Big\{\frac{\sqrt{d-\log\delta}}{\sqrt{c_1}}, \frac{96\sqrt{\lambda_1^*}(\sqrt{k(\lambda_1^*)^2+E}+\sqrt{k}\sigma_\xi)\sqrt{d-\log\delta}}{\sigma_1(\mathbf{\Theta}_{res}^0)\Delta\sqrt{c_1}}\Big\}, \tag{47}$$

with probability at least $1 - ct\delta$ for some constant c, we have

$$\sigma_1(\mathbf{\Theta}_{res}^\tau) \leq \Big(1 + \frac{\eta}{2}\lambda_{k+1}^* + \frac{\eta}{8}\Delta\Big)^\tau\sigma_1(\mathbf{\Theta}_{res}^0)$$

holds for all $\tau \leq t$.

Proof of Lemma A.17. Suppose $\eta \leq 1/6\lambda_1^*$ and $\sigma_1(\mathbf{\Theta}^\tau) \leq \sqrt{2\lambda_1^*}$. We have

$$\begin{aligned}\sigma_1(\mathbf{\Theta}_{\text{res}}^\tau) &\leq \Big(1 + \frac{\eta}{2}\big(\lambda_{k+1}^* - \sigma_1^2(\mathbf{\Theta}_{\text{res}}^{\tau-1}) - \sigma_k^2(\mathbf{\Theta}_k^{\tau-1})\big)\Big)\sigma_1(\mathbf{\Theta}_{\text{res}}^{\tau-1}) + \sigma_1(\mathbf{R}_{2\overline{d}-k}^\tau)\\ &\leq \Big(1 + \frac{\eta}{2}\lambda_{k+1}^*\Big)\sigma_1(\mathbf{\Theta}_{\text{res}}^{\tau-1}) + \sigma_1(\mathbf{R}^\tau).\end{aligned}$$

Combining Lemma A.13 and Lemma A.14, when N satisfies Equation (47), we have $\sigma_1(\mathbf{R}^t) \leq \frac{\eta}{8}\Delta\sigma_1(\mathbf{\Theta}_{\text{res}}^0)$. The lemma follows by induction. □

Lemma A.18. *If $\|\mathbf{\Theta}(\mathbf{\Theta})^\top - \text{diag}(\tilde{\mathbf{\Lambda}}_k, \mathbf{0})\|_F \leq \delta$ for some $\delta > 0$, then $\|\mathbf{BW} - \text{diag}(\mathbf{\Lambda}_k, \mathbf{0})\|_F \leq \delta$.*

Proof of Lemma A.18. Note that $\|\mathbf{\Theta}(\mathbf{\Theta})^T - \text{diag}(\tilde{\mathbf{\Lambda}}_k, \mathbf{0})\|_F \leq \delta$ implies that

$$\Big\|\Big(\frac{\mathbf{B}+\mathbf{W}^\top}{\sqrt{2}}\Big)\Big(\frac{\mathbf{B}+\mathbf{W}^\top}{\sqrt{2}}\Big)^\top - 2\,\text{diag}(\mathbf{\Lambda}_k, \mathbf{0})\Big\|_F \leq \delta,$$

and

$$\Big\|\Big(\frac{\mathbf{B}-\mathbf{W}^\top}{\sqrt{2}}\Big)\Big(\frac{\mathbf{B}-\mathbf{W}^\top}{\sqrt{2}}\Big)^\top\Big\|_F \leq \delta.$$

Then,

$$\begin{aligned}&\|\mathbf{BW} + \mathbf{W}^\top\mathbf{B}^\top - 2\,\text{diag}(\mathbf{\Lambda}_k, \mathbf{0})\|_F\\ &= \Big\|\Big(\frac{\mathbf{B}+\mathbf{W}^\top}{\sqrt{2}}\Big)\Big(\frac{\mathbf{B}+\mathbf{W}^\top}{\sqrt{2}}\Big)^\top - 2\,\text{diag}(\mathbf{\Lambda}_k, \mathbf{0}) - \Big(\frac{\mathbf{B}-\mathbf{W}^\top}{\sqrt{2}}\Big)\Big(\frac{\mathbf{B}-\mathbf{W}^\top}{\sqrt{2}}\Big)^\top\Big\|_F\\ &\leq \Big\|\Big(\frac{\mathbf{B}+\mathbf{W}^\top}{\sqrt{2}}\Big)\Big(\frac{\mathbf{B}+\mathbf{W}^\top}{\sqrt{2}}\Big)^\top - 2\,\text{diag}(\mathbf{\Lambda}_k, \mathbf{0})\Big\|_F + \Big\|\Big(\frac{\mathbf{B}-\mathbf{W}^\top}{\sqrt{2}}\Big)\Big(\frac{\mathbf{B}-\mathbf{W}^\top}{\sqrt{2}}\Big)^\top\Big\|_F\\ &\leq 2\delta.\end{aligned}$$

Combining with the fact $\|\mathbf{BW} + \mathbf{W}^\top\mathbf{B}^\top - 2\,\text{diag}(\mathbf{\Lambda}_k, \mathbf{0})\|_F = 2\|\mathbf{BW} - \text{diag}(\mathbf{\Lambda}_k, \mathbf{0})\|_F$, the proof is complete. □

Lemma A.19 (Theorem 2.13 in Davidson & Szarek (2001)). *Let $N \geq n$ and $\mathbf{A}$ be an $N \times n$ matrix whose entries are IID standard Gaussian random variables. Then, for any $\epsilon \geq 0$, with probability at least $1 - 2\exp(-\epsilon^2/2)$, we have*

$$\sqrt{N} - \sqrt{n} - \epsilon \leq \sigma_{\min}(\mathbf{A}) \leq \sigma_{\max}(\mathbf{A}) \leq \sqrt{N} + \sqrt{n} + \epsilon.$$

