PA^{3}Fed: Period-Aware Adaptive Aggregation for Improved Federated Learning

Chengxiang Huang[1], **Bingyan Liu**[1] *

[1]Beijing University of Posts and Telecommunications
{huangchengxiang2021, bingyanliu}@bupt.edu.cn

Abstract

Federated Learning (FL) is a distributed approach that enables collaborative model training while safeguarding client data privacy. Nevertheless, FL encounters difficulties due to statistical heterogeneity from the varied data distributions across numerous clients, which can affect overall efficiency and performance. Existing state-of-the-art FL methods often concentrate on optimizing interactions between clients, neglecting the potential insights from individual clients during training. Additionally, these approaches generally assume that every period of training has an equal impact on the final model's performance. To address these issues, this paper introduces a novel method, PA^{3}Fed, which conducts period-aware adaptive aggregation for improved federated learning. The key idea is to identify the most critical periods, *i.e., those with the highest information content and entropy*, where we leverages each client's own performance variations during training for adaptive aggregation. Furthermore, because it operates independently of inter-client optimization approaches, it can be easily incorporated into other baselines for improved performance. Experimental results show that our method improves accuracy by up to 15% and significantly enhances stability.

Introduction

Federated Learning (FL) (McMahan et al. 2017; Li et al. 2020; Karimireddy et al. 2020) has revolutionized the field of distributed machine learning by effectively addressing the challenges posed by real-world data heterogeneity and privacy concerns (Kairouz et al. 2021). Classical Federated Learning methods primarily rely on the Federated Averaging (FedAvg) (McMahan et al. 2017) algorithm, which can be seen as the first state-of-the-art method in FL. In each communication round, FedAvg utilizes local computations at each client to train models on their unique datasets and then relies on a centralized server to aggregate these local models for a better global model.

Despite the success of FL in various real-world applications (Wang, Liu, and Li 2024; Long et al. 2020; Liu et al. 2024), Federated Learning continues to grapple with the statistical heterogeneity challenge (Imteaj et al. 2022, 2021; Liu et al. 2023): the challenge posed by non-i.i.d. (non-identical and independent distribution) datasets can lead to a pronounced divergence in the gradient updates from local clients (Liu et al. 2021; Liu, Guo, and Chen 2021). This divergence causes the global model's updates to stray from the reasonable gradient direction. As a result, given the federated global model in each round, the loss function (such as CrossEntropy Loss) on individual clients can become highly unstable, experiencing considerable fluctuations during the update process. These fluctuations are not merely minor deviations but can represent substantial swings that undermine the consistency of local training. Consequently, the overall stability of the federated learning system is compromised, making it challenging to converge to an optimal status, and ultimately reducing the effectiveness and reliability of the model in real-world applications.

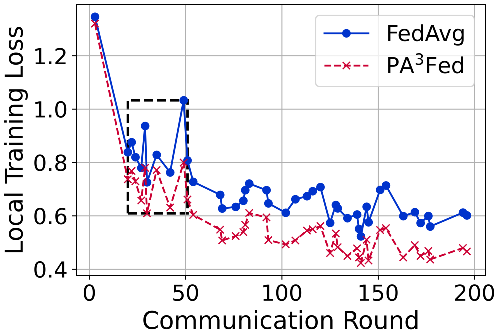


Figure 1: **Illustration of the Loss Curve for A Single Local Client.** The FedAvg algorithm curve shows significant oscillations in the early stages, whereas our method is not only much smoother but also achieves a lower overall loss.

Many state-of-the-art methods have been proposed to address these challenges. For instance, FedProx (Li et al. 2020) and SCAFFOLD (Karimireddy et al. 2020) introduce global constraints to help local models align more closely with the global model's updates, thereby achieving more stable and consistent training. Additionally, techniques such as clustering (Ruan and Joe-Wong 2022), contrastive learning (Mu et al. 2023), and knowledge distillation (Lin et al. 2020; Fan et al. 2023) have been employed to further enhance the infer-

[*] Corresponding author.

Metric	**FedAvg**	**PA^{3}Fed**
Standard Deviation	0.089	**0.066**
Moving Average Standard Deviation	0.019	**0.014**
Mean Absolute Deviation	0.069	**0.058**
Range	0.307	**0.189**
Maximum Change Rate	0.270	**0.170**

Table 1: **Comparison of Various Metrics for Stability: FedAvg vs PA^{3}Fed.** These metrics are widely used to evaluate the stability of a system (Bashan et al. 2008; Buckley and Voorhees 2017). PA^{3}Fed demonstrates superior stability in various aspects, highlighting its robustness and effectiveness compared to the traditional FedAvg approach.

ence capability of the global model, ensuring that it can generalize effectively across diverse data distributions. These approaches contribute to mitigating the negative effects of heterogeneity, ultimately improving the overall performance and robustness of federated learning systems.

The aforementioned methods have made significant progress in addressing the statistical heterogeneity issue. However, they still have limitations. Most FL approaches primarily focus on inter-client optimization (Li et al. 2020; Karimireddy et al. 2020; Liang et al. 2019), which often leads to the oversight of valuable information embedded within the training processes of individual clients. As shown in Figure 1, when examining the training process of a single client, especially under highly non-i.i.d. conditions, we observe that the loss curve exhibits significant fluctuations, particularly during the early stages of training. Inspired by this observation, our question is: *Could we design an FL method that focuses on the fluctuations in the training process of individual clients, allowing for adaptations based on each client's specific data, while ensuring compatibility with inter-client optimization methods to maintain scalability?*

