FEDCDA: FEDERATED LEARNING WITH CROSS-ROUND DIVERGENCE-AWARE AGGREGATION

Haozhao Wang[1], **Haoran Xu**[2], **Yichen Li**[3], **Yuan Xu**[1], **Ruixuan Li**[3,*], **Tianwei Zhang**[4]
[1]S-Lab, Nanyang Technological University [2]Zhejiang University
[3]Department of Computer Science, Huazhong University of Science and Technology
[4]Nanyang Technological University
{hz_wang, rxli}@hust.edu.cn, tianwei.zhang@ntu.edu.sg

ABSTRACT

In Federated Learning (FL), model aggregation is pivotal. It involves a global server iteratively aggregating client local trained models in successive rounds without accessing private data. Traditional methods typically aggregate the local models from the current round alone. However, due to the statistical heterogeneity across clients, the local models from different clients may be greatly diverse, making the obtained global model incapable of maintaining the specific knowledge of each local model. In this paper, we introduce a novel method, FedCDA, which selectively aggregates cross-round local models, decreasing discrepancies between the global model and local models. The principle behind FedCDA is that due to the different global model parameters received in different rounds and the non-convexity of deep neural networks, the local models from each client may converge to different local optima across rounds. Therefore, for each client, we select a local model from its several recent local models obtained in multiple rounds, where the local model is selected by minimizing its divergence from the local models of other clients. This ensures the aggregated global model remains close to all selected local models to maintain their data knowledge. Extensive experiments conducted on various models and datasets reveal our approach outperforms state-of-the-art aggregation methods.

1 INTRODUCTION

Federated Learning (FL) has emerged as a key framework for training deep neural networks (DNNs) through client collaboration without the need to share original datasets (McMahan et al., 2017b; Wang et al., 2022; Li et al., 2022b). It has been extensively utilized in areas like medical image processing (Liu et al.; Guo et al.; Xu et al.) and recommendation systems (Ramaswamy et al., 2019; Ammad-ud-din et al., 2019). FL is an iterative procedure in which each round involves the local model training across various individual clients, and then aggregating these models centrally on a server (McMahan et al., 2017a).

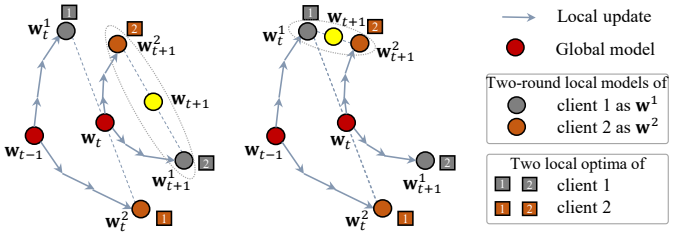
In this paper, we focus on the *aggregation* of FL, which is the critical step to obtain the global model from multiple local models. The typical aggregation method is FedAvg, which computes the coordinate-wise weighted average of parameters of local models with the weight as the ratio of the data size (McMahan et al., 2017b). Although the implementation of this method is straightforward, some works (Yurochkin et al., 2019a; Li et al., 2022b; Liu et al., 2022; Wang et al., 2020a) consider that the coordinate-wise average will reduce the performance due to the NonIID (i.e., not independently and identically) data among clients. Specifically, they identify that the parameter ordering of different local models may be varied due to the permutation invariance of neural network (NN) parameters. Thus, they propose re-ordering the parameters before applying the weighted average. Another type of work considers that the NonIID data also affects the aggregation weights and they propose adaptively setting the weights using a learnable approach (Li et al., 2023). Although these

[*]Haozhao Wang, Haoran Xu, and Yichen Li contribute equally to this work. Ruixuan Li is the Corresponding author.

methods have achieved great success separately, they mainly aggregate the local models from the current single round, which may limit the improvement of FL performance.

Orthogonal to these works, in this paper, we focus on the aggregation of cross-round local models to further unleash the potential aggregation performance. Intuitively, to acquire the data knowledge of some specific client, it is necessary for the global model to be close to its locally trained model (Kirkpatrick et al., 2017). Nevertheless, the local models of different clients in the same single round may have a large divergence from each other due to the statistical heterogeneity. Thus, as shown in Figure 1(a), the aggregated global model may greatly deviate from these local models. To tackle this challenge, we consider a common fact that each client is usually able to achieve convergence in different rounds after the startup training stage, especially when FL prefers a larger interval for the local training process to save the communication cost (Sun et al., 2023a). In addition, due to receiving different global models in different rounds and the existence of multiple local optima in deep neural networks (Wu et al., 2017; Kawaguchi, 2016; Xie et al., 2021), each client often converges to different models in different rounds, each of which can usually learn local data well, especially when training data using the most advanced optimizer (Loshchilov & Hutter, 2019; Chaudhari et al., 2017). Therefore, a natural idea is that the global model can also fit the local data of some specific client once it approaches any of the different local models in multiple rounds. Motivated by this, the global model can essentially be obtained by aggregating selected local models from different rounds to reduce their divergence and maintain their knowledge.

Based on the above motivation, we propose a novel aggregation method named `FedCDA`, which selectively aggregates cross-round local models. More specifically, we design a divergence-aware selection strategy that selects local models from multiple rounds with minimum divergence to their aggregated model and only aggregates the selected local models to obtain the global model. In this way, as shown in Figure 1(b), the global model approaches selected local models and thus maintains the data knowledge of clients. Considering the selection problem is a combinatorial optimization problem with a large search space, we further design an approximation version by selecting local models in a batch way to reduce the selection cost. Then, we establish theories to provide a better understanding and guarantee the convergence of our method. We conduct extensive experiments on various datasets, and the results show that `FedCDA` outperforms state-of-the-art baselines. Our contributions are:



(a) Single round (b) Cross round (c) Legend

Figure 1: Comparison of different aggregation methods on a two-client FL. (a) The global model $\mathbf{w}_{t+1}$ is the aggregation of the local model $\mathbf{w}_{t+1}^1$ of client 1 and $\mathbf{w}_{t+1}^2$ of the client 2 in the same $t+1$-th round. (b) $\mathbf{w}_{t+1}$ is the aggregation of the t-th round local model $\mathbf{w}_t^1$ and the $t+1$-th round local model $\mathbf{w}_{t+1}^2$. $\mathbf{w}_{t+1}$ *obtained through cross-round aggregation is close to the local optimum 1 of client 1 and the local optimum 2 of client 2, while the single-round aggregation is distant from any local optima.* A practical example is in Appendix A.

- To the best of our knowledge, this paper is the first to study the aggregation of cross-round local models. We identify that cross-round local models among clients may have a smaller divergence than those in the single round and selectively aggregating them can make the global model approach local models more closely, thus maintaining their local data knowledge.
- We propose a new cross-round aggregation method named `FedCDA`. It obtains the global model by aggregating local models selected from multiple rounds based on the criterion of the minimum divergence. Besides, we design an approximation strategy to reduce the cost of selection.
- We establish comprehensive theories for our method. Specifically, we provide theoretical insights for understanding our algorithm and show that the approximation selection error is bounded by the convergence of the local model. Besides, we also prove the convergence of our method.
- We conduct extensive experiments over various deep-learning models and datasets. The efficiency superiority of `FedCDA` is demonstrated by comparing our proposed aggregation method with traditional aggregation methods, which achieves the best performance.

2 RELATED WORKS

Many previous methods have been proposed to improve the performance of FL. For example, some works propose regularizing the update of the local model to mitigate the NonIID issue (Li et al., 2020; Sun et al., 2023b). Orthogonal to these works, this paper focuses on the aggregation of local models. Generally, there are three main types of FL aggregation methods.

Aggregation Weights One of the typical researches is to determine adaptive aggregation weights (Li et al., 2022a; Rehman et al., 2023). For instance, AUTO-FEDAVG (Xia et al., 2021) tailors weights based on distinct institutional medical datasets to enable personalized medicine, whereas L2C (Li et al., 2022a) identifies similar peers in decentralized FL by adapting weights using local data. While these approaches have proven effective, they primarily emphasize the creation of personalized models for individual clients. In contrast, our work centers on acquiring a global model. Recently, FedLAW (Li et al., 2023) aims to obtain a global model by learning the weights. Nevertheless, all of these methods rely on the proxy dataset in the server while our aggregation method does not.

Model Fusion Due to the permutation invariance of neural network parameters, some works consider that the parameters ordering of different local models across clients may be varied especially when their data is NonIID (Yu et al., 2021; Singh & Jaggi, 2020; Li et al., 2022b). In this case, the coordinate-wise average of local models will lead to a mismatch between the same-position parameters of cross-client local models, degrading the performance of the aggregated model. Hence, these works seek to fuse these local models by re-ordering the parameters to match them across clients such as using Hungarian matching algorithm (Wang et al., 2020a), Bayesian approach (Yurochkin et al., 2019b), or a graph matching algorithm (Liu et al., 2022).

Federated Distillation Different from the two above types of methods that compute the average of model parameters, federated distillation employs an ensemble distillation computing the average of their logits over the aggregation of local models (Wu & Gong, 2021; Guo et al., 2020; Bistritz et al., 2020; Wang et al., 2023). Notably, Lin et al. (2020); Chen & Chao (2021) initially introduced a technique that harnesses knowledge distillation on the server side. This approach transfers knowledge from multiple local models to the global model using an unlabeled proxy dataset. However, these methods depend on the availability of an auxiliary dataset on the server, which may not be present in real-world scenarios. In response to this limitation, recent studies (Zhu et al., 2021; Zhang et al., 2022; Wang et al., 2023) proposed replacing the proxy dataset with generated data, enabling ensemble federated distillation in a data-free manner. We in this paper focus on the average of model parameters, which is orthogonal to these works.

3 SETUP

Federated learning allows N clients with a server to solve the following optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} F_n(\mathbf{w}), \quad s.t., F_n(\mathbf{w}) = \mathbb{E}_{\xi\sim\mathcal{D}_n} f_n(\mathbf{w};\xi) \tag{1}$$

to obtain the global model $\mathbf{w}$. The function $F_n(\mathbf{w}) : \mathbb{R}^d \to \mathbb{R}$ denotes the expected loss over the data distribution of client n. $\mathcal{D}_n$ denotes the data distribution of the n-th client. $f_n(\mathbf{w};\xi)$ denotes the loss value with respect to model $\mathbf{w}$ and random data sample ξ. Without causing confusion, we use $f_n(\mathbf{w})$ to denote a mini-batch of $f_n(\mathbf{w};\xi)$ for simplicity. Besides, we make the following assumptions for these objectives which are widely adopted in FL (Dinh et al., 2020; Wang et al., 2020b).

Assumption 1 ***(L-smoothness)****. The objective function F_n is L-smooth with Lipschitz constant $L > 0$, i.e., $\|\nabla F_n(\mathbf{w}) - \nabla F_n(\mathbf{w}')\|_2 \leq L\|\mathbf{w} - \mathbf{w}'\|_2$ for all $\mathbf{w}$, $\mathbf{w}'$.*

Assumption 2 ***(Bounded Variance)****. For all parameters $\mathbf{w}$, the variance of the local stochastic gradient in each client is bounded by σ_l^2: $\mathbb{E}(\|\nabla f_n(\mathbf{w}) - \nabla F_n(\mathbf{w})\|^2) \leq \sigma_l^2$. Besides, the global variance of gradients among clients is bounded by σ_g^2: $\frac{1}{N}\sum_{n=1}^{N}\|\nabla F_n(\mathbf{w}) - \nabla F(\mathbf{w})\|^2 \leq \sigma_g^2$.*

Assumption 3 ***(Bounded Gradient)****. For all parameters $\mathbf{w}$, the stochastic gradient with respect to the loss is bounded by a constant M: $\mathbb{E}(\|\nabla f_n(\mathbf{w})\|^2) \leq M^2$.*

4 METHODOLOGY

In this part, we will introduce our proposed aggregation method. To minimize the objective (1), we first apply Assumption 1 to each local loss function $F_n(\mathbf{w})$:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} F_n(\mathbf{w}) \leq \frac{1}{N}\sum_{n=1}^{N}\left[F_n(\mathbf{w}^n) + \nabla_{\mathbf{w}^n} F_n(\mathbf{w}^n)(\mathbf{w}-\mathbf{w}^n) + L\|\mathbf{w}-\mathbf{w}^n\|_2^2\right]. \quad (2)$$

Then, we turn to minimize the upper bound of the objective function (1), which corresponds to the right-hand side term of the inequality (2), by aggregating local models $\mathbf{w}^n$ of each client n from multiple rounds. Given the set of recent K local models $\mathcal{W}_t^n = \{\mathbf{w}_{t_1}^n, \ldots, \mathbf{w}_{t_K}^n\}$ of each client n obtained in multiple rounds, the server seeks to solve the following objective:

$$\min_{\mathbf{w}\in\mathbb{R}^d, \mathbf{w}^1\in\mathcal{W}_t^1,\ldots,\mathbf{w}^N\in\mathcal{W}_t^N} \frac{1}{N}\sum_{n=1}^{N}\left[F_n(\mathbf{w}^n) + \nabla_{\mathbf{w}^n} F_n(\mathbf{w}^n)^T(\mathbf{w}-\mathbf{w}^n) + \frac{L}{2}\|\mathbf{w}-\mathbf{w}^n\|_2^2\right]. \quad (3)$$

