# A Coalition Formation Game Approach for Personalized Federated Learning

**Leijie Wu**[1*] , **Song Guo**[1] , **Yaohong Ding**[1] , **Yufeng Zhan**[2] and **Jie Zhang**[1]
[1]The Hong Kong Polytechnic University
[2]Beijing Institute of Technology
lei-jie.wu@connect.polyu.hk, song.guo@polyu.edu.hk, yaohong.ding@connect.polyu.hk,
yu-feng.zhan@bit.edu.cn, 18104473r@connect.polyu.hk

## Abstract

Facing the challenge of statistical diversity in client local data distribution, personalized federated learning (PFL) has become a growing research hotspot. Although the state-of-the-art methods with model similarity based pairwise collaboration have achieved promising performance, they neglect the fact that model aggregation is essentially a collaboration process within the coalition, where the complex multiwise influences take place among clients. In this paper, we first apply Shapley value (SV) from coalition game theory into the PFL scenario. To measure the multiwise collaboration among a group of clients on the personalized learning performance, SV takes their marginal contribution to the final result as a metric. We propose a novel personalized algorithm: pFedSV, which can 1. identify each client's optimal collaborator coalition and 2. perform personalized model aggregation based on SV. Extensive experiments on various datasets (MNIST, Fashion-MNIST, and CIFAR-10) are conducted with different Non-IID data settings (Pathological and Dirichlet). The results show that pFedSV can achieve superior personalized accuracy for each client, compared to the state-of-the-art benchmarks.

## 1 Introduction

Federated learning (FL) is a recent promising distributed machine learning technique, which can collaboratively train a shared model among multiple clients with data privacy protection [McMahan *et al.*, 2017]. The effectiveness of this shared-model scheme highly depends on similar local data distribution among clients, which is called independent and identically distribution (IID). However, in the vast majority of real-world scenarios, the data distribution of clients is Non-IID with significant heterogeneity, also called *statistical diversity*. This phenomenon is particularly evident in the cross-silo FL [Kairouz *et al.*, 2019], where even the label distribution of each client is distinctly different.

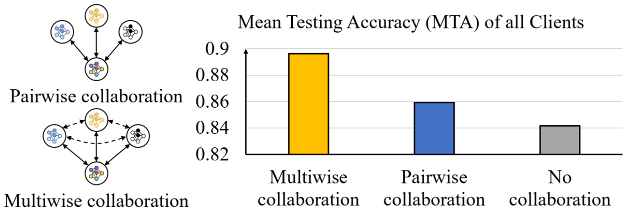


Figure 1: The schematic of Multiwise vs. Pairwise collaboration and the pre-experimental results on CIFAR-10 dataset with the pathological Non-IID setting.

To deal with the above challenges, some researchers have noted that statistical diversity may be an advantage when viewing the issue from the perspective of personalized federated learning (PFL). Initially, some methods try to add an extra fine-tuning step to the trained shared-model by only using local data [Cortes and Mohri, 2014; Mansour *et al.*, 2020; Wang *et al.*, 2019]. In a further step, to facilitate cooperation between clients for mutual progress, they encourage collaboration between clients based on their pairwise model similarity (or loss differences), which determine the respective aggregation weights [Huang *et al.*, 2021; Zhang *et al.*, 2020a; Collins *et al.*, 2021]. In this way, clients with similar model features will make more contribution in the collaboration.

However, latest research reveals the fact that model aggregation is essentially a coalition game [Donahue and Kleinberg, 2021]. Thus, the collaboration within the coalition needs to consider the multiwise influences among clients. In Fig. 1, we illustrate the schematic of Multiwise vs. Pairwise collaboration and the pre-experimental results, which empirically prove the superiority of multiwise collaboration. More specifically, we argue that pairwise collaboration for PFL cannot address two critical challenges, which will be elaborated later in Sec. 3.2 with detailed experiments analysis. **First is data relevance**, each client actually wants to collaborate with others whose local data distribution is relevant to themselves, but the simple one-to-one model parameter similarity testing cannot guarantee clients with similar models will also have higher relevance on their local data, since they ignore the fact that client relevance needs to be analyzed in a multiwise collaboration. **Second is multiwise aggregation weights**, when performing personalized model aggregation, the collaboration within the client coalition must consider their multiwise influences on the final result, while

[*] Contact Author

the aggregation weights in pairwise collaboration only rely on myopia one-to-one model similarity.

Carrying the above insights, in this paper, we are the first to dissect PFL problems via the lens of coalition game theory [Myerson, 2013]. We introduce **Shapley Value (SV)** from coalition game theory to evaluate each client's marginal contribution to the personalized performance, which is naturally based on multiwise influences analysis within the coalition. Specifically, the desirable properties of SV ensure us to simultaneously address above critical issues with a multiwise collaboration solution. Our key technical contributions are:

- As far as we know, we first provide an in-depth analysis to PFL scenario via the lens of coalition game theory. We introduce SV to evaluate each client's marginal contribution during the personalization process, which is based on multiwise influences analysis within the collaborator coalition.
- We propose a novel pFedSV algorithm that exploits the unique properties of SVs to simultaneously address the key issues in PFL, where the *Null player* and *Symmetry* properties are applied to identify data-relevant client coalition, and the *Linearity* and *Group rationality* properties are employed for each client's personalized model aggregation with multiwise collaboration.
- Eventually, we conduct extensive experiments to evaluate the performance of pFedSV on several realistic datasets with different Non-IID data settings. The results show that pFedSV can outperform the state-of-the-art methods in the personalized accuracy of each client.

