CAUSALLY MOTIVATED COLLABORATIVE LEARNING ACROSS HETEROGENEOUS DATA DISTRIBUTIONS

XUEYANG TANG

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Computing

Causally Motivated Collaborative Learning Across Heterogeneous Data Distributions

Xueyang Tang

A thesis submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
September 2024

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: ____________________

Name of Student: Xueyang Tang

Abstract

As data privacy is attached more and more importance to in the field of machine learning, federated learning (FL) gains increasing attention and dramatic development in recent years. Federated learning allows the participation of a massive number of data holders (i.e., clients) that possess limited data to collaboratively train learning models in a privacy-preserving manner with raw data preserved locally. Traditional FL approaches develop a shared global model by the periodical model aggregation to fit all the local datasets, which can work well when the local data instances among different clients are independent and identically distributed (IID). The performance of the produced model can be significantly degraded if the data distributions across participants are heterogeneous (i.e., if a data distribution shift exists among local clients). On the one hand, the distribution shift across local training datasets can result in negative knowledge transfer between distant clients. On the other hand, the presence of distribution shift between training and test datasets can render the trained model incapable of generalizing effectively to unseen test data on each client. These challenges greatly impede the applicability of federated learning in practical scenarios. To address the challenge of data distribution shift in heterogeneous FL, we propose innovative frameworks for personalized federated learning in this thesis.

First, the prevalent personalized federated learning (PFL) can handle the distribution shift across local training datasets through building a personalized model for each client with the guidance of a shared global model. However, the sole global model

may easily transfer deviated context knowledge to some local models when multiple latent contexts exist across local datasets. We propose a concept called contextualized generalization (CG) to provide each client with fine-grained context knowledge that can better fit the local data distributions and facilitate faster model convergence. Theoretical analysis on convergence rate and generalization error shows our method CGPFL grants a $\mathcal{O}(\sqrt{K})$ speedup over most existing methods and achieves a better personalization-generalization trade-off than existing solutions. Moreover, our theoretical analysis further inspires a heuristic algorithm to find a near-optimal trade-off in CGPFL.

Second, modern machine learning model prefers to rely on shortcut which can perform well at training stage but fail to generalize to the unseen test data that presents distribution shift with regard to training data. The limited data diversity on federated clients can exacerbate this issue, making mitigating shortcut and meanwhile preserving personalization knowledge rather difficult. We formulate the structural causal models (SCMs) for heterogeneous federated clients, and derive two significant causal signatures which inspire a provable **shortcut discovery and removal** method. The proposed FedSDR is divided into two steps: 1) utilizing the available training data distributed among local clients to discover all the shortcut features in a collaborative manner. 2) developing the optimal personalized causally invariant predictor for each client by eliminating the discovered shortcut features. We provide theoretical analysis to prove that our method can draw complete shortcut features and produce the optimal personalized invariant predictor that can generalize to unseen test data on each client.

Third, while the preceding research makes a primary endeavor to address the challenge of train-test distribution shift, it exhibits two notable limitations: 1) FedSDR can offer theoretical guarantees solely within linear feature spaces; 2) the server necessitates access to local environmental knowledge in FedSDR. To mitigate these two limitations, we propose a crucial causal signature which can distinguish personalized

features from spurious features with global invariant features as the anchor. The novel causal signature is quantified as an information-theoretic constraint that facilitates the shortcut-averse personalized invariant learning on each client. Theoretical analysis demonstrates the novel method, FedPIN, can yield a tighter bound on generalization error than the prevalent PFL approaches when train-test distribution shift exists on clients. Moreover, we provide a theoretical guarantee on the convergence rate of the proposed FedPIN.

In summary, we address the data distribution shift in heterogeneous federated learning by proposing three novel PFL methods. The experimental results on diverse settings demonstrate the effectiveness of the proposed methods compared to the existing PFL approaches. Given that data distribution shift is prevalent in practical federated learning scenarios, our methods can not only contribute to the academic community of federated learning but also facilitate the deployment of federated learning in real-world applications.

Publications Arising from the Thesis

* indicates equal contribution (co-first authors).

7. Xueyang Tang, Song Guo, Xiaosong Ma, Haoxi Li, Jie Zhang and Yue Yu. "Causally Motivated Logic Alignment for Prompt Tuning in Vision-Language Models", submitted and under review.

6. Haoxi Li*, Xueyang Tang*, Jie Zhang, Song Guo, Sikai Bai, Peiran Dong and Yue Yu. "Causally Motivated Sycophancy Mitigation for Large Language Models", in *The Thirteenth International Conference on Learning Representations (ICLR)*, April 24-28, 2025, Singapore.

5. Xueyang Tang, Song Guo, Jingcai Guo, Jie ZHANG and Yue Yu, "Causally Motivated Personalized Federated Invariant Learning with Shortcut-Averse Information-Theoretic Regularization", in *Forty-first International Conference on Machine Learning (ICML)*, July 21-27 2024, Vienna, Austria.

4. Jie Zhang, Xiaosong Ma, Song Guo, Peng Li, Wenchao Xu, Xueyang Tang and Zicong Hong, "Amend to Alignment: Decoupled Prompt Tuning for Mitigating Spurious Correlation in Vision-Language Models", in *Forty-first International Conference on Machine Learning (ICML)*, July 21-27 2024, Vienna, Austria.

3. Xueyang Tang, Song Guo, Jie ZHANG and Jingcai Guo, "Learning Personalized Causally Invariant Representations for Heterogeneous Federated Clients", in *The Twelfth International Conference on Learning Representations (ICLR)*, May 7-11 2024, Vienna, Austria.

2. Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang and Wenchao Xu, "Promptfl: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model", *IEEE Transactions on Mobile Computing (TMC)*, vol. 23, no. 5, pp. 5179-5194, May 2024.

1. Xueyang Tang, Song Guo and Jingcai Guo, "Personalized Federated Learning with Contextualized Generalization", in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, (IJCAI-22)*, July 23-29, 2022, Vienna, Austria.

Acknowledgments

This thesis would not have been completed without the generous assistance and invaluable contributions of many individuals. It is a great pleasure to have the opportunity to express my sincere gratitude to all those people who provided support throughout my PhD journey.

First and foremost, I am deeply grateful to my supervisor, Prof. Song Guo, for his thoughtful support and encouraging guidance throughout the four-year journey towards earning my Ph.D. degree at the Hong Kong Polytechnic University. I regard Prof. Guo as a lifelong role model for his consistently exemplified excellence in academic ethic, professional expertise and research insight. Moreover, Prof. Guo is always ready to offer warm encouragement, unwavering support, and constructive suggestions whenever I encounter difficulties in both my personal life and academic studies. Through numerous discussions with Prof. Guo, I have gained valuable insights into literature review, idea development, paper writing, and presentation skills. These valuable lessons have shaped my research perspective and will benefit my career development in diverse aspects. I am fortunate and honored to be supervised by Prof. Guo and to be a member of PEILab.

Then, special thanks go to Dr. Jingcai Guo from the Hong Kong Polytechnic University for his encouraging supervision and constant support during my final year of PhD study. I would also like to extend my gratitude to my collaborators during my PhD studies for their full support and constructive discussions. Thanks to Prof. Yue

Yu from the Peng Cheng Laboratory, and Prof. Bin Xiao from the Hong Kong Polytechnic University for their generous assistance and inspiring discussions. I would also like to thank Dr. Jie Zhang, Mr. Xiaosong Ma, Ms. Tao Guo and Dr. Wenchao Xu from the Hong Kong Polytechnic University for their contributions.

Additionally, I would like to acknowledge my wonderful teammates and friends from the Hong Kong Polytechnic University and Hong Kong University of Science and Technology. Thanks to Kang Wei, Jing Li, Haozhao Wang, Qihua Zhou, Jiewei Zhang, Enyuan Zhou, Haoxi Li, Junxiao Wang, Xin Xie, Jinwen Liang, Yan Hong, Yue Zeng, Yi Liu, Bingjie Wang, Chuan He, Fushuo Huo, Jinyu Chen, Junbao Pan, Leijie Wu, Peiran Dong, Rui Zhang, Yingchun Wang, Yunfeng Fan, Ziming Liu, Feijie Wu, Sikai Bai, Zibo Jiang, Xiaocheng Lu who gave me kind assistance and warm encouragement during my PhD study period.

Last but not least, I would like to express my profound gratitude to my parents and my beloved girlfriend, who have provided unwavering understanding, encouragement and love throughout my entire study period. Without their constant support, I would not have been able to complete this long journey of my PhD.

To my family.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

As data privacy becomes increasingly emphasized, federated learning (FL) has emerged and achieved significant success in both academic and industrial fields. Federated learning, known as a collaborative learning paradigm, enables participants to jointly develop machine learning models while keeping their raw data preserved locally. Since the participants in a federated learning system come from diverse environments, the data they provide is usually subject to heterogeneous distributions [112, 38, 46]. In this thesis, we focus on addressing the critical challenge of data distribution shifts in heterogeneous federated learning to ensure both personalization and generalization performance for every participant in the system.

As a starting point, we will outline the primary research problem of this thesis in Section 1.1. The outstanding challenges in addressing the research problem is discussed in Section 1.2. Then, we will list the main contributions of this thesis in Section 1.3. Finally, in Section 1.4, we provide an overview of the thesis organization.

1.1 Overview

In the era of artificial intelligence (AI), data enriched with human knowledge serves as the cornerstone for the success of advanced machine learning models in both academic and industrial fields. Powerful AI models require large-scale training datasets to explore, understand, and utilize the knowledge about the world that is embedded in raw data. The diversity of data is highly associated with a model's generalization performance, which reflects its ability to apply the learned knowledge to unseen test contexts [72, 107, 105]. Therefore, collecting and exploiting diverse data plays a crucial role in the development of modern machine learning models.

In the real world, vast amounts of data are generated from a wide array of sources, including individuals, digital devices, and interconnected systems. This data originates from diverse activities such as social media interactions, sensor readings, online transactions, and communication networks, reflecting the complex and dynamic nature of human behavior and technological environments. As privacy protection attracts increasing attention, how to make use of the data distributed across different parties in a privacy-preserving manner has become a significant research topic [103, 46]. In this context, federated learning is proposed as a promising paradigm for achieving collaborative machine learning while preserving user privacy [70].

A typical federated learning system comprises a number of local clients, each with limited data samples and computational resources, and a central server that coordinates the collaboration among the participating clients [70]. The objective of federated learning is to leverage the data on local clients to obtain a shared model that can perform well on each client. During the training process, clents retain their raw data locally while the server can only have access to the model updates (i.e., parameters or gradients) uploaded by local clients. As illustrated in Figure 1.1, local clients update model parameters using their own data samples in parallel, after which the server aggregates the updates received from the participating clients to obtain the global model. Once global aggregation is completed, the server broadcasts the updated global model parameters to local clients, which serve as the initialization point for local optimization in the subsequent round. Therefore, federated learning can effectively prevent data privacy leakage compared to the centralized machine learning paradigm which requires access to users' raw data.

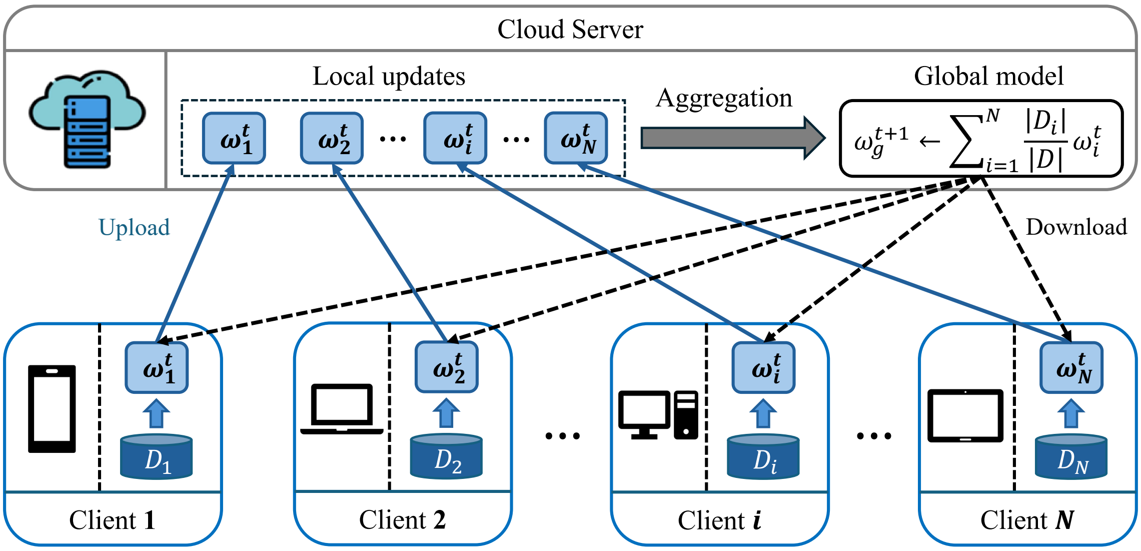

Cloud Server
Local updates
ω_1^t ω_2^t … ω_i^t … ω_N^t
Aggregation
Global model
$\omega_g^{t+1} \leftarrow \sum_{i=1}^{N} \frac{|D_i|}{|D|} \omega_i^t$
Upload
Download
ω_1^t D_1 Client 1
ω_2^t D_2 Client 2
ω_i^t D_i Client i
ω_N^t D_N Client N

Figure 1.1: Illustration of the typical federated learning system. There is one cloud server and N local clients, each of which has a local dataset D_i, $i \in [N]$. ω_i^t denotes local model parameters at communication round t and ω_g^{t+1} represents global model parameters at communication round $t+1$. The solid arrows indicate the upload link while the dotted arrow indicate the download link between the server and local clients.

Although the distributed learning scheme in federated learning enables the retention of raw data locally, it also introduces new challenges. In this thesis, we focus on one of the most significant challenge in federated learning, i.e., data distribution shift. Since local datasets are generated by different users, data instances are inherently non-independent and identically distributed (Non-IID) across local clients. In other words, there exists data distribution shift among local clients in federated learning. It has been demonstrated from both empirical and theoretical perspectives that data distribution shifts across clients can significantly degrade the performance of federated learning models [38, 48]. In short, when there is a distribution shift across local datasets, the global model struggles to fit all local data distributions well. Personalized federated learning (PFL) emerges as a promising approach to cope with data distribution shift by generating a personalized model for every participating client [4, 47, 95].

In fact, we believe that data distribution shifts across local clients can be a double-edged sword. On one hand, it creates a discrepancy between global and local optimal models, which can adversely affect the performance and convergence of the global model. On the other hand, data distribution shifts introduce data diversity that is closely associated with improved generalization capability of the trained models. Better generalization performance signifies that a model can achieve superior results on unseen test data distributions. Since model performance during the testing stage determines its suitability for deployment in real-world applications, generalization is a critical factor that must be considered in modern machine learning.

To tackle the data distribution shift across federated clients, we conduct an in-depth analysis and categorize the shift into two classes: train-train distribution shift and train-test distribution shift. From the inter-client perspective, the distribution shift across training datasets on local clients is defined as “train-train distribution shift”. From the intra-client perspective, the distribution shift between training and test datasets on a single client is referred to as “train-test distribution shift”. This thesis investigates how to address these two types of shifts by designing three novel federated learning frameworks (i.e., CGPFL, FedSDR and FedPIN) and makes significant contributions in both empirical and theoretical aspects. In summary, the overall framework of this thesis is illustrated in Figure 1.1.

Specifically, CGPFL is proposed to handle the train-train distribution shift across federated clients by optimizing the trade-off between personalization and generalization. The prevalent personalized federated learning (PFL) trains a personalized

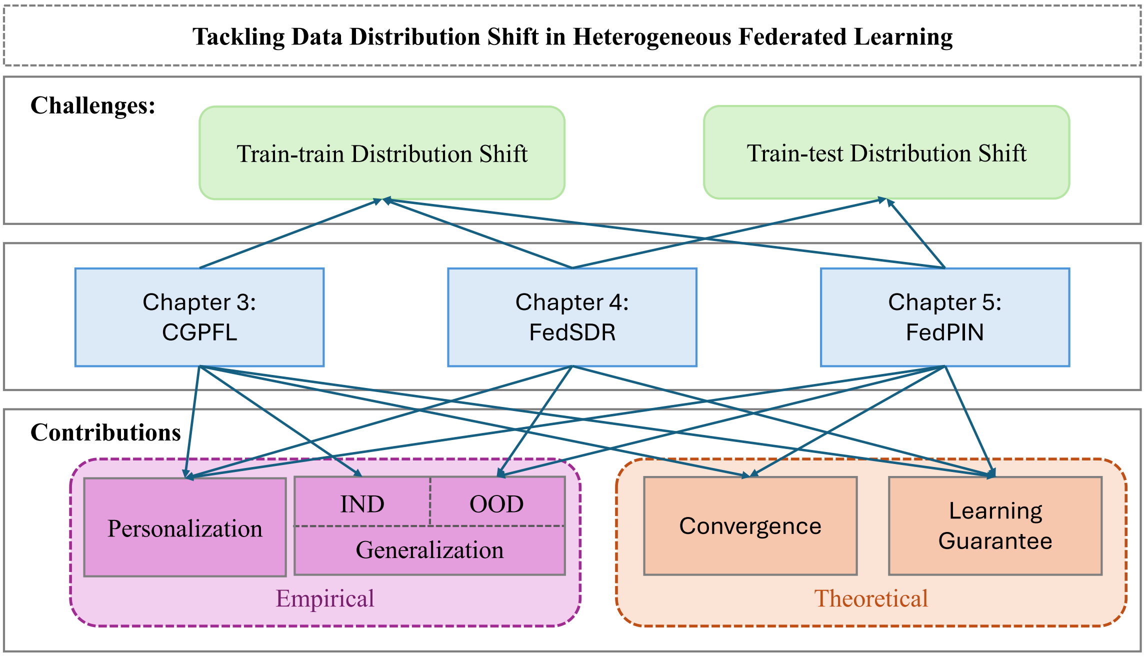


Figure 1.2: Illustration of the thesis framework. The goal of this thesis is to address the data distribution shift across federated clients. “Train-train Distribution shift” represents data distribution shift among training datasets on local clients, while “Train-test Distribution Shift” indicates data distribution shift between training and test datasets on each client. Besides, “IND” denotes “in-distribution” and “OOD” represents “out-of-distribution”.

model for each client with the guidance of some shared global knowledge (such as a global model) [33, 32, 95, 57]. The training process pushes personalized models to fit local data distributions as accurately as possible, while also leveraging shared contextual knowledge among clients to enhance generalization. In this way, personalized federated learning seeks to strike a balance between personalization and generalization to achieve superior model accuracy compared to traditional federated learning approaches. Inspired by the scheme adopted to address negative transfer in multi-task learning [111, 89], we propose "contextualized generalization (CG)" to provide fine-grained generalization knowledge and avoid negative knowledge transfer between latent contexts. In the CGPFL framework, local clients are dynamically clustered into multiple latent contexts, and the personalized model on each client is guided by a corresponding global model that incorporates contextualized generalization information relevant to that context. The training process of the global and personalized models in CGPFL is formulated as a bi-level optimization problem. By providing fine-grained generalization knowledge for each client, CGPFL facilitates improved accuracy and faster convergence of the obtained personalized models. Both empirical and theoretical results support the superiority of the proposed CGPFL framework.

When there is a distribution shift between training and test datasets, modern machine learning models can be prone to shortcut learning where spurious correlations between shortcut features and labels lead to poor generalization on unknown test datasets. Although shortcut learning is found pervasive in modern machine learning [26], it is rarely considered in prevalent personalized federated learning. To the best of our knowledge, we are the first to analyze the shortcut problem in personalized federated learning from a causal modeling perspective. We construct structured causal models (SCMs) [75] to simulate heterogeneous data generation across federated clients and propose two key causal signatures that underpin a provable method for discovering and removing shortcuts in personalized federated learning. At the first stage, the proposed FedSDR extracts shortcut features using available environmental information from local clients in a collaborative manner. At the subsequent stage, FedSDR learns personalized invariant representations for each client with a carefully crafted shortcut removal constraint. Results from extensive experiments demonstrate the effectiveness of the proposed FedSDR, compared with the state-of-the-art PFL competitors. In addition, the theoretical analysis proves that FedSDR can yield the optimal personalized invariant predictor for each client in linear cases.

Even though FedSDR successfully completes an initial step in addressing the train-test distribution shift in personalized federated learning, it also has two notable limitations: 1) the shortcut discovery method in FedSDR necessitates that the server has access to information about the training environments on each client, thereby increasing the risk of privacy leakage in federated learning; 2) theoretical analysis can only guarantee the effectiveness of FedSDR within a linear feature space. To extend personalized invariant learning to broader scenarios, we improve the structured causal models (SCMs) for heterogeneous federated clients and propose an environment-independent causal signature. By formulating the key causal signature as a shortcut-averse information-theoretic constraint, we develop a novel algorithm for achieving personalized invariant learning across federated clients. Theoretical analysis shows that the proposed FedPIN can output the optimal personalized invariant predictor for each client in non-linear cases, even without explicit environmental information from local clients. Moreover, FedPIN can achieve a tighter generalization error bound compared with the state-of-the-art personalized federated learning methods. Evaluation results on diverse datasets validate the superiority of FedPIN on out-of-distribution (OOD) generalization performance.

1.2 Outstanding Challenges

When addressing the train-train distribution shift across local clients, the shift can lead to negative knowledge transfer between clients, thereby degrading model performance. However, it can also improve the generalization capability of the developed models. Therefore, strategically leveraging the distribution shift to strike an optimal balance between personalization and generalization is the key point for achieve better performance. The proposed "contextualized generalization" balances personalization and generalization by clustering local clients into multiple contexts, within which positive knowledge transfer outweighs negative knowledge transfer among clients. However, in the absence of prior knowledge about data distributions on local clients, effectively identifying latent contexts poses a significant challenge.

The pervasive issue of shortcut learning can undermine out-of-distribution generalization performance when train-test distribution shift exists in personalized federated learning. Worse still, the limited data diversity on federated clients can exacerbate shortcut learning in personalized federated learning. In contrast to the centralized learning paradigm which does not require personalization, personalized federated learning (PFL) must disentangle invariant features from shortcut/spurious features due to their similar variability across heterogeneous clients. Moreover, centralized invariant learning typically requires explicit environmental labels, which is risky in federated learning since it can increase the potential for privacy leakage. When environmental information is not available, eliminating shortcut features while preserving personalized invariant features can be particularly challenging because of their close entanglement in personalized federated learning.

1.3 Thesis Contributions

To address the challenge of data distribution shift in heterogeneous FL, this thesis primarily makes the following contributions:

1. **Personalized Federated Learning with Contextualized Generalization.** we propose the concept of contextualized generalization (CG) to provide fine-

grained generalization and seek a better trade-off between personalization and generalization in PFL, and further formulate the training as a bi-level optimization problem that can be solved effectively by our designed CGPFL algorithm. Detailed theoretical analysis is conducted to provide the convergence guarantee and prove that CGPFL can obtain a $\mathcal{O}(\sqrt{K})$ times acceleration over the convergence rate of most existing algorithms for non-convex and smooth case. We further derive the generalization bound of CGPFL and demonstrate that the proposed contextualized generalization can constantly help reach a better trade-off between personaliztion and generalization in terms of generalization error against the state-of-the-arts. We provide a heuristic improvement of CGPFL, dubbed CGPFL-Heur, by minimizing the generalization bound in the theoretical analysis, to find a near-optimal trade-off between personalization and generalization. It can achieve a near-optimal accuracy with negligible additional computation in the server, while retaining the same convergence rate as that of CGPFL. Experimental results on multiple real-world datasets demonstrate that our proposed methods can achieve higher model accuracy than the state-of-the-art PFL methods in both convex and non-convex cases.

2. **Learning Personalized Causally Invariant Representations for Heterogeneous Federated Clients.**
   To the best of our knowledge, we are the first to consider the shortcut trap problem in personalized federated learning and analyse it by formulating the structural causal models for heterogeneous clients. Based on the proposed SCMs, we design a provable shortcut discovery and removal method to develop the optimal personalized invariant predictor which can generalize to unseen local test distribution for each client. The elaborated shortcut discovery and removal method can cooperate with most of the existing FL and PFL methods to improve the OOD generalization performance. Theoretically, we demonstrate that the designed shortcut discovery method can draw all the latent shortcut

components, then the shortcut removal method can eliminate the discovered shortcut features and produce the optimal personalized invariant predictor for each client. Empirically, we conduct experiments on several commonly used out-of-distribution datasets and the results validate the superiority of our method on out-of-distribution generalization performance, compared with the state-of-the-art competitors.

3. **Causally Motivated Personalized Federated Invariant Learning with Shortcut-Averse Information-Theoretic Regularization.** We improve the heterogeneous structured causal model to interpret Non-IID data distributions across federated clients, and propose a crucial causal signature which is quantified as a shortcut-averse information-theoretic constraint in the local objective to achieve personalized invariant learning on each client. Besides, an effective algorithm FedPIN is proposed to solve the devised optimization problem. Theoretically, we demonstrate that FedPIN can develop the optimal personalized invariant predictor for each client and provide a tighter generalization error bound compared with the state-of-the-art PFL methods. Moreover, we prove FedPIN can achieve a convergence rate on the same order as FedAvg [70]. Experimental results on diverse datasets validate the superiority of FedPIN on OOD generalization performance, in comparison with the state-of-the-art federated learning and personalized federated learning baselines.

1.4 Thesis Organization

This section will outline the organization of the remaining parts of this thesis. Before interpreting our works, we first provide the background and preliminary knowledge about federated learning, data distribution shift and generalization in Chapter 2. We will discuss the proposed CGPFL which handles train-train distribution shift across federated clients and facilitates personalized models toward higher accuracy and faster convergence with contextualized generalization in Chapter 3. Next, in Chapter 4, we interpret the algorithm FedSDR that is designed to learn personalized invariant representations for local clients in order to address the challenge of train-test distribution shifts. Then, in Chapter 5, we will elucidate the algorithm FedPIN which overcomes the existing limitations of FedSDR and promotes the application of personalized federated invariant learning to real-world scenarios. Finally, we conclude the thesis and explore potential directions for future research in Chapter 6.

Chapter 2

Background Review

In this chapter, we introduce the background and preliminary knowledge necessary for understanding federated learning, data distribution shift, personalization and generalization. Specifically, we explain the prevalent workflow of federated learning in Section 2.1. The data distribution shift problem is discussed in Section 2.2. At the conclusion of this chapter, foundational knowledge about generalization is presented in Section 2.4.

2.1 Federated Learning

As illustrated in Figure 1.1, a typical federated learning (FL) framework is composed of one global server and many local clients. Suppose there are N clients in the concerned FL system and each client has a local dataset D_i, where i denotes the index of federated clients and $i \in \{1, 2, ..., N\}$. For simplicity, we will use $[N]$ to represent $1, 2, \ldots, N$ throughout this thesis without additional description. The data instance in dataset D_i is described by $(X, y) \in D_i$, where X is the input and y denotes the corresponding label. The size of dataset D_i can be denoted by $|D_i|$. At communication round t, the model parameter on client i is represented by ω_i^t while the global model parameter is denoted by ω_g^t. When we input X, the prediction given by model ω can be expressed by $\hat{y} = f_\omega(X)$ where f_ω indicates the mapping function parameterized by ω. The expected empirical loss for model f_ω on dataset D is denoted as $\mathcal{R}(f_\omega; D) := \mathbb{E}_{(X,y)\in D}[\ell(f_\omega(X), y)]$ where ℓ is the loss function.

Taking the traditional federated learning algorithm (FedAvg [70]) as an example, the objective function of FL can be expressed as:

$$\min_{\omega} \sum_{i=1}^{N} \frac{|D_i|}{|D|} \mathcal{R}(f_\omega; D_i), \tag{2.1}$$

where $|D|$ describes the total number of data samples in the federated learning system. That is $|D| = \sum_{i=1}^{N} |D_i|$. The detailed algorithm designed to solve this objective can be divided into two components: server-side update and client-side update.

Sever-side Update If the communication round $t = 0$, server initializes the global model parameter as ω_g^0. Then, server broadcasts the model parameter ω_g^0 to all local clients for model initialization.

When the communication round $t \in \{1, 2, ..., T\}$ where T denotes the total number of communication round, server receives local updates $\{\omega_t^i \mid i \in [N]\}$ from the participating clients and then conducts global aggregation using the following expression:

$$\omega_g^t = \sum_{i=1}^{N} \frac{|D_i|}{|D|} \omega_i^t \tag{2.2}$$

Then, server randomly selects a client subset $\mathcal{A}_t$ and sends the global model parameter ω_g^t to them.

Client-side Update After receiving the global model parameter ω_g^t, $t \in \{0, 1, ..., T\}$ from the server, client i ($i \in [N]$) initializes the local model by $\omega_i^t = \omega_g^t$. Then, each client conducts local update for R epochs. At each local epoch, client i randomly samples a data batch B_i and updates the local model parameter by

$$\omega_i^t \leftarrow \omega_i^t - \eta \nabla \mathcal{R}(f_{\omega_i^t}; B_i), \tag{2.3}$$

where ∇ denotes the learning rate. When completing R local epochs, client i uploads the local update ω_i^t to the server. Notably, this local process runs in parallel in federated learning.

2.2 Distribution Shift

This section provides the preliminary insights on data distribution shift in federated learning.

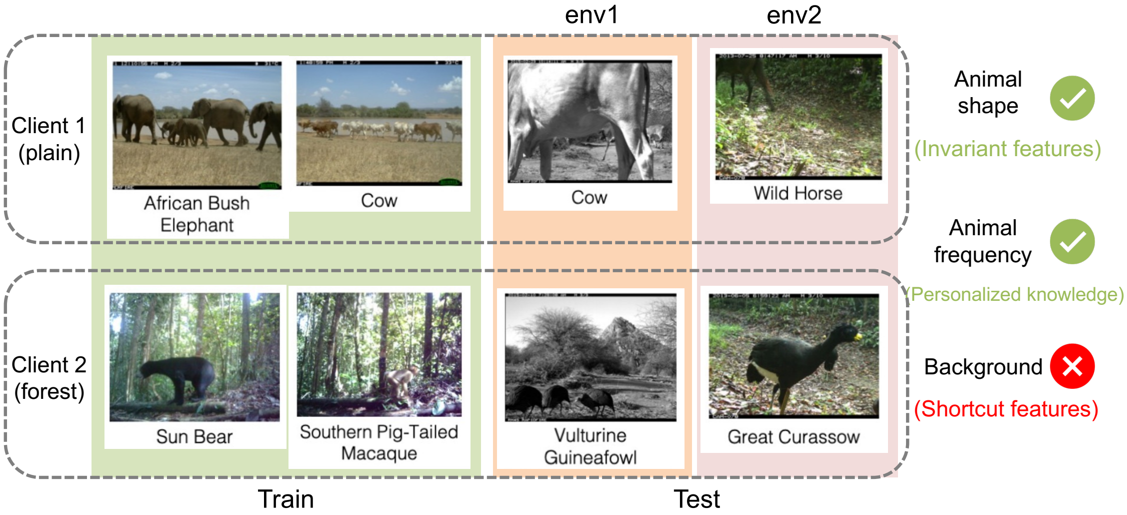


Figure 2.1: An illustration of "train-train distribution shift", "train-test distribution shift" and spurious features (also called shortcut features) using an example classification task. "env1" and "env2" indicate environment 1 and environment 2, respectively. In invariant learning, an environment corresponds to an latent data distribution.

In machine learning, data instances from a dataset can be regarded as subject to an underlying data distribution. A data distribution shift occurs when the probability density functions of two data distributions are distinct. In federated learning, each local client has their training and test datasets. From the inter-client perspective, there might be shifts in data distribution among local training datasets. From the intra-client viewpoint, there can also exists data distribution shift between the training and test datasets on each client. Therefore, we classify data distribution shifts in federated learning into two categories:

- **Train-train distribution shift** denotes the inter-client distribution shift, describing the distribution shift across local training datasets;
- **Train-test distribution shift** represents the intra-client distribution shift, describing the distribution shift between the training and test datasets on each client.

As shown in Figure 2.2, we provide an example to illustrate the data distribution shift in a practical federated learning system. The task is to classify images captured by cameras deployed in natural environments into their corresponding "animal class". Federated clients are located in different geographical regions and local data samples on each client are collected from multiple cameras distributed its region. Since wild animals appear with varying frequencies across different geographical regions, there is train-train distribution shift across federated clients. On the other hand, pictures captured by a camera can define an **environment**. The term 'environment' is influenced by various factors, including the natural surroundings, lighting conditions, and camera parameters. When the test data is generated by a different camera than the training data on a local client, a train-test distribution shift occurs on that client.

2.3 Shortcut and Invariant Learning

Modern machine learning models are prone to relying on spurious correlations (correlations between spurious features and the target/label, also known as shortcuts) in a variety of vision and language tasks [26]. We consider a binary classification task for illustration where a learning model needs to differentiate between pictures of "cow" and "camel" [7]. Because most cows stand with grass backgrounds and the majority of camels appear in desert backgrounds in the practical training dataset, there is a shortcut from background representation to target. The trained learning model prefers to choose background (spurious feature) rather than the shape of animals (intended feature) as the discriminative feature. When images with camels standing in grass backgrounds arrive at inference stage, they will be categorized as "cow" because the spurious correlation is no longer applicable.

Since shortcuts are unstable across diverse data distributions, models that perform well on training data can experience significant performance degradation on test data when a train-test distribution shift exists. Hence, mitigating shortcuts is of vital significance for equipping learning models with out-of-distribution generalization capabilities. Invariant learning emerges as a promising approach for learning intended features while eliminating shortcut features, and has attracted significant attention in centralized scenarios [6, 19, 40, 104, 99, 16, 77].

Invariant learning distinguishes between shortcut features (i.e., spurious features) and intended features (referred to as invariant features) from the perspective of causality. Because invariant features are the direct cause of the target while spurious correlations vary with respect to environment, they can be identified by applying an invariance constraint [6]:

$$\mathbb{P}(Y|\Phi(X) = z, e) = \mathbb{P}(Y|\Phi(X) = z, e'), \forall z \in \mathcal{Z}, \forall e, e' \in \mathcal{E}_{all}, \tag{2.4}$$

where e and e' indicate two distinct environments respectively. $\mathcal{E}_{all}$ denotes the set of all possible environments in the concerned task and $\mathcal{Z}$ denotes the feature space. $\mathbb{P}(\cdot \mid \cdot, e)$ represents the conditional probability distribution under environment e.

Based on the extracted invariant features, invariant learning can develop an invariant predictor by leveraging the consistent causal relationship between invariant features and the target across varied environments. Due to the stable conditional probability between invariant features and the target, the invariant predictor can achieve consistent performance across diverse environments, including those that are unknown at test time. In other words, invariant learning can effectively mitigate spurious correlations and facilitate out-of-distribution generalization, enabling models to generalize to test data distributions that exhibit a distribution shift relative to the training data.

2.4 Generalization

In general, generalization refers to the capability of a machine learning model to perform well on unseen test data instances. On one hand, the test data samples can differ from the training samples but still adhere to the same data distribution as the training samples. On the other hand, the test data samples may follow a distribution that exhibits a shift relative to the training data distribution. Therefore, we can categorize generalization into two classes:

- **In-distribution (IND) generalization** refers to the ability of learning models to generalize to unseen test data that adheres to the same distribution as the training data.
- **Out-of-distribution (OOD) generalization** describes the ability of learning models to generalize to unseen test data that exhibits a distribution shift compared to the training data.

2.5 Learning Guarantee

To demonstrate the theoretical advantages of the proposed algorithms, we provide learning guarantees for the algorithm CGPFL, FedSDR and FedPIN in Chapters 3, 4 and 5, respectively. In summary, we present two schemes for evaluating the learning guarantees of the proposed algorithms. The first ensures that the optimal solution of the proposed algorithm leads to optimal personalized invariant predictors, while the second measures the generalization error bound to provide a guarantee for the test performance of the learned models. Specifically, Chapter 3 gives the generalization error bound as learning guarantee for CGPFL. Chapter 4 proves that FedSDR can produce the optimal personalized invariant predictors. Both optimality analysis and generalization error bound are provided for FedPIN in Chapter 5. As foundational knowledge, this section presents the mathematical descriptions for optimality analysis and generalization error bound.

Optimality Analysis Before analyzing the optimal solutions of the proposed methods, we first provide a formal definition of optimal personalized invariant predictors in federated learning.

Definition 2.5.1 (**Optimal Personalized Invariant Predictor**). *The optimal personalized invariant predictor for client u is elicited based on the complete invariant features which are informative for the target in the task that client u concentrates on, i.e., $\Phi_u^\star \in \arg\max_{\Phi_u} I(Y; \Phi_u(X))$, where Φ_u satisfies that $\mathbb{P}(Y|\Phi_u(X) = z, e) = \mathbb{P}(Y|\Phi_u(X) = z, e'), \forall z \in \mathcal{Z}, \forall e, e' \in \mathcal{E}_{all}^u$.*

Based on this mathematical definition, we can examine whether the proposed algorithms are capable of developing an optimal personalized invariant predictor for each federated client.

Generalization Error Bound Suppose the expected loss for model f_ω on test data distribution $\mathcal{D}_\mathcal{T}$ is $\mathcal{R}(f_\omega; \mathcal{D}_\mathcal{T}) = \mathbb{E}_{(X,y)\in\mathcal{D}_\mathcal{T}}[\ell(f_\omega(X), y)]$ where ℓ is the loss function, and the empirical loss for model f_ω on training dataset D is denoted by $\mathcal{R}(f_\omega; D) = \mathbb{E}_{(X,y)\in D}[\ell(f_\omega(X), y)]$. The generalization error for model f_ω can be expressed as:

$$\mathcal{R}(f_\omega; \mathcal{D}_\mathcal{T}) - \mathcal{R}(f_\omega; D) \leq \text{Error Bound}. \tag{2.5}$$

According to the definitions, $\mathcal{R}(f_\omega; \mathcal{D}_\mathcal{T})$ measures the test performance of learning model f_ω while $\mathcal{R}(f_\omega; D)$ represents the training performance of model f_ω. Therefore, a small “Error Bound” in inequality 2.5 indicates that the learning model f_ω can guarantee consistent performance on training and test data. In other words, the learning model f_ω exhibits strong generalization performance.

Chapter 3

Personalized Federated Learning with Contextualized Generalization

3.1 Introduction

Recently, personalized federated learning (PFL) has emerged as an alternative to conventional federated learning (FL) to cope with the statistical heterogeneity of local datasets (a.k.a., Non-I.I.D. data). Different from conventional FL that focuses on training a shared global model to explore the global optima of the whole system, i.e., minimizing the averaged loss of clients, the PFL aims at developing a personalized model (distinct from the individually trained local model which usually fail to work due to the insufficient local data and the limited diversity of local dataset) for each client to properly cover diverse data distributions. To develop the personalized model, each user needs to incorporate some context information into the local data, since the insufficient local data cannot present the complete context which the personalized model will be applied to [47]. However, the context is generally latent and can be hardly featurized in practice, especially when the exchange of raw data is forbidden. In the existing PFLs, the latent context knowledge can be considered to be transferred to the local users via the global model update. During the PFL training, the personalization usually requires personalized models to fit local data distributions as well as possible, while the generalization needs to exploit the common context knowledge among clients by collaborative training. Thus, the PFL is indeed pursuing a trade-off between them to achieve better model accuracy than the traditional FL. More specifically, the server-side model is trained by aggregating local model updates from each client and hence can obtain the common context knowledge covering diverse data distributions. Such knowledge can then be offloaded to each client and contributes to the generalization of personalized models.

Despite the recent PFL approaches have reported better performance against conventional FL methods, they may still be constrained in personalization by using sole global model as the guidance during the training process. Concretely, our intuition is that: If there exists multiple latent contexts across local data distributions, then contextualized generalization can provide fine-grained context knowledge and further facilitate the personalized models toward better recognition accuracy and faster model convergence. We thus argue one potential bottleneck of current PFL methods is the loss of generalization diversity with only one global model. Worse still, the global model may also easily degrade the overall performance of PFL models due to negative knowledge transfers between the disjoint contexts.

In this work, we design a novel PFL training framework, dubbed CGPFL, by involving the proposed concept, i.e., *contextualized generalization (CG)*, to handle the challenge of the context-level heterogeneity. More specifically, we suppose the participating clients can be covered by several latent contexts based on their statistical characteristics and each latent context can be corresponded to a generalized model maintained in the server. The personalized models are dynamically associated with the most pertinent generalized model and guided by it with fine-grained contextualized generalization in an iterative manner. We formulate the process as a bi-level optimization problem considering both the global models with contextualized generalization maintained in the server and the personalized models trained locally in clients.

The main contributions of this work are summarized as follows:

- To the best of our knowledge, we are the first to propose the concept of *contextualized generalization (CG)* to provide fine-grained generalization and seek a better trade-off between personalization and generalization in PFL, and further formulate the training as a bi-level optimization problem that can be solved effectively by our designed CGPFL algorithm.

- We conduct detailed theoretical analysis to provide the convergence guarantee and prove that CGPFL can obtain a $\mathcal{O}(\sqrt{K})$ times acceleration over the convergence rate of most existing algorithms for non-convex and smooth case. We further derive the generalization bound of CGPFL and demonstrate that the proposed *contextualized generalization* can constantly help reach a better trade-off between personalization and generalization in terms of generalization error against the state-of-the-arts.

- We provide a heuristic improvement of CGPFL, dubbed CGPFL-Heur, by minimizing the generalization bound in the theoretical analysis, to find a near-optimal trade-off between personalization and generalization. CGPFL-Heur can achieve a near-optimal accuracy with negligible additional computation in the server, while retaining the same convergence rate as that of CGPFL.

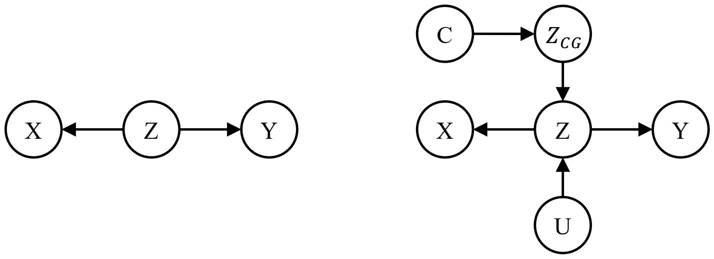
- Experimental results on multiple real-world datasets demonstrate that our proposed methods, i.e., CGPFL and CGPFL-Heur, can achieve higher model accuracy than the state-of-the-art PFL methods in both convex and non-convex cases.

3.2 Related Work

Clustered Federated learning. Considering that one shared global model can hardly fit the heterogeneous data distributions, some recent FL works [27, 82, 11, 69] try to cluster the participating clients into multiple groups and develop corresponding number of shared global models by aggregating the local updates. After the training process, the obtained global models are offloaded to the corresponding clients for inference. Since these methods only reduce the FL training into several sub-groups, of which each global model is still shared by their in-group clients, the personalization is scarce and the offloaded models can still hardly cover the heterogeneous data distributions across the in-group clients. Specifically, *IFCA* [27] requires each client to calculate the losses on all global models to estimate its cluster identity during each iteration, and result in significantly higher computation cost. *CFL* [82] demonstrates that the conventional FL even cannot converge in some Non-I.I.D. settings and provides intriguing perspective for clustered FL with bi-partitioning clustering. However, it can only work for some special Non-I.I.D. case described as *'same feature & different labels'* [39]. *FL+HC* [11] divides the clients clustering and the model training processes separately, and only conducts the clustering once at a manually defined step, while the training remains the same as conventional FL. Last, three effective PFL approaches are proposed in [69], of which the user clustering method is very similar to *IFCA* [27].

Personalized Federated Learning. Most recently, the PFL approaches have attracted increasing attention [47]. Among them, a branch of works [33, 32, 20] propose to mix the global model on the server with local models to acquire the personalized models. More concretely, Hanzely *et al.* [32, 33] formulate the mixture problem as a combined optimization of the local and global models, while *APFL* [20] straightforwardly mixes them with an adaptive weight. *KT-pFL* [108] exploits the knowledge distillation (KD) to transfer the generalization information to local models and allows the training of heterogeneous models in FL setting. Differently, *FedPer* [4] splits the personalized models into two separate parts, of which the base layers are shared by all the clients and trained on the server, and the personalization layers are trained to adapt to individual data and maintain the privacy properties on local devices. *MOCHA* [89] considers the model training on the clients as relevant tasks and formulate this problem as a distributed multi-task learning objective. Fallah *et al.* [23] make use of the model agnostic meta learning (*MAML*) to implement the PFL, of which the obtained meta-model contains the generalization information and can be utilized as a good initialization point of training.

3.3 Causal Insight



(a) machine learning (b) personalized FL

Figure 3.1: Graph (a) presents a typical structural causal model (SCM) adopted in machine learning, while (b) show the SCM utilized to analyse the train-train distribution shift in personalized federated learning. Z denotes the latent representation while Z_{CG} indicates the contextualized generalization knowledge which is indexed by the latent context variable C. Besides, U is the indicator of user/client.

In order to demonstrate the effect of the proposed “contextualized generalization” on the learning process of personalized models from the causal perspective, we adopt the structured causal models [75] to analyse the inter-client data distribution shift in personalized federated learning. As shown in Figure 3.1, we introduce a observable variable U to indicate the index of local clients and a latent variable C to represent the latent context. Accordingly, variable Z_{CG} denotes the contextualized generalization knowledge. With the SCM displayed in Figure 3.1(b), we can drive the following lemma to validate the effectiveness of the proposed "contextualized generalization".

Lemma 3.3.1. *If the data generating mechanism of the federated learning system complies with the causal graph in Figure 3.1(b), and $I(\cdot;\cdot \mid \cdot)$ denotes the conditional mutual information, then for any $i \in [N]$, the following inequality always holds:*

$$I(Y; X \mid U = i) < I(Y; X, C \mid U = i), \tag{3.1}$$

where i represents the index of client i.

Proof. For any client i, according to the definition of conditional mutual information, we can write that

$$\begin{aligned}
& I(Y; X, C \mid U = i) \\
&= \sum_x \sum_y \sum_c \mathbb{P}_{YXC}(y, x, c, i) \log \frac{\mathbb{P}(U = i)\mathbb{P}_{YXC}(y, x, c, i)}{\mathbb{P}_Y(y, i)\mathbb{P}_{XC}(x, c, i)} \\
&= \sum_x \sum_y \sum_c \mathbb{P}_{YXC}(y, x, c, i) \log \left[\frac{\mathbb{P}_X(x, i)\mathbb{P}_{YXC}(y, x, c, i)}{\mathbb{P}_{YX}(y, x, i)\mathbb{P}_{XC}(x, c, i)} \cdot \frac{\mathbb{P}(U = i)\mathbb{P}_{YX}(y, x, i)}{\mathbb{P}_Y(y, i)\mathbb{P}_X(x, i)}\right] \\
&= \sum_x \sum_y \sum_c \mathbb{P}_{YXC}(y, x, c, i) \log \frac{\mathbb{P}(U = i)\mathbb{P}_{YX}(y, x, i)}{\mathbb{P}_Y(y, i)\mathbb{P}_X(x, i)} \\
&\quad + \sum_x \sum_y \sum_c \mathbb{P}_{YXC}(y, x, c, i) \log \frac{\mathbb{P}_X(x, i)\mathbb{P}_{YXC}(y, x, c, i)}{\mathbb{P}_{YX}(y, x, i)\mathbb{P}_{XC}(x, c, i)} \\
&= \sum_x \sum_y \mathbb{P}_{YX}(y, x, i) \log \frac{\mathbb{P}(U = i)\mathbb{P}_{YX}(y, x, i)}{\mathbb{P}_Y(y, i)\mathbb{P}_X(x, i)} \\
&\quad + \sum_x \sum_y \sum_c \mathbb{P}_{YXC}(y, x, c, i) \log \frac{\mathbb{P}_X(x, i)\mathbb{P}_{YXC}(y, x, c, i)}{\mathbb{P}_{YX}(y, x, i)\mathbb{P}_{XC}(x, c, i)} \\
&= I(Y; X \mid U = i) + I(Y; C \mid X, U = i)
\end{aligned}$$

Using the d-separation criterion in [75], we can know that the variable set $[X, U]$ cannot block all paths between variable Y and C. That is, we have $Y \not\perp\!\!\!\perp C \mid [X, U]$. It is known that mutual information is non-negative, i.e., $I(Y; C \mid X, U = i) \geq 0$, and

$I(Y; C \mid X, U = i) = 0$ if and only if $Y \perp\!\!\!\perp C \mid [X, U = i]$ holds. Therefore, we can have that $I(Y; C \mid X, U = i) > 0$ always holds for any $i \in [N]$.

