Out-of-Distribution Generalization of Federated Learning via Implicit Invariant Relationships

Yaming Guo [*1] **Kai Guo** [*1] **Xiaofeng Cao** [1] **Tieru Wu** [1] **Yi Chang** [1]

Abstract

Out-of-distribution generalization is challenging for non-participating clients of federated learning under distribution shifts. A proven strategy is to explore those invariant relationships between input and target variables, working equally well for non-participating clients. However, learning invariant relationships is often in an explicit manner from data, representation, and distribution, which violates the federated principles of privacy-preserving and limited communication. In this paper, we propose FEDIIR, which implicitly learns invariant relationships from parameter for out-of-distribution generalization, adhering to the above principles. Specifically, we utilize the prediction disagreement to quantify invariant relationships and implicitly reduce it through inter-client gradient alignment. Theoretically, we demonstrate the range of non-participating clients to which FEDIIR is expected to generalize and present the convergence results for FEDIIR in the massively distributed with limited communication. Extensive experiments show that FEDIIR significantly outperforms relevant baselines in terms of out-of-distribution generalization of federated learning.

1. Introduction

With the growth of storage and computational capabilities on devices within distributed networks, **federated learning** has emerged as a popular distributed learning paradigm(Yang et al., 2019; Kairouz et al., 2021; Li et al., 2020a). In the federated learning scenario, multiple **clients** collaborate to solve machine learning problems under the coordination of a **server**, where each client's raw data is stored locally and is not exchanged or transferred. The federated networks are usually comprised of a large number of clients that generate and collect data in a non-identical distribution manner, most of which may never participate in training. Since the distribution shift probably exists between participating and non-participating (unseen) clients, models that follow empirical risk minimization (ERM) may perform poorly on the non-participating clients, known as the **out-of-distribution (OOD) generalization** problem(Mohri et al., 2019; Yuan et al., 2022). The OOD generalization is challenging under the principles specific to federated learning.

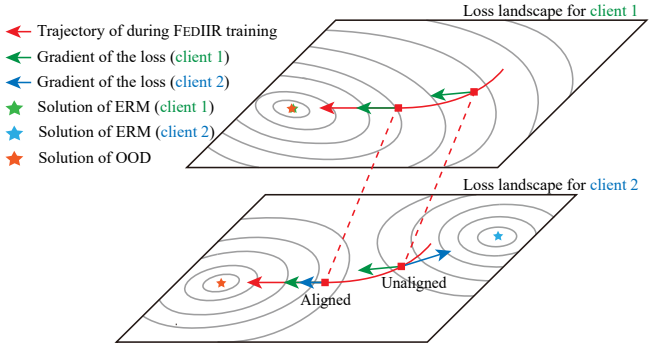

Figure 1. Illustration of inter-client gradient alignment with two clients. If the inter-client gradient is aligned, the model's local learning on one client will also improve its performance on other clients. This indicates that the model implicitly learns invariant relationships that work equally for all clients. In this way, FEDIIR will converge to the OOD solution instead of the ERM one.

A proven strategy in the OOD generalization literature(Peters et al., 2016; Arjovsky et al., 2019) is to learn the **invariant relationships** that are stable across distributions and build a model that works equally well over OOD. Intuitively, an invariant relationship is a statistical relationship between inputs and target variables that is maintained across all data distributions. A typical example(Beery et al., 2018) is training a model to classify camels from cows, which fails when the background is switched. The reason is that the model classifies relying on the spurious relationship (i.e., background color vs. label) rather than the invariant relationship (i.e., animal feature vs. label). Therefore, when the client's data are drawn from different distributions, the model should make predictions using invariant relationships instead of relying on varying spurious relationships.

[*]Equal contribution [1]School of Artificial Intelligence, Jilin University, Changchun, China. Correspondence to: Xiaofeng Cao <xiaofengcao@jlu.edu.cn>, Tieru Wu <wutr@jlu.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Question We find that there exists an open discussion: *could the current techniques for learning invariant relationships adhere entirely to the federated principles of* ***privacy-preserving*** *and* ***limited communication****?*

An Explicit Perspective Most existing work concentrates on learning invariant relationships explicitly from three angles: data, representation, and distribution(Shen et al., 2021; Wang et al., 2022). The methods that rely on data or representation, such as IRM(Arjovsky et al., 2019) and REx(Krueger et al., 2021), require a centralized setting where data or representation is shared across clients, putting clients' privacy at risk. The distribution-based methods, such as FedADG(Zhang et al., 2021) and FedSR(Nguyen et al., 2022), can protect privacy by matching distributions. Still, in the scenario where clients are massively distributed with limited communication [1], these methods may fail because they usually assume the presence of only a small number of participating clients, most of which are involved in each round of communication.

Motivation However, those practical explicit methods may not entirely adhere to the federated principles of privacy-preserving and limited communication. Considering that the model parameter is usually the only interaction between the client and the server, we thus stand on a new perspective, i.e., restrict the method to the parameter space for learning invariant relationships implicitly.

A New Perspective: Implicit From this perspective, an implicit method does not need to communicate anything other than the parameter, which can offer better protection for client privacy than explicit methods. Additionally, these implicit methods can be analyzed in the stochastic optimization framework like standard federated techniques(Wang et al., 2021), which helps to examine its convergence behavior in scenarios with a large number of clients. In fact, convergence analysis is usually neglected in the OOD generalization literature(Nagarajan et al., 2021; Ahuja et al., 2021). But it is important to know the convergence rate of the method in federated learning, as communication is the primary bottleneck(Wang et al., 2021; Kairouz et al., 2021).

Motivated by the above analysis, this paper proposes ***Fed****erated Learning with* ***I****mplicit* ***I****nvariant* ***R****elationships* (FedIIR), which implicitly learns invariant relationships for OOD generalization while adhering to the federated principles of privacy-preserving and limited communication. Specifically, we theoretically introduce prediction disagreement to quantify the invariant relationships and obtain its surrogate in the parameter space by parameterization, i.e., the maximum gap of the gradient across clients. By aligning the inter-client gradient, FedIIR implicitly reduces prediction disagreement, encouraging the model to use invariant relationships for making predictions. As seen in Figure 1, when the inter-client gradient is unaligned, the model's local learning on client 2 degrades its performance on client 1, which indicates the model absorbs spurious relationships detrimental to client 1. After aligning the inter-client gradient, the model is able to learn invariant relationships that work equally for all clients implicitly. We summarize our main contributions below.

1. We propose a method for implicitly learning invariant relationships, called FedIIR, designed to address the problem of OOD generalization under the challenges specific to federated learning.
2. We theoretically demonstrate that FedIIR is expected to generalize to non-participating clients whose distributions can be written as an affine combination of participating clients' distributions. We also present the convergence results for FedIIR in the massively distributed with limited communication scenario, including both μ-PL inequality and non-convex cases.
3. We validate the effectiveness of the proposed method using two scenarios: a small number of clients and a large number of clients (limited communication). The experimental results show that FedIIR significantly outperforms existing federated learning methods in terms of OOD generalization.

2. Preliminaries

In this section, we introduce the federated learning scenario and formalize its OOD generalization problem.

Notation Let $\mathcal{X}$ and $\mathcal{Y}$ represent the input space and target space, respectively. $\mathcal{C}_{\text{all}}$ is the (possibly infinite) collection of all the possible clients. We denote by $\mathcal{C}_{\text{par}} \subseteq \mathcal{C}_{\text{all}}$ the participating clients, where $\mathcal{C}_{\text{par}}$ is drawn from a distribution $\mathcal{Q}_{\text{par}}$. Each client $c \in \mathcal{C}_{\text{all}}$ holds a local dataset denoted as $D_c = \{(x_i^c, y_i^c)\}_{i=1}^{n_c}$ i.i.d. drawn from the distribution $\mathbb{P}_c$ over $\mathcal{X} \times \mathcal{Y}$, and the random variables that follow $\mathbb{P}_c$ are represented by $(X^c, Y^c) \sim \mathbb{P}_c$. The model we consider is formalized as $f = w \circ \Phi$, where $\Phi : \mathcal{X} \to \mathcal{Z} \subseteq \mathbb{R}^n$ denotes the (feature) representation and $w : \mathcal{Z} \to \hat{\mathcal{Y}}$ denotes the classifier[2]. The set of all models is given as $\mathcal{F}$. We assume that the model is parameterized as $f_\theta = w_\omega \circ \Phi_\phi$, where $\theta = (\phi, \omega)$. We denote the loss function as $\ell(f(x), y) : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}_+$, e.g., cross-entropy loss in classification problems, mean-squared error loss in regression problems, etc. We use $\mathcal{R}(f) = \mathbb{E}_{c \sim \mathcal{Q}_{\text{par}}} \mathcal{R}_c(f)$ to denote the global expected risk of model f, where $\mathcal{R}_c(f) = \mathbb{E}_{X^c, Y^c} \ell(f(X^c), Y^c)$ is the expected risk w.r.t. client c. With finite samples,

[1] In this scenario, only a fraction of clients is available during one communication round, as participating clients are sampled from an enormous population.

[2] In a regression problem, the term "classifier" means the last layer of the regression model.

the empirical risk of model f w.r.t. client c is defined as $R_c(f) = \frac{1}{n_c}\ell(f(x_i^c), y_i^c)$ and the global empirical risk is $R(f) = \mathbb{E}_{c\sim\mathcal{Q}_{\text{par}}} R_c(f)$. Unless otherwise stated, $\|\cdot\|$ denotes the Euclidean L_2-norm.

Federated Learning This work is interested in the scenario where clients are massively distributed with limited communication, referred to as the cross-device federated learning in Kairouz et al. (2021). In this scenario, the clients are a very large number of highly unreliable mobile devices, e.g., phones or sensors, in which only a fraction of clients participate in each round.

The federated learning scenario involves two levels of sampling, drawing client c from the distribution $\mathcal{Q}_{par}$ and then drawing sample (x, y) from that client's local data distribution $\mathbb{P}_c$. In the standard federated learning paradigm, the goal is to learn a global model that minimizes the average expected risk over all participating clients, i.e., the ***Empirical Risk Minimization*** (ERM) principle(Vapnik, 1991; McMahan et al., 2017). Mathematically, ERM is formalized as the following optimization problem.

$$\min_f \mathop{\mathbb{E}}_{c\sim\mathcal{Q}_{\text{par}}} \left[\mathcal{R}_c(f) = \mathop{\mathbb{E}}_{X^c,Y^c} \ell(f(X^c), Y^c)\right] \quad \text{(ERM)}$$

A standard algorithm for solving (ERM) is FEDAVG(McMahan et al., 2017), whose pseudo-code is presented in Algorithm 1. Intrinsically, (ERM) minimizes the risk of the model over the mixture distribution $\bigcup_{c\in\mathcal{C}_{\text{par}}} \mathbb{P}_c$. The distribution of non-participating clients may be, in general, quite different from $\bigcup_{c\in\mathcal{C}_{\text{par}}} \mathbb{P}_c$ because there are distribution shifts across clients(Mohri et al., 2019). Thus, models following the ERM principle may perform poorly on non-participating clients.

Out-of-distribution Generalization In order to generalize the model appropriately to non-participating clients, we examine the problem of ***Out-Of-Distribution*** (OOD) generalization in federated learning. Following Arjovsky et al. (2019), we formulate it as finding the model that performs optimally in the worst case. Within the terminology of federated learning, OOD generalization attempts to minimize the maximum risk of the model over all possible clients, formalized as the following optimization problem.

$$\min_f \max_{c\in\mathcal{C}_{all}} \left[\mathcal{R}_c(f) = \mathop{\mathbb{E}}_{X^c,Y^c} \ell(f(X^c), Y^c)\right] \quad \text{(OOD)}$$

The (OOD) cannot be directly solved because we cannot observe all potential clients in $\mathcal{C}_{all}$. To generalize to non-participating clients, a key of existing methods(Peters et al., 2016; Ahuja et al., 2020) is to learn invariant relationships between input X and target Y, which is maintained across all clients. The model using invariant relationships removes spurious correlations that depend on specific clients and thus can be robustly generalized to non-participating clients.

Why need an implicit method? As previously mentioned, explicit methods for learning invariant relationships can hardly adhere to the federated principles of privacy-preserving and limited communication. The first reason is that most explicit methods require sharing data or representations across clients, violating privacy-preserving. The second reason is that explicit methods usually require clients to participate in each round, which violates limited communication. Because the model parameter is typically the only interaction between the client and the server, we argue that methods for learning invariant relationships should be restricted to the parameter space. In this way, implicit methods provide better privacy protection and can also be analyzed efficiently for convergence in the scenario where clients are massively distributed with limited communication.

3. The Proposed FEDIIR Method

In this section, we present our FEDIIR method. We first introduce prediction disagreement theoretically to quantify the invariant relationship and obtain its surrogate objective in the parameter space by parameterization. Based on this surrogate objective, we formulate the optimization objective of FEDIIR, which learns the invariant relationships implicitly through inter-client gradient alignment. Finally, we outline the optimization process for FEDIIR in the massively distributed with limited communication scenario.

3.1. Implicit Invariant Relationships

Without prior knowledge or structural assumptions, the (OOD) problem would be pointless because it is impossible to characterize the samples from the non-participating client. A commonly used assumption in the invariant learning literature(Rojas-Carulla et al., 2018; Arjovsky et al., 2019; Liu et al., 2021) is as follows.

Assumption A. There exists a representation $\Phi(\cdot)$ such that for all $c, c' \in \mathcal{C}_{\text{all}}$ and for all z in the intersection of the supports $\text{supp}\Big(\mathbb{P}(\Phi(X^c))\Big) \cap \text{supp}\Big(\mathbb{P}(\Phi(X^{c'}))\Big)$, we have

$$\mathop{\mathbb{E}}_{X^c,Y^c}[Y^c|\Phi(X^c) = z] = \mathop{\mathbb{E}}_{X^{c'},Y^{c'}}[Y^{c'}|\Phi(X^{c'}) = z].$$

This assumption shows that the relationship between representation $\Phi(X)$ and target Y is fixed across distributions in $\mathcal{C}_{\text{all}}$, i.e., using $\Phi(X)$ to predict Y is invariant. We call such a relationship between $\Phi(X)$ and Y an **invariant relationship**, and a model that only uses the invariant relationship to predict is called the **invariant predictor**. The invariant

predictor captures the latent invariant relationship between the input variable X and the target variable Y, which works equally well across all clients in $\mathcal{C}_{\text{all}}$.

We attempt to learn the invariant relationships implicitly because the model parameter is the only interaction between the clients and the server. To do so, we consider a representation $\Phi(\cdot)$ on which the optimal classifier for the client c is specified as $w_c^*(\cdot)$. For some loss functions, such as cross-entropy and mean-squared error, the optimal classifiers can be written as conditional expectations, i.e., $w_c^*(z) = \mathbb{E}[Y^c|\Phi(X^c) = z]$. Recall the invariant relationship w.r.t. Assumption A, which shows that the optimal classifiers of all clients have the same prediction for the representation z. Based on the above observations, we can quantify the invariant relationship using the prediction disagreement between the optimal classifier of clients, being formalized as follows.

Definition 1 (Prediction Disagreement). Given the collection $\mathcal{C}$ of clients, for the representation $\Phi(\cdot)$, let the optimal classifier for client c be $w_c^*(\cdot)$. The prediction disagreement w.r.t. $\mathcal{C}$ and $\Phi(\cdot)$ is defined as:

$$\mathcal{I}(\Phi, \mathcal{C}) = \sup_{z \in \mathrm{U}(\Phi, \mathcal{C})} \sup_{(c,c') \in \mathcal{C}^2} |w_c^*(z) - w_{c'}^*(z)|,$$

where $\mathrm{U}(\Phi, \mathcal{C}) = \cup_{c \in \mathcal{C}} \mathrm{supp}\big(\mathbb{P}(\Phi(X^c))\big)$ is the union of the supports.

The prediction disagreement $\mathcal{I}(\Phi, \mathcal{C})$ gives the maximum prediction gap among the client's optimal classifiers. The following theorem relates the prediction disagreement to the invariant predictor, proved in Appendix B.

Theorem 2. *Given the collection $\mathcal{C}_{\text{all}}$ of clients, for the representation $\Phi(\cdot)$, let the optimal classifier for client c be $w_c^*(\cdot)$. If $\mathcal{I}(\Phi, \mathcal{C}_{\text{all}}) = 0$, then $f = w \circ \Phi$ is an invariant predictor, where $w(\cdot) = w_c^*(\cdot)$ for any $c \in \mathcal{C}_{\text{all}}$.*

According to this theorem, a representation with the lowest prediction disagreement elicits an invariant predictor, i.e., the model that only predicts using invariant relationships. Thus, we can lead to an invariant predictor by reducing the prediction disagreement w.r.t. Definition 1. Suppose that the current global model is f with parameters $\theta = (\phi, \omega)$. The client c initializes the local model using f and performs a gradient descent $\omega_c = \omega - \nabla_\omega \mathcal{R}_c(\theta)$. We approximate $w(\cdot; \omega_c)$ via its first-order Taylor expansion w.r.t. $-\nabla_\omega \mathcal{R}_c(\theta)$ around ω, obtaining

$$\begin{aligned}
&\sup_{(c,c') \in \mathcal{C}^2} |w(z; \omega - \nabla_\omega \mathcal{R}_c(\theta)) - w(z; \omega - \nabla_\omega \mathcal{R}_{c'}(\theta))| \\
\approx &\sup_{(c,c') \in \mathcal{C}^2} |w(z;\omega) - [\nabla_\omega w(z;\omega)]^\top \nabla_\omega \mathcal{R}_c(\theta) \\
&\qquad - w(z;\omega) + [\nabla_\omega w(z;\omega)]^\top \nabla_\omega \mathcal{R}_{c'}(\theta)| \\
\leq &\sup_{(c,c') \in \mathcal{C}^2} \|\nabla_\omega w(z;\omega)\| \underbrace{\|\nabla_\omega \mathcal{R}_c(\theta) - \nabla_\omega \mathcal{R}_{c'}(\theta)\|}_{\mathcal{A}}.
\end{aligned}$$

The first thing to see is that the term "$\mathcal{A}$" only involves the parameter space, which is the inter-client gradient gap concerning the classifier. The term "$\mathcal{A}$" can be viewed as a surrogate objective for prediction disagreement in the parameter space. Thus, we can implicitly reduce prediction disagreement by aligning the inter-client gradient of the classifier, thereby encouraging the model to use invariant relationships for making predictions. In contrast to the work(Arjovsky et al., 2019; Ahuja et al., 2020) that explicitly enforces the invariance constraint w.r.t. Assumption A, this strategy of implicitly promoting invariance is more suitable for federated learning scenarios.

