EXPLOITING PERSONALIZED INVARIANCE FOR BETTER OUT-OF-DISTRIBUTION GENERALIZATION IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, data heterogeneity among the training datasets on the local clients (a.k.a., Non-IID data) has attracted intense interest in Federated Learning (FL), and many personalized federated learning methods have been proposed to handle it. However, the distribution shift between the training dataset and testing dataset on each client is never considered in FL, despite it being general in real-world scenarios. We notice that the distribution shift (a.k.a., out-of-distribution generalization) problem under Non-IID federated setting becomes rather challenging due to the entanglement between personalized and spurious information. To tackle the above problem, we elaborate a general dual-regularized learning framework to explore the *personalized invariance*, compared with the exsiting personalized federated learning methods which are regularized by a single baseline (usually the global model). Utilizing the personalized invariant features, the developed personalized models can efficiently exploit the most relevant information and meanwhile eliminate spurious information so as to enhance the out-of-distribution generalization performance for each client. Both the theoretical analysis on convergence and OOD generalization performance and the results of extensive experiments demonstrate the superiority of our method over the existing federated learning and invariant learning methods, in diverse out-of-distribution and Non-IID data cases.

1 INTRODUCTION

As data privacy is attached more and more importance to in the field of machine learning, federated learning (McMahan et al., 2017) (FL) gains increasing attention and dramatic development in recent years. Federated learning allows the participation of a massive number of data holders (usually called clients) to train the learning models in a collaborative manner. Most importantly, the participating clients can preserve their data locally during the collaborative training process, and hence prevent the leakage of private data. Traditional federated learning approaches develop a shared gloal model by model aggregation to fit all the local datasets, which can work well when the local data instances on clients subjects to independent and identically distribution (IID). Conversely, the Non-IID setting where the local datasets don't satisfy the IID assumption can degrade the performance (both model accuracy and convergence rate) of the global model dramatically (Hsieh et al., 2020). To settle the challenge of Non-IID data, personalized federated learning (PFL) is proposed to train a personalized model for each client. Contrary to the individually trained local model which is often trapped in overfitting due to the insufficient local data, the personalized model obtains some generalization knowledge from the sever or the cooperation with other clients.

Despite the exsiting personalized federated learning methods achieve great success recently, we notice that all of them focus on ,the model performance on in-distribution testing dataset. In this way, the produced models can hardly adapt to the scenarios where the testing dataset is out-of-distribution in terms of the training dataset. Here, we refer to out-of-distribution (OOD) scenario as the case where there exists distribution shift between the training and testing dataset. In fact, the out-of-distribution scenario is more genreal and practical in machine learning applications, and has drawn lots of efforts in the literature on OOD generalization (Arjovsky et al., 2019; Sagawa et al., 2019; Creager et al., 2021; Ye et al., 2021; Shen et al., 2021; Krueger et al., 2021; Zhang et al., 2021c). However, to the best of our knowledge, the out-of-distribution generalization problem is mostly

studied in centralized setting where all data can be accessed, but has never been considered in the federated learning. Although some of the methods designed for centralized setting can be naively applied into the federated learning to generate a shared global model, they cannot cast off the Non-IID data quagmire because the gained model drops the useful information about personalization. In summary, the existing federated learning works can hardly handle the out-of-distribution scenarios, and meanwhile the centralized methods for OOD generalization will be troubled by the Non-IID data even if they can be implemented in federated learning setting.

What's worse, we find that the simple combination of the existing PFL and OOD methods, instead of solving the OOD generalization problem under Non-IID federated setting, will produce much worse performance than the better one of themselves (the finding will be shown and discussed in the experiment part). To handle the challenging OOD problem under Non-IID federated setting, we disscting the data heterogeneity (Non-IID) from two orthogonal perspectives and then introduce a novel concept: *personalized invariance*. The proposed personalized invariance can preserve the personal information, and in the meantime eliminate the spurious information to equip the obtained personalized models with out-of-distribution generalization. In contrast with the naive combination of PFL and OOD methods which can easily 'pick up' the spurious correlation during the process of personalization because of the entanglement of personalized and spurious information, we propose a general learning framework that is regularized by two baselines: global invariance and local invariance. The global invariance can prevent the personalized models from being stuck in overfitting due to the insufficiency of local data. However, global invariance will discard both personalized and spurious information. Therefore, the constraint of local invariance is utilized to exclude the spurious correlation. With the guidance of the designed dual-constraint, the personalized models can effectively exploit personalized invariance to solve the challenging OOD problem for every client in federated setting. The main contributions of this paper are summarized as follows:

- To the best of our knowledge, we are the first to consider and handle the challenging out-of-distribution problem under Non-IID federated learning setting. We propose the novel concept of *personalized invariance* and theoretically analyze the advantage of personalized invariance over the naive global invariance.
- We elaborate a dual-regularized learing framework to explore the personalized invariant features which can include important personalization information and meanwhile exclude the spurious correlation. Formally, the OOD problem under Non-IID setting is formulated as a bi-level optimization where the client-level objective is constrained by a dual-baseline regularization. Besides, we provide theoretical analysis on convergence rate and OOD generalization performance which proves the effectiveness of our method.
- We conduct extensive experiments on diverse datasets to compare the performance of our method with the existing federated learning (including personalized federated learning) and invariant learning methods. The results show that our method outperforms the state-of-the-art ones in varied out-of-distribution and Non-IID settings.

2 Related work

2.1 Personalized Federated Learning

One of the most challenging problems in federated learning (McMahan et al., 2017) is the data heterogeneity (a.k.a, Non-IID) across the local clients (Hsieh et al., 2020). To handle the Non-IID data quagmire, personalized federated learning (PFL) methods have been widely explored in many real-world applications. Jeong et al. (Jeong et al., 2018; Duan et al., 2020) focus on data augmentation methods by generating additional data to augment its local data towards yielding an IID dataset. However, these methods usually require the FL server to know some statistical information about the local data distributions (e.g., class sizes, mean and standard deviation), which may potentially violate privacy policy (Tan et al., 2021). Another line of work designs special client selection mechanisms to simulate homogeneous data distribution (Wang et al., 2020; Yang et al., 2020; Lyu et al., 2020).

On the other hand, many model-based PFL methods focus on producing customized model structures or parameters for different clients, which can also be divided into two types: single-model, multiple-model based approaches. Single-model based methods extended from the conventional FL algorithms like FedAvg (McMahan et al., 2017) combine the optimization of the local models and global model,

including: local fine-tuning (Wang et al., 2019; Schneider & Vlachos, 2019; Arivazhagan et al., 2019), regularization (T Dinh et al., 2020; Hanzely & Richtárik, 2020; Hanzely et al., 2020), model mixture (Mansour et al., 2020; Deng et al., 2020), meta learning (Fallah et al., 2020; Jiang et al., 2019) and parameter decomposition (Bui et al., 2019; Collins et al., 2021; Arivazhagan et al., 2019). Considering the diversity and inherent relationship of local data, multi-model-based approaches where multiple global models are trained for heterogeneous clients are more reasonable. Some researchers (Ghosh et al., 2020; Mansour et al., 2020; Tang et al., 2022) propose to train multiple global models at the server, where similar clients are clustered into a group to share the in-distribution generalization knowledge. Another strategy is to collaboratively train a personalized model without explicit global models, e.g., FedAMP (Huang et al., 2021), FedFomo (Zhang et al., 2021b), MOCHA (Smith et al., 2017), and KT-pFL (Zhang et al., 2021a). However, all of these works focus on the discrepancy between training data distributions rather than that between training and testing distributions.

2.2 Out-of-distribution Generalization

Out-of-Distribution (OOD) generalization problem refers to a challenging setting where the testing distribution is unknown and different from the training. In order to deal with the OOD generalization problem, tremendous efforts have been made and vary greatly ranging from causality to representation learning and from structure-based to optimization-based. The existing methods can be categorized into three parts: unsupervised representation learning (e.g., to analyse causal relationships between data) (Schölkopf et al., 2021; Shen et al., 2020a; Yang et al., 2021; Träuble et al., 2021; Locatello et al., 2019), supervised model learning (e.g., design various model architectures and learning strategies to achieve OOD generalization, including domain generalization (Wang et al., 2021; Zhao et al., 2020; Garg et al., 2021; Du et al., 2020), causal & invariant learning (Gamella & Heinze-Deml, 2020; Oberst et al., 2021; Krueger et al., 2021; Kamath et al., 2021; Arjovsky et al., 2019) and stable learning (Shen et al., 2018; 2020b; Zhang et al., 2021c)) and optimization approaches (focus on distributed robust optimization across different distributions (Liu et al., 2021c; Sagawa et al., 2019) or capturing invariant features (Liu et al., 2021a; Chang et al., 2020; Koyama & Yamaguchi, 2020)). All of the above OOD generalization methods are studied in centralized scenarios, where all training data can be accessed. In some real-world applications, the data are usually generated locally and the data owners are not willing to share data with others due to the concern about privacy. Therefore, we are motivated to investigate the OOD generalization problem under the federated setting.

3 Problem formulation

In this section, we first introduce the preliminary knowledge of invariant learning and then deconstruct the data heterogeneity into two orthogonal levels to investigate what information is necessary and what is spurious for solving the OOD problem under the Non-IID federated learning. For the purpose of fully exploiting the useful invariant information, we propose the personalized invariance which retains the important personalized information and concurrently excludes the spurious information to equip the obtained models with both personalization and out-of-distribution generalization.

