Connecting Low-Loss Subspace for Personalized Federated Learning

Seok-Ju Hahn
seokjuhahn@unist.ac.kr
Ulsan National Institute of Science and Technology & Kakao Enterprise
South Korea

Minwoo Jeong
lloyd.ai@kakaoenterprise.com
Kakao Enterprise
South Korea

Junghye Lee
junghyelee@unist.ac.kr
Ulsan National Institute of Science and Technology
South Korea

ABSTRACT
Due to the curse of statistical heterogeneity across clients, adopting a personalized federated learning method has become an essential choice for the successful deployment of federated learning-based services. Among diverse branches of personalization techniques, a model mixture-based personalization method is preferred as each client has their own personalized model as a result of federated learning. It usually requires a local model and a federated model, but this approach is either limited to partial parameter exchange or requires additional local updates, each of which is helpless to novel clients and burdensome to the client’s computational capacity. As the existence of a connected subspace containing diverse low-loss solutions between two or more independent deep networks has been discovered, we combined this interesting property with the model mixture-based personalized federated learning method for improved performance of personalization. We proposed SuPerFed, a personalized federated learning method that induces an explicit connection between the optima of the local and the federated model in weight space for boosting each other. Through extensive experiments on several benchmark datasets, we demonstrated that our method achieves consistent gains in both personalization performance and robustness to problematic scenarios possible in realistic services.

CCS CONCEPTS
• Computing methodologies → Distributed algorithms; Neural networks.

KEYWORDS
federated learning, non-IID data, personalization, personalized federated learning, label noise, mode connectivity

ACM Reference Format:
Seok-Ju Hahn, Minwoo Jeong, and Junghye Lee. 2022. Connecting Low-Loss Subspace for Personalized Federated Learning. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539254

Figure 1: Overview of the model mixture-based personalized federated learning.

1 INTRODUCTION
Individuals and institutions are now data producers as well as data keepers, thanks to advanced communication and computation technologies. Therefore, training a machine learning model in a data-centralized setting is sometimes not viable due to many realistic limitations, such as the existence of massive clients generating their own data in real-time or the data privacy issue restricting the collection of data. Federated learning (FL) [46] is a solution to solve this problematic situation as it enables parallel training of a machine learning model across clients or institutions without sharing their private data, usually under the orchestration of the server (e.g., service providers). In the most common FL setting, such as FedAvg [46], when the server prepares and broadcasts a model appropriate for a target task, each participant trains the model with its own data and transmits the resulting model parameters to the server. Then, the server receives the locally updated model parameters to aggregate (i.e., weighted averaging) into a new global model. The new global model is broadcast again to some fraction of participating clients. This collaborative learning process is repeated until convergence. However, obtaining a single global model through FL is not enough to provide satisfiable experiences to all clients, due to the curse of data heterogeneity. Since there exists an inherent difference among clients, the data residing in each client is not statistically identical. That is, a different client has their own different data distribution, or the data is not independent and identically distributed (non-IID). Thus, applying a simple FL method like FedAvg cannot avoid a pitfall of the high generalization
error or even the divergence of model training [32]. That is, it is sometimes possible that the performance of the global model is lower than the locally trained model without collaboration. When this occurs, clients barely have the motivation to participate in the collaborative learning process to train a single global model.

Hence, it is natural to rethink the scheme of FL. As a result, personalized FL (PFL) has become an essential component of FL, which has yielded many related studies. Many approaches are proposed for PFL, based on multi-task learning [12, 44, 51], regularization technique [11, 23, 38]; meta-learning [14, 31], clustering [20, 43, 49], and a model mixture method [2, 8, 10, 38, 41, 43]. We focus on the model mixture-based PFL method in this paper as it shows promising performance as well as lets each client have their own model, which is a favorable trait in terms of personalization as each participating client eventually has their own model. This method assumes that each client has two distinct parts: a local (sub-)model and a federated (sub-)model \(^1\) (i.e., federated (sub-)model is a global model transferred to a client). In the model mixture-based PFL method, it is expected that the local model captures the information of heterogeneous client data distribution by staying only at the client-side, while the federated model focuses on learning common information across clients by being communicated with the server. Each federated model built at the client is sent to the server, where it is aggregated into a single global model (Figure 1). At the end of each PFL iteration, a participating client ends up having a personalized part (i.e., a local model), which potentially relieves the risks from the non-iidness as well as benefits from the collaboration of other clients.

The model mixture-based PFL method has two sub-branches. One is sharing only a partial set of parameters of one single model as a federated model [2, 8, 41] (e.g., only exchanging weights of a penultimate layer), the other is having two identically structured models as a local and a federated model [10, 11, 23, 38, 43]. However, each of them has some limitations. In the former case, new clients cannot immediately exploit the model trained by FL, as the server only has a partial set of model parameters. In the latter case, all of them require separate (or sequential) updates for the local and the federated model each, which is possibly burdensome to some clients having low computational capacity.

Here comes our main motivation: the development of a model mixture-based PFL method that can jointly train a whole model communicated with the server and another model for personalization. We attempted to achieve this goal through the lens of connectivity. As many studies on the deep network loss landscape have progressed, one intriguing phenomenon is being actively discussed in the field: the connectivity [13, 15, 16, 18, 19, 54] between deep networks. Though deep networks are known to have many local minima, it has been revealed that each of these local minima has a similar performance to each other [7]. Additionally, it has recently been discovered that two different local minima obtained by two independent deep networks can be connected through the linear path [17] or the non-linear path [13, 19] in the weight space, where all weights on the path have a low loss. Note that the two endpoints of the path are two different local minima reached from the two different deep networks. These findings have been extended to a multi-dimensional subspace [4, 54] with only a few gradient steps on two or more independently initialized deep networks. The resulting subspace, including the one-dimensional connected path, contains functionally diverse models having high accuracy. It can also be viewed as an ensemble of deep networks in the weight space, thereby sharing similar properties such as good calibration performance and robustness to the label noise. By introducing the connectivity to a model mixture-based PFL method, such good properties can also be absorbed advantageously.