To address this question, in this paper, we propose **PA^{3}Fed**, aiming to achieve period-aware adaptive aggregation for improved federated learning. Instead of relying on inter-client optimization, PA^{3}Fed takes advantage of the training status of individual clients to guide the final FL aggregation in a period-aware manner. Our method is mainly driven by the following observations: (1) As aforementioned, the loss curves under FedAvg exhibits significant oscillations. Besides, the various metrics for training stability also have the potential to optimize (illustrated in Table 1). Based on these observations, our method attempts to utilize a performance score to evaluate each client's performance during every selected round, using the difference in these scores to measure performance discrepancies. This discrepancy serves as one of the factors for adjusting the aggregation coefficients. (2) We also find that the loss curves of local clients experience considerable volatility during the early phase of training. In the context of Federated Learning, this period is referred to as the Critical Learning Periods (CLP) (Yan, Wang, and Li 2021; Yan et al. 2023), underscoring the importance of the early phase in the training process. During this time, the FL system's information is most abundant, with the information entropy reaching its peak (measured by the Fisher Information Matrix). Therefore, we believe applying adaptation during this period can maximize the impact on the model's performance with minimal intervention during the overall training process. Specifically, we first check whether the current period is in CLP and then employ the adaptive aggregation with our computed coefficients accordingly.

Since PA^{3}Fed pays more attention to the intra-client design, it is orthogonal to other baselines and can be easily integrated for better performance. We conduct extensive experiments on various datasets and models and validate the excellent performance of our method by incorporating it into several state-of-the-art approaches, including FedAvg (McMahan et al. 2017), FedProx (Li et al. 2020), FedNova (Wang et al. 2020b), VRL-SGD (Liang et al. 2019), and FedMut (Hu et al. 2024). The results indicate that our method can improve performance by up to 15% under the high non-IID conditions compared to the original methods. In addition, we conduct a series of in-depth empirical analyses to prove the effectiveness of the proposed method.

Our main contributions are summarized as follows:

- We design an effective and easy-to-compute method to dynamically calculate the aggregation coefficient for each client in each round, with the help of the intra-client information. This is orthogonal to other inter-client optimization methods.
- We propose Period-Aware Adaptive Aggregation, namely PA^{3}Fed, a streamlined and computationally efficient module for optimizing global model aggregation during critical learning periods. This approach notably enhances the convergence rate and performance of the global model, particularly under severe non-IID conditions.
- The results demonstrate that PA^{3}Fed can effectively enhance the performance atop other state-of-the-arts.

Related Work

Federated Learning. To address the needs of secure distributed machine learning, Google introduces the classical federated learning (FL) approach, known as FedAvg (McMahan et al. 2017). FedAvg learns a single global model by averaging the local models sent by randomly selected clients. It is well-known for its effectiveness on i.i.d. datasets and its privacy-protection characteristics. However, there are still many challenges to tackle in real-world applications, including the number of communication rounds (McMahan et al. 2016), gradient leakage (Chu et al. 2022), and statistical heterogeneity (Li et al. 2020; Karimireddy et al. 2020; Cui et al. 2022).

Many methods have been developed in recent years to address these challenges. For example, FedProx (Li et al. 2020) introduces a proximal term in the local loss function to ensure that local client updates remain closer to the global model, leading to more stable updates, particularly in non-i.i.d. situations. VRL-SGD (Liang et al. 2019) reduces communication overhead compared to local SGD algorithms by incorporating a variance reduction technique. FedOpt (Asad, Moustafa, and Ito 2020) develops a federated optimization method to address convergence issues and the lack of adaptability in federated learning by using an adaptive optimizer as the server optimizer while keeping the client optimizers simple. SCAFFOLD (Karimireddy et al. 2020) uses a control variate method to mitigate client drift caused by client data heterogeneity, achieving faster convergence than FedAvg. Apart from the methods mentioned above, clustering (Ruan and Joe-Wong 2022), contrastive learning (Li, He, and Song 2021), and knowledge distillation (Lin et al. 2020; Fan et al. 2023) are widely used in FL. Clustering groups similar data distributions to improve local training and aggregation. Contrastive learning aligns similar data points and differentiates dissimilar ones to enhance generalization. Knowledge distillation transfers knowledge from teacher to student models, ensuring consistent performance across diverse nodes. Together, these techniques significantly mitigate the impact of data heterogeneity in FL models.

Critical Learning Periods. Critical Learning Periods (CLP) (Achille, Rovere, and Soatto 2018; Jastrzębski et al. 2018) are initially discussed in centralized learning, emphasizing the pivotal role of early training rounds. Recently, the concept of CLP in federated learning (Yan, Wang, and Li 2021; Yan et al. 2023) has been explored, demonstrating its significant impact on convergence time and final model accuracy. For example, CriticalFL (Yan et al. 2023) capitalizes on this by dynamically adjusting the number of selected clients—gradually increasing them during the CLP and decreasing them afterward—thereby enhancing FL model performance and reducing communication overhead.

Limitations of current practice. While previous work has indeed improved the performance of FL models, several limitations remain. Firstly, previous research has predominantly focused on inter-client interactions, often overlooking the valuable information inherent in the variations of training rounds within a single client. For instance, if a client's accuracy at round t is greater than at round $t + \sigma$ ($\sigma \geq 1$), it may indicate that the client's updates are not aligned with the global model's direction. Secondly, previous work has generally ignored the inherent constraints of federated learning, optimizing throughout the entire training phase. This can easily lead to training noise, especially in cases where the degree of statistical heterogeneity is high. To the best of our knowledge, PA^{3}Fed is the first method introduced with the primary aim of overcoming these two key limitations. With the intra-client based adaptive aggregation, our method not only brings considerable improvements to the model's stability and performance but also harmonizes well with inter-client optimization strategies in federated learning, demonstrating its excellent scalability and compatibility.

