AggEnhance: Aggregation Enhancement by Class Interior Points in Federated Learning with Non-IID Data

JINXIANG OU, YUNHENG SHEN, and FENG WANG, Tsinghua University, China
QIAO LIU, Stanford University, USA
XUEGONG ZHANG and HAIRONG LV, Tsinghua University, China

Federated learning (FL) is a privacy-preserving paradigm for multi-institutional collaborations, where the aggregation is an essential procedure after training on the local datasets. Conventional aggregation algorithms often apply a weighted averaging of the updates generated from distributed machines to update the global model. However, while the data distributions are non-IID, the large discrepancy between the local updates might lead to a poor averaged result and a lower convergence speed, i.e., more iterations required to achieve a certain performance. To solve this problem, this article proposes a novel method named AggEnhance for enhancing the aggregation, where we synthesize a group of reliable samples from the local models and tune the aggregated result on them. These samples, named class interior points (CIPs) in this work, bound the relevant decision boundaries that ensure the performance of aggregated result. To the best of our knowledge, this is the first work to explicitly design an enhancing method for the aggregation in prevailing FL pipelines. A series of experiments on real data demonstrate that our method has noticeable improvements of the convergence in non-IID scenarios. In particular, our approach reduces the iterations by 31.87% on average for the CIFAR10 dataset and 43.90% for the PASCAL VOC dataset. Since our method does not modify other procedures of FL pipelines, it is easy to apply to most existing FL frameworks. Furthermore, it does not require additional data transmitted from the local clients to the global server, thus holding the same security level as the original FL algorithms.

CCS Concepts: • **Security and privacy**; • **Computing methodologies** → **Artificial intelligence**; **Supervised learning by classification**; **Distributed computing methodologies**;

Additional Key Words and Phrases: Federated learning, aggregation enhancement, class interior points, non-IID, communication

ACM Reference format:
Jinxiang Ou, Yunheng Shen, Feng Wang, Qiao Liu, Xuegong Zhang, and Hairong Lv. 2022. AggEnhance: Aggregation Enhancement by Class Interior Points in Federated Learning with Non-IID Data. *ACM Trans. Intell. Syst. Technol.* 13, 6, Article 104 (September 2022), 25 pages.
https://doi.org/10.1145/3544495

Jinxiang Ou and Yunheng Shen contributed equally to this research.
This article is supported in part by National Natural Science Foundation of China under grant Nos. 42050101 and 61721003.
Authors' addresses: J. Ou, Y. Shen, F. Wang, X. Zhang, and H. Lv (corresponding author), Beijing National Research Center for Information Science and Technology, Dept. of Automation, Tsinghua University, Haidian, Beijing, China; emails: {ojx19, shenyh19, wangf19}@mails.tsinghua.edu.cn, {zhangxg, lvhairong}@tsinghua.edu.cn; Q. Liu, Department of Statistics, Stanford University, Stanford, California 94305; email: liuqiao@stanford.edu.

2157-6904/2022/09-ART104
https://doi.org/10.1145/3544495

1 INTRODUCTION

Deep learning models typically need training on large-scale samples to obtain excellent performance. However, in the real world, data are often distributed at a large number of devices (e.g., smartphones and wearable sensors). It is challenging to collect data from these devices due to the limitations of communication, power, and privacy issues. To tackle this problem, Google proposed the concept of **federated learning (FL)** [19] recently. In FL scenarios, data are distributed at multiple devices, but at the training mode, only the model weights instead of the original data are transmitted to the server. Since the sizes of model weights are much smaller than the training data and do not contain personal information, FL alleviates the communication and privacy issues. Federated learning has been applied to smartphone applications such as keyword spotting [14] and keyboard prediction [6]. In the medical field, References [10, 16, 20, 21] reported potential federated learning applications in medical image processing and digital health.

Figure 1 illustrates the fundamental pipelines of federated learning. First, the server broadcasts the weights of the global model to a number of selected clients. Initialized by the global weights, each client updates the model by its local data and uploads the updated parameters to the server. After collecting all updates, the server applies an aggregation operator to update the global model and then broadcasts its new weights for the next round of training. There is no doubt that aggregation is an essential step in federated learning. In conventional aggregation methods, represented by **Federated-Averaging (FedAvg)** [19], the server takes a weighted average of the local weights as the new weights of the global model. However, these methods take effect based on the assumption that the discrepancy between updates from different clients is small. In heterogeneous scenarios, where the training data are **non-independent identically distributed (non-IID)** on the clients, the assumption does not hold. Since the weighted average of feasible solutions may be infeasible for complex models, the averaging-based aggregation methods converge slowly. A proper aggregation method is a significant factor in the convergence of federated learning with non-IID data, where it is harder to converge and requires more iterations to train, transmit, and aggregate to reach a certain performance than IID data settings.

This article focuses on enhancing the aggregation without any sensitive data. We propose a novel method, **AggEnhance (aggregation enhancement)**, to enhance the aggregation results. As Figure 1 shows, our method inserts an enhancing procedure in FL after updating the global model and before the next training round, which means that it is compatible with most federated learning frameworks. The proposed method generates a group of reliable synthetic samples (named class interior points, CIPs) for each class in the enhancing procedure, which contain class-specific information. Then, we tune the global model on these synthetic samples to modify the decision boundaries and help the global model reach a better solution. Our contributions are as follows:

(1) We point out that the vanilla weighted average is not a good aggregation strategy, which will cause slow convergence of FL in Non-IID scenarios, and then we propose a novel method named AggEnhance to overcome it. In AggEnhance, we generate a certain amount of consentaneous CIP samples and fine-tune the aggregation results on them and achieve a non-linear aggregation of local model weights through the above aggregation enhancement. Furthermore, our method does not require any additional local privacy data, while the existing methods (e.g., knowledge distillation) require the server to access the local intermediate results or share certain real data among all clients.
(2) We explore an available and easy-to-follow method to generate the CIPs and experimentally verify the validity of CIPs. The method generates a group of samples based on the local model that do not reveal privacy, so the gradient of the optimization function at these few points is approximated over the entire local data. We also propose a heuristic initialization