Lemma A.20 (Eigenvalue Interlacing Theorem (Hwang, 2004)). *For a symmetric matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, let $\mathbf{B} \in \mathbb{R}^{k\times k}$, $k < d$, be a principal matrix of $\mathbf{A}$. Denote the eigenvalues of $\mathbf{A}$ as $\lambda_1 \geq \cdots \geq \lambda_d$ and the eigenvalues of $\mathbf{B}$ as $\mu_1 \geq \cdots \geq \mu_d$. Then, for any $i \in [k]$, it holds that*

$$\lambda_{i+d-k} \leq \mu_i \leq \lambda_i.$$

B. General `FLUTE`

B.1. Details of General `FLUTE`

Algorithm 2 General `FLUTE`

Input: Learning rates η_l and η_r, regularization parameter λ, communication round T
Initialization: Server initializes model parameters $\mathbf{B}^0, \{b_i^0\}, \{\mathbf{H}_i^0\}$
for $t = \{0, \cdots, T-1\}$ **do**
  Server samples a batch of clients $\mathcal{I}^{t+1}$
  Server sends $\mathbf{B}^t$, and $\mathbf{H}_i^t$ to all client $i \in \mathcal{I}^{t+1}$
  for client $i \in [M]$ in parallel **do**
    if $i \in \mathcal{I}^{t+1}$ **then**
      $\mathbf{B}_i^{t,0} \leftarrow \mathbf{B}^t$, $b_i^{t,0} \leftarrow b_i^t$ and $\mathbf{H}_i^{t,0} \leftarrow \mathbf{H}_i$
      for $\tau = \{0, \cdots, \mathcal{T}-1\}$ **do**
        $\mathbf{H}_i^{t,\tau+1} \leftarrow \text{GRD}(\mathcal{L}_i(\mathbf{B}_i^{t,\tau}, b_i^{t,\tau}, \mathbf{H}_i^{t,\tau}); \mathbf{H}_i^{t,\tau}, \eta_l)$
        $b_i^{t,\tau+1} \leftarrow \text{GRD}(\mathcal{L}_i(\mathbf{B}_i^{t,\tau}, b_i^{t,\tau}, \mathbf{H}_i^{t,\tau}); b_i^{t,\tau}, \eta_l)$
        $\mathbf{B}_i^{t,\tau+1} \leftarrow \text{GRD}(\mathcal{L}_i(\mathbf{B}_i^{t,\tau}, b_i^{t,\tau}, \mathbf{H}_i^{t,\tau}); \mathbf{B}^{t,\tau}, \eta_l)$
      end for
      $\mathbf{B}_i^{t+1} \leftarrow \mathbf{B}_i^{t,\mathcal{T}}$, $b_i^{t+1} \leftarrow b_i^{t,\mathcal{T}}$ and $\mathbf{H}_i^{t+1} \leftarrow \mathbf{H}_i^{t,\mathcal{T}}$
      Sends $\mathbf{B}_i^{t+1}$, b_i^{t+1} and $\mathbf{H}_i^{t+1}$ to the server
    else
      $b_i^{t+1} \leftarrow b_i^t$
    end if
  end for
  Server updates:
    $\mathbf{B}^{t+1} = \frac{1}{rM}\sum_{i\in\mathcal{I}^{t+1}} \mathbf{B}_i^{t+1}$
    $\{\mathbf{H}_i^{t+1}\}_{i\in\mathcal{I}^{t+1}} \leftarrow \text{GRD}(R(\{\mathbf{H}_i^{t+1}\}_{i\in\mathcal{I}^{t+1}}, \mathbf{B}^{t+1}); \{\mathbf{H}_i^{t+1}\}_{i\in\mathcal{I}^{t+1}}, \eta_r)$
    $\mathbf{H}_i^{t+1} \leftarrow \mathbf{H}_i^t, \forall i \notin \mathcal{I}^{t+1}$
end for

The General `FLUTE` is presented in Algorithm 2, where $\text{GRD}(f; \theta, \alpha)$ denotes the update of variable θ using the gradient of the function f with respect to θ and the step size α. The local loss function $\mathcal{L}_i$ is defined as

$$\mathcal{L}_i(\mathbf{B}, b, \mathbf{H}) = \frac{1}{N}\sum_{(x,y)\in\mathcal{D}_i} \mathcal{L}\big(\mathbf{H}^\top f_{\mathbf{B}}(x) + b, y\big). \tag{48}$$

In this work, we instantiate the general `FLUTE` by a federated multi-class classification problem. In this case, the local loss function is specialized as

$$\mathcal{L}_i(\mathbf{B}, b, \mathbf{H}) = \frac{1}{N}\sum_{(x,y)\in\mathcal{D}_i} \mathcal{L}_{\text{CE}}\big(\mathbf{H}_i^\top f_{\mathbf{B}}(x) + b_i, y\big) + \lambda_1\|f_{\mathbf{B}}(x)\|_2^2 + \lambda_2\|\mathbf{H}_i\|_F^2 + \lambda_3\mathcal{NC}_i(\mathbf{H}_i), \tag{49}$$

where $y \in \mathbb{R}^m$ is a one-hot vector whose k-th entry is 1 if the corresponding x belongs to class k and 0 otherwise, and λ_1, λ_2 and λ_3 are non-negative regularization parameters. $\mathcal{L}_{\text{CE}}(\cdot)$ is the cross-entropy loss, where for a one-hot vector y whose

k-th entry is 1, we have:

$$\mathcal{L}_{\text{CE}}(\hat{y}, y) = -\log\left(\frac{\exp(\hat{y}_k)}{\sum_{i\in[c]}\exp(\hat{y}_i)}\right). \tag{50}$$

