Personalized Federated Learning under Local Supervision

Supplementary Material

7. Appendix-Experiments

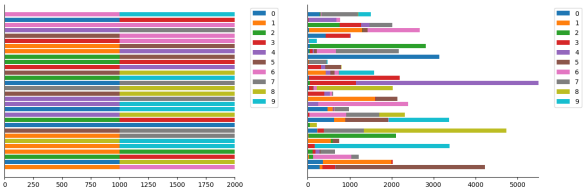
7.1. Datasets

In this work, we perform experiments on three datasets, i.e., CIFAR10, CIFAR100 and IMAGENET. CIFAR10 and CIFAR100 are among the most classic image classification tasks, both containing 60, 000 images, evenly distributed across 10 and 100 categories, respectively. FEMNIST is a dataset with 62 different character categories (including numbers and uppercase and lowercase English letters), with a total of 805,263 samples. IMAGENET is a significantly more complex image classification dataset compared to CIFAR10/100, featuring a larger number of images, higher resolution, and a greater variety of categories. We used the ILSVRC2012 training set (138 GB, 1,331,167 images), assigning approximately 13,000 images per client under the Dirichlet(100, 0.1) split. Images were resized to 64×64 using a single-crop method, matching Tiny ImageNet. Since this differs from the sizes in CIFAR and FEMNIST, we adjusted the linear layers in LeNet-5 to fit the new convolutional output, while keeping the convolutional layer parameters unchanged.

7.2. Data partitioning

Our data partitioning only considers the label differences between clients.

Pathological Non-IID partition. In pathological distribution, we first need to determine the number of categories c to be distributed to each client. We will partition the data based on the total amount of data, the number of categories, the number of clients, ensuring that each piece of data does not appear more than once and that all data is utilized. We present our partitioning on CIFAR10, as shown in Figure 5a.



(a) Pathological Non-IID partition (b) Dirichlet Non-IID partition

Figure 5. Partitioning on CIFAR10

Dirichlet Non-IID partition. In the Dirichlet distribution, the distribution for each client is independent. Assume that the distribution for client is governed by a vector q ($q_i > 0, i \in [1, M], \|q\|_1 = 1$) of length M, where M represents the number of classes. The vector q is sampled from a Dirichlet distribution

$$q \sim Dir(\alpha p) \tag{19}$$

Dir p ($p_i > 0, i \in [1, M], \|p\|_1 = 1$) represents the prior class distribution that we manually set. Here, we define them as $p_i = \frac{1}{M}, i \in [1, M]$. The parameter α is a concentration parameter, which can be simply understood as determining the probability that a sample belongs to the prior p. When each element in p is the same, the probability density function of the Dirichlet distribution is given by:

$$Dir\left(q \mid \alpha p\right) = \frac{1}{B\left(\alpha p\right)} \prod_{i=1}^{M} q_i^{\alpha p_i - 1}, \tag{20}$$

$$B\left(\alpha p\right) = \frac{\prod_{i=1}^{M} \Gamma\left(\alpha p_i\right)}{\Gamma\left(\sum_{i=1}^{M} \alpha p_i\right)}. \tag{21}$$

And $E\left(q_i\right) = p_i$. We can see from Eq. (20) that when αp_i is large, our samples are nearly $q_i = \frac{1}{M}, i \in [1, M]$, whereas when αp_i is small, only one category appears in the samples. Therefore, we can set the size of αp to control the degree of Non-IID data. Since each element in p is the same and we are only concerned with the size of αp, we can set just one variable α to automatically normalize p and control the generation of the desired data.

However, this partitioning method still presents some issues. First, different clients may have overlapping data, or certain data in the dataset may not be utilized. Second, the number of samples for each client is predetermined and the same across all clients, which is almost impossible in real-world scenarios because clients vary in their ability to collect data. Therefore, we apply the Dirichlet distribution to the data for each class, where q and p become vectors of size n, where n is the number of clients. During the partitioning process, we need to ensure that a larger portion of the data is allocated to clients with fewer overall data points to maintain a Non-IID distribution. However, a problem arises when there are too many clients: insufficient data may result in some clients having too little data after all categories have been split. In this case, we can repartition the data until the client with the least amount of data reaches the required threshold. We present our partitioning on CIFAR10, as shown in Fig. 5b.

7.3. Baselines and training details

Details of Baselines. These methods are selected for their relevance to our approach (see Related Work) or their strong performance.In FedProx [26], a proximal term is used to improve stability. Per-FedAvg [14] proposes using the

	CIFAR10				CIFAR100			
No. of Clients (Dir)	100 (0.1)	50 (0.1)	100 (0.5)	50 (0.5)	100 (0.1)	50 (0.1)	100 (0.5)	50 (0.5)
FedSimSup	.892(12)	.882(21)	.725(207)	.736(174)	.503(107)	.546(92)	.331(327)	.385(245)
FedSimSup*	87	96	342	253	276	245	572	473
FedSimSup**	.874	.843	.712	.720	.473	.496	.307	.327

Table 5. * denotes "without similarity information" and ** denotes "using the serial architecture". Values in () for FedSimSup and in the * column indicate the number of epochs to reach 0.60 and 0.20 acc, showing that similarity-based aggregation accelerates convergence. All other values are final accuracy.

MAML framework to obtain an initial model that quickly adapts to clients. FedRep [10] sets up a unique head for each client to enhance personalization capability. FedProto [45] aggregates the local prototypes to avoid gradient misalignment. FedPac [48] performs explicit local-global feature alignment by leveraging global semantic knowledge. pFedFda [32] employs a generative classifier for global representation learning while adapting the classifier to the local client distributions. FedAs [50] leverages the alignment of client-side parameters and the synchronization of server-side clients to overcome the challenge of intra-client and inter-client inconsistency in pFL methods, respectively.