Method

Overview

As depicted in Figure 2 and detailed in Algorithm 1, in this paper, we propose the PA^{3}Fed approach, which dynamically adjusts the aggregation coefficient based on the performance variations of local clients and the stage of training. The process of PA^{3}Fed involves several key steps on the client side and server side. We elaborate on them as follows:

- *Calculate Aggregation Coefficient*: As illustrated in Step 3 of Figure 2, each local client computes its aggregation coefficient α_i based on the changes in its model's performance.
- *Period Check:* Once the central server receives the data uploaded by the clients, it conducts a period check to determine the current training period of the FL system.
- *Adjust Aggregation Coefficient*: Based on the results of the period check, the central server adjusts the aggregation coefficient from α_i to $\alpha_i^{'}$.
- *Adaptive Aggregation:* With the updated coefficients $\alpha_i^{'}$, the system proceeds to perform adaptive aggregation.

Problem Definition

Consider a federated learning (FL) system with a central server S and K clients, denoted as $\{\mathrm{C_1}, \mathrm{C_2}, \mathrm{C_3}, \ldots, \mathrm{C_K}\}$. Each client $\mathrm{C_i}$ possesses its own private dataset $\mathrm{D_i}$, containing $\mathrm{n_i}$ data points, where $\mathrm{i} \in \{1, 2, 3, \ldots, K\}$. The dataset $\mathrm{D_i}$ comprises input-output pairs $\{\mathcal{X}_i, \mathcal{Y}_i\}$, where $\mathcal{X}_i = \left\{\mathrm{x_1^i}, \mathrm{x_2^i}, \mathrm{x_3^i}, \ldots, \mathrm{x_n^i}\right\}$ represents the input space, and $\mathcal{Y}_i = \left\{\mathrm{y_1^i}, \mathrm{y_2^i}, \mathrm{y_3^i}, \ldots, \mathrm{y_n^i}\right\}$ represents the output space. A single data point can be expressed as $\left\{\mathrm{x_n^i}, \mathrm{y_n^i}\right\}$. We define the relative size of a client's dataset as $\mathfrak{m}_i = \frac{\mathrm{n_i}}{\sum_{j=1}^{K} \mathrm{n_j}}$, which represents the proportion of the total data held by client $\mathrm{C_i}$. Most FL methods utilize the aggregation approach introduced in FedAvg (McMahan et al. 2017) to obtain the global model. This method averages selected client models based on the relative size of the local dataset $\mathfrak{m}_i$, to learn a global model by solving the following optimization problem:

$$\min_{\mathrm{W}} \mathcal{L}\left(\mathrm{w}, \mathcal{D}\right) := \sum_{i=1}^{K} \mathfrak{m}_i \mathfrak{l}_i \left(\mathrm{w}, \mathrm{D}_i\right), \tag{1}$$

where w denotes the model parameters, $\mathcal{D} = \bigcup_{i=1}^{K} \mathrm{D}_i$ is the whole training dataset, $\mathfrak{l}_i$ is the local loss function.

However, FedAvg, in its aggregation process, solely considers the size of the datasets across different clients to determine each client's contribution, thereby overlooking the issue of statistical heterogeneity and the distinct needs of various training periods. Our goal is to assign period-aware aggregation coefficient $\alpha_i(t)$ to different training round $t \in \{1, 2, 3, \ldots, T\}$. This coefficient allows FL system to adjust the contribution of each client, $\mathfrak{m}_i$, during the aggregation process to enhance the model's stability and performance.

CLP in Federated Learning

In the context of FL, Critical Learning Periods (CLP) are typically identified by analyzing the trace of the Federated Fisher Information Matrix (Yan, Wang, and Li 2021). The Fisher Information Matrix (FIM) is instrumental in evaluating the quantity of information contained within the FL system. It serves as a local gauge to measure how perturbations in the weights affect the model's output. Furthermore, the FIM can be regarded as an approximation of the Hessian matrix of the loss function, which provides insight into

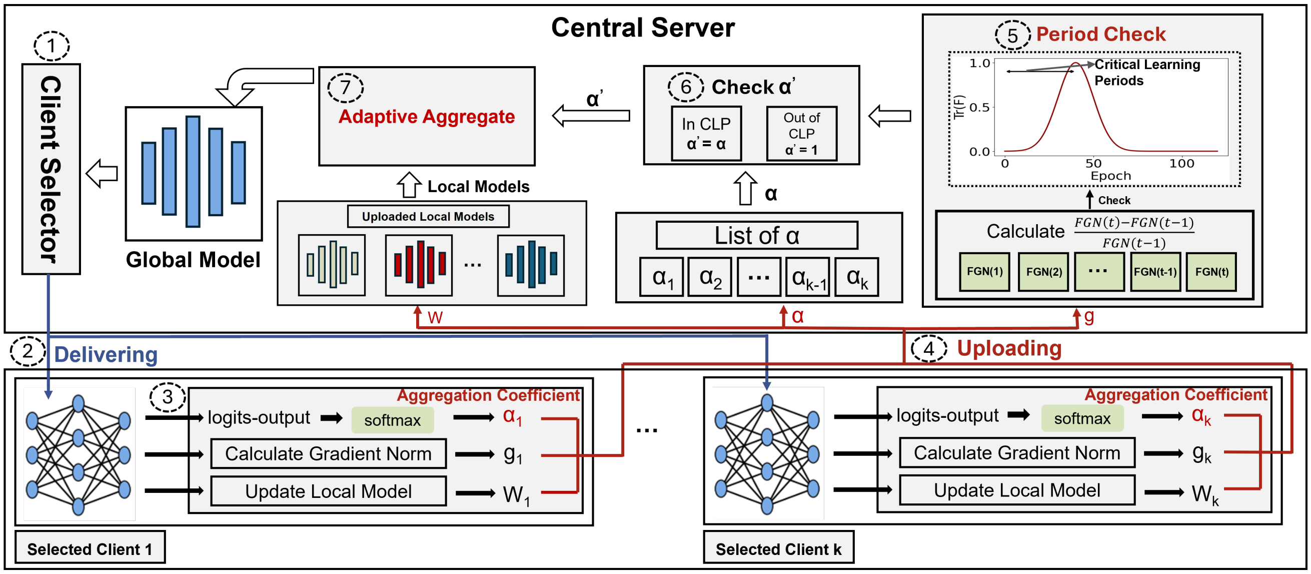


Figure 2: **Overview of PA^{3}Fed.** The diagram provides a comprehensive view of the proposed pipeline. It illustrates the main components and workflow in the client side and server side.

