
FedSPD: A Soft-clustering Approach for Personalized Decentralized Federated Learning

I-Cheng Lin
Carnegie Mellon University

Osman Yağın
Carnegie Mellon University

Carlee Joe-Wong
Carnegie Mellon University

Abstract

Federated learning has recently gained popularity as a framework for distributed clients to collaboratively train a machine learning model using local data. While traditional federated learning relies on a central server for model aggregation, recent advancements adopt a decentralized framework, enabling direct model exchange between clients and eliminating the single point of failure. However, existing decentralized frameworks often assume all clients train a shared model. Personalizing each client’s model can enhance performance, especially with heterogeneous client data distributions. We propose FedSPD, an efficient personalized federated learning algorithm for the decentralized setting, and show that it learns accurate models even in *low-connectivity* networks. To provide theoretical guarantees on convergence, we introduce a clustering-based framework that enables consensus on models for distinct data clusters while personalizing to unique mixtures of these clusters at different clients. This flexibility, allowing selective model updates based on data distribution, substantially reduces communication costs compared to prior work on personalized federated learning in decentralized settings. Experimental results on real-world datasets show that FedSPD outperforms multiple decentralized variants of personalized federated learning algorithms, especially in scenarios with *low-connectivity* networks.

erate within a centralized federated learning (CFL) framework, where a central server coordinates the training process.¹ In CFL, each client independently trains a model on its local data and then sends the model parameters to a central server for aggregation, which is subsequently broadcast back to the clients to begin a new training round. However, communication delays and bottlenecks often arise when managing numerous mobile or IoT (Internet-of-Things) clients, hampering CFL’s efficiency. Furthermore, this centralized structure poses risks of attacks and failures due to the single point of failure (Lalitha et al., 2018).

Decentralized Federated Learning (DFL) addresses these limitations by adopting a fully decentralized architecture where clients share their locally trained model parameters directly with neighboring clients, eliminating the need for a central server (Lalitha et al., 2018). This approach allows for substantial reductions in communication and computational costs (Beltrán et al., 2023) while mitigating vulnerabilities associated with a central server. However, most existing DFL methods focus on learning a single global model for all clients, aiming for consensus across clients. This global model, however, may under-perform on clients with non-IID (independent and identically distributed) local data. To address this challenge, we aim to design an efficient **personalized, decentralized federated learning algorithm** that personalizes models to each client’s data distribution without relying on a central server and preserves DFL’s communication benefits by limiting the required communication between clients. Personalized DFL can be particularly useful when clients are IoT devices using device-to-device communication protocols, e.g., vehicles learning personalized models of human driver preferences (Nakanoya et al., 2021).

Personalization of a shared global model has shown to improve performance in CFL settings (Ruan and

1 Introduction

Federated Learning (FL) is a popular approach for distributed clients to collaboratively learn from their local data. The most popular FL algorithm, FedAvg (McMahan et al., 2017), and most of its variants op-

¹Note that clients in the CFL setting still train their models in a distributed manner; the term “centralized” simply refers to the presence of a central server managing the clients’ interactions.

Joe-Wong, 2022; Marfoq et al., 2021). However, extending such personalization methods to DFL poses significant **technical challenges**. DFL algorithms typically strive for consensus by sharing local models among neighboring clients, which represent only a subset of all clients. Ensuring that all clients can benefit from each other’s updates despite limited communication is a key challenge (Beltrán et al., 2023). In contrast, learning personalized models requires intentionally maintaining differences in clients’ models, particularly for non-IID data. This makes it difficult to distinguish whether model disparities are due to communication issues or differences in local data distributions. We overcome this challenge by *quantifying similarities between client data* using a clustering-based method, allowing the training of distinct models for different data clusters, which are then personalized to each client’s unique data mixture.

Prior works that seek to personalize models in DFL settings, including cluster-based methods, are typically straightforward extensions of personalization methods designed for CFL settings, which do not take into account the distinct communication patterns in DFL and thus perform poorly when the client network has poor connectivity. For example, a naïve clustering method assigns each client to a single cluster based on its data distribution (Ghosh et al., 2020). However, such "hard" clustering assumes identical distributions within the same cluster, which is rarely the case. Instead, we adopt a **soft clustering** approach, as explored in CFL settings (Ruan and Joe-Wong, 2022; Marfoq et al., 2021), where each client’s data is modeled as an unknown mixture of distributions, and a model is trained for each cluster in this mixture. Existing DFL soft clustering approaches require clients to train models for all clusters in every round (Marfoq et al., 2021), imposing **significant training and communication overhead** that scales linearly with the number of clusters. This is particularly problematic in DFL scenarios, where clients often have limited communication and computation capacity (Nguyen et al., 2021). Therefore, we introduce a training algorithm that (i) learns each client’s mixture coefficients, (ii) ensures consensus on models for each cluster, and (iii) unlike prior work, avoids communication resource requirements that scale with the number of clusters. Our **contributions** are as follows:

- We propose **FedSPD**, a novel FL algorithm for clients that utilizes soft clustering to train personalized models in a decentralized manner. FedSPD allows clients to reach a consensus on cluster-specific models and adapt their cluster mixture estimates over time, while requiring each client to train only **one** cluster model per training round,

significantly reducing communication costs.

- We **prove the convergence of FedSPD** in Theorem 4. This proof adopts a different approach from prior work on soft clustering in DFL, which typically requires clients to train models for every cluster in each round (Marfoq et al., 2021).
- We demonstrate through experiments on real-world datasets that **FedSPD outperforms existing DFL algorithms** (both personalized and non-personalized) and, in some cases, approaches the accuracy of centralized training algorithms. Furthermore, we show that FedSPD’s performance remains robust across different client communication topologies, making it **particularly effective in networks with low connectivity**.

Following a review of related work in Section 2, we present our DFL model in Section 3 and introduce the FedSPD algorithm in Section 4. We then provide a convergence proof in Section 5 and demonstrate the algorithm’s superior performance in Section 6, before concluding in Section 7.

2 Related Work

Decentralized Federated Learning has its roots in decentralized optimization (Nedic and Ozdaglar, 2009; Wei and Ozdaglar, 2012; Zhang et al., 2021) and in particular decentralized Stochastic Gradient Descent (SGD) (Lian et al., 2017). Several methods have been explored for decentralized optimization (Nedic and Ozdaglar, 2009; Wu et al., 2017; Lü et al., 2020), while the convergence analysis of decentralized SGD was first presented by Yuan et al. (2016) and Sirb and Ye (2018) with delayed information, highlighting decentralized SGD’s advantages over centralized methods (Lian et al., 2017). This literature establishes conditions on client connectivity such that all local models will converge to a consensus model (Lian et al., 2017). The effects of client communication topologies in DFL (Lalitha et al., 2018; Warnat-Herresthal et al., 2021) have also been studied, and gradient tracking techniques based on push-sum algorithms have been proposed to relax the assumptions on client connectivity needed to show consensus (Nedić and Olshevsky, 2014, 2016; Assran et al., 2019).

Personalization in CFL is generally motivated by highly non-IID client data (McMahan et al., 2017; Collins et al., 2021), which can impede convergence and lead to a global model performing poorly at some clients, which may discourage them from participating in the FL process (Huang et al., 2020). Common techniques include local finetuning (Sim et al.,

2019), model interpolation (Mansour et al., 2020), meta-learning (Fallah et al., 2020), pFedME (T Dinh et al., 2020) adding regularization terms, and multi-task learning (Smith et al., 2017; Yousefi et al., 2019; Li et al., 2021). Clustered FL in particular includes hard clustering, which partitions clients into clusters based on their data’s similarity (Ghosh et al., 2020) and its variations (Xie et al., 2021; Briggs et al., 2020; Duan et al., 2021; Mansour et al., 2020). In soft clustered FL, one instead assumes that each client’s data conforms to a mixture of distributions (Marfoq et al., 2021; Ruan and Joe-Wong, 2022). Like these prior works, we use models learned for each cluster as guides for a personalized model; unlike them, we add a final personalization step to ensure good performance. We discuss this comparison in more detail in Section 4.

Some prior works have considered **combining personalization and DFL**. Jeong and Kountouris (2023) proposed a distillation-based algorithm, while Ma et al. (2022) proposed a communication-efficient algorithm with model pruning and neighbor selection. Sadiev et al. (2022) proves lower bounds of personalized DFL algorithms under specific objectives. Unlike these works, we provide theoretical convergence guarantees under more general learning objectives. Some centralized personalization algorithms also include decentralized versions, such as FedEM (Marfoq et al., 2021) and hard-clustered FL (Ghosh et al., 2020). We experimentally show (Section 6) that FedSPD outperforms both FedEM and hard-clustered FL, particularly in low-connectivity settings. Moreover, we *only require each client to train one cluster model at a time*, which leads to significantly smaller computational and communication overhead than FedEM.

3 Problem Formulation

We illustrate our **system model** in Figure 1 and summarize our notation in Table 1. We suppose there are N clients that are connected to each other via a graph with adjacency matrix \mathbf{A} and use \mathcal{N}_i to denote the set of client i ’s neighbors. Each client $i = 1, 2, \dots, N$ has a fixed set \mathcal{D}_i of training data. Clients with a shared edge can directly communicate with each other, e.g., to send model parameters.

Each data point $d \in \mathcal{D}_i$ on each client i is randomly sampled from one of S unique probability distributions (clusters) denoted as P_1, P_2, \dots, P_S , as illustrated in Figure 1. Consistent with standard clustering methods, we take S as a hyperparameter predetermined (Ruan and Joe-Wong, 2022). Letting \mathbf{x} denote the parameters of a machine learning model, we define the loss function $l(\mathbf{x}; D)$ as measuring the sum of the model losses with parameters \mathbf{x} over all

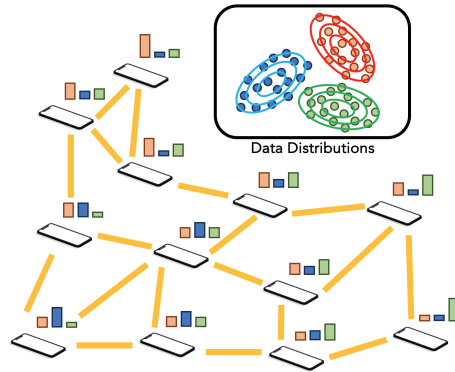


Figure 1: Illustration of the mixture of data distribution at clients in DFL.

points d in a dataset D . Cross-entropy loss, for example, is a typical loss function for classification problems. The *risk* of cluster s can then be written as: $F_s(\mathbf{x}) = \mathbb{E}_{D \sim P_s}[l(\mathbf{x}; D)]$. Our goal is for the clients to collectively find the optimal (i.e., risk-minimizing) model parameters for each cluster, which we also call the *cluster centers* and can be written as: $\mathbf{c}_s^* = \operatorname{argmin}_{\mathbf{x}} F_s(\mathbf{x})$, for $s = 1, 2, \dots, S$.