**Our contributions.** Inspired by this observation, we propose SuPerFed, a connected low-loss subspace construction method for the personalized federated learning, adapting the concept of connectivity to the model mixture-based PFL method. This method aims to find a low-loss subspace between a single global model and many different local models at clients in a mutually beneficial way. Accordingly, we aim to achieve better personalization performance while overcoming the aforementioned limitations of model mixture-based PFL methods. Adopting the connectivity to FL is non-trivial as it obviously differs from the setting where the connectivity is first considered: each deep network observes different data distributions and only the global model can be connected to other local models on clients. The main contributions of the work are as follows:

- We propose SuPerFed, a model mixture-based PFL method that aims to achieve good personalization performance of each local model on clients by harmoniously connecting them with the global model as a pivot, under diverse non-iid scenarios.
- Our proposed method extends the limitations of existing model mixture-based PFL methods by jointly updating both federated and local models in clients and exchanging the entire federated model with the server, allowing novel clients to benefit from it.
- Personalized models trained by adopting our method are well-calibrated and also robust to potential label noise that is a common problem in realistic FL-driven services.

## 2 RELATED WORKS

### 2.1 FL with Non-IID Data

After [46] proposed the basic FL algorithm (FedAvg), handling non-iid data across clients is one of the major points to be resolved in the FL field. While some methods are proposed such as sharing a subset of client’s local data at the server[56], accumulating previous model updates at the server [26]. These are either unrealistic assumptions for FL or not enough to handle a realistic level of statistical heterogeneity. Other branches to aid the stable convergence of a single global model include modifying a model aggregation method at the server [40, 48, 53, 55] and adding a regularization to the optimization [1, 32, 39]. However, a single global model may still not be sufficient to provide a satisfactory experience to clients using FL-driven services in practice.

### 2.2 PFL Methods

As an extension of the above, PFL methods shed light on the new perspective of FL. PFL aims to learn a client-specific personalized
Figure 2: An illustration of the proposed method SuPerFed. (A) Local update of SuPerFed: at every federated learning round, a selected client receives a global model from the server and sets it to be a federated model. After being mixed by a randomly generated $\lambda$, two models are jointly updated with regularization. (B) Communication with the server: only the updated federated model is uploaded to the server (dotted arrow in crimson) to be aggregated (e.g., weighted averaging) as a new global model, and it is broadcast to clients in the next round (arrow in gray). (C) Exploitation of trained models: (Case 1) FL clients can sample and use any model on the connected subspace (e.g., $W(0.2)$, $W(0.8)$) because it only contains low-loss solutions. (Case 2) Novel clients can download and use the server’s trained global model $w^g$.

model, and many methodologies for PFL have been proposed. Multi-task learning-based PFL [12, 44, 51] treats each client as a different task and learns a personalized model for each client. Local fine-tuning-based PFL methods [14, 31] adopt a meta-learning approach for a global model to be adapted promptly to a personalized model for each client, and clustering-based PFL methods [20, 43, 49] mainly assume that similar clients may reach a similar optimal global model. Model mixture-based PFL methods [2, 8, 10, 11, 23, 38, 41] divide a model into two parts: one (i.e., a local model) for capturing local knowledge, and the other (i.e., a federated model) for learning common knowledge across clients. In these methods, only the federated model is shared with the server while the local model resides locally on each client. [2] keeps weights of the last layer as a local model (i.e., the personalization layers), and [8] is similar to this except it requires a separate update on a local model before the update of a federated model. In contrast, [41] retains lower layer weights in clients and only exchanges higher layer weights with the server. Due to the partial exchange, new clients in the scheme of [2, 8, 41] should train their models from scratch or need at least some steps of fine-tuning. In [10, 11, 23, 38, 43], each client holds at least two separate models with the same structure: one for a local model and the other for a federated model. In [10, 23, 43], they are explicitly interpolated in the form of a convex combination after the independent update of two models. In [11, 38], the federated model affects the update of the local model in the form of proximity regularization.

2.3 Connectivity of Deep Networks

The existence of either linear path [17] or non-linear path [13, 19] between two minima derived by two different deep networks has been recently discovered through extensive studies on the loss landscape of deep networks [13, 15, 16, 18, 19]. That is, there exists a low-loss subspace (e.g., line, simplex) connecting two or more deep networks independently trained on the same data. Though there exist some studies on constructing such a low loss subspace [5, 28–30, 36, 42], they require multiple updates of deep networks. After it is observed that independently trained deep networks have a low cosine similarity with each other in the weight space [15], a recent study by [54] proposes a straightforward and efficient method for explicitly inducing linear connectivity in a single training run. Our method adopts this technique and adapts it to construct a low-loss subspace between each local model and a federated model (which will later be aggregated into a global model at the server) suited for an effective PFL in heterogeneously distributed data.

3 PROPOSED METHOD

3.1 Overview

In the standard FL scheme [46], the server orchestrates the whole learning process across participating clients with iterative communication of model parameters. Since our method is essentially a model mixture-based PFL method, we need two models same in structure per client (one for a federated model, the other for a local model), of which initialization is different from each other. As a result, the client’s local models are trained using different random initializations than its federated model counterpart. Note that different initialization is not common in FL [46]; however, this is intended behavior in our scheme for promoting construction of
We first define arguments required for the federated optimization as follows: a total number of communication rounds \( R \), a personalization start round \( L \), a local batch size \( B \), a number of local epochs \( E \), a total number of clients \( K \), a fraction of clients selected at each round \( \lambda \), and a learning rate \( \eta \). Each parameter \( \mu, \nu, \gamma \) is a regularization constant, \( \gamma \) is a constant for inducing connectivity along a subspace between local and federated models \( w^f \) and \( w^f \).