Settings for Baselines. In the FedAvg method, we set the client participation rate to 0.1, the global communication rounds to 1000, and the local epochs to 5. For other methods, unless otherwise specified, the parameters remain the same. In the FedProx method, we set the coefficient of the proximal term μ to 1 to improve stability. In the FedPac method, we set the hyper-parameter to balance supervised loss and regularization loss λ to 1. In the Per-FedAvg method, we set steps of stochastic gradient descent locally τ to 4 and stepsize α to 0.001, and use Per-FedAvg (HF). During testing, each client performs fine-tuning for 3 epochs. In the FedRep method, we set the classification head as the personalized layer, training the classification head for 2 epochs and the representation layer for 3 epochs. In the FedProto method, we set the importance weight λ to 1. For local training, we randomly select clients at a proportional rate in each round and conduct training, but do not perform aggregation. This means that the client's model will only change after the client participates in communication.

Training Details. We set the global communication rounds T to 1,000 and the local training epochs to 5, with $\tau_\theta = 2$ epochs dedicated to training the inter-learning model and $\tau_s = 2$ epochs for training the supervisor. Parameters C and γ in (7) are set to 40 and 3/7, respectively. For CIFAR10 and CIFAR100, we set the number of clients n to 50 and 100 with a participation rate r of 0.1 per round. For FEMNIST we maintain its original setup with a total of 3,597 clients to ensure that our method remains effective under a large number of clients and the participation rate r is set to 0.1 for local training and 0.01 for the nine methods besides local training. We set the batch size for SGD to 32 and the learning rate to 0.1.

7.4. Ablation Experiments

We conduct ablation experiments under all Dirichlet configurations on CIFAR-10 and CIFAR-100. Tab. 5 presents the results.The results confirm the effectiveness of our design. Using similarity-based aggregation leads to faster convergence compared to naive averaging, and the proposed parallel architecture outperforms the serial alternative.

7.5. Supervisor Assistance

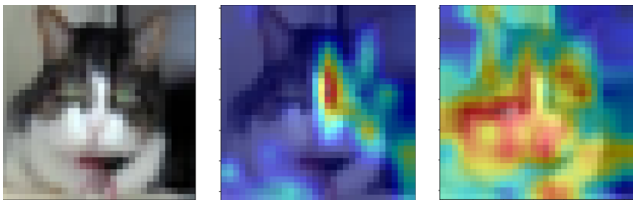


Figure 6. CAM of the inter-learning model (middle) and the supervisor (right).

We verify the assistance effect of the supervisor using Class Activation Map (CAM) [39] in image classification tasks. As shown in Fig. 6, the image on the left is the original classification task image, the middle one is the CAM of the inter-learning model, and the one on the right is the CAM of the supervisor. It can be observed that, when trying to recognize the image as a cat, the inter-learning model, possibly influenced by information learned from other clients, tends to focus on scattered details, such as the cat's eyes or nose. In contrast, the supervisor focuses on the entire body of the cat, helping to prevent the inter-learning model's attention from deviating too much. Thus, we conclude that the supervisor and inter-learning model in our FedSimSup have different focuses, enhancing the explainability of the model's behavior. To further demonstrate the auxiliary role of the supervisor, we evaluated the supervisor's impact in a new inter-learning model by testing non-participating clients after each round. The results show that clients with supervisors performed better and more stably, while those without supervisors exhibited greater fluctuations.

8. Convergence Analysis: Full Proofs

We give the full convergence proofs here. The outline of this section is:

- Sec. 8.1: Review of assumptions and main theorem;
- Sec. 8.2: The full proof of Theorem 1;
- Sec. 8.3: Claims used in the analysis.

8.1. Assumptions and Main theorem

For integrity, we rewrite the assumptions from the main paper as follows. Note that these assumptions are standard and widely used in convergence analysis in federated learning [14, 16, 26, 27, 35].

Assumption 1 (Bounded Loss)**.** *There exists constant $F^* \in \mathbb{R}$ such that for any client $i \in \{1, \dots, n\}$, f_i is bounded from below by F^*, $f_i(s, \theta) > F^*, \forall s, \theta$.*

Assumption 2 (Smoothness)**.** *There exists $L > 0$ such that for any client $i \in \{1, \dots, n\}$, $\nabla_s f_i(\cdot, \theta)$, $\nabla_s f_i(s, \cdot)$, $\nabla_\theta f_i(\cdot, \theta)$ and $\nabla_\theta f_i(s, \cdot)$ are L-Lipschitz.*

Assumption 3 (Bounded Gradient)**.** *For all $i \in \{1, \dots, n\}$, the gradient of loss function f_i is bounded. There exists $G > 0$ such that*

$$\|\nabla_s f_i(s, \theta)\| \leq G, \quad \|\nabla_\theta f_i(s, \theta)\| \leq G, \quad \forall s, \theta. \tag{22}$$

Assumption 4 (Unbiasedness)**.** *SGD estimator is unbiased. There exists $\sigma > 0$ such that for any client $i \in \{1, \dots, n\}$,*

$$\begin{aligned} \mathbb{E}[SGD(f_i(s, \theta), s)] &= \nabla_s f_i(s, \theta), \quad \forall s, \theta, \\ \mathbb{E}[SGD(f_i(s, \theta), \theta)] &= \nabla_\theta f_i(s, \theta), \quad \forall s, \theta. \end{aligned} \tag{23}$$

Assumption 5 (Bounded Variance)**.** *The variance of SGD estimator is bounded. That is, for any client $i \in \{1, \dots, n\}$,*

$$\mathbb{E}\Big[\|SGD(f_i(s, \theta), s) - \nabla_s f_i(s, \theta)\|^2\Big] \leq \sigma^2, \forall s, \theta. \tag{24}$$

$$\mathbb{E}\Big[\|SGD(f_i(s, \theta), \theta) - \nabla_\theta f_i(s, \theta)\|^2\Big] \leq \sigma^2, \forall s, \theta. \tag{25}$$

8.2. Convergence analysis in FedSimSup

With the above assumptions in Sec. 8.1, we restate the convergence of the proposed FedSimSup as follows.

