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

Xueyang Tang [1 2] Song Guo [3] Jingcai Guo [1 4] Jie Zhang [1] Yue Yu [2]

Abstract

Exploiting invariant relations and mitigating spurious correlation (a.k.a., shortcut) between representation and target across varied data distributions can tackle the challenging out-of-distribution (OOD) generalization problem. In personalized federated learning (PFL), heterogeneous data distribution across local clients offers the inherent prerequisites to extract the invariant features that maintain invariant relation with target. Nevertheless, personalized features are closely entangled with spurious features in PFL since they exhibit similar variability across different clients, which makes preserving personalization knowledge and eliminating shortcuts two conflicting objectives in PFL. To address the above challenge, we analyse the heterogeneous data generation on local clients through the lens of structured causal model and propose a crucial causal signature which can distinguish personalized features from spurious features with global invariant features as the anchor. Then the causal signature is quantified as an information-theoretic constraint that facilitates the shortcut-averse personalized invariant learning on each client. Theoretical analysis demonstrates our 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 FedPIN in this paper. The results of extensive experiments show that our method can achieve superior OOD generalization performance compared with the state-of-the-art competitors.

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 (Geirhos et al., 2020). 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” (Beery et al., 2018). 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

[1]The Hong Kong Polytechnic University. [2]Peng Cheng Laboratory. [3]The Hong Kong University of Science and Technology. [4]Hong Kong Polytechnic University Shenzhen Research Institute. Correspondence to: Song Guo <songguo@cse.ust.hk>, Jingcai Guo <jc-jingcai.guo@polyu.edu.hk>, Jie Zhang <jie-comp.zhang@polyu.edu.hk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

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

distributions (Hsieh et al., 2020). 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 constitute 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., Guo et al. (2023)) 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., T Dinh et al. (2020); Luo et al. (2022); Xu et al. (2023)). 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 (Tang et al., 2024) 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 (Pearl, 2009)) 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, 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 (McMahan et al., 2017).
- 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.

2. Related Work

A more comprehensive review is included in Appendix A.

Invariant Learning (IL) Attaining causally invariant predictors over varied data distributions is proposed in the field of causal inference (Peters et al., 2016), and introduced into machine learning to tackle the OOD generalization problem by IRM (Arjovsky et al., 2019). Some subsequent works focus on achieving invariant learning when the environment label is unavailable, e.g., EIIL (Creager et al., 2021), HRM (Liu et al., 2021a), EDNIL (Huang et al., 2022) and ZIN (Lin et al., 2022).IFM (Chen et al., 2022a) lowers the requirement on the number of available environments, while iCaRL (Lu et al., 2022) extends IL to non-linear feature space. Another branch (Ahuja et al., 2021; Chen et al., 2022b; Huh & Baidya, 2022) completes the constraints that IRM misses. These works focus on centralized scenarios where all training data is accessible.

Heterogeneous Federated Learning (FL) Traditional FL develops a shared global model with raw data maintained on local clients, e.g., FedAvg (McMahan et al., 2017), DRFA (Deng et al., 2020), FedSR (Nguyen et al., 2022) and FedIIR (Guo et al., 2023). In contrast, PFL targets at producing a personalized model to fit the target dataset on each client. A typical strand of PFL methods trains personalized models with the guidance of a global model embedding shared knowledge (T Dinh et al., 2020; Hanzely et al., 2020; Fallah et al., 2020; Li et al., 2021; Tang et al., 2022; Cheng et al., 2023). DFL (Luo et al., 2022) disentangles shared features from the client-specific ones to achieve accurate aggregation on shared knowledge. FedRep (Collins et al., 2021), FedRoD (Chen & Chao, 2022) and FedPAC (Xu et al., 2023) employ the shared/aligned feature extractor to capture global knowledge and personalized classifiers to encode personalization information. Besides, FedSDR (Tang et al., 2024) devises a shortcut discovery and removal method to extract personalized invariant features in linear feature space with explicit environment labels available on local clients.

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 paper unless noted otherwise.

3.1. Invariant Learning (IL)

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

$$\mathbb{P}(Y|\Phi(X) = z, e) = \mathbb{P}(Y|\Phi(X) = z, e'), \forall z \in \mathcal{Z}, \forall e, e' \in \mathcal{E}_{all}. \quad (1)$$

Hence, the generic objective of invariant learning is to build an invariant feature extractor that fits the above invariance constraint. As Eq. (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 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. (1).

3.2. Causal Setup

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 (Pearl, 2009). When we study causal invariance in PFL, the heterogeneity among data generating mechanisms on local clients needs to be considered.

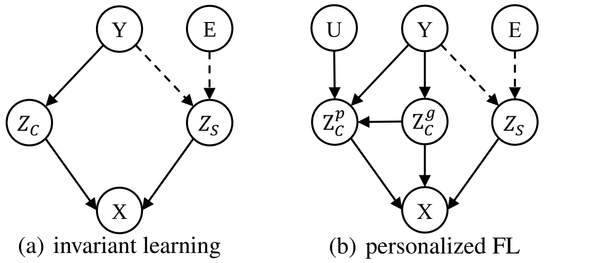


Figure 1. (a) presents the structural causal model (SCM) generally adopted in invariant learning, e.g., (Rosenfeld et al., 2021; Jiang & Veitch, 2022; Huh & Baidya, 2022), while (b) and show the SCM proposed in this paper. Z_C and Z_S denote the invariant and spurious features respectively. E is the indicator of shortcut while U is the indicator of user/client. Dotted arrows indicate unstable causal relations that can vary in different environments.

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 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 3.1 in invariant learning, we provide the definition of the optimal personalized invariant predictor in PFL.

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

4. Methodology

To handle the outstanding challenge that personalization information is closely entangled with spurious features, we resort to causal characteristics to differentiate them. Due to space limitations, we defer the detailed proofs of the theoretical analyses presented in this section to Appendix B.

Lemma 4.1. *If the data generating mechanism on each federated client complies with the causal graph in Figure 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.

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 4.1 demonstrates the feasibility of achieving personalized invariant learning under FL.

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 (Guo et al., 2023) 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 (Arjovsky et al., 2019) 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:

$$\max_{\Phi_g} I(Y; \Phi_g(X)) - \alpha I(E, U; Y \mid \Phi_g(X)), \tag{2}$$

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 4.2.

Proposition 4.2. *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*

$$\begin{aligned} I(E, U; Y \mid \Phi_g(X)) = &\min_{\omega_g} \mathbb{E}_u[\mathcal{R}(\omega_g(\Phi_g); D_u)] \\ &- \min_{\omega_a} \mathbb{E}_u[\mathcal{R}(\omega_a(\Phi_g, u); D_u)] \end{aligned} \tag{3}$$

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.*

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

$$\Phi_g^\star, \omega_g^\star, \omega_a^\star = \arg\min_{\Phi_g, \omega_g, \omega_a} \mathcal{L}_{glob}(\Phi_g, \omega_g, \omega_a), \tag{4}$$

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

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

Theorem 4.3. *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 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. (4) with an appropriately chosen hyper-parameter* α.

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 4.4. *For any representation* $h(X)$ *and* $h'(X)$ *where* h *and* h' *are two functions, under the SCM in Figure 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*

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

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 4.4 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 4.4 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:

$$\min_{\Phi_u, \omega_u} \mathcal{R}(\omega_u(\Phi_u); D_u) - \lambda I(\Phi_u(X); \Phi_g^\star(X)|Y) + \gamma H(\Phi_u(X)) \quad (5)$$

where λ and γ are non-negative balancing weights.

We provide formal theoretical analysis on the effectiveness of the local objective (5) in the subsequent Theorem 4.5.

