SPFL: Sequential Updates with Parallel Aggregation for Enhanced Federated Learning Under Category and Domain Shifts

Haoyuan Liang[1*], **Shilei Cao**[1*] , **Guowen Li**[1] , **Zhiyu Ye**[3] , **Haohuan Fu**[2,3] , **Juepeng Zheng**[1,2†]
[1] School of Artificial Intelligence, Sun Yat-sen University, China
[2] National Supercomputing Center in Shenzhen, China
[3] Tsinghua Shenzhen International Graduate School, China
{lianghy68 , caoshlei , ligw8}@mail2.sysu.edu.cn
zhengjp8@mail.sysu.edu.cn, yezy25@mails.tsinghua.edu.cn

Abstract

Federated Learning (FL) has recently emerged as the primary approach to overcoming data silos, enabling collaborative model training without sharing sensitive or proprietary data. Parallel Federated Learning (PFL) aggregates models trained independently on each client's local data, which could prevent the model from converging to the optimal solution due to limited data exposure. In contrast, Sequential Federated Learning (SFL) allows models to traverse client datasets sequentially, enhancing data utilization. However, SFL effectiveness is limited in real-world Non-IID scenarios characterized by category shift (inconsistent class distributions) and domain shift (distribution discrepancies). These shifts cause two critical issues: *update order sensitivity*, where model performance varies significantly with the sequence of client updates; and *catastrophic forgetting*, where the model forgets previously learned features when trained on new client data. Therefore, based on SFL, we propose a novel updating framework, **SPFL** (**S**equential updates with **P**arallel aggregation **F**ederated **L**earning), that can be integrated into existing PFL methods. It integrates sequential updates with parallel aggregation to enhance data utilization and ease update order sensitivity. Meanwhile, we give the convergence analysis of SPFL under strong convex, general convex, and non-convex conditions, proving that this update scheme is significantly better than PFL and SFL. Additionally, we introduce the GLAM (**G**lobal-**L**ocal **A**lignment **M**odule) to mitigate catastrophic forgetting by aligning the predictions of the local model with those of previous models and the global model during training. Our extensive experiments demonstrate that integrating the SPFL framework into existing PFL methods significantly improves performance under category and domain shifts.

1 Introduction

Deep learning models have demonstrated immense potential across various vision tasks, typically depending on large-scale data training [1, 2, 3]. As privacy concerns regarding the centralized collection of extensive data continue to escalate, traditional centralized training methods increasingly fail to address clients' privacy needs. In response, Federated Learning (FL) [4, 5, 6] has emerged as a decentralized solution, designed specifically to prioritize data privacy by allowing models to be trained collaboratively without centralizing sensitive information. Aligning with the foundational

[*] Equal contributions.
[†] Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

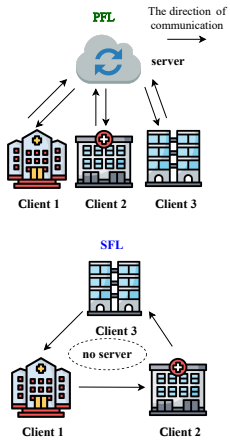


Figure 1: **Top**: Parallel Federated Learning (**PFL**): The server aggregates the local models trained independently on each client. **Bottom**: Sequential Federated Learning (**SFL**): The local model will be trained on the clients sequentially without a centralized sever.

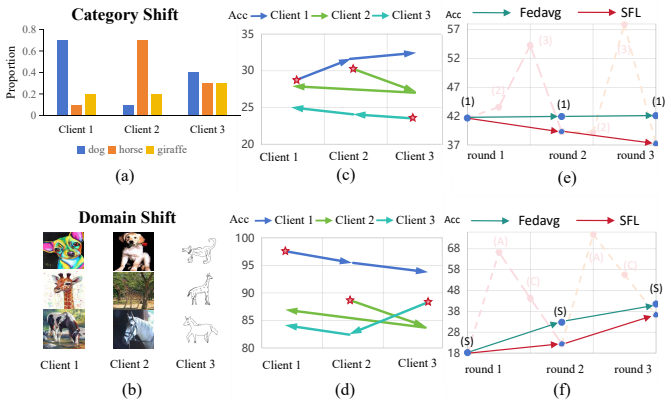


Figure 2: (a) and (b) present cases of category shift and domain shift. (c) and (d) illustrate the performance of three starting points for client training (*i.e.*, *Client 1*→*Client 2*→*Client 3*, *Client 2*→*Client 3*→*Client 1*, and *Client 3*→*Client 1*→*Client 2*) on the three clients' data under category and domain shift, respectively. The solid lines in (e) and (f) compare the performance of *Client 1*'s data between FedAvg, which repeatedly learns solely on *Client 1*'s data, and SFL, which iteratively learns sequentially across different client datasets under category and domain shift, respectively. The dashed lines represent the sequential training process of SFL.

method FedAvg [7], existing FL methods [8, 9] commonly employ Parallel Federated Learning (PFL), which aggregates models trained independently on each client at a central server, as shown in Fig. 1 (top). While PFL effectively enhances the global model's performance through iterative aggregation, it does not fully exploit all available data since each client's model is trained solely on its local data, potentially leading to poor model generalization. This raises the question: **what strategies can be explored to leverage client data better and improve the model generalization?**

One promising strategy is to allow the model to traverse each client's dataset sequentially, as demonstrated in Fig. 1 (bottom), leading to Sequential Federated Learning (SFL) [10, 11, 12]. Although current SFL methods perform effectively in **I**ndependently and **I**dentically **D**istributed (IID) settings across clients to enhance generalization and leverage more data, their performance remains constrained in real-world Non-IID scenarios. The challenges arise because the Non-IID nature across clients mainly manifests in two forms in image tasks: **category shift** and **domain shift**. The category shift refers to inconsistent class distributions across clients. For instance, dog samples constitute 70% of the *Client 1*'s data but only 10% of the *Client 2*'s, as shown in Fig. 2 (a). The domain shift involves distribution discrepancies arising from different domains, such as images sampled from paintings in one client and real photos in another, as depicted in Fig. 2 (b).

To analyze the impact of two types of shifts in SFL, we test SFL on two datasets characterized by the category shift [13, 14] (*i.e.* CIFAR-10) and the domain shift [15, 16] (*i.e.* PACS) following previous work, respectively. The results presented in Fig. 2 reveal two critical issues impacting SFL under Non-IID conditions: (1) **Update Order Sensitivity**: The model becomes highly sensitive to the order in which clients are updated. Different update sequences can lead to inconsistent results in the same client dataset, as illustrated in Fig. 2 (c) and (d). For example, Fig. 2 (c) demonstrates that starting the training process from *Client 2* (*Client 2*→*Client 3*→*Client 1*) leads to better performance in the *Client 1* dataset than starting from *Client 3* (*Client 3*→*Client 1*→*Client 2*) under category shift. This may be because the latter update order causes the model to become trapped in a saddle point, resulting in consistently poor performance. In addition, we present two examples to theoretically demonstrate the existence of this phenomenon in Appendix Section H. However, solving this by exhaustively exploring all possible update sequences to determine the optimal order is impractical due to the high computational and time costs, especially in real-world applications involving a large

number of clients. (2) **Catastrophic Forgetting**: As the model updates sequentially across clients with different data distributions, it tends to forget previously learned features when trained on new client data, a phenomenon known as catastrophic forgetting in continual learning [17, 15, 18]. This issue is evident in Fig. 2 (e) and (f), where SFL performs worse in *Client 1*'s data after repeated iterations under category or domain shift, compared to SFL, which learns repeatedly on a single client's dataset.

To effectively leverage the available client data and overcome these challenges under category shift and domain shift, we propose a novel FL framework named Sequential updates with Parallel aggregation Federated Learning (SPFL) and a Global-Local Alignment Module (GLAM).

Specifically, SPFL harnesses the strengths of both parallel and sequential learning. The server initially distributes the global model to all clients. Each client trains the model on its local data before passing the updated model to the next client for further training. After completing the sequential update round starting from different clients, the server aggregates these models in parallel to form a new global model for the next iteration. **This hybrid approach maximizes data utilization and mitigates the risk of suboptimal update sequences, improving model robustness within a single computation round.** Moreover, SPFL eliminates the need to share dataset sizes for weighting during aggregation, which increases privacy in applications where dataset sizes may leak sensitive information. Additionally, SPFL can be integrated into existing FL optimization approaches as a new updating strategy to address category shift, such as Scaffold [19] and FedDyn [20]. Furthermore, our proposed GLAM tackles catastrophic forgetting by continuously aligning the predictions of the global model with the local model during training. It also aligns the predictions of the local model with those from the previous client, effectively preserving learned knowledge among clients. The GLAM demonstrates significant effectiveness in improving the performance of SPFL under domain shift. **We also provide a convergence analysis of SPFL under strong convex, general convex, and non-convex conditions, demonstrating that this update scheme outperforms both PFL and SFL in Appendix Section G.** In summary, the contributions of our paper are summarized as follows:

• We identify and analyze the limitations of existing SFL methods in Non-IID scenarios, highlighting the issues of update order sensitivity and catastrophic forgetting under category and domain shifts.

• We propose SPFL, a novel federated learning framework that combines sequential updates with parallel aggregation. SPFL improves data utilization and model robustness by aggregating models with different starting clients to traverse each client's dataset sequentially. We also design GLAM to align predictions of the local model, both with global model and previous local model to mitigate catastrophic forgetting, ensuring that the model retains knowledge across different client updates.

• We provide a convergence analysis under strong convex, general convex, and non-convex conditions. Extensive experiments demonstrate the effectiveness of SPFL under Non-IID conditions (domain shift and category shift), as well as their compatibility with traditional PFL methods.

2 Related Work

2.1 Parallel Federated Learning

As privacy concerns grow, companies become increasingly hesitant to upload their data to the cloud for centralized model training. To effectively utilize these distributed data, FedAvg [7] pioneer federated learning community that serves as the foundation of Parallel Federated Learning (PFL) [21]. Building upon this work, most federated optimization algorithms have adopted the PFL paradigm, improving FedAvg through various improvements [22, 23, 19, 24, 25]. For example, SCAFFOLD [19] mitigates client drift in federated learning through control variates. FedSAM [25] addresses the decline in global model generalization caused by inconsistent data distributions across clients by employing the Sharpness Aware Minimization optimizer. Although these FL methods significantly enhance the handling of traditional Non-IID issues, they are less effective on complex image tasks involving Non-IID data, such as domain shift [26] and category shift [13]. Therefore, to address domain shift, FedCSA [27] addresses the inconsistency in the global model due to Non-IID data, using a model bias-based client data clustering method. Furthermore, to address the category shift, FedDisco [28] tackles the poor convergence of the global model under the category shift. However, existing PFL methods fall short in fully exploiting the available client data, since each local model is

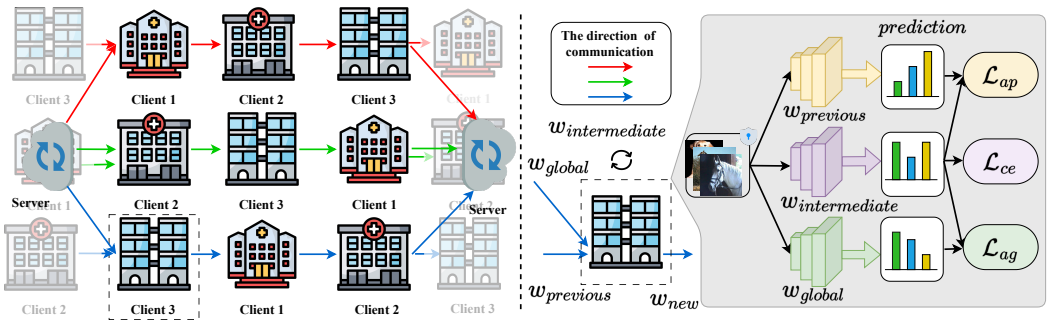


Figure 3: (**Left**) The process of our update method Sequential Updates with Parallel Aggregation (SPFL). The serial numbers indicate the steps to be executed sequentially. (**Right**) The module that solves the forgetting problem when the domain differences of each client are too large, that is, the module that Global-Local Alignment Module (GLAM). Lines of different colors represent the update order of different starts, and the black represents aggregation to the central server.

limited access to its corresponding client data. This limitation may result in suboptimal convergence of model training due to insufficient data exposure, which motivates our proposed SPFL.

2.2 Sequential Federated Learning

Nowadays, studies on Sequential Federated Learning (SFL) remain relatively limited. Some research addresses the issue of information leakage in federated learning by performing membership inference attacks in the SFL setting [29]. Although other research uses sequential updates for privacy protection [29], they do not address the issue of poor model performance on Non-IID data. FedGSP [30] seeks to address the decline in model performance caused by Non-IID data in federated learning through a dynamic collaborative training approach that transitions from sequential to grouped parallel updates. However, this approach deviates from the standard SFL setting. Similarly, FedSeq [10] mitigates the issues of slow convergence and performance degradation caused by computational heterogeneity in federated learning by sequentially training on heterogeneous client groups, referred to as super-clients. Although some experimental results demonstrate the advantages of FedSeq on Non-IID data, and some scholars theoretically prove the convergence of SFL under Non-IID conditions [31], these models still perform poorly on complex tasks involving domain shift and category shift.

3 Method

In this section, we first introduce the problem formulation. Then, we provide the convergence analysis of SPFL, comparing it with PFL and SFL. Next, to tackle the challenges associated with domain shift for SPFL, we propose a Global-Local Alignment Module (GLAM). The workflow of SPFL is illustrated in Fig. 3. Limited by space, notations are summarized in Appendix D. Meanwhile, the general assumptions for convergence are provided in Appendix E.

3.1 Preliminaries

In our paper, we adopt a similar setting to the traditional FL and assume that each client m has its dataset $D_k = \{(x_i^m, y_i^m)\}_{i=1}^{q_m}$ where represents the amount of data from client kand performs model updates in this way where $x \in \mathcal{X}$ is the input and $y \in \mathcal{Y}$ its corresponding label. For each client k, we have separate models $f(\cdot; \boldsymbol{w}_k)$. The optimization objective of PFL can be written as (1):

$$\min_{w} \sum_{m=1}^{M} \sum_{k=1}^{K} \frac{q_m}{\sum_{j'=1}^{M} q_{j'}} \sum_{i=1}^{q_m} \mathcal{L}_m \left(f(x_i^m; \boldsymbol{w}_m), y_i^m \right) \tag{1}$$

Unlike PFL, the server is no longer needed in SFL, and our optimization objective is written as (2):

$$\min_{w} \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{i=1}^{q_m} \mathcal{L}_m \left(f(x_i^m; \boldsymbol{w}_m), y_i^m \right) \tag{2}$$

By comparing (1) and (2), we observe that the gradient of $\mathcal{L}_m$ is solely dependent on the dataset of client m, whereas the gradient of $\mathcal{L}_m$ in (2) is influenced by all the datasets. Meanwhile, we can see that SFL does not need to introduce the dataset size in the optimization. However, as we analyzed in Fig. 2, inconsistent starting points have a great impact on model performance. To solve this problem, we proposed Sequential updates with Parallel aggregation (SPFL), which can avoid the worst solution of the model, but of course, this does not represent the best model performance. To distinguish the different models at each starting point, we add conditional probability to the optimization target to fully consider this issue. Meanwhile, because SPFL goes through each client, the dataset that undergoes gradient descent is the same size, and we no longer need to consider the impact of the dataset size on the model, which could be proved in Fig. 4c.

3.2 Sequential updates with Parallel aggregation

Our paper proposes a new framework of updating Sequential updates with Parallel aggregation (SPFL), similar to the client setting in PFL. Unlike PFL, our updating framework can continuously learn from *different client datasets*. From the perspective of global optimization, we are different from PFL because the aggregation process is different from the SFL aggregation process, as shown:

$$\min_{w} \frac{1}{M} \sum_{n=1}^{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{i=1}^{q_{\pi_m^n,k}} \nabla \mathcal{L}_{\pi_m^n,k} \left(f(x_i^{\pi_m^n,k}; \boldsymbol{w}_{\pi_m^n,k}), y_i^{\pi_m^n,k} | \sum_{j=1}^{m-1} \nabla \mathcal{L}_{\pi_j^n,k} \right) \tag{3}$$

However, to compare the subsequent convergence analysis, according to Assumption 1, this deviation can be incorporated into the σ in Assumption 3. For client m, the update method is simplified as (4). See (45) for more details.

$$\boldsymbol{w}^{(r+1)} = \boldsymbol{w}^{(r)} - \eta \frac{1}{M} \sum_{n=1}^{M} \sum_{m=1}^{M} \sum_{k=0}^{K-1} \mathbf{g}_{\pi_m^n,k}^{(r)} \tag{4}$$

To help readers better understand SPFL, we provide the pseudo-code in Appendix Algorithm 1.

3.3 Convergence analysis of SPFL

Theorem 1. *For SPFL, there exists a constant effective learning rate $\tilde{\eta} := MK\eta$ and weights θ_r, such that the weighted average of the global parameters $\bar{\boldsymbol{w}}^{(R)} = \frac{1}{W_R} \sum_{r=0}^{R} \theta_r \boldsymbol{w}^{(r)}$ (where $W_R = \sum_{r=0}^{R} \theta_r$) satisfies the following upper bounds:*

Strongly convex: *Under Assumptions 2, 3, and 5, there exists a constant effective learning rate $\frac{1}{\mu R} \leq \tilde{\eta} \leq \frac{1}{6L}$ and weights $\theta_r = \left(1 - \frac{\mu\tilde{\eta}}{2}\right)^{-(r+1)}$, such that the following holds:*

$$\mathbb{E}\left[\mathcal{L}(\bar{\boldsymbol{w}}^{(R)}) - \mathcal{L}(\boldsymbol{w}^*)\right] \leq \frac{9}{2}\mu\mathcal{A}^2 \exp\left(-\frac{1}{2}\mu\tilde{\eta}R\right) + \frac{12\tilde{\eta}\sigma^2}{M^2K} + \frac{18L\tilde{\eta}^2\sigma^2}{MK} + \frac{18L\tilde{\eta}^2\zeta_*^2}{MK} \tag{5}$$

General convex: *Under Assumptions 2, 3, and 5, there exists a constant effective learning rate $\tilde{\eta} \leq \frac{1}{6L}$ and weights $\theta_r = 1$, such that the following holds:*

$$\mathbb{E}\left[\mathcal{L}(\bar{\boldsymbol{w}}^{(R)}) - \mathcal{L}(\boldsymbol{w}^*)\right] \leq \frac{3\mathcal{A}^2}{\tilde{\eta}R} + \frac{12\tilde{\eta}\sigma^2}{M^2K} + \frac{18L\tilde{\eta}^2\sigma^2}{MK} + \frac{18L\tilde{\eta}^2\zeta_*^2}{MK} \tag{6}$$

Non-convex: *Under Assumptions 2, 3, and 4, there exists a constant effective learning rate $\tilde{\eta} \leq \frac{1}{6L(\beta+1)}$ and weights $\theta_r = 1$, such that the following holds:*

$$\min_{0 \leq r \leq R} \mathbb{E}\left[\left\|\nabla \mathcal{L}(\boldsymbol{w}^{(r)})\right\|^2\right] \leq \frac{3\mathcal{B}}{\tilde{\eta}R} + \frac{3L\tilde{\eta}\sigma^2}{M^2K} + \frac{27L^2\tilde{\eta}^2\sigma^2}{8MK} + \frac{27L^2\tilde{\eta}^2\zeta^2}{8M} \tag{7}$$

$\mathcal{A} := \|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\|$ for the convex cases and $\mathcal{B} := \mathcal{L}(\boldsymbol{w}^{(0)}) - \mathcal{L}^*$ for the non-convex case.

We include Corollary 1 in the appendix. Additionally, we compare SPFL, SFL, and PFL. As shown in Tab. 1, SPFL achieves a smaller convergence upper bound than both SFL and PFL, indicating a faster convergence rate. Compared to PFL and SFL, our proposed SPFL achieves a tighter upper

Table 1: The upper bounds in the strongly convex case, with absolute constants and polylogarithmic factors omitted, are presented. The general convex and non-convex cases are included in Corollary 1.