Given the cluster centers and mixture coefficients u_{is} , which represent the proportion of cluster s in client i ’s data, each client can find a personalized model for its local data mixture (Section 4). By focusing on common cluster centers, personalized learning can be reframed as achieving consensus on these centers, addressing a key challenge in personalized DFL. However, clients cannot directly determine the cluster centers using their local data \mathcal{D}_i since it is a *mixture* of clusters, and they do not know the cluster assignments of their data points. In the next section, we present an algorithm for clients to estimate the cluster centers and use them to derive personalized models.

4 Proposed FedSPD Algorithm

At each round $t = 1, 2, \dots, T$, each client i maintains two types of parameters: (i) its estimate of the cluster center \mathbf{c}_{is}^t for each cluster s , and (ii) the cluster to which each data point $d \in \mathcal{D}_i$ is associated, and the corresponding fraction of its data belonging to each cluster s , denoted by u_{is}^t . In each round t , clients update these parameters based on their local data and information received from their neighbors.

Each round of training consists of **four steps**: (1) local training, (2) parameter exchange, (3) parameter (i.e., cluster center) update, and (4) data clustering. Following the last training round, we conduct a **final personalization step**, which involves a local training update to each client’s personalized model. The entire

Name	Notation	Domain	Description
Number of Clients / Clusters	N, S	$N, S \in \mathbb{N}$	The total number of clients / clusters
Learning Rate	η_t	$\eta_t \in \mathbb{R}, 0 < \eta < 1$	Define the learning rate at time t of the task
Number of Local Updates	τ	$\tau \in \mathbb{N}$	Number of local updates in each training round
Client Neighbors	\mathcal{N}_i	$\mathcal{N}_i \in \mathcal{P}(N)$	Indices (in $\{1, 2, \dots, N\}$) of client i 's neighbors
Final Model Parameters	\mathbf{x}_i	$\mathbf{x}_i \in \mathbb{R}^{1 \times X}$	Final model parameters of client i
Final Concatenated Model Parameters	\mathbf{X}	$\mathbf{X} \in \mathbb{R}^{N \times X}$	Concatenated personalized model parameters
Local Dataset	$\mathcal{D}_{i,s}^t$	$\mathcal{D}_{i,s}^t \subseteq \mathcal{D}_i^t$, client i 's data	Data at client i associated with cluster s
Cluster Selection	s_i^t	$s_i^t \in \{1, 2, \dots, S\}$	Index of cluster that client i trains in round t
Portion of Clusters	$u_{i,s}^t$	$u_{i,s}^t \in \mathbb{R}, 0 < u_{i,s} \leq 1$	Portion of data for client i of cluster s at time t
Concatenated Portions of Clusters	$\mathbf{U}(t)$	$\mathbf{U} \in \mathbb{R}^{N \times S}$	Concatenated portion of data of all clients
Average Cluster Centers	$\bar{\mathbf{c}}_s^t$	$\bar{\mathbf{c}}_s^t \in \mathbb{R}^X$	Average center of cluster s of all clients at time t
Concatenated Cluster Centers	\mathbf{C}_s^t	$\mathbf{C}_s^t \in \mathbb{R}^{N \times X}$	Concatenated centers of cluster s at time t
Collection of Cluster Centers	$\mathcal{C}(t)$	$\mathcal{C}(t) \in \mathbb{R}^{S \times N \times X}$	$\mathcal{C}(t) = \{\mathbf{C}_1^t, \mathbf{C}_2^t, \dots, \mathbf{C}_S^t\}$
Weight Matrix	\mathbf{W}_s^t	$\mathbf{W}_s^t \in \mathbb{R}^{N \times N}$	Weight matrix of cluster s at time t
Augmented Adjacency Matrix	\mathbf{A}	$\mathbf{A} \in \mathbb{R}^{N \times N}$	Augmented adjacency matrix with diagonal elements equal to 1
Concatenated Gradients	\mathbf{G}_s^t	$\mathbf{G}_s^t \in \mathbb{R}^{N \times X}$	Concatenated gradients at time t for cluster s , $\mathbf{G}_s^t := [\nabla F_1, \dots, \nabla F_N]$

Table 1: Mathematical notations used in the paper.

training algorithm is shown in Algorithm 1.

Step 1: Local training (line 13 in Algorithm 1). In round t , each client i has an estimated portion $u_{i,s}^t$ of its data coming from cluster s , where $\sum_{s=1}^S u_{i,s}^t = 1$. These values are computed at the end of the previous round (step 4). Client i then selects cluster s to update with probability $u_{i,s}$, ensuring that clients contribute more to clusters where they have more data. By selecting only one cluster per round, *FedSPD* keeps the training overhead independent of the number of clusters S , as each client always trains a single model.

Once a cluster s is selected, the client performs τ SGD updates on its current estimate $\mathbf{c}_{i,s}^t$ for the cluster center using learning rate η . Gradients are computed on the risk of the data associated with the selected cluster, $\mathcal{D}_{i,s}^t$, as $\nabla_{\mathbf{c}} \ell(\mathbf{c}; d)$, where d is sampled uniformly at random from $\mathcal{D}_{i,s}^t$. The dataset $\mathcal{D}_{i,s}^t$ is formed in the previous round's clustering step, which assigns each data point $d \in \mathcal{D}_i$ to a cluster.

Step 2: Parameter exchange (line 20 in Algorithm 1). Let s_i^t be the cluster selected by client i in round t , resulting in an updated value for $\mathbf{c}_{i,s_i^t}^t$. Client i then broadcasts s_i^t and $\mathbf{c}_{i,s_i^t}^t$ to its neighbors j where $j \in \mathcal{N}_i$. Consequently, each client i receives the communications $\{s_j^t, \mathbf{c}_{j,s_j^t}^t\}_{j \in \mathcal{N}_i}$ from all its neighbors.

Step 3: Cluster center updates (line 26 in Algorithm 1). After receiving the updated cluster center parameters and indices from its neighbors, each client i updates its estimate of each cluster center s using the average of the received updates for cluster s :

$$\mathbf{c}_{i,s}^{t+1} = \frac{1}{|j \in \mathcal{N}[i] \cap s_j^t = s|} \sum_{j \in \mathcal{N}[i] \cap s_j^t = s} \mathbf{c}_{j,s}^t \quad (1)$$

Here, $\mathcal{N}[i]$ is the closed neighborhood, including client

i and its neighboring clients. $|j \in \mathcal{N}[i] \cap s_j^t = s|$ represents the number of clients j that both selected cluster s for updating and belong to $\mathcal{N}[i]$. The client applies Eq. (1) for *all* clusters s for which it received at least one update. If no updates for cluster s are received in round t , i.e., none of the neighbors selected it, the estimated cluster center remains unchanged: $\mathbf{c}_{i,s}^{t+1} = \mathbf{c}_{i,s}^t$. This update rule can be expressed in matrix form as $\mathbf{C}_s^{t+1} = \mathbf{W}_s^t \mathbf{C}_s^t$, where \mathbf{W}_s^t is the weight matrix for cluster s at time t , and $\mathbf{C}_s^t = [\mathbf{c}_{1s}^t, \dots, \mathbf{c}_{N_s}^t]$ contains the concatenated cluster centers.

Step 4: Data clustering (line 33 in Algorithm 1). After updating the cluster centers, each client i associates its data points $d \in \mathcal{D}_i$ with a cluster. It calculates the loss $\ell(\mathbf{c}_{i,s}^{t+1}, d)$ for each cluster s and assigns data point d to the cluster with the lowest loss. Using these new associations, $u_{i,s}^{t+1}$, the fraction of data points linked to cluster s , is computed. This step enables *FedSPD* to adapt the mixture coefficients as cluster center estimates evolve. With clustering complete, the process moves to the next round, $t + 1$, starting again with local training.

Final Step: Personalization (line 42 in Algorithm 1). After T rounds, each client i computes a personalized model as a weighted sum of its cluster centers:

$$\mathbf{x}_i = \sum_{s=1}^S u_{i,s}^T \mathbf{c}_{i,s}^T \quad (2)$$

Marfoq et al. (2021) show that this weighted sum provides the optimal personalized model for client i when the loss function ℓ is convex. However, since most practical loss functions, such as cross-entropy for neural networks, are not convex, this aggregated model may not perform optimally in practice. To address this, each client runs a few additional local training iterations, starting from \mathbf{x}_i computed in Eq. (2), using

its entire local dataset \mathcal{D}_i .

Algorithm 1 Our Proposed FedSPD Algorithms

```

1: procedure FEDSPD( $\eta, \tau, S, T, \mathbf{K}_0, \mathbf{W}_s^t$ )
2:   for  $t = 1, 2, \dots, T\tau$  do
3:     LOCALUPDATE( $\mathcal{C}(t)$ )
4:     if  $t \bmod \tau = 0$  then
5:       PARAMETEREXCHANGE( $\mathcal{C}(t), \mathbf{A}$ )
6:       PARAMETERUPDATE( $\mathcal{C}(t), \mathbf{A}$ )
7:       DATACLUSTERING( $\mathcal{C}(t), \mathbf{A}$ )
8:     end if
9:   end for
10:  FINALPHASE( $\mathcal{C}(t), \mathbf{u}(t)$ )
11: end procedure
12:
13: procedure LOCALUPDATE( $\mathcal{C}(t)$ )
14:   for  $i = 1, 2, \dots, N$  do
15:     Client  $i$  select cluster  $s_i^t$  to update
16:      $\mathbf{c}_{s_i^t}^{t+1} = \mathbf{c}_{s_i^t}^t - \eta_t \nabla f_{is}(\mathbf{c}_{s_i^t}^t)$ 
17:   end for
18: end procedure
19:
20: procedure PARAMETEREXCHANGE( $\mathcal{C}(t), \mathbf{A}$ )
21:   for  $i = 1, 2, \dots, N$  do
22:     For each client  $i$ , exchange the updated parameter  $\mathbf{c}_{is}$  and the selected cluster  $s$  with client  $j \in \mathcal{N}_i$ 
23:   end for
24: end procedure
25:
26: procedure PARAMETERUPDATE( $\mathcal{C}(t), \mathbf{A}$ )
27:   Construct  $\mathbf{W}_s^t$  for each cluster  $s$ . If client  $i$  is not selected to update cluster  $s$ , the row  $i$  and column  $i$  will only have diagonal element equal to 1, else equal to 0, meaning the model parameter will remain the same as it was in the previous epoch.
28:   for  $s = 1, 2, \dots, S$  do
29:      $\mathbf{C}_s^{t+1} = \mathbf{W}_s^t \mathbf{C}_s^{t+1}$ 
30:   end for
31: end procedure
32:
33: procedure DATACLUSTERING( $\mathcal{C}(t), \mathbf{A}$ )
34:   for  $i = 1, 2, \dots, N$  do
35:     for  $d_k \in \mathcal{D}_i$  do
36:       Label data  $d_k$  with the least loss of all the model parameters among all clusters.
37:     end for
38:     For  $s = 1, \dots, S$  update  $u_{i,s}^t$  for client  $i$ 
39:   end for
40: end procedure
41:
42: procedure FINALPHASE( $\mathcal{C}(t), \mathbf{u}^t$ )
43:   for  $i = 1, 2, \dots, N$  do
44:      $\mathbf{X}_i = \sum_{s=1}^S u_{i,s}^t \mathbf{C}_s^t(i, :)$ 
45:   end for
46:   LOCALUPDATE( $\mathbf{X}$ )  $\triangleright$  Do the local update using all data of the client for the aggregated training
47: end procedure

