# ENHANCING CLUSTERED FEDERATED LEARNING: INTEGRATION OF STRATEGIES AND IMPROVED METHODOLOGIES

**Yongxin Guo**[1] **Xiaoying Tang**[1,2,3,*] **Tao Lin**[4,5]
[1]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China
[2]Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS), Shenzhen, China
[3]Guangdong Provincial Key Laboratory of Future Networks of Intelligence, Shenzhen, China
[4]School of Engineering, Westlake University
[5]Research Center for Industries of the Future, Westlake University

## ABSTRACT

Federated Learning (FL) is an evolving distributed machine learning approach that safeguards client privacy by keeping data on edge devices. However, the variation in data among clients poses challenges in training models that excel across all local distributions. Recent studies suggest clustering as a solution to address client heterogeneity in FL by grouping clients with distribution shifts into distinct clusters. Nonetheless, the diverse learning frameworks used in current clustered FL methods create difficulties in integrating these methods, leveraging their advantages, and making further enhancements. To this end, this paper conducts a thorough examination of existing clustered FL methods and introduces a four-tier framework, named HCFL, to encompass and extend the existing approaches. Utilizing the HCFL, we identify persistent challenges associated with current clustering methods in each tier and propose an enhanced clustering method called HCFL$^+$ to overcome these challenges. Through extensive numerical evaluations, we demonstrate the effectiveness of our clustering framework and the enhanced components. Our code is available at https://github.com/LINs-lab/HCFL.

## 1 INTRODUCTION

Federated Learning (FL) is a privacy-focused distributed machine learning approach. In FL, the server shares the model with clients for local training, and the clients send parameter updates back to the server. The clients will not share their raw data with servers, ensuring privacy. However, the non-iid client data distribution leads to significant performance drops for FL algorithms (McMahan et al., 2016; Li et al., 2018; Karimireddy et al., 2020; 2019). To address data heterogeneity, traditional FL focuses on training a single global model that performs well across all local distributions (Li et al., 2021; 2018; Tang et al., 2022; Guo et al., 2023a). However, relying solely on a global model may not adequately handle the heterogeneous client distributions. As a remedy, clustered FL methods have been proposed to group clients into different clusters based on their local distributions [1]. Numerous studies have demonstrated the superiority of clustered FL methods over single-model FL approaches (Long et al., 2023; Sattler et al., 2020b; Ghosh et al., 2020; Marfoq et al., 2021; Guo et al., 2023b).

**Diverse learning frameworks pose challenges on enhancing the clustered FL.** Despite the success of current clustered FL methods, the use of diverse learning frameworks poses challenges in integrating different algorithms, gathering their advantages, and achieving further improvements. For instance, FedEM (Marfoq et al., 2021) excels in addressing complex mixture distribution scenarios and performs admirably on challenging tasks. However, it necessitates a predefined number of clusters, constraining its practicality. In contrast, adaptive clustering techniques such as CFL (Sattler et al., 2020b) can autonomously determine the number of clusters. Nonetheless, CFL cannot be

*Corresponding author.

[1]In this study, we address the issue of supervised clustered FL, which may differ from the unsupervised clustered FL examined by Ding et al. (2023); Qiao et al. (2024).

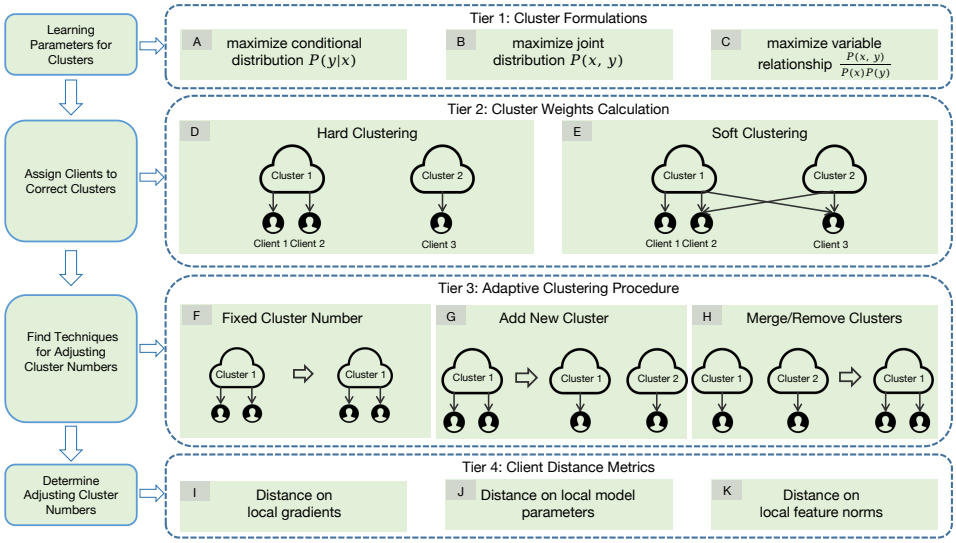



Figure 1: **Overview of the HCFL.** The HCFL encompasses the existing clustered FL algorithms through the design of four tiers, including *cluster formulations*, which maximize conditional distribution, joint distribution, or variable relationships; *cluster weights calculation*, including soft clustering and hard clustering; *adaptive clustering procedure*, including using a predefined number of clusters, automatically adding new clusters, or merge and remove existing clusters; *client distance metrics*, including using distance on clients' local gradients, clients' local model parameters, or clients' local feature norms. The four tiers collaborate to form a comprehensive clustered FL learning process, as shown in the left part of the figure. For instance, CFL can be described by the A, D, G, and J, while A, E, and F cover FedEM.

seamlessly integrated with soft clustering methods like FedEM, thereby limiting its effectiveness in handling complex mixture distribution tasks.

**Consolidating existing methods as a solution.** To tackle these challenges, we believe there is a need to develop a holistic learning framework for supervised clustered FL methods, allowing us to seamlessly combine their advantages. In this paper, we introduce HCFL, a holistic clustered FL algorithm framework incorporating (1) a unified clustering objective function that handles both soft and hard clustering and (2) a unified, four-tier clustering procedure paradigm [2] (as shown in Figure 1).

The HCFL framework enables the flexible combination of existing techniques within each tier, unlocking new benefits beyond the mere recovery of traditional methods. For example, FedRC Guo et al. (2023b) demonstrates strong generalization performance but cannot automatically determine the number of clusters, whereas CFL Sattler et al. (2020a) excels at determining the number of clusters but lacks generalization. The HCFL framework allows for the integration of the strengths of both FedRC and CFL, achieving both superior generalization and personalization performance (see Table 1). Additionally, enabling both cluster removal and addition yields comparable—or even better—performance than baseline algorithms, with a significantly reduced number of clusters (see Table 1).

**Enhancing clustered FL methods by improved methodologies.** In light of the HCFL, we have identified the remaining challenges within each tier that were previously overlooked by existing clustered FL, as illustrated in Figure 2 in Section 4.1. We then introduce HCFL$^+$, an enhanced algorithm designed to tackle these remaining challenges. Numerical results confirm that HCFL$^+$ effectively extends existing methods, achieving a superior balance between personalization and generalization while delivering strong performance. We summarize the contribution of this paper as follows:

- We introduce HCFL, a holistic framework for clustered FL that encompass the existing methods. The HCFL represents a feasible approach in achieving the integration of benefits from existing methods by adjusting the techniques at each tier.

[2]This study focuses on data heterogeneity. While other techniques may enhance privacy and communication efficiency, they are not directly related to our approach.

- We identify four remaining challenges within each tier of HCFL, and introduce an improved algorithm called HCFL$^+$ to address these challenges.
- Extensive experiments on different datasets (CIFAR10, CIFAR100, and Tiny-Imagenet) and various architectures (MobileNet-V2 and ResNet18) demonstrate the effectiveness of our framework and the improved components of HCFL$^+$.

## 2 Related Works

In the field of Federated Learning, FedAvg serves as the de-facto algorithm, employing local Stochastic Gradient Descent (local SGD) techniques (McMahan et al., 2016; Lin et al., 2020) to reduce communication costs and protect client privacy. However, FL faces significant challenges due to distribution shifts among clients, which can hinder the performance of FL algorithms (Li et al., 2018; Wang et al., 2020; Karimireddy et al., 2020; Jiang & Lin, 2023; Guo et al., 2021). To address these challenges, researchers have introduced clustered FL algorithms to enhance the performance of FL algorithms.

Clustered FL groups clients based on their local data distribution, addressing the distribution shift problem. Most methods employ hard clustering with a fixed number of clusters, grouping clients by measuring their similarities (Ghosh et al., 2020; Long et al., 2023; Wang et al., 2022b; Stallmann & Wilbik, 2022; Ma et al., 2022). However, hard clustering may not adequately capture complex relationships between local distributions, and soft clustering paradigms have been proposed to address this issue (Marfoq et al., 2021; Wu et al., 2023; Ruan & Joe-Wong, 2022; Guo et al., 2023b; Reisser et al., 2021; Ruan & Joe-Wong, 2022). In this paper, we propose a generalized formulation for clustered FL that encompasses current methods and improves them by addressing issues related to intra-client inconsistency and efficiency.

Another line of research focuses on automatically determining the number of clusters. Current methods utilize hierarchical clustering (Sattler et al., 2020b;a; Zhao et al., 2020; Briggs et al., 2020; Zeng et al., 2023; Duan et al., 2021a;b), which measures client dissimilarity using model parameters or local gradient distances. Some papers enhance these distance metrics by employing various techniques, such as eigenvectors (Yan et al., 2023) and local feature norms (Wei & Huang, 2023). FEDCOLLAB (Bao et al., 2023) quantifies client similarity through client discriminators. However, the requirement for discriminators between every client pair in FEDCOLLAB hinders scalability for cross-device scenarios with numerous clients. In this paper, we concentrate on cross-device settings, introducing a holistic adaptive clustering framework enabling cluster splitting and merging. We also present enhanced weight updating for soft clustering and finer distance metrics for various clustering principles. For further discussions on related works, please refer to Appendix C.

## 3 HCFL: Revisiting and Extending Clustered FL Methods

Current clustered FL methods typically employ diverse learning frameworks. As a result, existing methods often face challenges in gathering the advantages of different algorithms for potential enhancements. To address this issue, as shown in Figure 1, we introduce the HCFL, consisting of four tiers designed to tackle the primary tasks of clustering methods: (1) Cluster Learning and Assignment (tiers 1 and 2): Identify which clients should belong to the same clusters and learns parameters for each cluster. (2) Cluster Number Determinant (tiers 3 and 4): Decide the number of clusters. As a result, the four tiers of the HCFL form a comprehensive learning process (Algorithm 1), enabling flexible improvements and the integration of advantages from different algorithms (Table 1).

### 3.1 Tiers 1 & 2: The Cluster Formulations and Cluster Weights Calculation

We introduce the first two tiers: Cluster Formulations and Cluster Weights Calculation. Cluster Formulations defines the objective functions of the clustering methods, aiming to learn the underlying distributions of each cluster. Cluster Weights Calculation orthogonally helps find the suitable clusters for each client, whereas hard clustering assigns each client to one cluster, while soft clustering allows clients to contribute to multiple clusters. We propose the following optimization framework to encompass these two tiers.

**Optimization framework of clustered FL methods.** The clustered FL methods can be expressed as a dual-variable optimization problem that maximizes $\mathcal{L}(\mathbf{\Theta}, \mathbf{\Omega})$, with $K$ clusters and $M$ data sources represented as $\mathcal{D}_1, \cdots, \mathcal{D}_M$:

$$\begin{aligned} &\mathcal{L}(\mathbf{\Theta}, \mathbf{\Omega}) = \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \log \left( \sum_{k=1}^{K} \omega_{i;k} \mathcal{L}_k(\mathbf{x}_{i,j}, y_{ij}; \boldsymbol{\theta}_k) \right) , \\ &\text{s.t.} \quad \sum_{k=1}^{K} \omega_{i;k} = 1, \forall i , \end{aligned} \tag{1}$$

where $N = \sum_{i=1}^{M} N_i$, $N_i := |\mathcal{D}_i|$, $(\mathbf{x}_{i,j}, y_{ij})$ are sampled from $\mathcal{D}_i$, and $\mathcal{L}_k(\mathbf{x}_{i,j}, y_{ij}; \boldsymbol{\theta}_k)$ is the local objective function. The parameters to be optimized are clustering weights $\mathbf{\Omega} = [\omega_{1;1}, \cdots, \omega_{M,K}]$, and model parameters $\mathbf{\Theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K]$.

**Tier 1: Incorporate existing Cluster Formulations.** The existing methods employ clustering to address diverse tasks, which results in the proposal of various formulations. Our Algorithm 1 encompasses the existing clustering formulations by selecting $\mathcal{L}_k(\mathbf{x}_{i,j}, y_{ij}; \boldsymbol{\theta}_k)$ as follows:

- $\mathcal{P}_{\boldsymbol{\theta}_k}(y_{i,j}|\mathbf{x}_{i,j})$. Most existing methods (Marfoq et al., 2021; Ghosh et al., 2020; Long et al., 2023) can be recovered using this conditional distribution (likelihood functions).
- $\mathcal{P}_{\boldsymbol{\theta}_k}(\mathbf{x}_{i,j}, y_{i,j})$. FedGMM (Wu et al., 2022) uses joint probability. [3]
- $\frac{\mathcal{P}_{\boldsymbol{\theta}_k}(\mathbf{x}_{i,j}, y_{i,j})}{\mathcal{P}_{\boldsymbol{\theta}_k}(\mathbf{x}_{i,j})\mathcal{P}_{\boldsymbol{\theta}_k}(y_{i,j})}$. FedRC (Guo et al., 2023b) relies on correlations between variables $\mathbf{x}$ and $y$.

**Tier 2: Incorporate existing Cluster Weights Calculation.** Various methods employ distinct mechanisms for calculating clustering weights $\omega_{i;k}$. The choice of $\omega_{i;k}$, with either binary values $\omega_{i;k} \in \{0, 1\}$ or continuous values $\omega_{i;k} \in [0, 1]$, characterizes the dynamic clustering procedure.

- Hard clustering methods employ binary values $\omega_{i;k} \in 0, 1$. In these methods, $\omega_{i;k}$ is determined using heuristic techniques, such as parameter distance (Long et al., 2023; Zeng et al., 2023; Sattler et al., 2020b), local loss function values (Ghosh et al., 2020) or label distributions (Diao et al., 2024).
- Soft clustering approaches permit $\omega_{i;k} \in [0, 1]$, determined by maximizing $\mathcal{L}(\mathbf{\Theta}, \mathbf{\Omega})$(Marfoq et al., 2021; Guo et al., 2023b; Wu et al., 2023), or by normalizing local loss values(Ruan & Joe-Wong, 2022). Soft clustering methods do not assume separated clients' local distributions and can thus handle complex scenarios, such as mixture distributions (Marfoq et al., 2021; Wu et al., 2023).

### 3.2 Tiers 3 & 4: The Adaptive Clustering Procedure and Distance Metrics

Tiers 3 and 4 illustrate the techniques for Cluster Number Determination. In detail, the adaptive clustering procedures automatically adjust the number of clusters, while distance metrics control the adaptive clustering procedures, determining whether clusters should split or merge. The HCFL allows for different techniques at each tier, enhancing flexibility in choosing the optimal adaptive clustering methods or converting methods that rely on fixed cluster numbers to adaptive ones.

**Tier 3: Adaptive clustering procedures demonstrate how to modify cluster numbers.** To automatically determine the number of clusters, current approaches can be categorized into two orthogonal methods: (1) Splitting clusters to increase the number of clusters (Sattler et al., 2020b;a). (2) Merging clusters to reduce the number of clusters (Zeng et al., 2023). We unify these approaches at tier 3.

**Tier 4: Client distance metrics dictate when cluster numbers should be adjusted.** The client's distances are utilized to determine whether the current number of clusters should be adjusted. For instance, when the distance within a cluster is large, the cluster will divide into sub-clusters. Conversely, if the distances between two clusters are small, these two clusters should be merged. Existing clustering methods use various metrics such as cosine similarity of local gradients (Sattler et al., 2020b), gradients from a globally shared network (Zeng et al., 2023), and local feature norms (Wei & Huang, 2023).

## 4 HCFL$^+$: Tackling Remaining Challenges in Clustered FL

Section 3 introduces a holistic clustering framework with four tiers to encompass existing methods. However, each tier still presents challenges that current methods cannot address. In this section, we

[3] In FedGMM (Wu et al., 2023), $\boldsymbol{\theta}_k$ is split into $[\boldsymbol{\theta}_{k_1}, \boldsymbol{\nu}_{k_2}]$, and it uses $\mathcal{L}_k(\mathbf{x}_{i,j}, y_{i,j}; \boldsymbol{\theta}_k) = \mathcal{P}_{\boldsymbol{\theta}_{k_1}}(y_{i,j}|\mathbf{x}_{i,j})\mathcal{P}_{\boldsymbol{\nu}_{k_2}}(\mathbf{x}_{i,j})$ to model the joint probability.

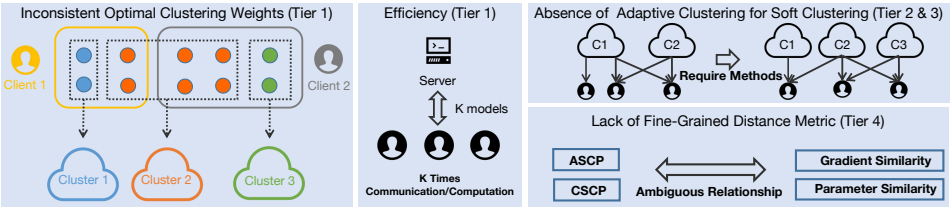



Figure 2: **Remaining challenges in clustered FL methods.** We identify four key issues in clustered FL algorithms: (1) inconsistent intra-client clustering weights, (2) efficiency concerns, (3) the absence of adaptive clustering for soft clustering methods, and (4) the lack of fine-grained distance metrics for various clustering principles. Clustering principles ASCP and CSCP differ in their approach as follows: ASCP assigns clients with any shifts into different clusters, while CSCP only assigns clients with concept shifts to different clusters."

outline four key remaining challenges in Figure 2 and introduce HCFL$^+$ to tackle them. Due to space constraints, we summarize the improved algorithm in Algorithm 2.

While the HCFL$^+$ employs standard FL training process without introducing extra privacy concerns compared to traditional methods. Existing privacy-preserving techniques, like secure aggregation, can be integrated as plugins. We plan to explore this possibility further in the future work.

### 4.1 Remaining Challenges of the Clustering in FL

In this subsection, we identify four remaining challenges of the HCFL. We categorize these challenges by tiers in the HCFL, as shown in Figure 2. The details are provided below.

**Challenges on tier 1: Inconsistent intra-client clustering weights and efficiency concerns.** These challenges can be addressed by improving the clustering formulations.

- *Inconsistent intra-client clustering weights.* Existing approaches use the same clustering weights $\omega_{i;k}$ for all the samples belonging to client $i$ (Sattler et al., 2020b; Ghosh et al., 2020; Marfoq et al., 2021; Guo et al., 2023b). However, they overlook cases where the optimal clustering weights of different samples within the same client can be inconsistent, implying that $\omega_{i,j_1;k} \neq \omega_{i,j_2;k}$ for certain samples $(\mathbf{x}_{i,j_1}, y_{i,j_1})$ and $(\mathbf{x}_{i,j_2}, y_{i,j_2})$. See our example here[4].
- *Efficiency.* The current clustered FL methods (Marfoq et al., 2021; Long et al., 2023; Guo et al., 2023b) require K-fold higher communication or computation costs, hindering overall algorithm efficiency during deployment.

**Challenges on tiers 2 & 3: The absence of adaptive clustering for soft clustering methods.** Current adaptive clustering methods primarily address hard clustering (Sattler et al., 2020b;a; Zeng et al., 2023). Hence, there exists a gap between research and practice, as there is a need to automatically determine the number of clusters for soft clustering methods (Marfoq et al., 2021; Guo et al., 2023b).

**Challenges on tier 4: Lack of fine-grained distance metrics for various clustering principles.** The clustering principles determine which clients should be assigned to the same clusters. Existing clustering methods may use different clustering principles, as described by ASCP and CSCP below:

- ASCP (Any Shift Type Clustering Principle): clients with any distribution shifts are placed into separate clusters (Marfoq et al., 2021; Wu et al., 2023).
- CSCP (Concept Shift Only Clustering Principle): only clients with concept shifts are assigned to separate clusters (Guo et al., 2023b).

As discussed in Section 3.2, client distances determine whether the current number of clusters should be changed, aligning with the role of clustering principles: if the current number of clusters cannot meet the requirements of the clustering principles, the cluster number should be adjusted. Consequently, we advocate for distance metrics to be closely tied to distribution shifts, ultimately aligning with clustering principles. Unfortunately, existing distance metrics, such as those based on local gradients or local model parameters (Sattler et al., 2020b; Zeng et al., 2023; Long et al., 2023;

[4]In real-world scenarios, clients may have varying local data distributions, but they can still share certain data, like common knowledge or location information. While client-specific data may belong to different clusters, shared data should be in the same clusters, leading to inconsistent intra-client clustering weights.

**Algorithm 1** Algorithm Framework of HCFL.

**Require:** Number of communication rounds $T$, initial number of clusters $K^0$, initial parameters $\boldsymbol{\phi}^0$, and $\boldsymbol{\Theta}^0$.
**Ensure:** Number of clusters $K^T$, trained parameters $\boldsymbol{\phi}^T$, and $\boldsymbol{\Theta}^T$.

```
for t = 0, ··· , T − 1 do
    Sample a subset of clients S^t, and send Θ^{t+1} to the clients.
    for Client i in S^t do
        Do local SGD by solving (1) for τ local epochs..                ▷ Tiers 1 and 2
        Upload local model updates g_{i;k}^{t+1} to the server.
    θ_k^{t+1} = θ_k^t − η_g Σ_{i∈S^t} g_{i;k}^{t+1}, ∀k.
    Calculate distance matrix D^t, and D_k^t, ∀k.                       ▷ Tier 4
    if Detect cluster k_s need to be split then                         ▷ Tier 3
        Split cluster k_s into sub-clusters S_{s,1} and S_{s,2}.
        θ_{k_s}^{t+1} = θ_k^t − η_g Σ_{i∈S_{s,1}} g_{i;k}^{t+1}.
        θ_{K^t+1}^{t+1} = θ_k^t − η_g Σ_{i∈S_{s,2}} g_{i;k}^{t+1}.
        Update ω_{i;k} for corresponding clients.
    if Detect cluster k_d need to be deleted then                       ▷ Tier 3
        Delete cluster k_d.
        Update ω_{i;k} for corresponding clients.
    Update K^{t+1} by the current number of clusters.
```

Yan et al., 2023), cannot establish a clear link to distribution shifts. As a result, current methods struggle to satisfy diverse and detailed clustering principles.

### 4.2 Improve Tier1: Inconsistency and Efficiency Aware Objective Functions

To address tier 1 challenges, specifically, (i) inconsistent intra-client clustering weights and (ii) efficiency, we propose an extension of the objective function (Eq. (1)), which is defined as $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}})$ and includes the parameters $\boldsymbol{\phi}$, $\boldsymbol{\Theta}$, $\boldsymbol{\Omega}$, and $\tilde{\boldsymbol{\Omega}}$.

- *Shared feature extractor $\boldsymbol{\phi}$, and cluster-specific predictors $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K]$.* Dividing the feature extractor $\boldsymbol{\phi}$ and the predictors $\{\boldsymbol{\theta}_k\}$ reduces communication and computation costs since the predictors are lightweight architectures, like linear classifier layers.
- *Sample-wise clustering weights $\boldsymbol{\Omega} = [\omega_{1,1;1} \cdots, \omega_{M,N_M;K}]$ for enhanced training stage and client-wise clustering weights $\tilde{\boldsymbol{\Omega}} = [\tilde{\omega}_{1;1} \cdots, \tilde{\omega}_{M;K}]$ for testing stage* [5]. We employ sample-specific clustering weights ($\omega_{i,j;k}$) during training to ensure that data samples from the same clients can contribute to different cluster models, resolving the issue of inconsistent intra-client clustering weights. Furthermore, during testing, when test-time label information is unavailable, we utilize client-specific weights ($\tilde{\omega}_{i;k}$) for each client and cluster.
- *Enhance the optimization of $\omega_{i,j;k}$ by regularizing the distance between $\tilde{\omega}_{i;k}$ and $\omega_{i,j;k}$.* Motivated by the intuition that "if data from the same clients have similar distributions, the corresponding clustering weights should be similar", we encourage $\tilde{\omega}_{i;k}$ and $\omega_{i,j;k}$ to be close to each other.

The following objective function is designed to meet our requirements.

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}}) = \underbrace{\frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \log \left( \sum_{k=1}^{K} \omega_{i,j;k} \mathcal{L}_k(\mathbf{x}_{i,j}, y_{ij}) \right)}_{\mathcal{A}_1} - \underbrace{\mu \sum_{i=1}^{M} \sum_{j=1}^{N_i} \left( \sum_{k=1}^{K} \tilde{\omega}_{i;k} \log \frac{\tilde{\omega}_{i;k}}{\omega_{i,j;k}} \right)}_{\mathcal{A}_2} \tag{2}$$

$$\text{s.t.} \sum_{k=1}^{K} \omega_{i,j;k} = 1, \forall i, j, \sum_{k=1}^{K} \tilde{\omega}_{i;k} = 1, \forall i, \tilde{\boldsymbol{\Omega}} = \arg\min_{\tilde{\boldsymbol{\Omega}}} \left| \max_{\boldsymbol{\Omega}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}}) - \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \tilde{\boldsymbol{\Omega}}, \tilde{\boldsymbol{\Omega}}) \right|, \tag{3}$$

where $\mathcal{L}_k(\mathbf{x}_{i,j}, y_{ij})$ is the simplified notation of $\mathcal{L}_k(\mathbf{x}_{i,j}, y_{ij}; \boldsymbol{\phi}, \boldsymbol{\theta}_k)$. $\mathcal{A}_1$ term is extended from (1) by using the global shared feature extractor $\boldsymbol{\phi}$ and the sample-wise weights $\omega_{i,j;k}$. $\mathcal{A}_2$ focuses on regularizing the difference between the sample-wise clustering weights $\omega_{i,j;k}$ and the client-wise clustering weights $\tilde{\omega}_{i;k}$. The $\mu$ controls the strength of this regularization. We obtain $\tilde{\omega}_{i;k}$ by solving (3), where we aim to minimize the impact of replacing $\omega_{i,j;k}$ with $\tilde{\omega}_{i;k}$.

[5] Experiments on the effectiveness of sample-wise clustering weights in Figures 3a and 3b.