Theorem 4.5. *If $f_{\theta_u}^\star \triangleq \omega_u^\star(\Phi_u^\star)$ is the minimizer of objective (5) with the hyper-parameter λ and γ chosen appropriately, then $f_{\theta_u}^\star$ is the optimal personalized invariant predictor that satisfies Definition 3.2 for the client u, $\forall u \in [N]$.*

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

$$\begin{aligned} & -\lambda I(\Phi_u(X); \Phi_g^\star(X) \mid Y) + \gamma H(\Phi_u(X)) \\ & \le \lambda \mathcal{L}_{con}^B(\Phi_u; \Phi_g^\star, \Psi_u^\star) + \gamma Var(\Phi_u(X)) - \lambda \log(|B_u| + 1), \end{aligned} \quad (6)$$

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

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

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 (Chen et al., 2020).

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

$$\min_{\Phi_u, \omega_u} \mathcal{R}(\omega_u(\Phi_u); D_u) + \lambda \mathcal{L}_{con}^B(\Phi_u; \Phi_g^\star, \Psi_u^\star) + \gamma Var(\Phi_u(X)). \quad (7)$$

4.3. Algorithm: FedPIN

Algorithm 1 FedPIN: **P**ersonalized **I**nvariant Lear**N**ing

```
Input: $T$, $R$, $K$, $\beta$, $\eta$, $\alpha$, $\lambda$, $\gamma$.
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 select a client subset $\mathcal{A}_t$, and broadcast global models $\omega_g^t(\Phi_g^t)$ and $\omega_a^t$ to them.
  for each client $u \in \mathcal{A}_t$ in parallel do
    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 $\Psi_u$:
      $\omega_u(\Phi_u) = \omega_u(\Phi_u) - \eta \nabla \mathcal{L}_{loc}^u(\omega_u(\Phi_u); \Phi_g^t, \Psi_u)$
    Solve the sub-problem of $\mathcal{L}_{glob}(\Phi_g, \omega_g, \omega_a)$:
    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
      $\tilde{\omega}_g^u, \tilde{\Phi}_g^u, \tilde{\omega}_a^u = \tilde{\omega}_g^u, \tilde{\Phi}_g^u, \tilde{\omega}_a^u - \beta \nabla \mathcal{L}_g^u(\tilde{\omega}_g^u, \tilde{\Phi}_g^u, \tilde{\omega}_a^u)$
    end for
    Send $\tilde{\omega}_g^u(\tilde{\Phi}_g^u)$ and $\tilde{\omega}_a^u$ back to the server.
  end for
  Server aggregates $\{\tilde{\omega}_g^u(\tilde{\Phi}_g^u), \tilde{\omega}_a^u | u \in \mathcal{A}_t\}$:
  $\omega_g^{t+1}(\Phi_g^{t+1}) = \frac{1}{|\mathcal{A}_t|} \sum_{u \in \mathcal{A}_t} \tilde{\omega}_g^u(\tilde{\Phi}_g^u)$
  $\omega_a^{t+1} = \frac{1}{|\mathcal{A}_t|} \sum_{u \in \mathcal{A}_t} \tilde{\omega}_a^u$
end for
return personalized invariant models $\{\omega_u(\Phi_u) | u \in [N]\}$.
```

In federated learning system, the global objective in Eq. (4) 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.

As regard to the local objective, it can be solved locally with the received global invariant feature extractor Φ_g^t. To simplify the expressions, we denote the local objective by

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

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 1.

4.4. Theoretical Analysis

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 4.7. *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, there exists a constant C that makes the following inequality hold with a probability at least $1 - \delta$:*

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

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

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

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

Remark 4.7. 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). 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 1, the 'OOD generalization term' can be eliminated by our FedPIN since the minimizer of objective (5) (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 4.5). 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. Detailed proof of Theorem 4.7 is provided in Appendix C.

Convergence Rate In this chapter, we will derive the convergence rate of FedPIN shown in Algorithm 1. The complete proofs are placed in Appendix D.

Assumption 4.8. Variance of local gradients to the aggregated average is upper bounded by a finite constant δ_L^2:

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

Theorem 4.9. *Suppose loss function $\mathcal{L}_g^u(\Phi_g, \omega_g, \omega_a)$, $\forall u \in [N]$ is L-smooth and assumption 4.8 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 $\{0, 1, ..., T-1\}$.

Theorem 4.9 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 4.10. *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 1 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 $K \to \infty$.

5. Experiments

5.1. Datasets

Colored-MNIST (CMNIST) (Arjovsky et al., 2019) is constructed based on MNIST (LeCun et al., 1998) 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

Table 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 (Ours)	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**

a shortcut in this 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 and two digit sub-classes labeled 1 without replacement.

Colored-FMNIST (CFMNIST) (Ahuja et al., 2020) is constructed using the same strategy as Colored-MNIST, but the original images come from Fashion-MNIST (Xiao et al., 2017). Hence, dataset CFMNIST possesses a more complex feature space compared to colored-MNIST.

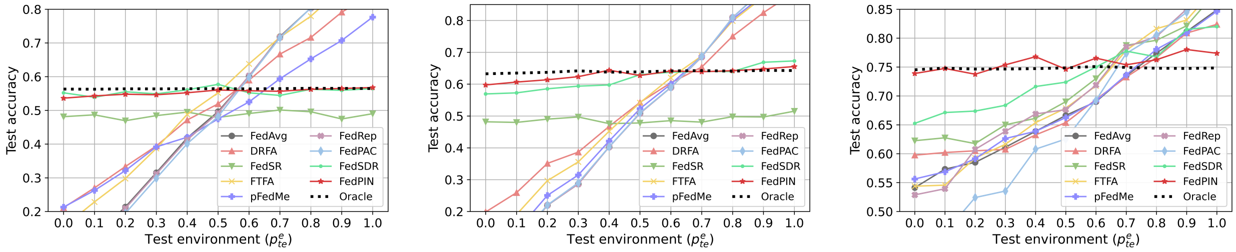
WaterBird (Sagawa et al., 2019) considers a real-world scenario where the photographs of waterbirds usually have water backgrounds while the photographs of landbirds usually have land backgrounds because of the distinct habitats. It makes learning models easily trapped by "background" shortcut when classify "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 (Li et al., 2017) 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.

5.2. Implementation

Model Selection 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 (He et al., 2016) serves as the learning model, with the preceding layers acting as the feature extractor and the final fully-connected layer functioning as classifier.

Competitors We compare our method (FedPIN) with 11 state-of-the-art algorithms: four federated learning methods (FedAvg (McMahan et al., 2017), DRFA (Deng et al., 2020), FedSR (Nguyen et al., 2022) and FedIIR (Guo et al., 2023)); and seven PFL methods (pFedMe (T Dinh et al., 2020), Ditto (Li et al., 2021), FTFA (Cheng et al., 2023), FedRep (Collins et al., 2021), FedRoD (Chen & Chao, 2022), FedPAC (Xu et al., 2023)) and FedSDR (Tang et al., 2024).



(a) CMNIST

(b) CFMNIST

(c) WaterBird

Figure 2. The relationship between test accuracy and test distribution specified by p^e_{te} on three dataset where explicit shortcuts exists.

Local clients and a centralized server are simulated on one workstation (Intel(R) Core(TM) i9-12900K CPU @ 3.20GHz with one NVIDIA GeForce RTX 3090 GPU).

5.3. Experimental Results

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 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 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.

Mitigation of Spurious Correlations As mentioned in section 5.1, 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 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^e_{te} changes.