Theorem 1 (Convergence of FedSimSup)**.** *Suppose Assumptions 1 to 5 hold, and the learning rates in FedSimSup are chosen as*

$$\eta_s^t = \eta / \sqrt{T} L \tau_s, \quad \eta_\theta^t = \eta / \sqrt{T} L \tau_\theta \tag{26}$$

with $\eta < 1$, then we have the following bound.

$$\begin{aligned} &\frac{1}{T} \sum_{i=0}^{T-1} \left[\left\| \nabla_s f_i(s_i^{t,0}, \theta_i^{t,0}) \right\|^2 + 2 \left\| \nabla_\theta f_i(s_i^{t,0}, \theta_i^{t,0}) \right\|^2 \right] \\ \leq\ & \frac{L}{2T^{1/2} r \eta} (f_i(s_i^{0,0}, \theta_i^{0,0}) - F^*) + \frac{3\eta^2 G^2}{2T} + \frac{\eta \sigma^2}{T^{1/2} \bar{\tau}} \\ &+ \frac{(1-r)\lambda_i G^2}{2r} \Big(- C^2 (1 - 2\gamma) T^{2\gamma - 1} \ln T \\ &+ C^2 (1 + 2 \ln C) T^{2\gamma - 1} \\ &+ 3C T^{\gamma - 1} + 1/T - C^2 T^{2\gamma - 2} \Big) \\ &+ \frac{(1-r)\lambda_i^2 \eta^2 G^2}{2rL} \Big(3C T^{3\gamma - 3/2} + 9C^2 T^{2\gamma - 3/2} \\ &+ 3C T^{\gamma - 3/2} - 3C^4 T^{4\gamma - 9/2} \Big). \end{aligned} \tag{27}$$

Here, $\bar{\tau} = 2/(1/\tau_s + 1/\tau_\theta)$ and $\lambda_i = \max_{|\mathcal{N}| \subset \{1, \dots, n\}} (\sum_{j \in \mathcal{N}} m_j) / (\sum_{j \in \mathcal{N}} m_j + K m_i)$.

Proof. In the communication round t, the probability of each client i being sampled is r. Therefore, function value $f_i(s_i^{t,\tau_s}, \theta_i^{t,\tau_\theta})$ can be computed as follows,

$$\begin{aligned} &\mathbb{E}_t \big[f_i(s_i^{t,\tau_s}, \theta_i^{t,\tau_\theta}) \big] \\ =\ & \mathbb{E}_t \big[f_i(s_i^{t,\tau_s}, \theta_i^{t,\tau_\theta}) \mid i \in \mathcal{N}(t) \big] \Pr[i \in \mathcal{N}(t)] \\ &+ \mathbb{E}_t \big[f_i(s_i^{t,\tau_s}, \theta_i^{t,\tau_\theta}) \mid i \notin \mathcal{N}(t) \big] \Pr[i \notin \mathcal{N}(t)] \\ =\ & r \mathbb{E}_t \big[f_i(s_i^{t,\tau_s}, \theta_i^{t,\tau_\theta}) \mid i \in \mathcal{N}(t) \big] \\ &+ (1 - r) \mathbb{E}_t \big[f_i(s_i^{t,\tau_s}, \theta_i^{t,\tau_\theta}) \mid i \notin \mathcal{N}(t) \big], \end{aligned} \tag{28}$$

where $\mathbb{E}_t = \mathbb{E}\Big[\cdot \mid s_i^{t,0}, \theta_i^{t,0} \Big]$.

We plug Claims 1 and 2 into Eq. (28), then we can upper bound $\mathbb{E}_t \big[f_i(s_i^{t,\tau_s}, \theta_i^{t,\tau_\theta}) \big]$ as follows.

$$\begin{aligned} &\mathbb{E}_t \big[f_i(s_i^{t,\tau_s}, \theta_i^{t,\tau_\theta}) \big] \\ \leq\ & f_i(s_i^{t,0}, \theta_i^{t,0}) + r \cdot \frac{3\eta^3 G^2}{T^{3/2} L} + r \cdot \frac{\eta^2 \sigma^2}{TL\bar{\tau}} \\ &- r \cdot \frac{\eta}{\sqrt{T} L} \left\| \nabla_s f_i(s_i^{t,0}, \theta_i^{t,0}) \right\|^2 \\ &- r \cdot \frac{2\eta}{\sqrt{T} L} \left\| \nabla_\theta f_i(s_i^{t,0}, \theta_i^{t,0}) \right\|^2 \\ &+ (1 - r) \cdot \frac{\lambda_i \beta^t (2t + 1) \eta G^2}{T^{1/2} L} \\ &+ (1 - r) \cdot \frac{(\lambda_i \beta^t)^2 (2t + 1)^2 \eta^2 G^2}{TL^2} \end{aligned} \tag{29}$$

Note that $s_i^{t+1,0} = s_i^{t,\tau_s}, \theta_i^{t+1,0} = \theta_i^{t,\tau_\theta}$. Therefore, sum

across $t = 0, \ldots, T-1$, we have

$$\begin{aligned}
&\mathbb{E}\Big[f_i(s_i^{T,0}, \theta_i^{T,0})\Big] \\
&\leq f_i(s_i^{0,0}, \theta_i^{0,0}) + \frac{3\eta^3 G^2}{T^{1/2}L} + \frac{\eta^2\sigma^2}{L\bar{\tau}} \\
&\quad - \frac{r\eta}{\sqrt{T}L}\sum_{i=0}^{T-1}\Big\|\nabla_s f_i(s_i^{t,0}, \theta_i^{t,0})\Big\|^2 \\
&\quad - \frac{2r\eta}{\sqrt{T}L}\sum_{i=0}^{T-1}\Big\|\nabla_\theta f_i(s_i^{t,0}, \theta_i^{t,0})\Big\|^2 \\
&\quad + \frac{(1-r)\lambda_i\eta G^2}{T^{1/2}L}\sum_{i=0}^{T-1}\beta^t(2t+1) \\
&\quad + \frac{(1-r)\lambda_i^2\eta^2 G^2}{TL^2}\sum_{i=0}^{T-1}(\beta^t)^2(2t+1)^2
\end{aligned} \tag{30}$$

The inequality can be re-written as follows.

$$\begin{aligned}
&\frac{1}{T}\sum_{i=0}^{T-1}\left[\Big\|\nabla_s f_i(s_i^{t,0}, \theta_i^{t,0})\Big\|^2 + 2\Big\|\nabla_\theta f_i(s_i^{t,0}, \theta_i^{t,0})\Big\|^2\right] \\
&\leq \frac{L}{2T^{1/2}r\eta}(f_i(s_i^{0,0}, \theta_i^{0,0}) - F^*) + \frac{3\eta^2 G^2}{2T} + \frac{\eta\sigma^2}{T^{1/2}\bar{\tau}} \\
&\quad + \frac{(1-r)\lambda_i G^2}{2Tr}\sum_{i=0}^{T-1}\beta^t(2t+1) \\
&\quad + \frac{(1-r)\lambda_i^2\eta^2 G^2}{2T^{3/2}rL}\sum_{i=0}^{T-1}(\beta^t)^2(2t+1)^2
\end{aligned} \tag{31}$$

Invoking Claims 4 and 5, we can obtain the final result. □

8.3. Proof for Claims

The analysis for each communication round is given in the following claims.