3.2. Optimization Objective

As analyzed in the previous section, we can learn invariant relationships implicitly by aligning the inter-client gradient of the classifier. To enable the optimization objective to be separable across clients, we propose to align the classifier's local and global gradient, i.e., $\|\nabla_\omega \mathcal{R}_c(f) - \nabla_\omega \mathcal{R}(f)\|$, where $\mathcal{R}(f) = \mathbb{E}_{c \sim \mathcal{Q}_{\text{par}}} \mathcal{R}_c(f)$ is the global expected risk. Therefore, FEDIIR attempts to minimize the empirical risk and align the local and global gradients of the classifier, formalized as the following optimization objective.

$$\min_f \mathop{\mathbb{E}}_{c \sim \mathcal{Q}_{\text{par}}} \Big[\mathcal{R}_c(f) + \frac{\gamma}{2} \|\nabla_\omega \mathcal{R}_c(f) - \nabla_\omega \mathcal{R}(f)\|^2\Big] \qquad \text{(FEDIIR)}$$

The optimization objective of FEDIIR is quite direct: it adds a penalty term concerning the gradient of the classifier to the local sub-problem. The penalty factor γ controls the balance between the reducing risk term and the aligning gradient term, with $\gamma = 0$ recovering (ERM). In the practice of deep learning, these objectives may be at odds with each other, as the reducing risk term greedily maximizes the speed of learning(Parascandolo et al., 2021). Therefore, FEDIIR prioritizes learning invariant relationships by trading off some convergence speed, resulting in an improved OOD generalization capability of the model.

By minimizing risk and aligning gradient, FEDIIR forces the local learning to progress consistently across clients. If the inter-client gradient is aligned, the model's local learning on one client will also improve its performance on other clients, indicating that the model learns invariant relationships that work equally for all clients. Further, FEDIIR also prompts the model to converge on a client-shared solution. Specifically, we assume that the model $f = w \circ \Phi$ converges to the stationary point of global expected risk, i.e., $\nabla \mathcal{R}(f) = \mathbb{E}_{c \sim \mathcal{Q}_{\text{par}}} \nabla \mathcal{R}_c(f) = \mathbf{0}$. If classifier $w(\cdot)$ has an equal gradient on all client, then $\nabla_\omega \mathcal{R}_c(f) = \mathbf{0}$ holds for all $c \in \mathcal{C}_{\text{par}}$. Notice that the classifier of the invariant predictor also reaches optimality for all clients simultaneously (see

Definition 6) since the invariant features have the same joint distributions with the targets across all clients.

Discussion of related work Our objective is similar to that of IGA(Koyama & Yamaguchi, 2020), a centralized approach that attempts to align the gradient of the model $f = w \circ \Phi$ rather than only the classifier w. In fact, IGA is much more expensive to optimize than FEDIIR because, most of the time, $|\omega| \ll |\theta|$. We show that satisfying the invariance of the classifier w is sufficient, which can be effectively optimized in limited resources of communication. Moreover, our objective is similar to FEDPROX(Li et al., 2020b), which adds a penalty term to the local sub-problem regarding the model parameter. The penalty term of FEDPROX can be roughly written as $\frac{\gamma}{2}\|\theta_c - \overline{\theta}\|^2$, where $\overline{\theta}$ is the current global model parameter. A drawback of FEDPROX is that it may lead to $\theta_c \approx \overline{\theta}$ in local learning, inherently limiting the potential of local learning. Our method does not confine local learning to the area around the global model and is more concerned with the consistency of the learning process. A more extensive discussion of related work is provided in Appendix A.

3.3. Optimization in Limited Communication

We describe below the optimization process for the proposed objective in the massively distributed with limited communication scenario. A widespread practice derived from FEDAVG(McMahan et al., 2017) is performing multiple local updates on the clients to reduce communication costs. To do so, we maintain a server-level state that remains fixed throughout the local update of clients. We summarize the pseudo-code of FEDAVG and FEDIIR in Algorithm 1.

We start by rewriting some notations under finite samples only in terms of θ, instead of f. We denote the global empirical risk as $R(\theta) = \mathbb{E}_{c\sim\mathcal{Q}_{\text{par}}} R_c(\theta)$ w.r.t. all participating clients $\mathcal{C}_{\text{par}}$, where $R_c(\theta)$ is the empirical risk of model f (corresponding to θ) w.r.t. client c. In particular, for a mini-batch $\zeta = \{(x_i, y_i)\}_{i=1}^n$ with size n, we denote by $R(\theta;\zeta) = \frac{1}{n}\ell(f(x_i), y_i)$ the empirical risk of θ over ζ.

In the t-th communication round, the server randomly samples a subset $\mathcal{C}$ of clients with $|\mathcal{C}| = C$. For the global model parameter θ^{t-1}, we then use the average full-batch gradient $\overline{g}_\omega = \frac{1}{C}\sum_{c\in\mathcal{C}} \nabla_\omega R_c(\theta^{t-1})$ as the unbiased estimate[3] of the global gradient $\nabla_\omega R(\theta^{t-1})$. Note that $\overline{g}_\omega$ remains fixed throughout the local update of the clients. For each $c \in \mathcal{C}$, client c initializes the local model parameter $\theta_c^t = \theta^{t-1}$ and performs K local updates:

$$\theta_c^t \leftarrow \theta_c^t - \eta_l g_c,$$

where η_l denotes the local step-size, $g_c \leftarrow \nabla R_c(\theta_c^r;\zeta)$ is the stochastic gradient of FEDAVG and $g_c \leftarrow \nabla[R_c(\theta_c^r;\zeta) + \frac{\gamma}{2}\|\nabla_\omega R_c(\theta_c^r;\zeta) - \overline{g}_\omega\|^2]$ is the stochastic gradient of FEDIIR. Finally, the server aggregates the parameter updates of the sampled clients:

$$\theta^t = \theta^{t-1} + \eta_g \frac{1}{C}\sum_{c\in\mathcal{C}} \Delta_c,$$

where η_g denotes the global step-size and $\Delta_c = \theta_c^t - \theta^{t-1}$ denotes the the parameter updates on client c.

[3] Appendix C presents a discussion in detail on the additional computation introduced and the stability of the global gradient estimate in FEDIIR.

Algorithm 1 FEDAVG and FEDIIR

Initialize: θ^0
for $t = 1, \dots, T$ **do**
  sample subset $\mathcal{C}$ of clients with $|\mathcal{C}| = C$
  $\overline{g}_\omega := \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \nabla_\omega R_c(\theta^{t-1})$
  on client $c \in \mathcal{C}$ **in parallel do**
  initialize local model $\theta_c^t = \theta^{t-1}$
  for $k = 1, \dots, K$ **do**
    sample mini-batch ζ from local data D_c
    $g_c \leftarrow \nabla R_c(\theta_c^t;\zeta)$
    $g_c \leftarrow \nabla[R_c(\theta_c^t;\zeta) + \frac{\gamma}{2}\|\nabla_\omega R_c(\theta_c^t;\zeta) - \overline{g}_\omega\|^2]$
    update $\theta_c^t \leftarrow \theta_c^t - \eta_l g_c$
  end for
  $\Delta_c = \theta_c^t - \theta^{t-1}$
  end on client
  $\theta^t = \theta^{t-1} + \eta_g \frac{1}{C}\sum_{c\in\mathcal{C}} \Delta_c$
end for

4. Theoretical Analysis

In this section, we present our main theoretical results. We first establish a generalization risk bound, providing the range of non-participating clients to which FEDIIR is expected to generalize. We also present the convergence results for FEDIIR in the scenario where clients are massively distributed with limited communication, guaranteeing that the global empirical risk can converge to a stationary point.

4.1. Generalization Analysis

When the number of participating clients is finite in practice, what is the range of non-participating clients to which FEDIIR is expected to generalize? We provide a simple generalization risk bound that gives some clues for this question. We consider the binary classification setting $\mathcal{Y} = \{0, 1\}$, and $\hat{\mathcal{Y}} = [0, 1]$ denotes the estimated probability of the true label being 1. Given a collection of participating clients $\mathcal{C}_{\text{par}}$, we use $\mathcal{R}(f) = \mathbb{E}_{c\sim\mathcal{Q}_{\text{par}}} \mathcal{R}_c(f)$ to denote the global expected risk. For $\lambda \in \Lambda_\nu := \{\{\lambda_c : c \in \mathcal{C}_{\text{par}}\} | \lambda_c \geq -\nu, \sum_{c\in\mathcal{C}_{\text{par}}} \lambda_c = 1\}$, we define a non-participating client λ with distribution $\mathbb{P}_\lambda$, where $\mathbb{P}_\lambda = \sum_{c\in\mathcal{C}_{\text{par}}} \lambda_c \mathbb{P}_c$ is an

affine combination. For this non-participating client λ, its expected risk is denoted as $\mathcal{R}_\lambda(f) = \sum_{c\in\mathcal{C}_{\text{par}}} \lambda_c \mathcal{R}_c(f)$. If $\nu = 0$, the non-participating clients fall within the convex hull of participating clients; if $\nu > 0$, the non-participating clients fall outside of that convex hull. The following theorem presents a sufficient condition for the model to generalize to the affine combination of participating clients, proved in Appendix D.

Theorem 3. *Given the collection $\mathcal{C}_{\text{par}}$ of clients, let's assume that $\ell(\cdot,\cdot) \le M$. Then for all $f = w \circ \Phi \in \mathcal{F}$, we have the following risk bound for the affine combination of participating clients:*

$$\sup_{\lambda\in\Lambda_\nu} \mathcal{R}_\lambda(f) \le \mathcal{R}(f) + \widetilde{M}\mathcal{I}(\Phi, \mathcal{C}_{\text{par}}) + \widetilde{M} \sup_{(c,c')\in\mathcal{C}_{\text{par}}^2} \rho(\mathbb{P}_c(X), \mathbb{P}_{c'}(X)),$$

where $\widetilde{M} = (1 + |\mathcal{C}_{\text{par}}|\nu)M$ is monotonic in ν, and $\rho(\mathbb{P}_c(X), \mathbb{P}_{c'}(X)) = \sup_X |\mathbb{P}_c(X) - \mathbb{P}_{c'}(X)|$ is the total variation distance.

The upper bound of risk on the affine combinations of participating clients consists of three terms: the first term is the global expected risk; the second term corresponds to the prediction disagreement; and the third term measures the distance between marginal distributions. Most upper bounds for OOD generalization(Albuquerque et al., 2019; Arjovsky, 2020) involve the invariance constraint or the covariate shift across $\mathcal{C}_{\text{all}}$, thus they do not provide insights about the range of generalization. In contrast, Theorem 3 shows that the model is expected to perform as well on non-participating clients as on participating clients if the invariance constraint and covariate shift in $\mathcal{C}_{\text{par}}$ are sufficiently small, where the distribution of non-participating clients can be written as an affine combination of participating clients' distributions. As a result, FEDIIR implicitly reduces the invariance constraint by aligning the inter-client gradient, thereby promising to generalize to non-participating clients included in the affine combination of participating clients.

4.2. Convergence Analysis

Because communication is the primary bottleneck in federated learning, we do not want to sacrifice convergence speed alone for OOD generalization. To this end, we analyze the convergence of the FEDIIR algorithm w.r.t. global empirical risk $R(\theta) = \mathbb{E}_{c\sim\mathcal{Q}_{\text{par}}} R_c(\theta)$, including the **$\mu$-PL inequality** and the **general non-convex** two classes of functions. Although the gradient alignment term also affects the optimization of representation $\Phi(\cdot)$, we fix Φ as the identity mapping to clearly show the convergence results. We first state a few customary assumptions on the function, and their formal versions can be found in Appendix E.1.

Assumption B (Smoothness)**.** For all clients c, we assume that $R_c(\omega)$ is L-smoothness and Moral-smoothness.

Assumption C (Bounded Statistical Heterogeneity)**.** For all clients c, we assume that when there is no perturbation, the variance of the local gradient w.r.t. the global gradient is bounded by G.

Assumption D (Bounded Intra-client Variance)**.** For all clients c, we assume that $\nabla R_c(\omega;\zeta)$, $\nabla^2 R_c(\omega;\zeta)$, and $\nabla^2 R_c(\omega;\zeta)\nabla R_c(\omega;\zeta)$ are unbiased estimates of $\nabla R_c(\omega)$, $\nabla^2 R_c(\omega)$, and $\nabla^2 R_c(\omega)\nabla R_c(\omega)$, respectively, with variances bounded by σ^2.

Assumption B states the smoothness of the local risk function, which is standard in the optimization literature(Crane & Roosta, 2019; Elgabli et al., 2022). Because FEDIIR involves the second-order gradient, we also assume that $R_c(\omega)$ is Moral-smoothness. As proved in Roosta et al. (2022), the Moral-smoothness is strictly weaker than the gradient and Hessian Lipschitz continuous, which are commonly used in second-order methods. Assumption C bounds the variance of local gradients relative to the global gradient, a technique widely used for quantifying statistical heterogeneity in the federated learning literature(Karimireddy et al., 2021; Wang et al., 2021). Assumption D bounds the variance of the stochastic gradient and stochastic Hessian, which is common in stochastic optimization analysis(Fallah et al., 2020).

We now present the convergence results of FEDIIR for the μ-PL inequality case. We say $R(\omega)$ satisfies the μ-PL inequality (formalized as Assumption E) for $\mu > 0$ if $\|\nabla R(\omega)\|^2 \ge 2\mu(R(\omega) - R^*)$ $(\forall\omega)$, where $R^* := \min R(\omega)$. The μ-PL inequality is a generalization of strong convexity, which is much weaker than the standard notion of strong convexity and can even satisfy some non-convex functions. For the sake of the presentation, we made some simplifications, and its full version and proof can be found in Appendix E.3.

Theorem 4. *Let Assumption B, C, D and E hold and* FEDIIR *updates with constant local and global step-size such that $\eta_l \le \frac{1}{4KL\sqrt{1+\gamma^2}}$, $\tilde{\eta} = K\eta_g\eta_l < \frac{1}{2\alpha\mu}$. Then, the sequence of iterates generated by* FEDIIR *satisfies*

$$\mathbb{E}[R(\omega^t) - R^*] \le (1 - 2\alpha\mu\tilde{\eta})^t [R(\omega^0) - R^*] + \eta_l \frac{\beta_1 G^2 + \beta_2\gamma^2\sigma^2 + \beta_3\gamma^2 G^2\sigma^2}{2\alpha\mu},$$

where $\alpha > 0$ is a constant, and $\beta_1, \beta_2, \beta_3$ are the polynomials in η_l.

For the μ-PL inequality case, FEDIIR has a linear convergence rate up to a solution that is proportional to η_l. This shows that FEDIIR converges even with limited communication resources, where the penalty factor γ affects the suboptimality of the solution. If the algorithm appears to have stalled, we can halve the step-size and thus halve the

suboptimality. Remark that our result was established with a constant step-size, a strategy more commonly used in practice. We below present the convergence results of FEDIIR for the general non-convex case.

Theorem 5. *Let Assumption B, C, and D hold and* FEDIIR *updates with constant local and global step-size such that* $\eta_l \leq \frac{1}{4KL\sqrt{1+\gamma^2}}$. *Then, the sequence of iterates generated by* FEDIIR *satisfies*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla R(\omega^{t-1})\|^2 \leq \frac{R(\omega^0)-R^*}{\alpha\tilde{\eta}T} + \eta_l\frac{\beta_1 G^2+\beta_2\gamma^2\sigma^2+\beta_3\gamma^2G^2\sigma^2}{\alpha},$$

where $\alpha > 0$ *is a constant, and* $\beta_1, \beta_2, \beta_3$ *are the polynomials in* η_l. *If we choose the step-sizes* $\eta_l = \frac{1}{\sqrt{TKL}}, \eta_g = \sqrt{CK}$ *and omitting the larger order of each part, we have the convergence rates of* FEDIIR *as follows*

$$\mathcal{O}\left(\frac{(R(\omega^0)-R^*)L^2}{\sqrt{TCK}}, \frac{\sqrt{CK}L^2G^2}{\sqrt{T}}, \frac{\gamma^2\sigma^2}{\sqrt{TCK}}, \frac{\gamma^2G^2\sigma^2}{\sqrt{TCK}}\right).$$

For the general non-convex case, FEDIIR converges to a stationary point that is proportional to η_l at $\mathcal{O}(\frac{1}{T})$ rate. By correctly choosing the step-size, FEDIIR has a $\mathcal{O}(\frac{1}{\sqrt{T}})$ convergence rate. Since multiple terms in the bound increase with penalty factor γ, a larger penalty factor will slow down the convergence of the algorithm. Nevertheless, FEDIIR has a good convergence rate in the scenario where clients are massively distributed with limited communication.

5. Experiments

In this section, we conduct experiments to evaluate the performance of our proposed FEDIIR and answer the following questions. **Q1**: In a small number of clients scenario, can FEDIIR achieve better performance compared with explicit methods? **Q2**: How effective is the proposed FEDIIR in the scenario where clients are massively distributed with limited communication?