The problem (3) is strongly convex in terms of $\mathbf{w}$ for any combination of the local models $\mathbf{w}^n$. Therefore, the global model $\mathbf{w}$ has a closed-form solution with respective to the local models $\mathbf{w}^n$:

$$\mathbf{w} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{w}^n - \frac{1}{LN}\sum_{n=1}^{N}\nabla_{\mathbf{w}^n} F_n(\mathbf{w}^n). \quad (4)$$

Given equation (4), the problem (3) is equivalent to a combinatorial optimization problem to select local models $\mathbf{w}^n$. However, solving this problem requires computing the full gradient $\nabla_{\mathbf{w}^n} F_n(\mathbf{w}^n)$ on the local dataset of each client n, leading to extra expensive computation and communication cost. Considering the local model $\mathbf{w}^n$ may nearly approach one of the local optima or saddle point $\mathbf{w}^{n,*}$ especially at the end of the FL training stage or when the number of local epochs is large, we take an approximation as $\nabla_{\mathbf{w}^n} F_n(\mathbf{w}^n) \approx 0$. The problem (3) can be re-formulated as:

$$\min_{\mathbf{w}^1\in\mathcal{W}_t^1,\ldots,\mathbf{w}^N\in\mathcal{W}_t^N} \frac{1}{N}\sum_{n=1}^{N} F_n(\mathbf{w}^n) + \frac{L}{2N}\sum_{n=1}^{N}\|\mathbf{w}-\mathbf{w}^n\|_2^2, \quad s.t., \mathbf{w} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{w}^n \quad (5)$$

$$\Leftrightarrow \min_{\mathbf{w}^1\in\mathcal{W}_t^1,\ldots,\mathbf{w}^N\in\mathcal{W}_t^N} \frac{1}{N}\sum_{n=1}^{N} F_n(\mathbf{w}^n) + \frac{L}{2N}\sum_{n=1}^{N}\|\mathbf{w}^n\|_2^2 - \frac{L}{2}\|\mathbf{w}\|_2^2, \quad s.t., \mathbf{w} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{w}^n. \quad (6)$$

Equation (5) reveals that the criterion for choosing local models can be understood as the selection of cross-round local models that exhibit minimal divergence among each other, i.e., variance $\frac{1}{N}\sum_{n=1}^{N}\|\mathbf{w}-\mathbf{w}^n\|_2^2$, particularly when the difference in loss $F_n(\mathbf{w}^n)$ tends to be small among clients. Although solving (5) can obtain the optimal combination of cross-round local models, the computation complexity and memory cost are large. An approach to reducing the computation cost is to utilize the equivalent version of (5), i.e., (6), which is derived using $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$ with $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$. In this way, the l_2 norm of $\|\mathbf{w}^n\|_2^2$ can be cached once it is computed to avoid repeated computations. Yet, the search space for all combinations is still large. Denoting the model size as C, we have the following conclusion.

Proposition 1 *The computation complexity of solving (6) is $\mathcal{O}(K^N)$ and the memory cost is KNC.*

Due to the exponential complexity of computation, directly solving (6) is not affordable even by a cloud for large K and N. Therefore, we further propose selecting local models with approximately minimum divergence to reduce the cost. Our strategy includes two steps.

First, selection for partial clients. We propose only selecting local models from P clients $n \in \mathcal{P}_t$ that participate in the current round t and fixing the local models of other clients $n \in \mathcal{N}-\mathcal{P}_t$ by using those selected in previous rounds:

$$\min_{\mathbf{w}^n\in\mathcal{W}_t^n,\forall n\in\mathcal{P}_t} \frac{1}{N}\sum_{n\in\mathcal{P}_t}\mathcal{L}_n(\mathbf{w}^n) + \underbrace{\frac{1}{N}\sum_{n\in\mathcal{N}-\mathcal{P}_t}\mathcal{L}_n(\mathbf{w}^n)}_{\text{Fixed in current round}} - \frac{L}{2}\|\mathbf{w}\|_2^2, \quad (7)$$

where the aggregated model $\mathbf{w}$ remains the same as $\mathbf{w} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{w}^n$ and $\mathcal{L}_n(\mathbf{w}^n)$ denotes $\mathcal{L}_n(\mathbf{w}^n) = F_n(\mathbf{w}^n) + \frac{L}{2}\|\mathbf{w}^n\|_2^2$. As the local models of non-participating clients are fixed in the current round, this leads to a great reduction in computational complexity and memory requirements.

Proposition 2 *The computation complexity of solving (7) is $\mathcal{O}(K^P)$ and the memory cost is KPC.*

Second, batch-based selection. To further reduce the computation complexity, we propose selecting local models in a stochastic greedy manner. Specifically, we randomly group participated clients into B equal-size batches $\mathcal{P}_t = \mathcal{P}_t^1 \cup \cdots \cup \mathcal{P}_t^B$ and select local models for these clients batch by batch. When selecting local models for clients in the b-th batch, the local models of clients in batches 1 to $b-1$ are fixed and in batches $b+1$ to P are excluded, and the objective is:

$$\min_{\mathbf{w}^n \in \mathcal{W}_t^n, \forall n \in \mathcal{P}_t^b} \frac{1}{N-P+\frac{bP}{B}} \Bigg(\sum_{n \in \mathcal{P}_t^b} \mathcal{L}_n(\mathbf{w}^n) + \underbrace{\sum_{n \in (\mathcal{N}-\mathcal{P}_t) \cup \mathcal{P}_t^1 \cup \cdots \cup \mathcal{P}_t^{b-1}} \mathcal{L}_n(\mathbf{w}^n)}_{\text{Fixed in the } b\text{-th subset selection}} \Bigg) - \frac{L}{2}\|\mathbf{w}\|_2^2, \quad (8)$$

where the model $\mathbf{w}$ is aggregated by computing the average of local models of non-participated clients and the 1-st to b-th batch of participated clients, i.e., $\mathbf{w} = \frac{1}{N-P+\frac{bP}{B}} \sum_{n \in (\mathcal{N}-\mathcal{P}_t) \cup \mathcal{P}_t^1 \cup \cdots \cup \mathcal{P}_t^b} \mathbf{w}^n$. This can also be viewed as selecting local models that are close to that of clients participating in previous rounds and thus maintaining the memory of their data. The complexity of the computation is further reduced.

Proposition 3 *The computation complexity of solving (8) is $\mathcal{O}(BK^{\frac{P}{B}})$ and the memory cost is KPC.*

While the computational complexity remains exponential, we retain the flexibility to manually adjust the value of B for control. In practice, we can maintain $\frac{P}{B}$ as a constant, effectively reducing the complexity to an acceptable level. In an extreme scenario, we can set $B = P$, resulting in linear complexity with respect to the value of K. Our experiments have shown that even a small value of K, such as $K = 3$, produces satisfactory performance, rendering the computational complexity acceptable for practical applications. Additionally, the memory cost KPC is also manageable when K is small, because the number of sampled clients P is usually a small ratio of the total clients. The local models of non-participated clients can be stored on the disk which has sufficient storage space. For example, a 1TB hard drive can store approximately $20,000$ copies of ResNet-18, which is widely adopted on the edge. Given that even a mobile phone is equipped with 1 TB storage, we believe that the cost is within the budget of the aggregation node which is typically hosted by a cloud.

As compared to other aggregation methods like weight setting (Li et al., 2023) or ensemble distillation Lin et al. (2020); Chen & Chao (2021), our approach has a distinct advantage. We do not depend on an additional public dataset, which can be challenging to acquire due to the requirement for a similar distribution as the global dataset. Moreover, our method does not introduce higher computational complexity compared to existing methods. Many existing aggregation methods involve performing gradient descent (Li et al., 2023; Chen & Chao, 2021) or solve maximum bipartite matching problems (Wang et al., 2020a), which can be computationally intensive.

4.1 FedCDA Algorithm

The complete procedure of our method is given in Algorithm 1 by assuming the base algorithm is FedAvg (McMahan et al., 2017a). `FedCDA` differs from FedAvg primarily in lines 9 and 10 of its implementation. When it receives local models from a subset of clients, the server updates its cached local models. This update involves replacing the oldest round's local model with the most recently received one, as indicated in line 9. After this update, the server selects local models for aggregation by solving the problem (8) in line 10. Finally, the global model is obtained by averaging both selected and fixed local models to retain knowledge contributed by all clients.

In practice, *we usually apply `FedCDA` after a warmup training stage* using FedAvg or other baselines to ensure that the local models can approach the convergence during the local training process. It is also worthwhile to note that most existing methods usually employ improved techniques over the step of line 11, e.g., re-setting aggregation weights (Li et al., 2023) or using ensemble distillation (Lin et al., 2020; Chen & Chao, 2021), which are orthogonal to us.

5 Theoretical Analysis

In this section, we provide theories for better understanding the principles and bounding the error of the proposed algorithm. We first prove that the selection error of using the approximated objective

Algorithm 1 `FedCDA` Algorithm

Input: Number of cached local models K, number of subsets B, learning rate η, number of sampling clients P, and total communication rounds T.
Output: Converged global model $\mathbf{w}$.

```
Initialize the model parameter w_0;
Distribute w_0 to all clients;
for each communication round t ∈ {1, 2, ..., T} do
    Randomly select a set of clients P_t;
    for each selected client n ∈ P_t in parallel do
        Initialize the local model with the received global model: w^n = w_t;
        Solve the local problem by updating w_n for E local mini-batch SGD steps and accumulate the local loss F_n(w^n) in the last local epoch: w^n = w^n − η∇_{w^n} f_n(w^n);
    Update the cached set W_t^n for n ∈ P_t by replacing the oldest model with received w^n;
    Select local models w^n for each client n ∈ P_t by solving the problem (8);
    Aggregate both selected and fixed local models to obtain: w_{t+1} = (1/N) Σ_{n=1}^N w^n;
return global model w_T
```

(5) to the exact objective (3) is bounded by the convergence degree of local models. Then, we present the benefits of `FedCDA` on idealized cases and conditions. Finally, we establish the convergence theories for our algorithm.

Theorem 1 *(**Approximation Selection Error**) Define the local optima closest to $\mathbf{w}_t^n$ as $\mathbf{w}_t^{n,*}$ and the maximum distance between any two local optima that are close to cached local models across clients and rounds as D, i.e., $D = max_{n\in[N],n'\in[N],n\neq n',i\in[K],i'\in[K]}(\|\mathbf{w}_{t_i}^{n,*} - \mathbf{w}_{t_i'}^{n',*}\|)$. If the distance between the local model $\mathbf{w}_t^n$ and its approximated critical point $\mathbf{w}_t^{n,*}$ is limited by a constant $\epsilon > 0$, i.e., $\|\mathbf{w}_t^n - \mathbf{w}_t^{n,*}\| \leq \epsilon$, then the disparity in the global loss between aggregating local models selected using (5) and (3) is constrained by $\varepsilon \leq 4L\epsilon^2 + 2LD\epsilon$.*

The proof can be found in Appendix B.1. The theorem indicates that the approximation error of using (5) to (3) becomes smaller when local models are convergent. It implicitly reveals that our algorithm may obtain a better global model when the local models approach convergence, i.e., with large local iterations or large warmup rounds, which are verified by our experimental results in Figure 2(c) and Figure 6.2. Further, we seek to show that minimizing (3) leads to a lower global loss than naively aggregating the local models in the newest current round. We define the divergence among local models $\mathbf{w}^n, \forall n = 1 \ldots, N$ as $\text{Var}(\mathbf{w}^n) = \frac{1}{N}\sum_{n=1}^N \|\frac{1}{N}\sum_{n=1}^N \mathbf{w}^n - \mathbf{w}^n\|_2^2$. We denote $\mathbf{w}_{t_*}, \mathbf{w}_{t_*}^n, n = 1, \ldots, N$ as the solution of objective (3). Similarly, we denote $\mathbf{w}_t$ as the t-th round global model aggregated from all t-th round local models $\mathbf{w}_t^n, \mathbf{w}_{t_*}^n, n = 1, \ldots, N$. Subsequently, we demonstrate that the global loss $F(\mathbf{w}_{t_*})$ can be assured to be lower than $F(\mathbf{w}_t)$ under the condition that the divergence $\text{Var}(\mathbf{w}_{t_*}^n)$ is less than $\text{Var}(\mathbf{w}_t^n)$ by a certain value.