Method	**Bound** ($\mathcal{A} = \|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\|$)
PFL (FedAvg) [31]	$\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 KR^2} + \frac{L\zeta_*^2}{\mu^2 R^2} + \mu\mathcal{A}^2 \exp\left(-\frac{\mu R}{L}\right)$
PFL (Scaffold) [19]	$\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 KR^2} + \frac{L\zeta_*^2}{\mu^2 R^2} + \mu\mathcal{A}^2 \exp\left(-\frac{\mu R}{L}\right)$
SFL [31]	$\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 MKR^2} + \frac{L\zeta_*^2}{\mu^2 MR^2} + \mu\mathcal{A}^2 \exp\left(-\frac{\mu R}{L}\right)$
SPFL (ours)	$\frac{\sigma^2}{\mu M^2 KR} + \frac{L\sigma^2}{\mu^2 MKR^2} + \frac{L\zeta_*^2}{\mu^2 MR^2} + \mu\mathcal{A}^2 \exp\left(-\frac{\mu R}{L}\right)$

bound in the strongly convex case, notably improving the variance term from $\tilde{\mathcal{O}}(1/(MKR))$ to $\tilde{\mathcal{O}}(1/(M^2KR))$. Therefore, increasing the number of *Clients* can enhance the convergence speed of the model compared to other methods.

As shown in Fig. 3, we set an update order for the startup of each client and perform aggregation after each step, enabling more effective integration with traditional personalized federated learning (PFL). This sequential federated learning (SFL) strategy allows the model to traverse a larger volume of data, leading to a 3% improvement in handling category shift. However, SFL alone remains insufficient for addressing domain shift due to the inherently uneven data distributions across clients. While SPFL inherits the PFL-style weighted aggregation, which helps maintain competitive performance under domain shift, it still faces limitations. To mitigate these issues, we introduce a global-local alignment module (GLAM) specifically designed to enhance robustness against domain discrepancies.

3.4 Global-Local Alignment Module (GLAM)

In SFL with client data, if the model does not learn enough in the current domain and quickly transitions to the next domain, it may cause the model to forget the features learned in the previous round when moving to the next. For example, in the PACS task, when the model transitions from learning comics to the photo domain, we aim to offset the style features during the learning process. However, it seems that some content information is also lost, making the SFL method significantly inferior to the PFL method in domain transfer.Parameter are summarized in Appendix D.

In real-world federated learning, the data distribution of each client is unknown, so the model needs to be robust enough to perform well across a variety of potential scenarios. Traditional solutions to the forgetting problem typically store part of the source domain data, which is not permissible in federated learning. **In SPFL, each client maintains not only its own trained model but also a global model and a model passed from the previous client—an advantage that neither SFL nor PFL possesses.** To save resources as much as possible and fully leverage the advantages of SPFL, $f(\boldsymbol{w}_g;\cdot)$ and $f(\boldsymbol{w}_p;\cdot)$ are used solely for inference prediction, while they assist model $(\boldsymbol{w}_i;\cdot)$ in performing gradient descent. We fully leverage this advantage and develop a global-local alignment module (GLAM) based on these three models. To prevent the local model from excessively forgetting previous information, we designed two additional *loss* terms.

To ensure the convergence of the training process, we first use the cross-entropy loss function as the basis for gradient descent. The specific loss function used during the training is as follows:

$$\mathcal{L}_{ce} = -\frac{1}{B}\sum_{i=1}^{B} \boldsymbol{y}_i{}^T \log(f(\boldsymbol{w}_i; \boldsymbol{x}_i^{\pi_m^n,k})) \tag{8}$$

To mitigate catastrophic forgetting during training, we incorporate the predictions of the global model with those of the model under training by using a mutual cross-entropy loss, as shown in (9). This loss $\mathcal{L}_{ag}$ is named Approach to the global model, ensures consistency at the output level, and facilitates alignment between local and global models.

$$\mathcal{L}_{ag} = -\frac{1}{B}\sum_{i=1}^{B} f(\boldsymbol{w}_g; x_i^{\pi_m^n,k})^T \log(f(\boldsymbol{w}_i; x_i^{\pi_m^n,k})) \tag{9}$$

For a client, we can also use the model passed from the previous client to align, thereby reducing forgetting of the previous client. Unlike before, we use KL divergence to make the distributions of

the two outputs consistent. We refer to this loss as the Local Alignment with the Previous Output $\mathcal{L}_{ap}$, which is defined as follows:

$$\mathcal{L}_{ap} = -\frac{1}{2B}\sum_{i=1}^{B} KL(P||Q) + KL(Q||P), s.t.\ P = f(\boldsymbol{w}_i; x_i^{\pi_m^n,k}), Q = f(\boldsymbol{w}_p; x_i^{\pi_m^n,k}) \quad (10)$$

Table 2: Comparison of Classic and State-of-the-Art Algorithms in Federated Learning with and without SPFL on Cifar-10, Cifar-100, CINIC-10 and Fmnist in Category Shift

Method	Cifar-10 resnet18 (num=10)	Cifar-100 resnet18 (num=10)	CINIC-10 simple-cnn (num=10)	Fmnist resnet18 (num=10)	Avg
FedAvg [7]	75.00	57.55	39.74	81.13	63.35
+ SPFL	78.88 (**+3.88**)	64.33 (**+6.78**)	41.01 (**+1.27**)	84.75(**+3.62**)	67.24 (**+3.89**)
FedDc [32]	80.52	64.44	40.92	83.44	67.33
+ SPFL	85.90(**+5.38**)	69.88 (**+5.44**)	44.14 (**+3.23**)	87.67(**+4.23**)	71.90 (**+4.57**)
FedDyn [20]	77.91	64.11	40.91	81.35	66.03
+ SPFL	80.90(**+2.99**)	64.24 (**+0.13**)	43.06 (**+2.15**)	83.23(**+1.88**)	67.86 (**+1.83**)
FedNova [33]	76.02	57.64	39.82	81.38	63.46
+ SPFL	79.31(**+3.29**)	64.47 (**+6.83**)	41.32 (**+1.50**)	84.85(**+3.47**)	67.49(**+4.03**)
FedProx [34]	76.04	57.56	39.65	81.23	63.62
+ SPFL	77.25(**+1.21**)	60.46 (**+2.90**)	40.27 (**+0.62**)	83.89(**+2.66**)	65.47 (**+1.85**)
MOON [35]	78.70	58.44	40.11	81.29	63.68
+ SPFL	79.98 (**+1.28**)	63.54 (**+5.20**)	41.11 (**+1.00**)	84.83(**+3.54**)	67.37(**+3.69**)
SCAFFOLD [19]	76.38	56.15	36.00	80.71	62.31
+ SPFL	78.73 (**+2.35**)	65.05 (**+8.90**)	39.52 (**+3.52**)	85.80(**+5.09**)	67.26 (**+4.95**)
FedDisco [28]	76.32	57.50	39.67	81.24	63.68
+ SPFL	78.58(**+2.26**)	64.44 (**+6.94**)	40.79 (**+1.12**)	84.27(**+3.02**)	67.13(**+3.45**)

Table 3: Comparison of Classic Algorithms in FL with and without SPFL on Cifar-10 in IID

Method	Fedavg [7]	Feddc [32]	FedDyn [20]	FedNova [33]	FedProx [34]	MOON [35]	SCAFFOLD [19]	FedDisco [28]
PFL	83.81	81.75	80.66	83.39	83.94	84.03	82.62	84.03
SPFL	89.85 (**+6.54**)	93.08 (**+11.33**)	81.77 (**+1.11**)	90.00 (**+6.61**)	86.12(**+2.18**)	90.16 (**+6.13**)	91.97 (**+9.53**)	89.92 (**+5.89**)

This approach helps to maintain a consistent output distribution across clients, thus facilitating a smoother transition and reducing catastrophic forgetting. Combining (9) , (10) , and (8), we can derive the final loss function $\mathcal{L}$ for our model, as shown below:

$$\mathcal{L} = \tau\mathcal{L}_{ap} + \rho\mathcal{L}_{ag} + \mathcal{L}_{ce} \quad (11)$$

where τ and ρ are hyperparameters used to control the influence of the previous client's model and the global model on the current gradient descent, which affects the convergence of the model.

In this paper, we enhance SFL to develop SPFL and introduce the GLAM module, making SPFL a more effective update method than PFL. SPFL achieves significant improvements under category shift and delivers domain shift performance comparable to that of PFL. These insights are empirically validated through extensive experiments.

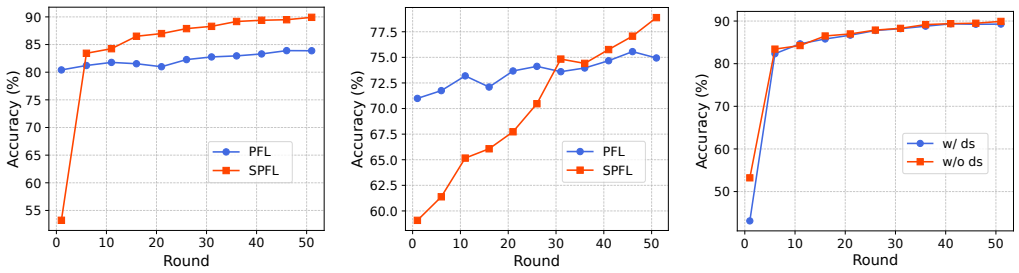
4 Experiments

4.1 Set up

In this section, we introduce the setup. Unless otherwise specified, all experiments will be conducted under the following conditions. We include the dataset descriptions and important experimental details in Appendix Sections I.1 and I.2. Here, we present only the comparison methods:

Comparing Methods. We compared our proposed method, SPFL, with the most classic algorithms in federated learning, both with and without integration. This comparison aimed to evaluate the performance enhancement that the integration of SPFL brings to the federated learning process. Including: (1) Fedavg [7] is the first algorithm for FL and PFL; (2) FedProx [34], SCAFFOLD [19], FedDyn [20], MOON [35], and FedDC[32] focus on dynamically adjusting the client models; (3) FedNova [33] adjusts the iteration counts from a global perspective to optimize the federated learning process; (4) FedDisco [28] is the first paper that focuses on addressing category shift in FL.

To comprehensively evaluate the impact of the SPFL update framework under domain shift, we compare not only with traditional PFL methods but also with domain generalization approaches (e.g.,



(a) IID (b) Non-IID (c) Without/With weighted average

Figure 4: (a) Comparison of the performance of SPFL and PFL on IID data in the cifar 10; (b) Comparison of the performance of SPFL and PFL on Non-IID in the cifar 10; (c) indicates that the Weighted Average shows no significant difference from the Arithmetic Average in SPFL.

Table 4: Comparison between PFL method and *GLAM* under the updated framework of SPFL

Method	is PFL?	with SPFL	***P**hoto*	***A**rt*	***C**artoon*	***S**ketch*	Avg
Ditto [40]	✓	✓	92.16	74.22	69.03	63.09	74.63
Fedavg [7]	✓	✓	90.96	76.17	72.31	65.84	76.32
Fedprox [34]	✓	✓	91.62	76.07	71.20	67.68	76.64
Scaffold [19]	✓	✓	91.32	76.61	67.45	67.12	75.62
AM [36]	✓	✓	92.16	81.01	67.15	68.01	77.08
RSC [37]	✓	✓	89.82	75.05	72.40	66.94	76.05
CCNet [39]	✓	✓	91.92	75.10	65.10	**70.78**	75.73
Fedseq [41]	×	×	70.54	59.01	58.53	52.28	60.04
FedDG-GA [38]	✓	✓	90.78	69.63	69.58	63.76	73.44
SPFL-GLAM (ours)	×	✓	**92.28**	**77.44**	**74.42**	67.32	**77.87**

AM [36], RSC [37]) and federated domain generalization methods (e.g., FedDG-GA [38], CCNet [39]). Ditto [40] is a personalized FL method that incorporates model distillation. Since our approach also involves distillation components, we include it in our comparisons as well. As shown in Tab. 1, our proposed GLAM effectively mitigates the impact of domain shift within the SPFL framework. Currently, the performance of most federated sequential learning methods heavily depends on data distribution and hierarchical aggregation strategies. For comparison, we include FedSeq [41], which demonstrates limited effectiveness under domain shift scenarios.

4.2 Results

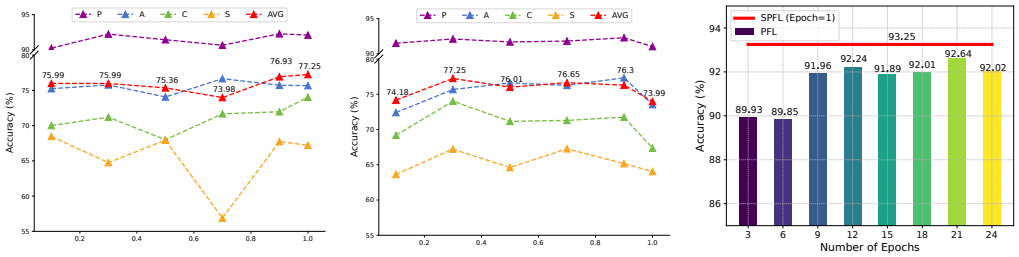
The performance on IID. From Tab. 3, it can be observed that for client data under IID, SPFL can fully leverage its capability to encounter more data. SPFL can perfectly adapt to each method, achieving an average improvement of over 6%. Additionally, SPFL demonstrates a fast convergence rate; the model can reach from 40% accuracy to 70% in just one round, as shown in Fig. 4a. However, most real-world data is Non-IID. So we focus more on the performance of SPFL in the domain and category shift. In Fig. 4b, SPFL also shows strong performance in the category shift.

Performance under Category Shift. We evaluate SPFL on four datasets (CIFAR-10 [42], CIFAR-100 [42], CINIC-10 [43], and FMNIST [44]), as shown in Tab. 2. SPFL can be effectively integrated with existing federated learning methods, yielding an average performance gain of approximately 3%. This integration highlights the compatibility and potential of SPFL in enhancing federated learning algorithms across diverse data distributions under category shift.

Performance under Domain Shift. In Tab. 5, the proposed SPFL framework alone does not perform well under domain shift. To address this limitation, we introduce the Global-Local Alignment Module (GLAM), which significantly narrows the performance gap. With the integration of GLAM, SPFL's performance improved from 2.05% below the expected baseline to just 0.5% below, demonstrating the module's effectiveness in enhancing SPFL's robustness to domain shift. Additionally, most PFL methods integrated into the SPFL framework show suboptimal performance, as illustrated in Tab. 4.

4.3 Ablation Studies

Impact of Hyperparameters τ and ρ in the model. As shown in Fig. 5a and 5b, we observe that setting τ and ρ too high leads the model to overly align with the global models and previous



(a) Sensitivity to ρ (b) Sensitivity to τ (c) The efficiency of SPFL

Figure 5: (a) and (b) illustrate the sensitivity to GLAM hyperparameters; (c) demonstrates increasing the number of PFL's local Epoch under category shift does not lead to significant improvement.

Table 5: The combination experiments of the two different loss functions $\mathcal{L}_{ap}$ and $\mathcal{L}_{ag}$ were conducted on the PACS dataset using a pre-trained ResNet-18 (Best in **bold**)

Method	$\mathcal{L}_{ag}$	$\mathcal{L}_{ap}$	*Photo*	*Art*	*Cartoon*	*Sketch*	Avg
PFL (Fedavg)	×	×	90.48	76.95	74.87	**71.21**	**78.37**
SFL (Fedseq)	×	×	70.54	59.01	58.53	52.28	60.04
SPFL	×	×	90.96	76.17	72.31	65.84	76.32
SPFL	✓	×	91.50	73.73	72.40	66.58	76.05
SPFL	×	✓	92.22	76.12	71.46	64.34	76.04
SPFL	✓	✓	*92.28*	**77.44**	**74.42**	67.32	77.87

rounds' models, which in turn degrades local training performance.Through our sensitivity analysis, we observe that the combination $\tau = 0.3$ and $\rho = 1$ yields the best model performance.

Impact of the GLAM Module. As shown in Tab. 5, SPFL alone is insufficient to effectively address domain shift. However, with the introduction of the proposed GLAM module, the SPFL update method becomes competitive under domain shift scenarios. Each component of GLAM ($\mathcal{L}_{ap}$ and $\mathcal{L}_{ag}$) contributes to the overall performance improvement. In addition, we added in the appendix Tab. 10 that if PFL is used first to improve the overall ability of the model, and then SPFL is used, the generalization of the model can be further improved.

The Impact of Uploaded Dataset Size on the Model. In PFL, the contribution of each client is weighted based on the size of its dataset. In contrast, SPFL iteratively updates the model over all client datasets, effectively treating them as having equal size. **As a result, SPFL does not require uploading client data, offering stronger privacy protection compared to PFL.** As shown in Fig. 4c, our experimental results further confirm that the presence or absence of client data upload has no significant impact on model performance in SPFL.

The impact of Epoch on the SPFL framework. In our setting, we trained for 1 epoch with 10 clients, resulting in a computational cost that is 10 times higher than PFL. However, as shown in Fig. 5c, when the number of epochs is set to 10, the resource consumption of PFL and SPFL becomes comparable. In Fig. 5c, we conduct 200 rounds of training to approximate the theoretical upper bound. It can be observed that SPFL improves the upper limit of model convergence when addressing category shift, which is consistent with the results reported in Tab. 1. **Notably, even when the number of PFL epochs is increased to 21—doubling the computational cost compared to SPFL—the model performance of PFL remains 0.6% lower than that of SPFL!**

5 Communication cost analysis

When directly applying SPFL to approximate the performance of PFL, it often incurs significant communication and computational overhead. To demonstrate the advantage of SPFL in achieving a lower convergence bound, the pre-trained model reported in Tab. 6 and Tab. 7 is trained on CIFAR-10 with FedAvg for 200 rounds across 10 clients. Furthermore, to show that SPFL can attain a superior convergence upper bound, we use SPFL to match the effect of PFL within only 10 rounds, surpassing the performance of PFL trained for thousands of rounds, while simultaneously

Table 6: Efficiency Comparison on CIFAR-10 (Dirichlet α=0.1, Number of clients=10)

Method	Accuracy(%)	Rounds	Comm. Cost (GB)	Total Time(s)	Compute Cost(GB)
PFL	79.15	500	116.90	3419.91	175.35
SPFL(Ours)	84.17	15	21.042	678.6	87.675

Table 7: Efficiency Comparison on CIFAR-10 (Dirichlet α=0.1, Number of clients=100)

Method	Accuracy(%)	Rounds	Comm. Cost (TB)	Total Time(s)	Compute Cost(TB)
PFL	81.38	2000	18.30	30259.82	27.46
SPFL(Ours)	86.68	10	4.68	14713.13	22.85

reducing both communication and computational costs. This provides a new direction for the practical application of SPFL.

6 Conclusion

In this paper, we first identify and analyze the limitations of existing Sequential Federated Learning (SFL) methods under non-independent and identically distributed (Non-IID) scenarios, with a focus on addressing category and domain shifts caused by update order sensitivity and catastrophic forgetting. To tackle these challenges, we propose SPFL, a novel framework that integrates seamlessly with existing federated learning (FL) methods. SPFL combines sequential updates with parallel aggregation to enhance data utilization and reduce sensitivity to update order. In this framework, the model is distributed to all clients, where each client performs local training and sequentially passes the model. The server then aggregates the updated models from each starting client to generate a new global model. We provide a convergence analysis of the SPFL update scheme, showing that it achieves faster convergence than both PFL and SFL across strongly convex, general convex, and non-convex settings. Additionally, we introduce the Global-Local Alignment Module (GLAM) to address catastrophic forgetting by aligning the predictions of the global model with those of the local and previous models during training. Extensive experiments demonstrate the effectiveness of SPFL and GLAM under Non-IID conditions—specifically domain and category shifts—and confirm their compatibility with traditional Parallel Federated Learning (PFL) methods.

7 Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant T2125006 and Grant 42401415; in part by Shenzhen Science and Technology Program under Grant KCXFZ20240903093759004 and Grant KJZD20230923115106012; in part by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University, under Project 24xkjc002; and in part by Jiangsu Innovation Capacity Building Program under Project BM2022028.

References

[1] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[4] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021.

[5] Yichen Li, Haozhao Wang, Wenchao Xu, Tianzhe Xiao, Hong Liu, Minzhu Tu, Yuying Wang, Xin Yang, Rui Zhang, Shui Yu, Song Guo, and Ruixuan Li. Unleashing the power of continual learning on non-centralized devices: A survey, 2024.

[6] Haoyuan Liang, Xinyu Zhang, Shilei Cao, Guowen Li, and Juepeng Zheng. Tta-feddg: Leveraging test-time adaptation to address federated domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18658–18666, 2025.

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

[8] Haozhao Wang, Yichen Li, Wenchao Xu, Ruixuan Li, Yufeng Zhan, and Zhigang Zeng. Dafkd: Domain-aware federated knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20412–20421, 2023.

[9] Yichen Li, Yijing Shan, Yi Liu, Haozhao Wang, Wei Wang, Yi Wang, and Ruixuan Li. Personalized federated recommendation for cold-start users via adaptive knowledge fusion. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2700–2709, New York, NY, USA, 2025. Association for Computing Machinery.