5.1. Experimental Setup

Benchmark Datasets. We conduct extensive experiments on four widely used datasets, including RotatedMNIST(Ghifary et al., 2015), VLCS(Fang et al., 2013), PACS(Li et al., 2017) and OfficeHome(Venkateswara et al., 2017). These datasets are classic OOD generalization benchmarks for classification. For all datasets, we perform the "leave-one-domain-out" strategy. More specifically, we set aside one domain as the test domain and the rest as the training domains. For a small number of clients scenario, each training domain is treated as a separate participating client. For scenarios with a large number of clients, we further split each training domain into multiple sub-domains, each of which is treated as a separate participating client. This allows some clients to share a single training domain, but no client has data from multiple domains simultaneously. See Appendix F.1 for more details.

Implementation. For the RotatedMNIST dataset, we use a network architecture comprising four 3×3 convolutional layers and one average pooling layer as the feature representation $\Phi(\cdot)$, and a single linear layer as the classifier $w(\cdot)$. For the VLCS and PACS datasets, we employ ResNet-18 as the feature representation $\Phi(\cdot)$, replacing the last fully connected layer with a 512-dimensional linear layer. And we utilize two fully connected layers as the classifier $w(\cdot)$. For the OfficeHome dataset, we use ResNet-50 as the feature representation $\Phi(\cdot)$, where the last fully connected layer is replaced by a 2048-dimensional linear layer. The classifier $w(\cdot)$ employed on OfficeHome is identical to the one used for the VLCS dataset. For each dataset, we only tune hyperparameters via grid search in the scenario with a small number of clients and do not modify them for a larger number of client scenarios (see Appendix F.3). In all experiments, we train the global model using the global step-size $\eta_g = 1$ for 100 communication rounds, where the local model on the client is trained with stochastic gradient descent (SGD) for one epoch. Per common practice, we allocate 90% of the available data for training and 10% for validation. We chose the global model that maximizes accuracy on the overall validation set as the final model (no data leakage from the test domain). We run the experiments three times and report the average performance of the final model on the test domain. Unless otherwise stated, the performance of the methods on the dataset refers to the average result obtained when each domain is treated as a test domain once. Our code will be released at https://github.com/YamingGuo98/FedIIR.

5.2. Results

Results on a small number of clients scenario. In response to **Q1**, we compare the proposed method with baselines to evaluate the OOD generalization performance on four datasets. We take FEDAVG(McMahan et al., 2017), FEDADG(Zhang et al., 2021), and FEDSR(Nguyen et al., 2022) as the baselines. The summarized results of the experiments are presented in Table 1, while detailed results for each domain can be found in Appendix F.4. From our findings, we draw the following conclusions:

i) Our method FEDIIR consistently outperforms the baseline FEDAVG on all datasets. It is worth emphasizing that compared to FEDAVG, FEDIIR achieves this performance improvement by merely introducing the inter-client gradient alignment term;

Algorithm	RotatedMNIST	VLCS	PACS	OfficeHome	Average
	ConvNet	ResNet-18	ResNet-18	ResNet-50	
FEDAVG	94.5±0.1	76.3±0.4	83.1±0.0	68.5±0.1	80.6
FEDADG	94.7±0.0	**77.1±0.1**	83.1±0.2	68.4±0.2	80.8
FEDSR	94.7±0.1	75.8±0.4	83.4±0.3	69.1±0.2	80.8
FEDIIR	**95.0±0.2**	76.6±0.6	**83.7±0.3**	**69.2±0.0**	**81.1**

Table 1. Average test accuracy (%) using leave-one-out domain validation in the scenario with a small number of clients. Each training domain is treated as a separate participating client, and all participating clients are sampled in each round of communication.

Algorithm	RotatedMNIST	VLCS	PACS	OfficeHome	Average
	ConvNet	ResNet-18	ResNet-18	ResNet-50	
FEDAVG	90.8±0.6	65.0±1.1	68.8±0.7	60.5±0.2	71.3
FEDADG	92.2±0.4	60.7±2.2	72.3±0.6	60.1±0.2	71.3
FEDSR	91.2±1.1	60.0±0.2	72.7±0.6	55.3±0.7	69.8
FEDIIR	**93.0±0.3**	**74.1±0.3**	**75.4±0.6**	**65.6±0.3**	**77.0**

Table 2. Average test accuracy (%) using leave-one-out domain validation in the scenario with a large number of clients. The total number of participating clients is 50, and the number of sampled clients in one communication round matches the number of training domains.

ii) In comparison to state-of-the-art methods, the proposed FEDIIR achieves superior performance on the RotatedMNIST, PACS, and OfficeHome datasets. Remarkably, our method employs a simple regularization technique, making it more accessible and easier to implement in practice compared to other approaches.

Analysis Our proposed method, FEDIIR, not only consistently outperforms FEDAVG, but also achieves competitive results when compared to methods specifically designed for the scenario with a small number of clients. The reason is that FEDIIR utilizes inter-client gradient alignment to effectively encourage the model to learn invariant relationships implicitly. In short, FEDIIR demonstrates excellent OOD generalization performance in the scenario with a small number of clients.

Results on a large number of clients scenario. To answer the **Q2**, we study the OOD generalization performance of the proposed method on four datasets, considering the increase in the total number of participating clients. In all experiments, the number of sampled clients during one communication round remains fixed, corresponding to the number of training domains: 5 for RotatedMNIST and 3 for VLCS, PACS, and OfficeHome. Figure 2 visually displays the experimental results, with the horizontal axis denoting the total number of participating clients and the vertical axis representing the average test accuracy. Additionally, Table 2 presents full results with 50 participating clients, accompanied by detailed domain-specific outcomes in Appendix F.4. Based on the results of our experiments, we make the following conclusions:

i) The effectiveness of all baselines noticeably decreases as the total number of participating clients increases. However, our method FEDIIR produces comparable performance for scenarios with both a small and large number of clients;

ii) Our proposed FEDIIR significantly outperforms all baselines even when the number of participating clients reaches 50. For instance, FEDIIR achieves a substantial average improvement of 14% and 8% compared to the baselines on VLCS and OfficeHome datasets, respectively.

Analysis As the participation rate of clients decreases in one communication round, explicit methods inevitably incur performance degradation. The fact that these approaches explicitly learn invariant relationships necessitates relatively reliable participating clients and is impractical for a large number of clients. In contrast, our method implicitly learns invariant relationships and converges quickly in scenarios with a large number of clients population. In summary, the proposed FEDIIR does indeed perform excellently for OOD generalization performance in the scenario where the clients are massively distributed with limited communication.

5.3. Visualization of the Convergence Process

To visualize the difference in convergence speed between methods, we examine their average test accuracy versus communication round. We consider a scenario with 50 participating clients, where the number of sampled clients in one communication round coincides with the number of training domains. Figure 3 presents the results for the VLCS and PACS datasets, and Figure 5 displays the results for RotatedMNIST and OfficeHome. The experimental findings

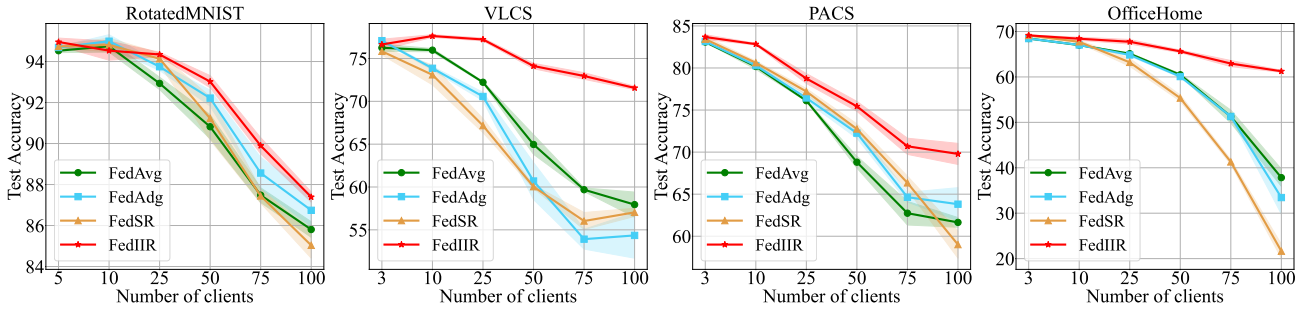

Figure 2. Average test accuracy (%) versus the total number of participating clients, with the number of sampled clients in one communication round matches the number of training domains.

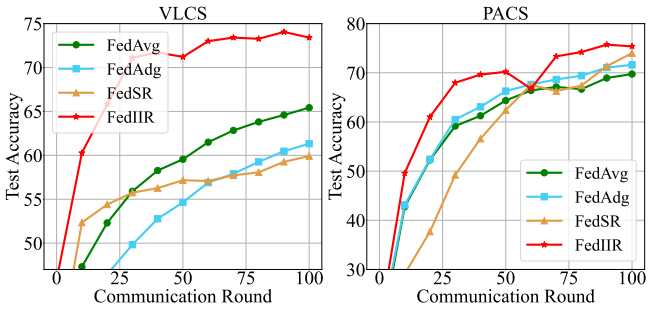

Figure 3. Average test accuracy (%) versus communication round on VLCS (**left**) and PACS (**right**) dataset with 50 participating clients, where the number of sampled clients in one communication round matches the number of training domains.

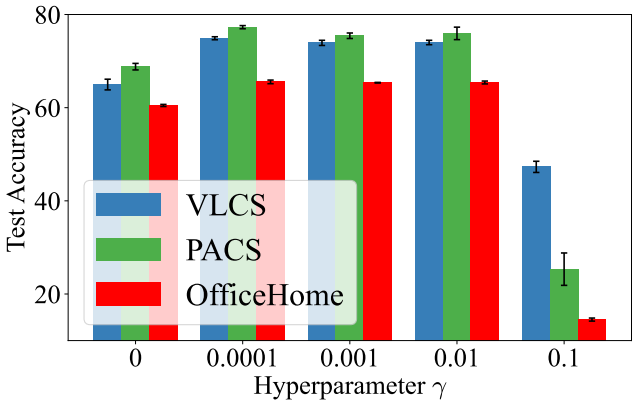

Figure 4. Average test accuracy (%) for various values of the hyperparameter γ in FEDIIR, with 50 participating clients and 3 sampled clients in one communication.

demonstrate that FEDIIR exhibits faster convergence compared to the baselines, particularly during the first 50 rounds of communication. As demonstrated in Section 4.2, FEDIIR has good convergence speed when appropriate learning step-sizes are used, whereas the alternative methods do not provide any theoretical insight into the convergence speed. This property is crucial in federated learning scenarios, where clients are massively distributed and communication becomes a primary bottleneck.

5.4. Sensitivity of γ

Here, we investigate the sensitivity of FEDIIR w.r.t. the hyperparameter γ. With 50 participating clients and 3 sampled clients in one communication round, we examine the performance of FEDIIR on VLCS, PACS, and OfficeHome datasets for various values of hyperparameter γ, where $\gamma \in \{0, 0.0001, 0.001, 0.01, 0.1\}$. The experimental results are presented in Figure 4. We observe that FEDIIR outperforms FEDAVG when $\gamma \in \{0.0001, 0.001, 0.01\}$, but is worse than FEDAVG when $\gamma = 0.1$. This can be attributed to the fact that a smaller local step-size η_l is necessary for the convergence of FEDIIR with a higher γ (Section 4.2). In other words, achieving optimal performance of FEDIIR requires striking a delicate balance between the hyperparameter γ and the local step-size η_l, highlighting the importance of careful consideration in tuning.

6. Conclusion

In this paper, we study OOD generalization of federated learning with a novel perspective that implicitly learns invariant relationships from the parameter. To this end, we propose FEDIIR, a simple federated learning method that performs better in OOD generalization by aligning inter-client gradient. This method adheres entirely to the federated principles of privacy-preserving and limited communication. The theoretical results demonstrate that FEDIIR is expected to generalize to non-participating clients included in the affine combination of participating clients and also has a good convergence speed under limited communication resources. Extensive experiments show that FEDIIR provides better OOD generalization performance than the relevant baseline, especially in the scenario where clients are massively distributed with limited communication.

Acknowledgements

This work is supported by the National Key R&D Program of China (No. 2022YFB3103700, No. 2022YFB3103702) and the National Natural Science Foundation of China (No. 62206108, No. 61976102, and No. U19A2065).

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.

Albuquerque, I., Monteiro, J., Darvishi, M., Falk, T. H., and Mitliagkas, I. Generalizing to unseen domains via distribution matching. arXiv preprint arXiv:1911.00804, 2019.

Arjovsky, M. Out of distribution generalization in machine learning. PhD thesis, New York University, 2020.

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

Bai, R., Bagchi, S., and Inouye, D. I. Benchmarking algorithms for domain generalization in federated learning, 2023. URL https://openreview.net/forum?id=IsCg7qoy8i9.

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.

Bellot, A. and van der Schaar, M. Accounting for unobserved confounding in domain generalization. arXiv preprint arXiv:2007.10653, 2020.

Blanchard, G., Deshmukh, A. A., Dogan, Ü., Lee, G., and Scott, C. Domain generalization by marginal transfer learning. The Journal of Machine Learning Research, 22 (1):46–100, 2021.

Chen, H.-Y. and Chao, W.-L. Fed{be}: Making bayesian model ensemble applicable to federated learning. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=dgtpE6gKjHn.

Crane, R. and Roosta, F. Dingo: Distributed newton-type method for gradient-norm optimization. Advances in Neural Information Processing Systems, 32, 2019.

Deng, Y., Kamani, M. M., and Mahdavi, M. Distributionally robust federated averaging. Advances in neural information processing systems, 33:15111–15122, 2020.

Elgabli, A., Issaid, C. B., Bedi, A. S., Rajawat, K., Bennis, M., and Aggarwal, V. Fednew: A communication-efficient and privacy-preserving newton-type method for federated learning. In International Conference on Machine Learning, pp. 5861–5877. PMLR, 2022.

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.

Fang, C., Xu, Y., and Rockmore, D. N. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In Proceedings of the IEEE International Conference on Computer Vision, pp. 1657–1664, 2013.

Ghifary, M., Kleijn, W. B., Zhang, M., and Balduzzi, D. Domain generalization for object recognition with multi-task autoencoders. In Proceedings of the IEEE international conference on computer vision, pp. 2551–2559, 2015.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Gupta, S., Ahuja, K., Havaei, M., Chatterjee, N., and Bengio, Y. Fl games: A federated learning framework for distribution shifts. arXiv preprint arXiv:2205.11101, 2022.

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

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In Joint European conference on machine learning and knowledge discovery in databases, pp. 795–811. Springer, 2016.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In International Conference on Machine Learning, pp. 5132–5143. PMLR, 2020.

Karimireddy, S. P., Jaggi, M., Kale, S., Mohri, M., Reddi, S., Stich, S. U., and Suresh, A. T. Breaking the centralized barrier for cross-device federated learning. Advances in Neural Information Processing Systems, 34:28663–28676, 2021.

Koyama, M. and Yamaguchi, S. When is invariance useful in an out-of-distribution generalization problem? arXiv preprint arXiv:2008.01883, 2020.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In International Conference on Machine Learning, pp. 5815–5826. PMLR, 2021.

Lee, G., Jeong, M., Shin, Y., Bae, S., and Yun, S.-Y. Preservation of the global knowledge by not-true distillation in federated learning. In Advances in Neural Information Processing Systems, 2022.

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, Q., Diao, Y., Chen, Q., and He, B. Federated learning on non-iid data silos: An experimental study. In 2022 IEEE 38th International Conference on Data Engineering (ICDE), pp. 965–978. IEEE, 2022.

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

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, 2020b.

Lin, T., Kong, L., Stich, S. U., and Jaggi, M. Ensemble distillation for robust model fusion in federated learning. Advances in Neural Information Processing Systems, 33: 2351–2363, 2020.

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

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 Proceedings of the 39th International Conference on Machine Learning, pp. 14527–14541, 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.

Mendieta, M., Yang, T., Wang, P., Lee, M., Ding, Z., and Chen, C. Local learning matters: Rethinking data heterogeneity in federated learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8397–8406, 2022.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In International Conference on Machine Learning, pp. 4615–4625. PMLR, 2019.

Nagarajan, V., Andreassen, A., and Neyshabur, B. Understanding the failure modes of out-of-distribution generalization. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=fSTD6NFIW_b.

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.

Parascandolo, G., Neitz, A., ORVIETO, A., Gresele, L., and Schölkopf, B. Learning explanations that are hard to vary. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=hb1sDDSLbV.

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), 78(5):947–1012, 2016.

Qu, Z., Li, X., Duan, R., Liu, Y., Tang, B., and Lu, Z. Generalized federated learning via sharpness aware minimization. In International Conference on Machine Learning, pp. 18250–18280. PMLR, 2022.

Rame, A., Dancette, C., and Cord, M. Fishr: Invariant gradient variances for out-of-distribution generalization. In International Conference on Machine Learning, pp. 18347–18377. PMLR, 2022.

Reisizadeh, A., Farnia, F., Pedarsani, R., and Jadbabaie, A. Robust federated learning: The case of affine distribution shifts. Advances in Neural Information Processing Systems, 33:21554–21565, 2020.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Invariant models for causal transfer learning. The Journal of Machine Learning Research, 19(1):1309–1342, 2018.

Roosta, F., Liu, Y., Xu, P., and Mahoney, M. W. Newton-mr: Inexact newton method with minimum residual sub-problem solver. EURO Journal on Computational Optimization, 10:100035, 2022.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In International Conference on Learning Representations, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. Towards out-of-distribution generalization: A survey. arXiv preprint arXiv:2108.13624, 2021.

Tenison, I., Francis, S., and Rish, I. Gradient masked federated optimization. arXiv preprint arXiv:2104.10322, 2021.

Vapnik, V. Principles of risk minimization for learning theory. Advances in neural information processing systems, 4, 1991.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5018–5027, 2017.

Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., et al. A field guide to federated optimization. arXiv preprint arXiv:2107.06917, 2021.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., and Yu, P. Generalizing to unseen domains: A survey on domain generalization. IEEE Transactions on Knowledge and Data Engineering, 2022.