**Optimization of the proposed objective function.** Different from heuristic methods used in most studies to optimize (2) [6], we aim to introduce a more interpretable approach by maximizing the objective functions (Eq. (2)). Detailed proof refer to Appendix A. Specifically, we can update $\tilde{\omega}_{i;k}$, $\omega_{i,j;k}$, $\boldsymbol{\theta}_k$, and $\boldsymbol{\phi}$ by (4)–(7).

$$\gamma_{i,j;k}^{t+1} = \frac{\omega_{i,j;k}^{t}\mathcal{L}_k^t(\mathbf{x}_{i,j},y_{i,j})}{\sum_{n=1}^{K}\omega_{i,j;n}^{t}\mathcal{L}_k^t(\mathbf{x}_{i,j},y_{i,j})}\,,\tilde{\gamma}_{i,j;k}^{t+1} = \frac{\tilde{\omega}_{i;k}^{t}\mathcal{L}_k^t(\mathbf{x}_{i,j},y_{i,j})}{\sum_{n=1}^{K}\omega_{i;n}^{t}\mathcal{L}_k^t(\mathbf{x}_{i,j},y_{i,j})}\,, \tag{4}$$

$$\tilde{\omega}_{i;k}^{t+1} = \frac{1}{N_i}\sum_{j=1}^{N_i}\tilde{\gamma}_{i,j;k}^{t+1}\,,\omega_{i,j;k}^{t+1} = \frac{\gamma_{i,j;k}^{t+1}}{1+\mu N} + \frac{\mu N}{1+\mu N}\tilde{\omega}_{i;k}^{t+1} = \tilde{\mu}\gamma_{i,j;k}^{t+1} + (1-\tilde{\mu})\tilde{\omega}_{i;k}^{t+1}\,, \tag{5}$$

$$\boldsymbol{\theta}_k^{t+1} = \boldsymbol{\theta}_k^{t} - \eta\sum_{i=1}^{M}\sum_{j=1}^{N_i}\frac{\gamma_{i,j;k}^{t+1}}{\mathcal{L}_k^t(\mathbf{x}_{i,j},y_{i,j})}\nabla_{\boldsymbol{\theta}_k}\mathcal{L}_k^t(\mathbf{x}_{i,j},y_{i,j})\,, \tag{6}$$

$$\boldsymbol{\phi}^{t+1} = \boldsymbol{\phi}^{t} - \eta\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\frac{\gamma_{i,j;k}^{t+1}}{\mathcal{L}_k^t(\mathbf{x}_{i,j},y_{i,j})}\nabla_{\boldsymbol{\phi}}\mathcal{L}_k^t(\mathbf{x}_{i,j},y_{i,j})\,, \tag{7}$$

where $\mathcal{L}_k^t(\mathbf{x}_{i,j},y_{ij})$ is the simplified notation of local objective function $\mathcal{L}_k(\mathbf{x}_{i,j},y_{ij};\boldsymbol{\phi}^t,\boldsymbol{\theta}_k^t)$, $\gamma_{i,j;k}$ and $\tilde{\gamma}_{i,j;k}$ are intermediate results for calculating $\omega_{i,j;k}$ and $\tilde{\omega}_{i;k}$. More detailed proofs can be found in Appendix A. $\tilde{\mu} = \frac{1}{1+\mu N}$ serves as a hyperparameter to control the strength of the penalty term in Equation (2). We provide the theoretical analysis on linear representation learning case in Appendix B.

### 4.3 Improve Tiers 2 & 3: Adaptive Clustering for Soft Clustering Paradigms

Given the limitations of existing adaptive clustering methods, we have extended the clustering weight update mechanisms to incorporate soft clustering and have verified its effectiveness in Figures 3c and 3d. The overall process is summarized in Algorithms 4 and 5. In Algorithm 4, the clustering weights are adjusted after splitting cluster $k$ into two sub-clusters, denoted by $k_1$ and $k_2$. Then we set $\omega_{i,j,k_1} = \omega_{i,j,k_2} = \omega_{i,j,k}/2$ for all $i$ and $j$. In Algorithm 5, the clustering weights are updated when removing cluster $k$. For all $k^{'} \neq k$, we modify $\omega_{i,j;k'}$ as $\omega_{i,j;k'} = \frac{\omega_{i,j;k'}}{\sum_{n\neq k}\omega_{i,j;n}}$.

We use the hyperparameter $\rho$ to control cluster splitting. As evidenced in Table 1, a higher $\rho$ results in fewer clusters, signifying enhanced generalization but reduced personalization. In detail, the cluster $k$ will split if the following condition is met:

$$\max(\mathbf{D}_k) - \text{mean}(\mathbf{D}_k) \geq \rho\,, \tag{8}$$

where $\mathbf{D}_k$ is the distance matrix of cluster $k$. We identify the need for cluster removal when the cluster no longer receives the highest clustering weights from any clients. Additional details about the enhanced adaptive process can be found in Algorithm 2.

### 4.4 Improve Tier4: Fine-Grained Distance Metric Design

Due to the page limitations, we include most of the details about the method design and practical implements in Appendix D. As discussed in Section 4.1, various algorithms may group clients into different clusters based on different clustering principles. Therefore, in this section, we design the following fine-grained distance metrics for these different clustering principles. In detal, we have $\mathbf{D}_{i,j}^k =$

$$\begin{cases} \max\{d_c, d_{lf}\}\,\mathbb{E}_{D_i}\left[\tilde{\mathcal{L}}_k^t(\mathbf{z},y)\right]\mathbb{E}_{D_j}\left[\tilde{\mathcal{L}}_k^t(\mathbf{z},y)\right]\,, & \text{ASCP}\,, \\ d_c\mathbb{E}_{D_i}\left[\tilde{\mathcal{L}}_k^t(\mathbf{z},y)\right]\mathbb{E}_{D_j}\left[\tilde{\mathcal{L}}_k^t(\mathbf{z},y)\right]\,, & \text{CSCP}\,, \end{cases} \tag{9}$$

where

$$d_c = \max_y\left\{\text{dist}\left(\mathbb{E}_{D_i}\left[\mathcal{P}(\mathbf{z}|\mathbf{x},y;\boldsymbol{\phi})\right],\mathbb{E}_{D_j}\left[\mathcal{P}(\mathbf{z}|\mathbf{x},y;\boldsymbol{\phi})\right]\right)\right\}\,,\, d_{lf} = \text{dist}\left(\mathbb{E}_{D_i}\left[\mathcal{P}(\mathbf{z}|\mathbf{x};\boldsymbol{\phi})\right],\mathbb{E}_{D_j}\left[\mathcal{P}(\mathbf{z}|\mathbf{x};\boldsymbol{\phi})\right]\right)$$

where $\tilde{\mathcal{L}}_k^t(\mathbf{z},y)$ is the simplified notation of local objective function $\tilde{\mathcal{L}}_k^t(\mathbf{z},y;\boldsymbol{\theta}_k)$ given extracted feature $\mathbf{z}$, dist is the cos-similarity. The distances above become large only when the following

[6] IFCA (Ghosh et al., 2020) sets $\omega_{i,j;k_{i,\text{min}}} = 1, \forall j$ when $k_{i,\text{min}} = \arg\min_k \mathbb{E}_{D_i}\left[f_{i;k}(\mathbf{x}_{i,j},y_{i,j},\boldsymbol{\phi},\boldsymbol{\theta}_k)\right]$, where $f_{i;k}$ is the local loss function. FeSEM (Long et al., 2023) sets $k_{i,\text{min}} = \arg\min_k \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_i\|_2$, where $\boldsymbol{\theta}_k, \boldsymbol{\theta}_i$ represents the model parameters of cluster $k$ and client $i$, respectively.

conditions occur together: (1) Large values of $d_c$ indicate concept shifts between clients $i$ and $j$; (2) Large $d_{lf}$ indicate significant feature and label distribution differences. (3) Large values of $\mathbb{E}_{D_i}\left[\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)\right] \mathbb{E}_{D_j}\left[\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)\right]$ indicate incorrect clustering weights with high confidence. The effectiveness of the above distance metrics design is evidenced in Table 2.

**Approximation of the distance metrics in practice.** When calculating the distance metrics (Equation (9)) in practice, to avoid training extra generative networks and transmitting more data between servers and clients, we substitute $\tilde{\omega}_{i;k}$ for $\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)$ since $\tilde{\omega}_{i;k}$ is positively correlated with $\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)$ (Marfoq et al., 2021; Guo et al., 2023b). Additionally, we approximate $\mathbb{E}_{D_i}\left[\mathcal{P}(\mathbf{z}|\mathbf{x}, y; \boldsymbol{\phi})\right]$ and $\mathbb{E}_{D_i}\left[\mathcal{P}(\mathbf{z}|\mathbf{x}, y; \boldsymbol{\phi})\right]$ using feature prototypes. The prototypes are defined by the following equation:

$$\tilde{d}_c = Dist(\mathbf{P}_{c,i}, \mathbf{P}_{c,j}), \tilde{d}_{lf} = Dist(\mathbf{P}_{lf,i}, \mathbf{P}_{lf,j}), \tag{10}$$

where

$$\mathbf{P}_{c,i} \in \mathbb{R}^{d \times C} = [\frac{1}{N_{i,1}} \sum_{j=1}^{N_i} \mathbf{1}_{y_{i,j}=1} g(\mathbf{x}_{i,j}, \boldsymbol{\phi}), \cdots, \frac{1}{N_{i,C}} \sum_{j=1}^{N_i} \mathbf{1}_{y_{i,j}=C} g(\mathbf{x}_{i,j}, \boldsymbol{\phi})], \tag{11}$$

$$\mathbf{P}_{lf,i} \in \mathbb{R}^{d} = \frac{1}{N_i} \sum_{j=1}^{N_i} g(\mathbf{x}_{i,j}, \boldsymbol{\phi}), \tag{12}$$

$N_{i,c} = \sum_{j=1}^{N_i} \mathbf{1}_{y_{i,j}=c}$, $g(\mathbf{x}_{i,j}, \boldsymbol{\phi})$ is the function parameterized by $\boldsymbol{\phi}$, $Dist$ is a function to measure the distance between prototypes, which we use the cosine similarity as an example in this paper.

## 5 Numerical Results

In this section, we evaluate the performance of HCFL$^+$ and other clustered FL methods. Additional experiment results, including hyper-parameter ablation studies, different model architectures, additional scenarios, efficiency comparision, and clusteirng quality illustration can be found in Appendix E.

### 5.1 Datasets and Experiment settings

**Diverse distribution shifts scenarios.** We establish clients with three types of distribution shifts. For label distribution shifts, we employ LDA with $\alpha = 1.0$, as introduced by (Yoshida et al., 2019; Hsu et al., 2019; Reddi et al., 2021). For feature distribution shifts, we adopt the methodology from CIFAR10-C and CIFAR100-C creation (Hendrycks & Dietterich, 2019). Regarding concept shift, we draw inspiration from (Guo et al., 2023b; Jothimurugesan et al., 2023), and selectively swap labels based on the parameter $\beta$. For example, with $\beta = 0.1$ for CIFAR10, two labels per concept are swapped, while the remaining eight labels remain unchanged. By default, we create three concepts in the experiments. More details about the construction of scenarios are included in Appendix E.1.

**Baselines.** We use FedAvg (McMahan et al., 2016) as a single-model FL example. We consider the most recently published clustered FL methods as our baselines. For clustered FL with fixed cluster number, we select IFCA (Ghosh et al., 2020), FedEM (Marfoq et al., 2021), FeSEM (Long et al., 2023), and FedRC (Guo et al., 2023b). For the adaptive clustering FL methods, we choose CFL (Sattler et al., 2020b), ICFL (Yan et al., 2023), and StoCFL (Zeng et al., 2023). For the HCFL$^+$ variants, we enhance backbones that necessitate a predetermined number of clusters (FeSEM, FedEM, and FedRC) by integrating them with the adaptive soft clustering method and the feature extractor-classifier split mechanism of HCFL$^+$. The HCFL$^+$ (FedSoft) is devised by substituting the local training component of HCFL$^+$ (FedEM) with proximal regularized local updates as outlined in FedSoft (Ruan & Joe-Wong, 2022). Further elaboration can be found in Appendix E.2.

**Experiment settings.** Unless specifically mentioned, we divide the datasets into 100 clients and execute all algorithms for 200 communication rounds. Additional settings are provided in Appendix E.2. We conducted all experiments using MobileNet-V2 (Sandler et al., 2018) and results on ResNet18 defer to Table 8 of Appendix E.

Table 1: **Performance of the adaptive clustering methods** on CIFAR10, CIFAR100, and Tiny-Imagenet datasets. For each algorithm, we present the best Validation and Test accuracies averaged over three trials. For clustering methods that require a fixed number of clusters, we set $K = 3$. The hyperparameters $\text{tol}_1$, $\text{tol}_2$, $\alpha^*(0)$, $\tau$, and $\rho$ in adaptive clustering methods govern the balance between personalization and generalization, as well as the cluster number. For instance, lower $\tau$ in StoCFL or lower $\rho$ in HCFL$^+$ indicate improved personalization and reduced generalization. $K^T$ denotes the cluster number in the final training round, where a larger $K^T$ suggests improved personalization but reduced generalization and communication efficiency. We emphasize the best results in **bold** and the worst results in blue.

| Algorithm | CIFAR10, $\beta = 0.2$ | | | CIFAR100, $\beta = 0.2$ | | | Tiny-Imagenet, $\beta = 0.2$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Val | Test | $K^T$ | Val | Test | $K^T$ | Val | Test | $K^T$ |
| **Pre-defined** $K$ | | | | | | | | | |
| FedAvg | 49.19 ±2.15 | 45.42 ±2.42 | 1 | 26.01 ±1.15 | 27.87 ±2.12 | 1 | 38.83 ±0.20 | 39.07 ±0.44 | 1 |
| FeSEM | 45.30 ±0.40 | 29.01 ±0.79 | 3 | 26.37 ±0.64 | 24.50 ±0.28 | 3 | 37.10 ±0.80 | 30.00 ±2.02 | 3 |
| IFCA | 34.46 ±2.06 | 23.18 ±2.55 | 3 | 26.99 ±3.89 | 26.20 ±1.56 | 3 | 38.52 ±0.30 | 29.92 ±0.30 | 3 |
| FedEM | 66.49 ±0.69 | 53.64 ±1.61 | 3 | 29.75 ±0.47 | 24.18 ±0.03 | 3 | 42.00 ±0.74 | 39.25 ±0.31 | 3 |
| FedRC | 63.65 ±2.95 | 59.41 ±0.19 | 3 | 34.56 ±0.79 | **37.62** ±0.16 | 3 | 38.93 ±0.18 | 39.73 ±0.04 | 3 |
| **Adaptive** $K$ | | | | | | | | | |
| CFL | | | | | | | | | |
| $\text{tol}_1 = 0.4, \text{tol}_2 = 1.6$ | 61.55 ±1.74 | 46.88 ±0.35 | 6 | 35.05 ±0.35 | 24.84 ±2.50 | 4 | 37.41 ±1.87 | 30.25 ±0.55 | 3 |
| $\text{tol}_1 = 0.4, \text{tol}_2 = 0.8$ | 65.06 ±3.34 | 45.74 ±4.01 | 9 | 36.98 ±3.37 | 22.00 ±1.88 | 5 | 40.36 ±3.55 | 28.82 ±0.71 | 4 |
| $\text{tol}_1 = 0.2, \text{tol}_2 = 0.8$ | 58.92 ±2.09 | 55.02 ±0.97 | 4 | 37.73 ±7.68 | 31.47 ±0.09 | 3 | 35.74 ±0.57 | 34.41 ±1.92 | 1 |
| ICFL | | | | | | | | | |
| $\alpha^*(0) = 0.85$ | 77.59 ±0.04 | 57.38 ±1.91 | 98 | 52.73 ±1.03 | 32.77 ±0.28 | 100 | 64.72 ±0.30 | 34.73 ±0.39 | 87 |
| $\alpha^*(0) = 0.98$ | 60.58 ±1.07 | 61.18 ±0.78 | 14 | 41.49 ±4.11 | 33.57 ±1.56 | 40 | 53.05 ±2.57 | 35.09 ±0.25 | 42 |
| StoCFL | | | | | | | | | |
| $\tau = 0.05$ | 59.79 ±1.34 | 57.35 ±0.92 | 15 | 29.97 ±0.47 | 31.40 ±2.16 | 4 | 31.85 ±0.08 | 31.39 ±0.87 | 1 |
| $\tau = 0.10$ | 70.84 ±1.58 | 51.72 ±0.07 | 54 | **69.76** ±2.57 | 9.42 ±0.07 | 89 | **67.48** ±1.53 | 13.03 ±0.67 | 91 |
| HCFL$^+$ (FeSEM) | | | | | | | | | |
| $\rho = 0.05$ | **87.77** ±1.11 | 41.85 ±4.11 | 58 | **69.25** ±0.69 | 14.24 ±1.93 | 67 | 60.44 ±0.86 | 23.14 ±1.46 | 32 |
| $\rho = 0.1$ | 85.08 ±0.11 | 43.34 ±0.94 | 44 | 62.32 ±0.23 | 16.67 ±2.97 | 38 | 52.18 ±2.90 | 32.97 ±1.27 | 14 |
| $\rho = 0.3$ | 79.31 ±3.95 | 47.62 ±2.90 | 17 | 44.49 ±1.57 | 28.03 ±0.85 | 8 | 45.76 ±0.09 | 36.08 ±1.25 | 4 |
| HCFL$^+$ (FedEM) | | | | | | | | | |
| $\rho = 0.05$ | 82.45 ±0.13 | 57.73 ±1.70 | 22 | 60.36 ±1.47 | 22.95 ±1.44 | 40 | 63.41 ±0.05 | 34.24 ±0.33 | 33 |
| $\rho = 0.1$ | 84.64 ±1.47 | 60.90 ±0.61 | 16 | 62.98 ±0.42 | 26.17 ±1.22 | 34 | 59.88 ±0.11 | 37.17 ±0.37 | 20 |
| $\rho = 0.3$ | 83.67 ±0.72 | 62.43 ±0.71 | 10 | 50.72 ±2.97 | 32.13 ±0.18 | 9 | 45.53 ±0.53 | 38.64 ±0.23 | 3 |
| HCFL$^+$ (FedRC) | | | | | | | | | |
| $\rho = 0.05$ | 69.16 ±0.65 | 67.37 ±0.42 | 8 | 39.20 ±0.31 | 34.38 ±0.64 | 11 | 43.78 ±0.31 | 38.75 ±0.54 | 10 |
| $\rho = 0.1$ | 71.67 ±0.83 | 68.64 ±0.76 | 8 | 39.56 ±0.14 | 34.62 ±0.78 | 8 | 44.26 ±0.10 | 38.82 ±0.77 | 6 |
| $\rho = 0.3$ | 69.33 ±0.24 | **69.67** ±1.27 | 3 | 39.97 ±0.21 | **36.50** ±0.28 | 4 | 42.60 ±0.21 | **40.65** ±0.36 | 3 |

**Evaluation metrics.** We present the following metrics to evaluate the personalization and generalization abilities of the algorithms: (1) Validation Accuracy for evaluating personalization: The average accuracy on local validation datasets that match the distribution of local training sets. (2) Test Accuracy evaluating generalization: The average accuracy on global shared test datasets.

### 5.2 Results on diverse distribution shifts scenarios

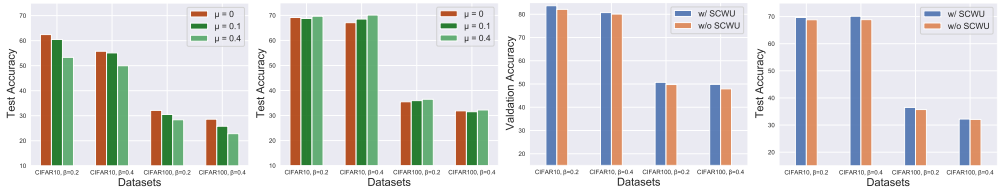
In this section, we compare the performance of HCFL$^+$ with other clustered FL methods. We also perform ablation studies to confirm the effectiveness of HCFL$^+$'s proposed components.

**HCFL$^+$ achieves comparable performance and better personalization-generalization trade-offs.** We highlight some key observations in Table 1. **A.** HCFL$^+$ consistently achieves superior test accuracy, with validation accuracy surpassing that of baseline methods with a similar number of clusters. This demonstrates improved efficiency and a better balance between personalization and generalization. **B.** Soft clustering methods like HCFL$^+$ (FedEM) and HCFL$^+$ (FedRC) outperform hard clustering methods in test accuracy, showcasing their superior generalization capabilities. **C.** While baseline methods may achieve higher validation accuracy by separating every client into different clusters (namely when the value of $K^T$ is close to 100), these trained clusters tend to overfit local distributions, resulting in significantly lower test accuracy. As detailed in Table 9, ICFL exhibits lower test accuracy when achieving the best personalization, while its personalization cannot outperform HCFL$^+$ when achieving its best test performance. **D.** The extended algorithms, namely HCFL$^+$ (FeSEM), HCFL$^+$ (FedEM), and HCFL$^+$ (FedRC), outperform the original methods that rely on fixed cluster numbers significantly. Additionally, these extended algorithms can automatically adjust the number of clusters, making the algorithms more practical, as we illustrated in Figure 4. **E.** As shown in Table 3, it is possible to enhance the personalized performance of HCFL$^+$ by employing a smaller $\rho$. However, this comes at the expense of reduced generalization performance.

**Ablation studies on Sec 4.2.** We perform ablation studies on $\tilde{\mu}$, which control the distance between sample-wise weights $\omega_{i,j;k}$ and client-wise weights $\tilde{\omega}_{i;k}$ in Figures 3a and 3b. A larger $\tilde{\mu}$ signifies a greater difference between $\omega_{i,j;k}$ and $\tilde{\omega}_{i;k}$. Our results show that HCFL$^+$ (FedEM) prefers smaller

Table 2: **Ablation studies on Sec 4.4.** We conducted experiments on the CIFAR10 and CIFAR100 datasets, showcasing the highest test accuracies, the maximum number of clusters during training ($\max_t K^t$), and the final number of clusters ($K^T$) or each algorithm while maintaining a fixed value of $\rho = 0.3$. We used FedRC as the backbone, with 3 clusters identified as ideal by CSCP. ① Measuring client distance by gradient similarity; ② Excluding $\mathbb{E}_{D_i}[\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)]$ and $\mathbb{E}_{D_i}[\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)]$ from Eq (9); ③ Averaging $d_c$ over label $y$ instead of maximizing it over $y$.

| Algorithm | CIFAR10, $\beta = 0.2$ | | | CIFAR10, $\beta = 0.4$ | | | CIFAR100, $\beta = 0.2$ | | | CIFAR100, $\beta = 0.4$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Test Acc | $\max_t K^t$ | $K^T$ | Test Acc | $\max_t K^t$ | $K^T$ | Test Acc | $\max_t K^t$ | $K^T$ | Test Acc | $\max_t K^t$ | $K^T$ |
| HCFL$^+$ | **69.67** ±1.27 | **4.5** | **3.0** | **70.13** ±0.42 | **7.0** | **6.0** | **36.50** ±0.28 | **3.5** | **3.5** | **32.22** ±0.20 | 5.0 | **4.0** |
| + ① | 67.83 ±1.70 | 9.5 | 7.0 | 64.53 ±0.23 | 10.5 | 10.0 | **36.77** ±0.67 | 9.5 | 8.5 | 31.33 ±2.12 | 11.0 | 7.5 |
| + ② | 56.14 ±8.11 | 10.5 | 5.5 | 50.87 ±2.26 | 12.5 | 8.5 | 34.11 ±1.58 | 10.0 | 6.0 | **32.75** ±0.67 | 7.5 | 6.5 |
| + ② | 68.52 ±0.64 | 8.0 | 6.0 | 69.47 ±0.15 | 11.0 | 8.5 | 34.65 ±1.16 | 8.5 | 5.5 | 31.61 ±0.54 | 11.0 | 7.5 |
| + ③ | 68.82 ±0.59 | 5.5 | 3.5 | 65.74 ±0.09 | **7.0** | 7.0 | 35.97 ±0.80 | 4.0 | **3.5** | 31.72 ±0.59 | **4.5** | **4.0** |




(a) HCFL$^+$ (FedEM) (b) HCFL$^+$ (FedRC) (c) Validation Acc (d) Test Acc

Figure 3: **Ablation studies on Sections 4.2 and 4.3.** For Sec 4.2, we evaluated test accuracies of HCFL$^+$ using different backbones (FedEM and FedRC) and varying values of $\tilde{\mu}$, as shown in Figures 3a and 3b. For Sec 4.3, we present the best Val and Test accuracy achieved by HCFL$^+$ with either FedEM or FedRC as backbones. "w/ SCWU" indicates the use of soft clustering weight updating mechanisms introduced in Section 4.3. More detailed results can be found in Tables 6 and 7 in Appendix E.

Table 3: **Performance of HCFL$^+$ with small $\rho$.** We set the $\rho = 0.01$ for HCFL$^+$, and use FedEM as the backbone algorithm.

| Algorithms | CIFAR100 Val | CIFAR100 Test | Tiny-ImageNet Val | Tiny-ImageNet Test |
|---|---|---|---|---|
| Baselines (best) | 69.76 | 37.62 | 67.48 | 39.73 |
| HCFL+ ($\rho = 0.01$) | 78.98 | 14.33 | 77.56 | 21.53 |
| HCFL+ (best in Table 1) | 69.25 | 36.50 | 63.41 | 40.65 |

distance between $\omega_{i,j;k}$ and $\tilde{\omega}_{i;k}$. However, HCFL$^+$ (FedRC) prefers larger $\tilde{\mu}$ values, highlighting the necessity of different clustering weights among samples within the same clients.

**Ablation studies on Sec 4.3.** In Figures 3c and 3d, we perform ablation studies on the soft clustering weight updating mechanism (w/ SCWU) introduced in Section 4.3. The term w/o SCWU refers to using the traditional clustering weight updating mechanism as described in (Sattler et al., 2020b; Zeng et al., 2023). The results demonstrate that our proposed SCWU consistently achieves better performance in terms of both validation and test accuracies.

**Ablation studies on techniques in Sec 4.4.** We perform ablation studies to demonstrate the effectiveness of the designed distance metrics in Sec 4.4. The ablation studies include: ① Using gradient similarity, as in previous works (Sattler et al., 2020b; Yan et al., 2023), instead of distance on $\mathcal{P}(\mathbf{z}|x; \boldsymbol{\phi})$ and $\mathcal{P}(\mathbf{z}|x, y; \boldsymbol{\phi})$, as we proposed in Equation 9; ② Remove $\mathbb{E}_{D_i}[\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)]$ and $\mathbb{E}_{D_i}[\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)]$ in (9); ③ Using mean distances instead of maximum distances in (9). The results show that HCFL$^+$ consistently achieves the highest test accuracy and produces a number of clusters closer to the ideal number than other ablation studies.

## 6 Conclusion and Limitations

In this paper, we introduce HCFL, a comprehensive clustered FL framework that unifies existing methods while enabling the integration of diverse algorithms to gather the advantages of various clustered FL approaches. Additionally, we identify persistent challenges unaddressed by current algorithms and propose HCFL$^+$ as a solution. The HCFL is flexible and can generate numerous clustered FL methods by altering techniques in each tier. Though we have chosen some typical components and demonstrated their effectiveness, conducting further performance verification with more choices in each tier would be beneficial.

## ACKNOWLEDGMENTS

This work is supported in part by the funding from Shenzhen Institute of Artificial Intelligence and Robotics for Society, in part by the Shenzhen Key Lab of Crowd Intelligence Empowered Low-Carbon Energy Network (Grant No. ZDSYS20220606100601002), in part by Shenzhen Stability Science Program 2023, and in part by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001). This work is also supported in part by the Research Center for Industries of the Future (RCIF) at Westlake University, and Westlake Education Foundation.

## REFERENCES

Wenxuan Bao, Haohan Wang, Jun Wu, and Jingrui He. Optimizing the collaboration structure in cross-silo federated learning. 2023.

Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9. IEEE, 2020.

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

Yiqun Diao, Qinbin Li, and Bingsheng He. Exploiting label skews in federated learning with model concatenation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11784–11792, 2024.

Shifei Ding, Chao Li, Xiao Xu, Lili Guo, Ling Ding, and Xindong Wu. Horizontal federated density peaks clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Moming Duan, Duo Liu, Xinyuan Ji, Renping Liu, Liang Liang, Xianzhang Chen, and Yujuan Tan. FedGroup: Efficient federated learning via decomposed similarity-based clustering. In *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pp. 228–237. IEEE, 2021a.

Moming Duan, Duo Liu, Xinyuan Ji, Yu Wu, Liang Liang, Xianzhang Chen, Yujuan Tan, and Ao Ren. Flexible clustered federated learning for client-level data distribution shift. *IEEE Transactions on Parallel and Distributed Systems*, 33(11):2661–2674, 2021b.

Xiuwen Fang and Mang Ye. Robust federated learning with noisy and heterogeneous clients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10072–10081, 2022.

Shaoduo Gan, Akhil Mathur, Anton Isopoussu, Fahim Kawsar, Nadia Berthouze, and Nicholas Lane. FRuDA: Framework for distributed adversarial domain adaptation. *IEEE Transactions on Parallel and Distributed Systems*, 2021.

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

Yongxin Guo, Tao Lin, and Xiaoying Tang. Towards federated learning on time-evolving heterogeneous data. *arXiv preprint arXiv:2112.13246*, 2021.