$\mathcal{NC}_i(\mathbf{H}_i)$, inspired by the concept of neural collapse (Papyan et al., 2020), is defined as

$$\mathcal{NC}_i(\mathbf{H}_i) = \left\| \frac{\mathbf{H}_i^\top \mathbf{H}_i}{\|\mathbf{H}_i^T \mathbf{H}_i\|_F} - \frac{1}{\sqrt{m-1}}\mathbf{u}_i\mathbf{u}_i^\top \odot \left(\mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top\right)\right\|_F, \tag{51}$$

where $\mathbf{u}_i$ is an m-dimensional one-hot vector whose c-th entry is 1 if $c \in \mathcal{C}_i$ and 0 otherwise. Also, we specialize the regularization term optimized on the server side as $R(\{\mathbf{H}_i\}) = \sum_i \mathcal{NC}_i(\mathbf{H}_i)$.

B.2. Additional Definition

Definition B.1 (k-Simplex ETF, Definition 2.2 in Tirer & Bruna (2022))**.** The standard simplex equiangular tight frame (ETF) is a collection of points in $\mathbb{R}^k$ specified by the columns of

$$\mathbf{M} = \sqrt{\frac{k}{k-1}}\left(\mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^\top\right). \tag{52}$$

Consequently, the standard simplex EFT obeys

$$\mathbf{M}^\top\mathbf{M} = \mathbf{M}\mathbf{M}^\top = \frac{k}{k-1}\left(\mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^\top\right). \tag{53}$$

In this work, we consider a (general) simplex ETF as a collection of points in $\mathbb{R}^d$, $d \geq k$ specified by the columns of $\tilde{\mathbf{M}} \propto \sqrt{\frac{k}{k-1}}\mathbf{P}\left(\mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^\top\right)$, where $\mathbf{P} \in \mathbb{R}^{d\times k}$ is an orthonormal matrix. Consequently, $\tilde{\mathbf{M}}^\top\tilde{\mathbf{M}} \propto \mathbf{M}\mathbf{M}^\top = \frac{k}{k-1}\left(\mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^\top\right)$.

B.3. More Discussion on General `FLUTE`

Firstly, we explain the concept of *neural collapse*.

Neural collapse. Neural collapse (NC) was experimentally identified in Papyan et al. (2020), and they outlined four elements in the neural collapse phenomenon:

- (NC1) Features learned by the model (output of the representation layers) for samples within the same class tend to converge toward their average, essentially causing the within-class variance to diminish;
- (NC2) When adjusted for their overall average, the final means of different classes display a structure known as a simplex equiangular tight frame (ETF);
- (NC3) The weights of the final layer, which serves as the classifier, align with this simplex ETF structure;
- (NC4) Consequently, after this collapse occurs, classification decisions are made based on measuring the nearest class center in the feature space.

Next, we discuss some observations on the vanilla multi-classification problem, i.e., no additional regularization term and no client-side optimization, which is given as

$$\mathcal{L}_i(\mathbf{B}, b, \mathbf{H}) = \frac{1}{N}\sum_{(x,y)\in\mathcal{D}_i}\mathcal{L}_{\text{CE}}\left(\mathbf{H}_i^\top f_{\mathbf{B}}(x) + b_i, y\right) + \lambda_1\|f_{\mathbf{B}}(x)\|_2^2 + \lambda_2\|\mathbf{H}_i\|_F^2. \tag{54}$$

The first observation, which directly comes from Theorem 3.2 in Tirer & Bruna (2022), describes the phenomena of local neural collapse, which could happen when the model is locally trained for long epochs.

Observation B.2. *When $f_{\mathbf{B}}(\cdot)$ is sufficiently expressive such that $f_{\mathbf{B}}(x)$ can be viewed as a free variable. and the feature dimension k is no smaller than the number of total classes m, locally learned $\mathbf{B}$ and $\mathbf{H}_i$ that optimize the objective function (54) must satisfy:*

$$f_{\mathbf{B}^*}(x_1) = f_{\mathbf{B}^*}(x_2), \quad \forall x_1, x_2 \in \mathcal{D}_i^c, c \in \mathcal{C}_i \tag{55}$$

$$\frac{\langle f_{\mathbf{B}^*}(x), h_{i,c}^* \rangle}{\|f_{\mathbf{B}^*}(x)\| \cdot \|h_{i,c}^*\|} = 1, \quad \forall x \in \mathcal{D}_i^c, c \in \mathcal{C}_i \tag{56}$$

$$\frac{\mathbf{H}_i^\top \mathbf{H}_i}{\|\mathbf{H}_i^\top \mathbf{H}_i\|_F} = \frac{1}{\sqrt{m'-1}} \mathbf{u}_i \mathbf{u}_i^\top \odot \left(\mathbf{I}_m - \frac{1}{m'} \mathbf{1}_m \mathbf{1}_m^\top \right), \tag{57}$$

where $\mathbf{u}_i$ is a m-dimensional one-hot vector whose c-th entry is 1 if $c \in \mathcal{C}_i$ and 0 otherwise, and m' is the number of classes per client.