Claim 1 (Sufficient Decrease for Sampled Clients)**.** *Consider the setting of Theorem 1, we have the following inequality for all sampled client $i \in \mathcal{N}(t)$.*

$$\begin{aligned}
\mathbb{E}_t\big[f_i(s_i^{t,\tau_s}, \theta_i^{t,\tau_\theta})\big] \leq f_i(s_i^{t,0}, \theta_i^{t,0}) &+ \frac{3\eta^3 G^2}{T^{3/2}L} + \frac{\eta^2\sigma^2}{TL\bar{\tau}} \\
&- \frac{2\eta}{\sqrt{T}L}\Big\|\nabla_\theta f_i(s_i^{t,0}, \theta_i^{t,0})\Big\|^2 \\
&- \frac{\eta}{\sqrt{T}L}\Big\|\nabla_s f_i(s_i^{t,0}, \theta_i^{t,0})\Big\|^2,
\end{aligned} \tag{32}$$

where $\mathbb{E}_t = \mathbb{E}[\cdot \mid \mathcal{N}(t)]$ denotes the expectation conditioned on client sampling $\mathcal{N}(t)$ with respect to SGD steps, and $\bar{\tau} = 2/(1/\tau_s + 1/\tau_\theta)$ is the harmonic mean.

Proof. In communication round t, for the $\tau+1$-th ($0 \leq \tau \leq \tau_s - 1$) SGD step on parameter s of each sampled client $i \in \mathcal{N}(t)$, we use Assumption 2 to upper bound $f_i(s_i^{t,\tau+1}, \theta_i^{t,0})$ as follows.

$$\begin{aligned}
f_i(s_i^{t,\tau+1}, \theta_i^{t,0}) \leq f_i(s_i^{t,\tau}, \theta_i^{t,0}) &+ \underbrace{\Big\langle\nabla_s f_i(s_i^{t,\tau}, \theta_i^{t,0}), s_i^{t,\tau+1} - s_i^{t,\tau}\Big\rangle}_{\mathcal{T}_1} \\
&+ \frac{L}{2}\underbrace{\Big\|s_i^{t,\tau+1} - s_i^{t,\tau}\Big\|^2}_{\mathcal{T}_2},
\end{aligned} \tag{33}$$

where $s_i^{t,\tau+1}$ and $s_i^{t,\tau}$ is the parameter of supervisor after $\tau+1$-th and τ-th SGD step, respectively.

We first calculate the expectation of term $\mathcal{T}_1$

$$\begin{aligned}
\mathbb{E}_s^{t,\tau}[\mathcal{T}_1] &= \mathbb{E}_s^{t,\tau}\Big\langle\nabla_s f_i(s_i^{t,\tau}, \theta_i^{t,0}), -\eta_s \text{SGD}(f_i(s_i^{t,\tau}, \theta_i^{t,0}))\Big\rangle \\
&= -\eta_s\Big\langle\nabla_s f_i(s_i^{t,\tau}, \theta_i^{t,0}), \mathbb{E}_s^{t,\tau}\Big[\text{SGD}(f_i(s_i^{t,\tau}, \theta_i^{t,0}))\Big]\Big\rangle \\
&= -\eta_s\Big\|\nabla_s f_i(s_i^{t,\tau}, \theta_i^{t,0})\Big\|^2.
\end{aligned} \tag{34}$$

Here, $\mathbb{E}_s^{t,\tau} = \mathbb{E}\Big[\cdot \mid s_i^{t,\tau}, \theta_i^{t,0}\Big]$ means that the expectation is conditioned on $s_i^{t,\tau}$ and $\theta_i^{t,0}$.

Then we bound the expectation term $\mathcal{T}_2$,

$$\begin{aligned}
\mathbb{E}_s^{t,\tau}[\mathcal{T}_2] &= \eta_s^2\mathbb{E}_s^{t,\tau}\Big\|\text{SGD}(f_i(s_i^{t,\tau}, \theta_i^{t,0}))\Big\|^2 \\
&= \eta_s^2\mathbb{E}_s^{t,\tau}\Big\|\text{SGD}(f_i(s_i^{t,\tau}, \theta_i^{t,0})) - \nabla_s f_i(s_i^{t,\tau}, \theta_i^{t,0})\Big\|^2 \\
&\quad + \eta_s^2\Big\|\nabla_s f_i(s_i^{t,\tau}, \theta_i^{t,0})\Big\|^2 \\
&\leq \eta_s^2\sigma^2 + \eta_s^2\Big\|\nabla_s f_i(s_i^{t,\tau}, \theta_i^{t,0})\Big\|^2.
\end{aligned} \tag{35}$$

Note that $\Big\|\nabla_s f_i(s_i^{t,\tau}, \theta_i^{t,0})\Big\|^2$ can be lower bound by $\Big\|\nabla_s f_i(s_i^{t,0}, \theta_i^{t,0})\Big\|^2$ as follows.