```

Comparison to prior soft clustering algorithms. Marfoq et al. (2021) and Ruan and Joe-Wong (2022) use soft clustering to learn cluster centers and personalized models without this final personalization step, directly learning personalized models in each iteration,

with a central server estimating the cluster centers. In DFL, achieving consensus on cluster models is difficult due to the extensive parameter exchanges needed for model propagation, particularly when clients have few neighbors. Marfoq et al. (2021) proposes a decentralized algorithm that sets the personalized model as a weighted sum of the cluster centers *at each round's end*, which can be sub-optimal for non-convex loss functions. Such a framework can lead to overfitting in DFL, as clients have low connectivity and thus cannot rely on receiving many other clients' updates in each training round. Adding a final personalization step, as we use in FedSPD, is likely to exacerbate overfitting, as cluster center gradients already incorporate personalized models. In Section 6, we demonstrate that *FedSPD* outperforms the FedEM algorithm by (Marfoq et al., 2021), which requires each client to train all models per round, incurring significantly higher computational and communication costs than *FedSPD*.

5 Convergence Analysis

We prove that *FedSPD* converges in Theorem 4. We first outline our technical assumptions and then present our main results. All proof details can be found in Appendix A due to space limitations.

Assumptions. Our analysis relies on the following assumptions on the risk function and gradient estimates, which are common in the literature (Marfoq et al., 2021; Ghosh et al., 2020; Koloskova et al., 2020).

Assumption 1 (*Strong convexity and smoothness*) The risk function F_s for each cluster s is μ -strongly convex and L -smooth. That is, for some $L > 0$:

$$\begin{aligned} \|\nabla F_s(\mathbf{x}) - \nabla F_s(\mathbf{y})\| &\leq L\|\mathbf{x} - \mathbf{y}\|; \\ \nabla F_s(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2 &\leq F_s(\mathbf{y}) - F_s(\mathbf{x}) \end{aligned} \quad (3)$$

Assumption 2 (*Bounded risk function*) The risk function F_s for each cluster s is lower-bounded by some $F_{inf} > 0$, i.e., $F_s(\mathbf{x}) \geq F_{inf}$.

Assumption 3 (*Unbiased gradient estimation*) The gradient is unbiased, i.e., $\mathbb{E}[\nabla f(\mathbf{x})] = \nabla F(\mathbf{x})$.

Assumption 4 (*Bounded gradient*) We have $\mathbb{E}\|\nabla f(\mathbf{x})\|^2 \leq \sigma^2$ for some $\sigma^2 > 0$.

Assumption 5 (*Bounded variance of gradient estimation*) The gradient estimation is bounded:

$$\mathbb{E}\|\nabla f(\mathbf{x}) - \nabla F(\mathbf{x})\|^2 \leq v^2, \text{ for some } v^2 > 0. \quad (4)$$

Assumption 6 (*Bounded cluster error*) Following (Ruan and Joe-Wong, 2022; Ghosh et al., 2020), at

a certain training step, all estimated cluster centers have bounded distance to the optimal centers. That is:

$$\|\mathbf{c}_{i_s}^t - \mathbf{c}_s^*\| \leq (0.5 - \alpha_0) \sqrt{\frac{\mu}{L}} \delta, \forall s \in 1, 2, \dots, S \quad (5)$$

where $0 < \alpha_0 \leq 0.5$. Without loss of generality, we also assume for all s , $\|\mathbf{c}_s^*\| \leq 1$.

Note that this assumption will always hold for some value of δ ; however, a larger δ , and thus larger cluster error, will also lead to slower convergence.

We finally follow Koloskova et al. (2020) in assuming that clients communicate sufficiently for consensus:

Assumption 7 (Expected consensus rate) For some constant $p \in (0, 1]$ and integer $\beta \geq 1$, such that \mathbf{C}_s , the concatenated model parameter matrix of cluster s and all non-negative integer $l \leq \frac{T}{\beta}$ we have:

$$\mathbb{E} \left\| \mathbf{C}_s \prod_{t=l\beta}^{(l+1)\beta-1} \mathbf{W}_s^t - \bar{\mathbf{C}}_s \right\|_F^2 \leq (1-p) \|\mathbf{C}_s - \bar{\mathbf{C}}_s\|_F^2 \quad (6)$$

where $\bar{\mathbf{C}}_s := \underbrace{[\bar{\mathbf{c}}_s, \dots, \bar{\mathbf{c}}_s]}_{\text{total } N \text{ terms}}$ is the matrix with every column equal to the average of the model parameters.

For simplicity, we further assume that all clients have the same amount of data (i.e., \mathcal{D}_i has the same number of data points for all clients i) and that the number of local updates $\tau = 1$ in the remainder of this section. These can be easily relaxed if needed.

Results. Without loss of generality, we present our results for a specific cluster i , where $i = 1, \dots, S$. Since the convergence proof is identical for all S clusters, we omit the cluster index for clarity. Let n be the number of clients chosen to update the selected cluster. If the total data across clients is roughly uniform for each cluster, then $n \approx \frac{N}{S}$. We begin by bounding the distance of the average cluster center to its optimality:

Theorem 1 (Descent lemma) The distance $\mathbb{E} \left\| \bar{\mathbf{c}}^{(t+1)} - \mathbf{c}^* \right\|^2$ between the average cluster center and its optimum \mathbf{c}^* satisfies the bound (7) with proper choice of learning rate η_t :

$$\begin{aligned} &\leq \frac{\eta_t(L+\mu)}{n} \sum_{i=1}^{n_1} \|\bar{\mathbf{c}}^{(t)} - \mathbf{c}_i^{(t)}\|^2 + \frac{18L^2\epsilon_N^2\eta_t^2}{n^2} + v^2\eta_t^2 \\ &+ \left(1 - \eta_t\mu + \frac{\eta_t\mu\epsilon_N}{n}\right) \|\bar{\mathbf{c}}^{(t)} - \mathbf{c}^*\|^2 + \frac{2\epsilon_N(S-1)v^2\eta_t^2}{n^2} \\ &+ \left(\frac{4\eta_t^2(n-\epsilon_N)^2L}{n^2} + 2\eta_t - \frac{2\eta_t\epsilon_N}{n}\right) \left(f(\bar{\mathbf{c}}^{(t)}) - f(\mathbf{c}^*)\right) \end{aligned} \quad (7)$$

Here ϵ_N is the bound of the expected number of clients using the wrong data in Lemma 2.

We then derive an expression for the cluster centers estimated by individual clients.

Theorem 2 (Update rule) Clients' estimated centers of the cluster after time t can be written as:

$$\mathbf{C}^t = \mathbf{C}^{l\beta} \prod_{m=l\beta}^{t-1} \mathbf{W}^m - \sum_{m=l\beta}^{t-1} \left(\eta_t \mathbf{G}^m \prod_{r=t-1}^m \mathbf{W}^r \right) \quad (8)$$

Here $l \in \mathbb{N}$ and β is the the constant in Assumption 7. Given this expression, we can relate the clients' cluster center estimates to their average, showing that they eventually reach a near-consensus:

Theorem 3 (Consensus distance) Define $\mathbf{E}_t = \frac{1}{N} \sum_{i=1}^N \mathbb{E} \|\mathbf{c}_i^{(t)} - \bar{\mathbf{c}}^{(t)}\|^2$, the expected squared distance of the model parameters of client i to the average model parameter. It is upper-bounded by

$$\begin{aligned} \left(1 - \frac{p}{2}\right) \mathbf{E}_{m,\beta} + \sum_{j=m\beta}^{t-1} \left(\frac{p\mathbf{E}_j}{16\beta} + \frac{18\beta n\sigma^2 + nv^2p}{Np} \eta_j^2 \right. \\ \left. + \left(\frac{36Ln\beta\eta_t^2}{pN} \right) (f(\bar{\mathbf{c}}) - f(\mathbf{c}^*)) \right) \end{aligned} \quad (9)$$

Here p is the constant defined in Assumption 7.

Theorem 4 (Cluster convergence rate) For given target accuracy ϵ , there exists a constant learning rate for which ϵ accuracy can be reached after T iterations.

$$\begin{aligned} \left[1 + \left(\frac{n-\epsilon_N}{n}\right) \eta L\right] \sum_{t=0}^T \frac{r_t}{R_T} (\mathbb{E}f(\bar{\mathbf{c}}^t) - f(\mathbf{c}^*)) \\ + \mu \mathbb{E} \|\bar{\mathbf{c}}^{(T+1)} - \mathbf{c}^*\|^2 \leq \epsilon \end{aligned} \quad (10)$$

Here w_t is a sequence of positive weights defined in Lemma 3 in Appendix A.4 and $R_T = \sum_{t=1}^T r_t$. Rearranging, we find that the number of required iterations T is at the order:

$$\begin{aligned} \tilde{O} \left(\frac{\sqrt{L+\mu}}{\sqrt{\epsilon}\mu} \left(\sigma + \frac{\sqrt{n}}{\sqrt{N}} v \right) + \frac{L\beta n^{\frac{3}{2}}}{\mu p \sqrt{N}(n-\epsilon_N)} \ln\left(\frac{1}{\epsilon}\right) \right. \\ \left. + v^2 \frac{n^2 + L^2\epsilon_N^2 + \epsilon_N(S-1)}{\mu n^2 \epsilon} \right) \end{aligned} \quad (11)$$

The convergence rate, asymptotically requiring $O(1/\sqrt{\epsilon})$ training rounds to reach an error ϵ , aligns with previous works on DFL without personalization (Koloskova et al., 2020), leading us to conjecture that FedSPD will converge well. We note that the network connectivity appears in this bound through the constant $p \in (0, 1]$ (Assumption 7), where higher connectivity indicates a larger p . However, the second term in the convergence rate that involves p is not the dominant term. Thus, as long as the network is connected, we expect that the effect of network connectivity on convergence will be relatively minor. Our simulation results in later Section 6.2 support this observation.