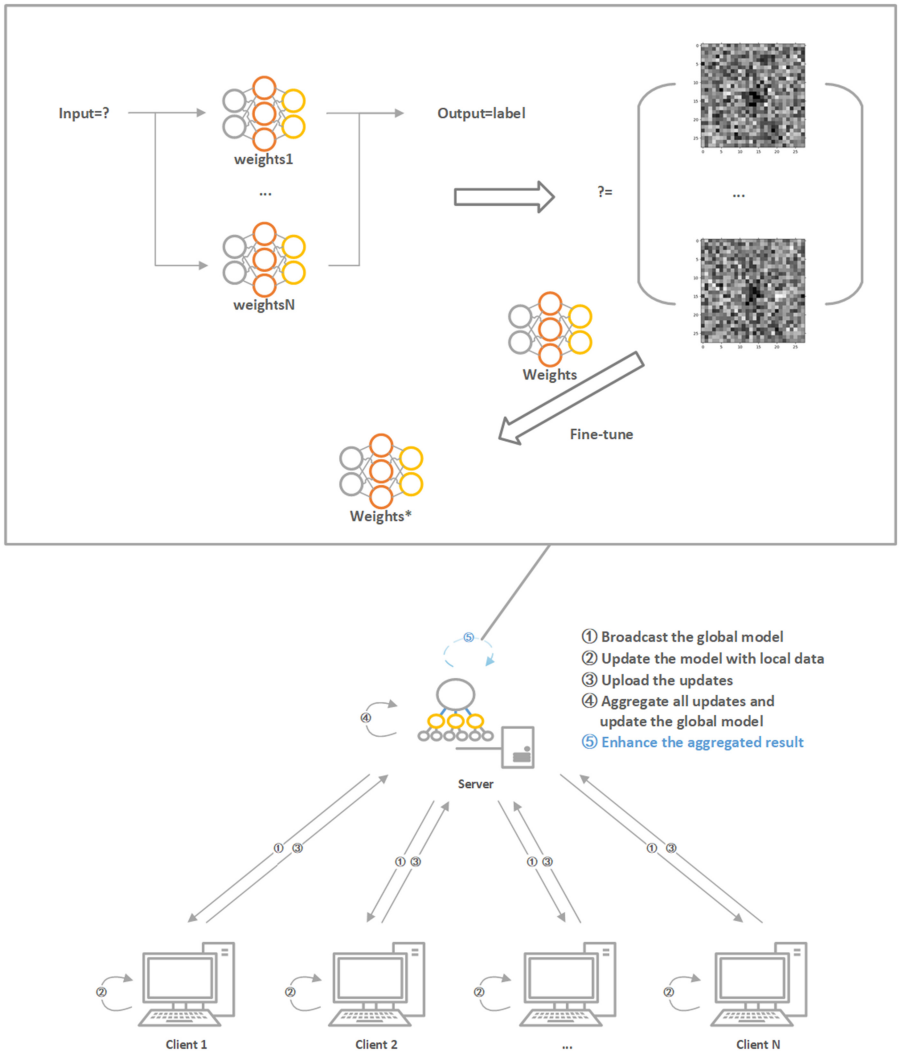


Fig. 1. Illustration of our aggregation enhancement method in federated learning. The sequence of steps in basic FL framework is ①→②→③→④→①. Our method only needs to insert a module ⑤ between ④ and ① to enhance the aggregated result.

and consensus CIP screening method to help obtain more efficient CIPs. Moreover, our experiments show that the model trained only on the CIPs set could achieve 78.2% prediction accuracy on the test set of MNIST, and the global model enhanced by CIPs is even better than taking the same amounts of raw data, which means that the CIPs contain enough effective information about the decision boundary.

(3) We perform expensive experiments on image classification tasks to show that our method makes the convergence faster to achieve a certain predictive performance, i.e., less communication cost. In non-IID scenarios, our method reduces the communication rounds by 31.87% on average to achieve 50% predictive accuracy on CIFAR10, 43.90% to achieve 30% accuracy on the PASCAL VOC dataset, and 18.64% to achieve 40% accuracy on the FER dataset.

(4) We provide discussion about the privacy security of our method. The experiments show that it is impossible to distinguish the CIPs by the naked eye, and our method is still applicative to FL even though we apply differential privacy to prevent deep leakage. And it requires only a few changes to apply to secure aggregation-based FL. The AggEnhance algorithm is able to improve FL without hindering the privacy-preserving capability of FL.

2 RELATED WORK

Federated learning aims to train models among multiple machines collaboratively without exposing sensitive data to each other. Fully-batch gradient descent method is available to address the problem. Denote w as the model weights, N as the number of devices, and n_i as the number of samples in the ith device. The global objective function is to minimize the average of all clients' training loss as:

$$\min_w \mathcal{L}\left(w; \bigcup_{i=1}^{N} D_i\right) \approx \sum_{i=1}^{N} \alpha_i \mathcal{L}(w; D_i), \tag{1}$$

where the local objective $\mathcal{L}(w; D_i)$ is usually the mean of loss function on the local data distribution D_i, e.g., cross-entropy in classification problems, and the averaging coefficients α is the proportion vector of data volumes:

$$\alpha_i = \frac{n_i}{\sum_{j=1}^{N} n_j}. \tag{2}$$

Whereas, in many scenes, there are such massive devices in the federation that it is hard to optimize the global objective. Inspired by mini-batch **stochastic gradient descent (SGD)** method, the FedSGD algorithm [19] provides an approximate solution. For each round, FedSGD randomly selects $K (K \leq N)$ devices, $d_1, \ldots, d_K$. The server broadcasts the global model weights to the selected clients for computing the mean of gradients on all local private data. Then each selected device d_k sends its local gradient $\nabla_w \mathcal{L}(w; D_k)$ to the global server. After receiving the gradients from all selected devices, the server updates the global weights by:

$$w \leftarrow w - \eta \frac{1}{\sum_{k \in \{d_1, \ldots, d_K\}} n_k} \sum_{k \in \{d_1, \ldots, d_K\}} n_k \nabla_w \mathcal{L}(w; D_k), \tag{3}$$

where η is the learning rate. After certain rounds of communication, the global model converges to a local minimum of Equation (1). Since the local model w_k is initialized by w and updated on the training process by:

$$w_k \leftarrow w - \eta \nabla_w \mathcal{L}(w; D_k), \tag{4}$$

Equation (3) is equivalent to the following Equation (5):

$$w \leftarrow \frac{1}{\sum_{k \in \{d_1, \ldots, d_K\}} n_k} \sum_{k \in \{d_1, \ldots, d_K\}} n_k w_k. \tag{5}$$

Nevertheless, FedSGD needs to communicate between the server and the selected clients for each training iteration, leading to massive communication costs. In the real world, particularly on edge networks and mobile networks, communication between devices is usually limited. The FedAvg framework [19] alleviates it by the extension of FedSGD. The local machine trains the model on local data for E epochs in FedAvg, instead of computing the gradients for only one time. FedAvg retains the aggregation step by Equation (5). References [11, 17] provided necessary conditions to guarantee its convergence for FedAvg.