Table 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**

Effect of Information-Theoretic Constraint In this paragraph, we analyse the effect of each part in the proposed information-theoretic regularizer and the results are depicted in Table 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 (Cheng et al., 2023) and $L2$-norm regularization (Li et al., 2020; T Dinh et al., 2020; Hanzely et al., 2020; T Dinh et al., 2020), 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 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 2. The results indicate that the conditional mu-

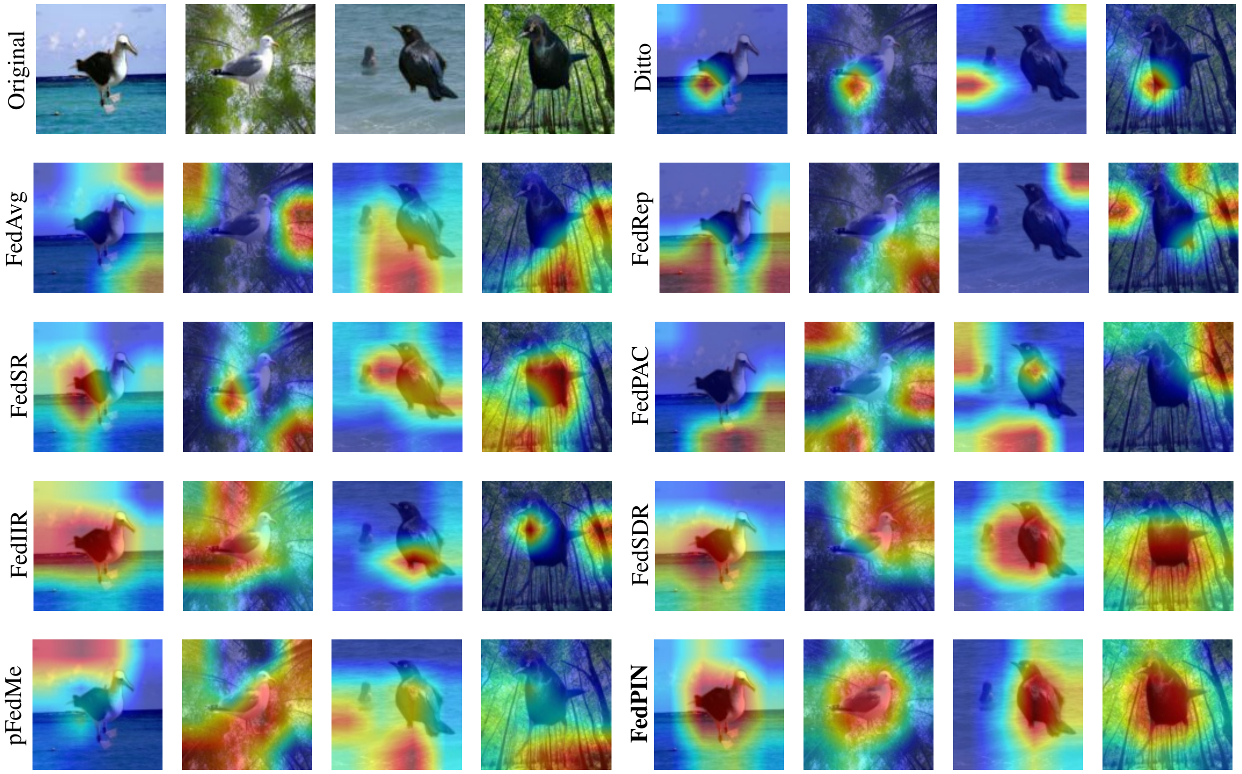


Figure 3. The visualization results of various federated learning (FL) and personalized FL methods on WaterBird dataset are generated by using Grad-CAM (Selvaraju et al., 2017). The red regions in the pictures correspond to high importance score for the predicted class. For optimal viewing, refer to the figure in color.

tual 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.

Table 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**

Effect of Local Epochs 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 3. Our method FedPIN exhibits robust performance across a range of R, as evidenced by the outcomes.

Visualization 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 (Selvaraju et al., 2017). 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 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 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.

6. Conclusion

In this paper, 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.

Acknowledgements

This research was supported by fundings from the Key-Area Research and Development Program of Guangdong Province (No. 2021B0101400003), Hong Kong RGC Research Impact Fund (No. R5060-19, No. R5034-18), Areas of Excellence Scheme (AoE/E-601/22-R), General Research Fund (No. 152203/20E, 152244/21E, 152169/22E, 152228/23E, No. 152211/23E), the National Natural Science Foundation of China (No. 62102327), the PolyU Internal Fund (No. P0043932).

Impact Statement

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

References

Ahuja, K., Shanmugam, K., Varshney, K., and Dhurandhar, A. Invariant risk minimization games. In *International Conference on Machine Learning*, pp. 145–155. PMLR, 2020.

Ahuja, K., Caballero, E., Zhang, D., Gagnon-Audet, J.-C., Bengio, Y., Mitliagkas, I., and Rish, I. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Beery, S., Van Horn, G., and Perona, P. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Boudiaf, M., Rony, J., Ziko, I. M., Granger, E., Pedersoli, M., Piantanida, P., and Ayed, I. B. A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. In *European conference on computer vision*, pp. 548–564. Springer, 2020.

Chen, H.-Y. and Chao, W.-L. On bridging generic and personalized federated learning for image classification. In *International Conference on Learning Representations*, 2022.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chen, Y., Rosenfeld, E., Sellke, M., Ma, T., and Risteski, A. Iterative feature matching: Toward provable domain generalization with logarithmic environments. *Advances in Neural Information Processing Systems*, 35:1725–1736, 2022a.

Chen, Y., Xiong, R., Ma, Z.-M., and Lan, Y. When does group invariant learning survive spurious correlations? *Advances in Neural Information Processing Systems*, 35: 7038–7051, 2022b.

Chen, Y., Zhang, Y., Bian, Y., Yang, H., Kaili, M., Xie, B., Liu, T., Han, B., and Cheng, J. Learning causally invariant representations for out-of-distribution generalization on graphs. *Advances in Neural Information Processing Systems*, 35:22131–22148, 2022c.

Cheng, G., Chadha, K., and Duchi, J. Federated asymptotics: a model to compare federated learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pp. 10650–10689. PMLR, 2023.

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

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Deng, Y., Kamani, M. M., and Mahdavi, M. Distributionally robust federated averaging. *Advances in Neural Information Processing Systems*, 33:15111–15122, 2020.

Dieuleveut, A., Fort, G., Moulines, E., and Robin, G. Federated-em with heterogeneity mitigation and variance reduction. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29553–29566. Curran Associates, Inc., 2021.

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

Farnia, F. and Tse, D. A minimax approach to supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Guo, Y., Guo, K., Cao, X., Wu, T., and Chang, Y. Out-of-distribution generalization of federated learning via implicit invariant relationships. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 11905–11933, 2023.

Hanzely, F. and Richtárik, P. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

Hanzely, F., Hanzely, S., Horváth, S., and Richtarik, P. Lower bounds and optimal algorithms for personalized federated learning. *Advances in Neural Information Processing Systems*, 33, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hsieh, K., Phanishayee, A., Mutlu, O., and Gibbons, P. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pp. 4387–4398. PMLR, 2020.

Huang, B.-W., Liao, K.-T., Kao, C.-S., and Lin, S.-D. Environment diversification with multi-head neural network for invariant learning. *Advances in Neural Information Processing Systems*, 35:915–927, 2022.

Huang, Y., Chu, L., Zhou, Z., Wang, L., Liu, J., Pei, J., and Zhang, Y. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7865–7873, 2021.

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

Hyeon-Woo, N., Ye-Bin, M., and Oh, T.-H. Fedpara: Low-rank hadamard product for communication-efficient federated learning. In *International Conference on Learning Representations*, 2022.

Jeong, W. and Hwang, S. J. Factorized-fl: Personalized federated learning with parameter factorization & similarity matching. In *Advances in Neural Information Processing Systems*, 2022.

Jiang, Y. and Veitch, V. Invariant and transportable representations for anti-causal domain shifts. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 20782–20794. Curran Associates, Inc., 2022.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. SCAFFOLD: Stochastic controlled averaging for federated learning. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 13–18 Jul 2020.

Kirsch, A., Lyle, C., and Gal, Y. Unpacking information bottlenecks: Unifying information-theoretic objectives in deep learning. *arXiv preprint arXiv:2003.12537*, 2020.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Li, H., Zhang, Z., Wang, X., and Zhu, W. Learning invariant graph representations for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, 2022.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Li, T., Hu, S., Beirami, A., and Smith, V. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021.