6 Simulation Results

In this section, we evaluate the performance of our proposed algorithms and compare them with existing methods. We also analyze how different network connectivity and topology influence the performance.

Datasets and models. Unless specified, we use $N = 100$ clients for all experiments on hand-written character recognition (MNIST and EMNIST datasets (Cohen et al., 2017)) and $N = 25$ clients for all experiments on image classification (CIFAR-10 and CIFAR-100 datasets (Krizhevsky et al., 2009)). We use a CNN (convolutional neural network) model for each client with data from a mixture of $S = 2$ distributions, \mathcal{D}_A and \mathcal{D}_B . Each client draws 10% to 90% of its data from \mathcal{D}_A and the remainder from \mathcal{D}_B with unbalanced class (Marfoq et al., 2021) or image rotation (Ruan and Joe-Wong, 2022) or both. We follow Ruan and Joe-Wong (2022) and Marfoq et al. (2021) for other parameter settings. Details are described in the Appendix B. The test accuracy is evaluated on each client’s local test dataset, which is unseen during training.

Client communications. Unless specified, the client graph is a connected Erdos-Renyi (ER) random graph (Erdos et al., 1960) with average degree from 5 to 12; more specifics are in the Appendix. To avoid the label switching problem (Stephens, 2000), we compute the cosine similarities of the model parameters received from other clients to ensure cluster consensus.

Baselines. We compare *FedSPD* with: (i) centralized and decentralized **FedAvg** (McMahan et al., 2017); (ii) centralized and decentralized **FedEM** (Marfoq et al., 2021), a prior soft clustering method; (iii) centralized and decentralized versions of **FedSoft** (Ruan and Joe-Wong, 2022), which also uses soft clustering; (iv) centralized and decentralized **IFCA** (Ghosh et al., 2020) using hard clustering; (v) centralized and decentralized **pFedMe** (T Dinh et al., 2020) another state-of-the-art FL personalization approach without clustering; and (vi) **local** training on local dataset only.

Additional results on the impact of (i) τ (local update rounds), (ii) the number of epochs in the last phase, and (iii) S (the number of clusters); and (iv) extended experiments are included in Appendix B.

6.1 Comparison with Baselines

We first compare our method with other decentralized personalized methods. Our results on EMNIST, CIFAR-10, and CIFAR-100 are shown in Table 2. *FedSPD* obtains higher test accuracy among all DFL methods, approaching the accuracy of CFL. The centralized methods still outperform decentralized methods, as expected from prior literature (Sun et al.,

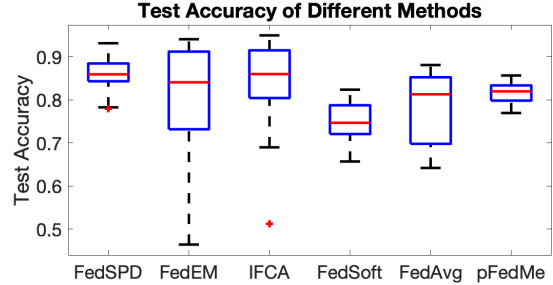


Figure 2: Box-plot for accuracy across clients on EMNIST dataset. *FedSPD* has much lower variance in test accuracy across clients.

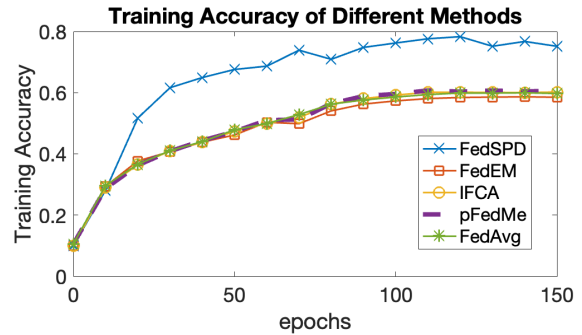


Figure 3: Training accuracy of different DFL methods versus number of epochs on CIFAR-10 ($N = 25$). *FedSPD* converges faster in terms of training accuracy compared to all other DFL methods.

2023). However, decentralized methods offer advantages such as lower communication traffic and increased robustness, as they do not rely on a single point of failure like a centralized server. Figure 3 shows the training accuracy versus number of epochs on the CIFAR-10 dataset. *FedSPD* converges faster than all other DFL algorithms in terms of training accuracy. This shows that each of the clusters in *FedSPD* does converge as desired. Note that compared to *FedEM*, another soft clustering method, our *FedSPD* needs half the communication cost, since *FedEM* clients exchange the information of all $S = 2$ clusters.

To guarantee the **fairness across clients**, we show the box plot of the final test accuracy across different clients on EMNIST in Figure 2. *FedSPD* has much less variance in accuracy across different clients, validating that its improvement in average accuracy does not come from high accuracy in a few clients.

6.2 Effects of Network Connectivity

In this section, we investigate how the performance varies with the connectivity of the client network.

Figure 4 shows the test accuracy of different DFL

Dataset	DFL						CFL					Local
	FedSPD	FedEM	IFCA	FedAvg	FedSoft	pFedMe	FedEM	IFCA	FedAvg	FedSoft	pFedMe	
EMNIST	84.16	80.63	83.82	77.87	74.81	81.61	88.83	89.42	88.81	84.97	90.95	57.64
CIFAR-10	68.71	50.02	51.75	49.10	42.38	50.29	79.93	79.76	79.44	76.62	79.12	42.38
CIFAR-100	41.12	18.12	17.36	17.34	13.17	18.31	44.54	43.82	43.33	39.76	7.26 ²	13.14

Table 2: FedSPD achieves higher test accuracy than other DFL algorithms and at times comparable test accuracy to CFL algorithms on EMNIST, CIFAR-10 and CIFAR-100 datasets. Accuracy in percentage (%)

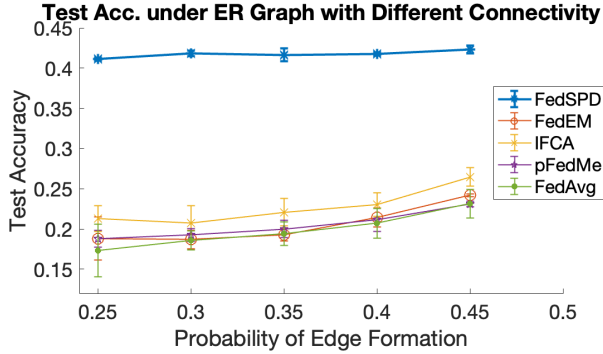


Figure 4: Test accuracy of different methods under different connectivity of ER Random Graph on CIFAR-100 dataset ($N = 15$). FedSPD shows consistently high test accuracies compare to other DFL methods.

methods under different connectivity on the CIFAR-100 dataset using the ER Random Graph, over three experimental runs. FedSPD consistently shows the highest test accuracies, though other methods’ performance begins to increase as the graph becomes more connected (a higher probability of link formation).

Tables 3 and 4 show the test accuracy of FedSPD in different type of networks and connectivity. We use three different network topologies: the ER Random Graph; the Barabasi-Albert (BA) Model (Albert and Barabási, 2002) with preferential attachment representing the network following the power law; and the Random Geometrical Graph (RGG) (Penrose, 2003), which is often used in wireless communication and Internet-of-Things (IoT) scenarios with high clustering effect (Penrose, 2003). We observe that the final test accuracy does not vary significantly across different network topologies and levels of connectivity in MNIST. In EMNIST, the test accuracy slightly increases when the average degree increases. The test accuracy is more stable in RGG under different connectivity, which we conjecture is due to RGG’s highly clustered nature. Thus, as long as the network is connected, FedSPD performs well in both high and low connectivity scenarios and across various types of networks, as we expect from Theorem 4, FedSPD converges regardless of the network topology.

²The centralized pFedMe on CIFAR-100 does not converge in the various settings of hyperparameters we tried.

Average Degree	6	8	10	12	14
ER	92.86	92.93	93.37	93.31	93.26
BA	93.06	92.58	92.56	92.87	93.17
RGG	92.86	92.61	92.84	93.49	92.97

Table 3: FedSPD shows consistently high test accuracies on MNIST data for $N = 50$ clients.

Average Degree	8	12	16	20
ER	79.79	82.26	84.28	84.49
BA	79.45	82.13	84.58	84.73
RGG	82.26	83.49	84.06	84.08

Table 4: FedSPD shows consistently high test accuracies on EMNIST data for $N = 50$ clients.

7 Conclusion

We propose FedSPD, a soft clustering approach that enables federated training of personalized models in a decentralized setting. FedSPD models each FL client’s data as a mixture of cluster distributions and aims to learn a distinct model for each cluster. In the final phase, all models are aggregated and further personalized for each client. Importantly, FedSPD requires each client to train only one cluster model per training round, ensuring scalability with the number of clusters and works well when communication resource is limited. We theoretically demonstrate that FedSPD can achieve consensus within each cluster. Our experiments on real-world datasets show that FedSPD outperforms previous algorithms for personalized, decentralized FL and performs well even in *low-connectivity* networks. For future extensions, this work can serve as a foundation for various applications, such as environmental monitoring in IoT, object identification in AR/VR, or autonomous driving, all of which benefit from the low latency of direct communication and data similarity among adjacent devices.

References

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

Assran, M., Loizou, N., Ballas, N., and Rabbat, M. (2019). Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning*, pages 344–353. PMLR.