Yongxin Guo, Xiaoying Tang, and Tao Lin. FedBR: Improving federated learning on heterogeneous data via local learning bias reduction. 2023a.

Yongxin Guo, Xiaoying Tang, and Tao Lin. FedRC: Tackling diverse distribution shifts challenge in federated learning by robust clustering. *arXiv preprint arXiv:2301.12379*, 2023b.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

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

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674, 2013.

Liangze Jiang and Tao Lin. Test-time robust personalization for federated learning. In *International Conference on Learning Representations*, 2023.

Ellango Jothimurugesan, Kevin Hsieh, Jianyu Wang, Gauri Joshi, and Phillip B Gibbons. Federated learning under distributed concept drift. pp. 5834–5853, 2023.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning, 2019. URL https://arxiv.org/abs/1910.06378.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Shuqi Ke, Chao Huang, and Xin Liu. Quantifying the impact of label noise on federated learning. *arXiv preprint arXiv:2211.07816*, 2022.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

Tao Lin, Sebastian U. Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local sgd. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1eyO1BFPr.

Guodong Long, Ming Xie, Tao Shen, Tianyi Zhou, Xianzhi Wang, and Jing Jiang. Multi-center federated learning: clients clustering for better personalization. *World Wide Web*, 26(1):481–500, 2023.

Jie Ma, Guodong Long, Tianyi Zhou, Jing Jiang, and Chengqi Zhang. On the convergence of clustered federated learning. *arXiv preprint arXiv:2202.06187*, 2022.

Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, and Richard Vidal. Federated multi-task learning under a mixture of distributions. *Advances in Neural Information Processing Systems*, 34:15434–15447, 2021.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. 2016. doi: 10.48550/ARXIV.1602.05629. URL https://arxiv.org/abs/1602.05629.

Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation, 2019. URL https://arxiv.org/abs/1911.02054.

Dong Qiao, Chris Ding, and Jicong Fan. Federated spectral clustering via secure similarity reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=LkFG3lB13U5.

Matthias Reisser, Christos Louizos, Efstratios Gavves, and Max Welling. Federated mixture of experts. *arXiv preprint arXiv:2107.06724*, 2021.

Yichen Ruan and Carlee Joe-Wong. FedSoft: Soft clustered federated learning with proximal local updating. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8124–8131, 2022.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

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

Felix Sattler, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. On the byzantine robustness of clustered federated learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8861–8865. IEEE, 2020b.

Yan Shen, Jian Du, Han Zhao, Benyu Zhang, Zhanghexuan Ji, and Mingchen Gao. FedMM: Saddle point optimization for federated adversarial domain adaptation. *arXiv preprint arXiv:2110.08477*, 2021.

Morris Stallmann and Anna Wilbik. Towards federated clustering: A federated fuzzy $c$-means algorithm (ffcm). *arXiv preprint arXiv:2201.07316*, 2022.

Yuwei Sun, Ng Chong, and Ochiai Hideya. Multi-source domain adaptation based on federated knowledge alignment. *arXiv preprint arXiv:2203.11635*, 2022.

Zhenheng Tang, Yonggang Zhang, Shaohuai Shi, Xin He, Bo Han, and Xiaowen Chu. Virtual homogeneity learning: Defending against data heterogeneity in federated learning. *International Conference on Machine Learning*, pp. 21111–21132, 2022.

Isidoros Tziotis, Zebang Shen, Ramtin Pedarsani, Hamed Hassani, and Aryan Mokhtari. Straggler-resilient personalized federated learning. *arXiv preprint arXiv:2206.02078*, 2022.

Saeed Vahidian, Mahdi Morafah, Weijia Wang, Vyacheslav Kungurtsev, Chen Chen, Mubarak Shah, and Bill Lin. Efficient distribution similarity identification in clustered federated learning via principal angles between client data subspaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10043–10052, 2023.

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

Bin Wang, Gang Li, Chao Wu, WeiShan Zhang, Jiehan Zhou, and Ye Wei. A framework for self-supervised federated domain adaptation. *EURASIP Journal on Wireless Communications and Networking*, 2022(1):1–17, 2022a.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020.

Zhiyuan Wang, Hongli Xu, Jianchun Liu, Yang Xu, He Huang, and Yangming Zhao. Accelerating federated learning with cluster construction and hierarchical aggregation. *IEEE Transactions on Mobile Computing*, 2022b.

Xiao-Xiang Wei and Hua Huang. Edge devices clustering for federated visual classification: A feature norm based framework. *IEEE Transactions on Image Processing*, 32:995–1010, 2023.

Shanshan Wu, Tian Li, Zachary Charles, Yu Xiao, Ziyu Liu, Zheng Xu, and Virginia Smith. Motley: Benchmarking heterogeneity and personalization in federated learning. *arXiv preprint arXiv:2206.09262*, 2022.

Yue Wu, Shuaicheng Zhang, Wenchao Yu, Yanchi Liu, Quanquan Gu, Dawei Zhou, Haifeng Chen, and Wei Cheng. Personalized federated learning under mixture of distributions. 2023.

Yihan Yan, Xiaojun Tong, and Shen Wang. Clustered federated learning in heterogeneous environment. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Naoya Yoshida, Takayuki Nishio, Masahiro Morikura, Koji Yamamoto, and Ryo Yonetani. Hybrid-FL: Cooperative learning mechanism using non-iid data in wireless networks. *arXiv preprint arXiv:1905.07210*, 2019.

Dun Zeng, Xiangjing Hu, Shiyu Liu, Yue Yu, Qifan Wang, and Zenglin Xu. Stochastic clustered federated learning. *arXiv preprint arXiv:2303.00897*, 2023.

Fengpan Zhao, Yan Huang, Akshita Maradapu Vera Venkata Sai, and Yubao Wu. A cluster-based solution to achieve fairness in federated learning. In *2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pp. 875–882. IEEE, 2020.

## CONTENTS OF APPENDIX

**A Proof of Optimization Steps** **15**

**B Theoretical Study on Linear Representation Case** **17**
B.1 Definition and Assumptions . . . . . 17
B.2 Preliminary and Lemmas . . . . . 18
B.3 Main Results . . . . . 27

**C Related Works** **36**

**D Algorithms** **37**

**E Experiment Results** **41**
E.1 Datasets and Models . . . . . 41
E.2 Baselines and Hyper-Parameter Settings . . . . . 41
E.3 Additional Experiment Results . . . . . 42

Below are the key takeaways of the appendix:

- **Appendix A:** We provide a proof showing how (4)–(7) are derived, demonstrating how they solve the EM-like objective function (2)– (3) in practice.
- **Appendix B:** We analyze the HCFL framework in the context of linear representation learning, without assuming an equal number of samples across clusters. Our findings indicate that an imbalance in sample sizes and a higher level of drift can slow down convergence.
- **Appendix C:** A more detailed discussion of related works.
- **Appendix D:** We explain the rationale behind the design of the distance metric and its practical implementation. Specifically, for each pair of clients, we compute the distance between their local class prototypes for each class, as well as the distance between their feature means. The maximum of these two distances is then selected as the final distance metric for the client pair.
- **Appendix D:** A detailed version of the HCFL$^+$ algorithm is provided.
- **Appendix E:** We report the experimental settings used in our study in detail.
-
- **Appendix E:** Ablation studies on the impact of hyperparameters and algorithmic components show that HCFL$^+$ is robust to variations in hyperparameters, and that all proposed components provide individual performance gains.

## A PROOF OF OPTIMIZATION STEPS

**Theorem A.1.** *Given objective function* $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}})$

$$\begin{aligned}\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}}) = &\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\log\left(\sum_{k=1}^{K}\omega_{i,j;k}\mathcal{L}_k(\mathbf{x}_{i,j}, y_{ij}; \boldsymbol{\phi}, \boldsymbol{\theta}_k)\right)\\ &+\sum_{i=1}^{M}\sum_{j=1}^{N_i}\lambda_{i,j}\left(\sum_{k=1}^{K}\omega_{i,j;k}-1\right)\\ &-\mu\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(\sum_{k=1}^{K}\tilde{\omega}_{i;k}\log\frac{\tilde{\omega}_{i;k}}{\omega_{i,j;k}}\right), \end{aligned} \tag{13}$$

*and we define* $\tilde{\boldsymbol{\Omega}} = \{\tilde{\omega}_{i;k} | \forall i, k\}$*, then* $\tilde{\boldsymbol{\Omega}}$ *is obtained by*

$$\tilde{\boldsymbol{\Omega}} = \arg\min_{\tilde{\boldsymbol{\Omega}}}\left|\max_{\boldsymbol{\Omega}}\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}) - \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \tilde{\boldsymbol{\Omega}}, \tilde{\boldsymbol{\Omega}})\right|. \tag{14}$$

*Then E-M steps are obtained by maximizing $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}})$.*

$$\gamma_{i,j;k}^{t+1} = \frac{\omega_{i,j;k}^{t} \mathcal{L}_k(\mathbf{x}_{i,j}, y_{ij}; \boldsymbol{\theta}_k^t)}{\sum_{n=1}^{K} \omega_{i,j;n}^{t} \mathcal{L}_k(\mathbf{x}_{i,j}, y_{ij}; \boldsymbol{\theta}_n^t)}\,, \tag{15}$$

$$\tilde{\gamma}_{i,j;k}^{t+1} = \frac{\tilde{\omega}_{i;k}^{t} \mathcal{L}_k(\mathbf{x}_{i,j}, y_{ij}; \boldsymbol{\theta}_k^t)}{\sum_{n=1}^{K} \omega_{i;n}^{t} \mathcal{L}_k(\mathbf{x}_{i,j}, y_{ij}; \boldsymbol{\theta}_n^t)}\,, \tag{16}$$

$$\tilde{\omega}_{i;k}^{t+1} = \frac{1}{N_i} \sum_{j=1}^{N_i} \tilde{\gamma}_{i,j;k}^{t+1}\,, \tag{17}$$

$$\omega_{i,j;k}^{t+1} = \frac{\gamma_{i,j;k}^{t+1}}{1+\mu N} + \frac{\mu N}{1+\mu N} \tilde{\omega}_{i;k}^{t+1}\,, \tag{18}$$

$$\boldsymbol{\theta}_k^{t+1} = \boldsymbol{\theta}_k^t - \eta \sum_{i=1}^{M} \sum_{j=1}^{N_i} \frac{\gamma_{i,j;k}^{t+1}}{\mathcal{L}_k(\mathbf{x}_{ij}, y_{ij}, \boldsymbol{\phi}^t, \boldsymbol{\theta}_k^t)} \nabla_{\boldsymbol{\theta}_k} \mathcal{L}_k(\mathbf{x}_{ij}, y_{ij}, \boldsymbol{\phi}^t, \boldsymbol{\theta}_k^t)\,, \tag{19}$$

$$\boldsymbol{\phi}^{t+1} = \boldsymbol{\phi}^t - \eta \sum_{i=1}^{M} \sum_{j=1}^{N_i} \sum_{k=1}^{K} \frac{\gamma_{i,j;k}^{t+1}}{\mathcal{L}_k(\mathbf{x}_{ij}, y_{ij}, \boldsymbol{\phi}^t, \boldsymbol{\theta}_k^t)} \nabla_{\boldsymbol{\phi}} \mathcal{L}_k(\mathbf{x}_{ij}, y_{ij}, \boldsymbol{\phi}^t, \boldsymbol{\theta}_k^{t+1}) \tag{20}$$

*Proof.* Let's begin by assuming $\tilde{\omega}_{i;k}^{t+1}$ is given for each round $t$, then we will discuss how to compute $\omega_{i,j;k}^{t+1}$. Consider the objective function $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}})$, we have

$$\frac{\partial \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}})}{\partial \omega_{i,j;k}} = \frac{1}{N} \frac{\mathcal{L}_k(\mathbf{x}_{i,j}, y_{ij}; \boldsymbol{\theta}_n)}{\sum_{n=1}^{K} \omega_{i,j;n} \mathcal{L}_n(\mathbf{x}_{i,j}, y_{ij}; \boldsymbol{\theta}_n)} + \lambda_{i,j} + \mu \frac{\tilde{\omega}_{i;k}}{\omega_{i,j;k}}\,. \tag{21}$$

Then define

$$\gamma_{i,j;k} = \frac{\omega_{i,j;k} \mathcal{L}_k(\mathbf{x}_{i,j}, y_{ij}; \boldsymbol{\theta}_n)}{\sum_{n=1}^{K} \omega_{i,j;n} \mathcal{L}_n(\mathbf{x}_{i,j}, y_{ij}; \boldsymbol{\theta}_n)}\,, \tag{22}$$

and take $\frac{\partial \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Omega})}{\partial \omega_{i,j;k}} = 0$ we have

$$\frac{\gamma_{i,j;k}}{N} + \mu \tilde{\omega}_{i;k} = -\lambda_{i,j} \omega_{i,j;k}\,. \tag{23}$$

Then we have

$$\omega_{i,j;k} = -\frac{1}{\lambda_{i,j}} \left( \frac{\gamma_{i,j;k}}{N} + \mu \tilde{\omega}_{i;k} \right)\,. \tag{24}$$

Because we have $\sum_{k=1}^{K} \omega_{i,j;k} = 1$, we have

$$1 = -\frac{1}{\lambda_{i,j}} \left( \frac{1}{N} + \mu \right) \tag{25}$$

$$\lambda_{i,j} = -\frac{1+\mu N}{N}\,. \tag{26}$$

Then we have

$$\omega_{i,j;k} = \frac{\gamma_{i,j;k}}{1+\mu N} + \frac{\mu N}{1+\mu N} \tilde{\omega}_{i;k}\,. \tag{27}$$

Then consider to optimize $\boldsymbol{\theta}_k$, we have,

$$\begin{aligned} &\frac{\partial \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}})}{\partial \boldsymbol{\theta}_k} \\ &= \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \frac{\omega_{i,j;k}}{\sum_{n=1}^{K} \omega_{i,j;n} \mathcal{L}_n(\mathbf{x}_{ij}, y_{ij}; \boldsymbol{\phi}, \boldsymbol{\theta}_n)} \cdot \frac{\partial \mathcal{L}_k(\mathbf{x}_{ij}, y_{ij}; \boldsymbol{\phi}, \boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k}\,, \qquad (28) \\ &= -\frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \frac{\gamma_{i,j;k}^{t+1}}{\mathcal{L}_k(\mathbf{x}_{ij}, y_{ij}; \boldsymbol{\phi}, \boldsymbol{\theta}_k)} \nabla_{\boldsymbol{\theta}_k} \mathcal{L}_k(\mathbf{x}_{ij}, y_{ij}; \boldsymbol{\phi}, \boldsymbol{\theta}_k)\,. \qquad (29) \end{aligned}$$

Finally, consider to optimize $\boldsymbol{\phi}$, we have

$$\frac{\partial \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}})}{\partial \boldsymbol{\phi}}$$

$$= \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \sum_{k=1}^{K} \frac{\omega_{i,j;k}}{\sum_{n=1}^{K} \omega_{i,j;n} \mathcal{L}_n(\mathbf{x}_{ij}, y_{ij}; \boldsymbol{\phi}, \boldsymbol{\theta}_n)} \cdot \frac{\partial \mathcal{L}_k(\mathbf{x}_{ij}, y_{ij}; \boldsymbol{\phi}, \boldsymbol{\theta}_k)}{\partial \boldsymbol{\phi}}, \tag{30}$$

$$= -\frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \sum_{k=1}^{K} \frac{\gamma_{i,j;k}^{t+1}}{\mathcal{L}_k(\mathbf{x}_{ij}, y_{ij}; \boldsymbol{\phi}, \boldsymbol{\theta}_k)} \nabla_{\boldsymbol{\phi}} \mathcal{L}_k(\mathbf{x}_{ij}, y_{ij}; \boldsymbol{\phi}, \boldsymbol{\theta}_k). \tag{31}$$

Because if hard to find a close-form solution to $\frac{\partial \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}})}{\partial \boldsymbol{\theta}_k} = 0$ when $\boldsymbol{\theta}_k$ is the parameter of deep neural networks, we use gradient ascent to optimize $\boldsymbol{\theta}_k$. The same method is used for feature extractors $\boldsymbol{\phi}$.

Then the remaining thing is how to decide $\tilde{\omega}_{i;k}$. From the formulation of objective function, the $\tilde{\omega}_{i;k}$ is decided by

$$\tilde{\boldsymbol{\Omega}} = \arg\min_{\tilde{\boldsymbol{\Omega}}} \left| \max_{\boldsymbol{\Omega}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}) - \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \tilde{\boldsymbol{\Omega}}, \tilde{\boldsymbol{\Omega}}) \right|. \tag{32}$$

To solve this problem, firstly, we can transform the definition of $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \tilde{\boldsymbol{\Omega}}, \tilde{\boldsymbol{\Omega}})$ to

$$\begin{aligned} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \tilde{\boldsymbol{\Omega}}, \tilde{\boldsymbol{\Omega}}) = & \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \log \left( \sum_{k=1}^{K} \omega_{i,j;k} \mathcal{L}_k(\mathbf{x}_{i,j}, y_{ij}; \boldsymbol{\phi}, \boldsymbol{\theta}_k) \right) \\ & + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \lambda_{i,j} \left( \sum_{k=1}^{K} \omega_{i,j;k} - 1 \right) \\ & - \mu \sum_{i=1}^{M} \sum_{j=1}^{N_i} \left( \sum_{k=1}^{K} \tilde{\omega}_{i;k} \log \frac{\tilde{\omega}_{i;k}}{\omega_{i,j;k}} \right), \end{aligned} \tag{33}$$

$$s.t. \;\; \omega_{i,j;k} = \tilde{\omega}_{i;k}, \; \forall i, j, k. \tag{34}$$

Then by removing the constrains, we can always find

$$\max_{\boldsymbol{\Omega}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}}) \geq \tilde{\mathcal{L}}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \tilde{\boldsymbol{\Omega}}, \tilde{\boldsymbol{\Omega}}). \tag{35}$$

Then we have

$$\tilde{\boldsymbol{\Omega}} = \arg\min_{\tilde{\boldsymbol{\Omega}}} \left| \max_{\boldsymbol{\Omega}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}}) - \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \tilde{\boldsymbol{\Omega}}, \tilde{\boldsymbol{\Omega}}) \right| \tag{36}$$

$$= \arg\min_{\tilde{\boldsymbol{\Omega}}} \left( \max_{\boldsymbol{\Omega}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}}) - \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \tilde{\boldsymbol{\Omega}}, \tilde{\boldsymbol{\Omega}}) \right) \tag{37}$$

$$= \arg\max_{\tilde{\boldsymbol{\Omega}}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Theta}, \tilde{\boldsymbol{\Omega}}, \tilde{\boldsymbol{\Omega}}) \tag{38}$$

Then we can obtain the results by directly use the proof in (Guo et al., 2023b) and (Marfoq et al., 2021). □

# B Theoretical Study on Linear Representation Case

## B.1 Definition and Assumptions

In this section, we target on the effectiveness of split-feature-classifier method. Therefore, we focus on a case study that clients are solving a linear representation learning problem, similar to the analysis in (Collins et al., 2021; Tziotis et al., 2022), and assume the optimal $\omega_{i,j;k}$ is already given. In detail, we assume local data $\mathbf{x}_{i,j} \in \mathbb{R}^d$, and $\boldsymbol{\phi}$ is a global shared projection onto a $c$-dimensional subspace $\mathbb{R}^d$, which is parameterized by matrix $\mathbf{B} \in \mathbb{R}^{d \times c}$. Besides, for each underlying distribution,

we have $\boldsymbol{\theta}_k \in \mathbb{R}^c$, and the labels of each sample $\mathbf{x}_{i,j}$ belongs to the distribution $k$ is given by $y_{i,j} = \boldsymbol{\theta}_k^{*T}\mathbf{B}^{*T}x_{i,j} + z_k$, where $\boldsymbol{\theta}_k^*$ and $\mathbf{B}^*$ are ground truth parameters, and $z_k \sim \mathcal{N}(0, \sigma^2)$ is to capture the heterogeneous between $K$ underlying distributions. Under these assumptions, the global empirical risk is

$$\min_{\mathbf{B},\boldsymbol{\Theta}} \frac{1}{2N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \left( y_{i,j} - \sum_{k=1}^{K} \omega_{i,j;k} \boldsymbol{\theta}_k^T \mathbf{B}^T \mathbf{x}_{i,j} \right)^2 \tag{39}$$

The distance that measuring the distance between sub-spaces is defined as follows (Jain et al., 2013; Collins et al., 2021; Tziotis et al., 2022),

**Definition B.1.** *The principal angle distance between the column spaces of* $\mathbf{B}_1, \mathbf{B}_2 \in \mathbb{R}^{d\times c}$ *is given by,*

$$dist(\mathbf{B}_1, \mathbf{B}_2) = \left\| \hat{\mathbf{B}}_{1,\perp}^T \hat{\mathbf{B}}_2 \right\|_2 , \tag{40}$$

*where* $\hat{\mathbf{B}}_{1,\perp}$ *and* $\hat{\mathbf{B}}_2$ *are orthogonal matrices satisfying* $span(\hat{\mathbf{B}}_{1,\perp}) = span(\mathbf{B}_1)^{\perp}$*, and* $span(\hat{\mathbf{B}}_2) = span(\mathbf{B}_2)$.

**Definition B.2** ($\|\mathbf{A}\|_2$-sub-Gaussian)**.** *For a random vector* $\mathbf{x} \in \mathbb{R}^d$ *and a fixed matrix* $\mathbf{A} \in \mathbb{R}^{d\times c}$*, the vector* $\mathbf{A}^T\mathbf{x}$ *is called* $\|\mathbf{A}\|_2$*-sub-Gaussian if* $\mathbf{y}^T\mathbf{A}^T\mathbf{x}$ *is sub-Gaussian with sub-Gaussian norm* $\|\mathbf{A}\|_2 \|\mathbf{y}\|_2$ *for all* $\mathbf{y} \in \mathbb{R}^c$*, i.e.,* $\mathbb{E}\left[\exp(\mathbf{y}^T\mathbf{A}^T\mathbf{x})\right] \le \exp(\|\mathbf{A}\|_2^2 \|\mathbf{y}\|_2^2 /2)$.

**Assumption 1** (Sub-Gaussian design)**.** *The samples* $\mathbf{x}_{i,j} \in \mathbb{R}^d$ *are i.i.d. with mean* $\mathbf{0}$*, covariance* $\mathbf{I}_d$*, and are* $\mathbf{I}_d$*-sub-Gaussian, i.e.,* $\mathbb{E}\left[e^{\mathbf{v}^T\mathbf{x}_{i,j}}\right] \le e^{\|\mathbf{v}\|_2^2/2}$ *for all* $\mathbf{v} \in \mathbb{R}^d$.

**Assumption 2** (Underlying distribution diversity)**.** *Let* $\bar{\sigma}_{min,*}$ *be the minimum singular value of any matrix* $\bar{W} \in \mathbb{R}^{K\times c}$ *with rows being a* $K$*-sized subset of ground-truth distribution-specific parameters* $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K\}$*. Then* $\bar{\sigma}_{min,*} > 0$.

**Assumption 3** (Client normalization)**.** *The ground-truth distribution-specific parameters satisfy* $\|\boldsymbol{\theta}_k^*\|_2 = \sqrt{c}$ *for all* $k \in [K]$*, and* $\mathbf{B}^*$ *has orthogonal columns.*

**Assumption 4** (Binary weights)**.** *We assume the value of* $\omega_{i,j;k}$ *is given in* $\{0, 1\}$.

Assumption 1, 2, 3 are widely used assumptions in the analysis of linear representation learning problem (Collins et al., 2021; Tziotis et al., 2022). We introduce Assumption 4 to simplify the analysis, and this assumption match the observation in our numerical experiments that the max weights $\max_k[\omega_{i,j;k}]$ usually close to 1 when converge.

### B.2 Preliminary and Lemmas

We first introduce some important lemmas that we would like to use in the following parts of proof.

**Lemma B.3.** *Given orthogonal matrix* $\hat{\mathbf{B}} \in \mathbb{R}^{d\times c}$*, matrix* $\mathbf{X} \in \mathbb{R}^{N\times d}$ *that each row of* $\mathbf{X}$ *is sub-Gaussian (Assumption 1), and matrix* $\bar{\boldsymbol{\Omega}}_k \in \mathbb{R}^{N\times d}$ *valued by* $\{0, 1\}$*. Then let* $\delta_k = \frac{12\mathcal{C}_{1;k}c^{3/2}\sqrt{\log(M)}}{\sqrt{\hat{N}_k}}$*, we have*

$$\sigma_{\min}\left( \frac{1}{N}\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}} \right) \ge \frac{\hat{N}_k}{N}\left(1 - \delta_k\right) , \tag{41}$$

*with probability at least* $1 - \exp(121c^3\log(M))$*, and* $\hat{N}_k = \sum_{i=1}^{M}\sum_{j=1}^{N_i} \mathbf{1}_{\omega_{i,j;k}=1}$.