[10] Riccardo Zaccone, Andrea Rizzardi, Debora Caldarola, Marco Ciccone, and Barbara Caputo. Speeding up heterogeneous federated learning with sequentially trained superclients. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3376–3382. IEEE, 2022.

[11] Andrea Silvi, Andrea Rizzardi, Debora Caldarola, Barbara Caputo, and Marco Ciccone. Accelerating federated learning via sequential training of grouped heterogeneous clients. *IEEE Access*, 2024.

[12] Xuyang Li, Weizhuo Zhang, Yue Yu, Wei-Shi Zheng, Tong Zhang, and Ruixuan Wang. Sift: A serial framework with textual guidance for federated learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 655–665. Springer, 2024.

[13] Edvin Listo Zec, Adam Breitholtz, and Fredrik D Johansson. Overcoming label shift in targeted federated learning. *arXiv preprint arXiv:2411.03799*, 2024.

[14] Zhuang Qi, Lei Meng, Zhaochuan Li, Han Hu, and Xiangxu Meng. Cross-silo feature space alignment for federated learning on clients with imbalanced data. In *The 39th Annual AAAI Conference on Artificial Intelligence (AAAI-25)*, pages 19986–19994, 2025.

[15] Yichen Li, Wenchao Xu, Yining Qi, Haozhao Wang, Ruixuan Li, and Song Guo. Sr-fdil: Synergistic replay for federated domain-incremental learning. *IEEE Transactions on Parallel and Distributed Systems*, 35(11):1879–1890, 2024.

[16] Zhuang Qi, Sijin Zhou, Lei Meng, Han Hu, Han Yu, and Xiangxu Meng. Federated deconfounding and debiasing learning for out-of-distribution generalization. *arXiv preprint arXiv:2505.04979*, 2025.

[17] Yichen Li, Yuying Wang, Haozhao Wang, Yining Qi, Tianzhe Xiao, and Ruixuan Li. Fedssi: Rehearsal-free continual federated learning with synergistic synaptic intelligence. In *Forty-second International Conference on Machine Learning*.

[18] Xuankun Rong, Jianshu Zhang, Kun He, and Mang Ye. Can: Leveraging clients as navigators for generative replay in federated continual learning. In *Forty-second International Conference on Machine Learning*, 2025.

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

[20] Cheng Jin, Xuandong Chen, Yi Gu, and Qun Li. Feddyn: A dynamic and efficient federated distillation approach on recommender system. In *2022 IEEE 28th international conference on parallel and distributed systems (ICPADS)*, pages 786–793. IEEE, 2023.

[21] Xuezheng Liu, Zhicong Zhong, Yipeng Zhou, Di Wu, Xu Chen, Min Chen, and Quan Z Sheng. Accelerating federated learning via parallel servers: A theoretically guaranteed approach. *IEEE/ACM Transactions on Networking*, 30(5):2201–2215, 2022.

[22] Yichen Li, Qunwei Li, Haozhao Wang, Ruixuan Li, Wenliang Zhong, and Guannan Zhang. Towards efficient replay in federated incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12820–12829, June 2024.

[23] Haozhao Wang, Haoran Xu, Yichen Li, Yuan Xu, Ruixuan Li, and Tianwei Zhang. FedCDA: Federated learning with cross-rounds divergence-aware aggregation. In *The Twelfth International Conference on Learning Representations*, 2024.

[24] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

[25] Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *International conference on machine learning*, pages 18250–18280. PMLR, 2022.

[26] Yichen Li, Wenchao Xu, Haozhao Wang, Yining Qi, Jingcai Guo, and Ruixuan Li. Personalized federated domain-incremental learning based on adaptive knowledge matching. In *European conference on computer vision*, pages 127–144. Springer, 2024.

[27] Zhen Wang, Daniyal M Alghazzawi, Li Cheng, Gaoyang Liu, Chen Wang, Zeng Cheng, and Yang Yang. Fedcsa: Boosting the convergence speed of federated unlearning under data heterogeneity. In *2023 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 388–393. IEEE, 2023.

[28] Rui Ye, Mingkai Xu, Jianyu Wang, Chenxin Xu, Siheng Chen, and Yanfeng Wang. Feddisco: Federated learning with discrepancy-aware collaboration. In *International Conference on Machine Learning*, pages 39879–39902. PMLR, 2023.

[29] Anastasia Pustozerova and Rudolf Mayer. Information leaks in federated learning. In *Proceedings of the network and distributed system security symposium*, volume 10, page 122, 2020.

[30] Shenglai Zeng, Zonghang Li, Hongfang Yu, Yihong He, Zenglin Xu, Dusit Niyato, and Han Yu. Heterogeneous federated learning via grouped sequential-to-parallel training. In *International Conference on Database Systems for Advanced Applications*, pages 455–471. Springer, 2022.

[31] Yipeng Li and Xinchen Lyu. Convergence analysis of sequential federated learning on heterogeneous data. *Advances in Neural Information Processing Systems*, 36, 2024.

[32] Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10112–10121, 2022.

[33] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

[34] Xiaotong Yuan and Ping Li. On convergence of fedprox: Local dissimilarity invariant bounds, non-smoothness and beyond. *Advances in Neural Information Processing Systems*, 35:10752–10765, 2022.

[35] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722, 2021.

[36] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14383–14392, 2021.

[37] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part II 16*, pages 124–140. Springer, 2020.

[38] Ruipeng Zhang, Qinwei Xu, Jiangchao Yao, Ya Zhang, Qi Tian, and Yanfeng Wang. Federated domain generalization with generalization adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3954–3963, 2023.

[39] Ahmed Radwan and Mohamed Shehata. Fedpartwhole: federated domain generalization via consistent part-whole hierarchies. *Pattern Analysis and Applications*, 28(2):61, 2025.

[40] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pages 6357–6368. PMLR, 2021.

[41] Zhikun Chen, Daofeng Li, Rui Ni, Jinkang Zhu, and Sihai Zhang. Fedseq: A hybrid federated learning framework based on sequential in-cluster training. *IEEE Systems Journal*, 17(3):4038–4049, 2023.

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

[43] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

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

[45] Guowen Li, Xintong Liu, Shilei Cao, Haoyuan Liang, Mengxuan Chen, Lixian Zhang, Jinxiao Zhang, Jiuke Wang, Meng Jin, Juepeng Zheng, et al. Tianquan-climate: A subseasonal-to-seasonal global weather model via incorporate climatology state. *arXiv preprint arXiv:2504.09940*, 2025.

[46] Lei Meng, Zhuang Qi, Lei Wu, Xiaoyu Du, Zhaochuan Li, Lizhen Cui, and Xiangxu Meng. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36, 2024.

[47] Zhuang Qi, Lei Meng, Zitan Chen, Han Hu, Hui Lin, and Xiangxu Meng. Cross-silo prototypical calibration for federated learning with non-iid data. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3099–3107, 2023.

[48] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1013–1023, 2021.

[49] Meirui Jiang, Zirui Wang, and Qi Dou. Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1087–1095, 2022.

[50] Junming Chen, Meirui Jiang, Qi Dou, and Qifeng Chen. Federated domain generalization for image recognition via cross-client style transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 361–370, 2023.

[51] Enyi Jiang, Yibo Jacky Zhang, and Sanmi Koyejo. Principled federated domain adaptation: Gradient projection and auto-weighting. *arXiv preprint arXiv:2302.05049*, 2023.

[52] Trong-Binh Nguyen, Minh-Duong Nguyen, Jinsun Park, Quoc-Viet Pham, and Won Joo Hwang. Federated domain generalization with data-free on-server matching gradient. *arXiv preprint arXiv:2501.14653*, 2025.

[53] Xingyu Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018.

[54] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International conference on artificial intelligence and statistics*, pages 4519–4529. PMLR, 2020.

[55] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International conference on machine learning*, pages 5381–5393. PMLR, 2020.

[56] Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

[57] Francesco Orabona. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2019.

[58] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[59] S. Boyd, L. Vandenberghe, and L. Faybusovich. Convex optimization. *IEEE Transactions on Automatic Control*, 51(11):1859–1859, 2006.

[60] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[61] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[62] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[63] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[64] Farhad Pourpanah, Mahdiyar Molahasani, Milad Soltany, Michael Greenspan, and Ali Etemad. Federated unsupervised domain generalization using global and local alignment of gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19948–19958, 2025.

[65] Yichen Li, Haozhao Wang, Yining Qi, Wei Liu, and Ruixuan Li. Re-fed+: A better replay strategy for federated incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the scope of applicability and the computation and communication overhead in detail in section J.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

We have provided the full set of assumptions and a complete (and correct) proof. Please refer to Section 3 in the main paper and Section G in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided detailed implementation details in Section 4 of the main paper and Section I.1 and I.2 of the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have provided detailed implementation details in Section 4 of the main paper, as well as in Sections I.1 and I.2 of the Appendix. The source code is included in the supplementary material. All datasets used are publicly available and properly cited in the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have provided detailed implementation details in Section 4 of the main paper, as well as in Sections I.1 and I.2 of the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Our experimental results are computed three times and report the average result.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: My research group supports me in computer resources.The specific hardware used for the experiments is described in Appendix Section I.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper fully adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discussed both potential positive societal impacts and negative societal impacts in Section K of the Appendix.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The data and code used in this paper have obtained legal permissions.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: The large language model (LLM) used in this paper is solely employed for text polishing and formatting.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

SPFL: Sequential Updates with Parallel Aggregation for Enhanced Federated Learning Under Category and Domain Shifts (Supplementary material)

Table of Contents in Appendix

A Pseudocode

Here we show our algorithm process.

Algorithm 1 Sequential updates With Parallel aggregation Federated learning (SPFL)

Require: Initial global model $w = w_0$, the number of clients is k , the client dataset $D = \{D_1, D_2, ..., D_k\}$, (Hyperparameters:local epoch T,total aggregation round R)
Ensure: Final global model w_R

Server:
  Deploy the global model w_0 to each client to obtain each client model:$w_0^k = w_0$
  for each round $i = 1, 2, ..., R$ **do**
    for each client $e = 1, 2, ..., k$ **in parallel do**
      for each client $a = e, ..., k\&1, ..., e-1$ **do**
        if $e + 1 \leq k$:
          $w_i^{e+1} \leftarrow$ **Client** (e, w_i)
        else:
          $w_{i+1}^e \leftarrow$ **Client** (e, w_i)
      $w_{i+1} \leftarrow \frac{1}{k} \sum_{e=1}^{k} w_{i+1}^e$
    Deploy w_{i+1} to all clients.
Client:
  for each local epoch $t = 1, 2, ..., T$ **do**
    for Batch $b = 1, 2, ..., B$ **do**
      $\nabla \mathcal{L}(w_{i+(t)}; x_{B_q}^{D_k}) = \nabla \mathcal{L}_{ce} + \tau \nabla \mathcal{L}_{ag} + \rho \nabla \mathcal{L}_{ap}$
      $w_{i+(t+1)} \leftarrow w_{i+(t)} - \eta \nabla \mathcal{L}(w_{i+(t)}; x_{B_q}^{D_k})$
  if $k \,\%\, R$:
    return $w_{i+1} = w_{i+(T)}$ to **Next Client**
  else :
    return $w_{i+1} = w_{i+(T)}$ to **Server**

B More Related Work

Domain generalization is also an important criterion for evaluating the model. Examining the generalization ability of the global model is crucial, as it better reflects the overall performance of the model rather than its tendency to overfit. Therefore, we introduce related work on Federated Domain Generalization in this section.

B.1 Federated Domain Generalization

In recent years, Domain Generalization (DG) has developed rapidly, aiming to learn models that generalize well to unseen domains. However, in practice, multi-source data often exhibit significant domain shifts—for example, between general-purpose datasets and meteorological datasets [45]. Federated Domain Generalization (FedDG) [46, 47] is an emerging field where each domain trains a local model, which is then aggregated, considering the generalization ability of the global model on target clients (domains). Domain shift is one of the most important issues in FedDG, which is consistent with this paper. Despite recent progress, research in this area is still limited. For example, ELCFS [48] generalizes the model by sharing spectra across domains, yet this approach risks privacy leaks. [49] further proposes a flatness-aware optimization method for better generalization on local updates. CCST [50] could lead to more uniform distributions of source clients and make each local model learn to fit the image styles of all the clients to avoid the different model biases. An adaptive weighting method, named GA [38], is proposed to achieve a tighter generalization bound through explicit re-weighted aggregation. However, the abovementioned studies only focus on the Federated Aggregation. To this end, AutoFedGP [51] tackles the domain adaptation (DA) problem by projecting local gradients and assigning adaptive weights to align source and target domains. For domain generalization (DG), FedOMG [52] enhances robustness to unseen domains by aligning gradient directions across clients without access to target data. However, both approaches focus on model-level alignment, which may be insufficient when domain shifts induce severe forgetting during sequential

client updates. In contrast, our proposed SPFL explicitly targets this issue by introducing GLAM. This global-local alignment mechanism concurrently aligns each client with both its predecessor and the global model, effectively mitigating the forgetting problem caused by domain shift.

C Corollary 1

Due to space limitations, we introduce Corollary 1 here.

Corollary 1. *Applying the results of Theorem 1, and by selecting an appropriate learning rate (see the proof of Theorem 1 in the Appendix Section G), we derive the following convergence bounds for SPFL:*

Strongly convex: *Under Assumptions 2, 3, and 5, there exists a constant effective learning rate* $\frac{1}{\mu R} \leq \tilde{\eta} \leq \frac{1}{6L}$ *and weights* $\theta_r = \left(1 - \frac{\mu\tilde{\eta}}{2}\right)^{-(r+1)}$*, such that the following holds:*

$$\mathbb{E}\left[\mathcal{L}(\bar{\boldsymbol{w}}^{(R)}) - \mathcal{L}(\boldsymbol{w}^*)\right] = \tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu M^2 K R} + \frac{L\sigma^2}{\mu^2 M K R^2} + \frac{L\zeta_*^2}{\mu^2 M R^2} + \mu\mathcal{A}^2 \exp\left(-\frac{\mu R}{12L}\right)\right) \quad (12)$$

General convex: *Under Assumptions 2, 3, and 5, there exists a constant effective learning rate* $\tilde{\eta} \leq \frac{1}{6L}$ *and weights* $\theta_r = 1$*, such that the following holds:*

$$\mathbb{E}\left[\mathcal{L}(\bar{\boldsymbol{w}}^{(R)}) - \mathcal{L}(\boldsymbol{w}^*)\right] = \mathcal{O}\left(\frac{\sigma\mathcal{A}}{\sqrt{M^2 K R}} + \frac{\left(L\sigma^2\mathcal{A}^4\right)^{1/3}}{(MK)^{1/3}R^{2/3}} + \frac{\left(L\zeta_*^2\mathcal{A}^4\right)^{1/3}}{M^{1/3}R^{2/3}} + \frac{L\mathcal{A}^2}{R}\right) \quad (13)$$

Non-convex: *Under Assumptions 2, 3, and 4, there exists a constant effective learning rate* $\tilde{\eta} \leq \frac{1}{6L(\beta+1)}$ *and weights* $\theta_r = 1$*, such that the following holds:*

$$\min_{0\leq r\leq R} \mathbb{E}\left[\|\nabla\mathcal{L}(\boldsymbol{w}^{(r)})\|^2\right] = \mathcal{O}\left(\frac{\left(L\sigma^2\mathcal{B}\right)^{1/2}}{\sqrt{M^2 K R}} + \frac{\left(L^2\sigma^2\mathcal{B}^2\right)^{1/3}}{(MK)^{1/3}R^{2/3}} + \frac{\left(L^2\zeta^2\mathcal{B}^2\right)^{1/3}}{M^{1/3}R^{2/3}} + \frac{L\beta\mathcal{B}}{R}\right) \quad (14)$$

where $\mathcal{O}$ omits absolute constants, $\tilde{\mathcal{O}}$ omits absolute constants and polylogarithmic factors, $\mathcal{A} := \|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\|$ for the convex cases, and $\mathcal{B} := \mathcal{L}(\boldsymbol{w}^{(0)}) - \mathcal{L}^*$ for the non-convex case.

D Notations

Tab.8 summarizes the notations appearing in this paper. We follow the same settings as in [31] to compare the convergence bounds better.

E Assumption

We assume that:

- $\mathcal{L}$ is lower bounded by $\mathcal{L}^*$ for all cases and there exists a minimizer $\boldsymbol{w}^*$ such that $\mathcal{L}(\boldsymbol{w}^*) = \mathcal{L}^*$ for strongly and generally convex cases;
- each local objective function is *L-smooth*(Assumption.2);
- Furthermore, we need to make assumptions about the diversities: the assumptions on the stochasticity bounding the diversity of $\{\mathcal{L}_m(\cdot;\xi_m^i) : i \in [|\mathcal{D}_m|]\}$ with respect to i inside each client (Assumption 3);
- the assumptions on the heterogeneity bounding the diversity of local objectives $\{\mathcal{L}_m : m \in [M]\}$ with respect to m across clients (Assumptions 4, 5).

Assumption 1 (Gradient Boundedness)**.** *There exists a constant* $G > 0$ *such that for all clients* m*, any parameter* $x \in \mathbb{R}^d$*, and any sample* ξ*, the following holds:*

$$\|\nabla\mathcal{L}_m(\boldsymbol{w};\xi)\| \leq G. \quad (15)$$

Table 8: Summary of key notations.

Symbol	Description
R, r	number, index of training rounds
M, m	number, index of clients
K, k	number, index of local update steps
q^m	Number of datasets for client m
B	Number of batch size
N	number of participating clients
n	Client n is the starting $Client$
S	S represents S clients selected from M clients. When all clients participate,$M = S$
π	$\{\pi_1, \pi_2, \ldots, \pi_M\}$ is a permutation of $\{1, 2, \ldots, M\}$
$\pi^{(r+1)}$	$\pi^{(r+1)} = \mathcal{R}\left(\pi^{(r)}\right) := \left[\pi_2^{(r)}, \pi_3^{(r)}, \ldots, \pi_M^{(r)}, \pi_1^{(r)}\right]$
η	learning rate (or stepsize)
$\tilde{\eta}$	effective learning rate ($\tilde{\eta} := MK\eta$ in SFL and SPFL and $\tilde{\eta} := K\eta$ in PFL)
μ	μ-strong convexity constant
L	L-smoothness constant (Asm. 2)
σ	upper bound on variance of stochastic gradients at each client (Asm. 3)
β, ζ	constants in Asm. 3a to bound heterogeneity everywhere
ζ_*	constants in Asm. 3b to bound heterogeneity at the optima
$\mathcal{L}/\mathcal{L}_m$	global objective/local objective of client m
$\boldsymbol{w}^{(r)}$	global model parameters in the r-th round
$\boldsymbol{w}_g$	the global model $f(\cdot; w_g)$ where w_g represents the global model parameters.
$\boldsymbol{w}_i$	Represents the parameters of the model during training,and $\boldsymbol{w}_i = \boldsymbol{w}_{intermediate}$
$\boldsymbol{w}_p$	Indicates the parameters of the model passed from the previous client training, and $\boldsymbol{w}_p = \boldsymbol{w}_{previous}$
$\boldsymbol{w}_{m,k,n}^{(r)}$	When client n is the starting client, local model parameters of the m-th client after k local steps in the r-th round
$\mathbf{g}_{\pi_m^n,k}^{(r)}$	When client n is the starting client , $\mathbf{g}_{\pi_m^n,k}^{(r)} := \nabla f_{\pi_m^n}(\boldsymbol{w}_{m,k,n}^{(r)}; \xi)$ denotes the stochastic gradients of $\mathcal{L}_{\pi_m^n}$ regarding $\boldsymbol{w}_{m,k,n}^{(r)}$

Assumption 2 (L-Smoothness)**.** *Each local objective function $\mathcal{L}_m$ is L-smooth,$m \in \{1, 2, \ldots, M\}$, i.e. , there exists a constant L > 0 such that $\|\nabla\mathcal{L}_m(\boldsymbol{x}) - \nabla\mathcal{L}_m(\mathbf{y})\| \leq L\|\boldsymbol{x} - \mathbf{y}\|$,for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}$.*

Assumption 3. *The variance of the stochastic gradient at each client is bounded:*

$$\mathbb{E}_{\xi\sim\mathcal{D}_{\pi_m^n}}\left[\left\|\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w};\xi) - \nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w})\right\|^2 |\boldsymbol{w}\right] \leq \sigma^2, \quad \forall m \in \{1, 2, \ldots, M\} \tag{16}$$

Assumption 4. *For strongly convex and generally convex functions, there exist constants β^2 and ζ^2 such that*

$$\frac{1}{M}\sum_{m=1}^{M}\|\nabla\mathcal{L}_m(\boldsymbol{w}) - \nabla\mathcal{L}(\boldsymbol{w})\|^2 \leq \beta^2\|\nabla\mathcal{L}(\boldsymbol{w})\|^2 + \zeta^2 \tag{17}$$

Assumption 5. *For non-convex functions, there exists one constant ζ_*^2 such that*

$$\frac{1}{M}\sum_{m=1}^{M}\|\nabla\mathcal{L}_m(\boldsymbol{w}^*)\|^2 = \zeta_*^2 \tag{18}$$

F Technical Lemmas

F.1 Basic Identities and Inequalities

These identities and inequalities are mostly from [53],[54],[55],[19],[56] and [31].Thanks for the analytical ideas provided by these works.