Theorem 2 *(**Impact of Divergence of Local Optima**) Let the definition of the local optima $\mathbf{w}_t^{n,*}$ and distance ϵ be the same as Theorem 1. Consider the loss function $F_n(\mathbf{w})$ is strongly convex with a parameter μ within the region spanning from the local optima $\mathbf{w}_t^{n,*}$ to the global model $\mathbf{w}_t$ and the local loss achieves equivalent values on local optima in different rounds, i.e., $F_n(\mathbf{w}_t^{n,*}) = F_n(\mathbf{w}_{t'}^{n,*})$. If the divergence among selected local models is small enough, i.e., satisfying $Var(\mathbf{w}_{t_*}^{n,*}) \leq \frac{\mu}{L}Var(\mathbf{w}_t^{n,*}) - (\frac{\mu}{L}+1)\epsilon^2$, then the global loss of using selected global model $\mathbf{w}_{t_*}$ is smaller than that of using t-th round global model $\mathbf{w}_t$, i.e., $F(\mathbf{w}_{t_*}) \leq F(\mathbf{w}_t)$.*

The proof can be found in Appendix B.2. Although the conditions of Theorem 2 may be idealized in practical settings, it provides some insights for understanding our method. Smaller divergence among local models leads to a smaller loss of the aggregated model. An ideal case is that the divergence is reduced to 0 where the local optima of all local models across clients are the same. In fact, such an ideal case can widely exist in overparameterized deep neural networks, where a large model may achieve 0 loss in the local dataset of each client and hence is the local optima of all clients. Therefore, our method may prefer large models. The experimental results in Table 1 also verify our statement, where the improvement is higher for the larger models. Although our

motivation mainly comes from the non-convex functions where there are multiple local optima, our algorithm is also applicable to convex cases. More discussions can be found in Appendix B.3. Finally, we present the convergence of our algorithm. Noting that even though our algorithm does not achieve faster theoretical convergence using existing optimization analytical tools, our algorithm demonstrates great empirical benefits.

Theorem 3 ***(Convergence on Non-convex Functions)*** *Consider problem (1) under Assumption 1,2, and 3. If the learning rate η satisfies $0 < \eta \leq \frac{1}{LE}$, then the global model $\mathbf{w}_{t_*}$ solved by (3) achieves asymptotic convergence, i.e., $\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(\mathbf{w}_{t_*})\|_2^2 = \mathcal{O}(\frac{1}{\sqrt{T}})$.*

Ideas of Proof: Our proof mainly includes two parts. First, we prove that the difference between the loss of the global model $\mathbf{w}_{t_*}$ obtained by (3) and that of the reference global model $\mathbf{w}_t$ obtained by aggregating the newest local models is bounded. Then, we prove that the loss of the global model $\mathbf{w}_t$ achieves convergence, which in turn indicates the convergence of the global model $\mathbf{w}_{t_*}$. Detailed derivations are deferred to Appendix B.4.

6 EVALUATION

6.1 EXPERIMENTAL SETUP

Datasets and Models: We consider three popular datasets in experiments: Fashion-MNIST (Xiao et al., 2017), CIFAR-10 (Krizhevsky et al., 2009) and CIFAR-100 (Krizhevsky et al., 2009), which contains 10, 10, 100 classes respectively. For CIFAR-10 and CIFAR-100 datasets, we use ResNet-18 (He et al., 2016) as the backbone to train and test the performance while for Fashion-MNIST we use a simple CNN instead. The simple CNN has two 5x5 convolution layers (the first with 32 channels, the second with 64, each followed with 2x2 max pooling), a fully connected layer with 512 units and ReLu activation, and a final fully output connected layer.

Data Partition: To evaluate the performance of our work in a heterogeneous scenario, we specify two Non-IID data partition methods called Shards (McMahan et al., 2017a) and Dirichlet (Lin et al., 2020). In the Shards setting, the sorted samples are shuffled into $N * S$ shards, and assigned to N clients randomly. Each client owns an equal number of pieces. In the second setting, data distribution over clients satisfies the Dirichlet distribution by using α to characterize the degree of heterogeneity. We set α of Dirichlet: $\{0.1, 0.3, 0.5\}$ and shards for each client: $\{2, 4, 8\}$.

Baselines: Beside of FedAvg (McMahan et al., 2017a), we also compare against various types of efficient federated learning approaches with the proposed method in our experiments. The first main type includes typical non-aggregation methods that speed FL in the local process or tuning learning rate, including FedProx (Li et al., 2020), FedExP (Jhunjhunwala et al., 2023), and FedSAM (Qu et al., 2022). The methods of the second type can be divided into three main representative aggregation categories: ensemble distillation including FedDF (Lin et al., 2020) and FedGEN (Chen & Chao, 2021); model fusion including FedMA (Wang et al., 2020a) and GAMF (Liu et al., 2022); weights setting including FedLAW (Li et al., 2023).

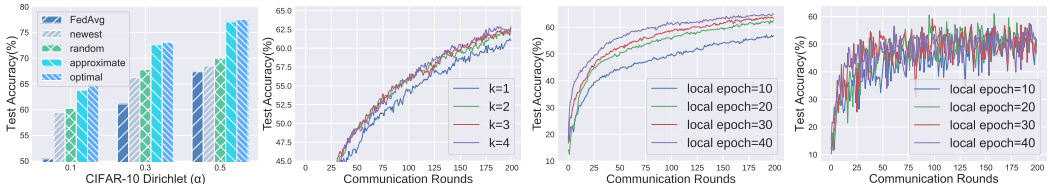
Implementation: We implement the whole experiment in a simulation environment based on PyTorch 2.0 and 8 NVIDIA GeForce RTX 3090 GPUs. We use 20 clients in total and randomly choose 20% each round for local training. We set the local epoch to 20, batch size to 64, and learning rate to $1e-3$. We employ SGD optimizer with momentum of $1e-4$ and weight decay of $1e-5$ for all methods and datasets. At the same time, we set the number of global communication rounds to 200. Each experiment setting is run twice and we take each run's final 10 rounds' accuracy and calculate the average value and standard variance. For our method, we also need to set the memory size of the client K to 3, batch number B to 3, and the number of warmup rounds to 50. Besides, we simply assume $L = 1$ for all clients to save the computation cost.

6.2 EXPERIMENT RESULTS

Performance Comparison. We report the comparison results with other baselines in Table 1. The results with a broader range of hyperparameters can be found in Appendix D. In order to demonstrate the generalization of our method, we compare them on two different Non-IID settings, Shards and Dirichlet distribution. We apply different data distributions on different datasets. We can see that our proposed `FedCDA` achieves the best performance on almost all settings. It demonstrates the effectiveness and benefit of cross-round divergence-aware aggregation. Specifically, on relatively

Table 1: The comparison of test accuracy of different methods. The best results are **bolded**.

Method	Fashion-MNIST(%)			CIFAR-10(%)			CIFAR-100(%)		
Shards (S)	2	4	8	2	4	8	2	4	8
FedAvg	64.69±5.62	74.78±4.55	76.81±3.33	28.10±3.96	59.83±2.94	70.87±1.91	11.86±1.19	15.87±1.00	21.91±0.55
FedProx	64.21±4.11	70.76±3.89	72.19±4.16	26.39±4.16	53.03±2.29	70.91±1.87	10.87±0.58	15.37±0.46	24.16±0.33
FedExP	65.24±3.47	69.31±4.62	76.66±5.04	26.84±4.75	59.31±3.61	69.53±1.94	11.59±0.81	16.47±0.99	23.58±1.36
FedSAM	59.28±0.15	75.19±0.10	76.07±0.09	29.31±0.32	57.12±0.08	61.56±0.31	11.19±0.16	15.95±0.15	22.44±0.16
FedDF	64.72±2.11	74.16±1.52	**85.51**±0.95	32.37±2.39	60.08±5.67	71.52±2.67	11.63±0.67	17.13±1.12	25.84±1.02
FedGEN	63.50±3.27	69.42±4.09	80.17±4.71	27.21±3.12	57.16±2.71	68.93±1.75	10.07±0.19	15.26±0.29	21.49±0.17
FedMA	64.71±4.92	74.98±5.03	77.13±4.10	28.61±1.39	59.97±0.96	70.91±1.02	11.89±0.57	15.90±0.92	22.02±0.82
GAMF	64.97±3.93	75.21±4.05	77.34±3.78	28.92±1.52	60.23±1.93	71.44±1.75	11.98±0.99	16.76±0.77	24.15±0.49
FedLAW	60.34±4.39	73.93±4.91	77.53±3.52	26.32±2.80	46.81±3.61	61.08±2.61	11.57±1.61	15.99±0.49	22.37±0.68
Ours	**66.30**±0.07	**76.59**±0.25	78.99±0.13	**34.97**±0.31	**62.81**±0.28	**72.04**±0.23	**12.20**±0.13	**19.98**±0.25	**28.16**±0.30
Dirichlet (α)	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
FedAvg	71.81±5.61	75.97±3.21	79.73±1.94	50.43±1.68	61.11±2.68	67.37±1.69	30.13±0.70	35.73±0.56	38.86±0.35
FedProx	70.44±3.87	72.17±4.10	75.24±2.19	38.98±5.91	61.64±1.92	70.16±2.03	32.96±1.18	40.81±0.41	42.53±0.48
FedExP	73.42±4.22	76.57±3.39	80.22±3.78	60.63±4.32	70.22±2.40	74.37±1.91	36.76±1.18	44.18±0.53	47.80±0.58
FedSAM	71.65±0.07	75.91±0.06	77.67±0.10	49.96±0.20	59.53±0.19	64.54±0.21	21.54±0.12	24.72±0.18	28.59±0.19
FedDF	**80.03**±1.04	84.42±0.62	86.84±1.93	54.28±2.39	69.85±5.67	73.76±2.67	34.76±0.67	39.42±1.12	42.31±1.02
FedGEN	73.02±1.87	77.48±3.50	81.76±4.21	47.09±3.12	64.90±2.71	68.74±1.75	29.02±0.19	38.54±0.29	40.81±0.17
FedMA	71.87±4.28	75.89±4.15	80.12±3.23	49.98±2.01	61.32±2.17	68.42±1.95	30.02±0.58	36.21±0.83	39.55±0.52
GAMF	72.11±5.16	76.24±3.67	80.55±2.06	51.21±1.37	63.45±1.03	70.14±1.81	31.12±0.69	37.26±0.78	41.25±0.74
FedLAW	71.93±8.23	76.88±2.80	79.98±1.09	48.91±3.59	61.50±2.29	67.08±1.75	32.01±2.61	38.80±2.20	40.11±1.17
Ours	78.63±0.14	**84.67**±0.12	**87.01**±0.08	**62.46**±0.22	**70.27**±0.29	**74.96**±0.17	**39.38**±0.25	**45.86**±0.22	**49.31**±0.22



(a) Aggregation strategies (b) Memory size K (c) Local ep.s (FedCDA) (d) Local ep.s (FedAvg)

Figure 2: (a) shows the effect of different aggregation strategies. (b) shows the impact of the memory size K on FedCDA. (c) and (d) present the impact of local epochs (ep.s) on FedCDA and FedAvg.

larger datasets such as CIFAR-10 Dirichlet 0.1, FedCDA with ResNet-18 achieves 62.46% accuracy whereas the best baseline method FedExP achieves 60.63% accuracy. In addition, FedCDA with simple CNN also makes improvements on relatively smaller datasets, although the improvement is less than in large models. At the same time, we can also see that the results of our method on relatively small datasets and simple CNN are not the best, which may be because the features of models with different rounds are more similar on small datasets and simple models, and can not provide more aggregation features to accelerate convergence. In conclusion, we can notice our FedCDA makes more improvements on the large model and complex datasets.

More Comparison Results with Different Hyper-parameters. To compare with baselines in a comprehensive way, we further conduct experiments on different hyper-parameters. The number of clients is 100 with the sample ratio being 10%. The learning rate is set to be 0.1 with the weight decay being 1e-3, and the number of local epochs is 5. The local optimizer is SGD without momentum. The experiment is conducted by running the ResNet18 on the CIFAR-100 dataset. The results are shown in Table 2. As can be seen, our method still performs the best. Specifically, when the data is the most heterogeneous, i.e., with Dirichlet $\alpha = 0.1$, FedCDA achieves the accuracy of 47.38% which outperforms the best baseline method FedDF by 6.34%.

Table 2: Results of FL with 100 clients.