The above observation states that NC1, NC2, and NC3 happen locally, implying: 1) $h_{i,c} = \mathbf{0}$ if $c \notin \mathcal{C}_i$; and 2) the sub-matrix of $\mathbf{H}_i$ constructed by columns $h_{i,c}$ with $c \in \mathcal{C}_i$ will form a K-Simplex ETF (c.f. Definition B.1) up to some scaling and rotation. We conclude that if there exist $\mathbf{B}$ and $\mathbf{H}_1, \cdots, \mathbf{H}_M$ such that they are the optimal models for all clients, then the data from the same class may be mapped to different points in the feature space by $f_{\mathbf{B}^*}$ when data are drawn from different clients. However, this condition usually cannot be satisfied in the under-parameterized regime, due to the less expressiveness of the under-parameterized model.

To further demonstrate the phenomenon in the under-parameterized regime, we assume that in the under-parameterized regime, a well-performed representation $f_{\mathbf{B}}$ should map data from the same class but different clients to the same feature mean:

Condition 1. For client i and j, if class $c \in \mathcal{C}_i$ and $c \in \mathcal{C}_j$, then $\frac{1}{|\mathcal{D}_i^c|} \sum_{x:(x,y)\in\mathcal{D}_i^c} f_{\mathbf{B}}(x) = \frac{1}{|\mathcal{D}_j^c|} \sum_{x:(x,y)\in\mathcal{D}_j^c} f_{\mathbf{B}}(x)$.

With this condition, we have the following observation that also comes from Theorem 3.1 in Zhu et al. (2021), which describes the neural collapse in the under-parameterized regime.

Observation B.3. *When Condition 1 holds and the feature dimension k is no smaller than the number of total classes m, any global optimizer $\mathbf{B}^*, \mathbf{H}_1^*, \cdots, \mathbf{H}_M^*$ of (54) satisfies*

$$f_{\mathbf{B}^*}(x_1) = f_{\mathbf{B}^*}(x_2), \quad \forall x_1, x_2 \in \mathcal{D}_i^c, i \in [M], c \in \mathcal{C}_i, \tag{58}$$

$$\frac{\langle f_{\mathbf{B}^*}(x), h_{i,c}^* \rangle}{\|f_{\mathbf{B}^*}(x)\| \cdot \|h_{i,c}^*\|} = 1, \quad \forall x \in \mathcal{D}_i^c, i \in [M], c \in \mathcal{C}_i, \tag{59}$$

$$\frac{\mathbf{H}_i^\top \mathbf{H}_i}{\|\mathbf{H}_i^\top \mathbf{H}_i\|_F} = \frac{1}{\sqrt{m-1}} \mathbf{u}_i \mathbf{u}_i^\top \odot \left(\mathbf{I}_m - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^\top \right), \quad \forall i \in [M], \tag{60}$$

where $\mathbf{u}_i$ is a m-dimensional one-hot vector whose c-th entry is 1 if $c \in \mathcal{C}_i$ and 0 otherwise.

Comparing these two observations, we conclude that in the under-parameterized case, the optimal models $h_{i,c}$ and $h_{j,c}$ are of the same direction when class c is included in both $\mathcal{C}_i$ and $\mathcal{C}_j$. It implies that the globally optimized model performs differently compared with the locally learned model. In Figure 2, we present an example to illustrate how $\mathbf{H}$ performs differently when it is globally or locally optimized.

In Figure 2, we consider the scenario that the number of clients $M = 3$, total number of data classes $m = 3$, number of data classes per client $m' = 2$, client 1 contains data of class 1 and class 2, client 2 contains data of class 1 and class 3, and client 3 contains data of class 2 and class 3. The first row of the three sub-figures shows the structure of normalized columns of $\mathbf{H}_1$, $\mathbf{H}_2$, and $\mathbf{H}_3$ when they are locally optimized, and the second row of the three sub-figures shows those optimize (54). We observe that under this setting, the locally optimized heads are in opposite directions, which perform differently compared with the global optimal heads.

Inspired by such observations, we add $\mathcal{NC}_i$ to the local loss function and also optimize $R(\{\mathbf{H}_i\})$, to ensure that the personalized heads also contribute to the global performance. This principle aligns with our motivation to design the linear `FLUTE`.

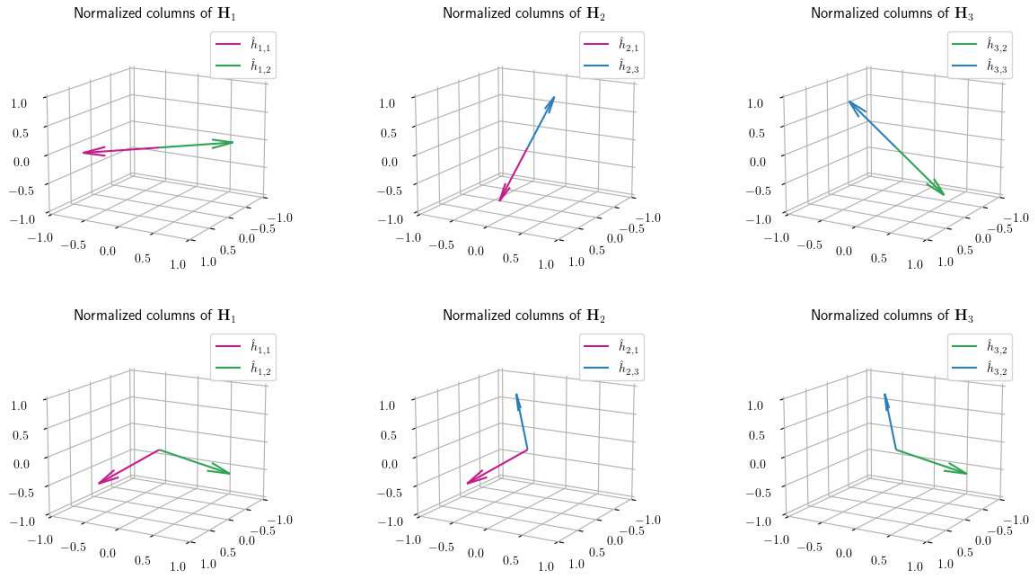

Figure 2. Behavior of locally optimized heads and globally optimized heads.