Xie, C., Ye, H., Chen, F., Liu, Y., Sun, R., and Li, Z. Risk variance penalization. arXiv preprint arXiv:2006.07544, 2020.

Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T., and Yu, H. Federated learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 13(3):1–207, 2019.

Yuan, H., Morningstar, W. R., Ning, L., and Singhal, K. What do we mean by generalization in federated learning? In International Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=VimqQq-i_Q.

Zhang, L., Lei, X., Shi, Y., Huang, H., and Chen, C. Federated learning with domain generalization. arXiv preprint arXiv:2111.10487, 2021.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. Federated learning with non-iid data. arXiv preprint arXiv:1806.00582, 2018.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

Zhu, H., Xu, J., Liu, S., and Jin, Y. Federated learning on non-iid data: A survey. Neurocomputing, 465:371–390, 2021a.

Zhu, Z., Hong, J., and Zhou, J. Data-free knowledge distillation for heterogeneous federated learning. In International Conference on Machine Learning, pp. 12878–12889. PMLR, 2021b.

A. Related Work

Federated Learning Federated learning is a popular distributed learning paradigm that enables collaborative training of a global machine learning model over multiple clients involving data silos in a privacy-preserving manner(Yang et al., 2019; Kairouz et al., 2021). FEDAVG(McMahan et al., 2017) defines a standard optimization method in which clients perform multiple epochs of stochastic gradient descent (SGD) on their local data. Numerous studies have shown that this simple method inevitably produces performance degradation when there are distribution shifts across clients(Zhao et al., 2018; Zhu et al., 2021a; Li et al., 2022). Many existing works mainly address distribution shifts from the following two perspectives. The first perspective focuses on stabilizing local model updates: FEDPROX(Li et al., 2020b) adds a proximal term to the local subproblem for limiting the distance between the local and global models. SCAFFOLD(Karimireddy et al., 2020) uses an additional control variable to correct for client-drift in local updates. FEDNTD(Lee et al., 2022) performs local-side distillation only for not-true classes to prevent forgetting global knowledge corresponding to regions outside the local distribution. FEDALIGN(Mendieta et al., 2022) regularizes the Lipschitz constants of the final block in a network with respect to its representations, thereby improving local learning generality during training. Another perspective aims to improve the efficacy of model aggregation: FEDDF(Lin et al., 2020) uses ensemble distillation techniques to aggregate knowledge from heterogeneous local models on a proxy dataset. FEDBE(Chen & Chao, 2021) generates a series of models from the Bayesian perspective using local models and combining them via Bayesian model Ensemble. FEDGEN(Zhu et al., 2021b) learns a lightweight generator to ensemble knowledge of local models and regulates local training using the learned knowledge. However, these studies focus on the distribution shifts across participating clients and ignore the generalization to non-participating clients who are also affected by distribution shifts. Unlike these techniques, our methods attempt to learn latent invariant relationships that work equally well for all clients, expecting to achieve OOD generalization.

Out-of-distribution Generalization OOD generalization aims to learn a model that can be generalized to unseen environments (corresponding to clients) under distribution shifts, which is critical in practice for model deployment in the wild(Shen et al., 2021; Zhou et al., 2022; Wang et al., 2022). Recent research has attempted to build an invariant predictor by learning invariant causal relationships, working equally well over OOD. IRM(Arjovsky et al., 2019) tries to find a representation such that the optimal linear classifier, on top of that representation, is the same for all environments. IRM GAME(Ahuja et al., 2020) reformulates IRM as finding the Nash equilibrium of an ensemble game among several environments, holding for a large class of nonlinear classifiers. GROUPDRO(Sagawa et al., 2020) conducte an analysis of group DRO in overparameterized neural networks, demonstrating the importance of regularization for worst-case group generalization. REX(Krueger et al., 2021) learns invariant predictor by reducing differences in risk across environments, which essentially minimizes the variance of training risks(Xie et al., 2020). MATCHDG(Mahajan et al., 2021) constructs a representation using contrastive learning, where representations of the same object across environments are invariant. IGA(Koyama & Yamaguchi, 2020) builds the invariant predictor based on information theory, which looks for the mutual-information maximizing feature amongst the invariant features. These explicit methods require a centralized setting where data or representation is shared across clients, violating the federated principles of privacy-preserving. In contrast, our methods do not need to communicate anything other than the model parameter, which can offer better protection for client privacy than the above methods.

Generalization in Federated Learning Following the terminology introduced in Yuan et al. (2022), we aim to bridge the participation gap introduced by non-participating (unseen) clients. AFL(Mohri et al., 2019) and DRFA(Deng et al., 2020) suggests that the global modal is optimized for any target distribution formed by a mixture of client distributions instead of a specific target distribution. FLRA(Reisizadeh et al., 2020) attempts to find a global model that minimizes the total loss induced by the worst-case local affine transformation. FEDGMA(Tenison et al., 2021) assigns higher importance to the consistent components of the gradient for promoting the global model converges to a consistent minimum across clients. FEDSAM(Qu et al., 2022) focuses on local learning generality, where each local client trains the local model with the same perturbation bound. These techniques can, at best, generalize to non-participating clients included in the convex hull of participating clients but fail to extrapolate well, i.e., generalize to non-participating clients that fall outside of that convex hull. To achieve a more ambitious OOD generalization, FEDADG(Zhang et al., 2021) employs the federated adversarial learning approach to align the distributions among different clients for learning universal features. FEDSR(Nguyen et al., 2022) enforces regularization on representation and conditional mutual information to encourage the model for learning only the necessary information, which helps to ignore spurious relationships. FL GAME(Gupta et al., 2022) designs a game-theoretic framework that allows parallel computation for learning causal features that are invariant across clients. Typically, these explicit methods require clients to participate in each round, which violates the federated principle of limited communication. In contrast, our approach involves only regularization in the parameter space, which can be efficiently optimized in scenarios where clients are massively distributed with limited communication.

Comparison with DFL(Luo et al., 2022) in "invariance" terminology We and DFL both understand the term "invariant" in the context of causal learning, but our definitions differ slightly. In their work, a feature z is deemed "invariant" if $\mathbb{P}_c(Z=z)=\mathbb{P}_{c'}(Z=z)$, where c and c' denote different clients. This definition responds to the problem of attribute skewness, where the distribution of some features shifts across clients. On the other hand, we define z as an invariant feature if $\mathbb{P}_c(Y|Z=z)=\mathbb{P}_{c'}(Y|Z=z)$, where Y is the target variable. This definition supports OOD generalization because the invariant features constitute the direct cause of the target. Based on different assumptions, we approach the problem differently. DFL addresses the attribute skewness issue by separating features in the dataset that do not have skew and successfully mitigating negative transfer. However, it might also take on spurious features that are not skewed, which may lead to a failure to generalize to OOD. In contrast, we use inter-client gradient alignment to encourage the model to learn invariant relationships and achieve OOD generalization.

B. More Details on Section 3.1

In this section, we supplement what was omitted in Section 3.1, including the formal definition of invariant predictor and the proof of Theorem 2.

The formal definition of the invariant predictor, derived from Arjovsky et al. (2019), is presented as follows.

Definition 6 (Invariant Predictor). We say that a representation Φ elicits an *invariant predictor* $f=w\circ\Phi:\mathcal{X}\to\mathcal{Y}$ across clients $\mathcal{C}_{\text{all}}$ if there is a classifier w simultaneously optimal for all clients, that is, $w\in\arg\min_w \mathcal{R}_c(w\circ\Phi)$ for all $c\in\mathcal{C}_{\text{all}}$.

If features have different joint distributions with the targets across clients, a fixed classifier on top of them will not be optimal in all clients. This suggests that the invariant predictor w.r.t. Definition 6 uses only invariant relationships for predicting the target variable, which captures the latent invariant relationship between the input variable X and the target variable Y.

We now prove Theorem 2, thus relating the prediction disagreement to the invariant predictor.

Theorem 2*. *Given the collection $\mathcal{C}_{\text{all}}$ of clients, for the representation $\Phi(\cdot)$, let the optimal classifier for client c be $w_c^*(\cdot)$. If $\mathcal{I}(\Phi,\mathcal{C}_{\text{all}})=0$, then $f=w\circ\Phi$ is an invariant predictor, where $w(\cdot)=w_c^*(\cdot)$ for any $c\in\mathcal{C}_{\text{all}}$.*

Proof. Notice that $w_c^*(z)=\mathbb{E}[Y^c|\Phi(X^c)=z]$ holds for all $c\in\mathcal{C}_{\text{all}}$. If $\mathcal{I}(\Phi,\mathcal{C}_{\text{all}})=0$, then $\Phi(\cdot)$ satisfies following invariant constraint for any pair $(c,c')\in\mathcal{C}_{\text{all}}^2$ and all $z\in\cap_{c\in\mathcal{C}_{\text{all}}}\text{supp}(\mathbb{P}(\Phi(X^c)))$:

$$\mathop{\mathbb{E}}_{X^c,Y^c}[Y^c|\Phi(X^c)=z]=\mathop{\mathbb{E}}_{X^{c'},Y^{c'}}[Y^{c'}|\Phi(X^{c'})=z].$$

Thus, $w(z):=w_c^*(z)=\mathbb{E}[Y^c|\Phi(X^c)=z]$ for any c such that $z\in\text{supp}(\mathbb{P}(\Phi(X^c)))$ is well defined, which indicates that $w(\cdot)$ is simultaneously optimal for all clients, i.e., $f=w\circ\Phi$ is an invariant predictor. □

C. More Discussion on Section 3.3

This section discusses the additional computation introduced by FEDIIR and how to stabilize the estimate of the global gradient in FEDIIR.

FEDIIR introduces extra computation because it needs to compute the average full-batch gradient of sampled clients in each round of communication. However, we believe that the additional computation is manageable for most practical federated settings, especially in scenarios with a large number of clients. Firstly, the number of sampled clients in one round of communication is typically small, and the number of samples held by each client is also small. Additionally, FEDIIR only aligns the gradients of the classifier, which generally consists of the last few layers of the model, and thus does not require significant computational overhead. Overall, we argue that the additional computation introduced by FEDIIR is reasonable.

As the number of sampled clients in one round of communication is usually small, relying solely on the average full-batch gradient of sampled clients to estimate the global gradient can result in significant errors. To mitigate this issue, we adopt the exponential moving average (ema) technique to stabilize the estimate of the global gradient. Specifically, in the t-th round of communication, we align $\widetilde{g}_\omega^t=\upsilon\widetilde{g}_\omega^{t-1}+(1-\upsilon)\overline{g}_\omega$ instead of $\overline{g}_\omega$, where υ is a hyperparameter to control the update speed. FEDIIR using the ema technique enables a smoother estimation of the global gradient throughout training. It is worth noting that similar techniques have been widely employed in OOD generalization(Rame et al., 2022; Blanchard et al., 2021).

D. Generalization Analysis

In this section, we restate Theorem 3 and provide proof.

Theorem 3*. *Given the collection $\mathcal{C}_{\text{par}}$ of clients, let's assume that $\ell(\cdot,\cdot) \leq M$. Then for all $f = w \circ \Phi \in \mathcal{F}$, we have the following risk bound for the affine combination of participating clients:*

$$\sup_{\lambda \in \Lambda_\nu} \mathcal{R}_\lambda(f) \leq \mathcal{R}(f) + \widetilde{M}\mathcal{I}(\Phi, \mathcal{C}_{\text{par}}) + \widetilde{M} \sup_{(c,c') \in \mathcal{C}_{\text{par}}^2} \rho(\mathbb{P}_c(X), \mathbb{P}_{c'}(X)),$$

where $\widetilde{M} = (1 + |\mathcal{C}_{\text{par}}|\nu)M$ is monotonic in ν, and $\rho(\mathbb{P}_c(X), \mathbb{P}_{c'}(X)) = \sup_X |\mathbb{P}_c(X) - \mathbb{P}_{c'}(X)|$ is the total variation distance.

Proof. Taking a cue from Bellot & van der Schaar (2020), we first rewrite the affine combination of risk:

$$\begin{aligned}
\sup_{\lambda \in \Lambda_\nu} \mathcal{R}_\lambda(f) &= \sup_{\lambda \in \Lambda_\nu} \sum_{c \in \mathcal{C}_{\text{par}}} \lambda_c \mathcal{R}_c(f) \\
&\overset{(a)}{\leq} (1 + |\mathcal{C}_{\text{par}}|\nu) \sup_{c \in \mathcal{C}_{\text{par}}} \mathcal{R}_c(f) - \nu \sum_{c \in \mathcal{C}_{\text{par}}} \mathcal{R}_c(f) \\
&= \frac{1}{|\mathcal{C}_{\text{par}}|} \sum_{c \in \mathcal{C}_{\text{par}}} \mathcal{R}_c(f) + (1 + |\mathcal{C}_{\text{par}}|\nu) \sup_{c \in \mathcal{C}_{\text{par}}} \mathcal{R}_c(f) - (1 + |\mathcal{C}_{\text{par}}|\nu) \frac{1}{|\mathcal{C}_{\text{par}}|} \sum_{c \in \mathcal{C}_{\text{par}}} \mathcal{R}_c(f) \\
&\overset{(b)}{=} \mathcal{R}(f) + (1 + |\mathcal{C}_{\text{par}}|\nu) \left(\sup_{c \in \mathcal{C}_{\text{par}}} \mathcal{R}_c(f) - \mathcal{R}(f) \right) \\
&\leq \mathcal{R}(f) + (1 + |\mathcal{C}_{\text{par}}|\nu) \sup_{(c,c') \in \mathcal{C}_{\text{par}}^2} \big(\mathcal{R}_c(f) - \mathcal{R}_{c'}(f) \big),
\end{aligned}$$

where (a) comes from the fact that $\sup_{c \in \mathcal{C}_{\text{par}}} \mathcal{R}_c(f) - \mathcal{R}_c(f) \geq 0$; (b) is from the definition of global expected risk $\mathcal{R}(f)$. We will bound the second term of the above inequality below. For any $(c, c') \in \mathcal{C}_{\text{par}}^2$, there exists

$$\begin{aligned}
\mathcal{R}_c(f) &= \mathop{\mathbb{E}}_{x \sim \mathbb{P}_c(X)} \underbrace{\left[\mathop{\mathbb{E}}_{y \sim \mathbb{P}_c(Y|\Phi(X)=\Phi(x))} [\ell(w(\Phi(x)), y)] \right]}_{g(x)} \\
&= \underbrace{\mathop{\mathbb{E}}_{x \sim \mathbb{P}_{c'}(X)} \left[\mathop{\mathbb{E}}_{y \sim \mathbb{P}_c(Y|\Phi(X)=\Phi(x))} [\ell(w(\Phi(x)), y)] \right]}_{\mathcal{A}_1} + \underbrace{\mathop{\mathbb{E}}_{x \sim \mathbb{P}_c(X)} [g(x)] - \mathop{\mathbb{E}}_{x \sim \mathbb{P}_{c'}(X)} [g(x)]}_{\mathcal{A}_2}.
\end{aligned} \tag{1}$$

For the first term $\mathcal{A}_1$, we have

$$\begin{aligned}
\mathcal{A}_1 = &\mathop{\mathbb{E}}_{x \sim \mathbb{P}_{c'}(X)} \Bigg[\mathop{\mathbb{E}}_{y \sim \mathbb{P}_c(Y|\Phi(X)=\Phi(x))} [\ell(w(\Phi(x)), y)] + \mathop{\mathbb{E}}_{y \sim \mathbb{P}_{c'}(Y|\Phi(X)=\Phi(x))} [\ell(w(\Phi(x)), y)] \\
&\qquad - \mathop{\mathbb{E}}_{y \sim \mathbb{P}_{c'}(Y|\Phi(X)=\Phi(x))} [\ell(w(\Phi(x)), y)] \Bigg] \\
\leq &\mathop{\mathbb{E}}_{x \sim \mathbb{P}_{c'}(X)} \left[\mathop{\mathbb{E}}_{y \sim \mathbb{P}_{c'}(Y|\Phi(X)=\Phi(x))} [\ell(w(\Phi(x)), y)] \right] \\
&+ \mathop{\mathbb{E}}_{x \sim \mathbb{P}_{c'}(X)} \left[\mathop{\mathbb{E}}_{y \sim \mathbb{P}_c(Y|\Phi(X)=\Phi(x))} [\ell(w(\Phi(x)), y)] - \mathop{\mathbb{E}}_{y \sim \mathbb{P}_{c'}(Y|\Phi(X)=\Phi(x))} [\ell(w(\Phi(x)), y)] \right] \\
\overset{(a)}{\leq} &\mathcal{R}_{c'}(f) + M \mathop{\mathbb{E}}_{x \sim \mathbb{P}_{c'}(X)} [\rho(\mathbb{P}_c(Y|\Phi(X) = \Phi(x)), \mathbb{P}_{c'}(Y|\Phi(X) = \Phi(x)))] \\
\overset{(b)}{\leq} &\mathcal{R}_{c'}(f) + M \sup_{z \in \mathrm{U}(\Phi, \mathcal{C})} |w_c^*(z) - w_{c'}^*(z)|,
\end{aligned}$$

where (a) is from the condition $\ell(\cdot,\cdot) \leq M$, (b) comes from the setting of Y is a binary variable, and $\mathrm{U}(\Phi,\mathcal{C}) = \cup_{c\in\mathcal{C}}\mathrm{supp}(\mathbb{P}(\Phi(X^c)))$ is the union of the supports.