- Beltrán, E. T. M., Pérez, M. Q., Sánchez, P. M. S., Bernal, S. L., Bovet, G., Pérez, M. G., Pérez, G. M., and Celdrán, A. H. (2023). Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges. *IEEE Communications Surveys & Tutorials*.
- Briggs, C., Fan, Z., and Andras, P. (2020). Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE.
- Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. (2017). Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE.
- Collins, L., Hassani, H., Mokhtari, A., and Shakkottai, S. (2021). Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pages 2089–2099. PMLR.
- Duan, M., Liu, D., Ji, X., Liu, R., Liang, L., Chen, X., and Tan, Y. (2021). Fedgroup: Efficient federated learning via decomposed similarity-based clustering. In *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 228–237. IEEE.
- Erdos, P., Rényi, A., et al. (1960). On the evolution of random graphs. *Publ. math. inst. hung. acad. sci.*, 5(1):17–60.
- Fallah, A., Mokhtari, A., and Ozdaglar, A. (2020). Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*.
- Ghosh, A., Chung, J., Yin, D., and Ramchandran, K. (2020). An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597.
- Huang, T., Lin, W., Wu, W., He, L., Li, K., and Zomaya, A. Y. (2020). An efficiency-boosting client selection scheme for federated learning with fairness guarantee. *IEEE Transactions on Parallel and Distributed Systems*, 32(7):1552–1564.
- Jeong, E. and Kountouris, M. (2023). Personalized decentralized federated learning with knowledge distillation.
- Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. (2020). A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR.
- Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.
- Lalitha, A., Shekhar, S., Javidi, T., and Koushanfar, F. (2018). Fully decentralized federated learning. In *Third workshop on bayesian deep learning (NeurIPS)*, volume 2.
- Li, T., Hu, S., Beirami, A., and Smith, V. (2021). Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pages 6357–6368. PMLR.
- Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. (2017). Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30.
- Lü, Q., Liao, X., Li, H., and Huang, T. (2020). A computation-efficient decentralized algorithm for composite constrained optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 6:774–789.
- Ma, Z., Xu, Y., Xu, H., Liu, J., and Xue, Y. (2022). Like attracts like: Personalized federated learning in decentralized edge computing. *IEEE Transactions on Mobile Computing*, pages 1–17.
- Mansour, Y., Mohri, M., Ro, J., and Suresh, A. T. (2020). Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*.
- Marfoq, O., Neglia, G., Bellet, A., Kameni, L., and Vidal, R. (2021). Federated multi-task learning under a mixture of distributions. *Advances in Neural Information Processing Systems*, 34:15434–15447.
- McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.
- Nakanoya, M., Im, J., Qiu, H., Katti, S., Pavone, M., and Chinchali, S. (2021). Personalized federated learning of driver prediction models for autonomous driving. *arXiv preprint arXiv:2112.00956*.
- Nedić, A. and Olshevsky, A. (2014). Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615.
- Nedić, A. and Olshevsky, A. (2016). Stochastic gradient-push for strongly convex functions on time-varying directed graphs. *IEEE Transactions on Automatic Control*, 61(12):3936–3947.
- Nedic, A. and Ozdaglar, A. (2009). Distributed subgradient methods for multi-agent optimization.

- IEEE Transactions on Automatic Control*, 54(1):48–61.
- Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., and Poor, H. V. (2021). Federated learning for internet of things: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 23(3):1622–1658.
- Penrose, M. (2003). *Random geometric graphs*, volume 5. OUP Oxford.
- Ruan, Y. and Joe-Wong, C. (2022). Fedsoft: Soft clustered federated learning with proximal local updating. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8124–8131.
- Sadiev, A., Borodich, E., Beznosikov, A., Dvinskikh, D., Chezhegov, S., Tappenden, R., Takáč, M., and Gasnikov, A. (2022). Decentralized personalized federated learning: Lower bounds and optimal algorithm for all personalization modes. *EURO Journal on Computational Optimization*, 10:100041.
- Sim, K. C., Beaufays, F., Benard, A., Guliani, D., Kabel, A., Khare, N., Lucassen, T., Zadrazil, P., Zhang, H., Johnson, L., et al. (2019). Personalization of end-to-end speech recognition on mobile devices for named entities. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 23–30. IEEE.
- Sirb, B. and Ye, X. (2018). Decentralized consensus algorithm with delayed and stochastic gradients. *SIAM Journal on Optimization*, 28(2):1232–1254.
- Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. S. (2017). Federated multi-task learning. *Advances in neural information processing systems*, 30.
- Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.
- Sun, Y., Shen, L., and Tao, D. (2023). Which mode is better for federated learning? centralized or decentralized. *arXiv preprint arXiv:2310.03461*.
- T Dinh, C., Tran, N., and Nguyen, J. (2020). Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405.
- Warnat-Herresthal, S., Schultze, H., Shastry, K. L., Manamohan, S., Mukherjee, S., Garg, V., Sarveswara, R., Händler, K., Pickkers, P., Aziz, N. A., et al. (2021). Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 594(7862):265–270.
- Wei, E. and Ozdaglar, A. (2012). Distributed alternating direction method of multipliers. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5445–5450. IEEE.
- Wu, T., Yuan, K., Ling, Q., Yin, W., and Sayed, A. H. (2017). Decentralized consensus optimization with asynchrony and delays. *IEEE Transactions on Signal and Information Processing over Networks*, 4(2):293–307.
- Xie, M., Long, G., Shen, T., Zhou, T., Wang, X., Jiang, J., and Zhang, C. (2021). Multi-center federated learning. *arXiv preprint arXiv:2108.08647*.
- Yousefi, F., Smith, M. T., and Alvarez, M. (2019). Multi-task learning for aggregated data using gaussian processes. *Advances in Neural Information Processing Systems*, 32.
- Yuan, K., Ling, Q., and Yin, W. (2016). On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854.
- Zhang, J., Ling, Q., and So, A. M.-C. (2021). A newton tracking algorithm with exact linear convergence for decentralized consensus optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 7:346–358.

Supplementary Material

October 25, 2024

A Proof of the Theorems

A.1 Proof of Theorem 1

Without loss of generality, we select a single cluster, cluster 1 for analysis; the same analysis applies to the other $S - 1$ clusters. For readability, we eliminate the subscription indicating the cluster number 1. Consider each client running single step of SGD, we use n to indicate the number of clients selected to update this cluster and n_1 and n_0 to indicate the number of clients using the correct data and incorrect data, respectively (i.e. the data is drawn from this selected cluster is consider a correct data.), so that $n_1 + n_0 = n$. S indicates the set of the client selected to update this cluster, S^* indicates the set of clients using the correct data, and $\overline{S^*}$ indicates the set of clients using the incorrect data.

Lemma 1 (*Doubly-stochastic weight matrix preserves the average*) *At the communication step, if the model of each client in the network is updated according to a doubly-stochastic weight matrix \mathbf{W}^t then the average after the communication step remains the same. Formally, we have:*

$$\mathbf{C}^{t+1} \frac{\mathbf{1}\mathbf{1}^T}{N} = \mathbf{C}^t \mathbf{W}^t \frac{\mathbf{1}\mathbf{1}^T}{N} = \mathbf{C}^t \frac{\mathbf{1}\mathbf{1}^T}{N} \quad (12)$$

From Lemma 1, we can write the left-hand side of Theorem 1 as:

$$\begin{aligned} \left\| \bar{\mathbf{c}}^{(t+1)} - \mathbf{c}^* \right\|^2 &= \left\| \bar{\mathbf{c}}^{(t)} - \frac{\eta t}{n} \sum_{i=1}^n \nabla F_i(\mathbf{c}_i^{(t)}, D_i^{(t)}) - \mathbf{c}^* \right\|^2 \\ &= \left\| \bar{\mathbf{c}}^{(t)} - \mathbf{c}^* - \frac{\eta t}{n} \sum_{i \in S \cap S^*} \nabla F_i(\mathbf{c}_i^{(t)}) - \frac{\eta t}{n} \sum_{i \in S \cap \overline{S^*}} \nabla F_i(\mathbf{c}_i^{(t)}) \right\|^2 \\ &= \left\| \bar{\mathbf{c}}^{(t)} - \mathbf{c}^* - \frac{\eta t}{n} \sum_{i \in S \cap S^*} \nabla F_i(\mathbf{c}_i^{(t)}) \right\|^2 + \left\| \frac{\eta t}{n} \sum_{i \in S \cap \overline{S^*}} \nabla F_i(\mathbf{c}_i^{(t)}) \right\|^2 \\ &\quad - \frac{2\eta t}{n} \left\langle \bar{\mathbf{c}}^{(t)} - \mathbf{c}^* - \frac{\eta t}{n} \sum_{i \in S \cap S^*} \nabla F_i(\mathbf{c}_i^{(t)}), \sum_{i \in S \cap \overline{S^*}} \nabla F_i(\mathbf{c}_i^{(t)}) \right\rangle \end{aligned} \quad (13)$$

We let the first and the second term on the right-hand side as $\|T_1\|^2$ and $\|T_2\|^2$ respectively. Thus the above equation can be written as:

$$\left\| \bar{\mathbf{c}}^{(t+1)} - \mathbf{c}^* \right\|^2 = \|T_1\|^2 + \|T_2\|^2 + 2 \langle T_1, T_2 \rangle \leq (1 + \alpha) \|T_1\|^2 + (1 + \alpha^{-1}) \|T_2\|^2 \quad (14)$$

for all $\alpha > 0$.