Notations. Let $\mathcal{X}$ denotes the input space, $\mathcal{Y}$ denotes the target space, and correspondingly $(X, y) \in (\mathcal{X}, \mathcal{Y})$ is the data instance. The sets of training and testing environments are represented by $\mathcal{E}_{train}$ and $\mathcal{E}_{test}$ respectively. We use $\mathcal{E}_{all}$ as the set of all possible environments in the task concerned, i.e., $\mathcal{E}_{train}, \mathcal{E}_{test} \subset \mathcal{E}_{all}$. Suppose that there are N clients and the local dataset $D_i, i \in [N]$ contains m_i data instances on client i. For convenience, we separate the learning model or parameterized mapping from the input space $\mathcal{X}$ to $\mathcal{Y}$ into two consecutive parts: the feature extractor Φ and the classifier w. Specifically, the feature extractor Φ maps from the input space $\mathcal{X}$ to the latent feature space $\mathcal{H}$, i.e., $\Phi(X) \in \mathcal{H}$, and the classifier w generates a hard prediction $\hat{y}$ from a latent feature $\Phi(X)$. Therefore, the overall learning model is denoted by $f_\theta = \Phi \circ w$, where f_θ indicates the function f parameterized by θ. In this paper, we define the expected empirical loss on dataset D as $\mathcal{R}(f_\theta; D) := \mathbb{E}_{(X,y)\in D}[\ell(f_\theta(X), y)]$, where ℓ is the loss function.

3.1 Invariant Learning

Invraint learning has been emerged as a promising approach for handling OOD generalization problem (Arjovsky et al., 2019; Ahuja et al., 2020; Liu et al., 2021a;b; Creager et al., 2021). The

Table 1: The coverage of the related works.

Methods	Collaboration	Personalization	OOD Generalization
Federated learning, e.g., (McMahan et al., 2017)	✓	✗	✗
Personalized FL, e.g., (T Dinh et al., 2020; Fallah et al., 2020; Li et al., 2021; Cheng et al., 2021)	✓	✓	✗
Invariant learning, e.g., (Arjovsky et al., 2019; Sagawa et al., 2019; Creager et al., 2021)	✗	✗	✓
PerInvFL (Ours)	✓	✓	✓

prevalent invariant learning assumes that there exists some invariant feature $\Phi(X)$ satisfying the **invariance property**:

$$\mathbb{E}[y|\Phi(X)=h,e]=\mathbb{E}[y|\Phi(X)=h,e'],\forall h\in\mathcal{H},\forall e,e'\in\mathcal{E}_{all} \tag{Invariance}$$

To discover the invariant representation $\Phi:\mathcal{X}\to\mathcal{H}$ that elicits an invariant predictor $w\circ\Phi$ across all environments, ***IRM*** (Arjovsky et al., 2019) is proposed as a constrained optimization problem based on the empirical data from multiple accessible training environments $\mathcal{E}_{train}$. Out of efficiency concerns, a practical version of ***IRM*** is offered in that work:

$$\min_{f_\theta:\mathcal{X}\to\mathcal{Y}}\sum_{e\in\mathcal{E}_{train}}\mathcal{R}(f_\theta;e)+\lambda\cdot\|\nabla_{\bar{w}|\bar{w}=1.0}\mathcal{R}(\bar{w}\cdot f_\theta;e)\|^2, \tag{IRMv1}$$

where $\mathcal{R}(f_\theta;e)$ is the expected empirical loss on environment e. In particular, f_θ is used as the entire invariant predictor, and $\bar{w}$ is a scalar and fixed "dummy" classifier as in (Arjovsky et al., 2019). We define that $\mathcal{L}_{IRM}(\theta;\mathcal{E}_{train})=\sum_{e\in\mathcal{E}_{train}}\mathcal{R}(f_\theta;e)+\lambda\cdot\|\nabla_{\bar{w}|\bar{w}=1.0}\mathcal{R}(\bar{w}\cdot f_\theta;e)\|^2$ in this paper.

Apart from ***IRM***, another branch of approaches to OOD problem is worst-case generalization (Sagawa et al., 2019). The target of worst-case generalization is to minimize the expected risk of the training enviroment which produces the worst risk such that the generated predictor can be simultaneously optimal for all environments. One typical method is ***GroupDRO*** given in (Sagawa et al., 2019), of which the objective is:

$$\min_{f_\theta:\mathcal{X}\to\mathcal{Y}}\max_{q\in\Delta_g}\sum_{e\in\mathcal{E}_{train}}q_e\cdot\mathcal{R}(f_\theta;e), \tag{GroupDRO}$$

where $g=|\mathcal{E}_{train}|$ is the numbers of accessible training environments. Similarly, we define that $\mathcal{L}_{GroupDRO}(\theta;\mathcal{E}_{train})=\max_{q\in\Delta_g}\sum_{e\in\mathcal{E}_{train}}q_e\cdot\mathcal{R}(f_\theta;e)$ in this paper.

3.2 PERSONALIZED INVARIANCE

In order to study the OOD generalization problem under Non-IID federated setting, we first dissect the data heterogeneity from two orthogonal perspectives and categorize the related works in Table 1 according to the coverage of their methods.

Client-level heterogeneity. The distributions of training datasets are usually heterogeneous across the clients. Since the local data on every client is scarce and of limited diversity, the clients in FL framework need to collaborate to mine the common/generalization knowledge with some shared/global objective. However, the client-level heterogeneity can result in the discrepancy between the optimas of local objective and global objective. In other words, the client-independent representation (i.e., global knowledge) can be biased for some participating clients. Therefore, preserving the personalization information is of great importance for dealing with the client-level data heterogeneity.

Context-level heterogeneity. In addition to the heterogeneity among the training datasets, the heterogeneity between the training and testing dataset is also worthy of attentions. In FL, the local training and testing dataset on each client may be generated in distinct contexts/environments[1]. For example, the training image samples on a client are mainly captured from the local cameras, while the testing images may come from the web and with different styles. From the perspective of context, the target learning models must be equipped with OOD generalization so that they can perform both well on unknown and distinct contexts.

As discussed above, we can conclude that solving the OOD problem under Non-IID federated setting needs to discover the invariant representation without dropping the important personalized information. When we consider to implement the distributed version of invariant learning, such as ***IRM*** and ***GroupDRO***, each client can be regarded as an environment. In this way, both the context-specific (i.e., spurious) representation and the client-related (personalized) representation will be

[1]In this paper, environment, context and domain are used equivalently.

treated as spurious representations. Only the common invariant representation across the clients is extracted and utilized to elicit the invariant predictors. Consequently, the important personalized representation is dropped.

Assumption 1 *(**Heterogeneity of invariance**) For each client i, there exists a set of invariant representations Φ_i that satisfies the invariance property on the local set of environments $\mathcal{E}^i_{all}$ as defined in equation Invariance. There are $K = pN$ clients ($[N_K] \subset [N]$) on each of which the corresponding set of invariant representations $\Phi_j (j \in [N_K])$ contain some representations Z_j satisfying:*

*(i) **Heterogeneity:** Z_j is a subset of Φ_j and not contained in the intersection $\bigcap_{i\neq j, i\in[N]} \Phi_i$,*

*(ii) **Informativeness:** $\max_{z\subset Z_j} I(Y; z) = \delta > 0$, where $I(\cdot;\cdot)$ measures the mutual information (**MI**) between two random variables.*

The heterogeneity condition in Assumption 1 is consistent with the Non-IID assumption in FL, and the informativeness assumption claims the existence of meaningful personalization. If we define *global invariant* representation as $\Phi_g = \bigcap_{i\in[N]} \Phi_i$ and *personalized invariant* representation as Φ_i, $i \in [N]$. In the prediction task, the optimal global and personalized invariant representations are given by $\Phi^\star_i = \arg\max_{S\subset\Phi_i} I(Y; S)$ and $\Phi^\star_g = \arg\max_{S\subset\Phi_g} I(Y; S)$, respectively.

Theorem 1 *If the Assumption 1 holds in FL, the proposed personalized invariant representations can be constantly more informative than the global invariant representations obtained by the distributed implementation of existing invariant learning, for the prediction performance. That is,*

$$\frac{1}{N}\sum_{i=1}^{N} I(Y; \Phi^\star_i) \geq \frac{1}{N}\sum_{i=1}^{N} I(Y; \Phi^\star_g) + p\delta, \tag{1}$$

where $0 < p < 1$ is a constant and δ is a positive constant that is independent of N.

Theorem 1 demonstrates the importance of the proposed personalized invariance in improving the average prediction performance for OOD problems under Non-IID federated setting. The detailed proof of Theorem 1 is provided in the Appendix.

4 Dual-regularization constrained learning framework

Adhering to the prevalent setup of invariant learning, we suppose there are totally K training contexts in the federated learning system. On client i, the training data $(X, y) \in D_i$ is generated from K_i contexts while the testing data comes from unknown contexts. That is, $\mathcal{E}^i_{train} = \{e^i_j | j \in S_i\}$, where S_i is a subset of $[K]$ and the size of S_i equals K_i.

4.1 Algorithm Design

In ideal circumstances, we can elicit the personalized invariant predictor for each client i ($i \in [N]$) using $\theta^\star_i \in \arg\min_{\theta_i} \mathcal{L}_{INV}(\theta_i; \mathcal{E}^i_{train})$, where $\mathcal{L}_{INV} \in \{\mathcal{L}_{IRM}, \mathcal{L}_{GroupDRO}\}$. We refer to 'ideal' circumstances as the scenarios where the following two conditions are met: **1)** the number of training contexts on every client is sufficient for deriving the invariant predictor, **2)** the data samples in every training context are of sufficient diversity. Unfortunately, both of the above two conditions can hardly be satisfied in FL setting since the local data is usually scarce (McMahan et al., 2017). On the one hand, the latest efforts prove that the minimizer of $\mathcal{L}_{IRM}$ will necessarily utilize the non-invariant features and therefore cannot universally generalize to unknown testing contexts when the number of training contexts is not sufficient (Rosenfeld et al., 2021). On the other hand, the limited data diversity in each training contexts will make the obtained model easily trapped by overfitting.