### 3.3 Problem Statement

Consider that each client \( c_i \) \((i \in [K])\) has its own dataset \( \mathcal{D}_i = \{(x_i, y_i)\} \) of size \( n_i \), as well as a model \( W_i \). We assume that datasets are non-IID across clients in PFL, which means that each dataset \( \mathcal{D}_i \) is sampled independently from a corresponding distribution \( \mathcal{P}_i \) on both input and label spaces \( \mathcal{X} \times \mathcal{Y} \). A hypothesis \( h \in \mathcal{H} \) can be learned by the objective function \( l : \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}^+ \) in the form of \( l(h(x; W), y) \). We denote the expected loss as \( \mathcal{L}_{\mathcal{P}_i}(h_i, W_i) = \mathbb{E}_{(x_i, y_i) \sim \mathcal{P}_i} [l(h_i(x_i; W_i), y_i)] \), and the empirical loss as \( \hat{\mathcal{L}}_{\mathcal{D}_i}(h_i, W_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} l(h_i(x_j^i; W_i), y_j^i) \); a regularization term \( \Omega(W_i) \) can also be incorporated here. Then, the global objective of PFL is to optimize (1).

\[
\min_{h_i \in \mathcal{H}} \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}_{\mathcal{P}_i}(h_i, W_i) \tag{1}
\]

We can minimize this global objective through the empirical loss minimization (2).

\[
\min_{W_1, \ldots, W_K} \frac{1}{K} \sum_{i=1}^{K} \hat{\mathcal{L}}_{\mathcal{D}_i}(h_i, W_i) = \frac{1}{K} \sum_{i=1}^{K} \left[ \frac{1}{n_i} \sum_{j=1}^{n_i} l(h_i(x_j^i; W_i), y_j^i) + \Omega(W_i) \right] \tag{2}
\]

A model mixture-based PFL assumes each client \( c_i \) has a set of paired parameters \( w_i^f, w_i^f \leq W_i \). Note that each of which is previously defined as a federated model and a local model. Both of them are grouped as a client model \( W_i = G(w_i^f, w_i^f) \) with the grouping operator \( G \). This operator can be a concatenation (e.g., stacking semantically separated layers like feature extractor or classifier [2, 8, 41]):

\[
G(w_i^f, w_i^f) = \text{Concat}(w_i^f, w_i^f), \text{ a simple enumeration of parameters having the same structure [11, 38]}: G(w_i^f, w_i^f) = \{w_i^f, w_i^f\}, \text{ or a convex combination [10, 23, 43]} \text{ given a constant } \lambda \in [0, 1]: G(w_i^f, w_i^f, \lambda) = (1 - \lambda)w_i^f + \lambda w_i^f.
\]

### 3.4 Local Update

In SuPerFed, we suppose both models \( w_i^f, w_i^f \) have the same structure and consider \( G \) to be a function for constructing a convex combination. We will write \( \lambda \) as an explicit input to the client model with a slight abuse of notation: \( W_i(\lambda) := G(w_i^f, w_i^f; \lambda) = (1-\lambda)w_i^f + \lambda w_i^f \). The main goal of the local update in SuPerFed is to find a flat wide minima between the local model and the federated model. In the initial round, each client receives a global model transmitted from the server, sets it as a federated model, and copies the structure to set a local model. As different initialization is required, clients initialize their local model again with its fresh random seed. During the local update, a client mixes two models using \( \lambda \sim \text{Unif}(0, 1) \) as \( W_i(\lambda) = (1 - \lambda)w_i^f + \lambda w_i^f \) each time with a fresh mini-batch including \( (x, y) \) sampled from local dataset \( \mathcal{D}_i \). The local model and the federated model are reduced to be interpolated diversely during local training. Finally, it is optimized with a loss function
\( l(\cdot) \) specific to the task, for example, cross-entropy, with additional regularization \( \Omega(\cdot) \) that is critical to induce the connectivity.

**Remark 1.** Connectivity \([17, 54]\). There exists a low-loss subspace spanned by \( W(\lambda) \) between two independently trained (i.e., trained from different random initializations) deep networks \( w_1 \) and \( w_2 \), in the form of a linear combination formed by \( \lambda : \mathbf{W}(\lambda) = (1-\lambda)w_1 + \lambda w_2, \lambda \in [0, 1] \).

The local objective is

\[
\min_{W_i} \hat{L}_D(h, W_i(\lambda))
\]

where \( \hat{L}_D(h, W_i(\lambda)) \) is denoted as:

\[
\hat{L}_D(h, W_i(\lambda)) = \frac{1}{|D|} \sum_{(x,y) \in D} l(h(x; W_i(\lambda)), y) + \Omega(W_i(\lambda))
\]

\[
= \frac{1}{|D|} \sum_{(x,y) \in D} l(h(x; W_i(\lambda)), y) + \mu \left\| w_i^f - w^g \right\|_2^2 + \nu \cos^2 \left( \frac{w_i^f}{w_i} \right)
\]

Note that the global model \( w^g \) is fixed here for the proximity regularization of \( w_i^f \). Denote the local objective (3) simply as \( \hat{L} \). Then, the update of each endpoint \( w_i^f \) and \( w_i^l \) is done using the following estimates. Note that two endpoints (i.e., \( W_i(0) = w_i^f \) and \( W_i(1) = w_i^l \)) can be jointly updated at the same time in a single training run.

\[
\frac{\partial \hat{L}}{\partial w_i^f} = (1-\lambda) \frac{\partial l}{\partial W_i(\lambda)} + \mu \frac{\partial \left\| w_i^f - w^g \right\|_2^2}{\partial w_i^f} + \nu \frac{\partial \cos^2 \left( \frac{w_i^f}{w_i} \right)}{\partial w_i^f}
\]

\[
\frac{\partial \hat{L}}{\partial w_i^l} = \lambda \frac{\partial l}{\partial W_i(\lambda)} + \nu \frac{\partial \cos^2 \left( \frac{w_i^f}{w_i} \right)}{\partial w_i^l}
\]