However, the statistical heterogeneity is still a huge challenge of federated learning, primarily referring to **non-independent identically distributed (non-IID)** data across different clients. Reference [26] showed that the non-IID data might cause a significant decrease in the

performance. More iterations are needed for convergence on non-IID datasets than that on the IID distributions, i.e., more communication is required, and the predictive performance of the final model is also declined. To address this problem, Reference [18] introduced local representation learning and proposed **LocalGlobal-Federated-Averaging (LG-FedAvg)**. The local representation learner and the classifier are trained collaboratively at the training phase in LG-FedAvg. The global server only aggregates the local classifiers at the federated aggregation phase, which enables LG-FedAvg to alleviate data heterogeneity by using flexible local representation learners. FedProx [15] adds a proximal term to the local objective function that constrains the local update to be small. The performance of FedAvg is degraded in non-IID scenarios due to the large discrepancy among all local updates. Hence, the local constraints of FedProx make the averaging result more robust than FedAvg. Besides, Reference [14] found that using Adam optimizer could significantly reduce communication rounds.

Client selection is another direction to tackle this problem. Since the data on different devices are in various levels of quality, filtering the poor-quality clients out might be helpful for federated learning. The principle is to select the clients that have data distribution as close to IID as possible; e.g., Reference [3] applied active learning to federated learning for selecting efficient clients. Their results showed that the method reduced the training iterations by 20%–70% to achieve the same accuracy.

The above methods to deal with non-IID data primarily consider the training process on local devices. From the perspective of the server, Reference [5] developed a one-shot federated learning algorithm where the server distills all local models for the ensemble to obtain a useful student model. This process only needs one round of communication. However, this algorithm requires the server to store data for distillation, which might cause insecurity for data privacy. Reference [27] applied dataset distillation techniques to each local dataset. The server collects all distilled datasets from clients and trains the model only on them. Reference [25] aggregated all updates by the median-based and trimmed-mean-based gradient descent algorithms. Reference [2] applied clustering algorithms to detect and abandon aggregated models' outliers. Due to these robust aggregation methods, the aggregated results are more robust to handle the poor-quality data, e.g., non-IID data to some extent. Handling the permutation invariance of neurons by matching the weights in a layer-wise manner, **Federated-Matched-Averaging (FedMA)** [23] decreases the discrepancy of local updates and improves the performance on non-IID settings.

In this work, we design a novel framework for aggregation enhancement for federated learning pipelines. As shown in Figure 1, after the aggregation step of each round, we design a procedure to enhance the aggregation result. We define a type of data, the class interior points, which helps refine the decision boundaries. Our method synthesizes these data from all collected local models and tunes the aggregated result on them in the enhancing procedure. This method requires no additional modification of the above existing FL algorithms and is easy to be applied to them. We provide a series of experimental results to verify the efficiency of our method.

3 METHODS

3.1 AggEnhance

This work addresses the problem where deep learning models are not easy to converge in federated systems with non-IID data in a novel perspective. Due to the non-IID data, the local models trained on different data distributions might not have a consistent update direction, which causes their weighted average uncertain to move towards the global optimum after each communication round. Moreover, the aggregated global model is the initial model in the next local update round, and a bad aggregation result might make the training process of deep models unstable in the averaging-based frameworks. The above factors cause that FL converges slowly with non-IID data.

Therefore, we propose a novel method named AggEnhance to enhance the aggregation result rather than to modify the training process or to adjust the weighted coefficients to tackle this problem, which is shown in Algorithm 1.

ALGORITHM 1: AggEnhance algorithm

Input: number of communication rounds T, number of clients K, number of local epochs E, batch size B, learning rate η, number of AggEnhance epochs E_s, AggEnhance learning rate η_s
Output: the global model $\boldsymbol{w}^T$

Initialize the global model $\boldsymbol{w}^0$, the class interior points set S
for communicate round $t = 0, 1, 2, \ldots, T-1$ **do**
  for client $k = 1, 2, .., K$ **do**
    $\boldsymbol{w}_k^t \leftarrow \boldsymbol{w}^t$
    $\boldsymbol{w}_k^{t+1} \leftarrow$ **ModelTraining**$(\boldsymbol{w}_k^t, D_k, E, \eta)$
  end for
  $S \leftarrow$ **UpdateCIPSet**$(\boldsymbol{w}_1^{t+1}, \ldots, \boldsymbol{w}_K^{t+1}; S)$ **(Follows Algorithm 2)**
  $\boldsymbol{w}^{t+1} \leftarrow \sum_i \frac{|D_i|}{|D|}(\boldsymbol{w}_i^{t+1})$
  $\boldsymbol{w}^{t+1} \leftarrow$ **ModelTraining**$(\boldsymbol{w}^{t+1}, S, E_s, \eta_s)$
end for
return $\boldsymbol{w}^T$

ModelTraining ($\boldsymbol{w}$,D,E,η)**:**
for epoch $e = 1, 2, \ldots, E$ **do**
  for each batch $\mathcal{B} \subset D$ **do**
    $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \nabla \ell(\boldsymbol{w}, \mathcal{B})$
  end for
end for
return $\boldsymbol{w}$

We consider constructing a global set of synthetic samples S and fine-tuning the averaging results on it by:

$$w' \leftarrow w - \eta \nabla_w \mathcal{L}(w; S), \tag{6}$$

where w is the averaging result and w' is the final one.

Denote D as the whole federated system's data distribution, which approximates to the union of all training samples in the federation:

$$D \approx \bigcup_{i=1}^{N} D_i. \tag{7}$$

The Taylor approximation of the objective function is:

$$\mathcal{L}(w'; D) \approx \mathcal{L}(w; D) - \eta \nabla_w^T \mathcal{L}(w; S) \nabla_w \mathcal{L}(w; D). \tag{8}$$

While S contains enough information of D to guarantee that $\mathcal{L}(w; S)$ and $\mathcal{L}(w; D)$ hold similar gradient directions at the solution w, i.e.,

$$\nabla_w^T \mathcal{L}(w; S) \nabla_w \mathcal{L}(w; D) \geq 0, \tag{9}$$

we obtain that:

$$\mathcal{L}(w'; D) \leq \mathcal{L}(w; D). \tag{10}$$

In other words, if we find a set of data S whose distribution approximates to the whole distribution D, then tuning the result of averaging on S may enhance the aggregation process. We will introduce in detail how to construct an appropriate S in the next section.