To handle the challenges of insufficient training contexts and limited data diversity in each available context . In this paper, we elaborate a dual-regularization constrained optimization framework to handle these challenges of OOD generalization problems in federated scenarios. The overall objective on client $i (i \in [N])$ is:

$$\min_{\theta_i} \mathcal{L}_{INV}(\theta_i; \mathcal{E}^i_{train}) + \beta\|\theta_i - \nu^\star\|^2 \tag{2}$$

$$\text{s.t.} \quad \nu^\star \in \arg\min_{\nu} \mathcal{L}_{INV}(\nu; \mathcal{E}^D) \tag{3}$$

where $\mathcal{E}^{\mathcal{D}} = \{D_i | i = 1, 2, ..., N\}$ and $\mathcal{L}_{INV} \in \{\mathcal{L}_{IRM}, \mathcal{L}_{GroupDRO}\}$. Different from the popular single-regularized (i.e., global-regularized) personalized federated learning schemes (T Dinh et al., 2020; Hanzely & Richtárik, 2020; Deng et al., 2020; Hanzely et al., 2020; Li et al., 2021), the objective on each client is constrained by two regularizations. One is the 'local invariance' expressed by the first term $\mathcal{L}_{INV}(\theta_i; \mathcal{E}^i_{train})$ and nother is the 'global invariance' represented by the second term $\|\theta_i - \nu^\star\|^2$. The 'local invariance' is necessary for discovering the personalized invariant predictor, while the 'global invariance' is adopted to deal with the challenges of insufficient training contexts and limited data diversity in each context. Without loss of generality, we focus on studying the case where $\mathcal{L}_{INV} = \mathcal{L}_{IRM}$ in this paper. Therefore, the detailed objective on client i can be written as

$$\min_{\theta_i} \sum_{j=1}^{K_i} \mathcal{R}(f_{\theta_i}; e^i_j) + \lambda \cdot \|\nabla_{\bar{\omega}|\bar{\omega}=1.0} \mathcal{R}(\bar{\omega} \cdot f_{\theta_i}; e^i_j)\|^2 + \beta \|\theta_i - \nu^\star\|^2 \tag{4}$$

$$\text{s.t.} \quad \nu^\star \in \arg\min_{\nu} \sum_{i=1}^{N} \mathcal{R}(f_\nu; D_i) + \lambda \cdot \|\nabla_{\bar{\omega}|\bar{\omega}=1.0} \mathcal{R}(\bar{\omega} \cdot f_\nu; D_i)\|^2 \tag{5}$$

On each client, the outer problem in objective (2) can be solved locally with the guidance of the global model $\nu^\star$ downloaded from the server. In order to solve the inner problem in a collaborative manner, the objective (3) needs to be decomposed into N subproblems that can be individually and parallelly solved at the local clients. In fact, the objective of ***IRM*** ($\mathcal{L}_{IRM}(\nu; \mathcal{E}_D)$) can be naturally decomposed into N subproblems: $\mathcal{L}^i_{IRM} = \mathcal{R}(f_\nu; D_i) + \lambda \cdot \|\nabla_{\bar{w}|\bar{w}=1.0} \mathcal{R}(\bar{w} \cdot f_\nu; D_i)\|^2, i \in [N]$. As regard to ***GroupDRO***, the objective $\mathcal{L}_{GroupDRO}(\nu; \mathcal{E}_{\mathcal{D}})$ can be decomposed into $\mathcal{L}^i_{GroupDRO} = q_i \cdot \mathcal{R}(f_\nu; D_i), i \in [N]$. The weight q_i can be updated on the server using the uploaded $\mathcal{R}(f_\nu; D_i), i \in [N]$ following the update law in (Sagawa et al., 2019). Our algorithm is shown in Algorithm 1.

Algorithm 1 *PerInvFL*: Personalized Invariant Federated Learning

Input: $T, R, S, \beta, \eta, \gamma, \alpha$.

Initialize the models ν^0, $\{\theta^0_i | i \in [N]\}$.
for $t = 0$ to $T - 1$ **do**
  Server sends the global model (ν^t) to the participating clients.
  for local device $i = 1$ to N in parallel **do**
    Initialization: $\nu^t_{i,0} \leftarrow \nu^t$.
    for $r = 0$ to $R - 1$ **do**
      for $s = 0$ to $S - 1$ **do**
        Update the personalized model: $\theta^{s+1}_i \leftarrow \theta^s_i - \eta\big(\nabla \mathcal{L}_{INV}(\theta^s_i; \mathcal{E}^i_{train}) + 2\beta(\theta^s_i - \nu^t_{i,r})\big)$.
      Local update for global model: $\nu^t_{i,r+1} \leftarrow \nu^t_{i,r} - \gamma \nabla \mathcal{L}_{INV}(\nu^t_{i,r}; D_i)$.
  Global aggregation: $\nu^{t+1} \leftarrow \nu^t - \alpha\big(\nu^t - \frac{1}{N}\sum_{i \in [N]} \nu^t_{i,R}\big)$.
return the personalized models $\{\theta^S_i | i \in [N]\}$ and global model ν^T.

4.2 THEORETICAL ANALYSIS

In this section, we will provie detailed theoretical analysis on convergence rate and out-of-distribution generalization guarantee of the proposed algorithm, and discuss how to improve the practical performance of our method based on the theoretical outcomes. For simplicity, we define $\mathcal{L}(\nu) := \frac{1}{N}\mathcal{L}_{IRM}(\nu; \mathcal{E}^D)$ in this section. It's mentioned that the penalty term of ***IRM*** is generally non-convex in (Arjovsky et al., 2019). Therefore, we figure out the convergence rate of our algorithm under the non-convex and smooth case.

Assumption 2 (Bounded variance) The variance of local gradients to the aggregated average is upper bounded by $\frac{1}{N}\sum_{i=1}^{N} \|\nabla \mathcal{L}^i_{IRM}(\nu; D_i) - \nabla \mathcal{L}_{IRM}(\nu; \mathcal{E}^D)\|^2 \leq \delta_L^2$, where δ_L is a finite constant.

Theorem 2 ***(Convergence)*** *Suppose the IRM loss function $\mathcal{L}_{IRM}(\theta; e^i_j)$ is L_F-smooth, $\forall$ i, j, and Assumption 2 holds. If $\gamma \leq \frac{\gamma_0}{\alpha R}$ $\forall$ $\alpha \leq 1$, $R \geq 1$ and $\gamma_0 := \frac{\alpha}{\sqrt{32(R+3)L_F}}$. Then we have*

1. *The convergence rate of the global model is given by:*

$$\mathbb{E}\big[\|\nabla \mathcal{L}(\nu_{t^\star})\|^2\big] \leq \mathcal{O}\big(\mathbb{E}\big[\|\nabla \mathcal{L}(\nu_{t^\star})\|^2\big]\big) := \mathcal{O}\Big(\frac{\Delta_{\mathcal{L}} R^{\frac{1}{2}} L_F}{\beta T} + \frac{(\Delta_{\mathcal{L}} L_F)^{\frac{3}{4}} (R\delta_L^2)^{\frac{1}{4}}}{\alpha^{\frac{1}{2}} T^{\frac{3}{4}}} + \frac{(\Delta_{\mathcal{L}} \delta_L L_F)^{\frac{2}{3}}}{\alpha^{\frac{2}{3}} T^{\frac{2}{3}}}\Big).$$

2. *And, the convergence rate of the personalized models is given by:*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\big[\|\theta_i^S(\nu_{t^\star})-\nu_{t^\star}\|^2\big]\leq\mathcal{O}\Big(\mathbb{E}\big[\|\nabla\mathcal{L}(\nu_{t^\star})\|^2\big]\Big)+C,$$

where $\Delta_{\mathcal{L}}:=\mathcal{L}(\nu_0)-\mathcal{L}(\nu_T)$, C is a finite positive constant, and $t^\star$ is uniformly sampled from $\{0,1,...,T-1\}$. The constant C exists since the number of local iterations S is finite.

Corollary 1 Theorem 2 proves that the proposed algorithm achieves a convergence rate of $\mathcal{O}(1/T^{2/3})$ under the non-convex and smooth case. The detailed proof is provided in the Appendix.

Theorem 3 ***(Out-of-distribution Generalization)*** *As discussed in Section 3, the learning model can be written as $f=\Phi\cdot\omega$. Suppose Assumption 1 is satisfied. Let the global invariant featurizer $\Phi_g^\star\in\mathbb{R}^{d\times d}$ have rank $\tau>0$, and the personalized invariant featurizer $\Phi_i^\star$ have rank $\tau+q_i$ where $q_i>0$ $(i\in[N])$ and $d>\max_{i\in[N]}q_i+\tau$. Denote the minimizers of our designed objective by $\{f_{\hat{\theta}_i}=\hat{\Phi}_i\cdot\hat{\omega}_i\}$. Then, if there are at least $d-\tau+\frac{d}{\tau}$ local datasets in $\mathcal{E}^D$ lying in a linear general position of degree τ, the following holds on each client i $(i\in[N])$:*

1. *When there are at least $d-(\tau+q_i)+\frac{d}{\tau+q_i}$ training contexts in $\mathcal{E}_{train}^i$ lying in a linear general position of degree $\tau+q_i$, with β appropriately chosen, we can guarantee that $\mathbb{E}[Y|\hat{\Phi}_i(X),e]=\mathbb{E}[Y|\hat{\Phi}_i(X),e']$, for all $e,e'\in\mathcal{E}_{all}^i$, and the test error $\mathbb{E}_{(X,Y)\sim\mathbb{P}^e}[\ell(f_{\hat{\theta}_i}(X),Y)]=\min_{\omega,\Phi_i^\star}\mathbb{E}_{(X,Y)\sim\mathbb{P}^e}[\ell(\omega(\Phi_i^\star(X)),Y)]$ for any $e\in\mathcal{E}_{all}^i$.*

2. *Otherwise, with β appropriately chosen, we can guarantee that $\mathbb{E}[Y|\hat{\Phi}_i(X),e]=\mathbb{E}[Y|\hat{\Phi}_i(X),e']$, for all $e,e'\in\mathcal{E}_{all}^i$.*

Remark 1 Theorem 3 provides the personalized models with performance guarantee on any unseen contexts, even if both challenges of insufficient training contexts and limited data diversity in each available context exist. The detailed proof of Theorem 3 can be found in the Appendix. In particular, we find that choosing a appropriate value for the balancing weight β is significant for guaranteeing the out-of-distribution performance on each client. Empirically, the less training contexts a client has, the smaller the value of β should be. Therefore, personalized β can be adopted on heterogeneous clients by themselves, according to the number of available training contexts.