## 2 Related Work

Recently, to address the statistical diversity challenge of clients with Non-IID data, Personalized Federated Learning (PFL) has emerged as a solution scenario that is attracting more attention. Initially, additional fine-tuning step on the client's local dataset is a natural strategy for personalization [Mansour *et al.*, 2020; Wang *et al.*, 2019], and some prior studies have attempted to enhance the robustness of global model under severe non-IID level by regularization [T Dinh *et al.*, 2020] or add a proximal term [Li *et al.*, 2020]. However, they are all adjusted on single global model scheme which cannot satisfy the personalized demand of individual clients at the local data level, as the target distribution of clients in severe Non-IID setting can be fairly different from the global average aggregation [Jiang *et al.*, 2019].

With the above challenges, some recent methodes consider to train a personalized model for each client that perfectly adapts to their local targets. pFedHN employs a hypernetwork to directly generate personalized parameters for each client's model [Shamsian *et al.*, 2021]. To promote cooperation between clients with relevant local target distributions to achieve mutual progress, FedFomo [Zhang *et al.*, 2020a] and FedAMP [Huang *et al.*, 2021] encourage pairwise collaboration for clients with similar model features, where the former uses loss similarity and the latter adopts parameter similarity.

Although pairwise collaboration methods have achieved good results, they ignore the fact that model aggregation is a coalition game, which requires considering the multiwise influences among collaborators. Different from existing work, we introduce SV from coalition game theory to analyze the multiwise influences among clients, which is based on their marginal contribution to the personalization of others.

| Client index | **A** | **B** | **C** | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Label distribution | $[\mathbf{0}, \mathbf{1}]$ | $[\mathbf{1}, \mathbf{2}]$ | $[\mathbf{0}, \mathbf{3}]$ | $[2, 7]$ | $[3, 6]$ | $[6, 7]$ | $[4, 9]$ | $[5, 8]$ | $[4, 8]$ | $[5, 9]$ |

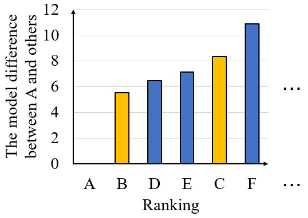


Figure 2: Pre-experiment on CIFAR-10 with the pathological Non-IID setting, where the table shows the details of client data distribution and the bar chart shows the model different $||\theta_A - \theta_i||^2$ between client $A$ and other $i \in \{N\}$.

## 3 The Essence of PFL Problem

In this section, we first introduce the objectives of PFL and the corresponding problem formulation [Kairouz *et al.*, 2019; Zhao *et al.*, 2018]. Then, we will delve into the root causes of the problems through various pre-experiments analysis and present our multiwise collaboration solution: Shapley value, to address these problems.

### 3.1 Problem Formulation

PFL aims to customize personalized models for each client to accommodate their private data distribution through collaboration between a set of clients. Considering $n$ clients $C_1, C_2, \cdots, C_n$ with the same structure of model $\mathcal{M}$ but parameterized by different weights $\theta_1, \theta_2, \cdots, \theta_n$, their respective personalized models can be denoted by $\mathcal{M}(\theta_i)$. Unlike traditional federated learning, the private dataset $\mathcal{D}_i$ of each client $i$ is uniformly sampled from their own distinct data distribution $\mathcal{P}_i$. Let $\ell_i$ denote the corresponding loss function for client $i$, and $\mathcal{L}_i$ the average loss over the private dataset $\mathcal{D}_i$ is denoted by $\mathcal{L}_i(\theta_i) = \frac{1}{d_i} \sum_{j \in \mathcal{D}_i} \ell_i(j, \theta_i)$, where $d_i$ is the data size of $\mathcal{D}_i$ and $j$ is one of the data samples in $\mathcal{D}_i$. The optimization objective of PFL is

$$\Theta^* = \arg\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(\theta_i), \tag{1}$$

where $\Theta$ is the set of personalized model parameter $\{\theta_i\}_{i=1}^n$.