3.2 Update CIP Set

One essential step in the proposed algorithm is to obtain an appropriate S that contains enough global information in the server without additional training data from the clients. This work provides an alternative solution by generating the class interior points from the collected models, which are defined by Definition 3.1. We define S to be a continuously updated dataset that is made up by the class interior points of all labels, and the whole algorithm about obtaining S is described as Algorithm 2. This section elaborates on the motivation and theoretical reliability of Algorithm 2.

Denote $\mathcal{F}$ as the feature encoder that maps the input x to a d-dimensional hidden space, C as the classifier that constructs a decision surface in a d-dimensional feature space and outputs the category probability of a sample. The entire model is represented as $C(\mathcal{F}(\cdot))$ and parameterized by $w = \{w_{\mathcal{F}}, w_C\}$. Denote $C_j(\mathcal{F}(\cdot))$ as the jth element of $C(\mathcal{F}(\cdot))$, a sample x is discriminated as the category c_i when $i = \arg\max_j C_j(\mathcal{F}(x))$. Specifically, if $C_i(\mathcal{F}(x)) > 0.5$, then the input x must be discriminated as the category c_i.

Definition 3.1. For a decision model $C(\mathcal{F}(\cdot))$, an input s is defined as a **class interior point (CIP)** of the category c_i if and only if $\exists \mathcal{R} > 0$, s.t.$\forall \Delta f \in \mathbb{R}^d, \|\Delta f\| < \mathcal{R}$, $\arg\max_j C_j(\mathcal{F}(s) + \Delta f) = i$.

Obviously, $i = \arg\max_j C_j(\mathcal{F}(s))$ is a necessary condition for s to be a CIP of the category c_i, i.e., s belongs to the category c_i.

Corollary 3.2. *Let $\mathcal{L}(\cdot;\cdot)$ be a continuous loss function, for all feasible $w = \{w_{\mathcal{F}}, w_C\}$, $\exists \epsilon > 0$, s is determined to be a class interior point of category c_i when $\mathcal{L}(w; \{(s, c_i)\}) < \epsilon$.*

Proof. Let $f = \mathcal{F}(s)$, and regard $\mathcal{L}(w; \{(s, c_i)\})$ as a function of f. Performing a first-order Taylor expansion on $\mathcal{L}(f)$, we have

$$\mathcal{L}(f + \Delta f) \approx \mathcal{L}(f) + \nabla_f^T \mathcal{L}(f) \Delta f \leq \mathcal{L}(f) + \mathcal{R} \|\nabla_f \mathcal{L}(f)\|, \forall \|\Delta f\| < \mathcal{R}.$$

The value of ϵ is given by the specific form of the loss function, as long as it satisfies that $\arg\max_j C_j(\mathcal{F}(s)) = i$ when $\mathcal{L}(w; \{(s, c_i)\}) < \epsilon$. Taking the cross-entropy loss as an example, we can choose $\epsilon = \log 2$, because $\mathcal{L}(w; \{(s, c_i)\}) < \log 2$ is equivalent to $C_i(\mathcal{F}(s)) > 0.5$.

When $\mathcal{L}(f) < \epsilon$, note that if we take $\mathcal{R} = \frac{\epsilon - \mathcal{L}(f)}{2\|\nabla_f \mathcal{L}(f)\|}$, then we have $\mathcal{L}(f + \Delta f) \leq \frac{\epsilon + \mathcal{L}(f)}{2} \leq \epsilon$, which means that $C_i(\mathcal{F}(s) + \Delta f) > 0.5$ and $\arg\max_j C_i(\mathcal{F}(s) + \Delta f) = i$. By Definition 3.1, s is a class interior point of category c_i. □

Therefore, we identify a CIP by calculating $\mathcal{L}(w; \{(s, c_i)\})$ according to Corollary 3.2. Furthermore, we hope that these data are able to represent a certain category of samples effectively and contain some information about their embedding space. Let Ω_i be the region in the embedding space that is discriminated as the category c_i; it is considered that the point at the center of Ω_i is more representative of a certain category of samples. For example, if the feature of a certain category obeys the Gaussian distribution that is usually assumed in machine learning theory, then the center point of the embedding space is obviously the mean of this Gaussian distribution. To effectively find the class interior points that are more representative of a certain category of sample, we define the confidence of CIP as shown in Definition 3.3.

Definition 3.3. Define $\tilde{\mathcal{R}}(s, c_i)$ as the confidence of CIP, where $\tilde{\mathcal{R}}(s, c_i) = \sup\{\mathcal{R} > 0 \mid \forall \Delta f \in \mathbb{R}^d, \|\Delta f\| < \mathcal{R}, C_i(\mathcal{F}(s) + \Delta f) > 0.5\}$.

ALGORITHM 2: UpdateCIPSet algorithm in the server

Input: The local models' weights $w_1, \ldots, w_K$, the class interior points set S, and the numbers of categories C.
Output: The updated class interior points set $\tilde{S}$.

Fetch the last C points from $S \rightarrow \{\hat{s}^1, \ldots, \hat{s}^C\}$.
for $c = 1$ to C **do**
  Sample a random noise r^c from the uniform distribution with the same shape of $\hat{s}^c$.
  Initialize a new point by $\hat{s}^c \leftarrow (\hat{s}^c + r^c)/2$.
  for $k = 1$ to K **do**
    $s_k^c \leftarrow$ optimize $\hat{s}^c$ by Equation (11).
  end for
end for
for $c = 1$ to C, $k = 1$ to K **do**
  $score_k^c = 0$.
  for $i = 1$ to K **do**
    $(prob, \hat{c}) \leftarrow$ ModelPredict(w_i,s_k^c).
    if $\hat{c}$ is c **then**
      $score_k^c \leftarrow score_k^c + prob$.
    end if
  end for
end for
$\tilde{S} \leftarrow S$.
for $c = 1$ to C **do**
  Select the top-1 point $\tilde{s}^c$ from $\{s_k^c | k = 1, \ldots, K\}$ according to their scores.
  $\tilde{S} \leftarrow \tilde{S} \cup \{\tilde{s}^c\}$.
end for
return $\tilde{S}$

Corollary 3.4. *From the proof of Corollary 3.2, we obtain that* $\tilde{\mathcal{R}}(s, c_i) = \frac{\epsilon - \mathcal{L}(w, \{(s, c_i)\})}{\|\nabla_f \mathcal{L}\|}$.