5 Experiments

5.1 Experimental Setup

Datasets We evaluate the effectiveness of our method on three datasets that are frequently adopted in the related literature: **1) Rotated CMNIST (RC-MNIST)**, a variant of CMNIST (Arjovsky et al., 2019). Firstly, We construct the CMNIST dataset according to the same law as in (Arjovsky et al., 2019): each image having label 0 is colored green with probability p^e and colored red with probability $1-p^e$. In contrary, each image having label 1 is colored red with probability p^e and colored green with probability $1-p^e$. In the training and testing contexts, we set $p^e=p_{train}^e$ and $p^e=p_{test}^e$ to generate the data respectively. On the basis of CMNIST, we add rotation property to simulate the personal writing style and regard the rotation property as useful personalized information. To distribute the data to the clients in a Non-IID scheme, we construct four training contexts with $p_{train}^e=0.95,0.90,0.85,0.80$ and assign each context to a client as training data. For the testing data, we constrcut a out-of-distribution context with $p_{test}^e=0.10$ for every client. The data used for constructing the four training/testing contexts is randomly sampled from the train-set/test-set of MNIST (LeCun et al., 1998) without replacement. In addition, all images (from both training and testing sets) in the four clients are rotated by 0°, 90°, 180° and 270° respectively. **2) Rotated Colored Fashion-MNIST (RC-FMNIST)** (Ahuja et al., 2020), which is constructed using the same strategy as RC-MNIST, except that the original images come from FashionMNIST dataset. **3) WaterBird** (Sagawa et al., 2019), which is constructed by placing the waterbird photographs onto the water background with probability p^e and onto water background with probability $1-p^e$. In the meanwhile, the landbird photographs are placed onto the land background with probability p^e and

onto water background with probability $1 - p^e$. In this paper, we construct four training contexts with $p^e_{train} = 0.95, 0.90, 0.85, 0.80$ and the testing contexts with $p^e_{test} = 0.10$. To distribute the data in a Non-IID manner, we construct the four training contexts using the bird photograghs of disjoint bird classes from the bird dataset and background photographs randomly selected from the background dataset without replacement. Each training context is assigned to a client as training data. For the testing data in each client, we construct a testing contexts with probability $p^e_{test} = 0.10$, using the bird photographs of same bird classes (but distinct instances) as the corresponding training context and the background photographs are randomly selected from the background dataset.

Model selection and Competitors For RC-MNIST and RC-FMNIST dataset, we adopt a deep neural network with two hidden layers as the feature extractor and an subsequent fully-connected layer as the classifier. In the evaluation on WaterBird dataset, we adopt the ResNet-18 (He et al., 2016) as learning model. Similarly, the part before the last fully-connected layer works as the feature extractor and the last fully-connected layer works as the classifier. We compare our method (***PerInvFL***) with six state-of-the-art algorithms in federated learning and invariant learning, including one traditional federated learning method (***FedAvg*** (McMahan et al., 2017)), three personalized learning methods (***pFedMe*** (T Dinh et al., 2020), ***Ditto*** (Li et al., 2021), and ***FTFA*** (Cheng et al., 2021)), and two invariant learning methods (***IRM*** (Arjovsky et al., 2019) and ***GroupDRO*** (Sagawa et al., 2019)).

Table 2: Results on Rotated-Colored MNIST dataset

p^e_{test}	OOD case 1 0.10	OOD case 2 0.20	OOD case 3 0.30	OOD case 4 0.40	OOD case 5 0.50	Average
FedAvg	20.77(±0.0)	29.66(±0.0)	37.74(±0.0)	45.15(±0.0)	54.12(±0.0)	37.49(±11.6)
Ditto	23.62(±0.0)	31.85(±0.0)	39.57(±0.2)	46.40(±0.0)	**54.92(±0.0)**	39.27(±10.9)
FTFA	19.86(±0.1)	28.87(±0.1)	37.37(±0.0)	43.75(±0.1)	53.69(±0.0)	36.71(±11.7)
pFedMe	30.28(±0.0)	38.98(±0.0)	40.05(±0.0)	44.48(±0.0)	49.67(±0.0)	40.69(±6.43)
GroupDRO	29.32(±0.1)	35.68(±0.0)	41.50(±0.0)	46.93(±0.0)	53.77(±0.2)	41.44(±8.51)
IRM	51.48(±0.0)	51.70(±0.1)	52.17(±0.1)	**51.53(±0.0)**	51.18(±0.0)	51.61(±0.33)
PerInvFL	**53.46(±0.2)**	**52.75(±0.1)**	**53.63(±0.1)**	51.41(±0.3)	52.73(±0.1)	**52.80(±0.78)**

Table 3: Results on Rotated Colored-FMNIST dataset

p^e_{test}	OOD case 1 0.10	OOD case 2 0.20	OOD case 3 0.30	OOD case 4 0.40	OOD case 5 0.50	Average
FedAvg	10.70(±0.0)	20.80(±0.0)	29.85(±0.0)	41.84(±0.0)	50.54(±0.0)	30.75(±14.3)
Ditto	16.35(±0.0)	25.75(±0.1)	34.11(±0.0)	45.09(±0.0)	52.59(±0.0)	34.78(±13.0)
FTFA	17.32(±0.1)	26.48(±0.0)	34.36(±0.0)	45.43(±0.0)	53.03(±0.0)	35.32(±12.8)
pFedMe	20.32(±0.0)	27.40(±0.1)	34.40(±0.0)	42.85(±0.0)	48.63(±0.0)	34.72(±10.2)
GroupDRO	30.13(±0.0)	36.00(±0.0)	41.33(±0.0)	47.55(±0.1)	52.59(±0.0)	41.52(±7.99)
IRM	47.35(±0.1)	48.99(±0.1)	50.53(±0.1)	51.57(±0.0)	52.66(±0.1)	50.22(±1.88)
PerInvFL	**51.71(±0.2)**	**52.10(±0.1)**	**50.77(±0.2)**	**53.12(±0.1)**	**53.64(±0.1)**	**52.27(±1.02)**

Table 4: Results on WaterBird dataset.

p^e_{test}	OOD case 1 0.10	OOD case 2 0.20	OOD case 3 0.30	OOD case 4 0.40	OOD case 5 0.50	Average
FedAvg	51.33(±0.77)	60.08(±0.47)	57.75(±0.21)	62.50(±0.61)	59.25(±1.62)	58.18(±3.76)
Ditto	51.17(±1.31)	60.58(±0.68)	58.74(±0.53)	63.50(±0.74)	60.17(±0.66)	58.83(±4.13)
FTFA	51.67(±0.69)	61.50(±0.71)	59.00(±0.43)	63.00(±0.83)	60.25(±0.94)	59.08(±3.94)
pFedMe	52.17(±0.12)	56.42(±0.12)	58.75(±0.01)	65.00(±0.03)	62.25(±0.00)	58.92(±4.47)
GroupDRO	61.33(±0.31)	63.25(±0.35)	65.75(±1.03)	69.05(±1.08)	62.58(±0.31)	64.39(±2.74)
IRM	62.75(±0.41)	62.75(±0.00)	65.25(±0.23)	66.42(±0.12)	63.42(±0.59)	64.12(±1.47)
PerInvFL	**71.17(±0.44)**	**74.15(±0.37)**	**73.67(±0.21)**	**75.13(±0.11)**	**75.83(±0.34)**	**73.99(±1.60)**

5.2 EXPERIMENTAL RESULTS

In order to comprehensively compare the performance of our method with that of the baselines, we conduct evaluation on five out-of-distribution (OOD) testing cases for each dataset, including $p^e_{test} = 0.10, 0.20, 0.30, 0.40$ and 0.50. All hyperparameters of the algorithms are tuned optimal, and more details are provided in the Appendix. We run each algorithm three times with different random seeds to record the mean and standard deviation of the test accuracy in every OOD case.

Performance Comparison. The overall results on RC-MNIST, RC-FMNIST and WaterBird are displayed in Table 2, Table 3 and Table 4, respectively. In each table, the last column gives the mean and standard deviation of the results in the five OOD cases. From the results shown in the three tables, we can find that 1) our method *PerInvFL* can constantly outperform the baseline methods on both worst-case accuracy and average accuracy for diverse datasets. In particular, our method achieves an average accuracy of 73.99% which is about $\mathbf{9}\%$ **higher** than the second highest one on WaterBird dataset. 2) As shown in the last columns, the performance deviation of our method is much smaller than the existing federated learning methods and approximate to the invariant learning approach *IRM*, demonstrating that our method can effectively obtain the invariant predictor and hence achieve consistent performance in different out-of-distribution cases. These findings verify that the proposed method can effecrively extract the invarinat features and concurrently exploit the personalized information to improve the performance on out-of-distribution testing data.

Table 5: The effect of the elaborated dual-regularization

	Rotated-CMNIST					WaterBird				
p^e_{test}	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50
IRM	51.48	51.70	52.17	**51.53**	51.18	62.75	62.75	65.25	66.42	63.42
IRM-$L2$	31.19	29.35	36.53	43.06	51.83	56.83	62.75	58.75	66.00	59.42
IRM-FT	21.43	22.37	32.42	40.32	50.96	55.67	62.50	65.05	66.00	63.75
PerInvFL	**53.46**	**52.75**	**53.63**	51.41	**52.73**	**71.17**	**74.15**	**73.67**	**75.13**	**75.83**

Effect of dual-regularized optimization. As explained in section 4, the elaborated dual-regularized optimization is formulated as the objective (2). To demonstrate the effect of the designed dual-regularized franework, we replace the dual-regularized optimization with some prevalent single-regularized methods that are usually adopted in the state-of-the-art personalized federated learning, and keep the rest part of the overall objective (shown in equation (2) and (3)) unchanged. Concretely, we implement two typical personalization skills with the global model being achieved by distributional *IRM* that we introduce in section 4.1. One is the $L2$-norm regularizer which is widely used in PFL (T Dinh et al., 2020; Hanzely & Richtárik, 2020; Hanzely et al., 2020; Li et al., 2021), and we call this implementation *IRM*-$L2$. Another one is the local-finetune skill which is proved simple and effective in (Cheng et al., 2021). In particular, the local-finetune skill can be viewed as a 'none' regularizer. We name the implementation with local-finetune skill as *IRM-FT*.