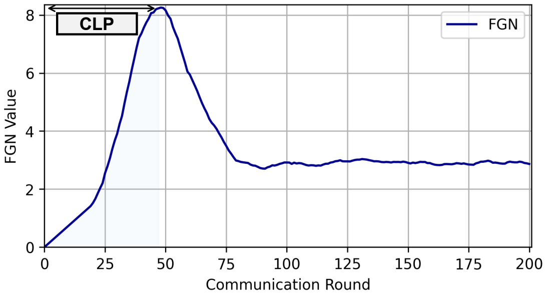


Figure 3: **Illustration of Critical Learning Periods (CLP).** In the diagram, the light blue shaded area represents the CLP regions.

the curvature of the loss landscape at a specific point during training (Martens 2020). This approximation allows us to understand the FIM as a reflection of the local behavior of the optimization process, linking it directly to the dynamics of how the model learns from the data. For a single client $i \in K$, consider the local classification model $\phi^i(y \mid x)$, with w, the parameter downloaded from the central server, the local FIM can be denoted as:

$$\mathcal{F}_i(\mathrm{w}) = \mathbb{E}_{\mathrm{x}_\mathrm{n}^\mathrm{i} \sim \mathcal{X}_i, \mathrm{y}_\mathrm{n}^\mathrm{i} \sim \phi_\mathrm{w}^i(y|x)} \left\| g\left(\mathrm{x}_\mathrm{n}^\mathrm{i}, \mathrm{y}_\mathrm{n}^\mathrm{i}\right)\right\|^2 \tag{2}$$

where $\left\{\mathrm{x}_\mathrm{n}^\mathrm{i}, \mathrm{y}_\mathrm{n}^\mathrm{i}\right\}$ represents a data sample and $g\left(\mathrm{x}_\mathrm{n}^\mathrm{i}, \mathrm{y}_\mathrm{n}^\mathrm{i}\right)$ is the gradient, which can be calculated as $g\left(\mathrm{x}_\mathrm{n}^\mathrm{i}, \mathrm{y}_\mathrm{n}^\mathrm{i}\right) = \frac{\partial \mathfrak{l}_i\left(\mathrm{w}; \mathrm{x}_\mathrm{n}^\mathrm{i}, \mathrm{y}_\mathrm{n}^\mathrm{i}\right)}{\partial \mathrm{w}}$. Then, the Federated FIM can be denoted as:

$$\mathcal{F}(\mathrm{w}) = \sum_{i=1}^{K} \mathfrak{m}_i \mathcal{F}_i(\mathrm{w}) \tag{3}$$

However, calculating the trace of the FIM is computationally expensive, which poses a challenge for real-time application during training. To address this, and drawing inspiration from (Yan et al. 2023), we identify the CLP utilizing Federated Gradient Norm (FGN). With the learning rate η, we start by expressing the variance in training losses as follows:

$$\Delta \mathfrak{l}_i\left(\mathrm{w}; \mathrm{x}_\mathrm{n}^\mathrm{i}, \mathrm{y}_\mathrm{n}^\mathrm{i}\right) = -\eta \left\| g\left(\mathrm{x}_\mathrm{n}^\mathrm{i}, \mathrm{y}_\mathrm{n}^\mathrm{i}\right)\right\|^2, \tag{4}$$

which is approximated using the Taylor expansion. The FGN is subsequently computed by the equation:

$$FGN\left(t\right) = \sum_{i=1}^{K} \mathfrak{m}_i \Delta \mathfrak{l}_i\left(\mathrm{w}; \mathrm{x}_\mathrm{n}^\mathrm{i}, \mathrm{y}_\mathrm{n}^\mathrm{i}\right)^{(t)}, \tag{5}$$

where t is the communication round. As shown in Figure 3, the calculation of FGN generates a curve that initially rises and then declines, providing an approximate estimation of the duration of CLP. The light blue shaded area in the figure represents the CLP interval. Crucially, FGN can be computed in real-time, whereas FIM calculations necessitate the completion of the entire training process. This distinction allows for immediate adjustment during the CLP, which is not feasible when relying solely on FIM.

Period-Aware Adaptive Aggregation

As aforementioned, for Figure 1, when the degree of statistical heterogeneity is high, the loss of a single client exhibits significant fluctuations in the early stages of training. A natural and intuitive solution is to increase the proportion of clients with lower volatility during aggregation while reducing the proportion of those with higher volatility, thereby enhancing the overall stability of the FL system. However, the drawback of this approach is evident—it may cause FL system to lose too much information from the highly volatile clients, ultimately compromising the model's performance. In our work, we only focus on the specific period for adaptive aggregation, maintaining all the useful knowledge of each client as well as alleviating negative impact. Next, we describe the two key steps in detail.