### 3.5 Regularization

The regularization term \( \cos(\cdot, \cdot) \) stands for cosine similarity defined as: \( \cos(w_i^f, w_i^l) = \langle w_i^f, w_i^l \rangle / \|w_i^f\| \|w_i^l\| \). This aims to induce connectivity between the local model and the federated model, thereby allowing the connected subspace between them to contain weights yielding high accuracy for the given task. Its magnitude is controlled by the constant \( \nu \). Since merely applying the mixing strategy between the local and the federated models has little benefit \([54]\), it is required to set \( \nu > 0 \). In \([15, 54]\), it is observed that weights of independently trained deep networks show dissimilarity in terms of cosine similarity, which has functional diversity as a result. Moreover, in \([3, 45]\), inducing orthogonality (i.e., forcing cosine similarity to zero) between weights prevents deep networks from learning redundant features given their learning capacity. In this context, we expect the local model of each client \( w_i^f \) to learn client-specific knowledge while the federated model \( W_i^f \) learns client-agnostic knowledge in a complementary manner and is harmoniously combined with the local model. The other regularization term adjusted by \( \mu \) is \( L2 \)-norm which controls the proximity of the update of the federated model \( w_i^f \) from the global model \( w^g \) \([39]\). By doing so, we expect each local update of the federated model not to be derailed from the global model, which prevents divergence of the aggregated global model during the process of FL. Note that an inordinate degree of proximity regularization can hinder the local model’s adaptation and thereby cause the divergence of a global model \([11, 39]\). Thus, a moderate degree of \( \mu \) is required. Note that our method is designed to be reduced to the existing FL methods by adjusting hyperparameters, and thus it can be guaranteed to have at least their performance. In detail, when fixing \( \lambda = 0, \nu = 0, \mu = 0 \), the objective of SuPerFed is reduced to FedAvg \([46]\), when \( \lambda = 0, \nu = 0, \mu > 0 \), it is equivalent to FedProx \([39]\).

### 3.6 Communication with the Server

At each round, the server samples a fraction of clients at every round (suppose \( N \) clients are sampled), and requests a local update to each client. Only the federated model \( w_i^f \) and the amounts of consumed training data per client \( n_i \) are transferred to the server after the local update is completed in parallel across selected clients. The server then aggregates them to create an updated global model \( w^g \) as \( w^g = \frac{1}{\sum n_i} \sum n_i w_i^f \). In the next round, only the updated global model \( w^g \) is transmitted to another group of selected clients, and it is set to be a new federated model at each client: \( w_i^f = w^g \). It is worth noting that the communication cost remains constant as with the single model-based FL methods \([39, 46]\). See Algorithm 2 for understanding the whole process of SuPerFed.

### 3.7 Variation in Mixing Method

Until now, we only considered the setting of applying \( \lambda \) identically to the whole layers of \( w^f \) and \( W \). We name this setting \textit{Model Mixing}, in short, \textit{MM}. On the one hand, it is also possible to sample different \( \lambda \) for weights of different layers, i.e., mixing two models in a layer-wise manner. We also adopt this setting and name it \textit{Layer Mixing}, in short, \textit{LM}.

### Table 1: Experimental results on the pathological non-IID setting (MNIST and CIFAR10 datasets) compared with other FL and PFL methods. The top-1 accuracy is reported with a standard deviation.

| Dataset   | MNIST     | CIFAR10   |
|-----------|-----------|-----------|
|           | 100 clients | 100 clients | 100 clients | 100 clients |
|           | 50 | 100 | 500 | 100 | 500 | 100 | 500 | 100 | 500 |
| FedAvg    | 98.50 ± 0.04 | 99.90 ± 0.45 | 99.44 ± 0.43 | 97.32 ± 0.13 | 95.60 ± 0.18 | 46.16 ± 0.56 |
| FedProx   | 94.87 ± 2.27 | 93.91 ± 4.45 | 96.33 ± 4.24 | 93.08 ± 2.25 | 94.50 ± 1.37 | 43.95 ± 6.27 |
| SCAFFOLD  | 96.71 ± 1.42 | 95.41 ± 3.99 | 75.43 ± 1.45 | 58.99 ± 1.40 | 59.05 ± 2.84 | 43.21 ± 2.84 |
| LbFedAvg   | 98.15 ± 2.08 | 96.15 ± 3.00 | 95.92 ± 4.80 | 90.97 ± 1.09 | 97.75 ± 5.92 | 95.71 ± 13.38 |
| LbFedProx  | 99.17 ± 4.45 | 99.20 ± 1.00 | 98.36 ± 3.02 | 90.95 ± 4.48 | 97.89 ± 6.08 | 92.57 ± 11.06 |
| APFL      | 99.34 ± 0.60 | 98.59 ± 8.78 | 98.72 ± 2.72 | 92.92 ± 3.35 | 95.96 ± 4.77 | 91.58 ± 7.51 |
| APFL-Side  | 97.85 ± 1.67 | 97.43 ± 2.44 | 96.69 ± 7.70 | 93.22 ± 3.76 | 92.80 ± 4.71 | 90.89 ± 4.81 |
| Ditto     | 97.28 ± 1.99 | 97.68 ± 1.35 | 96.39 ± 2.39 | 93.22 ± 2.78 | 90.09 ± 6.46 | 91.58 ± 7.51 |
| FedRep    | 99.90 ± 0.94 | 99.92 ± 1.00 | 98.33 ± 4.42 | 91.17 ± 6.08 | 96.04 ± 12.89 | 79.78 ± 11.95 |
| SuPerFed-MM| 98.37 ± 0.64 | 99.42 ± 0.85 | 98.14 ± 2.12 | 94.02 ± 3.32 | 95.36 ± 3.51 | 93.71 ± 6.59 |
| SuPerFed-MM | 99.31 ± 0.64 | 99.25 ± 1.00 | 98.72 ± 2.50 | 94.05 ± 3.59 | 95.24 ± 4.09 | 93.12 ± 6.59 |