In other words, we can conclude that $I(Y; X \mid U = i) < I(Y; X, C \mid U = i), \forall i \in [N]$. Proof ends. □

Proposition 3.3.1 (Lemma 2 in [10]). *When we train a classifier conditioned on a feature extractor Φ with the data distribution $\mathcal{D}$, minimizing the cross-entropy loss $\mathcal{R}(\omega(\Phi); \mathcal{D})$ is equivalent to maximizing the mutual information $I(Y; \Phi(X))$ on $\mathcal{D}$.*

Remark 3.3.1. *Based on the conclusion given in Proposition 3.3.1, the inequality proved in Lemma 3.3.1 indicates that knowing the context information (i.e., C) can always provide additional information for the classification task on each client. This insight demonstrate the effectiveness of the proposed "contextualized generalization" knowledge on the development of personalized models on federated clients.*

Connection to Algorithm Design: Although theoretical results show that leveraging context information can improve the performance of personalized models, especially for classification tasks, obtaining explicit context information can be prohibitive in practical federated learning settings. True context information is jointly determined by complex and diverse factors, such as the geographical region where the client is located, the current environment, and user preferences. Consequently, such information is rarely directly available for use, and obtaining tailored annotations can incur prohibitive costs. Moreover, many factors included in context information (e.g., the client's geographical area and user preferences) are typically considered private data. Therefore, leveraging explicit context information can therefore increase the risk of data privacy leakage for federated clients. In summary, exploiting explicit context information is often impractical in real-world federated learning systems.

Considering the applicability to real-world federated learning systems where context information is latent and unobservable, we design our algorithm to address more practical and challenging scenarios in which the server has no access to explicit context information. To avoid increasing the risk of privacy leakage compared to traditional federated learning approaches, we aim to exploit latent contexts using only the uploaded model parameters or gradient updates from local clients. Specifically, the server dynamically classifies local clients into several latent contexts based on their uploaded model parameters, without requiring any explicit context information that could increase the risk of data privacy leakage. The detailed algorithm design will be presented and discussed in the following section.

3.4 Problem Formulation

We start by formalizing the FL task and then introduce our proposed method. Given N clients and the their Non-I.I.D. datasets $\widetilde{D}_1, ..., \widetilde{D}_i, ..., \widetilde{D}_N$ that subject to the underlying distributions as $D_1, ..., D_i, ..., D_N$ ($D_i \in \mathbb{R}^{d \times n_i}$ and $i \in [N]$). Every client i has m_i instances $z^{i,j} = (\mathbf{x^{i,j}}, y^{i,j})$, $j \in [m_i]$, where $\mathbf{x}$ is the data features and y denotes the label. Hence, the objective function of the conventional FL can be described as [57]:

$$\min_{\omega \in \mathbb{R}^d} \{G(\omega) := G\big(\mathcal{R}_1(\omega; \widetilde{D}_1), ..., \mathcal{R}_N(\omega; \widetilde{D}_N)\big)\}, \tag{3.2}$$

where ω is the global model and $\mathcal{R}_i : \mathbb{R}^d \to \mathbb{R}, i \in [N]$ denotes the expected loss function over the data distribution of client i: $\mathcal{R}_i(\omega; \widetilde{D}_i) = \mathbb{E}_{z^{i,j} \in \widetilde{D}_i}[\ell(\omega; z^{i,j})]$. The function $G(\cdot)$ denotes the aggregation method to obtain the global model ω. For example, *FedAvg* [70] applies $G(\omega) = \sum_{i=1}^{N} \frac{m_i}{m} \mathcal{R}_i(\omega)$ to do the aggregation, where m is the total number of instances on local devices.

To handle the challenge of rich statistical diversities in PFL, especially in the cases where the local datasets belong to several latent contexts, our CGPFL propose to maintain K context-level generalized models in the server to guide the training of personalized models on the clients. During training, the local training process based on its local dataset can *push* the personalized model to fit its local data distribution as well as possible. Meanwhile, the regularizer will dynamically *pull* the personalized model as close as possible to the most pertinent generalized model during the iterative algorithm, from which the fine-grained context knowledge can be transferred to each personalized model to better balance the generalization and personalization. Hence, the overall objective function of CGPFL can be described as a bi-level optimization problem as:

$$\min_{\Theta \in \mathbb{R}^{d\times N}} \frac{1}{N}\sum_{i=1}^{N}\Big\{F_i(\theta_i) := \mathcal{R}_i(\theta_i) + \lambda r(\theta_i, \omega_k^*)\Big\}, i \in C_k^*,$$
$$s.t. \quad \Omega^*, C_K^* = \mathop{\arg\min}_{\Omega\in\mathbb{R}^{d\times K}, C_K} G(\omega_1, ..., \omega_K; C_K),$$

where θ_i $(i \in [N])$ denotes the personalized model on client i and $\Theta = [\theta_1, ..., \theta_N]$. The context-level generalized models are denoted by $\Omega = [\omega_1, ..., \omega_K]$. λ is a hyper-parameter and C_k denotes the corresponding context that client i belongs to. Considering the latent contexts are represented in disjoint subspaces respectively, the function $G(\cdot)$ can be decomposed as $G(\omega_1, ..., \omega_K; C_K) = \frac{1}{K}\sum_{k=1}^{K} G_k(\omega_k; C_k)$.

In general, there exists two alternative strategies to generate the context-level generalized models. The intuitive one is to solve the inner-level objective $\min_{\Omega\in\mathbb{R}^{d\times K}} G(\omega_1, ..., \omega_K)$ based on local datasets, which is similar to *IFCA* [27]. However, the computation overhead is high in the local devices while their available computation resources are usually limited. Comparing the local objective that trains a generalized model ω_k based on local dataset, i.e., $\omega_i^* = \mathop{\arg\min}_{\omega} \mathcal{R}_i(\omega; \widetilde{D}_i)$, with that of the personalized model, i.e., $\theta_i^* = \mathop{\arg\min}_{\theta_i}\{\mathcal{R}_i(\theta_i; \widetilde{D}_i) + \lambda r(\theta_i, \omega_k^*)\}$, we notice that the locally obtained θ_i^* can be regarded as the distributed estimation of ω_k^*. In this way, the regularizer $r(\theta_i^*, \omega_k^*)$ can be used to evaluate the estimation error, and we can further derive the context-level generalized models by minimizing the average estimation error. In this work, we use $L2$-norm i.e., $r(\theta_i, \omega_k) = \frac{1}{2}\|\theta_i - \omega_k\|^2$ as the regularizer, which is also adopted in various prevalent PFL methods [33, 32, 95, 57] and has been empirically demonstrated to be superior over other regularizers, e.g., the symmetrized KL divergence in [57]. Hence, we formulate our overall objective as:

$$\min_{\Theta \in \mathbb{R}^{d \times N}} \frac{1}{N} \sum_{i=1}^{N} \Big\{ F_i(\theta_i) := \mathcal{R}_i(\theta_i) + \frac{\lambda}{2} \|\theta_i - \omega_k^*\|^2 \Big\}, i \in C_k^*,$$

$$s.t. \quad \Omega^*, C_K^* = \underset{\Omega \in \mathbb{R}^{d \times K}, C_K}{\arg\min} \sum_{k=1}^{K} q_k \sum_{j \in C_k} p_{k,j} \|\theta_j - \omega_k\|^2. \tag{3.3}$$

We adopt $p_{k,j} = \frac{1}{|C_k|}$ and $q_k = \frac{|C_k|}{N}$ in this work, where $C_k (k \in [K])$ denotes the latent context k, and $|C_k|$ is the number of clients that belong to the context k. Intriguingly, the inner-level objective is exactly the classic objective of k-means clustering [64]. We notice that when $K = 1$, the above objective is equivalent to the overall objective in [95], which means that the objective in [95] can be regarded as a *special case* $(K = 1)$ of ours.

3.5 Methodology

3.5.1 Overview

In this section, we introduce our proposed CGPFL in detail. As shown in Figure 3.2, the key idea is to dynamically relate the clients to K latent contexts based on their uploaded local model updates, and then develop a generalized model for each context by aggregating the updates from each user group. These generalized models are utilized to guide the training directions of personalized models and transfer contextualized generalization to them. Both the personalized models and the generalized models are trained in parallel, so we can denote the model parameters in matrix form. The generalized models can be written as $\Omega_K := [\omega_1, \dots, \omega_k, \dots, \omega_K] \in \mathbb{R}^{d \times K}$, and the corresponding local approximations are $\Omega_{I,R} := [\tilde{\omega}_{1,R}, \dots, \tilde{\omega}_{i,R}, \dots, \tilde{\omega}_{N,R}]$, where R is the number of local iterations and $\tilde{\omega}_{i,R}, \omega_k \in \mathbb{R}^d, \forall i \in [N], k \in [K]$. In this work, we use capital characters to represent matrices unless stated otherwise.

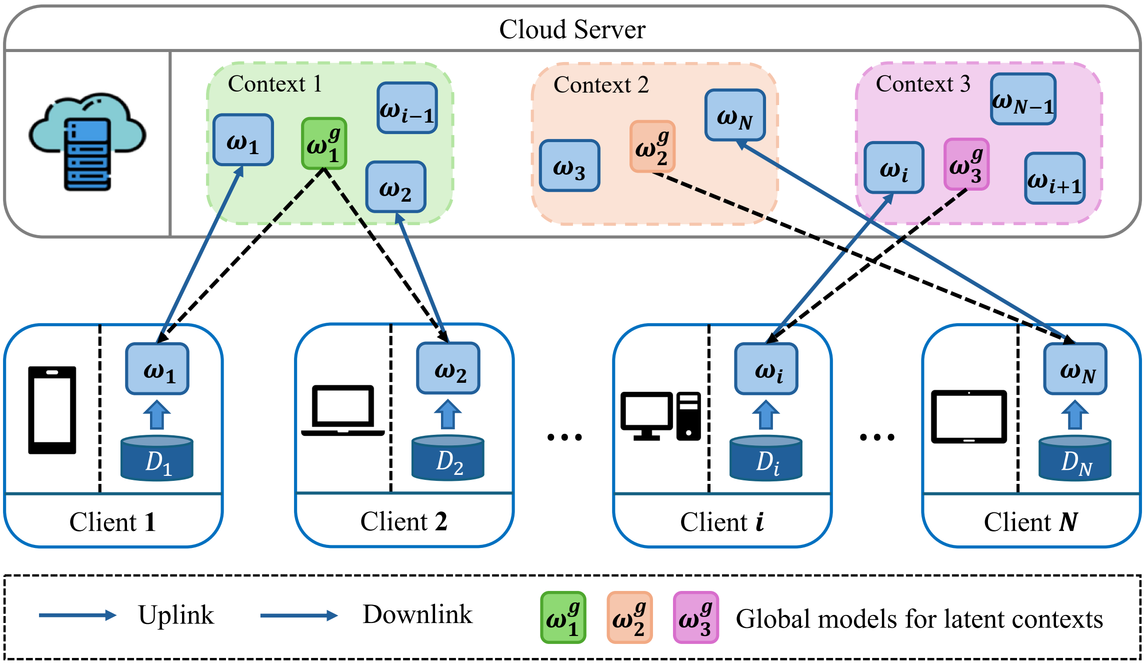


Figure 3.2: Illustration of the overall framework designed for the proposed method CGPFL. In this example, there are three latent contexts among the federated clients.

3.5.2 Algorithm Design

We design an effective alternating optimization framework to minimize the overall objective in (3.3). Specifically, the upper-level problem can be decomposed into N separate sub-problems with fixed generalized models and to be solved on local devices in parallel. Next, we can further settle the inner-level problem to derive the generalized models with fixed personalized models. Since the solution to the sub-problems of the upper-level objective has been well-explored in recent PFL methods [95, 57, 32], we hereby mainly focus on the inner-level problem. We alternately update the context-level generalized models Ω_K and the context indicator C_K to obtain the optimal generalized models. We view the personalized models, i.e., $\Theta_I = [\theta_i, ..., \theta_N]$, as private data, and distributionally update the context-level generalized models Ω_K on clients with fixed context indicator C_K. During each server round, the server conducts k-means clustering on uploaded local parameters $\Omega_{I,R}^t$ to cluster the clients into K latent contexts, and the clustering results C_K are re-arranged to the matrix form as $P^t \in \mathbb{R}^{N \times K}$. For example, if client $i, i \in [N]$ is clustered into the context $C_j, j \in [K]$ (where $C_j, j \in [K]$ are sets, the union $\bigcup_{j \in [K]} C_j$ and intersection $\bigcap_{j \in [K]} C_j$ are the set $[N]$ and empty set, respectively), the element $(P^t)_{i,j}$ is defined as $\frac{1}{|C_j|}$, or set 0 otherwise. In this way, the elements of every column in P^t amount to 1, i.e. $\sum_{i=1}^{N} (P^t)_{i,j} = 1$, $\forall j, t$.

Algorithm 1 CGPFL: Personalized Federated Learning with Contextualized Generalization

Input: Initialized models and hyper-parameters $\Theta_I^0, \Omega_K^0, P^0, T, R, S, K, \lambda, \eta, \alpha, \beta$.

Output: Personalized models Θ_I^T.

```
for $t = 0$ to $T - 1$ do
    Server sends $\Omega_K^t$ to clients according to $P^t$.
    for local device $i = 1$ to $N$ in parallel do
        Initialization: $\Omega_{I,0}^t = \Omega_K^t J^t$.
        Local update for the sub-problem of $G(\Theta_I, \Omega_K)$:
        for $r = 0$ to $R - 1$ do
            for $s = 0$ to $S - 1$ do
                Update personalized model: $\theta_i^{s+1} = \theta_i^s - \eta \nabla F_i(\theta_i^s)$.
            end for
            Local update: $\tilde{\omega}_{i,r+1}^t = \tilde{\omega}_{i,r}^t - \beta \nabla_{\omega_i} G(\tilde{\theta}_i(\tilde{\omega}_{i,r}^t), \tilde{\omega}_{i,r}^t)$.
        end for
    end for
    Clients send back $\tilde{\omega}_{i,R}^t$ and server conducts clustering (e.g., $k$-means++) on models $\Omega_{I,R}^t$ to obtain $P^{t+1}$.
    Global aggregation: $\Omega_K^{t+1} = \Omega_K^t - \alpha(\Omega_K^t - \Omega_{I,R}^t P^{t+1})$.
end for
return The personalized models $\Theta_I^T$.
```

When considering the relationship between the consecutive P^t, we can formulate the iterate as $P^{t+1} = P^tQ^t$, where $Q^t \in \mathbb{R}^{K\times K}$ is a square matrix. We can find that to maintain the above property of P^t $(\forall t)$, the matrix Q^t must satisfies that:

$$\sum_{j=1}^{K}(Q^t)_{j,k} = 1, \, \forall k, t \qquad \text{and} \qquad \sum_{k=1}^{K}(Q^t)_{j,k} = 1, \, \forall j, t. \tag{3.4}$$

It is noticed that the clustering is based on the latest model parameters Ω_I^{t+1} that depends on Ω_I^t, and the latest gradient updates given by clients. Hence, P^{t+1} is determined by and only by P^t and Q^t. Then we can consider this global iteration as a discrete-time Markov chain and Q^t corresponds the transition probability matrix.

During each local round, the clients need to first utilize local datasets to solve the regularized optimization objective, i.e., the upper-level objective in (3.3) with fixed $\tilde{\omega}_{i,r}^t$ to obtain a δ-approximate solution $\tilde{\theta}_i(\tilde{\omega}_{i,r}^t)$. Then, each client is required to calculate the gradients $\nabla_{\omega_i}G(\tilde{\theta}_i(\tilde{\omega}_{i,r}^t), \tilde{\omega}_{i,r}^t)$ with fixed $\tilde{\theta}_i(\tilde{\omega}_{i,r}^t)$ and update the model using $\tilde{\omega}_{i,r+1}^t = \tilde{\omega}_{i,r}^t - \beta\nabla_{\omega_i}G(\tilde{\theta}_i(\tilde{\omega}_{i,r}^t), \tilde{\omega}_{i,r}^t)$, where β is the learning rate and $\nabla_{\omega_i}G(\tilde{\theta}_i(\tilde{\omega}_{i,r}^t), \tilde{\omega}_{i,r}^t) = \frac{2}{N}\nabla r\big(\tilde{\theta}_i(\tilde{\omega}_{i,r}^t), \tilde{\omega}_{i,r}^t\big)$. To reduce the communication overhead, our CGPFL allows the clients to process several local iterations before uploading the latest model parameters to the server. The details of CGPFL is given in algorithm 1, from which we can summarize the parameters update process as:

$$\Omega_{I,R}^{t-1} \xrightarrow{P^t} \Omega_K^t \xrightarrow{J^t} \Omega_{I,0}^t \xrightarrow{H_I^t} \Omega_{I,R}^t \xrightarrow{P^{t+1}} \Omega_K^{t+1}, \tag{3.5}$$

where $P^{t+1} = P^tQ^t$ and $J^tP^t = I_K$ $(J^t \in \mathbb{R}^{K\times N}$ and I_K is an identity matrix), $\forall t$.

3.6 Theoretical Analysis

To demonstrate the effectiveness of the proposed method from the theoretical perspectives, we provide the convergence rate (in section 3.6.1) and generalization error bound (in section 3.6.2) of CGPFL in the following part.

3.6.1 Convergence Rate

Since the inner-level objective in (3.3) is non-convex, we focus on analyzing the convergence rate under the smooth case. Firstly, we can write the local updates as:

$$\Omega_{I,R}^t = \Omega_{I,0}^t - \beta R H_I^t, \tag{3.6}$$

where $H_I^t = \frac{1}{R}\sum_{r=0}^{R-1} H_{I,r}^t$ and $H_{I,r}^t = \frac{2}{N}\big(\Omega_{I,r}^t - \widetilde{\Theta}_I(\Omega_{I,r}^t)\big)$. Based on (3.6) and the update process in (3.5), we can obtain the global updates as:

$$\Omega_K^{t+1} = (1-\alpha)\Omega_K^t + \alpha\Omega_{I,R}^t P^{t+1} = \Omega_K^t[(1-\alpha)I_K + \alpha Q^t] - \alpha\beta R H_I^t P^t Q^t.$$

Definition 3.6.1 (L-smooth). *If a function f satisfies $\|\nabla f(\omega) - \nabla f(\omega')\| \leq L\|\omega - (\omega)'\|$, $\forall \omega$, ω', we say f is L-smooth.*

Assumption 3.6.1 (Smoothness). *The loss functions $\mathcal{R}_i$ is L-smooth and $G(\omega_k)$ is L_G-smooth, $\forall i$, k.*

Assumption 3.6.2 (Bounded intra-context diversity). *The variance of local gradients to the corresponding context-level generalized models is upper bounded by:*

$$\frac{1}{|C_k|}\sum_{i \in C_k} \|\nabla G_{k,i}(\omega_k) - \nabla G_k(\omega_k)\|^2 \leq \delta_G^2, \forall k \in [K], \tag{3.7}$$

where $G_{k,i}(\omega_k) := r(\theta_i, \omega_k)$.

Assumption 3.6.3 (Bounded parameters and gradients). *The generalized model parameters Ω_K^t and the gradients $\nabla G_K(\Omega_K^t)$ are upper bounded by ρ_Ω and ρ_g, respectively.*

$$\left\|\Omega_K^t\right\|^2 \leq \rho_\Omega^2 \qquad \textit{and} \qquad \left\|\nabla G_K(\Omega_K^t)\right\|^2 \leq \rho_g^2, \quad \forall t \tag{3.8}$$

where ρ_Ω and ρ_g are finite non-negative constants, and the gradients $\nabla G_K(\Omega_K^t)$ is defined as $\nabla G_K(\Omega_K^t) := [\nabla G_1(\omega_1^t), ..., \nabla G_k(\omega_k^t), ..., \nabla G_K(\omega_K^t)]$.

Proposition 3.6.1. *[95] The deviation between the δ-approximate and the optimal solution is upper bounded by δ. That is:*

$$\mathbb{E}\Big[\big\|\widetilde{\Theta}_I(\Omega_{I,r}^t) - \widehat{\Theta}_I(\Omega_{I,r}^t)\big\|^2\Big] \leq N\delta^2, \forall r, t, \tag{3.9}$$

where $\widetilde{\Theta}_I$ is the δ-approximate solution and $\widehat{\Theta}_I$ is the matching optimal solution.

Assumption 3.6.1 provides typical conditions for convergence analysis, and assumption 3.6.2 is common in analyzing algorithms that are built on SGD. As for assumption 3.6.3, the model parameters are easily bounded by using projection during the model training process, while the gradients can be bounded with the smooth condition and bounded model parameters. To evaluate the convergence of the proposed CGPFL, we adopt the technique used in [95] to define that:

$$\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega_K^{t^*})\big\|^2\Big] := \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega_K^{t})\big\|^2\Big],$$

where t^* is uniformly sampled from the set $\{0, 1, \ldots, T-1\}$.

Theorem 3.6.1 (Convergence of CGPFL)**.** *Suppose Assumption 3.6.1, 3.6.2 and 3.6.3 hold. If $\beta \leq \frac{1}{2\sqrt{R(R+1)L_G^2}}, \forall R \geq 1$, $\alpha \leq 1$, and $\hat{\alpha}_0 := \min\left\{\frac{8\alpha^2\rho_\Omega^2}{K\Delta_G}, \sqrt{\frac{4}{3}\frac{\alpha\rho_\Omega}{\rho_g}}, \sqrt{\frac{1}{416L_G{}^2}\alpha}\right\}$, where Δ_G is defined as $\Delta_G := \mathbb{E}\Big[\frac{1}{K}\sum_{k=1}^{K} G_k(\omega_k^0) - \frac{1}{K}\sum_{k=1}^{K} G_k(\omega_k^T)\Big]$, we have:*

- *The convergence of the generalized models:*

$$\frac{1}{K}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^{t^*})\big\|^2\Big] \leq \mathcal{O}\bigg(\frac{48\alpha^2(\rho_\Omega^2/K)}{\hat{\alpha}_0^2 T} + \frac{80(26(\rho_\Omega^2/K)L_G^2\delta^2)^{\frac{1}{2}}}{\sqrt{NKRT}} + \frac{52\delta^2}{KN}\bigg).$$

- *The convergence of the personalized models:*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big[\big\|\widetilde{\Theta}_I^{t^*} - \Omega_K^{t^*}J^{t^*}\big\|^2\Big] \leq \mathcal{O}\Big(\frac{1}{K}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^{t^*})\big\|^2\Big]\Big) + \mathcal{O}\Big(\frac{\delta_G^2}{\lambda^2} + \delta^2\Big).$$

Proof. According to algorithm 3.5.2, the local update is given as follows:

$$\omega_{i,r+1}^t = \omega_{i,r}^t - \beta\nabla G_i(\omega_{i,r}^t) = \omega_{i,r}^t - \beta\underbrace{\frac{2}{N}(\omega_{i,r}^t - \tilde{\theta}_i(\omega_{i,r}^t))}_{:=h_{i,r}^t},$$

Summing the local iterates, we can get

$$\beta \sum_{r=0}^{R-1} h_{i,r}^t = \sum_{r=0}^{R-1} (\omega_{i,r}^t - \omega_{i,r+1}^t) = \omega_{i,0}^t - \omega_{i,R}^t.$$

According to the algorithm, we have

$$\Omega_K^{t+1} - \Omega_K^t = -\alpha(\Omega_K^t - \Omega_{I,R}^t P^{t+1}).$$

Therefore, we can get the model parameters of the global models as follows:

$$\begin{aligned}
\Omega_K^{t+1} &= (1-\alpha)\Omega_K^t + \alpha \Omega_{I,R}^t P^{t+1} \\
&= (1-\alpha)\Omega_K^t + \alpha(\Omega_{I,0}^t - \beta R \underbrace{\frac{1}{R}\sum_{r=0}^{R-1} H_{I,r}^t}_{:=H_I^t}) P^{t+1} \\
&= (1-\alpha)\Omega_K^t + \alpha \Omega_K^t J^t P^{t+1} - \underbrace{\alpha\beta R}_{:=\hat{\alpha}} H_I^t P^{t+1} \\
&= (1-\alpha)\Omega_K^t + \alpha \Omega_K^t J^t P^t Q^t - \hat{\alpha} H_I^t P^{t+1} \\
&= (1-\alpha)\Omega_K^t + \alpha \Omega_K^t Q^t - \hat{\alpha} H_I^t P^{t+1} \\
&= \Omega_K^t[(1-\alpha)I_K + \alpha Q^t] - \hat{\alpha} H_I^t P^t Q^t
\end{aligned}$$

That is

$$\Omega_K^t - \Omega_K^{t+1} = \alpha \Omega_K^t (I_K - Q^t) + \hat{\alpha} H_I^t P^{t+1}. \tag{3.10}$$

It's noted that

$$G_k(\omega_k^t) := \big[G_K(\Omega_K^t)\big]_k,$$

where $[G_K(\Omega_K^t)]_k$ denotes the k-th element of the row vector $G_K(\Omega_K^t)$.

In the following part, we derive the **Convergence Rate** of CGPFL as

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{k=1}^{K} G_k(\omega_k^{t+1}) - \sum_{k=1}^{K} G_k(\omega_k^{t})\Big] \\
&= \mathbb{E}\Big[\sum_{k=1}^{K}[G_I(\Omega_K^{t+1})P^{t+1}]_k - \sum_{k=1}^{K}[G_I(\Omega_K^{t})P^{t}]_k\Big] \\
&= \mathbb{E}\Big[\sum_{k=1}^{K}[G_I(\Omega_K^{t+1})P^{t+1} - G_I(\Omega_K^{t})P^{t}]_k\Big] \\
&= \mathbb{E}\Big[\sum_{k=1}^{K}\big[\big(G_I(\Omega_K^{t+1}) - G_I(\Omega_K^{t})\big)P^{t}\big]_k + \sum_{k=1}^{K}\big[G_I(\Omega_K^{t+1})P^{t}(Q^t - I_K)\big]_k\Big] \\
&\leq \underbrace{\mathbb{E}\Big[\big\langle \nabla G_K(\Omega_K^t), \Omega_K^{t+1} - \Omega_K^t\big\rangle\Big] + \frac{L_G}{2}\mathbb{E}\Big[\big\|\Omega_K^{t+1} - \Omega_K^t\big\|^2\Big]}_{\mathbf{A}} \\
&\quad + \underbrace{\mathbb{E}\Big[\sum_{k=1}^{K}\big[G_I(\Omega_K^{t+1})P^{t}(Q^t - I_K)\big]_k\Big]}_{\mathbf{B}},
\end{aligned}
$$

where we assume that $L_G := max_{k\in[K]}\, L_{G_k}$. We first deal with the part **A** in above inequality. According to the above derivation, we have

$$
\begin{aligned}
\mathbf{A} &= \mathbb{E}\Big[\big\langle \nabla G_K(\Omega_K^t), \Omega_K^{t+1} - \Omega_K^t\big\rangle\Big] + \frac{L_G}{2}\mathbb{E}\Big[\big\|\Omega_K^{t+1} - \Omega_K^t\big\|^2\Big] \\
&= -\hat{\alpha}\mathbb{E}\Big[\big\langle \nabla G_K(\Omega_K^t), \frac{1}{\hat{\alpha}}\big(\Omega_K^t - \Omega_K^{t+1}\big) - \nabla G_K(\Omega_K^t) + \nabla G_K(\Omega_K^t)\big\rangle\Big] \\
&\quad + \frac{L_G}{2}\mathbb{E}\Big[\big\|\Omega_K^{t+1} - \Omega_K^t\big\|^2\Big] \\
&= -\hat{\alpha}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{L_G}{2}\mathbb{E}\Big[\big\|\Omega_K^{t+1} - \Omega_K^t\big\|^2\Big] \\
&\quad - \hat{\alpha}\mathbb{E}\Big[\big\langle \nabla G_K(\Omega_K^t), \frac{1}{\hat{\alpha}}\big(\Omega_K^t - \Omega_K^{t+1}\big) - \nabla G_K(\Omega_K^t)\big\rangle\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\leq -\hat{\alpha}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\frac{1}{\hat{\alpha}}\big(\Omega_K^t - \Omega_K^{t+1}\big) - \nabla G_K(\Omega_K^t)\big\|^2\Big] \\
&\quad + \frac{L_G}{2}\mathbb{E}\Big[\big\|\Omega_K^{t+1} - \Omega_K^t\big\|^2\Big] \\
&= -\frac{\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \underbrace{\frac{L_G}{2}\mathbb{E}\Big[\big\|\Omega_K^{t+1} - \Omega_K^t\big\|^2\Big]}_{\mathbf{A_1}} \\
&\quad + \underbrace{\frac{\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\frac{1}{\hat{\alpha}}\big(\Omega_K^t - \Omega_K^{t+1}\big) - \nabla G_K(\Omega_K^t)\big\|^2\Big]}_{\mathbf{A_2}}
\end{aligned}
$$

Plugging equation (3.10) into above inequality, we can get

$$
\begin{aligned}
\mathbf{A_1} &= \frac{L_G}{2}\mathbb{E}\Big[\big\|\alpha\Omega_K^t(I_K - Q^t) + \hat{\alpha}H_I^tP^{t+1}\big\|^2\Big] \\
&= \frac{L_G}{2}\mathbb{E}\Big[\big\|\alpha\Omega_K^t(I_K - Q^t) + \hat{\alpha}H_I^tP^{t+1} - \hat{\alpha}\nabla G_I(\Omega_{I,0}^t)P^{t+1} + \hat{\alpha}\nabla G_I(\Omega_{I,0}^t)P^tQ^t\big\|^2\Big] \\
&\leq \frac{3\alpha^2 L_G}{2}\mathbb{E}\Big[\big\|\Omega_K^t(I_K - Q^t)\big\|^2\Big] + \frac{3\hat{\alpha}^2 L_G}{2}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)Q^t\big\|^2\Big] \\
&\quad + \frac{3\hat{\alpha}^2 L_G}{2}\mathbb{E}\Big[\big\|\big(H_I^t - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big]
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbf{A_2} &= \frac{\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\frac{\alpha}{\hat{\alpha}}\Omega_K^t(I_K - Q^t) + H_I^tP^{t+1} - \nabla G_I(\Omega_{I,0}^t)P^{t+1} + \nabla G_I(\Omega_{I,0}^t)P^tQ^t - \nabla G_K(\Omega_K^t)\big\|^2\Big] \\
&\leq \frac{3\alpha^2}{2\hat{\alpha}}\mathbb{E}\Big[\big\|\Omega_K^t(I_K - Q^t)\big\|^2\Big] + \frac{3\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)(I_K - Q^t)\big\|^2\Big] \\
&\quad + \frac{3\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\big(H_I^t - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big]
\end{aligned}
$$

Proposition 3.6.2. *For any vector $x_i \in \mathbb{R}^d, i = 1, 2, \ldots, M$, according to Jensen's inequality, we have*

$$
\Big\|\sum_{i=1}^{M} x_i\Big\|^2 \leq M\sum_{i=1}^{M}\|x_i\|^2.
$$

And because the real function $\varphi(y) = y^2, y \in \mathbb{R}$ is convex, if some constants satisfy that $\lambda_i \geq 0, \forall i = 1, 2, \ldots, M$, and $\sum_{i=1}^{M}\lambda_i = 1$, we have

$$
\Big\|\sum_{i=1}^{M} \lambda_i y_i\Big\|^2 \leq \sum_{i=1}^{M}\lambda_i\|y_i\|^2.
$$

Lemma 3.6.1. *We can obtain that $\mathbb{E}\Big[\big\|XP^{t+1}\big\|^2\Big] \leq \mathbb{E}\Big[\big\|X\big\|^2\Big]$, and $\mathbb{E}\Big[\big\|YQ^t\big\|^2\Big] \leq \mathbb{E}\Big[\big\|Y\big\|^2\Big]$ for any matrices $X \in \mathbb{R}^{d\times N}$ and $Y \in \mathbb{R}^{d\times K}$, as long as the P^{t+1} and Q^t satisfy that $\sum_{i=1}^{N} P_{i,k}^{t+1} = 1$, $\sum_{j=1}^{K} Q_{j,k}^t = 1, \forall k, t$, and $\sum_{k=1}^{K} Q_{j,k}^t = 1, \forall j, t$. Especially in this work, we have $P_{i,k}^{t+1} = \begin{cases} \frac{1}{|C_k|}, & \text{if } i \in C_k \\ 0, & \text{otherwise} \end{cases}$.*

Proof. We provide the proof of above useful lemma here.

$$\begin{aligned}
\mathbb{E}\Big[\|XP^{t+1}\|^2\Big] &= \sum_{l=1}^{d}\sum_{k=1}^{K}\big[(XP^{t+1})_{l,k}\big]^2 \\
&= \sum_{l=1}^{d}\sum_{k=1}^{K}\Big[\sum_{i=1}^{N} X_{l,i}P_{i,k}^{t+1}\Big]^2 \\
&\leq \sum_{l=1}^{d}\sum_{k=1}^{K}\sum_{i=1}^{N} {X_{l,i}}^2 P_{i,k}^{t+1} \\
&= \sum_{l=1}^{d}\sum_{i=1}^{N}\sum_{k=1}^{K} {X_{l,i}}^2 P_{i,k}^{t+1} \\
&= \sum_{l=1}^{d}\sum_{i=1}^{N} {X_{l,i}}^2 \sum_{k=1}^{K} P_{i,k}^{t+1} \\
&\leq \sum_{l=1}^{d}\sum_{i=1}^{N} {X_{l,i}}^2 = \mathbb{E}\big[\|X\|^2\big]
\end{aligned}$$

Similarly, we can write that

$$\begin{aligned}
\mathbb{E}\big[\|YQ^t\|^2\big] &= \sum_{l=1}^{d}\sum_{k=1}^{K}\big[(YQ^t)_{l,k}\big]^2 \\
&= \sum_{l=1}^{d}\sum_{k=1}^{K}\Big[\sum_{j=1}^{K} Y_{l,j}Q_{j,k}^{t}\Big]^2 \\
&\leq \sum_{l=1}^{d}\sum_{k=1}^{K}\sum_{j=1}^{K} {Y_{l,j}}^2 Q_{j,k}^{t} = \sum_{l=1}^{d}\sum_{k=1}^{K}\sum_{j=1}^{K} {Y_{l,j}}^2 Q_{j,k}^{t} \\
&= \sum_{l=1}^{d}\sum_{j=1}^{N} {Y_{l,j}}^2 \sum_{k=1}^{K} Q_{j,k}^{t} \\
&= \sum_{l=1}^{d}\sum_{j=1}^{N} {Y_{l,j}}^2 = \mathbb{E}\big[\|Y\|^2\big]
\end{aligned}$$

Therefore, the proof of Lemma 3.6.1 is complete. □

In the next part, we will first cope with the term $\mathbb{E}\Big[\big\|H_I^t - \nabla G_I(\Omega_{I,0}^t)\big\|^2\Big]$.

$$
\begin{aligned}
&\mathbb{E}\Big[\big\|\big(G_I^t - \nabla F_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{R}\sum_{r=0}^{R-1}\big(H_{I,r}^t - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\Big\|^2\Big] \\
&\le \frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\big(H_{I,r}^t - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&= \frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\big(H_{I,r}^t - \nabla G_I(\Omega_{I,r}^t) + \nabla G_I(\Omega_{I,r}^t) - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&\le \frac{2}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\big(H_{I,r}^t - \nabla G_I(\Omega_{I,r}^t)\big)P^{t+1}\big\|^2\Big] + \frac{2}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\big(\nabla G_I(\Omega_{I,r}^t) - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&\le \frac{2}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|H_{I,r}^t - \nabla G_I(\Omega_{I,r}^t)\big\|^2\Big] + \frac{2}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\big(\nabla G_I(\Omega_{I,r}^t) - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&\le \frac{2}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\Big\|\frac{2}{N}(\widetilde{\Theta}_i(\Omega_{I,r}^t) - \widehat{\Theta}_i(\Omega_{I,r}^t))\Big\|^2\Big] + \frac{2{L_G}^2}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\big(\Omega_{I,r}^t - \Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big] \\
&\le \frac{8}{N}\delta^2 + \frac{2{L_G}^2}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\big(\Omega_{I,r}^t - \Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big]
\end{aligned}
$$

In above inequality, we can can bound the term $\mathbb{E}\Big[\big\|\big(\Omega_{I,r}^t - \Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big]$ by

$$
\begin{aligned}
&\mathbb{E}\Big[\big\|\big(\Omega_{I,r}^t - \Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big] \\
&= \mathbb{E}\Big[\big\|\big(\Omega_{I,r-1}^t - \Omega_{I,0}^t - \beta H_{I,r-1}^t\big)P^{t+1}\big\|^2\Big] \\
&= \mathbb{E}\Big[\big\|\big(\Omega_{I,r-1}^t - \Omega_{I,0}^t - \beta\nabla G_I(\Omega_{I,0}^t) + \beta\nabla G_I(\Omega_{I,0}^t) - \beta H_{I,r-1}^t\big)P^{t+1}\big\|^2\Big] \\
&\le (1+\frac{1}{R})\mathbb{E}\Big[\big\|\big(\Omega_{I,r-1}^t - \Omega_{I,0}^t - \beta\nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&\quad + (1+R)\beta^2\mathbb{E}\Big[\big\|\big(\nabla G_I(\Omega_{I,0}^t) - H_{I,r-1}^t\big)P^{t+1}\big\|^2\Big] \\
&\le (1+\frac{1}{R})(1+\frac{1}{2R})\mathbb{E}\Big[\big\|\big(\Omega_{I,r-1}^t - \Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big] + (1+\frac{1}{R})(1+2R)\beta^2\mathbb{E}\Big[\big\|\nabla G_I(\Omega_{I,0}^t)P^tQ^t\big\|^2\Big] \\
&\quad + \beta^2(1+R)\Big(\frac{8}{N}\delta^2 + 2{L_G}^2\mathbb{E}\Big[\big\|\big(\Omega_{I,r-1}^t - \Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big]\Big) \\
&= (1+\frac{1}{R})\big(1+\frac{1}{2R} + 2(1+R)\beta^2{L_G}^2\big)\mathbb{E}\Big[\big\|\big(\Omega_{I,r-1}^t - \Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big] \\
&\quad + (1+\frac{1}{R})(1+2R)\beta^2\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)Q^t\big\|^2\Big] + \frac{8(1+R)\beta^2}{N}\delta^2
\end{aligned}
$$

When $\beta^2 \le \frac{1}{4R(1+R){L_G}^2}$, which implies that $2(1+R)\beta^2 {L_G}^2 \le \frac{1}{2R}$, we can drive that

$$\mathbb{E}\Big[\big\|\big(\Omega_{I,r}^t - \Omega_{I,0}^t\big) P^{t+1}\big\|^2\Big] \le (1+\frac{1}{R})^2 \mathbb{E}\Big[\big\|\big(\Omega_{I,r-1}^t - \Omega_{I,0}^t\big) P^{t+1}\big\|^2\Big] + (1+\frac{1}{R})(1+2R)\beta^2 \mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{8(1+R)\beta^2}{N}\delta^2.$$

By unrolling the above result recursively, we can bound the term $\mathbb{E}\Big[\big\|\big(\Omega_{I,r}^t - \Omega_{I,0}^t\big) P^{t+1}\big\|^2\Big]$ as follows:

$$\begin{aligned}
&\mathbb{E}\Big[\big\|\big(\Omega_{I,r}^t - \Omega_{I,0}^t\big) P^{t+1}\big\|^2\Big] \\
&\le \Big\{(1+\frac{1}{R})(1+2R)\beta^2 \mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{8(1+R)\beta^2}{N}\delta^2\Big\} \sum_{\hat{r}=0}^{r-2} \big(1+\frac{1}{R}\big)^{2\hat{r}} \\
&\le \Big\{(1+\frac{1}{R})(1+2R)\beta^2 \mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{8(1+R)\beta^2}{N}\delta^2\Big\} \frac{(1+\frac{1}{R})^{2(r-1)} - 1}{(1+\frac{1}{R})^2 - 1} \\
&\le \Big\{(1+\frac{1}{R})(1+2R)\beta^2 \mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{8(1+R)\beta^2}{N}\delta^2\Big\} \frac{(1+\frac{1}{R})^{2(r-1)}}{(1+\frac{1}{R})^2 - 1}.
\end{aligned}$$

With this inequality, we can write that

$$\begin{aligned}
&\mathbb{E}\Big[\big\|\big(H_I^t - \nabla G_I(\Omega_{I,0}^t)\big) P^{t+1}\big\|^2\Big] \\
&\le \frac{8}{N}\delta^2 + \frac{2{L_G}^2}{R} \sum_{r=0}^{R-1} \mathbb{E}\Big[\big\|\big(\Omega_{I,r}^t - \Omega_{I,0}^t\big) P^{t+1}\big\|^2\Big] \\
&\le \frac{8}{N}\delta^2 + \frac{2\beta^2{L_G}^2}{R}\Big\{(1+\frac{1}{R})(1+2R) \mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{8(1+R)}{N}\delta^2\Big\} \sum_{r=0}^{R-1} \frac{(1+\frac{1}{R})^{2(r-1)}}{(1+\frac{1}{R})^2 - 1} \\
&\le \frac{8}{N}\delta^2 + \frac{2\beta^2{L_G}^2}{R}\Big\{(1+\frac{1}{R})(1+2R) \mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{8(1+R)}{N}\delta^2\Big\} \frac{(1+\frac{1}{R})^{2R} - 1}{(1+\frac{1}{R})^2 - 1} \\
&\le \frac{8}{N}\delta^2 + \frac{2\beta^2{L_G}^2}{R}\Big\{(1+\frac{1}{R})(1+2R) \mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{8(1+R)}{N}\delta^2\Big\} \frac{e^2 - 1}{(1+\frac{1}{R})^2 - 1} \\
&\le \frac{8}{N}\delta^2 + \frac{2\beta^2{L_G}^2}{R}\Big\{(1+\frac{1}{R})(1+2R) \mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{8(1+R)}{N}\delta^2\Big\} \frac{8R^2}{1+2R} \\
&= \frac{8}{N}\delta^2 + \frac{128R(1+R)\beta^2{L_G}^2\delta^2}{(1+2R)N} + 16(1+R)\beta^2{L_G}^2 \mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] \\
&\le \frac{8}{N}\delta^2 + \frac{128R\beta^2{L_G}^2\delta^2}{N} + 32R\beta^2{L_G}^2 \mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]
\end{aligned}$$

Finally, we can rewrite the term A as

$$
\begin{aligned}
\mathbf{A} &= -\frac{\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \mathbf{A_1} + \mathbf{A_2} \\
&\leq \Big(-\frac{\hat{\alpha}}{2} + \frac{3\hat{\alpha}^2 L_G}{2}\Big)\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \Big(\frac{3\hat{\alpha}^2 L_G}{2} + \frac{3\hat{\alpha}}{2}\Big)\mathbb{E}\Big[\big\|\big(H_I^t - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&\quad + \Big(\frac{3\alpha^2 L_G}{2} + \frac{3\alpha^2}{2\hat{\alpha}}\Big)\underbrace{\mathbb{E}\Big[\big\|\Omega_K^t(I_K - Q^t)\big\|^2\Big]}_{\mathbf{B_1}} + \frac{3\hat{\alpha}}{2}\underbrace{\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)(I_K - Q^t)\big\|^2\Big]}_{\mathbf{B_2}}
\end{aligned}
$$

For simplicity, we can write that

$$
\begin{aligned}
\mathbf{A} &\leq -\frac{\hat{\alpha}}{2}(1 - 3\hat{\alpha}L_G)\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{3\hat{\alpha}}{2}(1 + \hat{\alpha}L_G)\mathbb{E}\Big[\big\|\big(H_I^t - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&\quad + \frac{3\alpha^2}{2}\big(L_G + \frac{1}{\hat{\alpha}}\big)\mathbf{B_1} + \frac{3\hat{\alpha}}{2}\mathbf{B_2} \\
&\leq -\frac{\hat{\alpha}}{2}(1 - 3\hat{\alpha}L_G)\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{3\alpha^2}{2}\big(L_G + \frac{1}{\hat{\alpha}}\big)\mathbf{B_1} + \frac{3\hat{\alpha}}{2}\mathbf{B_2} \\
&\quad + \frac{3\hat{\alpha}}{2}(1 + \hat{\alpha}L_G)\Big\{\frac{8}{N}\delta^2 + \frac{128R\beta^2 {L_G}^2\delta^2}{N} + 32R\beta^2 {L_G}^2\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]\Big\} \\
&= -\frac{\hat{\alpha}}{2}\big(1 - 3\hat{\alpha}L_G - 96R\beta^2 {L_G}^2(1 + \hat{\alpha}L_G)\big)\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] \\
&\quad + \frac{3\alpha^2}{2}\big(L_G + \frac{1}{\hat{\alpha}}\big)\mathbf{B_1} + \frac{3\hat{\alpha}}{2}\mathbf{B_2} + \frac{192\hat{\alpha}^3\delta^2 {L_G}^2(1 + \hat{\alpha}L_G)}{NR\alpha^2} + \frac{12\hat{\alpha}(1 + \hat{\alpha}L_G)\delta^2}{N} \\
&= -\frac{\hat{\alpha}}{2}\underbrace{\big(1 - 3\hat{\alpha}L_G - 96R\beta^2 {L_G}^2(1 + \hat{\alpha}L_G)\big)}_{\geq\frac{1}{2}\text{ when }\beta^2 {L_G}^2 \leq \frac{1}{416R^2}\text{ and }\alpha\leq 1}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] \\
&\quad + \frac{3\alpha^2}{2}\big(L_G + \frac{1}{\hat{\alpha}}\big)\mathbf{B_1} + \frac{3\hat{\alpha}}{2}\mathbf{B_2} + \frac{192\hat{\alpha}^3\delta^2 {L_G}^2(1 + \hat{\alpha}L_G)}{NR\alpha^2} + \frac{12\hat{\alpha}(1 + \hat{\alpha}L_G)\delta^2}{N} \\
&\leq -\frac{\hat{\alpha}}{4}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{192\hat{\alpha}^3\delta^2 {L_G}^2(1 + \hat{\alpha}L_G)}{NR\alpha^2} + \frac{12\hat{\alpha}(1 + \hat{\alpha}L_G)\delta^2}{N} \\
&\quad + \frac{3\alpha^2}{2\hat{\alpha}}(\hat{\alpha}L_G + 1)\mathbf{B_1} + \frac{3\hat{\alpha}}{2}\mathbf{B_2} \\
&\leq -\frac{\hat{\alpha}}{4}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{2\alpha^2}{\hat{\alpha}}\mathbf{B_1} + \frac{3\hat{\alpha}}{2}\mathbf{B_2} + \frac{208\hat{\alpha}^3\delta^2 {L_G}^2}{NR\alpha^2} + \frac{13\hat{\alpha}\delta^2}{N}
\end{aligned}
$$

with $\beta^2 {L_G}^2 \leq \frac{1}{416R^2} \leq \frac{1}{8R^2} \leq \frac{1}{4R(1+R)}$, $\forall R \geq 1$ and $\alpha \leq 1$.