C. Additional Experimental Results

C.1. Synthetic Datasets

Implementation Details. In the experiments conducted on synthetic datasets shown in Figure 3, $\mathbf{\Lambda} \in \mathbb{R}^{\underline{d} \times \underline{d}}$ is generated by setting the i-th singular value to be $\frac{2d}{i+1}$. We randomly generate $\mathbf{U} \in \mathbb{R}^{d \times \underline{d}}$ with $\underline{d}$ orthonormal columns and $\mathbf{V} \in \mathbb{R}^{\underline{d} \times M}$ with $\underline{d}$ orthonormal rows. The ground-truth model is then $\mathbf{\Phi} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$, where each column ϕ_i represents the local ground-truth model for client i. Each client generates N samples (x, y) from $y = x^\top \phi_i + \xi_i$, where x is sampled from a standard Gaussian distribution and every entry of ξ_i is IID sampled from $\mathcal{N}(0, 0.3)$. The learning rate is set to $\eta = 0.03$, and for random initialization, we set $\alpha = \frac{1}{10d}$.

Parameter Settings. For experiments on synthetic datasets shown in Figure 3, we set $d = 10$. We select the value of k from the set $\{2, 4, 6, 8\}$, M from the set $\{15, 30\}$, and N from the set $\{12, 20\}$.

Experimental Results. From the experiments in Figure 3, we observe that, with the dimensions d, M, and N fixed, an increase in k results in a diminishing discrepancy in convergence speeds between `FLUTE` and FedRep. This trend demonstrates `FLUTE`'s superior performance in under-parameterized settings. Furthermore, keeping d, k, and N unchanged while increasing the number of clients M, we see a reduction in the average error of models generated by `FLUTE`. This observation aligns with our theoretical findings presented in Theorem 5.5.

Varying γ_1 and γ_2. In Figure 4, we report the results of the following experiments where $d = 10$, $k = 6$, $M = 10$, and N selected from the set $\{8, 9, 10, 11\}$. For comparison, we use three pairs of γ_1 and γ_2: $\gamma_1 = 2\gamma_2$, $\gamma_1 = \gamma_2$, and $\gamma_1 = \frac{2}{3}\gamma_2$. We do not set $\gamma_1 > 2\gamma_2$ because in this setting, $\|\mathbf{BW}\|_F$ usually diverges. From the experimental results, we observe that when $N = 8$, 9, or 10, $\gamma_1 = \gamma_2$ shows the best performance among the three settings of γ_1 and γ_2.