Period Check To enhance the stability of the FL system, instead of adjusting the client aggregation coefficients throughout the entire training process, we dynamically adjust each client's contribution to the model aggregation based on their performance during the CLP, a period where information entropy is the highest and has the most significant impact on the model's final accuracy. As depicted in Figure 3, the presence of the CLP is indicated by a continuous rise in terms of FGN. This distinctive growth trend enables the application of a slope-based detection method to confirm the occurrence of the CLP. Motivated by this, we detect the presence of a CLP by verifying if the following inequality holds:

$$\frac{FGN\left(t\right) - FGN\left(t-1\right)}{FGN\left(t-1\right)} > \delta, \tag{6}$$

where δ is a hyperparameter to adjust the sensitivity for detecting the CLP.

We find that the Period Check module leads to a considerable enhancement in model performance. Our ablation studies in the experiment section validate the effectiveness of the Period Check module in PA^{3}Fed.

Adaptive Aggregation After the clients upload α_i to the central server, the server will perform *Period Check*. If the system is within the CLP, the aggregation coefficient is set to $\alpha_i^{'} = \alpha_i$. If it is not, the aggregation coefficient defaults to $\alpha_i^{'} = 1$. Finally, the server performs *Adaptive Aggregation* with $\alpha_i^{'}$.

In this part, we illustrate how we design adaptive aggregation to replace the traditional FedAvg. Formally, the selected clients first download w^{t}, the global model, from the previous round to conduct local training. For a single data point, we can define the output logits as:

$$f\left(\mathrm{x}_{\mathrm{n}}^{\mathrm{i}}\right) = \phi^{i}\left(\mathrm{w}^{\mathrm{t}}; \mathrm{x}_{\mathrm{n}}^{\mathrm{i}}\right) \tag{7}$$

Then, we can evaluate the performance of a single client by a score and accuracy defined through the whole local dataset $\mathrm{D_i} = \{\mathcal{X}_i, \mathcal{Y}_i\}$ as follows:

$$\mathrm{s}_{\mathrm{i}}^{\mathrm{t}} = \mathbb{E}_{\mathrm{x}_{\mathrm{n}}^{\mathrm{i}} \sim \mathcal{X}_i}\left[-log\left(softmax(f\left(\mathrm{x}_{\mathrm{n}}^{\mathrm{i}}\right))_{\mathrm{y}_{\mathrm{n}}^{\mathrm{i}}}\right)\right], \tag{8}$$

$$Accuracy = \mathbb{E}_{\mathrm{x}_{\mathrm{n}}^{\mathrm{i}} \sim \mathcal{X}_i}\left[\mathbb{1}_{\mathrm{y}_{\mathrm{n}}^{\mathrm{i}} = y_p}\right], \tag{9}$$

where $\mathrm{y}_{\mathrm{n}}^{\mathrm{i}}$ represents the ground truth label, and y_p is the prediction. The score serves as an indicator of a client's performance throughout the training process, and ideally, it should show a gradual increase. However, due to statistical heterogeneity, achieving this in an FL system is challenging. To address this issue, the aggregation coefficient is adjusted based on the difference in scores between consecutive rounds. The concrete aggregation coefficient α can be defined as:

$$\alpha_i = exp\left(-\beta * \left(\mathrm{s}_{\mathrm{i}}^{\mathrm{t}-\sigma} - \mathrm{s}_{\mathrm{i}}^{\mathrm{t}}\right)\right), \tag{10}$$

where β is a hyperparameter that controls the degree of adjustment, σ represents the interval between rounds in which a single client is selected. By applying α_i, clients with greater stability are given more weight in the model aggregation, while those with lower stability are assigned reduced weight. In this way, we are able to achieve adaptive aggregation for improved FL performance.

Algorithm 1: Pipeline of PA^3Fed

```
for $t = 1, 2, \cdots, T$ do
    Server randomly selects $K^t$ clients
    $\mathrm{w}_1^{\mathrm{t}}, \mathrm{w}_2^{\mathrm{t}}, \mathrm{w}_3^{\mathrm{t}}, \cdots, \mathrm{w}_{\mathrm{K}^{\mathrm{t}}}^{\mathrm{t}} \leftarrow \mathrm{w}_{\mathrm{global}}^{\mathrm{t}}$
    // Client Side
    for $i = 1, 2, \cdots, K^t$ do
        $\hat{\mathrm{w}}_{\mathrm{i}}^{\mathrm{t}} \leftarrow LocalTraining\left(\mathrm{w}_{\mathrm{i}}^{\mathrm{t}}\right)$
        Calculate $\mathrm{s}_{\mathrm{i}}^{\mathrm{t}}$ using equation (8)
        Calculate $\alpha_i^t$ using equation (10)
        Calculate Local Gradient Norm $g_i^t$
        Upload $\alpha_i^t, g_i^t$ and $\hat{\mathrm{w}}_{\mathrm{i}}^{\mathrm{t}}$ to Server
    end for
    // Server Side
    Calculate $FGN(t)$ using equation (5)
    Perform Period Check using equation(6)
    if $In\ CLP$ then
        $\alpha_i^{'} = \alpha_i$
    else
        $\alpha_i^{'} = 1$
    end if
    Aggregate $\hat{\mathrm{w}}_{\mathrm{i}}^{\mathrm{t}}$ with factor $\mathfrak{m}_i * \alpha_i^{'}$
end for
```

Experiment

Experiments Setup

Datasets and Heterogeneity Setting

We employ three of the most commonly used image classification datasets in federated learning: CIFAR-10 (Krizhevsky, Hinton et al. 2009), CIFAR-100 (Krizhevsky, Hinton et al. 2009), and Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017). To optimize performance across these datasets, we select neural network models best suited to each: AlexNet (Krizhevsky, Sutskever, and Hinton 2012) for CIFAR-10 and FMNIST, and ResNet-18 (He et al. 2016) for CIFAR-100. The batch size is set to 32, with an initial learning rate of 0.01 that decays by a constant factor of 0.8 after each communication round. Additionally, we applied a momentum of 0.9 and a weight decay of 10^{-5}. All our experiments are simulated and executed on a server configured with three NVIDIA GeForce RTX 3090 GPUs, 48 Intel Xeon CPU cores, and 128GB of RAM.