Since $\ell(\cdot,\cdot) \leq M$, $|g(x)| \leq M$ holds for all x. For the second term $\mathcal{A}_2$, the following inequality exist

$$\mathcal{A}_2 \leq M\rho(\mathbb{P}_c(X), \mathbb{P}_{c'}(X)).$$

Plugging back the bounds on $\mathcal{A}_1$ and $\mathcal{A}_2$, obtaining

$$\begin{aligned}\sup_{(c,c')\in\mathcal{C}^2_{\mathrm{par}}} \big(\mathcal{R}_c(f) - \mathcal{R}_{c'}(f)\big) &\leq M \sup_{z\in\mathrm{U}(\Phi,\mathcal{C})} \sup_{(c,c')\in\mathcal{C}^2_{\mathrm{par}}} |w^*_c(z) - w^*_{c'}(z)| + M \sup_{(c,c')\in\mathcal{C}^2_{\mathrm{par}}} \rho(\mathbb{P}_c(X), \mathbb{P}_{c'}(X)) \\ &= M\mathcal{I}(\Phi, \mathcal{C}_{\mathrm{par}}) + M \sup_{(c,c')\in\mathcal{C}^2_{\mathrm{par}}} \rho(\mathbb{P}_c(X), \mathbb{P}_{c'}(X)).\end{aligned}$$

Let $\widetilde{M} = (1 + |\mathcal{C}_{\mathrm{par}}|\nu)M$ to finish the proof. □

E. Convergence Analysis of FEDIIR

In this section, we add more details about the convergence analysis of FEDIIR. We first examine the assumptions and definitions, then introduce some technical lemmas, and finally prove the convergence result of FEDIIR.

E.1. Assumptions and Definitions

We reiterate the assumptions required for the proof and explain some of their implications.

Assumption B* (Smoothness)**.** For all clients c, we assume that $R_c(\omega)$ is **L-smoothness**, i.e., there exist a constant $L \geq 0$ such that

$$\|\nabla R_c(\omega) - \nabla R_c(\omega')\| \leq L\|\omega - \omega'\| \qquad (\forall \omega, \omega');$$

we also assume that $R_c(\omega)$ is **Moral-smoothness**, i.e., there exist a constant $L \geq 0$ such that

$$\|\nabla^2 R_c(\omega)\nabla R_c(\omega) - \nabla^2 R_c(\omega')\nabla R_c(\omega')\| \leq L\|\omega - \omega'\| \qquad (\forall \omega, \omega').$$

Further, if $R_c(\omega)$ is twice-differentiable, L-smoothness implies that$\|\nabla^2 R_c(\omega)\| \leq L$.

Assumption C* (Bounded Statistical Heterogeneity)**.** For all clients c, we assume that when there is no perturbation, the variance of the local gradient w.r.t. the global gradient is bounded by G, i.e., there exists a constant $G \geq 0$ such that

$$\|\nabla R_c(\omega) - \nabla R(\omega)\|^2 \leq G^2 \qquad (\forall \omega).$$

Assumption D* (Bounded Intra-client Variance)**.** For all clients c, we assume that $\nabla R_c(\omega;\zeta)$, $\nabla^2 R_c(\omega;\zeta)$, and $\nabla^2 R_c(\omega;\zeta)\nabla R_c(\omega;\zeta)$ are unbiased estimates of $\nabla R_c(\omega)$, $\nabla^2 R_c(\omega)$, and $\nabla^2 R_c(\omega)\nabla R_c(\omega)$, respectively, with variances bounded by σ^2, i.e., there exist a constant $\sigma \geq 0$ such that

$$\|\nabla R_c(\omega;\zeta) - \nabla R_c(\omega)\|^2 \leq \sigma^2$$

$$\|\nabla^2 R_c(\omega;\zeta) - \nabla^2 R_c(\omega)\|^2 \leq \sigma^2$$

$$\|\nabla^2 R_c(\omega;\zeta)\nabla R_c(\omega;\zeta) - \nabla^2 R_c(\omega)\nabla R_c(\omega)\|^2 \leq \sigma^2$$

Assumption E (μ-PL Inequality)**.** We say that $R(\omega)$ satisfies the μ-PL if the following holds for $\mu > 0$:

$$\|\nabla R(\omega)\|^2 \geq 2\mu(R(\omega) - R^*),$$

where $R^* = \min R(\omega)$.

Note that the PL inequality is much weaker than the standard notion of strong convexity and can even satisfy some non-convex functions(Karimi et al., 2016).

Now, let us rewrite the FEDIIR updates using notation convenient for analysis. In t-th communication round, server sample clients $\mathcal{C}$ with $|\mathcal{C}| = C$. We define the average gradient $\bar{g}^t$ across the sampled clients as:

$$\bar{g}^t = \frac{1}{C}\sum_{c\in\mathcal{C}} \nabla R_c(\omega^{t-1}).$$

Client initialize local model parameter $\omega_{c,0}^t = \omega^{t-1}$ and perform K local updates:

$$\omega_{c,k}^t = \omega_{c,k-1}^t - \eta_l(g_c(\omega_{c,k-1}^t) + \gamma H_c(\omega_{c,k-1}^t)(g_c(\omega_{c,k-1}^t) - \bar{g}^t)),$$

where $g_c(\omega_{c,k-1}^t) = \nabla R_c(\omega_{c,k-1}^t;\zeta)$ is stochastic gradient and $H_c(\omega_{c,k-1}^t) = \nabla^2 R_c(\omega_{c,k-1}^t;\zeta)$ is stochastic Hessian.

Server aggregates the new global model parameter as:

$$\omega^t = \omega^{t-1} + \eta_g \frac{1}{C}\sum_{c\in\mathcal{C}}(\omega_{c,K}^t - \omega_{c,0}^t).$$

We introduce below some additional definitions to better describe the various errors we need to track. We define the effective step-size to be

$$\tilde{\eta} := K\eta_g\eta_l.$$

We define the server update error as how much the server has moved from its starting point:

$$\begin{aligned}\|\omega^t - \omega^{t-1}\|^2 :&= \| -\eta_g\eta_l \frac{1}{C}\sum_{c\in\mathcal{C}}\sum_{k\in[K]} \left(g_c(\omega_{c,k-1}^t) + \gamma H_c(\omega_{c,k-1}^t)(g_c(\omega_{c,k-1}^t) - \bar{g}^t)\right)\|^2 \\ &= \|\frac{\tilde{\eta}}{CK}\sum_{c,k}\left((I + \gamma H_c(\omega_{c,k-1}^t))g_c(\omega_{c,k-1}^t) - \gamma H_c(\omega_{c,k-1}^t)\bar{g}^t\right)\|^2.\end{aligned}$$

We define the client drift error to be how much the clients have moved from their starting point:

$$\begin{aligned}\mathcal{E}^t :&= \frac{1}{CK}\sum_{c\in\mathcal{C}}\sum_{k\in[K]} \mathbb{E}\|\omega_{c,k-1}^t - \omega^{t-1}\|^2 \\ &= \frac{1}{CK}\sum_{c,k}\mathbb{E}\|\omega_{c,k-1}^t - \omega^{t-1}\|^2.\end{aligned}$$

E.2. Technical Lemmas

We now present some technical lemmas involved in later proofs, where the proofs of Lemma 7 and Lemma 8 can be found in Karimireddy et al. (2020).

Lemma 7 (Relaxed triangle inequality)**.** *Let $\{v_1, v_2, \ldots, v_\tau\}$ be τ vectors in $\mathbb{R}^d$. Then the following inequalities are true for the squared L_2-norm:*

1. $\|v_i + v_j\|^2 \le (1+a)\|v_i\|^2 + (1+\frac{1}{a})\|v_j\|^2 \quad$ *for $a > 0$;*
2. $\|\sum_{i=1}^\tau v_i\|^2 \le \tau\sum_{i=1}^\tau \|v_i\|^2$.

Lemma 8 (Separating mean and variance)**.** *Let $\{\Xi_1, \Xi_2, \ldots, \Xi_\tau\}$ be τ random variables in $\mathbb{R}^d$ which are not necessarily independent. First suppose that their mean is $\mathbb{E}[\Xi_i] = \xi_i$ and variance is bounded as $\mathbb{E}[\|\Xi_i - \xi_i\|^2] \le \sigma^2$. Then the following holds*

$$\mathbb{E}[\|\sum_{i=1}^\tau \Xi\|^2] \le \|\sum_{i=1}^\tau \xi\|^2 + \tau^2\sigma^2$$

Now instead suppose that their conditional mean is $\mathbb{E}[\Xi_i|\Xi_{i-1},\dots,\Xi_1]=\xi_i$, i.e., the variables $\{\Xi_i-\xi_i\}$ form a martingale difference sequence, and the variance is bounded by $\mathbb{E}[\|\Xi_i-\xi_i\|^2]\le\sigma^2$ as before. Then we can show the tighter bound

$$\mathbb{E}[\|\sum_{i=1}^{\tau}\Xi\|^2]\le 2\|\sum_{i=1}^{\tau}\xi\|^2+2\tau\sigma^2$$

Lemma 9 (Separating mean and variance for FEDIIR)**.** *Suppose that $\{R_c\}$ satisfies Assumption D, we then have the following inequality*

$$\mathbb{E}\|\frac{1}{CK}\sum_{c,k}\left(I+\gamma H_c(\omega_{c,k-1}^t)\right)g_c(\omega_{c,k-1}^t)\|^2\le 2\mathbb{E}\|\frac{1}{CK}\sum_{c,k}\left(I+\gamma\nabla^2 R_c(\omega_{c,k-1}^t)\right)\nabla R_c(\omega_{c,k-1}^t)\|^2+\frac{8\gamma^2\sigma^2}{CK}.$$

Proof. For any c,k, we have upper bounds on the variance:

$$\begin{aligned}
&\mathbb{E}\|\left(I+\gamma H_c(\omega_{c,k-1}^t)\right)g_c(\omega_{c,k-1}^t)-\left(I+\gamma\nabla^2 R_c(\omega_{c,k-1}^t)\right)\nabla R_c(\omega_{c,k-1}^t)\|^2\\
=&\mathbb{E}\|g_c(\omega_{c,k-1}^t)-\nabla R_c(\omega_{c,k-1}^t)+\gamma H_c(\omega_{c,k-1}^t)g_c(\omega_{c,k-1}^t)-\gamma\nabla^2 R_c(\omega_{c,k-1}^t)\nabla R_c(\omega_{c,k-1}^t)\|^2\\
\overset{(a)}{\le}&2\mathbb{E}\|g_c(\omega_{c,k-1}^t)-\nabla R_c(\omega_{c,k-1}^t)\|^2+2\gamma^2\mathbb{E}\|H_c(\omega_{c,k-1}^t)g_c(\omega_{c,k-1}^t)-\nabla^2 R_c(\omega_{c,k-1}^t)\nabla R_c(\omega_{c,k-1}^t)\|^2\\
\overset{(b)}{\le}&4\gamma^2\sigma^2,
\end{aligned}$$

where (a) is from Lemma 7 with $a=1$, (b) is from Assumption D. Recalling that $\tau=CK$ finishes the lemma. □

Lemma 10 (Smoothness for FEDIIR)**.** *Suppose that $\{R_c\}$ satisfies Assumption B, we then have the following inequality*

$$\|\left(I+\gamma\nabla^2 R_c(\omega_{c,k-1}^t)\right)\nabla R_c(\omega_{c,k-1}^t)-\left(I+\gamma\nabla^2 R_c(\omega^{t-1})\right)\nabla R_c(\omega^{t-1})\|^2\le 2(1+\gamma^2)L^2\|\omega_{c,k-1}^t-\omega^{t-1}\|^2.$$

Proof. Using Lemma 7, there exists

$$\begin{aligned}
&\|\left(I+\gamma\nabla^2 R_c(\omega_{c,k-1}^t)\right)\nabla R_c(\omega_{c,k-1}^t)-\left(I+\gamma\nabla^2 R_c(\omega^{t-1})\right)\nabla R_c(\omega^{t-1})\|^2\\
=&\|\nabla R_c(\omega_{c,k-1}^t)-\nabla R_c(\omega^{t-1})+\gamma\nabla^2 R_c(\omega_{c,k-1}^t)\nabla R_c(\omega_{c,k-1}^t)-\gamma\nabla^2 R_c(\omega^{t-1})\nabla R_c(\omega^{t-1})\|^2\\
\le&2\|\nabla R_c(\omega_{c,k-1}^t)-\nabla R_c(\omega^{t-1})\|^2+2\gamma^2\|\nabla^2 R_c(\omega_{c,k-1}^t)\nabla R_c(\omega_{c,k-1}^t)-\nabla^2 R_c(\omega^{t-1})\nabla R_c(\omega^{t-1})\|^2\\
\le&2(1+\gamma^2)L^2\|\omega_{c,k-1}^t-\omega^{t-1})\|^2.
\end{aligned}$$

The last inequality comes from L-smoothness and Moral-smoothness of $R_c(\omega)$ w.r.t. Assumption B. □

Lemma 11 (Statistical Heterogeneity for FEDIIR)**.** *Suppose that $\{R_c\}$ satisfies Assumption C, we then have the following inequality*

$$\mathbb{E}\|\nabla R_c(\omega^{t-1})\|^2\le 2\|\nabla R(\omega^{t-1})\|^2+2G^2.$$

Proof. Using Lemma 7 and Assumption C, we have

$$\begin{aligned}
&\mathbb{E}\|\nabla R_c(\omega^{t-1})\|^2\\
=&\mathbb{E}\|\nabla R_c(\omega^{t-1})-\nabla R(\omega^{t-1})+\nabla R(\omega^{t-1})\|^2\\
\le&2G^2+2\mathbb{E}\|\nabla R(\omega^{t-1})\|^2.
\end{aligned}$$

□

E.3. Convergence Result of FEDIIR

We now present the proof of the convergence theorem for FEDIIR, including both μ-PL inequality and non-convex cases. We first bound the server update error in Lemma 12, then bound the amount of client drift error in Lemma 13, and finally prove the progress made in each round in Lemma 14. Based on this progress, we can derive the required rate of convergence.

Bounding the server update error We started looking at ways to bounding the server update error.

Lemma 12. *Suppose that $\{R_c\}$ satisfies Assumption B, C and D, for all $t \in [T]$, we can bound the server update error as follows*

$$\mathbb{E}\|\omega^t - \omega^{t-1}\|^2 \leq 8\tilde{\eta}^2(3\gamma^2L^2 + 4\gamma L + 2 + \frac{\gamma^2\sigma^2}{CK})\mathbb{E}\|\nabla R(\omega^{t-1})\|^2 + 8\tilde{\eta}^2(3\gamma^2L^2 + 4\gamma L + 2 + \frac{\gamma^2\sigma^2}{CK})G^2 + 16\tilde{\eta}^2(1+\gamma^2)L^2\mathcal{E}^t + \frac{16\tilde{\eta}^2\gamma^2\sigma^2}{CK}.$$

Proof. In t-th communication round, the server update error can be expanded as follows

$$\begin{aligned}
\mathbb{E}\|\omega^t - \omega^{t-1}\|^2 &= \tilde{\eta}^2\mathbb{E}\|\frac{1}{CK}\sum_{c,k}\left(I + \gamma H_c(\omega_{c,k-1}^t)\right) g_c(\omega_{c,k-1}^t) - \frac{\gamma}{CK}\sum_{c,k} H_c(\omega_{c,k-1}^t)\bar{g}^t\|^2 \\
&\overset{(a)}{\leq} 2\tilde{\eta}^2\mathbb{E}\|\frac{1}{CK}\sum_{c,k}\left(I + \gamma H_c(\omega_{c,k-1}^t)\right) g_c(\omega_{c,k-1}^t)\|^2 + 2\tilde{\eta}^2\mathbb{E}\|\frac{\gamma}{CK}\sum_{c,k} H_c(\omega_{c,k-1}^t)\bar{g}^t\|^2 \\
&\overset{(b)}{\leq} \underbrace{4\tilde{\eta}^2\mathbb{E}\|\frac{1}{CK}\sum_{c,k}\left(I + \gamma\nabla^2 R_c(\omega_{c,k-1}^t)\right)\nabla R_c(\omega_{c,k-1}^t)\|^2}_{\mathcal{A}_1} \\
&\quad + \underbrace{2\tilde{\eta}^2\mathbb{E}\|\frac{\gamma}{CK}\sum_{c,k} H_c(\omega_{c,k-1}^t)\bar{g}^t\|^2}_{\mathcal{A}_2} + \frac{16\tilde{\eta}^2\gamma^2\sigma^2}{CK},
\end{aligned}$$

where (a) is from Lemma 7 with $a = 1$; (b) is from Lemma 9.

For the first term $\mathcal{A}_1$, we repeatedly apply the relaxed triangle inequality w.r.t. Lemma 7:

$$\begin{aligned}
\mathcal{A}_1 &= 4\tilde{\eta}^2\mathbb{E}\Big\|\frac{1}{CK}\sum_{c,k}\Big(\left(I + \gamma\nabla^2 R_c(\omega_{c,k-1}^t)\right)\nabla R_c(\omega_{c,k-1}^t) - \left(I + \gamma\nabla^2 R_c(\omega^{t-1})\right)\nabla R_c(\omega^{t-1}) \\
&\qquad + \left(I + \gamma\nabla^2 R_c(\omega^{t-1})\right)\nabla R_c(\omega^{t-1})\Big)\Big\|^2 \\
&\overset{(c)}{\leq} 8\tilde{\eta}^2\mathbb{E}\|\frac{1}{CK}\sum_{c,k}\left(\left(I + \gamma\nabla^2 R_c(\omega_{c,k-1}^t)\right)\nabla R_c(\omega_{c,k-1}^t) - \left(I + \gamma\nabla^2 R_c(\omega^{t-1})\right)\nabla R_c(\omega^{t-1})\right)\|^2 \\
&\quad + 8\tilde{\eta}^2\mathbb{E}\|\frac{1}{C}\sum_{c}\left(I + \gamma\nabla^2 R_c(\omega^{t-1})\right)\nabla R_c(\omega^{t-1})\|^2 \\
&\overset{(d)}{\leq} \frac{8\tilde{\eta}^2}{CK}\sum_{c,k}\mathbb{E}\|\left(I + \gamma\nabla^2 R_c(\omega_{c,k-1}^t)\right)\nabla R_c(\omega_{c,k-1}^t) - \left(I + \gamma\nabla^2 R_c(\omega^{t-1})\right)\nabla R_c(\omega^{t-1})\|^2 \\
&\quad + \frac{8\tilde{\eta}^2}{C}\sum_{c}\mathbb{E}\|\left(I + \gamma\nabla^2 R_c(\omega^{t-1})\right)\nabla R_c(\omega^{t-1})\|^2 \\
&\overset{(e)}{\leq} \frac{16\tilde{\eta}^2(1+\gamma^2)L^2}{CK}\sum_{c,k}\mathbb{E}\|\omega_{c,k-1}^t - \omega^{t-1}\|^2 + \frac{8\tilde{\eta}^2(1+\gamma L)^2}{C}\sum_{c}\mathbb{E}\|\nabla R_c(\omega^{t-1})\|^2 \\
&\overset{(f)}{\leq} \frac{16\tilde{\eta}^2(1+\gamma^2)L^2}{CK}\sum_{c,k}\mathbb{E}\|\omega_{c,k-1}^t - \omega^{t-1}\|^2 + 16\tilde{\eta}^2(1+\gamma L)^2\mathbb{E}\|\nabla R(\omega^{t-1})\|^2 + 16\tilde{\eta}^2(1+\gamma L)^2G^2,
\end{aligned}$$

where (c) is from Lemma 7 with $a = 1$; (d) is from Lemma 7; (e) is from Lemma 10; (f) is from Lemma 11.