$$\begin{aligned}
\Big\|\nabla_s f_i(s_i^{t,0}, \theta_i^{t,0})\Big\|^2 &= \|\nabla_s f_i(s_i^{t,0}, \theta_i^{t,0}) - \nabla_s f_i(s_i^{t,\tau}, \theta_i^{t,0}) \\
&\quad + \nabla_s f_i(s_i^{t,\tau}, \theta_i^{t,0})\| \\
&\leq \frac{1}{2}\Big\|\nabla_s f_i(s_i^{t,0}, \theta_i^{t,0}) - \nabla_s f_i(s_i^{t,\tau}, \theta_i^{t,0})\Big\|^2 \\
&\quad + \frac{1}{2}\Big\|\nabla_s f_i(s_i^{t,\tau}, \theta_i^{t,0})\Big\|^2 \\
&\leq \frac{1}{2}(L\tau_s\eta_s G)^2 + \frac{1}{2}\Big\|\nabla_s f_i(s_i^{t,\tau}, \theta_i^{t,0})\Big\|^2.
\end{aligned} \tag{36}$$

That is

$$\Big\|\nabla_s f_i(s_i^{t,\tau}, \theta_i^{t,0})\Big\|^2 \geq 2\Big\|\nabla_s f_i(s_i^{t,0}, \theta_i^{t,0})\Big\|^2 - L^2{\tau_s}^2\eta_s^2 G^2. \tag{37}$$

Combine the Eqs. (33) to (35) and (37), we have

$$\begin{aligned}\mathbb{E}_s^{t,\tau}\left[f_i(s_i^{t,\tau+1},\theta_i^{t,0})\right] &\le f_i(s_i^{t,\tau},\theta_i^{t,0}) + \frac{L}{2}\eta_s^2\sigma^2 \\ &\quad - (\eta_s - L\eta_s^2/2)\left\|\nabla_s f_i(s_i^{t,\tau},\theta_i^{t,0})\right\|^2 \\ &\le f_i(s_i^{t,\tau},\theta_i^{t,0}) + \frac{L}{2}\eta_s^2\sigma^2 \\ &\quad - 2(\eta_s - L\eta_s^2/2)\left\|\nabla_s f_i(s_i^{t,0},\theta_i^{t,0})\right\|^2 \\ &\quad + (\eta_s - L\eta_s^2/2)L^2{\tau_s}^2\eta_s^2G^2.\end{aligned} \tag{38}$$

Note that when learning rate is $\eta_s^t = \eta/\sqrt{T}L\tau_s$ and $\tau \le 1$,

$$\frac{\eta}{2\sqrt{T}L\tau_s} \le (\eta_s - L\eta_s^2/2) = \eta_s(1 - L\eta_s/2) \le \eta_s = \frac{\eta}{\sqrt{T}L\tau_s}. \tag{39}$$

Then $\mathbb{E}_s^{t,\tau}\left[f_i(s_i^{t,\tau+1},\theta_i^{t,0})\right]$ can be further bounded as follows.

$$\begin{aligned}\mathbb{E}_s^{t,0}\left[f_i(s_i^{t,\tau+1},\theta_i^{t,0})\right] &\le f_i(s_i^{t,\tau},\theta_i^{t,0}) + \frac{\eta^3G^2}{T^{3/2}L\tau_s} + \frac{\eta^2\sigma^2}{2TL{\tau_s}^2} \\ &\quad - \frac{\eta}{T^{1/2}L\tau_s}\left\|\nabla_s f_i(s_i^{t,0},\theta_i^{t,0})\right\|^2\end{aligned} \tag{40}$$

We sum Eq. (40) over $\tau = 0,\dots,\tau_s$, we have

$$\begin{aligned}\mathbb{E}_s^{t,0}\left[f_i(s_i^{t,\tau_s},\theta_i^{t,0})\right] &\le f_i(s_i^{t,0},\theta_i^{t,0}) + \frac{\eta^3G^2}{T^{3/2}L} + \frac{\eta^2\sigma^2}{2TL\tau_s} \\ &\quad - \frac{\eta}{\sqrt{T}L}\left\|\nabla_s f_i(s_i^{t,0},\theta_i^{t,0})\right\|^2.\end{aligned} \tag{41}$$

Similarly, for personalization parameter θ, we have

$$\begin{aligned}\mathbb{E}_\theta^{t,0}\left[f_i(s_i^{t,\tau_s},\theta_i^{t,\tau_\theta})\right] &\le f_i(s_i^{t,\tau_s},\theta_i^{t,0}) + \frac{\eta^3G^2}{T^{3/2}L} + \frac{\eta^2\sigma^2}{2TL\tau_\theta} \\ &\quad - \frac{\eta}{\sqrt{T}L}\left\|\nabla_\theta f_i(s_i^{t,\tau_s},\theta_i^{t,0})\right\|^2.\end{aligned} \tag{42}$$

where $\mathbb{E}_\theta^{t,\tau} = \mathbb{E}\left[\cdot \mid s_i^{t,\tau_s},\theta_i^{t,0}\right]$ denotes the expectation conditioned on s_i^{t,τ_s} and $\theta_i^{t,0}$.

Similarly we can bound $\left\|\nabla_\theta f_i(s_i^{t,\tau_s},\theta_i^{t,0})\right\|^2$ as follows.

$$\begin{aligned}\left\|\nabla_\theta f_i(s_i^{t,0},\theta_i^{t,0})\right\|^2 &= \|\nabla_\theta f_i(s_i^{t,0},\theta_i^{t,0}) - \nabla_\theta f_i(s_i^{t,\tau_s},\theta_i^{t,0}) \\ &\quad + \nabla_\theta f_i(s_i^{t,\tau_s},\theta_i^{t,0})\| \\ &\le \frac{1}{2}\left\|\nabla_\theta f_i(s_i^{t,0},\theta_i^{t,0}) - \nabla_\theta f_i(s_i^{t,\tau_s},\theta_i^{t,0})\right\|^2 \\ &\quad + \frac{1}{2}\left\|\nabla_\theta f_i(s_i^{t,\tau_s},\theta_i^{t,0})\right\|^2 \\ &\le \frac{1}{2}(L\tau_\theta\eta_\theta G)^2 + \frac{1}{2}\left\|\nabla_\theta f_i(s_i^{t,\tau_s},\theta_i^{t,0})\right\|^2.\end{aligned} \tag{43}$$

We plug Eq. (43) into Eq. (42), we can obtain

$$\begin{aligned}\mathbb{E}_\theta^{t,0}\left[f_i(s_i^{t,\tau_s},\theta_i^{t,\tau_\theta})\right] &\le f_i(s_i^{t,\tau_s},\theta_i^{t,0}) + \frac{2\eta^3G^2}{T^{3/2}L} + \frac{\eta^2\sigma^2}{2TL\tau_\theta} \\ &\quad - \frac{2\eta}{\sqrt{T}L}\left\|\nabla_\theta f_i(s_i^{t,0},\theta_i^{t,0})\right\|^2.\end{aligned} \tag{44}$$

Combining Eqs. (41) and (44), we complete the proof for this claim.