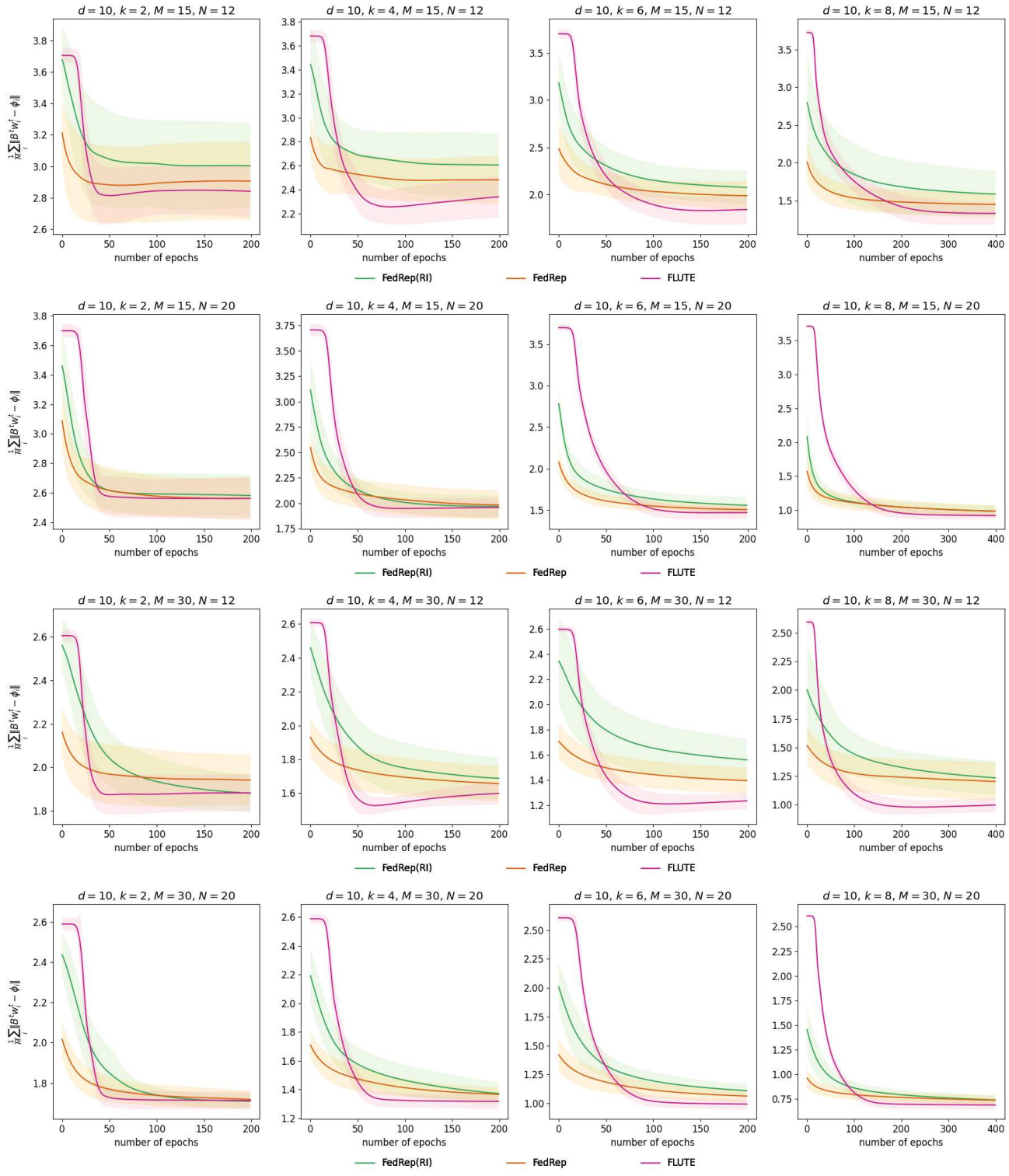

Figure 3. Experimental results with synthetic datasets.

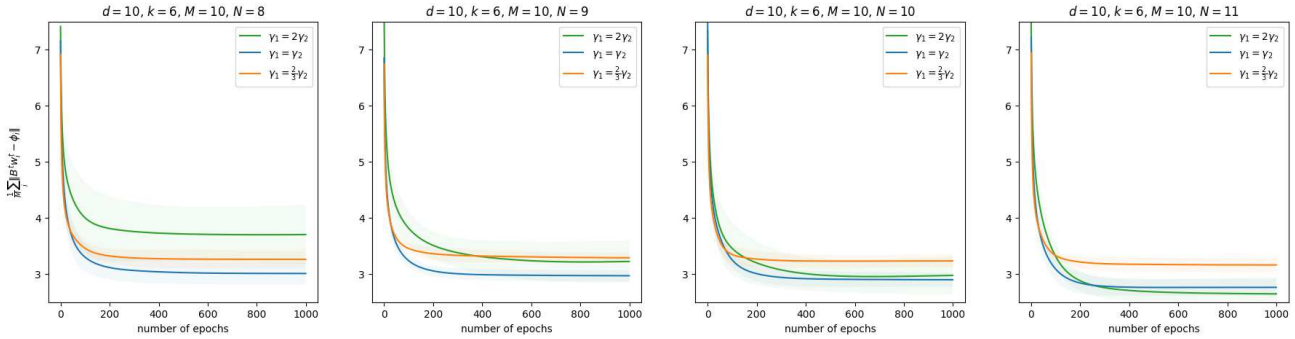

Figure 4. Experimental results with synthetic datasets.

C.2. Real-world Datasets

Implementation Details. For our experiments on the CIFAR-10 dataset, we employ a 5-layer CNN architecture. It begins with a convolutional layer `Conv2d(3, 64, 5)`, followed by a pooling layer `MaxPool2d(2, 2)`. The second convolutional layer is `Conv2d(64, 64, 5)`, which precedes three fully connected layers: `Linear(64*5*5, 120)`, `Linear(120, 64)`, and `Linear(64, 10)`. In contrast, for the CIFAR-100 dataset, we also use a 5-layer CNN, but with some modifications to accommodate the higher complexity of the dataset. The initial layer is `Conv2d(3, 64, 5)`, followed by pooling and dropout layers: `MaxPool2d(2, 2)` and `nn.Dropout(0.6)`. The subsequent convolutional layer is `Conv2d(64, 128, 5)`. This is succeeded by three fully connected layers: `Linear(128*5*5, 256)`, `Linear(256, 128)`, and `Linear(128, 100)`.

Experimental Results. In this section, we plot Figure 5 to Figure 12 to illustrate the detailed convergence behavior of the test accuracy of the trained models reported in Table 1 as a function of the training epochs. We augment the test accuracy results by introducing two different metrics. The first one is *Global NC2*, which is measured by

$$\frac{1}{M}\sum_{i\in[M]}\left\|\frac{\mathbf{H}_i^\top\mathbf{H}_i}{\|\mathbf{H}_i^\top\mathbf{H}_i\|_F}-\frac{1}{\sqrt{m-1}}\mathbf{u}_i\mathbf{u}_i^\top\odot\left(\mathbf{I}_m-\frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top\right)\right\|_F.$$

The second one is *Averaged Local NC2*, referred to as

$$\frac{1}{M}\sum_{i\in[M]}\left\|\frac{\mathbf{H}_i^\top\mathbf{H}_i}{\|\mathbf{H}_i^\top\mathbf{H}_i\|_F}-\frac{1}{\sqrt{m'-1}}\mathbf{u}_i\mathbf{u}_i^\top\odot\left(\mathbf{I}_m-\frac{1}{m'}\mathbf{1}_m\mathbf{1}_m^\top\right)\right\|_F.$$

These two metrics are inspired by Observation B.3 and Observation B.2, respectively. *Global NC2* aims to measure the similarity between the learned models and the optimal under-parameterized global model. In contrast, *Averaged Local NC2* assesses the similarity between the learned models and the optimal local models. Note that these two metrics are positively correlated, meaning that when one is small, the other is usually small as well. In some results, such as those shown in Figure 5 and Figure 7, the gaps between FedRep* and `FLUTE`* in terms of *Averaged Local NC2* are significantly larger than those in terms of *Global NC2*, suggesting that the models learned by `FLUTE`* are closer to the global optimizer than those learned by FedRep*.