*Proof.* Firstly, we can rewrite

$$\frac{1}{N}\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}} = \frac{\hat{N}_k}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(\frac{\omega_{i,j;k}}{\sqrt{\hat{N}_k}}\hat{\mathbf{B}}^T\mathbf{x}_{i,j}\right)\left(\frac{\omega_{i,j;k}}{\sqrt{\hat{N}_k}}\hat{\mathbf{B}}^T\mathbf{x}_{i,j}\right)^T , \tag{42}$$

$$= \frac{\hat{N}_k}{N}\sum_{\omega_{i,j;k}=1}^{N}\left(\frac{1}{\sqrt{\hat{N}_k}}\hat{\mathbf{B}}^T\mathbf{x}_{i,j}\right)\left(\frac{1}{\sqrt{\hat{N}_k}}\hat{\mathbf{B}}^T\mathbf{x}_{i,j}\right)^T , \tag{43}$$

where $\hat{N}_k = \sum_{i=1}^{M}\sum_{j=1}^{N_i} \mathbf{1}_{\omega_{i,j;k}=1}$. Define $\mathbf{v}_{i,j;k} = \frac{\omega_{i,j;k}}{\sqrt{\hat{N}_k}}\hat{\mathbf{B}}^T\mathbf{x}_{i,j}$, by Definition B.2 and Assumption 1, we can observe that each $\mathbf{v}_{i,j;k}$ is either $\left\|\frac{1}{\sqrt{\hat{N}_k}}\hat{\mathbf{B}}^T\right\|_2$-sub-Gaussian or $\mathbf{v}_{i,j;k} = \mathbf{0}$. Then we can define $\mathbf{V}_k \in \mathbb{R}^{\hat{N}_k \times c}$, and each row of $\mathbf{V}_k$ is $\mathbf{v}_{i,j;k}$ that $\mathbf{v}_{i,j;k} \neq 0$. Then we have

$$\frac{1}{N}\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}} = \frac{\hat{N}_k}{N}\mathbf{V}_k^T\mathbf{V}_k\,. \tag{44}$$

Then based on Theorem 4.6.1, Equation (4.22) in (Vershynin, 2018), we have

$$\sigma_{min}\left(\frac{1}{N}\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\right) = \frac{\hat{N}_k}{N}\sigma_{min}\left(\mathbf{V}_k^T\mathbf{V}_k\right) \geq \frac{\hat{N}_k}{N}\left(1-\delta_k\right) \tag{45}$$

, with probability at least $1-\exp(-\left(\delta_k\sqrt{N_k}/\mathcal{C}_{1;k}-\sqrt{C}\right)^2)$ for constant $\mathcal{C}_{1;k}$. We then set $\delta_k = \frac{12\mathcal{C}_{1;k}c^{3/2}\sqrt{\log(M)}}{\sqrt{\hat{N}_k}}$, and we have

$$1-\exp(-\left(\delta_k\sqrt{N_k}/\mathcal{C}_{1;k}-\sqrt{C}\right)^2) \geq 1-\exp(121c^3\log(M))\,, \tag{46}$$

which finish the proof. □

**Lemma B.4.** *Given the objective function defined in Equation (39)*

$$\min_{\mathbf{B},\mathbf{\Theta}} \frac{1}{2N}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(y_{i,j}-\sum_{k=1}^{K}\omega_{i,j;k}\boldsymbol{\theta}_k^T\mathbf{B}^T\mathbf{x}_{i,j}\right)^2\,, \tag{47}$$

*define the matrix form of $\mathbf{x}_{i,j}$, $y_{i,j}$, and $\omega_{i,j;k}$ by $\mathbf{X}\in\mathbb{R}^{N\times d}$, $\mathbf{Y}\in\mathbb{R}^N$, $\mathbf{\Omega}_k\in\mathbb{R}^N$, and define $\bar{\mathbf{\Omega}}_k\in\mathbb{R}^{N\times d}$ by repeat $\mathbf{\Omega}_k$ for $d$ times, we can have the following optimization steps to solve the above objective function*

$$\boldsymbol{\theta}_k^{t+1} = \left(\frac{1}{N}[\hat{\mathbf{B}}^t]^T\left(\bar{\mathbf{\Omega}}_k^T\odot\mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}^t\right)^{-1}\left(\frac{1}{N}[\hat{\mathbf{B}}^t]^T\left(\bar{\mathbf{\Omega}}_k^T\odot\mathbf{X}^T\right)\left(\mathbf{Y}\odot\mathbf{\Omega}_k\right)\right)\,, \tag{48}$$

$$\mathbf{B}^{t+1} = \hat{\mathbf{B}}^t - \sum_{k=1}^{K}\frac{\eta}{N}\left(\left(\bar{\mathbf{\Omega}}_k^T\odot\mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\boldsymbol{\theta}_k - \left(\bar{\mathbf{\Omega}}_k^T\odot\mathbf{X}^T\right)\left(\mathbf{Y}\odot\mathbf{\Omega}_k\right)\right)\boldsymbol{\theta}_k^T\,, \tag{49}$$

$$\hat{\mathbf{B}}^{t+1} = \mathbf{B}^{t+1}\left(\mathbf{R}^{t+1}\right)^{-1}\,, \tag{50}$$

*where $\hat{\mathbf{B}}^{t+1}\mathbf{R}^{t+1}$ is the QR factorization of $\mathbf{B}^{t+1}$.*

*Proof.* We first extend the empirical function via matrices $\mathbf{X}_i\in\mathbb{R}^{N_i\times d}$, $\mathbf{Y}_i\in\mathbb{R}^{N_i}$, and $\mathbf{\Omega}_{i;k}\in\mathbb{R}^{N_i}$, $\bar{\mathbf{\Omega}}_{i;k}\in\mathbb{R}^{N_i\times d}$ is repeat $\mathbf{\Omega}_{i;k}\in\mathbb{R}^{N_i}$ for $d$ times.

$$\mathcal{L}(\mathbf{B},\mathbf{\Theta}) = \frac{1}{2N}\sum_{i=1}^{M}\left\|\sum_{k=1}^{K}\mathbf{\Omega}_{i;k}\odot\left(\mathbf{Y}_i-\mathbf{X}_i\hat{\mathbf{B}}\boldsymbol{\theta}_k\right)\right\|_2^2\,. \tag{51}$$

Define $\mathbf{X}\in\mathbb{R}^{N\times d} = [\mathbf{X}_1^T,\cdots,\mathbf{X}_M^T]^T$, $\bar{\mathbf{\Omega}}_k\in\mathbb{R}^{N\times d} = [\bar{\mathbf{\Omega}}_{1,k}^T,\cdots,\bar{\mathbf{\Omega}}_{M,k}^T]^T$, and $\mathbf{\Omega}_k\in\mathbb{R}^N = [\mathbf{\Omega}_{1,k}^T,\cdots,\mathbf{\Omega}_{M,k}^T]^T$, then compute the gradients of $\hat{B}$ and $\boldsymbol{\theta}_k$,

$$\frac{\partial\mathcal{L}(\mathbf{B},\mathbf{\Theta})}{\partial\hat{B}} = -\frac{1}{N}\sum_{i=1}^{M}\sum_{k=1}^{K}\left(\bar{\mathbf{\Omega}}_{i;k}\odot\mathbf{X}_i\right)^T\left(\mathbf{\Omega}_{i;k}\odot\left(\mathbf{Y}_i-\mathbf{X}_i\hat{\mathbf{B}}\boldsymbol{\theta}_k\right)\right)\boldsymbol{\theta}_k^T\,, \tag{52}$$

$$= \frac{1}{N}\sum_{k=1}^{K}\left(\left(\bar{\mathbf{\Omega}}_k^T\odot\mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\boldsymbol{\theta}_k - \left(\bar{\mathbf{\Omega}}_k^T\odot\mathbf{X}^T\right)\left(\mathbf{Y}\odot\mathbf{\Omega}_k\right)\right)\boldsymbol{\theta}_k^T\,, \tag{53}$$

$$\frac{\partial\mathcal{L}(\mathbf{B},\mathbf{\Theta})}{\partial\boldsymbol{\theta}_k} = -\frac{1}{N}\sum_{i=1}^{M}\left(\bar{\mathbf{\Omega}}_{i;k}\odot\mathbf{X}_i\hat{\mathbf{B}}\right)^T\left(\mathbf{\Omega}_{i;k}\odot\left(\mathbf{Y}_i-\mathbf{X}_i\hat{\mathbf{B}}\boldsymbol{\theta}_k\right)\right)\,, \tag{54}$$

$$= \frac{1}{N}\left(\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T\odot\mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\boldsymbol{\theta}_k - \hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T\odot\mathbf{X}^T\right)\left(\mathbf{Y}\odot\mathbf{\Omega}_k\right)\right)\,. \tag{55}$$

Then we can define the following optimization steps based on the above gradients, similar to the analysis in (Collins et al., 2021).

$$\boldsymbol{\theta}_k^{t+1} = \left(\frac{1}{N}[\hat{\mathbf{B}}^t]^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}}^t\right)^{-1}\left(\frac{1}{N}[\hat{\mathbf{B}}^t]^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Y} \odot \boldsymbol{\Omega}_k\right)\right), \quad (56)$$

$$\mathbf{B}^{t+1} = \hat{\mathbf{B}}^t - \sum_{k=1}^{K}\frac{\eta}{N}\left(\left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}}\boldsymbol{\theta}_k - \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Y} \odot \boldsymbol{\Omega}_k\right)\right)\boldsymbol{\theta}_k^T, \quad (57)$$

$$\hat{\mathbf{B}}^{t+1} = \mathbf{B}^{t+1}\left(\mathbf{R}^{t+1}\right)^{-1}, \quad (58)$$

where $\hat{\mathbf{B}}^{t+1}\mathbf{R}^{t+1}$ is the QR factorization of $\mathbf{B}^{t+1}$. Then the $\hat{\mathbf{B}}$ will be orthogonal matrix at the end of each optimization step. From Lemma B.3 we know that $\left(\frac{1}{N}[\hat{\mathbf{B}}^t]^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}}^t\right)$ is invertible with high probability, which indicates the feasibility of the above optimization steps. □

**Lemma B.5.** *The optimization step of $\boldsymbol{\theta}_k$ can be expressed by the following equation*

$$\boldsymbol{\theta}_k = \hat{\mathbf{B}}^T\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* + \mathbf{F}_k + \mathbf{G}_k, \quad (59)$$

*where $\mathbf{F}_k, \mathbf{G}_k \in \mathbb{R}^c$ are given in Equation (65) and (66).*

*Proof.* From Lemma B.4 we have

$$\boldsymbol{\theta}_k = \left(\frac{1}{N}\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1}\left(\frac{1}{N}\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Y} \odot \boldsymbol{\Omega}_k\right)\right). \quad (60)$$

Based on the fact that $\mathbf{Y} = \sum_{k=1}^{K}\boldsymbol{\Omega}_k \odot \left(\mathbf{X}\hat{B}^*\boldsymbol{\theta}_k^* + \mathbf{Z}\right)$, $\boldsymbol{\Omega}_k \odot \boldsymbol{\Omega}_k = \boldsymbol{\Omega}_k$, $\boldsymbol{\Omega}_k \odot \boldsymbol{\Omega}_{k'} = \mathbf{0}$ we have

$$\begin{aligned}
\boldsymbol{\theta}_k &= \left(\frac{1}{N}\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1}\left(\frac{1}{N}\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Y} \odot \boldsymbol{\Omega}_k\right)\right), && (61)\\
&= \left(\frac{1}{N}\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1} \\
&\left(\frac{1}{N}\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\sum_{n=1}^{K}\boldsymbol{\Omega}_k \odot \boldsymbol{\Omega}_n \odot \left(\mathbf{X}\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* + \mathbf{Z}\right)\right)\right), && (62)\\
&= \left(\frac{1}{N}\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1}\left(\frac{1}{N}\left(\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)\right) \\
&+ \left(\frac{1}{N}\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1}\left(\frac{1}{N}\left(\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Z} \odot \boldsymbol{\Omega}_k\right)\right)\right), && (63)\\
&= \hat{\mathbf{B}}^T\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* + \left(\frac{1}{N}\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1} \\
&\left(\frac{1}{N}\left(\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)\right) \\
&- \left(\frac{1}{N}\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1} \\
&\left(\frac{1}{N}\left(\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}}\hat{\mathbf{B}}^T\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)\right) \\
&+ \left(\frac{1}{N}\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1}\left(\frac{1}{N}\left(\hat{\mathbf{B}}^T \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Z} \odot \boldsymbol{\Omega}_k\right)\right)\right). && (64)
\end{aligned}$$

Then define

$$\mathbf{F}_k = \left(\frac{1}{N}\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1} \left(\frac{1}{N}\left(\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)\right) - \left(\frac{1}{N}\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1} \left(\frac{1}{N}\left(\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\hat{\mathbf{B}}^T\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)\right), \tag{65}$$

$$\mathbf{G}_k = \left(\frac{1}{N}\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1}\left(\frac{1}{N}\left(\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Z}\odot\mathbf{\Omega}_k\right)\right)\right). \tag{66}$$

□

**Corollary B.6.** *Define $\mathbf{\Theta}, \mathbf{F}, \mathbf{G}$ by the matrices that each row is $\boldsymbol{\theta}_k, \mathbf{F}_k, \mathbf{G}_k$, respectively, we have*

$$\mathbf{\Theta}^{t+1} = \mathbf{\Theta}^*[\hat{\mathbf{B}}^*]^T\hat{\mathbf{B}}^t + \mathbf{F}^t + \mathbf{G}^t. \tag{67}$$

*Proof.* We can easily obtain this result by Lemma B.5. □

**Lemma B.7.** *Define*

$$\mathbf{H}_k = \left(\frac{1}{\sqrt{N}}\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\right)\frac{1}{\sqrt{N}}\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\left(\mathbf{I}_d - \hat{\mathbf{B}}\hat{\mathbf{B}}^T\right)\hat{\mathbf{B}}^* \tag{68}$$

*, $\hat{N}_k = \mathbf{\Omega}_k^T\mathbf{\Omega}_k$, and $\delta = \mathcal{C}\frac{c^{3/2}\sqrt{\log(M)}}{\sqrt{\min_k \hat{N}_k}}$ for constant $\mathcal{C}$, we have*

$$\|\mathbf{H}_k\|_2 \le \frac{\delta\hat{N}_k}{\sqrt{c}N} dist\left(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*\right), \tag{69}$$

$$\sum_{k=1}^K \|\mathbf{H}_k\boldsymbol{\theta}_k^*\|_2^2 \le \left(\frac{\sum_{k=1}^K \hat{N}_k^2}{KN^2}\right)\delta^2\|\mathbf{\Theta}^*\|_2^2 dist^2\left(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*\right), \tag{70}$$

*with probability at least $1 - \exp(-111c^2\log(M))$.*

*Proof.* Because we have

$$\mathbf{H}_k = \sum_{i=1}^M\sum_{j=1}^{N_i}\left(\frac{\omega_{i,j;k}}{\sqrt{N}}\hat{\mathbf{B}}^T\mathbf{x}_{i,j}\right)\left(\frac{\omega_{i,j;k}}{\sqrt{N}}[\hat{\mathbf{B}}^*]^T\left(\mathbf{I}_d - \hat{\mathbf{B}}\hat{\mathbf{B}}^T\right)\mathbf{x}_{i,j}\right)^T, \tag{71}$$

$$= \sum_{i=1}^M\sum_{\omega_{i,j;k}=1}^{N_i}\left(\frac{1}{\sqrt{N}}\hat{\mathbf{B}}^T\mathbf{x}_{i,j}\right)\left(\frac{1}{\sqrt{N}}[\hat{\mathbf{B}}^*]^T\left(\mathbf{I}_d - \hat{\mathbf{B}}\hat{\mathbf{B}}^T\right)\mathbf{x}_{i,j}\right)^T, \tag{72}$$

$$= \left(\frac{1}{\sqrt{N}}\hat{\mathbf{B}}^T\hat{\mathbf{X}}_k^T)\right)\frac{1}{\sqrt{N}}\hat{\mathbf{X}}_k\left(\mathbf{I}_d - \hat{\mathbf{B}}\hat{\mathbf{B}}^T\right)\hat{\mathbf{B}}^*, \tag{73}$$

$$= \frac{\hat{N}_k}{N}\left(\frac{1}{\sqrt{\hat{N}_k}}\hat{\mathbf{B}}^T\hat{\mathbf{X}}_k^T)\right)\frac{1}{\sqrt{\hat{N}_k}}\hat{\mathbf{X}}_k\left(\mathbf{I}_d - \hat{\mathbf{B}}\hat{\mathbf{B}}^T\right)\hat{\mathbf{B}}^* \tag{74}$$

where $\hat{\mathbf{X}}_k \in \mathbb{R}^{\hat{N}_k\times d}$ with rows the concatenation of $\mathbf{x}_{i,j}$ that $\omega_{i,j;k} = 1$. Here we define $\hat{N}_k = \sum_{i=1}^M\sum_{j=1}^{N_i}\mathbf{1}_{\omega_{i,j;k}=1}$. Then directly use Lemma 4 of (Collins et al., 2021), and define $\delta = \mathcal{C}\frac{c^{3/2}\sqrt{\log(M)}}{\sqrt{\min_k \hat{N}_k}}$ for constant $\mathcal{C}$, we have

$$\|\mathbf{H}_k\|_2 \le \frac{\delta\hat{N}_k}{\sqrt{c}N} dist\left(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*\right). \tag{75}$$

with probability at least $1 - \exp(-111c^2 \log(M))$. Then we have

$$\sum_{k=1}^{K} \|\mathbf{H}_k \boldsymbol{\theta}_k^*\|_2^2 \leq \sum_{k=1}^{K} \|\mathbf{H}_k\|_2^2 \|\boldsymbol{\theta}_k^*\|_2^2 , \tag{76}$$

$$\leq \frac{c}{K} \|\boldsymbol{\Theta}^*\|_2^2 \sum_{k=1}^{K} \|\mathbf{H}_k\|_2^2 , \tag{77}$$

$$\leq \left( \frac{\sum_{k=1}^{K} \hat{N}_k^2}{KN^2} \right) \delta^2 \|\boldsymbol{\Theta}^*\|_2^2 \, dist^2 \left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) . \tag{78}$$

Then the proof finished. □

**Lemma B.8.** *Given* $\mathbf{F}$ *defined in Corollary B.6, define* $\delta = \mathcal{C} \frac{c^{3/2}\sqrt{\log(M)}}{\sqrt{\min_k \hat{N}_k}}$, *we have*

$$\|\mathbf{F}_k\|_2 \leq \frac{\delta}{(1-\delta)\sqrt{c}} \|\boldsymbol{\theta}_k^*\|_2 \, dist \left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) , \tag{79}$$

$$\|\mathbf{F}\|_F \leq \frac{\delta}{(1-\delta)\sqrt{K}} \|\boldsymbol{\Theta}^*\|_2 \, dist \left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) , \tag{80}$$

*with probability at least* $1 - \exp(-111c^2 \log(M))$.

*Proof.* From Lemma B.3, we have

$$\left\| \left( \frac{1}{N} \hat{\mathbf{B}}^T \left( \bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T \right) \left( \mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k \right) \hat{\mathbf{B}} \right)^{-1} \right\|_2^2 \leq \frac{N^2}{\hat{N}_k^2 (1-\delta)^2} . \tag{81}$$

Then we consider to bound $\mathbf{F}_k$ first, by Lemma B.7 we have

$$\|\mathbf{F}_k\|_2^2 \leq \left\| \left( \frac{1}{N} \hat{\mathbf{B}}^T \left( \bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T \right) \left( \mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k \right) \hat{\mathbf{B}} \right)^{-1} \right\|_2^2 \|\mathbf{H}_k\|_2^2 \|\boldsymbol{\theta}_k^*\|_2^2 , \tag{82}$$

$$\leq \frac{1}{(1-\delta)^2} \frac{\delta^2}{c} dist^2 \left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) \|\boldsymbol{\theta}_k^*\|_2^2 , \tag{83}$$

for $\delta = \mathcal{C} \frac{c^{3/2}\sqrt{\log(M)}}{\sqrt{\min_k \hat{N}_k}}$ with probability at least $1 - \exp(-111c^2 \log(M))$. Then we consider to bound $\mathbf{F}$, and we have

$$\|\mathbf{F}\|_F^2 = \sum_{k=1}^{K} \|\mathbf{F}_i\|_2^2 , \tag{84}$$

$$\leq \sum_{k=1}^{K} \left\| \left( \frac{1}{N} \hat{\mathbf{B}}^T \left( \bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T \right) \left( \mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k \right) \hat{\mathbf{B}} \right)^{-1} \right\|_2^2 \|\mathbf{H}_k \boldsymbol{\theta}_k\|_2^2 , \tag{85}$$

$$\leq \frac{1}{(1-\delta)^2} \sum_{k=1}^{K} \frac{N^2}{\hat{N}_k^2} \|\mathbf{H}_k \boldsymbol{\theta}_k\|_2^2 , \tag{86}$$

$$\leq \frac{c}{(1-\delta)^2 K} \|\boldsymbol{\Theta}^*\|_2^2 \sum_{k=1}^{K} \frac{N^2}{\hat{N}_k^2} \|\mathbf{H}_k\|_2^2 , \tag{87}$$

$$\leq \frac{\delta^2}{(1-\delta)^2 K} \|\boldsymbol{\Theta}^*\|_2^2 \, dist^2 \left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) \tag{88}$$

with probability at least $1 - \exp(-111c^2 \log(M))$. The last equation comes from Lemma B.7. □

**Lemma B.9.** *Given* $\mathbf{G}_k$ *defined by*

$$\mathbf{G}_k = \left(\frac{1}{N}\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1}\left(\frac{1}{N}\left(\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Z}\odot\mathbf{\Omega}_k\right)\right)\right) \tag{89}$$

*and* $\mathbf{G}$ *defined in Corollary B.6. Define* $\delta = \mathcal{C}\frac{c^{3/2}\sqrt{\log(M)}}{\sqrt{\min_k \hat{N}_k}}$*, we have*

$$\|\mathbf{G}_k\|_2 \leq \frac{\delta}{1-\delta}\sigma^2\,, \tag{90}$$

$$\|\mathbf{G}\|_F \leq \sqrt{K}\frac{\delta}{(1-\delta)}\sigma^2\,, \tag{91}$$

*with probability at least* $1-\exp(-110c^2\log(M))$.

*Proof.* We can rewrite $\mathbf{G}_k$ by

$$\mathbf{G}_k = \left(\frac{1}{N}\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1}\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\omega_{i,j;k}z_k\hat{\mathbf{B}}^T\mathbf{x}_{i,j}\,, \tag{92}$$

$$= \left(\frac{1}{N}\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1}\frac{1}{N}\sum_{\omega_{i,j;k}=1}^{N}z_k\hat{\mathbf{B}}^T\mathbf{x}_{i,j}\,, \tag{93}$$

$$= \left(\frac{1}{N}\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1}\frac{1}{N}\left(\hat{\mathbf{B}}^T\hat{\mathbf{X}}_k^T\hat{\mathbf{Z}}_k\right)\,. \tag{94}$$

where $\hat{\mathbf{X}}_k \in \mathbb{R}^{\hat{N}_k\times d}$ with rows the concatenation of $\mathbf{x}_{i,j}$ that $\omega_{i,j;k}=1$. Here we define $\hat{N}_k = \sum_{i=1}^{M}\sum_{j=1}^{N_i}\mathbf{1}_{\omega_{i,j;k}=1}$. Then directly use Lemma A.7 of (Tziotis et al., 2022), and define $\delta = \mathcal{C}\frac{c^{3/2}\sqrt{\log(M)}}{\sqrt{\min_k \hat{N}_k}}$, we have

$$\left\|\frac{1}{\hat{N}_k}\hat{\mathbf{B}}^T\hat{\mathbf{X}}_k^T\hat{\mathbf{Z}}_k\right\|_2 \leq \sigma^2\delta\,. \tag{95}$$

with probability at least $1-\exp(-113c^2\log(M))$. Then consider $\mathbf{G}_k$, we have

$$\|\mathbf{G}_k\|_2 \leq \frac{\hat{N}_k}{N}\left\|\left(\frac{1}{N}\hat{\mathbf{B}}^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\right)^{-1}\right\|\left\|\frac{1}{\hat{N}_k}\hat{\mathbf{B}}^T\hat{\mathbf{X}}_k^T\hat{\mathbf{Z}}_k\right\|_2\,, \tag{96}$$

$$\leq \frac{\delta}{1-\delta}\sigma^2\,. \tag{97}$$

Then consider $\mathbf{G}$, we have

$$\|\mathbf{G}\|_F^2 = \sum_{k=1}^{K}\|\mathbf{G}_k\|_2^2 \leq K\left(\frac{\delta}{(1-\delta)}\right)^2\sigma^4\,. \tag{98}$$

□

**Lemma B.10.** *Define* $\delta = \mathcal{C}\frac{c^{3/2}\sqrt{\log(M)}}{\sqrt{\min_k \hat{N}_k}}$*, we have*

$$\|\boldsymbol{\theta}_k\|_2 \leq \sqrt{c} + \frac{\delta}{1-\delta}dist\left(\hat{\mathbf{B}},\hat{\mathbf{B}}^*\right) + \frac{\delta}{1-\delta}\sigma^2\,, \tag{99}$$

$$\left\|\hat{\mathbf{B}}\boldsymbol{\theta}_k - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right\|_2 \leq \left(\sqrt{c}+\frac{\delta}{1-\delta}\right)dist\left(\hat{\mathbf{B}},\hat{\mathbf{B}}^*\right) + \frac{\delta}{1-\delta}\sigma^2\,, \tag{100}$$

$$\left\|\frac{1}{N}\sum_{k=1}^{K}\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Z}\odot\mathbf{\Omega}_k\right)\boldsymbol{\theta}_k^T\right\|_2$$

$$\leq \mathcal{C}_1\sigma^2\frac{\sqrt{d+c}}{\sqrt{N}}\left(\sqrt{c}+\frac{\delta}{1-\delta}dist\left(\hat{\mathbf{B}},\hat{\mathbf{B}}^*\right)+\frac{\delta}{1-\delta}\sigma^2\right)\,, \tag{101}$$

*with probability at least* $1-\exp(-105(d+c))-\exp(-105c^2\log(M))$ *for some constant* $\mathcal{C}_1$.

*Proof.* Define

$$\mathbf{q}_k = \left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\boldsymbol{\theta}_k - \left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Y} \odot \mathbf{\Omega}_k\right) , \tag{102}$$

and $\mathbf{Q} \in \mathbb{R}^{d\times K}$ with rows the concatenation of $\mathbf{q}_k$. With the fact that $\mathbf{Y} = \sum_{k=1}^K \mathbf{\Omega}_k \odot \left(\mathbf{X}\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* + \mathbf{Z}\right)$, we have

$$\begin{aligned}
\mathbf{q}_k &= \left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\boldsymbol{\theta}_k - \left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Y} \odot \mathbf{\Omega}_k\right) , && (103)\\
&= \left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\boldsymbol{\theta}_k - \left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{\Omega}_k \odot \sum_{n=1}^K \mathbf{\Omega}_n \odot \left(\mathbf{X}\hat{\mathbf{B}}^*\boldsymbol{\theta}_n^* + \mathbf{Z}\right)\right) , && (104)\\
&= \left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\boldsymbol{\theta}_k - \left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{\Omega}_k \odot \left(\mathbf{X}\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* + \mathbf{Z}\right)\right) , && (105)\\
&= \left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}\boldsymbol{\theta}_k - \left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\mathbf{\Omega}}_k\right)\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\\
&\quad - \left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Z} \odot \mathbf{\Omega}_k\right) , && (106)\\
&= \left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\mathbf{\Omega}}_k\right)\left(\hat{\mathbf{B}}\boldsymbol{\theta}_k - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right) - \left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Z} \odot \mathbf{\Omega}_k\right) . && (107)
\end{aligned}$$

Then we would like to consider the $\left(\hat{\mathbf{B}}\boldsymbol{\theta}_k - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)$ first. From Lemma B.5, we have

$$\boldsymbol{\theta}_k = \hat{\mathbf{B}}^T\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* + \mathbf{F}_k + \mathbf{G}_k . \tag{108}$$

Therefore we have

$$\begin{aligned}
\left\|\hat{\mathbf{B}}\boldsymbol{\theta}_k - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right\|_2 &= \left\|\hat{\mathbf{B}}\hat{\mathbf{B}}^T\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* + \hat{\mathbf{B}}\mathbf{F}_k + \hat{\mathbf{B}}\mathbf{G}_k - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right\|_2 , && (109)\\
&\le \left\|\left(\hat{\mathbf{B}}\hat{\mathbf{B}}^T - \mathbf{I}_d\right)\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right\|_2 + \left\|\hat{\mathbf{B}}\mathbf{F}_k\right\|_2 + \left\|\hat{\mathbf{B}}\mathbf{G}_k\right\|_2 , && (110)\\
&\le dist\left(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*\right)\|\boldsymbol{\theta}_k^*\|_2 + \|\mathbf{F}_k\|_2 + \|\mathbf{G}_k\|_2 , && (111)\\
&\le \sqrt{c}dist\left(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*\right) + \frac{\delta}{(1-\delta)\sqrt{c}}\|\boldsymbol{\theta}_k^*\|_2\, dist\left(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*\right) + \frac{\delta}{1-\delta}\sigma^2 , && (112)\\
&= \left(\sqrt{c} + \frac{\delta}{1-\delta}\right) dist\left(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*\right) + \frac{\delta}{1-\delta}\sigma^2 , && (113)
\end{aligned}$$

with probability at least $1 - \exp(-110c^2\log(M))$, and $\delta = \mathcal{C}\frac{c^{3/2}\sqrt{\log(M)}}{\sqrt{\min_k \hat{N}_k}}$.

Then we consider to bound $\boldsymbol{\theta}_k$, and we have

$$\begin{aligned}
\|\boldsymbol{\theta}_k\|_2 &= \left\|\hat{\mathbf{B}}^T\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* + \mathbf{F}_k + \mathbf{G}_k\right\|_2 , && (114)\\
&\le \|\boldsymbol{\theta}_k^*\|_2 + \|F_k\|_2 + \|\mathbf{G}_k\|_2 , && (115)\\
&\le \sqrt{c} + \frac{\delta}{1-\delta}dist\left(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*\right) + \frac{\delta}{1-\delta}\sigma^2 , && (116)
\end{aligned}$$

with probability at least $1 - \exp(-110c^2\log(M))$.