Method	Dir(0.1)	Dir(0.3)	Dir(0.5)
FedAvg	38.89%±0.85%	40.38%±0.55%	42.23%±0.30%
FedProx	39.86%±0.45%	39.48%±0.37%	40.18%±0.46%
FedExP	38.04%±3.37%	44.10%±1.69%	41.79%±1.62%
FedSAM	16.35%±0.25%	20.53%±0.25%	25.70%±0.28%
FedDF	41.04%±0.57%	47.06%±0.74%	47.63%±0.53%
FedGEN	39.91%±1.72%	41.65%±1.35%	43.39%±1.08%
FedMA	39.12%±0.52%	40.42%±0.61%	42.89%±0.21%
GAMF	39.89%±0.67%	40.98%±0.32%	43.25%±0.34%
FedLAW	40.88%±0.66%	41.77%±0.78%	41.89%±0.33%
Ours	**47.38%**±0.23%	**49.96%**±0.21%	**50.04%**±0.19%

Different Aggregation Strategies. We compare five aggregation strategies on the CIFAR-10 datasets. Because the *optimal* selection for updates method is an exponential method, we only sample 10 clients in each round, where the sample ratio is 0.3 and the client memory size is $K = 3$. As shown in Figure 2(a), we compare FedAvg with different aggregation strategies in our method. The

clients in FedAvg do not require memory. The *newest* strategy is that only the newest local model of each client is aggregated during the server aggregation phase. The *random* strategy is that during the server aggregation phase, we randomly select a local model from multiple rounds to aggregate. Finally, the *approximate* strategy is as 8 shown above and the optimal one is as 3. We can find that the approximate and optimal strategies have huge performance improvement over FedAvg, newest and random strategies with ResNet-18 on CIFAR-10 Dirichlet 0.1, 0.3 and 0.5. At its peak, there is an almost 10% increase over FedAvg. We can also see the performance of approximate performance is about the same as the optimal one, but the convergence time of the former is much smaller than that of the latter. In fact, the former is actually a greedy algorithmic approximation of the latter, so the computation of the solution is greatly reduced. We also compare the average polymerization time of each round of these aggregation strategies. Details are in the Appendix C.

Hyperparameters Sensitivity. As shown in the following figure 2(b) , We compare the test accuracy of client memory size K for 1, 2, 3, 4 on CIFAR-10 Dirichlet 0.1. As K increases, the final test accuracy increases which confirms our theory. The increase of K value gives cross-round polymerization more choices and possibilities.

Comparison with Different Epochs for `FedCDA` and FedAvg. By comparing the results in Figure 2(c) and Figure 2(d), from the vertical perspective, our `FedCDA` eventually converges with increasing test accuracy with more local epochs while the final convergence accuracy of Fedavg remains roughly unchanged. *This proves that our algorithm can tolerate large local interactions to save communication cost*. Horizontally, `FedCDA` converges rapidly and stably, whereas the convergence curve of FedAvg is very oscillatory. The reason is that our method excludes the negative impact of sampling clients while FedAvg cannot. Therefore, *the convergence of our method is more stable than FedAvg.*

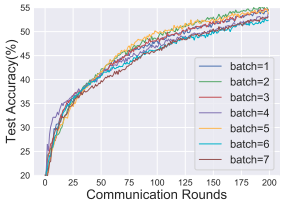


Figure 3: Impact of Batch

Effect of Batch Number. As we can see in Figure 6.2, different batch numbers have little effect on the final precision result. Our experiment setup 50 clients and 10 sample clients on the CIFAR-10 Dirichlet 0.1. We compare the results for batch $B = 1, 2, 3, 4, 5, 6, 7$. The results show that the approximation selection can keep the accuracy closer to the optimal selection.

Warmup Analysis. The experiments of Figure 6.2 are conducted on settings of CIFAR-10, 20 clients, and the sample ratio 0.2. `FedCDA` with no warmup rounds is worse than some with warmup rounds. This is because the local models in the early rounds can not approach convergence. *The combination with not-well-converged local models in old rounds may prevent the training of the global model*. Therefore, it is similar to FedAvg in the startup training stages. Its advantages gradually exhibit with the training proceeds and outperforms FedAvg (warmup=200). Yet, there is a threshold for raising warmup rounds. Specifically, we can see that FL with 150 warmup rounds has worse performance than 100 warmup rounds. The principle behind it is that the local models and the global model have approached convergence in the final training stage. *The difference between local models in different rounds is greatly reduced and thus the combination of them gains little benefits.*

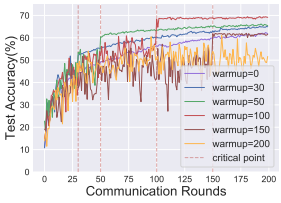


Figure 4: Impact of Warmup

7 Conclusion

This paper targets aggregation in federated learning, addressing the issue that traditional single-round methods may not preserve locally learned knowledge due to statistical heterogeneity. Recognizing clients' convergence post-startup stage and local models' consistent data fitting across rounds, we propose `FedCDA`- a new method that selectively aggregates cross-round models with minimum divergence. To enhance efficiency, we introduce an approximation selection algorithm. Theoretical convergence is proven and empirical results show our method outperforms state-of-the-art baselines.

This paper addresses the ideal scenario where smoothness value is equal among clients. The goal is to improve our method for cases with varying smoothness by refining objectives and incorporating sharpness in cross-round aggregation, as we currently treat all local models equally without considering the sharpness of their local optima, despite flatter optima often correlating with better generalization.

ACKNOWLEDGMENTS

The research is supported under the National Key R&D Program of China (2022ZD0160201) and the RIE2020 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contributions from the industry partner(s). This work is supported by National Natural Science Foundation of China under grants U1836204, U1936108, 62206102, and Science and Technology Support Program of Hubei Province under grant 2022BAA046.

REFERENCES

Muhammad Ammad-ud-din, Elena Ivannikova, Suleiman A. Khan, Were Oyomno, Qiang Fu, Kuan Eeik Tan, and Adrian Flanagan. Federated collaborative filtering for privacy-preserving personalized recommendation system. *CoRR*, abs/1901.09888, 2019.

Ilai Bistritz, Ariana Mann, and Nicholas Bambos. Distributed distillation for on-device learning. In *Proceedings of Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2020.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Hong-You Chen and Wei-Lun Chao. Fedbe: Making bayesian model ensemble applicable to federated learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

Canh T. Dinh, Nguyen H. Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. In *Proceedings of Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2020.

Pengfei Guo, Puyang Wang, Jinyuan Zhou, Shanshan Jiang, and Vishal M. Patel. Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 2423–2432.

Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 11020–11029, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, 2016.

Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. Fedexp: Speeding up federated averaging via extrapolation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5*, 2023.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 586–594, 2016.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

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

Shuangtong Li, Tianyi Zhou, Xinmei Tian, and Dacheng Tao. Learning to collaborate in decentralized learning of personalized models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 9756–9765, 2022a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems 2020, MLSys 2020, Austin, TX, USA, March 2-4, 2020*, 2020.

Xin-Chun Li, Yichu Xu, Shaoming Song, Bingshuai Li, Yinchuan Li, Yunfeng Shao, and De-Chuan Zhan. Federated learning with position-aware neurons. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10072–10081, 2022b.

Zexi Li, Tao Lin, Xinyi Shang, and Chao Wu. Revisiting weighted aggregation in federated learning with neural networks. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pp. 19767–19788, 2023.

Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33:2351–2363, 2020.

Chang Liu, Chenfei Lou, Runzhong Wang, Alan Yuhan Xi, Li Shen, and Junchi Yan. Deep neural network fusion via graph matching with applications to model ensemble and federated learning. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pp. 13857–13869, 2022.

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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 1013–1023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2017a.

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*, pp. 1273–1282. PMLR, 2017b.

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, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pp. 18250–18280, 2022.

Swaroop Ramaswamy, Rajiv Mathews, Kanishka Rao, and Françoise Beaufays. Federated learning for emoji prediction in a mobile keyboard. *CoRR*, abs/1906.04329, 2019.

Yasar Abbas Ur Rehman, Yan Gao, Pedro Porto Buarque de Gusmão, Mina Alibeigi, Jiajun Shen, and Nicholas D. Lane. L-DAWA: layer-wise divergence aware weight aggregation in federated self-supervised visual representation learning. *CoRR*, abs/2307.07393, 2023.

Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Yan Sun, Li Shen, Tiansheng Huang, Liang Ding, and Dacheng Tao. Fedspeed: Larger local interval, less communication round, and higher generalization accuracy. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5*, 2023a.

Yan Sun, Li Shen, and Dacheng Tao. Understanding how consistency works in federated learning via stage-wise relaxed initialization. *Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, LA, USA, Dec 10-16*, 2023b.

Chunnan Wang, Xiang Chen, Junzhe Wang, and Hongzhi Wang. ATPFL: automatic trajectory prediction model design under federated learning framework. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, New Orleans, LA, USA,June 18-24*, pp. 6553–6562, 2022.

Haozhao Wang, Yichen Li, Wenchao Xu, Ruixuan Li, Yufeng Zhan, and Zhigang Zeng. Dafkd: Domain-aware federated knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 20412–20421, 2023.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris S. Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020a.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

Guile Wu and Shaogang Gong. Peer collaborative learning for online knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI*, 2021.

Lei Wu, Zhanxing Zhu, and Weinan E. Towards understanding generalization of deep learning: Perspective of loss landscapes. *CoRR*, abs/1706.10239, 2017.

Yingda Xia, Dong Yang, Wenqi Li, Andriy Myronenko, Daguang Xu, Hirofumi Obinata, Hitoshi Mori, Peng An, Stephanie A. Harmon, Evrim Turkbey, Baris Turkbey, Bradford J. Wood, Francesca Patella, Elvira Stellato, Gianpaolo Carrafiello, Anna Ierardi, Alan L. Yuille, and Holger Roth. Auto-fedavg: Learnable federated averaging for multi-institutional medical image segmentation. *CoRR*, abs/2104.10195, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

An Xu, Wenqi Li, Pengfei Guo, Dong Yang, Holger Roth, Ali Hatamizadeh, Can Zhao, Daguang Xu, Heng Huang, and Ziyue Xu. Closing the generalization gap of cross-silo federated medical image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 20834–20843.

Fuxun Yu, Weishan Zhang, Zhuwei Qin, Zirui Xu, Di Wang, Chenchen Liu, Zhi Tian, and Xiang Chen. Fed2: Feature-aligned federated learning. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pp. 2066–2074, 2021.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan H. Greenewald, and Trong Nghia Hoang. Statistical model aggregation via parameter matching. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 10954–10964, 2019a.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan H. Greenewald, Trong Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 7252–7261, 2019b.

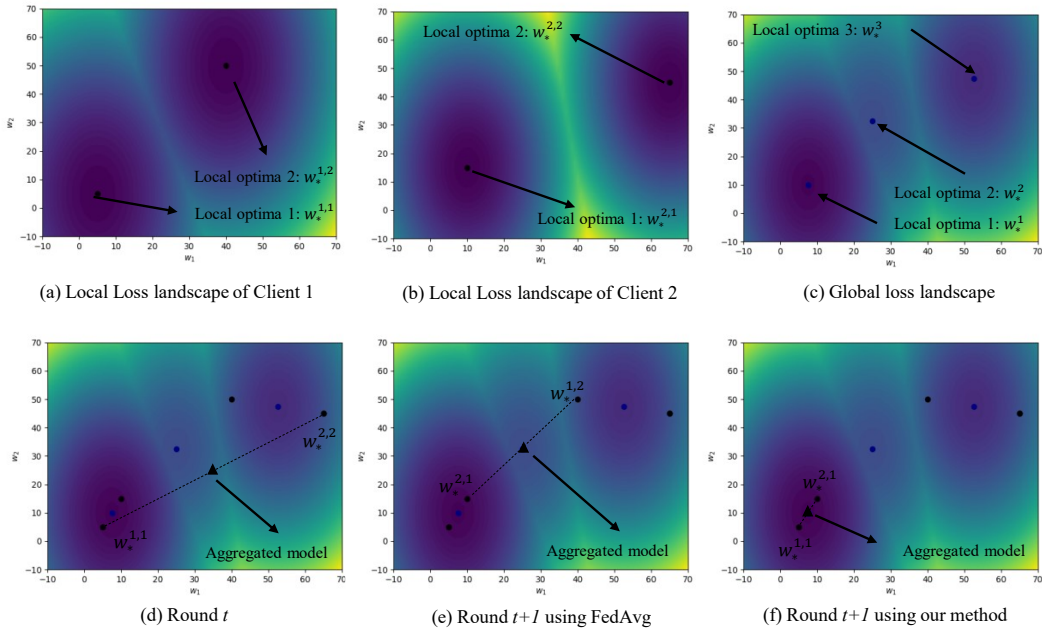
Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling-Yu Duan. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10164–10173, 2022.

Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International Conference on Machine Learning*, pp. 12878–12889. PMLR, 2021.

A MORE DETAILS ABOUT EXPERIMENTS

Consider a two-client FL system where the loss functions are $F_1(\mathbf{w}) = min\big((w_1-5)^2+(w_2-5)^2), 6(w_1-40)^2+1.5(w_2-50)^2\big)$ and $F_2(\mathbf{w}) = min\big((w_1-10)^2+(w_2-15)^2), 6(w_1-65)^2+1.5(w_2-45)^2\big)$, where $\mathbf{w}=[w_1,w_2]$ and the landscape of the local loss in two clients and the global loss are shown in Figure 5(a), 5(b), and 5(c) respectively.