### 4 EXPERIMENTS

#### 4.1 Setup

We focus on two points in our experiments: (i) personalization performance in various non-IID scenarios, (ii) robustness to label noise.
Table 2: Experimental results on the Dirichlet distribution-based non-IID setting (CIFAR100 and TinyImageNet) compared with other FL and PFL methods. The top-5 accuracy is reported with a standard deviation.

| Dataset          | CIFAR100 | TinyImageNet |
|------------------|----------|--------------|
|                  | 1        | 100          | 100 |
| CIFARAvg         | 51.3 ± 0.84 | 52.8 ± 0.78 | 52.8 ± 0.73 |
| FedProx          | 51.60 ± 0.84 | 54.67 ± 0.45 | 54.67 ± 0.45 |
| SCAFFOLD         | 52.78 ± 0.49 | 53.1 ± 0.56 | 54.68 ± 0.28 |
| FedAvg           | 58.85 ± 0.27 | 58.92 ± 0.27 | 59.05 ± 0.25 |
| SuPerFed-LM      | 56.71 ± 0.21 | 47.78 ± 0.65 | 47.00 ± 0.72 |
| SuPerFed-MM      | 49.96 ± 0.41 | 36.13 ± 0.52 | 35.80 ± 0.17 |
| FedRep           | 24.75 ± 10.79 | 25.62 ± 15.95 | 24.41 ± 13.55 |
| Ditto            | 56.65 ± 0.57 | 56.76 ± 0.56 | 56.53 ± 0.52 |
| FedPer           | 47.10 ± 0.10 | 47.96 ± 0.27 | 47.96 ± 0.47 |
| SuPerFed-LM      | 58.18 ± 0.54 | 57.81 ± 0.87 | 56.81 ± 0.12 |
| SuPerFed-MM      | 59.43 ± 0.94 | 59.54 ± 0.11 | 59.57 ± 0.32 |

Table 3: Experimental results on the realistic non-IID setting (FEMINIST and Shakespeare) compared with other FL and PFL methods. The top-1 accuracy is reported with a standard deviation.

| Dataset          | FEMINIST | Shakespeare |
|------------------|----------|-------------|
|                  | 700      | 600         |      |
|                  | Top-1    | Top-5       |      |
|                  |          | Top-1       | Top-5 |
| FedAvg           | 82.93 ± 10.09 | 98.57 ± 2.88 | 53.36 ± 0.94 | 83.27 ± 6.14 |
| FedProx          | 81.40 ± 10.30 | 98.40 ± 3.15 | 53.14 ± 0.75 | 83.67 ± 5.70 |
| SCAFFOLD         | 81.33 ± 10.41 | 98.56 ± 2.94 | 50.73 ± 7.35 | 80.86 ± 7.09 |
| LG-FedAvg        | 48.96 ± 21.99 | 74.17 ± 20.49 | 29.35 ± 10.88 | 61.09 ± 15.65 |
| FedPer           | 71.99 ± 12.29 | 84.83 ± 10.02 | 50.07 ± 9.09 | 79.75 ± 9.27 |
| APFL             | 84.31 ± 10.90 | 98.41 ± 3.01 | 58.23 ± 6.84 | 86.71 ± 4.52 |
| pFedMe           | 6.64 ± 4.75 | 27.82 ± 12.53 | 19.19 ± 5.53 | 47.97 ± 6.50 |
| Ditto            | 81.00 ± 17.09 | 96.61 ± 15.46 | 49.12 ± 13.88 | 78.34 ± 18.30 |
| FedRep           | 80.65 ± 14.95 | 76.14 ± 14.02 | 45.72 ± 11.95 | 75.50 ± 15.52 |
| SuPerFed-MM      | 85.22 ± 2.84 | 98.76 ± 2.74 | 58.82 ± 0.61 | 88.41 ± 3.72 |
| SuPerFed-LM      | 83.90 ± 5.23 | 98.27 ± 3.11 | 59.17 ± 5.72 | 87.01 ± 4.18 |

through extensive benchmark datasets. Details of each experiment are provided in Appendix A. Throughout all experiments, if not specified, we set 5 clients to be sampled at every round and used stochastic gradient descent (SGD) optimization with a momentum of 0.9 and a weight decay factor of 0.0001. As SuPerFed is a model mixture-based PFL method, we compare its performance to other model mixture-based PFL methods: FedPer [2], LG-FedAvg [41], APFL [10], pFedMe [11], Ditto [38], FedRep [8], along with basic single-model based FL methods, FedAvg [46] FedProx [39] and SCAFFOLD [32]. We let the client randomly split their data into a training set and a test set with a fraction of 0.2 to estimate the performance of each FL algorithm on each client’s test set with its own personalized model (if possible) with various metrics: top-1 accuracy, top-5 accuracy, expected calibration error (ECE [21]), and maximum calibration error (MCE [34]). Note that we evaluated all participating clients after the whole round of FL is finished. For each variation of the mixing method (i.e., LM & MM) stated in section 3.7, we subdivide our method into two, SuPerFed-MM for Model Mixing, and SuPerFed-LM for Layer Mixing.

4.2 Personalization

For the estimation of the performance of SuPerFed as PFL methods, we simulate three different non-IID scenarios. (i) a pathological non-IID setting proposed by [46], which assumes most clients have samples from two classes for a multi-class classification task. (2) Dirichlet distribution-based non-IID setting proposed by [27], in which the Dirichlet distribution with its concentration parameter \(\alpha\) determines the label distribution of each client. All clients have samples from only one class when using \(\alpha \to 0\), whereas \(\alpha \to \infty\) divides samples into an identical distribution. (3) Realistic non-IID setting proposed in [6], which provides several benchmark datasets for PFL.

Pathological non-IID setting. In this setting, we used two multi-class classification datasets, MNIST [9] and CIFAR10 [33] and both have 10 classes. We used the two-layered fully-connected network for the MNIST dataset and the two-layered convolutional neural network (CNN) for the CIFAR10 dataset as proposed in [46]. For each dataset, we set the number of total clients (\(K=50, 100, 500\)) to check the scalability of the PFL methods. The results of the pathological non-IID setting are shown in Table 1. It is notable that our proposed method beats most of the existing model mixture-based PFL methods with a small standard deviation.