Then we consider to bound $\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Z} \odot \mathbf{\Omega}_k\right)$, and we have

$$\begin{aligned}
\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Z} \odot \mathbf{\Omega}_k\right) &= \sum_{\omega_{i,j;k}=1}^N z_k\mathbf{x}_{i,j} , && (117)\\
&= \hat{\mathbf{X}}_k^T\hat{\mathbf{Z}}_k , && (118)
\end{aligned}$$

where rows of $\hat{\mathbf{X}}_k \in \mathbb{R}^{\hat{N}_k\times d}$ and $\hat{\mathbf{Z}}_k \in \mathbb{R}^{\hat{N}_k}$ are $\mathbf{x}_{i,j}$ and $z_k$ subject to $\omega_{i,j;k} = 1$. Then let $\mathcal{S}^{d-1}, \mathcal{S}^{c-1}$ denote the unit spheres in $d$ and $c$ dimensions, and $\mathcal{N}_d, \mathcal{N}_k$ denote the $\frac{1}{4}$-nets of cardinality

$9^d$ and $9^k$, respectively. Then by Equation 4.13 of (Vershynin, 2018), we have

$$\left\| \frac{1}{N} \sum_{k=1}^{K} \hat{\mathbf{X}}_k^T \hat{\mathbf{Z}}_k \boldsymbol{\theta}_k^T \right\|_2 \leq 2 \max_{\mathbf{p} \in \mathcal{N}_d, \mathbf{y} \in \mathcal{N}_k} \mathbf{p}^T \left( \frac{1}{N} \sum_{k=1}^{K} \hat{\mathbf{X}}_k^T \hat{\mathbf{Z}}_k \boldsymbol{\theta}_k^T \right) \mathbf{y} \,, \tag{119}$$

$$= 2 \max_{\mathbf{p} \in \mathcal{N}_d, \mathbf{y} \in \mathcal{N}_k} \sum_{k=1}^{K} \sum_{i,j}^{\hat{N}_k} \left( \frac{z_k}{N} \langle \mathbf{x}_{i,j}, \mathbf{p} \rangle \langle \boldsymbol{\theta}_k, \mathbf{y} \rangle \right) \,. \tag{120}$$

Notice that for any fixed $\mathbf{p}, \mathbf{y}$, the random variables $\frac{z_k}{N} \langle \mathbf{x}_{i,j}, \mathbf{p} \rangle \langle \boldsymbol{\theta}_k, \mathbf{y} \rangle$ are i.i.d. zero-mean sub-exponentials with the norm at most $\mathcal{C}_1 \frac{\sigma^2 \|\boldsymbol{\theta}_k\|}{N}$ for some constant $\mathcal{C}$. Then consider the event

$$\mathcal{E} = \bigcap_{k=1}^{K} \left\{ \|\boldsymbol{\theta}_k\|_2 \leq \sqrt{c} + \frac{\delta}{1-\delta} dist\left(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*\right) + \frac{\delta}{1-\delta} \sigma^2 \right\} \,, \tag{121}$$

which holds with probability at least $1 - \exp(-105c^2 \log(M))$. Then use the Bernstein's inequality we have

$$\begin{aligned} &\Pr\left( \sum_{k=1}^{K} \sum_{i,j}^{\hat{N}_k} \frac{z_k}{N} \langle \mathbf{x}_{i,j}, \mathbf{p} \rangle \langle \boldsymbol{\theta}_k, \mathbf{y} \rangle \geq s \;\mid\; \mathcal{E} \right) \\ &\leq \exp(-\mathcal{C}_1^{'} N \min\{ \frac{s^2}{\sigma^4 \left( \sqrt{c} + \frac{\delta}{1-\delta} dist\left(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*\right) + \frac{\delta}{1-\delta}\sigma^2 \right)^2}, \\ &\frac{s}{\sigma^2 \left( \sqrt{c} + \frac{\delta}{1-\delta} dist\left(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*\right) + \frac{\delta}{1-\delta}\sigma^2 \right)} \}) \,. \end{aligned} \tag{122}$$

Setting

$$s = \mathcal{C}_2 \frac{\sigma^2 \sqrt{d+c} \left( \sqrt{c} + \frac{\delta}{1-\delta} dist\left(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*\right) + \frac{\delta}{1-\delta}\sigma^2 \right)}{\sqrt{N}} \,, \tag{123}$$

we have

$$\begin{aligned} &\Pr(\sum_{k=1}^{K} \sum_{i,j}^{\hat{N}_k} \frac{z_k}{N} \langle \mathbf{x}_{i,j}, \mathbf{p} \rangle \langle \boldsymbol{\theta}_k, \mathbf{y} \rangle \geq \\ &\mathcal{C}_2 \sigma^2 \frac{\sqrt{d+c}}{\sqrt{N}} \left( \sqrt{c} + \frac{\delta}{1-\delta} dist\left(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*\right) + \frac{\delta}{1-\delta}\sigma^2 \right) \;\mid\; \mathcal{E}) \\ &\leq \exp(-\mathcal{C}_1^{'} \mathcal{C}_2 d) \leq exp(-110(d+c)) \,, \end{aligned} \tag{124}$$

for $\mathcal{C}_2$ large enough. Taking the union bound over all points $\mathbf{p}, \mathbf{y}$ on the $\mathcal{N}_d, \mathcal{N}_k$, we have

$$\begin{aligned} &\Pr(\left\| \frac{1}{N} \sum_{k=1}^{K} \hat{\mathbf{X}}_k^T \hat{\mathbf{Z}}_k \boldsymbol{\theta}_k^T \right\|_2 \geq \\ &2\mathcal{C}_2 \sigma^2 \frac{\sqrt{d+c}}{\sqrt{N}} \left( \sqrt{c} + \frac{\delta}{1-\delta} dist\left(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*\right) + \frac{\delta}{1-\delta}\sigma^2 \right) \;\mid\; \mathcal{E}) \\ &\leq 9^{d+c} \exp(-110(d+c)) \leq \exp(-105(d+c)) \,. \end{aligned} \tag{125}$$

Removing the conditional on $\mathcal{E}$, we have

$$\begin{aligned} &\Pr\left( \left\| \frac{1}{N} \sum_{k=1}^{K} \hat{\mathbf{X}}_k^T \hat{\mathbf{Z}}_k \boldsymbol{\theta}_k^T \right\|_2 \geq 2\mathcal{C}_2 \sigma^2 \frac{\sqrt{d+c}}{\sqrt{N}} \left( \sqrt{c} + \frac{\delta}{1-\delta} dist\left(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*\right) + \frac{\delta}{1-\delta}\sigma^2 \right) \right) \\ &\leq \exp(-105(d+c)) + \Pr(\mathcal{E}^C) \leq \exp(-105(d+c)) + \exp(-105c^2 \log(M)). \end{aligned} \tag{126}$$

□

**Lemma B.11.** *Define* $\delta = \mathcal{C}\frac{c^{3/2}\sqrt{\log(M)}}{\sqrt{\min_k \hat{N}_k}}$, *we have*

$$\left\| \sum_{k=1}^{K} \frac{1}{N} \left( \left( \bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T \right) \left( \mathbf{X} \odot \bar{\mathbf{\Omega}}_k \right) \left( \hat{\mathbf{B}}\boldsymbol{\theta}_k - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* \right) - \hat{N}_k \left( \hat{\mathbf{B}}\boldsymbol{\theta}_k - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* \right) \right) \boldsymbol{\theta}_k^T \right\|_2 \leq \mathcal{C}\frac{\sqrt{d+c}}{\sqrt{N}}\epsilon\,, \tag{127}$$

*where*

$$\epsilon = \left( \sqrt{c}\frac{\delta}{1-\delta} + \frac{\delta^2}{(1-\delta)^2} \right) dist^2\left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) + \left( c + (\sigma^2+1)\sqrt{c}\frac{\delta}{1-\delta} + \frac{2\delta^2}{(1-\delta)^2}\sigma^2 \right) dist\left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) + \sqrt{c}\frac{\delta}{1-\delta}\sigma^2 + \frac{\delta^2}{(1-\delta)^2}\sigma^4\,, \tag{128}$$

*with probability at least* $1 - \exp(-100(d+c)) - \exp(-105c^2\log(M))$.

*Proof.* Define

$$\mathbf{q}_k = \frac{1}{N} \left( \bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T \right) \left( \mathbf{X} \odot \bar{\mathbf{\Omega}}_k \right) \left( \hat{\mathbf{B}}\boldsymbol{\theta}_k - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* \right) - \frac{\hat{N}_k}{N} \left( \hat{\mathbf{B}}\boldsymbol{\theta}_k - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* \right)\,, \tag{129}$$

$$= \frac{1}{N}\hat{X}_k^T\hat{X}_k \left( \hat{\mathbf{B}}\boldsymbol{\theta}_k - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* \right) - \frac{\hat{N}_k}{N} \left( \hat{\mathbf{B}}\boldsymbol{\theta}_k - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* \right)\,, \tag{130}$$

where rows of $\hat{\mathbf{X}}_k \in \mathbb{R}^{\hat{N}_k \times d}$ are $\mathbf{x}_{i,j}$ subject to $\omega_{i,j;k} = 1$. Then we would like to define the event

$$\mathcal{E} = \bigcap_{k=1}^{K} \left\{ \mathcal{A}_k \bigcap \mathcal{B}_k \right\}\,, \tag{131}$$

$$\mathcal{A}_k = \left\{ \|\boldsymbol{\theta}_k\|_2 \leq \sqrt{c} + \frac{\delta}{1-\delta} dist\left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) + \frac{\delta}{1-\delta}\sigma^2 \right\}\,, \tag{132}$$

$$\mathcal{B}_k = \left\{ \left\| \hat{\mathbf{B}}\boldsymbol{\theta}_k - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* \right\|_2 \leq \left( \sqrt{c} + \frac{\delta}{1-\delta} \right) dist\left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) + \frac{\delta}{1-\delta}\sigma^2 \right\}\,, \tag{133}$$

happens with the probability at least $1 - \exp(-105(d+c)) - \exp(-105c^2\log(M))$ by Lemma B.10. Define $\mathbf{g}_k = \hat{\mathbf{B}}\boldsymbol{\theta}_k - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*$, we have

$$\sum_{k=1}^{K} \mathbf{q}_k\boldsymbol{\theta}_k^T = \frac{1}{N} \left( \sum_{k=1}^{K} \sum_{i,j}^{\hat{N}_k} \left( \langle \mathbf{x}_{i,j}, \mathbf{g}_k \rangle \mathbf{x}_{i,j}\boldsymbol{\theta}_k^T - \mathbf{g}_k\boldsymbol{\theta}_k^T \right) \right)\,. \tag{134}$$

Let $\mathcal{S}^{d-1}, \mathcal{S}^{c-1}$ denote the unit spheres in $d$ and $c$ dimensions and $\mathcal{N}_d, \mathcal{N}_k$ the $\frac{1}{4}$-nets of cardinality $9^d$ and $9^k$, respectively. By Equation 4.13 in (Vershynin, 2018), we have

$$\left\| \sum_{k=1}^{K} \mathbf{q}_k\boldsymbol{\theta}_k^T \right\|_2 \leq \frac{2}{N} \max_{\mathbf{p}\in\mathcal{N}_d, \mathbf{y}\in\mathcal{N}_k} \mathbf{p}^T \left( \sum_{k=1}^{K} \sum_{i,j}^{\hat{N}_k} \langle \mathbf{x}_{i,j}, \mathbf{g}_k \rangle \mathbf{x}_{i,j}\boldsymbol{\theta}_k^T - \sum_{k=1}^{K} \mathbf{g}_k\boldsymbol{\theta}_k^T \right) \mathbf{y}\,, \tag{135}$$

$$= \frac{2}{N} \max_{\mathbf{p}\in\mathcal{N}_d, \mathbf{y}\in\mathcal{N}_k} \sum_{k=1}^{K} \sum_{i,j}^{\hat{N}_k} \left( \langle \mathbf{x}_{i,j}, \mathbf{g}_k \rangle \langle \mathbf{p}, \mathbf{x}_{i,j} \rangle \langle \boldsymbol{\theta}_k, \mathbf{y} \rangle - \langle \mathbf{p}, \mathbf{g}_k \rangle \langle \boldsymbol{\theta}_k, \mathbf{y} \rangle \right)\,. \tag{136}$$

Then for any $\mathbf{p}, \mathbf{y}$, the inner products $\langle \mathbf{x}_{i,j}, \mathbf{g}_k \rangle, \langle \mathbf{p}, \mathbf{x}_{i,j} \rangle$ are sub-gaussians with norm at most $\mathcal{C}_1 \|\mathbf{g}_i\|_2$ and $\mathcal{C}_2 \|\mathbf{p}\|_2 = \mathcal{C}_2$, respectively for some constants $\mathcal{C}_1, \mathcal{C}_2$. Then under the condition that $\mathcal{E}$ holds we have $\frac{1}{N}\langle \mathbf{x}_{i,j}, \mathbf{g}_k \rangle \langle \mathbf{p}, \mathbf{x}_{i,j} \rangle \langle \boldsymbol{\theta}_k, \mathbf{y} \rangle$ is sub-exponential with norm at most

$$\frac{\mathcal{C}_3}{N}\epsilon = \frac{\mathcal{C}_3}{N} \left( \sqrt{c}\frac{\delta}{1-\delta} + \frac{\delta^2}{(1-\delta)^2} \right) dist^2\left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) + \frac{\mathcal{C}_3}{N} \left( \left( c + (\sigma^2+1)\sqrt{c}\frac{\delta}{1-\delta} + \frac{2\delta^2}{(1-\delta)^2}\sigma^2 \right) dist\left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) + \sqrt{c}\frac{\delta}{1-\delta}\sigma^2 + \frac{\delta^2}{(1-\delta)^2}\sigma^4 \right)\,. \tag{137}$$

The same thing can be observed for $\langle \mathbf{p}, \mathbf{g}_k \rangle \langle \boldsymbol{\theta}_k, \mathbf{y} \rangle$. Besides, we can observe that

$$\mathbb{E}\left[\langle \mathbf{x}_{i,j}, \mathbf{g}_k \rangle \langle \mathbf{p}, \mathbf{x}_{i,j} \rangle \langle \boldsymbol{\theta}_k, \mathbf{y} \rangle - \langle \mathbf{p}, \mathbf{g}_k \rangle \langle \boldsymbol{\theta}_k, \mathbf{y} \rangle\right] = 0\,. \tag{138}$$

Then we are dealing with $N$ zero-mean, sub-exponential random variables. Using Bernstein's inequality we have

$$\Pr\left(\frac{1}{N}\sum_{k=1}^{K}\sum_{i,j}^{\hat{N}_k}\langle \mathbf{x}_{i,j}, \mathbf{g}_k \rangle \langle \mathbf{p}, \mathbf{x}_{i,j} \rangle \langle \boldsymbol{\theta}_k, \mathbf{y} \rangle - \langle \mathbf{p}, \mathbf{g}_k \rangle \langle \boldsymbol{\theta}_k, \mathbf{y} \rangle \geq s \mid \mathcal{E}\right) \leq$$
$$\exp\left(-\mathcal{C}_4 N \min\left\{\frac{s^2}{\epsilon^2}, \frac{s}{\epsilon}\right\}\right)\,. \tag{139}$$

Setting $s = \frac{\sqrt{\mathcal{C}_5(d+c)}\epsilon}{\sqrt{N}}$, for constant $\mathcal{C}_5$ that satisfy $\mathcal{C}_5 \leq \frac{N}{d+c}$, and taking union bound over all $\mathbf{p}, \mathbf{y}$, we have

$$\Pr\left(\left\|\sum_{k=1}^{K}\mathbf{q}_k\boldsymbol{\theta}_k^T\right\|_2 \geq \frac{2\sqrt{\mathcal{C}_5(d+c)}\epsilon}{\sqrt{N}} \mid \mathcal{E}\right)$$
$$\leq 9^{d+c}\exp\left(-\mathcal{C}_4\mathcal{C}_5(d+c)\right) \leq \exp(-105(d+c))\,. \tag{140}$$

Then by removing the conditional on $\mathcal{E}$, we have

$$\left\|\sum_{k=1}^{K}\mathbf{q}_k\boldsymbol{\theta}_k^T\right\|_2 \leq \mathcal{C}\frac{\sqrt{d+c}\epsilon}{\sqrt{N}}\,, \tag{141}$$

with probability at least $1 - \exp(-100(d+c)) - \exp(-105c^2\log(M))$. □

### B.3 Main Results

**Theorem B.12.** *Under Assumption 1- 4, when we have* $N \geq \frac{K^2}{d+c}$, *and* $\min_k \hat{N}_k \geq \mathcal{C}\frac{c^3(1+\sigma^2)^4\log(M)}{E_0^2}\min\left\{\frac{1}{\kappa^2}, \bar{\sigma}_{\min}^2\right\}$ *for some constant* $\mathcal{C}$, *we have*

$$dist(\hat{\mathbf{B}}^{t+1}, \hat{\mathbf{B}}^*) \leq dist(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*)\left(1 - c_{min} + \frac{57}{200}c_{max}\right)\left(1 - \frac{1}{2}c_{max}\right)^{-1/2}$$
$$+ \left(\frac{7}{100}c_{max}\right)\left(1 - \frac{1}{2}c_{max}\right)^{-1/2}\,, \tag{142}$$

*with the probability at least* $1 - \exp(-90(d+c)) - \exp(-90c^2\log(M))$. *Here* $\hat{N}_k = \sum_{i=1}^{M}\sum_{j=1}^{N_i}\omega_{i,j;k}$, $E_0 = 1 - dist^2(\hat{\mathbf{B}}^0, \hat{\mathbf{B}}^*)$, $c_{min} = \eta K\frac{\min_k \hat{N}_k}{N}\bar{\sigma}_{\min,*}^2 E_0$, *and* $c_{max} = \eta K\frac{\max_k \hat{N}_k}{N}\bar{\sigma}_{\min,*}^2 E_0$.

*Proof.* From the optimization steps we have

$$\mathbf{B}^{t+1} = \hat{\mathbf{B}}^t - \sum_{k=1}^{K}\frac{\eta}{N}\left(\left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t - \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Y} \odot \boldsymbol{\Omega}_k\right)\right)(\boldsymbol{\theta}_k^t)^T\,, \tag{143}$$

$$= \hat{\mathbf{B}}^t - \sum_{k=1}^{K}\frac{\eta}{N}(\left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)$$
$$- \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Z} \odot \boldsymbol{\Omega}_k\right))(\boldsymbol{\theta}_k^t)^T\,, \tag{144}$$

$$= \hat{\mathbf{B}}^t - \eta(\sum_{k=1}^{K}(\frac{1}{N}\left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k\right)\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)$$
$$- \frac{\hat{N}_k}{N}\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right))(\boldsymbol{\theta}_k^t)^T)$$
$$- \frac{\eta}{N}\sum_{k=1}^{K}\hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)(\boldsymbol{\theta}_k^t)^T + \frac{\eta}{N}\sum_{k=1}^{K}\left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Z} \odot \boldsymbol{\Omega}_k\right)(\boldsymbol{\theta}_k^t)^T\,. \tag{145}$$

Multiplying both sides by $(\hat{\mathbf{B}}_\perp^*)^T$, we have

$$
\begin{aligned}
&(\hat{\mathbf{B}}_\perp^*)^T\mathbf{B}^{t+1} = (\hat{\mathbf{B}}_\perp^*)^T\hat{\mathbf{B}}^t \\
&- \eta(\hat{\mathbf{B}}_\perp^*)^T(\sum_{k=1}^{K}(\frac{1}{N}\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right) \\
&- \frac{\hat{N}_k}{N}\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right))(\boldsymbol{\theta}_k^t)^T) \\
&- \frac{\eta}{N}\sum_{k=1}^{K}\hat{N}_k\left((\hat{\mathbf{B}}_\perp^*)^T\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t - (\hat{\mathbf{B}}_\perp^*)^T\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)(\boldsymbol{\theta}_k^t)^T \\
&+ \frac{\eta}{N}\sum_{k=1}^{K}(\hat{\mathbf{B}}_\perp^*)^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Z}\odot\mathbf{\Omega}_k\right)(\boldsymbol{\theta}_k^t)^T, \qquad (146)\\
&= (\hat{\mathbf{B}}_\perp^*)^T\hat{\mathbf{B}}^t\left(\mathbf{I}_c - \frac{\eta}{N}\sum_{k=1}^{K}\hat{N}_k\boldsymbol{\theta}_k^t(\boldsymbol{\theta}_k^t)^T\right) + \frac{\eta}{N}\sum_{k=1}^{K}(\hat{\mathbf{B}}_\perp^*)^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Z}\odot\mathbf{\Omega}_k\right)(\boldsymbol{\theta}_k^t)^T \\
&- \frac{\eta}{N}(\hat{\mathbf{B}}_\perp^*)^T\left(\sum_{k=1}^{K}\left(\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right) - \hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)\right)(\boldsymbol{\theta}_k^t)^T\right). \qquad (147)
\end{aligned}
$$

Because we have $\hat{\mathbf{B}}^{t+1} = \mathbf{B}^{t+1}\left(\mathbf{R}^{t+1}\right)^{-1}$, multiplying both sides by $\left(\mathbf{R}^{t+1}\right)^{-1}$ we have

$$
\begin{aligned}
&dist\left(\hat{\mathbf{B}}^{t+1}, \hat{\mathbf{B}}^*\right) \\
&\leq dist\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right)\left\|\mathbf{I}_c - \eta\sum_{k=1}^{K}\boldsymbol{\theta}_k^t(\boldsymbol{\theta}_k^t)^T\right\|_2 \left\|(\mathbf{R}^{t+1})^{-1}\right\|_2 \\
&+ \left\|\frac{\eta}{N}\sum_{k=1}^{K}(\hat{\mathbf{B}}_\perp^*)^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Z}\odot\mathbf{\Omega}_k\right)(\boldsymbol{\theta}_k^t)^T\right\|_2 \left\|(\mathbf{R}^{t+1})^{-1}\right\|_2 \\
&+ \left\|\frac{\eta}{N}(\hat{\mathbf{B}}_\perp^*)^T\left(\sum_{k=1}^{K}\left(\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right) - \hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)\right)(\boldsymbol{\theta}_k^t)^T\right)\right\|_2 \\
&\left\|(\mathbf{R}^{t+1})^{-1}\right\|_2. \qquad (148)
\end{aligned}
$$

Then we can define

$$
\begin{aligned}
A_1 &= dist\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right)\left\|\mathbf{I}_c - \frac{\eta}{N}\sum_{k=1}^{K}\hat{N}_k\boldsymbol{\theta}_k^t(\boldsymbol{\theta}_k^t)^T\right\|_2, \qquad (149)\\
A_2 &= \left\|\frac{\eta}{N}\sum_{k=1}^{K}(\hat{\mathbf{B}}_\perp^*)^T\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{Z}\odot\mathbf{\Omega}_k\right)(\boldsymbol{\theta}_k^t)^T\right\|_2, \qquad (150)\\
A_3 &= \|\frac{\eta}{N}(\hat{\mathbf{B}}_\perp^*)^T(\sum_{k=1}^{K}(\left(\bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\mathbf{\Omega}}_k\right)\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right) \\
&\quad - \hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right))(\boldsymbol{\theta}_k^t)^T)\|_2. \qquad (151)
\end{aligned}
$$

Then the inequality become

$$
dist\left(\hat{\mathbf{B}}^{t+1}, \hat{\mathbf{B}}^*\right) \leq (A_1 + A_2 + A_3)\left\|(\mathbf{R}^{t+1})^{-1}\right\|_2. \qquad (152)
$$

For the following parts of the proof, we consider the following events hold

$$\mathcal{E}_1 = \bigcap_{k=1}^{K} \left\{ \mathcal{A}_k \bigcap \mathcal{B}_k \right\} , \tag{153}$$

$$\mathcal{E}_2 = \left\{ \left\| \mathbf{F}^t \right\|_F \leq \frac{\delta}{(1-\delta)\sqrt{K}} \left\| \mathbf{\Theta}^* \right\|_2 dist\left( \hat{\mathbf{B}}^t, \hat{\mathbf{B}}^* \right) \bigcap \left\| \mathbf{G}^t \right\|_F \leq \sqrt{K} \frac{\delta}{(1-\delta)} \sigma^2 \right\} , \tag{154}$$

$$\begin{aligned} \mathcal{E}_3 = \{ & \left\| \frac{1}{N} \sum_{k=1}^{K} \left( \bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T \right) \left( \mathbf{Z} \odot \mathbf{\Omega}_k \right) \left( \boldsymbol{\theta}_k^t \right)^T \right\|_2 \\ & \leq \mathcal{C}_1 \sigma^2 \frac{\sqrt{d+c}}{\sqrt{N}} \left( \sqrt{c} + \frac{\delta}{1-\delta} dist\left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) + \frac{\delta}{1-\delta} \sigma^2 \right) \} , \end{aligned} \tag{155}$$

$$\begin{aligned} \mathcal{E}_4 = \{ & \left\| \sum_{k=1}^{K} \frac{1}{N} \left( \left( \bar{\mathbf{\Omega}}_k^T \odot \mathbf{X}^T \right) \left( \mathbf{X} \odot \bar{\mathbf{\Omega}}_k \right) \left( \hat{\mathbf{B}} \boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^* \boldsymbol{\theta}_k^* \right) - \hat{N}_k \left( \hat{\mathbf{B}} \boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^* \boldsymbol{\theta}_k^* \right) \right) \left( \boldsymbol{\theta}_k^t \right)^T \right\|_2 \\ & \leq \mathcal{C}_2 \frac{\sqrt{d+c}}{\sqrt{N}} \epsilon \} , \end{aligned} \tag{156}$$

$$\tag{157}$$

where

$$\mathcal{A}_k = \left\{ \left\| \boldsymbol{\theta}_k^t \right\|_2 \leq \sqrt{c} + \frac{\delta}{1-\delta} dist\left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) + \frac{\delta}{1-\delta} \sigma^2 \right\} , \tag{158}$$

$$\mathcal{B}_k = \left\{ \left\| \hat{\mathbf{B}} \boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^* \boldsymbol{\theta}_k^* \right\|_2 \leq \left( \sqrt{c} + \frac{\delta}{1-\delta} \right) dist\left( \hat{\mathbf{B}}^t, \hat{\mathbf{B}}^* \right) + \frac{\delta}{1-\delta} \sigma^2 \right\} , \tag{159}$$

$$\begin{aligned} \epsilon = & \left( \sqrt{c} \frac{\delta}{1-\delta} + \frac{\delta^2}{(1-\delta)^2} \right) dist^2\left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) \\ & + \left( c + (\sigma^2+1)\sqrt{c} \frac{\delta}{1-\delta} + \frac{2\delta^2}{(1-\delta)^2} \sigma^2 \right) dist\left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) + \sqrt{c} \frac{\delta}{1-\delta} \sigma^2 + \frac{\delta^2}{(1-\delta)^2} \sigma^4 . \end{aligned} \tag{160}$$

which hold with probability at least $1 - \exp(-90(d+c)) - \exp(-90c^2 \log(M))$ for some constants $\mathcal{C}_1, \mathcal{C}_2$ by Lemma B.8, B.9, B.10, B.11. Then we consider to bound $A_1, A_2, A_3$, respectively. Then we consider to bound $A_1$ first, and we have

$$\lambda_{\max}((\mathbf{\Theta}^t)^T \mathbf{\Theta}^t) = \left\| \mathbf{\Theta}^t \right\|_2^2 = \left\| \mathbf{\Theta}^* (\hat{\mathbf{B}}^*)^T \hat{\mathbf{B}}^t + \mathbf{F}^t + \mathbf{G}^t \right\|_2^2 , \tag{161}$$

$$\leq 2 \left\| \mathbf{\Theta}^* \right\|_2^2 + 2 \left\| \mathbf{F}^t \right\|_2^2 + 2 \left\| \mathbf{G}^t \right\|_2^2 , \tag{162}$$

$$\leq 2 \left\| \mathbf{\Theta}^* \right\|_2^2 + \frac{2\delta^2}{(1-\delta)^2 K} \left\| \mathbf{\Theta}^* \right\|_2^2 dist^2\left( \hat{\mathbf{B}}^t, \hat{\mathbf{B}}^* \right) + 2K \frac{\delta}{(1-\delta)^2} \sigma^4 , \tag{163}$$

$$\leq \left( 2 + \frac{2\delta^2}{(1-\delta)^2 K} \right) \left\| \mathbf{\Theta}^* \right\|_2^2 + 2K \frac{\delta}{(1-\delta)^2} \sigma^4 , \tag{164}$$

$$\leq \left( 2K + \frac{2\delta^2}{(1-\delta)^2} \right) \bar{\sigma}_{\max,*}^2 + 2K \frac{\delta}{(1-\delta)^2} \sigma^4 . \tag{165}$$