The T_1 part is the typical decentralized SGD items. Inspired by (Koloskova et al., 2020), we write T_1 as:

$$\begin{aligned}
 \left\| \bar{\mathbf{c}}^{(t)} - \mathbf{c}^* - \frac{\eta_t}{n} \sum_{i \in S \cap S^*} \nabla F_i(\mathbf{c}_i^{(t)}) \right\|^2 &\leq \left\| \bar{\mathbf{c}}^{(t)} - \mathbf{c}^* \right\|^2 + \underbrace{\eta_t^2 \frac{n_1^2}{n^2} \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \nabla f_i(\mathbf{c}_i^{(t)}) \right\|^2}_{T_{11}} \\
 &\quad + 2\eta_t \frac{n_1}{n} \underbrace{\left\langle \bar{\mathbf{c}}^{(t)} - \mathbf{c}^*, \frac{-1}{n_1} \sum_{i=1}^{n_1} \nabla f_i(\mathbf{c}_i^{(t)}) \right\rangle}_{T_{12}} + \eta_t^2 v^2
 \end{aligned} \tag{15}$$

We can bound T_{11} and T_{12} separately as:

$$\begin{aligned}
 T_{11} &= \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \left(\nabla f_i(\mathbf{c}_i^{(t)}) - \nabla f_i(\bar{\mathbf{c}}^{(t)}) + \nabla f_i(\bar{\mathbf{c}}^{(t)}) - \nabla f_i(\mathbf{c}^*) \right) \right\|^2 \\
 &\leq \frac{2}{n_1} \sum_{i=1}^{n_1} \left\| \nabla f_i(\mathbf{c}_i^{(t)}) - \nabla f_i(\bar{\mathbf{c}}^{(t)}) \right\|^2 + 2 \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \nabla f_i(\bar{\mathbf{c}}^{(t)}) - \frac{1}{n_1} \sum_{i=1}^{n_1} \nabla f_i(\mathbf{c}^*) \right\|^2 \\
 &= \frac{2L^2}{n_1} \sum_{i=1}^{n_1} \left\| \mathbf{c}_i^{(t)} - \bar{\mathbf{c}}^{(t)} \right\|^2 + 4L \left(f(\bar{\mathbf{c}}^{(t)}) - f(\mathbf{c}^*) \right)
 \end{aligned} \tag{16}$$

$$\begin{aligned}
 -T_{12} &= -\frac{1}{n_1} \sum_{i=1}^{n_1} \left[\left\langle \bar{\mathbf{c}}^{(t)} - \mathbf{c}_i^{(t)}, \nabla f_i(\mathbf{c}_i^{(t)}) \right\rangle + \left\langle \mathbf{c}_i^{(t)} - \mathbf{c}^*, \nabla f_i(\mathbf{c}_i^{(t)}) \right\rangle \right] \\
 &\leq -\frac{1}{n_1} \sum_{i=1}^{n_1} \left[f_i(\bar{\mathbf{c}}^{(t)}) - f_i(\mathbf{c}_i^{(t)}) - \frac{L}{2} \left\| \bar{\mathbf{c}}^{(t)} - \mathbf{c}_i^{(t)} \right\|^2 + f_i(\mathbf{c}_i^{(t)}) - f_i(\mathbf{c}^*) + \frac{\mu}{2} \left\| \mathbf{c}_i^{(t)} - \mathbf{c}^* \right\|^2 \right] \\
 &\leq -\left(f(\bar{\mathbf{c}}^{(t)}) - f(\mathbf{c}^*) \right) + \frac{L + \mu}{2n_1} \sum_{i=1}^{n_1} \left\| \bar{\mathbf{c}}^{(t)} - \mathbf{c}_i^{(t)} \right\|^2 - \frac{\mu}{4} \left\| \bar{\mathbf{c}}^{(t)} - \mathbf{c}^* \right\|^2
 \end{aligned} \tag{17}$$

Now we deal with T_2 . From (Ruan and Joe-Wong, 2022) and (Ghosh et al., 2020) we have the following Lemma:

Lemma 2 (*Mis-classified probability*) For a data point belongs to cluster j , the probability of error classification $\mathbb{P}(\epsilon^{j,j'})$ to cluster $j' \neq j$ can be bound as:

$$\mathbb{P}(\epsilon^{j,j'}) \leq \frac{c_1}{\alpha_0^2 \delta^4} \tag{18}$$

And by union bound, the error probability is bounded as:

$$\mathbb{P}(\bar{\epsilon}) \leq \frac{c_1 S}{\alpha_0^2 \delta^4} \tag{19}$$

The expected number of clients using wrong cluster of data is bounded as:

$$\mathbb{E}[S \cap \bar{S}^*] \leq \frac{c_1 N}{\alpha_0^2 \delta^4} = \epsilon_N \tag{20}$$

for some constant c_1 . We define this bound as ϵ_N

Inspired by (Ghosh et al., 2020), define T_{2k} as the clients selecting the mis-classified data points that should be belongs to cluster k where $k \neq 1$ (The correct cluster). That is:

$$T_{2k} = \sum_{i \in S \cap \bar{S}^* \cap S_k^*} \nabla F_i(\mathbf{c}_i) \tag{21}$$

For each T_{2k} , we use n_k to indicate the number of clients using mis-classified data that should be belongs to cluster k . We have:

$$T_{2k} = \sum_{i=1}^{n_k} \nabla F_i^k(\mathbf{c}_i) + \sum_{i=1}^{n_k} \nabla F_i(\mathbf{c}_i) - \nabla F_i^k(\mathbf{c}_i) \quad (22)$$

Taking the expectation and by Markov's inequality:

$$\begin{aligned} \|T_{2k}\| &= \left\| \sum_{i=1}^{n_k} \nabla F_i^k(\mathbf{c}_i) + \sum_{i=1}^{n_k} \nabla F_i(\mathbf{c}_i) - \nabla F_i^k(\mathbf{c}_i) \right\| \\ &\leq 3n_k L + \frac{\sqrt{n_k} v}{\theta_1} \end{aligned} \quad (23)$$

For any $\theta_1 \in (0, 1)$ with probability equal or greater than $1 - \theta_1$. The above used Lemma 2, Assumption 5 and Assumption 6 and the Markov inequality.

Using the union bound we see that $T_2 = \sum_k T_{2k}$ is bounded as the following with probability greater or equal to $1 - (S - 1)\theta_1 - \theta_2$:

$$\begin{aligned} \|T_2\|^2 &= \left\| \sum_{k=2}^S T_{2k} \right\|^2 \leq (S - 1) \sum_{k=2}^S \|T_{2k}\|^2 \\ &\leq \frac{18L^2 \epsilon_N^2}{\theta_2^2} + \frac{2\epsilon_N (S - 1)v^2}{\theta_1^2 \theta_2} \end{aligned} \quad (24)$$

When $\sum_{k=2}^S n_k \leq \frac{\epsilon_N}{\theta_2}$ with probability at least $1 - \theta_2$.

Combining the above three terms and Lemma 2, we have:

$$\begin{aligned} \mathbb{E} \left\| \bar{\mathbf{c}}^{(t+1)} - \mathbf{c}^* \right\|^2 &\leq (1 - \eta_t \mu + \eta_t \mu \frac{\epsilon_N}{n}) \left\| \bar{\mathbf{c}}^{(t)} - \mathbf{c}^* \right\|^2 + \frac{18L^2 \epsilon_N^2 \eta_t^2}{n^2} + \frac{2\epsilon_N (S - 1)v^2 \eta_t^2}{n^2} + \eta_t^2 v^2 \\ &\quad + \frac{\eta_t (L + \mu)}{n} \sum_{i=1}^{n_1} \left\| \bar{\mathbf{c}}^{(t)} - \mathbf{c}_i^{(t)} \right\|^2 + \left(\frac{4\eta_t^2 (n - \epsilon_N)^2 L}{n^2} + 2\eta_t - \frac{2\eta_t \epsilon_N}{n} \right) \left(f(\bar{\mathbf{c}}^{(t)}) - f(\mathbf{c}^*) \right) \end{aligned} \quad (25)$$

A.2 Proof of Theorem 2

For time $t - 1$, after the local updating round, the cluster parameters can be expressed as:

$$\mathbf{C}^{t-1'} = \mathbf{C}^{t-1} - \eta_t \mathbf{G}^{t-1} \quad (26)$$

After the communication round, the parameters can be expressed as:

$$\mathbf{C}^t = \mathbf{C}^{t-1'} \mathbf{W}^{t-1} = \mathbf{C}^{t-1} \mathbf{W}^{t-1} - \eta_t \mathbf{G}^{t-1} \mathbf{W}^{t-1} \quad (27)$$

Thus, recursively expanding the parameters at time t back to $l\beta$, we can get the final form:

$$\mathbf{C}^t = \mathbf{C}^{l\beta} \prod_{m=l\beta}^{t-1} \mathbf{W}^m - \sum_{m=l\beta}^{t-1} \left(\eta_t \mathbf{G}^m \prod_{r=t-1}^m \mathbf{W}^r \right) \quad (28)$$

A.3 Proof of Theorem 3

Following the same flow of Lemma 9 in (Koloskova et al., 2020), applying Theorem 3, we have for all $\alpha > 0$:

$$\begin{aligned}
 \mathbb{E}\|\mathbf{C}^t - \bar{\mathbf{C}}^t\|_F^2 &= N\mathbf{E}_t \leq \mathbb{E} \left\| \mathbf{C}^{(m\beta)} \prod_{i=t-1}^{m\beta} \mathbf{W}^{(i)} - \bar{\mathbf{C}}^{(m\beta)} + \sum_{j=m\beta}^{t-1} \eta_j \nabla \mathbf{F}(\mathbf{C}^{(j)}) \prod_{i=t-1}^j \mathbf{W}^{(i)} \right\|_F^2 \\
 &\leq \mathbb{E} \left\| \mathbf{C}^{(m\beta)} \prod_{i=t-1}^{m\beta} \mathbf{W}^{(i)} - \bar{\mathbf{C}}^{(m\beta)} + \sum_{j=m\beta}^{t-1} \eta_j \left(\nabla \mathbf{F}(\mathbf{C}^{(j)}) - \nabla \mathbf{F}(\mathbf{C}^*) + \nabla \mathbf{f}(\mathbf{C}^*) \right) \prod_{i=t-1}^j \mathbf{W}^{(i)} \right\|_F^2 \\
 &\quad + \left\| \sum_{j=m\beta}^{t-1} \eta_j \left(\nabla \mathbf{F}(\mathbf{C}^*) + \nabla \mathbf{f}(\mathbf{C}^*) \right) \prod_{i=t-1}^j \mathbf{W}^{(i)} \right\|_F^2 \\
 &\leq (1 + \alpha) \mathbb{E} \left\| \mathbf{C}^{(m\beta)} \prod_{i=t-1}^{m\beta} \mathbf{W}^{(i)} - \bar{\mathbf{C}}^{(m\beta)} \right\|_F^2 \\
 &\quad + (1 + \alpha^{-1}) \mathbb{E} \left\| \sum_{j=m\beta}^{t-1} \eta_j \left(\nabla \mathbf{F}(\mathbf{C}^{(j)}) - \nabla \mathbf{F}(\mathbf{C}^*) + \nabla \mathbf{f}(\mathbf{C}^*) \right) \prod_{i=t-1}^j \mathbf{W}^{(i)} \right\|_F^2 \\
 &\quad + \left\| \sum_{j=m\beta}^{t-1} \eta_j \left(\nabla \mathbf{F}(\mathbf{C}^*) + \nabla \mathbf{f}(\mathbf{C}^*) \right) \prod_{i=t-1}^j \mathbf{W}^{(i)} \right\|_F^2
 \end{aligned} \tag{29}$$