Dirichlet distribution-based non-IID setting. In this setting, we used other multi-class classification datasets to simulate a more challenging setting than the pathological non-IID setting. We use CIFAR100[33] and TinyImageNet[37] datasets, having 100 classes and 200 classes each. We also selected other deep architectures, ResNet[24] for CIFAR10 and MobileNet[25] for TinyImageNet. For each dataset, we adjust the concentration parameter \(\alpha = 1, 10, 100\) to control the degree of statistical heterogeneity across clients. It is a more natural choice for simulating statistical heterogeneity than pathological non-IID setting, making it more challenging. Note that the smaller the \(\alpha\), the more heterogeneous each client’s data distribution is. Table 2 presents the results of the Dirichlet distribution-based non-IID setting. Both of our methods are less affected by the degree of statistical heterogeneity (i.e., non-IIDness) determined by \(\alpha\).

Realistic non-IID setting. We used FEMINIST and Shakespeare datasets in the LEAF benchmark [6]. As the purpose of these datasets is to simulate a realistic FL scenario, each dataset is naturally split for a certain data mining task. The selected two datasets are for multi-class classification (62 classes) and next character prediction (80 characters) given a sentence, respectively. We used VGG [50] for the FEMINIST dataset and networks with two stacked LSTM layers were used in [46] for the Shakespeare dataset. The results of the realistic non-IID setting are shown in Table 3. In this massively distributed setting, our methods showed a consistent gain in terms of personalization performance.

Robustness to label noise. In realistic FL services, it is possible that each client suffers from noisy labels. We anticipate that SuPerFed will benefit from the strength of ensemble learning because our method can be viewed as implicit ensemble of two models, \(w_f\) and \(w^t\). One of the strengths is robustness to the label noise, as each of the model components of the ensemble method can view diverse data distributions [35, 47]. In SuPerFed, the federated model and the local model are induced to be independent of each other (i.e., having low cosine similarity in the weight space). We assume each of them should inevitably learn from a different view of data distribution. To prove this, we adopted two representative methods to simulate the noisy labels: pair flipping [22] and symmetric flipping [52], and their detailed descriptions are in Appendix B. We
Table 4: Experimental results on the label noise simulation with Dirichlet distribution-based non-IID setting (α = 100) on (MNIST and CIFAR10) compared with other FL and PFL methods. Per each cell, expected calibration error (ECE), maximum calibration error (MCE), and the best averaged top-1 accuracy (in parentheses) are enumerated from top to bottom. Note that the lower ECE and MCE, the better the model calibration is.

| Dataset | MNIST | CIFAR10 |
|---------|-------|---------|
| Noise type | pair | symmetric | pair | symmetric |
| Noise ratio | 0.1 | 0.4 | 0.6 | 0.1 | 0.4 | 0.6 |
| FedAvg | 0.17 ± 0.03 | 0.38 ± 0.03 | 0.29 ± 0.04 | 0.42 ± 0.04 | 0.46 ± 0.04 | 0.57 ± 0.04 | 0.52 ± 0.05 | 0.59 ± 0.05 |
| FedProx | 0.58 ± 0.08 | 0.67 ± 0.04 | 0.66 ± 0.07 | 0.75 ± 0.05 | 0.80 ± 0.06 | 0.87 ± 0.04 | 0.81 ± 0.04 | 0.84 ± 0.04 |
| SCAFFOLD | 0.16 ± 0.03 | 0.38 ± 0.03 | 0.29 ± 0.03 | 0.42 ± 0.05 | 0.47 ± 0.05 | 0.70 ± 0.05 | 0.53 ± 0.05 | 0.59 ± 0.05 |
| LG-FedAvg | 0.23 ± 0.04 | 0.50 ± 0.03 | 0.34 ± 0.04 | 0.45 ± 0.05 | 0.59 ± 0.06 | 0.69 ± 0.05 | 0.63 ± 0.05 | 0.66 ± 0.05 |
| FedProx | 0.58 ± 0.07 | 0.78 ± 0.05 | 0.66 ± 0.07 | 0.74 ± 0.05 | 0.80 ± 0.05 | 0.87 ± 0.04 | 0.81 ± 0.05 | 0.84 ± 0.05 |
| APFL | 0.23 ± 0.03 | 0.42 ± 0.05 | 0.28 ± 0.05 | 0.40 ± 0.05 | 0.45 ± 0.06 | 0.60 ± 0.06 | 0.51 ± 0.06 | 0.56 ± 0.06 |
| Ditto | 0.22 ± 0.04 | 0.42 ± 0.05 | 0.31 ± 0.04 | 0.41 ± 0.04 | 0.54 ± 0.05 | 0.60 ± 0.06 | 0.56 ± 0.07 | 0.58 ± 0.06 |
| FedRep | 0.20 ± 0.04 | 0.53 ± 0.03 | 0.30 ± 0.05 | 0.43 ± 0.05 | 0.62 ± 0.05 | 0.76 ± 0.04 | 0.66 ± 0.05 | 0.64 ± 0.05 |
| SuPerFed-MM | 0.16 ± 0.03 | 0.28 ± 0.04 | 0.27 ± 0.03 | 0.30 ± 0.04 | 0.28 ± 0.06 | 0.27 ± 0.05 | 0.31 ± 0.06 | 0.28 ± 0.06 |
| SuPerFed-LM | 0.14 ± 0.03 | 0.35 ± 0.04 | 0.27 ± 0.04 | 0.32 ± 0.04 | 0.29 ± 0.04 | 0.40 ± 0.04 | 0.36 ± 0.05 | 0.37 ± 0.06 |

Table 5: Experiment on selecting L, the starting round of personalization. The best averaged top-1 accuracy (top) and loss (bottom) among models that can be realized by $\lambda \in [0, 1]$ is reported with a standard deviation.