□

Claim 2 (Upper Bound for Non-Sampled Clients)**.** *In each communication round t, for any non-sampled client $i \notin \mathcal{N}(t)$, the following inequality holds*

$$\begin{aligned}f_i(s_i^{t+1},\theta_i^{t+1}) &\le f_i(s_i^t,\theta_i^t) + \frac{\lambda_i\beta^t(2t+1)\eta G^2}{T^{1/2}L} \\ &\quad + \frac{(\lambda_i\beta^t)^2(2t+1)^2\eta G^2}{TL^2}\end{aligned} \tag{45}$$

Proof. In communication round t, for each non-sampled client $i \notin \mathcal{N}(t)$, we use Assumption 2 and have

$$\begin{aligned}f_i(s_i^{t+1},\theta_i^{t+1}) &= f_i(s^t,\theta_i^{t+1}) \\ &\le f_i(s_i^t,\theta_i^t) + \underbrace{\left\langle\nabla_\theta f_i(s^t,\theta_i^t),\theta_i^{t+1} - \theta_i^t\right\rangle}_{\mathcal{T}_3} \\ &\quad + \frac{L_\theta}{2}\left\|\theta_i^{t+1} - \theta_i^t\right\|^2\end{aligned} \tag{46}$$

For notation simplicity, we rewrite **??** as follows.

$$\theta_i^{t+1} = \lambda_i\beta^t\theta_i^t + \sum_{j\in\mathcal{N}(t)} w_{ij}^t\theta_j^{t+1}. \tag{47}$$

We apply Claim 3 to upper bound $\left\|\theta_i^{t+1} - \theta_i^t\right\|$ as follows.

$$\begin{aligned}\left\|\theta_i^{t+1} - \theta_i^t\right\| &= \left\|\sum_{j\in\mathcal{N}(t)} w_{ij}^t(\theta_j^{t+1} - \theta_j^t) + \sum_{j\in\mathcal{N}(t)} w_{ij}^t(\theta_j^t - \theta_i^t)\right\| \\ &\le \sum_{j\in\mathcal{N}(t)} w_{ij}^t\left\|\theta_j^{t+1} - \theta_j^t\right\| + \sum_{j\in\mathcal{N}(t)} w_{ij}^t\left\|\theta_j^t - \theta_i^t\right\| \\ &\le \lambda_i\beta^t\cdot\tau_\theta\eta_\theta G + \lambda_i\beta^t\cdot 2t\tau_\theta\eta_\theta G \\ &= \lambda_i\beta^t(2t+1)\tau_\theta\eta_\theta G.\end{aligned} \tag{48}$$

We bound term $\mathcal{T}_3$ as follows.

$$\begin{aligned}\mathcal{T}_3 &\le \left\|\nabla_\theta f_i(s^t,\theta_i^t)\right\|\cdot\left\|\theta_i^{t+1} - \theta_i^t\right\| \\ &\le G\cdot\lambda_i\beta^t(2t+1)\tau_\theta\eta_\theta G.\end{aligned} \tag{49}$$

Combine the above results, we have

$$\begin{aligned}f_i(s_i^{t+1},\theta_i^{t+1}) &\le f_i(s_i^t,\theta_i^t) + \lambda_i\beta^t(2t+1)\tau_\theta\eta_\theta G^2 \\ &\quad + (\lambda_i\beta^t)^2(2t+1)^2{\tau_\theta}^2\eta_\theta^2G^2.\end{aligned} \tag{50}$$

Plugging $\eta_\theta = \eta/\sqrt{T}\tau_\theta L$ into Eq. (50) completes the proof.

□

The analysis of client drift is given in the next result.

Claim 3 (Client Drift)**.** *Consider the setting of Theorem 1, we have*

$$\left\|\theta_i^t - \theta_j^t\right\| \leq 2t\tau_s\eta_s G, \quad \forall i, j \in \{1, \ldots, n\}. \tag{51}$$

Proof. We first prove the following inequality.

$$\left\|\theta_i^t - \theta^0\right\| \leq t\tau_s\eta_s G, \quad \forall i, j \in \{1, \ldots, n\}, \tag{52}$$

where θ^0 is the initialization of inter-learning models q_θ.

We prove Eq. (52) by induction. When $t = 0$, Eq. (52) holds since all clients share the same initialization θ^0.

Assume Eq. (52) holds for communication round t, we show that it holds for communication round $t + 1$.

For any sampled client $i \in \mathcal{N}(t)$, according to Assumption 3, we have

$$\begin{aligned}\left\|\theta_i^{t+1} - \theta^0\right\| &\leq \left\|\theta_i^{t+1} - \theta_i^t\right\| + \left\|\theta_i^t - \theta^0\right\| \\ &\leq \tau_s\eta_s G + t\tau_s\eta_s G = (t+1)\tau_s\eta_s G\end{aligned} \tag{53}$$

For any non-sampled client $i \notin \mathcal{N}(t)$, We rewrite **??** for notation simplicity as follows.

$$\theta_i^{t+1} = \lambda_i\beta^t\theta_i^t + \sum_{j\in\mathcal{N}(t)} w_{ij}^t\theta_j^{t+1}. \tag{54}$$

Then we have

$$\begin{aligned}\left\|\theta_i^{t+1} - \theta^0\right\| &= \left\|\lambda_i\beta^t(\theta_i^t - \theta^0) + \sum_{j\in\mathcal{N}(t)} w_{ij}^t(\theta_j^{t+1} - \theta^0)\right\| \\ &\leq \lambda_i\beta^t\left\|\theta_i^t - \theta^0\right\| + \sum_{j\in\mathcal{N}(t)} w_{ij}^t\left\|\theta_j^{t+1} - \theta^0\right\| \\ &\leq \lambda_i\beta^t t\tau_s\eta_s G + \sum_{j\in\mathcal{N}(t)} w_{ij}^t(t+1)\tau_s\eta_s G \\ &\leq (t+1)\tau_s\eta_s G.\end{aligned} \tag{55}$$

This completes the proof for Eq. (52).