The comparison of results is shown in Table 5. We can find that combining the existing personalization terms with *IRM* cannot improve the performance of OOD problem under federated setting. On the contrary, the widely adopted personalization term ($L2$-norm and local-finetune) can even degrade the OOD performance, compared with the distributional version of *IRM*. The underlying reason is that the local objective with these personalization terms is to minimize the expected loss as long as the target model is not too different from the global model. However, when the realtionship with the global model is satisfied, minimizing the expected loss can readily make the trained model pick up the spurious correlations and hence decrease the performance on OOD testing data. By comparison, the designed dual-regularized optimization in our algorithm contains two constraints: 1) being close/similar to the 'global invariance' and 2) satisfying the 'local invariance' to exclude the spurious features. Therefore, our method can effectively exploit the personalized information and avoid 'picking up' the spurious information concurrently, so as to improve the OOD performance.

6 Conclusion

In this paper, we are the first to investigate the out-of-distribution problem under the Non-IID federated setting. Atfer formally analyzing the challenges, we propose the novel concept *personalized invariance* which can improve the model performance, by preserving the important personalized information and meanwhile eliminating the spurious correlations. To explore the optimal personalized invariance, we propose a objective: dual-regularization constrained optimization and design a practical algorithm *PerInvFL* to solve it. Both the theoretical and experimental results demonstrate the superiority of our method over the state-of-the-art federated learning and invariant learning methods.

REFERENCES

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

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

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

Duc Bui, Kshitiz Malik, Jack Goetz, Honglei Liu, Seungwhan Moon, Anuj Kumar, and Kang G Shin. Federated user representation learning. *arXiv preprint arXiv:1909.12535*, 2019.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In *International Conference on Machine Learning*, pp. 1448–1458. PMLR, 2020.

Gary Cheng, Karan Chadha, and John Duchi. Fine-tuning is fine in federated learning. *arXiv preprint arXiv:2108.07313*, 2021.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139, pp. 2089–2099, 2021.

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

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

Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees GM Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. In *European Conference on Computer Vision*, pp. 200–216. Springer, 2020.

Moming Duan, Duo Liu, Xianzhang Chen, Renping Liu, Yujuan Tan, and Liang Liang. Self-balancing federated learning with global imbalanced data in mobile systems. *IEEE Transactions on Parallel and Distributed Systems*, 32(1):59–71, 2020.

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

Juan L Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. *Advances in Neural Information Processing Systems*, 33:15464–15475, 2020.

Vikas Garg, Adam Tauman Kalai, Katrina Ligett, and Steven Wu. Learn to expect the unexpected: Probably approximately correct domain generalization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3574–3582. PMLR, 2021.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In *Proceedings of Advances in Neural Information Processing Systems, NeurIPS*, 2020.

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

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

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

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

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

Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*, 2018.

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

Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pp. 4069–4077. PMLR, 2021.

Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. 2020.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

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

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

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

Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Integrated latent heterogeneity and invariance learning in kernel space. *Advances in Neural Information Processing Systems*, 34, 2021b.

Jiashuo Liu, Zheyan Shen, Peng Cui, Linjun Zhou, Kun Kuang, Bo Li, and Yishi Lin. Stable adversarial learning under distributional shifts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8662–8670, 2021c.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Lingjuan Lyu, Jiangshan Yu, Karthik Nandakumar, Yitong Li, Xingjun Ma, Jiong Jin, Han Yu, and Kee Siong Ng. Towards fair and privacy-preserving federated deep models. *IEEE Transactions on Parallel and Distributed Systems*, 31(11):2524–2541, 2020.

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

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

Michael Oberst, Nikolaj Thams, Jonas Peters, and David Sontag. Regularizing towards causal invariance: Linear models with proxies. In *International Conference on Machine Learning*, pp. 8260–8270. PMLR, 2021.

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

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

Johannes Schneider and Michail Vlachos. Personalization of deep learning. *arXiv preprint arXiv:1909.02803*, 2019.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Disentangled generative causal representation learning. *arXiv preprint arXiv:2010.02637*, 2020a.

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li, and Peixuan Chen. Causally regularized learning with agnostic data selection bias. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 411–419, 2018.

Zheyan Shen, Peng Cui, Jiashuo Liu, Tong Zhang, Bo Li, and Zhitang Chen. Stable learning via differentiated variable decorrelation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2185–2193, 2020b.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S. Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems, NeurIPS*, 2017.

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

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *arXiv preprint arXiv:2103.00710*, 2021.

Xueyang Tang, Song Guo, and Jingcai Guo. Personalized federated learning with contextualized generalization. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 2241–2247, 2022. Main Track.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In *International Conference on Machine Learning*, pp. 10401–10412. PMLR, 2021.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Hao Wang, Zakhary Kaplan, Di Niu, and Baochun Li. Optimizing federated learning on non-iid data with reinforcement learning. In *INFOCOM*, pp. 1698–1707. IEEE, 2020.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Wenjun Zeng, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. *arXiv preprint arXiv:2103.03097*, 2021.

Kangkang Wang, Rajiv Mathews, Chloé Kiddon, Hubert Eichner, Françoise Beaufays, and Daniel Ramage. Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252*, 2019.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021.

Miao Yang, Akitanoshou Wong, Hongbin Zhu, Haifeng Wang, and Hua Qian. Federated learning with class imbalance reduction. *arXiv preprint arXiv:2011.11266*, 2020.

Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang. Towards a theoretical framework of out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

Jie Zhang, Song Guo, Xiaosong Ma, Haozhao Wang, Wenchao Xu, and Feijie Wu. Parameterized knowledge transfer for personalized federated learning. *Advances in Neural Information Processing Systems*, 34, 2021a.

Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M. Alvarez. Personalized federated learning with first order model optimization. In *Proceedings of the 9th International Conference on Learning Representations, ICLR 2021*, 2021b.

Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5372–5382, 2021c.

Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. *Advances in Neural Information Processing Systems*, 33:16096–16107, 2020.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

A Appendix

In this appendix, we provide more details that are simplified in the main text duo to page limitation, including the complemte proofs, more details on experimental setup, and some further discussions.

A.1 Proof of Theorem 1

Theorem 1 If the Assumption 1 holds in FL, the proposed *personalized invariant* representations can be constantly more informative than the global invariant representations obtained by the distributed implementation of existing invariant learning, for the prediction performance. That is,

$$\frac{1}{N}\sum_{i=1}^{N} I(Y;\Phi_i^\star) \geq \frac{1}{N}\sum_{i=1}^{N} I(Y;\Phi_g^\star) + p\delta, \tag{6}$$

where $0 < p < 1$ is a constant and δ is a positive constant that is independent of N.

Proof: According to the definition, there are N clients in the federated learning system. The set of all possible environments on each client i is given by $\mathcal{E}_{all}^i$, $i \in [N]$. Then, the set of all possible environments on all the clients is the union:

$$\mathcal{E}_{all} = \bigcup_{i\in[N]} \mathcal{E}_{all}^i \tag{7}$$

Because of the data heterogeneity across the clients, we have $\mathcal{E}_{all}^i \neq \mathcal{E}_{all}$, i.e., $\mathcal{E}_{all}^i \subset \mathcal{E}_{all}, \forall i \in [N]$. For every invariant feature $\phi_g \in \Phi_g$ across all the environments $\mathcal{E}_{all}$, we can get

$$\phi_g \in \Phi_i, \forall i \in [N]. \tag{8}$$

But, the converse is not true, due to the result in equation 7. Therefore, we have

$$\Phi_g \subseteq \Phi_i, \forall i \in [N] \tag{9}$$

When the **Heterogeneity** condition in Assumption 1 is satisfied, we can conclude that

$$\begin{cases} \Phi_g \subset \Phi_j, \forall j \in [N_K], \\ \Phi_g = \Phi_j, \forall j \notin [N_K] \text{ and } j \in [N]. \end{cases} \tag{10}$$

When the **Informativeness** condition in Assumption 1 is satisfied, we can derive the average mutual-information among the clients by

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} I(Y;\Phi_i^\star) &= \frac{1}{N}\sum_{i=1}^{N}\max_{S_i\subseteq\Phi_i} I(Y;S_i) \\
&= \frac{1}{N}\sum_{i\in[N_K]}\max_{S_i\subseteq\Phi_i} I(Y;S_i) + \frac{1}{N}\sum_{i\notin[N_K]}\max_{S_i\subseteq\Phi_i} I(Y;S_i) \\
&\geq \frac{1}{N}\sum_{i\in[N_K]}\big\{\max_{S_{i,1}\subseteq\Phi_g} I(Y;S_{i,1}) + \max_{S_{i,2}\subseteq Z_i} I(Y;S_{i,2})\big\} + \frac{1}{N}\sum_{i\notin[N_K]}\max_{S_i\subseteq\Phi_i} I(Y;S_i) \\
&\geq \frac{1}{N}\sum_{i\in[N_K]}\max_{S_{i,1}\subseteq\Phi_g} I(Y;S_{i,1}) + \frac{K}{N}\delta + \frac{1}{N}\sum_{i\notin[N_K]}\max_{S_i\subseteq\Phi_g} I(Y;S_i) \\
&= \frac{1}{N}\sum_{i\in[N_K]} I(Y;\Phi_g^\star) + \frac{K}{N}\delta + \frac{1}{N}\sum_{i\notin[N_K]} I(Y;\Phi_g^\star) \\
&= \frac{1}{N}\sum_{i=1}^{N} I(Y;\Phi_g^\star) + p\delta.
\end{aligned}$$

Proof ends.