E-norm identity. (1) For any random variable **x**, letting the variance can be decomposed as

$$\mathbb{E}\left[\|\boldsymbol{w}-\mathbb{E}[\boldsymbol{w}]\|^2\right]=\mathbb{E}\left[\|\boldsymbol{w}\|^2\right]-\|\mathbb{E}[\boldsymbol{w}]\|^2 \tag{19}$$

(2) In particular, its version for vectors with finite number of values gives

$$\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i-\bar{\boldsymbol{x}}\|^2=\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^2-\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\right\|^2 \tag{20}$$

where vectors $\boldsymbol{w}_1,\ldots,\boldsymbol{w}_n\in\mathbb{R}^d$ are the values of **w** and their average is $\bar{\boldsymbol{w}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{w}_i$.

Lemma 1. *Under standard discrete summation rules, the following closed-form expressions hold for any integer $K\geq 1$:*

$$\sum_{k=1}^{K-1}k=\frac{(K-1)K}{2},\quad \sum_{k=1}^{K-1}k^2=\frac{(K-1)K(2K-1)}{6},\quad \sum_{k=1}^{K-1}k^3=\left(\frac{(K-1)K}{2}\right)^2 \tag{21}$$

Lemma 2 (Subadditivity of Concave Power Functions)**.** *Let $0<n<1$ and $a,b>0$. The power function $f(x)=x^n$ is concave, and the following inequality holds:*

$$(a+b)^n\leq a^n+b^n$$

Jensen's inequality. For any convex function *h* and any vectors $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ we have

$$h\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\right)\leq\frac{1}{n}\sum_{i=1}^{n}h(\boldsymbol{x}_i) \tag{22}$$

As a special case with $h(\boldsymbol{x})=\|\boldsymbol{x}\|^2$,we obtain

$$\left\|\frac{1}{n}\sum_{i=1}^{n}x_i\right\|^2\leq\frac{1}{n}\sum_{i=1}^{n}\|x_i\|^2 \tag{23}$$

Smoothness and general convexity, strong convexity. There are some useful inequalities concerning L-smoothness (Assumption 2), convexity, and $\mu-strong\ convexity$. Their proofs can be found in [53] and [56].

Bregman Divergence associated with function *h* and arbitrary $\boldsymbol{x},\boldsymbol{y}$ is denoted as

$$D_h(\boldsymbol{x},\boldsymbol{y}):=h(\boldsymbol{x})-h(\boldsymbol{y})-\langle\nabla h(\boldsymbol{y}),\boldsymbol{x}-\boldsymbol{y}\rangle \tag{24}$$

The divergence is strictly non-negative when the function *h* is convex. A more formal definition can be found in [57]. One corollary[58] called the three-point identity, is

$$D_h(\boldsymbol{z},\boldsymbol{x})+D_h(\boldsymbol{x},\boldsymbol{y})-D_h(\boldsymbol{z},\boldsymbol{y})=\langle\nabla h(\boldsymbol{y})-\nabla h(\boldsymbol{x}),\boldsymbol{z}-\boldsymbol{x}\rangle \tag{25}$$

where $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}$ are three points in the set.

When *h* is *L-smooth*, with the definition of *Bregman divergence*, a consequence of *L-smoothness* is

$$D_h(\boldsymbol{x},\boldsymbol{y})=h(\boldsymbol{x})-h(\boldsymbol{y})-\langle\nabla h(\boldsymbol{y}),\boldsymbol{x}-\boldsymbol{y}\rangle\leq\frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 \tag{26}$$

Further, if *h* is *L-smooth* and lower bounded by h_*, then

$$\|\nabla h(\boldsymbol{x})\|^2\leq 2L\left(h(\boldsymbol{x})-h_*\right) \tag{27}$$

If *h* is smooth and convex in L (the definition of convexity can be found in [59]), then

$$D_h(\boldsymbol{x},\boldsymbol{y})\geq\frac{1}{2L}\|\nabla h(\boldsymbol{x})-\nabla h(\boldsymbol{y})\|^2 \tag{28}$$

The function $h:\mathbb{R}^d\to\mathbb{R}$ is $\mu-strongly$ convex if and only if there exists a convex function $g:\mathbb{R}^d\to\mathbb{R}$ such that $h(\boldsymbol{x})=g(\boldsymbol{x})+\frac{\mu}{2}\|\boldsymbol{x}\|^2$.

If *h* is $\mu-strongly$ convex, it holds that

$$\frac{\mu}{2}\|x-\boldsymbol{y}\|^2\leq D_h(\boldsymbol{x},\boldsymbol{y}) \tag{29}$$

F.2 Technical Lemmas

We obtain a recurrence relation suitable for SPFL based on previous work[31, 19], as shown below: Lemma .3 ,Lemma .4 and lemma .5. Two nonnegative sequences $\{r_t\}_{r\geq 0}$, $\{s_t\}_{r\geq 0}$, which satisfy the relation

$$r_{t+1} \leq (1 - a\gamma_t)r_t - b\gamma_t s_t + c_1\gamma_t^2 + c_2\gamma_t^3, \tag{30}$$

for all $t \geq 0$ and for parameters $b > 0$, $a, c_1, c_2 \geq 0$ and non-negative learning rates $\{\gamma_t\}_{r\geq 0}$ with $\gamma_t < \frac{1}{d}$, $\forall t \geq 0$, for a parameter $d \geq a$, $d > 0$.

Lemma 3 (linear convergence rate and Constant Step sizes)**.** *$\{r_t\}_{r\geq 0}$, $\{r_t\}_{r\geq 0}$ as in (30) and $a > 0$. Then there exists a constant step size $\gamma_t = \gamma \leq \frac{1}{d}$ such that for weights $\theta_t := (1 - a\gamma)^{-(t+1)}$ and $W_t := \sum_{t=0}^{T} w_t$ it holds:*

$$\begin{aligned}\Psi_T = \frac{b}{W_T}\sum_{t=0}^{T} s_t\theta_t &\leq 3ar_0(1 - a\gamma)^{(T+1)} + c_1\gamma + c_2\gamma^2 \\ &\leq 3ar_0 \exp\left[-a\gamma(T + 1)\right] + c_1\gamma + c_2\gamma^2\end{aligned} \tag{31}$$

$$\frac{b}{W_T}\sum_{t=0}^{T} s_t\theta_t + ar_{T+1} = \tilde{\mathcal{O}}\left(dr_0 \exp\left[-\frac{aT}{d}\right] + \frac{c_1}{aT} + \frac{c_2}{a^2T^2}\right) \tag{32}$$

Proof. We start by rearranging (30) and multiplying both sides with θ_t:

$$bs_t\theta_t \leq \frac{\theta_t(1 - a\gamma)r_t}{\gamma} - \frac{\theta_t r_{t+1}}{\gamma} + c_1\gamma\theta_t + c_2\gamma^2\theta_t = \frac{w_{t-1}r_t}{\gamma} - \frac{\theta_t r_{t+1}}{\gamma} + c_1\gamma\theta_t + c_2\gamma^2\theta_t$$

By summing from $t = 0$ to $t = T$, we obtain a telescoping sum:

$$\frac{b}{W_T}\sum_{t=0}^{T} s_t\theta_t \leq \frac{1}{\gamma W_T}\left(w_0(1 - a\gamma)r_0 - w_T r_{T+1}\right) + c_1\gamma + c_2\gamma^2$$

and hence

$$\Psi_T = \frac{b}{W_T}\sum_{t=0}^{T} s_t\theta_t \leq \frac{b}{W_T}\sum_{t=0}^{T} s_t\theta_t + \frac{w_T r_{T+1}}{\gamma W_T} \leq \frac{r_0}{\gamma W_T} + c_1\gamma + c_2\gamma^2$$

With the estimates

- $W_T = (1 - a\gamma)^{-(T+1)}\sum_{t=0}^{T}(1 - a\gamma)^t \leq \frac{w_T}{a\gamma}$ (here we leverage $a\gamma \leq \frac{a}{d} \leq 1$)
- and $W_T \geq w_T = (1 - a\gamma)^{-(T+1)}$

We can further simplify the left and right-hand sides:

$$\frac{b}{W_T}\sum_{t=0}^{T} s_t w_t + ar_{T+1} \leq (1 - a\gamma)^{(T+1)}\frac{r_0}{\gamma} + c_1\gamma + c_2\gamma^2 \leq \frac{r_0}{\gamma}\exp\left[-a\gamma(T + 1)\right] + c_1\gamma + c_2\gamma^2$$

Now the lemma follows by carefully tuning $\gamma.\gamma = \frac{\ln(\max\{2, a^2r_0T^2/c_1\})}{aT}$ and $\gamma = \frac{\ln(\max\{2, a^3r_0T^3/c_2\})}{aT}$,yielding two choices of γ.Consider the three cases:

- If $\frac{1}{d} \geq \frac{\ln(\max\{2, a^2r_0T^2/c_1\})}{aT}$ then we choose $\gamma = \frac{\ln(\max\{2, a^2r_0T^2/c_1\})}{aT}$ and get that Eq.32:
  $\tilde{\mathcal{O}}\left(ar_0T\exp[-\ln(\max\{2, a^2r_0T^2/c_1\})]\right) + \tilde{\mathcal{O}}\left(\frac{c_1}{aT} + \frac{c_2}{a^2T^2}\right) = \tilde{\mathcal{O}}\left(\frac{c_1}{aT} + \frac{c_2}{a^2T^2}\right)$
  as in case $2 \geq a^2r_0T^2/c_1$ it holds $ar_0T \leq \frac{2c_1}{aT}$.
- If $\frac{1}{d} \geq \frac{\ln(\max\{2, a^3r_0T^3/c_2\})}{aT}$ then we choose $\gamma = \frac{\ln(\max\{2, a^3r_0T^3/c_2\})}{aT}$ and get that Eq.32:
  $\tilde{\mathcal{O}}\left(ar_0T\exp[-\ln(\max\{2, a^3r_0T^3/c_2\})]\right) + \tilde{\mathcal{O}}\left(\frac{c_1}{aT} + \frac{c_2}{a^2T^2}\right) = \tilde{\mathcal{O}}\left(\frac{c_1}{aT} + \frac{c_2}{a^2T^2}\right)$
  as in case $2 \geq a^3r_0T^3/c_2$ it holds $ar_0T \leq \frac{2c_2}{a^2T^2}$.

- If otherwise $\frac{1}{d} < \frac{\ln(\max\{2, a^2 r_0 T^2/c_1, a^3 r_0 T^3/c_2\})}{aT}$ then we pick $\gamma = \frac{1}{d}$ and get that Eq.32:

$$\begin{aligned} & dr_0 \exp\left[-\frac{aT}{d}\right] + \frac{c_1}{d} + \frac{c_2}{d^2} \\ & \leq dr_0 \exp\left[-\frac{aT}{d}\right] + \frac{c_1 \ln(\max\{2, a^2 r_0 T^2/c_1, a^3 r_0 T^3/c_2\})}{aT} \\ & \quad + \frac{c_2 \ln^2(\max\{2, a^2 r_0 T^2/c_1, a^3 r_0 T^3/c_2\})}{a^2 T^2} \\ & = \tilde{\mathcal{O}}\left(dr_0 \exp\left[-\frac{aT}{d}\right] + \frac{c_1}{aT} + \frac{c_2}{a^2 T^2}\right) \end{aligned}$$

Lemma 4 (linear convergence rate and Decreasing Step sizes)**.** *$\{r_t\}_{r\geq 0}$, $\{s_t\}_{r\geq 0}$ as in (30) and $a > 0$. Then there exist step sizes $\gamma_t = \gamma \leq \frac{1}{d}$ and weights $\theta_t \geq 0$, $W_t := \sum_{t=0}^{T} w_t$, such that:*

$$\frac{b}{W_T} \sum_{t=0}^{T} s_t \theta_t + a r_{T+1} \leq 32 d r_0 \exp\left[-\frac{aT}{2d}\right] + \frac{36 c_1}{aT} + \frac{36 c_2}{adT} \tag{33}$$

Proof. Let $\{r_t\}_{r\geq 0}$, $\{s_t\}_{r\geq 0}$ be as in (30) for $a > 0$ and for constant stepsizes $\gamma_t := \gamma = \frac{1}{d}$, $\forall t > 0$. Then, it holds for all $T \geq 0$. We have

$$\begin{aligned} r_T \leq (1 - a\gamma) r_{T-1} + c_1 \gamma^2 + c_2 \gamma^3 & \leq (1 - a\gamma)^T r_0 + c_1 \gamma^2 \sum_{t=0}^{T-1} (1 - a\gamma)^t + + c_2 \gamma^3 \sum_{t=0}^{T-1} (1 - a\gamma)^t \\ & \leq (1 - a\gamma)^T r_0 + \frac{c_1 \gamma}{a} + + \frac{c_2 \gamma^2}{a} \\ & \leq r_0 \exp\left[-\frac{aT}{d}\right] + \frac{c_1}{ad} + + \frac{c_2}{ad^2} \end{aligned} \tag{34}$$

Let $\{r_t\}_{r\geq 0}$, $\{s_t\}_{r\geq 0}$ as in (30) for $a > 0$and for decreasing steps $\gamma_t := \frac{2}{a(\kappa + t)}$, $\forall t > 0$, with parameter $\kappa := \frac{2d}{a}$, weights $w_t := (\kappa + t)$ and $W_T := \sum_{t=0}^{T} \theta_t$. Then

$$\begin{aligned} b s_t w_t & \leq \frac{w_t (1 - a\gamma_t) r_t}{\gamma_t} - \frac{w_t r_{t+1}}{\gamma_t} + c_1 \gamma_t w_t + + c_2 \gamma_t^2 w_t \\ & = a(\kappa + t)(\kappa + t - 2) r_t - a(\kappa + t)^2 r_{t+1} + \frac{c_1}{a} + \frac{c_2}{a} \gamma_t \\ & \leq a(\kappa + t - 1)^2 r_t - a(\kappa + t)^2 r_{t+1} + \frac{c_1}{a} + \frac{c_2}{ad} \end{aligned} \tag{35}$$

where the equality follows from the definition of γ_t and θ_t, and the inequality from $(\kappa+t)(\kappa+t-2) = (\kappa + t - 1)^2 - 1 \leq (\kappa + t - 1)^2$. Again, we have a telescoping sum:

$$\begin{aligned} \frac{b}{W_T} \sum_{t=0}^{T} s_t w_t + \frac{a(\kappa + T)^2 r_{T+1}}{W_T} & \leq \frac{a\kappa^2 r_0}{W_T} + \frac{c_1 (T + 1)}{a W_T} + \frac{c_2 (T + 1)}{ad W_T} \\ & \leq \frac{2a\kappa^2 r_0}{T^2} + \frac{2c_1}{aT} + + \frac{2c_2}{adT} \end{aligned} \tag{36}$$

with

- $W_T = \sum_{t=0}^{T} \theta_t = \sum_{t=0}^{T} (\kappa + t) = \frac{(2\kappa + T)(T+1)}{2} \geq \frac{T(T+1)}{2} \geq \frac{T^2}{2}$,
- and $W_T = \frac{(2\kappa+T)(T+1)}{2} \leq \frac{2(\kappa+T)(1+T)}{2} \leq (\kappa + T)^2$ for $\kappa = \frac{2d}{a} \geq 1$

By applying (34) and (36), we conclude the proof.

For the integer $T \geq 0$, we choose the step sizes and weights as follows:

$$\textit{if} \quad T \leq \tfrac{d}{a}, \qquad \gamma_t = \tfrac{1}{d}, \qquad \theta_t = (1 - a\gamma_t)^{-(t+1)} = \left(1 - \tfrac{a}{d}\right)^{-(t+1)},$$

$$\textit{if} \quad T > \tfrac{d}{a} \text{and} t < t_0, \quad \gamma_t = \tfrac{1}{d}, \quad \theta_t = 0,$$
$$\textit{if} \quad T > \tfrac{d}{a} \text{and } t \geq t_0, \quad \gamma_t = \tfrac{2}{a(\kappa+t-t_0)}, \quad \theta_t = (\kappa + t - t_0)^2,$$

for $\kappa = \frac{2d}{a}$ and $t_0 = [\frac{T}{2}]$. We will now show that these choices imply the claimed result.

We start with the case $T \leq \frac{d}{a}$. This case is similar to the proof of the Lemma. 32 and it suffices to consider Eq .35 for the choice $\gamma_t = \frac{1}{d}$. We observe that Eq .35 simplifies to

$$dr_0 \exp\left[-\frac{aT}{d}\right] + \frac{c_1}{d} + \frac{c_2}{d^2} \leq dr_0 \exp\left[-\frac{aT}{d}\right] + \frac{c_1}{aT} + \frac{c_2}{a^2T^2}$$

If $T > \frac{d}{a}$, then from Eq.34 we obtain the following :

$$r_{t_0} \leq r_0 \exp\left[-\frac{aT}{2d}\right] + \frac{c_1}{ad} + + \frac{c_2}{ad^2}$$

From Eq .36 we have for the second half of the iterates:

$$\frac{b}{W_T}\sum_{t=0}^{T} s_t\theta_t + ar_{T+1} = \frac{b}{W_T}\sum_{t=t_0}^{T} s_t\theta_t + ar_{T+1} \leq \frac{8a\kappa^2 r_{t_0}}{T^2} + \frac{4c_1}{aT} + \frac{4c_2}{adT}$$

Now we observe that the restart condition r_{t_0} satisfies:

$$\frac{a\kappa^2 r_{t_0}}{T^2} = \frac{a\kappa^2 r_0 \exp\left(-\frac{aT}{2d}\right)}{T^2} + \frac{\kappa^2 c}{dT^2} + \frac{\kappa^2 c_2}{d^2T^2} \leq 4ar_0 \exp\left[-\frac{aT}{2d}\right] + \frac{4c_1}{aT} + \frac{4c_2}{adT}$$

Because $T \geq \frac{d}{a}$. These inequalities show the claim:

$$\frac{b}{W_T}\sum_{t=0}^{T} s_t\theta_t + ar_{T+1} \leq 32dr_0 \exp\left[-\frac{aT}{2d}\right] + \frac{36c_1}{aT} + \frac{36c_2}{adT}$$

Lemma 5 (Sub-linear Convergence rate from [31]). *$\{r_t\}_{r\geq 0}$, $\{r_t\}_{r\geq 0}$ as in (30) and $a = 0$. Then there exists the step size $\gamma_t \leq \frac{1}{d}$ such that for weights $\theta_t := 1$ and $W_t := \sum_{t=0}^{T} w_t$ it holds:*

When our step size is a constant, that is, $\gamma_t = \gamma \leq \frac{1}{d}$, we have

$$\Psi_T := \frac{b}{T+1}\sum_{t=0}^{T} s_t \leq \frac{r_0}{\gamma(T+1)} + c_1\gamma + c_2\gamma^2$$

When we dynamically adjust the step size γ_t, we have

$$\Psi_T \leq 2c_1^{\frac{1}{2}}\left(\frac{r_0}{T+1}\right)^{\frac{1}{2}} + 2c_2^{\frac{1}{3}}\left(\frac{r_0}{T+1}\right)^{\frac{2}{3}} + \frac{dr_0}{T+1} \tag{37}$$

Proof. For constant learning rates $\gamma_t = \gamma \leq \frac{1}{d}$ we can derive the estimate

$$\Psi_T = \frac{1}{\gamma(T+1)}\sum_{t=0}^{T}(r_t - r_{t+1}) + c_1\gamma + c_2\gamma^2 \leq \frac{r_0}{\gamma(T+1)} + c_1\gamma + c_2\gamma^2$$

which is the first result of this lemma. 5. Let $\frac{r_0}{\gamma(T+1)} = c_1\gamma$ and $\frac{r_0}{\gamma(T+1)} = c_2\gamma^2$, yielding two choices of γ ,$\gamma = \left(\frac{r_0}{c_1(T+1)}\right)^{\frac{1}{2}}$ and $\gamma = \left(\frac{r_0}{c_2(T+1)}\right)^{\frac{1}{3}}$ .Then choosing $\gamma = \min\left\{\left(\frac{r_0}{c_1(T+1)}\right)^{\frac{1}{2}}, \left(\frac{r_0}{c_2(T+1)}\right)^{\frac{1}{3}}, \frac{1}{d}\right\} \leq \frac{1}{d}$, there are three cases:

If $\gamma = \frac{1}{d}$, which implies that $\gamma = \frac{1}{d} \leq \left(\frac{r_0}{c_1(T+1)}\right)^{\frac{1}{2}}$ and $\gamma = \frac{1}{d} \leq \left(\frac{r_0}{c_2(T+1)}\right)^{\frac{1}{3}}$,then:

$$\Psi_T \leq \frac{dr_0}{T+1} + \frac{c_1}{d} + \frac{c_2}{d^2} \leq \frac{dr_0}{T+1} + c_1^{\frac{1}{2}} \left(\frac{r_0}{T+1}\right)^{\frac{1}{2}} + c_2^{\frac{1}{3}} \left(\frac{r_0}{T+1}\right)^{\frac{2}{3}}$$

If $\gamma = \left(\frac{r_0}{c_1(T+1)}\right)^{\frac{1}{2}}$, which implies that $\gamma = \left(\frac{r_0}{c_1(T+1)}\right)^{\frac{1}{2}} \leq \left(\frac{r_0}{c_2(T+1)}\right)^{\frac{1}{3}}$, then:

$$\Psi_T \leq 2c_1 \left(\frac{r_0}{c_1(T+1)}\right)^{\frac{1}{2}} + c_2 \left(\frac{r_0}{c_1(T+1)}\right) \leq 2c_1^{\frac{1}{2}} \left(\frac{r_0}{T+1}\right)^{\frac{1}{2}} + c_2^{\frac{1}{3}} \left(\frac{r_0}{T+1}\right)^{\frac{2}{3}}$$

If $\gamma = \left(\frac{r_0}{c_2(T+1)}\right)^{\frac{1}{3}}$, which implies that $\gamma = \left(\frac{r_0}{c_2(T+1)}\right)^{\frac{1}{3}} \leq \left(\frac{r_0}{c_1(T+1)}\right)^{\frac{1}{2}}$, then:

$$\Psi_T \leq c_1 \left(\frac{r_0}{c_2(T+1)}\right)^{\frac{1}{3}} + 2c_2^{\frac{1}{3}} \left(\frac{r_0}{T+1}\right)^{\frac{2}{3}} \leq c_1^{\frac{1}{2}} \left(\frac{r_0}{T+1}\right)^{\frac{1}{2}} + 2c_2^{\frac{1}{3}} \left(\frac{r_0}{T+1}\right)^{\frac{2}{3}}$$

Combining these three cases, we get the second result of this lemma :

$$\Psi_T \leq 2c_1^{\frac{1}{2}} \left(\frac{r_0}{T+1}\right)^{\frac{1}{2}} + 2c_2^{\frac{1}{3}} \left(\frac{r_0}{T+1}\right)^{\frac{2}{3}} + \frac{dr_0}{T+1}$$

Lemma 6 (Simple Random Sampling from [31])**.** *Let $\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_n}$ be fixed units (e.g., vectors). The population mean and population variance are given as*

$$\overline{w} := \frac{1}{n}\sum_{i=1}^{n} w_i \quad \zeta^2 := \frac{1}{n}\sum_{i=1}^{n} \|w_i - \overline{w}\|^2 \tag{38}$$

Draw $s \in [n] = \{1, 2, \ldots, n\}$ random units $\boldsymbol{w}_{\pi_1}, \boldsymbol{w}_{\pi_2}, \ldots \boldsymbol{w}_{\pi_s}$ randomly from the population. There are two possible ways of simple random sampling, well known as "sampling with replacement (SWR)" and "sampling without replacement (SWOR)". For these two ways, the expectation and variance of the sample mean $\overline{\boldsymbol{w}}_\pi := \frac{1}{s}\sum_{p=1}^{s} \boldsymbol{w}_{\pi_p}$ satisfy

$$SWR \quad : \quad \mathbb{E}[\overline{\boldsymbol{w}}_\pi] = \overline{\boldsymbol{w}} \qquad \mathbb{E}\left[\|\overline{\boldsymbol{w}}_\pi - \overline{\boldsymbol{w}}\|^2\right] = \frac{\zeta^2}{s} \tag{39}$$

$$SWOR: \quad \mathbb{E}[\overline{\boldsymbol{w}}_\pi] = \overline{w} \qquad \mathbb{E}\left[\|\overline{\boldsymbol{w}}_\pi - \overline{\boldsymbol{w}}\|^2\right] = \frac{n-s}{s(n-1)}\zeta^2 \tag{40}$$

Proof. We can easily get the relationship between variance and expectation, as well as Eq.19 and Eq.20. If you want this proof, refer to [31].

Lemma 7. *Under the same conditions of Lemma 6,use the way "sampling without replacement" and let $b_{m,k}(i) = \begin{cases} K-1, & i \leq m-1 \\ k-1, & i = m \end{cases}$ Then for $S \leq M (M \geq 2)$, it holds that*

$$\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)} (\boldsymbol{w}_{\pi_i^n} - \overline{\boldsymbol{w}})\right\|^2\right] \leq \frac{1}{2}S^2K^3\zeta^2 \tag{41}$$

Proof. As shown below, the idea refers to [31]:

$$\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)} (\boldsymbol{w}_{\pi_i^n} - \overline{\boldsymbol{w}})\right\|^2\right] = \mathbb{E}\left[\left\|K\sum_{i=1}^{m-1}\left(\boldsymbol{w}_{\pi_i^n} - \overline{\boldsymbol{w}}\right) + k\left(\boldsymbol{w}_{\pi_m^n} - \overline{\boldsymbol{w}}\right)\right\|^2\right]$$

$$= K^2\mathbb{E}\left[\left\|\sum_{i=1}^{m-1}\left(\boldsymbol{w}_{\pi_i^n} - \overline{w}\right)\right\|^2\right] + k^2\mathbb{E}\left[\left\|\boldsymbol{w}_{\pi_m^n} - \overline{w}\right\|^2\right] + 2Kk\mathbb{E}\left[\left\langle\sum_{i=1}^{m-1}\left(\boldsymbol{w}_{\pi_i^n} - \overline{w}\right), \left(\boldsymbol{w}_{\pi_m^n} - \overline{w}\right)\right\rangle\right]$$

For the first term on the right-hand side in Eq.40, using Lemma.6, we have

$$K^2\mathbb{E}\left[\left\|\sum_{i=1}^{m-1}\left(\boldsymbol{w}_{\pi_i^n}-\overline{\boldsymbol{w}}\right)\right\|^2\right]\overset{(6)}{=}\frac{(m-1)(M-(m-1))}{M-1}K^2\zeta^2$$

For the second term on the right-hand side in Eq.40, we have

$$k^2\mathbb{E}\left[\left\|\boldsymbol{w}_{\pi_m^n}-\overline{\boldsymbol{w}}\right\|^2\right]=k^2\mathbb{E}\left[\left\|\boldsymbol{w}_{\pi_m^n}-\overline{\boldsymbol{w}}\right\|^2\right]=k^2\zeta^2$$

For the third term on the right-hand side in Eq.40, we have

$$2Kk\mathbb{E}\left[\left\langle\sum_{i=1}^{m-1}\left(\boldsymbol{w}_{\pi_i^n}-\overline{\boldsymbol{w}}\right),\left(\boldsymbol{w}_{\pi_m^n}-\overline{\boldsymbol{w}}\right)\right\rangle\right]=2Kk\sum_{i=1}^{m-1}\mathbb{E}\left[\left\langle\boldsymbol{w}_{\pi_i^n}-\overline{\boldsymbol{w}},\boldsymbol{w}_{\pi_m^n}-\overline{\boldsymbol{w}}\right\rangle\right]\overset{(6)}{=}-\frac{2(m-1)}{M-1}Kk\zeta^2$$

where we use Lemma.6 in the last equality, since $i\in\{1,2,\ldots,m-1\}\neq m$. With these three preceding equations, we get

$$\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\left(\boldsymbol{w}_{\pi_i^n}-\overline{w}\right)\right\|^2\right]=\frac{(m-1)(M-(m-1))}{M-1}K^2\zeta^2+k^2\zeta^2-\frac{2(m-1)}{M-1}Kk\zeta^2$$

Then summing the preceding terms over m and k, we can get

$$\begin{aligned}&\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\left(\boldsymbol{w}_{\pi_i^n}-\overline{\boldsymbol{w}}\right)\right\|^2\right]\\=&\frac{MK^3\zeta^2}{S(M-1)}\sum_{n=1}^{S}\sum_{m=1}^{S}(m-1)-\frac{K^3\zeta^2}{S(M-1)}\sum_{n=1}^{S}\sum_{m=1}^{S}(m-1)^2+S\zeta^2\sum_{k=0}^{K-1}k^2\\&-\frac{2K\zeta^2}{S(M-1)}\sum_{n=1}^{S}\sum_{m=1}^{S}(m-1)\sum_{k=0}^{K-1}k\end{aligned}$$

Then, applying the lemma.1, we can simplify the preceding equation as follows:

$$\begin{aligned}&\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\left(\boldsymbol{w}_{\pi_i^n}-\overline{\boldsymbol{w}}\right)\right\|^2\\&=\frac{1}{2}SK^2(SK-1)-\frac{1}{6}SK(K^2-1)-\frac{1}{M-1}(S-1)S\left(\frac{1}{6}(2S-1)K-\frac{1}{2}\right)\leq\frac{1}{2}S^2K^3\zeta^2\end{aligned}$$

Lemma 8 (from [19])**.** *Let $\{\xi_i\}_{i=1}^n$ be a sequence of random variables. And the random sequence $\{\boldsymbol{w}_i\}_{i=1}^n$ satisfy that $\boldsymbol{w}_i\in\mathbb{R}^d$ is a function of $\xi_i,\xi_{i-1},\ldots,\xi_1$ for all i. Suppose that the conditional expectation is $\mathbb{E}_{\xi_i}\left[\boldsymbol{w}_i|\xi_{i-1},\ldots,\xi_1\right]=\mathbf{e}_i$ (i.e., the vectors $\{\boldsymbol{w}_i-\mathbf{e}_i\}$ form a martingale difference sequence with respect to $\{\xi_i\}$), and the variance is bounded by $\mathbb{E}_{\xi_i}\left[\left\|\boldsymbol{w}_i-\mathbf{e}_i\right\|^2\middle|\xi_{i-1},\ldots,\xi_1\right]\leq\sigma^2$. Then it holds that*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n}(\boldsymbol{w}_i-\mathbf{e}_i)\right\|^2\right]=\sum_{i=1}^{n}\mathbb{E}\left[\left\|\boldsymbol{w}_i-\mathbf{e}_i\right\|^2\right]\leq n\sigma^2\tag{42}$$

Proof. For details, please refer to [19]'s Lemma 4 and [31]'s Lemma 1.

Lemma 9 (from [19])**.** *The following holds for any L-smooth and μ-strongly convex function h, and any $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}$ in the domain of h:*

$$\langle\nabla h(\boldsymbol{x}),\boldsymbol{z}-\boldsymbol{y}\rangle\geq h(\boldsymbol{z})-h\left(\mathbf{y}\right)+\frac{\mu}{4}\left\|\boldsymbol{y}-\boldsymbol{z}\right\|^2-L\left\|\boldsymbol{z}-\boldsymbol{x}\right\|^2\tag{43}$$

Proof. It can be easily obtained using the three-point identity (24) and Jensen's inequality (22). For details, please refer to [31]. Here, we only use relevant conclusions to assist our proof.

G Proofs of Theorem 1

In this section, we provide the proof of Theorem 1 for the strongly convex, general convex, and non-convex cases in G.1, G.2 and G.3, respectively.

In the following proof, we consider the partial client participation setting. So we assume that $\pi^n = \{\pi_1^n, \pi_2^n, \ldots, \pi_M^n\}$ is a permutation of $\{1, 2, \ldots, M\}$in a certain training round and only the first S selected clients $\{\pi_1^n, \pi_2^n, \ldots, \pi_S^n\}$ will participate in this round. Without otherwise stated, we use E[·] to represent the expectation concerning both types of randomness (i.e., sampling data samples ζ and sampling clients π). $\mathcal{R}$ represents rotation, that is, $\pi^{(n+1)} = \left[\pi_2^{(n+1)}, \pi_3^{(n+1)}, \ldots, \pi_M^{(n+1)}, \pi_1^{(n+1)}\right] = \mathcal{R}\left(\pi^{(n)}\right) := \left[\pi_2^{(n)}, \pi_3^{(n)}, \ldots, \pi_M^{(n)}, \pi_1^{(n)}\right]$,where n represents the client starting with client n.

G.1 Strongly Convex Case

G.1.1 Finding the recursion

Lemma 10. *Let Assumptions 2, 3, and 5 hold, and assume that all the local objectives are μ-strongly convex. If the learning rate satisfies $\eta \leq \frac{1}{6LSK}$, then it holds that*

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{w}^{(r+1)} - \boldsymbol{w}^*\right\|^2\right] \leq \left(1 - \frac{\mu SK\eta}{2}\right)\|\boldsymbol{w} - \boldsymbol{w}^*\|^2 + 4K\eta^2\sigma^2 + 4S^2K^2\eta^2\frac{M-S}{S(M-1)}\zeta_*^2 \\ - \frac{2}{3}SK\eta D_{\mathcal{L}}(\boldsymbol{w}, \boldsymbol{w}^*) + \frac{8}{3}\frac{L\eta}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\boldsymbol{w}_{m,k,n} - \boldsymbol{w}\|^2\right]\end{aligned} \tag{44}$$

Proof. According to the pseudocode of our algorithm, the overall model updates of SFL after one complete training round are shown in (45).However, to compare the subsequent convergence analysis, according to Assumption 1, this deviation can be incorporated into the σ in Assumption 3. For client m, the update method is simplified as (45):

$$\begin{aligned}\boldsymbol{w}^{(r+1)} &= \boldsymbol{w}^{(r)} - \eta\frac{1}{M}\sum_{n=1}^{M}\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{i=1}^{q_{\pi_m^n,k}}\nabla\mathcal{L}_{\pi_m^n,k}^{(r)}\left(f(x_i^{\pi_m^n,k}; \boldsymbol{w}_{\pi_m^n,k}), y_i^{\pi_m^n,k}|\sum_{j=1}^{m-1}\nabla\mathcal{L}_{\pi_j^n,k}\right) \\ &= \boldsymbol{w}^{(r)} - \eta\frac{1}{M}\sum_{n=1}^{M}\sum_{m=1}^{M}\sum_{k=0}^{K-1}\mathbf{g}_{\pi_m^n,k}^{(r)}\end{aligned} \tag{45}$$

More generally, when S represents the number of clients selected to participate, we have

$$\Delta\boldsymbol{w} = \boldsymbol{w}^{(r+1)} - \boldsymbol{w}^{(r)} = -\eta\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbf{g}_{\pi_m^n,k}^{(r)}$$

where $\mathbf{g}_{\pi_m^n,k}^{(r)} = \nabla f_{\pi_m^n}(\boldsymbol{w}_{m,k,n}^{(r)}; \xi)$ is the stochastic gradient of $\mathcal{L}_{\pi_m^n}$ regarding the vector $\boldsymbol{w}_{m,k,n}^{(r)}$. Thus,

$$\mathbb{E}\left[\Delta\boldsymbol{w}\right] = -\eta\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})\right]$$

In the following, we focus on the recurrence of adjacent training rounds, so we omit the superscript r for a while, e.g., writing $\boldsymbol{w}_{m,k,n}^r$ as $\boldsymbol{w}_{m,k,n}$. In particular, we would like to use $\boldsymbol{w}$ to replace $\boldsymbol{w}_{1,0,1}$.Without otherwise stated, the expectation is conditioned on $\boldsymbol{w}^r$.

We start by substituting the overall updates:

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{w}+\Delta\boldsymbol{w}-\boldsymbol{w}^*\|^2\right] \\
&=\|\boldsymbol{w}-\boldsymbol{w}^*\|^2+2\mathbb{E}\left[\langle\boldsymbol{w}-\boldsymbol{w}^*,\Delta\boldsymbol{w}\rangle\right]+\mathbb{E}\left[\|\Delta\boldsymbol{w}\|^2\right] \\
&=\|\boldsymbol{w}-\boldsymbol{w}^*\|^2-2\eta\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\langle\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n}),\boldsymbol{w}-\boldsymbol{w}^*\rangle\right]+\eta^2\mathbb{E}\left[\left\|\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbf{g}_{\pi_m^n,k}\right\|^2\right]
\end{aligned} \tag{46}$$

We can apply Lemma.9 with $\boldsymbol{x}=\boldsymbol{w}_{m,k,n}$, $\boldsymbol{y}=\boldsymbol{w}^*$, $\boldsymbol{z}=\boldsymbol{w}$ and $h=\mathcal{L}_{\pi_m^n}$ for the second term on the right-hand side in (46):

$$\begin{aligned}
&-2\eta\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\langle\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n}),\boldsymbol{w}-\boldsymbol{w}^*\rangle\right] \\
&\leq-2\eta\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\mathcal{L}_{\pi_m^n}(\boldsymbol{w})-\mathcal{L}_{\pi_m^n}(\boldsymbol{w}^*)+\frac{\mu}{4}\|\boldsymbol{w}-\boldsymbol{w}^*\|^2-L\|\boldsymbol{w}_{m,k,n}-\boldsymbol{w}\|^2\right] \\
&\leq-2SK\eta D_{\mathcal{L}}(\boldsymbol{w},\boldsymbol{w}^*)-\frac{1}{2}\mu SK\eta\|\boldsymbol{w}-\boldsymbol{w}^*\|^2+2L\eta\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\boldsymbol{w}_{m,k,n}-\boldsymbol{w}\|^2\right]
\end{aligned} \tag{47}$$

For the third term on the right-hand side in (47), using Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbf{g}_{\pi_m^n,k}\right\|^2\right] \\
&\leq 4\mathbb{E}\left[\left\|\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\left(\mathbf{g}_{\pi_m^n,k}-\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})\right)\right\|^2\right] \\
&+4\mathbb{E}\left[\left\|\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\left(\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})-\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w})\right)\right\|^2\right] \\
&+4\mathbb{E}\left[\left\|\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\left(\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w})-\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}^*)\right)\right\|^2\right]+4\mathbb{E}\left[\left\|\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}^*)\right\|^2\right]
\end{aligned} \tag{48}$$

Seeing the data sample $\xi_{m,k,n}$, the stochastic gradient $\mathbf{g}_{\pi_m^n,k}$, the gradient $\nabla\mathcal{L}_{\pi_m^n}(\xi_{m,k,n})$ as ξ_i,$\boldsymbol{w}_i$ and $\mathbf{e}_i$ in Lemma.6 respectively and applying the result of Lemma6, the first term on the right-hand side in (48) can be bounded by $4K\sigma^2$:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\left(\mathbf{g}_{\pi_m^n,k}-\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})\right)\right\|^2\right] \\
&\overset{(23)}{\leq}4\frac{1}{S^2}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}||\mathbb{E}\left(\mathbf{g}_{\pi_m^n,k}-\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})\right)||^2 \\
&\leq 4K\sigma^2
\end{aligned} \tag{49}$$

For the second term on the right-hand side in (48), we have

$$\begin{aligned}
&4\mathbb{E}\left[\left\|\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\left(\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})-\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w})\right)\right\|^2\right]\\
&\overset{(23)}{\leq} 4S^2K\frac{1}{S^2}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})-\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w})\right\|^2\right]\\
&\overset{\text{Asm.2}}{\leq} 4L^2K\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\boldsymbol{w}_{m,k,n}-\boldsymbol{w}\right\|^2\right]
\end{aligned}\tag{50}$$

For the third term on the right-hand side in 48, we have

$$\begin{aligned}
&4\mathbb{E}\left[\left\|\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\left(\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w})-\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}^*)\right)\right\|^2\right]\\
&\overset{(23)}{\leq} 4S^2K\frac{1}{S^2}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w})-\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}^*)\right\|^2\right]\\
&\overset{(29)}{\leq} 8LK\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[D_{\mathcal{L}_{\pi_m^n}}(\boldsymbol{w},\boldsymbol{w}^*)\right]\\
&\overset{(26)}{=} 8LS^2K^2D_{\mathcal{L}}(\boldsymbol{w},\boldsymbol{w}^*),
\end{aligned}\tag{51}$$