For the second term $\mathcal{A}_2$, we have

$$\begin{aligned}
\mathcal{A}_2 &\overset{(g)}{\leq} 2\tilde{\eta}^2\mathbb{E}\left[\|\frac{\gamma}{CK}\sum_{c,k}H_c(\omega_{c,k-1}^t)\|^2\cdot\|\bar{g}^t\|^2\right] \\
&\overset{(h)}{\leq} 2\tilde{\eta}^2\left[2\gamma^2\mathbb{E}\|\frac{1}{CK}\sum_{c,k}\nabla^2 R_c(\omega_{c,k-1}^t)\|^2+\frac{2\gamma^2\sigma^2}{CK}\right]\left[\mathbb{E}\|\frac{1}{C}\sum_{c}\nabla R_c(\omega^{t-1})\|^2\right] \\
&\overset{(i)}{\leq} (4\tilde{\eta}^2\gamma^2L^2+\frac{4\tilde{\eta}^2\gamma^2\sigma^2}{CK})\left[\frac{1}{C}\sum_{c}\mathbb{E}\|\nabla R_c(\omega^{t-1})\|^2\right] \\
&\overset{(j)}{\leq} \left(4\tilde{\eta}^2\gamma^2L^2+\frac{4\tilde{\eta}^2\gamma^2\sigma^2}{CK}\right)\left(2G^2+2\mathbb{E}\|\nabla R(\omega^{t-1})\|^2\right) \\
&= 8\tilde{\eta}^2\gamma^2L^2\mathbb{E}\|\nabla R(\omega^{t-1})\|^2+\frac{8\tilde{\eta}^2\gamma^2\sigma^2}{CK}\mathbb{E}\|\nabla R(\omega^{t-1})\|^2+8\tilde{\eta}^2\gamma^2L^2G^2+\frac{8\tilde{\eta}^2\gamma^2G^2\sigma^2}{CK},
\end{aligned}$$

where (g) is from that L_2-norm satisfies the compatibility; (h) is from Lemma 8 and the definition of $\bar{g}^t$; (i) is from Assumption B and Lemma 7; (j) is from Lemma 11. Plugging back the bounds on $\mathcal{A}_1$ and $\mathcal{A}_2$, there exist

$$\begin{aligned}
\mathbb{E}\|\omega^t-\omega^{t-1}\|^2 &\leq 16\tilde{\eta}^2(1+\gamma L)^2\mathbb{E}\|\nabla R(\omega^{t-1})\|^2+8\tilde{\eta}^2\gamma^2L^2\mathbb{E}\|\nabla R(\omega^{t-1})\|^2+\frac{8\tilde{\eta}^2\gamma^2\sigma^2}{CK}\mathbb{E}\|\nabla R(\omega^{t-1})\|^2 \\
&\quad+16\tilde{\eta}^2(1+\gamma L)^2G^2+8\tilde{\eta}^2\gamma^2L^2G^2+\frac{8\tilde{\eta}^2\gamma^2G^2\sigma^2}{CK} \\
&\quad+\frac{16\tilde{\eta}^2(1+\gamma^2)L^2}{CK}\sum_{c,k}\mathbb{E}\|\omega_{c,k-1}^t-\omega^{t-1}\|^2+\frac{16\tilde{\eta}^2\gamma^2\sigma^2}{CK} \\
&= 8\tilde{\eta}^2(3\gamma^2L^2+4\gamma L+2+\frac{\gamma^2\sigma^2}{CK})\mathbb{E}\|\nabla R(\omega^{t-1})\|^2+8\tilde{\eta}^2(3\gamma^2L^2+4\gamma L+2+\frac{\gamma^2\sigma^2}{CK})G^2 \\
&\quad+16\tilde{\eta}^2(1+\gamma^2)L^2\mathcal{E}^t+\frac{16\tilde{\eta}^2\gamma^2\sigma^2}{CK}.
\end{aligned}$$

□

Bounding the client drift error We will now bound the client-drift error.

Lemma 13. *Suppose that $\{R_c\}$ satisfies Assumption B, C and D, the* FEDIIR *updates with constant local and global step-size such that $\eta_l \leq \frac{1}{4KL\sqrt{1+\gamma^2}}$. Then, for all $t\in[T]$, we can bound the client drift error as follows*

$$\mathcal{E}^t \leq 36K^2\eta_l^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2)\|\nabla R(\omega^{t-1})\|^2+36K^2\eta_l^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2)G^2+144K^2\eta_l^2\gamma^2\sigma^2.$$

Proof. Assuming $K=1$, we have $\mathcal{E}^t=0$ since $\omega_{c,0}^t=\omega^{t-1}$ and the right-hand side are both positive. Thus, the lemma is trivially true if $K=1$. Without loss of generality, we assume below that $K>1$. We first build a recursive bound for the client drift error by expanding the clients' update equation:

$$\begin{aligned}
\mathbb{E}\|\omega_{c,k}^t-\omega^{t-1}\|^2 &= \mathbb{E}\|\omega_{c,k-1}^t-\omega^{t-1}-\eta_l(g_c(\omega_{c,k-1}^t)+\gamma H_c(\omega_{c,k-1}^t)(g_c(\omega_{c,k-1}^t)-\bar{g}^t))\|^2 \\
&\overset{(a)}{\leq} (1+\frac{1}{K-1})\mathbb{E}\|\omega_{c,k-1}^t-\omega^{t-1}\|^2+\underbrace{K\eta_l^2\mathbb{E}\|(I+\gamma H_c(\omega_{c,k-1}^t))g_c(\omega_{c,k-1}^t)-\gamma H_c(\omega_{c,k-1}^t)\bar{g}^t\|^2}_{\mathcal{A}_1}
\end{aligned}$$

where (a) is from Lemma 7 with $a = K-1$. For the first term $\mathcal{A}_1$, we have

$$
\begin{aligned}
\mathcal{A}_1 &\overset{(b)}{\leq} 2K\eta_l^2\mathbb{E}\|(I+\gamma H_c(\omega_{c,k-1}^t))g_c(\omega_{c,k-1}^t)\|^2 + 2K\eta_l^2\gamma^2\mathbb{E}\|H_c(\omega_{c,k-1}^t)\bar{g}^t\|^2 \\
&\overset{(c)}{\leq} 4K\eta_l^2\mathbb{E}\|(I+\gamma\nabla^2 R_c(\omega_{c,k-1}^t)\nabla R_c(\omega_{c,k-1}^t)\|^2 + 16K\eta_l^2\gamma^2\sigma^2 + 2K\eta_l^2\gamma^2\mathbb{E}\left[\|H_c(\omega_{c,k-1}^t)\|^2\cdot\|\bar{g}^t\|^2\right] \\
&\overset{(d)}{\leq} 4K\eta_l^2\mathbb{E}\Big\|(I+\gamma\nabla^2 R_c(\omega_{c,k-1}^t)\nabla R_c(\omega_{c,k-1}^t) - \left(I+\gamma\nabla^2 R_c(\omega^{t-1})\right)\nabla R_c(\omega^{t-1}) \\
&\qquad + \left(I+\gamma\nabla^2 R_c(\omega^{t-1})\right)\nabla R_c(\omega^{t-1})\Big\|^2 \\
&\quad + 16K\eta_l^2\gamma^2\sigma^2 + 2K\eta_l^2\gamma^2\left[\mathbb{E}\|\nabla^2 R_c(\omega_{c,k-1}^t)\|^2 + \sigma^2\right]\left[\mathbb{E}\|\frac{1}{C}\sum_c \nabla R_c(\omega^{t-1})\|^2\right] \\
&\overset{(e)}{\leq} 8K\eta_l^2\mathbb{E}\|(I+\gamma\nabla^2 R_c(\omega_{c,k-1}^t)\nabla R_c(\omega_{c,k-1}^t) - \left(I+\gamma\nabla^2 R_c(\omega^{t-1})\right)\nabla R_c(\omega^{t-1})\|^2 \\
&\quad + 8K\eta_l^2\mathbb{E}\|\left(I+\gamma\nabla^2 R_c(\omega^{t-1})\right)\nabla R_c(\omega^{t-1})\|^2 \\
&\quad + 16K\eta_l^2\gamma^2\sigma^2 + 2K\eta_l^2\gamma^2\left(L^2+\sigma^2\right)\left[\frac{1}{C}\sum_c \mathbb{E}\|\nabla R_c(\omega^{t-1})\|^2\right] \\
&\overset{(f)}{\leq} 16K\eta_l^2(1+\gamma^2)L^2\mathbb{E}\|\omega_{c,k-1}^t - \omega^{t-1}\|^2 + 8K\eta_l^2(1+\gamma L)^2\mathbb{E}\|\nabla R_c(\omega^{t-1})\|^2 + 16K\eta_l^2\gamma^2\sigma^2 \\
&\quad + 2K\eta_l^2\gamma^2\left(L^2+\sigma^2\right)\left[\frac{1}{C}\sum_c \mathbb{E}\|\nabla R_c(\omega^{t-1})\|^2\right] \\
&\overset{(g)}{\leq} 16K\eta_l^2(1+\gamma^2)L^2\mathbb{E}\|\omega_{c,k-1}^t - \omega^{t-1}\|^2 + 8K\eta_l^2(1+\gamma L)^2\left(2\|\nabla R(\omega^{t-1})\|^2 + 2G^2\right) + 16K\eta_l^2\gamma^2\sigma^2 \\
&\quad + 2K\eta_l^2\gamma^2\left(L^2+\sigma^2\right)\left(2\|\nabla R(\omega^{t-1})\|^2 + 2G^2\right) \\
&= 16K\eta_l^2(1+\gamma^2)L^2\mathbb{E}\|\omega_{c,k-1}^t - \omega^{t-1}\|^2 + 4K\eta_l^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2)\|\nabla R(\omega^{t-1})\|^2 \\
&\quad + 4K\eta_l^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2)G^2 + 16K\eta_l^2\gamma^2\sigma^2,
\end{aligned}
$$

where (b) is from Lemma 7 with $a=1$; (c) is from Lemma 9 with $C=1$ and $K=1$; (d) is from Assumption D and the definition of $\bar{g}^t$; (e) is from Lemma 7 with $a=1$ and Assumption B; (f) is from Lemma 10 and that L_2-norm satisfies the compatibility; (g) is from Lemma 11. Plugging back the bounds on $\mathcal{A}_1$, we obtain the recursive bound of the client drift error

$$
\begin{aligned}
\mathbb{E}\|\omega_{c,k}^t - \omega^{t-1}\|^2 &\leq (1+\frac{1}{K-1} + 16K\eta_l^2(1+\gamma^2)L^2)\mathbb{E}\|\omega_{c,k-1}^t - \omega^{t-1}\|^2 \\
&\quad + 4K\eta_l^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2)\|\nabla R(\omega^{t-1})\|^2 \\
&\quad + 4K\eta_l^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2)G^2 + 16K\eta_l^2\gamma^2\sigma^2 \\
&\leq (1+\frac{2}{K-1})\mathbb{E}\|\omega_{c,k-1}^t - \omega^{t-1}\|^2 + 4K\eta_l^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2)\|\nabla R(\omega^{t-1})\|^2 \\
&\quad + 4K\eta_l^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2)G^2 + 16K\eta_l^2\gamma^2\sigma^2.
\end{aligned}
$$

The last inequality comes from the condition on local step-size that $\eta_l \leq \frac{1}{4KL\sqrt{1+\gamma^2}}$, which implies $16K\eta_l^2(1+\gamma^2)L^2 \leq \frac{1}{K-1}$. Unrolling the recursion above, there exists

$$
\begin{aligned}
\mathbb{E}\|\omega_{c,k}^t - \omega^{t-1}\|^2 \leq \Big(&4K\eta_l^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2)\|\nabla R(\omega^{t-1})\|^2 + 4K\eta_l^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2)G^2 \\
&+ 16K\eta_l^2\gamma^2\sigma^2\Big)\sum_{i=1}^{k}(1+\frac{2}{K-1})^i.
\end{aligned}
$$

Note that $(1+\frac{2}{K-1})^i \leq 9$, we have $\sum_{t=1}^{k}(1+\frac{2}{K-1})^i \leq 9K$ for all $k \in [K]$. Averaging then over c and k, we get

$$
\mathcal{E}^t \leq 36K^2\eta_l^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2)\|\nabla R(\omega^{t-1})\|^2 + 36K^2\eta_l^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2)G^2 + 144K^2\eta_l^2\gamma^2\sigma^2.
$$

□

Progress made in each round Now that we have the bound on server update and client drift error, we can describe the progress made in each round of FEDIIR.

Lemma 14. *Suppose that $\{R_c\}$ satisfies Assumption B, C and D, the* FEDIIR *updates with constant local and global step-size such that $\eta_l \leq \frac{1}{4KL\sqrt{1+\gamma^2}}$. Then, for all $t \in [T]$, under condition $\frac{1}{2} - 4\tilde{\eta}L(3\gamma^2L^2 + 4\gamma L + 2 + \frac{\gamma^2\sigma^2}{CK}) - 18\eta_l^2K^2L^2(1+\gamma L)^2(5\gamma^2L^2 + 8\gamma L + 4 + \gamma^2\sigma^2) - 288\tilde{\eta}\eta_l^2K^2(1+\gamma^2)L^3(5\gamma^2L^2 + 8\gamma L + 4 + \gamma^2\sigma^2) > 0$, the* FEDIIR *makes progress in each round as follows:*

$$\mathbb{E}R(\omega^t) \leq \mathbb{E}R(\omega^{t-1}) - \alpha\tilde{\eta}\|\nabla R(\omega^{t-1})\|^2 + \tilde{\eta}\eta_l\left(\beta_1G^2 + \beta_2\gamma^2\sigma^2 + \beta_3\gamma^2G^2\sigma^2\right), \tag{2}$$

where $\alpha > 0$ is a constant, $\beta_1 = 4KL(3\gamma^2L^2+4\gamma L+2)\eta_g + 18K^2L^2(1+\gamma L)^2(5\gamma^2L^2+8\gamma L+4)\eta_l + 288K^2(1+\gamma^2)L^3(5\gamma^2L^2+8\gamma L+4)\tilde{\eta}\eta_l$, $\beta_2 = \frac{8KL}{CK}\eta_g + 72K^2L^2(1+\gamma L)^2\eta_l + 1152K^2(1+\gamma^2)L^3\tilde{\eta}\eta_l$, $\beta_3 = \frac{4KL}{CK}\eta_g + 18K^2L^2(1+\gamma L)^2\eta_l + 288K^2(1+\gamma^2)L^3\tilde{\eta}\eta_l$ are the polynomials in η_l.