As discussed in the main content, the Theorem 1 demonstrates the superiority of personalized invariance on improving the average prediction performance in OOD and Non-IID federated setting, compared with the global invariance which is elicited by the distributional implementation of existing invariant learning methods. In paticular, the improvement $p\delta$ is independent of the number of clients and p is the percentage of the clients with heterogeneous invariance. In other words, p can represent the extent of data heterogeneity across the clients. Therefore, we can conclude that the performance improvement caused by the proposed personalized invariance can be enlarged as the herteroheneity of invariance among the clients increases.

A.2 Proof of Theorem 2

Theorem 2 Suppose the IRM loss function $\mathcal{L}_{IRM}(\theta; e_j^i)$ is L_F-smooth, $\forall\ i,\ j$, and Assumption 2 holds. If $\gamma \leq \frac{\gamma_0}{\alpha R}\ \forall\ \alpha \leq 1$, $R \geq 1$ and $\gamma_0 := \frac{\alpha}{\sqrt{32(R+3)}L_F}$. Then we have

1. The convergence rate of the global model is given by:

$$\mathbb{E}\big[\|\nabla\mathcal{L}(\nu_{t^\star})\|^2\big] \leq \mathcal{O}\big(\mathbb{E}\big[\|\nabla\mathcal{L}(\nu_{t^\star})\|^2\big]\big) := \mathcal{O}\Big(\frac{\Delta_{\mathcal{L}} R^{\frac{1}{2}} L_F}{\beta T} + \frac{(\Delta_{\mathcal{L}} L_F)^{\frac{3}{4}}(R\delta_L^2)^{\frac{1}{4}}}{\alpha^{\frac{1}{2}} T^{\frac{3}{4}}} + \frac{(\Delta_{\mathcal{L}}\delta_L L_F)^{\frac{2}{3}}}{\alpha^{\frac{2}{3}} T^{\frac{2}{3}}}\Big).$$

2. And, the convergence rate of the personalized models is given by:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\big[\|\theta_i^S(\nu_{t^\star}) - \nu_{t^\star}\|^2\big] \leq \mathcal{O}\Big(\mathbb{E}\big[\|\nabla\mathcal{L}(\nu_{t^\star})\|^2\big]\Big) + C,$$

where $\Delta_{\mathcal{L}} := \mathcal{L}(\nu_0) - \mathcal{L}(\nu_T)$, C is a finite positive constant, and $t^\star$ is uniformly sampled from $\{0, 1, ..., T-1\}$. The constant C exists since the number of local iterations S is finite.

Proof: We know the local update of global model follows

$$\nu_{i,r+1}^t = \nu_{i,r}^t - \gamma \underbrace{\nabla\mathcal{L}_{IRM}(\nu_{i,r}^t)}_{=:g_{i,r}^t} \tag{11}$$

and the global update is

$$\begin{aligned}\nu_{t+1} &= \nu_t - \alpha(\nu_t - \frac{1}{N}\sum_{i=1}^{N}\nu_{i,R}^t) \\ &= \nu_t - \frac{\alpha}{N}\sum_{i=1}^{N}(\nu_t - \nu_{i,R}^t) \\ &= \nu_t - \underbrace{\alpha\gamma R}_{=:\hat{\gamma}}\underbrace{\frac{1}{NR}\sum_{i=1}^{N}\sum_{r=0}^{R-1} g_{i,r}^t}_{=:g_t},\end{aligned}$$

Since the loss function is L-smooth, we have

$$\begin{aligned}&\mathbb{E}[\mathcal{L}(\nu_{t+1}) - \mathcal{L}(\nu_t)] \\ &\leq \mathbb{E}[\langle\nabla\mathcal{L}(\nu_t), \nu_{t+1} - \nu_t\rangle] + \frac{L_F}{2}\mathbb{E}[\|\nu_{t+1} - \nu_t\|^2] \\ &= -\hat{\gamma}\mathbb{E}[\langle\nabla\mathcal{L}(\nu_t), g_t\rangle] + \frac{\hat{\gamma}^2 L_F}{2}\mathbb{E}[\|g_t\|^2] \\ &= -\hat{\gamma}\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] - \hat{\gamma}\mathbb{E}[\langle\nabla\mathcal{L}(\nu_t), g_t - \nabla\mathcal{L}(\nu_t)\rangle] + \frac{\hat{\gamma}^2 L_F}{2}\mathbb{E}[\|g_t\|^2] \\ &\leq -\hat{\gamma}\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] + \frac{\hat{\gamma}}{2}\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] + \frac{\hat{\gamma}}{2}\mathbb{E}[\|\frac{1}{NR}\sum_{i,r}^{N,R} g_{i,r}^t - \nabla\mathcal{L}(\nu_t)\|^2] + \frac{\hat{\gamma}^2 L_F}{2}\mathbb{E}[\|g_t\|^2] \\ &= -\frac{\hat{\gamma}}{2}\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] + \frac{\hat{\gamma}}{2}\mathbb{E}[\|\frac{1}{NR}\sum_{i,r}^{N,R} g_{i,r}^t - \nabla\mathcal{L}^i(\nu_t)\|^2] + \frac{\hat{\gamma}^2 L_F}{2}\mathbb{E}[\|g_t\|^2] \\ &= -\frac{\hat{\gamma}}{2}\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] + \frac{\hat{\gamma}}{2}\mathbb{E}[\|\frac{1}{NR}\sum_{i,r}^{N,R} g_{i,r}^t - \nabla\mathcal{L}^i(\nu_t)\|^2] \\ &\quad + \frac{\hat{\gamma}^2 L_F}{2}\mathbb{E}[\|\frac{1}{NR}\sum_{i,r}^{N,R} g_{i,r}^t - \nabla\mathcal{L}^i(\nu_t) + \frac{1}{N}\sum_{i=1}^{N}\nabla\mathcal{L}^i(\nu_t)\|^2] \\ &\leq -\frac{\hat{\gamma}}{2}\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] + \frac{\hat{\gamma}}{2}\mathbb{E}[\|\frac{1}{NR}\sum_{i,r}^{N,R} g_{i,r}^t - \nabla\mathcal{L}^i(\nu_t)\|^2] \\ &\quad + \frac{\hat{\gamma}^2 L_F(1+R)}{2}\mathbb{E}[\|\frac{1}{NR}\sum_{i,r}^{N,R} g_{i,r}^t - \nabla\mathcal{L}^i(\nu_t)\|^2] + \frac{\hat{\gamma}^2 L_F(1+R)}{2R}\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] \\ &\leq -\frac{\hat{\gamma}}{2}(1 - \frac{(R+1)\hat{\gamma}L_F}{R})\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] + \frac{\hat{\gamma}[1 + (1+R)\hat{\gamma}L_F]}{2}\frac{1}{NR}\sum_{i,r}^{N,R}\mathbb{E}[\|g_{i,r}^t - \nabla\mathcal{L}^i(\nu_t)\|^2]\end{aligned}$$

Because

$$\mathbb{E}[\|g_{i,r}^t - \nabla\mathcal{L}^i(\nu_t)\|^2] = \mathbb{E}[\|\nabla\mathcal{L}^i(\nu_{i,r}^t) - \nabla\mathcal{L}^i(\nu_t)\|^2] \leq L_F^2\mathbb{E}[\|\nu_{i,r}^t - \nu_t\|^2], \tag{12}$$

we first deal with $\mathbb{E}[\|\nu_{i,r}^t - \nu_t\|^2]$. We can figure out that

$$
\begin{aligned}
&\mathbb{E}[\|\nu_{i,r}^t - \nu_t\|^2] \\
&= \mathbb{E}[\|\nu_{i,r-1}^t - \nu_t - \gamma\nabla\mathcal{L}^i(\nu_t) + \gamma\nabla\mathcal{L}^i(\nu_t) - \gamma g_{i,r-1}^t\|^2] \\
&\leq (1+\frac{1}{R})\mathbb{E}[\|\nu_{i,r-1}^t - \nu_t - \gamma\nabla\mathcal{L}^i(\nu_t)\|^2] + (1+R)\gamma^2\mathbb{E}[\|g_{i,r-1}^t - \nabla\mathcal{L}^i(\nu_t)\|^2] \\
&\leq (1+\frac{1}{R})(1+\frac{1}{2R})\mathbb{E}[\|\nu_{i,r-1}^t - \nu_t\|^2] + (1+\frac{1}{R})(1+2R)\gamma^2\mathbb{E}[\|\nabla\mathcal{L}^i(\nu_t)\|^2] \\
&\quad + (1+R)\gamma^2 L_F^2\mathbb{E}[\|\nu_{i,r-1}^t - \nu_t\|^2] \\
&= (1+\frac{1}{R})(1+\frac{1}{2R}+R\gamma^2L_F^2)\mathbb{E}[\|\nu_{i,r-1}^t - \nu_t\|^2] + (1+\frac{1}{R})(1+2R)\gamma^2\mathbb{E}[\|\nabla\mathcal{L}^i(\nu_t)\|^2]
\end{aligned}
$$

Suppose $\gamma^2 < \frac{1}{2R^2L_F^2}$, then we can get

$$
\begin{aligned}
&\mathbb{E}[\|g_{i,r}^t - \nabla\mathcal{L}^i(\nu_t)\|^2] \\
&\leq L_F^2\mathbb{E}[\|\nu_{i,r}^t - \nu_t\|^2] \\
&\leq R(R+1)\{[(1+\frac{1}{R})^2]^r - 1\}\gamma^2L_F^2\mathbb{E}[\|\nabla\mathcal{L}^i(\nu_t)\|^2] \\
&\leq R(R+1)[(1+\frac{1}{R})^2]^r\gamma^2L_F^2\mathbb{E}[\|\nabla\mathcal{L}^i(\nu_t)\|^2]
\end{aligned}
$$