We explain it as follows, because $D_{\mathcal{L}}(\boldsymbol{w},\boldsymbol{w}^*)$ is linear concerning $\mathcal{L}(\boldsymbol{w})$, so it satisfies the formula:

$$\mathbb{E}\left[D_{\mathcal{L}_{\pi_m}}(\boldsymbol{w},\boldsymbol{w}^*)\right]=D_{\mathcal{L}}(\boldsymbol{w},\boldsymbol{w}^*)$$

The fourth term on the right hand side in 48 can be bounded by Lemma.6 as follows:

$$4\mathbb{E}\left[\left\|\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}^*)\right\|^2\right]\overset{(Lem.6)}{\leq}4S^2K^2\frac{M-S}{S(M-1)}\zeta_*^2\tag{52}$$

With the preceding four inequalities, we can bound the third term on the right hand side in (48):

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbf{g}_{\pi_m^n,k}\right\|^2\right]\\
&\leq 4K\sigma^2+4L^2K\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\boldsymbol{w}_{m,k,n}-\boldsymbol{w}\right\|^2\right]\\
&\quad+8LS^2K^2D_{\mathcal{L}}(\boldsymbol{w},\boldsymbol{w}^*)+4S^2K^2\frac{M-S}{S(M-1)}\zeta_*^2
\end{aligned}\tag{53}$$

Then substituting (47) and (53) into (46), we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{w}+\Delta\boldsymbol{w}-\boldsymbol{w}^*\right\|^2\right]\leq&\left(1-\frac{\mu SK\eta}{2}\right)\left\|\boldsymbol{w}-\boldsymbol{w}^*\right\|^2+4K\eta^2\sigma^2+4S^2K^2\eta^2\frac{M-S}{S(M-1)}\zeta_*^2\\
&+2\frac{L\eta}{S}(1+2LSK\eta)\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\boldsymbol{w}_{m,k,n}-\boldsymbol{w}\right\|^2\right]\\
&-2SK\eta(1-4LSK\eta)D_{\mathcal{L}}(\boldsymbol{w},\boldsymbol{w}^*)
\end{aligned}$$

Here we substitute $\eta\leq\frac{1}{6LSK}$ to get Lemma.10, as follows:

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{w}+\Delta\boldsymbol{w}-\boldsymbol{w}^*\right\|^2\right]\leq&\left(1-\frac{\mu SK\eta}{2}\right)\left\|\boldsymbol{w}-\boldsymbol{w}^*\right\|^2+4K\eta^2\sigma^2+4S^2K^2\eta^2\frac{M-S}{S(M-1)}\zeta_*^2\\
&-\frac{2}{3}SK\eta D_{\mathcal{L}}(\boldsymbol{w},\boldsymbol{w}^*)+\frac{8}{3}\frac{L\eta}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\boldsymbol{w}_{m,k,n}-\boldsymbol{w}\right\|^2\right]
\end{aligned}$$

G.1.2 Bounding the client drift with Assumption. 5

Similar to the "client drift" in PFL [19] and SFL [31], we define the client drift in SPFL:

$$E_r := \frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\boldsymbol{w}_{m,k,n}^{(r)}-\boldsymbol{w}^{(r)}\right\|^2\right] \tag{54}$$

Lemma 11. *When Assumptions. 2, 3 and 5 hold, and assuming that all local objective functions are μ-strongly convex, then if the learning rate satisfies $\eta \leq \frac{1}{6LSK}$, the client drift is bounded:*

$$E_r \leq \frac{9}{4}S^2K^2\eta^2\sigma^2 + \frac{9}{4}S^2K^3\eta^2\zeta^2 + 3LS^3K^3\eta^2\mathbb{E}\left[D_{\mathcal{L}}(\boldsymbol{w}^{(r)},\boldsymbol{w}^*)\right] \tag{55}$$

Proof. According to our SPFL pseudo code, we can get the model updates of SPFL from $\boldsymbol{w}^{(r)}$ to $\boldsymbol{w}_{m,k,n}^{(r)}$ is

$$\boldsymbol{w}_{m,k,n}^{(r)} - \boldsymbol{w}^{(r)} = -\eta\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\mathbf{g}_{\pi_i^n,j}^{(r)}$$

with $b_{m,k}(i) := \begin{cases} K-1, & i \leq m-1 \\ k-1, & i=m \end{cases}$. In the following, we focus on a single training round, and hence we drop the superscript r for a while, e.g., writing $\boldsymbol{w}_{m,k,n}$ to replace $\boldsymbol{w}_{m,k,n}^{(r)}$.In particular, we would like to use $\boldsymbol{w}$ to replace $\boldsymbol{w}_{1,0,1}$.Without otherwise stated, the expectation is conditioned on $\boldsymbol{w}^r$. We use Jensen's inequality to bound the term $\mathbb{E}\left[\|\boldsymbol{w}_{m,k,n}-\boldsymbol{w}\|^2\right] = \eta^2\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\mathbf{g}_{\pi_i^n,j}\right\|^2\right]$:

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{w}_{m,k,n}-\boldsymbol{w}\|^2\right] \\
&\leq 4\eta^2\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\left(\mathbf{g}_{\pi_i^n,j}-\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w}_{i,j,n})\right)\right\|^2\right] \\
&+ 4\eta^2\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\left(\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w}_{i,j,n})-\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w})\right)\right\|^2\right] \\
&+ 4\eta^2\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\left(\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w})-\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w}^*)\right)\right\|^2\right] + 4\eta^2\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w}^*)\right\|^2\right]
\end{aligned}$$

Applying Lemma.8 to the first term and Jensen's inequality to the second, third terms on the right hand side in the preceding inequality, respectively, we can get

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{w}_{m,k,n}-\boldsymbol{w}\|^2\right] \\
&\leq 4\eta^2\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\mathbb{E}\left[\left\|\mathbf{g}_{\pi_i^n,j}-\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w}_{i,j,n})\right\|^2\right] \\
&+ 4\eta^2\mathcal{C}_{m,k}\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w}_{i,j,n})-\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w})\right\|^2\right] \\
&+ 4\eta^2\mathcal{C}_{m,k}\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\mathbb{E}\left[\|\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w})-\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w}^*)\|^2\right] + 4\eta^2\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w}^*)\right\|^2\right]
\end{aligned} \tag{56}$$

where $\mathcal{C}_{m,k} := \sum_{i=1}^{m} \sum_{i=0}^{b_{m,k}(i)} 1 = (m-1)K + k$. The first term on the right-hand side in (56) is bounded by $4\mathcal{C}_{m,k}\eta^2\sigma^2$. For the second term on the right-hand side in (56), we have

$$\mathbb{E}\left[\|\nabla \mathcal{L}_{\pi_i^n}(\boldsymbol{w}_{i,j,n}) - \nabla \mathcal{L}_{\pi_i^n}(\boldsymbol{w})\|^2\right] \overset{\text{Asm.2}}{\leq} L^2 \mathbb{E}\left[\|\boldsymbol{w}_{i,j,n} - \boldsymbol{w}\|^2\right]$$

For the third term on the right-hand side in (56), we have

$$\mathbb{E}\left[\|\nabla \mathcal{L}_{\pi_i^n}(\boldsymbol{w}) - \nabla \mathcal{L}_{\pi_i^n}(\boldsymbol{w}^*)\|^2\right] \overset{(23)}{\leq} 2L\mathbb{E}\left[D_{\mathcal{L}_{\pi_i^n}}(\boldsymbol{w}, \boldsymbol{w}^*)\right] = 2LD_{\mathcal{L}}(\boldsymbol{w}, \boldsymbol{w}^*)$$

Since $\boldsymbol{w}^*$ is the optimal solution, its gradient $\nabla \mathcal{L}(\boldsymbol{w}^*) = 0$. As a result, we can get

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{w}_{m,k,n} - \boldsymbol{w}\|^2\right] \leq 4\mathcal{C}_{m,k}\eta^2\sigma^2 + 4L^2\eta^2\mathcal{C}_{m,k}\sum_{i=1}^{m}\sum_{j=0}^{b(i)}\mathbb{E}\left[\|\boldsymbol{w}_{i,j,n} - \boldsymbol{w}\|^2\right] + 8L\eta^2\mathcal{C}_{m,k}^2 D_{\mathcal{L}}(\boldsymbol{w}, \boldsymbol{w}^*) \\ + 4\eta^2\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\nabla \mathcal{L}_{\pi_i^n}(\boldsymbol{w}^*)\right\|^2\right]\end{aligned} \tag{57}$$

Then, returning to $E_r := \frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\boldsymbol{w}_{m,k,n} - \boldsymbol{w}\|^2\right]$, we have

$$\begin{aligned}E_r \leq 4\eta^2\sigma^2\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathcal{C}_{m,k} + 4L^2\eta^2\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathcal{C}_{m,k}\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\mathcal{E}\left[\|\boldsymbol{w}_{i,j,n} - \boldsymbol{w}\|^2\right] \\ + 8L\eta^2\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathcal{C}_{m,k}^2 D_{\mathcal{L}}(\boldsymbol{w}, \boldsymbol{w}^*) + 4\eta^2\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\nabla \mathcal{L}_{\pi_i^n}(\boldsymbol{w}^*)\right\|^2\right]\end{aligned}$$

Applying Lemma.7 with $\boldsymbol{w}_{\pi_i^n} = \nabla \mathcal{L}_{\pi_i^n}(\boldsymbol{w}^*)$ and $\overline{w} = \nabla \mathcal{L}(\boldsymbol{w}^*) = 0$ and Lemma.1 that

$$\begin{aligned}\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathcal{C}_{m,k} &= \frac{1}{2}SK(SK-1) \leq \frac{1}{2}S^2K^2, \\ \frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathcal{C}_{m,k}^2 &= \frac{1}{3}(SK-1)SK(SK-\frac{1}{2}) \leq \frac{1}{3}S^3K^3\end{aligned}$$

we can simplify the preceding inequality:

$$E_r \leq 2S^2K^2\eta^2\sigma^2 + 2L^2S^2K^2\eta^2E_r + \frac{8}{3}LS^3K^3\eta^2D_{\mathcal{L}}(\boldsymbol{w}, \boldsymbol{w}^*) + 2S^2K^3\eta^2\zeta_*^2$$

After rearranging the preceding inequality, we get

$$(1 - 2L^2S^2K^2\eta^2)E_r \leq 2S^2K^2\eta^2\sigma^2 + 2S^2K^3\eta^2\zeta_*^2 + \frac{8}{3}LS^3K^3\eta^2D_{\mathcal{L}}(\boldsymbol{w}, \boldsymbol{w}^*)$$

Finally, using the condition that $\eta \leq \frac{1}{6LSK}$, which implies $1 - 2L^2S^2K^2\eta^2 \geq \frac{8}{9}$, we have

$$E_r \leq \frac{9}{4}S^2K^2\eta^2\sigma^2 + \frac{9}{4}S^2K^3\eta^2\zeta_*^2 + 3LS^3K^3\eta^2D_{\mathcal{L}}(\boldsymbol{w}, \boldsymbol{w}^*)$$

The claim follows after recovering the superscripts and taking unconditional expectations.

G.1.3 Proof of strongly convex case of Theorem 1

Proof of the strongly convex case of Theorem 1. Substituting Lemma .11 into Lemma .10 and using $\eta \leq \frac{1}{6LSK}$, we can simplify the recursion as follows:

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{w}^{(r+1)}-\boldsymbol{w}^{*}\right\|^{2}\right] &\leq\left(1-\frac{\mu S K \eta}{2}\right)\left\|\boldsymbol{w}-\boldsymbol{w}^{*}\right\|^{2}+4 K \eta^{2} \sigma^{2}+4 S^{2} K^{2} \eta^{2} \frac{M-S}{S(M-1)} \zeta_{*}^{2} \\ &-\frac{2}{3} S K \eta D_{\mathcal{L}}\left(\boldsymbol{w}, \boldsymbol{w}^{*}\right)+\frac{8}{3} L \eta E_{r} \\ &\leq\left(1-\frac{\mu S K \eta}{2}\right) \mathbb{E}\left[\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{*}\right\|^{2}\right]-\frac{1}{3} S K \eta \mathbb{E}\left[D_{\mathcal{L}}\left(\boldsymbol{w}^{(r)}, \boldsymbol{w}^{*}\right)\right] \\ &+4 K \eta^{2} \sigma^{2}+4 S^{2} K^{2} \eta^{2} \frac{M-S}{S(M-1)} \zeta_{*}^{2}+6 L S^{2} K^{2} \eta^{3} \sigma^{2}+6 L S^{2} K^{3} \eta^{3} \zeta_{*}^{2}\end{aligned}$$

Let $\tilde{\eta}=MK\eta$, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{w}^{(r+1)}-\boldsymbol{w}^{*}\right\|^{2}\right] \leq\left(1-\frac{\mu \tilde{\eta}}{2}\right) \mathbb{E}\left[\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{*}\right\|^{2}\right]-\frac{\tilde{\eta}}{3} \mathbb{E}\left[D_{\mathcal{L}}(\boldsymbol{w}^{(r)}, \boldsymbol{w}^{*})\right] \\ +\frac{4 \tilde{\eta}^{2} \sigma^{2}}{S^{2} K}+\frac{4 \tilde{\eta}^{2}(M-S) \zeta_{*}^{2}}{S(M-1)}+\frac{6 L \tilde{\eta}^{3} \sigma^{2}}{S K}+\frac{6 L \tilde{\eta}^{3} \zeta_{*}^{2}}{S}\end{aligned} \tag{58}$$

Applying Lemma 3 with $t=r(T=R), \gamma=\tilde{\eta}, r_t=\mathbb{E}\left[\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{*}\right\|^{2}\right], a=\frac{\mu}{2}, b=\frac{1}{3}$, $s_t=\mathbb{E}\left[D_{\mathcal{L}}(\boldsymbol{w}^{(r)}, \boldsymbol{w}^{*})\right], \theta_t=(1-\frac{\mu\tilde{\eta}}{2})^{-(r+1)}, c_1=\frac{4\sigma^2}{S^2K}+\frac{4(M-S)\zeta_*^2}{S(M-1)}, c_2=\frac{6L\sigma^2}{SK}+\frac{6L\zeta_*^2}{SK}$ and $\frac{1}{d}=\frac{1}{6L}(\tilde{\eta}=MK\eta\leq\frac{1}{6L})$, it follows that

$$\begin{aligned}&\mathbb{E}\left[\mathcal{L}(\bar{\boldsymbol{w}}^{(R)})-\mathcal{L}(\boldsymbol{w}^{*})\right] \leq \frac{1}{W_R} \sum_{r=0}^{R} \theta_r \mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^{(r)})-\mathcal{L}(\boldsymbol{w}^{*})\right] \\ &\leq \frac{9}{2} \mu\left\|\boldsymbol{w}^{(0)}-\boldsymbol{w}^{*}\right\|^{2} \exp\left(-\frac{1}{2} \mu \tilde{\eta} R\right)+\frac{12 \tilde{\eta} \sigma^{2}}{S^{2} K}+\frac{12 \tilde{\eta}(M-S) \zeta_{*}^{2}}{S(M-1)}+\frac{18 L \tilde{\eta}^{2} \sigma^{2}}{S K}+\frac{18 L \tilde{\eta}^{2} \zeta_{*}^{2}}{S K}\end{aligned} \tag{59}$$

where $\bar{\boldsymbol{w}}^{(R)}=\frac{1}{W_R}\sum_{r=0}^{R}\theta_r\boldsymbol{w}^{(r)}$ and we use Jensen's inequality ($\mathcal{L}$ is convex) in the first inequality. Applying Lemma .3 to Eq.59 and using a suitable dynamic learning rate yields:

$$\begin{aligned}&\mathbb{E}\left[\mathcal{L}(\bar{\boldsymbol{w}}^{(R)})-\mathcal{L}(\boldsymbol{w}^{*})\right] \\ &=\tilde{\mathcal{O}}\left(\mu \mathcal{A}^{2} \exp\left(-\frac{\mu R}{12 L}\right)+\frac{\sigma^{2}}{\mu S^{2} K R}+\frac{(M-S) \zeta_{*}^{2}}{\mu S R(M-1)}+\frac{L \sigma^{2}}{\mu^{2} S K R^{2}}+\frac{L \zeta_{*}^{2}}{\mu^{2} S R^{2}}\right)\end{aligned} \tag{60}$$

where $\mathcal{A}:=\left\|\boldsymbol{w}^{(0)}-\boldsymbol{w}^{*}\right\|$. Eq. (59) and Eq . 60 are the upper bounds with partial client participation. In particular, when S = M, we can get the claim of the strongly convex case of Theorem 1.

G.2 General Convex Case

G.2.1 Proof of general convex case of Theorem 1

Proof of the general convex case of Theorem 1. Letting $\mu=0$ in Eq. (58), we get the recursion of the general convex case,

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{w}^{(r+1)}-\boldsymbol{w}^{*}\right\|^{2}\right] \leq \mathbb{E}\left[\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{*}\right\|^{2}\right]-\frac{\tilde{\eta}}{3} \mathbb{E}\left[D_{\mathcal{L}}(\boldsymbol{w}^{(r)}, \boldsymbol{w}^{*})\right] \\ +\frac{4 \tilde{\eta}^{2} \sigma^{2}}{S^{2} K}+\frac{4 \tilde{\eta}^{2}(M-S) \zeta_{*}^{2}}{S(M-1)}+\frac{6 L \tilde{\eta}^{3} \sigma^{2}}{S K}+\frac{6 L \tilde{\eta}^{3} \zeta_{*}^{2}}{S}\end{aligned}$$

Applying Lemma.5 with $t=r(T=R), \gamma=\tilde{\eta}, r_t=\mathbb{E}\left[\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}^{*}\right\|^{2}\right], a=0, b=\frac{1}{3}$, $s_t=\mathbb{E}\left[D_{\mathcal{L}}(\boldsymbol{w}^{(r)}, \boldsymbol{w}^{*})\right], \theta_t=(1-\frac{\mu\tilde{\eta}}{2})^{-(r+1)}, c_1=\frac{4\sigma^2}{S^2K}+\frac{4(M-S)\zeta_*^2}{S(M-1)}, c_2=\frac{6L\sigma^2}{SK}+\frac{6L\zeta_*^2}{SK}$ and $\frac{1}{d}=$

$\frac{1}{6L}(\tilde{\eta} = MK\eta \leq \frac{1}{6L})$,it follows that

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\bar{\boldsymbol{w}}^{(R)}) - \mathcal{L}(\boldsymbol{w}^*)\right] &\leq \frac{1}{W_R}\sum_{r=0}^{R}\theta_r\left(\mathcal{L}(\boldsymbol{w}^{(r)}) - \mathcal{L}(\boldsymbol{w}^*)\right) \\ &\leq \frac{3\left\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\right\|^2}{\tilde{\eta}R} + \frac{12\tilde{\eta}\sigma^2}{S^2K} + \frac{12\tilde{\eta}(M-S)\zeta_*^2}{S(M-1)} + \frac{18L\tilde{\eta}^2\sigma^2}{SK} + \frac{18L\tilde{\eta}^2\zeta_*^2}{SK}\end{aligned} \tag{61}$$

where $\bar{\boldsymbol{w}}^{(R)} = \frac{1}{W_R}\sum_{r=0}^{R}\theta_r\boldsymbol{w}^{(r)}$ and we use Jensen's inequality ($\mathcal{L}$ is convex) in the first inequality. By using a suitable dynamic learning rate, we get

$$\begin{aligned}\mathcal{L}(\bar{\boldsymbol{w}}^{(R)}) - \mathcal{L}(\boldsymbol{w}^*) \leq& 2\left(\frac{4\sigma^2}{S^2K} + \frac{4(M-S)\zeta_*^2}{S(M-1)}\right)^{\frac{1}{2}}\left(\frac{\mathbb{E}\left[\left\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\right\|^2\right]}{R+1}\right)^{\frac{1}{2}} \\ &+ 2\left(\frac{6L\sigma^2}{SK} + \frac{6L\zeta_*^2}{SK}\right)^{\frac{1}{3}}\left(\frac{\mathbb{E}\left[\left\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\right\|^2\right]}{R+1}\right)^{\frac{2}{3}} + \frac{6L\mathbb{E}\left[\left\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\right\|^2\right]}{R+1}\end{aligned}$$

Due to the concave nature of the power function(Lemma.2), we can get

$$\begin{aligned}\mathcal{L}(\bar{\boldsymbol{w}}^{(R)}) - \mathcal{L}(\boldsymbol{w}^*) \leq& 2\left[\left(\frac{4\sigma^2}{S^2K}\right) + \left(\frac{4(M-S)\zeta_*^2}{S(M-1)}\right)\right]^{\frac{1}{2}}\left(\frac{\mathbb{E}\left[\left\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\right\|^2\right]}{R+1}\right)^{\frac{1}{2}} \\ &+ 2\left[\left(\frac{6L\sigma^2}{SK}\right) + \left(\frac{6L\zeta_*^2}{SK}\right)\right]^{\frac{1}{3}}\left(\frac{\mathbb{E}\left[\left\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\right\|^2\right]}{R+1}\right)^{\frac{2}{3}} + \frac{6L\mathbb{E}\left[\left\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\right\|^2\right]}{R+1}\end{aligned}$$

After finishing, we can get:

$$\mathcal{L}(\bar{\boldsymbol{w}}^{(R)}) - \mathcal{L}(\boldsymbol{w}^*) = \mathcal{O}\left(\frac{\sigma\mathcal{A}}{\sqrt{S^2KR}} + \sqrt{1-\frac{S}{M}}\cdot\frac{\zeta_*\mathcal{A}}{\sqrt{SR}} + \frac{\left(L\sigma^2\mathcal{A}^4\right)^{1/3}}{(SK)^{1/3}R^{2/3}} + \frac{\left(L\zeta_*^2\mathcal{A}^4\right)^{1/3}}{S^{1/3}R^{2/3}} + \frac{L\mathcal{A}^2}{R}\right) \tag{62}$$

where $\mathcal{A} := \left\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\right\|$. Eq. (61) and Eq. (62) are the upper bounds with partial client participation. In particular, when $S = M$, we can claim the strongly convex case of Theorem 1 and Corollary 1.