Lin, Y., Zhu, S., Tan, L., and Cui, P. Zin: When and how to learn invariance without environment partition? *Advances in Neural Information Processing Systems*, 35: 24529–24542, 2022.

Liu, J., Hu, Z., Cui, P., Li, B., and Shen, Z. Heterogeneous risk minimization. In *International Conference on Machine Learning*, pp. 6804–6814. PMLR, 2021a.

Liu, J., Hu, Z., Cui, P., Li, B., and Shen, Z. Integrated latent heterogeneity and invariance learning in kernel space. *Advances in Neural Information Processing Systems*, 34: 21720–21731, 2021b.

Liu, Q., Chen, C., Qin, J., Dou, Q., and Heng, P.-A. 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)*, pp. 1013–1023, June 2021c.

Lu, C., Wu, Y., Hernández-Lobato, J. M., and Schölkopf, B. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2022.

Luo, Z., Wang, Y., Wang, Z., Sun, Z., and Tan, T. Disentangled federated learning for tackling attributes skew via invariant aggregation and diversity transferring. In *International Conference on Machine Learning*, pp. 14527–14541. PMLR, 2022.

Mahajan, D., Tople, S., and Sharma, A. Domain generalization using causal matching. In *International Conference on Machine Learning*, pp. 7313–7324. PMLR, 2021.

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

Nguyen, A. T., Torr, P., and Lim, S.-N. Fedsr: A simple and effective domain generalization method for federated learning. In *Advances in Neural Information Processing Systems*, 2022.

Pearl, J. *Causality*. Cambridge university press, 2009.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012, 2016.

Peyrard, M., Ghotra, S., Josifoski, M., Agarwal, V., Patra, B., Carignan, D., Kiciman, E., Tiwary, S., and West, R. Invariant language modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5728–5743, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Rosenfeld, E., Ravikumar, P., and Risteski, A. The risks of invariant risk minimization. In *International Conference on Learning Representations*, volume 9, 2021.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shamir, O., Sabato, S., and Tishby, N. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

Sharma, P., Panda, R., Joshi, G., and Varshney, P. Federated minimax optimization: Improved convergence analyses and algorithms. In *International Conference on Machine Learning*, pp. 19683–19730. PMLR, 2022.

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. S. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

Sordoni, A., Dziri, N., Schulz, H., Gordon, G., Bachman, P., and Des Combes, R. T. Decomposed mutual information estimation for contrastive representation learning. In *International Conference on Machine Learning*, pp. 9859–9869. PMLR, 2021.

Sun, Z. and Wei, E. A communication-efficient algorithm with linear convergence for federated minimax learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 6060–6073. Curran Associates, Inc., 2022.

T Dinh, C., Tran, N., and Nguyen, J. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

Tang, X., Guo, S., and Guo, J. Personalized federated learning with contextualized generalization. In Raedt, L. D. (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 2241–2247. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

Tang, X., Guo, S., ZHANG, J., and Guo, J. Learning personalized causally invariant representations for heterogeneous federated clients. In *The Twelfth International Conference on Learning Representations*, 2024.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xu, J., Tong, X., and Huang, S.-L. Personalized federated learning with feature alignment and classifier collaboration. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhang, J., Guo, S., Ma, X., Wang, H., Xu, W., and Wu, F. Parameterized knowledge transfer for personalized federated learning. *Advances in Neural Information Processing Systems*, 34:10092–10104, 2021.

Zhang, J., Li, Z., Li, B., Xu, J., Wu, S., Ding, S., and Wu, C. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning*, pp. 26311–26329. PMLR, 2022.

In this appendix, we will provide the complete proofs for the theoretical analyses appearing in the main text and more details about the experimental setups.

A. Related Work

Invariant Learning (IL) Attaining causally invariant predictors over varied data distributions is proposed in the field of causal inference (Peters et al., 2016), and introduced into machine learning to tackle the OOD generalization problem by IRM (Arjovsky et al., 2019). Then, many efforts are dedicated to facilitating the application of IL to more general scenarios. Some works focus on achieving invariant learning when the environment label is unavailable, e.g., EIIL (Creager et al., 2021), HRM (Liu et al., 2021a), KerHRM (Liu et al., 2021b), EDNIL (Huang et al., 2022) and ZIN (Lin et al., 2022).IFM (Chen et al., 2022a) lowers the requirement on the number of available environments. Another branch (Ahuja et al., 2021; Chen et al., 2022b; Huh & Baidya, 2022) completes the constraints that IRM misses. Besides, iCaRL (Lu et al., 2022) extends IL to non-linear causal representations while ACTIR (Jiang & Veitch, 2022) extends IL to anti-causal scenarios. IL is also applied to graph representation learning (Li et al., 2022; Chen et al., 2022c) and natural language modeling (Peyrard et al., 2022). These methods are devised for centralized scenarios where all training data is accessible.

Federated Learning (FL) The classic FedAvg (McMahan et al., 2017) can perform well when local datasets are IID. A number of methods (e.g., SCAFFOLD (Karimireddy et al., 2020), FedEM (Dieuleveut et al., 2021) and FedLC (Zhang et al., 2022)) delve into alleviating the negative impact of training data heterogeneity on convergence rate, while another line (Deng et al., 2020; Sharma et al., 2022; Sun & Wei, 2022) targets at reducing the performance bias of global model on local clients. Few works (Liu et al., 2021c; Nguyen et al., 2022; Guo et al., 2023) 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 local 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 (T Dinh et al., 2020; Hanzely et al., 2020; Hanzely & Richtárik, 2020; Fallah et al., 2020; Li et al., 2021; Tang et al., 2022; Cheng et al., 2023), while another branch studies the parameterized knowledge transfer between similar clients, e.g., MOCHA (Smith et al., 2017), FedAMP (Huang et al., 2021) and KT-pFL (Zhang et al., 2021). DFL (Luo et al., 2022) disentangles the shared features from the client-specific ones to achieve accurate aggregation on shared knowledge. Similarly, pFedPara (Hyeon-Woo et al., 2022) and Factorized-FL (Jeong & Hwang, 2022) factorizes the model parameters into the shared and personalized parts. FedRep (Collins et al., 2021), FedRoD (Chen & Chao, 2022) and FedPAC (Xu et al., 2023) employ the shared/aligned feature extractor to capture global knowledge and personalized classifiers to encode personalization information. Besides, FedSDR (Tang et al., 2024) 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.

B. Objective Design and Theoretical Guarantees

B.1. Proof of Lemma 4.1 and Proposition 4.2

Lemma 4.1. If the data generating mechanism on each federated client complies with the causal graph in Figure 1(b), 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_S 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 (Pearl, 2009) 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. □

Proposition 4.2. 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

$$I(E, U; Y \mid \Phi_g(X)) = \min_{\omega_g} \mathbb{E}_u[\mathcal{R}(\omega_g(\Phi_g); D_u)] - \min_{\omega_a} \mathbb{E}_u[\mathcal{R}(\omega_a(\Phi_g, u); D_u)] \tag{8}$$

where the global invariant classifier ω_g accepts global features $\Phi(X)$ as input while the auxiliary classifier ω_a takes both global features $\Phi(X)$ 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

$$I(E, U; Y \mid \Phi_g(X)) = H(Y \mid \Phi_g(X)) - H(Y \mid E, U, \Phi_g(X)) \tag{9}$$

As discussed in (Farnia & Tse, 2016), 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. □

B.2. Proof of Theorem 4.3 and Lemma 4.4

Theorem 4.3. 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 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. (4) 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 $\min_{\omega_g} \mathcal{R}(\omega_\}(\oplus_\}^\star); \rceil, \sqcap) = \mathbb{E}[Y \mid \Phi_g^\star(X), u, e]$ (Mahajan et al., 2021). 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. (4) is equivalent to minimizing the following objective:

$$\begin{aligned} &\Phi_g^\star \in \arg\min_{\Phi_g, \omega_g} \mathbb{E}_u[\mathcal{R}(\omega_g(\Phi_g); D_u)] \\ &s.t. \qquad I(E, U; Y \mid \Phi_g(X)) = 0. \end{aligned} \tag{10}$$

Therefore, the two statements in Theorem 4.3 is equivalent to each other.