Thus, with Assumption 2, we can get

$$
\begin{aligned}
&\frac{1}{NR}\sum_{i,r}^{N,R}\mathbb{E}[\|g_{i,r}^t - \nabla\mathcal{L}^i(\nu_t)\|^2] \\
&\leq \frac{1}{NR}\sum_{i,r}^{N,R}R(R+1)\gamma^2L_F^2\mathbb{E}[\|\nabla\mathcal{L}^i(\nu_t)\|^2][(1+\frac{1}{R})^2]^r \\
&\leq \frac{1}{N}\sum_{i=1}^{N}(R+1)\gamma^2L_F^2\frac{(e^2-1)}{2R+1}\mathbb{E}[\|\nabla\mathcal{L}^i(\nu_t)\|^2] \\
&\leq 8R^2\gamma^2L_F^2\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla\mathcal{L}^i(\nu_t)\|^2] \\
&= 8R^2\gamma^2L_F^2\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla\mathcal{L}^i(\nu_t) - \nabla\mathcal{L}(\nu_t) + \nabla\mathcal{L}(\nu_t)\|^2] \\
&\leq 8R^2\gamma^2L_F^2\frac{1}{N}\sum_{i=1}^{N}\left\{2\mathbb{E}[\|\nabla\mathcal{L}^i(\nu_t) - \nabla\mathcal{L}(\nu_t)\|^2] + 2\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2]\right\} \\
&\leq 16R^2\gamma^2L_F^2\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla\mathcal{L}^i(\nu_t) - \nabla\mathcal{L}(\nu_t)\|^2] + 16R^2\gamma^2L_F^2\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] \\
&\leq 16R^2\gamma^2L_F^2\delta_L^2 + 16R^2\gamma^2L_F^2\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2]
\end{aligned}
$$

Therefore, we can write

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}(\nu_{t+1}) - \mathcal{L}(\nu_t)] \\
&\leq -\frac{\hat{\gamma}}{2}\big(1 - \frac{(1+R)\hat{\gamma}L_F}{R}\big)\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] \\
&\quad + \frac{\hat{\gamma}[1+(1+R)\hat{\gamma}L_F]}{2}\big\{16R^2\gamma^2L_F^2\delta_L^2 + 16R^2\gamma^2L_F^2\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2]\big\} \\
&= -\frac{\hat{\gamma}}{2}\Big\{1 - \frac{(R+1)\hat{\gamma}L_F}{R} - \frac{16[1+(1+R)\hat{\gamma}L_F]\hat{\gamma}^2L_F^2}{\alpha^2}\Big\}\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] \\
&\quad + \frac{8\hat{\gamma}[1+(1+R)\hat{\gamma}L_F]\hat{\gamma}^2L_F^2\delta_L^2}{\alpha^2}
\end{aligned}$$

Let $\gamma \leq \frac{\gamma_0}{\alpha R}\ \forall\, \alpha \leq 1, R \geq 1$ and $\gamma_0 := \frac{\alpha}{\sqrt{32(R+3)L_F}}$, then $1 - \frac{(R+1)\hat{\gamma}L_F}{R} - \frac{16[1+(1+R)\hat{\gamma}L_F]\hat{\gamma}^2L_F^2}{\alpha^2} > \frac{1}{2}$. Furthermore, we can get

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}(\nu_{t+1}) - \mathcal{L}(\nu_t)] \\
&\leq -\frac{\hat{\gamma}}{4}\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] + \frac{8\hat{\gamma}^3L_F^2\delta_L^2}{\alpha^2} + \frac{8(1+R)\hat{\gamma}^4L_F^3\delta_L^2}{\alpha^2}
\end{aligned}$$

That is

$$\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] \leq \frac{4\big\{\mathbb{E}[\mathcal{L}(\nu_t) - \mathcal{L}(\nu_{t+1})]\big\}}{\hat{\gamma}} + \frac{32\hat{\gamma}^2L_F^2\delta_L^2}{\alpha^2} + \frac{32(1+R)\hat{\gamma}^3L_F^3\delta_L^2}{\alpha^2}.$$

We can get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] \leq \frac{4\sum_{t=0}^{T-1}\mathbb{E}[\mathcal{L}(\nu_t) - \mathcal{L}(\nu_{t+1})]}{\hat{\gamma}} + \frac{32\hat{\gamma}^2L_F^2\delta_L^2}{\alpha^2} + \frac{32(1+R)\hat{\gamma}^3L_F^3\delta_L^2}{\alpha^2}. \\
&= \frac{C_1}{\hat{\gamma}T} + \frac{C_2\hat{\gamma}^2}{\alpha^2} + \frac{C_3\hat{\gamma}^3}{\alpha^2},
\end{aligned}$$

where $C_1 = 4(\mathcal{L}(\nu_0) - \mathcal{L}(\nu_T))$, $C_2 = 32\delta_L^2L_F^2$ and $C_3 = 32(1+R)\delta_L^2L_F^3$.

We consider the following two cases:

- When $\gamma_0 \geq \min\Big\{\Big(\frac{C_1\alpha^2}{C_3T}\Big)^{\frac{1}{4}}, \Big(\frac{C_1\alpha^2}{C_2T}\Big)^{\frac{1}{3}}\Big\}$, we choose $\hat{\gamma} = \min\Big\{\Big(\frac{C_1\alpha^2}{C_3T}\Big)^{\frac{1}{4}}, \Big(\frac{C_1\alpha^2}{C_2T}\Big)^{\frac{1}{3}}\Big\}$, we have
$$\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] \leq \frac{C_1^{\frac{3}{4}}C_3^{\frac{1}{4}}}{\alpha^{\frac{1}{2}}T^{\frac{3}{4}}} + \frac{C_1^{\frac{2}{3}}C_2^{\frac{1}{3}}}{\alpha^{\frac{2}{3}}T^{\frac{2}{3}}}.$$
- When $\gamma_0 \leq \min\Big\{\Big(\frac{C_1\alpha^2}{C_3T}\Big)^{\frac{1}{4}}, \Big(\frac{C_1\alpha^2}{C_2T}\Big)^{\frac{1}{3}}\Big\}$, we can choose $\hat{\gamma} = \gamma_0$. We can get
$$\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] \leq \frac{C_1}{\gamma_0T} + \frac{C_1^{\frac{3}{4}}C_3^{\frac{1}{4}}}{\alpha^{\frac{1}{2}}T^{\frac{3}{4}}} + \frac{C_1^{\frac{2}{3}}C_2^{\frac{1}{3}}}{\alpha^{\frac{2}{3}}T^{\frac{2}{3}}}.$$

Therefore, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\mathcal{L}(\nu_t)\|^2] \leq \mathcal{O}\Big(\frac{C_1}{\gamma_0T} + \frac{C_1^{\frac{3}{4}}C_3^{\frac{1}{4}}}{\alpha^{\frac{1}{2}}T^{\frac{3}{4}}} + \frac{C_1^{\frac{2}{3}}C_2^{\frac{1}{3}}}{\alpha^{\frac{2}{3}}T^{\frac{2}{3}}}\Big). \tag{13}$$

As regard to the convergence rate of the personalized models, the personalized models are locally updated via

$$\theta_i^\star(\nu) \in \arg\min_\theta \mathcal{L}_{IRM}(\theta; \mathcal{E}_{train}^i) + \beta\|\theta - \nu\|^2. \tag{14}$$

When $\mathcal{L}_{IRM}(\theta; \mathcal{E}^i_{train})$ is L_F-smooth, after S local iterations, we have (T Dinh et al., 2020)

$$\mathbb{E}[\|\theta_i^S(\nu) - \theta_i^\star(\nu)\|^2] \leq \delta^2,$$

where δ is a finite constant. We can easily get from the proof of Theorem 2 in (T Dinh et al., 2020) that

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\big[\|\theta_i^S(\nu_{t^\star}) - \nu_{t^\star}\|^2\big] \leq \mathcal{O}\Big(\mathbb{E}\big[\|\nabla\mathcal{L}(\nu_{t^\star})\|^2\big]\Big) + C, \tag{15}$$

where C is a positive constant and it exists since the number of local iterations S is finite.

Proof ends.

A.3 Proof of Theorem 3

Adopting the similar assumptions in (Arjovsky et al., 2019), we first define a structural equation model (SEM) for each client to interpret the generating process of local data. On client $i (i \in [N])$, the data is generated according to:

$$\begin{aligned} &Y_i^e = H_i^e \cdot \rho_i + \epsilon_i^e, \qquad H_i^e \perp \epsilon_i^e, \qquad \mathbb{E}[\epsilon_i^e] = 0, \\ &X_i^e = G_i(H_i^e, Z_i^e) \end{aligned}$$

where $\rho \in \mathbb{R}^c$. The random variables X_i^e, H_i^e and Z_i^e take values in $\mathbb{R}^d, \mathbb{R}^c$ and $\mathbb{R}^l$, respectively. Besides, the H component of G_i is invertible. That is, there exists $\widetilde{G}_i$ satisfying that $\widetilde{G}_i(G_i(h, z)) = h$, for all $h \in \mathbb{R}^c$, $z \in \mathbb{R}^l$. Note that the above assumptions about linearity, statistical independence between the causal variables H_i^e and the noise ϵ_i^e, and zero-mean noise are also required in ***IRM*** (Arjovsky et al., 2019).