For a suitable loss function that promotes correct classification, such as cross-entropy loss, we have that $\|\nabla_f \mathcal{L}\| \propto \|C(f) - \mathbb{I}_i\| \propto \mathcal{L}(w, \{(s, c_i)\})$, where $\mathbb{I}_i$ is the one-hot encoding of label c_i. Therefore, according to Corollary 3.4, $\tilde{\mathcal{R}}(s, c_i)$ and $\mathcal{L}(w, \{(s, c_i)\})$ are negatively correlated, which guides us to minimize $\mathcal{L}(w, \{(s, c_i)\})$ to obtain a larger $\tilde{\mathcal{R}}(s, c_i)$.

Motivated by the above, for the model with weights w_k, a method to generate the relevant class interior point of category c is to minimize the loss function by:

$$s_k^c \leftarrow \arg\min_s \mathcal{L}(w_k; \{(s, c)\}). \tag{11}$$

More concretely, given a category c, we fix the model weights w_k and view the input s as the objective variable. Then, we optimize Equation (11) by an optimizer (e.g., SGD-based optimizer) with respect to the input s. After certain iterations (e.g., 100 iterations), we get a feasible input that is usually a local optimum of the loss function. According to Corollaries 3.2 and 3.4, if the loss function reaches a low value, then the relevant input is a class interior point of category c. Moreover, the lower the loss, the larger the confidence of CIP.

Applying the above method, We synthesize a sequence of local class interior points $\{s_k^1, \ldots, s_k^C\}$ for each collected model w_k, where $k \in \{d_1, \ldots, d_K\}$. Since these points contain coordinate information about the relevant class spaces in some way, their relative positions provide constraints of the decision boundaries, making the updates approximate to the gradient direction of the real data.

Corollary 3.5. *Let* $S = \{(s_i, c_i) | i = 1, 2, \ldots, C\}$ *be a great CIP set for the local model* w_k *where* $\tilde{\mathcal{R}}(s_i, c_i)$ *is as large as possible, we have* $\nabla_w^T \mathcal{L}(w; S) \nabla_w \mathcal{L}(w; D_k) \geq 0$.

A simple proof of Corollary 3.5 is that $-\nabla_w \mathcal{L}(w; D_k)$ is the direction of the negative gradient pointing to the optimal model w_k, while $-\nabla_w \mathcal{L}(w; S)$ points to the model where each s_i is successfully discriminated as c_i with large category probability. However, s_i is exactly the point that maximizes the category probability of c_i by Equation (11) related to w_k. Therefore, $-\nabla_w \mathcal{L}(w; D_k)$ and $-\nabla_w \mathcal{L}(w; S)$ point to the same optimal model w_k, which means $\nabla_w^T \mathcal{L}^T(w; S) \nabla_w \mathcal{L}(w; D_k) \geq 0$.

However, in federated learning, each client trains a biased local model w_k due to the non-IID data, causing the category embedding spaces and decision boundaries of models to not be completely overlapping. Nevertheless, The intersection of all CIP sets, related to local models, represents the generalization characteristics of multiple local datasets. Specifically, the global class interior points defined by Definition 3.6 contain information about the whole category embedding space.

Definition 3.6. Client k with data D_k produces a local model w_k. Define s as a global class interior point of category c_i, when s is also a class interior point of c_i for all local models.

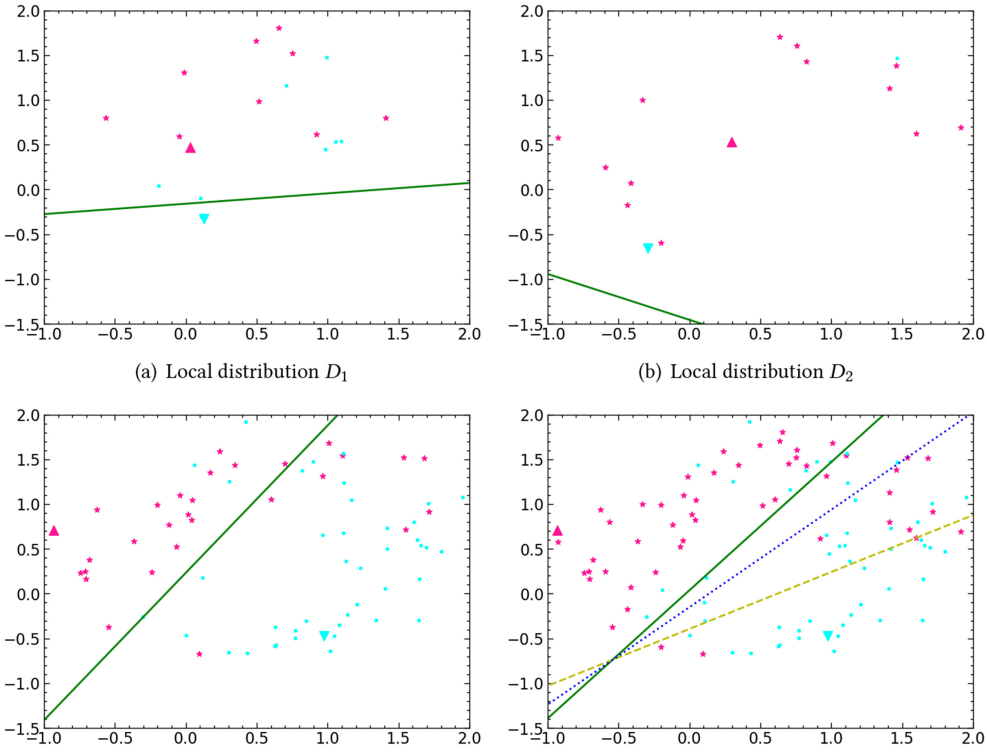
Corollary 3.7. *Let $S = \{(s_i, c_i) | i = 1, 2, \ldots, C\}$ be a global CIP set where $\tilde{\mathcal{R}}(s_i, c_i)$ is as large as possible for each w_k. Note that $\nabla_w \mathcal{L}(w; D) = \sum_k \frac{\|D_k\|}{\|D\|} \nabla_w \mathcal{L}(w; D_k)$, from Corollary 3.5, we have $\nabla_w^T \mathcal{L}(w; S) \nabla_w \mathcal{L}(w; D) \geq 0$, i.e., meeting the Equation (9).*

According to Corollary 3.7, we obtain a greater benefit from AggEnhance when these class interior points meet more properties of global class interior points. Thus, we should select the most representative points that have highest probability to be the global class interior points. According to Definition 3.3 and Corollary 3.4, we score each local point in $\{s_k^1, \ldots, s_k^C | k = d_1, \ldots, d_K\}$ based on the principle of voting. Then for each class, we only select one point with the top-1 score and add them to the global class interior points set S.