Then when $\eta$ is small enough, it's simple to promise that $\mathbf{I}_c - \frac{\eta}{N}\sum_{k=1}^{K}\boldsymbol{\theta}_k\boldsymbol{\theta}_k^T$ is positive definite. Define diagonal matrix $\mathbf{W} \in \mathbb{R}^{K\times K}$, and $\mathbf{W}_{k,k} = \frac{\hat{N}_k}{N}$. Then we have

$$\left\|\mathbf{I}_c - \frac{\eta}{N}\sum_{k=1}^{K}\hat{N}_k\boldsymbol{\theta}_k^t(\boldsymbol{\theta}_k^t)^T\right\|_2 = \left\|\mathbf{I}_c - \eta(\mathbf{W}\boldsymbol{\Theta}^t)^T\boldsymbol{\Theta}^t\right\|_2 \tag{166}$$

$$\leq 1 - \eta\lambda_{\min}\left((\mathbf{W}\boldsymbol{\Theta}^t)^T\boldsymbol{\Theta}^t\right), \tag{167}$$

$$\begin{aligned}
&\leq 1 - \eta\left(\sigma_{\min}(\mathbf{W})\,\sigma_{\min}^2\left(\boldsymbol{\Theta}^*(\hat{\mathbf{B}}^*)^T\hat{\mathbf{B}}^t\right) - \sigma_{\min}((\mathbf{W}\mathbf{F}^t)^T\mathbf{F}^t) - \sigma_{\min}((\mathbf{W}\mathbf{G}^t)^T\mathbf{G}^t)\right)\\
&+ 2\eta\left(\sigma_{\max}\left((\mathbf{W}\mathbf{F}^t)^T\boldsymbol{\Theta}^*(\hat{\mathbf{B}}^*)^T\hat{\mathbf{B}}^t\right) + \sigma_{\max}\left((\mathbf{W}\mathbf{F}^t)^T\mathbf{G}\right) + \sigma_{\max}\left((\mathbf{W}\mathbf{G}^t)^T\boldsymbol{\Theta}^*(\hat{\mathbf{B}}^*)^T\hat{\mathbf{B}}^t\right)\right),
\end{aligned} \tag{168}$$

$$\begin{aligned}
&\leq 1 - \eta\left(\frac{\min_k\hat{N}_k}{N}\right)\sigma_{\min}^2(\boldsymbol{\Theta}^*)\sigma_{\min}^2\left((\hat{\mathbf{B}}^*)^T\hat{\mathbf{B}}^t\right) + \eta\left(\frac{\max_k\hat{N}_k}{N}\right)\sigma_{\min}^2(\mathbf{F}^t)\\
&+ \eta\left(\frac{\max_k\hat{N}_k}{N}\right)\sigma_{\min}^2(\mathbf{G}^t)\\
&+ \frac{2\eta}{N}\left(\frac{\max_k\hat{N}_k}{N}\right)\left(\sigma_{\max}\left((\mathbf{F}^t)^T\boldsymbol{\Theta}^*\right) + \sigma_{\max}\left((\mathbf{G}^t)^T\boldsymbol{\Theta}^*\right)\right)\\
&+ 2\eta\left(\frac{\max_k\hat{N}_k}{N}\right)\sigma_{\max}(\mathbf{F}^t)\sigma_{\max}(\mathbf{G}^t),
\end{aligned} \tag{169}$$

$$\begin{aligned}
&\leq 1 - \eta K\left(\frac{\min_k\hat{N}_k}{N}\right)\bar{\sigma}_{\min,*}^2\sigma_{\min}^2\left((\hat{\mathbf{B}}^*)^T\hat{\mathbf{B}}^t\right) + 2\eta\left(\frac{\max_k\hat{N}_k}{N}\right)\left(\left\|\mathbf{F}^t\right\|_2 + \left\|\mathbf{G}^t\right\|_2\right)\left\|\boldsymbol{\Theta}^*\right\|_2\\
&+ 2\eta\left(\frac{\max_k\hat{N}_k}{N}\right)\left\|\mathbf{F}^t\right\|_2\left\|\mathbf{G}^t\right\|_2 + \eta\left(\frac{\max_k\hat{N}_k}{N}\right)\left\|\mathbf{F}^t\right\|_2^2 + \eta\left(\frac{\max_k\hat{N}_k}{N}\right)\left\|\mathbf{G}^t\right\|_2^2,
\end{aligned} \tag{170}$$

$$\begin{aligned}
&= 1 - \eta K\left(\frac{\min_k\hat{N}_k}{N}\right)\bar{\sigma}_{\min,*}^2\sigma_{\min}^2\left((\hat{\mathbf{B}}^*)^T\hat{\mathbf{B}}^t\right) + 2\eta\left(\frac{\max_k\hat{N}_k}{N}\right)\left(\left\|\mathbf{F}^t\right\|_2 + \left\|\mathbf{G}^t\right\|_2\right)\left\|\boldsymbol{\Theta}^*\right\|_2\\
&+ \eta\left(\frac{\max_k\hat{N}_k}{N}\right)\left(\left\|\mathbf{F}^t\right\|_2 + \left\|\mathbf{G}^t\right\|_2\right)^2.
\end{aligned} \tag{171}$$

Then under condition $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4$, we have

$$
\begin{aligned}
&\left\| \mathbf{I}_c - \frac{\eta}{N} \sum_{k=1}^{K} \hat{N}_k \boldsymbol{\theta}_k^t (\boldsymbol{\theta}_k^t)^T \right\|_2 \\
&\leq 1 - \eta K \left( \frac{\min_k \hat{N}_k}{N} \right) \bar{\sigma}_{\min,*}^2 \sigma_{\min}^2 \left( (\hat{\mathbf{B}}^*)^T \hat{\mathbf{B}}^t \right) \\
&+ 2\eta \left( \frac{\max_k \hat{N}_k}{N} \right) \left( \frac{\delta}{(1-\delta)\sqrt{K}} \left\| \boldsymbol{\Theta}^* \right\|_2 dist \left( \hat{\mathbf{B}}^t, \hat{\mathbf{B}}^* \right) + \sqrt{K} \frac{\delta}{1-\delta} \sigma^2 \right) \left\| \boldsymbol{\Theta}^* \right\|_2 \\
&+ \eta \left( \frac{\max_k \hat{N}_k}{N} \right) \left( \frac{\delta}{(1-\delta)\sqrt{K}} \left\| \boldsymbol{\Theta}^* \right\|_2 dist \left( \hat{\mathbf{B}}^t, \hat{\mathbf{B}}^* \right) + \sqrt{K} \frac{\delta}{1-\delta} \sigma^2 \right)^2 , \qquad (172) \\
&\leq 1 - \eta K \left( \frac{\min_k \hat{N}_k}{N} \right) \bar{\sigma}_{\min,*}^2 \sigma_{\min}^2 \left( (\hat{\mathbf{B}}^*)^T \hat{\mathbf{B}}^t \right) \\
&+ 2\eta \left( \frac{\max_k \hat{N}_k}{N} \right) \left( \frac{\delta\sqrt{K}}{1-\delta} \bar{\sigma}_{\max,*}^2 + \frac{\delta K \sigma^2}{1-\delta} \bar{\sigma}_{\max,*} \right) \\
&+ \eta \left( \frac{\max_k \hat{N}_k}{N} \right) \frac{\delta^2}{(1-\delta)^2} \left( 2\bar{\sigma}_{\max,*}^2 + 2K\sigma^4 \right) , \qquad (173) \\
&\leq 1 - \eta K \left( \frac{\min_k \hat{N}_k}{N} \right) \bar{\sigma}_{\min,*}^2 E_0 + \eta K \left( \frac{\max_k \hat{N}_k}{N} \right) \frac{\delta}{1-\delta} \left( 2\bar{\sigma}_{\max,*} + \frac{1}{2}\sigma^2 \right)^2 , \qquad (174) \\
&\leq 1 - \eta K \left( \frac{\min_k \hat{N}_k}{N} \right) \bar{\sigma}_{\min,*}^2 E_0 + \frac{\eta K}{5} \left( \frac{\max_k \hat{N}_k}{N} \right) \bar{\sigma}_{\min,*}^2 E_0 \\
&+ \frac{\eta K}{20} \left( \frac{\max_k \hat{N}_k}{N} \right) \bar{\sigma}_{\min,*}^2 E_0 . \qquad (175)
\end{aligned}
$$

where $E_0 = 1 - dist^2 \left( \hat{\mathbf{B}}^0, \hat{\mathbf{B}}^* \right) \leq \sigma_{\min}^2 \left( (\hat{\mathbf{B}}^*)^T \hat{\mathbf{B}}^t \right)$. The Equation (174) holds when $\hat{N}_k$ is large enough that makes the following equation holds

$$
\frac{\delta}{1-\delta} \leq 2\delta \leq \frac{E_0}{20(1+\sigma^2)^2} \max \left\{ \frac{1}{\kappa^2}, \bar{\sigma}_{\min}^2 \right\} , \qquad (176)
$$

where $\kappa = \frac{\bar{\sigma}_{\max,*}}{\bar{\sigma}_{\min,*}}$, and the equation holds when $\hat{N}_k$ satisfy

$$
\min_k \hat{N}_k \geq \mathcal{C}_0 \frac{c^3 (1+\sigma^2)^4 \log(M)}{E_0^2} \min \left\{ \kappa^4, \frac{1}{\bar{\sigma}_{\min}^4} \right\} . \qquad (177)
$$

Then consider $A_1$, we will have

$$
A_1 \leq dist \left( \hat{\mathbf{B}}^t, \hat{\mathbf{B}}^* \right) \left( 1 - \left( \left( \frac{\min_k \hat{N}_k}{N} \right) - \frac{1}{4} \left( \frac{\max_k \hat{N}_k}{N} \right) \right) \eta K \bar{\sigma}_{\min,*}^2 E_0 \right) . \qquad (178)
$$

The consider $A_2$, because $\left\| \hat{\mathbf{B}}_\perp^* \right\|_2 = 1$, and $\mathcal{E}_3$ holds, we have

$$
\begin{aligned}
A_2 &\leq \eta \mathcal{C}_1 \sigma^2 \frac{\sqrt{d+c}}{\sqrt{N}} \left( \sqrt{c} + \frac{\delta}{1-\delta} dist \left( \hat{\mathbf{B}}, \hat{\mathbf{B}}^* \right) + \frac{\delta}{1-\delta} \sigma^2 \right) , \qquad (179) \\
&\leq \eta \mathcal{C}_1 \sigma^2 \frac{\sqrt{d+c}}{\sqrt{N}} \left( \sqrt{c} + \frac{1}{10} \right) . \qquad (180)
\end{aligned}
$$

Similarly, for $A_3$, we have

$$A_3 \leq \eta \mathcal{C}_2 \frac{\sqrt{d+c}}{\sqrt{N}} \epsilon \,, \tag{181}$$

$$\leq \eta \mathcal{C}_2 \frac{\sqrt{d+c}}{\sqrt{N}} \left( \left( \sqrt{c} + \frac{1}{\sqrt{10}} \right)^2 dist\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) + \left( \frac{1}{400} + \frac{\sqrt{c}}{20} \right) \left(\sigma^2 + 1\right) \right) . \tag{182}$$

Combining Equation (152), (178), (180), and (182), and choose $\mathcal{C} = \max\{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_1\mathcal{C}_2 + \mathcal{C}_2\}$, we have

$$\begin{aligned} & dist\left(\hat{\mathbf{B}}^{t+1}, \hat{\mathbf{B}}^*\right) \leq dist\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) \\ & (1 - \eta K \left( \left( \frac{\min_k \hat{N}_k}{N} \right)^2 - \frac{1}{4} \left( \frac{\max_k \hat{N}_k}{N} \right)^2 \right) \bar{\sigma}^2_{\min,*} E_0 \\ & + \eta \mathcal{C} \frac{\sqrt{d+c}}{\sqrt{N}} \left( \sqrt{c} + \frac{1}{\sqrt{10}} \right)^2 ) \left\| (\mathbf{R}^{t+1})^{-1} \right\|_2 \\ & + \eta \mathcal{C} \frac{\sqrt{d+c}}{\sqrt{N}} \left(\sigma^2 + 1\right) \left( 2\sqrt{c} + \frac{1}{5} \right) \left\| (\mathbf{R}^{t+1})^{-1} \right\|_2 . \end{aligned} \tag{183}$$

Then the remaining thing is to bound $(\mathbf{R}^{t+1})^{-1}$. Firstly we define

$$\mathbf{S}^t = \sum_{k=1}^{K} \left( \bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T \right) \left( \mathbf{X} \odot \bar{\boldsymbol{\Omega}}_k \right) \left( \hat{\mathbf{B}}^t \boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^* \boldsymbol{\theta}_k^* \right) (\boldsymbol{\theta}_k^t)^T \,, \tag{184}$$

$$\mathbf{E}^t = \sum_{k=1}^{K} \left( \bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T \right) \left( \mathbf{Z} \odot \boldsymbol{\Omega}_k \right) (\boldsymbol{\theta}_k^t)^T \,. \tag{185}$$

Then we have

$$\mathbf{B}^{t+1} = \hat{\mathbf{B}}^t - \frac{\eta}{N} \mathbf{S}^t + \frac{\eta}{N} \mathbf{E}^t \,. \tag{186}$$

Because $(\mathbf{R}^{t+1})^T \mathbf{R}^{t+1} = (\mathbf{B}^{t+1})^T \mathbf{B}^{t+1}$, and we have

$$\begin{aligned} (\mathbf{B}^{t+1})^T \mathbf{B}^{t+1} &= (\hat{\mathbf{B}}^t)^T \hat{\mathbf{B}}^t - \frac{\eta}{N} \left( (\hat{\mathbf{B}}^t)^T \mathbf{S}^t + (\mathbf{S}^t)^T \hat{\mathbf{B}}^t \right) + \frac{\eta}{N} \left( (\hat{\mathbf{B}}^t)^T \mathbf{E}^t + (\mathbf{E}^t)^T \hat{\mathbf{B}}^t \right) \\ &\quad + \frac{\eta^2}{N^2} (\mathbf{S}^t)^T \mathbf{S}^t - \frac{\eta^2}{N^2} \left( (\mathbf{E}^t)^T \mathbf{S}^t + (\mathbf{S}^t)^T \mathbf{E}^t \right) + \frac{\eta^2}{N^2} (\mathbf{E}^t)^T \mathbf{E}^t \,, \end{aligned} \tag{187}$$

$$\begin{aligned} &= \mathbf{I}_c - \frac{\eta}{N} \left( (\hat{\mathbf{B}}^t)^T \mathbf{S}^t + (\mathbf{S}^t)^T \hat{\mathbf{B}}^t \right) + \frac{\eta}{N} \left( (\hat{\mathbf{B}}^t)^T \mathbf{E}^t + (\mathbf{E}^t)^T \hat{\mathbf{B}}^t \right) \\ &\quad + \frac{\eta^2}{N^2} (\mathbf{S}^t)^T \mathbf{S}^t - \frac{\eta^2}{N^2} \left( (\mathbf{E}^t)^T \mathbf{S}^t + (\mathbf{S}^t)^T \mathbf{E}^t \right) + \frac{\eta^2}{N^2} (\mathbf{E}^t)^T \mathbf{E}^t \,. \end{aligned} \tag{188}$$

By Weyl's inequality, we have

$$\begin{aligned} & \sigma^2_{\min}\left(\mathbf{R}^{t+1}\right) \\ & \geq 1 - \frac{\eta}{N} \lambda_{\max} \left( (\hat{\mathbf{B}}^t)^T \mathbf{S}^t + (\mathbf{S}^t)^T \hat{\mathbf{B}}^t \right) \\ & - \frac{\eta}{N} \lambda_{\max} \left( (\hat{\mathbf{B}}^t)^T \mathbf{E}^t + (\mathbf{E}^t)^T \hat{\mathbf{B}}^t \right) - \frac{\eta^2}{N^2} \lambda_{\max} \left( (\mathbf{E}^t)^T \mathbf{S}^t + (\mathbf{S}^t)^T \mathbf{E}^t \right) . \end{aligned} \tag{189}$$

Then we can define

$$R_1 = \frac{\eta}{N} \lambda_{\max} \left( (\hat{\mathbf{B}}^t)^T \mathbf{S}^t + (\mathbf{S}^t)^T \hat{\mathbf{B}}^t \right) , \tag{190}$$

$$R_2 = \frac{\eta}{N} \lambda_{\max} \left( (\hat{\mathbf{B}}^t)^T \mathbf{E}^t + (\mathbf{E}^t)^T \hat{\mathbf{B}}^t \right) , \tag{191}$$

$$R_3 = \frac{\eta^2}{N^2} \lambda_{\max} \left( (\mathbf{E}^t)^T \mathbf{S}^t + (\mathbf{S}^t)^T \mathbf{E}^t \right) . \tag{192}$$

Then we have

$$\sigma_{\min}^2\left(\mathbf{R}^{t+1}\right) \geq 1 - R_1 - R_2 - R_3\,. \tag{193}$$

Then we consider to bound $R_1, R_2$, and $R_3$, respectively. Consider $R_1$ first, and we have

$$\begin{aligned}
&R_1\\
&= \frac{2\eta}{N}\max_{\|\mathbf{p}\|_2=1}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\mathbf{S}^t\mathbf{p}\,, && (194)\\
&= \frac{2\eta}{N}\max_{\|\mathbf{p}\|_2=1}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\left(\sum_{k=1}^K\left(\bar{\boldsymbol{\Omega}}_k^T\odot\mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\boldsymbol{\Omega}}_k\right)\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t-\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)(\boldsymbol{\theta}_k^t)^T\right)\mathbf{p}\,, && (195)\\
&\leq \frac{2\eta}{N}\max_{\|\mathbf{p}\|_2=1}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T(\sum_{k=1}^K(\left(\bar{\boldsymbol{\Omega}}_k^T\odot\mathbf{X}^T\right)\left(\mathbf{X}\odot\bar{\boldsymbol{\Omega}}_k\right)\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t-\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)\\
&-\hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t-\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right))(\boldsymbol{\theta}_k^t)^T)\mathbf{p}\\
&+\frac{2\eta}{N}\max_{\|\mathbf{p}\|_2=1}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\left(\sum_{k=1}^K\hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t-\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)(\boldsymbol{\theta}_k^t)^T\right)\mathbf{p}\,. && (196)
\end{aligned}$$

Under the condition $\mathcal{E}_4$, we have

$$\begin{aligned}
R_1 &\leq 2\eta\mathcal{C}_2\left\|\hat{\mathbf{B}}^t\right\|_2\frac{\sqrt{d+c}}{\sqrt{N}}\epsilon+\frac{2\eta}{N}\max_{\|\mathbf{p}\|_2=1}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\left(\sum_{k=1}^K\hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t-\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)(\boldsymbol{\theta}_k^t)^T\right)\mathbf{p}\,, && (197)\\
&\leq\left(2\eta\mathcal{C}_2\frac{\sqrt{d+c}}{\sqrt{N}}\right)\left(\left(\sqrt{c}+\frac{1}{\sqrt{10}}\right)^2 dist\left(\hat{\mathbf{B}}^t,\hat{\mathbf{B}}^*\right)+\left(\frac{1}{400}+\frac{\sqrt{c}}{20}\right)\left(\sigma^2+1\right)\right)\\
&+\frac{2\eta}{N}\max_{\|\mathbf{p}\|_2=1}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\left(\sum_{k=1}^K\hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t-\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)(\boldsymbol{\theta}_k^t)^T\right)\mathbf{p}\,, && (198)\\
&\leq 4\eta\left(\mathcal{C}_2\frac{\sqrt{d+c}}{\sqrt{N}}\right)\left(\left(\sigma^2+1\right)\left(\sqrt{c}+\frac{1}{\sqrt{10}}\right)^2\right)\\
&+\frac{2\eta}{N}\max_{\|\mathbf{p}\|_2=1}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\left(\sum_{k=1}^K\hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t-\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)(\boldsymbol{\theta}_k^t)^T\right)\mathbf{p} && (199)
\end{aligned}$$

The remaining thing is to bound $\frac{2\eta}{N}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\left(\sum_{k=1}^K\hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t-\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)(\boldsymbol{\theta}_k^t)^T\right)\mathbf{p}$, and we have

$$\begin{aligned}
&\frac{2\eta}{N}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\left(\sum_{k=1}^K\hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t-\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)(\boldsymbol{\theta}_k^t)^T\right)\mathbf{p}\\
&=\frac{2\eta}{N}tr\left[\sum_{k=1}^K\hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t-\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)(\boldsymbol{\theta}_k^t)^T\mathbf{p}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\right]\,, && (200)\\
&=\frac{2\eta}{N}tr\left[\sum_{k=1}^K\hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t-\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)\left((\hat{\mathbf{B}}^t)^T\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*+\mathbf{F}_k^t+\mathbf{G}_k^t\right)^T\mathbf{p}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\right]\,. && (201)
\end{aligned}$$

Then we can define

$$T_1=\frac{2\eta}{N}tr\left[\sum_{k=1}^K\hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t-\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)(\boldsymbol{\theta}_k^*)^T(\hat{\mathbf{B}}^*)^T\hat{\mathbf{B}}^t\mathbf{p}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\right]\,, \tag{202}$$

$$T_2=\frac{2\eta}{N}tr\left[\sum_{k=1}^K\hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t-\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)(\mathbf{F}_k^t)^T\mathbf{p}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\right]\,, \tag{203}$$

$$T_3=\frac{2\eta}{N}tr\left[\sum_{k=1}^K\hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t-\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)(\mathbf{G}_k^t)^T\mathbf{p}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\right]\,. \tag{204}$$

Consider $T_1$ first, we have

$$
\begin{aligned}
T_1 &= \frac{2\eta}{N} tr[\sum_{k=1}^{K} \hat{N}_k \left(\hat{\mathbf{B}}^t(\hat{\mathbf{B}}^t)^T\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* + \hat{\mathbf{B}}^t\mathbf{F}_k^t + \hat{\mathbf{B}}^t\mathbf{G}_k^t - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right) \\
&\quad (\boldsymbol{\theta}_k^*)^T(\hat{\mathbf{B}}^*)^T\hat{\mathbf{B}}^t\mathbf{p}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T] \,, && (205)\\
&= \frac{2\eta}{N} tr\left[\left(\hat{\mathbf{B}}^t(\hat{\mathbf{B}}^t)^T - \mathbf{I}_d\right)\sum_{k=1}^{K}\hat{N}_k\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*(\boldsymbol{\theta}_k^*)^T(\hat{\mathbf{B}}^*)^T\hat{\mathbf{B}}^t\mathbf{p}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\right] \\
&\quad + \frac{2\eta}{N} tr\left[\sum_{k=1}^{K}\hat{N}_k\left(\hat{\mathbf{B}}^t\mathbf{F}_k^t(\boldsymbol{\theta}_k^*)^T\right)(\hat{\mathbf{B}}^*)^T\hat{\mathbf{B}}^t\mathbf{p}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\right] \\
&\quad + \frac{2\eta}{N} tr\left[\sum_{k=1}^{K}\hat{N}_k\left(\hat{\mathbf{B}}^t\mathbf{G}_k^t(\boldsymbol{\theta}_k^*)^T\right)(\hat{\mathbf{B}}^*)^T\hat{\mathbf{B}}^t\mathbf{p}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\right] \,, && (206)\\
&= 2\eta tr\left[(\hat{\mathbf{B}}^t)^T\hat{\mathbf{B}}^t\left((\mathbf{F}^t)^T + (\mathbf{G}^t)^T\right)\mathbf{W}\boldsymbol{\Theta}^*(\hat{\mathbf{B}}^*)^T\hat{\mathbf{B}}^t\mathbf{p}\mathbf{p}^T\right] \,, && (207)\\
&\le 2\eta\frac{\max_k N_k}{N}\left(\left\|\mathbf{F}^t\right\|_F + \left\|\mathbf{G}^t\right\|_F\right)\left\|\boldsymbol{\Theta}^*\right\|_2 \,. && (208)
\end{aligned}
$$

Because $\mathcal{E}_2$ holds, we have

$$
\begin{aligned}
T_1 &\le 2\eta\frac{\max_k N_k}{N}\left(\frac{\delta}{(1-\delta)\sqrt{K}}\left\|\boldsymbol{\Theta}^*\right\|_2^2 + \sqrt{K}\frac{\delta}{1-\delta}\sigma^2\left\|\boldsymbol{\Theta}^*\right\|_2\right) \,, && (209)\\
&\le 2\eta K\frac{\max_k N_k}{N}\left(\frac{\delta}{1-\delta}\bar{\sigma}_{\max,*}^2 + \frac{\delta}{1-\delta}\sigma^2\bar{\sigma}_{\max,*}\right) \,, && (210)\\
&\le \frac{\eta K}{5}\frac{\max_k N_k}{N}E_0\bar{\sigma}_{\min,*}^2 \,. && (211)
\end{aligned}
$$

Then consider $T_2$, we have

$$
\begin{aligned}
T_2 &= \frac{2\eta}{N} tr\left[\sum_{k=1}^{K}\hat{N}_k\left(\hat{\mathbf{B}}^t\boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)(\mathbf{F}_k^t)^T\mathbf{p}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\right] \,, && (212)\\
&= \frac{2\eta}{N} tr\left[\sum_{k=1}^{K}\hat{N}_k\left(\hat{\mathbf{B}}^t(\hat{\mathbf{B}}^t)^T\hat{\mathbf{B}}^*\boldsymbol{\theta}_k^* + \hat{\mathbf{B}}^t\mathbf{F}_k^t + \hat{\mathbf{B}}^t\mathbf{G}^t - \hat{\mathbf{B}}^*\boldsymbol{\theta}_k^*\right)(\mathbf{F}_k^t)^T\mathbf{p}\mathbf{p}^T(\hat{\mathbf{B}}^t)^T\right] \,, && (213)\\
&= 2\eta tr\left[(\hat{\mathbf{B}}^t)^T\hat{\mathbf{B}}^t\left((\mathbf{F}^t)^T\mathbf{F}^t + (\mathbf{G}^t)^T\mathbf{F}^t\right)\mathbf{W}\mathbf{p}\mathbf{p}^T\right] \,, && (214)\\
&\le 2\eta\frac{\max_k N_k}{N}\left(\left\|\mathbf{F}^t\right\|_F^2 + \left\|\mathbf{F}^t\right\|_F\left\|\mathbf{G}^t\right\|_F\right) \,, && (215)\\
&\le 2\eta\frac{\max_k N_k}{N}\left(\frac{\delta^2}{(1-\delta)^2}\bar{\sigma}_{\max,*}^2 + \frac{\delta^2}{(1-\delta)^2}\sqrt{K}\bar{\sigma}_{\max,*}\sigma^2\right) \,, && (216)\\
&\le \frac{\eta K}{100}\frac{\max_k N_k}{N}E_0\bar{\sigma}_{\min,*}^2 \,. && (217)
\end{aligned}
$$

Finally, consider $T_3$, we have

$$T_3 = \frac{2\eta}{N} tr\left[\sum_{k=1}^{K} \hat{N}_k \left(\hat{\mathbf{B}}^t \boldsymbol{\theta}_k^t - \hat{\mathbf{B}}^* \boldsymbol{\theta}_k^*\right) (\mathbf{G}_k^t)^T \mathbf{p}\mathbf{p}^T (\hat{\mathbf{B}}^t)^T\right], \tag{218}$$

$$= \frac{2\eta}{N} tr\left[\sum_{k=1}^{K} \hat{N}_k \left(\hat{\mathbf{B}}^t (\hat{\mathbf{B}}^t)^T \hat{\mathbf{B}}^* \boldsymbol{\theta}_k^* + \hat{\mathbf{B}}^t \mathbf{F}_k^t + \hat{\mathbf{B}}^t \mathbf{G}^t - \hat{\mathbf{B}}^* \boldsymbol{\theta}_k^*\right) (\mathbf{G}_k^t)^T \mathbf{p}\mathbf{p}^T (\hat{\mathbf{B}}^t)^T\right], \tag{219}$$

$$= 2\eta tr\left[(\hat{\mathbf{B}}^t)^T \hat{\mathbf{B}}^t \left((\mathbf{G}^t)^T \mathbf{G}^t + (\mathbf{F}^t)^T \mathbf{G}^t\right) \mathbf{W}\mathbf{p}\mathbf{p}^T\right], \tag{220}$$

$$\leq 2\eta \frac{\max_k N_k}{N} \left(\left\|\mathbf{G}^t\right\|_F^2 + \left\|\mathbf{F}^t\right\|_F \left\|\mathbf{G}^t\right\|_F\right), \tag{221}$$

$$\leq 2\eta \frac{\max_k N_k}{N} \left(\frac{\delta^2}{(1-\delta)^2} K\sigma^4 + \frac{\delta^2}{(1-\delta)^2} \sqrt{K}\bar{\sigma}_{\max,*}\sigma^2\right), \tag{222}$$

$$\leq \frac{\eta K}{100} \frac{\max_k N_k}{N} E_0 \bar{\sigma}_{\min,*}^2 . \tag{223}$$