| L/R | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 |
|-----|-----|-----|-----|-----|-----|
| SuPerFed-MM | 88.10 ± 0.25 | 92.13 ± 0.49 | 95.19 ± 0.60 | 91.64 ± 0.78 | 86.84 ± 14.88 |
| SuPerFed-LM | 90.89 ± 0.45 | 91.78 ± 0.53 | 92.10 ± 0.44 | 92.15 ± 0.03 | 89.60 ± 11.56 |

4.3 Ablation Studies

Constants for regularization terms. Our proposed method has two hyperparameters, $v$ and $\mu$. Each hyperparameter is supposed to do two things: (1) create a connected low-loss subspace between the federated model and the local model in clients by making the weights of the two models orthogonal; and (2) penalize local updates from going in the wrong direction, away from the direction of global updates. We ran experiments with $v = \{0.0, 1.0, 2.0, 5.0\}$ and $\mu = \{0.0, 0.01, 0.1, 1.0\}$ (a total of 16 combinations) under the Model Mixing and Layer Mixing regimes, with $R = 100$ and $L = [0.4R] = 40$, which yielded the best personalization performance on the CIFAR10 dataset with pathological non-IID Experimental results on these combinations are displayed in Figure 3.

There are noticeable observations supporting the efficacy of our method. First, a learning strategy to find a connected low loss
This implies that allowing some rounds to a federated model to
(i.e., a federated model (\(\lambda \in (0, 1)\)) outperforms two single models from both endpoints (i.e., a federated model (\(\lambda = 0\)) and a local model (\(\lambda = 1\)). Second, a moderate degree of orthogonality regularization (adjusted by \(\nu\)) boosts the personalization performance of the local model. Third, the proximity regularization restricts learning too much when \(\mu\) is large, but often helps a federated model not to diverge when it is properly selected. Since this phenomenon has already been pointed out in [11, 39], a careful selection of \(\mu\) is required.

Start round of personalization. We conducted an experiment on selecting \(L\), which determines the initiation of mixing operation, in other words, sampling \(\lambda \sim \text{Unif}(0, 1)\). As our proposed method can choose a mixing scheme from either Model Mixing or Layer Mixing, we designed separate experiments per scheme. We denote each scheme by SuperFed-MM and SuperFed-LM. By adjusting \(L\) by 0.1 increments from 0 to 0.9 of \(R\), we evaluated the top-1 accuracy of local models. This experiment was done only using the CIFAR10 dataset with \(R = 200\), \(v = 2\), and \(\mu = 0.01\). Experimental results are summarized in Table 5.

We see that there are some differences according to the mixing scheme. In terms of the personalization performance, setting \(L/R = 0.4\) shows the best performance with a small standard deviation. This implies that allowing some rounds to a federated model to concentrate on learning knowledge is a valid strategy. While \(L\) is a tunable hyperparameter, it can be adjusted for the purpose of FL. To derive a good global model, choosing a large \(L\) close to \(R\) is a valid choice, while in terms of personalization, it should be far from \(R\).

5 CONCLUSION

In this paper, we propose a simple and scalable personalized federated learning method SuperFed under the scheme of the model mixture method. Its goal is to improve personalization performance while also providing robustness to label noise. In each client, the method aims to build a connected low-loss subspace between a local and a federated model, inspired by the special property of deep networks, connectivity. Through this, two models are connected in a mutually cooperative manner so that each client has a personalized model with good performance and well-calibrated, which is robust to possible label noises in its dataset. With extensive experiments, we empirically demonstrated that SuperFed, with its two variants in the mixing scheme, outperforms the existing model mixture-based PFL methods and three basic FL methods.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MSIT) under Grant No. 2020R1C1C1011063 and by Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government (MSIT) (No. 2020-0-01336, Artificial Intelligence Graduate School support (UNIST)). This work is completed during internship at Kakao Enterprise.

REFERENCES

[1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. 2020. Federated learning based on dynamic regularization. In International Conference on Learning Representations.

[2] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aditya Kumar Singh, and Sunav Choudhary. 2019. Federated learning with personalization layers. arXiv preprint arXiv:1912.00818 (2019).

[3] Babajide O Ayinde, Tamer Inanc, and Jacek M Zurada. 2019. Regularizing deep neural networks by enhancing diversity in feature extraction. IEEE transactions on neural networks and learning systems 30, 9 (2019), 2650–2661.

[4] Gregory Benton, Wesley Maddox, Sanae Lotfi, and Andrew Gordon Gordon Wilson. 2021. Loss surface simplices for mode connecting volumes and fast ensembling. In International Conference on Machine Learning. PMLR, 769–779.

[5] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In International Conference on Machine Learning. PMLR, 1613–1622.

[6] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H Brendan McMahahan, Virginia Smith, and Ameet Talwalkar. 2018. Leaf: A benchmark for federated settings. arXiv preprint arXiv:1812.01097 (2018).
A APPENDIX

A.1 Experimental Details
Table A1 summarizes the details of each experiment: the number of clients K, total rounds R, local batch size B, number of local epochs E, learning rate η, and the network. Note that the fraction of sampled clients, C is adjusted to be 5 clients, and the start of personalization L is set to 40% of total rounds. For a stable convergence of a global model, we applied applied learning rate decay by 1% per each round. Each network architecture is borrowed from the original paper it proposed: TwoNN, TwoCNN, StackedLSTM [46], ResNet [24], MobileNet [25], and VGG [50].

A.2 Loss Landscape of a Global Model
The loss landscape of global models obtained from the proposed methods SuperFed-MM and SuperFed-LM are visualized in Figure A1 along with one from FedAvg by following [19]. As the proposed regularization term adjusted by ν induces connected low-loss subspace, it is natural that the wider minima is acquired from both SuperFed-MM and SuperFed-LM compared to that of FedAvg. As the global model learns to be connected with local models in weight spaces per each federation round, flat minima of the global model are induced, i.e., the connectivity is observed.