In this case, the local model of each client only has two local optima denoted by $\mathbf{w}_*^{n,i}$ for the i-th local optima in client n. Now, considering the training of t-round. We compare the aggregation process of FedAvg and our method by allowing them to start from the same global model in the t-th round, as shown in in Figure 5(d), and then present their distinct aggregated global models in Figure 5(e) and Figure 5(f) respectively. It can be found that our method can achieve the global minima (the global model in the most blue area) while FedAvg falls into a bad area. The essence is that our method aggregates the local optima $\mathbf{w}_*^{1,1}$ of client 1 obtained in the round t and the local optima $\mathbf{w}_*^{2,1}$ of client 2 obtained in the round $t+1$ together into the global model. While FedAvg only computes an average of $\mathbf{w}_*^{1,2}$ of client 1 obtained in the round $t+1$ and the local optima $\mathbf{w}_*^{2,1}$ of client 2 obtained in the in the round $t+1$.



(a) Local Loss landscape of Client 1 (b) Local Loss landscape of Client 2 (c) Global loss landscape

(d) Round t (e) Round $t+1$ using FedAvg (f) Round $t+1$ using our method

Figure 5: Illustration of training a model with multiple local optima. The loss is smaller when the color becomes more blue. Black triangle denotes the aggregated global model. (a)-(c) denotes the loss landscapes. (d) presents the location of the global model in the t-th round. (e) presents the location of the global model aggregated by FedAvg in the $t+1$-th round. (f) presents the location of the global model aggregated by our method in the $t+1$-th round.

B PROOFS

B.1 PROOFS FOR THEOREM 1

We denote $\mathbf{w}_{t_*}, \mathbf{w}_{t_*}^n, n=1,\ldots,N$ as the solution of objective (3) and $\mathbf{w}_{t'_*}, \mathbf{w}_{t'_*}^n, n=1,\ldots,N$ as the solution of objective (5). Then, the approximation error ε can be calculated as:

$$\varepsilon = \frac{1}{N}\sum_{n=1}^{N}\big[F_n(\mathbf{w}_{t'_*}^n) + \nabla_{\mathbf{w}_{t'_*}^n} F_n(\mathbf{w}_{t'_*}^n)(\mathbf{w}_{t'_*} - \mathbf{w}_{t'_*}^n) + \frac{L}{2}\|\mathbf{w}_{t'_*} - \mathbf{w}_{t'_*}^n\|_2^2\big]$$

$$
\begin{aligned}
&- \frac{1}{N}\sum_{n=1}^{N}\big[F_n(\mathbf{w}_{t_*}^n) + \nabla_{\mathbf{w}_{t_*}^n} F_n(\mathbf{w}_{t_*}^n)(\mathbf{w}_{t_*} - \mathbf{w}_{t_*}^n) + \frac{L}{2}\|\mathbf{w}_{t_*} - \mathbf{w}_{t_*}^n\|_2^2\big] \\
=&\frac{1}{N}\sum_{n=1}^{N}\nabla_{\mathbf{w}_{t'_*}^n} F_n(\mathbf{w}_{t'_*}^n)(\mathbf{w}_{t'_*} - \mathbf{w}_{t'_*}^n) - \frac{1}{N}\sum_{n=1}^{N}\nabla_{\mathbf{w}_{t_*}^n} F_n(\mathbf{w}_{t_*}^n)(\mathbf{w}_{t_*} - \mathbf{w}_{t_*}^n) \\
&+ \frac{1}{N}\sum_{n=1}^{N}\big[F_n(\mathbf{w}_{t'_*}^n) + \frac{L}{2}\|\mathbf{w}_{t'_*} - \mathbf{w}_{t'_*}^n\|_2^2\big] - \frac{1}{N}\sum_{n=1}^{N}\big[F_n(\mathbf{w}_{t_*}^n) + \frac{L}{2}\|\mathbf{w}_{t_*} - \mathbf{w}_{t_*}^n\|_2^2\big] \\
\leq&\frac{1}{N}\sum_{n=1}^{N}\nabla_{\mathbf{w}_{t'_*}^n} F_n(\mathbf{w}_{t'_*}^n)(\mathbf{w}_{t'_*} - \mathbf{w}_{t'_*}^n) - \frac{1}{N}\sum_{n=1}^{N}\nabla_{\mathbf{w}_{t_*}^n} F_n(\mathbf{w}_{t_*}^n)(\mathbf{w}_{t_*} - \mathbf{w}_{t_*}^n) \\
\leq&\frac{1}{N}\sum_{n=1}^{N}\nabla_{\mathbf{w}_{t'_*}^n} F_n(\mathbf{w}_{t'_*}^n)(\mathbf{w}_{t'_*} - \mathbf{w}_{t'_*}^n) - \frac{1}{N}\sum_{n=1}^{N}\nabla_{\mathbf{w}_{t_*}^n} F_n(\mathbf{w}_{t_*}^n)(\mathbf{w}_{t_*} - \mathbf{w}_{t_*}^n). \qquad (9)
\end{aligned}
$$

Further, let the constant $\epsilon > 0$ be the bound between the local model $\mathbf{w}_t^n$ of each client n and one of its saddle points or local optima $\mathbf{w}_t^{n,*}$ at each round t, i.e., $\|\mathbf{w}_t^n - \mathbf{w}_t^{n,*}\| \leq \epsilon$. Then, the approximation error ε in (9) can be bounded as:

$$
\begin{aligned}
\varepsilon \leq&\frac{1}{N}\sum_{n=1}^{N}\|\nabla_{\mathbf{w}_{t'_*}^n} F_n(\mathbf{w}_{t'_*}^n)\|\|(\mathbf{w}_{t'_*} - \mathbf{w}_{t'_*}^n)\| + \frac{1}{N}\sum_{n=1}^{N}\|\nabla_{\mathbf{w}_{t_*}^n} F_n(\mathbf{w}_{t_*}^n)\|\|(\mathbf{w}_{t_*} - \mathbf{w}_{t_*}^n)\| \\
=&\frac{1}{N}\sum_{n=1}^{N}\|\nabla_{\mathbf{w}_{t'_*}^n} F_n(\mathbf{w}_{t'_*}^n) - \nabla_{\mathbf{w}_{t'_*}^{n,*}} F_n(\mathbf{w}_{t'_*}^{n,*})\|\|\mathbf{w}_{t'_*} - \mathbf{w}_{t'_*}^n\| \\
&+ \frac{1}{N}\sum_{n=1}^{N}\|\nabla_{\mathbf{w}_{t_*}^n} F_n(\mathbf{w}_{t_*}^n) - \nabla_{\mathbf{w}_{t_*}^{n,*}} F_n(\mathbf{w}_{t_*}^{n,*})\|\|\mathbf{w}_{t_*} - \mathbf{w}_{t_*}^n\| \\
\leq&\frac{1}{N}\sum_{n=1}^{N}L\|\mathbf{w}_{t'_*}^n - \mathbf{w}_{t'_*}^{n,*}\|\|\mathbf{w}_{t'_*} - \mathbf{w}_{t'_*}^n\| + \frac{1}{N}\sum_{n=1}^{N}L\|\mathbf{w}_{t_*}^n - \mathbf{w}_{t_*}^{n,*}\|\|\mathbf{w}_{t_*} - \mathbf{w}_{t_*}^n\| \\
\leq&L\epsilon\frac{1}{N}\sum_{n=1}^{N}\|\mathbf{w}_{t'_*} - \mathbf{w}_{t'_*}^n\| + L\epsilon\frac{1}{N}\sum_{n=1}^{N}\|\mathbf{w}_{t_*} - \mathbf{w}_{t_*}^n\| \\
=&L\epsilon\frac{1}{N}\sum_{n=1}^{N}\|\frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}_{t'_*}^i - \mathbf{w}_{t'_*}^{i,*} + \mathbf{w}_{t'_*}^{i,*}) - \mathbf{w}_{t'_*}^{n,*} + \mathbf{w}_{t'_*}^{n,*} - \mathbf{w}_{t'_*}^n\| \\
&+ L\epsilon\frac{1}{N}\sum_{n=1}^{N}\|\frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}_{t_*}^i - \mathbf{w}_{t_*}^{i,*} + \mathbf{w}_{t_*}^{i,*}) - \mathbf{w}_{t_*}^{n,*} + \mathbf{w}_{t_*}^{n,*} - \mathbf{w}_{t_*}^n\| \\
\leq&L\epsilon\frac{1}{N}\sum_{n=1}^{N}\Big(\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_{t'_*}^{i,*} - \mathbf{w}_{t'_*}^{n,*}\| + \|\frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}_{t'_*}^i - \mathbf{w}_{t'_*}^{i,*})\| + \|\mathbf{w}_{t'_*}^{n,*} - \mathbf{w}_{t'_*}^n\|\Big) \\
&+ L\epsilon\frac{1}{N}\sum_{n=1}^{N}\Big(\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_{t_*}^{i,*} - \mathbf{w}_{t_*}^{n,*}\| + \|\frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}_{t_*}^i - \mathbf{w}_{t_*}^{i,*})\| + \|\mathbf{w}_{t_*}^{n,*} - \mathbf{w}_{t_*}^n\|\Big) \\
\leq&2L\epsilon(D + 2\epsilon) = 4L\epsilon^2 + 2LD\epsilon. \qquad (10)
\end{aligned}
$$

The proof is done.

B.2 Proof of Theorem 2

We tend to prove that the global loss $F(\mathbf{w}_{t_*})$ obtained by using cross-round aggregated model $\mathbf{w}_{t_*}$ of solving (3) is guaranteed to be smaller than the global loss $F(\mathbf{w}_t)$ where $\mathbf{w}_t$ is obtained by aggregating local models in the t-th round, under some conditions.

Like the analysis for Theorem 1, we also define the local optima that is close to the local model $\mathbf{w}_t^n$ as $\mathbf{w}_t^{n,*}$ for each round t and the distance between them is ϵ, i.e., $\|\mathbf{w}_t^n - \mathbf{w}_t^{n,*}\| \leq \epsilon$. Also, we define

$\bar{\mathbf{w}}_t^{n,*} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{w}_t^{n,*}$ for simplicity of presentation. By denoting $\mathbf{w}_{t_*}, \mathbf{w}_{t_*}^n, n = 1, \ldots, N$ as the solution of objective (3), we have

$$\begin{aligned}
F(\mathbf{w}_{t_*}) =& \frac{1}{N}\sum_{n=1}^{N} F(\mathbf{w}_{t_*}) \leq \frac{1}{N}\sum_{n=1}^{N} \big[F_n(\mathbf{w}_{t_*}^{n,*}) + \nabla_{\mathbf{w}_{t_*}^{n,*}} F_n(\mathbf{w}_{t_*}^{n,*})(\mathbf{w}_{t_*} - \mathbf{w}_{t_*}^{n,*}) + \frac{L}{2}\|\mathbf{w}_{t_*} - \mathbf{w}_{t_*}^{n,*}\|_2^2\big] \\
=& \frac{1}{N}\sum_{n=1}^{N} \big[F_n(\mathbf{w}_{t_*}^{n,*}) + \frac{L}{2}\|\frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}_{t_*}^{n} - \mathbf{w}_{t_*}^{n,*} + \mathbf{w}_{t_*}^{n,*}) - \mathbf{w}_{t_*}^{n,*}\|_2^2\big] \\
\leq& \frac{1}{N}\sum_{n=1}^{N} \big[F_n(\mathbf{w}_{t_*}^{n,*}) + \frac{L}{2}\|\bar{\mathbf{w}}_{t_*}^{n,*} - \mathbf{w}_{t_*}^{n,*}\|_2^2 + \frac{L}{2}\|\frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}_{t_*}^{n} - \mathbf{w}_{t_*}^{n,*})\|_2^2\big] \\
\leq& \frac{1}{N}\sum_{n=1}^{N} \big[F_n(\mathbf{w}_{t_*}^{n,*}) + \frac{L}{2}\|\bar{\mathbf{w}}_{t_*}^{n,*} - \mathbf{w}_{t_*}^{n,*}\|_2^2 + \frac{L\epsilon^2}{2}\big] \qquad (11)
\end{aligned}$$