When the above conditions are satisfied, we can have

$$
\hat{\alpha}L_G = R\alpha\beta L_G \leq \frac{R}{\sqrt{416R^2}} \leq \frac{1}{12}, \tag{3.11}
$$

and the term $96R\beta^2 {L_G}^2(1 + \hat{\alpha}L_G)$ is bounded by

$$
96R\beta^2 {L_G}^2(1 + \hat{\alpha}L_G) \leq \frac{96R}{416R^2}(1 + \frac{1}{12}) = \frac{1}{4R} \leq \frac{1}{4}, \forall R \geq 1. \tag{3.12}
$$

In summary we can drive that,

$$(1 - 3\hat{\alpha}L_G - 96R\beta^2 {L_G}^2(1 + \hat{\alpha}L_G)) \geq 1 - \frac{1}{4} - \frac{1}{4R} \geq \frac{1}{2}.$$

Lemma 3.6.2. *When Assumption 3.6.1 is satisfied, the following two statements are granted:*

1. *Equation* $\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \underbrace{\mathbb{E}\left[\left\|\Omega_K^t(Q^t - I_K)\right\|^2\right]}_{\mathbf{B_1}} = 0$ *is equivalent to the equation* $\lim_{T\to\infty} \left\|Q^T - I_K\right\|^2 = 0$*, and we can get* $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\Omega_K^t(Q^t - I_K)\right\|^2\right] \leq \mathcal{O}(\frac{1}{T})$*;*
2. *Equation* $\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \underbrace{\mathbb{E}\left[\left\|\nabla G_K(\Omega_K^t)(Q^t - I_K)\right\|^2\right]}_{\mathbf{B_2}} = 0$ *is equivalent to the equation* $\lim_{T\to\infty} \left\|Q^T - I_K\right\|^2 = 0$*, and* $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla G_K(\Omega_K^t)(Q^t - I_K)\right\|^2\right] \leq \mathcal{O}(\frac{1}{T})$.

Proof. Firstly, we prove the **Sufficiency** in the first statement by contradiction. Now, we have the equation:

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\Omega_K^t(Q^t - I_K)\right\|^2\right] = 0.$$

Assuming that $\lim_{T\to\infty} \left\|Q^T - I_K\right\|^2 \neq 0$, we can get

$$\exists j, k \in [K], \lim_{T\to\infty} \left|\left(Q^T - I_K\right)_{j,k}\right| \neq 0.$$

In other words,

$$\forall T,\ \exists j_T, k_T \in [K] \text{ and } \delta_T > 0,\ \left|\left(Q^T - I_K\right)_{j_T,k_T}\right| > \delta_T.$$

Because we can always find some Ω_K^t making that

$$\left|\sum_{j=1}^{K} (\Omega_K^t)_{l,j}(Q^t - I_K)_{j,k}\right| = \sum_{j=1}^{K} \left|(\Omega_K^t)_{l,j}(Q^t - I_K)_{j,k}\right|,$$

we can derive that

$$
\begin{aligned}
\Big|\sum_{t=0}^{T-1} \left\|\Omega_K^t(Q^t - I_K)\right\|^2\Big| &= \sum_{t=0}^{T-1}\sum_{l=1}^{d}\sum_{k=1}^{K} \left[\Omega_K^t(Q^t - I_K)\right]_{j,k}{}^2 \\
&= \sum_{t=0}^{T-1}\sum_{l=1}^{d}\sum_{k=1}^{K} \Big[\sum_{j=1}^{K}(\Omega_K^t)_{l,j}(Q^t - I_K)_{j,k}\Big]^2 \\
&\geq \sum_{t=0}^{T-1}\sum_{l=1}^{d}\sum_{k=1}^{K} \Big[\sum_{j=1}^{K}(\Omega_K^t)_{l,j}{}^2(Q^t - I_K)_{j,k}{}^2\Big] \\
&\geq \sum_{t=0}^{T-1}\sum_{l=1}^{d}(\Omega_K^t)_{l,j_t}{}^2(Q^t - I_K)_{j_t,k_t}{}^2 \\
&\geq \sum_{t=0}^{T-1}\delta_{\Omega max}{}^2\delta_t{}^2
\end{aligned}
$$

where $\delta_{\Omega max}{}^2 = \min_{t\in[T]}\max_{l\in[d]}\left\{(\Omega_K^t)_{l,j_t}{}^2\right\}$ and $\delta_{\Omega max}{}^2 > 0$ (Otherwise, $(\Omega_K^t)_{l,j_t} = 0, \forall l$. Thus, the j_t-th global model is invalid). Then we have

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\left\|\Omega_K^t(Q^t - I_K)\right\|^2\Big] \geq \frac{1}{T}\sum_{t=0}^{T-1}\delta_{\Omega max}{}^2\delta_t{}^2 > 0.
$$

In summary, we can get

$$
\forall T,\ \exists\delta = \frac{1}{T}\sum_{t=0}^{T-1}\delta_{\Omega max}{}^2\delta_t{}^2 > 0,\ \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\left\|\Omega_K^t(Q^t - I_K)\right\|^2\Big] > \delta,
$$

which means that

$$
\underset{T\to\infty}{lim}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\left\|\Omega_K^t(Q^t - I_K)\right\|^2\Big] \neq 0.
$$

It contradicts the assumption. The proof of **Sufficiency** ends.

Next, we prove the **Necessity** in the first statement. Now, we have

$$
\lim_{T\to\infty}\left\|Q^T - I_K\right\|^2 = 0,
$$

which indicates that $\forall j,k$ and $\varepsilon_0 > 0$, $\exists T_0 > 0$, making $\forall T > T_0$, $\left|(Q^T - I_K)_{j,k}\right| < \varepsilon_0$.

We know that $\lim_{T\to\infty}\frac{T_1}{T} = 0$, $\forall T_1$, which means that

$$
\forall\varepsilon_1 > 0,\ \exists T_2,\ \text{making}\ \forall T > T_2,\ \frac{T_0+1}{T} < \varepsilon_1.
$$

When $T_3 = \max\{T_0, T_2\}$, $\forall T > T_3$, we can have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T_0} \left\|\Omega_K^t(Q^t - I_K)\right\|^2 \\
&= \frac{1}{T}\sum_{t=0}^{T_0}\sum_{l=1}^{d}\sum_{k=1}^{K}\Big[\sum_{j=1}^{K}(\Omega_K^t)_{l,j}(Q^t - I_K)_{j,k}\Big]^2 \\
&= \frac{1}{T}\sum_{t=0}^{T_0}\sum_{l=1}^{d}\sum_{k=1}^{K}\Big[\sum_{j=1}^{K}(\Omega_K^t)_{l,j}(Q^t)_{j,k} - \sum_{j=1}^{K}(\Omega_K^t)_{l,j}(I_K)_{j,k}\Big]^2 \\
&\leq \frac{2}{T}\sum_{t=0}^{T_0}\sum_{l=1}^{d}\sum_{k=1}^{K}\Big\{\Big[\sum_{j=1}^{K}(\Omega_K^t)_{l,j}(Q^t)_{j,k}\Big]^2 + \Big[\sum_{j=1}^{K}(\Omega_K^t)_{l,j}(I_K)_{j,k}\Big]^2\Big\} \\
&\leq \frac{2}{T}\sum_{t=0}^{T_0}\sum_{l=1}^{d}\sum_{k=1}^{K}\Big\{\sum_{j=1}^{K}(\Omega_K^t)_{l,j}{}^2(Q^t)_{j,k} + (\Omega_K^t)_{l,k}{}^2\Big\} \\
&= \frac{2}{T}\sum_{t=0}^{T_0}\sum_{l=1}^{d}\sum_{j=1}^{K}(\Omega_K^t)_{l,j}{}^2\sum_{k=1}^{K}(Q^t)_{j,k} + \frac{2}{T}\sum_{t=0}^{T-1}\sum_{l=1}^{d}\sum_{k=1}^{K}(\Omega_K^t)_{l,k}{}^2 \\
&\leq \frac{4}{T}\sum_{t=0}^{T_0}\sum_{l=1}^{d}\sum_{k=1}^{K}(\Omega_K^t)_{l,k}{}^2 \\
&\leq \frac{4\rho_\Omega^2(T_0+1)}{T}
\end{aligned}$$

Plugging the above inequality into the term $\left|\frac{1}{T}\sum_{t=0}^{T-1}\left\|\Omega_K^t(Q^t - I_K)\right\|^2\right|$, we can get

$$\begin{aligned}
&\Big|\frac{1}{T}\sum_{t=0}^{T-1}\left\|\Omega_K^t(Q^t - I_K)\right\|^2\Big| \\
&= \frac{1}{T}\sum_{t=0}^{T_0}\left\|\Omega_K^t(Q^t - I_K)\right\|^2 + \frac{1}{T}\sum_{t=T_0+1}^{T-1}\left\|\Omega_K^t(Q^t - I_K)\right\|^2 \\
&\leq \frac{4\rho_\Omega^2(T_0+1)}{T} + \frac{1}{T}\sum_{t=T_0+1}^{T-1}\sum_{l=1}^{d}\sum_{k=1}^{K}K\Big[\sum_{j=1}^{K}(\Omega_K^t)_{l,j}{}^2(Q^t - I_K)_{j,k}{}^2\Big] \\
&\leq \frac{4\rho_\Omega^2(T_0+1)}{T} + \frac{1}{T}\sum_{t=T_0+1}^{T-1}\sum_{l=1}^{d}\sum_{k=1}^{K}K\varepsilon_0{}^2\sum_{j=1}^{K}(\Omega_K^t)_{l,j}{}^2 \\
&\leq \rho_\Omega^2\Big(\frac{4(T_0+1)}{T} + \frac{T-T_0-1}{T}K^2\varepsilon_0{}^2\Big) \\
&< \underbrace{\rho_\Omega^2\big(4\varepsilon_1 + K^2\varepsilon_0{}^2\big)}_{:=\varepsilon}
\end{aligned}$$

That is, $\forall \varepsilon > 0$, $\exists T_3 = max\{T_0, T_2\}$, making $\forall T > T_3$, $\left| \frac{1}{T} \sum_{t=0}^{T-1} \left\| \Omega_K^t (Q^t - I_K) \right\|^2 \right| < \varepsilon$, which is the definition of

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \Omega_K^t (Q^t - I_K) \right\|^2 \right] = 0.$$

Thus, the proof of **Necessity** in the first statement is complete.

Finally, we prove the inequality $\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \Omega_K^t (Q^t - I_K) \right\|^2 \right] \leq \mathcal{O}(\frac{1}{T})$ in the first statement. From the analysis of the algorithm CGPFL, we know the iterates of the global models are

$$\Omega_K^0 \longrightarrow \cdots \longrightarrow \Omega_K^t \longrightarrow \Omega_K^{t+1}$$

At any global round t, we consider a client i which belongs to the cluster k at current round, i.e., $i \in C_k^t$. At the next round $t+1$, we focus on any cluster j, where $j \in [K]$. According to the definition of P^t, we have

$$P_{i,j}^{t+1} = \sum_{p=1}^{K} P_{i,p}^t (Q^t)_{p,j} \tag{3.13}$$

Since we focus on the disjoint cluster structure, i.e., $P_{i,k}^t = \begin{cases} \frac{1}{|C_k^t|}, & \text{if } i \in C_k^t \\ 0, & \text{otherwise} \end{cases}$, we can get that $P_{i,j}^{t+1} = \frac{1}{|C_k^t|} (Q^t)_{k,j}$. We know that the k-means clustering partitions the data points into different groups according to the distances between the data points and the centers of the clusters, i.e., $P_{i,k}^t = Probability\big(k = \arg\min_{p \in [K]} \|\omega_{i,R}^{t-1} - \omega_p^t\|^2\big)$. Because the global models are initialized from a same point, under the non-IID case, the distances between these models will necessarily become larger than certain tiny positive constants δ_d^2 after one global steps. Then the models can be separated into different clusters, and gradually the cluster structure will remain invariant since the updates of model parameters become smaller and smaller as the learning rate shrinks. Therefore, as long as the index of the selected initialization centroid in k-means clustering keeps unchanged (e.g., k-means++. This is the reason why we adopt k-means++ in our algorithm to conduct clustering) during the algorithm, Q^t will keep equal to I_K after the first few global rounds. And we can get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\Omega_K^t(Q^t - I_K)\big\|^2\Big] \leq \mathcal{O}\Big(\frac{4\rho_\Omega^2}{T}\Big) \tag{3.14}$$

Similarly, the inequality in the second statement can be granted as

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)(Q^t - I_K)\big\|^2\Big] \leq \mathcal{O}\Big(\frac{4\rho_g^2}{T}\Big) \tag{3.15}$$

Here, the proof of Lemma 3.6.2 is complete. □

In the next part, we will first deal with $\mathbf{B} = \mathbb{E}\Big[\sum_{k=1}^K \big[G_I(\Omega_K^{t+1})P^t(Q^t - I_K)\big]_k\Big]$ and give the proof of $\mathbf{B} = 0$.

$$\begin{aligned}
&\sum_{k=1}^K \big[G_I(\Omega_K^{t+1})P^t(Q^t - I_K)\big]_k \\
=&\sum_{k=1}^K\sum_{j=1}^K \big[G_I(\Omega_K^{t+1})P^t\big]_j(Q^t - I_K)_{j,k} \\
=&\sum_{j=1}^K \big[G_I(\Omega_K^{t+1})P^t\big]_j\sum_{k=1}^K(Q^t - I_K)_{j,k} \\
=&\sum_{j=1}^K \big[G_I(\Omega_K^{t+1})P^t\big]_j\Big[\sum_{k=1}^K(Q^t)_{j,k} - \sum_{k=1}^K(I_K)_{j,k}\Big] \equiv 0,
\end{aligned}$$

no matter what value $G_I(\Omega_K^{t+1})P^t$ takes. Therefore, we can conclude

$$\mathbf{B} = \mathbb{E}\Big[\sum_{k=1}^K \big[G_I(\Omega_K^{t+1})P^t(Q^t - I_K)\big]_k\Big] = 0. \tag{3.16}$$

In summary, we can bound $\mathbb{E}\Big[\sum_{k=1}^K G_k(\omega_k^{t+1}) - \sum_{k=1}^K G_k(\omega_k^t)\Big]$ as follows:

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{k=1}^K G_k(\omega_k^{t+1}) - \sum_{k=1}^K G_k(\omega_k^t)\Big] \\
\leq& -\frac{\hat{\alpha}}{4}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{2\alpha^2}{\hat{\alpha}}\mathbf{B_1} + \frac{3\hat{\alpha}}{2}\mathbf{B_2} + \frac{208\hat{\alpha}^3\delta^2 {L_G}^2}{NR\alpha^2} + \frac{13\hat{\alpha}\delta^2}{N}.
\end{aligned}$$

With this inequality, we can write that

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] \\
&\leq \frac{4}{\hat{\alpha}T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\sum_{k=1}^{K}G_k(\omega_k^t) - \frac{1}{K}\sum_{k=1}^{K}G_k(\omega_k^{t+1})\Big] \\
&\quad + \frac{8\alpha^2}{K\hat{\alpha}^2T}\sum_{t=0}^{T-1}\mathbf{B_1} + \frac{6}{KT}\sum_{t=0}^{T-1}\mathbf{B_2} + \frac{832\hat{\alpha}^2\delta^2 {L_G}^2}{KNR\alpha^2} + \frac{52\delta^2}{KN} \\
&\leq \frac{4\mathbb{E}\Big[\frac{1}{K}\sum_{k=1}^{K}G_k(\omega_k^0) - \frac{1}{K}\sum_{k=1}^{K}G_k(\omega_k^T)\Big]}{\hat{\alpha}T} + \frac{32\alpha^2\rho_\Omega^2}{K\hat{\alpha}^2T} + \frac{24\rho_g^2}{KT} + \frac{832\hat{\alpha}^2\delta^2 {L_G}^2}{KNR\alpha^2} + \frac{52\delta^2}{KN}
\end{aligned}
$$

We define that $\Delta_G := \mathbb{E}\Big[\frac{1}{K}\sum_{k=1}^{K}G_k(\omega_k^0) - \frac{1}{K}\sum_{k=1}^{K}G_k(\omega_k^T)\Big]$ which is a constant with finite value, $C_1 := \frac{32\rho_\Omega^2}{K}$, $C_2 := \frac{24\rho_g^2}{K}$ and $C_3 := \frac{832\delta^2 {L_G}^2}{KNR}$, then we get

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] \leq \frac{4\Delta_G}{\hat{\alpha}T} + \frac{C_1\alpha^2}{\hat{\alpha}^2T} + \frac{C_2}{T} + \frac{C_3\hat{\alpha}^2}{\alpha^2} + \frac{52\delta^2}{KN}. \tag{3.17}
$$

With $\hat{\alpha}_0 := min\Big\{\frac{C_1\alpha^2}{4\Delta_G}, \sqrt{\frac{C_1}{C_2}}\alpha, \sqrt{\frac{1}{416{L_G}^2}}\alpha\Big\}$, we consider two cases as [49, 5, 95] do.

If $\hat{\alpha}_0 \leq \alpha\Big(\frac{C_1}{C_3T}\Big)^{\frac{1}{4}}$, we choose $\hat{\alpha} = \hat{\alpha}_0$. Thus we have

$$
\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] \leq \frac{3C_1\alpha^2}{2\hat{\alpha}_0^2T} + \frac{(C_1C_3)^{\frac{1}{2}}}{2\sqrt{T}} + \frac{26\delta^2}{KN}. \tag{3.18}
$$

If $\hat{\alpha}_0 \geq \alpha\Big(\frac{C_1}{C_3T}\Big)^{\frac{1}{4}}$, we choose $\hat{\alpha} = \alpha\Big(\frac{C_1}{C_3T}\Big)^{\frac{1}{4}}$. Thus we have

$$
\begin{aligned}
\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] &\leq \frac{3C_1\alpha^2}{2\hat{\alpha}^2T} + \frac{C_3\hat{\alpha}^2}{2\alpha^2} + \frac{26\delta^2}{KN} \\
&= \frac{2(C_1C_3)^{\frac{1}{2}}}{\sqrt{T}} + \frac{26\delta^2}{KN}.
\end{aligned} \tag{3.19}
$$

Combining these two cases, we can obtain

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] &\leq \frac{3C_1\alpha^2}{2\hat{\alpha}_0^2T} + \frac{5(C_1C_3)^{\frac{1}{2}}}{2\sqrt{T}} + \frac{52\delta^2}{KN} \\
&\leq \frac{3C_1\alpha^2}{2\hat{\alpha}_0^2T} + \frac{80\sqrt{26\delta^2{L_G}^2(\rho_\Omega^2/K)}}{\sqrt{KNRT}} + \frac{52\delta^2}{KN}
\end{aligned} \tag{3.20}
$$

As regard to the relationship between the personalized models and the global models, we adopt the process of the corresponding proof in [95], and can get that

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\tilde{\theta}_i^t-\omega_j^t\big\|^2\Big] \leq \mathcal{O}\Big(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]\Big)+\mathcal{O}\Big(\frac{\delta_G^2}{\lambda^2}+\delta^2\Big).$$

The complete proof of Theorem 3.6.1 ends. □

Remark 3.6.1. *Theorem 3.6.1 shows that the proposed CGPFL can achieve a convergence rate of $\mathcal{O}\big(1/\sqrt{KNRT}\big)$, which is $\mathcal{O}(\sqrt{K})$ times faster than what most of the state-of-the-art works [49, 20, 79] achieved (i.e., $\mathcal{O}\big(1/\sqrt{NRT}\big)$) in non-convex federated learning setting.*

3.6.2 Generalization Error Bound

We analyse the generalization error of CGPFL in this section. Before starting the analysis, we first introduce two important definitions as follows. For simplicity, we define $\mathcal{L}_D(h)=\mathbb{E}_{(x,y)\in D}[\ell(h(x),y)]$ where ℓ is the adopted loss function in this section.

Definition 3.6.2 (Complexity)**.** *Let $\mathcal{H}$ be a hypothesis class (corresponding to $\omega\in\mathbb{R}^d$ in neural network), and $|D|$ be the size of dataset D, the complexity of $\mathcal{H}$ can be expressed by the maximum disagreement between two hypotheses on a dataset D:*

$$\lambda_{\mathcal{H}}(D)=\sup_{h_1,h_2\in\mathcal{H}}\frac{1}{|D|}\sum_{(x,y)\in D}|h_1(x)-h_2(x)|. \tag{3.21}$$

Definition 3.6.3 (Label-discrepancy)**.** *Consider a hypothesis class $\mathcal{H}$, the label-discrepancy between two data distributions D_1 and D_2 is given by:*

$$disc_{\mathcal{H}}(D_1,D_2)=\sup_{h\in\mathcal{H}}|\mathcal{L}_{D_1}(h)-\mathcal{L}_{D_2}(h)|, \tag{3.22}$$

where $\mathcal{L}_D(h)=\mathbb{E}_{(x,y)\in D}[\ell(h(x),y)]$.

Theorem 3.6.2 (Generalization error bound of CGPFL)**.** *When Assumption 3.6.1 is satisfied, with probability at least $1-\delta$, the following holds:*

$$\sum_{i=1}^{N}\frac{m_i}{m}\Big\{\mathcal{L}_{D_i}(\hat{h}_i^*)-\min_{h\in\mathcal{H}}\mathcal{L}_{D_i}(h)\Big\}\leq 2\sqrt{\frac{\log\frac{N}{\delta}}{m}}+\sqrt{\frac{dK}{m}\log\frac{em}{d}}+(\lambda+\frac{L}{2})cost(\Theta^*,\Omega^*;K)$$
$$+\sum_{i=1}^{N}\frac{m_i}{m}\big\{2B\,\lambda_{\mathcal{H}}(D_i)+disc(D_i,\widetilde{D}_i)\big\},$$

where B is a positive constant with $\big|\mathcal{L}_D(h_1)-\mathcal{L}_D(h_2)\big|\leq B\,\lambda_{\mathcal{H}}(D)$, $\forall h_1,\, h_2\in\mathcal{H}$. Besides, $\hat{h}_i^*$ is given by $\hat{h}_i^*=\arg\min_{\theta_i}\big\{\mathcal{L}_{\tilde{D}_i}(h(\theta_i))+\frac{\lambda}{2}\|\theta_i-\omega_k^*\|^2\big\}$ and $cost(\Theta^*,\Omega^*;K)=\sum_{i=1}^{N}\frac{m_i}{m}\min_{k\in[K]}\|\theta_i^*-\omega_k^*\|^2$.

Proof. Before we start the proof of the generalization bound, we first give some definitions which will be used in the following proof.

$$\begin{aligned}
&h=h(\theta),\quad g=g(\omega)\\
&\hat{h}_i^*=\hat{h}_i(\theta_i^*)=\arg\min_{\theta_i}\big\{\mathcal{L}_{\tilde{D}_i}\big(h(\theta_i)\big)+\frac{\lambda}{2}\|\theta_i-\omega_k^*\|^2\big\}\\
&h_i^*=h_i(\theta_i^*)=\arg\min_{\theta_i}\big\{\mathcal{L}_{D_i}\big(h(\theta_i)\big)+\frac{\lambda}{2}\|\theta_i-\omega_k^*\|^2\big\}\\
&\hat{h}_{i,loc}^*=\hat{h}_{i,loc}(\theta_{i,loc}^*)=\arg\min_{\theta_{i,loc}}\big\{\mathcal{L}_{\tilde{D}_i}\big(h(\theta_{i,loc})\big)\big\}\\
&h_{i,loc}^*=h_{i,loc}(\theta_{i,loc}^*)=\arg\min_{\theta_{i,loc}}\big\{\mathcal{L}_{D_i}\big(h(\theta_{i,loc})\big)\big\}
\end{aligned}\tag{3.23}$$

In this way, we can bound the generalization error of the obtained personalized model θ_i^*, $i\in[N]$ as follows:

$$\begin{aligned}
&\sum_{i=1}^{N}\frac{m_i}{m}\Big\{\mathcal{L}_{D_i}(\hat{h}_i^*)-\min_{h\in\mathcal{H}}\mathcal{L}_{D_i}(h)\Big\}\\
&=\sum_{i=1}^{N}\frac{m_i}{m}\Big\{\mathcal{L}_{D_i}(\hat{h}_i^*)-\mathcal{L}_{D_i}(h_{i,loc}^*)\Big\}\\
&=\sum_{i=1}^{N}\frac{m_i}{m}\Big\{\mathcal{L}_{D_i}(\hat{h}_i^*)-\mathcal{L}_{D_i}(\hat{g}_k^*)+\mathcal{L}_{D_i}(\hat{g}_k^*)-\mathcal{L}_{\tilde{D}_i}(\hat{g}_k^*)+\mathcal{L}_{\tilde{D}_i}(\hat{g}_k^*)\\
&\quad-\mathcal{L}_{\tilde{D}_i}(\hat{h}_i^*)+\mathcal{L}_{\tilde{D}_i}(\hat{h}_i^*)-\mathcal{L}_{D_i}(\hat{h}_i^*)+\mathcal{L}_{D_i}(\hat{h}_i^*)-\mathcal{L}_{D_i}(h_{i,loc}^*)\Big\}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \frac{m_i}{m} \Big\{ \mathcal{L}_{D_i}(\hat{g}_k^*) - \mathcal{L}_{\tilde{D}_i}(\hat{g}_k^*) \Big\} + \sum_{i=1}^{N} \frac{m_i}{m} \Big\{ \mathcal{L}_{\tilde{D}_i}(\hat{g}_k^*) - \mathcal{L}_{\tilde{D}_i}(\hat{h}_i^*) \Big\} \\
&\quad + \sum_{i=1}^{N} \frac{m_i}{m} \Big\{ \mathcal{L}_{D_i}(\hat{h}_i^*) - \mathcal{L}_{D_i}(\hat{g}_k^*) \Big\} + \sum_{i=1}^{N} \frac{m_i}{m} \Big\{ \mathcal{L}_{\tilde{D}_i}(\hat{h}_i^*) - \mathcal{L}_{D_i}(h_{i,loc}^*) \Big\}
\end{aligned}
$$

The above function is divided into four parts. In the following section, we will bound them sequentially. To deal with the first part, we define that $k = \psi(i)$, where $i \in [N]$ and $k \in [K]$.

$$
\begin{aligned}
&\sum_{i=1}^{N} \frac{m_i}{m} \Big\{ \mathcal{L}_{D_i}(\hat{g}_k^*) - \mathcal{L}_{\tilde{D}_i}(\hat{g}_k^*) \Big\} \\
&\leq \max_{g_1,\cdots,g_K} \sum_{i=1}^{N} \frac{m_i}{m} \max_{\psi(i)} \Big\{ \mathcal{L}_{D_i}(\hat{g}_{\psi(i)}^*) - \mathcal{L}_{\tilde{D}_i}(\hat{g}_{\psi(i)}^*) \Big\} \\
&\leq \max_{\psi} \max_{g_1,\cdots,g_K} \sum_{i=1}^{N} \frac{m_i}{m} \Big\{ \mathcal{L}_{D_i}(\hat{g}_{\psi(i)}^*) - \mathcal{L}_{\tilde{D}_i}(\hat{g}_{\psi(i)}^*) \Big\}
\end{aligned}
$$

Since the results of k-means++ depend on the selection of the first initialization centroid, the possible number of clustering results is N. By the McDiarmid's inequality, with probability at least $1-\delta$, we have

$$
\begin{aligned}
&\max_{g_1,\cdots,g_K} \sum_{i=1}^{N} \frac{m_i}{m} \Big\{ \mathcal{L}_{D_i}(\hat{g}_{\psi(i)}^*) - \mathcal{L}_{\tilde{D}_i}(\hat{g}_{\psi(i)}^*) \Big\} \\
&\leq \mathbb{E}\Big[\max_{g_1,\cdots,g_K} \sum_{i=1}^{N} \frac{m_i}{m} \Big(\mathcal{L}_{D_i}(\hat{g}_{\psi(i)}^*) - \mathcal{L}_{\tilde{D}_i}(\hat{g}_{\psi(i)}^*) \Big) \Big] + 2\sqrt{\frac{\log \frac{N}{\delta}}{m}}
\end{aligned}
$$

Utilizing the results in [69], we can get

$$
\begin{aligned}
&\mathbb{E}\Big[\max_{g_1,\cdots,g_K} \sum_{i=1}^{N} \frac{m_i}{m} \Big(\mathcal{L}_{D_i}(\hat{g}_{\psi(i)}^*) - \mathcal{L}_{\tilde{D}_i}(\hat{g}_{\psi(i)}^*) \Big) \Big] \\
&\leq \frac{1}{m} \mathbb{E}\Big\{ \sum_{k=1}^{K} \max_{g_k} \Big[m_{C_k} \Big(\mathcal{L}_{D_{C_k}}(g_k) - \mathcal{L}_{\tilde{D}_{C_k}}(g_k) \Big) \Big] \Big\} \\
&\leq \sum_{k=1}^{K} \frac{m_{C_k}}{m} \mathfrak{R}_{D_{C_k}, m_{C_k}}(\mathcal{H}) \leq \sqrt{\frac{dK}{m} \log \frac{em}{d}}
\end{aligned}
$$

Therefore, we can get

$$
\sum_{i=1}^{N} \frac{m_i}{m} \Big\{ \mathcal{L}_{D_i}(\hat{g}_k^*) - \mathcal{L}_{\tilde{D}_i}(\hat{g}_k^*) \Big\} \leq 2\sqrt{\frac{\log \frac{N}{\delta}}{m}} + \sqrt{\frac{dK}{m} \log \frac{em}{d}}. \tag{3.24}
$$

When Assumption 1 is satisfied, we know that $\mathcal{L}_{\tilde{D}_i}(h(\omega))$ is L-Lipschitz smooth. Thus, we have

$$\mathcal{L}_{\tilde{D}_i}(\hat{g}_k^*) - \mathcal{L}_{\tilde{D}_i}(\hat{h}_i^*) \le \left\langle \nabla \mathcal{L}_{\tilde{D}_i}(\hat{h}_i(\theta_i^*)), \omega_k^* - \theta_i^* \right\rangle + \frac{L}{2}\left\|\theta_i^* - \omega_k^*\right\|^2 \tag{3.25}$$

Because $\hat{h}_i^*(\theta_i^*)$ is obtained by solving $h_i(\theta_i^*) = \arg\min_{\theta_i} \left\{\mathcal{L}_{D_i}\big(h(\theta_i)\big) + \frac{\lambda}{2}\|\theta_i - \omega_k^*\|^2\right\}$, we can get that $\nabla \mathcal{L}_{\tilde{D}_i}(\hat{h}_i(\theta_i^*)) + \lambda(\theta_i^* - \omega_k^*) = 0$, that is $\nabla \mathcal{L}_{\tilde{D}_i}(\hat{h}_i(\theta_i^*)) = -\lambda(\theta_i^* - \omega_k^*)$ Thus, we have

$$\sum_{i=1}^{N} \frac{m_i}{m}\left\{\mathcal{L}_{\tilde{D}_i}(\hat{g}_k^*) - \mathcal{L}_{\tilde{D}_i}(\hat{h}_i^*)\right\} \le (\lambda + \frac{L}{2}) \sum_{i=1}^{N} \frac{m_i}{m}\left\|\theta_i^* - \omega_k^*\right\|^2 \tag{3.26}$$

Finally, according to the definitions of Complexity and Label-discrepancy, we can know that

$$\begin{aligned}
&\sum_{i=1}^{N} \frac{m_i}{m}\left\{\mathcal{L}_{D_i}(\hat{h}_i^*) - \mathcal{L}_{D_i}(\hat{g}_k^*)\right\} + \sum_{i=1}^{N} \frac{m_i}{m}\left\{\mathcal{L}_{\tilde{D}_i}(\hat{h}_i^*) - \mathcal{L}_{D_i}(h_{i,loc}^*)\right\} \\
&\le 2B \sum_{i=1}^{N} \frac{m_i}{m}\, \lambda_{\mathcal{H}}(D_i) + \sum_{i=1}^{N} \frac{m_i}{m}\, disc(D_i, \tilde{D}_i) \\
&= \sum_{i=1}^{N} \frac{m_i}{m}\left\{2B\, \lambda_{\mathcal{H}}(D_i) + disc(D_i, \tilde{D}_i)\right\}
\end{aligned}$$

where the constant B satisfies that $\left|\mathcal{L}_D(h_1) - \mathcal{L}_D(h_2)\right| \le B\, \lambda_{\mathcal{H}}(D)$ for $h_1,\, h_2 \in \mathcal{H}$. Summarizing the obtained results, we can get

$$\begin{aligned}
\sum_{i=1}^{N} \frac{m_i}{m}\left\{\mathcal{L}_{D_i}(\hat{h}_i^*) - \min_{h\in\mathcal{H}} \mathcal{L}_{D_i}(h)\right\} \le 2\sqrt{\frac{\log \frac{N}{\delta}}{m}} + \sqrt{\frac{dK}{m} \log \frac{em}{d}} + (\lambda + \frac{L}{2}) \sum_{i=1}^{N} \frac{m_i}{m}\left\|\theta_i^* - \omega_k^*\right\|^2 \\
+ \sum_{i=1}^{N} \frac{m_i}{m}\left\{2B\, \lambda_{\mathcal{H}}(D_i) + disc(D_i, \tilde{D}_i)\right\}
\end{aligned}$$

The complete proof of Theorem 3.6.2 ends. □

Remark 3.6.2. *Theorem 3.6.2 gives the generalization error bound of CGPFL. When $K = 1$, it yields the error bound of PFL with single global model [57, 95, 33, 32]. As the number of contexts increases, the second terms become larger, while the last term get smaller. Hence, our CGPFL can always reach better personalization-generalization trade-off by adjusting the number of contexts K, and further achieve higher accuracy than the existing PFL methods.*

3.7 An Heuristic Improvement: CGPFL-Heur

As discussed, Theorem 4.2 indicates that there exists a optimal K^* ($K^* \in [K]$) to achieve the minimal generalization error bound that corresponds to the highest model accuracy. Theoretically, the optimal K^* can be obtained by minimizing the generalization bound in Theorem 4.2. We can find that the first and the third term have no relationship with the number of latent contexts, that is, they are irrelevant to K. Therefore, we can obtain an optimal K^* by minimizing the following expression:

$$e(K) := \sqrt{\frac{dK}{m} \log \frac{em}{d}} + \mu \cdot cost(\Theta^*, \Omega^*; K), \tag{3.27}$$

where μ is a hyper-parameter which is induced by the unknown constant L. The above objective can be solved in the server along with the clustering. In the down-to-earth experiments, we notice that the latent context structure can be learned efficiently in the first few rounds. Based on this observation, we believe that CGPFL-Heur can efficiently figure out a near-optimal solution $\hat{K}$ by operating the solver of (3.27) only in the first few rounds (in the experimental part, we only operate the solver in the first global round), and after that, the obtained $\hat{K}$ will no longer be updated. In this way, CGPFL-Heur can reach a near-optimal trade-off (corresponding to the near-optimal $\hat{K}$) between generalization and personalization with negligible additional computation in the server. Moreover, in view of the fact that we only need to operate the solver in the first few rounds, CGPFL-Heur can retain the same convergence rate as CGPFL.

3.8 Experiments

3.8.1 Experimental Setup

Dataset Setup: Three datasets including MNIST [52], CIFAR10 [51], and Fashion-MNIST (FMNIST) [101] are used in our experiments. To generate Non-I.I.D. datasets for the clients, we split the whole dataset as follows. 1) MNIST: we distribute the train-set containing $60,000$ digital instances into 40 clients, and each of them is only provided with 3 classes out of total 10. The number of instances obtained by each client is randomly chosen from the range of $[400, 5000]$, of which 75% are used for training and the remaining 25% for testing. 2) CIFAR10: We distribute the whole dataset containing $60,000$ instances into 40 clients, and each of them is also provided with 3 classes out of total 10. The number of instances obtained by each client is randomly chosen from the range of $[400, 5000]$. The train/test split remains 75%/25%. 3) Fashion-MNIST: It's a more challenging replacement of MNIST, and the Non-I.I.D. splitting is the same as MNIST.

Baseline Methods: We compare our CGPFL and CGPFL-Heur with seven state-of-the-art works: one traditional FL method, *FedAvg* [70]; one typical cluster-based FL method, *IFCA* [27]; and five most recent PFL models, *APFL* [20], *Per-FedAvg* [23], *L2SGD* [33], *pFedMe* [95], and *Ditto* [57].

Model Architectures: 1) For strongly convex case, we use a l_2-regularized multinomial logistic regression model (MLR) with the softmax layer and cross-entropy loss, in line with [95]; 2) For the non-convex case, we apply a neural network (DNN) with one hidden layer of size 128 and a softmax layer at the end for evaluation. In addition, we apply a CNN that has two convolutional layers and two fully connected layers for the CIFAR10. All competitors and our algorithms are based on the same configurations and fine-tuned to their best performances.

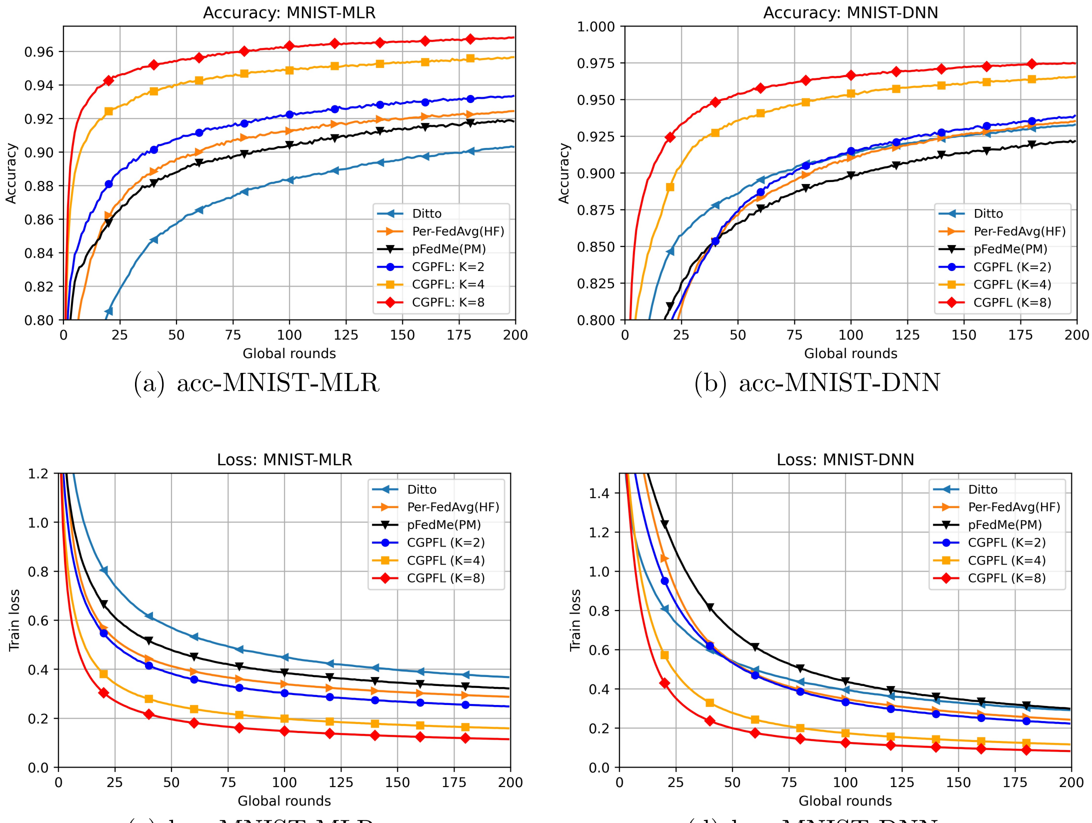
3.8.2 Overall Performance

The comprehensive comparison results of our CGPFL and CGPFL-Heur are shown in Table 3.1. It can be observed that our methods outperform the competitors with large margins for both non-convex and convex cases on all datasets, even if *IFCA* works with a good initialization. Besides, although we only provide the proof of convergence rate under non-convex case, as shown in Figure 3.3 and Figure 3.4,

Table 3.1: Comparison of test accuracy. We set $N = 40$, $\alpha = 1$, $\lambda = 12$, $S = 5$, $lr = 0.005$ and $T = 200$ for MNIST and Fashion-MNIST (FMNIST), and $T = 300$, $lr = 0.03$ for CIFAR10, where lr denotes the learning rate.

Method	MNIST		FMNIST		CIFAR10
	MLR	DNN	MLR	DNN	CNN
FedAvg [70]	88.63	91.05	82.44	83.45	46.34
IFCA ($K = 4$) [27]	95.27	96.19	91.55	92.56	60.22
L2SGD [33]	89.46	92.48	88.59	90.64	58.68
APFL [20]	92.69	95.59	92.60	93.76	72.12
pFedMe (PM) [95]	91.90	92.20	85.49	86.87	68.88
Per-FedAvg (HF) [23]	92.44	93.54	87.17	87.57	71.46
Ditto [57]	89.96	92.85	88.62	90.56	69.56
CGPFL (K = 4)	95.65	96.55	92.65	93.56	72.78
CGPFL-Heur	**97.41**	**98.03**	**95.18**	**96.00**	**74.75**

the extensive experiments further demonstrate that our methods constantly obtain better performance against multiple state-of-the-art PFL methods (*pFedMe*, *Ditto*, and *Per-FedAvg*) with faster convergence rate under both strongly-convex and non-convex cases. Specifically, the figures in Figure 3.3 show the results for MNIST dataset on MLR and DNN model, while the figures in Figure 3.4 give the results for Fashion-MNIST dataset on MLR and DNN model.



(a) acc-MNIST-MLR

(b) acc-MNIST-DNN

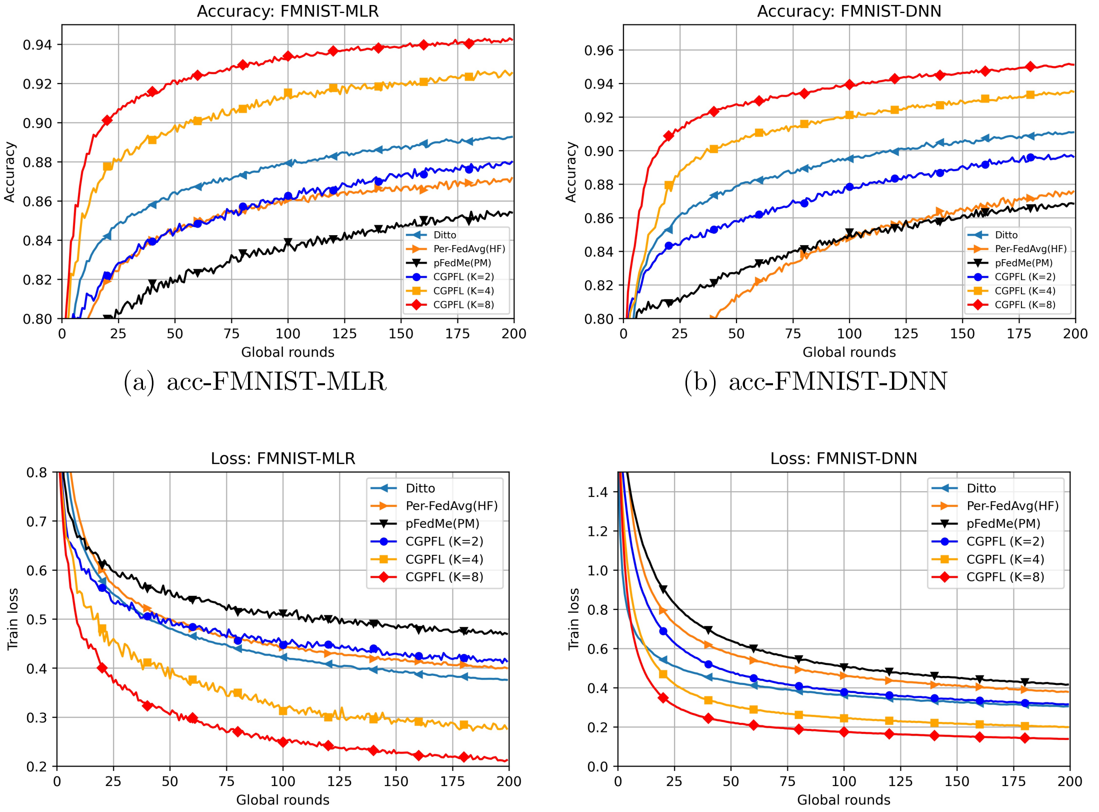
(c) loss-MNIST-MLR

(d) loss-MNIST-DNN

Figure 3.3: Performance on MNIST for different K with $N = 40$, $\alpha = 1$, $\lambda = 12$, $R = 10$, and $S = 5$.

3.8.3 Further Evaluation on CGPFL-Heur

To further evaluate the performance of CGPFL-Heur, on the one hand, we conduct the CGPFL training with different number of contexts (i.e., K) varying form 1 to $N/2$



(a) acc-FMNIST-MLR

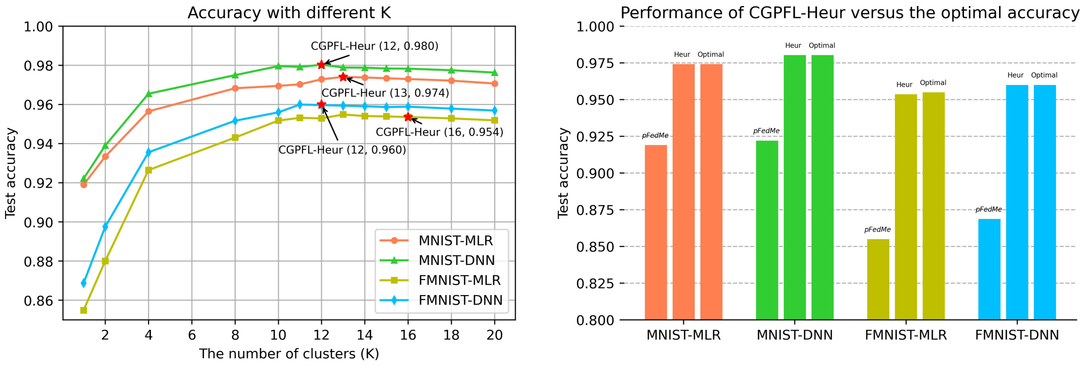
(b) acc-FMNIST-DNN

(c) loss-FMNIST-MLR

(d) loss-FMNIST-DNN

Figure 3.4: Performance on FMNIST for different K with $N = 40$, $\alpha = 1$, $\lambda = 12$, $R = 10$, and $S = 5$.

on MINST and FMNIST, respectively. In particular, we set the maximal value of K no more than $N/2$ to avoid overfitting. By collating the model accuracy with different K, we can find out the optimal K which corresponds to the optimal personalization-generalization trade-off in CGPFL. The results are demonstrated in Figure 3.5(a). On the other hand, we conduct the CGPFL-Heur training with an appropriate μ and keep other parameters same as that of the above evaluation. As shown in Figure 3.5(a), we distinguish the results of CGPFL-Heur using red-star points. Besides, we make comparisons between the performance of a state-of-the-art PFL algorithm, *pFedMe* [95] with our proposed CGPFL and CGPFL-Heur in Figure 3.5(b). The results in Figure 3.5(a) and Figure 3.5(b) demonstrate that our designed heuristic algorithm CGPFL-Heur can effectively reach a near-optimal trade-off and consequently achieve the near-optimal model accuracy.



(a) CGPFL with variable K

(b) CGPFL-Heur

Figure 3.5: Further evaluation for CGPFL-Heur on MNIST and FMNIST datasets

3.8.4 Effects of Balancing Weight

As mentioned that the hyper-parameter λ can balance the weight of personalization and generalization in several state-of-the-art PFL algorithms [95, 32, 57], we also conduct experiments to compare the performance of our CGPFL and CGPFL-Heur with a state-of-the-art PFL algorithm, *pFedMe* [95], on different values of λ. Specifically, the range of λ is properly chosen as in Table 3.2 to avoid the divergence in *pFedMe*. The experimental results in Table 3.2 show that our methods can constantly achieve better performance than *pFedMe* when λ varies, which indicates a better trade-off between personalization and generalization in PFL.

Table 3.2: Comparisons with various λ. We set $N = 40$, $\beta = 1$, $R = 10$, $S = 5$, $lr = 0.005$ and $T = 200$ for MNIST and Fashion-MNIST (FMNIST), where lr denotes the learning rate.