Combining Equation (211), (217), and (223), we have

$$R_1 \leq 4\eta \left(\mathcal{C}_2 \frac{\sqrt{d+c}}{\sqrt{N}}\right) \left((\sigma^2+1)\left(\sqrt{c} + \frac{1}{\sqrt{10}}\right)^2\right) + \frac{11\eta K}{50} \frac{\max_k N_k}{N} E_0 \bar{\sigma}_{\min,*}^2 . \tag{224}$$

Then consider $R_2$, because condition $\mathcal{E}_3$ holds, we have

$$R_2 = \frac{2\eta}{N} \max_{\|\mathbf{p}\|_2=1} \mathbf{p}^T (\hat{\mathbf{B}}^t)^T \mathbf{E}^t \mathbf{p}, \tag{225}$$

$$= \frac{2\eta}{N} \max_{\|\mathbf{p}\|_2=1} \mathbf{p}^T (\hat{\mathbf{B}}^t)^T \left(\sum_{k=1}^{K} \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right) \left(\mathbf{Z} \odot \boldsymbol{\Omega}_k\right) \left(\boldsymbol{\theta}_k^t\right)^T\right) \mathbf{p}, \tag{226}$$

$$\leq 2\eta \left\|\frac{1}{N} \sum_{k=1}^{K} \left(\bar{\boldsymbol{\Omega}}_k^T \odot \mathbf{X}^T\right) \left(\mathbf{Z} \odot \boldsymbol{\Omega}_k\right) \left(\boldsymbol{\theta}_k^t\right)^T\right\|_2, \tag{227}$$

$$\leq 2\eta \mathcal{C}_1 \sigma^2 \frac{\sqrt{d+c}}{\sqrt{N}} \left(\sqrt{c} + \frac{1}{10}\right). \tag{228}$$

Then consider $R_3$, when condition $\mathcal{E}_3, \mathcal{E}_4, \mathcal{E}_5$, and $\mathcal{E}_6$ hold, we have

$$R_3 = \frac{\eta^2}{N^2} \max_{\|\mathbf{p}\|_2=1} \mathbf{p}^T (\mathbf{E}^t)^T \mathbf{S}^t \mathbf{p}, \tag{229}$$

$$\leq \frac{2\eta^2}{N} \left(\mathcal{C}_1 \sigma^2 \frac{\sqrt{d+c}}{\sqrt{N}} \left(\sqrt{c} + \frac{1}{10}\right)\right) \left(\mathcal{C}_2 \frac{\sqrt{d+c}}{\sqrt{N}} + 1\right) \left((\sigma^2+1)\left(\sqrt{c} + \frac{1}{\sqrt{10}}\right)^2\right), \tag{230}$$

$$\leq \mathcal{C} \frac{2\eta^2 \sqrt{d+c}}{N^{3/2}} (\sigma^2+1)^2 \left(\sqrt{c} + \frac{1}{\sqrt{10}}\right)^3. \tag{231}$$

The last equation holds when $N$ satisfy $N \geq \frac{K^2}{d+c}$. Then combine Equation (193), (199), (228), and (231), because $N \geq (\sqrt{c}+1)(\sigma^2+1)$ holds, we have

$$\sigma_{\min}^2(\mathbf{R}^{t+1}) \geq 1 - R_1 - R_2 - R_3, \tag{232}$$

$$\geq 1 - 8\eta\mathcal{C} \left(\frac{\sqrt{d+c}}{\sqrt{N}}\right) (\sigma^2+1) \left(\sqrt{c} + \frac{1}{\sqrt{10}}\right)^2 - \frac{11\eta K}{50} \frac{\max_k N_k}{N} E_0 \bar{\sigma}_{\min,*}^2 . \tag{233}$$

Combining Equation (183) and (233), we have

$$
\begin{aligned}
& dist\left(\hat{\mathbf{B}}^{t+1}, \hat{\mathbf{B}}^{*}\right) \leq dist\left(\hat{\mathbf{B}}^{t}, \hat{\mathbf{B}}^{*}\right) \\
& \left(1-\eta K\left(\left(\frac{\min_k \hat{N}_k}{N}\right)-\frac{1}{4}\left(\frac{\max_k \hat{N}_k}{N}\right)\right) \bar{\sigma}_{\min,*}^2 E_0+\eta \mathcal{C} \frac{\sqrt{d+c}}{\sqrt{N}}\left(\sqrt{c}+\frac{1}{\sqrt{10}}\right)^2\right) \\
& \sigma_{\min}(\mathbf{R}^{t+1})^{-1} \\
& +\eta \mathcal{C} \frac{\sqrt{d+c}}{\sqrt{N}}\left(\sigma^2+1\right)\left(2\sqrt{c}+\frac{1}{5}\right) \sigma_{\min}(\mathbf{R}^{t+1})^{-1} . \qquad (234)
\end{aligned}
$$

When $\frac{\max_k \hat{N}_k}{N}$ satisfy

$$
\frac{\max_k \hat{N}_k}{N} \geq \frac{200}{7} \mathcal{C} \frac{\sqrt{d+c}}{\sqrt{N}} \frac{(\sigma^2+1)(\sqrt{c}+1)^2}{K E_0 \bar{\sigma}_{\min,*}^2}, \qquad (235)
$$

we will have

$$
\sigma_{\min}^2(\mathbf{R}^{t+1}) \geq 1-\frac{1}{2} \eta K\left(\frac{\max_k \hat{N}_k}{N}\right) \bar{\sigma}_{\min,*}^2 E_0, \qquad (236)
$$

$$
\eta \mathcal{C} \frac{\sqrt{d+c}}{\sqrt{N}}\left(\sqrt{c}+\frac{1}{\sqrt{10}}\right)^2 \leq \frac{7}{200} \eta K\left(\frac{\max_k \hat{N}_k}{N}\right) \bar{\sigma}_{\min,*}^2 E_0 . \qquad (237)
$$

Then we have

$$
\begin{aligned}
& dist\left(\hat{\mathbf{B}}^{t+1}, \hat{\mathbf{B}}^{*}\right) \leq dist\left(\hat{\mathbf{B}}^{t}, \hat{\mathbf{B}}^{*}\right) \\
& \left(1-\eta K\left(\left(\frac{\min_k \hat{N}_k}{N}\right)-\frac{57}{200}\left(\frac{\max_k \hat{N}_k}{N}\right)\right) \bar{\sigma}_{\min,*}^2 E_0\right) \\
& \left(1-\frac{1}{2} \eta K\left(\frac{\max_k \hat{N}_k}{N}\right) \bar{\sigma}_{\min,*}^2 E_0\right)^{-1/2} \\
& +\left(\frac{7}{100} \eta K\left(\frac{\max_k \hat{N}_k}{N}\right) \bar{\sigma}_{\min,*}^2 E_0\right)\left(1-\frac{1}{2} \eta K\left(\frac{\max_k \hat{N}_k}{N}\right) \bar{\sigma}_{\min,*}^2 E_0\right)^{-1/2} . \qquad (238)
\end{aligned}
$$

□

**Remark B.13.** *Different from previous studies that assuming clusters to have the same number of data points (Collins et al., 2021; Tziotis et al., 2022), we analysed the impact of imbalanced sampled numbers among clusters in Theorem B.12. We can see that* HCFL$^+$*'s convergence is strongly influenced by* $\max_k \hat{N}_k$ *and* $\min_k \hat{N}_k$*. In other words,* HCFL$^+$ *converges faster as* $\frac{\min_k \hat{N}_k}{\max_k \hat{N}_k}$ *increases. This suggests that* HCFL$^+$ *performs better when the number of samples is evenly distributed among all* $K$ *underlying clusters. In contrary, if the number of samples is extremely imbalanced, the* HCFL$^+$ *may hard to converge to optimum.*

## C Related Works

**Federated Learning.** As the de-facto algorithm in FL, FedAvg employs local SGD (McMahan et al., 2016; Lin et al., 2020) to reduce communication costs and protect client privacy. However, distribution shifts among clients pose a significant challenge in FL and hinder the performance of FL algorithms (Li et al., 2018; Wang et al., 2020; Karimireddy et al., 2020; Jiang & Lin, 2023; Guo et al., 2021). Traditional FL methods primarily aim to improve the convergence speed of global models and incorporate bias reduction techniques (Tang et al., 2022; Guo et al., 2023a; Li et al., 2021; 2018). At the same time, some studies investigate feature distribution shifts using domain generalization techniques (Peng et al., 2019; Wang et al., 2022a; Shen et al., 2021; Sun et al., 2022; Gan et al., 2021). However, single-model approaches are inadequate for handling heterogeneous data distributions, especially when dealing with concept shifts (Ke et al., 2022; Guo et al., 2023b; Jothimurugesan et al., 2023). To tackle these challenges, clustered FL algorithms are introduced to enhance FL algorithm performance.

**Clustered FL with fixed cluster numbers.** Clustered FL groups clients based on their local data distribution, tackling the distribution shift problem. Most methods employ hard clustering with a fixed number of clusters, grouping clients by various similarity metrics, such as local loss values (Ghosh et al., 2020), local model parameter differences (Long et al., 2023), communication time/local calculation time (Wang et al., 2022b), and fuzzy $c$-Means (Stallmann & Wilbik, 2022). However, hard clustering may not capture complex relationships between local distributions adequately, and soft clustering paradigms have been proposed to address this issue. For instance, FedEM (Marfoq et al., 2021) employs Expectation-Maximization techniques to maximize likelihood functions. FedGMMcitepwu2023personalized suggests using joint distributions instead of conditional distributions. FedRC(Guo et al., 2023b) introduces Robust Clustering, assigning clients with concept shifts to different clusters to enhance model generalization. FedSoft (Ruan & Joe-Wong, 2022) calculates weights based on the distances between clients' local model parameters and cluster model parameters, with smaller distances indicating larger weights for that cluster. In this paper, we propose a generalized formulation for clustered FL that encompasses the current methods and improves them by addressing issues related to intra-client inconsistency and efficiency.

**Clustered FL with adaptive clustering numbers.** Another line of research focuses on automatically determining the number of clusters. Current methods utilize hierarchical clustering, which measures client dissimilarity using model parameters or local gradient distances. Most current methods modify cluster numbers by splitting them when client distances within clusters are large (Sattler et al., 2020b;a; Zhao et al., 2020; Briggs et al., 2020; Duan et al., 2021a;b). Recently, StoCFL (Zeng et al., 2023) suggests initially setting cluster numbers equal to the client count and merging clusters with small distances. In addition to model parameter distances, some papers employ alternative distance metrics for improved performance. For instance, (Yan et al., 2023) employ principal eigenvectors of model parameters. (Vahidian et al., 2023) use truncated singular value decomposition (SVD) to obtain a reduced set of principal vectors for distance measurement. Meanwhile, (Wei & Huang, 2023) focus on the distance of normalized local features. FEDCOLLAB (Bao et al., 2023) focuses on cross-silo scenarios with a limited number of clients and quantifies client similarity by training client discriminators. However, the need for discriminators between every pair of clients in FEDCOLLAB makes it challenging to expand to cross-device scenarios with numerous clients. In this paper, we concentrate on cross-device settings, introducing a holistic adaptive clustering framework enabling cluster splitting and merging. We also present enhanced weight updating for soft clustering and finer distance metrics for various clustering principles.

## D ALGORITHMS

**Details of the HCFL$^+$.** In Algorithm 2, we present a concise summary of the comprehensive algorithm that integrates all the enhanced components of HCFL$^+$, as introduced in Section 4. Specifically, during each communication round, the algorithm performs the following steps: (1) Randomly selects a subset of clients. (2) Calculates prototypes using Equations (11) and (12). (3) Performs local updates using Algorithm 3. (4) The server aggregates local updates, updates cluster model parameters, and computes client distance metrics using Equation (9) for each cluster $k$. (5) Identifies $k_{max}$ as the cluster with the highest average distance. (6) Checks if the maximum distance within $k_{max}$ significantly exceeds the average distance in this cluster. (7) If the following condition is met, splits the clusters using Algorithm 4.

$$\max(D^t_{k_{max}}) - \text{mean}(D^t_{k_{max}}) \geq \rho\,. \tag{239}$$

(8) Mark and remove the empty clusters no clients will assign large clustering weights to using Algorithm 5.

**Intuitions on the distance metrics design.** From the objective function (Eq. (2)), we should assign higher clustering weights $\omega_{i,j;k}$ to clusters with greater $\mathcal{L}_k$ to maximize the objective function. Because the ultimate goals of the clustering algorithms are solving the objective functions, we analyse the $\mathcal{L}_k$ to check the key factors influencing the value of $\mathcal{L}_k$, and the relationships between these factors and the clustering principles.

We use the following algorithms as examples. For FedEM (Marfoq et al., 2021) and IFCA (Ghosh et al., 2020), $\mathcal{L}_k(\mathbf{x}, y, \boldsymbol{\phi}, \boldsymbol{\theta}_k) = \mathcal{P}_{\boldsymbol{\phi},\boldsymbol{\theta}_k}(y|\mathbf{x})$; For FedRC (Guo et al., 2023b), $\mathcal{L}_k(\mathbf{x}, y, \boldsymbol{\phi}, \boldsymbol{\theta}_k) =$

$\frac{\mathcal{P}_{\boldsymbol{\phi},\boldsymbol{\theta}_k}(\mathbf{x},y)}{\mathcal{P}_{\boldsymbol{\phi},\boldsymbol{\theta}_k}(\mathbf{x})\mathcal{P}_{\boldsymbol{\phi},\boldsymbol{\theta}_k}(y)}$. Defining $\mathbf{z} = g(\mathbf{x};\boldsymbol{\phi})$ as the local features extracted by $\boldsymbol{\phi}$, assuming a $\mathbf{x} \to \mathbf{z} \to y$ Probabilistic Graphical Model (with $\mathbf{x}$ and $y$ being independent given $\mathbf{z}$), we obtain:

$$\mathcal{L}_k(\mathbf{x}, y, \boldsymbol{\phi}, \boldsymbol{\theta}_k) = \left\{ \begin{array}{ll} \mathcal{P}(y|\mathbf{x};\boldsymbol{\phi},\boldsymbol{\theta}_k) = \frac{\mathcal{P}(y|\mathbf{z};\boldsymbol{\theta}_k)\mathcal{P}(\mathbf{z}|\mathbf{x};\boldsymbol{\phi})}{\mathcal{P}(\mathbf{z}|\mathbf{x},y;\boldsymbol{\phi})} & \text{(FedEM, IFCA)} \\ \frac{\mathcal{P}(\mathbf{x},y;\boldsymbol{\phi},\boldsymbol{\theta}_k)}{\mathcal{P}(y;\boldsymbol{\phi},\boldsymbol{\theta}_k)\mathcal{P}(\mathbf{x};\boldsymbol{\phi},\boldsymbol{\theta}_k)} = \frac{\mathcal{P}(y|\mathbf{z};\boldsymbol{\theta}_k)\mathcal{P}(\mathbf{z}|\mathbf{x};\boldsymbol{\phi})}{\mathcal{P}(y;\boldsymbol{\phi},\boldsymbol{\theta}_k)\mathcal{P}(\mathbf{z}|\mathbf{x},y;\boldsymbol{\phi})} & \text{(FedRC)} \end{array} \right\}$$
$$= \frac{\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)\mathcal{P}(\mathbf{z}|\mathbf{x};\boldsymbol{\phi})}{\mathcal{P}(\mathbf{z}|\mathbf{x}, y; \boldsymbol{\phi})}.$$

Then we aim to give the following explanations of the three terms $\mathcal{P}(\mathbf{z}|\mathbf{x};\boldsymbol{\phi})$, $\mathcal{P}(\mathbf{z}|\mathbf{x};\boldsymbol{\phi})$, and $\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)$, which align with the terms considered in Sec 4.4.

- $\mathcal{P}(\mathbf{z}|\mathbf{x};\boldsymbol{\phi})$ **for feature and label shifts**. Feature shifts introduce significant distances in $\mathbf{x}$. Additionally, $\mathbf{x}$ with different $y$ values generally exhibit substantial distances in the feature space. Without this, classifiers cannot distinguish samples with different labels. Hence, we employ $\mathcal{P}(\mathbf{z}|\mathbf{x};\boldsymbol{\phi})$ to assess both feature and label shifts.
- $\mathcal{P}(\mathbf{z}|\mathbf{x}, y;\boldsymbol{\phi})$ **for concept shifts.** Concept shifts signify alerted $\mathbf{x} - y$ correlations. Hence, samples with concept shifts but have the same $y$ should exhibit a significant difference in $\mathcal{P}(\mathbf{z}|\mathbf{x}, y;\boldsymbol{\phi})$.
- $\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)$ **for the quality of clustering.** The $\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)$ is defined using features $\mathbf{z} = g(\mathbf{x};\boldsymbol{\phi})$ instead of data $\mathbf{x}$ in $\mathcal{L}_k(\mathbf{x}, y, \boldsymbol{\phi}, \boldsymbol{\theta}_k)$. This term evaluates if features $\mathbf{z}$ can be correctly assigned to clusters given the current $\boldsymbol{\Theta}$; otherwise, the objectives in (2) cannot be achieved.

Finally, we propose the following distance metric:

$$\mathbf{D}^k_{i,j} = \left\{ \begin{array}{ll} \max\{d_c, d_{lf}\}\, \mathbb{E}_{D_i}\left[\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)\right] \mathbb{E}_{D_j}\left[\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)\right], & \text{ASCP} \\ d_c \mathbb{E}_{D_i}\left[\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)\right] \mathbb{E}_{D_j}\left[\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)\right], & \text{CSCP} \end{array} \right. \tag{240}$$

where dist is the cos-similarity in this paper, and

$$d_c = \max_y \left\{ \text{dist}\left(\mathbb{E}_{D_i}\left[\mathcal{P}(\mathbf{z}|\mathbf{x}, y;\boldsymbol{\phi})\right], \mathbb{E}_{D_j}\left[\mathcal{P}(\mathbf{z}|\mathbf{x}, y;\boldsymbol{\phi})\right]\right)\right\}, \tag{241}$$

$$d_{lf} = \text{dist}\left(\mathbb{E}_{D_i}\left[\mathcal{P}(\mathbf{z}|\mathbf{x};\boldsymbol{\phi})\right], \mathbb{E}_{D_j}\left[\mathcal{P}(\mathbf{z}|\mathbf{x};\boldsymbol{\phi})\right]\right) \tag{242}$$

The distances above become large only when the following conditions occur together: (1) Large values of $d_c$ indicate concept shifts between clients $i$ and $j$; (2) Large $d_{lf}$ indicate significant feature and label distribution differences. (2) Large values of $\mathbb{E}_{D_i}\left[\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)\right] \mathbb{E}_{D_j}\left[\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)\right]$ indicate incorrect clustering weights with high confidence.

**Approximation of the distance metrics in practice.** When calculating the distance metrics (Equation (9)) in practice, to avoid training extra generative networks and transmitting more data between servers and clients, we substitute $\tilde{\omega}_{i;k}$ for $\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)$ since $\tilde{\omega}_{i;k}$ is positively correlated with $\tilde{\mathcal{L}}_k(\mathbf{z}, y; \boldsymbol{\theta}_k)$ (Marfoq et al., 2021; Guo et al., 2023b). Additionally, we approximate $\mathbb{E}_{D_i}\left[\mathcal{P}(\mathbf{z}|\mathbf{x}, y;\boldsymbol{\phi})\right]$ and $\mathbb{E}_{D_i}\left[\mathcal{P}(\mathbf{z}|\mathbf{x}, y;\boldsymbol{\phi})\right]$ using feature prototypes. The prototypes are defined by the following equation:

$$\tilde{d}_c = Dist(\mathbf{P}_{c,i}, \mathbf{P}_{c,j}), \tilde{d}_{lf} = Dist(\mathbf{P}_{lf,i}, \mathbf{P}_{lf,j}), \tag{243}$$

where

$$\mathbf{P}_{c,i} \in \mathbb{R}^{d\times C} = [\frac{1}{N_{i,1}} \sum_{j=1}^{N_i} \mathbf{1}_{y_{i,j}=1} g(\mathbf{x}_{i,j}, \boldsymbol{\phi}), \cdots, \frac{1}{N_{i,C}} \sum_{j=1}^{N_i} \mathbf{1}_{y_{i,j}=C} g(\mathbf{x}_{i,j}, \boldsymbol{\phi})], \tag{244}$$

$$\mathbf{P}_{lf,i} \in \mathbb{R}^{d} = \frac{1}{N_i} \sum_{j=1}^{N_i} g(\mathbf{x}_{i,j}, \boldsymbol{\phi}), \tag{245}$$

$N_{i,c} = \sum_{j=1}^{N_i} \mathbf{1}_{y_{i,j}=c}$, $g(\mathbf{x}_{i,j}, \boldsymbol{\phi})$ is the function parameterized by $\boldsymbol{\phi}$, $Dist$ is a function to measure the distance between prototypes, which we use the cosine similarity as an example in this paper.

**Algorithm 2** Algorithm Framework of HCFL$^+$

**Require:** Local datasets $D_1, \ldots, D_N$, number of local iterations $\mathcal{T}$, number of communication rounds $T$, number of clients chosen in each round $S$, initial number of clusters $K^0$, number of classes $C$, and hyper-parameter $\rho$.

**Ensure:** Trained global feature extractor $\boldsymbol{\phi}^T$, final number of clusters $K^T$, and cluster-specific predictors $\boldsymbol{\Theta}^T = [\boldsymbol{\theta}_1^T, \cdots, \boldsymbol{\theta}_{K^T}^T]$.

Initialize $\boldsymbol{\phi}^0, \boldsymbol{\Theta}^0 = [\boldsymbol{\theta}_1^0, \cdots, \boldsymbol{\theta}_{K^0}^0]$.
**for** $t = 0, \ldots, T-1$ **do**
  Choose a subset of clients $\mathcal{S}^t$, where $|\mathcal{S}^t| = S$.
  **for** chosen client $i \in \mathcal{S}^t$ **do**
    Calculate client prototypes $\mathbf{P}_i^t$ by Equation (11)- (12).
    $\mathcal{F}_i^{t+1}, \tilde{\omega}_{i;k}^{t+1}, \boldsymbol{\phi}_i^{\mathcal{T}}, \boldsymbol{\theta}_{k,i}^{\mathcal{T}} \leftarrow$ Local updates by Algorithm 3.
    Send $\mathbf{P}_i^t$, $\mathcal{F}_i^{t+1}$, and $\tilde{\omega}_{i;k}^{t+1}, \boldsymbol{\phi}_i^{\mathcal{T}}, \boldsymbol{\theta}_{k,i}^{\mathcal{T}}, \forall k \leq K^t$ to the server.
  $\mathcal{F}_i^{latest} \leftarrow \mathcal{F}_i^{t+1}$.
  $\boldsymbol{\phi}^{t+1} = \frac{1}{\sum_{i \in \mathcal{S}^t}} \sum_{i \in \mathcal{S}^t} N_i \boldsymbol{\phi}_i^{\mathcal{T}}$.
  $\boldsymbol{\theta}_k^{t+1} = \frac{1}{\sum_{i \in \mathcal{S}^t}} \sum_{i \in \mathcal{S}^t} N_i \boldsymbol{\theta}_{k,i}^{\mathcal{T}}, \forall k \leq K^t$.
  $\mathcal{F}_g^{t+1} \leftarrow [\mathcal{F}_{g,1}^{t+1}, \cdots, \mathcal{F}_{g,K^t}^{t+1}]$, where $\mathcal{F}_{g,k}^{t+1} \leftarrow [\sum_i \mathcal{F}_{i,k,1}^{latest}, \cdots, \sum_i \mathcal{F}_{i,k,C}^{latest}]$.
  Initialize $\mathcal{C}_k^t = \emptyset, \forall k \leq K^t$.
  **for** all client $i$ **do**
    $c_i \leftarrow \arg\max_k \tilde{\omega}_{i;k}$.
    $\mathcal{C}_{c_i}^t \leftarrow \mathcal{C}_{c_i}^t \cup i$.
  $\mathcal{R}^t \leftarrow \emptyset$.
  **for** $k \leq K^t$ **do**
    **if** $\mathcal{C}_k^t$ is empty **then**
      $\mathcal{R}^t \leftarrow \mathcal{R}^t \cup k$.
    Get the cluster-specific distance matrix $\mathbf{D}_k \in \mathbb{R}^{|\tilde{\mathcal{S}}_k^t| \times |\tilde{\mathcal{S}}_k^t|}, \forall k \leq K^t$ by Equation (9).
  $k_{min} \leftarrow \arg\max_k \max(\mathbf{D}_k^t)$.
  **if** $\max(\mathbf{D}_{k_{min}}^t) - mean(\mathbf{D}_{k_{min}}^t) \geq \rho$ **then**
    Split $\mathcal{C}_{k_{min}}^t$ into two clusters $\mathcal{C}_{k_{min},1}^t$ and $\mathcal{C}_{k_{min},2}^t$.
    $\boldsymbol{\theta}_{k_{min}}^{t+1} = \frac{1}{\sum_{i \in \mathcal{C}_{k_{min},1}^t}} \sum_{i \in \mathcal{C}_{k_{min},1}^t} N_i \boldsymbol{\theta}_{k_{min},i}^{\mathcal{T}}$.
    Add new cluster and update $\mathcal{F}_g$ by server side of Algorithm 4.
    $K^{t+1} \leftarrow K^t + 1$.
  **else**
    $K^{t+1} \leftarrow K^t$.
  **for** cluster $k_r \in \mathcal{R}^t$ **do**
    Remove cluster $k_r$ and update $\mathcal{F}_g$ by server side of Algorithm 5.
    $K^{t+1} \leftarrow K^{t+1} - 1$.
  Send $\boldsymbol{\phi}^{t+1}$, $\boldsymbol{\Theta}^{t+1} = [\boldsymbol{\theta}_1^{t+1}, \cdots, \boldsymbol{\theta}_{K^{t+1}}^{t+1}]$, and information about add/remove cluster to clients.

**Algorithm 3** Local Updates of HCFL$^+$

**Require:** Number of local iterations $\mathcal{T}$, current number of clusters $K^t$, number of classes $C$, local dataset $D_i$, global feature extractor $\boldsymbol{\phi}^t$, cluster-specific predictors $\boldsymbol{\Theta}^t = [\boldsymbol{\theta}_1^t, \cdots, \boldsymbol{\theta}_{K^t}^t]$.

**Ensure:** Trained feature extractor $\boldsymbol{\phi}_i^{\mathcal{T}}$, predictors $\boldsymbol{\Theta}_i^{\mathcal{T}} = [\boldsymbol{\theta}_{i,1}^{\mathcal{T}}, \cdots, \boldsymbol{\theta}_{i,K^t}^{\mathcal{T}}]$, $\tilde{\boldsymbol{\Omega}}_i^{t+1} = [\tilde{\omega}_{i;k}, \cdots, \tilde{\omega}_{i;K^t}]$, and $\mathcal{F}_i^{t+1} = [\mathcal{F}_{i,1}^{t+1}, \cdots, \mathcal{F}_{i,K^t}^{t+1}]$, where $\mathcal{F}_{i,k}^{t+1} = [\mathcal{F}_{i,k,1}^{t+1}, \cdots, \mathcal{F}_{i,k,C}^{t+1}]$.