Proof. Starting from the smoothness of $R(\omega)$, we have

$$\begin{aligned}
\mathbb{E}R(\omega^t) &\leq \mathbb{E}R(\omega^{t-1}) + \mathbb{E}\langle\nabla R(\omega^{t-1}), \omega^t - \omega^{t-1}\rangle + \frac{L}{2}\mathbb{E}\|\omega^t - \omega^{t-1}\|^2 \\
&= \mathbb{E}R(\omega^{t-1}) + \frac{L}{2}\mathbb{E}\|\omega^t - \omega^{t-1}\|^2 \\
&\quad + \tilde{\eta}\mathbb{E}\langle\nabla R(\omega^{t-1}), -\frac{1}{CK}\sum_{c,k}\left(I + \gamma\nabla^2R_c(\omega_{c,k-1}^t)\right)\left(\nabla R_c(\omega_{c,k-1}^t) - \nabla R_c(\omega^{t-1})\right) - \nabla R(\omega^{t-1})\rangle \\
&\overset{(a)}{\leq} \mathbb{E}R(\omega^{t-1}) - \tilde{\eta}\|\nabla R(\omega^{t-1})\|^2 + \frac{\tilde{\eta}}{2}\|\nabla R(\omega^{t-1})\|^2 + \frac{L}{2}\mathbb{E}\|\omega^t - \omega^{t-1}\|^2 \\
&\quad + \frac{\tilde{\eta}}{2}\mathbb{E}\|\frac{1}{CK}\sum_{c,k}\left(I + \gamma\nabla^2R_c(\omega_{c,k-1}^t)\right)\left(\nabla R_c(\omega_{c,k-1}^t) - \nabla R_c(\omega^{t-1})\right)\|^2 \\
&\overset{(b)}{\leq} \mathbb{E}R(\omega^{t-1}) - \frac{\tilde{\eta}}{2}\|\nabla R(\omega^{t-1})\|^2 + \frac{L}{2}\mathbb{E}\|\omega^t - \omega^{t-1}\|^2 \\
&\quad + \frac{\tilde{\eta}}{2CK}\sum_{c,k}\mathbb{E}\|\left(I + \gamma\nabla^2R_c(\omega_{c,k-1}^t)\right)\left(\nabla R_c(\omega_{c,k-1}^t) - \nabla R_c(\omega^{t-1})\right)\|^2 \\
&\overset{(c)}{\leq} \mathbb{E}R(\omega^{t-1}) - \frac{\tilde{\eta}}{2}\|\nabla R(\omega^{t-1})\|^2 + \frac{\tilde{\eta}(1+\gamma L)^2}{2CK}\sum_{c,k}\mathbb{E}\|\nabla R_c(\omega_{c,k-1}^t) - \nabla R_c(\omega^{t-1})\|^2 + \frac{L}{2}\mathbb{E}\|\omega^t - \omega^{t-1}\|^2 \\
&\overset{(d)}{\leq} \mathbb{E}R(\omega^{t-1}) - \frac{\tilde{\eta}}{2}\|\nabla R(\omega^{t-1})\|^2 + \frac{\tilde{\eta}(1+\gamma L)^2L^2}{2CK}\sum_{c,k}\mathbb{E}\|\omega_{c,k-1}^t - \omega^{t-1}\|^2 + \frac{L}{2}\mathbb{E}\|\omega^t - \omega^{t-1}\|^2 \\
&= \mathbb{E}R(\omega^{t-1}) - \frac{\tilde{\eta}}{2}\|\nabla R(\omega^{t-1})\|^2 + \frac{\tilde{\eta}(1+\gamma L)^2L^2}{2}\mathcal{E}^t + \frac{L}{2}\mathbb{E}\|\omega^t - \omega^{t-1}\|^2,
\end{aligned}$$

where (a) is from that $\langle\mathbf{a},\mathbf{b}\rangle \leq \frac{1}{2}(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2)$; (b) is from Lemma 7, (c) is from that L_2-norm satisfies the compatibility; (d) is from Assumption B.

Combining the above results with Lemma 12 and Lemma 13 yields

$$
\begin{aligned}
\mathbb{E}R(\omega^t) &\leq \mathbb{E}R(\omega^{t-1}) - \frac{\tilde{\eta}}{2}\|\nabla R(\omega^{t-1})\|^2 + \frac{\tilde{\eta}(1+\gamma L)^2L^2}{2}\mathcal{E}^t \\
&\quad + \frac{L}{2}\Big(8\tilde{\eta}^2(3\gamma^2L^2+4\gamma L+2+\frac{\gamma^2\sigma^2}{CK})\mathbb{E}\|\nabla R(\omega^{t-1})\|^2 + 8\tilde{\eta}^2(3\gamma^2L^2+4\gamma L+2+\frac{\gamma^2\sigma^2}{CK})G^2 \\
&\quad + 16\tilde{\eta}^2(1+\gamma^2)L^2\mathcal{E}^t + \frac{16\tilde{\eta}^2\gamma^2\sigma^2}{CK}\Big) \\
&= \mathbb{E}R(\omega^{t-1}) - \frac{\tilde{\eta}}{2}\|\nabla R(\omega^{t-1})\|^2 + \tilde{\eta}\Big(4\tilde{\eta}L(3\gamma^2L^2+4\gamma L+2+\frac{\gamma^2\sigma^2}{CK})\Big)\mathbb{E}\|\nabla R(\omega^{t-1})\|^2 \\
&\quad + \frac{\tilde{\eta}}{2}\big(16\tilde{\eta}(1+\gamma^2)L^3 + L^2(1+\gamma L)^2\big)\mathcal{E}^t \\
&\quad + 4\tilde{\eta}^2L\big(3\gamma^2L^2+4\gamma L+2+\frac{\gamma^2\sigma^2}{CK}\big)G^2 + \frac{8\tilde{\eta}^2L\gamma^2\sigma^2}{CK} \\
&\leq \mathbb{E}R(\omega^{t-1}) - \frac{\tilde{\eta}}{2}\|\nabla R(\omega^{t-1})\|^2 + \tilde{\eta}\Big(4\tilde{\eta}L(3\gamma^2L^2+4\gamma L+2+\frac{\gamma^2\sigma^2}{CK})\Big)\mathbb{E}\|\nabla R(\omega^{t-1})\|^2 \\
&\quad + \frac{\tilde{\eta}}{2}\Big(16\tilde{\eta}(1+\gamma^2)L^3 + L^2(1+\gamma L)^2\Big)\Big(36K^2\eta_l^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2)\|\nabla R(\omega^{t-1})\|^2 \\
&\qquad + 36K^2\eta_l^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2)G^2 + 144K^2\eta_l^2\gamma^2\sigma^2\Big) \\
&\quad + 4\tilde{\eta}^2L\big(3\gamma^2L^2+4\gamma L+2+\frac{\gamma^2\sigma^2}{CK}\big)G^2 + \frac{8\tilde{\eta}^2L\gamma^2\sigma^2}{CK} \\
&= \mathbb{E}R(\omega^{t-1}) - \frac{\tilde{\eta}}{2}\|\nabla R(\omega^{t-1})\|^2 + \tilde{\eta}\Big(4\tilde{\eta}L(3\gamma^2L^2+4\gamma L+2+\frac{\gamma^2\sigma^2}{CK}) + 18\eta_l^2K^2L^2(1+\gamma L)^2 \\
&\quad \cdot (5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2) + 288\tilde{\eta}\eta_l^2K^2(1+\gamma^2)L^3(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2)\Big)\mathbb{E}\|\nabla R(\omega^{t-1})\|^2 \\
&\quad + \tilde{\eta}\eta_l\underbrace{\Big(4KL(3\gamma^2L^2+4\gamma L+2)\eta_g + 18K^2L^2(1+\gamma L)^2(5\gamma^2L^2+8\gamma L+4)\eta_l + 288K^2(1+\gamma^2)L^3(5\gamma^2L^2+8\gamma L+4)\tilde{\eta}\eta_l\Big)}_{\beta_1}G^2 \\
&\quad + \tilde{\eta}\eta_l\underbrace{\Big(\frac{8KL}{CK}\eta_g + 72K^2L^2(1+\gamma L)^2\eta_l + 1152K^2(1+\gamma^2)L^3\tilde{\eta}\eta_l\Big)}_{\beta_2}\gamma^2\sigma^2 \\
&\quad + \tilde{\eta}\eta_l\underbrace{\Big(\frac{4KL}{CK}\eta_g + 18K^2L^2(1+\gamma L)^2\eta_l + 288K^2(1+\gamma^2)L^3\tilde{\eta}\eta_l\Big)}_{\beta_3}\gamma^2G^2\sigma^2 \\
&\leq \mathbb{E}R(\omega^{t-1}) - \alpha\tilde{\eta}\|\nabla R(\omega^{t-1})\|^2 + \tilde{\eta}\eta_l\left(\beta_1G^2 + \beta_2\gamma^2\sigma^2 + \beta_3\gamma^2G^2\sigma^2\right),
\end{aligned}
$$

where the last inequality holds because there exists a constant α such that $\frac{1}{2} - 4\tilde{\eta}L(3\gamma^2L^2+4\gamma L+2+\frac{\gamma^2\sigma^2}{CK}) - 18\eta_l^2K^2L^2(1+\gamma L)^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2) - 288\tilde{\eta}\eta_l^2K^2(1+\gamma^2)L^3(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2) > \alpha > 0$. □

Convergence results of FEDIIR for μ-PL inequality case We first study the convergence of FEDIIR for the μ-PL inequality case.

Theorem 4*. *Let Assumption B, C, D and E hold and* FEDIIR *updates with constant local and global step-size such that* $\eta_l \leq \frac{1}{4KL\sqrt{1+\gamma^2}}, \tilde{\eta} < \frac{1}{2\alpha\mu}$. *Then, under condition* $\frac{1}{2} - 4\tilde{\eta}L(3\gamma^2L^2+4\gamma L+2+\frac{\gamma^2\sigma^2}{CK}) - 18\eta_l^2K^2L^2(1+\gamma L)^2(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2) - 288\tilde{\eta}\eta_l^2K^2(1+\gamma^2)L^3(5\gamma^2L^2+8\gamma L+4+\gamma^2\sigma^2) > 0$, *the sequence of iterates generated by* FEDIIR *satisfies*

$$
\mathbb{E}[R(\omega^t) - R^*] \leq (1-2\alpha\mu\tilde{\eta})^t[R(\omega^0) - R^*] + \eta_l\frac{\beta_1G^2 + \beta_2\gamma^2\sigma^2 + \beta_3\gamma^2G^2\sigma^2}{2\alpha\mu},
$$

where $\alpha > 0$ *is a constant,* $\beta_1 = 4KL(3\gamma^2L^2+4\gamma L+2)\eta_g + 18K^2L^2(1+\gamma L)^2(5\gamma^2L^2+8\gamma L+4)\eta_l + 288K^2(1+\gamma^2)L^3(5\gamma^2L^2+8\gamma L+4)\tilde{\eta}\eta_l$, $\beta_2 = \frac{8KL}{CK}\eta_g + 72K^2L^2(1+\gamma L)^2\eta_l + 1152K^2(1+\gamma^2)L^3\tilde{\eta}\eta_l$, $\beta_3 = \frac{4KL}{CK}\eta_g + 18K^2L^2(1+\gamma L)^2\eta_l + 288K^2(1+\gamma^2)L^3\tilde{\eta}\eta_l$ *are the polynomials in* η_l.

Proof. Using the μ-PL inequality to Equation (2), we have

$$\mathbb{E}R(\omega^t) \leq \mathbb{E}R(\omega^{t-1}) - 2\alpha\mu\tilde{\eta}(R(\omega^{t-1}) - R^*) + \tilde{\eta}\eta_l\left(\beta_1 G^2 + \beta_2\gamma^2\sigma^2 + \beta_3\gamma^2 G^2\sigma^2\right).$$

Subtracting R^* from both sides yields:

$$\begin{aligned}
\mathbb{E}[R(\omega^t) - R^*] &\leq (1 - 2\alpha\mu\tilde{\eta})[R(\omega^{t-1}) - R^*] + \tilde{\eta}\eta_l\left(\beta_1 G^2 + \beta_2\gamma^2\sigma^2 + \beta_3\gamma^2 G^2\sigma^2\right) \\
&\leq (1 - 2\alpha\mu\tilde{\eta})^t[R(\omega^0) - R^*] + \tilde{\eta}\eta_l\left(\beta_1 G^2 + \beta_2\gamma^2\sigma^2 + \beta_3\gamma^2 G^2\sigma^2\right)\sum_{i=0}^{t}(1 - 2\alpha\mu\tilde{\eta})^i \\
&\leq (1 - 2\alpha\mu\tilde{\eta})^t[R(\omega^0) - R^*] + +\tilde{\eta}\eta_l\left(\beta_1 G^2 + \beta_2\gamma^2\sigma^2 + \beta_3\gamma^2 G^2\sigma^2\right)\sum_{i=0}^{\infty}(1 - 2\alpha\mu\tilde{\eta})^i \\
&= (1 - 2\alpha\mu\tilde{\eta})^t[R(\omega^0) - R^*] + \eta_l\frac{\beta_1 G^2 + \beta_2\gamma^2\sigma^2 + \beta_3\gamma^2 G^2\sigma^2}{2\alpha\mu},
\end{aligned}$$

where the last line uses that $\tilde{\eta} < \frac{1}{2\alpha\mu}$ and the limit of the geometric series. □

Convergence results of FEDIIR for general non-convex case We study below the convergence of FEDIIR for the general non-convex case.

Theorem 5*. *Let Assumption B, C, and D hold and* FEDIIR *updates with constant local and global step-size such that* $\eta_l \leq \frac{1}{4KL\sqrt{1+\gamma^2}}$. *Then, under condition* $\frac{1}{2} - 4\tilde{\eta}L(3\gamma^2L^2 + 4\gamma L + 2 + \frac{\gamma^2\sigma^2}{CK}) - 18\eta_l^2K^2L^2(1+\gamma L)^2(5\gamma^2L^2 + 8\gamma L + 4 + \gamma^2\sigma^2) - 288\tilde{\eta}\eta_l^2K^2(1+\gamma^2)L^3(5\gamma^2L^2 + 8\gamma L + 4 + \gamma^2\sigma^2) > 0$, *the sequence of iterates generated by* FEDIIR *satisfies*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla R(\omega^{t-1})\|^2 \leq \frac{R(\omega^0) - R^*}{\alpha\tilde{\eta}T} + \eta_l\frac{\beta_1 G^2 + \beta_2\gamma^2\sigma^2 + \beta_3\gamma^2 G^2\sigma^2}{\alpha},$$

where $\alpha > 0$ *is a constant,* $\beta_1 = 4KL(3\gamma^2L^2 + 4\gamma L + 2)\eta_g + 18K^2L^2(1+\gamma L)^2(5\gamma^2L^2 + 8\gamma L + 4)\eta_l + 288K^2(1+\gamma^2)L^3(5\gamma^2L^2 + 8\gamma L + 4)\tilde{\eta}\eta_l$, $\beta_2 = \frac{8KL}{CK}\eta_g + 72K^2L^2(1+\gamma L)^2\eta_l + 1152K^2(1+\gamma^2)L^3\tilde{\eta}\eta_l$, $\beta_3 = \frac{4KL}{CK}\eta_g + 18K^2L^2(1+\gamma L)^2\eta_l + 288K^2(1+\gamma^2)L^3\tilde{\eta}\eta_l$ *are the polynomials in* η_l. *If we choose the step-sizes* $\eta_l = \mathcal{O}(\frac{1}{\sqrt{T}KL}), \eta_g = \sqrt{CK}$ *and omitting the larger order of each part, we have the convergence rates of* FEDIIR *as follows*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla R(\omega^{t-1})\|^2 = \mathcal{O}\left(\frac{(R(\omega^0) - R^*)L^2}{\sqrt{TCK}}, \frac{\sqrt{CK}L^2G^2}{\sqrt{T}}, \frac{\gamma^2\sigma^2}{\sqrt{TCK}}, \frac{\gamma^2G^2\sigma^2}{\sqrt{TCK}}\right).$$

Proof. Summing up all the T inequalities in Equation (2) for $t \in [T]$ and dividing both sides by $\alpha\tilde{\eta}T$, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla R(\omega^{t-1})\|^2 &= \frac{\sum_{t=1}^{T}\left(R(\omega^{t-1}) - R(\omega^t)\right)}{\alpha\tilde{\eta}T} + \eta_l\frac{\beta_1 G^2 + \beta_2\gamma^2\sigma^2 + \beta_3\gamma^2 G^2\sigma^2}{\alpha} \\
&\leq \frac{R(\omega^0) - R^*}{\alpha\tilde{\eta}T} + \eta_l\frac{\beta_1 G^2 + \beta_2\gamma^2\sigma^2 + \beta_3\gamma^2 G^2\sigma^2}{\alpha},
\end{aligned}$$

where the last inequality is the fact that $R^* \leq R(\omega^T)$. If we choose the step-sizes $\eta_l = \mathcal{O}(\frac{1}{\sqrt{T}KL}), \eta_g = \sqrt{CK}$, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla R(\omega^{t-1})\|^2 = \mathcal{O}\Bigg(&\frac{(R(\omega^0) - R^*)L^2}{\sqrt{TCK}}, \frac{\sqrt{CK}P_2(L)G^2}{\sqrt{T}}, \frac{P_4(L)G^2}{T}, \frac{\sqrt{CK}P_2(L)G^2}{T^{\frac{3}{2}}}, \frac{\sqrt{CK}L^2G^2}{\sqrt{T}}, \\
&\frac{\gamma^2\sigma^2}{\sqrt{TCK}}, \frac{P_2(L)\gamma^2\sigma^2}{T}, \frac{\sqrt{CK}\gamma^2\sigma^2}{T^{\frac{3}{2}}}, \frac{\gamma^2G^2\sigma^2}{\sqrt{TCK}}, \frac{P_2(L)\gamma^2G^2\sigma^2}{T}, \frac{\sqrt{CK}\gamma^2G^2\sigma^2}{T^{\frac{3}{2}}}\Bigg),
\end{aligned}$$

where $P_n(L)$ the n-th degree polynomial in L. Omitting the larger order of each part, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla R(\omega^{t-1})\|^2 = \mathcal{O}\left(\frac{(R(\omega^0) - R^*)L^2}{\sqrt{TCK}}, \frac{\sqrt{CK}L^2G^2}{\sqrt{T}}, \frac{\gamma^2\sigma^2}{\sqrt{TCK}}, \frac{\gamma^2G^2\sigma^2}{\sqrt{TCK}}\right).$$

□

F. Experiments

In this section, we provide additional details about the experimental setup, including the datasets, baseline methods, and hyperparameters used. Then, we supplement the experimental results that were omitted in the main text.

F.1. Datasets

We first give more details about the datasets used in our experiments: RotatedMNIST(Ghifary et al., 2015), VLCS(Fang et al., 2013), PACS(Li et al., 2017), and OfficeHome(Venkateswara et al., 2017). These datasets are classical OOD generalization benchmarks for classification.

RotatedMNIST is a common variant of original MNIST where domain $d \in \{0, 15, 30, 45, 60, 75\}$ contains digits rotated by d degrees. The dataset contains 70,000 samples of dimension $(1, 28, 28)$ and ten classes.

VLCS aggregates photographs from four domains: "VOC2007", "LabelMe", "Caltech101", and "SUN09". The dataset contains 10,729 samples of dimension $(3, 224, 224)$ and five classes ('bird', 'car', 'chair', 'dog', and 'person').

PACS includes images from four domains: "Photos", "Art", "Cartoons", and "Sketches". The dataset contains 9991 samples of dimension $(3, 224, 224)$ and seven classes ('dog', 'elephant', 'giraffe', 'guitar', 'horse', 'house', and 'person').