According to Eq. (52), we have

$$\left\|\theta_i^t - \theta_j^t\right\| \leq \left\|\theta_i^t - \theta^0\right\| + \left\|\theta^0 - \theta_j^t\right\| \leq 2t\tau_s\eta_s G, \tag{56}$$

which completes the proof of this claim.

□

Note that when inter-learning models are initialized differently, it only adds a constant to the RHS of Eq. (51), which is an order smaller than $2t\tau_s\eta_s G$ and will not affect the convergence.

Claim 4. *Consider the setting of Theorem 1, we have the following upper bound,*

$$\begin{aligned}\sum_{i=0}^{T-1}\beta^t(2t+1) \leq\ & -C^2(1-2\gamma)T^{2\gamma}\ln T && (57)\\ & + C^2(1+2\ln C)T^{2\gamma} && (58)\\ & + 3CT^\gamma + 1 - C^2T^{2\gamma-1}. && (59)\end{aligned}$$

Proof. Recall that

$$\beta^t = \begin{cases}1, & t < CT^\gamma \\ \left(\frac{CT^\gamma}{t}\right)^2, & t \geq CT^\gamma,\end{cases} \tag{60}$$

Here, $C \geq 0$, $0 < \gamma < 1/2$.

Therefore, $\beta^t(2t+1)$ is increasing in $[0, CT^\gamma)$ and decreasing in $[CT^\gamma, T)$. Thus we have the following upper bound.

$$\begin{aligned}&\sum_{i=0}^{T-1}\beta^t(2t+1) \\ =\ & \sum_{i=0}^{CT^\gamma-1}\beta^t(2t+1) + (1+2CT^\gamma) + \sum_{i=CT^\gamma+1}^{T-1}\beta^t(2t+1) \\ \leq\ & \int_0^{CT^\gamma}(2t+1)\,\mathrm{d}t + (1+2CT^\gamma) + \int_{CT^\gamma}^{T}\frac{C^2T^{2\gamma}}{t^2}\cdot(2t+1)\,\mathrm{d}t \\ =\ & (t^2+t)\,|_0^{CT^\gamma} + (1+2CT^\gamma) + C^2T^{2\gamma}(-2\ln t - \frac{1}{t})\,|_{CT^\gamma}^{T} \\ =\ & -C^2(1-2\gamma)T^{2\gamma}\ln T + C^2(1+2\ln C)T^{2\gamma} \\ & + 3CT^\gamma + 1 - C^2T^{2\gamma-1}.\end{aligned} \tag{61}$$

□

Claim 5. *Consider the setting of Theorem 1, we have the following upper bound.*

$$\begin{aligned}\sum_{i=0}^{T-1}(\beta^t)^2(2t+1)^2 \leq\ & 3CT^{3\gamma} + 9C^2T^{2\gamma} && (62)\\ & + 3CT^\gamma - 3C^4T^{4\gamma-3}. && (63)\end{aligned}$$

Proof. Recall that

$$\beta^t = \begin{cases}1, & t < CT^\gamma \\ \left(\frac{CT^\gamma}{t}\right)^2, & t \geq CT^\gamma,\end{cases} \tag{64}$$

Here, $C \geq 0$, $0 < \gamma < 1/2$.

Therefore, $\beta^t(2t+1)$ is increasing in $[0, CT^\gamma)$ and decreasing in $[CT^\gamma, T)$. Thus we have the following upper bound.

$$\begin{aligned}&\sum_{i=0}^{T-1}(\beta^t)^2(2t+1)^2 && (65)\\ \leq\ & \sum_{i=0}^{T-1}(\beta^t)^2(3t)^2 && (66)\\ =\ & 9\sum_{i=0}^{CT^\gamma-1}(\beta^t)^2t^2 + 9C^2T^{2\gamma} + 9\sum_{i=CT^\gamma+1}^{T-1}(\beta^t)^2t^2 && (67)\\ \leq\ & 9\int_0^{CT^\gamma}t^2\,\mathrm{d}t + 9C^2T^{2\gamma} + 9\int_{CT^\gamma}^{T}\frac{C^4T^{4\gamma}}{t^4}\cdot t^2\,\mathrm{d}t && (68)\\ =\ & 3t^3\,|_0^{CT^\gamma} + 9C^2T^{2\gamma} - 3C^4T^{4\gamma}t^{-3}\,|_{CT^\gamma}^{T} && (69)\\ =\ & 3CT^{3\gamma} + 9C^2T^{2\gamma} + 3CT^\gamma - 3C^4T^{4\gamma-3}. && (70)\end{aligned}$$