One thing to note here is that the minima is far wider in layer-wise mixing scheme (i.e., SuperFed-LM) than that of model-wise mixing scheme (i.e., SuperFed-MM). It is presumably due to the fact that the diversity of model combinations is richer in the layer-wise mixing scheme than in the model-wise mixing scheme. In other words, the global model should be connected to the larger number of different models.

Meanwhile, since the global model has wide minima, it can be expected that it is friendly to be exploited by unseen clients even if their data distribution is heterogeneous from existing clients’ data distributions. It is because the global model contains many candidates of low-loss solutions thanks to the induced connectivity, the chances of finding a suitable model are higher than the global model derived from other FL methods.

A.3 Difference from the Existing PFL Methods with Model Interpolation
In previous model mixture-based PFL methods, there exist similar approaches that interpolate the local model and the federated model in the form of a linear combination of a federated model and a local model: (i) APFL [10], (ii) mapper [43] and (iii) L2GD [23].

In (i) APFL, the federated model and the local model are updated separately with a fixed λ, which is not true in SuperFed as it jointly updates both federated and local models at once. While authors of [10] proposed the method to dynamically tune λ through another step of gradient descent, SuperFed does not require finding one optimal λ. It aims to find a good set of models wᵢ and wℓ that can be combined to yield good performance regardless of the value of λ.

In (ii) mapper [43], on the other hand, it iteratively selects only one client (an unrealistic assumption) and requests to find the best local models wᵢ and wℓ based on the transmitted global model wθ. The federated model is then updated with the the local model and the optimal λ is tuned. In SuperFed, it is not required to find a fixed value of λ and updates of both local and federated models can be completed in a single run of back-propagation using gradient descent. No separate training on each model is required as in other methods.

In (iii) L2GD proposed by [23], the interpolation of the global model and the local model is intermittently occurred at the server. Moreover, it requires all clients to participate in each round, which is a fairly unrealistic assumption in FL.

We only compared with the methods applicable for realistic FL settings, APFL. For the comparison, we observed dynamics of performances by varying λ ∈ [0, 1] in both SuperFed (SuperFed-MM & SuperFed-LM) and APFL (See Figure A2). While SuperFed-MM and SuperFed-LM consistently perform well in terms of personalization regardless of the value λ, APFL performs best only when λ is close to 0.25. This supports the evidence of induced connectivity between the federated and local models by the regularization. As endpoints (i.e., λ = 0 for a federated model, λ = 1 for a local model) are updated in the way to be connected, both are trained in a mutually beneficial way.

A.4 Label Noise Scenario
Since we simulated noisy label scenarios using clean-labeled datasets (MNIST and CIFAR10), we manually obfuscate the labels of training samples following two schemes: pairwise flipping [22] and symmetric flipping [52]. Both are based on the defining label transition matrix T, where each element of Tᵢⱼ indicates the probability of transitioning from an original label y = i to an obfuscated label ŷ = j, i.e., Tᵢⱼ = ℓ(y = j | y = i).

A.4.1 Pairwise Flipping for Label Noise. The pairwise flipping emulates a situation when labelers make mistakes only within similar classes. In this case, the transition matrix T for a pairwise flip given the noise ratio ℓ is defined as follows [22].

\[
T = \begin{bmatrix}
1 - \ell & \ell & 0 & \cdots & 0 \\
0 & 1 - \ell & \ell & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & 1 - \ell & \ell \\
\ell & 0 & \cdots & 0 & 1 - \ell
\end{bmatrix}
\]

A.4.2 Symmetric Flipping for Label Noise. The symmetric flipping assumes a probability of mislabeling of a clean label as other labels are uniformly distributed. For n classes, the transition matrix for symmetric flipping is defined as follows [52].

\[
T = \begin{bmatrix}
1 - \ell & \ell & \cdots & \ell \\
\ell & 1 - \ell & \cdots & \ell \\
\vdots & \vdots & \ddots & \vdots \\
\ell & \ell & \cdots & 1 - \ell
\end{bmatrix}
\]
Table A1: Details of training configurations

| Setting | Pathological non-IID | Dirichlet distribution-based non-IID | Realistic non-IID | Label noise (pair/symmetric) |
|---------|----------------------|--------------------------------------|------------------|-------------------------------|
| Dataset | MNIST  500 | CIFAR10  500 | CIFAR100  1000 | TinyImageNet  1000 | FEMNIST  10 | Shakesspeare  5 | MNIST  500 | CIFAR10  500 |
| R       | 10       | 10       | 10       | 10       | 10       | 50       | 10       | 10       |
| E       | 10       | 10       | 10       | 10       | 10       | 50       | 10       | 10       |
| B       | 10       | 10       | 10       | 10       | 10       | 50       | 10       | 10       |
| K       | 50,100,500 | 50,100,500 | 100       | 200       | 730       | 660       | 100       | 100       |
| η       | 0.01     | 0.01     | 0.01     | 0.01     | 0.01     | 0.8      | 0.01     | 0.01     |
| Model   | TwoNN    | TwoCNN   | ResNet   | MobileNet | VGG      | StackedLSTM | TwoNN    | TwoCNN   |

Figure A1: The loss surface of a deep network TwoCNN on CIFAR10 dataset calculated from l2-regularized cross-entropy loss as a function of network parameters in a two-dimensional subspace. The top rows indicate three-dimensional visualization of loss landscapes and the bottom rows indicates contours of derived loss landscapes First column: loss landscape obtained from FedAvg. Second column: loss landscape obtained from SuPerFed-MM. First column: loss landscape obtained from SuPerFed-LM.

Figure A2: Comparison of the personalization performance between APFL (left), SuPerFed-MM (middle), and SuPerFed-LM (right) by varying $\lambda \in [0, 1]$. Note that each method is trained on MNIST dataset with TwoNN model assuming pathological non-IID setting ($K = 500$ and $R = 500$).