Using Assumption 7, the above can be further simplified:

$$\begin{aligned}
 \mathbb{E}\|\mathbf{C}^t - \bar{\mathbf{C}}^t\|_F^2 &\leq (1 + \alpha)(1 - p) \mathbb{E} \left\| \mathbf{C}^{(m\beta)} - \bar{\mathbf{C}}^{(m\beta)} \right\|_F^2 \\
 &\quad + (1 + \alpha^{-1}) 2\beta \sum_{j=m\beta}^{t-1} \eta_j^2 \mathbb{E} \left\| \left(\nabla \mathbf{F}(\mathbf{C}^{(j)}) - \nabla \mathbf{F}(\mathbf{C}^*) + \nabla \mathbf{f}(\mathbf{C}^*) \right) \right\|_F^2 \\
 &\quad + \sum_{j=m\beta}^{t-1} \eta_j^2 \mathbb{E} \left\| \left(\nabla \mathbf{F}(\mathbf{C}^*) + \nabla \mathbf{f}(\mathbf{C}^*) \right) \right\|_F^2 \\
 &\leq (1 + \alpha)(1 - p) \mathbb{E} \left\| \mathbf{C}^{(m\beta)} - \bar{\mathbf{C}}^{(m\beta)} \right\|_F^2 \\
 &\quad + (1 + \alpha^{-1}) 2\beta \sum_{j=m\beta}^{t-1} \eta_j^2 \mathbb{E} \left\| \left(\nabla \mathbf{F}(\mathbf{C}^{(j)}) - \nabla \mathbf{F}(\mathbf{C}^*) + \nabla \mathbf{f}(\mathbf{C}^*) \right) \right\|_F^2 \\
 &\quad + \sum_{j=m\beta}^{t-1} \eta_j^2 n v^2
 \end{aligned} \tag{30}$$

The expectation of the second term on the right-hand side can be bounded as:

$$\begin{aligned}
 &\mathbb{E} \left\| \left(\nabla \mathbf{F}(\mathbf{C}^{(j)}) - \nabla \mathbf{F}(\mathbf{C}^*) + \nabla \mathbf{f}(\mathbf{C}^*) \right) \right\|_F^2 \\
 &= \mathbb{E} \left\| \left(\nabla \mathbf{F}(\mathbf{C}^{(j)}) - \nabla \mathbf{F}(\bar{\mathbf{C}}) + \nabla \mathbf{F}(\bar{\mathbf{C}}) - \nabla \mathbf{F}(\mathbf{C}^*) + \nabla \mathbf{f}(\mathbf{C}^*) \right) \right\|_F^2 \\
 &\leq 3 \frac{n}{N} L^2 \|\mathbf{C}^{(j)} - \bar{\mathbf{C}}^{(j)}\|_F^2 + 3n\sigma^2 + 6nL(f(\bar{\mathbf{c}}^j) - f(\mathbf{c}^*)) \\
 &\leq 3 \frac{n}{N} L^2 \|\mathbf{C}^{(j)} - \bar{\mathbf{C}}^{(j)}\|_F^2 + 3n\sigma^2 + 6nL(f(\bar{\mathbf{c}}^j) - f(\mathbf{c}^*))
 \end{aligned} \tag{31}$$

Putting the above equations together and setting a proper α to make the first term become $1 - \frac{p}{2}$, similar to (Koloskova et al., 2020) with stepsize $\eta_j \leq \frac{p\sqrt{N}}{12\sqrt{2n\beta}L}$, we can get the desired bound:

$$\begin{aligned} \mathbf{E}_t &\leq \left(1 - \frac{p}{2}\right)\mathbf{E}_{m\beta} + \frac{p}{16\beta} \sum_{j=m\beta}^{t-1} \mathbf{E}_j + \frac{36Ln\beta}{pN} \sum_{j=m\beta}^{t-1} \eta_j^2 (f(\bar{\mathbf{c}}^j) - f(\mathbf{c}^*)) \\ &\quad + \left(\frac{18\beta n}{Np}\sigma^2 + \frac{n}{N}v^2\right) \sum_{j=m\beta}^{t-1} \eta_j^2 \end{aligned} \quad (32)$$

A.4 Proof of Theorem 4

We adapted the following Lemma 3 from (Koloskova et al., 2020):

Lemma 3 (*Simplify the Recursive Equations*) For a bound of the cluster distance to the optimal $d_t = \mathbb{E}\|\bar{\mathbf{c}}^{(t)} - \mathbf{c}^*\|^2$ in the following form:

$$d_{t+1} \leq (1 - a\eta_t) d_t - b\eta_t e_t + c\eta_t^2 + \eta_t B \mathbf{E}_t, \quad (33)$$

and for any non-negative sequences $\{\mathbf{E}_t\}_{t \geq 0}$, $\{e_t\}_{t \geq 0}$, $\{\eta_t\}_{t \geq 0}$ that satisfy the following form:

$$\mathbf{E}_t \leq \left(1 - \frac{p}{2}\right) \mathbf{E}_{m\beta} + \frac{p}{16\beta} \sum_{j=m\beta}^{t-1} \mathbf{E}_j + D \sum_{j=m\beta}^{t-1} \eta_j^2 e_j + A \sum_{j=m\beta}^{t-1} \eta_j^2, \quad (34)$$

then if the learning rate $\{\eta_t\}_{t \geq 0}$ and $\{r_t\}_{t \geq 0}$ are respectively a $\frac{8\beta}{p}$ -slow decreasing sequence and $\frac{16\beta}{p}$ -slow increasing non-negative sequence, then for some constant $E > 0$ with learning rate $\eta_t \leq \frac{1}{16} \sqrt{\frac{pb}{DB\beta}}$ the following holds:

$$E \sum_{t=0}^T r_t \mathbf{E}_t \leq \frac{b}{2} \sum_{t=0}^T r_t e_t + 64BA \frac{\beta}{p} \sum_{t=0}^T r_t \eta_t^2 \quad (35)$$

By combining the above equations we have:

$$\frac{1}{2R_T} \sum_{t=0}^T b r_t e_t \leq \frac{1}{R_T} \sum_{t=0}^T \left(\frac{(1 - a\eta_t) r_t}{\eta_t} d_t - \frac{r_t}{\eta_t} d_{t+1} \right) + \frac{c}{R_T} \sum_{t=0}^T r_t \eta_t + \frac{64BA}{R_T} \sum_{t=0}^T r_t \eta_t^2 \quad (36)$$

Where $R_T = \sum_{t=0}^T r_t$

Following the previous Lemma, we adapt Lemma 13 from (Koloskova et al., 2020) as the following Lemma 4

Lemma 4 (*Main Recursion*) The main recursion can be bounded as the following with a constant step-size $\eta_t = \eta < \frac{1}{h}$:

$$\frac{1}{2R_T} \sum_{t=0}^T b e_t r_t + a d_{T+1} \leq \tilde{\mathcal{O}} \left(d_0 h \exp \left[-\frac{a(T+1)}{h} \right] + \frac{c}{aT} + \frac{BA}{a^2 T^2} \right) \quad (37)$$

For the following two cases, tuning η we have: If $\frac{1}{h} \geq \frac{\ln(\max\{2, a^2 d_0 T^2 / c\})}{aT}$ η is chosen to be equal to this value and that:

$$\tilde{\mathcal{O}} \left(a d_0 T \exp \left[-\ln(\max\{2, a^2 d_0 T^2 / c\}) \right] \right) + \tilde{\mathcal{O}} \left(\frac{c}{aT} \right) + \tilde{\mathcal{O}} \left(\frac{BA}{a^2 T^2} \right) = \tilde{\mathcal{O}} \left(\frac{c}{aT} \right) + \tilde{\mathcal{O}} \left(\frac{BA}{a^2 T^2} \right) \quad (38)$$

If else choose $\eta = \frac{1}{h}$ and that:

$$\tilde{\mathcal{O}} \left(d_0 h \exp \left[-\frac{a(T+1)}{h} \right] + \frac{c}{h} + \frac{BA}{h^2} \right) \leq \tilde{\mathcal{O}} \left(d_0 h \exp \left[-\frac{a(T+1)}{h} \right] + \frac{c}{aT} + \frac{BA}{a^2 T^2} \right) \quad (39)$$

Using the above Lemma 3, Lemma 4 and Theorem 1 and Theorem 3, we can get the final bound.

B Simulation Details and Additional Simulations

B.1 Experiment Details

The following shows the detailed settings of our experiments. We largely follow Marfoq et al. (2021); Ruan and Joe-Wong (2022) in our experiment settings.

B.1.1 MNIST/EMNIST Data

Half of the dataset was selected to undergo a 90-degree rotation. Each client received the same amount of data, but the ratio of rotated to non-rotated data was set uniformly at random in the range from 10% and 90%. The number of clients was fixed at $N = 100$ for comparison with the baselines. A CNN (convolutional neural network) model was employed, consisting of two convolutional layers with kernel size and padding set to 5 and 2, respectively. Each convolutional layer was followed by a max-pooling layer with a kernel size of 2. After the convolutional layers, fully connected layers were used, with a dropout layer of size 50. The ReLU activation function was applied to each convolution layer and fully-connected layer. All clients utilized SGD as the optimizer. The number of local epochs was set to 5, with the initial step having double the local epochs to accelerate the initial learning, leading to a faster reduction in global loss. The initial learning rate was $5e-2$, with a decay factor of 0.80. Training was carried out over 150 global epochs.