### 3.2 Root Causes of PFL Problems

**Data Relevance.** According to extensive previous work for data Non-IID in FL [Zhao *et al.*, 2018; Li *et al.*, 2020; Li *et al.*, 2019; Karimireddy *et al.*, 2020], the model performance degradation in the Non-IID case is due to the significant local level data distribution differences among clients. By the same token, for better model personalization in PFL scenario, each client should seek to collaborate with others whose local data distribution is truly relevant to their own. In the table of Fig. 2, we explain what is data relevance using the MNIST dataset (ten labeled digits from 0 to 9) with pathological Non-IID, i.e., each client randomly has two types of labels with equal data size. Take client $A$ with labels $[0, 1]$ as an

一对一算相似度不能反映真正的相关

example, client $B$ with labels $[1, 2]$ and client $C$ with labels $[0, 3]$ are its data-relevant clients, which are the clients that $A$ really wants to collaborate with in its own model personalization process. However, the one-to-one model similarity testing is myopia and ineffective as it completely ignores the multiwise influences among clients. Still using client $A$ as an example, we adopt $||\theta_A - \theta_i||^2, i \in \{N\}$ to measure the model parameter difference between $A$ and other models. If the model similarity theory is true, the model differences of $B$ and $C$ should be the smallest among all clients, which is contrary to our experiment results in Fig. 2.

**Multiwise Collaboration Weights.** Another key-point in PFL is the personalized model aggregation targeting each client's local data level features. The previous methods adopted pairwise collaboration by comparing model similarities one-to-one and assigning proportional aggregation weights based on their magnitudes, which is demonstrated in Fig. 1. However, imagine a scenario where the client's current model is a carriage, and every other client's model is a force that moves the carriage in a certain direction, and the destination of the carriage is the client's optimal personalized model. Obviously, the movement of carriage is the result of multiple forces combination, which indicates that the multiwise influences among collaborators must be considered when generating the personalized model aggregation weights. Under the same conditions that their respective data-related customers are informed in advance, we conduct extensive pre-experiments where the only variable is the collaboration methods among clients when generating aggregated weights. The results in Fig. 1 indicating that multiwise collaboration outperforms pairwise collaboration.

### 3.3 The Coalition Game Solution: Shapley Value

The SV from coalition game theory help us to evaluate each client's marginal contribution to the personalization of others, its calculation process involves the performance analysis of clients in different combinations, which naturally incorporates the multiwise influences analysis we need within the coalition. Therefore, we adopt SV as our multiwise collaboration solution to address two above critical challenges.

**Preliminaries of SV.** Consider each client as a player in the coalition game, where $N = \{1, 2, \cdots, n\}$ denotes the set of players. A *utility function* $v(S) : 2^n \to \mathbb{R}$ assigns to every coalition $S \subseteq N$ a real number representing the gain obtained by the coalition as a whole. By convention, we assume that $v(\emptyset) = 0$. Formally, let $\pi \in \Pi(N)$ denote a permutation of clients in $N$, and $C_\pi(i) = \{j \in \pi : \pi(j) < \pi(i)\}$ is a coalition containing all predecessors of client $i$ in $\pi$. The SV of client $i$ is defined as the average marginal contribution to all possible coalitions $C_\pi(i)$ formed by other clients:

$$\varphi_i(v) = \frac{1}{|N|!} \sum_{\pi \in \Pi} [v(C_\pi(i) \cup \{i\}) - v(C_\pi(i))]. \tag{2}$$

The formula in (2) can also be rewritten as:

$$\varphi_i(v) = \sum_{S \in N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)]. \tag{3}$$

The importance of SV in addressing the two root problems stems from its desirable and unique properties as follows.

**SV for Data Relevance.**

- *Symmetry*: Two clients who have the same contribution to the coalition should have the same value. That is, if client $i$ and $j$ are equivalent in the sense of $v(S \cup \{i\}) = v(S \cup \{j\}), \forall S \subseteq N \setminus \{i, j\}$, then $\varphi_i = \varphi_j$.
- *Null Player*: Client with zero marginal contributions to all possible coalitions is null player and receive zero payoff, i.e., $\varphi_i = 0$ if $v(S \cup \{i\}) = 0$ for all $S \subseteq N \setminus \{i\}$.

The *Symmetry* and *Null Player* properties in SV can assist us in precisely identifying their data-relevant clients, where those irrelevant clients will be identified as *null* or even *negative*. For PFL scenario, in each communication round $t$, each client $i$ will first upload their local updated model parameter $\theta_i^t$ to the server, forming a model pool $\{\theta_i^t\}_{i=1}^n$ on the server-side. Then, they will download the relevant clients' model parameters from the model pool for their personalization. Of course, at the beginning, they must have no idea who are their data-relevant clients in the model pool. Therefore, we construct a model download vector for each client based on the relevance score, which can guarantee their data-relevant clients will be identified within a certain rounds .