Proof ends. □

Lemma 4.4. For any representation $h(X)$ and $h'(X)$ where h and h' are two functions, under the SCM in Figure 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

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

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

Proof of the first conclusion: As claimed in Theorem 4.3, 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 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

$$\begin{aligned} I(h'(X); \Phi_g^\star(X) \mid Y) &\le I(Z_C^p, Z_S; \Phi_g^\star \mid Y) \\ &= H(Z_C^p, Z_S \mid Y) + H(\Phi_g^\star \mid Y) - H(Z_C^p, Z_S, \Phi_g^\star \mid Y) \end{aligned} \tag{11}$$

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

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

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 1(b), we can have that $Z_S \perp\!\!\!\perp Z_C^p \mid [\Phi_g^\star, Y]$. Thus, we can get that

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

Substituting the above two equations into the inequality (11), 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 ends. □

B.3. Proof of Theorem 4.5 and Proposition 4.6

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

Proposition B.1 (Lemma 2 (Boudiaf et al., 2020))**.** *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 $\mathcal{D}$.*

Theorem 4.5. If $f_{\theta_u}^\star = \omega_u^\star(\Phi_u^\star)$ is the minimizer of objective (5) with the hyper-parameter λ and γ chosen appropriately, then $f_{\theta_u}^\star$ is the optimal personalized invariant predictor that satisfies Definition 3.2 for the client u, $\forall u \in [N]$.

Proof. 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:

$$\hat{\Phi}_u = \min_{\Phi_u} -I(\Phi_u(X); \Phi_g^\star(X) \mid Y) + \rho H(\Phi_u(X)) \tag{12}$$

We justify this claim by constructing contradiction:

Using the d-separate criterion in Figure 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 a 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

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

Using the d-separate criterion, we have that $\hat{Z}_C \perp\!\!\!\perp \hat{Z}_S \mid Y$, therefore,

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

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 1(b), we can have that $\hat{Z}_S \perp\!\!\!\perp \hat{Z}_C \mid [\Phi_g^\star, Y]$. Thus, we can get that

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

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. (12). Contradiction appears. Therefore, $B = 0$ must hold if $\hat{\Phi}_u$ is the minimizer of Eq. (12).

Because $\hat{\Phi}_u(X)$ cannot depend on any components of Z_S, using the d-separate criterion in Figure 1(b), we can get that $Y \perp\!\!\!\perp E \mid \hat{\Phi}_u$ which indicates that $\mathbb{P}(Y|\hat{\Phi}_u(X) = z, e) = \mathbb{P}(Y|\hat{\Phi}_u(X) = z, e'), \forall z \in \mathcal{Z}, \forall e, e' \in \mathcal{E}^u_{all}$.

Meanwhile, according to Proposition B.1, 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) (denoted by $\Phi^\star_u$) can guarantee that

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

Proof ends. □

Proposition 4.6. When the local batch on client u is B_u and $\Psi^\star_u = \min_{\Psi_u, \omega_{\psi_u}} \mathcal{R}(\omega_{\psi_u}(\Psi_u); D_u), \forall u \in [N]$, the following inequality holds:

$$-\lambda I(\Phi_u(X); \Phi^\star_g(X) \mid Y) + \gamma H(\Phi_u(X)) \leq \lambda \mathcal{L}^B_{con}(\Phi_u; \Phi^\star_g, \Psi^\star_u) + \gamma Var(\Phi_u(X)) - \lambda \log(|B_u| + 1) \tag{13}$$

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

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

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 (Chen et al., 2020).

Proof. In each local batch B_u, the contrastive loss is constructed via regarding $\Phi^\star_u(X)$ and $\Phi_u(X)$ as positive pair while adopting $\Psi^\star_u(X), X \in B_u$ 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 (Sordoni et al., 2021), we can get that the conditional mutual information satisfies $I(\Phi_u(X); \Phi^\star_g(X) \mid Y) \geq \log(|B_u| + 1) - \mathcal{L}^B_{con}(\Phi_u; \Phi^\star_g, \Psi^\star_u)$. On the other hand, according to the proof of Proposition 4 in (Kirsch et al., 2020), 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:

$$-\lambda I(\Phi_u(X); \Phi^\star_g(X) \mid Y) + \gamma H(\Phi_u(X)) \leq \lambda \mathcal{L}^B_{con}(\Phi_u; \Phi^\star_g, \Psi^\star_u) + \gamma Var(\Phi_u(X)) - \lambda \log(|B_u| + 1)$$

Proof ends. □

C. Generalization Error Bound

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 $\mathcal{D}^{\mathcal{T}}_u$ 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 $D^{\mathcal{T}}_u$ by $p_u, \forall u \in [N]$.

Proposition C.1 (Lemma 11 (Shamir et al., 2010)). *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 inquality holds:*

$$\|p - \hat{p}\| \leq \frac{2 + \sqrt{2\log(1/\delta)}}{\sqrt{m}} \tag{14}$$

Theorem 4.7. In general, 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, there exists a constant C that makes the following inequality hold with a probability at least $1 - \delta$:

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

where $m_u \ge \frac{C}{4}\log(|\mathcal{Y}|/\delta)|\mathcal{X}|e^2$ 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

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

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

Proof. 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 process of proof, we define a useful real-valued function ξ as follows:

$$\xi(x) = \begin{cases} 0, & x = 0 \\ x \log(\frac{1}{x}), & 0 < x \le \frac{1}{e} \\ \frac{1}{e}, & x > \frac{1}{e} \end{cases}. \tag{15}$$

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

$$\begin{aligned}|\hat{I}_{\mathcal{S}}(Y; \Phi_u(X)) - I_{\mathcal{T}}(Y; \Phi_u(X))| &\triangleq |\hat{I}_{\mathcal{S}}(Y; \Phi_u) - I_{\mathcal{T}}(Y; \Phi_u)| \\ &= |\hat{I}_{\mathcal{S}}(Y; \Phi_u) - I_{\mathcal{S}}(Y; \Phi_u) + I_{\mathcal{S}}(Y; \Phi_u) - I_{\mathcal{T}}(Y; \Phi_u)| \\ &\le \underbrace{|\hat{I}_{\mathcal{S}}(Y; \Phi_u) - I_{\mathcal{S}}(Y; \Phi_u)|}_{\mathcal{A}_1} + \underbrace{|I_{\mathcal{S}}(Y; \Phi_u) - I_{\mathcal{T}}(Y; \Phi_u)|}_{\mathcal{A}_2}\end{aligned} \tag{16}$$

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

$$I(Y; \Phi) \triangleq H(\Phi) - H(\Phi \mid Y) \tag{17}$$

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

$$\begin{aligned}\mathcal{A}_1 &= \left|\hat{H}_{\mathcal{S}}(\Phi_u) - H_{\mathcal{S}}(\Phi_u) + H_{\mathcal{S}}(\Phi_u \mid Y) - \hat{H}_{\mathcal{S}}(\Phi_u \mid Y)\right| \\ &\le \left|H_{\mathcal{S}}(\Phi_u \mid Y) - \hat{H}_{\mathcal{S}}(\Phi_u \mid Y)\right| + \left|\hat{H}_{\mathcal{S}}(\Phi_u) - H_{\mathcal{S}}(\Phi_u)\right|\end{aligned} \tag{18}$$