MNIST-MLR	λ	11	12	13	14	15	16	17	18	19
	pFedMe (PM)	91.46	91.90	92.19	92.54	92.80	93.00	93.15	93.16	93.04
	CGPFL (K=2)	93.43	93.34	93.62	93.88	94.16	93.69	93.52	93.52	93.31
	CGPFL (K=4)	95.49	95.65	95.19	95.47	95.60	95.77	96.49	94.85	94.53
	CGPFL-Heur	97.46	97.41	96.27	96.32	96.34	96.33	96.32	96.33	96.25
MNIST-DNN	λ	9	10	11	12	13	14	15	16	17
	pFedMe (PM)	91.21	91.54	91.86	92.21	92.43	92.79	93.05	93.30	93.24
	CGPFL (K=2)	94.11	94.42	94.71	93.90	94.14	94.36	94.49	93.34	93.36
	CGPFL (K=4)	96.17	96.37	96.57	96.55	95.87	95.99	96.01	95.45	95.49
	CGPFL-Heur	97.69	97.86	98.00	98.03	97.95	97.96	98.20	98.14	98.16
FMNIST-MLR	λ	9	10	11	12	13	14	15	16	17
	pFedMe (PM)	85.03	85.26	85.42	85.49	85.49	85.28	85.16	84.76	84.22
	CGPFL (K=2)	90.29	87.70	87.93	88.00	87.72	87.53	87.65	86.94	85.19
	CGPFL (K=4)	92.50	92.84	92.94	92.65	92.63	92.44	92.42	92.17	92.20
	CGPFL-Heur	95.46	95.44	95.45	95.36	94.61	94.40	94.35	94.41	94.19
FMNIST-DNN	λ	7	8	9	10	11	12	13	14	15
	pFedMe (PM)	84.65	85.20	85.86	86.28	86.70	86.87	87.09	87.10	86.66
	CGPFL (K=2)	87.69	88.15	88.72	89.13	89.59	89.75	91.15	89.25	88.93
	CGPFL (K=4)	92.26	92.89	92.71	92.86	93.03	93.56	93.21	93.44	92.83
	CGPFL-Heur	95.60	95.73	95.84	95.94	95.98	96.00	95.98	95.95	95.83

3.9 Remarks

In this work, we propose a novel personalized federated learning framework, dubbed CGPFL, to handle the challenge of statistical heterogeneity (Non-I.I.D.), especially contextual heterogeneity in the federated setting. To the best of our knowledge, we are the first to propose the concept of contextualized generalization (CG) for personalized federated learning and further formulate it to a bi-level optimization problem that is solved effectively. Our method provides fine-grained generalization for personalized models which can prompt higher test accuracy and facilitate faster model convergence. Experimental results on real-world datasets demonstrate the effectiveness of our method over the state-of-the-art works.

Although both empirical and theoretical results demonstrate the effectiveness of the proposed CGPFL in addressing data heterogeneity in federated learning systems, CGPFL is specifically tailored to tackle the train–train data distribution shift across federated clients. However, addressing the train–test data distribution shift on each client, when a train–train data distribution shift simultaneously exists within the federated learning system, remains an open problem. This challenging issue will be further investigated in the next chapter.

Chapter 4

Learning Personalized Causally Invariant Representations for Heterogeneous Federated Clients

4.1 Introduction

Federated learning (FL) allows the participation of a massive number of data holders (i.e., clients) that possess limited data to collaboratively train learning models in a privacy-preserving manner [70]. From the view of the heterogeneity of target datasets across local clients, we can divide the literature on FL into two branches. 1) Federated learning aims at training a global model to fit the local data distributions and perform well when the local target datasets are subject to independent and identically distribution (IID). In particular, some works [21, 63, 73] (including robust federated learning and federated domain generalization) focus on training a global model that can tackle the distribution/domain shift across local training datasets. Unfortunately, the shared global model can diverge from the optimal local solutions when the target datasets are heterogeneous or not IID (i.e., Non-IID) across local clients [39], since the useful information about personalization is dropped. 2) Personalized federated learning develops a personalized model for each client to handle the discrepancy among the local optima when the target datasets across local clients are Non-IID. Despite succeeding in handling Non-IID target datasets, all the existing PFL methods neglect the shortcut trap problem which is attracting more and more interest in centralized machine learning.

Shortcut trap is found pervasive in modern machine learning [26] where models prefer to rely on the shortcut to solve problems due to the bias of training dataset. The utilized shortcut can perform well on training data but fails to generalize to unseen test data that is out-of-distribution (OOD) with respect to the training data. For example, there is a binary image classification task where the model needs to recognize the pictures of cows and camels [7]. Deep learning model can classify the picture of a cow in a desert background as "camel" at test time, if most of cows appear in grass backgrounds and most of camels stand in desert backgrounds in training environments (environments are data subsets that have different data distributions). This dataset bias makes the obtained model choose the background rather than the shape of animals in the pictures as the discriminative feature. The similar shortcut trap exists in diverse real-world scenarios [26]. Although many efforts have been attracted to the shortcut trap problem in centralized situations, they focus on mitigating shortcut by extracting environment-invariant (a.k.a. invariant) features. When applying these schemes into PFL, the invariance constraint will eliminate all heterogeneous features, including shortcut and personalized features. Therefore, the existing invariant learning schemes can hardly tackle the shortcut trap problem in PFL.

What's worse, we find the trivial combination of the existing PFL and centralized invariant learning schemes, instead of solving the shortcut trap problem in PFL, can even induce worse performance than the better one of themselves (discussed in the evaluation part). To handle the challenging shortcut trap problem in PFL, we firstly formulate the structural causal models (SCMs) to simulate the heterogeneous data

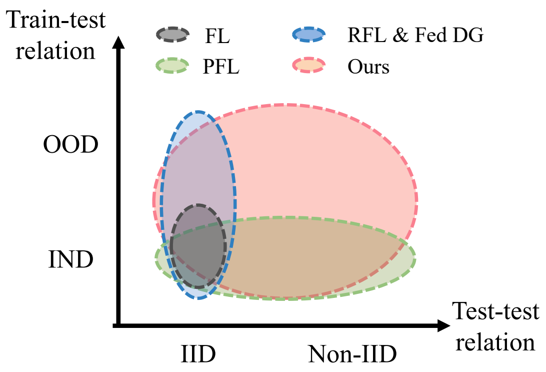


Figure 4.1: The coverage of FedSDR and the related works. a) FL: **F**ederated **L**earning; b) PFL: **P**ersonalized **F**ederated **L**earning; c) RFL: **R**obust **F**ederated **L**earning; d) Fed DG: **Fed**erated **D**omain **G**eneralization. Besides, IND denotes in-distribution.

generating processes on local clients. From the SCMs, we derive a causal signature which reveals that the shortcut is statistical independent with the client/user indicator conditional on label and environment indicator. Inspired by this finding, we design a collaborative shortcut discovery method which can work well even if there is only one available training environment on each client. Then, the personalized causally invariant representations are extracted by utilizing another causal signature that describes the conditional independence between the personalized invariant features and the shortcut features. Finally, the optimal personalized invariant predictors can be elicited from the extracted personalized causally invariant features. The comparison between the coverage of our approach and the related works is illustrated in Figure 4.1. The main contributions of this work are summarized as follows:

- To the best of our knowledge, we are the first to consider the shortcut trap problem in personalized federated learning and analyse it by formulating the structural causal models for heterogeneous clients. Based on the proposed SCMs, we design a provable shortcut discovery and removal method to develop the optimal personalized invariant predictor which can generalize to unseen local test distribution for each client.

- Theoretically, we demonstrate that the designed shortcut discovery method can draw all the latent shortcut components, then the shortcut removal method can eliminate the discovered shortcut features and produce the optimal personalized invariant predictor for each client.
- Empirically, we conduct experiments on several commonly used out-of-distribution datasets and the results validate the superiority of our method on out-of-distribution generalization performance, compared with the state-of-the-art competitors.

4.2 Related Work

Federated learning. The classic FedAvg [70] performs well if local training datasets are IID. Some methods ([48, 22, 110, 29]) mitigate the negative impact of training data heterogeneity on convergence rate, while another branch ([21, 87, 93]) targets at reducing the performance bias of global model on local clients. Besides, few works ([63, 73, 30]) investigate the scenarios where the training data heterogeneity appears to be domain shift. All the above methods produce a shared global model which can diverge from the local optimal solutions when local target datasets are Non-IID.

Personalized federated learning. Many PFLs ([94, 32, 24, 57, 96, 17, 28]) train the personalized models with the guidance of a global model which embeds in the shared knowledge. Some researchers study the parameterized knowledge transfer between similar clients, e.g., MOCHA [90], FedAMP [41] and KT-pFL [109]. DFL [66] disentangles the shared features from the client-specific ones to achieve accurate aggregation on shared knowledge. Similarly, pFedPara [43] and Factorized-FL [44] factorizes the model parameters into the shared and personalized parts. Another branch

([18, 12, 102]) employs the shared/aligned feature extractor to capture global knowledge and personalized classifiers to encode the personalization information. All of them don't cover the situations where there exists shortcut in local training datasets.

Shortcut and Invariant learning (IL). Causally invariant predictor is proposed in [76], and then applied into deep learning in IRM [6] to mitigate shortcut. Subsequently, [80] prove that IRM and its variants can be still trapped by shortcut when training environments are insufficient. IFM [15] lowers the requirement and demands only logarithmic training environments. Some works focus on settling IL problem when the environment label is unavailable, e.g., EIIL [19], HRM [61, 62], EDNIL [40] and ZIN [59]. Another branch ([1, 14, 42]) completes the constraints that IRM misses to improve the performance. The iCaRL [65] extends IL to non-linear causal representations while ACTIR [45] extends IL to anti-causal scenarios. All these methods are devised for centralized scenarios where all training data is accessed and training environments are sufficient.

4.3 Problem Formulation

Notations. Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{E}$ denote the input, target and environment space respectively. Data instance is $(X, y, e) \in (\mathcal{X}, \mathcal{Y}, \mathcal{E})$. Suppose there are N clients and the local dataset D_u on client u contains M_u samples, $u \in [N]$. The sets of training and test environments on client u are denoted by $\mathcal{E}_{tr}^u$ and $\mathcal{E}_{te}^u$ respectively. We use $\mathcal{E}_{all}^u$ as the set of all possible environments in the task that client u concentrates on, i.e., $\mathcal{E}_{tr}^u$, $\mathcal{E}_{te}^u \subset \mathcal{E}_{all}^u$, $\forall u \in [N]$. In federated learning system, the overall environment sets are denoted by $\mathcal{E}_{tr} := \bigcup_u \mathcal{E}_{tr}^u$ and $\mathcal{E}_{all} := \bigcup_u \mathcal{E}_{all}^u$. For convenience, we separate the learning model or parameterized mapping from $\mathcal{X}$ to $\mathcal{Y}$ into two consecutive parts: **1)** the feature extractor (Φ and Ψ denote the invariant and spurious feature extractors respectively) maps from input space $\mathcal{X}$ to latent feature space $\mathcal{Z}$, i.e., $\Phi(X) \in \mathcal{Z}$ and

$\Psi(X) \in \mathcal{Z}$; **2)** the classifier ω outputs a prediction $\hat{y}$ from a latent feature $z \in \mathcal{Z}$. For example, the overall model based on the invariant feature extractor is denoted by $f_\theta(\cdot) = \omega(\Phi(\cdot))$ where f_θ indicates the function f parameterized by θ. We define the expected empirical loss for model f_θ on dataset D as $\mathcal{R}(f_\theta; D) := \mathbb{E}_{(X,y)\in D}[\ell(f_\theta(X), y)]$ where ℓ is the loss function.

4.3.1 Invariant Learning

Succeeding in mitigating shortcut and solving the OOD generalization problem, invariant learning assumes that there exists some invariant feature $\Phi(X)$ satisfying the ***invariance constraint***:

$$\mathbb{P}(Y|\Phi(X) = z, e) = \mathbb{P}(Y|\Phi(X) = z, e'), \forall z \in \mathcal{Z}, \forall e, e' \in \mathcal{E}_{all}. \tag{4.1}$$

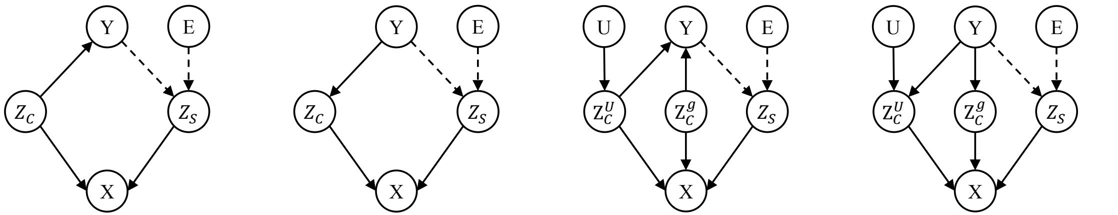
[80] proved that IRM [6] and its variants need at least $d_S + 1$ (d_S is the dimension of shortcut features) training environments to eliminate all shortcut features and elicit the optimal invariant predictor, under linear scenarios.

Definition 4.3.1 (Optimal Invariant Predictor)**.** *The optimal invariant predictor is elicited based on the complete invariant features which are informative for the target label in the task concerned, i.e.,* $\Phi^\star = \arg\max_\Phi I(Y; \Phi(X))$*, where* $I(\cdot;\cdot)$ *denotes the Shannon mutual information between two random variables and* Φ *satisfies the above invariance constraint.*

4.3.2 Causal Setup

In invariant learning (IL), researchers usually formulate a structural causal model to simulate the data generating process in the target task. A valid SCM is depicted by a directed acyclic graph where each node represents a random variable and each edge describe a directed functional relationship between the corresponding variables [75].

When we study the invariant learning in federated setting, the latent heterogeneity of data generating mechanisms among local clients need to be considered.



(a) Causal IL (b) Anti-causal IL (c) Causal FedSDR (d) Anti-causal FedSDR

Figure 4.2: Graph (a) [6, 40] and (b) [80, 42] give the structural causal models (SCMs) commonly adopted in invariant learning, while (c) and (d) show the SCMs proposed in FedSDR. Z_C and Z_S denote the invariant and shortcut features respectively. E is the indicator of shortcut while U is the indicator of user/client. Dotted arrows indicate unstable causal relations that can vary in different environments.

Therefore, we propose the SCMs in federated learning by adding the **u**ser/client indicator U and deconstructing the invariant features into two separate parts: the personalized invariance Z_C^U and the shared/global invariance Z_C^g. The detailed SCMs are shown in Figure 4.2. As discussed in the literature on invariant learning, Z_S is the latent shortcut feature. The functional relation between Z_S and label Y can vary across different environments. That is, $\forall Z_S$ there always exists some $e, e' \in \mathcal{E}_{all}$ that make $\mathbb{P}(Y|Z_S, e) \neq \mathbb{P}(Y|Z_S, e')$ hold. By analogy with the optimal invariant predictor in invariant learning, we provide the definition of the optimal personalized invariant predictor in PFL.

Definition 4.3.2 (Optimal Personalized Invariant Predictor)**.** *The optimal personalized invariant predictor for client u is elicited based on the complete invariant features which are informative for target label in the task that client u concentrates on. That is, $\Phi_u^\star = \arg\max_{\Phi_u} I(Y; \Phi_u(X))$, where Φ_u satisfies that $\mathbb{P}(Y|\Phi_u(X) = z, e) = \mathbb{P}(Y|\Phi_u(X) = z, e'), \forall z \in \mathcal{Z}, \forall e, e' \in \mathcal{E}_{all}^u$.*

4.4 Methodology

When the training environments on each client are insufficient, locally invariant learning can fail as discussed before. *How about a collaborative manner?* Unfortunately, personalized invariant features can cause deviation of the invariance constraint as shortcut features do if training environments are collected from different clients. As a result, the collaborative invariant learning can eliminate/preserve both personalized invariant and shortcut features with the same probability. *Nonetheless, how about combining collaborative invariant learning with PFL methods?* Even though we can get the global invariant features via collaborative IL and conduct local adaptation as in many PFL schemes (e.g., fine-tuning [17] and L2-regularization [94, 32, 57]), local adaptation can pick up both personalized invariant and shortcut features again since local training environments are insufficient. It turns out the trivial combination can hardly outperform the superior individual one on OOD generalization performance.

FedSDR. In view of the above failure, we turn to the complementary perspective: discovering the shortcut features and removing them instead of straightly constraining invariance. The feasibility of this tack is guaranteed by the causal signatures that we derive from the SCMs in Figure 4.2(c) and 4.2(d).

Lemma 4.4.1. *If the data generating mechanism of each federated client obeys the causal graph in Figure 4.2(c) or the anti-causal graph in Figure 4.2(d), we can have:*

- $Z_S \perp\!\!\!\perp U \mid Y, E$ *which means that the shortcut features* Z_S *are conditionally independent of the personalization indicator* U *given* Y *and* E*.*
- $Z_C^g \perp\!\!\!\perp Z_S \mid Y$ *and* $Z_C^U \perp\!\!\!\perp Z_S \mid Y$*, which means that both the global* (Z_C^g) *and personalized* (Z_C^U) *invariant features are conditionally independent of the shortcut features* Z_S *given* Y*.*

Proof. According to the causal Markov condition (Theorem 1.4.1) proved in [75], we know that the variable Z_S is independent of all its nondescendants, given its parents in the (Markov) causal graph. Since Y and E are the parent variables of Z_S and U is a nondescendant of Z_S, the first causal signature in Lemma 4.4.1 is guaranteed. Besides, based on the d-separation criterion in [75] we can find the variable Y d-separates Z_C^g from Z_S and d-separates Z_C^U from Z_S in the SCMs. Therefore, we get the second causal signature in Lemma 4.4.1. □

Remark 4.4.1. *The first causal signature in Lemma 4.4.1 indicates that we can discover the shortcut features using training environments across local clients even if the data generating mechanisms are heterogeneous among them. The second causal signature makes it possible to develop the optimal personalized invariant predictors with the discovered shortcut features even though there is just one training environment on each client, since the relationships between Z_C^g, Z_C^U and Z_S are independent of environment E.*

In the following sections, we will introduce the two-stage implementation of our method in detail. The overall framework of FedSDR is given in Figure 4.3.

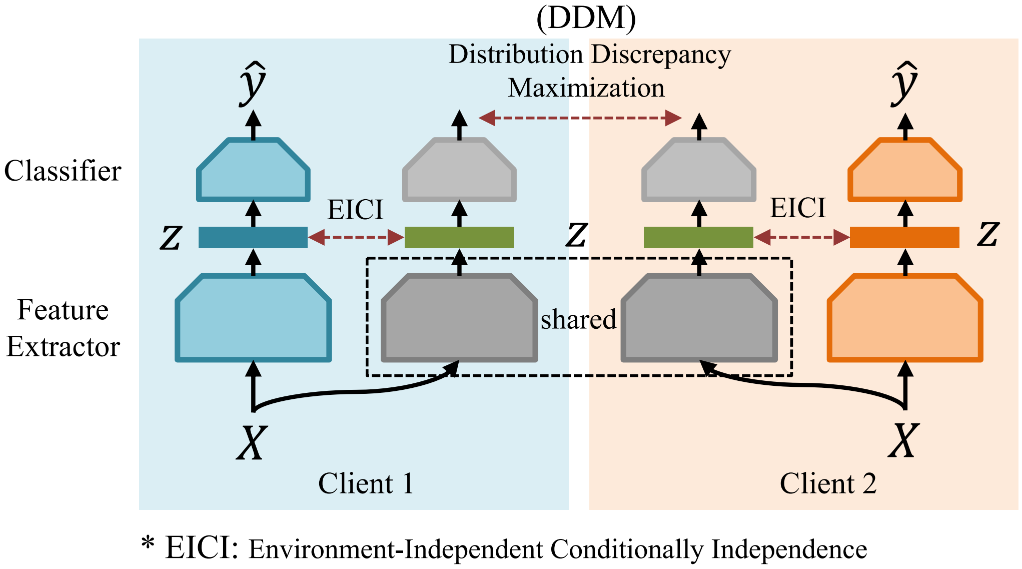

(DDM)
Distribution Discrepancy Maximization
$\hat{y}$
$\hat{y}$
Classifier
EICI
EICI
Z
Z
Z
Feature Extractor
shared
X
X
Client 1
Client 2
* EICI: Environment-Independent Conditionally Independence

Figure 4.3: Illustration of the overall framework designed for the proposed FedSDR.

4.4.1 The Provable Shortcut Discovery

At the first stage, we need to capture the complete shortcut features in a collaborative manner. Recalling the difference between the definitions of shortcut features and invariant features, we design the following objective to extract the complete shortcut features in a collaborative manner:

$$\omega_{\Psi}^{\star}, \Psi^{\star} = \underset{\substack{\Psi:\mathcal{X}\to\mathcal{H} \\ \omega:\mathcal{H}\to\mathcal{Y}}}{\arg\min} \frac{1}{N}\sum_{u=1}^{N}\{\ell_{SD}^{u}(\Psi; D_u) := \mathcal{R}(\omega(\Psi); D_u) - \lambda\,\ell_{dis}(\Psi; D_u)\}, \tag{4.2}$$

where the first term $\mathcal{R}(\omega(\Psi); D_u)$ is adopted to exclude the uninformative features (e.g., noise). λ is the balancing weight and the second term $\ell_{dis}(\Psi; D_u)$ is designed for extracting the complete shortcut features. Specifically, we define that

$$\ell_{dis}(\Psi, D_u) := \mathbb{E}_{X\in D_u}\big[\sum_{e_i\in\mathcal{E}_{tr}}\sum_{e_j\in\mathcal{E}_{tr}}\mathcal{KL}\big(\mathbb{P}_{\omega_i^{\star}}(Y \mid \Psi(X), e_i)\big\|\mathbb{P}_{\omega_j^{\star}}(Y \mid \Psi(X), e_j)\big)\big], \tag{4.3}$$

where $\mathcal{KL}(\mathbb{P}\|\mathbb{Q})$ denotes the Kullback–Leibler divergence between two probability distributions. $\mathbb{P}_{\omega_i^{\star}}(Y \mid \Psi, e_i)$ means that $\mathbb{P}(Y \mid \Psi, e_i)$ is parameterized by the classifier $\omega_i^{\star}$ which is trained by:

$$\omega_i^{\star} = \underset{\omega_i:\mathcal{H}\to\mathcal{Y}}{\arg\min}\sum_{u=1}^{N}\rho_u^i\mathcal{R}(\omega_i(\Psi); e_i), \forall e_i \in \mathcal{E}_{tr}, \tag{4.4}$$

where $\rho_u^i = 1$ when client u has data samples from environment e_i and $\rho_u^i = 0$ otherwise.

Since $\mathbb{P}_{\omega_i^{\star}}(Y \mid \Psi, e_i)$ is parameterized by the classifier $\omega_i^{\star}$ to be a distribution around $\omega_i^{\star}(\Psi)$ for any given Ψ and e_i, we adopt a simple and effective measure to compute the divergence $\mathcal{KL}(\mathbb{P}_{\omega_i^{\star}}(Y \mid \Psi, e_i)\|\mathbb{P}_{\omega_j^{\star}}(Y \mid \Psi, e_j))$. That is $\mathcal{KL}(\mathbb{P}_{\omega_i^{\star}}(Y \mid \Psi, e_i)\|\mathbb{P}_{\omega_j^{\star}}(Y \mid \Psi, e_j)) = \frac{1}{2}\|\omega_i^{\star}(\Psi) - \omega_j^{\star}(\Psi)\|^2$, $\forall e_i, e_j \in \mathcal{E}_{all}$. In this way, we can rewrite the overall objective 4.2 for shortcut discovery as the following bi-level optimization:

$$\omega_{\Psi}^{\star}, \Psi^{\star} = \underset{\Psi,\omega}{\arg\min}\frac{1}{N}\sum_{u=1}^{N}\{\ell_{SD}^{u}(\Psi; D_u) := \mathcal{R}(\omega(\Psi); D_u) - \lambda\,\ell_{dis}(\Psi; D_u)\} \tag{4.5}$$

$$\text{s.t.} \qquad \omega_i^{\star} = \underset{\omega_i:\mathcal{H}\to\mathcal{Y}}{\arg\min}\sum_{u=1}^{N}\rho_u^i\mathcal{R}(\omega_i(\Psi); e_i), \forall e_i \in \mathcal{E}_{tr}, \tag{4.6}$$

where $\ell_{dis}(\Psi; D_u) = \mathbb{E}_{X \in D_u}[\frac{1}{2}\sum_{e_i \in \mathcal{E}_{tr}} \sum_{e_j \in \mathcal{E}_{tr}} \|\omega_i^\star(\Psi(X)) - \omega_j^\star(\Psi(X))\|^2]$. This bi-level optimization can be solved by alternatively updating the solutions of the outer and inner objective. Under federated learning, both the outer and inner objective can be divided into N sub-problems that can be settled on N local clients respectively. The sever can aggregate the update from local clients to gain the solution $\Psi^\star$. To avoid the outer objective being dominated by maximizing $\ell_{dis}(\Psi; D_u)$, we replace ℓ_{dis} with $\min(\alpha, \lambda\, \ell_{dis}(\Psi; D_u))$ in the practical version, where α is a positive threshold.

Theoretical Analysis Before continuing to introduce the shortcut removal method, we formally analyse the optimal solution of the Eq. 4.5. In this theoretical analysis part, we consider the linear data model that [80] adopted. Specifically, we focus on the logistic regression problem where label $y \in \{\pm 1\}$. As in [80], we suppose both the invariant features $Z_C = [Z_C^g, Z_C^U]$ and shortcut features Z_S of sample y are drawn from the following Gaussian:

$$Z_C \sim \mathcal{N}(y \cdot \mu_c, \sigma_c^2 I), \qquad Z_S \sim \mathcal{N}(y \cdot \mu_s, \sigma_s^2 I),$$

where $\mu_c \in \mathbb{R}^{d_c}$ and $\mu_s \in \mathbb{R}^{d_s}$. Samples of observation X are generated by $X = g(Z_C, Z_S)$ where $g(\cdot)$ is a non-parameterized function. Parameters μ_c, σ_c and function g are independent of environment while μ_s varies as environment changes.

Assumption 4.4.1. *The distribution of label Y satisfies the following two conditions: 1) $\mathbb{P}(Y \mid u) = \mathbb{P}(Y), \forall u \in U$; 2) $\mathbb{P}(y \mid e) = \mathbb{P}(y' \mid e), \forall y, y' \in Y$ and $\forall e \in \mathcal{E}_{tr}$.*

Theorem 4.4.1. *If the Assumption 4.4.1 holds, the function g is linear and the total number of training environments in the federated learning system satisfies $|\mathcal{E}_{tr}| > d_s$, then the following two statements are equivalent:*

- *$\Psi^\star(X)$ depends and only depends on the complete shortcut features Z_S. That is, $\Psi^\star(X)$ is a function of Z_S alone;*
- *$\Psi^\star$ is the optima of the objective 4.5 with an appropriately chosen value of the hyper-parameter λ.*

Proof. We write the linear feature extractors Ψ that can recover the latent features ($[Z_C, Z_S]$) from the observation X as $\Psi(X) = \Psi(g(Z_C, Z_S)) = AZ_C + BZ_S$, where A and B are fixed transformation matrices. This formulation is also adopted in the theoretical analysis in [80] and [99]. For the concerned logistic regression, we can get a closed form for the distribution $\mathbb{P}(Y \mid \Psi, e)$ as:

$$\begin{aligned}\mathbb{P}(Y \mid \Psi, e) &:= \mathbb{P}^e(Y \mid AZ_C + BZ_S)\\ &= \frac{\mathbb{P}^e(AZ_C + BZ_S \mid Y)\mathbb{P}^e(Y)}{\mathbb{P}^e(AZ_C + BZ_S)}\\ &= \frac{\mathbb{P}^e(AZ_C + BZ_S \mid Y)\mathbb{P}^e(Y)}{\sum_y \mathbb{P}^e(Y = y)\mathbb{P}^e(AZ_C + BZ_S \mid Y = y)}\end{aligned}$$

Since Assumption 4.4.1 holds, we have $\mathbb{P}^e(Y = y) = \mathbb{P}^e(Y = y'), \forall y \in Y$. We can obtain

$$\begin{aligned}\mathbb{P}^e(y \mid AZ_C + BZ_S) &= \frac{\mathbb{P}^e(AZ_C + BZ_S \mid y)}{\sum_y \mathbb{P}^e(AZ_C + BZ_S \mid Y = y)}\\ &= \frac{\mathbb{P}^e(AZ_C + BZ_S \mid y)}{\mathbb{P}^e(AZ_C + BZ_S \mid Y = y) + \mathbb{P}^e(AZ_C + BZ_S \mid Y = -y)}\\ &= \frac{1}{1 + \frac{\mathbb{P}^e(AZ_C + BZ_S \mid Y = -y)}{\mathbb{P}^e(AZ_C + BZ_S \mid Y = y)}}, \quad \forall y \in \{\pm 1\}.\end{aligned}$$

Because we have $Z_C \perp\!\!\!\perp Z_S \mid Y$ from Lemma 4.4.1, we can get the probability density of $AZ_C + BZ_S$ as follows:

$$AZ_C + BZ_S \mid y \sim \mathcal{N}(y \cdot \mu_z, \Sigma_z), \tag{4.7}$$

where $\mu_z = A\mu_c + B\mu_s$ and $\Sigma_z = AA^T\sigma_c^2 + BB^T\sigma_s^2$. Thus, we can get $\mathbb{P}(Y \mid \Psi, e)$ as:

$$\begin{aligned}\mathbb{P}^e(y \mid \Psi) &= \frac{1}{1 + \frac{\mathbb{P}^e(AZ_C + BZ_S \mid Y = -y)}{\mathbb{P}^e(AZ_C + BZ_S \mid Y = y)}}\\ &= \frac{1}{1 + \exp(-y \cdot 2\Psi^T \Sigma_z^{-1} \mu_z)}, \quad \forall y \in \{\pm 1\},\end{aligned}$$

where Σ_z^{-1} represents the generalized inverse of Σ_z, i.e., $\Sigma_z^{-1}\Sigma_z = I$.

According to Lemma F.2. proved in the appendix of [80], the optimal classifier based on the feature extractor $\Psi(X) = AZ_C + BZ_S$ is sufficiently and necessarily given by

$2(AA^T\sigma_c^2 + BB^T\sigma_s^2)^{-1}(A\mu_c + B\mu_s)$. That is, we have $\mathbb{P}_{\omega_i^\star}(y \mid \Psi) = \frac{1}{1+\exp(-y\cdot 2\Psi^T\Sigma_z^{-1}\mu_z)}$, $\forall y \in \{\pm 1\}$, if and only if $\omega_i^\star \in \arg\min_{\omega_i:\mathcal{H}\to\mathcal{Y}} \sum_{u=1}^N \rho_u^i \mathcal{R}(\omega_i(\Psi); e_i), \forall e_i \in \mathcal{E}_{tr}$.

Therefore, we can calculate the KL-divergence between $\mathbb{P}_{\omega_i^\star}(Y \mid \Psi, e_i)$ and $\mathbb{P}_{\omega_j^\star}(Y \mid \Psi, e_j)$ by

$$
\begin{aligned}
&\mathcal{KL}\big(\mathbb{P}_{\omega_i^\star}(Y \mid \Psi, e_i) \big\| \mathbb{P}_{\omega_j^\star}(Y \mid \Psi, e_j)\big)\\
&= \sum_{y\in\{\pm1\}} \mathbb{P}_{\omega_i^\star}(y \mid \Psi, e_i) \log \frac{\mathbb{P}_{\omega_i^\star}(y \mid \Psi, e_i)}{\mathbb{P}_{\omega_j^\star}(y \mid \Psi, e_j)}\\
&= \sum_{y\in\{\pm1\}} \frac{1}{1+\exp(-y\cdot 2\Psi^T\Sigma_{z_i}^{-1}\mu_z^i)} \log \frac{1+\exp(-y\cdot 2\Psi^T\Sigma_{z_j}^{-1}\mu_z^j)}{1+\exp(-y\cdot 2\Psi^T\Sigma_{z_i}^{-1}\mu_z^i)}\\
&= \frac{1}{1+\exp(-2\Psi^T\Sigma_{z_i}^{-1}\mu_z^i)} \log \frac{1+\exp(-2\Psi^T\Sigma_{z_j}^{-1}\mu_z^j)}{1+\exp(-2\Psi^T\Sigma_{z_i}^{-1}\mu_z^i)}\\
&\quad + \frac{1}{1+\exp(2\Psi^T\Sigma_{z_i}^{-1}\mu_z^i)} \log \frac{1+\exp(2\Psi^T\Sigma_{z_j}^{-1}\mu_z^j)}{1+\exp(2\Psi^T\Sigma_{z_i}^{-1}\mu_z^i)}\\
&= \frac{1}{1+\exp(-2\Psi^T\Sigma_{z_i}^{-1}\mu_z^i)} \left\{ \log \frac{1+\exp(2\Psi^T\Sigma_{z_j}^{-1}\mu_z^j)}{1+\exp(2\Psi^T\Sigma_{z_i}^{-1}\mu_z^i)} + \log \frac{\exp(2\Psi^T\Sigma_{z_i}^{-1}\mu_z^i)}{\exp(2\Psi^T\Sigma_{z_j}^{-1}\mu_z^j)} \right\}\\
&\quad + \frac{1}{1+\exp(2\Psi^T\Sigma_{z_i}^{-1}\mu_z^i)} \log \frac{1+\exp(2\Psi^T\Sigma_{z_j}^{-1}\mu_z^j)}{1+\exp(2\Psi^T\Sigma_{z_i}^{-1}\mu_z^i)}\\
&= \log \frac{1+\exp(2\Psi^T\Sigma_{z_j}^{-1}\mu_z^j)}{1+\exp(2\Psi^T\Sigma_{z_i}^{-1}\mu_z^i)} + \frac{2\Psi^T(\Sigma_{z_i}^{-1}\mu_z^i - \Sigma_{z_j}^{-1}\mu_z^j)}{1+\exp(-2\Psi^T\Sigma_{z_i}^{-1}\mu_z^i)}
\end{aligned}
$$

Similarly, we can get that

$$
\begin{aligned}
\mathcal{KL}\big(\mathbb{P}_{\omega_j^\star}(Y \mid \Psi, e_j) \big\| \mathbb{P}_{\omega_i^\star}(Y \mid \Psi, e_i)\big) &= \sum_{y\in\{\pm1\}} \mathbb{P}_{\omega_j^\star}(y \mid \Psi, e_j) \log \frac{\mathbb{P}_{\omega_j^\star}(y \mid \Psi, e_j)}{\mathbb{P}_{\omega_i^\star}(y \mid \Psi, e_i)}\\
&= \log \frac{1+\exp(2\Psi^T\Sigma_{z_i}^{-1}\mu_z^i)}{1+\exp(2\Psi^T\Sigma_{z_j}^{-1}\mu_z^j)} + \frac{2\Psi^T(\Sigma_{z_j}^{-1}\mu_z^j - \Sigma_{z_i}^{-1}\mu_z^i)}{1+\exp(-2\Psi^T\Sigma_{z_j}^{-1}\mu_z^j)}
\end{aligned}
$$

Combining the above results, we can get that

$$
\begin{aligned}
&\mathcal{KL}\big(\mathbb{P}_{\omega_i^\star}(Y \mid \Psi, e_i) \big\| \mathbb{P}_{\omega_j^\star}(Y \mid \Psi, e_j)\big) + \mathcal{KL}\big(\mathbb{P}_{\omega_j^\star}(Y \mid \Psi, e_j) \big\| \mathbb{P}_{\omega_i^\star}(Y \mid \Psi, e_i)\big)\\
&= \underbrace{\left\{ \frac{1}{1+\exp(-2\Psi^T\Sigma_{z_i}^{-1}\mu_z^i)} - \frac{1}{1+\exp(-2\Psi^T\Sigma_{z_j}^{-1}\mu_z^j)} \right\}}_{T_1} \cdot \underbrace{\left\{2\Psi^T(\Sigma_{z_i}^{-1}\mu_z^i - \Sigma_{z_j}^{-1}\mu_z^j)\right\}}_{T_2}
\end{aligned}
$$

$\geq 0, \forall e_i, e_j \in \mathcal{E}_{all}, \forall \Psi \in \mathcal{H}$.

Since the absolute value of term T_1 (i.e., $|T_1|$) monotonically increases with term $|T_2|$ increasing, the objective $\max_\Psi \mathcal{KL}\big(\mathbb{P}_{\omega_i^\star}(Y \mid \Psi, e_i)\big\|\mathbb{P}_{\omega_j^\star}(Y \mid \Psi, e_j)\big) + \mathcal{KL}\big(\mathbb{P}_{\omega_j^\star}(Y \mid \Psi, e_j)\big\|\mathbb{P}_{\omega_i^\star}(Y \mid \Psi, e_i)\big)$ is equivalent to $\max_\Psi \big\|2\Psi^T(\Sigma_{z_i}^{-1}\mu_z^i - \Sigma_{z_j}^{-1}\mu_z^j)\big\|^2$. Therefore, the second term $\frac{1}{N}\sum_{u=1}^N \ell_{dis}(\Psi; D_u)$ in Eq. 4.5 can be written as

$$\begin{aligned}\frac{1}{N}\sum_{u=1}^N \ell_{dis}(\Psi; D_u) &= \mathbb{E}_\Psi\Big[\sum_{e_i\in\mathcal{E}_{tr}}\sum_{e_j\in\mathcal{E}_{tr}} \big\|2\Psi^T(\Sigma_{z_i}^{-1}\mu_z^i - \Sigma_{z_j}^{-1}\mu_z^j)\big\|^2\Big] \\ &= \sum_{e_i\in\mathcal{E}_{tr}}\sum_{e_j\in\mathcal{E}_{tr}} \frac{4\|(A\mu_c^i + B\mu_s^i) - (A\mu_c^j + B\mu_s^j)\|^2}{(AA^T\sigma_c^2 + BB^T\sigma_s^2)^2}\cdot\mathbb{E}_\Psi\|\Psi\|^2 \\ &= \sum_{e_i\in\mathcal{E}_{tr}}\sum_{e_j\in\mathcal{E}_{tr}} \frac{4\|B(\mu_s^i - \mu_s^j)\|^2}{(AA^T\sigma_c^2 + BB^T\sigma_s^2)^2}\cdot\mathbb{E}_\Psi\|\Psi\|^2\end{aligned}$$

According to the mentioned $AZ_C + BZ_S \mid y \sim \mathcal{N}(y\cdot\mu_z, \Sigma_z)$, we can get the density

$$\mathbb{P}(\Psi) = \sum_{y\in Y}\mathbb{P}(Y = y)\mathbb{P}(\Psi \mid Y = y)$$

With the Assumption 4.4.1 holding, we can get the mean $\mathbb{E}[\Psi] = 0$ and the variance $\mathbb{D}[\Psi] = AA^T\sigma_c^2 + BB^T\sigma_s^2$. Therefore, we have

$$\begin{aligned}\frac{1}{N}\sum_{u=1}^N \ell_{dis}(\Psi; D_u) &= \mathbb{E}_\Psi\Big[\sum_{e_i\in\mathcal{E}_{tr}}\sum_{e_j\in\mathcal{E}_{tr}} \big\|2\Psi^T(\Sigma_{z_i}^{-1}\mu_z^i - \Sigma_{z_j}^{-1}\mu_z^j)\big\|^2\Big] \\ &= \sum_{e_i\in\mathcal{E}_{tr}}\sum_{e_j\in\mathcal{E}_{tr}} \frac{4\|B(\mu_s^i - \mu_s^j)\|^2}{(AA^T\sigma_c^2 + BB^T\sigma_s^2)^2}\cdot\{\mathbb{D}[\Psi] + (\mathbb{E}[\Psi])^2\} \\ &= \sum_{e_i\in\mathcal{E}_{tr}}\sum_{e_j\in\mathcal{E}_{tr}} \frac{4\|B(\mu_s^i - \mu_s^j)\|^2}{AA^T\sigma_c^2 + BB^T\sigma_s^2} \\ &= \frac{4}{\frac{AA^T}{BB^T}\sigma_c^2 + \sigma_s^2}\sum_{e_i\in\mathcal{E}_{tr}}\sum_{e_j\in\mathcal{E}_{tr}} \|\mu_s^i - \mu_s^j\|^2\end{aligned}$$

We can find that maximizing the above objective will make $A = 0$ and $BB^T \neq 0$. Moreover, when $|\mathcal{E}_{tr}| > d_s$, maximizing $\sum_{e_i\in\mathcal{E}_{tr}}\sum_{e_j\in\mathcal{E}_{tr}}\|\mu_s^i - \mu_s^j\|^2$ will make $rank(B) = d_s$. In the meanwhile, satisfying $A = 0$ and $rank(B) = d_s$ will in turn maximize the objective $\frac{1}{N}\sum_{u=1}^N \ell_{dis}(\Psi; D_u)$.

In Eq. 4.5, we utilize a Lagrangian multiplier to solve the constrained optimization and the balancing weight is λ. Therefore, Theorem 4.4.1 gets proved. □

Remark 4.4.2. *Theorem 4.4.1 guarantees the elaborated Eq. 4.5 can yield the feature extractor that extracts complete shortcut features and excludes all invariant features. Note that Assumption 4.4.1 is about the label distributions in training datasets. Since the shortcut extractor works as an auxiliary model and is never part of the optimal personalized invariant predictors, we can sample some data subsets from local training datasets to train the shortcut extractor $\Psi^\star$. In this way, the sampled data subsets can easily satisfy Assumption 4.4.1. Besides, the causal signatures in Lemma 4.4.1 play critical parts in the proof of Theorem 4.4.1.*

4.4.2 Personalized Invariant Learning with Shortcut Removal

With the shortcut extractor that depends and only depends on the complete shortcut features Z_S, we can extract the most informative invariant features to elicit the optimal personalized invariant predictor for each client. Based on the second causal signature in Lemma 4.4.1, we design the following objective for each client to develop the optimal personalized invariant predictor:

$$\omega_u^\star(\Phi_u^\star) = \underset{\Phi_u,\, \omega_u}{\arg\min}\, \ell_{SR}^u(\omega_u(\Phi_u); D_u) := \{\mathcal{R}(\omega_u(\Phi_u); D_u) + \gamma \cdot I(\Phi_u; \Psi^\star \mid Y)\}, \forall u \in [N], \tag{4.8}$$

where $I(\cdot;\cdot \mid \cdot)$ denotes the conditional mutual information and γ is the balancing weight. The optimal personalized invariant predictor is given by $f_{\theta_u}^\star := \omega_u^\star(\Phi_u^\star)$.

Theorem 4.4.2. *Suppose $\Psi^\star(X)$ in the Eq. 4.8 depends and only depends on the complete shortcut features Z_S. If $f_{\theta_u}^\star (\forall u \in [N])$ is the optima of the Eq. 4.8 with the hyper-parameter γ chosen appropriately, then the $f_{\theta_u}^\star$ is the optimal personalized invariant predictor for the client u, $\forall u \in [N]$.*

Proof. We know that minimizing $\mathcal{R}(\omega_u(\Phi_u); D_u)$ is the sufficient condition of maximizing $I(Y; \Phi_u(X))$, and $I(\Phi_u; \Psi^\star \mid Y) = 0$ is equivalent to $\Phi_u \perp\!\!\!\perp \Psi^\star \mid Y$. According to the property of Lagrangian multiplier, the objective in Eq. 4.8 is equivalent to the constrained optimization where the constrain is $I(\Phi_u; \Psi^\star \mid Y) = 0$, with the appropriately chosen γ. Combining with the second causal signature in Lemma 4.4.1, Theorem 4.4.2 gets proved. □

Remark 4.4.3. *Theorem 4.4.2 guarantees that our method can produce the optimal personalized invariant predictor for every client. Note that $I(\Phi_u; \Psi^\star \mid Y) = 0$ is the necessary and sufficient condition for $\Phi_u \perp\!\!\!\perp \Psi^\star \mid Y$. Since $\Phi_u \perp\!\!\!\perp \Psi^\star \mid Y$ is independent of environment, our method can develop the optimal personalized invariant predictor for every client even though there is only one training environment on each federated client.*

In the practical implementation, it can be infeasible to compute the exact value of $I(\Phi_u; \Psi^\star \mid Y)$. Considering the limited computation resources on local clients, we adopt a simple approximating scheme used in [45] to measure $I(\Phi_u; \Psi^\star \mid Y)$. Specifically, we estimate it by $I(\Phi_u; \Psi^\star \mid Y) \approx \mathbb{E}[\Phi_u(X) \cdot (\Psi^\star(X) - \mathbb{E}[\Psi^\star(X) \mid Y])]$ because $I(\Phi_u; \Psi^\star \mid Y) = 0$ is the sufficient (but not necessary) condition for $\mathbb{E}[\Phi_u(X) \cdot (\Psi^\star(X) - \mathbb{E}[\Psi^\star(X) \mid Y])] = 0$. With the data samples on local clients, we estimate the conditional mutual information by

$$I(\Phi_u; \Psi^\star \mid Y) \approx \left\| \frac{1}{M_u} \sum_{m=1}^{M_u} \Phi_u(X_m) \Big(\Psi^\star(X_m) - \sum_{n=1}^{M_u} \frac{q_n^m}{\sum_{n \in [M_u]} q_n^m} \Psi^\star(X_n) \Big) \right\|_1$$

where $(X_m, y_m), m \in [M_u]$ is drawn from dataset D_u, $q_n^m = 1$ if $y_n = y_m$ and $q_n^m = 0$ otherwise.

Note that our method can easily cooperate with most of the existing PFL methods to improve their OOD generalization performance by adding $I(\Phi_u; \Psi^\star \mid Y)$ into their objectives as a regularization term, since $I(\Phi_u; \Psi^\star \mid Y) = 0$ can constrain the personalized models to eliminate all shortcut features even though each client has only one training environment.

4.4.3 Algorithm Design

In the following contents, we will discuss what the server and local clients need to conduct to develop the optimal personalized invariant predictor $f^{\star}_{\theta_u}$ for each client, $\forall u \in [N]$. The detailed pseudo-code of the designed algorithm FedSDR is provided in Algorithm 2.

Server Update. Before the algorithm starts, the server initializes the models with random parameters. At each communication round t, the server firstly selects a fraction of local clients ($u \in S^t$) and broadcast the current Ψ^t and $\{\omega_i^t \mid i = 1, 2, ..., |\mathcal{E}_{tr}|\}$ to them. After the selected local clients finish conducting the **client update** process, the server can receive the local update Ψ_u^{t+1} and $\{\omega_{i,u}^{t+1} \mid i = 1, 2, ..., |\mathcal{E}_{tr}^u|\}$ from the selected clients. Then it can update the global solutions by $\Psi^{t+1} = \frac{1}{|S^t|} \sum_{u \in S^t} \Psi_u^{t+1}$ and $\omega_i^{t+1} = \sum_{u \in S^t} \frac{\rho_u^i}{\sum_{u \in S^t} \rho_u^i} \omega_{i,u}^{t+1}, i = 1, 2, ..., |\mathcal{E}_{tr}|$.