B.1.2 CIFAR-10 & CIFAR-100 Data

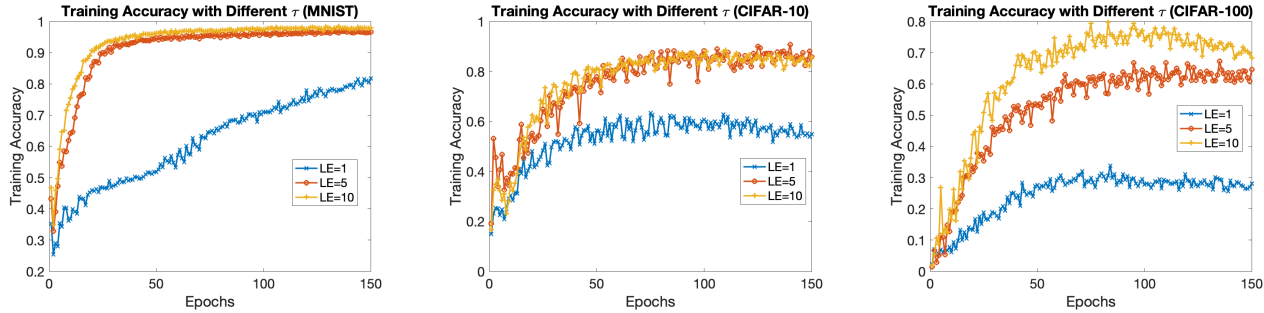
The dataset was divided into even and odd labels by its number of label marked in the dataset, and half of the data was randomly selected to undergo a 90-degree rotation. This process potentially created four different data distributions (rotated even, un-rotated even, rotated odd, un-rotated odd). Each client received an equal amount of data, but the proportion of odd-labeled and even-labeled data was randomly assigned, ranging uniformly at random from 10% to 90%. The number of clients was set to $N = 25$ for comparison with the baselines. A CNN model with four convolutional layers was used. The first two layers had a kernel size and padding of 5 and 2, respectively, while the last two layers had a kernel size and padding of 3 and 1, respectively. Each convolutional layer was followed by batch normalization. After the second and fourth convolutional layers, max-pooling with a kernel size of 2 and a dropout layer were applied. Following the convolutional layers, two fully connected layers with dropout and batch normalization were used, containing 1024 and 512 hidden neurons, respectively. The activation function was ReLU on each layer. All clients used SGD as the optimizer. The number of local epochs was set to 5, with the initial step doubling the local epochs. The initial learning rate was set to $5e-2$, with a decay factor of 0.85. Training was conducted for 150 global epochs.

B.2 Additional Simulation Results

B.2.1 Effect of τ

We conducted experiments for 150 epochs on the MNIST, CIFAR-10, and CIFAR-100 datasets. As shown in Figure 5, increasing the number of local epochs in FedSPD leads to faster convergence. For $\tau = 1$, the training did not converge even after 150 epochs on MNIST, and for CIFAR-10 and CIFAR-100, it seemed to converge to a lower training accuracy. We observed that as the dataset and model complexity increased, increasing the number of local epochs tended to improve performance.

Table 5 presents the final FedSPD testing accuracies for different numbers of local epochs across the datasets. On MNIST, the testing accuracies were 93.27% and 93.47%, respectively, showing only a slight difference, likely because the MNIST dataset is relatively simple, so the learning hyperparameters do not make much of a difference in model performance. For CIFAR-10, the testing accuracies for $\tau = 5$ and $\tau = 10$ were 70.61% and 66.52%, respectively, where a larger number of local epochs actually reduced the final performance. However, for CIFAR-100, $\tau = 10$ resulted in the best performance. This suggests that for more complex datasets, a higher number of local epochs can be beneficial, as indicated by the training accuracy curves. Nevertheless, it is important to note that setting τ too high may lead to overfitting to the local data, as was the case with $\tau = 10$ on the CIFAR-10 dataset. These findings are consistent with known results in general federated learning, where a higher number of local epochs can effectively increase the number of gradient steps taken, accelerating convergence as long as the local models do not diverge too much due to a large number of local steps.



(a) Training accuracy on MNIST. (b) Training accuracy on CIFAR-10. (c) Training accuracy on CIFAR-100.

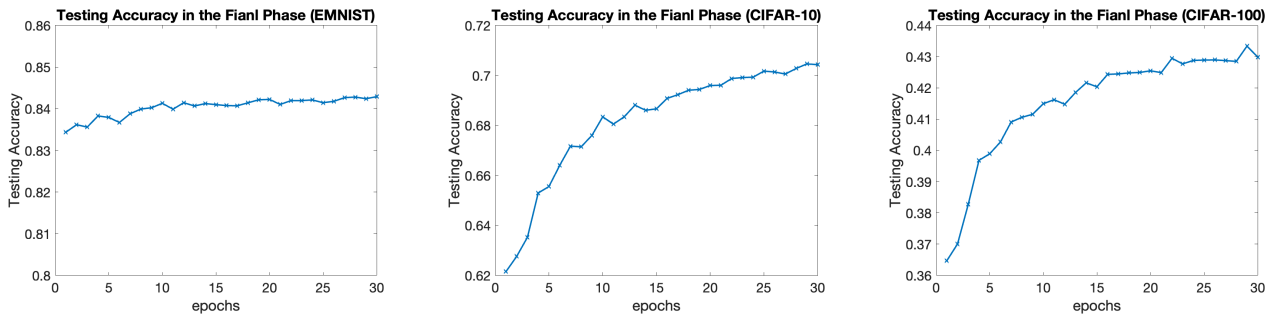
Figure 5: *FedSPD* training accuracy with different numbers of local steps τ . When the data become more complicated, increasing local epochs may be a better choice.

Local Epochs	1	5	10
MNIST	74.20	93.27	93.47
CIFAR-10	41.34	70.61	66.52
CIFAR100	19.86	43.35	44.99

Table 5: Final *FedSPD* testing accuracies for different number of local epochs.

B.2.2 Influence of the Final Phase

Our *FedSPD* algorithm uses a final phase that follows the typical federated learning training process. The optimal number of epochs for this final phase varies depending on the dataset and learning model. Due to the simplicity of EMNIST and its model, the testing accuracy is already sufficiently high after aggregation. In our EMNIST setup, using 10 epochs in the final phase increases performance by 0.5%, and beyond 10 epochs, the testing accuracy stabilizes. For CIFAR-10 and CIFAR-100, the testing accuracy improves by 7% and 6%, respectively, after 15 epochs. Around 30 epochs are sufficient to achieve optimal performance for both datasets. It is important to note that choosing the correct number of epochs and learning rate for this final phase is crucial. Too many epochs, or a learning rate that is too high (or with insufficient decay), may lead to overfitting to the local data. Since this final phase is trained locally without any communication overhead, it presents a key advantage of our *FedSPD* algorithm in communication-constrained settings. Additionally, note that for EMNIST, CIFAR-10, and CIFAR-100, our *FedSPD* already achieves higher accuracies compared to other methods, even without this final phase. Other algorithms like FedEM perform aggregation during the regular training phase, so adding extra local rounds in a final phase of training may lead to overfitting.



(a) Testing accuracy on EMNIST. (b) Testing accuracy on CIFAR-10. (c) Testing accuracy on CIFAR-100.

Figure 6: Testing accuracy of the final phase.

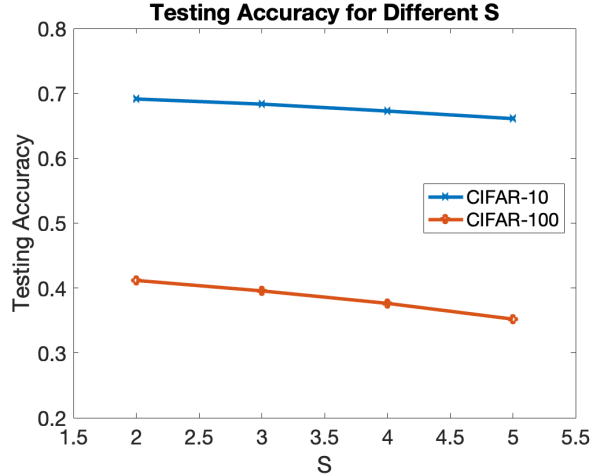


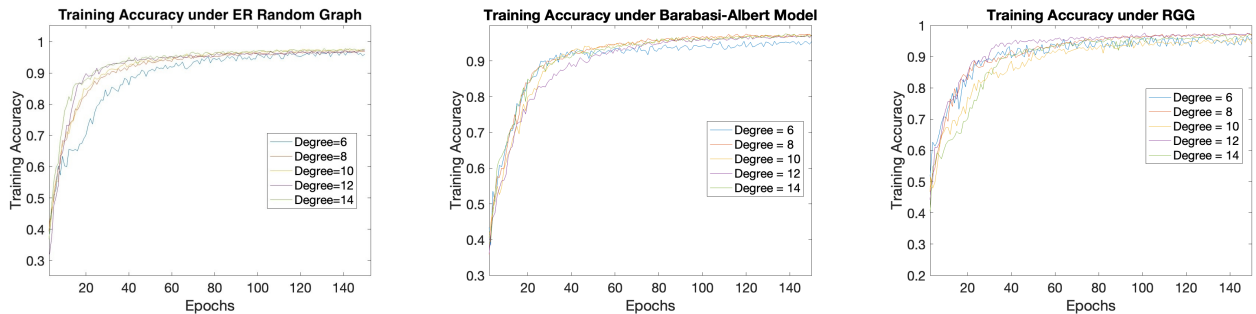
Figure 7: FedSPD testing accuracy for different numbers of clusters S .

B.2.3 Influence of the Hyperparameter S

The testing accuracy with different hyperparameters S (number of clusters) for the CIFAR-10 and CIFAR-100 datasets is shown in Figure 7. In the experimental settings, we potentially created four different distributions by using varying labels and image rotations. In our FedSPD algorithm, setting S too high does not necessarily improve performance. This may be because most practical loss functions, such as the cross-entropy used in neural networks, are non-convex, meaning that the aggregated model may not perform optimally in practice. Aggregating more models in the final phase can exacerbate this issue. However, in our FedSPD algorithm, setting $S = 2$ already yields excellent performance in terms of the final test accuracy.

B.2.4 Extra Details for Experiments with Different Graph Connectivity

FedSPD’s training accuracy versus epochs for MNIST across different topologies is shown in Figure 8. We observe that networks with lower connectivity typically converge more slowly than those with higher connectivity, in each topology. Additionally, RGG exhibits more oscillations compared to other topologies, likely due to its high clustering effect (Penrose, 2003). However, all topologies eventually reach the same level of training accuracy, regardless of the network structure, indicating that, as predicted by Theorem 4, *FedSPD* converges as long as the network is connected.



(a) Training accuracy of ER Graph.

(b) Training accuracy of BA Model.

(c) Training Accuracy of RGG.

Figure 8: *FedSPD* converges slightly faster on networks of higher average degree, with noisier convergence on highly clustered RGG graphs, on MNIST Data.

B.2.5 Additional Comments on the Main Results

As shown in Table 2, local learning performs the worst among all algorithms, validating that all other methods benefit from exchanging information between clients to learn a better model. Among the DFL algorithms,

FedAvg, the only one without personalization, typically performs the worst, indicating that personalization is beneficial in non-iid data distributions, as we would intuitively expect. However, an exception is observed with the FedSoft algorithm. In the CIFAR-10 and CIFAR-100 datasets, FedSoft performs poorly, nearing the accuracy of local training. We conjecture that this is due to the way FedSoft aggregates models, making it difficult to learn the correct cluster centers in a low-connectivity network, leading to suboptimal performance.