For the first term on the right side of Eq. 18, we can write that

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

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

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

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

$$\Big|\sum_y p_{\mathcal{S}}(y)\big(H_{\mathcal{S}}(\Phi_u \mid y) - \hat{H}_{\mathcal{S}}(\Phi_u \mid y)\big)\Big| \leq \sum_y p_{\mathcal{S}}(y)\sum_{\phi_u}\xi\Big(\big\|p_{\mathcal{S}}(X|y) - \hat{p}_{\mathcal{S}}(X|y)\big\| \cdot \sqrt{V(p_{\mathcal{S}}(\phi_u|X))}\Big) \tag{19}$$

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 (Shamir et al., 2010). Therefore,

$$\Big|\sum_y \big(p_{\mathcal{S}}(y) - \hat{p}_{\mathcal{S}}(y)\big)\hat{H}_{\mathcal{S}}(\Phi_u \mid y)\Big| \leq \big\|p_{\mathcal{S}}(Y) - \hat{p}_{\mathcal{S}}(Y)\big\|\Big|\sum_y \hat{H}_{\mathcal{S}}(\Phi_u)\Big| = \big\|p_{\mathcal{S}}(Y) - \hat{p}_{\mathcal{S}}(Y)\big\|\big(|Y|\hat{H}_{\mathcal{S}}(\Phi_u)\big) \tag{20}$$

Combining Eq. (19) and Eq. (20), we can get

$$\begin{aligned}
H_{\mathcal{S}}(\Phi_u \mid Y) - \hat{H}_{\mathcal{S}}(\Phi_u \mid Y)| \leq &\sum_y p_{\mathcal{S}}(y)\sum_{\phi_u}\xi\Big(\big\|p_{\mathcal{S}}(X|y) - \hat{p}_{\mathcal{S}}(X|y)\big\| \cdot \sqrt{V(p_{\mathcal{S}}(\phi_u|X))}\Big) \\
&+ \big(|Y| \cdot \hat{H}_{\mathcal{S}}(\Phi_u)\big) \cdot \big\|p_{\mathcal{S}}(Y) - \hat{p}_{\mathcal{S}}(Y)\big\|
\end{aligned} \tag{21}$$

On the other hand, we have

$$\begin{aligned}
\big|H_{\mathcal{S}}(\Phi_u) - \hat{H}_{\mathcal{S}}(\Phi_u)\big| &= \Big|\sum_{\phi_u}\big(p_{\mathcal{S}}(\phi_u)\log(p_{\mathcal{S}}(\phi_u)) - \hat{p}_{\mathcal{S}}(\phi_u)\log(\hat{p}_{\mathcal{S}}(\phi_u))\big)\Big| \\
&\leq \sum_{\phi_u}\xi\big(\big|p_{\mathcal{S}}(\phi_u) - \hat{p}_{\mathcal{S}}(\phi_u)\big|\big) \\
&= \sum_{\phi_u}\xi\Big(\Big|\sum_x p_{\mathcal{S}}(\phi_u|x)\big(p_{\mathcal{S}}(x) - \hat{p}_{\mathcal{S}}(x)\big)\Big|\Big) \\
&= \sum_{\phi_u}\xi\Big(\Big|\sum_x \big(p_{\mathcal{S}}(\phi_u|x) - A\big)\big(p_{\mathcal{S}}(x) - \hat{p}_{\mathcal{S}}(x)\big)\Big|\Big) \\
&\leq \sum_{\phi_u}\xi\Big(\big\|p_{\mathcal{S}}(X) - \hat{p}_{\mathcal{S}}(X)\big\| \cdot \sqrt{V(p_{\mathcal{S}}(\phi_u|X))}\Big)
\end{aligned} \tag{22}$$

where the constant A is chosen as $A \triangleq \frac{1}{|X|}\sum_x p_{\mathcal{S}}(\phi_u|x)$. Plugging Eq. (21) and Eq. (22) into Eq. (18), we can get

$$\begin{aligned}
\mathcal{A}_1 \leq &\sum_y p_{\mathcal{S}}(y)\sum_{\phi_u}\xi\Big(\big\|p_{\mathcal{S}}(X|y) - \hat{p}_{\mathcal{S}}(X|y)\big\| \cdot \sqrt{V(p_{\mathcal{S}}(\phi_u|X))}\Big) + \big(|Y| \cdot \hat{H}_{\mathcal{S}}(\Phi_u)\big) \cdot \big\|p_{\mathcal{S}}(Y) - \hat{p}_{\mathcal{S}}(Y)\big\| \\
&+ \sum_{\phi_u}\xi\Big(\big\|p_{\mathcal{S}}(X) - \hat{p}_{\mathcal{S}}(X)\big\| \cdot \sqrt{V(p_{\mathcal{S}}(\phi_u|X))}\Big)
\end{aligned} \tag{23}$$

Subsequently, we can apply the concentration bound given in Proposition C.1 to $\big\|p_{\mathcal{S}}(X|y) - \hat{y}_{\mathcal{S}}(X|y)\big\|$, $\big\|p_{\mathcal{S}}(X) - \hat{p}_{\mathcal{S}}(X)\big\|$ and $\big\|p_{\mathcal{S}}(Y) - \hat{p}_{\mathcal{S}}(Y)\big\|$ for any y in Eq. (23). To make sure the bounds hold simultaneously over these $|Y\| + 2$ quantities, we replace δ in Eq. (14) by $\delta/(|Y|+2)$ as in the proof of Theorem 3 in (Shamir et al., 2010). Hence, with a probability at least $1-\delta$ we have

$$\mathcal{A}_1 \leq 2\sum_{\phi_u} \xi\Bigg(\Big(2+\sqrt{2\log((|Y|+2)/\delta)}\Big)\sqrt{\frac{V\big(p_{\mathcal{S}}(\phi_u|X)\big)}{m}}\Bigg) + \frac{2+\sqrt{2\log\big((|Y|+2)/\delta\big)}}{\sqrt{m}}\cdot\big(|Y|\hat{H}_{\mathcal{S}}(\Phi_u)\big) \tag{24}$$

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

$$m \geq \frac{C}{4}\log(|Y|/\delta)|X|e^2 \tag{25}$$

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

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

Using the results proved in the proof of Theorem 3 in (Shamir et al., 2010), we can have that $\sum_{\phi_u}\sqrt{\mathcal{V}(\phi_u)} \leq \sqrt{|X||\Phi_u|}$. Therefore, we can write that

$$\sum_{\phi_u}\xi\Bigg(\sqrt{\frac{C\log(|Y|/\delta)V(p_{\mathcal{S}}(\phi_u|X))}{m}}\Bigg) \leq \frac{\sqrt{C\log(|Y|/\delta)|X||\Phi_u|}\log(m) + \frac{2}{e}|\Phi_u|}{2\sqrt{m}} \tag{26}$$

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

$$\begin{aligned}\mathcal{A}_1 &\leq \frac{\sqrt{C\log(|Y|/\delta)}|X|\log(m) + \frac{2}{e}|X|}{\sqrt{m}} + \frac{\sqrt{C\log(|Y|/\delta)}|Y|\hat{H}_{\mathcal{S}}(\Phi_u)}{\sqrt{m}}\\ &= \frac{\sqrt{C\log(|Y|/\delta)}\Big(|X|\log(m) + |Y|\hat{H}_{\mathcal{S}}(\Phi_u)\Big) + \frac{2}{e}|X|}{\sqrt{m}}\end{aligned} \tag{27}$$

As regard to the second term in Eq. (16), we can write that

$$\begin{aligned}\mathcal{A}_2 &= |I_{\mathcal{T}}(Y;\Phi_u) - I_{\mathcal{S}}(Y;\Phi_u)|\\ &= \Big|\sum_y\sum_{\phi_u} p_{\mathcal{T}}(y,\phi_u)\log\Big(\frac{p_{\mathcal{T}}(y,\phi_u)}{p_{\mathcal{T}}(y)p_{\mathcal{T}}(\phi_u)}\Big) - p_{\mathcal{S}}(y,\phi_u)\log\Big(\frac{p_{\mathcal{S}}(y,\phi_u)}{p_{\mathcal{S}}(y)p_{\mathcal{S}}(\phi_u)}\Big)\Big|\\ &= \Big|\sum_y\sum_{\phi_u}\Big(p_{\mathcal{T}}(y,\phi_u)\log\big(p_{\mathcal{T}}(y|\phi_u)\big) - p_{\mathcal{S}}(y,\phi_u)\log\big(p_{\mathcal{S}}(y|\phi_u)\big)\Big) + H_{\mathcal{T}}(Y) - H_{\mathcal{S}}(Y)\Big|\end{aligned} \tag{28}$$