Client Update. Before the algorithm starts, the client u initializes the personalized invariant model with random parameters $f^0_{\theta_u}$. After receiving the global model Ψ^t and $\{\omega_i^t \mid i = 1, 2, ..., |\mathcal{E}_{tr}|\}$ from the server, the local client u ($\forall u \in S^t$) needs to carry on the following two steps: **1)** update the personalized invariant model by

$$f^{t,k+1}_{\theta_u} = f^{t,k}_{\theta_u} - \eta \nabla \ell^u_{SR}(f^{t,k}_{\theta_u}; D_u)$$

for K steps and finally get $f^{t+1}_{\theta_u} = f^{t,K}_{\theta_u}$, where η is the personalized learning rate. **2)** The client can conduct R local iterations to update the local shortcut extractor. Before it starts, the client initializes the related models as $\Psi_u^{t,r=0} = \Psi^t$ and $\omega_{i,u}^{t,r=0} = \omega_i^t, i = 1, 2, ..., |\mathcal{E}_{tr}^u|$. During each local iteration r, the client firstly updates the local shortcut extractor by

$$\Psi_u^{t,r+1} = \Psi_u^{t,r} - \beta \nabla \ell^u_{SD}(\Psi_u^{t,r}; D_u)$$

for one epoch where β denotes the learning rate, and then get the near-optimal environment classifiers $\omega_{i,u}^{t,r+1}, i = 1, 2, ..., |\mathcal{E}_{tr}^u|$ by stochastic gradient descent (on

Algorithm 2 FedSDR: Federated Learning with Shortcut Discovery and Removal

Input: Hyper-parameters $T, R, K, \beta, \eta, \alpha, \lambda, \gamma$.

```
Initialize the models $\Psi^0$, $\{\omega_i^0 | i \in [|\mathcal{E}_{tr}|]\}$, $\{f^0_{\theta_u} | u \in [N]\}$.
for $t = 0$ to $T - 1$ do
    Server sends global models ($\Psi^t$, $\{\omega_i^t | i \in [|\mathcal{E}_{tr}|]\}$) to the participating clients.
    for local device $u = 1$ to $N$ in parallel do
        Initialization: $\Psi_u^{t,0} \leftarrow \Psi^t$, $\{\omega_{i,u}^t \leftarrow \omega_i^t | i \in [|\mathcal{E}_{tr}^u|]\}$.
        for $k = 0$ to $K - 1$ do
            Update personalized invariant model: $f^{t,k+1}_{\theta_u} = f^{t,k}_{\theta_u} - \eta \nabla \ell^u_{SR}(f^{t,k}_{\theta_u}; D_u)$.
        end for
        Initialization: $f^{t+1,0}_{\theta_u} \leftarrow f^{t,K}_{\theta_u}$.
        for $r = 0$ to $R - 1$ do
            Update the shortcut extractor: $\Psi_u^{t,r+1} = \Psi_u^{t,r} - \beta \nabla \ell^u_{SD}(\Psi_u^{t,r}; D_u)$.
            Update environment classifiers for $K$ epochs with $\nabla \mathcal{R}(\omega_{i,u}^{t,r}(\Psi_u^{t,r}); e_i)$.
        end for
    end for
    Randomly select a subset ($S^t$) of the users to upload the local approximation:
       $\Psi_u^{t+1} \leftarrow \Psi_u^{t,R}$ and $\{\omega_{i,u}^{t+1} \leftarrow \omega_{i,u}^{t,R} \mid i = 1, 2, ..., |\mathcal{E}_{tr}^u|\}$.
    Global aggregation:
       Shortcut extractor $\Psi^{t+1} = \frac{1}{|S^t|} \sum_{u \in S^t} \Psi_u^{t+1}$;
       Environment classifiers $\omega_i^{t+1} = \sum_{u \in S^t} \frac{\rho_u^i}{\sum_{u \in S^t} \rho_u^i} \omega_{i,u}^{t+1}, i = 1, 2, ..., |\mathcal{E}_{tr}|$.
end for
return the personalized invariant models $\{f^{T,0}_{\theta_u} | u \in [N]\}$.
```

$\nabla\mathcal{R}(\omega_{i,u}^{t,r}(\Psi_u^{t,r}); e_i))$ for L steps. When completing R local iterations, the client upload the local parameters $\Psi_u^{t+1} = \Psi_u^{t,R}$ and $\{\omega_{i,u}^{t+1} = \omega_{i,u}^{t,R} \mid i = 1, 2, ..., |\mathcal{E}_{tr}^u|\}$ to the server for **server update**.

4.5 Experiments

4.5.1 Empirical Validation of Theorem 4.4.2

To verify the provided theoretical guarantees under linear cases, we generate a synthetic dataset using the same strategy as in [80]. Specifically, it is a logistic regression task and the data instance X is generated by $X = g(Z_C^g, Z_C^U, Z_S)$, where the dimensionalities of Z_C^g, Z_C^U and Z_S are $d_C^g = 3$, $d_C^U = 3$ and $d_S = 6$ respectively. The linear function g is implemented by one fully-connected layer which has 12 neurons. The latent variables Z_C^g, Z_C^U and Z_S are subject to $\mathcal{N}(y \cdot \mu_{c,g}, \sigma_{c,g}^2 I)$, $\mathcal{N}(y \cdot \mu_{c,u}, \sigma_{c,u}^2 I)$ and $\mathcal{N}(y \cdot \mu_s, \sigma_s^2 I)$ respectively. Target variable y is taken from the distribution $\mathbb{P}(y = -1) = \mathbb{P}(y = 1) = 0.5$. Both $\mu_{c,g}$ and $\mu_{c,u}$ are randomly sampled from $\mathcal{N}(0, 1.5I)$ while μ_s is randomly sampled from $\mathcal{N}(0, 0.75I$. To make the shortcut representation Z_S easier to learn, we choose $\sigma_{c,g} = \sigma_{c,u} = 2$ and $\sigma_s = 1$ as in [80]. Each fixed value of μ_s indicates one specified environment. We generate 10 training environments and 5000 test environments to evaluate the out-of-distribution generalization performance. Each (training/test) environment contains 10000 data samples (X, y) and the training data samples are distributed onto totally 100 clients. The training and test data samples on each client are generated with an identical value of $\mu_{c,u}$. Besides, we choose the client sampling rate as 0.1. The experimental results on this synthetic dataset are shown in Table 4.1. In particular, when we manually select the causal features $[Z_C^g, Z_C^U]$ as the discriminating features, we find the optimal personalized classifiers achieve an stable accuracy around 97.5 in different test environments. Therefore, the results shown in Table 4.1 can demonstrate the effec-

Table 4.1: The performance of FedSDR and the competitors on the synthetic dataset.

Algorithm	FedAvg	DRFA	FedSR	FedIIR	FTFA	pFedMe	Ditto	FedRep	FedRoD	FedPAC	**FedSDR**
worst-case(%)	3.06	62.41	63.09	67.39	1.32	10.58	7.76	2.57	21.53	8.98	92.49
average(%)	85.56	69.64	70.53	70.75	96.26	95.72	96.50	97.80	97.24	98.77	96.07

tiveness of our FedSDR on developing the optimal personalized invariant predictors, compared with the state-of-the-art FL and PFL methods.

4.5.2 Experimental Setup

Colored-MNIST (CMNIST) [6] is constructed based on MNIST [53] via rearranging the images of digit 0-4 into a single class labeled 0 and the images of digit 5-9 into another class labeled 1. Each digit having label 0 is colored green/red with probability $p^e/1-p^e$ and each digit having label 1 is colored red/green with probability $p^e/1-p^e$, respectively. Thus "color" feature is shortcut in the dataset and the data distribution varies as p^e changes. We provide two training environments ($p^e_{tr} = 0.90$ and 0.80) as $\mathcal{E}_{tr}$ and every local client only has one training environment which is randomly sampled from $\mathcal{E}_{tr}$. To assess the model performance on different test distributions, the test environment on each client varies from $p^e_{te} = 0.00$ to 1.00. Considering the heterogeneous data generating process across local clients, the data instances used for constructing the training/test environments on each client are randomly sampled from only two digit sub-classes labeled 0 (e.g., digit 1, 2) and two digit sub-classes labeled 1 (e.g., digit 6, 7) without replacement.

Colored Fashion-MNIST (CFMNIST) [2] is constructed using the same strategy as Colored-MNIST, but the original images come from Fashion-MNIST [101]. Hence, CFMNIST dataset carries more complex feature space than colored-MNIST does.

WaterBird [81] considers a real-world scenario where the photographs of waterbirds usually have water backgrounds while the photographs of landbirds usually have land backgrounds because of the distinct habitats. It makes learning models easily trapped by "background" shortcut when classify "waterbird" and "landbird". In WaterBird, a waterbird is placed onto a water/land background with probability $p^e/1 - p^e$ and a landbird is placed onto a land/water background with probability $p^e/1 - p^e$ respectively. We setup two training environments ($p^e_{tr} = 0.95$ and 0.85) as $\mathcal{E}_{tr}$ and each client has only one training environment which is randomly sampled from $\mathcal{E}_{tr}$. The test environment varies from $p^e_{te} = 0.00$ to 1.00. We notice that the diverse geographic distributions of different bird species naturally accord with the heterogeneity of local data generating process if the federated clients are located in different geographic areas. Considering WaterBird includes 46 waterbird species and 154 landbird species, we distribute 15 (10 separated and 5 overlapped) waterbird species and 51 (34 separated and 17 overlapped) landbird species to each client. The training and test datasets on each client contain bird pictures that belong to the same bird species.

PACS [54] is a larger real-world dataset commonly-used in evaluating out-of-distribution (OOD) generalization. It consists of 7 classes distributed across 4 environments (or domains). We adopt the "leave-one-domain-out" strategy to evaluate the OOD generalization performance. Taking personalization into consideration, we split each training domain into two subsets according to classes (i.e., one subset consists of dog, elephant and giraffe and another subset consists of guitar, horse, house, and person), and then distribute these two subsets onto two clients respectively. Training and test datasets on each client come from different domains but consist of the same classes.

Baseline Methods: We compare our method (FedSDR) with 10 state-of-the-art algorithms: 4 federated learning methods (FedAvg [70], DRFA [21], FedSR [73] and FedIIR [30]), and 6 personalized federated learning methods (pFedMe [94], Ditto [57], FTFA [17], FedRep [18], FedRoD [12] and FedPAC [102]).

Model Architectures: For CMNIST and CFMNIST, we adopt the deep neural network with one hidden layer as feature extractor and an subsequent fully-connected layer as classifier. As regard to Waterbird and PACS, ResNet-18 [36] is used as the learning model where the part before the last fully-connected layer works as feature extractor and the last fully-connected layer works as classifier.

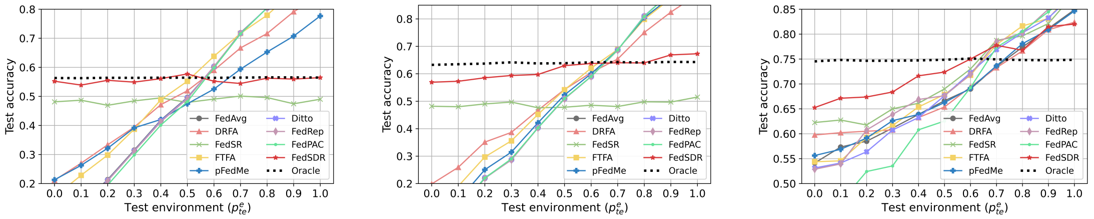
Selection of Hyper-parameters: The hyper-parameters of the competitors and our algorithm are tuned to make the accuracy on the validation environment (i.e., $p^e_{val} = 0.10$) as high as possible. Specifically, the mainly used hyper-parameters in the evaluation part are listed as follows: Global communication round: $T = 600$, Local iterations: $R = 10$, Personalized epochs to update the personalized invariant predictors: $K = 10$, Local batch size: $B = 50$, Global learning rate: $\beta = 0.0001$, Personalized learning rate: $\eta = 0.0001$, Discrepancy threshold: $\alpha = 1.0$, Balancing weight: $\lambda = 0.5$, Balancing weight: $\gamma = 1.4$, Optimizer: Adam.

Table 4.2: The overall comparison between the performance of our method and the baselines on four datasets.

Dataset	CMNIST		CFMNIST		WaterBird		PACS	
Test acc (%)	worst-case	average	worst-case	average	worst-case	average	worst-case	average
FedAvg	3.39	51.03	0.16	50.02	54.13	67.95	41.71	47.66
DRFA	21.15	52.81	19.84	53.88	59.75	68.39	42.48	48.95
FedSR	46.93	48.62	47.61	48.90	61.75	71.68	46.76	51.25
FedIIR	47.25	48.39	48.06	49.16	61.24	70.87	47.03	51.58
FTFA	15.42	54.96	11.35	53.52	54.38	69.68	40.89	48.79
pFedMe	21.30	48.53	4.22	51.26	55.63	68.24	45.24	51.33
Ditto	3.02	50.97	0.37	50.12	53.13	68.73	44.95	51.28
FedRep	2.76	50.83	0.11	50.01	52.88	70.23	49.27	53.75
FedRoD	9.09	50.84	1.23	51.57	52.36	70.86	48.16	52.92
FedPAC	1.01	50.05	0.16	50.13	45.08	65.57	49.93	54.20
FedSDR	**53.88**	**55.59**	**56.92**	**61.88**	**65.25**	**73.20**	**52.14**	**56.18**

4.5.3 Overall Performance

We summarize the test accuracy of all competitors on different unseen test distributions (11 test distributions in CMNIST, CFMNIST and WaterBird; 4 test distributions in PACS) and figure out the worst-case and average accuracy of each method in Table 4.2. We can find that our method FedSDR consistently outperform the baselines on both worst-case and average test accuracy. In particular, FedSDR achieves around **6.5**%, **9**%, **3.5**% and **2**% higher worst-case accuracy than the second best algorithm and in the meanwhile reaches the highest average accuracy on CMNIST, CFMNIST, WaterBird and PACS, respectively.



(a) CMNIST (b) CFMNIST (c) WaterBird

Figure 4.4: The relationship between the test accuracy and the test distribution.

4.5.4 Mitigation of Shortcut Features.

Since there exists definite correlation between shortcut features and label in CMNIST, CFMNIST and WaterBird, we can use these three datasets to evaluate how well a method can mitigate the shortcut features. The more highly a method relies on shortcut, the more approximate its test accuracy is to the corresponding p^e_{te}. In contrast, a method that eliminates the shortcut can produce consistent test accuracy across different p^e_{te}. We evaluate the competitors under diverse test environment (i.e., p^e_{te}) and show the relationships between test accuracy and p^e_{te} in Figure 4.4. In particular, "Oracle" represents the scheme where we manually remove the shortcut features (color in CMNIST and CFMNIST; background in WaterBird) from the whole dataset and then train the personalized models using the pre-processed dataset. Hence

"Oracle" provides an ideal performance for comparison. We can see that FedSDR can effectively mitigate the shortcut and achieve a more consistent test accuracy than most of the FL and PFL methods. Because FedSDR exploits the personalized invariant features, it consistently achieves a higher test accuracy than the federated domain generalization methods which drop the personalization information.

Table 4.3: Performance comparison between FedSDR and the trivial combination of invariant learning with personalized federated learning schemes.

Dataset	CMNIST		CFMNIST		WaterBird		PACS	
Test acc (%)	worst-case	average	worst-case	average	worst-case	average	worst-case	average
IRM†	46.38	49.14	47.76	49.41	60.38	68.63	46.35	50.83
IRM†-FT	14.32	54.27	11.09	53.48	60.25	69.46	43.18	50.04
IRM†-L2	45.68	49.04	47.92	49.46	61.25	68.93	48.57	51.98
FedSDR	**53.88**	**55.59**	**56.92**	**61.88**	**65.25**	**73.20**	**52.14**	**56.18**

4.5.5 Necessity of Shortcut Discovery and Removal.

At the beginning of Section 4.4, we analyse that trivial combination of invariant learning scheme with local adaptation (commonly used in PFL) can fail to generate the optimal personalized invariant predictors for local clients. To validate the superiority of our method on developing the personalized invariant predictors when local training environments are insufficient, we implement two typical personalization skills with the global model being trained by the distributional version of IRM (i.e., IRM† in Table 4.3). One is L2-norm regularizer used in PFL [94, 34, 32, 57], and we call this implementation IRM†-L2. Another one is local **F**ine-**T**uning which is proved simple and effective for personalization ([17]) and we name it IRM†-FT.

From the results in Table 4.3, we can find the combinations can hardly improve the OOD generalization performance. In particular, the local fine-tuning skill can even degrade the performance, compared with baseline IRM†. The underlying reason is that local adaptation can readily make the personalized model pick up the shortcut features when local training environments are insufficient. By contrast, our shortcut removal method is independent of environment and can effectively mitigate the shortcut features even though there is only one training environment on each client.

Table 4.4: Performance of FedSDR on WaterBird with different values of hyper-parameters λ and γ.

λ	**0.00**	**0.10**	**0.50**	**1.00**	**10.0**
worst-case (%)	61.88	62.51	65.25	61.68	61.74
average (%)	71.34	72.39	73.20	70.61	70.18
γ	**0.00**	**0.10**	**1.00**	**1.40**	**10.0**
worst-case (%)	43.75	44.16	57.64	65.25	48.29
average (%)	66.30	65.73	70.18	73.20	64.86

4.5.6 Effect of Effects of Balancing Weights.

We evaluate the effects of two significant hyper-parameters in the proposed objective (i.e., λ and γ) on model performance here. Since the results on other datasets present the similar tendency as on WaterBird, we herein focus on WaterBird. The results on other datasets are placed in the appendix. When evaluating the effect of λ, we fix $\gamma = 1.4$. When evaluating the effect of γ, we fix $\lambda = 0.5$. The results are shown in Table 4.4. When $\lambda = 0.0$, shortcut feature extractor is trained by empirical risk minimization (i.e., ERM). When $\gamma = 0.0$, the personalized models are trained by local ERM. Because models trained by ERM tend to rely on shortcut, the performance of FedSDR is more sensitive to the selection of γ than the selection of λ.

4.5.7 Scalability

In the evaluation part of the main text, we simulate 8 clients in the experiments on CMNIST, CFMNIST and WaterBird. The experiments on PACS are conducted on 6 clients. To further evaluate the scalability of FedSDR, we firstly partition CMNIST dateset into 8 subsets using the same strategy adopted to simulate 8 clients. And then, we randomly distribute each subset onto 10 clients. In this way, we totally construct 80 clients for CMNIST dataset. Similarly, we construct 80, 80, 60 clients for CFMNIST, WaterBird and PACS respectively. When evaluating the model performance on these four datasets, we adopt a client sampling rate of 0.1. The experimental results are shown in Table:

Table 4.5: The overall comparison between the performance of our method FedSDR and the baselines with a large number of clients.

Dataset	CMNIST		CFMNIST		WaterBird		PACS	
Test acc (%)	worst-case	average	worst-case	average	worst-case	average	worst-case	average
FedAvg	1.74	46.82	0.77	45.62	48.65	61.57	33.75	40.18
DRFA	14.94	47.24	15.51	47.14	52.34	60.43	36.17	41.75
FedSR	40.29	43.64	41.16	43.27	55.63	64.32	39.03	43.40
FedIIR	41.18	42.93	41.80	43.58	54.31	64.60	40.15	44.37
FTFA	11.51	49.28	7.20	47.57	50.25	63.39	34.65	42.19
pFedMe	17.28	44.13	2.42	47.95	50.01	61.97	41.06	45.84
Ditto	1.98	45.84	1.80	45.71	49.08	63.38	40.18	46.30
FedRep	1.56	46.20	0.83	46.14	48.12	64.52	42.16	47.58
FedRoD	6.53	46.86	1.60	47.43	49.56	65.49	42.68	46.61
FedPAC	0.38	45.64	0.23	44.88	42.61	63.81	44.19	49.71
FedSDR	**50.41**	**51.85**	**52.81**	**57.14**	**59.96**	**68.09**	**48.07**	**51.55**

The results show that FedSDR can still outperform the competitors when there are a large number of clients in the federated learning system, which can validate the scalability of the proposed FedSDR.

Algorithm 3 FedSDR (+FedAvg): Federated Learning with Shortcut Discovery and Removal

Input: Hyper-parameters $T, R, K, \beta, \eta, \alpha, \lambda, \gamma$.

```
Initialize the models $\Psi^0$, $f^0_\theta$ and $\{\omega^0_i | i \in [|\mathcal{E}_{tr}|]\}$.
for $t = 0$ to $T - 1$ do
    Server sends global models ($\Psi^t$, $f^t_\theta$ and $\{\omega^t_i | i \in [|\mathcal{E}_{tr}|]\}$) to participating clients.
    for local device $u = 1$ to $N$ in parallel do
        Initialization: $\Psi^{t,0}_u \leftarrow \Psi^t$, $f^{t,0}_{\theta_u} \leftarrow f^t_\theta$ and $\{\omega^t_{i,u} \leftarrow \omega^t_i | i \in |\mathcal{E}^u_{tr}|\}$.
        for $r = 0$ to $R - 1$ do
            Update personalized invariant model: $f^{t,r+1}_{\theta_u} = f^{t,r}_{\theta_u} - \eta \nabla \ell^u_{SR}(f^{t,r}_{\theta_u}; D_u)$.
            Update shortcut extractor: $\Psi^{t,r+1}_u = \Psi^{t,r}_u - \beta \nabla \ell^u_{SD}(\Psi^{t,r}_u; D_u)$.
            Update environment classifiers for $K$ epochs with $\nabla \mathcal{R}(\omega^{t,r}_{i,u}(\Psi^{t,r}_u); e_i)$.
        end for
    end for
    Randomly select a subset ($S^t$) of the users to upload the local approximation:
       $\Psi^{t+1}_u \leftarrow \Psi^{t,R}_u$, $f^{t+1}_{\theta_u} \leftarrow f^{t,R}_{\theta_u}$ and $\{\omega^{t+1}_{i,u} \leftarrow \omega^{t,R}_{i,u} \mid i = 1, 2, ..., |\mathcal{E}^u_{tr}|\}$.
    Global aggregation:
       Shortcut extractor $\Psi^{t+1} = \frac{1}{|S^t|} \sum_{u \in S^t} \Psi^{t+1}_u$;
       Environment classifiers $\omega^{t+1}_i = \sum_{u \in S^t} \frac{\rho^i_u}{\sum_{u \in S^t} \rho^i_u} \omega^{t+1}_{i,u}, i = 1, 2, ..., |\mathcal{E}_{tr}|$;
       Global invariant model $f^{t+1}_\theta = \frac{1}{|S^t|} \sum_{u \in S^t} f^{t+1}_{\theta_u}$.
end for
return the global and personalized invariant models $f^T_\theta$, $\{f^{T,R}_{\theta_u} | u \in [N]\}$.
```

4.5.8 Compatibility

We notice that the proposed shortcut discovery and removal method can easily cooperate with most of the existing federated and personalized federated learning method to improve the out-of-distribution generalization performance via adding shortcut discovery and removal as a regularization term. We provide an example combination of FedSDR with FedAvg [70] and pFedMe [94] in Algorithm 3 and Algorithm 4, respectively. Of course, the combinations with more other federated and personalized federated learning methods can be explored in the future.

We implement Algorithm 3 and Algorithm 4 mentioned above on three datasets (i.e., Colored-MNIST, Colored-FMNIST and WaterBird). The values of the generic hyper-parameters are set same as FedSDR. In particular, we choose $\lambda = 0.5$ and $\gamma = 1.0 * 10^4$ for Colored-MNIST and Colored-FMNIST. As to WaterBird dataset, we choose $\lambda = 0.5$ and $\gamma = 1.4$. The experimental results are shown in the following Table 4.6. Note that the performances of FedSDR+FedAvg and FedSDR+pFedMe are evaluated with the personalized invariant models output by Algorithm 3 and Algorithm 4, respectively. The results show that the model performance can be further improved when we combine the proposed shortcut discovery and removal method with the prevalent federated learning algorithms.

Table 4.6: The combinations of our method with other federated learning schemes.

Dataset	CMNIST		CFMNIST		WaterBird	
Test acc(%)	worst-case	average	worst-case	average	worst-case	average
FedSDR	53.88	55.59	56.92	61.88	65.25	73.20
FedSDR+FedAvg	51.76	56.01	57.14	61.56	66.73	74.27
FedSDR+pFedMe	53.54	55.93	56.69	63.22	67.32	74.38

Algorithm 4 FedSDR (+pFedMe): Federated Learning with Shortcut Discovery and Removal

Input: Hyper-parameters $T, R, K, \beta, \eta, \alpha, \lambda, \gamma$.

```
Initialize the models $\Psi^0$, $f_\theta^0$ and $\{\omega_i^0 | i \in [|\mathcal{E}_{tr}|]\}$.
for $t = 0$ to $T - 1$ do
    Server sends global models ($\Psi^t$, $f_\theta^t$ and $\{\omega_i^t | i \in [|\mathcal{E}_{tr}|]\}$) to participating clients.
    for local device $u = 1$ to $N$ in parallel do
        Initialization: $\Psi_u^{t,0} \leftarrow \Psi^t$, $f_\theta^{t,0} \leftarrow f_\theta^t$ and $\{\omega_{i,u}^t \leftarrow \omega_i^t | i \in |\mathcal{E}_{tr}^u|\}$.
        for $r = 0$ to $R - 1$ do
            for $k = 0$ to $K - 1$ do
                Update the personalized invariant model:
                $f_{\theta_u}^{r,k+1} = f_{\theta_u}^{r,k} - \eta(\nabla \ell_{SR}^u(f_{\theta_u}^{r,k}; D_u) + \gamma(f_{\theta_u}^{r,k} - f_\theta^{t,r}))$.
            end for
            Update global invariant model: $f_\theta^{t,r+1} = f_\theta^{t,r} - \beta\gamma(f_\theta^{t,r} - f_{\theta_u}^{r,K})$
            Update shortcut extractor: $\Psi_u^{t,r+1} = \Psi_u^{t,r} - \beta \nabla \ell_{SD}^u(\Psi_u^{t,r}; D_u)$.
            Update environment classifiers for $K$ epochs with $\nabla \mathcal{R}(\omega_{i,u}^{t,r}(\Psi_u^{t,r}); e_i)$.
        end for
    end for
    Randomly select a subset ($S^t$) of the users to upload the local approximation:
      $\Psi_u^{t+1} \leftarrow \Psi_u^{t,R}$, $f_{\theta_u}^{t+1} \leftarrow f_{\theta_u}^{t,R}$ and $\{\omega_{i,u}^{t+1} \leftarrow \omega_{i,u}^{t,R} \mid i = 1, 2, ..., |\mathcal{E}_{tr}^u|\}$.
    Global aggregation:
      Shortcut extractor $\Psi^{t+1} = \frac{1}{|S^t|} \sum_{u \in S^t} \Psi_u^{t+1}$;
      Environment classifiers $\omega_i^{t+1} = \sum_{u \in S^t} \frac{\rho_u^i}{\sum_{u \in S^t} \rho_u^i} \omega_{i,u}^{t+1}, i = 1, 2, ..., |\mathcal{E}_{tr}|$;
      Global invariant model $f_\theta^{t+1} = \frac{1}{|S^t|} \sum_{u \in S^t} f_{\theta_u}^{t+1}$.
end for
return the global and personalized invariant models $f_\theta^T$, $\{f_{\theta_u}^{R,K} | u \in [N]\}$.
```

4.6 Remark

In this chapter, we study the challenging shortcut trap problem in PFL. We formulate the SCMs to interpret the heterogeneous data generating mechanisms on federated clients and derive two significant causal signatures which inspire our provable shortcut discovery and removal method. Theoretical analysis proves the proposed FedSDR can draw all shortcut features and elicit the optimal personalized invariant predictor that can generalize to unseen target data for each client. FedSDR can cooperate with most of the existing PFL methods to improve their OOD generalization performance, which can facilitate the real-world application of PFL.

FedSDR is our first attempt to address the train–test data distribution shift on each client by employing a provable shortcut discovery and removal method based on causal modeling. However, FedSDR requires knowledge of the available environments on each client, and its theoretical guarantees apply only in linear feature spaces. In practice, the requirement for available environment information on clients can increase the risk of data privacy leakage, while the assumption of a linear feature space may limit the applicability of FedSDR to federated learning systems where local datasets exhibit more complex causal structures (i.e., non-linear feature spaces). To address these two significant limitations, we propose an improved causally motivated shortcut-averse method that tackles both train–train and train–test data distribution shifts in federated learning systems, which will be presented in the next chapter.

Chapter 5

Causally Motivated Personalized Federated Invariant Learning with Shortcut-Averse Information-Theoretic Regularization

5.1 Introduction

Modern machine learning models are prone to rely on spurious correlations (correlations between spurious features and target, a.k.a, shortcuts) in diverse vision and language tasks [26]. Since shortcuts are unstable over diverse data distributions, models performing well on training data can experience a significant degradation in performance on test data when distribution shift exists. We consider a binary classification task for illustration where a learning model needs to differentiate between pictures of "cow" and "camel" [7]. Because most cows stand with grass backgrounds and the majority of camels appear in desert backgrounds in the practical training dataset, there is a shortcut from background representation to target/label. The trained learning model prefers to choose background (spurious feature) rather than the shape of animals (intended feature) as the discriminative feature. When images with camels standing in grass backgrounds arrive at inference stage, they will be categorized as "cow" because the spurious correlation is no longer applicable.

With the aim of learning intended features and eliminating spurious features, invariant learning (IL) emerges as one of the most effective and promising directions recently. Intended features are regarded as features that have an invariant causal relation to the target across various data distributions, consequently, they are referred to as invariant features. The prevalent IL methods necessitate exposure to multiple training environments[1] (i.e., heterogeneous data distributions) for producing an invariant predictor elicited from the invariant features. The obtained model can generalize to diverse unknown data distributions, and therefore resolve the out-of-distribution (OOD) generalization problem.

When we shift our focus to federated learning where the local datasets are usually non-independently and identically distributed (i.e., Non-IID), exploiting invariant representation across different data distributions can be facilitated. However, the heterogeneous federated clients present an additional significant demand: personalization, due to the fact that a shared global model can fail to fit the diverse local data distributions [39]. Now, a question arises: **Is personalization still necessary when we consider OOD generalization in federated learning?** Affirmative, the answer is yes. For example, federated clients collaborate to train disease diagnosis models using their data samples gathered from various hospitals. One aspect to consider is the target model needs to exhibit OOD generalization across diverse hospitals since test data on each client can be collected from different hospitals/environments. On the flip side, the individualized physical characteristics of each user/client consti-

[1] Environment refers to a data distribution specified by a latent variable in invariant learning.

tute essential information for personalized disease diagnosis and should be preserved.

Regrettably, personalized features and spurious features are closely entangled under PFL due to their similar variability across heterogeneous clients. On the one hand, federated invariant learning (e.g., [31]) fails to develop personalized models because personalized features are dropped along with spurious features. On the other hand, existing PFL methods can hardly mitigate spurious correlation when preserving personalization information is necessary (e.g., [94, 67, 102]). Furthermore, empirical results indicate a concerning tendency of the prevalent personalization schemes to favor the selection of spurious features over personalized features (details are discussed in the evaluation part). In particular, FedSDR [97] devises a shortcut discovery and removal scheme to capture the personalized invariant features. However, the rigorous assumption that invariant and spurious features are separable in linear space hampers its effectiveness in more general scenarios.

To achieve provable personalized federated invariant learning (IL), we follow the solution concept of causally invariant learning and formulate heterogeneous structured causal model (SCM [75]) for federated clients. With the SCM extended from invariant learning, we propose a crucial causal signature where personalized invariant features can be distinguished from spurious features with global invariant features as the anchor. The global invariant features are captured through a global objective regularized by a constraint representing conditional independence that is commonly used in centralized IL. Subsequently, the principal causal signature is quantified as a shortcut-averse information-theoretic constraint which includes a conditional mutual information term and an information entropy term in the designed objective function. With this devised constraint, each client can effectively exploit the personalized invariant features and simultaneously exclude spurious correlations to achieve remarkable OOD generalization performance. Main contributions of this work are outlined as follows:

- We formulate heterogeneous structured causal model to interpret Non-IID data distributions across federated clients, and propose a crucial causal signature which is quantified as a shortcut-averse information-theoretic constraint in the local objective to achieve personalized invariant learning on each client. Besides, a practical and effective algorithm FedPIN is proposed to solve the devised optimization problem.

- Theoretically, we demonstrate that FedPIN can develop the optimal personalized invariant predictor for each client and provide a tighter generalization error bound compared with the state-of-the-art PFL methods. Moreover, we prove FedPIN can achieve a convergence rate on the same order as FedAvg [70].

- The experimental results on diverse datasets validate the superiority of FedPIN on OOD generalization performance, in comparison with the state-of-the-art FL and PFL competitors.

5.2 Related Work

Invariant Learning. Attaining causally invariant predictors over varied data distributions is proposed in the field of causal inference [76], and introduced into machine learning to tackle the OOD generalization problem by IRM [6]. Then, many efforts are dedicated to facilitating the application of IL to general scenarios. Some works focus on achieving invariant learning when environment label is unavailable, e.g., EIIL [19], HRM [61], KerHRM [62], EDNIL [40] and ZIN [59]. IFM [15] lowers the requirement on the number of available environments. Another branch [1, 14, 42] completes the constraints that IRM misses. Besides, iCaRL [65] extends IL to non-linear causal representations while ACTIR [45] extends IL to anti-causal scenarios. IL is also applied to graph representation learning [55, 16] and natural language modeling [77]. These methods are devised for centralized scenarios where all training data is accessible.

Federated Learning. The classic FedAvg [70] can perform well when local datasets are IID. A number of methods (e.g., SCAFFOLD [48], FedEM [22] and FedLC [110]) delve into alleviating the negative impact of training data heterogeneity on convergence rate, while another line [21, 87, 93] targets at reducing the performance bias of global model on local clients. Few works [63, 73, 31] investigate the scenarios where training data heterogeneity appears to be domain shift. These methods produce a shared global model which can hardly fit the Non-IID target datasets across clients.

Personalized Federated Learning (PFL). A typical strand of PFL methods train the personalized models with the guidance of a global model which embeds in the shared knowledge [94, 32, 34, 24, 57, 96, 17], while another branch studies the parameterized knowledge transfer between similar clients, e.g., MOCHA [90], FedAMP [41] and KT-pFL [109]. DFL [67] disentangles the shared features from the client-specific ones to achieve accurate aggregation on shared knowledge. Similarly, pFedPara [43] and Factorized-FL [44] factorizes the model parameters into the shared and personalized parts. FedRep [18], FedRoD [12] and FedPAC [102] employ the shared/aligned feature extractor to capture global knowledge and personalized classifiers to encode personalization information. Besides, FedSDR [97] proposes a provable shortcut discovery and removal method to extract personalized invariant features in linear feature space. However, the explicit shortcut discovery method renders that the server in FedSDR requires the knowledge of the available training environments on each client, which increases the risk of privacy leakage in federated learning.

5.3 Problem Formulation

Notations. Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{E}$ denote the input, target and environment space respectively. Data instance is $(X, y) \in (\mathcal{X}, \mathcal{Y})$. Suppose there are N clients and the local dataset on client u is D_u, $u \in [N]$. The sets of training and test environments on client u are denoted by $\mathcal{E}_{tr}^u$ and $\mathcal{E}_{te}^u$ respectively. We use $\mathcal{E}_{all}^u$ as the set of all possible environments in the task that client u concentrates on, i.e., $\mathcal{E}_{tr}^u$, $\mathcal{E}_{te}^u \subset \mathcal{E}_{all}^u$, $\forall u \in [N]$. In federated learning system, the overall environment sets are denoted by $\mathcal{E}_{tr} \triangleq \bigcup_u \mathcal{E}_{tr}^u$ and $\mathcal{E}_{all} \triangleq \bigcup_u \mathcal{E}_{all}^u$. For convenience, we separate the learning model or parameterized mapping from $\mathcal{X}$ to $\mathcal{Y}$ into two consecutive parts: **1)** the feature extractor (e.g., Φ denotes an invariant feature extractor) maps from input space $\mathcal{X}$ to latent feature space $\mathcal{Z}$, i.e., $\Phi(X) \in \mathcal{Z}$; **2)** the classifier ω outputs a prediction $\hat{y}$ from a latent feature $z \in \mathcal{Z}$. The overall model is denoted by $f_\theta(\cdot) = \omega(\Phi(\cdot))$ where f_θ indicates the function f parameterized by θ. We define the expected empirical loss for model f_θ on dataset D as $\mathcal{R}(f_\theta; D) := \mathbb{E}_{(X,y)\in D}[\ell(f_\theta(X), y)]$ where ℓ is the cross-entropy loss function in this chapter unless noted otherwise.

5.3.1 Invariant Learning (IL)

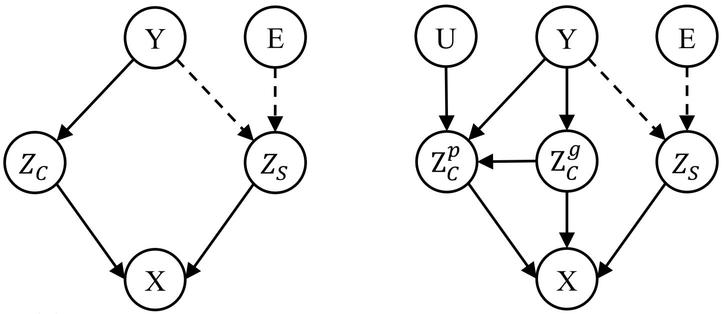
Invariant learning operates on an assumption that there exists invariant feature $\Phi(X)$ satisfying the ***invariance constraint:***

$$\mathbb{P}(Y|\Phi(X) = z, e) = \mathbb{P}(Y|\Phi(X) = z, e'), \forall z \in \mathcal{Z}, \forall e, e' \in \mathcal{E}_{all}. \tag{5.1}$$

Hence, the generic objective of invariant learning is to build an invariant feature extractor that fits the above invariance constraint. As Eq. (5.1) indicates a stable causal relation between invariant features $\Phi(X)$ and target Y, the invariant predictor elicited from the derived invariant feature extractor can tackle OOD generalization problem by achieving a consistent performance over various test data distributions. As a final point, we give the formal definition of the optimal invariant predictor in invariant learning.

Definition 5.3.1 (**Optimal Invariant Predictor**). *The optimal invariant predictor is elicited based on the complete invariant features that are informative for the target in the task, i.e.,* $\Phi^\star \in \arg\max_\Phi I(Y; \Phi(X))$ *where* $I(\cdot;\cdot)$ *denotes Shannon mutual* *information between two random variables and Φ satisfies the invariance constraint in Eq. (5.1).*

5.3.2 Causal Setup



(a) invariant learning (b) personalized FL

Figure 5.1: Graph (a) presents the structural causal model (SCM) generally adopted in invariant learning, e.g., [80, 45, 42], while (b) show the SCM proposed in FedPIN. Z_C and Z_S denote the invariant and spurious features respectively. E is the indicator of shortcut while U is the indicator of user/client. Dotted arrows indicate unstable causal relations that can vary in different environments.

Invariant learning usually formulates a structural causal model to simulate the data generating process in concerned task. A valid SCM is depicted by a directed acyclic graph where each node represents a random variable and each edge describes a directed functional relationship between the corresponding variables [75]. When we study causal invariance in PFL, the heterogeneity among data generating mechanisms on local clients needs to be considered.

Therefore, we construct the SCM in heterogeneous federated learning by adding the **U**ser/client indicator U which serves as the source of personalization information and extending the invariant features to two related parts: the personalized invariance Z_C^p and the shared/global invariance Z_C^g. The detailed SCM is shown in Figure 5.1. It is noted that the personalized invariance Z_C^p embeds all the invariant features on a local client, including both the exclusive individual invariant information that originates from variable U and the shared invariant knowledge represented by Z_C^g. Thus, there are causal relations from U to Z_C^p and from Z_C^g to Z_C^p. As discussed in IL, Z_S denotes spurious features. The functional relation between Z_S and Y can vary across different environments. By analogy with Definition 5.3.1 in invariant learning, we provide the definition of the optimal personalized invariant predictor in PFL.

Definition 5.3.2 (**Optimal Personalized Invariant Predictor**). *The optimal personalized invariant predictor for client u is elicited based on the complete invariant features which are informative for the target in the task that client u concentrates on, i.e., $\Phi_u^\star \in \arg\max_{\Phi_u} I(Y;\Phi_u(X))$, where Φ_u satisfies that $\mathbb{P}(Y|\Phi_u(X)=z,e)=\mathbb{P}(Y|\Phi_u(X)=z,e'),\forall z\in\mathcal{Z},\forall e,e'\in\mathcal{E}_{all}^u$.*

5.4 Methodology

To handle the outstanding challenge that personalization information is closely entangled with spurious features, we resort to causal characteristics to differentiate them.

Lemma 5.4.1. *If the data generating mechanism on each federated client complies with the causal graph in Figure 5.1(b) and the data distribution satisfies the Markov property, then the following two statements hold:*

- $[Z_C^p, Z_C^g] \perp\!\!\!\perp Z_S \mid Y$ *and* $Z_C^p \not\perp\!\!\!\perp Z_C^g \mid Y$*, which means both the global* (Z_C^g) *and personalized* (Z_C^p) *invariant features are conditionally independent of the shortcut features* Z_S *given* Y *while* Z_C^p *is not conditionally independent of* Z_C^g *given* Y*;*
- $[E,U] \perp\!\!\!\perp Y \mid Z_C^g$*, which means every component in the variable set* $[E,U]$ *is conditionally independent of the target* Y *given* Z_C^g*.*

Proof. According to the d-separation criterion in [75] we can find the variable Y d-separates Z_S from both Z_C^g and Z_C^p while the direct causal path from Z_C^g to Z_C^p is never blocked by variable Y in the given SCM. Therefore, the correctness of the first claim is granted. Besides, $[E, U] \perp\!\!\!\perp Y \mid Z_C^g$ holds since the variable Z_C^g d-separates Y from both the environment indicator E and the user/client indicator U. □

Upon the first claim, we can get the crucial causal signature: $Z_S \perp\!\!\!\perp Z_C^g \mid Y$ while $Z_C^p \not\perp\!\!\!\perp Z_C^g \mid Y$ to distinguish the personalized invariant features from spurious features with the anchor Z_C^g. Moreover, the second claim indicates the anchor Z_C^g (i.e., global invariant features) can be extracted via collaborative invariant learning among federated clients. In conclusion, Lemma 5.4.1 demonstrates the feasibility of achieving personalized invariant learning under FL.

5.4.1 Global Objective: Anchor Construction

Since the causal signature $[E, U] \perp\!\!\!\perp Y \mid Z_C^g$ is related to the client indicator U, the anchor Z_C^g needs to be captured in a collaborative manner. Although the recent work FedIIR [31] can develop a global invariant feature extractor, it can only guarantee to draw the global invariant features in linear feature space. This notable limitation is inherited from IRM [6] because the objective in FedIIR is a federated variant of that in IRM. Considering the above limitation can hinder the application of FedIIR to more complex cases, we choose to devise an information-theoretic regularization which can perform well in general cases to build the global invariant extractor.

Specifically, we quantify the causal signature $[E, U] \perp\!\!\!\perp Y \mid Z_C^g$ as a regularization term in the global objective function. Due to the equivalence of $[E, U] \perp\!\!\!\perp Y \mid Z_C^g$ to $I(E, U; Y \mid Z_C^g) = 0$, we can give a trivial global objective:

$$\max_{\Phi_g} I(Y; \Phi_g(X)) - \alpha I(E, U; Y \mid \Phi_g(X)), \tag{5.2}$$

where $I(\cdot; \cdot \mid \cdot)$ denotes the conditional mutual information, and α is a non-negative balancing weight. The first term in the above objective is utilized to filter out the non-informative components (e.g., noise) with regard to the target. We can achieve maximizing it via minimizing the cross-entropy loss in practical optimization. As regard to the second term $I(E, U; Y \mid \Phi_g(X))$, it can be computed effectively utilizing the equation provided in Proposition 5.4.1.

Proposition 5.4.1. *Suppose the heterogeneous data distributions across federated clients are independently caused by the variable U and E, that is $E \perp\!\!\!\perp U$ holds in the FL system, then we have*

$$I(E, U; Y \mid \Phi_g(X)) = \min_{\omega_g} \mathbb{E}_u[\mathcal{R}(\omega_g(\Phi_g); D_u)] - \min_{\omega_a} \mathbb{E}_u[\mathcal{R}(\omega_a(\Phi_g, u); D_u)] \tag{5.3}$$

where the global invariant classifier ω_g accepts global features $\Phi_g(X)$ as input while the auxiliary classifier ω_a takes both global features $\Phi_g(X)$ and user/client index u as input.

Proof. We know that the conditional mutual information $I(E, U; Y \mid \Phi_g(X))$ can be written as

$$I(E, U; Y \mid \Phi_g(X)) = H(Y \mid \Phi_g(X)) - H(Y \mid E, U, \Phi_g(X)) \tag{5.4}$$

As discussed in [25], with the universal approximation ability of neural networks, the first term in the above equation can be expressed by

$$H(Y \mid \Phi_g(X)) = \min_{\omega_g} \mathbb{E}_{(X,y)}[\ell(\omega_g(\Phi_g(X)), y)]$$

while the second term can be described using

$$H(Y \mid \Phi_g(X), E, U) = \min_{\omega} \mathbb{E}_u \mathbb{E}_e[\ell(\omega(\Phi_g(X), u, e), y)].$$

Since the heterogeneous data distributions across federated clients are independently caused by the variable U and E, we have that $\mathbb{E}_u[\mathcal{R}(f; D_u)] = \mathbb{E}_u \mathbb{E}_e[\ell(f(X), y; e)]$. Therefore, $\mathbb{E}_{(X,y)}[\ell(f(X), y)] = \mathbb{E}_u[\mathcal{R}(f; D_u)]$ and $\min_{\omega} \mathbb{E}_u \mathbb{E}_e[\ell(\omega(\Phi_g(X), u, e), y)] = \min_{\omega_a} \mathbb{E}_u[\mathcal{R}(\omega_a(\Phi_g(X), u); D_u)]$. To summarize, we can get

$$\begin{aligned} I(E, U; Y \mid \Phi_g(X)) &= H(Y \mid \Phi_g(X)) - H(Y \mid E, U, \Phi_g(X)) \\ &= \min_{\omega_g} \mathbb{E}_u[\mathcal{R}(\omega_g(\Phi_g); D_u)] - \min_{\omega_a} \mathbb{E}_u[\mathcal{R}(\omega_a(\Phi_g, u); D_u)], \forall u \in [N]. \end{aligned}$$

Proof ends. □

Therefore, the tractable global objective to construct the global invariant feature extractor ($\Phi_g^\star$) is given by

$$\Phi_g^\star, \omega_g^\star, \omega_a^\star = \arg\min_{\Phi_g, \omega_g, \omega_a} \mathcal{L}_{glob}(\Phi_g, \omega_g, \omega_a), \tag{5.5}$$

$$\mathcal{L}_{glob}(\Phi_g, \omega_g, \omega_a) \triangleq \mathbb{E}_u[\mathcal{R}(\omega_g(\Phi_g); D_u)] + \alpha I(E, U; Y \mid \Phi_g(X)).$$

The following theorem demonstrates the effectiveness of the above objective function.

Theorem 5.4.1. *Assuming that $\forall u \in [N]$, the data instance $(X, y) \in D_u$ is randomly taken from the joint distribution $\mathbb{P}(X, Y \mid U = u)$ which is subject to the SCM in Figure 5.1(b), then the following two statements are equivalent:*

- *$\Phi_g^\star(X)$ depends and only depends on the complete global invariant features Z_C^g. That is, $\Phi_g^\star(X)$ is a function of Z_C^g alone;*
- *$\Phi_g^\star$ is the minimizer of the objective in Eq. (5.5) with an appropriately chosen hyper-parameter α.*

Proof. We firstly prove that the regularization term $I(E, U; Y \mid \Phi_g(X)) = 0$ is equivalent to that $\Phi_g(X)$ depends and only depends on the complete global invariant features Z_C^g.

Necessity: When $\Phi_g(X)$ depends and only depends on the complete global invariant features Z_C^g, we have that $[U, E] \perp\!\!\!\perp Y \mid \Phi_g(X)$ since $[U, E] \perp\!\!\!\perp Y \mid Z_C^g$. We know that $I(E, U; Y \mid \Phi_g(X)) = 0$ is equivalent to $[U, E] \perp\!\!\!\perp Y \mid \Phi_g(X)$, therefore the necessity is justified.