Our FL system involves 50 clients, with 20% of them randomly selected in each round. Each client conducts 5 local training rounds and the global model is trained for a total of 250 rounds. The detection threshold of CLP is set to 0.01. The hyperparameter β, which controls the modification degree of the aggregation, is fixed at 0.3. To simulate a non-i.i.d. federated learning scenario, we implement a

Dataset	Model	Non-IID Degree	FedAvg		FedProx		FedNova		VRL-SGD		FedMut	
			Normal	Ours	Normal	Ours	Normal	Ours	Normal	Ours	Normal	Ours
CIFAR-10	AlexNet	h = 0.1	38.66	**53.99**	40.33	**55.02**	39.38	**54.74**	42.54	**57.75**	41.47	**55.72**
		h = 0.2	44.25	**58.85**	44.84	**58.85**	46.79	**58.90**	44.54	**58.16**	45.44	**59.56**
		h = 0.3	48.14	**58.94**	48.17	**58.66**	49.30	**59.19**	48.61	**60.30**	47.56	**60.36**
CIFAR-100	ResNet	h = 0.1	28.33	**36.97**	30.06	**36.59**	30.52	**37.01**	29.98	**37.12**	31.23	**38.27**
		h = 0.2	31.84	**41.01**	33.03	**40.86**	32.65	**41.21**	32.78	**41.34**	34.47	**42.23**
		h = 0.3	35.99	**42.23**	37.93	**43.66**	37.95	**42.08**	39.01	**43.57**	40.92	**44.79**
FMNIST	AlexNet	h = 0.1	74.40	**78.88**	74.67	**79.02**	74.54	**80.11**	74.62	**79.41**	75.01	**79.90**
		h = 0.2	76.94	**81.60**	76.99	**81.04**	76.24	**81.23**	76.38	**80.27**	77.09	**82.04**
		h = 0.3	77.30	**81.86**	77.58	**81.99**	77.42	**81.42**	78.21	**81.04**	78.12	**81.98**

Table 2: **Final Test Accuracy of State-of-the-Art FL Algorithms.** The table displays the final test accuracy for various state-of-the-art federated learning (FL) algorithms, labeled under the "Normal" columns, compared with the corresponding PA^{3}Fed augmentation, labeled under the "Ours" columns. This table highlights the performance improvements achieved by PA^{3}Fed over the conventional approaches.

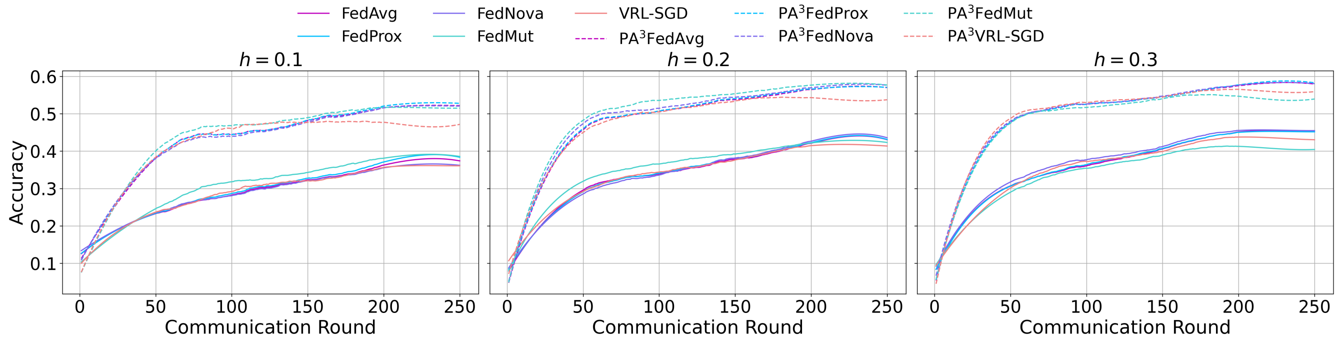


Figure 4: **Comparison of FL Methods and Their PA^{3}Fed Incorporated Versions on CIFAR-10.** The results highlight that PA^{3}Fed consistently enhances the model performance, showcasing significant improvements across non-i.i.d. conditions.

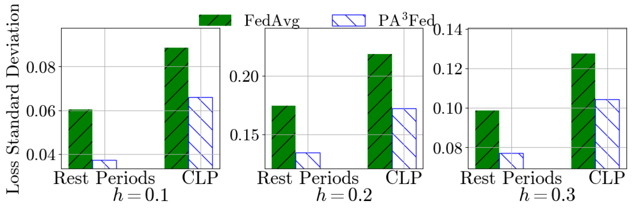


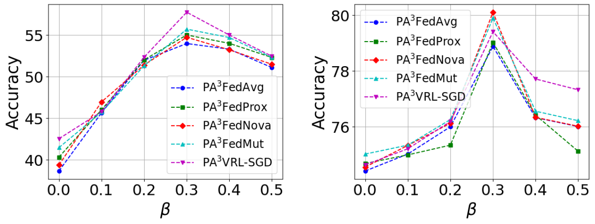
Figure 5: **Loss Standard Deviation under Different h.** The standard deviation of loss measures the stability of a model, with lower values indicating higher stability. Rest Periods refer to entire training periods excluding CLP.

heterogeneous partitioning strategy, ensuring that both the number of data points and class distributions are unbalanced across clients. Specifically, data are distributed among K clients, with client data proportions $\mathfrak{m}_i$ sampled from a Dirichlet distribution parameterized by h. We choose values of $h = 0.1, 0.2, 0.3$ in alignment with the methodology outlined in (Wang et al. 2020a,b; Yan et al. 2023). A lower value of h indicates a higher degree of non-i.i.d. in the data distribution, thereby creating more challenging and realistic scenarios for federated learning. By varying h, we are able to thoroughly examine how our method performs across different levels of data heterogeneity. All the experiments are conducted three times and we average them as the reported results.