By assuming the local loss achieves equivalent values on local optima in different rounds, i.e., $F_n(\mathbf{w}_{t_*}^{n,*}) = F_n(\mathbf{w}_t^{n,*})$, we have $\frac{1}{N}\sum_{n=1}^{N} F_n(\mathbf{w}_{t_*}^{n,*}) = \frac{1}{N}\sum_{n=1}^{N} F_n(\mathbf{w}_t^{n,*})$. Besides, given the condition $\mathrm{Var}(\mathbf{w}_{t_*}^{n,*}) \leq \frac{\mu}{L}\mathrm{Var}(\mathbf{w}_t^{n,*}) - (\frac{\mu}{L} + 1)\epsilon^2$, we can derive the following inequality from (11):

$$\begin{aligned}
F(\mathbf{w}_{t_*}) \leq& \frac{1}{N}\sum_{n=1}^{N} \big[F_n(\mathbf{w}_{t_*}^{n,*}) + \frac{L}{2}\|\bar{\mathbf{w}}_{t_*}^{n,*} - \mathbf{w}_{t_*}^{n,*}\|_2^2 + \frac{L\epsilon^2}{2}\big] \\
\leq& \frac{1}{N}\sum_{n=1}^{N} \big[F_n(\mathbf{w}_t^{n,*}) + \frac{\mu}{2}\|\bar{\mathbf{w}}_t^{n,*} - \mathbf{w}_t^{n,*}\|_2^2 - \frac{\mu\epsilon^2}{2}\big]. \qquad (12)
\end{aligned}$$

Further, we consider that the loss function $F_n(\mathbf{w})$ is strongly convex with a parameter μ within the region spanning from the local optima $\mathbf{w}_t^{n,*}$ to the aggregated model $\mathbf{w}_t = \frac{1}{N}\sum_{n=1}^{N} \mathbf{w}_t^n$. According to (12), we can immediately obtain

$$\begin{aligned}
F(\mathbf{w}_{t_*}) \leq& \frac{1}{N}\sum_{n=1}^{N} \big[F_n(\mathbf{w}_t^{n,*}) + \frac{\mu}{2}\|\bar{\mathbf{w}}_t^{n,*} - \mathbf{w}_t^{n,*}\|_2^2 - \frac{\mu\epsilon^2}{2}\big] \\
\leq& \frac{1}{N}\sum_{n=1}^{N} \big[F_n(\mathbf{w}_t^{n,*}) + \frac{\mu}{2}\|\bar{\mathbf{w}}_t^{n,*} - \mathbf{w}_t^{n,*}\|_2^2 - \frac{\mu}{2}\frac{1}{N}\sum_{n=1}^{N}\|(\mathbf{w}_t^{n} - \mathbf{w}_t^{n,*})\|_2^2\big] \\
\leq& \frac{1}{N}\sum_{n=1}^{N} \big[F_n(\mathbf{w}_t^{n,*}) + \frac{\mu}{2}\|\bar{\mathbf{w}}_t^{n,*} - \mathbf{w}_t^{n,*}\|_2^2 - \frac{\mu}{2}\|\frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}_t^{n} - \mathbf{w}_t^{n,*})\|_2^2\big] \\
\leq& \frac{1}{N}\sum_{n=1}^{N} \big[F_n(\mathbf{w}_t^{n,*}) + \frac{\mu}{2}\|\frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}_t^{n} - \mathbf{w}_t^{n,*} + \mathbf{w}_t^{n,*}) - \mathbf{w}_t^{n,*}\|_2^2\big] \\
=& \frac{1}{N}\sum_{n=1}^{N} \big[F_n(\mathbf{w}_t^{n,*}) + \nabla_{\mathbf{w}_t^{n,*}} F_n(\mathbf{w}_t^{n,*})(\mathbf{w}_t - \mathbf{w}_t^{n,*}) + \frac{\mu}{2}\|\mathbf{w}_t - \mathbf{w}_t^{n,*}\|_2^2\big] \\
\leq& \frac{1}{N}\sum_{n=1}^{N} F(\mathbf{w}_t) = F(\mathbf{w}_t), \qquad (13)
\end{aligned}$$

which completes the proof.

B.3 Case Study for Strongly Convex Objectives

Our method can also be applied to the convex cases where the local model of each client only has one local optimum, i.e., its global optimum. The intuition behind it is that the local converged model of each client will be stochastically located around the optimum due to the usage of the stochastic gradient descent (SGD) or its variants, and hence the combination of different stochastic local models also performs differently from each other. The selected aggregation of our method can

be viewed as selecting the most appropriate local model from all stochastic local models obtained in different rounds. To show this, we consider a case for the point estimation problem.

Theorem 4 *Consider a problem of point estimation, where a single-dimension value x is estimated. Each client n contains samples $\{x^n + z_1^n, \ldots, x^n + z_m^n\}$ where $x^n \sim$ Gaussian(x, σ^2) and $z_i^n \sim$ Gaussian$(0, \delta^2)$. The global loss function is defined as $F(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} F_n(\mathbf{w})$ with $F_n(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{w} - (x^n + z_i^n))^2$. Then, the global loss of the model $\mathbf{w}_{t_*}$ aggregated from local models selected by (3) is smaller than that of the model $\bar{\mathbf{w}}_t$ aggregated from the t-th round local models, i.e., $F(\mathbf{w}_{t_*}) \leq F(\mathbf{w}_t)$.*

Proof: According to the definition of the problem, we have

$$\begin{aligned} F(\mathbf{w}) &= \frac{1}{N}\sum_{n=1}^{N} F_n(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{m}\sum_{i=1}^{m}(\mathbf{w} - (x^n + z_i^n))^2 \\ &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{m}\sum_{i=1}^{m}\Big((\mathbf{w}^n - (x^n + z_i^n))^2 + 2(\mathbf{w}^n - (x^n + z_i^n))^T(\mathbf{w} - \mathbf{w}^n) + (\mathbf{w} - \mathbf{w}^n)^2\Big) \\ &= \frac{1}{N}\sum_{n=1}^{N}(F_n(\mathbf{w}^n) + \nabla_{F_n(\mathbf{w}^n)}F_n(\mathbf{w}^n)^T(\mathbf{w} - \mathbf{w}^n) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^n)^2) \end{aligned} \tag{14}$$

which has the same formula as (3). Therefore, by denoting the $\mathbf{w}_{t_*}, \mathbf{w}_{t_*}^1, \ldots, \mathbf{w}_{t_*}^N$ as the solution of (3), we have

$$\begin{aligned} F(\mathbf{w}_{t_*}) =& F_n(\mathbf{w}_{t_*}^n) + \nabla_{F_n(\mathbf{w}_{t_*}^n)}F_n(\mathbf{w}_{t_*}^n)^T(\mathbf{w}_{t_*} - \mathbf{w}_{t_*}^n) + \frac{1}{2}(\mathbf{w}_{t_*} - \mathbf{w}_{t_*}^n)^2 \\ \leq& F_n(\mathbf{w}_t^n) + \nabla_{F_n(\mathbf{w}_t^n)}F_n(\mathbf{w}_t^n)^T(\mathbf{w}_t - \mathbf{w}_t^n) + \frac{1}{2}(\mathbf{w}_t - \mathbf{w}_t^n)^2 \\ =& F(\mathbf{w}_t), \end{aligned} \tag{15}$$

which completes the proof.

B.4 Proof of Theorem 3

Our proof mainly includes two parts. First, we prove that the difference between the loss of the global model $\mathbf{w}_{t_*}$ obtained by (3) and that of the global model $\mathbf{w}_t$ obtained by aggregating the newest local models is bounded. Then, we prove that the loss of the global model $\mathbf{w}_t$ achieves convergence, which in turn indicates the convergence of the global model $\mathbf{w}_{t_*}$.

Lemma 1 *Let $\mathbf{w}_{t_*}$ be the global model solved by (3) and $\mathbf{w}_t$ be the global model aggregated from the t-th round local models. When considering all clients are sampled in each round, under the Assumption 1,2, and 3, given any fixed $\mathbf{w}_{t-1_*}$, the global loss $F(\mathbf{w}_{t_*})$ is bounded by*

$$\begin{aligned} \mathbb{E}F(\mathbf{w}_{t_*}) \leq \mathbb{E}\big(F(\mathbf{w}_{t-1_*}) + <\nabla F(\mathbf{w}_{t-1_*}), \mathbf{w}_t - \mathbf{w}_{t-1_*}> + \frac{L}{2}\|\mathbf{w}_t - \mathbf{w}_{t-1_*}\|_2^2\big) \\ + \frac{L\eta^2E^2M^2}{2} + 3L\eta^4E^4\sigma_l^2 + \frac{3E^2\eta^2\sigma_g^2}{2}. \end{aligned} \tag{16}$$

Proof: Based on the definition of selection strategy of (3), we can directly obtain the following conclusion when considering all clients participating in the training of each round:

$$\begin{aligned} F(\mathbf{w}_{t_*}) &\leq \min_{\mathbf{w}\in\mathbb{R}^d, \mathbf{w}^1\in\mathcal{W}_t^1, \ldots, \mathbf{w}^N\in\mathcal{W}_t^N} \frac{1}{N}\sum_{n=1}^{N}\big[F_n(\mathbf{w}^n) + \nabla_{\mathbf{w}^n}F_n(\mathbf{w}^n)^T(\mathbf{w} - \mathbf{w}^n) + \frac{L}{2}\|\mathbf{w} - \mathbf{w}^n\|_2^2\big] \\ &\leq \frac{1}{N}\sum_{n=1}^{N}\Big[F_n(\mathbf{w}_t^n) + \nabla_{\mathbf{w}_t^n}F_n(\mathbf{w}_t^n)^T(\mathbf{w}_t - \mathbf{w}_t^n) + \frac{L}{2}\|\mathbf{w}_t - \mathbf{w}_t^n\|_2^2\Big]. \end{aligned} \tag{17}$$

Our analysis identify the relationship between the $F(\mathbf{w}_{t_*})$ and $F(\mathbf{w}_t)$ with a fixed global model in previous round $t-1$, i.e., $\mathbf{w}_{t-1_*}$. For simplicity, we denote $\mathbf{w}_{t-1}$ as the previous-round model. It deserves noting that our conclusion also holds for $\mathbf{w}_{t-1_*}$.

We define $\mathbf{G}_t^n = \sum_{e=1}^{E} \nabla f_n(\mathbf{w}_{t,e}^n)$ with $\mathbf{w}_{t,1}^n = \mathbf{w}_{t-1_*}$, and $\mathbf{G}_t = \frac{1}{N}\sum_{n=1}^{N} \mathbf{G}_t^n$. For simplicity of representation, we also define $\tilde{\mathbf{G}}_t^n = \mathbb{E}_\xi \mathbf{G}_t^n = \sum_{e=1}^{E} \nabla f_n(\mathbf{w}_{t,e}^n)$ and $\tilde{\mathbf{G}}_t = \mathbb{E}_\xi \mathbf{G}_t$. The local update formula is $\mathbf{w}_t^n = \mathbf{w}_{t-1_*} - \eta \mathbf{G}_{t-1}^n$. According to Assumption 1, we can obtain the loss upper bound of the global model $\mathbf{w}_t$:

$$F(\mathbf{w}_t) \leq F(\mathbf{w}_{t-1_*}) + < \nabla F(\mathbf{w}_{t-1_*}), \mathbf{w}_t - \mathbf{w}_{t-1_*} > + \frac{L}{2}\|\mathbf{w}_t - \mathbf{w}_{t-1_*}\|_2^2. \tag{18}$$

To investigate the gap between (3) and (18), we can compute their expected difference as:

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\Big[F_n(\mathbf{w}_t^n) + \nabla_{\mathbf{w}_t^n} F_n(\mathbf{w}_t^n)^T(\mathbf{w}_t - \mathbf{w}_t^n) + \frac{L}{2}\|\mathbf{w}_t - \mathbf{w}_t^n\|_2^2\Big] \\
&- \mathbb{E}\Big[F(\mathbf{w}_{t-1_*}) + < \nabla F(\mathbf{w}_{t-1_*}), \mathbf{w}_t - \mathbf{w}_{t-1_*} > + \frac{L}{2}\|\mathbf{w}_t - \mathbf{w}_{t-1_*}\|_2^2\Big] \\
\leq &\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\Big[F_n(\mathbf{w}_{t-1_*}) + < \nabla F_n(\mathbf{w}_{t-1_*}), (\mathbf{w}_t^n - \mathbf{w}_{t-1_*}) > + \frac{L}{2}\|\mathbf{w}_t^n - \mathbf{w}_{t-1_*}\|_2^2 \\
&+ \nabla F_n(\mathbf{w}_t^n)^T(\mathbf{w}_t - \mathbf{w}_{t-1_*} + \eta\mathbf{G}_{t-1}^n) + \frac{L}{2}\|\mathbf{w}_t - (\mathbf{w}_{t-1_*} - \eta\mathbf{G}_{t-1}^n)\|_2^2\Big] \\
&- \mathbb{E}\Big[F(\mathbf{w}_{t-1_*}) + < \nabla F(\mathbf{w}_{t-1_*}), \mathbf{w}_t - \mathbf{w}_{t-1_*} > + \frac{L}{2}\|\mathbf{w}_t - \mathbf{w}_{t-1_*}\|_2^2\Big] \\
= &\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\Big[- \eta\nabla F_n(\mathbf{w}_{t-1_*})^T\mathbf{G}_{t-1}^n + \frac{L}{2}\|\eta\mathbf{G}_{t-1}^n\|_2^2 \\
&+ < \nabla F_n(\mathbf{w}_t^n), -\eta\mathbf{G}_{t-1} + \eta\mathbf{G}_{t-1}^n > + \frac{L}{2}\| - \eta\mathbf{G}_{t-1} + \eta\mathbf{G}_{t-1}^n\|_2^2\Big] \\
&- \mathbb{E}\Big[< \nabla F(\mathbf{w}_{t-1_*}), -\eta\mathbf{G}_{t-1} > + \frac{L}{2}\|\eta\mathbf{G}_{t-1}\|_2^2\Big] \\
= &\eta\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\Big[- < \nabla F_n(\mathbf{w}_{t-1_*}), \mathbf{G}_{t-1}^n > + < \nabla F_n(\mathbf{w}_t^n), -\mathbf{G}_{t-1} + \mathbf{G}_{t-1}^n > + < \nabla F(\mathbf{w}_{t-1_*}), \mathbf{G}_{t-1} > \Big] \\
&+ \frac{L\eta^2}{2}\mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^{N}\|\mathbf{G}_{t-1}^n\|_2^2 + \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{G}_{t-1} - \mathbf{G}_{t-1}^n\|_2^2 - \|\mathbf{G}_{t-1}\|_2^2\Big] \\
\underset{(a)}{=} &\eta\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\Big[- < \nabla F_n(\mathbf{w}_{t-1_*}), \mathbf{G}_{t-1}^n > + < \nabla F_n(\mathbf{w}_t^n), -\mathbf{G}_{t-1} + \mathbf{G}_{t-1}^n > + < \nabla F(\mathbf{w}_{t-1_*}), \mathbf{G}_{t-1} > \Big] \\
= &\eta\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\Big[< \nabla F_n(\mathbf{w}_t^n) - \nabla F_n(\mathbf{w}_{t-1_*}), \mathbf{G}_{t-1}^n > - < \nabla F_n(\mathbf{w}_t^n) - \nabla F(\mathbf{w}_{t-1_*}), \mathbf{G}_{t-1} > \Big] \\
\underset{(b)}{=} &\eta\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\Big[< \nabla F_n(\mathbf{w}_t^n) - \nabla F_n(\mathbf{w}_{t-1_*}), \mathbf{G}_{t-1}^n > - < \nabla F_n(\mathbf{w}_t^n) - \nabla F_n(\mathbf{w}_{t-1_*}), \mathbf{G}_{t-1} > \Big] \\
= &\eta\frac{1}{N}\sum_{n=1}^{N} \mathbb{E} < \nabla F_n(\mathbf{w}_t^n) - \nabla F_n(\mathbf{w}_{t-1_*}), \tilde{\mathbf{G}}_{t-1}^n - \tilde{\mathbf{G}}_{t-1} > \\
\leq &\frac{1}{2N}\sum_{n=1}^{N} \mathbb{E}\Big(\|\nabla F_n(\mathbf{w}_t^n) - \nabla F_n(\mathbf{w}_{t-1_*})\|_2^2 + \eta^2\|\tilde{\mathbf{G}}_{t-1}^n - \tilde{\mathbf{G}}_{t-1}\|_2^2\Big) \\
\leq &\frac{1}{2N}\sum_{n=1}^{N} \mathbb{E}\Big(L\eta^2\|\mathbf{G}_{t-1}^n\|_2^2 + \eta^2\|\tilde{\mathbf{G}}_{t-1}^n - \tilde{\mathbf{G}}_{t-1}\|_2^2\Big)
\end{aligned} \tag{19}$$

where (a) holds because $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$ with $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. (b) holds because $\nabla F(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \nabla F_n(\mathbf{w})$ and $\frac{1}{N}\sum_{n=1}^{N} < a^n - \frac{1}{N}\sum_{i=1}^{N} b^i, c > = \frac{1}{N}\sum_{n=1}^{N} < a^n -$

$b^n, c>$. According to Assumption 3, we have

$$\|\mathbf{G}_{t-1}^n\|_2^2 = \|\sum_{e=1}^{E} \nabla f_n(\mathbf{w}_{t-1,e}^n)\|_2^2 \leq E\sum_{e=1}^{E}\|\nabla f_n(\mathbf{w}_{t-1,e}^n)\|_2^2 \leq E^2M^2. \tag{20}$$

Further, we can derive that

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\|\tilde{\mathbf{G}}_{t-1}^n - \tilde{\mathbf{G}}_{t-1}\|_2^2 \\
=&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\|\sum_{e=1}^{E}\nabla F_n(\mathbf{w}_{t-1,e}^n) - \frac{1}{N}\sum_{i=1}^{N}\sum_{e=1}^{E}\nabla F_i(\mathbf{w}_{t-1,e}^n)\|_2^2 \\
\leq&\frac{E}{N}\sum_{e=1}^{E}\sum_{n=1}^{N}\mathbb{E}\|\big(\nabla F_n(\mathbf{w}_{t-1,e}^n) - \nabla F_n(\mathbf{w}_{t-1_*})\big) + \big(\nabla F_n(\mathbf{w}_{t-1_*}) - \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{w}_{t-1_*})\big) \\
&- \frac{1}{N}\sum_{i=1}^{N}\big(\nabla F_i(\mathbf{w}_{t-1,e}^i) - \nabla F_i(\mathbf{w}_{t-1_*})\big)\|_2^2 \\
\leq&\frac{3E}{N}\sum_{e=1}^{E}\sum_{n=1}^{N}\mathbb{E}\Big[\|\big(\nabla F_n(\mathbf{w}_{t-1,e}^n) - \nabla F_n(\mathbf{w}_{t-1_*})\big)\|_2^2 + \|\big(\nabla F_n(\mathbf{w}_{t-1_*}) - \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{w}_{t-1_*})\big)\|_2^2 \\
&+ \|\frac{1}{N}\sum_{i=1}^{N}\big(\nabla F_i(\mathbf{w}_{t-1,e}^i) - \nabla F_i(\mathbf{w}_{t-1_*})\big)\|_2^2\Big] \\
\leq&\frac{3E}{N}\sum_{e=1}^{E}\sum_{n=1}^{N}\mathbb{E}\Big[L\|\mathbf{w}_{t-1,e}^n - \mathbf{w}_{t-1_*}\|_2^2 + \|\big(\nabla F_n(\mathbf{w}_{t-1_*}) - \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{w}_{t-1_*})\big)\|_2^2 \\
&+ \|\frac{1}{N}\sum_{i=1}^{N}\big(\nabla F_i(\mathbf{w}_{t-1,e}^i) - \nabla F_i(\mathbf{w}_{t-1_*})\big)\|_2^2\Big] \\
\underset{(a)}{\leq}&\frac{3E}{N}\sum_{e=1}^{E}\sum_{n=1}^{N}\mathbb{E}\Big[L\|\eta\sum_{e=1}^{E}\nabla f_n(\mathbf{w}_{t-1,e}^n)\|_2^2 + \sigma_g^2 + \frac{1}{N}\sum_{i=1}^{N}\|\nabla F_i(\mathbf{w}_{t-1,e}^i) - \nabla F_i(\mathbf{w}_{t-1_*})\|_2^2\Big] \\
\leq&\frac{3E}{N}\sum_{e=1}^{E}\sum_{n=1}^{N}\mathbb{E}\Big[L\eta^2E\sum_{e=1}^{E}\|\nabla f_n(\mathbf{w}_{t-1,e}^n)\|_2^2 + \sigma_g^2 + \frac{L}{N}\sum_{i=1}^{N}\|\eta\sum_{e=1}^{E}\nabla f_n(\mathbf{w}_{t-1,e}^n)\|_2^2\Big] \\
\underset{(b)}{\leq}&\frac{3E}{N}\sum_{e=1}^{E}\sum_{n=1}^{N}\Big[L\eta^2E^2\sigma_l^2 + \sigma_g^2 + L\eta^2E^2\sigma_l^2\Big] \\
=&6L\eta^2E^4\sigma_l^2 + 3E^2\sigma_g^2,
\end{aligned} \tag{21}$$

where (a) and (b) are derived based on Assumption 2. Bringing (20) and (21) back to (19) obtains

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[F_n(\mathbf{w}_t^n) + \nabla_{\mathbf{w}_t^n}F_n(\mathbf{w}_t^n)^T(\mathbf{w}_t - \mathbf{w}_t^n) + \frac{L}{2}\|\mathbf{w}_t - \mathbf{w}_t^n\|_2^2\Big] \\
&- \mathbb{E}\Big[F(\mathbf{w}_{t-1_*}) + < \nabla F(\mathbf{w}_{t-1_*}), \mathbf{w}_t - \mathbf{w}_{t-1_*} > + \frac{L}{2}\|\mathbf{w}_t - \mathbf{w}_{t-1_*}\|_2^2\Big] \\
\leq&\frac{1}{2N}\sum_{n=1}^{N}L\eta^2E^2M^2 + \frac{\eta^2}{2N}\sum_{n=1}^{N}\|\mathbf{G}_{t-1}^n - \bar{\mathbf{G}}_{t-1}\|_2^2 \\
\leq&\frac{L\eta^2E^2M^2}{2} + 3L\eta^4E^4\sigma_l^2 + \frac{3\eta^2E^2\sigma_g^2}{2}
\end{aligned} \tag{22}$$

Joining (17) and (22) together, we can immediately obtain

$$\mathbb{E}F(\mathbf{w}_{t_*}) \leq \mathbb{E}\big(F(\mathbf{w}_{t-1_*}) + < \nabla F(\mathbf{w}_{t-1_*}), \mathbf{w}_t - \mathbf{w}_{t-1_*} > + \frac{L}{2}\|\mathbf{w}_t - \mathbf{w}_{t-1_*}\|_2^2\big) + \frac{L\eta^2 E^2 M^2}{2} + 3L\eta^4 E^4 \sigma_l^2 + \frac{3E^2\eta^2\sigma_g^2}{2}, \tag{23}$$

which completes the proof.

Theorem 4 *Consider problem* (1) *under Assumption 1,2, and 3. If the learning rate* η *satisfies* $0 < \eta \leq \frac{1}{LE}$ *and all clients are sampled in each round, then for all* $T \geq 1$*, the global model* $\mathbf{w}_{t_*}$ *solved by (3) achieves asymptotic convergence, i.e.,* $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_{t-1_*})\|_2^2 \to 0$ *as* $T \to \infty$.

Proof: We define $\mathbf{w}_t$ as the global model obtained by aggregating all local models $\mathbf{w}_t^n$, i.e., $\mathbf{w}_t = \frac{1}{N}\sum_{n=1}^{N}\mathbf{w}_t^n$. Each local model $\mathbf{w}_t^n$ is calculated from the global model $\mathbf{w}_{t-1_*}$ in previous round, with E local SGD mini-batch steps, $\mathbf{w}_{t-1,E}^n = \mathbf{w}_{t-1_*} - \eta\sum_{e=1}^{E}\nabla f_n(\mathbf{w}_{t-1,e}^n)$ with $\mathbf{w}_{t-1,1}^n) = \mathbf{w}_{t-1_*}$. Then, for any $t > 1$, according to the L-smoothness of function, we have

$$\mathbb{E}F(\mathbf{w}_t) \leq \mathbb{E}F(\mathbf{w}_{t-1_*}) + \mathbb{E} < \nabla F(\mathbf{w}_{t-1_*}), \mathbf{w}_t - \mathbf{w}_{t-1_*} > + \frac{L}{2}\mathbb{E}\|\mathbf{w}_t - \mathbf{w}_{t-1_*}\|_2^2. \tag{24}$$

Note that

$$\begin{aligned}
&\mathbb{E}\|\mathbf{w}_t - \mathbf{w}_{t-1_*}\|_2^2 \\
=&\eta^2\mathbb{E}\|\frac{1}{N}\sum_{n=1}^{N}\sum_{e=1}^{E}\nabla f_n(\mathbf{w}_{t-1,e}^n)\|_2^2 \\
=&\eta^2\mathbb{E}\|\frac{1}{N}\sum_{n=1}^{N}\sum_{e=1}^{E}\big(\nabla f_n(\mathbf{w}_{t-1,e}^n) - \nabla F_n(\mathbf{w}_{t-1,e}^n) + \nabla F_n(\mathbf{w}_{t-1,e}^n)\big)\|_2^2 \\
\underset{(a)}{=}&\eta^2\mathbb{E}\|\frac{1}{N}\sum_{n=1}^{N}\sum_{e=1}^{E}\big(\nabla f_n(\mathbf{w}_{t-1,e}^n) - \nabla F_n(\mathbf{w}_{t-1,e}^n)\|_2^2 + \eta^2\mathbb{E}\|\frac{1}{N}\sum_{n=1}^{N}\sum_{e=1}^{E}\nabla F_n(\mathbf{w}_{t-1,e}^n)\big)\|_2^2 \\
\underset{(b)}{=}&\eta^2\frac{1}{N^2}\sum_{n=1}^{N}\mathbb{E}\|\sum_{e=1}^{E}\big(\nabla f_n(\mathbf{w}_{t-1,e}^n) - \nabla F_n(\mathbf{w}_{t-1,e}^n)\|_2^2 + \eta^2\mathbb{E}\|\frac{1}{N}\sum_{n=1}^{N}\sum_{e=1}^{E}\nabla F_n(\mathbf{w}_{t-1,e}^n)\|_2^2 \\
\underset{(c)}{\leq}&\frac{\eta^2 E\sigma_l^2}{N} + \eta^2 E\frac{1}{N}\sum_{n=1}^{N}\sum_{e=1}^{E}\mathbb{E}\|\nabla F_n(\mathbf{w}_{t-1,e}^n)\|_2^2,
\end{aligned} \tag{25}$$

where (a) follows by $\mathbf{Ez} = \mathbf{E}\|\mathbf{z} - \mathbf{Ez}\|^2 + \|\mathbf{Ez}\|^2$; (b) follows because $\mathbb{E}f_n(\mathbf{w}) = F_n(\mathbf{w})$; (c) holds due to Assumption 3.