Sufficiency: Next, we will prove that $I(E, U; Y \mid \Phi_g(X)) = 0$ can guarantee $\Phi_g(X)$ is either a function of Z_C^g alone or a constant for all inputs. We will validate the sufficiency by constructing contradiction:

Assuming that there exists a feature extractor Φ_a such that $I(E, U; Y \mid \Phi_a(X)) = 0$ holds and $\Phi_a(X)$ depends on some $Z_a \subseteq [Z_C^p, Z_S]$ (and is not trivially a constant function). We know $I(E, U; Y \mid \Phi_a(X)) = 0$ is equivalent to $[U, E] \perp\!\!\!\perp Y \mid \Phi_a(X)$ which indicates that the following equation holds:

$$\mathbb{P}(Y \mid \Phi_a(X) = z, v) = \mathbb{P}(Y \mid \Phi_a(X) = z, v'), \forall z \in \mathcal{Z}, \forall v, v' \in [E, U]$$

For simplicity, we define that $V \triangleq [E, U]$. Since a cause of Z_C^p is U and E is a cause of Z_S, there exists at least one Z_C^g and some $v \in [E, U]$ make $0 < \mathbb{P}(Z_a = z_a \mid V = v, Z_C^g = z^\star) < 1$ hold. Now consider a set of input S_X such that $\Phi_a(X) = h(Z_C^g = Z^\star, Z_a)$ remains true for any $X \in S_X$, where h represents a deterministic mapping function. According to the definition of Z_a, we have that there always exists two v_1 and v_2 such that $\mathbb{P}(Y \mid Z_a = z_a, V = v_1) \neq \mathbb{P}(Y \mid Z_a = z_a, V = v_2), \forall z_a$. Because $h(\cdot)$ is a deterministic function and Z_C^g remains unchanged on S_X, we can derive that $\mathbb{P}(Y \mid \Phi_a(X), v_1) \neq \mathbb{P}(Y \mid \Phi_a(X), v_2)$ holds for any $X \in S_X$. Hence a contradiction with $[U, E] \perp\!\!\!\perp Y \mid \Phi_a(X)$ appears and a feature extractor satisfying $[U, E] \perp\!\!\!\perp Y \mid \Phi_g(X)$ cannot depends on any $Z_a \subseteq [Z_C^p, Z_S]$ and $\Phi_g(X)$ is a function of Z_C^g alone.

In the above part, we demonstrate the theoretical relation between Z_C^g and the regularization term $I(E, U; Y \mid \Phi_g(X)) = 0$. Following, we will prove that minimizing the expected risk $\mathbb{E}_u[\mathcal{R}(\omega_g(\Phi_g); D_u)]$ can guarantee the optimal solution $\omega_u^\star(\Phi_g^\star)$ ensures $\Phi_g^\star(X)$ only depends on Z_C^g and can maximize accuracy.

Since we adopt cross entropy as loss function ℓ, for any $u \in [N]$ and $e \in \mathcal{E}$, we can get $\min_{\omega_g} \mathcal{R}(\omega_g(\Phi_g^\star); e, u) = \mathbb{E}[Y \mid \Phi_g^\star(X), u, e]$ [68]. On the other hand, we have that $[U, E] \perp\!\!\!\perp Y \mid \Phi_g^\star(X)$. Therefore, for any u and e, we can get $\mathbb{E}[Y \mid \Phi_g^\star(X), u, e] = \mathbb{E}[Y \mid \Phi_g^\star(X)]$. Because $E \perp\!\!\!\perp U$ and the data instances in D_u is randomly sampled from some environment e, $\min_{\omega_g} \mathbb{E}_u[\mathcal{R}(\omega_g(\Phi_g^\star); D_u) = \mathbb{E}[Y \mid \Phi_g^\star(X)]$ holds. In other words, for any set of u and training dataset D_u that contains data samples from some environment e, $\mathbb{E}[Y \mid \Phi_g^\star(X)]$ is the optimal solution that minimizes the expected loss term $\mathbb{E}_u[\mathcal{R}(\omega_g(\Phi_g); D_u)]$.

Moreover, minimizing the expected loss, i.e., $\min_{\omega_g} \mathbb{E}_u[\mathcal{R}(\omega_g(\Phi_g^\star); D_u)$ can exclude the exception case where $\Phi_g^\star(X)$ is a constant for all input, although this exception case can also make the regularization term $I(E, U; Y \mid \Phi_g(X)) = 0$ hold.

Finally, using a Lagrangian multiplier, with an appropriately chosen value of α, minimizing the objective in Eq. (5.5) is equivalent to minimizing the following objective:

$$\begin{aligned} &\Phi_g^\star \in \arg\min_{\Phi_g, \omega_g} \mathbb{E}_u[\mathcal{R}(\omega_g(\Phi_g); D_u)] \\ &s.t. \qquad I(E, U; Y \mid \Phi_g(X)) = 0. \end{aligned} \tag{5.6}$$

Therefore, the two statements in Theorem 5.4.1 is equivalent to each other.

Proof ends. □

5.4.2 Local Objective: Personalized Invariant Learning

As mentioned above, the causal signature: $Z_S \perp\!\!\!\perp Z_C^g \mid Y$ while $Z_C^p \not\perp\!\!\!\perp Z_C^g \mid Y$ can be utilized to differentiate Z_C^p and Z_S. A question arises regarding how to exploit the derived anchor $\Phi_g^\star$ rather than the exact Z_C^g. The following lemma makes it possible to design a computable regularization for shortcut-averse personalized invariant learning, with the obtained anchor feature extractor $\Phi_g^\star$.

Lemma 5.4.2. *For any representation $h(X)$ and $h'(X)$ where h and h' are two functions, under the SCM in Figure 5.1(b), it can be concluded that:*

- *When $h(X)$ depends only on Z_C^p and $h'(X)$ depends only on Z_S, we can always obtain*

$$I(h(X); \Phi_g^\star(X) \mid Y) > I(h'(X); \Phi_g^\star(X) \mid Y) = 0.$$

- *When $h(X)$ depends only on Z_C^p and $h'(X)$ depends only on $[Z_C^p, Z_S]$, we can always obtain*

$$I(h'(X); \Phi_g^\star(X) \mid Y) \leq \max_h I(h(X); \Phi_g^\star(X) \mid Y).$$

Proof. We will provide detailed proofs of the two conclusions in this part sequentially.

Proof of the first conclusion: As claimed in Theorem 5.4.1, we know that $\Phi_g^\star(X)$ depends and only depends on the global invariant features Z_C^g. According to the proved causal signatures in Lemma 5.4.1, we have that $[Z_C^p, Z_C^g] \perp\!\!\!\perp Z_S \mid Y$ and $Z_C^p \not\perp\!\!\!\perp Z_C^g \mid Y$. Since function of independent variables are still independent, we can get $h(X) \not\perp\!\!\!\perp \Phi_g^\star(X) \mid Y$ and $h'(X) \perp\!\!\!\perp \Phi_g^\star(X) \mid Y$. Because $A \perp\!\!\!\perp B \mid C$ is equivalent to $I(A; B \mid C) = 0$ and mutual information is non-negative, we can write that $I(h(X); \Phi_g^\star(X) \mid Y) > I(h'(X); \Phi_g^\star(X) \mid Y) = 0$.

Proof of the second conclusion: According to the definition of conditional mutual information, for any function h' such that $h'(X)$ depends only on $[Z_C^p, Z_S]$, we can get

$$\begin{aligned} I(h'(X); \Phi_g^\star(X) \mid Y) &\leq I(Z_C^p, Z_S; \Phi_g^\star \mid Y) \\ &= H(Z_C^p, Z_S \mid Y) + H(\Phi_g^\star \mid Y) - H(Z_C^p, Z_S, \Phi_g^\star \mid Y) \end{aligned} \tag{5.7}$$

Using the d-separate criterion, we have that $Z_C^p \perp\!\!\!\perp Z_S \mid Y$. Furthermore, we can derive that

$$\begin{aligned} H(Z_C^p, Z_S \mid Y) &= \sum_y \sum_{z_c^p} \sum_{z_s} p(z_c^p, z_s, y) \log\left(p(z_c^p, z_s \mid y)\right) \\ &= \sum_y \sum_{z_c^p} \sum_{z_s} p(z_c^p, z_s, y) \log\left(p(z_c^p \mid y) p(z_s \mid y)\right) \\ &= \sum_y \sum_{z_c^p} \sum_{z_s} p(z_c^p, z_s, y) \log\left(p(z_c^p \mid y)\right) \\ &\quad + \sum_y \sum_{z_c^p} \sum_{z_s} p(z_c^p, z_s, y) \log\left(p(z_s \mid y)\right) \\ &= \sum_y \sum_{z_c^p} p(z_c^p, y) \log\left(p(z_c^p \mid y)\right) + \sum_y \sum_{z_s} p(z_s, y) \log\left(p(z_s \mid y)\right) \\ &= H(Z_C^p \mid Y) + H(Z_S \mid Y) \end{aligned}$$

Since $\Phi_g^\star(X)$ is a function of Z_C^g alone, we have that $Z_S \perp\!\!\!\perp \Phi_g^\star \mid Y$. Moreover, Using the d-separate criterion in Figure 5.1(b), we can have that $Z_S \perp\!\!\!\perp Z_C^p \mid [\Phi_g^\star, Y]$. Thus, we can get that

$$\begin{aligned}
H(Z_C^p, Z_S, \Phi_g^\star \mid Y) &= \sum_y \sum_{z_g} \sum_{z_c^p} \sum_{z_s} p(z_s, z_c^p, z_g, y) \log\left(\frac{p(z_s, z_c^p, z_g, y)}{p(y)}\right) \\
&= \sum_y \sum_{z_g} \sum_{z_c^p} \sum_{z_s} p(z_s, z_c^p, z_g, y) \log\left(\frac{p(z_s, z_c^p \mid z_g, y) p(z_g, y)}{p(y)}\right) \\
&= \sum_y \sum_{z_g} \sum_{z_c^p} \sum_{z_s} p(z_s, z_c^p, z_g, y) \log\left(\frac{p(z_s \mid z_g, y) p(z_c^p \mid z_g, y) p(z_g, y)}{p(y)}\right) \\
&= \sum_y \sum_{z_g} \sum_{z_c^p} \sum_{z_s} p(z_s, z_c^p, z_g, y) \log\left(\frac{p(z_s, z_g, y) p(z_c^p, z_g, y)}{p(y) p(z_g, y)}\right) \\
&= \sum_y \sum_{z_g} \sum_{z_c^p} \sum_{z_s} p(z_s, z_c^p, z_g, y) \Big(\log\left(p(z_s, z_g \mid y)\right) \\
&\quad + \log\left(p(z_c^p, z_g \mid y)\right) - \log\left(p(z_g \mid y)\right) \Big) \\
&= H(Z_S, \Phi_g^\star \mid Y) + H(Z_C^p, \Phi_g^\star \mid Y) - H(\Phi_g^\star \mid Y) \\
&= H(Z_S \mid Y) + H(\Phi_g^\star \mid Y) + H(Z_C^p, \Phi_g^\star \mid Y) - H(\Phi_g^\star \mid Y) \\
&= H(Z_S \mid Y) + H(Z_C^p, \Phi_g^\star \mid Y)
\end{aligned}$$

Substituting the above two equations into the inequality (5.7), we can get

$$\begin{aligned}
I(h'(X); \Phi_g^\star(X) \mid Y) &\leq H(Z_C^p \mid Y) + H(\Phi_g^\star \mid Y) - H(Z_C^p, \Phi_g^\star \mid Y) \\
&= I(Z_C^p; \Phi_g^\star(X) \mid Y) = \max_h I(h(X); \Phi_g^\star(X) \mid Y).
\end{aligned}$$

Proof of Lemma 5.4.2 ends. □

As part of a qualitative analysis, we can exclude the spurious features Z_S by adopting $I(\Phi_u(X); \Phi_g^\star(X) \mid Y) - H(\Phi_u(X))$ as a regularization term, where $H(\cdot)$ denotes the Shannon information entropy. On the one hand, the first conclusion in Lemma 5.4.2 signifies the rationality of maximizing the term $I(\Phi_u(X); \Phi_g^\star(X) \mid Y)$. On the other hand, the second conclusion in Lemma 5.4.2 suggests that adding any components of Z_S does not lead to an increase in $\max I(\Phi_u(X); \Phi_g^\star(X) \mid Y)$ but instead results in an increase in $H(\Phi_u(X))$. Therefore, maximizing the regularization $I(\Phi_u(X); \Phi_g^\star(X) \mid Y) - H(\Phi_u(X))$ can rule out the spurious features. Of course, the expected loss $\mathcal{R}(\omega_u(\Phi_u); D_u)$ is also necessary for leveraging as many invariant features as possible. Specifically, the devised local objective to fully extract personalized invariant features for client u ($\forall u \in [N]$) is:

$$\min_{\Phi_u, \omega_u} \mathcal{R}(\omega_u(\Phi_u); D_u) - \lambda I(\Phi_u(X); \Phi_g^\star(X)|Y) + \gamma H(\Phi_u(X)), \tag{5.8}$$

where λ and γ are non-negative balancing weights.

We provide formal theoretical analysis on the effectiveness of the local objective (5.8) in the subsequent Theorem 5.4.2.

Theorem 5.4.2. *If $f^\star_{\theta_u} \triangleq \omega_u^\star(\Phi_u^\star)$ is the minimizer of objective (5.8) with the hyper-parameter λ and γ chosen appropriately, then $f^\star_{\theta_u}$ is the optimal personalized invariant predictor that satisfies Definition 5.3.2 for the client u, $\forall u \in [N]$.*

Proof. Before starting the proof, we firstly provide a useful proposition as follows:

Proposition 5.4.2 (Lemma 2 in [10])**.** *When we train a classifier conditioned on a feature extractor Φ with the data distribution $\mathcal{D}$, minimizing the cross-entropy loss $\mathcal{R}(\omega(\Phi); \mathcal{D})$ is equivalent to maximizing the mutual information $I(Y; \Phi(X))$ on $\mathcal{D}$.*

Firstly, we will prove that there exists some positive constant ρ such that the optimal solution of the following objective cannot depends on any components of Z_S:

$$\hat{\Phi}_u = \min_{\Phi_u} -I(\Phi_u(X); \Phi_g^\star(X) \mid Y) + \rho H(\Phi_u(X)) \tag{5.9}$$

We justify this claim by constructing contradiction:

Using the d-separate criterion in Figure 5.1(b), we have that $[Z_C^p, Z_C^g] \perp\!\!\!\perp Z_S \mid Y$. For simplicity, we define that $Z_C \triangleq [Z_C^p, Z_C^g]$. Suppose $\hat{\Phi}_u(X)$ depends on both Z_C and Z_S such that it can be expressed as $\hat{\Phi}_u(X) = g_z(AZ_C, BZ_S)$ where A and B are two constant coefficient matrix and g_z is a deterministic function.

Suppose $B \neq 0$, i.e., $\hat{\Phi}_u(X)$ depends on both Z_C and Z_S. For simplicity, we denote that $\hat{Z}_C \triangleq AZ_C$ and $\hat{Z}_S \triangleq BZ_S$. We know that, for any deterministic function g_z, $I(g_z(\hat{Z}_C, \hat{Z}_S); \Phi_g^\star \mid Y) \leq I(\hat{Z}_C, \hat{Z}_S; \Phi_g^\star \mid Y)$ and $H(g_z(\hat{Z}_C, \hat{Z}_S)) \leq H(\hat{Z}_C, \hat{Z}_S)$ where equality is achieved if and only if g_z is an invertible function. When the balancing weight ρ is appropriately chosen, there exists an invertible function g_z renders that $\hat{\Phi}_u = g_z(\hat{Z}_C, \hat{Z}_S)$. In this way, we can derive that

$$\begin{aligned} I(\hat{\Phi}_u(X); \Phi_g^\star(X) \mid Y) &= I(\hat{Z}_C, \hat{Z}_S; \Phi_g^\star \mid Y) \\ &= H(\hat{Z}_C, \hat{Z}_S \mid Y) + H(\Phi_g^\star \mid Y) - H(\hat{Z}_C, \hat{Z}_S, \Phi_g^\star \mid Y) \end{aligned}$$

According to the d-separate criterion, we can have that $\hat{Z}_C \perp\!\!\!\perp \hat{Z}_S \mid Y$. With this conditional independence held, we can write that

$$\begin{aligned} H(\hat{Z}_C, \hat{Z}_S \mid Y) &= \sum_y \sum_{\hat{z}_c} \sum_{\hat{z}_s} p(\hat{z}_c, \hat{z}_s, y) \log \big(p(\hat{z}_c, \hat{z}_s \mid y)\big) \\ &= \sum_y \sum_{\hat{z}_c} \sum_{\hat{z}_s} p(\hat{z}_c, \hat{z}_s, y) \log \big(p(\hat{z}_c \mid y) p(\hat{z}_s \mid y)\big) \\ &= \sum_y \sum_{\hat{z}_c} \sum_{\hat{z}_s} p(\hat{z}_c, \hat{z}_s, y) \log \big(p(\hat{z}_c \mid y)\big) \\ &\quad + \sum_y \sum_{\hat{z}_c} \sum_{\hat{z}_s} p(\hat{z}_c, \hat{z}_s, y) \log \big(p(\hat{z}_s \mid y)\big) \\ &= \sum_y \sum_{\hat{z}_c} p(\hat{z}_c, y) \log \big(p(\hat{z}_c \mid y)\big) + \sum_y \sum_{\hat{z}_s} p(\hat{z}_s, y) \log \big(p(\hat{z}_s \mid y)\big) \\ &= H(\hat{Z}_C \mid Y) + H(\hat{Z}_S \mid Y) \end{aligned}$$

Since $\Phi_g^\star(X)$ is a function of Z_C^g alone, we have that $\hat{Z}_S \perp\!\!\!\perp \Phi_g^\star \mid Y$. Moreover, Using the d-separate criterion in Figure 5.1(b), we can have that $\hat{Z}_S \perp\!\!\!\perp \hat{Z}_C \mid [\Phi_g^\star, Y]$. Thus, we can get that

$$\begin{aligned}
H(\hat{Z}_C, \hat{Z}_S, \Phi_g^\star \mid Y) &= \sum_y \sum_{z_g} \sum_{\hat{z}_c} \sum_{\hat{z}_s} p(\hat{z}_s, \hat{z}_c, z_g, y) \log \left(\frac{p(\hat{z}_s, \hat{z}_c, z_g, y)}{p(y)} \right) \\
&= \sum_y \sum_{z_g} \sum_{\hat{z}_c} \sum_{\hat{z}_s} p(\hat{z}_s, \hat{z}_c, z_g, y) \log \left(\frac{p(\hat{z}_s, \hat{z}_c \mid z_g, y) p(z_g, y)}{p(y)} \right) \\
&= \sum_y \sum_{z_g} \sum_{\hat{z}_c} \sum_{\hat{z}_s} p(\hat{z}_s, \hat{z}_c, z_g, y) \log \left(\frac{p(\hat{z}_s \mid z_g, y) p(\hat{z}_c \mid z_g, y) p(z_g, y)}{p(y)} \right) \\
&= \sum_y \sum_{z_g} \sum_{\hat{z}_c} \sum_{\hat{z}_s} p(\hat{z}_s, \hat{z}_c, z_g, y) \log \left(\frac{p(\hat{z}_s, z_g, y) p(\hat{z}_c, z_g, y)}{p(y) p(z_g, y)} \right) \\
&= \sum_y \sum_{z_g} \sum_{\hat{z}_c} \sum_{\hat{z}_s} p(\hat{z}_s, \hat{z}_c, z_g, y) \Big(\log \left(p(\hat{z}_s, z_g \mid y) \right) \\
&\quad + \log \left(p(\hat{z}_c, z_g \mid y) \right) - \log \left(p(z_g \mid y) \right) \Big) \\
&= H(\hat{Z}_S, \Phi_g^\star \mid Y) + H(\hat{Z}_C, \Phi_g^\star \mid Y) - H(\Phi_g^\star \mid Y) \\
&= H(\hat{Z}_S \mid Y) + H(\Phi_g^\star \mid Y) + H(\hat{Z}_C, \Phi_g^\star \mid Y) - H(\Phi_g^\star \mid Y) \\
&= H(\hat{Z}_S \mid Y) + H(\hat{Z}_C, \Phi_g^\star \mid Y)
\end{aligned}$$

combining the above two equations, we can get

$$\begin{aligned}
I(\hat{\Phi}_u(X); \Phi_g^\star(X) \mid Y) &= H(\hat{Z}_C \mid Y) + H(\Phi_g^\star \mid Y) - H(\hat{Z}_C, \Phi_g^\star \mid Y) \\
&= I(\hat{Z}_C; \Phi_g^\star \mid Y)
\end{aligned}$$

On the other hand, $H(\hat{\Phi}_u) = H(\hat{Z}_C, \hat{Z}_S) \geq H(\hat{Z}_S)$ and equality is achieved if and only if $B = 0$. Therefore, we have that $-I(\hat{\Phi}_u(X); \Phi_g^\star(X) \mid Y) + \rho H(\hat{\Phi}_u(X)) > -I(\hat{Z}_C; \Phi_g^\star(X) \mid Y) + \rho H(\hat{Z}_C)$ for any positive ρ, which indicates $\hat{\Phi}_u$ is not the minimizer of Eq. (5.9). Contradiction appears. Therefore, $B = 0$ must hold if $\hat{\Phi}_u$ is the minimizer of Eq. (5.9).

Because $\hat{\Phi}_u(X)$ cannot depend on any components of Z_S, using the d-separate criterion in Figure 5.1(b), we can get that $Y \perp\!\!\!\perp E \mid \hat{\Phi}_u$ which indicates that $\mathbb{P}(Y|\hat{\Phi}_u(X) = z, e) = \mathbb{P}(Y|\hat{\Phi}_u(X) = z, e'), \forall z \in \mathcal{Z}, \forall e, e' \in \mathcal{E}_{all}^u$.

Meanwhile, according to Proposition 5.4.2, we know that when data instances in D_u are randomly sampled from the true data distribution, minimizing $\mathcal{R}(\omega_u(\Phi_u); D_u)$ can guarantee that $I(\Phi_u(X);Y)$ is maximized.

Finally, we integrate the above theoretical output. Using a Lagrangian multiplier, with an appropriately chosen value of λ and γ, the minimizer of the objective in Eq. (5.8) (denoted by $\Phi_u^\star$) can guarantee that

- $\mathbb{P}(Y|\Phi_u^\star(X) = z, e) = \mathbb{P}(Y|\Phi_u^\star(X) = z, e'), \forall z \in \mathcal{Z}, \forall e, e' \in \mathcal{E}_{all}^u$;
- $I(\Phi_u^\star(X);Y) = \max I(\Phi_u(X);Y)$.

Hence, the proof of Theorem 5.4.2 is complete. □

Considering both $I(\Phi_u(X);\Phi_g^\star(X) \mid Y)$ and $H(\Phi_u(X))$ are difficult to calculate in practice, we exploit a tractable upper bound of $-\lambda I(\Phi_u(X);\Phi_g^\star(X) \mid Y)+\gamma H(\Phi_u(X))$ to construct the practical objective function.

Proposition 5.4.3. *When the local batch on client u is B_u and $\Psi_u^\star, \omega_{\psi_u}^\star$ is optimized by $\Psi_u^\star, \omega_{\psi_u}^\star = \min_{\Psi_u,\omega_{\psi_u}} \mathcal{R}(\omega_{\psi_u}(\Psi_u); D_u)$, $\forall u \in [N]$, the following inequality holds:*

$$\begin{aligned} &- \lambda I(\Phi_u(X);\Phi_g^\star(X) \mid Y) + \gamma H(\Phi_u(X)) \\ &\leq \lambda \mathcal{L}_{con}^B(\Phi_u;\Phi_g^\star,\Psi_u^\star) + \gamma Var(\Phi_u(X)) - \lambda \log(|B_u|+1), \end{aligned} \tag{5.10}$$

where $Var(\Phi_u(X))$ represents the variance of $\Phi_u(X)$ and $|B_u|$ is the batch size. $\mathcal{L}_{con}^B(\Phi_u;\Phi_g^\star,\Psi_u^\star)$ is a contrastive loss defined by

$$-\mathop{\mathbb{E}}_{X\in D_u}\left[\log \frac{e^{sim(\Phi_u(X),\Phi_g^\star(X))/\tau}}{e^{sim(\Phi_u(X),\Phi_g^\star(X))/\tau} + \sum_{X\in B_u} e^{sim(\Phi_u(X),\Psi_u^\star(X))/\tau}}\right],$$

where $sim(z,z') = \frac{z^\top z'}{\|z\|\|z'\|}$ is the cosine similarity and τ denotes a temperature parameter. They are commonly used in the design of contrastive loss [13].

Proof. In each local batch B_u, the contrastive loss is constructed via regarding $\Phi_u^\star(X)$ and $\Phi_u(X)$ as positive pair while adopting $\Psi_u^\star(X), X \in B_u$ as negative samples. Therefore, the number of negative samples in the devised contrastive loss is $|B_u|$.

Using the results proved in Proposition 1 in [91], we can get that the conditional mutual information satisfies $I(\Phi_u(X); \Phi_g^\star(X) \mid Y) \geq \log(|B_u|+1) - \mathcal{L}_{con}^B(\Phi_u; \Phi_g^\star, \Psi_u^\star)$. On the other hand, according to the proof of Proposition 4 in [50], we know that $H(\Phi_u(X)) \leq Var(\Phi_u(X))$. Therefore, the proposed information-theoretic regularization term can be upper bounded as follows, for any non-negative constant λ and γ:

$$\begin{aligned} & -\lambda I(\Phi_u(X); \Phi_g^\star(X) \mid Y) + \gamma H(\Phi_u(X)) \\ & \leq \lambda \mathcal{L}_{con}^B(\Phi_u; \Phi_g^\star, \Psi_u^\star) + \gamma Var(\Phi_u(X)) - \lambda \log(|B_u|+1) \end{aligned}$$

Proof of Proposition 5.4.3 ends. □

In the proposed contrastive loss, we treat the personalized invariant feature $\Phi_u(X)$ and the global invariant feature $\Phi_g^\star(X)$ as a positive pair while the features drawn from the local batch by $\Psi_u^\star$ are regarded as negative examples. In consequence, the tractable local objective on client u is

$$\min_{\Phi_u, \omega_u} \mathcal{R}(\omega_u(\Phi_u); D_u) + \lambda \mathcal{L}_{con}^B(\Phi_u; \Phi_g^\star, \Psi_u^\star) + \gamma Var(\Phi_u(X)). \tag{5.11}$$

5.4.3 Algorithm Design

In federated learning system, the global objective in Eq. (5.5) can be partitioned into N sub-problems:

$$\min_{\Phi_g, \omega_g, \omega_a} \mathcal{L}_{glob}(\Phi_g, \omega_g, \omega_a) = \frac{1}{N} \sum_{u=1}^{N} \mathcal{L}_g^u(\Phi_g, \omega_g, \omega_a)$$

$$\mathcal{L}_g^u(\Phi_g, \omega_g, \omega_a) = (1+\alpha)\mathcal{R}(\omega_g(\Phi_g); D_u) - \alpha \mathcal{R}(\omega_a(\Phi_g, u); D_u).$$

Furthermore, the local update (e.g., model parameters and gradients) for the global objective is obtained by solving the sub-objective $\mathcal{L}_g^u(\Phi_g, \omega_g, \omega_a)$ based on local dataset D_u, $\forall u \in [N]$. After the selected local clients conduct stochastic gradient descent for several local iterations, the server will aggregate the uploaded local updates and then broadcast the aggregated global model to the participating clients as in most federated learning algorithms.

Algorithm 5 Fed**PIN**: **P**ersonalized **I**nvariant Lear**N**ing

```
Input: Hyper-parameters $T$, $R$, $K$, $\beta$, $\eta$, $\alpha$, $\lambda$, $\gamma$.
Initialize models: $\omega_g^0(\Phi_g^0)$, $\omega_a^0$ and $\{\omega_u^0(\Phi_u^0)|u \in [N]\}$.
for $t = 0$ to $T - 1$ do
    Server randomly selects a client subset $\mathcal{A}_t$.    ▷ Client selection
    Server broadcasts global models $\omega_g^t(\Phi_g^t)$ and $\omega_a^t$ to all clients in $\mathcal{A}_t$.
    for each client $u \in \mathcal{A}_t$ in parallel do    ▷ Local update
        Update $\omega_{\psi_u}(\Psi_u)$ for $K$ local steps:
            $\omega_{\psi_u}(\Psi_u) = \omega_{\psi_u}(\Psi_u) - \eta\nabla\mathcal{R}(\omega_{\psi_u}(\Psi_u); D_u)$
        Update $\omega_u(\Phi_u)$ for $K$ local steps with $\Phi_g^t$ and $\Psi_u$:
            $\omega_u(\Phi_u) = \omega_u(\Phi_u) - \eta\nabla\mathcal{L}_{loc}^u(\omega_u(\Phi_u); \Phi_g^t, \Psi_u)$
        Initialize $\tilde{\omega}_g^u(\tilde{\Phi}_g^u) = \omega_g^t(\Phi_g^t)$ and $\tilde{\omega}_a^u = \omega_a^t$.
        for $r = 0$ to $R - 1$ do    ▷ Solve the sub-problem of $\mathcal{L}_{glob}(\Phi_g, \omega_g, \omega_a)$
            $\tilde{\omega}_g^u, \tilde{\Phi}_g^u, \tilde{\omega}_a^u = \tilde{\omega}_g^u, \tilde{\Phi}_g^u, \tilde{\omega}_a^u - \beta\nabla\mathcal{L}_g^u(\tilde{\omega}_g^u, \tilde{\Phi}_g^u, \tilde{\omega}_a^u)$
        end for
        Send $\tilde{\omega}_g^u(\tilde{\Phi}_g^u)$ and $\tilde{\omega}_a^u$ back to the server.
    end for
    Server aggregates $\{\tilde{\omega}_g^u(\tilde{\Phi}_g^u), \tilde{\omega}_a^u|u \in \mathcal{A}_t\}$:    ▷ Global aggregation
      $\omega_g^{t+1}(\Phi_g^{t+1}) = \frac{1}{|\mathcal{A}_t|}\sum_{u\in\mathcal{A}_t}\tilde{\omega}_g^u(\tilde{\Phi}_g^u)$;
      $\omega_a^{t+1} = \frac{1}{|\mathcal{A}_t|}\sum_{u\in\mathcal{A}_t}\tilde{\omega}_a^u$.
end for
return personalized invariant models $\{\omega_u(\Phi_u)|u \in [N]\}$.
```

As for the local objective, it can be solved locally with the received global invariant feature extractor Φ_g^t. To simplify the expressions, we rewrite the local objective as

$$\mathcal{L}_{loc}^{u}(\omega_u(\Phi_u); \Phi_g^\star, \Psi_u^\star) \triangleq \mathcal{R}(\omega_u(\Phi_u); D_u) + \lambda \mathcal{L}_{con}^{B}(\Phi_u, \Phi_g^\star, \Psi_u^\star) + \gamma Var(\Phi_u(X)),$$

where the feature extractor $\Psi_u^\star$ is derived by minimizing $\mathcal{R}(\omega_{\psi_u}(\Psi_u); D_u)$ locally on client u, $\forall u \in [N]$. The detailed algorithm FedPIN is shown in Algorithm 5.

5.5 Theoretical Analysis

5.5.1 Generalization Error Bound

Along the information flow in a personalized learning model $\omega_u(\Phi_u)$, we can evaluate the effectiveness of the personalized feature extractor Φ_u in predicting the target Y using the mutual information $I(Y; \Phi_u(X))$. In practice, we can acquire the empirical estimation of $I(Y; \Phi_u(X))$ on the training dataset D_u, represented as $\hat{I}_{\mathcal{S}}(Y; \Phi_u(X))$. When the learning model is ready for deployment, we prioritize the performance of Φ_u on some unknown test data distribution, denoted by $I_{\mathcal{T}}(Y; \Phi_u(X))$. Since $I_{\mathcal{T}}(Y; \Phi_u(X))$ is inaccessible, bounding the generalization error $I_{\mathcal{T}}(Y; \Phi_u(X)) - \hat{I}_{\mathcal{S}}(Y; \Phi_u(X))$ is critical for analysing the generalization performance of $\omega_u(\Phi_u)$ in learning theory.

Theorem 5.5.1. *Suppose the training and test data distributions on each client u are denoted by $\mathbb{P}_{\mathcal{S}}(X, Y \mid U = u)$ and $\mathbb{P}_{\mathcal{T}}(X, Y \mid U = u)$, respectively. If the size of training dataset D_u is m_u, for any $u \in [N]$, there exists a constant C that makes the following inequality hold with a probability at least $1 - \delta$:*

$$\begin{aligned}&\left|\hat{I}_{\mathcal{S}}(Y; \Phi_u(X)) - I_{\mathcal{T}}(Y; \Phi_u(X))\right| \\ &\leq \underbrace{\frac{\sqrt{C \log(|\mathcal{Y}|/\delta)}\Big(|\mathcal{X}| \log(m_u) + |\mathcal{Y}| \hat{H}(\Phi_u(X))\Big) + \frac{2}{e}|\mathcal{X}|}{\sqrt{m_u}}}_{\textbf{\textit{IND generalization term}}} + \underbrace{\mathcal{J}(\Phi_u) + \sqrt{C|\mathcal{Y}|\mathcal{J}(\Phi_u)}}_{\textbf{\textit{OOD generalization term}}},\end{aligned}$$

where $m_u \geq \frac{C}{4}\log(|\mathcal{Y}|/\delta)|\mathcal{X}|e^2$ and $\hat{H}(\Phi_u(X))$ denotes the estimation of the entropy $H(\Phi_u(X))$ on training dataset D_u. 'IND' and 'OOD' represents 'in-distribution' and 'out-of-distribution' respectively. $\mathcal{J}(\Phi_u)$ denotes the Jeffrey's divergence defined by

$$\mathcal{J}(\Phi_u) \triangleq \mathcal{KL}\big(\mathbb{P}_{\mathcal{T}}(Y \mid \Phi_u(X)) \| \mathbb{P}_{\mathcal{S}}(Y \mid \Phi_u(X))\big) + \mathcal{KL}\big(\mathbb{P}_{\mathcal{S}}(Y \mid \Phi_u(X)) \| \mathbb{P}_{\mathcal{T}}(Y \mid \Phi_u(X))\big),$$

where $\mathcal{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence.

Proof. In practice, the available training data samples on each client are limited, that is the size of $D_u, \forall u \in [N]$ is finite. We will use $\mathcal{D}_u$ and $\mathcal{D}_u^{\mathcal{T}}$ to denote the true training and test data distributions that the training and test data instances are taken from, respectively. Besides, we denote the empirical probability distribution described by the training dataset D_u by $\hat{p}_u$ and the true probability distribution on $D_u^{\mathcal{T}}$ by p_u, $\forall u \in [N]$.

Proposition 5.5.1 (Lemma 11 [85]). *Let p be a distribution vector of arbitrary (possible countably infinite) cardinality, and $\hat{p}$ be an empirical estimation of p based on a dataset of size m. Then with a probability of at least $1-\delta$ over the samples, the following inequality holds:*

$$\|p - \hat{p}\| \leq \frac{2 + \sqrt{2\log(1/\delta)}}{\sqrt{m}} \tag{5.12}$$

For simplicity, we denote the empirical values of the statistical metrics by symbols with a hat while the true values of the statistical metrics by symbols without a hat (e.g., the empirical distribution $\hat{p}$ and the true distribution p). The values of the statistical metrics on the training data are represented by symbols with a subscript S while the values of the statistical metrics on the test data are represented by symbols with a subscript T. For example, we denote the mutual information between X and Y which is computed on data distribution $\hat{p}_{\mathcal{S}}$, $\hat{p}_{\mathcal{T}}$, $p_{\mathcal{S}}$ and $p_{\mathcal{T}}$ by $\hat{I}_{\mathcal{S}}(Y;X)$, $\hat{I}_{\mathcal{T}}(Y;X)$, $I_{\mathcal{S}}(Y;X)$ and $I_{\mathcal{T}}(Y;X)$, respectively.

Before starting the proof, we define a useful real-valued function ξ as follows:

$$\xi(x) = \begin{cases} 0, & x = 0 \\ x\log(\frac{1}{x}), & 0 < x \leq \frac{1}{e} \\ \frac{1}{e}, & x > \frac{1}{e} \end{cases} . \tag{5.13}$$

It is noted that $\xi(x)$ is a continuous, monotonically increasing and concave real-valued function.

In general, we consider a deterministic personalized feature extractor denoted by Φ_u. To enhance conciseness in written expression, we will use Φ_u to represent $\Phi_u(X)$ in this proof. Thus, we can write that

$$\begin{aligned} |\hat{I}_{\mathcal{S}}(Y;\Phi_u(X)) - I_{\mathcal{T}}(Y;\Phi_u(X))| &\triangleq |\hat{I}_{\mathcal{S}}(Y;\Phi_u) - I_{\mathcal{T}}(Y;\Phi_u)| \\ &= |\hat{I}_{\mathcal{S}}(Y;\Phi_u) - I_{\mathcal{S}}(Y;\Phi_u) + I_{\mathcal{S}}(Y;\Phi_u) - I_{\mathcal{T}}(Y;\Phi_u)| \\ &\leq \underbrace{|\hat{I}_{\mathcal{S}}(Y;\Phi_u) - I_{\mathcal{S}}(Y;\Phi_u)|}_{\mathcal{A}_1} + \underbrace{|I_{\mathcal{S}}(Y;\Phi_u) - I_{\mathcal{T}}(Y;\Phi_u)|}_{\mathcal{A}_2} \end{aligned} \tag{5.14}$$

We know that the mutual information $I(Y;\Phi)$ is defined by:

$$I(Y;\Phi) \triangleq H(\Phi) - H(\Phi \mid Y) \tag{5.15}$$

where $H(\cdot)$ represents the Shannon information entropy. We firstly deal with the first term in the above inequality:

$$\begin{aligned} \mathcal{A}_1 &= \left|\hat{H}_{\mathcal{S}}(\Phi_u) - H_{\mathcal{S}}(\Phi_u) + H_{\mathcal{S}}(\Phi_u \mid Y) - \hat{H}_{\mathcal{S}}(\Phi_u \mid Y)\right| \\ &\leq \left|H_{\mathcal{S}}(\Phi_u \mid Y) - \hat{H}_{\mathcal{S}}(\Phi_u \mid Y)\right| + \left|\hat{H}_{\mathcal{S}}(\Phi_u) - H_{\mathcal{S}}(\Phi_u)\right| \end{aligned} \tag{5.16}$$

For the first term on the right side of Eq. 5.16, we can write that

$$\begin{aligned} &|H_{\mathcal{S}}(\Phi_u \mid Y) - \hat{H}_{\mathcal{S}}(\Phi_u \mid Y)| \\ &= \left|\sum_y \big(p_{\mathcal{S}}(y)H_{\mathcal{S}}(\Phi_u \mid y) - \hat{p}_{\mathcal{S}}(y)\hat{H}_{\mathcal{S}}(\Phi_u \mid y)\big)\right| \\ &= \left|\sum_y \big(p_{\mathcal{S}}(y)H_{\mathcal{S}}(\Phi_u \mid y) - p_{\mathcal{S}}(y)\hat{H}_{\mathcal{S}}(\Phi_u \mid y) + p_{\mathcal{S}}(y)\hat{H}_{\mathcal{S}}(\Phi_u \mid y) - \hat{p}_{\mathcal{S}}(y)\hat{H}_{\mathcal{S}}(\Phi_u \mid y)\big)\right| \\ &\leq \left|\sum_y p_{\mathcal{S}}(y)\big(H_{\mathcal{S}}(\Phi_u \mid y) - \hat{H}_{\mathcal{S}}(\Phi_u \mid y)\big)\right| + \left|\sum_y \big(p_{\mathcal{S}}(y) - \hat{p}_{\mathcal{S}}(y)\big)\hat{H}_{\mathcal{S}}(\Phi_u \mid y)\right| \end{aligned}$$

The first term on the right side of the above inequality can be bounded by

$$\begin{aligned}
&\Big|\sum_y p_{\mathcal{S}}(y)\big(H_{\mathcal{S}}(\Phi_u \mid y) - \hat{H}_{\mathcal{S}}(\Phi_u \mid y)\big)\Big| \\
&\leq \Big|\sum_y p_{\mathcal{S}}(y)\sum_{\phi_u}\big(p_{\mathcal{S}}(\phi_u|y)\log(p_{\mathcal{S}}(\phi_u|y)) - \hat{p}_{\mathcal{S}}(\phi_u|y)\log(\hat{p}_{\mathcal{S}}(\phi_u|y))\big)\Big| \\
&\leq \sum_y p_{\mathcal{S}}(y)\sum_{\phi_u}\xi\big(|p_{\mathcal{S}}(\phi_u|y) - \hat{p}_{\mathcal{S}}(\phi_u|y)|\big) \\
&= \sum_y p_{\mathcal{S}}(y)\sum_{\phi_u}\xi\Big(\Big|\sum_x p_{\mathcal{S}}(\phi_u|x)\big(p_{\mathcal{S}}(x|y) - \hat{p}_{\mathcal{S}}(x|y)\big)\Big|\Big) \\
&= \sum_y p_{\mathcal{S}}(y)\sum_{\phi_u}\xi\Big(\Big|\sum_x \big(p_{\mathcal{S}}(\phi_u|x) - A\big)\big(p_{\mathcal{S}}(x|y) - \hat{p}_{\mathcal{S}}(x|y)\big)\Big|\Big) \\
&\leq \sum_y p_{\mathcal{S}}(y)\sum_{\phi_u}\xi\Big(\big\|p_{\mathcal{S}}(X|y) - \hat{p}_{\mathcal{S}}(X|y)\big\|\big\|p_{\mathcal{S}}(\phi_u|X) - A\big\|\Big)
\end{aligned}$$

where A can be any constant. When we set $A \triangleq \frac{1}{|X|}\sum_x p_{\mathcal{S}}(\phi_u|x)$, we can get

$$\begin{aligned}
&\Big|\sum_y p_{\mathcal{S}}(y)\big(H_{\mathcal{S}}(\Phi_u \mid y) - \hat{H}_{\mathcal{S}}(\Phi_u \mid y)\big)\Big| \\
&\leq \sum_y p_{\mathcal{S}}(y)\sum_{\phi_u}\xi\Big(\big\|p_{\mathcal{S}}(X|y) - \hat{p}_{\mathcal{S}}(X|y)\big\| \cdot \sqrt{V(p_{\mathcal{S}}(\phi_u|X))}\Big)
\end{aligned} \tag{5.17}$$

where $\frac{1}{|X|}V(p_{\mathcal{S}}(\phi_u|X))$ describes the variance of the vector $p_{\mathcal{S}}(\phi_u|X)$. It is known that $\hat{H}_{\mathcal{S}}(\Phi_u) \geq \hat{H}_{\mathcal{S}}(\Phi_u \mid y)$ for any y, since conditioning cannot increase entropy [85]. Therefore,

$$\begin{aligned}
\Big|\sum_y \big(p_{\mathcal{S}}(y) - \hat{p}_{\mathcal{S}}(y)\big)\hat{H}_{\mathcal{S}}(\Phi_u \mid y)\Big| &\leq \big\|p_{\mathcal{S}}(Y) - \hat{p}_{\mathcal{S}}(Y)\big\|\Big|\sum_y \hat{H}_{\mathcal{S}}(\Phi_u)\Big| \\
&= \big\|p_{\mathcal{S}}(Y) - \hat{p}_{\mathcal{S}}(Y)\big\|\big(|Y|\hat{H}_{\mathcal{S}}(\Phi_u)\big)
\end{aligned} \tag{5.18}$$

Combining Eq. (5.17) and Eq. (5.18), we can get

$$\begin{aligned}
H_{\mathcal{S}}(\Phi_u \mid Y) - \hat{H}_{\mathcal{S}}(\Phi_u \mid Y)| \leq &\sum_y p_{\mathcal{S}}(y)\sum_{\phi_u}\xi\Big(\big\|p_{\mathcal{S}}(X|y) - \hat{p}_{\mathcal{S}}(X|y)\big\| \cdot \sqrt{V(p_{\mathcal{S}}(\phi_u|X))}\Big) \\
&+ \big(|Y| \cdot \hat{H}_{\mathcal{S}}(\Phi_u)\big) \cdot \big\|p_{\mathcal{S}}(Y) - \hat{p}_{\mathcal{S}}(Y)\big\|
\end{aligned} \tag{5.19}$$

On the other hand, we have

$$
\begin{aligned}
\left|H_{\mathcal{S}}(\Phi_u) - \hat{H}_{\mathcal{S}}(\Phi_u)\right| &= \Big|\sum_{\phi_u} \left(p_{\mathcal{S}}(\phi_u)\log(p_{\mathcal{S}}(\phi_u)) - \hat{p}_{\mathcal{S}}(\phi_u)\log(\hat{p}_{\mathcal{S}}(\phi_u))\right)\Big| \\
&\leq \sum_{\phi_u} \xi\left(\left|p_{\mathcal{S}}(\phi_u) - \hat{p}_{\mathcal{S}}(\phi_u)\right|\right) \\
&= \sum_{\phi_u} \xi\Big(\Big|\sum_{x} p_{\mathcal{S}}(\phi_u|x)\left(p_{\mathcal{S}}(x) - \hat{p}_{\mathcal{S}}(x)\right)\Big|\Big) \\
&= \sum_{\phi_u} \xi\Big(\Big|\sum_{x} \left(p_{\mathcal{S}}(\phi_u|x) - A\right)\left(p_{\mathcal{S}}(x) - \hat{p}_{\mathcal{S}}(x)\right)\Big|\Big) \\
&\leq \sum_{\phi_u} \xi\Big(\left\|p_{\mathcal{S}}(X) - \hat{p}_{\mathcal{S}}(X)\right\| \cdot \sqrt{V(p_{\mathcal{S}}(\phi_u|X))}\Big)
\end{aligned}
\tag{5.20}
$$

where the constant A is chosen as $A \triangleq \frac{1}{|X|}\sum_x p_{\mathcal{S}}(\phi_u|x)$. Plugging Eq. (5.19) and Eq. (5.20) into Eq. (5.16), we can get

$$
\begin{aligned}
\mathcal{A}_1 \leq &\sum_{y} p_{\mathcal{S}}(y) \sum_{\phi_u} \xi\Big(\left\|p_{\mathcal{S}}(X|y) - \hat{p}_{\mathcal{S}}(X|y)\right\| \cdot \sqrt{V(p_{\mathcal{S}}(\phi_u|X))}\Big) \\
&+ \left(|Y| \cdot \hat{H}_{\mathcal{S}}(\Phi_u)\right) \cdot \left\|p_{\mathcal{S}}(Y) - \hat{p}_{\mathcal{S}}(Y)\right\| + \sum_{\phi_u} \xi\Big(\left\|p_{\mathcal{S}}(X) - \hat{p}_{\mathcal{S}}(X)\right\| \cdot \sqrt{V(p_{\mathcal{S}}(\phi_u|X))}\Big)
\end{aligned}
\tag{5.21}
$$

Subsequently, we can apply the concentration bound given in Proposition 5.5.1 to $\left\|p_{\mathcal{S}}(X|y) - \hat{y}_{\mathcal{S}}(X|y)\right\|$, $\left\|p_{\mathcal{S}}(X) - \hat{p}_{\mathcal{S}}(X)\right\|$ and $\left\|p_{\mathcal{S}}(Y) - \hat{p}_{\mathcal{S}}(Y)\right\|$ for any y in Eq. (5.21). To make sure the bounds hold simultaneously over these $|Y\| + 2$ quantities, we replace δ in Eq. (5.12) by $\delta/(|Y| + 2)$ as in the proof of Theorem 3 in [85]. Hence, with a probability at least $1 - \delta$ we have

$$
\begin{aligned}
\mathcal{A}_1 \leq 2 &\sum_{\phi_u} \xi\left(\left(2 + \sqrt{2\log((|Y| + 2)/\delta)}\right)\sqrt{\frac{V\left(p_{\mathcal{S}}(\phi_u|X)\right)}{m}}\right) \\
&+ \frac{2 + \sqrt{2\log\left((|Y| + 2)/\delta\right)}}{\sqrt{m}} \cdot \left(|Y|\hat{H}_{\mathcal{S}}(\Phi_u)\right)
\end{aligned}
\tag{5.22}
$$

There exists a small constant C that makes the following inequality hold:

$$
2 + \sqrt{2\log((|Y| + 2)/\delta)} \leq \sqrt{C\log(|Y|/\delta)}
$$

In addition, we know that the variance of any random variable that takes value in the range $[0,1]$ is at most $\frac{1}{4}$. Since $\frac{1}{|X|}\sum_x V\big(p_{\mathcal{S}}(\phi_u|X)\big)$ is the variance of the distribution vector $p_{\mathcal{S}}(\phi_u|X)$, we have that $V\big(p_{\mathcal{S}}(\phi_u|X)\big) \leq |X|/4$, $\forall \phi_u$.