Algorithm 2 describes this method generating interior points for each class. The lines 1 to 8 apply a heuristic algorithm to search the alternative local points, where we explore the search space around the solutions synthesized in the last round. While the first line fetches a group of points from the class interior points set S, the third and fourth lines add random noises to the points and take them out of the local optimal solutions (heuristic initialization step). These steps enable the algorithm to search more points in the space. The lines 9 to 17 score each generated sample (consentaneous samples' selection step), and the remaining lines are to add the selected points into the global class interior points set.

Our proposed method would not cause "deadlock" problem. Since the CIPs depend on all local models and they are used to fine-tune the global one, the CIPs and the global model are not coupled. Besides, the random seeds of local SGDs and stochastic noise on CIPs enable it to break the potential "deadlock". When the CIPs cannot influence the model's performance, the enhanced FedAvg degenerates into the original one. After a new round of local training and global aggregation, if the global model's performance is the same as or not much different from its previous version, then the FedAvg converges to a local optimum and the global model is a fixed point of FedAvg.

There is a toy example of our method about a two-class classification problem, as shown in Figure 2, where we split a set of synthetic data into three partitions. Since the data are imbalanced, it is difficult to get a reasonable decision boundary in the local distributions D_1 and D_2. However, the decision boundaries are in different directions due to the non-IID data. These factors lead to worse performance of the aggregation result than that of training on the whole set of data. The global class interior points selected by voting provide constraints for the decision boundary to avoid incorrect directions and locations, while some unreliable local CIPs might mislead it. Therefore, in Figure 2, the predictive accuracy increases after our enhancing procedure.



(a) Local distribution D_1 (b) Local distribution D_2

(c) Local distribution D_3 (d) The whole distribution D

Fig. 2. A toy example of our AggEnhance algorithm. In these illustrations, the solid green line represents the best decision boundary of the relevant data distribution, the dashed yellow line represents the decision boundary of the averaging-based aggregation result, and the dotted blue line represents that of the enhancing result. The points with triangular shapes are the class interior points. The prediction accuracies on the whole data distribution are 77%, 71%, and 76% of the best model, the aggregation result, and the enhanced result.

3.3 Convergence Analysis

In this section, we conduct a theoretical analysis on the convergence of the algorithm and prove that it converges on strongly convex and smooth functions and non-IID data. Let w^* and w_k^* be the optimum of global model and local model, respectively. We use the term $\Gamma = \mathcal{L}^* - \sum_k p_k \mathcal{L}_k^*$ for quantifying the degree of non-IID, where $\mathcal{L}^* = \mathcal{L}(w^*; D)$, $\mathcal{L}_k^* = \mathcal{L}(w_k^*; D_k)$ and p_k is the normalized weight of the kth device usually taken as $\|D_k\|/\|D\|$. From the global optimization problem defined by federated learning, we have that

$$\Gamma = \mathcal{L}^* - \sum_k p_k \mathcal{L}_k^* = \sum_k p_k (\mathcal{L}_k(w^*) - \mathcal{L}_k(w_k^*)) \geq 0. \tag{12}$$

If the data are IID, then $w^* = w_k^*$, which means that Γ obviously goes to zero as the number of samples grows. If the data are non-IID, then $\Gamma \geq 0$, and its magnitude reflects the heterogeneity of the data distribution.

We first give some common assumptions about the function $\mathcal{L}_k$ and $\nabla \ell_k(\boldsymbol{w}, \mathcal{B}_k)$, which is the unbiased stochastic gradient of $\mathcal{L}_k$.

Assumption 1. *For all k, $\mathcal{L}_k$ has the properties of μ-strong convexity and L-smooth:*

$$\mu\text{-strongly convex:} \mathcal{L}_k(\boldsymbol{v}) \geq \mathcal{L}_k(\boldsymbol{w}) + \langle (\boldsymbol{v}-\boldsymbol{w}), \nabla \mathcal{L}_k(\boldsymbol{w}) \rangle + \frac{\mu}{2}||\boldsymbol{v}-\boldsymbol{w}||_2^2,$$

$$\beta\text{-smooth: } \mathcal{L}_k(\boldsymbol{v}) \leq \mathcal{L}_k(\boldsymbol{w}) + \langle (\boldsymbol{v}-\boldsymbol{w}), \nabla \mathcal{L}_k(\boldsymbol{w}) \rangle + \frac{L}{2}||\boldsymbol{v}-\boldsymbol{w}||_2^2.$$

Assumption 2. *Bounded variances and second moments: There exists constants $\sigma > 0$ and $G > 0$ such that*

$$\mathbb{E}_{\mathcal{B}_k \sim \mathcal{D}_k} \left\| \nabla \ell_k \left(\boldsymbol{w}; \mathcal{B}_k\right) - \nabla \mathcal{L}_k(\boldsymbol{w}) \right\|_2^2 \leq \sigma^2, \forall \boldsymbol{w}, \forall k,$$

$$\mathbb{E}_{\mathcal{B}_k \sim \mathcal{D}_k} \left[\left\| \nabla \ell_k(\boldsymbol{w}, \mathcal{B}_k) \right\|_2^2 \right] \leq G^2, \forall \boldsymbol{w}, \forall k.$$

Let $\bar{w}^t$ be the average model in the tth communication round that is used as the aggregated result in vanilla FedAvg, and w^t be the enhanced model based on the CIPs data in our AggEnhance method. Let $\Delta_{t+1}^{Avg} = \mathbb{E}\left\|\bar{w}^t - w^*\right\|^2$ denote the gap between the optimal model and the global model in the tth round for vanilla FedAvg, and $\Delta_{t+1}^{Enh} = \mathbb{E}\left\|w^t - w^*\right\|^2$ denote the gap between the optimal model and the global model in the tth round for our AggEnhance method. Theorem 3.8 demonstrates that our method can converge faster than FedAvg with the help of aggregation enhancement. The detailed proof is in Appendix A.1.