□

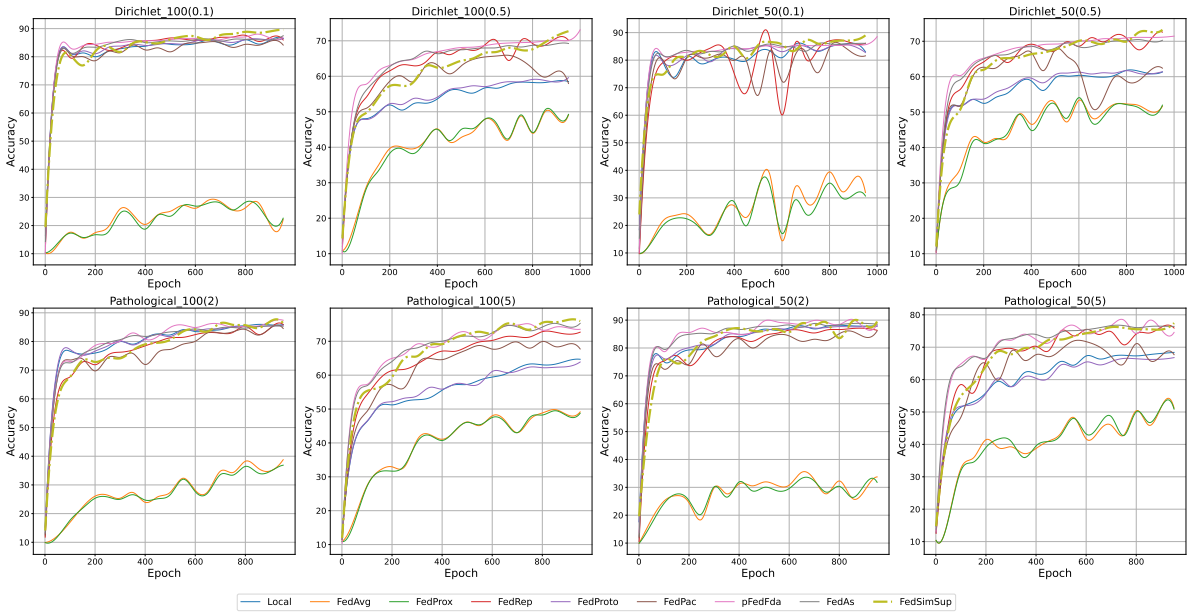


Figure 7. The accuracy of different method on CIFAR10

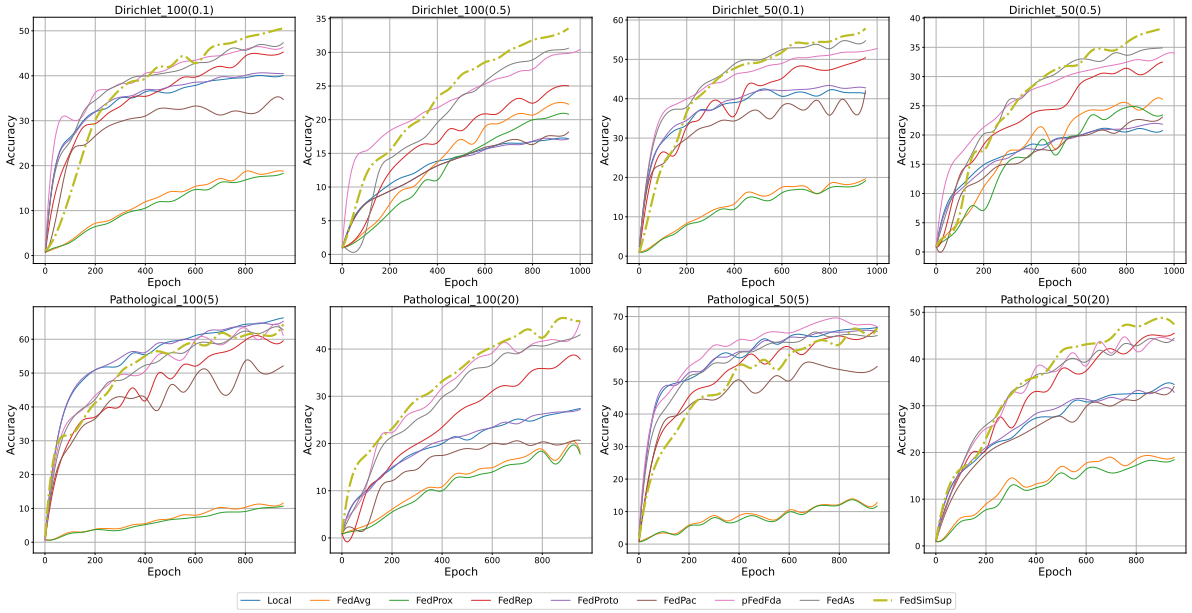


Figure 8. The accuracy of different method on CIFAR100

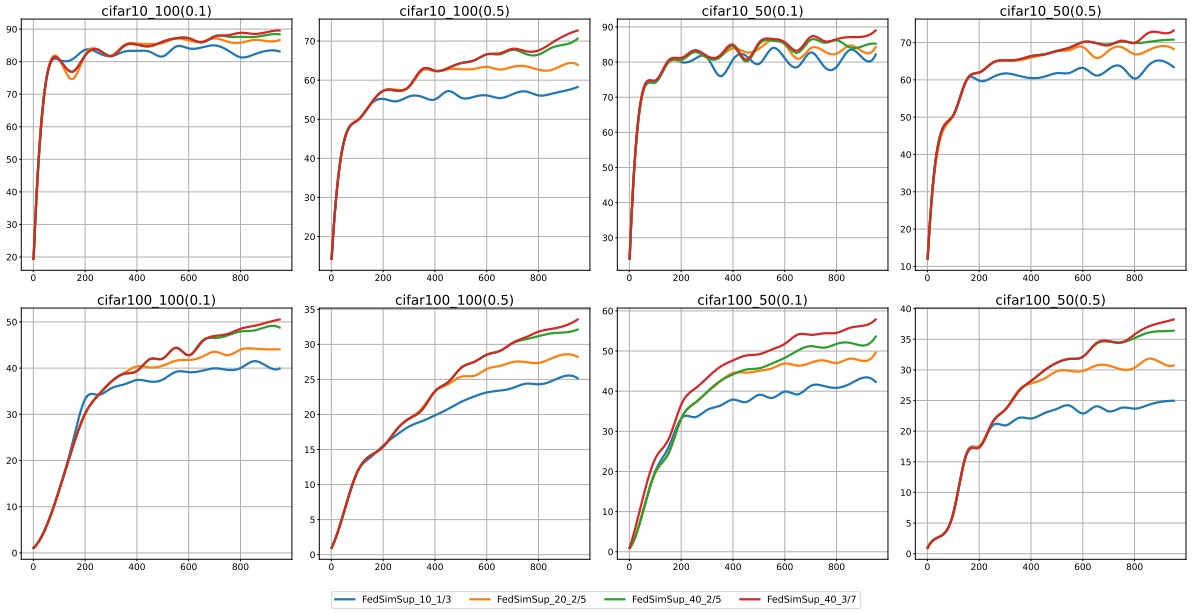


Figure 9. The performance of different parameters C and γ