At the beginning, for each client $i, i \in N$, it generates an $n$-dimensional all-zero relevance vector $\phi^{i,t}$, where $\phi_j^{i,t}$ denotes the relevance score of client $j$ to $i$ in $t$-th round and we have $\phi^{i,t=1} = \vec{0}$. Thus initially it will randomly download $k$ other model parameters, and later it will choose to download the top-$k$ with non-negative relevance. Then, we form a set $S_{i,k}^t$ with its own and downloaded $k$ model parameters. Now, client $i$ can compute the relevance score of the downloaded models by using the following coalition game and the local validation dataset $\mathcal{D}_{V_i}$. We define a coalition game $(\{\theta_j^t\}_{j \in S_{i,k}^t}, v)$, where $v$ is a utility function that assigns a value to each $X \subseteq S_{i,k}^t$. Here, we define the value using the performance $\mathcal{A}$ of the model with parameters $\theta_X^t$ generated from $X$ on the validation dataset $\mathcal{D}_{V_i}$ as follows.

$$\theta_X^t = \frac{1}{|X|} \sum_{j \in X} \theta_j^t, \text{and } v(X, \mathcal{D}_{V_i}) = \mathcal{A}(\theta_X^t, \mathcal{D}_{V_i}). \tag{4}$$

Then, we can obtain the SV $\varphi_j^t, j \in S_{i,k}^t$ of all downloaded model parameters from the coalition game $(\{\theta_j^t\}_{j \in S_{i,k}^t}, v)$ in $t$-th round according to Eq. (2). Next, client $i$ will updated its relevance vector to $\phi^{i,t+1}$ as below:

$$\phi_j^{i,t+1} = \alpha \phi_j^{i,t} + (1 - \alpha) \varphi_j^t, \forall j \in S_{i,k}^t. \tag{5}$$

Intuitively, a larger relevance score for client $j$ means that it contributes more to the personalized performance of client $i$, and therefore has a higher likelihood of being the data-relevant client. Besides, we notice that the relevance vector is unstable in the initial stage and requires several rounds of iterative updates. However, by definition, when other clients' model parameters negatively affect the personalized performance in the coalition game, its SV can be negative, so the irrelevant clients' scores will rapidly decrease in the iterations

**Algorithm 1** Shapley value based Personalized federated learning (pFedSV)

```
Input: $n$, $N$, $\{\theta_i\}_{i=1}^n$, $k$, $E$, $T$, $R$ and $\mathcal{D}_{V_i}$.
Onput: $\{\theta_i^*\}_{i=1}^n$: clients' personalized model parameters.
 1: Initialize the clients' model parameters $\{\theta_i\}_{i=1}^n$. 2
 2: Initialize clients' relevence vector: $\phi^{i,t=1} = \vec{0}, \forall i \in N$.
 3: for round $t = 1, 2, \cdots, T$ do
 4:     for client $i = 1, 2, \cdots, n$ do
 5:         update its model parameter to $\theta_i^t$ via $E$ local epochs and upload to the server.
 6:         download $k$ copies of other clients' model parameters from server according to the dynamic top-$k$ download mechanism.
 7:         $S_{i,k}^t \leftarrow \theta_i^t \cup \{k$ downloaded model parameters$\}$.
 8:         $\varphi_j^t \Leftarrow$ SV_evaluation($S_{i,k}^t, \mathcal{D}_{V_i}, R$), $\forall j \in S_{i,k}^t$.
 9:         $\phi_j^{i,t+1} = \alpha\phi_j^{i,t} + (1-\alpha)\varphi_j^t, \forall j \in S_{i,k}^t$
10:         $w_j^{t*} = \frac{w_j^t}{\sum_j w_j^t} \Leftarrow w_j^t = \frac{\max(\varphi_j^t, 0)}{\|\theta_i^t - \theta_j^t\|}, \forall j \in S_{i,k}^t$.
11:         $\theta_i^{t*} = \sum_j w_j^{t*}\theta_j^t, \forall j \in S_{i,k}^t$.
12:     end for
13: end for
```

and thus excluded. For example, if client $i$ has 2 data-relevant clients in total 20 clients and it downloads 5 other clients' model parameters per round, then it takes at most 5 rounds to identify all data-relevant clients (The Proof is in Appendix.1).

*Dynamic top-k download mechanism:* The above relevance vector has another crucial role in dynamically adjusting the number of model downloads $k$. With constant updates, only the score of data-relevant clients can remain non-negative, so when the remaining number does not match the current value of $k$, we dynamically adjust it to ensure that all downloads are for the necessary data-relevant clients.

**SV for Multiwise Collaboration Weights.**

- *Group Rationality*: The gain of the entire coalition $S$ is completely distributed among all clients in $S$, i.e., $v(S) = \sum_{i \in S} \varphi_i$.
- *Linearity*: The values under multiple utilities sum up to the value under a utility that is the sum of all these utilities: $\varphi_i(v) + \varphi_i(u) = \varphi_i(v+u)$. Also, for every $i \in N$ and any real number $a$, it has $\varphi_i(av) = a\varphi_i(v)$.

The *Group Rationality* and *Linearity* properties perfectly fit the demand of personalized model aggregation with multiwise collaboration in PFL. According to Eq. (2), the computation of SV requires exploring multiple permutations among clients within the coalition game, hence the process naturally takes into account the complex effects of multiwise collaborations among clients. Furthermore, the *Group Rationality* property guarantees that the target of all clients within the coalition is the same, i.e., to achieve the best performance for the current client $i$, which also means the optimal personalized model parameters. And the *Linearity* property naturally fits into the model aggregation process, i.e., the improvement of personalized accuracy by aggregating other client models into their own is fully reflected in the SV of the model, where a larger positive SV indicates a larger positive contribution to performance improvement and vise verse.