G.3 Nonconvex Case

Lemma 12. *Let Assumptions 2,3 and 4 hold. If the learning rate satisfies $\eta = \frac{1}{6LSK}$, then it holds that*

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^{(r+1)}) - \mathcal{L}(\boldsymbol{w}^{(r)})\right] \leq& -\frac{SK\eta}{2}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}^{(r)})\right\|^2\right] + LSK\eta^2\sigma^2 \\ &+ \frac{L^2\eta}{2}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\boldsymbol{w}_{m,k}^{(r)} - \boldsymbol{w}^{(r)}\right\|^2\right]\end{aligned} \tag{63}$$

Proof. According to the Pseudocode of SPFL, the overall model updates of SPFL after one complete training round (with S clients selected for training) is

$$\Delta\boldsymbol{w} = \boldsymbol{w}^{(r+1)} - \boldsymbol{w}^{(r)} = -\eta\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbf{g}_{\pi_m^n,k}^{(r)}$$

where $\mathbf{g}_{\pi_m^n,k}^{(r)} = \nabla f_{\pi_m^n}(\boldsymbol{w}_{m,k,n}^{(r)};\xi)$ is the stochastic gradient of $\mathcal{L}_{\pi_m^n}$ regarding the vector $\boldsymbol{w}_{m,k,n}^{(r)}$. Thus,

$$\mathbb{E}\left[\Delta\boldsymbol{w}\right] = -\eta\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})\right]$$

In the following, we focus on the recurrence of adjacent training rounds, so we omit the superscript r for a while, e.g., writing $\boldsymbol{w}_{m,k,n}^r$ as $\boldsymbol{w}_{m,k,n}$. In particular, we would like to use $\boldsymbol{w}$ to replace $\boldsymbol{w}_{1,0,1}$.Without otherwise stated, the expectation is conditioned on $\boldsymbol{w}^r$.

Starting from the smoothness of F (applying Eq. (53), $D_{\mathcal{L}}(\boldsymbol{x},\boldsymbol{y}) \leq \frac{L}{2}\left\|\boldsymbol{x}-\boldsymbol{y}\right\|^2$ with $x=\boldsymbol{w}+\Delta\boldsymbol{w}$, $\boldsymbol{y}=\boldsymbol{w}$,and substituting the overall updates, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}+\Delta\boldsymbol{w})-\mathcal{L}(\boldsymbol{w})\right] \\
&\leq \mathbb{E}\left[\langle\nabla\mathcal{L}(\boldsymbol{w}),\Delta\boldsymbol{w}\rangle\right]+\frac{L}{2}\mathbb{E}\left[\left\|\Delta\boldsymbol{w}\right\|^2\right] \\
&\leq -\eta\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\langle\nabla\mathcal{L}(\boldsymbol{w}),\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})\rangle\right]+\frac{L\eta^2}{2}\mathbb{E}\left[\left\|\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbf{g}_{\pi_m^n,k}\right\|^2\right]
\end{aligned} \tag{64}$$

For the first term on the right-hand side in Eq.(64), using the fact that $2\left\langle a,b\right\rangle=\left\|a\right\|^2+\left\|b\right\|^2-\left\|a-b\right\|^2$ with $a=\nabla\mathcal{L}(\boldsymbol{w})$ and $b=\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})$, we have

$$\begin{aligned}
&-\eta\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\langle\nabla\mathcal{L}(\boldsymbol{w}),\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})\rangle\right] \\
&=-\frac{\eta}{2}\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w})\right\|^2+\left\|\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})\right\|^2-\left\|\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})-\nabla\mathcal{L}(\boldsymbol{w})\right\|^2\right] \\
&\overset{\text{Asm.2}}{\leq}-\frac{SK\eta}{2}\left\|\nabla\mathcal{L}(\boldsymbol{w})\right\|^2-\frac{\eta}{2}\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})\right\|^2\right] \\
&\quad+\frac{L^2\eta}{2}\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\boldsymbol{w}_{m,k,n}-\boldsymbol{w}\right\|^2\right]
\end{aligned} \tag{65}$$

For the third term on the right hand side in Eq. (64), using Jensen's inequality, we have

$$\begin{aligned}
&\frac{L\eta^2}{2}\mathbb{E}\left[\left\|\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbf{g}_{\pi_m,k}\right\|^2\right] \\
&\leq L\eta^2\mathbb{E}\left[\left\|\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbf{g}_{\pi_m^n,k}-\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})\right\|^2\right] \\
&\quad+L\eta^2\mathbb{E}\left[\left\|\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})\right\|^2\right] \\
&\leq LK\eta^2\sigma^2+LSK\eta^2\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\pi_m}(\boldsymbol{w}_{m,k})\right\|^2\right],
\end{aligned} \tag{66}$$

where we apply Lemma .6 by seeing the data sample $\xi_{m,k,n}$, the stochastic gradient $\mathbf{g}_{\pi_m^n,k}$, the gradient $\nabla\mathcal{L}_{\pi_m^n}(\xi_{m,k,n})$ as $\xi_i,\boldsymbol{w}_i,\mathbf{e}_i$ respectively, in Lemma 6 for the first term and Jensen's inequality for the second term in the preceding inequality. Substituting Eq. (65) and Eq. (66) into Eq. (64), we have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}+\Delta\boldsymbol{w})-\mathcal{L}(\boldsymbol{w})\right]\leq&-\frac{SK\eta}{2}\left\|\nabla\mathcal{L}(\boldsymbol{w})\right\|^2+LK\eta^2\sigma^2+\frac{L^2\eta}{2}\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\boldsymbol{w}_{m,k,n}-\boldsymbol{w}\right\|^2\right] \\
&-\frac{\eta}{2}(1-2LSK\eta)\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\pi_m^n}(\boldsymbol{w}_{m,k,n})\right\|^2\right]
\end{aligned}$$

Since $\eta \leq \frac{1}{6LSK}$, the last term on the right-hand side in the preceding inequality is negative. Then

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}+\Delta\boldsymbol{w})-\mathcal{L}(\boldsymbol{w})\right] \leq -\frac{SK\eta}{2}\left\|\nabla\mathcal{L}(\boldsymbol{w})\right\|^2+LK\eta^2\sigma^2+\frac{L^2\eta}{2}\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\boldsymbol{w}_{m,k,n}-\boldsymbol{w}\right\|^2\right]$$

This conclusion can be obtained after restoring the superscript and taking the unconditional expectation.

G.3.1 Bounding the client drift with Assumption 4

Since the proof of Lemma. 11 uses Eq. (28), which is only applicable to convex functions; we cannot use the result of Lemma. 11. Next, we use Assumption 4 to bound the client drift (defined in Eq. (54)).

Lemma 13. *Assumptions 2, 3, and 4 hold. If the learning rate satisfies $\eta \leq \frac{1}{6LSK}$, the client drift is bounded*

$$E_r \leq \frac{9}{4}S^2K^2\eta^2\sigma^2+\frac{9}{4}S^2K^3\eta^2\zeta^2+\left(\frac{9}{4}\beta^2S^2K^3\eta^2+\frac{3}{2}S^3K^3\eta^2\right)\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}^{(r)})\right\|^2\right] \tag{67}$$

Proof. Similar to the "client drift" in PFL [19] and SFL [31], we define the client drift in SPFL:

$$E_r := \frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\boldsymbol{w}_{m,k,n}^{(r)}-\boldsymbol{w}^{(r)}\right\|^2\right] \tag{68}$$

with $b_{m,k}(i) := \begin{cases} K-1, & i \leq m-1 \\ k-1, & i=m \end{cases}$. In the following, we focus on a single training round, and hence we drop the superscript r for a while, e.g., writing $\boldsymbol{w}_{m,k,n}$ to replace $\boldsymbol{w}_{m,k,n}^{(r)}$.In particular, we would like to use $\boldsymbol{w}$ to replace $\boldsymbol{w}_{1,0,1}$.Without otherwise stated, the expectation is conditioned on $\boldsymbol{w}^r$. We use Jensen's inequality to bound the term $\mathbb{E}\left[\left\|\boldsymbol{w}_{m,k,n}-\boldsymbol{w}\right\|^2\right] = \eta^2\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\mathbf{g}_{\pi_i^n,j}\right\|^2\right]$:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{w}_{m,k,n}-\boldsymbol{w}\right\|^2\right] \\
&\leq 4\eta^2\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\left(\mathbf{g}_{\pi_i^n,j}-\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w}_{i,j,n})\right)\right\|^2\right] \\
&+ 4\eta^2\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\left(\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w}_{i,j,n})-\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w})\right)\right\|^2\right] \\
&+ \underbrace{4\eta^2\,\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\left(\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w})-\nabla\mathcal{L}(\boldsymbol{w})\right)\right\|^2\right]}_{T_1} + 4\eta^2\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\nabla\mathcal{L}(\boldsymbol{w})\right\|^2\right]
\end{aligned}$$

Applying Lemma. 8, Jensen's inequality and Jensen's inequality to the first, third, and fourth terms on the right side of the previous inequality, respectively, we can get

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{w}_{m,k,n}-\boldsymbol{w}\right\|^2\right] \\
&\leq 4\eta^2\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\mathbb{E}\left[\left\|\mathbf{g}_{\pi_i^n,j}-\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w}_{i,j,n})\right\|^2\right] \\
&+ 4\eta^2\mathcal{C}_{m,k}\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w}_{i,j,n})-\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w})\right\|^2\right] + 4\eta^2T_1 + 4\mathcal{C}_{m,k}^2\eta^2\left\|\nabla\mathcal{L}(\boldsymbol{w})\right\|^2
\end{aligned} \tag{69}$$

where $\mathcal{C}_{m,k} := \sum_{i=1}^{m} \sum_{i=0}^{b_{m,k}(i)} 1 = (m-1)K + k$. The first term on the right-hand side in (69) is bounded by $4\mathcal{C}_{m,k}\eta^2\sigma^2$ with Assumption 3. With Assumption 2, the second term on the right-hand side in (56) can be defined as

$$4\eta^2\mathcal{C}_{m,k}\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w}_{i,j,n}) - \nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w})\right\|^2\right] \leq 4L^2\eta^2\mathcal{C}_{m,k}\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\mathbb{E}\left[\|\boldsymbol{w}_{i,j,n} - \boldsymbol{w}\|^2\right]$$

Then, returning to $E_r := \frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\boldsymbol{w}_{m,k,n} - \boldsymbol{w}\|^2\right]$,we have

$$\begin{aligned}
E_r \leq\ & 4\eta^2\sigma^2\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathcal{C}_{m,k} + 4L^2\eta^2\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathcal{C}_{m,k}\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\mathbb{E}\left[\|\boldsymbol{w}_{i,j,n} - \boldsymbol{w}\|^2\right] \\
& + 4\eta^2\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sum_{j=0}^{b_{m,k}(i)}\left(\nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w}) - \nabla\mathcal{L}(\boldsymbol{w})\right)\right\|^2\right] \\
& + 4\eta^2\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathcal{C}_{m,k}^2\left\|\nabla\mathcal{L}(\boldsymbol{w})\right\|^2
\end{aligned} \tag{70}$$

Applying Lemma .7 with $\boldsymbol{w}_{\pi_i^n} = \nabla\mathcal{L}_{\pi_i^n}(\boldsymbol{w})$ and $\bar{x} = \nabla\mathcal{L}(\boldsymbol{w})$ to the third term and $\frac{1}{S}\sum_{n=1}^{M}\sum_{m=1}^{M}\sum_{k=0}^{K-1}\mathcal{C}_{m,k} \leq \frac{1}{2}S^2K^2$ and $\frac{1}{S}\sum_{n=1}^{S}\sum_{m=1}^{S}\sum_{k=0}^{K-1}\mathcal{C}_{m,k}^2 \leq \frac{1}{3}S^3K^3$ to the other terms on the right Hand side of the preceding inequality, we can simplify it:

$$\begin{aligned}
E_r \leq\ & 2S^2K^2\eta^2\sigma^2 + 2L^2S^2K^2\eta^2E_r + 2S^2K^3\eta^2\left(\frac{1}{M}\sum_{i=1}^{M}\left\|\nabla\mathcal{L}_i(\boldsymbol{w}) - \nabla\mathcal{L}(\boldsymbol{w})\right\|^2\right) \\
& + \frac{4}{3}S^3K^3\eta^2\left\|\nabla\mathcal{L}(\boldsymbol{w})\right\|^2 \\
\overset{Asm.2}{\leq}\ & 2S^2K^2\eta^2\sigma^2 + 2L^2S^2K^2\eta^2E_r + 2S^2K^3\eta^2\zeta^2 + 2\beta^2S^2K^3\eta^2\left\|\nabla\mathcal{L}(\boldsymbol{w})\right\|^2 \\
& + \frac{4}{3}S^3K^3\eta^2\left\|\nabla\mathcal{L}(\boldsymbol{w})\right\|^2
\end{aligned}$$

After rearranging the preceding inequality, we get

$$(1-2L^2S^2K^2\eta^2)E_r \leq 2S^2K^2\eta^2\sigma^2 + 2S^2K^3\eta^2\zeta^2 + 2\beta^2S^2K^3\eta^2\left\|\nabla\mathcal{L}(\boldsymbol{w})\right\|^2 + \frac{4}{3}S^3K^3\eta^2\left\|\nabla\mathcal{L}(\boldsymbol{w})\right\|^2$$

Finally, using the condition that $\eta \leq \frac{1}{6LSK}$, which implies $1 - 2L^2S^2K^2\eta^2 \geq \frac{8}{9}$, we have

$$E_r \leq \frac{9}{4}S^2K^2\eta^2\sigma^2 + \frac{9}{4}S^2K^3\eta^2\zeta^2 + \frac{9}{4}\beta^2S^2K^3\eta^2\left\|\nabla\mathcal{L}(\boldsymbol{w})\right\|^2 + \frac{3}{2}S^3K^3\eta^2\left\|\nabla\mathcal{L}(\boldsymbol{w})\right\|^2$$

The claim follows after recovering the superscripts and taking unconditional expectations.

G.3.2 Proof of nonconvex case of Theorem 1

Proof. Substituting Lemma.12 into Lemma.13 and using $\eta \leq \frac{1}{6LSK}min\{1, \frac{\sqrt{S}}{\beta}\}$ we can simplify the recursion as follows:

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^{(r+1)}) - \mathcal{L}(\boldsymbol{w}^{(r)})\right] \leq -\frac{1}{3}SK\eta\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}^{(r)})\right\|^2\right] + LK\eta^2\sigma^2 + \frac{9}{8}L^2S^2K^2\eta^3\sigma^2 + \frac{9}{8}L^2S^2K^3\eta^3\zeta^2$$

Letting $\tilde{\eta} := SK\eta$ Subtracting $\mathcal{L}^*$ from both sides and rearranging the terms, we have

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^{(r+1)}) - \mathcal{L}^*\right] \leq \mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^{(r)}) - \mathcal{L}^*\right] - \frac{\tilde{\eta}}{3}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}^{(r)})\right\|^2\right] + \frac{L\tilde{\eta}^2\sigma^2}{S^2K} + \frac{9L^2\tilde{\eta}^3\sigma^2}{8SK} + \frac{9L^2\tilde{\eta}^3\zeta^2}{8S}$$

Then applying Lemma .5 with $t = r(T = R), \gamma = \tilde{\eta}, r_t = \mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^{(r)}) - \mathcal{L}^*\right], b = \frac{1}{3}$, $s_t = \mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}^{(r)})\right\|^2\right], \theta_t = 1, c_1 = \frac{L\sigma^2}{S^2K}, c_2 = \frac{9L^2\sigma^2}{8SK} + \frac{9L^2\zeta^2}{8S}$ and $\frac{1}{d} = \frac{1}{6L}\min\left\{1, \frac{\sqrt{S}}{\beta}\right\}$ ($\tilde{\eta} = \frac{1}{6L(\beta+1)} \leq \min\left\{1, \frac{\sqrt{S}}{\beta}\right\}$) we have

$$\begin{aligned}\mathcal{L}(\bar{\boldsymbol{w}}^{(R)}) - \mathcal{L}(\boldsymbol{w}^*) \leq & 2\left(\frac{L\sigma^2}{S^2K}\right)^{\frac{1}{2}}\left(\frac{\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^{(0)}) - \mathcal{L}^*\right]}{R+1}\right)^{\frac{1}{2}} \\ & + 2\left(\frac{9L^2\sigma^2}{8SK} + \frac{9L^2\zeta^2}{8S}\right)^{\frac{1}{3}}\left(\frac{\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^{(0)}) - \mathcal{L}^*\right]}{R+1}\right)^{\frac{2}{3}} + \frac{6L\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^{(r)}) - \mathcal{L}^*\right]}{R+1}\end{aligned}$$

Due to the concave nature of the power function(Lemma.2), we can get

$$\begin{aligned}\mathcal{L}(\bar{\boldsymbol{w}}^{(R)}) - \mathcal{L}(\boldsymbol{w}^*) \leq & 2\left(\frac{L\sigma^2}{S^2K}\right)^{\frac{1}{2}}\left(\frac{\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^{(0)}) - \mathcal{L}^*\right]}{R+1}\right)^{\frac{1}{2}} \\ & + 2\left[\left(\frac{9L^2\sigma^2}{8SK}\right) + \left(\frac{9L^2\zeta^2}{8S}\right)\right]^{\frac{1}{3}}\left(\frac{\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^{(0)}) - \mathcal{L}^*\right]}{R+1}\right)^{\frac{2}{3}} + \frac{6L\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^{(r)}) - \mathcal{L}^*\right]}{R+1}\end{aligned}$$

After finishing, we can get:

$$\min_{0\leq r\leq R}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}^{(r)})\right\|^2\right] \leq \frac{3\left(\mathcal{L}(\boldsymbol{w}^0) - \mathcal{L}^*\right)}{\tilde{\eta}R} + \frac{3L\tilde{\eta}\sigma^2}{S^2K} + \frac{27L^2\tilde{\eta}^2\sigma^2}{8SK} + \frac{27L^2\tilde{\eta}^2\zeta^2}{8S} \tag{71}$$

where we use $\min_{0\leq r\leq R}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}^{(r)})\right\|^2\right] \leq \frac{1}{R+1}\sum_{r=0}^{R}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}^{(r)})\right\|^2\right]$ Then, using $\tilde{\eta} = \frac{1}{6L(\beta+1)} \leq \min\left\{1, \frac{\sqrt{S}}{\beta}\right\}$ and by dynamically adjusting the learning rate, we have

$$\min_{0\leq r\leq R}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}^{(r)})\right\|^2\right] = \mathcal{O}\left(\frac{\left(L\sigma^2\mathcal{B}\right)^{1/2}}{\sqrt{SKR}} + \frac{\left(L^2\sigma^2\mathcal{B}^2\right)^{1/3}}{(S^2K)^{1/3}R^{2/3}} + \frac{\left(L^2\zeta^2\mathcal{B}^2\right)^{1/3}}{S^{1/3}R^{2/3}} + \frac{L\beta\mathcal{B}}{R}\right) \tag{72}$$

where $\mathcal{B} := \mathcal{L}(\boldsymbol{w}^0) - \mathcal{L}^*$. Eq.71 and Eq. 72 are upper bounds for the case of partial client participation. In particular, when $S = M$, we obtain the conclusions of Theorem 1 in the non-convex case.