Baseline. To rigorously assess the effectiveness of our proposed method, we integrate it into five state-of-the-art federated learning approaches: (1) FedAvg (McMahan et al. 2017), (2) FedProx (Li et al. 2020), (3) FedNova (Wang et al. 2020b), (4) VRL-SGD (Liang et al. 2019), and (5) FedMut (Hu et al. 2024). For each approach, we conduct a comparative analysis between the original method and the version augmented with PA^{3}Fed. This comparison provides insight into the performance improvements introduced by PA^{3}Fed, allowing us to measure its impact on these well-established techniques.

Overall Performance

Towards a Higher Accuracy Performance. We evaluated our approach against five state-of-the-art methods. Setting $\beta = 0.3$, we tested across various image classification datasets and models, including CIFAR-10 (AlexNet), CIFAR-100 (ResNet-18), and FMNIST (AlexNet). As demonstrated in Table 2, our method shows significant improvement across all datasets. Notably, under high heterogeneity conditions ($i.e., h = 0.1$), our method achieves an approximate 15% improvement on CIFAR-10.



(a) CIFAR-10.

(b) FMNIST.

Figure 6: **The Impact of β.** This figure illustrates the effect of varying β values on model performance.

Towards a Faster Convergence. As shown in Figure 4, which illustrates the accuracy curves of all methods on the CIFAR-10 dataset, we observe that methods augmented with PA^{3}Fed converge more rapidly. This demonstrates that with adaptive aggregation, besides the final accuracy, the model convergence can also be facilitated.

Towards a More Stable Local Update. We evaluate our adaptive aggregation against FedAvg in terms of loss standard deviation to demonstrate the superiority of our approach. To assess the improvement in stability during local training, we record the loss trajectories of all 50 clients and calculate the average loss standard deviation over entire training periods excluding the CLP and specifically during the CLP. Our analysis reveals that the loss standard deviation is higher during CLP compared to rest periods. As shown in Figure 5, PA^{3}Fed, represented by the blue bar, consistently outperforms the original FedAvg, depicted by the green bar. Specifically, PA^{3}Fed demonstrates a lower loss standard deviation during both the CLP and rest periods, indicating enhanced stability in local updates. This result aligns with one of our key goals: improving stability in local training.

Ablation Studies

The hyperparameter β, which controls the adaptation degree during the aggregation process. We select β from the candidate values $\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. As β increases, the degree of adaptation also increases. As shown in Figure 6, we find that an optimal value for β is 0.3, which effectively balances the aggregation factor. A very large β can lead to a disproportionate influence from more stable clients, potentially causing the model to lose excessive information. Conversely, a very small β may result in insufficient influence from stable models, failing to guide the global model towards a better and flatter convergence region. However, no matter how to select the value, it performs better than the original ones, which validates the effectiveness of our adaptive aggregation.

The existence of CLP. To validate the efficiency of adaptive aggregation during the CLP, we conduct an ablation experiment. As shown in Figure 7, the bar chart represents, from left to right, FedAvg, PA^{3}Fed with adaptation applied throughout entire training periods, and PA^{3}Fed. We observe that the model with adaptation during the CLP achieves the best accuracy. Other FL methods exhibit similar performance. Furthermore, since the CLP is a relatively short period, applying adaptation during this period demonstrates the efficiency of our approach.

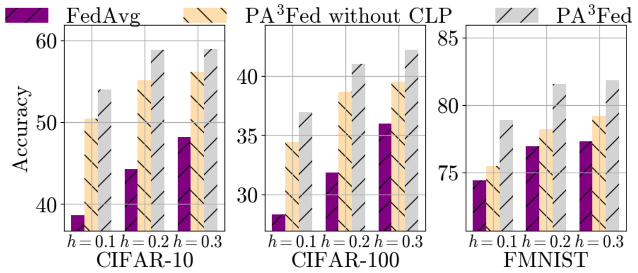


Figure 7: **The Impact of CLP.** This figure compares the model performance across different scenarios: no adaptation, adaptation throughout the entire training phase, and adaptation only during the Critical Learning Periods.

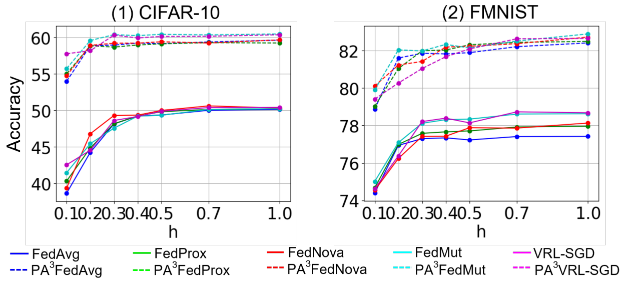


Figure 8: **The Impact of non-i.i.d. Degree.** This figure illustrates the effect of varying non-i.i.d. degrees on model performance.

The impact of non-i.i.d. degree. In this part, we conduct ablation experiments with various values of h (i.e., non-i.i.d. degree ranging from high to low). As shown in Figure 8, we observe that our method achieves greater performance improvements with higher non-i.i.d. degrees on both CIFAR-10 and FMNIST datasets, validating the effectiveness of our method in non-i.i.d. scenarios. Additionally, even as the non-i.i.d. degree decreases, our method continues to deliver strong results, demonstrating its robustness across different levels of data heterogeneity.