**Algorithm 2** Shapley value evaluation

```
Input: $S_{i,k}^t$, $\mathcal{D}_{V_i}$, $R$.
Onput: $\varphi_j^t, \forall j \in S_{i,k}^t$.
$P \leftarrow$ set of $R$ permutations of $S_{i,k}^t$.
for client $j \in S_{i,k}^t$ do
    for permutation $p \in P$ do
        $X_{p,j}^t = \{l | l \in S_{i,k}^t \wedge p(l) \leq j\}$
        $a_j^p \leftarrow v(\{X_{p,j}^t \cup j\}, \mathcal{D}_{V_i}) - v(X_{p,j}^t, \mathcal{D}_{V_i})$
        $\varphi_j^t \leftarrow \varphi_j^t + \frac{1}{|P|} a_j^p$.
    end for
end for
```

Based on the SV $\varphi_j^t$ for all $j \in S_{i,k}^t$ in Eq. (5), the downloaded model parameters are assigned a real number that represents their marginal contribution to the personalization of the current client $i$, where a positive number indicates a positive effect and vice versa. Therefore, we first need to exclude those model parameters of irrelevant clients with negative SV out of the multiwise collaboration in current round, and only compute the initial weights for data-relevant clients within the coalition as follows:

$$w_j^t = \frac{\max(\varphi_j^t, 0)}{\|\theta_i^t - \theta_j^t\|}, \tag{6}$$

where we adopt the model differences $\|\theta_i^t - \theta_j^t\|$ to further fine-tune the resulting weights. Then, we perform 0-1 normalization on the initial weights to obtain their respective aggregation weights $w_j^{t*} = \frac{w_j^t}{\sum_j w_j^t}$, which maintains $w_j^{t*} \in [0, 1]$ and $\sum_j w_j^{t*} = 1$. Finally, we generate the personalized model parameters of client $i$ in $t$-th round based on the following multiwise collaboration:

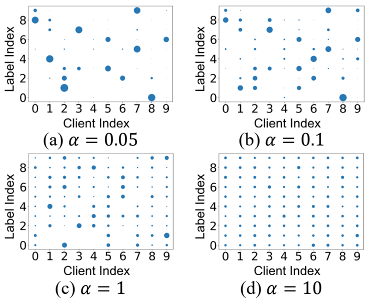
$$\theta_i^{t*} = \sum_j w_j^{t*}\theta_j^t, \quad \forall j \in S_{i,k}^t. \tag{7}$$

Note that we perform SV evaluations in each round to accommodate small changes in multiwise influences due to parameter changes after client local model training.

## 4 pFedSV Algorithm

Based on the above solution framework, we propose the pFedSV Algorithm shown below. The algorithm 1 demonstrates the personalization process of pfedSV for each client.

In the beginning, each client initialize their model parameters $\theta_i$ and the relevance vector $\phi^i$ (Line 1-2). Then in each round $t$, they update the model parameters to $\theta_i^t$ by $E$ local epochs training and upload them to the server (Line 5). Next, they download $k$ copies of other clients' model parameters according to the dynamic top-$k$ download mechanism (Line 6). At this point, the basic conditions of each client's coalition game for their own model personalization are available. First, they form a coalition game $(\{\theta_j^t\}_{j \in S_{i,k}^t}, v)$, where $S_{i,k}^t$ is the model parameter set consisting of $k$ downloaded



(a) $\alpha = 0.05$ (b) $\alpha = 0.1$ (c) $\alpha = 1$ (d) $\alpha = 10$

Figure 3: The visualization of client statistical diversity on MNIST dataset with the Dirichlet setting, where $x$-axis indicates the client index, $y$-axis indicates the label index, and the size of scattered points indicates the number of training samples owned by the client.