OfficeHome includes four distinct domains: "Art", "Clipart", "Product", and "Real". The dataset contains 15,588 samples of dimension $(3, 224, 224)$ and sixty five classes.

We start describing the data splitting in detail. We consider the most challenging domain separation setting, where each client contains only samples from a single training domain(Bai et al., 2023). For M training domains $\{S_m\}_m^M$, the number of samples w.r.t. S_m is denoted by $s_m := |S_m|$. With a slight abuse of notation, let's assume that there are P participating clients. For a small number of clients scenario ($P = M$), each training domain is treated as a separate participating client, i.e., $D_c = S_c$ for $c \in [M]$. For a large number of clients scenario ($P > M$), we further split the training domain to distribute it to more participating clients. Define the domain index of client c as $\text{Ind}(c)$, where $\text{Ind}(c) \in [M]$. We first iteratively split the largest domain $m^* = \arg\max_m \frac{s_m}{\sum_c^C \mathbb{1}[\text{Ind}(c)=m]}$, where $\mathbb{1}[\cdot]$ is the indicator function. We then treat each sub-domain as a separate participating client, i.e., $D_c = \frac{S_{\text{Ind}(c)}}{\sum_c^C \mathbb{1}[\text{Ind}(c)=m]}$[4]. This allows some clients to share a single training domain, but no client holds data from multiple domains at the same time. Additionally, this makes every effort to distribute the number of samples among the clients evenly. We summarize the pseudo-code for data splitting in Algorithm 2.

Algorithm 2 Data Splitting

Input: training domains $\{S_m\}_m^M$ and number of participating clients C
if $P = M$ **then**
  $\text{Ind}(c) = c$ for all $c \in [M]$
else if $P > M$ **then**
  $\text{Ind}(c) = c$ for all $c \in [M]$
  for $c = M + 1, \ldots, P$ **do**
    $m^* = \arg\max_m \frac{s_m}{\sum_c^C \mathbb{1}[\text{Ind}(c)=m]}$
    $\text{Ind}(c) = m^*$
  end for
end if
for $c = 1, \ldots, P$ **do**
  $D_c = \frac{S_{\text{Ind}(c)}}{\sum_c^C \mathbb{1}[\text{Ind}(c)=m]}$
end for
Return: $\{D_c\}_c^C$

F.2. Baselines

We compare our proposed method with the following representative classical federated learning methods.

[4]This means that the domain $S_{\text{Ind}(c)}$ is randomly split into $\sum_c^C \mathbb{1}[\text{Ind}(c) = m]$ sub-domains.

FedAvg(McMahan et al., 2017) is the most classical federated learning method in which clients perform multiple epochs of SGD on their local data. It is worth noting that this method lacks the capability for OOD generalization.

FedADG(Zhang et al., 2021) employs the federated adversarial learning approach to align the distribution across clients for learning universal features, which allows good generalization to non-participating clients.

FedSR(Nguyen et al., 2022) performs regularization on the representation and conditional mutual information to encourage the model to learn only the essential information, which helps ignore spurious relationships to achieve OOD generalization.

In addition to the federated learning methods described above, we also list some centralized methods as references in the following.

ERM(Vapnik, 1991) is a foundational approach in machine learning, which attempts to minimize the average empirical risk over all training environments (corresponding to clients).

IRM(Arjovsky et al., 2019) tries to find a representation such that the linear classifier on top of the representation is simultaneously optimal in all environments. Such a representation will discard spurious correlations and can be expected to generalize over OOD.

GroupDRO(Sagawa et al., 2020) apply strong regularization to distributionally robust optimization (DRO) to enhance the robustness of overparameterized neural networks, which significantly improve the performance of the worst group.

REx(Krueger et al., 2021) shows that reducing differences in risk across training environments can reduce a model's sensitivity to a wide range of extreme distributional shifts, resulting in better OOD generalization performance.

It is important to note that these centralized approaches are not direct competitors of our method because they do not apply to domain separation settings in federated learning. As a result, we highlight (in bold) the best-performing method for each of the centralized and federated settings, where the centralized method results are taken directly from Gulrajani & Lopez-Paz (2021).

F.3. Hyperparameters

As mentioned in the main text, we used grid search to tune the hyperparameters of FedIIR, summarized in Table 3.

Condition	Hyperparameter	Used value	Searched candidates
RotatedMNIST	local step-size η_l	$1e-2$	$\{1e-2, 5e-3, 2.5e-3, 1e-3, 5e-4, 2.5e-4, 1e-4\}$
	batch size	64	$\{32, 64\}$
	regularization strength γ	$1e-2$	$\{1e-2, 5e-3, 1e-3, 5e-4, 1e-4\}$
VLCS	local step-size η_l	$1e-3$	$\{1e-2, 5e-3, 2.5e-3, 1e-3, 5e-4, 2.5e-4, 1e-4\}$
	batch size	32	$\{32, 64\}$
	regularization strength γ	$5e-3$	$\{1e-2, 5e-3, 1e-3, 5e-4, 1e-4\}$
PACS	local step-size η_l	$2.5e-3$	$\{1e-2, 5e-3, 2.5e-3, 1e-3, 5e-4, 2.5e-4, 1e-4\}$
	batch size	32	$\{32, 64\}$
	regularization strength γ	$1e-3$	$\{1e-2, 5e-3, 1e-3, 5e-4, 1e-4\}$
OfficeHome	local step-size η_l	$1e-3$	$\{1e-2, 5e-3, 2.5e-3, 1e-3, 5e-4, 2.5e-4, 1e-4\}$
	batch size	32	$\{32, 64\}$
	regularization strength γ	$5e-4$	$\{1e-2, 5e-3, 1e-3, 5e-4, 1e-4\}$
ALL	number of rounds T	100	None
	global step-size η_g	1	None
	ema υ	0.95	$\{0.90, 0.95, 0.99\}$
	seed	$\{0, 1, 2\}$	None

Table 3. Hyperparameters of FedIIR used in the experiments.

Remark 15. These hyperparameters are only tuned in a small number of clients scenario ($P = M$) and remain unchanged across all the scenarios with a large number of clients ($P > M$), where P refers to the total number of participating clients and M represents the number of training domains.

F.4. Additional Experimental Results

Tables 4 to 7 provide detailed results for each domain in the scenario with a small number of clients, where the centralized methods are included as reference baselines. Tables 8 to 11 present detailed results for the specific domain with a large number of client scenarios, where the total number of participating clients is 50. Furthermore, Figure 5 visually depicts the test accuracy versus communication round on the RotatedMNIST and OfficeHome datasets. Finally, Figure 6 shows the validation accuracy versus communication round on all datasets, using the same experimental setup as in Section 5.3.

Algorithm		Backbone	RotatedMNIST (Number of participating clients is 5)						
			0	15	30	45	60	75	Average
Centralized Methods	ERM	ConvNet	**95.9**±**0.1**	98.9±0.0	98.8±0.0	98.9±0.0	**98.9**±**0.0**	96.4±0.0	**98.0**
	GROUPDRO	ConvNet	95.6±0.1	98.9±0.1	98.9±0.1	**99.0**±**0.0**	98.9±0.0	**96.5**±**0.2**	98.0
	IRM	ConvNet	95.5±0.1	98.8±0.2	98.7±0.1	98.6±0.1	98.7±0.0	95.9±0.2	97.7
	REX	ConvNet	95.9±0.2	**99.0**±**0.1**	**98.9**±**0.1**	98.9±0.1	98.7±0.1	96.2±0.2	97.9
Federated Methods	FEDAVG	ConvNet	82.7±0.3	98.2±0.1	99.0±0.1	99.1±0.0	98.2±0.1	89.9±0.4	94.5
	FEDADG	ConvNet	83.4±0.5	98.2±0.1	99.1±0.0	99.1±0.0	**98.7**±**0.1**	89.7±0.3	94.7
	FEDSR	ConvNet	**84.2**±**0.4**	98.0±0.1	98.9±0.1	99.0±0.0	98.3±0.1	90.0±0.3	94.7
	FEDIIR	ConvNet	83.8±1.3	**98.2**±**0.1**	**99.1**±**0.0**	**99.1**±**0.0**	98.5±0.1	**90.8**±**0.2**	**95.0**

Table 4. Average test accuracy (%) using leave-one-out domain validation on RotatedMNIST dataset with 5 participating clients.

Algorithm		Backbone	VLCS (Number of participating clients is 3)				
			C	L	S	V	Average
Centralized Methods	ERM	ResNet-50	97.7±0.4	64.3±0.9	73.4±0.5	74.6±1.3	77.5
	GROUPDRO	ResNet-50	97.3±0.3	63.4±0.9	69.5±0.8	76.7±0.7	76.7
	IRM	ResNet-50	**98.6**±**0.1**	**64.9**±**0.9**	73.4±0.6	**77.3**±**0.9**	**78.5**
	REX	ResNet-50	98.4±0.3	64.4±1.4	**74.1**±**0.4**	76.2±1.3	78.3
Federated Methods	FEDAVG	ResNet-18	95.3±1.0	62.6±0.9	73.0±0.3	74.1±0.8	76.3
	FEDADG	ResNet-18	95.2±0.5	**63.2**±**0.7**	**75.6**±**0.3**	74.3±0.6	**77.1**
	FEDSR	ResNet-18	93.8±1.1	62.3±0.3	74.4±0.6	72.8±0.3	75.8
	FEDIIR	ResNet-18	**96.3**±**0.4**	60.9±0.2	73.2±0.8	**76.1**±**1.4**	76.6

Table 5. Average test accuracy (%) using leave-one-out domain validation on VLCS dataset with 3 participating clients.

Algorithm		Backbone	PACS (Number of participating clients is 3)				
			A	C	P	S	Average
Centralized Methods	ERM	ResNet-50	84.7±0.4	**80.8±0.6**	**97.2±0.3**	**79.3±1.0**	**85.5**
	GROUPDRO	ResNet-50	83.5±0.9	79.1±0.6	96.7±0.3	78.3±2.0	84.4
	IRM	ResNet-50	84.8±1.3	76.4±1.1	96.7±0.6	76.1±1.0	83.5
	REX	ResNet-50	**86.0±1.6**	79.1±0.6	96.9±0.5	77.7±1.7	84.9
Federated Methods	FEDAVG	ResNet-18	82.6±0.4	**77.0±0.4**	94.3±0.2	78.5±0.4	83.1
	FEDADG	ResNet-18	81.7±0.3	76.8±0.6	**94.8±0.4**	79.3±0.9	83.1
	FEDSR	ResNet-18	82.8±1.5	75.2±0.5	94.0±0.6	81.7±0.8	83.4
	FEDIIR	ResNet-18	**82.9±0.8**	75.8±0.3	94.2±0.2	**81.9±0.8**	**83.7**

Table 6. Average test accuracy (%) using leave-one-out domain validation on PACS dataset with 3 participating clients.

Algorithm		Backbone	OfficeHome (Number of participating clients is 3)				
			A	C	P	R	Average
Centralized Methods	ERM	ResNet-50	**61.3±0.7**	52.4±0.3	**75.8±0.1**	**76.6±0.3**	**66.5**
	GROUPDRO	ResNet-50	60.4±0.7	52.7±1.0	75.0±0.7	76.0±0.7	66.0
	IRM	ResNet-50	58.9±2.3	52.2±1.6	72.1±2.9	74.0±2.5	64.3
	REX	ResNet-50	60.7±0.9	**53.0±0.9**	75.3±0.1	76.6±0.5	66.4
Federated Methods	FEDAVG	ResNet-50	64.5±0.1	54.0±0.2	76.8±0.1	**78.6±0.3**	68.5
	FEDADG	ResNet-50	64.3±0.5	54.1±0.6	**77.3±0.3**	78.1±0.1	68.4
	FEDSR	ResNet-50	**65.3±0.2**	**57.3±0.5**	76.2±0.1	77.8±0.1	69.1
	FEDIIR	ResNet-50	64.3±0.4	56.6±0.6	77.2±0.1	78.4±0.1	**69.2**

Table 7. Average test accuracy (%) using leave-one-out domain validation on OfficeHome dataset with 3 participating clients.

Algorithm		Backbone	RotatedMNIST (Number of participating clients is 50)						
			0	15	30	45	60	75	Avg
Federated Methods	FEDAVG	ConvNet	77.9±3.2	95.9±0.5	96.9±0.2	97.0±0.0	96.0±0.4	81.2±1.6	90.8
	FEDADG	ConvNet	80.9±3.9	96.3±0.4	96.9±0.3	97.2±0.3	96.4±0.4	**85.5±1.9**	92.2
	FEDSR	ConvNet	78.3±6.5	95.7±0.6	96.3±0.4	97.1±0.3	96.0±0.4	84.0±0.5	91.2
	FEDIIR	ConvNet	**84.0±1.7**	**96.8±0.4**	**97.7±0.0**	**97.7±0.2**	**97.4±0.2**	84.5±1.2	**93.0**

Table 8. Average test accuracy (%) using leave-one-out domain validation on RotatedMNIST dataset with 50 participating clients, where the number of sampled clients in one communication round is 5.

Algorithm		Backbone	VLCS (Number of participating clients is 50)				
			C	L	S	V	Average
Federated Methods	FEDAVG	ResNet-18	80.2±4.4	58.4±0.9	59.7±0.6	61.5±0.8	65.0
	FEDADG	ResNet-18	72.3±7.9	56.3±1.4	55.9±1.7	58.3±0.2	60.7
	FEDSR	ResNet-18	72.0±1.0	59.2±1.2	50.4±0.9	58.6±0.2	60.0
	FEDIIR	ResNet-18	**93.8±1.7**	**61.5±0.8**	**69.6±0.2**	**71.6±0.3**	**74.1**

Table 9. Average test accuracy (%) using leave-one-out domain validation on VLCS dataset with 50 participating clients, where the number of sampled clients in one communication round is 3.

Algorithm		Backbone	PACS (Number of participating clients is 50)				
			A	C	P	S	Average
Federated Methods	FedAvg	ResNet-18	69.9±0.9	59.0±2.1	90.9±0.6	55.4±1.7	68.8
	FedADG	ResNet-18	73.1±1.4	64.2±1.9	**92.2±0.6**	59.6±0.5	72.3
	FedSR	ResNet-18	70.9±2.0	**69.7±1.0**	86.3±3.6	64.1±3.6	72.7
	FedIIR	ResNet-18	**78.4±0.3**	67.9±1.8	88.8±1.4	**66.6±1.0**	**75.4**

Table 10. Average test accuracy (%) using leave-one-out domain validation on PACS dataset with 50 participating clients, where the number of sampled clients in one communication round is 3.

Algorithm		Backbone	OfficeHome (Number of participating clients is 50)				
			A	C	P	R	Average
Federated Methods	FedAvg	ResNet-50	58.7±0.4	45.4±1.1	67.5±1.4	70.2±1.1	60.5
	FedADG	ResNet-50	57.8±0.8	44.2±0.5	67.4±0.6	70.8±0.9	60.1
	FedSR	ResNet-50	53.8±1.0	41.1±0.8	59.7±1.7	66.8±1.2	55.3
	FedIIR	ResNet-50	**62.9±0.3**	**50.3±0.3**	**74.1±0.5**	**75.3±1.0**	**65.6**

Table 11. Average test accuracy (%) using leave-one-out domain validation on OfficeHome dataset with 50 participating clients, where the number of sampled clients in one communication round is 3.

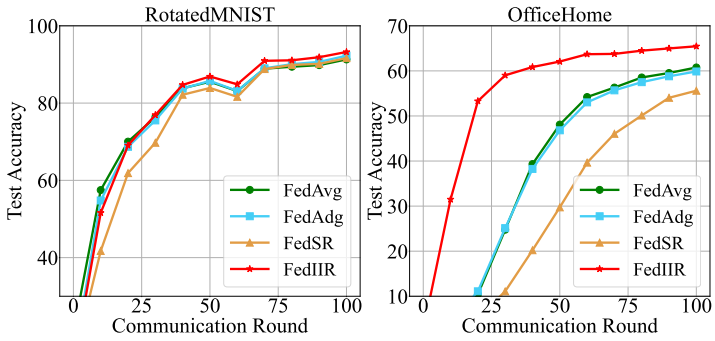

Figure 5. Average test accuracy (%) versus communication round on RotatedMNIST (**left**) and OfficeHome (**right**) dataset with 50 participating clients, where the number of sampled clients in one communication round matches the number of training domains.

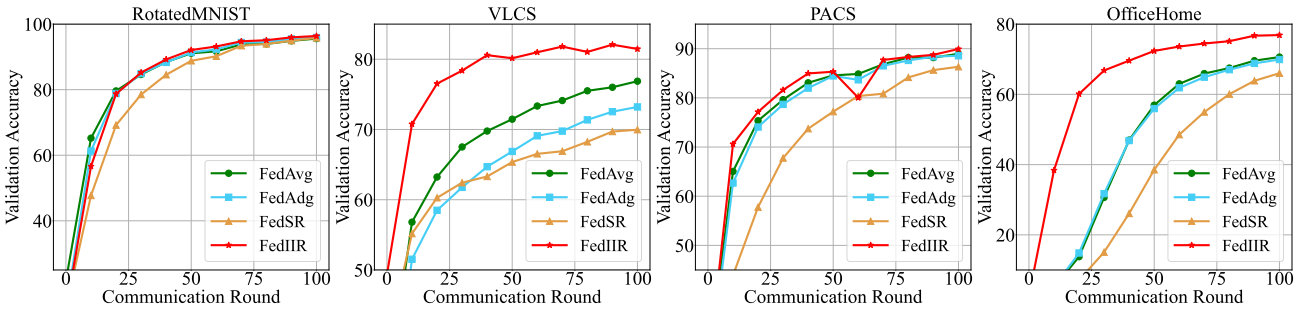

Figure 6. Average validation accuracy (%) versus communication round on four datasets with 50 participating clients, where the number of sampled clients in one communication round matches the number of training domains.