As shown in Figure 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

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

where C_S denotes a constant satisfying that $C_S = \max_{(\phi_u,y)\in(\Phi_u,Y)}\big|\log\big(p_{\mathcal{S}}(y|\phi_u)\big)\big|$. We know that $\log(\cdot)$ is a concave function, therefore we can get

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

Consequently, we can get that

$$\begin{aligned}
\mathcal{A}_2 &\leq \mathcal{KL}\big(p_{\mathcal{T}}(Y\mid\Phi_u)\big\|p_{\mathcal{S}}(Y\mid\Phi_u)\big) + \sqrt{|Y|C_S^2\Big(\mathcal{KL}\big(p_{\mathcal{T}}(Y\mid\Phi_u)\|p_{\mathcal{S}}(Y\mid\Phi_u)\big) + \mathcal{KL}\big(p_{\mathcal{S}}(Y\mid\Phi_u)\|p_{\mathcal{T}}(Y\mid\Phi_u)\big)\Big)} \\
&\leq \mathcal{J}\big(p_{\mathcal{T}}(Y\mid\Phi_u), p_{\mathcal{S}}(Y\mid\Phi_u)\big) + \sqrt{|Y|C_S^2\mathcal{J}\big(p_{\mathcal{T}}(Y\mid\Phi_u), p_{\mathcal{S}}(Y\mid\Phi_u)\big)}
\end{aligned} \tag{29}$$

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. (27) and l (29), we can conclude that

$$\begin{aligned}
|\hat{I}_{\mathcal{S}}(Y;\Phi_u(X)) - I_{\mathcal{T}}(Y;\Phi_u(X))| \leq{}& \frac{\sqrt{C\log(|Y|/\delta)}\Big(|X|\log(m) + |Y|\hat{H}_{\mathcal{S}}(\Phi_u)\Big) + \frac{2}{e}|X|}{\sqrt{m}} \\
&+ \mathcal{J}\big(p_{\mathcal{T}}(Y\mid\Phi_u), p_{\mathcal{S}}(Y\mid\Phi_u)\big) + \sqrt{|Y|C_S^2\mathcal{J}\big(p_{\mathcal{T}}(Y\mid\Phi_u), p_{\mathcal{S}}(Y\mid\Phi_u)\big)}
\end{aligned} \tag{30}$$

Thus, we complete the proof of Theorem 4.7. □

D. Convergence Analysis

In this section, we give the detailed proof for the convergence rate of the proposed algorithm FedPIN. 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.

Lemma D.1 (**Aggregation variance**). *When assumption 4.8 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*

$$\mathbb{E}_{S_t}\Big[\Big\|\frac{1}{M}\sum_{u \in S_t} \nabla \mathcal{L}_g^u(\theta_g^t) - \nabla \mathcal{L}_g(\theta_g^t)\Big\|^2\Big] \le \frac{N/M - 1}{N-1}\delta_L^2. \tag{31}$$

Proof. We can write that

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

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

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

Thus, we can write that

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

Therefore, we can derive that

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

□

Lemma D.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 $u\in[N]$:*

$$\frac{1}{R}\sum_{r=0}^{R-1}\big\|\theta_{g,u}^{t,r}-\theta_g^t\big\|^2\leq 8R^2\beta^2\|\nabla\mathcal{L}_g^u(\theta_g^t)\|^2. \tag{32}$$

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

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

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

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

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

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

Proof ends. □

Theorem 4.9. Suppose loss function $\mathcal{L}_g^u(\Phi_g, \omega_g, \omega_a), \forall u \in [N]$ is L-smooth and assumption 4.8 holds The number of the selected clients at each communication round is M. When the learning rate β satisfies that $\beta < \frac{1}{8RL}$, then the convergence rate of the **global model** is described by

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

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

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

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

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

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

$$\begin{aligned}&\mathbb{E}_{S_t}[\mathcal{L}_g(\theta_g^{t+1}) - \mathcal{L}_g(\theta_g^t)]\\ &\le \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]\\ &\le -\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]\\ &\le -\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]\\ &\le -\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]\\ &\le -\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}^3L^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}^3L^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]\end{aligned}$$

$$
\begin{aligned}
&\leq -\frac{\hat{\beta}}{2}\mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2] + \hat{\beta}^2 L\mathbb{E}_{S_t}[\|\psi^t - \nabla\mathcal{L}_g(\theta_g^t)\|^2] + \hat{\beta}^2 L\mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2] + 8\hat{\beta}^3 L^2\{\delta_L^2 + \mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2]\}\\
&= -\frac{\hat{\beta}}{2}\{1 - 2\hat{\beta}L - 16\hat{\beta}^2L^2\}\mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2] + 8\hat{\beta}^3L^2\delta_L^2 + \hat{\beta}^2 L\mathbb{E}_{S_t}[\|\psi^t - \nabla\mathcal{L}_g(\theta_g^t)\|^2].
\end{aligned}
$$

We firstly deal with the third term on the right side of above inequality as follows:

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

Using the inequalities in Lemma D.1 and Lemma D.2, we can get

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

Finally, we can get

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

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

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

Thus, we can derive that

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

In other words, we have

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

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

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

Using the schemes adopted in (Karimireddy et al., 2020; T Dinh et al., 2020; Tang et al., 2022), we consider the following two cases:

- When $\beta_0\leq\min\{(\frac{C_1}{C_2R^4T})^{\frac{1}{4}},(\frac{C_1}{C_3R^3T})^{\frac{1}{3}},(\frac{C_1}{C_4R^2T})^{\frac{1}{2}}\}$, we choose $\beta=\beta_0$. Then, we have

$$\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}_{S_t}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2]\leq\frac{C_1}{\beta_0RT}+\frac{C_1^{\frac{3}{4}}C_2^{\frac{1}{4}}}{T^{\frac{3}{4}}}+\frac{C_1^{\frac{2}{3}}C_3^{\frac{1}{3}}}{T^{\frac{2}{3}}}+\frac{C_1^{\frac{1}{2}}C_4^{\frac{1}{2}}}{T^{\frac{1}{2}}}.$$

- When $\beta_0\geq\min\{(\frac{C_1}{C_2R^4T})^{\frac{1}{4}},(\frac{C_1}{C_3R^3T})^{\frac{1}{3}},(\frac{C_1}{C_4R^2T})^{\frac{1}{2}}\}$, we choose $\beta=\min\{(\frac{C_1}{C_2R^4T})^{\frac{1}{4}},(\frac{C_1}{C_3R^3T})^{\frac{1}{3}},(\frac{C_1}{C_4R^2T})^{\frac{1}{2}}\}$. 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

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\mathcal{L}_g(\theta_g^t)\|^2]&\leq\mathcal{O}\Big(\frac{C_1}{\beta_0RT}+\frac{3C_1^{\frac{3}{4}}C_2^{\frac{1}{4}}}{T^{\frac{3}{4}}}+\frac{3C_1^{\frac{2}{3}}C_3^{\frac{1}{3}}}{T^{\frac{2}{3}}}+\frac{3C_1^{\frac{1}{2}}C_4^{\frac{1}{2}}}{T^{\frac{1}{2}}}\Big)\\&=\mathcal{O}\Big(\frac{\Delta_l}{\beta RT}+\frac{\Delta_l^{\frac{3}{4}}L^{\frac{3}{4}}\delta_L^{\frac{1}{2}}}{T^{\frac{3}{4}}}+\frac{\Delta_l^{\frac{2}{3}}L^{\frac{2}{3}}\delta_L^{\frac{2}{3}}}{T^{\frac{2}{3}}}+\sqrt{\frac{(N-M)\Delta_lL\delta_L^2}{M(N-1)T}}\Big)\end{aligned}$$

where $\Delta_l:=\mathbb{E}[\mathcal{L}_g(\theta_g^0)-\mathcal{L}_g(\theta_g^T)]$ and the learning rate β must satisfy $\beta\leq\frac{1}{8RL}$. □