Theorem 3 As discussed in Section 3, the learning model can be written as $f = \Phi \cdot \omega$. Suppose Assumption 1 is satisfied. Let the global invariant featurizer $\Phi_g^\star \in \mathbb{R}^{d\times d}$ have rank $\tau > 0$, and the personalized invariant featurizer $\Phi_i^\star$ have rank $\tau + q_i$ where $q_i > 0$ $(i \in [N])$ and $d > \max_{i\in[N]} q_i + \tau$. Denote the minimizers of our designed objective by $\{f_{\hat{\theta}_i} = \hat{\Phi}_i \cdot \hat{\omega}_i\}$. Then, if there are at least $d - \tau + \frac{d}{\tau}$ local datasets in $\mathcal{E}^D$ lying in a linear general position of degree τ, the following holds on each client i $(i \in [N])$:

1. When there are at least $d - (\tau + q_i) + \frac{d}{\tau+q_i}$ training contexts in $\mathcal{E}^i_{train}$ lying in a linear general position of degree $\tau + q_i$, with β appropriately chosen, we can guarantee that $\mathbb{E}[Y|\hat{\Phi}_i(X), e] = \mathbb{E}[Y|\hat{\Phi}_i(X), e']$, for all $e, e' \in \mathcal{E}^i_{all}$, and the test error $\mathbb{E}_{(X,Y)\sim\mathbb{P}^e}[\ell(f_{\hat{\theta}_i}(X), Y)] = \min_{\omega, Z\subset\Phi_i} \mathbb{E}_{(X,Y)\sim\mathbb{P}^e}[\ell(\omega(Z(X)), Y)]$ for any $e \in \mathcal{E}^i_{all}$.
2. Otherwise, with β appropriately chosen, we can guarantee that $\mathbb{E}[Y|\hat{\Phi}_i(X), e] = \mathbb{E}[Y|\hat{\Phi}_i(X), e'], for\ all\ e, e' \in \mathcal{E}^i_{all}$.

Proof: We denote the minimizer of the objective (5) by

$$f_{\nu^\star} \in \arg\min_{\nu} \sum_{i=1}^{N} \mathcal{R}(f_\nu; D_i) + \lambda \cdot \|\nabla_{\bar{\omega}|\bar{\omega}=1.0}\mathcal{R}(\bar{\omega} \cdot f_\nu; D_i)\|^2,$$

where $f_{\nu^\star} := \Phi_{\nu^\star} \cdot \omega_{\nu^\star}$.

When there are at least $d - \tau + \frac{d}{\tau}$ local datasets in $\mathcal{E}^D$ lying in a linear general position of degree τ, using Theorem 9 in (Arjovsky et al., 2019) we can conclude that $\Phi_{\nu^\star}$ elicits the global invariant prefictor $f_{\nu^\star} = \Phi_{\nu^\star} \cdot \omega_{\nu^\star}$ for all $e \in \mathcal{E}_{all}$, where $\mathcal{E}^i_{all} \subset \mathcal{E}_{all}$ holds for all $i \in [N]$.

Proof of conclusion 1 *Clue*: we can prove conclusion 1 by constructing a contradiction.
Firstly, when there are at least $d - (\tau + q_i) + \frac{d}{\tau+q_i}$ training contexts in $\mathcal{E}^i_{train}$ lying in a linear general position of degree $\tau + q_i$, we can conclude that $\hat{\Phi}_i$ elicits the invariant prefictor $f_{\hat{\theta}_i} = \hat{\Phi}_i \cdot \hat{\omega}_i$ for all $e \in \mathcal{E}^i_{all}$. For the test error guarantee,

Suppose there exists a model $f_{\bar{\theta}_i} = \bar{\Phi}_i \cdot \bar{\omega}_i$ that is the minimizer of upper objective (4), but there isn't any apppropriate β such that $f_{\bar{\theta}_i}$ guarantees the test error $\mathbb{E}_{(X,Y)\sim\mathbb{P}^e}[\ell(f_{\bar{\theta}_i}(X), Y)] = \min_{\omega, Z\subset\Phi_i} \mathbb{E}_{(X,Y)\sim\mathbb{P}^e}[\ell(\omega(Z(X)), Y)]$ for any $e \in \mathcal{E}^i_{all}$.

As regard to the objective (2): $\min_{\theta_i} \sum_{j=1}^{K_i} \mathcal{R}(f_{\theta_i}; e^i_j) + \lambda \cdot \|\nabla_{\bar{\omega}|\bar{\omega}=1.0} \mathcal{R}(\bar{\omega} \cdot f_{\theta_i}; e^i_j)\|^2 + \beta \|\theta_i - \nu^\star\|^2$. We know that there always exists an appropriate β makes it equivalent to

$$\min_{\theta_i} \sum_{j=1}^{K_i} \mathcal{R}(f_{\theta_i}; e^i_j) + \lambda \cdot \|\nabla_{\bar{\omega}|\bar{\omega}=1.0} \mathcal{R}(\bar{\omega} \cdot f_{\theta_i}; e^i_j)\|^2$$
$$s.t. \quad \|\theta_i - \nu^\star\|^2 \leq \delta_d^2$$

where δ_d is a constant. In other word, with appropriate β chosen, $f_{\bar{\theta}_i} = \bar{\Phi}_i \cdot \bar{\omega}_i$ can guarantee that $\mathbb{E}_{(X,Y)\sim\mathbb{P}^e}[\ell(f_{\bar{\theta}_i}(X), Y)]$ is minimized for all $e \in \mathcal{E}^i_{all}$. This result contradicts the previous assumption.

Therefore, the conclusion 1 in Theorem 2 is proved.

Proof of conclusion 2 When the training contexts on local client is insufficient, the minimizer of the objective (4) can necessarily rely on the spurious features. However, with an appropriate β chosen, the minimizers can be constrained to be $f_{\hat{\theta}_i} = \hat{\Phi}_i \cdot \hat{\omega}_i = f_{\nu^\star}$. It has been proved $f_{\nu^\star}$ can guarantee $\mathbb{E}[Y|\Phi_{\nu^\star}(X), e] = \mathbb{E}[Y|\Phi_{\nu^\star}(X), e'], for\ all\ e, e' \in \mathcal{E}_{all}$. Because $\mathcal{E}^i_{all} \subset \mathcal{E}_{all}$ holds for all $i \in [N]$, $f_{\nu^\star}$ can naturally guarantee that $\mathbb{E}[Y|\Phi_{\nu^\star}(X), e] = \mathbb{E}[Y|\Phi_{\nu^\star}(X), e'], for\ all\ e, e' \in \mathcal{E}^i_{all}$.

Proof ends.

A.4 Detailed experimental setup

In this chapter, we will provide the detailed experimental setup, including the dataset setup, model selection and more implementation details.

Dataset setup We give the detailed setup of the adopted datasets in Table 6. For the RC-MNIST dataset, we use the first 50000 images in the train-set of MNIST LeCun et al. (1998) to constrcut the train-sets for the clients, and use the other 10000 images in the train-set of MNIST LeCun et al. (1998) to constrcut the test-sets for the clients. Finally, in every trial, each client has 12500 data instances for training and 2500 data instances for testing. In the original MNIST dataset, the shape of the images is $28 * 28$. After the construction, the images in the obtained RC-MNIST dataset have the size of $28*28*2$. To save the computation cost, we downsample the images to $14*14*2$ before they are input into the first layer of the neural network. The RC-FMNIST dataset is constructed and processed using the same strategies as the RC-MNIST dataset. The WaterBird dataset is constructed using the bird dataset (Caltech-UCSD Birds-200-2011 Wah et al. (2011)) and the background dataset (Places Zhou et al. (2017)), as first introduced in Sagawa et al. (2019). We adopt the same law as in Sagawa et al. (2019) to construct our datasets. However, we only use the bird photographs in the train-set of Caltech-UCSD Birds-200-2011 which contains 5994 instances. For each client, we randomly sample 400 instances from the index range (shown in Table 6) of bird dataset to generate the train-set and 100 instances from the same range to generate the test-set. Besides, all of the corresponding background photographs are randomly sampled from the background dataset (Places Zhou et al. (2017)) without replacement.

Model selection For RC-MNIST and RC-FMNIST dataset, we adopt a deep neural network with two hidden layers (the dimension is 390 and 390 respectively) as the feature extractor and an subsequent fully-connected layer (the dimension is 390) as the classifier. Besides, after each haidden layer, there is a ReLU layer in the feature extractor. In the evaluation on WaterBird dataset, we adopt the standard ResNet-18 He et al. (2016) as learning model. Similarly, the part before the last fully-connected layer works as the feature extractor and the last fully-connected layer works as the classifier, and the output dimension of the feature extractor is 512.

Table 6: The detailed setup of the local datasets

		Client 1	Client 2	Client 3	Client 4
RC-MNIST	Train-set	$p^e_{train} = 0.95$ Rotated by 0°	$p^e_{train} = 0.90$ Rotated by 90°	$p^e_{train} = 0.85$ Rotated by 180°	$p^e_{train} = 0.80$ Rotated by 270°
	Test-set	$p^e_{test} = 0.10$ Rotated by 0°	$p^e_{test} = 0.10$ Rotated by 90°	$p^e_{test} = 0.10$ Rotated by 180°	$p^e_{test} = 0.10$ Rotated by 270°
RC-FMNIST	Train-set	$p^e_{train} = 0.95$ Rotated by 0°	$p^e_{train} = 0.90$ Rotated by 90°	$p^e_{train} = 0.85$ Rotated by 180°	$p^e_{train} = 0.80$ Rotated by 270°
	Test-set	$p^e_{test} = 0.10$ Rotated by 0°	$p^e_{test} = 0.10$ Rotated by 90°	$p^e_{test} = 0.10$ Rotated by 180°	$p^e_{test} = 0.10$ Rotated by 270°
WaterBird	Train-set	$p^e_{train} = 0.95$ Birds sampled from $[0, 1715]$	$p^e_{train} = 0.90$ Birds sampled from $[1716, 2908]$	$p^e_{train} = 0.85$ Birds sampled from $[2909, 3844]$	$p^e_{train} = 0.80$ Birds sampled from $[3845, 5993]$
	Test-set	$p^e_{test} = 0.10$ Birds sampled from $[0, 1715]$	$p^e_{test} = 0.10$ Birds sampled from $[1716, 2908]$	$p^e_{test} = 0.10$ Birds sampled from $[2909, 3844]$	$p^e_{test} = 0.10$ Birds sampled from $[3845, 5993]$

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 @ 3.50GHz with one NVIDIA GeForce RTX 3090 GPU).