Suppose that the size of training dataset (i.e., $m = |D_u|$) satisfying that

$$m \geq \frac{C}{4}\log(|Y|/\delta)|X|e^2 \tag{5.23}$$

Then, we can get

$$\sqrt{\frac{C\log(|Y|/\delta)V(p_{\mathcal{S}}(\phi_u|X))}{m}} \leq \sqrt{\frac{C\log(|Y|/\delta)|X|}{4m}} \leq \frac{1}{e}$$

We define that $\mathcal{V}(\phi_u) \triangleq C\log(|Y|/\delta)V(p_{\mathcal{S}}(\phi_u|X))$, then we have that

$$\begin{aligned}\sum_{\phi_u}\xi\Big(\sqrt{\frac{\mathcal{V}(\phi_u)}{m}}\Big) &= \sum_{\phi_u}\sqrt{\frac{\mathcal{V}(\phi_u)}{m}}\log\Big(\sqrt{\frac{\mathcal{V}(\phi_u)}{m}}\Big)\\ &= \sum_{\phi_u}\sqrt{\frac{\mathcal{V}(\phi_u)}{m}}\log(\sqrt{m}) + \sqrt{\frac{1}{m}}\sqrt{\mathcal{V}(\phi_u)}\log\Big(\frac{1}{\sqrt{\mathcal{V}(\phi_u)}}\Big)\\ &\leq \sum_{\phi_u}\left(\sqrt{\frac{\mathcal{V}(\phi_u)}{m}}\log(\sqrt{m}) + \frac{1}{\sqrt{m}e}\right)\end{aligned}$$

Using the results proved in the proof of Theorem 3 in [85], we can have that $\sum_{\phi_u}\sqrt{\mathcal{V}(\phi_u)} \leq \sqrt{|X||\Phi_u|}$. Therefore, we can write that

$$\sum_{\phi_u}\xi\left(\sqrt{\frac{C\log(|Y|/\delta)V(p_{\mathcal{S}}(\phi_u|X))}{m}}\right) \leq \frac{\sqrt{C\log(|Y|/\delta)|X||\Phi_u|}\log(m) + \frac{2}{e}|\Phi_u|}{2\sqrt{m}} \tag{5.24}$$

where $|\Phi_u|$ denote the size of the feature space from which ϕ_u takes value. Recalling that Φ_u is used to represent $\Phi_u(X)$ where Φ_u itself is a deterministic feature extractor, we can conclude that $|\Phi_u| \leq |X|$. Thus, we can get

$$\begin{aligned}\mathcal{A}_1 &\leq \frac{\sqrt{C\log(|Y|/\delta)}|X|\log(m) + \frac{2}{e}|X|}{\sqrt{m}} + \frac{\sqrt{C\log(|Y|/\delta)}|Y|\hat{H}_{\mathcal{S}}(\Phi_u)}{\sqrt{m}}\\ &= \frac{\sqrt{C\log(|Y|/\delta)}\Big(|X|\log(m) + |Y|\hat{H}_{\mathcal{S}}(\Phi_u)\Big) + \frac{2}{e}|X|}{\sqrt{m}}\end{aligned} \tag{5.25}$$

As regard to the second term in Eq. (5.14), we can write that

$$\begin{aligned}\mathcal{A}_2 &= |I_{\mathcal{T}}(Y;\Phi_u) - I_{\mathcal{S}}(Y;\Phi_u)| \\ &= \Big|\sum_y \sum_{\phi_u} p_{\mathcal{T}}(y,\phi_u)\log\Big(\frac{p_{\mathcal{T}}(y,\phi_u)}{p_{\mathcal{T}}(y)p_{\mathcal{T}}(\phi_u)}\Big) - p_{\mathcal{S}}(y,\phi_u)\log\Big(\frac{p_{\mathcal{S}}(y,\phi_u)}{p_{\mathcal{S}}(y)p_{\mathcal{S}}(\phi_u)}\Big)\Big| \\ &= \Big|\sum_y \sum_{\phi_u}\Big(p_{\mathcal{T}}(y,\phi_u)\log\big(p_{\mathcal{T}}(y|\phi_u)\big) - p_{\mathcal{S}}(y,\phi_u)\log\big(p_{\mathcal{S}}(y|\phi_u)\big)\Big) + H_{\mathcal{T}}(Y) - H_{\mathcal{S}}(Y)\Big|\end{aligned} \tag{5.26}$$

As shown in Figure 5.1, target variable Y is a exogenous node in the SCMs, which indicates that $p_{\mathcal{S}}(Y) = p_{\mathcal{T}}(Y)$. Therefore, we have that $\big|H_{\mathcal{S}}(Y) - H_{\mathcal{T}}(Y)\big| = 0$. Thus, we can write that

$$\begin{aligned}\mathcal{A}_2 &\leq \Big|\sum_y \sum_{\phi_u}\Big(p_{\mathcal{T}}(y,\phi_u)\log\big(p_{\mathcal{T}}(y|\phi_u)\big) - p_{\mathcal{S}}(y,\phi_u)\log\big(p_{\mathcal{S}}(y|\phi_u)\big)\Big)\Big| \\ &= \Big|\sum_y \sum_{\phi_u}\Big(p_{\mathcal{T}}(y,\phi_u)\log\big(p_{\mathcal{T}}(y|\phi_u)\big) - p_{\mathcal{T}}(y,\phi_u)\log\big(p_{\mathcal{S}}(y|\phi_u)\big) + p_{\mathcal{T}}(y,\phi_u)\log\big(p_{\mathcal{S}}(y|\phi_u)\big) - p_{\mathcal{S}}(y,\phi_u)\log\big(p_{\mathcal{S}}(y|\phi_u)\big)\Big)\Big| \\ &\leq \Big|\sum_y \sum_{\phi_u} p_{\mathcal{T}}(y,\phi_u)\log\Big(\frac{p_{\mathcal{T}}(y|\phi_u)}{p_{\mathcal{S}}(y|\phi_u)}\Big)\Big| + \Big|\sum_y \sum_{\phi_u}\big(p_{\mathcal{T}}(y,\phi_u) - p_{\mathcal{S}}(y,\phi_u)\big)\log\big(p_{\mathcal{S}}(y|\phi_u)\big)\Big| \\ &= \mathcal{KL}\big(p_{\mathcal{T}}(Y \mid \Phi_u)\big\|p_{\mathcal{S}}(Y \mid \Phi_u)\big) + \underbrace{\Big|\sum_y \sum_{\phi_u}\big(p_{\mathcal{T}}(y,\phi_u) - p_{\mathcal{S}}(y,\phi_u)\big)\log\big(p_{\mathcal{S}}(y|\phi_u)\big)\Big|}_{\mathcal{B}}\end{aligned}$$

According to the above equation, we have that

$$\mathcal{B}^2 = \Big\|\sum_y \sum_{\phi_u}\big(p_{\mathcal{T}}(y,\phi_u) - p_{\mathcal{S}}(y,\phi_u)\big)\log\big(p_{\mathcal{S}}(y|\phi_u)\big)\Big\|^2$$

Using the Jensen's inequality, we can get

$$\begin{aligned}\mathcal{B}^2 &\leq |Y|\sum_y \Big\|\sum_{\phi_u}\big(p_{\mathcal{T}}(y,\phi_u) - p_{\mathcal{S}}(y,\phi_u)\big)\log\big(p_{\mathcal{S}}(y|\phi_u)\big)\Big\|^2 \\ &\leq |Y|\sum_y \sum_{\phi_u} p(\phi_u)\Big\|\big(p_{\mathcal{T}}(y|\phi_u) - p_{\mathcal{S}}(y|\phi_u)\big)\log\big(p_{\mathcal{S}}(y|\phi_u)\big)\Big\|^2, \\ &\leq |Y|C_{\mathcal{S}}^2 \sum_y \sum_{\phi_u} p(\phi_u)\big\|p_{\mathcal{T}}(y|\phi_u) - p_{\mathcal{S}}(y|\phi_u)\big\|^2\end{aligned}$$

where $C_{\mathcal{S}}$ denotes a constant satisfying that $C_{\mathcal{S}} = \max_{(\phi_u,y)\in(\Phi_u,Y)}\big|\log\big(p_{\mathcal{S}}(y|\phi_u)\big)\big|$.

We know that $\log(\cdot)$ is a concave function, therefore we can get

$$\begin{aligned}\mathcal{B}^2 &\leq |Y|C_{\mathcal{S}}^2 \sum_{y}\sum_{\phi_u} p(\phi_u)\big\|p_{\mathcal{T}}(y|\phi_u) - p_{\mathcal{S}}(y|\phi_u)\big\|\big\|\log\big(p_{\mathcal{T}}(y|\phi_u)\big) - \log\big(p_{\mathcal{S}}(y|\phi_u)\big)\big\| \\ &= |Y|C_{\mathcal{S}}^2 \sum_{y}\sum_{\phi_u} p(\phi_u)\big(p_{\mathcal{T}}(y|\phi_u) - p_{\mathcal{S}}(y|\phi_u)\big)\Big(\log\big(p_{\mathcal{T}}(y|\phi_u)\big) - \log\big(p_{\mathcal{S}}(y|\phi_u)\big)\Big) \\ &= |Y|C_{\mathcal{S}}^2 \sum_{y}\sum_{\phi_u} p(\phi_u)\left(p_{\mathcal{T}}(y|\phi_u)\log\left(\frac{p_{\mathcal{T}}(y|\phi_u)}{p_{\mathcal{S}}(y|\phi_u)}\right) - p_{\mathcal{S}}(y|\phi_u)\log\left(\frac{p_{\mathcal{T}}(y|\phi_u)}{p_{\mathcal{S}}(y|\phi_u)}\right)\right) \\ &= |Y|C_{\mathcal{S}}^2\Big(\mathcal{KL}\big(p_{\mathcal{T}}(Y \mid \Phi_u)\|p_{\mathcal{S}}(Y \mid \Phi_u)\big) + \mathcal{KL}\big(p_{\mathcal{S}}(Y \mid \Phi_u)\|p_{\mathcal{T}}(Y \mid \Phi_u)\big)\Big).\end{aligned}$$

Consequently, we can get that

$$\begin{aligned}\mathcal{A}_2 &\leq \mathcal{KL}\big(p_{\mathcal{T}}(Y \mid \Phi_u)\big\|p_{\mathcal{S}}(Y \mid \Phi_u)\big) \\ &\quad + \sqrt{|Y|C_{\mathcal{S}}^2\Big(\mathcal{KL}\big(p_{\mathcal{T}}(Y \mid \Phi_u)\|p_{\mathcal{S}}(Y \mid \Phi_u)\big) + \mathcal{KL}\big(p_{\mathcal{S}}(Y \mid \Phi_u)\|p_{\mathcal{T}}(Y \mid \Phi_u)\big)\Big)} \\ &\leq \mathcal{J}\big(p_{\mathcal{T}}(Y \mid \Phi_u), p_{\mathcal{S}}(Y \mid \Phi_u)\big) + \sqrt{|Y|C_{\mathcal{S}}^2\mathcal{J}\big(p_{\mathcal{T}}(Y \mid \Phi_u), p_{\mathcal{S}}(Y \mid \Phi_u)\big)}\end{aligned} \tag{5.27}$$

where $\mathcal{J}(p,q)$ denotes the Jeffrey's divergence between probability p and q which is defined by

$$\mathcal{J}\big(p_{\mathcal{T}}(Y \mid \Phi_u), p_{\mathcal{S}}(Y \mid \Phi_u)\big) \triangleq \mathcal{KL}\big(p_{\mathcal{T}}(Y \mid \Phi_u)\|p_{\mathcal{S}}(Y \mid \Phi_u)\big) + \mathcal{KL}\big(p_{\mathcal{S}}(Y \mid \Phi_u)\|p_{\mathcal{T}}(Y \mid \Phi_u)\big)$$

With Eq. (5.25) and l (5.27), we can conclude that

$$\begin{aligned}&|\hat{I}_{\mathcal{S}}(Y;\Phi_u(X)) - I_{\mathcal{T}}(Y;\Phi_u(X))| \\ &\leq \frac{\sqrt{C\log(|Y|/\delta)}\Big(|X|\log(m) + |Y|\hat{H}_{\mathcal{S}}(\Phi_u)\Big) + \frac{2}{e}|X|}{\sqrt{m}} \\ &\quad + \mathcal{J}\big(p_{\mathcal{T}}(Y \mid \Phi_u), p_{\mathcal{S}}(Y \mid \Phi_u)\big) + \sqrt{|Y|C_{\mathcal{S}}^2\mathcal{J}\big(p_{\mathcal{T}}(Y \mid \Phi_u), p_{\mathcal{S}}(Y \mid \Phi_u)\big)}\end{aligned} \tag{5.28}$$

Thus, we complete the proof of Theorem 5.5.1. □

Remark 5.5.1. *For the 'IND generalization term' that will approach* 0 *as the size of training dataset grows towards infinity, it can be decreased by our FedPIN because minimizing* $\hat{H}(\Phi_u(X))$ *is included in the local objective as shown in Eq. (5.8). As regard to the 'OOD generalization term' caused by distribution shift, it can be unbounded and equals to* 0 *if and only if* $\mathcal{J}(\Phi_u) = 0$*. When the heterogeneity between* *training and test data distributions on each client u stems from the environment variable E as displayed in Figure 5.1, the 'OOD generalization term' can be eliminated by our FedPIN since the minimizer of objective (5.8) (i.e., $\omega_u^\star(\Phi_u^\star)$ ensures that $\mathbb{P}_{\mathcal{S}}(Y \mid \Phi_u^\star(X)) = \mathbb{P}_{\mathcal{T}}(Y \mid \Phi_u^\star(X))$ holds at all times (as discussed in Theorem 5.4.2). In summary, the personalized invariant models developed by our method can guarantee a tighter generalization error bound compared with the state-of-the-art PFL methods.*

5.5.2 Convergence Rate

In this section, we will derive the convergence rate of FedPIN shown in Algorithm 5. We start from the convergence analysis on the global models. For simplicity, we denotes the global model by $\theta_g \triangleq \{\Phi_g, \omega_g, \omega_a\}$ and define that $\mathcal{L}_g(\theta_g) \triangleq \mathcal{L}_{glob}(\Phi_g, \omega_g, \omega_a)$ and $\mathcal{L}_g^u(\theta_g) \triangleq \mathcal{L}_g^u(\Phi_g, \omega_g, \omega_a)$.

During each communication round t, the participating client u ($u \in S_t$) firstly initializes the model with $\theta_{g,u}^{t,0} = \theta_g^t$. Then, it conducts local gradient update for R iterations. At each local iteration r, the client u update the global model by $\theta_{g,u}^{t,r+1} = \theta_{g,u}^{t,r} - \beta \nabla \mathcal{L}_g^u(\theta_{g,u}^{t,r})$ using its local dataset D_u. After finishing local update for R iterations, client u ($u \in S_t$) uploads the local approximate model $\theta_{g,u}^{t,R}$ to the server which will aggregate the received local update $\{\theta_{g,u}^{t,R} | u \in S_t\}$ by $\theta_g^{t+1} = \frac{1}{M}\sum_{u \in S_t} \theta_{g,u}^{t,R}$. With the obtained θ_g^{t+1}, server can starts the next communication round.

Assumption 5.5.1. *Variance of local gradients to the aggregated average is upper bounded by a finite constant δ_L^2:*

$$\frac{1}{N}\sum_{u=1}^{N} \|\nabla \mathcal{L}_g^u(\Phi_g, \omega_g, \omega_a) - \nabla \mathcal{L}_{glob}(\Phi_g, \omega_g, \omega_a)\|^2 \leq \delta_L^2.$$

Lemma 5.5.1 (Aggregation variance)**.** *When assumption 5.5.1 holds and the number of selected clients at each communication round is $M = |S_t|$, the gradient bias caused by random client selection is upper-bounded by*

$$\mathbb{E}_{S_t}\left[\left\| \frac{1}{M}\sum_{u \in S_t} \nabla \mathcal{L}_g^u(\theta_g^t) - \nabla \mathcal{L}_g(\theta_g^t) \right\|^2\right] \leq \frac{N/M - 1}{N - 1}\delta_L^2. \tag{5.29}$$

Proof. We can write that

$$\begin{aligned}
&\mathbb{E}_{S_t}\Big[\Big\|\frac{1}{M}\sum_{u\in S_t}\nabla\mathcal{L}_g^u(\theta_g^t)-\nabla\mathcal{L}_g(\theta_g^t)\Big\|^2\Big]\\
&=\frac{1}{M^2}\mathbb{E}_{S_t}\Big[\Big\|\sum_{u\in S_t}\big(\nabla\mathcal{L}_g^u(\theta_g^t)-\nabla\mathcal{L}_g(\theta_g^t)\big)\Big\|^2\Big]\\
&=\frac{1}{M^2}\mathbb{E}_{S_t}\Big[\sum_{u\in S_t}\Big\|\nabla\mathcal{L}_g^u(\theta_g^t)-\nabla\mathcal{L}_g(\theta_g^t)\Big\|^2\\
&\quad+\sum_{u\in S_t}\sum_{\substack{v\neq u\\ v\in S_t}}\big\langle\nabla\mathcal{L}_g^u(\theta_g^t)-\nabla\mathcal{L}_g(\theta_g^t),\nabla\mathcal{L}_g^v(\theta_g^t)-\nabla\mathcal{L}_g(\theta_g^t)\big\rangle\Big]\\
&=\frac{1}{M^2}\mathbb{E}_{S_t}\Big[\sum_{u=1}^{N}I_{u\in S_t}\Big\|\nabla\mathcal{L}_g^u(\theta_g^t)-\nabla\mathcal{L}_g(\theta_g^t)\Big\|^2\\
&\quad+\sum_{\substack{u\in[N]\\ v\neq u}}I_{u\in S_t}I_{v\in S_t}\big\langle\nabla\mathcal{L}_g^u(\theta_g^t)-\nabla\mathcal{L}_g(\theta_g^t),\nabla\mathcal{L}_g^v(\theta_g^t)-\nabla\mathcal{L}_g(\theta_g^t)\big\rangle\Big],
\end{aligned}$$

where $I_{u\in S_t}=1$ if $u\in S_t$; $I_{u\in S_t}=0$ otherwise. Since every client $u\in[N]$ is randomly sampled with identical probability at each communication round t, we have $\mathbb{E}_{S_t}[I_{u\in S_t}]=p(u\in S_t)=\frac{M}{N}$ and $\mathbb{E}_{S_t}[I_{u\in S_t}I_{v\in S_t}]=p(u,v\in S_t \text{ and } u\neq v)=\frac{M(M-1)}{N(N-1)}$. According to the definition of $\mathcal{L}_g(\theta_g)$, we know that

$$\begin{aligned}
&\Big\|\frac{1}{N}\sum_{u=1}^{N}\nabla\mathcal{L}_g^u(\theta_g)-\nabla\mathcal{L}_g(\theta_g)\Big\|^2\\
&=\frac{1}{N^2}\sum_{u=1}^{N}\|\nabla\mathcal{L}_g^u(\theta)-\nabla\mathcal{L}_g(\theta_g)\|^2+\frac{1}{N^2}\sum_{u=1}^{N}\sum_{v\neq u}\big\langle\nabla\mathcal{L}_g^u(\theta_g)-\nabla\mathcal{L}_g(\theta_g),\nabla\mathcal{L}_g^v(\theta_g)-\nabla\mathcal{L}_g(\theta_g)\big\rangle\\
&=0
\end{aligned}$$

Thus, we can obtain that

$$\begin{aligned}
&\mathbb{E}_{S_t}\Big[\sum_{\substack{u\in[N]\\ v\neq u}}I_{u\in S_t}I_{v\in S_t}\big\langle\nabla\mathcal{L}_g^u(\theta_g)-\nabla\mathcal{L}_g(\theta_g),\nabla\mathcal{L}_g^v(\theta_g)-\nabla\mathcal{L}_g(\theta_g)\big\rangle\Big]\\
&=\sum_{\substack{u\in[N]\\ v\neq u}}\mathbb{E}_{S_t}[I_{u\in S_t}I_{v\in S_t}]\big\langle\nabla\mathcal{L}_g^u(\theta_g)-\nabla\mathcal{L}_g(\theta_g),\nabla\mathcal{L}_g^v(\theta)-\nabla\mathcal{L}_g(\theta_g)\big\rangle\\
&=\sum_{\substack{u\in[N]\\ v\neq u}}\frac{M(M-1)}{N(N-1)}\big\langle\nabla\mathcal{L}_g^u(\theta_g)-\nabla\mathcal{L}_g(\theta_g),\nabla\mathcal{L}_g^v(\theta_g)-\nabla\mathcal{L}_g(\theta_g)\big\rangle
\end{aligned}$$

$$= -\frac{M(M-1)}{N(N-1)} \sum_{u=1}^{N} \big\|\nabla\mathcal{L}_g^u(\theta_g) - \nabla\mathcal{L}_g(\theta_g)\big\|^2$$

Therefore, we can derive that

$$\begin{aligned}
&\mathbb{E}_{S_t}\Big[\Big\|\frac{1}{M}\sum_{u\in S_t} \nabla\mathcal{L}_g^u(\theta_g^t) - \nabla\mathcal{L}_g(\theta_g^t)\Big\|^2\Big] \\
&= \frac{1}{M^2}\Big[\sum_{u=1}^{N} \mathbb{E}_{S_t}[I_{u\in S_t}]\big\|\nabla\mathcal{L}_g^u(\theta_g^t) - \nabla\mathcal{L}_g(\theta_g^t)\big\|^2 - \frac{M(M-1)}{N(N-1)}\sum_{u=1}^{N}\big\|\nabla\mathcal{L}_g^u(\theta_g^t) - \nabla\mathcal{L}_g(\theta_g^t)\big\|^2\Big] \\
&= \Big(\frac{1}{M^2}\cdot\frac{M}{N} - \frac{1}{M^2}\cdot\frac{M(M-1)}{N(N-1)}\Big)\sum_{u=1}^{N}\big\|\nabla\mathcal{L}_g^u(\theta_g^t) - \nabla\mathcal{L}_g(\theta_g^t)\big\|^2 \\
&= \frac{N/M-1}{N-1}\cdot\frac{1}{N}\sum_{u=1}^{N}\big\|\nabla\mathcal{L}_g^u(\theta_u^t) - \nabla\mathcal{L}_g(\theta_g^t)\big\|^2 \\
&\leq \frac{N/M-1}{N-1}\delta_L^2.
\end{aligned}$$

Proof of Lemma 5.5.1 ends. □

Lemma 5.5.2 (Local update). *When $\mathcal{L}_g^u(\theta_g), \forall u \in [N]$ is L-smooth and the learning rate $\beta \leq \frac{1}{\sqrt{2}RL}$, if we denote the local approximate update of the global model parameter at local iteration r on client u by $\theta_{g,u}^{t,r}$ and $\theta_{g,u}^{t,r=0}$ is initialized as θ_g^t, the following inequality holds for any $u \in [N]$:*

$$\frac{1}{R}\sum_{r=0}^{R-1}\big\|\theta_{g,u}^{t,r} - \theta_g^t\big\|^2 \leq 8R^2\beta^2\|\nabla\mathcal{L}_g^u(\theta_g^t)\|^2. \tag{5.30}$$

Proof. We know $\theta_{g,u}^{t,r} = \theta_{g,u}^{t,r-1} - \beta\nabla\mathcal{L}_g^u(\theta_{g,u}^{t,r-1}), \forall r \geq 1$. Therefore, we can write

$$\begin{aligned}
&\big\|\theta_{g,u}^{t,r} - \theta_g^t\big\|^2 \\
&= \big\|\theta_{g,u}^{t,r-1} - \beta\nabla\mathcal{L}_g^u(\theta_{g,u}^{t,r-1}) - \theta_g^t\big\|^2 \\
&= \big\|\theta_{g,u}^{t,r-1} - \beta\nabla\mathcal{L}_g^u(\theta_{g,u}^{t,r-1}) + \beta\nabla\mathcal{L}_g^u(\theta_g^t) - \beta\nabla\mathcal{L}_g^u(\theta_g^t) - \theta_g^t\big\|^2 \\
&\leq (1+\frac{1}{R})\big\|\theta_{g,u}^{t,r-1} - \theta_g^t - \beta\nabla\mathcal{L}_g^u(\theta_g^t)\big\|^2 + (1+R)\big\|\beta\nabla\mathcal{L}_g^u(\theta_g^t) - \beta\nabla\mathcal{L}_g^u(\theta_{g,u}t, r-1)\big\|^2 \\
&\leq (1+\frac{1}{R})\{(1+\frac{1}{2R})\big\|\theta_{g,u}^{t,r-1} - \theta_g^t\big\|^2 + (1+2R)\big\|\beta\nabla\mathcal{L}_g^u(\theta_g^t)\big\|^2\} + (1+R)\beta^2L^2\big\|\theta_{g,u}^{t,r-1} - \theta_g^t\big\|^2 \\
&= (1+\frac{1}{R})(1+\frac{1}{2R} + R\beta^2L^2)\big\|\theta_{g,u}^{t,r-1} - \theta_g^t\big\|^2 + (1+\frac{1}{R})(1+2R)\beta^2\big\|\nabla\mathcal{L}_g^u(\theta_g^t)\big\|^2.
\end{aligned}$$

When $\beta \leq \frac{1}{8RL}$, we have $R\beta^2L^2 \leq \frac{1}{2R}$. Furthermore, we can get

$$\left\|\theta_{g,u}^{t,r} - \theta_g^t\right\|^2 \leq (1+\frac{1}{R})^2\left\|\theta_{g,u}^{t,r-1} - \theta_g^t\right\|^2 + (1+\frac{1}{R})(1+2R)\beta^2\left\|\nabla\mathcal{L}_g^u(\theta_g^t)\right\|^2.$$

Since $\theta_{g,u}^{t,0} = \theta_g^t$, we can derive the following inequality for any $r \geq 1$:

$$\begin{aligned}\left\|\theta_{g,u}^{t,r} - \theta_g^t\right\|^2 &\leq \sum_{s=0}^{r-1}(1+\frac{1}{R})^{2s}(1+\frac{1}{R})(1+2R)\beta^2\left\|\nabla\mathcal{L}_g^u(\theta_g^t)\right\|^2 \\ &= (1+\frac{1}{R})(1+2R)\beta^2\left\|\nabla\mathcal{L}_g^u(\theta_g^t)\right\|^2\frac{(1+\frac{1}{R})^{2r}-1}{(1+\frac{1}{R})^2-1} \\ &\leq (1+\frac{1}{R})(1+2R)\beta^2\left\|\nabla\mathcal{L}_g^u(\theta_g^t)\right\|^2\frac{(1+\frac{1}{R})^{2r}}{(\frac{2}{R}+\frac{1}{R^2})} \\ &= R^2(1+\frac{1}{R})(1+2R)\beta^2\left\|\nabla\mathcal{L}_g^u(\theta_g^t)\right\|^2\frac{(1+\frac{1}{R})^{2r}}{2R+1} \\ &= R(1+R)\beta^2\left\|\nabla\mathcal{L}_g^u(\theta_g^t)\right\|^2(1+\frac{1}{R})^{2r}.\end{aligned}$$

Therefore, based on the above inequality we can write that

$$\begin{aligned}\frac{1}{R}\sum_{r=0}^{R-1}\left\|\theta_{g,u}^{t,r} - \theta_g^t\right\|^2 &\leq R(1+R)\beta^2\left\|\nabla\mathcal{L}_g^u(\theta_g^t)\right\|^2\frac{1}{R}\sum_{r=0}^{R-1}(1+\frac{1}{R})^{2r} \\ &= (1+R)\beta^2\left\|\nabla\mathcal{L}_g^u(\theta_g^t)\right\|^2\frac{(1+\frac{1}{R})^{2R}-1}{(1+\frac{1}{R})^2-1} \\ &\leq (1+R)\beta^2\left\|\nabla\mathcal{L}_g^u(\theta_g^t)\right\|^2\frac{(1+\frac{1}{R})^{2R}}{\frac{2}{R}+\frac{1}{R^2}} \\ &= \frac{R^2(1+R)}{1+2R}\beta^2\left\|\nabla\mathcal{L}_g^u(\theta_g^t)\right\|^2(1+\frac{1}{R})^{2R} \\ &\leq \frac{1}{2}R(1+R)\beta^2\left\|\nabla\mathcal{L}_g^u(\theta_g^t)\right\|^2(1+\frac{1}{R})^{2R}.\end{aligned}$$

We know that $(1+\frac{1}{R})^R \leq \lim_{R\to\infty}(1+\frac{1}{R})^R = e$ and $e^2 < 8$. Thus, we can get that

$$\begin{aligned}\frac{1}{R}\sum_{r=0}^{R-1}\left\|\theta_{g,u}t,r - \theta_g^t\right\|^2 &\leq \frac{1}{2}R(1+R)\beta^2\left\|\nabla\mathcal{L}_g^u(\theta_g^t)\right\|^2e^2 \\ &\leq e^2R^2\beta^2\left\|\nabla\mathcal{L}_g^u(\theta_g^t)\right\|^2 \\ &< 8R^2\beta^2\left\|\nabla\mathcal{L}_g^u(\theta_g^t)\right\|^2, \forall R \geq 1.\end{aligned}$$

Proof of Lemma 5.5.2 ends. □

Theorem 5.5.2. *Suppose loss function $\mathcal{L}_g^u(\Phi_g, \omega_g, \omega_a)$, $\forall u \in [N]$ is L-smooth and assumption 5.5.1 holds. The number of the selected clients at each communication round is M. When the learning rate β satisfies that $\beta < \frac{1}{8RL}$, the convergence rate of the **global model** is described by*

$$\begin{aligned}&\mathbb{E}[\|\nabla \mathcal{L}_{glob}(\Phi_g^{t^\star}, \omega_g^{t^\star}, \omega_a^{t^\star})\|^2] \\ &\leq \mathcal{G}(T) \triangleq \mathcal{O}\left(\frac{\Delta_l}{\beta RT} + \frac{\Delta_l^{\frac{3}{4}} L^{\frac{3}{4}} \delta_L^{\frac{1}{2}}}{T^{\frac{3}{4}}} + \frac{\Delta_l^{\frac{2}{3}} L^{\frac{2}{3}} \delta_L^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \sqrt{\frac{(N-M)\Delta_l L \delta_L^2}{M(N-1)T}}\right),\end{aligned}$$

where $\Delta_l \triangleq \mathbb{E}[\mathcal{L}_{glob}(\Phi_g^0, \omega_g^0, \omega_a^0) - \mathcal{L}_{glob}(\Phi_g^T, \omega_g^T, \omega_a^T)]$ and $t^\star$ is uniformly sampled from the set $\{0, 1, ..., T-1\}$.

Proof. Since the local approximate model parameters are updated by $\theta_{g,u}^{t,r+1} = \theta_{g,u}^{t,r} - \beta \nabla \mathcal{L}_g^u(\theta_{g,u}^{t,r}), \forall u, r$, we can get $\sum_{r=0}^{R-1} \beta \mathcal{L}_g^u(\theta_{g,u}^{t,r}) = \theta_{g,u}^{t,0} - \theta_{g,u}^{t,R}$. That is,

$$\theta_{g,u}^{t,R} = \theta_g^t - \beta \sum_{r=0}^{R-1} \nabla \mathcal{L}_g^u(\theta_{g,u}^{t,r})$$

After the global aggregation, we can get the updated global model at communication round $t+1$ as

$$\begin{aligned}\theta_g^{t+1} = \frac{1}{M} \sum_{u \in S_t} \theta_{g,u}^{t,R} &= \frac{1}{M} \sum_{u \in S_t} \left\{\theta_g^t - \beta \sum_{r=0}^{R-1} \nabla \mathcal{L}_g^u(\theta_{g,u}^{t,r})\right\} \\ &= \theta_g^t - \frac{1}{M} \beta \sum_{u \in S_t} \sum_{r=0}^{R-1} \nabla \mathcal{L}_g^u(\theta_{g,u}^{t,r}) \\ &= \theta_g^t - \underbrace{\beta R}_{:=\hat{\beta}} \underbrace{\frac{1}{MR} \sum_{u \in S_t} \sum_{r=0}^{R-1} \nabla \mathcal{L}_g^u(\theta_{g,u}^{t,r})}_{:=\psi^t}.\end{aligned}$$

Since loss function $\mathcal{L}_g^u(\theta_g), \forall u \in [N]$ is L-smooth, we can get the following inequality:

$$\begin{aligned}\|\nabla \mathcal{L}_g(\theta_g) - \nabla \mathcal{L}_g(\theta_g')\| &= \left\|\frac{1}{N} \sum_{u=1}^{N} \nabla \mathcal{L}_g^u(\theta_g) - \frac{1}{N} \sum_{u=1}^{N} \nabla \mathcal{L}_g^u(\theta_g')\right\| \\ &\leq \frac{1}{N} \sum_{u=1}^{N} \|\nabla \mathcal{L}_g^u(\theta_g) - \nabla \mathcal{L}_g^u(\theta_g')\| \\ &\leq L\|\theta_g - \theta_g'\|, \forall \theta_g, \theta_g',\end{aligned}$$

which means that the global loss function $\mathcal{L}_g(\theta_g)$ is also L-smooth. Therefore, we can write that

$$
\begin{aligned}
&\mathbb{E}_{S_t}[\mathcal{L}_g(\theta_g^{t+1}) - \mathcal{L}_g(\theta_g^t)] \\
&\leq \mathbb{E}_{S_t}[\langle \nabla \mathcal{L}_g(\theta_g^t), \theta_g^{t+1} - \theta_g^t \rangle] + \frac{L}{2}\mathbb{E}_{S_t}[\|\theta_g^{t+1} - \theta_g^t\|^2] \\
&= \mathbb{E}_{S_t}[\langle \nabla \mathcal{L}_g(\theta_g^t), -\hat{\beta}\psi^t \rangle] + \frac{L}{2}\mathbb{E}_{S_t}[\|\hat{\beta}\psi^t\|^2] \\
&= \hat{\beta}\mathbb{E}_{S_t}[\langle \nabla \mathcal{L}_g(P\theta_g^t), \nabla \mathcal{L}_g(\theta_g^t) - \psi^t - \nabla \mathcal{L}_g(\theta_g^t) \rangle] + \frac{\hat{\beta}^2 L}{2}\mathbb{E}_{S_t}[\|\psi^t\|^2] \\
&= -\hat{\beta}\mathbb{E}_{S_t}[\|\nabla \mathcal{L}_g(\theta_g^t)\|^2] - \hat{\beta}\mathbb{E}_{S_t}[\langle \nabla \mathcal{L}_g(\theta_g^t), \psi^t - \nabla \mathcal{L}_g(\theta_g^t) \rangle] + \frac{\hat{\beta}^2 L}{2}\mathbb{E}_{S_t}[\|\psi^t\|^2] \\
&\leq -\hat{\beta}\mathbb{E}_{S_t}[\|\nabla \mathcal{L}_g(\theta_g^t)\|^2] + \frac{\hat{\beta}}{2}\mathbb{E}_{S_t}[\|\nabla \mathcal{L}_g(\theta_g^t)\|^2] + \frac{\hat{\beta}^2 L}{2}\mathbb{E}_{S_t}[\|\psi^t\|^2] \\
&\quad + \frac{\hat{\beta}}{2}\mathbb{E}_{S_t}\Big[\Big\|\frac{1}{NR}\sum_{u=1}^{N}\sum_{r=0}^{R-1}\nabla \mathcal{L}_g^u(\theta_{g,u}^{t,r}) - \frac{1}{N}\sum_{u=1}^{N}\nabla \mathcal{L}_g^u(\theta_g^t)\Big\|^2\Big] \\
&\leq -\frac{\hat{\beta}}{2}\mathbb{E}_{S_t}[\|\nabla \mathcal{L}_g(\theta_g^t)\|^2] + \frac{\hat{\beta}^2 L}{2}\mathbb{E}_{S_t}[\|\psi^t\|^2] + \frac{\hat{\beta}}{2}\mathbb{E}_{S_t}\Big[\frac{1}{NR}\sum_{u=1}^{N}\sum_{r=0}^{R-1}\|\nabla \mathcal{L}_g^u(\theta_{g,u}^{t,r}) - \nabla \mathcal{L}_g^u(\theta_g^t)\|^2\Big] \\
&\leq -\frac{\hat{\beta}}{2}\mathbb{E}_{S_t}[\|\nabla \mathcal{L}_g(\theta_g^t)\|^2] + \frac{\hat{\beta}^2 L}{2}\mathbb{E}_{S_t}[\|\psi^t\|^2] + \frac{\hat{\beta}L^2}{2}\mathbb{E}_{S_t}\Big[\frac{1}{NR}\sum_{u=1}^{N}\sum_{r=0}^{R-1}\|\theta_{g,u}^{t,r} - \theta_g^t\|^2\Big] \\
&\leq -\frac{\hat{\beta}}{2}\mathbb{E}_{S_t}[\|\nabla \mathcal{L}_g(\theta_g^t)\|^2] + \frac{\hat{\beta}^2 L}{2}\mathbb{E}_{S_t}[\|\psi^t\|^2] + 4\hat{\beta}^3 L^2\mathbb{E}_{S_t}\Big[\frac{1}{N}\sum_{u=1}^{N}\|\nabla \mathcal{L}_g^u(\theta_g^t)\|^2\Big] \\
&= -\frac{\hat{\beta}}{2}\mathbb{E}_{S_t}[\|\nabla \mathcal{L}_g(\theta_g^t)\|^2] + \frac{\hat{\beta}^2 L}{2}\mathbb{E}_{S_t}[\|\psi^t - \nabla \mathcal{L}_g(\theta_g^t) + \nabla \mathcal{L}_g(\theta_g^t)\|^2] \\
&\quad + 4\hat{\beta}^3 L^2\mathbb{E}_{S_t}\Big[\frac{1}{N}\sum_{u=1}^{N}\|\nabla \mathcal{L}_g^u(\theta_g^t) - \nabla \mathcal{L}_g(\theta_g^t) + \nabla \mathcal{L}_g(\theta_g^t)\|^2\Big] \\
&\leq -\frac{\hat{\beta}}{2}\mathbb{E}_{S_t}[\|\nabla \mathcal{L}_g(\theta_g^t)\|^2] + \hat{\beta}^2 L\mathbb{E}_{S_t}[\|\psi^t - \nabla \mathcal{L}_g(\theta_g^t)\|^2] + \hat{\beta}^2 L\mathbb{E}_{S_t}[\|\nabla \mathcal{L}_g(\theta_g^t)\|^2] \\
&\quad + 8\hat{\beta}^3 L^2\{\delta_L^2 + \mathbb{E}_{S_t}[\|\nabla \mathcal{L}_g(\theta_g^t)\|^2]\} \\
&= -\frac{\hat{\beta}}{2}\{1 - 2\hat{\beta}L - 16\hat{\beta}^2 L^2\}\mathbb{E}_{S_t}[\|\nabla \mathcal{L}_g(\theta_g^t)\|^2] + 8\hat{\beta}^3 L^2\delta_L^2 + \hat{\beta}^2 L\mathbb{E}_{S_t}[\|\psi^t - \nabla \mathcal{L}_g(\theta_g^t)\|^2].
\end{aligned}
$$

In the subsequent step, we firstly deal with the third term on the right side of above inequality as follows:

$$\begin{aligned}
&\mathbb{E}_{S_t}[\|\psi^t - \nabla\mathcal{L}_g(\theta_g^t)\|^2] \\
&= \mathbb{E}_{S_t}\Big[\Big\|\frac{1}{MR}\sum_{u\in S_t}\sum_{r=0}^{R-1}\nabla\mathcal{L}_g^u(\theta_{g,u}^{t,r}) - \frac{1}{N}\sum_{u=1}^{N}\nabla\mathcal{L}_g^u(\theta_g^t)\Big\|^2\Big] \\
&= \mathbb{E}_{S_t}\Big[\Big\|\frac{1}{MR}\sum_{u\in S_t}\sum_{r=0}^{R-1}\nabla\mathcal{L}_g^u(\theta_{g,u}^{t,r}) - \frac{1}{MR}\sum_{u\in S_t}\sum_{r=0}^{R-1}\nabla\mathcal{L}_g^u(\theta_g^t) + \frac{1}{MR}\sum_{u\in S_t}\sum_{r=0}^{R-1}\nabla\mathcal{L}_g^u(\theta_u^t) - \frac{1}{N}\sum_{u=1}^{N}\nabla\mathcal{L}_g^u(\theta_g^t)\Big\|^2\Big] \\
&\leq 2\mathbb{E}_{S_t}\Big[\Big\|\frac{1}{MR}\sum_{u\in S_t}\sum_{r=0}^{R-1}\nabla\mathcal{L}_g^u(\theta_{g,u}^{t,r}) - \frac{1}{MR}\sum_{u\in S_t}\sum_{r=0}^{R-1}\nabla\mathcal{L}_g^u(\theta_g^t)\Big\|^2\Big] \\
&\quad + 2\mathbb{E}_{S_t}\Big[\Big\|\frac{1}{MR}\sum_{u\in S_t}\sum_{r=0}^{R-1}\nabla\mathcal{L}_g^u(\theta_g^t) - \frac{1}{N}\sum_{u=1}^{N}\nabla\mathcal{L}_g^u(\theta_g^t)\Big\|^2\Big] \\
&= 2\mathbb{E}_{S_t}\Big[\Big\|\frac{1}{MR}\sum_{u\in S_t}\sum_{r=0}^{R-1}(\nabla\mathcal{L}_g^u(\theta_{g,u}^{t,r}) - \nabla\mathcal{L}_g^u(\theta_g^t))\Big\|^2\Big] + 2\mathbb{E}_{S_t}\Big[\Big\|\frac{1}{M}\sum_{u\in S_t}\nabla\mathcal{L}_g^u(\theta_g^t) - \frac{1}{N}\sum_{u=1}^{N}\nabla\mathcal{L}_g^u(\theta_g^t)\Big\|^2\Big] \\
&\leq 2\mathbb{E}_{S_t}\Big[\frac{1}{MR}\sum_{u\in S_t}\sum_{r=0}^{R-1}\|\nabla\mathcal{L}_g^u(\theta_{g,u}^{t,r}) - \nabla\mathcal{L}_g^u(\theta_g^t)\|^2\Big] + 2\mathbb{E}_{S_t}\Big[\Big\|\frac{1}{M}\sum_{u\in S_t}\nabla\mathcal{L}_g^u(\theta_g^t) - \frac{1}{N}\sum_{u=1}^{N}\nabla\mathcal{L}_g^u(\theta_g^t)\Big\|^2\Big] \\
&\leq 2\mathbb{E}_{S_t}\Big[\frac{1}{M}\sum_{u\in S_t}\frac{1}{R}\sum_{r=0}^{R-1}L^2\|\theta_{g,u}^{t,r} - \theta_g^t\|^2\Big] + 2\mathbb{E}_{S_t}\Big[\Big\|\frac{1}{M}\sum_{u\in S_t}\nabla\mathcal{L}_g^u(\theta_g^t) - \nabla\mathcal{L}_g(\theta_g^t)\Big\|^2\Big].
\end{aligned}$$

Using the inequalities in Lemma 5.5.1 and Lemma 5.5.2, we can get

$$\begin{aligned}
&\mathbb{E}_{S_t}[\|\psi^t - \nabla\mathcal{L}_g(\theta_g^t)\|^2] \\
&\leq 16R^2\beta^2L^2\mathbb{E}_{S_t}\Big[\frac{1}{M}\sum_{u\in S_t}\|\nabla\mathcal{L}_g^u(\theta_g^t)\|^2\Big] + \frac{2(N/M-1)}{N-1}\delta_L^2 \\
&\leq 16R^2\beta^2L^2\mathbb{E}_{S_t}\Big[\frac{1}{M}\sum_{u\in S_t}\|\nabla\mathcal{L}_g^u(\theta_g^t) - \nabla\mathcal{L}_g(\theta_g^t) + \nabla\mathcal{L}_g(\theta_g^t)\|^2\Big] + \frac{2(N/M-1)}{N-1}\delta_L^2 \\
&\leq 32R^2\beta^2L^2\mathbb{E}_{S_t}\Big[\frac{1}{M}\sum_{u\in S_t}\|\nabla\mathcal{L}_g^u(\theta_g^t) - \nabla\mathcal{L}_g(\theta_g^t)\|^2\Big] \\
&\quad + 32R^2\beta^2L^2\mathbb{E}_{S_t}\Big[\frac{1}{M}\sum_{u\in S_t}\|\nabla\mathcal{L}_g(\theta_g^t)\|^2\Big] + \frac{2(N/M-1)}{N-1}\delta_L^2 \\
&= 32R^2\beta^2L^2\frac{1}{M}\mathbb{E}_{S_t}\Big[\sum_{u=1}^{N}I_{u\in S_t}\|\nabla\mathcal{L}_g^u(\theta_g^t) - \nabla\mathcal{L}_g(\theta_g^t)\|^2\Big] \\
&\quad + 32R^2\beta^2L^2\mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2] + \frac{2(N/M-1)}{N-1}\delta_L^2
\end{aligned}$$

$$
\begin{aligned}
&= 32R^2\beta^2L^2\frac{1}{M}\sum_{u=1}^{N}\mathbb{E}_{S_t}[I_{u\in S_t}]\|\nabla\mathcal{L}_g^u(\theta_g^t)-\nabla\mathcal{L}_g(\theta_g^t)\|^2\\
&\quad + 32R^2\beta^2L^2\mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2] + \frac{2(N/M-1)}{N-1}\delta_L^2\\
&\le 32R^2\beta^2L^2\delta_L^2 + 32R^2\beta^2L^2\mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2] + \frac{2(N/M-1)}{N-1}\delta_L^2\\
&= 32R^2\beta^2L^2\mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2] + \big(32R^2\beta^2L^2 + \frac{2(N/M-1)}{N-1}\big)\delta_L^2\\
&= 32\hat{\beta}^2L^2\mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2] + \big(32\hat{\beta}^2L^2 + \frac{2(N/M-1)}{N-1}\big)\delta_L^2
\end{aligned}
$$

Finally, we can get

$$
\begin{aligned}
\mathbb{E}_{S_t}[\mathcal{L}_g(\theta_g^{t+1})-\mathcal{L}_g(\theta_g^t)] &\le -\frac{\hat{\beta}}{2}(1-2\hat{\beta}L-16\hat{\beta}^2L^2-64\hat{\beta}^3L^3)\mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2]\\
&\quad + 8\hat{\beta}^3L^2\delta_L^2 + 32\hat{\beta}^4L^3\delta_L^2 + \frac{2(N/M-1)\hat{\beta}^2L\delta_L^2}{N-1}.
\end{aligned}
$$

When $\beta \le \frac{1}{8RL}$, we have

$$
1-2\hat{\beta}L-16\hat{\beta}^2L^2-64\hat{\beta}^3L^3 \ge 1-\frac{1}{4}-\frac{1}{4}-\frac{1}{8} > \frac{1}{4}, \forall R\ge 1.
$$

Thus, we can derive that

$$
\begin{aligned}
&\mathbb{E}_{S_t}[\mathcal{L}_g(\theta_g^{t+1})-\mathcal{L}_g(\theta_g^t)]\\
&\le -\frac{\hat{\beta}}{8}\mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2] + 32\hat{\beta}^4L^3\delta_L^2 + 8\hat{\beta}^3L^2\delta_L^2 + \frac{2(N-M)\hat{\beta}^2L\delta_L^2}{M(N-1)}.
\end{aligned}
$$

In other words, we have

$$
\begin{aligned}
&\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2]\\
&\le \frac{4\mathbb{E}_{S_t}[\mathcal{L}_g(\theta_g^0)-\mathcal{L}_g(\theta_g^T)]}{\hat{\beta}T} + 128\hat{\beta}^3L^3\delta_L^2 + 32\hat{\beta}^2L^2\delta_L^2 + \frac{8(N-M)\hat{\beta}L\delta_L^2}{M(N-1)}.
\end{aligned}
$$

For simplicity, we define that $\beta_0 = \frac{1}{8RL}$, $C_1 = 4\mathbb{E}_{S_t}[\mathcal{L}_g(\theta_g^0)-\mathcal{L}_g(\theta_g^T)]$, $C_2 = 128L^3\delta_L^2$, $C_3 = 32L^2\delta_L^2$ and $C_4 = \frac{8(N-M)L\delta_L^2}{M(N-1)}$. Thus, we have

$$
\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2] \le \frac{C_1}{R\beta T} + C_2R^3\beta^3 + C_3R^2\beta^2 + C_4R\beta.
$$

Using the schemes adopted in [48, 94, 96], we consider the following two cases:

- When $\beta_0 \leq \min\{(\frac{C_1}{C_2 R^4 T})^{\frac{1}{4}}, (\frac{C_1}{C_3 R^3 T})^{\frac{1}{3}}, (\frac{C_1}{C_4 R^2 T})^{\frac{1}{2}}\}$, we choose $\beta = \beta_0$. Then, we have
$$\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2] \leq \frac{C_1}{\beta_0 RT} + \frac{C_1^{\frac{3}{4}}C_2^{\frac{1}{4}}}{T^{\frac{3}{4}}} + \frac{C_1^{\frac{2}{3}}C_3^{\frac{1}{3}}}{T^{\frac{2}{3}}} + \frac{C_1^{\frac{1}{2}}C_4^{\frac{1}{2}}}{T^{\frac{1}{2}}}.$$

- When $\beta_0 \geq \min\{(\frac{C_1}{C_2 R^4 T})^{\frac{1}{4}}, (\frac{C_1}{C_3 R^3 T})^{\frac{1}{3}}, (\frac{C_1}{C_4 R^2 T})^{\frac{1}{2}}\}$, we choose the value of β as $\beta = \min\{(\frac{C_1}{C_2 R^4 T})^{\frac{1}{4}}, (\frac{C_1}{C_3 R^3 T})^{\frac{1}{3}}, (\frac{C_1}{C_4 R^2 T})^{\frac{1}{2}}\}$. Then, we have
$$\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2] \leq \frac{2C_1^{\frac{3}{4}}C_2^{\frac{1}{4}}}{T^{\frac{3}{4}}} + \frac{2C_1^{\frac{2}{3}}C_3^{\frac{1}{3}}}{T^{\frac{2}{3}}} + \frac{2C_1^{\frac{1}{2}}C_4^{\frac{1}{2}}}{T^{\frac{1}{2}}}.$$

Combining these two cases, we can get
$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2] &\leq \mathcal{O}\Big(\frac{C_1}{\beta_0 RT} + \frac{3C_1^{\frac{3}{4}}C_2^{\frac{1}{4}}}{T^{\frac{3}{4}}} + \frac{3C_1^{\frac{2}{3}}C_3^{\frac{1}{3}}}{T^{\frac{2}{3}}} + \frac{3C_1^{\frac{1}{2}}C_4^{\frac{1}{2}}}{T^{\frac{1}{2}}}\Big)\\ &= \mathcal{O}\Big(\frac{\Delta_l}{\beta RT} + \frac{\Delta_l^{\frac{3}{4}}L^{\frac{3}{4}}\delta_L^{\frac{1}{2}}}{T^{\frac{3}{4}}} + \frac{\Delta_l^{\frac{2}{3}}L^{\frac{2}{3}}\delta_L^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \sqrt{\frac{(N-M)\Delta_l L\delta_L^2}{M(N-1)T}}\Big)\end{aligned}$$
where $\Delta_l := \mathbb{E}[\mathcal{L}_g(\theta_g^0) - \mathcal{L}_g(\theta_g^T)]$ and the learning rate β must satisfy $\beta \leq \frac{1}{8RL}$. Proof of Theorem 5.5.2 ends. □

Theorem 5.5.2 proves that our algorithm achieves a convergence rate of $\mathcal{O}(1/\sqrt{T})$ when only a subset of clients is selected at each communication round (i.e., $M < N$) and local data distributions are Non-IID (i.e., $\delta_L > 0$). In particular, the convergence rate can reach $\mathcal{O}(1/T^{\frac{2}{3}})$ if all clients are selected at each communication round.