Corollary 4.10 Assuming that local loss function $\mathcal{L}_{loc}^u(\omega_u(\Phi_u); \Phi_g^\star, \Psi_u^\star)$ is L-smooth and μ_l-strongly convex, and its gradient is upper bounded by a finite constant, $\forall u \in [N]$. If we define $f_{\theta_u} \triangleq \omega_u(\Phi_u)$, $f_{\theta_u}^\star = \arg\min_{\omega_u, \Phi_u} \mathcal{L}_{loc}^u(\omega_u(\Phi_u); \Phi_g^\star, \Psi_u^\star)$, and the output of Algorithm 1 after communication round T is $f_{\theta_u}^T$, 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 $C \geq \frac{\mathbb{E}[\|f_{\theta_u}^0 - f_{\theta_u}^\star\|^2]}{\mathcal{G}(0)}$, we have $\mathbb{E}[\|f_{\theta_u}^0 - f_{\theta_u}^\star\|^2] \leq C\mathcal{G}(0) + \epsilon_K^2$. Suppose $\mathbb{E}[\|f_{\theta_u}^t - f_{\theta_u}^\star\|^2] \leq C\mathcal{G}(t) + \epsilon_K^2$, for $t+1$, we can write

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

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 $\mathcal{L}_{loc}^u(f_{\theta_u})$ is L-smooth and μ_l-strongly convex, $\forall u \in [N]$, we have

$$\begin{aligned}\mathbb{E}[\langle \nabla \mathcal{L}_{loc}^u(f_{\theta_u}^t), f_{\theta_u}^\star - f_{\theta_u}^t\rangle] &\leq (\mathcal{L}_{loc}^u(f_{\theta_u}^\star) - \mathcal{L}_{loc}^u(f_{\theta_u}^t)) - \frac{1}{2L}\|\nabla \mathcal{L}_{loc}^u(f_{\theta_u}^\star) - \nabla \mathcal{L}_{loc}^u(f_{\theta_u}^t)\|^2 \\ &\leq (\mathcal{L}_{loc}^u(f_{\theta_u}^\star) - \mathcal{L}_{loc}^u(f_{\theta_u}^t)) - \frac{\mu_l^2}{2L}\|f_{\theta_u}^\star - f_{\theta_u}^t\|^2\end{aligned}$$

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

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

Using the similar scheme adopted during the proof of Theorem 10 in (Li et al., 2021), 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 $C \geq \max\{\frac{\mathbb{E}[\|f_{\theta_u}^0 - f_{\theta_u}^\star\|^2]}{\mathcal{G}(0)}, \frac{4NL^2G_u^2}{AM\mu_l^4}\}$. When we define that $\mathcal{L}_{loc}^u(f_{\theta_u}^\star) - \mathcal{L}_{loc}^u(f_{\theta_u}^t) \triangleq \frac{\mu_l^2}{2L}\epsilon_K^2$, with a personalized learning rate $\eta = \frac{2NL\mathcal{G}(t)}{AM\mu_l^2}$, we can derive that

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

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. □

E. More Details about Experiments

In this section, we will include more detailed setups and discussions on the evaluation part.

E.1. Non-IID data partition

For CMNIST and CFMNIST datasets, 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 the training environment set $\mathcal{E}_{tr}$. To assess the model performance on different test distributions, the test environment on each client varies across $p^e_{te} = 0.00, 0.10, ..., 0.90, 1.00$. 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 (1 separated and 1 overlapped) labeled 0 and two digit sub-classes (1 separated and 1 overlapped) labeled 1 without replacement. Specifically, we totally simulate eight local clients and one server in the federated learning system. For example, the data instances on client 1 are randomly sampled from digit 0, 1, 5, 6; the data instances on client 2 are randomly sampled from digit 1, 2, 6, 7; the data instances on client 3 are randomly sampled from digit 2, 3, 7, 8; and the data instances on client 8 are randomly sampled from digit 3, 4, 8, 9.

As regard to WaterBird, we distribute 15 (10 separated and 5 overlapped) waterbird species and 51 (34 separated and 17 overlapped) landbird species to each local client. Both the training and test data instances are constructed using bird photographs randomly sampled from the corresponding bird species in the bird dataset and background photographs randomly selected from the background dataset without replacement. Similarly, we totally simulate eight local clients and one server in the federated learning system.

PACS consists of 7 classes (i.e., dog, elephant, giraffe, guitar, horse, house, and person) distributed across 4 domains/environments (i.e., Art Painting, Cartoon, Photo and Sketch). We adopt the ”leave-one-domain-out” strategy to evaluate the out-of-distribution (OOD) generalization performance. For example, when we evaluate the performance on Art Painting domain, we use the remaining three domains (i.e., Cartoon, Photo and Sketch) as training environments. 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. The training and test datasets on each client come from different domains but consist of the same classes.

E.2. Implementation

Besides, the experiments are implemented in PyTorch. We simulate a set of clients and a centralized server on one deep learning workstation (Intel(R) Core(TM) i9-12900K CPU @ 3.20GHz with one NVIDIA GeForce RTX 3090 GPU).

E.3. Hyper-parameters

The hyper-parameters of the competitors and our algorithm are tuned to make the accuracy on the validation environment (i.e., $p^e_{val} = 0.10$ for CMNIST, CFMNIST and WaterBird; validation split in PACS) as high as possible. Specifically, the mainly used hyper-parameters in the evaluation part are listed as follows:

- **CMNIST:** Global communication round: $T = 600$, Local iterations: $R = 10$, Personalized epochs to update the personalized invariant predictors: $K = 10$, Local batch size: $B = 200$, Global learning rate: $\beta = 0.0001$, Personalized learning rate: $\eta = 0.0001$, Balancing weight: $\alpha = 1.0e5$, Balancing weight: $\lambda = 10.0$, Balancing weight: $\gamma = 6.0e{-}6$, Optimizer: Adam.
- **CFMNIST:** Global communication round: $T = 600$, Local iterations: $R = 10$, Personalized epochs to update the personalized invariant predictors: $K = 10$, Local batch size: $B = 200$, Global learning rate: $\beta = 0.0001$, Personalized learning rate: $\eta = 0.0001$, Balancing weight: $\alpha = 1.0e5$, Balancing weight: $\lambda = 10.0$, Balancing weight: $\gamma = 6.0e{-}6$, Optimizer: Adam.
- **WaterBird:** Global communication round: $T = 100$, 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: $\eta = 0.0001$, Balancing weight: $\alpha = 3.0e4$, Balancing weight: $\lambda = 9.0$, Balancing weight: $\gamma = 4.0e - 6$, Optimizer: Adam.
- **PACS:** Global communication round: $T = 600$, Local iterations: $R = 10$, Personalized epochs to update the personalized invariant predictors: $K = 10$, Local batch size: $B = 100$, Global learning rate: $\beta = 0.01$, Personalized learning rate: $\eta = 0.01$, Balancing weight: $\alpha = 1.0e4$, Balancing weight: $\lambda = 3.0$, Balancing weight: $\gamma = 1.0e - 5$, Optimizer: Adam.

E.4. Additional Experiments

In order to evaluate the computation cost empirically, we record the running time that each algorithm consumes to achieve the reported performance in Table 1 on WaterBird dataset (with the client sampling rate set as 0.1). The detailed results are listed as follows:

Table 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 4 and Table 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).