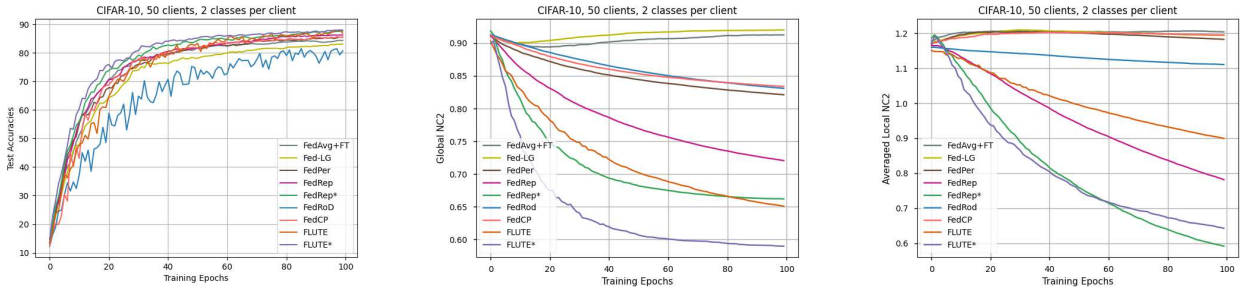

Figure 5. Experimental results for CIFAR10 when $M = 50, m' = 2$.

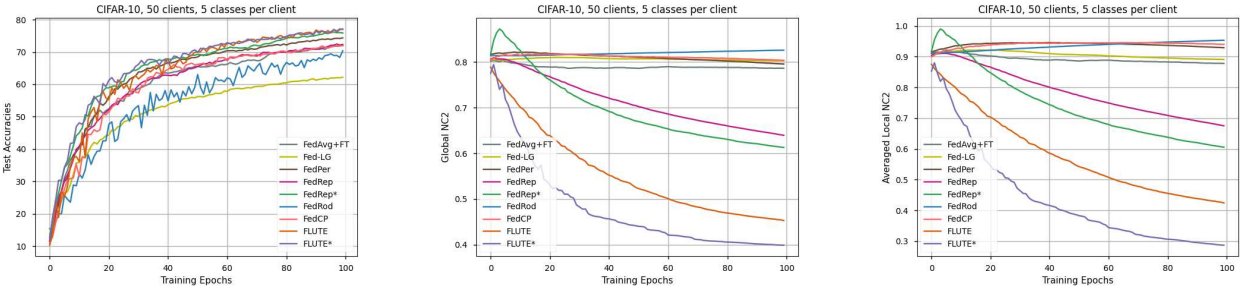

Figure 6. Experimental results for CIFAR10 when $M = 50, m' = 5$.

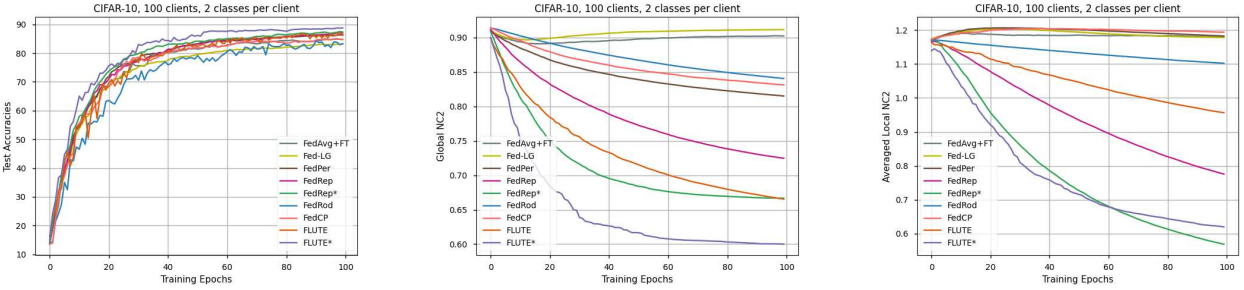

Figure 7. Experimental results for CIFAR10 when $M = 100, m' = 2$.

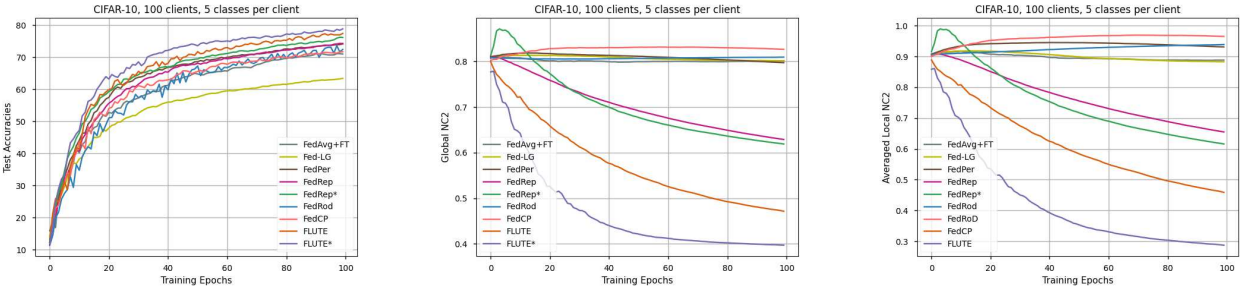

Figure 8. Experimental results for CIFAR10 when $M = 50, m' = 5$.