Corollary 5.5.1. *Assuming that the local loss function $\mathcal{L}_{loc}^u(\omega_u(\Phi_u); \Phi_g^\star, \Psi_u^\star)$ is L-smooth and strongly convex, and its gradient is upper bounded by a finite constant, $\forall u \in [N]$. If we define that $f_{\theta_u} \triangleq \omega_u(\Phi_u)$, $f_{\theta_u}^\star = \arg\min_{\omega_u,\Phi_u} \mathcal{L}_{loc}^u(\omega_u(\Phi_u); \Phi_g^\star, \Psi_u^\star)$, and the output of Algorithm 5 after communication round T is denoted as $f_{\theta_u}^T$, the convergence rate of* ***personalized model*** *is given by*
$$\mathbb{E}[\|f_{\theta_u}^T - f_{\theta_u}^\star\|^2] \leq C\mathcal{G}(T) + \epsilon_K^2, \forall u \in [N],$$
where both C and ϵ_K are finite constants and $\epsilon_K^2 \to 0$ as the personalization epochs $K \to \infty$.

Proof. We demonstrate this claim by induction. Firstly, when the constant $C \geq \frac{\mathbb{E}[\|f^0_{\theta_u} - f^\star_{\theta_u}\|^2]}{\mathcal{G}(0)}$, we have $\mathbb{E}[\|f^0_{\theta_u} - f^\star_{\theta_u}\|^2] \leq C\mathcal{G}(0) + \epsilon_K^2$. Suppose $\mathbb{E}[\|f^t_{\theta_u} - f^\star_{\theta_u}\|^2] \leq C\mathcal{G}(t) + \epsilon_K^2$, for $t+1$, we can write

$$\begin{aligned}
&\mathbb{E}[\|f^{t+1}_{\theta_u} - f^\star_{\theta_u}\|^2] \\
&= \mathbb{E}[\|f^t_{\theta_u} - \eta I_t \nabla \mathcal{L}^u_{loc}(f^t_{\theta_u}) - f^\star_{\theta_u}\|^2] \\
&= \mathbb{E}[\|f^t_{\theta_u} - f^\star_{\theta_u}\|^2] + \eta^2 \mathbb{E}[\|I_t \nabla \mathcal{L}^u_{loc}(f^t_{\theta_u})\|^2] + 2\eta \mathbb{E}[\langle I_t \nabla \mathcal{L}^u_{loc}(f^t_{\theta_u}), f^\star_{\theta_u} - f^t_{\theta_u}\rangle]
\end{aligned}$$

where I_t indicates whether client u is selected by server at communication round t. That is $I_t = 1$ when client u is selected by server at communication round t; and $I_t = 0$ otherwise. Hence, $\mathbb{E}[I_t] = \frac{M}{N}$. Because the local loss function $\mathcal{L}^u_{loc}(f_{\theta_u})$ is L-smooth and μ_l-strongly convex, $\forall u \in [N]$, we have

$$\begin{aligned}
\mathbb{E}[\langle \nabla \mathcal{L}^u_{loc}(f^t_{\theta_u}), f^\star_{\theta_u} - f^t_{\theta_u}\rangle] &\leq (\mathcal{L}^u_{loc}(f^\star_{\theta_u}) - \mathcal{L}^u_{loc}(f^t_{\theta_u})) - \frac{1}{2L}\|\nabla \mathcal{L}^u_{loc}(f^\star_{\theta_u}) - \nabla \mathcal{L}^u_{loc}(f^t_{\theta_u})\|^2 \\
&\leq (\mathcal{L}^u_{loc}(f^\star_{\theta_u}) - \mathcal{L}^u_{loc}(f^t_{\theta_u})) - \frac{\mu_l^2}{2L}\|f^\star_{\theta_u} - f^t_{\theta_u}\|^2
\end{aligned}$$

Besides, the gradient of $\mathcal{L}^u_{loc}(f_{\theta_u}), \forall u \in [N]$ is bounded by a finite constant. That is, there exists a finite constant G_u satisfying that $\mathbb{E}[\|\nabla \mathcal{L}^u_{loc}(f_{\theta_u})\|^2] \leq G_u^2, forallu \in [N]$. Therefore, we can write

$$\begin{aligned}
&\mathbb{E}[\|f^{t+1}_{\theta_u} - f^\star_{\theta_u}\|^2] \\
&= \mathbb{E}[\|f^t_{\theta_u} - \eta I_t \nabla \mathcal{L}^u_{loc}(f^t_{\theta_u}) - f^\star_{\theta_u}\|^2] \\
&\leq \mathbb{E}[\|f^t_{\theta_u} - f^\star_{\theta_u}\|^2] + \frac{M\eta^2}{N}\mathbb{E}[\|\nabla \mathcal{L}^u_{loc}(f^t_{\theta_u})\|^2] + \frac{2M\eta}{N}(\mathcal{L}^u_{loc}(f^\star_{\theta_u}) - \mathcal{L}^u_{loc}(f^t_{\theta_u})) - \frac{M\mu_l^2\eta}{NL}\|f^\star_{\theta_u} - f^t_{\theta_u}\|^2 \\
&= (1 - \frac{M\mu_l^2\eta}{NL})\mathbb{E}[\|f^t_{\theta_u} - f^\star_{\theta_u}\|^2] + \frac{M\eta^2}{N}\mathbb{E}[\|\nabla \mathcal{L}^u_{loc}(f^t_{\theta_u})\|^2] + \frac{2M\eta}{N}(\mathcal{L}^u_{loc}(f^\star_{\theta_u}) - \mathcal{L}^u_{loc}(f^t_{\theta_u}))
\end{aligned}$$

Using the similar scheme adopted during the proof of Theorem 10 in [57], we can suppose there exists a constant A such that $\frac{\mathcal{G}(t+1)}{\mathcal{G}(t)} \geq 1 - \frac{\mathcal{G}(t)}{A}$ and the constant C satisfies that $C \geq \max\{\frac{\mathbb{E}[\|f^0_{\theta_u} - f^\star_{\theta_u}\|^2]}{\mathcal{G}(0)}, \frac{4NL^2G_u^2}{AM\mu_l^4}\}$. When we define that $\mathcal{L}^u_{loc}(f^\star_{\theta_u}) - \mathcal{L}^u_{loc}(f^t_{\theta_u}) \triangleq$

$\frac{\mu_l^2}{2L}\epsilon_K^2$, with a personalized learning rate $\eta = \frac{2NL\mathcal{G}(t)}{AM\mu_l^2}$, we can derive that

$$\begin{aligned}
\mathbb{E}[\|f_{\theta_u}^{t+1} - f_{\theta_u}^{\star}\|^2] &\le (1 - \frac{M\mu_l^2\eta}{NL})\mathbb{E}[\|f_{\theta_u}^{t} - f_{\theta_u}^{\star}\|^2] + \frac{M\eta^2}{N}G_u^2 + \frac{M\mu_l^2\eta}{NL}\epsilon_K^2 \\
&\le (1 - \frac{M\mu_l^2\eta}{NL})(C\mathcal{G}(t) + \epsilon_K^2) + \frac{M\eta^2}{N}G_u^2 + \frac{M\mu_l^2\eta}{NL}\epsilon_K^2 \\
&= (1 - \frac{M\mu_l^2\eta}{NL})C\mathcal{G}(t) + \frac{4NL^2G_u^2}{A^2M\mu_l^4}\mathcal{G}(t)^2 + \epsilon_K^2 \\
&\le \big(1 - \frac{2}{A}\mathcal{G}(t))C\mathcal{G}(t) + \frac{C}{A}\mathcal{G}(t)^2 + \epsilon_K^2 \\
&= \big(1 - \frac{\mathcal{G}(t)}{A})C\mathcal{G}(t) + \epsilon_K^2 \\
&\le C\mathcal{G}(t+1) + \epsilon_K^2
\end{aligned}$$

Since $\mathcal{L}_{loc}^u(f_{\theta_u}^t) - \mathcal{L}_{loc}^u(f_{\theta_u}^{\star}) = \mathcal{L}_{loc}^u(f_{\theta_u}^t; \Phi_u^t) - \mathcal{L}_{loc}^u(f_{\theta_u}^{\star}; \Phi_u^t) \to 0$ as $K \to \infty$, we know that $\lim_{K\to\infty} \epsilon_K^2 \to 0$. Thus, we complete the proof of Corollary 5.5.1. □

5.6 Experiments

5.6.1 Experimental Setup

Colored-MNIST (CMNIST) [6] is constructed based on MNIST [53] via rearranging the images of digit 0-4 into a single class labeled 0 and the images of digit 5-9 into another class labeled 1. Each digit having label 0 is colored green/red with probability $p^e/1-p^e$ and each digit having label 1 is colored red/green with probability $p^e/1-p^e$, respectively. Thus "color" builds a shortcut in this dataset and the data distribution varies as p^e changes. We provide two training environments ($p_{tr}^e = 0.90$ and 0.80) as $\mathcal{E}_{tr}$ and every local client only has one training environment which is randomly sampled from $\mathcal{E}_{tr}$. To assess the model performance on different test distributions, the test environment on each client varies from $p_{te}^e = 0.00$ to 1.00. Considering the heterogeneous data generating process across local clients, the data instances used for constructing the training/test environments on each client are randomly sampled from only two digit sub-classes labeled 0 and two digit sub-classes labeled 1 without replacement.

Colored-FMNIST (CFMNIST) [2] is constructed using the same strategy as Colored-MNIST, but the original images come from Fashion-MNIST [101]. Hence, CFMNIST possesses a more complex feature space compared to colored-MNIST.

WaterBird [81] considers a real-world scenario where the photographs of waterbirds usually have water backgrounds while the photographs of landbirds usually have land backgrounds because of the distinct habitats. It makes learning models easily trapped by "background" shortcut when classify "waterbird" and "landbird". In WaterBird, a waterbird is placed onto a water/land background with probability $p^e/1-p^e$ and a landbird is placed onto a land/water background with probability $p^e/1-p^e$ respectively. We setup two training environments ($p^e_{tr}=0.95$ and 0.85) as $\mathcal{E}_{tr}$ and each client has only one training environment which is randomly sampled from $\mathcal{E}_{tr}$. The test environment varies from $p^e_{te}=0.00$ to 1.00. We notice that the diverse geographic distributions of different bird species naturally accord with the heterogeneity of local data generating process if the federated clients are located in different geographic areas. Considering WaterBird includes 46 waterbird species and 154 landbird species, we distribute 15 (10 separated and 5 overlapped) waterbird species and 51 (34 separated and 17 overlapped) landbird species to each client. The training and test datasets on each client contain bird pictures that belong to the same bird species.

PACS [54] is a larger real-world dataset commonly used for evaluating out-of-distribution (OOD) generalization. It consists of 7 classes distributed across 4 environments (or domains). We adopt the "leave-one-domain-out" strategy to evaluate the OOD generalization performance. Taking personalization into consideration, we split each training domain into two subsets according to classes (i.e., one subset consists of dog, elephant and giraffe; another subset consists of guitar, horse, house, and person), and then distribute these two subsets onto two clients respectively. The training and test datasets on each client come from distinct domains but consist of the same classes.

Baseline Methods: We compare our method (FedPIN) with 11 state-of-the-art algorithms: four federated learning methods (FedAvg [70], DRFA [21], FedSR [73] and FedIIR [31]); and seven PFL methods (pFedMe [94], Ditto [57], FTFA [17], FedRep [18], FedRoD [12], FedPAC [102]) and FedSDR [97].

Model Architectures: For CMNIST and CFMNIST, we adopt a deep neural network with one hidden layer as feature extractor and a consecutive fully-connected layer as classifier. As regard to Waterbird and PACS, ResNet-18 [36] serves as the learning model, with the preceding layers acting as the feature extractor and the final fully-connected layer functioning as classifier.

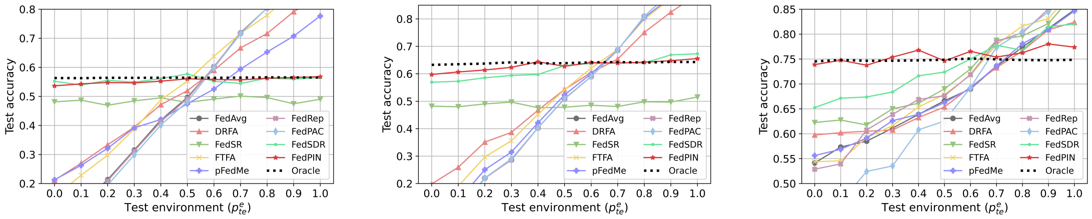
Table 5.1: The overall comparison between the performance of our method and the baselines on four datasets. When the number of clients is small, all clients are selected at each communication round. When the number of clients is large, the client sampling rate is set as 0.1.

Datasets	CMNIST				CFMNIST				WaterBird				PACS			
	8 clients		80 clients		8 clients		80 clients		8 clients		80 clients		6 clients		60 clients	
Test Acc (%)	Worst	Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst	Avg
FedAvg	3.4	51.0	1.7	46.8	0.2	50.0	0.8	45.6	54.1	68.0	48.7	61.6	41.7	47.7	33.8	40.2
DRFA	21.2	52.8	14.9	47.2	19.8	53.9	15.5	47.1	59.8	68.4	52.3	60.4	42.5	49.0	36.2	41.8
FedSR	46.9	48.6	40.3	43.6	47.6	48.9	41.2	43.3	61.8	71.7	55.6	64.3	46.8	51.3	39.0	43.4
FedIIR	47.3	48.4	41.2	42.9	48.1	49.2	41.8	43.6	61.2	70.9	54.3	64.6	47.0	51.6	40.2	44.4
FTFA	15.4	55.0	11.5	49.3	11.4	53.5	7.2	47.6	54.4	69.7	50.3	63.4	40.9	48.8	34.7	42.2
pFedMe	21.3	48.5	17.3	44.1	4.2	51.3	2.4	48.0	55.6	68.2	50.0	62.0	45.2	51.3	41.1	45.8
Ditto	3.0	51.0	2.1	45.8	0.4	50.1	1.8	45.7	53.1	68.7	49.1	63.4	44.9	51.3	40.2	46.3
FedRep	2.8	50.8	1.6	46.2	0.1	50.0	0.8	46.1	52.9	70.2	48.1	64.5	49.3	53.7	42.2	47.6
FedRoD	9.1	50.8	6.5	46.9	1.2	51.6	1.6	47.4	52.4	70.9	49.6	65.5	48.2	52.9	42.7	46.6
FedPAC	1.0	50.1	0.4	45.6	0.2	50.1	0.2	44.9	45.1	65.6	42.6	63.8	49.9	54.2	44.2	49.7
FedSDR	**53.9**	**55.6**	50.4	**51.8**	56.9	61.9	52.8	57.1	65.3	73.2	60.0	68.1	52.1	56.2	48.1	51.6
FedPIN	53.6	55.4	**50.8**	51.1	**59.8**	**63.1**	**56.4**	**59.5**	**73.8**	**75.8**	**67.9**	**71.3**	**55.4**	**58.6**	**52.3**	**54.8**

5.6.2 Overall Performance

To assess OOD generalization performance, we evaluate the test accuracy of the obtained models across a range of diverse test data distributions (11 test distributions in CMNIST, CFMNIST and WaterBird; 4 test distributions in PACS). Among them, the worst-case (Worst) accuracy and average (Avg) accuracy are summarized in Table 5.1. Since the test data distribution is unknown in practical scenarios, both the worst-case and average accuracy are significant for reflecting the OOD generalization performance of a model. As shown in Table 5.1, our method FedPIN outperforms the competitors on both worst-case and average test accuracy in three more complex datasets. In particular, FedPIN achieves around 3%, 8% and 3% higher worst-case accuracy than the second best algorithm on CFMNIST, WaterBird and PACS. Meanwhile, FedPIN achieves the highest average accuracy on these three datasets.

5.6.3 Mitigation of Spurious Correlations



(a) CMNIST

(b) CFMNIST

(c) WaterBird

Figure 5.2: The relationship between test accuracy and test distribution specified by p^e_{te} on three dataset where explicit shortcuts exists.

As mentioned in section 5.6.1, there exists explicit shortcuts in CMNIST, CFMNIST and WaterBird, and the degree of spurious correlations can be measured by the probability p^e. If a model abandons all correlations, it will achieve a consistent performance across varied test distributions specified by different p^e_{te}. Therefore, we show the relationship between test accuracy and p^e_{te} in Figure 5.2 to assess the efficacy of the concerned methods in mitigating spurious correlations. Moreover, we establish an oracle for comparison where the spurious features ('color' in CMNIST and CFMNIST; 'background' in WaterBird) are removed manually from the corresponding datasets. We can find the performance of our FedPIN closely matches that of the oracle on all three datasets, illustrating the effectiveness of FedPIN in mitigating spurious correlations. Conversely, the majority of state-of-the-art PFL methods struggle to eliminate spurious features since their performance varies dramatically as p_{te}^{e} changes.

5.6.4 Ablation Study

Table 5.2: The effect of the devised information-theoretic constraint in the local objective on achieving shortcut-averse personalization.

Datasets	CMNIST		CFMNIST		WaterBird		PACS	
Test Acc (%)	Worst	Avg	Worst	Avg	Worst	Avg	Worst	Avg
GM	47.5	49.6	48.2	50.1	63.4	71.5	47.2	51.5
GM-FT	15.1	54.8	10.7	56.2	62.8	72.3	45.5	52.3
GM-L2	46.9	50.0	48.6	50.8	64.9	73.0	48.2	53.4
PM ($\lambda = 0$)	20.2	54.5	18.2	55.7	64.0	72.8	46.1	53.0
PM ($\gamma = 0$)	52.8	55.2	58.6	63.2	69.8	75.4	52.9	56.3
PM	**53.6**	**55.4**	**59.8**	**63.1**	**73.8**	**75.8**	**55.4**	**58.6**

In this section, we analyse the effect of each part in the proposed information-theoretic regularizer and the results are depicted in Table 5.2. Specifically, 'GM' represents the performance of **G**lobal invariant **M**odel produced by FedPIN while 'PM' indicates the performance of **P**ersonalized invariant **M**odels developed by FedPIN. For comparison, we implement two effective personalization schemes in existing PFL: local **F**ine-**T**uning [17] and $L2$-norm regularization [58, 94, 32, 94], based on the global invariant model obtained by FedPIN. The results of these two baselines are labeled as

GM-FT and GM-$L2$ in Table 5.2. We can find these two schemes struggle to achieve personalization when the necessity of eliminating spurious correlation is considered. In particular, local fine-tuning can adversely impacts the OOD generalization performance of the global invariant model. The underlying reason is that these strategies cannot separate the personalized information from spurious features and preserving personalized features is accompanied with picking up spurious features.

In contrast, the proposed information-theoretic constraint can distinguish the personalized invariant features from spurious features and achieve shortcut-averse personalization. As regard to the two terms (conditional mutual information and entropy) in the constraint, we evaluate the isolated effects of them by independently setting $\lambda = 0$ and $\gamma = 0$ in Table 5.2. The results indicate that the conditional mutual information term weighted by λ is indispensable for excluding the spurious features. Of course, the entropy term weighted by γ can further improve the OOD generalization performance of the derived personalized invariant models.

5.6.5 Effect of Local Epochs

Table 5.3: Effect of the number of local epochs R in FedPIN.

# local epochs (R)	$R = 5$	$R = 10$	$R = 15$	$R = 20$
Worst-case Acc (%)	71.8	**73.8**	73.7	73.3
Average Acc (%)	74.4	75.8	76.2	**76.4**

Since allowing large number of local epochs can reduce the communication overhead in federated learning, we assess how varying the number of local epochs (i.e., R) impacts the performance of our method. The results on WaterBird dataset are presented in Table 5.3. Our method FedPIN exhibits robust performance across a range of R, as evidenced by the outcomes.

5.6.6 Visualization Results

For the purpose of verifying that the personalized models developed by our method FedPIN rely on the invariant features rather than spurious features, we randomly select one of the obtained personalized models and generate visual explanations for the selected model using Grad-CAM [83]. The commonly used Grad-CAM can produce a localization map which highlights the important regions in the input image for predicting the label. As shown in Figure 5.3, the pivotal features employed by various federated learning (FL) and personalized FL methods for prediction on WaterBird dataset are highlighted in red. The visualization results in Figure 5.3 support the claim that the personalized invariant features extracted by our method FedPIN are more related to the intended features (i.e., shape of the object), instead of the background.

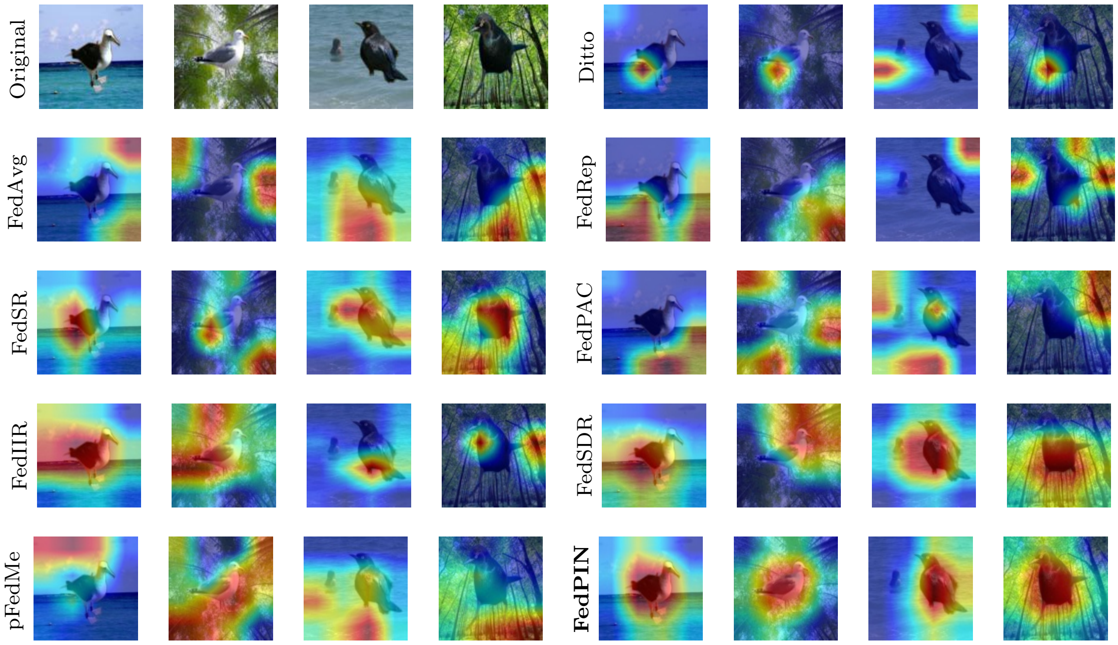


Figure 5.3: The visualization results of various federated learning (FL) and personalized FL methods on WaterBird dataset are generated by using Grad-CAM [83]. The red regions in the pictures correspond to high importance score for the predicted class. For optimal viewing, refer to the figure in color.

5.6.7 Computation Overhead

In order to evaluate the computation cost empirically, we record the running time that each algorithm consumes to achieve the reported performance in Table 5.1 on WaterBird dataset (with the client sampling rate set as 0.1). The detailed results are listed as follows:

Table 5.4: Empirical evaluation on computation cost of various algorithms.

Algorithm	FedAvg	DRFA	FedSR	FedIIR	FTFA	pFedMe	Ditto	FedRep	FedRoD	FedPAC	FedSDR	FedPIN
Running Time (s)	473	488	501	492	865	1490	1571	981	1379	1542	1565	1733

Combining the results in Table 5.4 and Table 5.1, we can find that our method FedPIN can achieve around 8% higher worst-case accuracy on WaterBird dataset than the second best baseline, with comparable computation cost over many state-of-the-art personalized federated learning approaches (e.g., pFedMe, Ditto, FedPAC and FedSDR).

5.7 Remark

In this chapter, a causal signature is proposed and quantified as an information-theoretic constraint to mitigate spurious correlations and achieve shortcut-averse personalized invariant learning under heterogeneous federated learning. The theoretical analysis demonstrates our method can guarantee a tighter generalization error bound in comparison with the state-of-the-art PFL methods and achieve a convergence rate on the same order as FedAvg. The results of extensive experiments affirm the superiority of the designed algorithm FedPIN over the competitors on out-of-distribution generalization performance. Moreover, FedPIN addresses the two major limitations of FedSDR presented in the previous chapter, thereby improving the applicability of our causally motivated personalized federated learning algorithm to real-world federated learning systems.

However, both FedSDR and FedPIN rely on elaborately constructed causal structural models (SCMs) as prior expert knowledge. Although the effectiveness of these SCMs has been demonstrated in related works and supported by our empirical results, their alignment with the true data distributions on each client has never been quantitatively evaluated. In practical large-scale federated learning systems, local datasets can exhibit highly diverse and even dynamic causal relationships, which can make predefined SCMs poorly suited to all heterogeneous clients. Extracting true and complex causal structural models from local datasets in large-scale heterogeneous federated learning systems using privacy-preserving methods remains an open problem, which we leave for future investigation.

Chapter 6

Conclusion and Future Work

As the final part, we offer a summary of this thesis and discussion on potential future research directions in this chapter. Specifically, section 6.1 concludes the research works and section 6.2 discusses several potential research directions that we can explore in the future.

6.1 Conclusion

This thesis investigates the problem of data distribution shift in collaborative learning, drawing inspiration from causal modeling. To thoroughly analyze this issue, we categorize distribution shifts in federated learning into two types: train-train distribution shift which describes inter-client distribution shift, and train-test distribution shift which represents intra-client distribution shift. In terms of train-train distribution shift, we design a personalized federated learning method with contextualized generalization (i.e., CGPFL), which can alleviate negative knowledge transfer among clients and facilitate faster model convergence. Regarding train-test distribution shift, we firstly propose a provable shortcut discovery and removal method (i.e., FedSDR) to extract personalized invariant representations with explicit environment information on clients. However, two outstanding limitations remain in the proposed FedSDR. In order to address these limitations, we present a personalized federated invariant learning method with a shortcut-averse information-theoretic constraint (i.e., FedPIN), capable of developing personalized invariant predictors for clients in more practical FL scenarios.

In contrast to existing personalized federated learning methods designed to tackle train-train distribution shift, CGPFL takes into account the latent contexts underlying federated clients and the negative knowledge transfer between distinct contexts. The proposed algorithm can cluster federated clients into multiple contexts and provide contextualized generalization knowledge to guide the training process of personalized models. Since the latent contexts provide fine-grained generalization knowledge and mitigate negative knowledge transfer among clients, CGPFL can enhance the accuracy and accelerate the convergence of the obtained personalized models. Theoretical analysis of the convergence rate indicates that CGPFL achieves faster model convergence compared to prevalent personalized federated learning methods. Moreover, the derived generalization error bound proves that CGPFL can achieve a tighter error bound in comparison to state-of-the-art personalized federated learning methods. Experimental results demonstrate that the proposed CGPFL achieves higher model accuracy and faster model convergence than baseline methods across diverse settings.

To the best of our knowledge, FedSDR is the first framework to address train-test distribution shift in personalized federated learning (PFL). Compared with the existing federated learning and personalized federated learning approaches, tackling train-test distribution shift in PFL necessitates simultaneously mitigating spurious correlations and preserving personalization information. FedSDR constructs structured causal models to simulate the heterogeneous data generation among federated clients and proposes two significant causal signatures. Inspired by these signatures, a provable shortcut discovery and removal method is designed to learn personalized invariant representations with available environment information on each client. Due to the stable causal relationship between personalized invariant representations and the target label, FedSDR achieves strong out-of-distribution generalization performance. Theoretical analysis ensures that FedSDR can produce optimal personalized invariant predictors for federated clients within linear representation spaces. The evaluation results demonstrate the superiority of FedSDR on out-of-distribution generalization compared to existing PFL methods.

Although FedSDR effectively addresses train-test distribution shift in personalized federated learning, it has two significant limitations: 1) the requirement for explicit environment information from each client can increase the risk of user privacy leakage; and 2) the theoretical guarantees are applicable only within linear representation spaces. FedPIN modifies the structured causal models for federated clients, based on which a shortcut-averse information-theoretic constraint is designed to achieve personalized invariant learning. Since the proposed shortcut-averse information-theoretic constraint is independent of environment information, FedPIN does not require environment labels on federated clients. Theoretical analysis proves that FedPIN can develop optimal personalized invariant predictors for clients in general representation spaces. Moreover, FedPIN achieves a tighter generalization error bound compared to existing personalized federated learning schemes. The evaluation results also demonstrate the effectiveness of FedPIN on addressing train-test distribution shift in personalized federated learning.

In summary, we address the data distribution shift in heterogeneous federated learning by proposing three innovative methods. Given that data distribution shift is prevalent in practical federated learning scenarios, our methods can not only contribute to the academic community of federated learning but also facilitate the deployment of federated learning in real-world applications.

6.2 Future Work

This thesis addresses the practical issue of heterogeneous data distribution in collaborative learning and inspires some potential research directions for future exploration, including learnability of personalization in collaborative learning, causal discovery in heterogeneous collaborative learning, personalized federated learning in the era of large-scale models and practical applicability in real-world federated learning systems.

Learnability of Personalization. As one of most important aspects in fundamental learning theory, learnability studies whether a specific concept is learnable with finite data samples using certain learning rules [98, 84]. Assessing the learnability of a machine learning problem can deepen researchers' understanding from a theoretical perspective. Although personalized federated learning has attracted considerable attention and gained significant success in recent years, characterizing the learnability of personalization in federated learning setting remains an open problem.

Compared to individual learning which trains a local model using only the local dataset, personalization requires collaboration across clients to acquire generalized knowledge. From the generalization error bound provided in Chapter 3 and Chapter 5, we can conclude that the performance of personalized federated learning is highly related to both train-train distribution shift and train-test distribution shift. Therefore, investigating the learnability of personalization in federated learning needs to take both inter-client distribution discrepancy and intra-client distribution shift into consideration. Moreover, the partial participating and random client selection strategy in federated learning setting can present unique challenges for evaluating the learnability of personalization in federated learning.

Causal Discovery in Heterogeneous Collaborative Learning. When we tackle the train-test distribution shift issue in Chapter 4 and Chapter 5, the useful causal graphs are provided as prior knowledge to simulate the heterogeneous data generation in federated learning. Discovering the true underlying causal graphs from local datasets can further improve the performance of the developed models and enhance the interpretability of personalized federated learning.

Despite extracting the underlying causal graph from the concerned dataset has been studied in the literature on causal discovery [92, 113, 60], these works focus on discovering causal graphs in centralized learning scenarios. The heterogeneous data distribution across federated clients and the requirement on data privacy-preserving introduces new challenges and opportunities to federated causal discovery. Therefore, integrating personalized invariant learning with federated causal discovery to address the train-test distribution shift problem presents a challenging and promising research direction for future investigation.

Personalized Federated Learning in the Era of Large-Scale Models. In the age of large-scale models, federated clients can also benefit from leveraging the pre-trained large-scale models. With their impressive ability to generalize across diverse applications, large-scale models have achieved significant success in both academic and industrial communities. Personalization of the pre-trained large-scale models can tailor models to individual user preferences and enhance user experience. A prevalent personalization scheme for pre-trained large-scale models is adapting with RLHF (i.e., reinforcement learning from human feedback) [74, 78, 56]. However, recent research has found that Reinforcement Learning with Human Feedback (RLHF) introduces a new problem, i.e., sycophancy, where fine-tuned models can prioritize user preferences at the expense of the correctness of output [37, 86]. Investigating whether heterogeneous user preference data in federated settings can address sycophancy in large-scale models represents a promising and intriguing research direction for future exploration.

Practical Applicability in Real-world Federated Learning Systems. Although causally motivated personalized federated learning frameworks proposed in this thesis have been evaluated in diverse federated settings, all of these evaluations have been conducted in simulated environments. To ensure the applicability of our algorithms to real-world federated learning systems, the following aspects need to be considered in future work:

1. **Scalability to large-scale federated learning system.** Due to limitations in available computational and data resources, the number of clients in our evaluation experiments was restricted to around one hundred. However, the number of participating clients in real-world federated learning systems can range from tens to millions, or even billions [88, 35, 8]. As the number of federated clients increases, data heterogeneity becomes more severe, particularly with respect to train–train data distribution shift across clients. Moreover, as discussed in Chapter 4 and Chapter 5, addressing the train–test data distribution shift is often interdependent with resolving train–train data distribution shift in practical federated learning systems. Therefore, applying the algorithms proposed in this thesis to real-world large-scale federated learning systems can introduce additional challenges. We leave this important research and industrial issue as a potential direction for future work.

2. **Complex, Dynamic, and Evolving Causal Structures.** As discussed in the third paragraph in this section, the structured causal models (SCMs) proposed in recent works, including our FedSDR and FedPIN, are employed as prior knowledge. However, in real-world federated learning systems, the underlying true SCMs governing local data distributions across clients can be more complex and even dynamic. Constructing SCMs directly from local datasets, rather than predefining approximate models, is crucial for enhancing the applicability of our causally motivated personalized federated learning algorithms in practical settings. Therefore, developing methods to extract complex, dynamic,

and evolving causal structures represents an important and valuable direction for future research.

3. **Privacy Preservation.** It is worth noting that all personalized federated learning algorithms presented in this thesis adopt the same global aggregation scheme as traditional federated learning methods (e.g., FedAvg [70]). This implies that our algorithms do not increase the risk of data privacy leakage compared to traditional FL approaches and remain compatible with prevalent schemes designed to enhance privacy protection during global communication, such as differential privacy [71, 100], secure aggregation [9], and homomorphic encryption [3, 106]. Therefore, integrating our causally motivated personalized learning frameworks with these established privacy-preserving methods to further strengthen privacy protection represents an interesting direction for future investigation.

References

[1] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

[2] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pages 145–155. PMLR, 2020.

[3] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE transactions on information forensics and security*, 13(5):1333–1345, 2017.

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

[5] Yossi Arjevani, Ohad Shamir, and Nathan Srebro. A tight convergence analysis for stochastic gradient descent with delayed updates. In *Algorithmic Learning Theory*, pages 111–132. PMLR, 2020.

[6] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[7] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[8] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, Brendan McMahan, et al. Towards federated learning at scale: System design. *Proceedings of machine learning and systems*, 1:374–388, 2019.

[9] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.

[10] Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. In *European conference on computer vision*, pages 548–564. Springer, 2020.

[11] 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)*, pages 1–9. IEEE, 2020.

[12] Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. In *International Conference on Learning Representations*, 2022.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[14] Yimeng Chen, Ruibin Xiong, Zhi-Ming Ma, and Yanyan Lan. When does group invariant learning survive spurious correlations? *Advances in Neural Information Processing Systems*, 35:7038–7051, 2022.

[15] Yining Chen, Elan Rosenfeld, Mark Sellke, Tengyu Ma, and Andrej Risteski. Iterative feature matching: Toward provable domain generalization with logarithmic environments. *Advances in Neural Information Processing Systems*, 35:1725–1736, 2022.

[16] Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. *Advances in Neural Information Processing Systems*, 35:22131–22148, 2022.

[17] Gary Cheng, Karan Chadha, and John Duchi. Federated asymptotics: a model to compare federated learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 10650–10689. PMLR, 2023.

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

[19] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.

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

[21] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. *Advances in Neural Information Processing Systems*, 33:15111–15122, 2020.

[22] Aymeric Dieuleveut, Gersende Fort, Eric Moulines, and Geneviève Robin. Federated-em with heterogeneity mitigation and variance reduction. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29553–29566. Curran Associates, Inc., 2021.

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

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

[25] Farzan Farnia and David Tse. A minimax approach to supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[26] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

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

[28] Tao Guo, Song Guo, and Junxiao Wang. Pfedprompt: Learning personalized prompt for vision-language models in federated learning. In *Proceedings of the ACM Web Conference 2023*, pages 1364–1374, 2023.

[29] Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. Promptfl: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*, 2023.

[30] Yaming Guo, Kai Guo, Xiaofeng Cao, Tieru Wu, and Yi Chang. Out-of-distribution generalization of federated learning via implicit invariant relationships. In *Proceedings of the 40th International Conference on Machine Learning*, pages 11905–11933, 2023.

[31] Yaming Guo, Kai Guo, Xiaofeng Cao, Tieru Wu, and Yi Chang. Out-of-distribution generalization of federated learning via implicit invariant relationships. In *Proceedings of the 40th International Conference on Machine Learning*, pages 11905–11933, 2023.

[32] Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtarik. Lower bounds and optimal algorithms for personalized federated learning. *Advances in Neural Information Processing Systems*, 33, 2020.

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

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

[35] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[37] Tom Hosking, Phil Blunsom, and Max Bartolo. Human feedback is not gold standard. In *The Twelfth International Conference on Learning Representations*, 2024.

[38] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pages 4387–4398. PMLR, 2020.

[39] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pages 4387–4398. PMLR, 2020.

[40] Bo-Wei Huang, Keng-Te Liao, Chang-Sheng Kao, and Shou-De Lin. Environment diversification with multi-head neural network for invariant learning. *Advances in Neural Information Processing Systems*, 35:915–927, 2022.

[41] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7865–7873, 2021.

[42] Dongsung Huh and Avinash Baidya. The missing invariance principle found – the reciprocal twin of invariant risk minimization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23023–23035. Curran Associates, Inc., 2022.

[43] Nam Hyeon-Woo, Moon Ye-Bin, and Tae-Hyun Oh. Fedpara: Low-rank hadamard product for communication-efficient federated learning. In *International Conference on Learning Representations*, 2022.

[44] Wonyong Jeong and Sung Ju Hwang. Factorized-fl: Personalized federated learning with parameter factorization & similarity matching. In *Advances in Neural Information Processing Systems*, 2022.

[45] Yibo Jiang and Victor Veitch. Invariant and transportable representations for anti-causal domain shifts. In S. Koyejo, S. Mohamed, A. Agarwal, D. Bel-

grave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 20782–20794. Curran Associates, Inc., 2022.

[46] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

[47] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[48] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR, 13–18 Jul 2020.

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

[50] Andreas Kirsch, Clare Lyle, and Yarin Gal. Unpacking information bottlenecks: Unifying information-theoretic objectives in deep learning. *arXiv preprint arXiv:2003.12537*, 2020.

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

[52] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[53] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[54] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[55] Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. Learning invariant graph representations for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, 2022.

[56] Lei Li, Yekun Chai, Shuohuan Wang, Yu Sun, Hao Tian, Ningyu Zhang, and Hua Wu. Tool-augmented reward modeling. In *The Twelfth International Conference on Learning Representations*, 2024.

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

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

[59] Yong Lin, Shengyu Zhu, Lu Tan, and Peng Cui. Zin: When and how to learn invariance without environment partition? *Advances in Neural Information Processing Systems*, 35:24529–24542, 2022.

[60] Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity constraints. *arXiv preprint arXiv:2107.10483*, 2021.

[61] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In *International Conference on Machine Learning*, pages 6804–6814. PMLR, 2021.

[62] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Integrated latent heterogeneity and invariance learning in kernel space. *Advances in Neural Information Processing Systems*, 34:21720–21731, 2021.

[63] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1013–1023, June 2021.

[64] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[65] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2022.

[66] Zhengquan Luo, Yunlong Wang, Zilei Wang, Zhenan Sun, and Tieniu Tan. Disentangled federated learning for tackling attributes skew via invariant aggregation and diversity transferring. In *International Conference on Machine Learning*, pages 14527–14541. PMLR, 2022.

[67] Zhengquan Luo, Yunlong Wang, Zilei Wang, Zhenan Sun, and Tieniu Tan. Disentangled federated learning for tackling attributes skew via invariant aggregation and diversity transferring. In *International Conference on Machine Learning*, pages 14527–14541. PMLR, 2022.

[68] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021.

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

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

[71] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.

[72] M Mohri. Foundations of machine learning, 2018.

[73] A Tuan Nguyen, Philip Torr, and Ser-Nam Lim. Fedsr: A simple and effective domain generalization method for federated learning. In *Advances in Neural Information Processing Systems*, 2022.

[74] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[75] Judea Pearl. *Causality*. Cambridge university press, 2009.

[76] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

[77] Maxime Peyrard, Sarvjeet Ghotra, Martin Josifoski, Vidhan Agarwal, Barun Patra, Dean Carignan, Emre Kiciman, Saurabh Tiwary, and Robert West. Invariant language modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5728–5743, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[78] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[79] Sashank J Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2020.

[80] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, volume 9, 2021.

[81] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[82] 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*, 2020.

[83] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations

from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[84] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

[85] Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

[86] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[87] Pranay Sharma, Rohan Panda, Gauri Joshi, and Pramod Varshney. Federated minimax optimization: Improved convergence analyses and algorithms. In *International Conference on Machine Learning*, pages 19683–19730. PMLR, 2022.

[88] Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):12598, 2020.

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

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

[91] Alessandro Sordoni, Nouha Dziri, Hannes Schulz, Geoff Gordon, Philip Bachman, and Remi Tachet Des Combes. Decomposed mutual information estimation for contrastive representation learning. In *International Conference on Machine Learning*, pages 9859–9869. PMLR, 2021.

[92] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. Springer, 2016.

[93] Zhenyu Sun and Ermin Wei. A communication-efficient algorithm with linear convergence for federated minimax learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 6060–6073. Curran Associates, Inc., 2022.

[94] Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

[95] Canh T Dinh, Nguyen Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33, 2020.

[96] Xueyang Tang, Song Guo, and Jingcai Guo. Personalized federated learning with contextualized generalization. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2241–2247. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[97] Xueyang Tang, Song Guo, Jie ZHANG, and Jingcai Guo. Learning personalized causally invariant representations for heterogeneous federated clients. In *The Twelfth International Conference on Learning Representations*, 2024.

[98] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[99] Haoxiang Wang, Haozhe Si, Bo Li, and Han Zhao. Provable domain generalization via invariant-feature subspace recovery. In *International Conference on Machine Learning*, pages 23018–23033. PMLR, 2022.

[100] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE transactions on information forensics and security*, 15:3454–3469, 2020.

[101] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[102] Jian Xu, Xinyi Tong, and Shao-Lun Huang. Personalized federated learning with feature alignment and classifier collaboration. In *The Eleventh International Conference on Learning Representations*, 2023.

[103] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[104] Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang. Towards a theoretical framework of out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

[105] Yu Yu, Shahram Khadivi, and Jia Xu. Can data diversity enhance learning generalization? In *Proceedings of the 29th international conference on computational linguistics*, pages 4933–4945, 2022.

[106] Chengliang Zhang, Suyi Li, Junzhe Xia, Wei Wang, Feng Yan, and Yang Liu. {BatchCrypt}: Efficient homomorphic encryption for {Cross-Silo} federated learning. In *2020 USENIX annual technical conference (USENIX ATC 20)*, pages 493–506, 2020.

[107] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[108] Jie Zhang, Song Guo, Xiaosong Ma, Haozhao Wang, Wenchao Xu, and Feijie Wu. Parameterized knowledge transfer for personalized federated learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[109] Jie Zhang, Song Guo, Xiaosong Ma, Haozhao Wang, Wenchao Xu, and Feijie Wu. Parameterized knowledge transfer for personalized federated learning. *Advances in Neural Information Processing Systems*, 34:10092–10104, 2021.

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

[111] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609, 2021.

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

[113] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.