| Methods | MNIST | | FMNIST | | CIFAR-10 | |
|---|---|---|---|---|---|---|
| | 10 clients | 100 clients | 10 clients | 100 clients | 10 clients | 100 clients |
| Seperate | $96.11 \pm 0.28$ | $93.27 \pm 3.68$ | $92.35 \pm 0.43$ | $91.42 \pm 2.69$ | $84.15 \pm 2.13$ | $75.57 \pm 4.08$ |
| FedAvg | $69.74 \pm 1.68$ | $60.46 \pm 1.14$ | $65.31 \pm 2.49$ | $57.63 \pm 4.73$ | $44.67 \pm 4.16$ | $36.64 \pm 4.75$ |
| FedProx | $75.12 \pm 0.73$ | $64.45 \pm 1.83$ | $72.16 \pm 3.05$ | $61.83 \pm 3.49$ | $47.68 \pm 2.67$ | $39.75 \pm 4.39$ |
| FedAvg+FT | $77.38 \pm 1.06$ | $68.51 \pm 1.67$ | $75.18 \pm 3.54$ | $66.49 \pm 4.51$ | $47.34 \pm 3.24$ | $42.13 \pm 5.68$ |
| pFedMe | $93.75 \pm 1.34$ | $84.57 \pm 2.61$ | $90.46 \pm 1.72$ | $81.39 \pm 2.97$ | $79.48 \pm 4.59$ | $64.15 \pm 5.86$ |
| FedFomo | $96.90 \pm 0.87$ | $93.71 \pm 2.05$ | $94.10 \pm 0.65$ | $92.78 \pm 1.92$ | $85.93 \pm 3.02$ | $74.36 \pm 2.15$ |
| FedAMP | $95.82 \pm 1.37$ | $92.59 \pm 1.88$ | $93.26 \pm 2.14$ | $91.46 \pm 2.04$ | $84.32 \pm 3.69$ | $72.91 \pm 2.83$ |
| pFedHN | $96.53 \pm 0.84$ | $94.16 \pm 1.38$ | $94.97 \pm 0.86$ | $93.69 \pm 1.58$ | $86.38 \pm 2.72$ | $76.62 \pm 3.05$ |
| pFedSV(Ours) | $\mathbf{98.01 \pm 0.83}$ | $\mathbf{96.94 \pm 1.75}$ | $\mathbf{96.16 \pm 0.58}$ | $\mathbf{94.68 \pm 2.36}$ | $\mathbf{89.64 \pm 1.88}$ | $\mathbf{80.65 \pm 3.78}$ |

Table 1: The MTA with the pathological Non-IID data setting, where bold indicates the best result among all methods.

model parameters and their own (Line 7). Then, the SV evaluation process is performed to obtain the SV of each model parameters in $S_{i,k}^t$ (Line 8), which will be elaborated in Algorithm 2 later. Next, the obtained SV are used to address two challenges: updating the relevance vector of each client for identifying their data-relevant clients (Line 9), and calculating the multiwise aggregation weights for model personalization (Line 10). Finally, each client performs the respective weighted aggregation to obtain new model parameters as the starting point for the next round $t + 1$.

Since the time complexity required to accurately evaluate SV is exponential to the number of players, we need an approximation algorithm to make the trade-off. According to Eq. (2), the calculation of SV can be viewed as an expectation calculation problem, so we adopt a widely accepted Monte Carlo sampling technique to approximate the SV [Mann and Shapley, 1962; Castro *et al.*, 2009; Maleki *et al.*, 2013]. The related details are elaborated in Algorithm 2. First, we randomly sample $R$ permutations of $S_{i,k}^t$ out of total $|S_{i,k}^t|!$ permutations to form a set $P$ (Line 1). Then, for each permutation, we scan it from the first element to the last and calculate the marginal contribution for every newly added model parameters (Line 3-5). Perform the same procedure for all $R$ permutations and the approximation of SV is the average of $R$ calculated marginal contributions (Line 6). As the number of samples $R$ gradually increases, Monte Carlo sampling will eventually be an unbiased estimate of the SV. Previous work has analyzed error bounds for this approximation and concluded that in practice, $R = 3|S_{i,k}^t| \ll |S_{i,k}^t|!$ Monte Carlo samples is sufficient for convergence [Maleki *et al.*, 2013].

# 5 Experiments

## 5.1 Experimental Setup

In this section we will show all the experiment setup, including hyperparameter settings, datasets, baselines, etc.

**Baselines & Evaluation Metric.** We evaluate the performance of pFedSV by comparing it with the state-of-the-art PFL algorithms, including pFedMe [T Dinh *et al.*, 2020], pFedHN[Shamsian *et al.*, 2021], FedFomo [Zhang *et al.*, 2020a],and FedAMP[Huang *et al.*, 2021]. For a more comprehensive understanding, we also add the classical single global model methods FedAvg [McMahan *et al.*, 2017], FedAvg with fine-tuning (FedAvg+FT) and FedProx [Li *et al.*, 2020], as well as the simplest separate local training named *separate*, where each client individually train their own model without collaboration. The performance of all algorithms is evaluated by the mean testing accuracy (MTA), which is the average of the testing accuracy on all clients and the ± indicates the error range of the MTA after 5 repeated experiments.

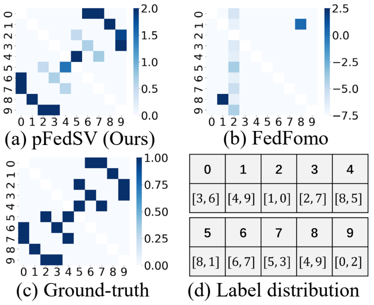
**Non-IID data setting.** We conduct experiments on three public benchmark datasets, MNIST [LeCun, 1998], FMNIST (Fashion-MNIST) [Xiao *et al.*, 2017] and CIFAR-10 [Krizhevsky *et al.*, 2009]. For each dataset, we adopt two different Non-IID settings: I. The pathological Non-IID data setting that each client is randomly assigned two types of labels and the privacy data is not similar between any two clients, which is shown as Fig. 1. II. The Dirichlet Non-IID data setting **Dir**($\alpha$), where a smaller $\alpha$ means higher data heterogeneity, as it makes the client label distribution more biased. We visualized the effect of different $\alpha$ on clients' statistical diversity for MNIST dataset in Fig. 3.