Update $\gamma_{i,j;k}^{t+1}, \tilde{\gamma}_{i,j;k}^{t+1}, \omega_{i,j;k}^{t+1}, \tilde{\omega}_{i;k}^{t+1}$ by Equations (4)-(5), $\forall j \leq N_i, k \leq K^t$. ▷ Tier 2
**for** $\tau = 1, \ldots, \mathcal{T}$ **do** ▷ Tier 1
  Update $\boldsymbol{\theta}_{k,i}^{\tau}$ by Equation (6), $\forall k \leq K^t$.
  Update $\boldsymbol{\phi}_i^{\tau}$ by Equation (7).
$\mathcal{F}_{i,k,c}^{t+1} \leftarrow \sum_{j=1}^{N_i} \mathbf{1}_{y_{i,j}=c} \gamma_{i,j;k}^{t+1}$.

**Algorithm 4** Cluster Adding of HCFL$^+$

**Require:** $k_{min}$, set of clients $\mathcal{C}^t_{k_{min},2}$, the corresponding $\boldsymbol{\theta}^{\mathcal{T}}_{k_{min},i}$ for each client $i \in \mathcal{C}^t_{k_{min},2}$, and $\mathcal{F}_g$.
**Ensure:** New $\mathcal{F}^{t+1}_g$, and predictor of the new cluster $\boldsymbol{\theta}^{t+1}_{K^t+1}$.
1: **Server Side:**
2: $\boldsymbol{\theta}^{t+1}_{K^t+1} = \frac{1}{\sum_{i \in \mathcal{C}^t_{k_{min},2}}} \sum_{i \in \mathcal{C}^t_{k_{min},2}} N_i \boldsymbol{\theta}^{\mathcal{T}}_{k_{min},i}$.
3: Add $\mathcal{F}_{g,K^t+1} \leftarrow \mathcal{F}_{g,k_{min}}$ to $\mathcal{F}^{t+1}_g$.
4: Add $\mathcal{F}^{latest}_{i,K^t+1} \leftarrow \mathcal{F}^{latest}_{i,k_{min}}$ to $\mathcal{F}^{latest}_i, \forall i$.
5: **Client Side:**
6: $\omega_{i,j;K^t+1} \leftarrow \omega_{i,j;k_{min}}/2, \forall j \leq N_i$.
7: $\omega_{i,j;k_{min}} \leftarrow \omega_{i,j;k_{min}}/2, \forall j \leq N_i$.
8: $\tilde{\omega}_{i;K^t+1} \leftarrow \tilde{\omega}_{i;k_min}/2$.
9: $\tilde{\omega}_{i;k_{min}} \leftarrow \tilde{\omega}_{i;k_min}/2$.

**Algorithm 5** Cluster Removing of HCFL$^+$

**Require:** The cluster needs to be removed $k_r$, and $\mathcal{F}^{t+1}_g$.
1: **Server Side:**
2: Remove $\mathcal{F}^{t+1}_{g,k_r}$ from $\mathcal{F}^{t+1}_g$.
3: Remove $\mathcal{F}^{latest}_{i,k_r}$ from $\mathcal{F}^{latest}_i, \forall i$.
4: **Client Side:**
5: $\omega_{i,j;k} \leftarrow \frac{\omega_{i,j;k}}{\sum_{n \neq k_r} \omega_{i,j;n}}, \forall j \leq N_i, k \neq k_r$.
6: $\tilde{\omega}_{i;k} \leftarrow \frac{\omega_{i;k}}{\sum_{n \neq k_r} \omega_{i;n}}, \forall k \neq k_r$.
7: Remove $\gamma_{i,j;k_r}, \tilde{\gamma}_{i,j;k_r}, \omega_{i,j;k_r}, \tilde{\omega}_{i,j;k_r}, \forall j \leq N_i$.

# E EXPERIMENT RESULTS

## E.1 DATASETS AND MODELS

**Diverse distribution shifts scenarios.** Similar to previous work (Guo et al., 2023b), the diverse distribution shift scenario construct clients with three types of distribution shifts with each other:

- **Label Distribution Shifts:** We use the idea introduced (Yoshida et al., 2019; Hsu et al., 2019; Reddi et al., 2021), where we leverage the Latent Dirichlet Allocation (LDA) with $\alpha = 1.0$. We split datasets to 100 clients by default.
- **Feature Distribution Shifts:** We utilize the idea of constructing CIFAR10-C, CIFAR100-C, and ImageNet-C (Hendrycks & Dietterich, 2019). In detail, we apply random augmentations to client samples, selecting from 20 types, including 'Original', 'Gaussian Noise', 'Shot Noise', 'Impulse Noise', 'Defocus Blur', 'Glass Blur', 'Motion Blur', 'Zoom Blur', 'Snow', 'Frost', 'Fog', 'Brightness', 'Contrast', 'Elastic', 'Pixelate', 'JPEG', 'Speckle Noise', 'Gaussian Blur', 'Spatter', and 'Saturate'. Augmentation types remain consistent within each client.
- **Concept Shifts:** For label $y \leq C_\beta$, it becomes $y$, $(1+y)\%C_\beta$, and $(2+y)\%C_\beta$ across concepts, where $C_\beta = \lfloor C * \beta \rfloor$, and $C$ is the number of classes.

**Noisy label scenarios.** We follow the methodology of previous works (Fang & Ye, 2022; Ke et al., 2022) to construct noisy label scenarios. Our approach involves two types of noisy labels: symmetric flip and pair flip. Symmetric flip entails randomly flipping the original class label to any wrong class label with equal probability. Pair flip involves flipping the original class label only to a very similar wrong category. We use the parameter $\chi$ to control the noisy rate, where $\chi = 0.1$ indicates that $10\%$ of the data have wrong labels.

## E.2 BASELINES AND HYPER-PARAMETER SETTINGS

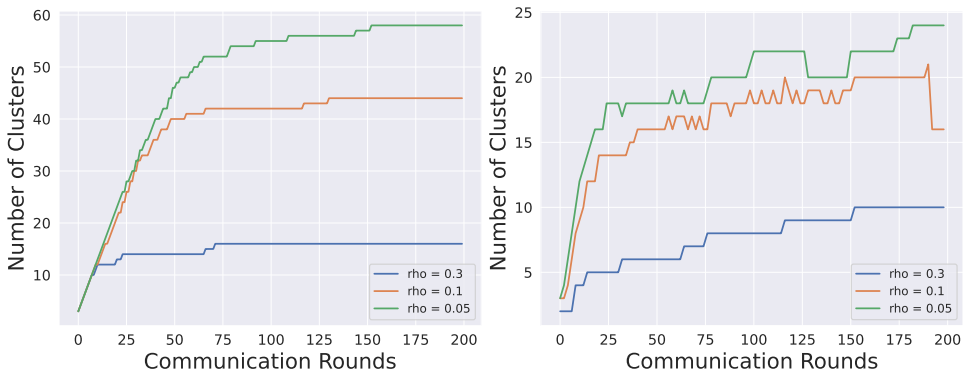
**Detailed implementations and hyper-parameter settings for all the algorithms** Unless special mentioned, we split each dataset to 100 clients with 3 concepts. The learning rates are chosen in $\{0.03, 0.06, 1.0\}$, and we report the best results for each algorithm. We run the algorithms for 200 communication rounds and set the number of local epochs to 1. The experiments are conducted on single NVIDIA RTX 3090 GPUs.

**Detailed implementations and hyper-parameter settings of baseline algorithms.** The details of the settings and hyper-parameters we used for the baseline methods a summarized below. We exclude the algorithms that do not require additional hyper-parameters here.

- **CFL** (Sattler et al., 2020b). We use the public code provided by (Marfoq et al., 2021) for the CFL algorithm. The hyper-parameters $\text{tol}_1$ and $\text{tol}_2$ are tuned, and we report how the hyper-parameters affect the results of the algorithm in Table 1.
- **ICFL** (Yan et al., 2023). Follow the same setting as the original paper, we set the hyper-parameter $\alpha * (0)$ to $\{0.85, 0.98\}$, and $\epsilon_1 = 4.0$.
- **stoCFL** (Zeng et al., 2023). We choose $\tau = \{0, 0.05, 0.1, 0.15\}$ to control the trade-off between personalization and generalization as suggested by the original paper. In addition, we choose $\lambda = 0.5$, which always achieve the best performance as reported in the original paper.
- **HCFL$^+$ (FedRC)** (Guo et al., 2023b). We set $\tilde{\mu} = 0.4$, and choose $\rho = \{0.05, 0.1, 0.3\}$. The distance between clients are calculated by Equation (9).
- **HCFL$^+$ (FedEM)** (Marfoq et al., 2021). We set $\tilde{\mu} = 0.4$, and choose $\rho = \{0.05, 0.1, 0.3\}$. The distance between clients are calculated by Equation (9).
- **HCFL$^+$ (FeSEM)** (Long et al., 2023). We choose $\rho = \{0.05, 0.1, 0.3\}$. Follow the original paper, we use hard clustering paradigms that does not require the hyper-parameter $\tilde{\mu}$. The model splitting process is the same as (Sattler et al., 2020b) that designed for hard clustering paradigms. The distance between clients are calculated by Equation (9).

Table 4: **Performance of algorithms on noisy data scenarios.** We evaluated the performance of algorithms using the CIFAR10 dataset split into 100 clients. For each algorithm, we report the best test accuracy for all 200 communication rounds.

| Algorithm | CIFAR10 (MobileNetV2) | | | |
|---|---|---|---|---|
| | *Pairflip*, $\chi = 0.1$ | *Pairflip*,$\chi = 0.2$ | *Symflip*, $\chi = 0.2$ | *Symflip*, $\chi = 0.4$ |
| FedAvg | 54.75 ±1.45 | 52.35 ±1.65 | 52.60 ±0.50 | 41.80 ±0.50 |
| FeSEM | 32.60 ±1.30 | 35.25 ±2.95 | 32.40 ±2.80 | 29.70 ±0.01 |
| IFCA | 24.95 ±7.05 | 20.55 ±4.65 | 30.35 ±2.05 | 36.05 ±4.45 |
| FedEM | 64.40 ±1.10 | 57.55 ±2.95 | 53.00 ±1.90 | 43.10 ±0.20 |
| FedRC | 67.90 ±1.00 | 59.95 ±1.05 | 55.25 ±2.05 | 42.00 ±0.40 |
| HCFL$^+$ | **66.70** ±0.40 | **62.70** ±0.30 | **59.95** ±1.15 | **47.20** ±0.20 |




(a) HCFL$^+$ (FeSEM) (b) HCFL$^+$ (FedEM)

Figure 4: **Number of clusters in HCFL$^+$ over communication rounds.** We illustrate changes in cluster numbers across communication rounds for various $\rho$ values using the CIFAR-10 dataset in our experiments.

### E.3 Additional Experiment Results

**Results on noisy data scenarios** In Table 4, we show the performance of clustered FL emthods on noisy data scenarios. Results show that HCFL$^+$ consistently outperform other methods by a large margin.

**Additional results on diverse distribution shift scenarios.** In Table 5, we show the performance of algorithms with $\beta = 0.4$. Results show HCFL$^+$ always achieve the best test accuracy, and achieve a good local-global balance.

**Ablation studies on only using feature extractor-classifier split mechanism.** In Table 10, we present a new ablation study HCFL$'$, only using the feature-extractor classifier split method on top of the naive HCFL framework (Using backbone algorithms for Tiers 1 & 2, and CFL for Tiers 3 & 4). The results indicate HCFL$'$ outperforming the original algorithms, notably on CIFAR100 [7].

**Number of clusters in HCFL$^+$ over communication rounds.** Figure 4 shows how the number of clusters evolves across communication rounds for different $\rho$ values using the CIFAR-10 dataset in our experiments. It is evident that HCFL$^+$ allows the algorithm to automatically determine the optimal number of clusters during training.

**Efficiency comparison.** In Table 11, we compare the efficiency of HCFL$^+$ with baseline methods. Notably, HCFL$^+$ (FedEM) demonstrates superior performance while using a significantly reduced number of clusters, resulting in lower simulation time and memory usage.

[7]Evaluating the HCFL without the split is challenging due to the 200+ million trainable parameters, but this can be reduced to 10 million with the split.

Table 5: **Performance of the adaptive clustering methods.** We evaluated algorithm performance on CIFAR10 and CIFAR100 datasets, employing 100 clients. For each algorithm, we present the highest validation and test accuracies across 200 communication rounds, and the final number of clusters during training denoted as $K^T$. All experiments utilized MobileNet-V2 (Sandler et al., 2018).

| Algorithm | CIFAR10, $\beta = 0.4$ | | | CIFAR100, $\beta = 0.4$ | | |
|---|---|---|---|---|---|---|
| | Val | Test | $K^T$ | Val | Test | $K^T$ |
| FedAvg | 48.16 ±1.64 | 49.93 ±0.80 | 3.0 | 22.77 ±0.01 | 24.62 ±0.55 | 3.0 |
| FeSEM | 46.08 ±4.54 | 35.99 ±4.59 | 3.0 | 23.56 ±1.52 | 22.31 ±1.08 | 3.0 |
| IFCA | 36.15 ±3.45 | 24.79 ±1.18 | 3.0 | 27.72 ±0.82 | 21.37 ±1.33 | 3.0 |
| FedEM | 60.26 ±1.10 | 54.44 ±0.04 | 3.0 | 25.80 ±0.20 | 22.88 ±0.19 | 3.0 |
| FedRC | 57.99 ±0.29 | 56.75 ±0.38 | 3.0 | 30.94 ±0.88 | 31.63 ±0.20 | 3.0 |
| CFL | | | | | | |
| $\text{tol}_1 = 0.4, \text{tol}_2 = 1.6$ | 61.86 ±5.29 | 51.15 ±0.82 | 6.0 | 34.11 ±6.35 | 21.04 ±2.21 | 5.0 |
| $\text{tol}_1 = 0.4, \text{tol}_2 = 0.8$ | 60.42 ±0.31 | 41.59 ±2.14 | 8.0 | 36.23 ±3.58 | 16.03 ±2.69 | 6.0 |
| $\text{tol}_1 = 0.2, \text{tol}_2 = 0.8$ | 49.14 ±6.11 | 49.88 ±4.21 | 3.0 | 34.20 ±7.13 | 26.42 ±0.73 | 2.5 |
| ICFL | | | | | | |
| $\alpha^*(0) = 0.85$ | 77.73 ±0.47 | 52.03 ±0.10 | 100.0 | 49.71 ±0.55 | 28.55 ±0.03 | 100.0 |
| $\alpha^*(0) = 0.98$ | 63.69 ±3.58 | 54.02 ±1.11 | 81.5 | 45.72 ±1.10 | 28.45 ±0.82 | 70.0 |
| StoCFL | | | | | | |
| $\tau = 0.00$ | 48.55 ±0.95 | 51.25 ±1.16 | 1.5 | 24.50 ±0.03 | 25.70 ±1.51 | 1.0 |
| $\tau = 0.05$ | 57.84 ±2.26 | 50.42 ±0.97 | 20.5 | 26.24 ±1.46 | 26.60 ±1.17 | 4.0 |
| $\tau = 0.10$ | 72.91 ±2.25 | 47.84 ±2.60 | 59.0 | 67.67 ±1.68 | 9.89 ±0.45 | 86.0 |
| $\tau = 0.15$ | 77.19 ±2.31 | 41.49 ±0.97 | 92.0 | 70.13 ±0.27 | 7.77 ±0.23 | 94.0 |
| HCFL$^+$ (FeSEM) | | | | | | |
| $\rho = 0.1$ | 85.30 ±1.05 | 45.20 ±0.28 | 47.0 | 58.61 ±4.14 | 18.29 ±2.38 | 35.5 |
| $\rho = 0.3$ | 80.34 ±1.33 | 48.25 ±2.72 | 20.5 | 44.65 ±0.35 | 21.73 ±1.27 | 12.0 |
| HCFL$^+$ (FedEM) | | | | | | |
| $\rho = 0.05$ | 80.31 ±1.60 | 53.62 ±4.36 | 18.5 | 62.19 ±1.54 | 21.15 ±0.88 | 44.5 |
| $\rho = 0.1$ | 82.89 ±0.92 | 56.27 ±1.08 | 26.5 | 59.08 ±0.06 | 21.29 ±0.87 | 31.5 |
| $\rho = 0.3$ | 80.72 ±1.90 | 55.77 ±1.93 | 10.0 | 49.84 ±6.85 | 28.62 ±0.78 | 11.0 |
| HCFL$^+$ (FedRC) | | | | | | |
| $\rho = 0.05$ | 68.48 ±0.25 | 66.77 ±0.28 | 9.5 | 38.75 ±0.98 | 30.45 ±0.07 | 10.0 |
| $\rho = 0.1$ | 68.56 ±3.56 | 65.75 ±5.40 | 6.0 | 40.30 ±1.19 | 30.23 ±0.85 | 11.0 |
| $\rho = 0.3$ | 70.86 ±0.31 | 70.13 ±0.42 | 5.5 | 39.62 ±0.34 | 32.22 ±0.20 | 5.0 |

Table 6: **Ablation studies on techniques in Sec 4.2.** We evaluated algorithm performance on CIFAR10 and CIFAR100 datasets, showcasing the top Validation and Test accuracies for each. We kept $\rho = 0.3$ consistent across all algorithms and varied $\tilde{\mu}$ to adjust the penalty term's strength in the objective function. The best results in each block are highlighted.

| Algorithm | CIFAR10, $\beta = 0.2$ | | CIFAR10, $\beta = 0.4$ | | CIFAR100, $\beta = 0.2$ | | CIFAR100, $\beta = 0.4$ | |
|---|---|---|---|---|---|---|---|---|
| | Val | Test | Val | Test | Val | Test | Val | Test |
| HCFL$^+$ (FedEM) | | | | | | | | |
| $\tilde{\mu} = 0.0$ | **83.67** ±0.72 | **62.43** ±0.71 | **80.72** ±1.90 | **55.77** ±1.93 | **50.72** ±2.97 | **32.13** ±0.18 | **49.84** ±6.85 | **28.62** ±0.78 |
| $\tilde{\mu} = 0.1$ | 81.60 ±0.59 | 60.48 ±0.50 | 80.36 ±2.40 | 55.10 ±1.75 | 48.78 ±0.62 | 30.50 ±0.33 | 48.56 ±1.10 | 25.80 ±1.17 |
| $\tilde{\mu} = 0.4$ | 79.52 ±0.11 | 53.33 ±2.97 | 76.50 ±0.34 | 49.97 ±2.26 | 44.85 ±0.48 | 28.39 ±0.12 | 41.52 ±0.08 | 22.83 ±0.42 |
| HCFL$^+$ (FedRC) | | | | | | | | |
| $\tilde{\mu} = 0.0$ | **70.82** ±0.25 | 69.15 ±0.35 | 69.95 ±1.99 | 67.09 ±1.01 | 39.55 ±1.29 | 35.49 ±0.16 | 38.77 ±1.20 | 31.87 ±1.13 |
| $\tilde{\mu} = 0.1$ | 69.91 ±0.16 | 68.77 ±1.56 | 69.53 ±0.21 | 68.54 ±1.08 | 39.38 ±0.40 | 35.95 ±0.59 | **39.77** ±2.33 | 31.52 ±0.45 |
| $\tilde{\mu} = 0.4$ | 69.33 ±0.24 | **69.67** ±1.27 | **70.86** ±0.31 | **70.13** ±0.42 | **39.97** ±0.21 | **36.50** ±0.28 | 39.62 ±0.34 | **32.22** ±0.20 |

**Performance of algorithms under various heterogeneous settings.** In Table 12, we show that HCFL$^+$ consistently outperforms the baseline algorithms across all settings, demonstrating its reliable performance improvement.

**Clustering quality illustration.** In Table 13, we present the clustering results to evaluate the clustering quality. It is evident that HCFL$^+$ successfully identifies a relatively optimal clustering by: (1) assigning all samples with the same label to the same cluster; (2) avoiding the creation of an excessive number of clusters, as observed in stoCFL and ICFL; and (3) achieving higher validation and test accuracy compared to the baseline methods.

Table 7: **Ablation studies on techniques in Sec 4.3.** We evaluated algorithm performance on CIFAR10 and CIFAR100 datasets, displaying their highest Validation and Test accuracies. We kept $\rho$ consistent at 0.3 for all algorithms. "w/ SCWU" denotes the use of soft clustering weight updating mechanisms designed in Section 4.3.

| Algorithm | CIFAR10, $\beta = 0.2$ | | CIFAR10, $\beta = 0.4$ | | CIFAR100, $\beta = 0.2$ | | CIFAR100, $\beta = 0.4$ | |
|---|---|---|---|---|---|---|---|---|
| | Val | Test | Val | Test | Val | Test | Val | Test |
| HCFL$^+$ (FedEM) | | | | | | | | |
| w/ SCWU | **83.67** $\pm 0.72$ | 62.43 $\pm 0.71$ | **80.72** $\pm 1.90$ | 55.77 $\pm 1.93$ | **50.72** $\pm 2.97$ | 32.13 $\pm 0.18$ | **49.84** $\pm 6.85$ | 28.62 $\pm 0.78$ |
| w/o SCWU | 82.11 $\pm 2.39$ | 63.84 $\pm 0.19$ | 80.08 $\pm 0.99$ | 58.83 $\pm 2.12$ | 49.77 $\pm 1.93$ | 32.90 $\pm 1.11$ | 47.91 $\pm 2.67$ | 27.40 $\pm 1.17$ |
| HCFL$^+$ (FedRC) | | | | | | | | |
| w/ SCWU | 69.33 $\pm 0.24$ | **69.67** $\pm 1.27$ | 70.86 $\pm 0.31$ | **70.13** $\pm 0.42$ | 39.97 $\pm 0.21$ | **36.50** $\pm 0.28$ | 39.62 $\pm 0.34$ | **32.22** $\pm 0.20$ |
| w/o SCWU | 69.88 $\pm 0.30$ | 68.83 $\pm 0.71$ | 70.77 $\pm 0.47$ | 68.87 $\pm 0.23$ | 40.96 $\pm 1.24$ | 35.72 $\pm 1.01$ | 39.18 $\pm 0.13$ | 32.08 $\pm 0.78$ |

Table 8: **Performance of algorithms with Resnet18.** We evaluated algorithm performance on CIFAR10 datasets with $\beta = 0.2$, displaying their highest Validation and Test accuracies. All algorithms utilize ResNet18 and run for 200 communication rounds.

| Algorithm | Val | Test |
|---|---|---|
| CFL | | |
| $\text{tol}_1 = 0.4, \text{tol}_2 = 0.6$ | 63.07 $\pm 7.42$ | 53.65 $\pm 2.33$ |
| $\text{tol}_1 = 0.4, \text{tol}_2 = 0.8$ | 61.14 $\pm 1.87$ | 54.87 $\pm 1.32$ |
| ICFL | | |
| $\alpha^*(0) = 0.85$ | 80.46 $\pm 0.99$ | 45.28 $\pm 6.56$ |
| $\alpha^*(0) = 0.98$ | 82.34 $\pm 0.28$ | 44.08 $\pm 0.40$ |
| StoCFL | | |
| $\tau = 0.1$ | 57.41 $\pm 6.69$ | 48.95 $\pm 1.95$ |
| $\tau = 0.15$ | 66.54 $\pm 1.05$ | 47.77 $\pm 0.14$ |
| HCFL$^+$ (FeSEM) | | |
| $\rho = 0.05$ | 86.90 $\pm 0.20$ | 50.34 $\pm 5.99$ |
| $\rho = 0.1$ | 85.55 $\pm 0.24$ | 49.38 $\pm 6.15$ |
| HCFL$^+$ (FedEM) | | |
| $\rho = 0.05$ | 83.88 $\pm 0.25$ | 58.92 $\pm 1.11$ |
| $\rho = 0.1$ | 83.83 $\pm 0.01$ | 60.27 $\pm 3.11$ |
| HCFL$^+$ (FedRC) | | |
| $\rho = 0.05$ | 67.72 $\pm 1.30$ | 64.13 $\pm 0.37$ |
| $\rho = 0.1$ | 67.51 $\pm 0.24$ | 63.15 $\pm 0.78$ |

Table 9: **More details** about the ICFL results.

| ICFL | CIFAR100 (Val) | CIFAR100 (Test) |
|---|---|---|
| Round on Best Val | 52.73 | 6.40 |
| Round on Best Test | 29.22 | 32.77 |

Table 10: **Ablation studied on only using Feature extractor-classifier split mechnism.** HCFL$'$ uses the feature-extractor classifier split mechanism and the optimization steps outlined in Eq (4)- (7).

| Algorithms | CIFAR10 (Val) | CIFAR10 (Test) | CIFAR100 (Val) | CIFAR100 (Test) |
|---|---|---|---|---|
| FedEM | 66.49 | 53.64 | 29.75 | 24.18 |
| FedRC | 63.65 | 59.41 | 34.56 | 37.62 |
| FedSoft | 83.62 | 34.70 | 73.58 | 4.60 |
| HCFL$'$ (FedEM) | 61.24 | 59.83 | 42.44 | 28.13 |
| HCFL$'$ (FedRC) | 63.58 | 63.57 | 37.46 | 37.27 |
| HCFL$'$ (FedSoft) | 85.76 | 52.20 | 76.70 | 13.47 |
| HCFL$^+$ (FedEM) | 83.67 | 62.43 | 50.72 | 32.13 |
| HCFL$^+$ (FedRC) | 69.33 | 69.67 | 39.97 | 36.50 |
| HCFL$^+$ (FedSoft) | 85.92 | 51.23 | 76.28 | 12.50 |

Table 11: **Efficiency comparison.** The CIFAR-10 dataset is divided among 100 clients, with each client holding data from 2 classes. The hyperparameters are chosen based on the best-performing configuration from Table 1.

| Algorithms | Memory (M) | Simulation Time (s/it) | Final Cluster Number | Val | Test |
|---|---|---|---|---|---|
| ICFL ($\alpha(0) = 0.85$) | 4190 | 124.71 | 100 | 83.96 | 46.00 |
| StoCFL ($\tau = 0.15$) | 2834 | 178.53 | 41 | 86.92 | 20.00 |
| HCFL$^+$ (FedEM, $\rho = 0.1$) | 2564 | 176.48 | 5 | 95.14 | 49.00 |

Table 12: **Performance Under Various Heterogeneous Settings.** We evaluated the algorithm's performance on the CIFAR-10 dataset, presenting the top Validation and Test accuracies for each setting. Results are reported with $\beta = 0.8$. Additionally, $C = 2$ and $C = 4$ indicate that each client contains data from 2 and 4 classes, respectively.

| Algorithm | $\beta = 0.9$ | | $C = 2$ | | $C = 4$ | |
|---|---|---|---|---|---|---|
| | Val | Test | Val | Test | Val | Test |
| ICFL | 72.94 | 24.98 | 83.96 | 46.00 | 72.34 | 72.30 |
| stoCFL | 60.0 | 11.97 | 86.92 | 20.00 | 70.38 | 32.60 |
| HCFL$^+$ (FedEM) | 73.68 | 28.63 | 95.14 | 49.00 | 89.80 | 62.30 |

Table 13: **Clustering quality illustration.** We present the clustering results of HCFL$^+$ in the $C = 2$ setting, where each client is assigned data from two classes. Specifically, HCFL$^+$ generates 5 clusters in this setting, and we report the number of samples from each class assigned to clusters 1 through 5.

| HCFL+(FedEM, $\rho = 0.1$) | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|---|
| Class 1 | 4483 | 0 | 0 | 0 | 0 |
| Class 2 | 0 | 4502 | 0 | 0 | 0 |
| Class 3 | 0 | 0 | 4464 | 0 | 0 |
| Class 4 | 0 | 0 | 4536 | 0 | 0 |
| Class 5 | 0 | 4498 | 0 | 0 | 0 |
| Class 6 | 0 | 0 | 0 | 0 | 4483 |
| Class 7 | 0 | 0 | 0 | 4491 | 0 |
| Class 8 | 4517 | 0 | 0 | 0 | 0 |
| Class 9 | 0 | 0 | 0 | 4509 | 0 |
| Class 10 | 0 | 0 | 0 | 0 | 4517 |