Theorem 3.8. *Let Assumptions 1 and 2 hold, and assume we can obtain a global CIP set, which means that $\cos(\bar{w}^t - w^*, \nabla_w \mathcal{L}(\bar{w}^t; S)) > 0$. Then let $B = \sum_k p_k^2 \sigma^2 + 6L\Gamma + 2I^2G^2$, where I denotes the iterations of SGD in each local update phase. If the learning rate $\eta_t < \min\{\frac{1}{\mu}, \frac{1}{4L}\}$, then we have $\exists \tau \in (0,1)$ s.t.*

$$\Delta_{t+1}^{Avg} \leq (1-\eta_t\mu)^I \Delta_t^{Avg} + \eta_t^2 IB, \tag{13}$$

$$\Delta_{t+1}^{Enh} \leq (1-\tau)(1-\eta_t\mu)^I \Delta_t^{Enh} + (1-\tau)\eta_t^2 IB. \tag{14}$$

In Theorem 3.8, the value of τ indicates that how our method AggEnhance converges faster than FedAvg, The larger the τ that can be taken, the faster our algorithm converges. Moreover, we point out in Appendix A.1 that $\tau \leq \cos^2(\bar{w}^t - w^*, \nabla_w \mathcal{L}(\bar{w}^t; S))$, which means The more efficient the global CIP set can bring the faster convergence. Further, we give the relevant analysis of the convergence of the above equation, as shown in Theorem 3.9, which shows that the algorithm converges at the rate of O(1/t) under the assumption of decreasing learning rate.

Theorem 3.9. *Assume the assumptions in Theorem 3.8 hold, and let $\gamma = \max\{2, \frac{8L}{\mu}\}$ and a diminishing learning rate $\eta_t = \frac{2}{\mu(\gamma+t)}$, then we have*

$$\Delta_{t+1}^{Enh} = \mathbb{E}\left\|w^t - w^*\right\|^2 \leq \frac{1-\tau}{\gamma+t}\left(\gamma\|w^0 - w^*\|^2 + \frac{4IB}{\mu^2}\right). \tag{15}$$

4 EXPERIMENTAL SETTINGS

4.1 Datasets

We test our method with image classification tasks on the following datasets:

— MNIST [13], which consists of 70,000 gray images about handwritten digits from "0" to "9."
— CIFAR10 [12], which contains 10 categories of images, and each class contains 6,000 32×32 color images.

— PASCAL VOC Dataset, a dataset we provide for classification task that derived from the object class recognition dataset PASCAL VOC2007 [1], where there are totally 8,243 32×32 color images for 20 categories in the training set and 5,000 images in the testing set.
— **Facial Expression Recognition Challenge Dataset (FER)** [4], where there are 28,709 48×48 color images in the training set and 7,178 testing images with seven different facial expression.

In many FL applications, only a few nodes store numerous data, and most client nodes keep a small number of samples. The sample sizes across clients subject to an approximate long-tailed distribution. We simulate this imbalanced property by a logarithmic normal distribution, i.e.,

$$\ln(|D_i| - bias) \sim N(\mu, \sigma^2), \tag{16}$$

where $N(\mu, \sigma^2)$ is a normal distribution with the mean value μ and variance σ^2, and $bias$ is an offset of this logarithmic normal distribution.

Besides, we utilize the splitting method [8] that applies a Dirichlet distribution to generate the non-IID dataset. First, we calculate each category's prior density, which is denoted as a weight vector q_i (sum to 1). Second, we sample a p_i from a Dirichlet distribution, where p_i is a proportion of all categories on the device i and γ is the distribution parameter, i.e.,

$$p_i \sim Dir(\gamma q_i). \tag{17}$$

We split the data into several subsets proportionally. Since the proportion generated randomly from the Dirichlet distribution assigns larger weights on some categories than the others, the distributions among different devices are non-IID.

In our experiments, we partition each of the above datasets into 100 non-IID and imbalanced slices to simulate 100 local clients. We set the parameter of Dirichlet distribution $\gamma = 1$ as default, and the parameters of logarithmic normal distribution as $\mu = 1$, $\sigma = 2$, and $bias = 10$.

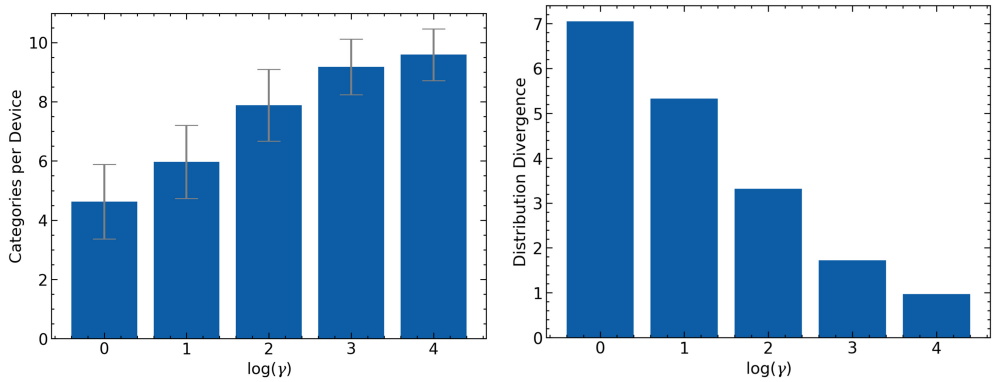
4.2 Hyper-parameter Settings

We use different depths of networks to test the performance of our method. For the MNIST dataset, we apply a multilayer perceptron to recognize the digit images. We use a simple **convolutional neural network (CNN)** on CIFAR10 with three convolutional blocks, each of which consists of a convolutional layer, a ReLU activation, and a Max Pooling layer. The sizes of the output channels are 16, 32, and 32, respectively. The same CNN framework is used for our PASCAL VOC dataset but with 16, 64, and 256 output channels. Another CNN with four convolutional blocks is applied to the FER dataset, where there is an additional normalization layer between the convolutional layer and the activation layer. Its output channels' sizes are 16, 32, 64, and 64. Since batch normalization [9] utilizes the local batches' distribution, it does poorly in distributed learning, especially in highly non-IID settings where the discrepancy among batches is large. Thus, we use GroupNorm [24] with four groups in place of BatchNorm here.

The proximal coefficient of FedProx is set as 1, 0.1, 0.1, and 0.1, successively. In the framework with trimmed-mean-based aggregation, we apply the averaging operator to the updates in $[-0.5, 0.5]$.

On each device, we train the above models for 20, 10, 10, and 10 epochs per round using SGD optimizers with the learning rates 0.01, 0.01, 0.005, and 0.001, and the batch sizes 10, 32, 32, 32.