**Implementation details.** We consider two FL scenarios with different client scales: total 10 clients with 100% participation and total 100 clients with 10% participation. We set the training parameters as 5 local epochs, the same number of communication rounds (20 rounds for the former and 100 rounds for the latter), and learning rates (0.01 for MNIST and FMNIST, 0.1 for CIFAR-10). For the SV related hyperparameters, we set the Monte Carlo sampling number as $R = 3|S_{i,k}^t|$, where the number of model parameters downloaded for each round is $k = 5$ in the beginning. Note that $k$ is dynamically adjusted according to the dynamic top-$k$ download mechanism in *SV for data relevance* of Sec. 3.3.

## 5.2 Performance Analysis

In this section, we will demonstrate the performance of our pFedSV compared to all the state-of-the-art benchmarks and analyze the experiment results in details.

**Results on the different Non-IID data setting.** Table 1 demonstrate the MTA of all methods with the pathological Non-IID data setting. Since each client has only two types of labels, which significantly simplifies the complexity of the classification task for each client, the high performance of separate on all datasets reflects the simplicity. However, the pathological Non-IID data setting is a great challenge for the single global model methods, we can observe that FedAvg and FedProx suffer from significant performance degradation



(a) pFedSV (Ours)

(b) FedFomo

(c) Ground-truth

| 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| [3, 6] | [4, 9] | [1, 0] | [2, 7] | [8, 5] |
| **5** | **6** | **7** | **8** | **9** |
| [8, 1] | [6, 7] | [5, 3] | [4, 9] | [0, 2] |

(d) Label distribution

Figure 4: The visualization of clients' relevance matrix $\phi^i, i \in N$ on different algorithms with the pathological MNIST Non-IID setting after convergence. $x$-axis and $y$-axis show the client index. The chart shows the client label distribution obtained from the omniscience perspective.

| Methods | MNIST | | FMNIST | | CIFAR-10 | |
|---|---|---|---|---|---|---|
| | 10 clients | 100 clients | 10 clients | 100 clients | 10 clients | 100 clients |
| Seperate | $74.05 \pm 2.11$ | $59.81 \pm 5.73$ | $60.18 \pm 6.42$ | $58.22 \pm 6.73$ | $40.53 \pm 7.20$ | $36.15 \pm 6.88$ |
| FedAvg | $43.57 \pm 3.75$ | $30.15 \pm 4.82$ | $40.58 \pm 4.16$ | $36.49 \pm 5.07$ | $33.81 \pm 5.07$ | $26.82 \pm 6.43$ |
| FedProx | $47.49 \pm 4.18$ | $44.76 \pm 5.49$ | $43.09 \pm 4.82$ | $40.34 \pm 4.72$ | $35.76 \pm 5.18$ | $29.91 \pm 5.58$ |
| FedAvg+FT | $55.72 \pm 3.84$ | $50.57 \pm 4.26$ | $50.27 \pm 4.13$ | $44.83 \pm 5.01$ | $42.42 \pm 5.29$ | $37.05 \pm 5.22$ |
| pFedMe | $64.39 \pm 4.08$ | $58.02 \pm 3.51$ | $60.27 \pm 3.59$ | $56.81 \pm 4.01$ | $50.73 \pm 4.29$ | $44.21 \pm 5.09$ |
| FedFomo | $72.54 \pm 2.18$ | $63.07 \pm 2.54$ | $64.75 \pm 3.42$ | $60.49 \pm 3.72$ | $53.83 \pm 4.57$ | $48.35 \pm 5.29$ |
| FedAMP | $70.15 \pm 3.02$ | $60.28 \pm 3.11$ | $62.28 \pm 2.53$ | $58.94 \pm 3.14$ | $51.57 \pm 4.03$ | $46.05 \pm 4.48$ |
| pFedHN | $73.35 \pm 2.04$ | $62.57 \pm 4.11$ | $62.95 \pm 3.44$ | $59.55 \pm 4.15$ | $52.82 \pm 3.88$ | $47.19 \pm 5.83$ |
| pFedSV(Ours) | $\mathbf{78.17 \pm 1.59}$ | $\mathbf{70.76 \pm 2.41}$ | $\mathbf{71.47 \pm 1.86}$ | $\mathbf{66.63 \pm 2.03}$ | $\mathbf{61.18 \pm 1.67}$ | $\mathbf{56.76 \pm 1.85}$ |

Table 2: The MTA with the Dirichlet Non-IID data setting ($\alpha = 0.1$), where bold indicates the best result among all methods.

on all datasets, since its global aggregation will contain models of data-irrelevant clients and thus lead to severe instability in the gradient optimization process [Zhang *et al.*, 2020b].