Further, we note that

$$\begin{aligned}
&\mathbb{E} < \nabla F(\mathbf{w}_{t-1_*}), \mathbf{w}_t - \mathbf{w}_{t-1_*} > \\
=&\eta\mathbb{E} < \nabla F(\mathbf{w}_{t-1_*}), -\frac{1}{N}\sum_{n=1}^{N}\sum_{e=1}^{E}\nabla f_n(\mathbf{w}_{t-1,e}^n) > \\
=&-\sum_{e=1}^{E}\eta\mathbb{E} < \nabla F(\mathbf{w}_{t-1_*}), \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(\mathbf{w}_{t-1,e}^n) > \\
\underset{(a)}{=}&-\frac{\eta}{2}\sum_{e=1}^{E}\mathbb{E}\Big[\|\nabla F(\mathbf{w}_{t-1_*})\|_2^2 + \|\frac{1}{N}\sum_{n=1}^{N}\nabla F_n(\mathbf{w}_{t-1,e}^n)\|_2^2 - \|\nabla F(\mathbf{w}_{t-1_*}) - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(\mathbf{w}_{t-1,e}^n)\|_2^2\Big] \\
\leq&-\frac{\eta}{2}\sum_{e=1}^{E}\mathbb{E}\Big[\|\nabla F(\mathbf{w}_{t-1_*})\|_2^2 + \frac{1}{N}\sum_{n=1}^{N}\|\nabla F_n(\mathbf{w}_{t-1,e}^n)\|_2^2 - \|\nabla F(\mathbf{w}_{t-1_*}) - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(\mathbf{w}_{t-1,e}^n)\|_2^2\Big],
\end{aligned} \tag{26}$$

where (a) holds because $2ab = a^2 + b^2 - (a-b)^2$. Besides, we have the following inequality

$$\begin{aligned}
&\|\nabla F(\mathbf{w}_{t-1_*}) - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(\mathbf{w}_{t-1,e}^n)\|_2^2 \\
=&\|\frac{1}{N}\sum_{n=1}^{N}\big[\nabla F_n(\mathbf{w}_{t-1_*}) - \nabla F_n(\mathbf{w}_{t-1,e}^n)\big]\|_2^2 \\
\leq&\frac{1}{N}\sum_{n=1}^{N}\|\nabla F_n(\mathbf{w}_{t-1_*}) - \nabla F_n(\mathbf{w}_{t-1,e}^n)\|_2^2 \\
\leq&\frac{L^2}{N}\sum_{n=1}^{N}\|\mathbf{w}_{t-1_*} - \mathbf{w}_{t-1,e}^n\|_2^2 \\
=&\frac{\eta^2L^2}{N}\sum_{n=1}^{N}\|\sum_{i=1}^{e}\nabla f_n(\mathbf{w}_{t-1,e}^n)\|_2^2 \\
\leq&\frac{e\eta^2L^2}{N}\sum_{n=1}^{N}\sum_{i=1}^{e}\|\nabla f_n(\mathbf{w}_{t-1,e}^n)\|_2^2 \\
\leq&\frac{e\eta^2L^2}{N}\sum_{n=1}^{N}\sum_{i=1}^{e}M^2 \\
=&e^2\eta^2L^2M^2,
\end{aligned} \tag{27}$$

where the last inequality is due to Assumption 3.

Bringing (27) back to (26) and then bringing the obtained inequality together with (25) to (24) derives

$$\begin{aligned}
\mathbb{E}F(\mathbf{w}_t) \leq&\mathbb{E}F(\mathbf{w}_{t-1_*}) + \mathbb{E} < \nabla F(\mathbf{w}_{t-1_*}), \mathbf{w}_t - \mathbf{w}_{t-1_*} > + \frac{L}{2}\mathbb{E}\|\mathbf{w}_t - \mathbf{w}_{t-1_*}\|_2^2 \\
\leq&\mathbb{E}F(\mathbf{w}_{t-1_*}) - \frac{\eta}{2}\sum_{e=1}^{E}\mathbb{E}\Big[\|\nabla F(\mathbf{w}_{t-1_*})\|_2^2 + \frac{1}{N}\sum_{n=1}^{N}\|\nabla F_n(\mathbf{w}_{t-1,e}^n)\|_2^2 - e^2\eta^2L^2M^2\Big] \\
&+ \frac{L}{2}\bigg(\frac{\eta^2E\sigma_l^2}{N} + \eta^2E\frac{1}{N}\sum_{n=1}^{N}\sum_{e=1}^{E}\mathbb{E}\|\nabla F_n(\mathbf{w}_{t-1,e}^n)\|_2^2\bigg) \\
=&\mathbb{E}F(\mathbf{w}_{t-1_*}) - \frac{\eta}{2}\sum_{e=1}^{E}\|\nabla F(\mathbf{w}_{t-1_*})\|_2^2 + \frac{\eta}{2}\sum_{e=1}^{E}e^2\eta^2L^2M^2 + \frac{L\eta^2E\sigma_l^2}{2N} \\
&- (\frac{\eta}{2} - \frac{L\eta^2E}{2})\frac{1}{N}\sum_{e=1}^{E}\sum_{n=1}^{N}\|\nabla F_n(\mathbf{w}_{t-1,e}^n)\|_2^2 \\
\leq&\mathbb{E}F(\mathbf{w}_{t-1_*}) - \frac{\eta E}{2}\|\nabla F(\mathbf{w}_{t-1_*})\|_2^2 + \frac{E^3\eta^3L^2M^2}{2} + \frac{L\eta^2E\sigma_l^2}{2N},
\end{aligned} \tag{28}$$

where the last inequality holds when $\eta \leq \frac{1}{LE}$. Based on Lemma 1, we have

$$\begin{aligned}
\mathbb{E}F(\mathbf{w}_{t_*}) \leq&\mathbb{E}F(\mathbf{w}_{t-1_*}) + \mathbb{E} < \nabla F(\mathbf{w}_{t-1_*}), \mathbf{w}_t - \mathbf{w}_{t-1_*} > + \frac{L}{2}\mathbb{E}\|\mathbf{w}_t - \mathbf{w}_{t-1_*}\|_2^2 \\
&+ \frac{L\eta^2E^2M^2}{2} + 3L\eta^4E^4\sigma_l^2 + \frac{3E^2\sigma_g^2}{2} \\
\leq&\mathbb{E}F(\mathbf{w}_{t-1_*}) - \frac{\eta E}{2}\|\nabla F(\mathbf{w}_{t-1_*})\|_2^2 + \frac{E^3\eta^3L^2M^2}{2} + \frac{L\eta^2E\sigma_l^2}{2N} \\
&+ \frac{L\eta^2E^2M^2}{2} + 3L\eta^4E^4\sigma_l^2 + \frac{3E^2\eta^2\sigma_g^2}{2}.
\end{aligned} \tag{29}$$

Dividing both sides of (29) by $\frac{\eta E}{2}$ and rearranging terms yields

$$\begin{aligned}\|\nabla F(\mathbf{w}_{t-1_*})\|_2^2 \leq & \mathbb{E}\frac{2(F(\mathbf{w}_{t-1_*}) - \mathbb{E}F(\mathbf{w}_{t_*}))}{\eta E} \\ & + E^2\eta^2L^2M^2 + \frac{L\eta\sigma_l^2}{N} + L\eta EM^2 + 6L\eta^3E^3\sigma_l^2 + 3E\eta\sigma_g^2. \end{aligned} \tag{30}$$

Summing both sides of (30) over $t = 1, \ldots, T$ and dividing by T derives

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_{t-1_*})\|_2^2 \leq & \mathbb{E}\frac{2(F(\mathbf{w}_0) - \mathbb{E}F(\mathbf{w}_{T_*}))}{\eta TE} \\ & + E^2\eta^2L^2M^2 + \frac{L\eta\sigma_l^2}{N} + L\eta EM^2 + 6L\eta^3E^3\sigma_l^2 + 3E\eta\sigma_g^2 \\ \leq & \mathbb{E}\frac{2(F(\mathbf{w}_0) - \mathbb{E}F(\mathbf{w}_*))}{\eta TE} \\ & + E^2\eta^2L^2M^2 + \frac{L\eta\sigma_l^2}{N} + L\eta EM^2 + 6L\eta^3E^3\sigma_l^2 + 3E\eta\sigma_g^2, \end{aligned} \tag{31}$$

where $F(\mathbf{w}_*)$ denotes the minimum of the function $F(\mathbf{w})$. Setting $\eta = \sqrt{\frac{N}{TE}}$ in (31) derives

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_{t-1_*})\|_2^2 \leq & \mathbb{E}\frac{2(F(\mathbf{w}_0) - \mathbb{E}F(\mathbf{w}_*))}{\sqrt{TNE}} \\ & + \frac{NEL^2M^2}{T} + \frac{L\sigma_l^2}{\sqrt{NTE}} + LM^2\sqrt{\frac{NE}{T}} + 6L\sigma_l^2(\frac{EN}{T})^{\frac{3}{2}} + 3\sigma_g^2\sqrt{\frac{EN}{T}}, \end{aligned} \tag{32}$$

which completes the proof.

C Discussion about the Efficiency

For different aggregation strategies about our FedCDA, we calculate the average polymerization time of each round. Details can be found in the table C below. It can be found that the cost of approximation selection is several times more than directly computing the average of all local models. Although the increased ratio is high as compared to FedAvg, the absolute value is still small and can be totally acceptable in practical settings.

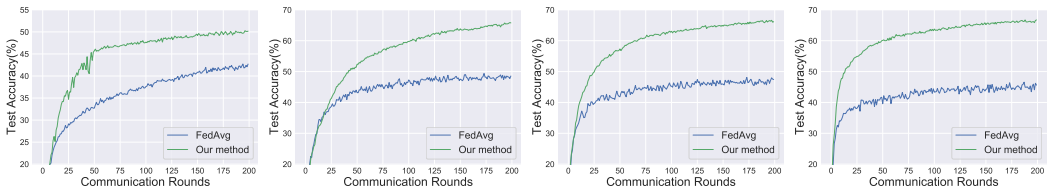
Method	Acc(%)			Time(ms)		
Dirichlet (n, α)	(20, 0.1)	(20, 0.3)	(20, 0.5)	(20, 0.1)	(20, 0.3)	(20, 0.5)
approximate	63.78±0.38	72.62±0.35	77.07±0.26	24	69	68
optimal	64.98±0.34	73.11±0.28	77.55±0.17	154	146	142
newest	59.49±0.15	66.29±0.11	68.51±0.12	33	27	19
random	60.78±1.22	67.67±1.51	70.81±0.89	21	14	19
FedAvg	50.43±1.68	61.11±2.68	67.37±1.69	18	15	11

D More Experiments

To demonstrate the effectiveness of our proposed method over a broader range of hyper-parameters. We conduct more experiments by using a different learning rate and smaller local epochs. The results are shown in the following table and figures. The dataset is CIFAR-100.

Table 3: The learning rate is 0.1, weight decay is 1e-3, local epoch is 5, and the optimizer is SGD without momentum.

Method	Accuracy
FedAvg	48.20% (±0.42%)
FedProx	54.46% (±1.01%)
FedExP	57.00% (±1.02%)
FedSAM	57.94% (±0.36%)
FedDF	58.83% (±0.81%)
FedGEN	55.29% (±1.37%)
FedMA	49.45% (±0.61%)
GAMF	50.87% (±0.27%)
FedLAW	51.78% (±0.97%)
FedCDA	**65.50%** (±0.25%)



(a) 100 clients (b) Local epoch = 5 (c) Local epoch = 10 (d) Local epoch = 20

Figure 6: (a) shows the result of 100 clients with a participation ratio of 0.1 on `FedCDA` and FedAvg. (b), (c) and (d) show the results of 20 clients with a participation ratio of 0.2 and local epoch of 5, 10 and 20 on `FedCDA` and FedAvg.