H Theoretical Analysis of Update Order Sensitivity

What differentiates PFL and SFL is that the current updated gradient is dependent on the gradient from the previous round; thus, we represent their dependency using conditional probability, as follows Eq. 73. This is why order sensitivity exists.

$$\mathcal{L}_{\pi_m^n,k}\left(f(x_i^{\pi_m^n,k}; \boldsymbol{w}_{\pi_m^n,k}), y_i^{\pi_m^n,k} | \sum_{j=1}^{m-1}\nabla\mathcal{L}_{\pi_j^n,k}\right) \neq \mathcal{L}_{\pi_m^n,k}\left(f(x_i^{\pi_m^n,k}; \boldsymbol{w}_{\pi_m^n,k}), y_i^{\pi_m^n,k}\right) \tag{73}$$

We provide two examples to illustrate that the update order in SFL introduces variance: one from the perspective of loss function differences, and the other from the data distribution perspective.

H.1 Loss Function Differences

We consider two clients A and B, each taking $K = 1$ steps of SGD in each training round, with a learning rate of η.

Definition:

- $x^{(r)}$: the global model at the beginning of round r;

- The local gradient of client A is $g_A(x) = \nabla F_A(x)$, and that of client B is $g_B(x) = \nabla F_B(x)$.

Order 1 (A $\to$ B):

$$x^{(r)} \xrightarrow{\text{A}} x_1 = x^{(r)} - \eta g_A(x^{(r)}) \xrightarrow{\text{B}} x^{(r+1)} = x_1 - \eta g_B(x_1)$$

Thus:

$$x^{(r+1)} = x^{(r)} - \eta g_A(x^{(r)}) - \eta g_B\left(x^{(r)} - \eta g_A(x^{(r)})\right)$$

Order 2 (B $\to$ A):

$$x^{(r)} \xrightarrow{\text{B}} x_1' = x^{(r)} - \eta g_B(x^{(r)}) \xrightarrow{\text{A}} x_{\text{alt}}^{(r+1)} = x_1' - \eta g_A(x_1')$$

Thus:

$$x_{\text{alt}}^{(r+1)} = x^{(r)} - \eta g_B(x^{(r)}) - \eta g_A(x^{(r)} - \eta g_B(x^{(r)}))$$

Here is the translated and refined version in academic style:

As an example, let $F_A(x) = \frac{1}{2}(x-1)^2$, $F_B(x) = \frac{1}{2}(x+1)^2$, representing left- and right-biased objectives, respectively:

That is,$\nabla F_A(x) = x - 1$, $\nabla F_B(x) = x + 1$. We get:

$$\begin{aligned}
x^{(r+1)} &= x^{(r)} - \eta g_A(x^{(r)}) - \eta g_B\left(x^{(r)} - \eta g_A(x^{(r)})\right) \\
&= x^{(r)} - \eta(x^{(r)} - 1) - \eta(x^{(r)} - \eta(x^{(r)} - 1) + 1) \\
&= (1 - 2\eta + \eta^2)x^{(r)} - \eta \\
x_{alt}^{(r+1)} &= x^{(r)} - \eta g_B(x^{(r)}) - \eta g_A(x^{(r)} - \eta g_B(x^{(r)})) \\
&= x^{(r)} - \eta(x^{(r)} + 1) - \eta(x^{(r)} - \eta(x^{(r)} + 1) - 1) \\
&= (1 - 2\eta + \eta^2)x^{(r)} + \eta
\end{aligned}$$

By substituting specific values (e.g., $x^{(r)} = 0$, $\eta = 0.1$), we can verify that the results differ under different update orders.Clearly, $x_{alt}^{(r+1)} \neq x^{(r+1)}$. When the loss functions are inconsistent, the update order in SFL influences the gradient used for model updates.

H.2 The Data Distribution Differences

Suppose two clients, A and B, have different local data distributions $\mathcal{D}_A \neq \mathcal{D}_B$. Then their corresponding local objective functions differ:

- $F_A(x) = \mathbb{E}_{\xi \sim \mathcal{D}_A}[f(x;\xi)]$
- $F_B(x) = \mathbb{E}_{\xi \sim \mathcal{D}_B}[f(x;\xi)]$

Therefore, for the same x, their gradients are also different:

$$\nabla F_A(x) \neq \nabla F_B(x) \tag{74}$$

This defines the mathematical nature of data heterogeneity: gradient diversity, i.e., inconsistency in gradient distributions.

Example: Suppose we use MSE loss (Mean Squared Error)

For a linear regression task:

- Client A's data: target value $y = ax$
- Client B's data: target value $y = bx$

Let the model be $\hat{y} = \theta x$. The loss function is:

$$f(x;\xi) = \frac{1}{2}(\hat{y} - y)^2 = \frac{1}{2}(wx - y)^2 \tag{75}$$

Then:

- The expected loss gradient for client A is:

$$\nabla F_A = \mathbb{E}_x \left[\frac{\partial}{\partial w} \left(\frac{1}{2}(wx - ax)^2 \right) \right] = \mathbb{E}_x \left[(wx - ax)x \right]$$

- The expected loss gradient for client B is:

$$\nabla F_B = \mathbb{E}_x \left[\frac{\partial}{\partial w} \left(\frac{1}{2}(wx - bx)^2 \right) \right] = \mathbb{E}_x \left[(wx - bx)x \right]$$

If the distribution of x is the same (e.g., x follows a standard normal distribution), then:

$$\nabla F_A(w) = (w - a)\mathbb{E}[x^2], \quad \nabla F_B(w) = (w - b)\mathbb{E}[x^2] \tag{76}$$

Order 1 (A $\rightarrow$ B):

$$\begin{aligned} w^{(r+1)} &= w^{(r)} - \eta \nabla F_A(w^{(r)}) - \eta \nabla F_B \left(w^{(r)} - \eta \nabla F_A(w^{(r)}) \right) \\ &= w^{(r)} - \eta(w^{(r)} - a)\mathbb{E}[x^2] - \eta(w^{(r)} - \eta(w^{(r)} - a)\mathbb{E}[x^2] - b)\mathbb{E}[x^2] \\ &= w - 2\eta w \mathbb{E}[x^2] + \eta^2 w \mathbb{E}^2[x^2] - (a + b)\eta \mathbb{E}[x^2] - a\eta^2 \mathbb{E}^2[x^2] \end{aligned}$$

Order 2 (B $\rightarrow$ A):

$$\begin{aligned} w_{alt}^{(r+1)} &= w^{(r)} - \eta \nabla F_B(w^{(r)}) - \eta \nabla F_A \left(w^{(r)} - \eta \nabla F_B(w^{(r)}) \right) \\ &= w^{(r)} - \eta(w^{(r)} - b)\mathbb{E}[x^2] - \eta(w^{(r)} - \eta(w^{(r)} - b)\mathbb{E}[x^2] - a)\mathbb{E}[x^2] \\ &= w - 2\eta w \mathbb{E}[x^2] + \eta^2 w \mathbb{E}^2[x^2] - (a + b)\eta \mathbb{E}[x^2] - b\eta^2 \mathbb{E}^2[x^2] \end{aligned}$$

By substituting specific values (e.g., $a = 1, b = 2$), we can verify that the results differ under different update orders.Clearly, $w_{alt}^{(r+1)} \neq w^{(r+1)}$. When the Data Distributions are inconsistent, the update order in SFL influences the gradient used for model updates.

I More Experimental Details

I.1 Dataset.

We researched our proposed method using two types of datasets. One category consists of datasets with a category shift, which necessitates our partitioning. Our partitioning is based on the Dirichlet distribution[35] Dir β. Where β (default 0.5) is an argument correlated with the heterogeneity level. In this problem, we set the default number of clients to 10. There are three data sets we need to divide here, CIFAR-10[42], CIFAR-100[42], and CINIC-10[43]. CIFAR-10 is a widely used dataset consisting of 60,000 32x32 color images in 10 classes, with 6,000 images per class, primarily used for image classification tasks.CIFAR-100 is similar to CIFAR-10, but with 100 classes containing 600 images each, providing a more granular challenge for image classification tasks.CINIC-10 is an extended version of CIFAR-10, containing 270,000 images split into 10 classes, designed to bridge the gap between CIFAR-10 and ImageNet[60] for improved training scalability.

For domain shift, we conducted experiments on two benchmark datasets: PACS [61] and Office-Home[62]. (*i*) The **PACS** [61] dataset contains four distinct domains (*Photos, Art Paintings, Cartoons, and Sketches*), with a total of 9,991 images. Each domain shares the same 7-class label space, despite the variations in image styles.

(*ii*) The **Office-Home** [62] dataset contains approximately 15,500 images across 65 categories from four domains (*Art, Clipart, Product, and Real-World*), offering a diverse range of categories that better test the robustness of our method. We have included the Office-Home experiment in the supplementary materials.

Table 9: Comparison of SPFL and PFL on Tiny-ImageNet and CIFAR-100 across various FL methods (Dirichlet $\alpha = 0.1$).

Method	Tiny-ImageNet	Tiny-ImageNet (SPFL)	CIFAR-100	CIFAR-100 (SPFL)
FedAvg	15.69	33.82 (+18.13)	57.55	**64.33 (+6.78)**
FedProx	15.47	16.65 (+1.18)	57.56	60.46 (+2.90)
MOON	15.42	16.47 (+1.05)	58.44	63.54 (+5.20)
FedDyn	15.22	17.66 (+2.44)	64.11	64.24 (+0.13)
Scaffold	14.58	18.16 (+3.58)	56.15	65.05 (+8.90)
FedDC	15.37	16.45 (+1.08)	**64.44**	**69.88 (+5.44)**
FedNova	15.53	16.45 (+0.98)	57.64	64.47 (+6.83)
FedDisco	**33.55**	**34.50 (+0.95)**	57.50	64.44 (+6.94)

Table 10: Accuracy (%) on PACS under Domain Shift Setting.

Method	P	A	C	S	Avg
PFL (120R)	91.02	68.61	**70.85**	66.84	74.33
PFL (100R)+SPFL (20R)	93.11	73.24	69.36	64.92	75.16
PFL (100R)+SPFL *with* GLAM (20R)	**93.71**	**74.45**	70.21	64.45	**75.71**

Table 11: Performance comparison of GLAM+SPFL with recent FL methods under domain shift.

Method	P	A	C	S	Avg
CAN [18]+SPFL	90.06	70.32	**77.63**	66.47	76.12
FedDA [51]+SPFL	90.41	73.83	69.01	**74.01**	76.82
FedGALA [64]+SPFL	87.58	68.58	68.17	66.56	72.72
GLAM+SPFL (Ours)	**92.28**	**77.44**	74.40	67.32	**77.86**

I.2 Implementation Details.

When we are in the domain shift setting, for local model training across the PACS and OfficeHome datasets, we utilize architectures ResNet18 and ResNet50, as detailed by [3], which are pre-trained on the ImageNet [60]. We adopt a leave-one-domain-out evaluation method for all benchmarks, where one domain is reserved for testing and the remaining domains are used for training and validation. To ensure consistency and fairness in experiments, we standardize the batch size and learning rate at 128 and 0.2 in local training. Furthermore, to guarantee that the local models reach convergence within each training phase, we set the number of local epochs E to 1 and define the communication rounds R as 100.

In the category shift setting, we utilized a pre-trained model from PFL (FedAvg) that was trained for 100 rounds. Subsequently, we ran our new algorithm for an additional 50 rounds. For datasets Cifar10 and Cifar100, we employed the ResNet18 model framework, while for dataset CINIC-10, we utilized a Simple-CNN [63] model framework. We consistently employed Accuracy (acc) as the performance metric for our evaluations.

In the training phase, we conducted our experiments on a single NVIDIA Tesla A800 GPU. We used four NVIDIA Tesla A800 GPUs during testing to obtain results more quickly. However, the experiments only required 5-10GB of GPU memory due to a batch size of 128.

I.3 More comparative experiments

Performance comparison on complex datasets. Tab. 9 compares SPFL and PFL on Tiny-ImageNet and CIFAR-100 under different FL methods. SPFL consistently outperforms PFL across all settings, showing significant gains in both accuracy and convergence speed. These results demonstrate that SPFL achieves better generalization with most FL methods.

Performance of SPFL and GLAM under hybrid strategies. As shown in Tab. 10, when adopting a hybrid strategy, SPFL maintains a lower convergence margin on the main shift but converges more

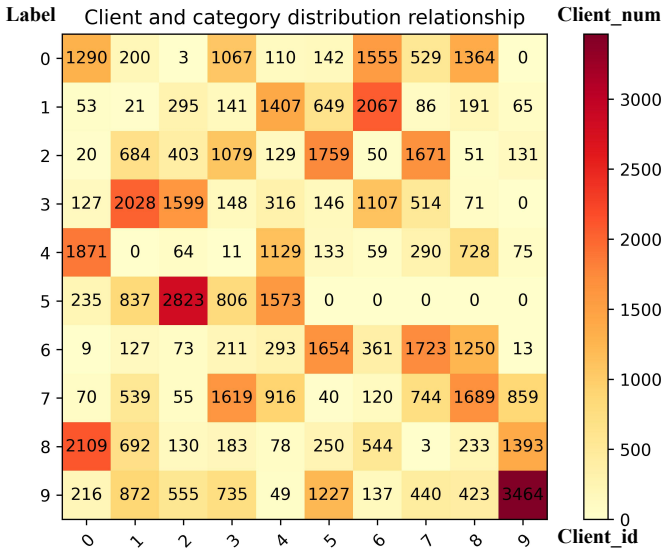


Figure 6: Distribution relationship between client and category

Table 12: The combination experiments of the two different loss functions were conducted on the Office-Home dataset using a pre-trained ResNet-18 (Best in **bold**)

Method	$\mathcal{L}_{ag}$	$\mathcal{L}_{ap}$	*Product*	*Art*	*Clipart*	*Real-World*	Avg
SPFL	×	×	57.22	43.10	46.07	59.22	51.40
SPFL	✓	×	**58.08**	43.88	47.10	59.10	52.04
SPFL	×	✓	56.75	44.27	47.62	**59.49**	52.03
SPFL	✓	✓	57.40	**44.47**	**47.40**	59.20	**52.14**

slowly. Therefore, first applying PFL to reach the boundary, followed by a low number of SPFL rounds, can achieve a lower convergence margin. Meanwhile, GLAM remains effective under this hybrid strategy.

Comparison with state-of-the-art methods. CAN [18] represents the latest federated continual learning approach [65, 26], while FedDA [51] and FedGala [64] are recent advances in federated domain generalization. However, unlike GLAM, these methods do not align the model transmitted from the previous client, leading to limited performance gains when adapting to new domains. In contrast, GLAM provides a more comprehensive alignment mechanism, allowing flexible adjustment of the comparison strength through parameters τ and ρ, as shown in Tab. 11.

I.4 More Ablation Study

Performance in Office-Home. Through ablation studies on PACS and Office-Home, the effectiveness of the GLAM module is thoroughly demonstrated, with each component contributing to the overall model performance, as shown in Fig. 12.

Comparison of different update methods. We compare the two existing update schemes, PFL and SFL, with our proposed SPFL. As shown in Tab. 13, SPFL consistently outperforms both methods across various FL frameworks, demonstrating its superior effectiveness and adaptability. To quickly converge, the initial model was initialized using a FedAvg pre-trained for 200 rounds on CIFAR-10, with 10 clients.

Comparison of different Dirichlet α. We conduct an ablation study on parameter α. As shown in Tab. 14, the performance improvement of SPFL becomes more pronounced as the degree of non-IID increases.To quickly converge, the initial model was initialized using a FedAvg pre-trained for 200 rounds on CIFAR-10, with 10 clients.

Table 13: Comparison of different update strategies (PFL, SFL, and our SPFL) under various FL methods on CIFAR-10 (Dirichlet $\alpha = 0.1$).

Update\Method	FedAvg	FedProx	FedDC	Moon
PFL	55.60	56.65	55.16	56.33
SFL	68.59	53.75	54.41	55.24
SPFL (ours)	**71.39**	**57.66**	**55.26**	**58.44**

Table 14: Accuracy (%) of SPFL and PFL (FedAvg) under different Dirichlet α settings.

Update\Dir	0.1	0.2	0.5	1	100
PFL (FedAvg)	39.25	43.40	42.85	46.03	53.37
SPFL (Ours)	**47.55**	**50.96**	**49.38**	**49.47**	**55.27**

Table 15: Accuracy (%) of PFL, multiple SFL variants, and SPFL under different Dirichlet α settings.

Method	PFL	SFL(1)	SFL(2)	SFL(3)	SFL(4)	SFL(5)	SPFL (ours)
Dir(0.1)	41.85	49.56	54.45	49.51	47.98	**55.56**	**55.33**
Dir(0.5)	43.60	50.22	47.95	**55.18**	54.20	49.52	**55.16**

Comparison of different update orders. We conduct experiments with five clients for training, where SFL(i) denotes using client i as the starting point. As shown in Tab. 15, different starting points result in significant performance variations, highlighting the update sensitivity problem in SFL. In contrast, SPFL effectively mitigates this issue, achieving more stable and consistent performance across clients.

Effect of client participation rate. Tab. 16 compares PFL and SPFL under different client participation rates (E) with 100 clients. As E increases, SPFL consistently outperforms PFL, showing greater stability and accuracy, which demonstrates its stronger adaptability to varying participation levels.

Effect under Category and Domain Shifts. As shown in Tab. 17, SPFL significantly improves performance under both category and domain shifts compared to PFL and SFL, demonstrating its stronger generalization capability. Furthermore, integrating the GLAM module further enhances performance, especially under domain shift, by effectively aligning inter-domain representations. This confirms that SPFL addresses update sensitivity, while GLAM strengthens cross-domain consistency and adaptability.

J Limitation

Although SPFL demonstrates strong performance in addressing category shift and achieves comparable results to PFL under domain shift, this improvement imposes stricter requirements on the deployment environment, specifically, it assumes that clients remain continuously online. Additionally, while increasing the number of clients can accelerate convergence and lower the theoretical convergence upper bound, it also introduces substantial communication overhead. As a result, SPFL is better suited for cross-silo scenarios rather than cross-device settings.

Moreover, cross-silo environments typically entail higher computational costs, whereas SPFL is able to match the performance of multi-epoch PFL using significantly fewer epochs. Therefore, although SPFL is theoretically applicable to cross-device scenarios, it is more practically aligned with cross-silo applications. A detailed comparison of computational and communication costs is presented in Tab. 18.

Where $\mathcal{M}$ is the model size, M is the number of clients, E is the epoch, and T is the number of communication rounds.

Table 16: Accuracy comparison between PFL and SPFL under different client participation rates (E) with 100 total clients.

Update\E	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
PFL	41.78	46.99	52.43	53.31	54.04	54.52	54.47	54.45	54.35	54.59
SPFL	51.72	63.57	72.33	76.19	80.84	82.68	84.43	85.41	85.98	**89.95**

Table 17: Performance under Category and Domain Shift with PFL, SFL, SPFL, and SPFL+GLAM.

Shift Type	**PFL**	**SFL**	**SPFL**	**SPFL+GLAM**
Category Shift	55.60	68.59	71.39	**71.71**
Domain Shift (Avg)	**78.37**	60.04	76.32	77.87

Table 18: Cost calculation analysis

Method	*Calculate costs*	*Communication costs*
PFL	$T \times M \times E \times 3 \times \mathcal{M} \times Batch\_Size$	$T \times K \times 2 \times \mathcal{M} \times 4$
SFL	$T \times M \times E \times 3 \times \mathcal{M} \times Batch\_Size$	$T \times M \times \mathcal{M} \times 4$
SPFL	$T \times M \times E \times 3 \times \mathcal{M} \times Batch\_Size \times K$	$T \times M \times 2 \times \mathcal{M} \times 4 \times (K + 1/2)$

K Broader impacts

The proposed SPFL framework enhances the robustness and generalization of federated learning systems under category and domain shifts, potentially benefiting applications involving privacy-sensitive and heterogeneous data, such as healthcare, finance, and education. By reducing the dependency on data centralization, it supports data sovereignty and regulatory compliance.

However, stronger model generalization across clients may inadvertently increase the risk of model inversion or membership inference attacks. Additionally, more complex update schemes could amplify computational inequality between clients with varying resources. Future work should consider fairness and security safeguards to mitigate these risks.