For the other PFL methods, FedAvg+FT, pFedMe, FedFomo, FedAMP, pFedHN and our pFedSV all realize a promising performance on all datasets. FedAvg+FT takes several local fine-tuning steps to tune the poor global model back to adapt the local Non-IID data distribution. The pFedMe proposes novel regularized loss functions based on Moreau envelopes to decouple the personalized optimization from the global model learning. The pFedHN is more specific in that it directly generates personalized parameters for each client's model through another hypernetwork. The good performance of FedFomo and FedAMP are achieved by adaptively encourage more pairwise collaboration between clients with similar models to form their own personalized model. Our pFedSV can outperform all other baselines since it consider the multiwise influences among clients to help them identify their data-relevant coalition and generate personalized aggregation weights with multiwise collaboration.

Table 2 illustrates the MTA of all methods on the Dirichlet Non-IID data setting ($\alpha = 0.1$). As we know from the visualization in Fig. 3, this setting (Dirichlet $\alpha = 0.1$) is much more challenge than pathological, which is reflected in the significant performance reduction of all methods. Nevertheless, our pFedSV is still guaranteed to outperform all other baselines.

**Relevance score & Multiwise collaboration weights.** The superior performance of pFedSV comes from the various desirable properties of SV, Fig. 3 visualize the relevance vector $\phi^i$ of each client after convergence on different algorithms, where FedFomo in Fig. 3(b) use the model similarity based weights to update the relevance vector while pFedSV in Fig. 3(a) use the computed SV. To illustrate the effectiveness of pFedSV, we also attach the visualized ground-truth of client relevance in Fig. 3(c) according to the client label distribution in Fig. 3(d) obtained from the omniscience perspective. Obviously, it is evident from ground-truth that symmetry is an important feature of client relevance matrix, pFedSV perfectly identifies the real data-relevant clients and assigns aggregation weights with multiwise collaboration in the coalition, which significantly outperforms other pairwise collaboration methods based on model similarity.

## 6 Discussion

We note the communication overhead and privacy concern that may arise from the model downloading required for each client to perform local SV evaluation. In this section, we will discuss the effective solutions for these potential issues.

**Communication Overhead.** To reduce the communication overhead of downloading other client models for local SV evaluation, except our dynamic download mechanism in *SV for Data Relevance* of Sec. 3.3, we further exploit the advantage of common representation between clients. Specifically, we note that for some learning tasks (i.e., image classification and next-word-prediction), some low-dimensional common representation with similar functions can be shared, such as feature extraction. The parts that each client really need to personalize are their unique local heads, such as classifier [Collins *et al.*, 2021]. Therefore, we only need to download other clients' local heads to generate respective models, which can significantly reduce the communication overhead. Please refer to the Appendix.2 for more experiment details.

**Model Privacy.** Although we can achieve anonymity by removing any information related to the client's identity from the downloaded model parameters during the entire process of pFedSV, it still may pose privacy issues in real-world scenarios. Therefore, we consider to adopt the $(\epsilon, \delta)$-differential privacy (DP) to address the privacy concerns [Abadi *et al.*, 2016]. We can add Gaussian noise into the model parameters after client's local training process, which can guarantee the model with DP. The additional Gaussian noise can make the model private but at the cost of performance degradation. Please refer to the Appendix.3 for more experiment details.

## 7 Conclusion

In this paper, we focus on the fact that model aggregation is a collaboration process within the coalition and first introduce SV from coalition game theory, to analyze the multiwise influences among clients and quantify their marginal contribution to the personalization of others. We propose a novel algorithm pFedSV to simultaneously address two critical challenges in PFL: identifying the optimal collaborator coalition and generating personalized aggregation weights based on multiwise influences analysis. We conducted extensive experiments to demonstrate the effectiveness of pfedSV and the results empirically illustrate the superiority of multiwise collaboration through the significant improvement on the personalized accuracy.

## References

[Abadi *et al.*, 2016] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[Castro *et al.*, 2009] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

[Collins *et al.*, 2021] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. *arXiv preprint arXiv:2102.07078*, 2021.

[Cortes and Mohri, 2014] Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

[Donahue and Kleinberg, 2021] Kate Donahue and Jon Kleinberg. Model-sharing games: Analyzing federated learning under voluntary participation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5303–5311, 2021.

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

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

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

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

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

[LeCun, 1998] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[Li *et al.*, 2019] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019.

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

[Maleki *et al.*, 2013] Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. Bounding the estimation error of sampling-based shapley value approximation. *arXiv preprint arXiv:1306.4265*, 2013.

[Mann and Shapley, 1962] Irwin Mann and Lloyd S Shapley. Values of large games. 6: Evaluating the electoral college exactly. Technical report, RAND CORP SANTA MONICA CA, 1962.

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

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

[Myerson, 2013] Roger B Myerson. *Game theory*. Harvard university press, 2013.

[Shamsian *et al.*, 2021] Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. *arXiv preprint arXiv:2103.04628*, 2021.

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

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

[Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[Zhang *et al.*, 2020a] Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. In *International Conference on Learning Representations*, 2020.

[Zhang *et al.*, 2020b] Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. Fedpd: A federated learning framework with optimal rates and adaptivity to non-iid data. *arXiv preprint arXiv:2005.11418*, 2020.

[Zhao *et al.*, 2018] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.