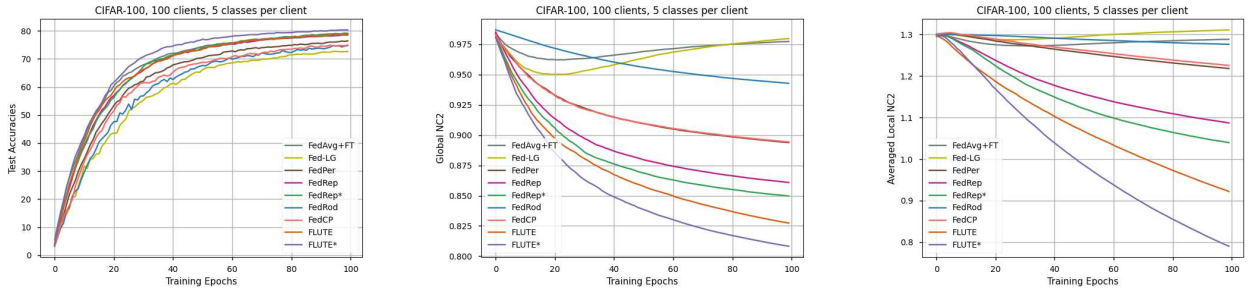

Figure 9. Experimental results for CIFAR100 when $M = 100, m' = 5$.

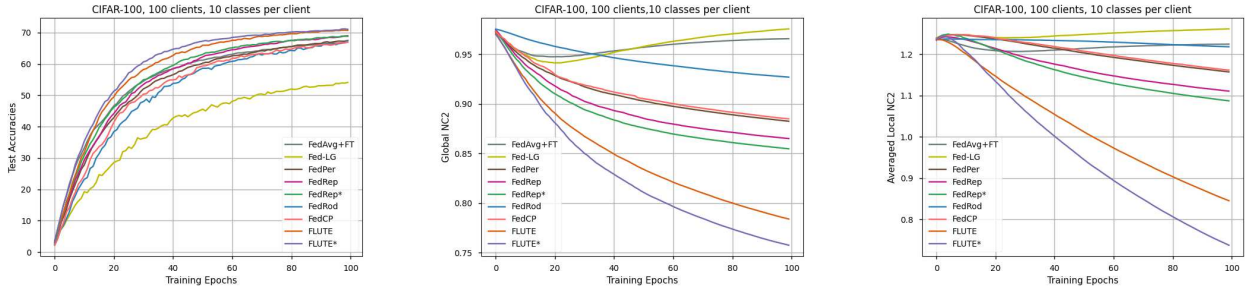

Figure 10. Experimental results for CIFAR100 when $M = 100, m' = 10$.

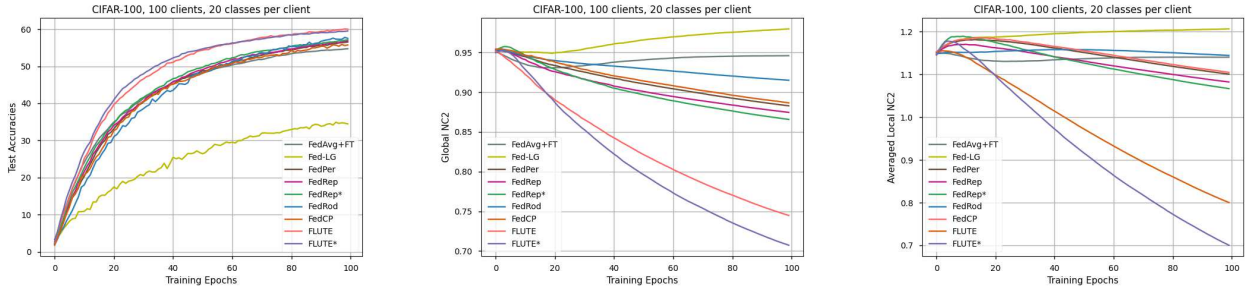

Figure 11. Experimental results for CIFAR100 when $M = 100, m' = 20$.

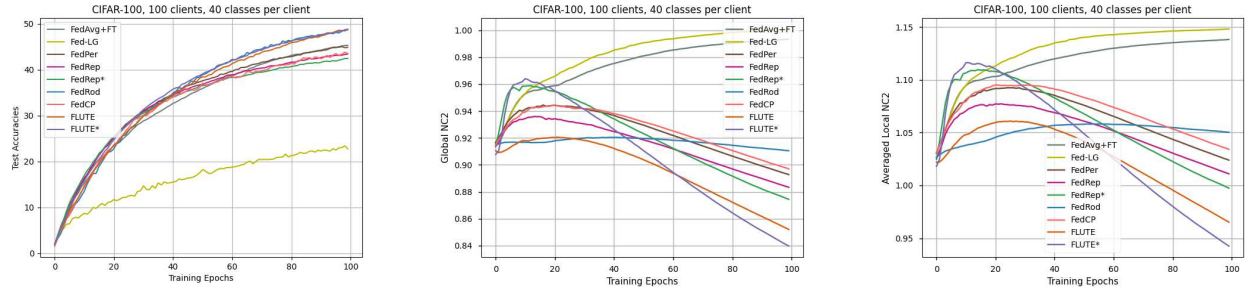

Figure 12. Experimental results for CIFAR100 when $M = 100, m' = 40$.