Conclusion

In this paper, we propose PA^{3}Fed, a method that leverages performance variations of individual clients (intra-client information) to assess model stability and perform adaptive aggregation. Experimental results demonstrate that PA^{3}Fed improves both performance and stability across multiple state-of-the-art methods.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback. This work was partly supported by the National Natural Science Foundation of China (62302054) and CCF-Huawei Populus Euphratica Fund (System Software Track, CCF-HuaweiSY202410).

References

Achille, A.; Rovere, M.; and Soatto, S. 2018. Critical learning periods in deep networks. In *International Conference on Learning Representations*.

Asad, M.; Moustafa, A.; and Ito, T. 2020. Fedopt: Towards communication efficiency and privacy preservation in federated learning. *Applied Sciences*, 10(8): 2864.

Bashan, A.; Bartsch, R.; Kantelhardt, J. W.; and Havlin, S. 2008. Comparison of detrending methods for fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, 387(21): 5080–5090.

Buckley, C.; and Voorhees, E. M. 2017. Evaluating evaluation measure stability. In *ACM SIGIR Forum*, volume 51, 235–242. ACM New York, NY, USA.

Chu, H.-M.; Geiping, J.; Fowl, L. H.; Goldblum, M.; and Goldstein, T. 2022. Panning for gold in federated learning: Targeted text extraction under arbitrarily large-scale aggregation. In *The Eleventh International Conference on Learning Representations*.

Cui, S.; Liang, J.; Pan, W.; Chen, K.; Zhang, C.; and Wang, F. 2022. Collaboration equilibrium in federated learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 241–251.

Fan, Y.; Xu, W.; Wang, H.; Zhu, J.; and Guo, S. 2023. Balanced Multi-modal Federated Learning via Cross-Modal Infiltration. *arXiv preprint arXiv:2401.00894*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, M.; Cao, Y.; Li, A.; Li, Z.; Liu, C.; Li, T.; Chen, M.; and Liu, Y. 2024. FedMut: Generalized Federated Learning via Stochastic Mutation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12528–12537.

Imteaj, A.; Mamun Ahmed, K.; Thakker, U.; Wang, S.; Li, J.; and Amini, M. H. 2022. Federated learning for resource-constrained iot devices: Panoramas and state of the art. *Federated and Transfer Learning*, 7–27.

Imteaj, A.; Thakker, U.; Wang, S.; Li, J.; and Amini, M. H. 2021. A survey on federated learning for resource-constrained IoT devices. *IEEE Internet of Things Journal*, 9(1): 1–24.

Jastrzębski, S.; Kenton, Z.; Ballas, N.; Fischer, A.; Bengio, Y.; and Storkey, A. 2018. On the relation between the sharpest directions of DNN loss and the SGD step length. *arXiv preprint arXiv:1807.05031*.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2021. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2): 1–210.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, 5132–5143. PMLR.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Li, Q.; He, B.; and Song, D. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10713–10722.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2: 429–450.

Liang, X.; Shen, S.; Liu, J.; Pan, Z.; Chen, E.; and Cheng, Y. 2019. Variance reduced local sgd with lower communication complexity. *arXiv preprint arXiv:1912.12844*.

Lin, T.; Kong, L.; Stich, S. U.; and Jaggi, M. 2020. Ensemble distillation for robust model fusion in federated learning. *Advances in neural information processing systems*, 33: 2351–2363.

Liu, B.; Cai, Y.; Bi, H.; Zhang, Z.; Li, D.; Guo, Y.; and Chen, X. 2023. Beyond Fine-Tuning: Efficient and Effective Fed-Tuning for Mobile/Web Users. In *Proceedings of the ACM Web Conference 2023*, 2863–2873.

Liu, B.; Cai, Y.; Zhang, Z.; Li, Y.; Wang, L.; Li, D.; Guo, Y.; and Chen, X. 2021. DistFL: Distribution-aware Federated Learning for Mobile Scenarios. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(4): 1–26.

Liu, B.; Guo, Y.; and Chen, X. 2021. PFA: Privacy-preserving Federated Adaptation for Effective Model Personalization. In *Proceedings of the Web Conference 2021*, 923–934.

Liu, B.; Lv, N.; Guo, Y.; and Li, Y. 2024. Recent advances on federated learning: A systematic survey. *Neurocomputing*, 128019.

Long, G.; Tan, Y.; Jiang, J.; and Zhang, C. 2020. Federated learning for open banking. In *Federated learning: privacy and incentive*, 240–254. Springer.

Martens, J. 2020. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146): 1–76.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

McMahan, H. B.; Yu, F.; Richtarik, P.; Suresh, A.; Bacon, D.; et al. 2016. Federated learning: Strategies for improving communication efficiency. In *Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain*, 5–10.

Mu, X.; Shen, Y.; Cheng, K.; Geng, X.; Fu, J.; Zhang, T.; and Zhang, Z. 2023. Fedproc: Prototypical contrastive federated learning on non-iid data. *Future Generation Computer Systems*, 143: 93–104.

Ruan, Y.; and Joe-Wong, C. 2022. Fedsoft: Soft clustered federated learning with proximal local updating. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 8124–8131.

Wang, H.; Yurochkin, M.; Sun, Y.; Papailiopoulos, D.; and Khazaeni, Y. 2020a. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*.

Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; and Poor, H. V. 2020b. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33: 7611–7623.

Wang, Q.; Liu, B.; and Li, Y. 2024. Traceable Federated Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12872–12881.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Yan, G.; Wang, H.; and Li, J. 2021. Critical learning periods in federated learning. *arXiv preprint arXiv:2109.05613*.

Yan, G.; Wang, H.; Yuan, X.; and Li, J. 2023. Criticalfl: A critical learning periods augmented client selection framework for efficient federated learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2898–2907.