We tune the averaging results on the global class interior points set for 10 epochs separately on the enhancing process, where batch sizes are both 32 and learning rates are the sample as the above. When generating the CIPs, each point is searching for 100 steps with a 0.1 learning rate.



(a) The number distribution of categories per device. (b) The mean divergence across all devices.

Fig. 3. The metrics to measure the non-IID degree across the local datasets. In the left figure, the mean number of categories increases as γ increases. In the right figure, the divergence of distribution displays the negative correlation with γ.

4.3 Metrics

We design two metrics to demonstrate the relationship between the Dirichlet distribution parameter γ and non-IID distribution. While the number of categories per device is small, each device holds only the data of few categories, and the combinations of categories are various, leading to the non-IID distributions across different devices. Therefore, we apply the mean of categories' number to show the non-IID degree intuitively. As Figure 3(a) shows, the mean value of categories' number is growing and the variance decreases as γ increases, i.e., the data are becoming more IID. Meanwhile, we use the mean of Kullback-Leibler divergence to quantify the non-IID degree:

$$Div = \frac{1}{N(N-1)} \sum_{i \neq j}^{N} p_i (log(p_i) - log(p_j)). \tag{18}$$

Shown in Figure 3(b), when we turn up the hyper-parameter γ, the divergence of local distributions decreases significantly, proving that the non-IID partition method works well in our experiments.

We study the convergence by the descending speed of loss curves and the ascending speed of accuracy curves qualitatively. Quantitatively, we compare the communication rounds required to achieve a certain accuracy to demonstrate the gain of our method's performance. Specifically, in MNIST dataset the target is to achieve 85% accuracy, and 50% in CIFAR10, 30% in PASCAL VOC, and 40% in FER datasets.

We estimate the information in the class interior points about the training sets in these experiments. After running the enhanced FedAvg framework, we save all generated class interior points. Then, we train the same model from scratch only on these points. If the latter model does a good decision in testing set, then we say these points contain sufficient information about the training set.

Besides, we design a series of procedures to quantify the information, as Figure 4 shows. First, we train a **Variational Auto-Encoder (VAE)** on the training set, whose encoder is applied to extract the features of data. We quantify the relativity of two sets by their cosine distance:

$$Distance(F_1, F_2) = 1 - \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{(F_1^i)^T F_2^j}{||F_1^i|| \cdot ||F_2^j||}, \tag{19}$$

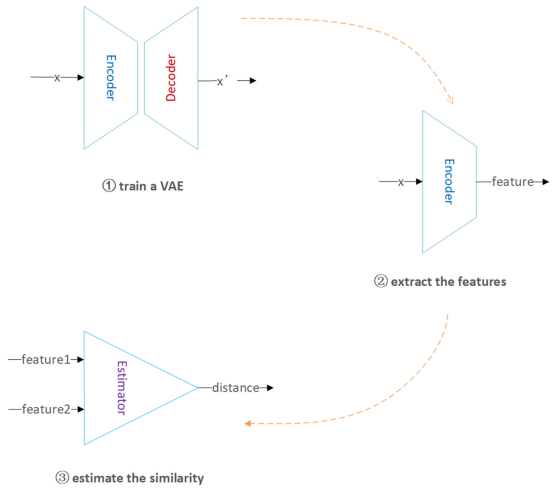


Fig. 4. The procedures to estimate the distance across two different datasets. We extract their features by the encoder from a VAE and then calculate the inter-dataset distance of these features.

where F_1, F_2 are two feature sets, consisting of N_1, N_2 feature vectors. The distance from the generated points to training data is expected to be lower than that from random noises significantly and approximate to the distance between training set and testing set.

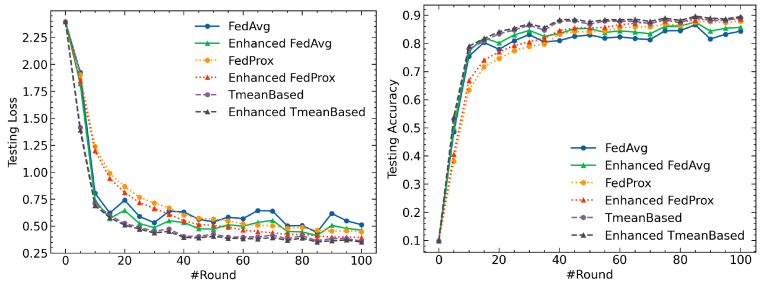
5 EXPERIMENTAL RESULTS

5.1 The Effectiveness of AggEnhance

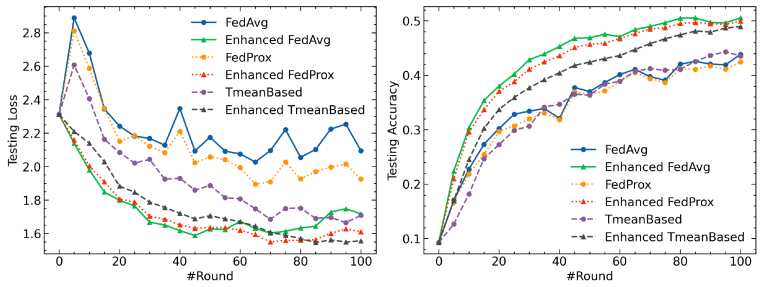
This work explores the effects of our AggEnhance by comparing the convergence of FedAvg with that of enhanced FedAvg, which are visualized by the relevant loss and accuracy curves. We also apply our method to other baseline frameworks of FL, e.g., FedProx, and trimmed-averaging-based aggregation. Figure 5 shows that the loss values and accuracy values vary as the communication rounds increase. We find that the federated learning frameworks enhanced by our method converge faster and better than the original pipelines without enhancing step, no matter how complex the datasets and the models are.

Table 1 summarizes the communication rounds required to obtain the certain accuracy mentioned in Section 4.3 for the first time. We find that the enhanced frameworks' communication costs are the same as those of the frameworks without enhancing process on MNIST, since the multilayer perceptron model and the MNIST dataset are too simple. Nevertheless, the enhanced methods still converge better than the original ones, shown in Figure 5(a). For more complex datasets and models, the enhanced FedAvg converges faster than its original version and reduces the communication rounds on average by 31.87% on CIFAR10, 43.90% on our PASCAL VOC dataset, and 18.64% on FER. Besides, the combinations with our AggEnhance also achieve better performance for other baseline frameworks than the frameworks alone in most cases, demonstrating the compatibility of our enhancing algorithm. Since the proposed method develops the module between the aggregation phase and the next training phase, it is compatible with most existing FL frameworks.

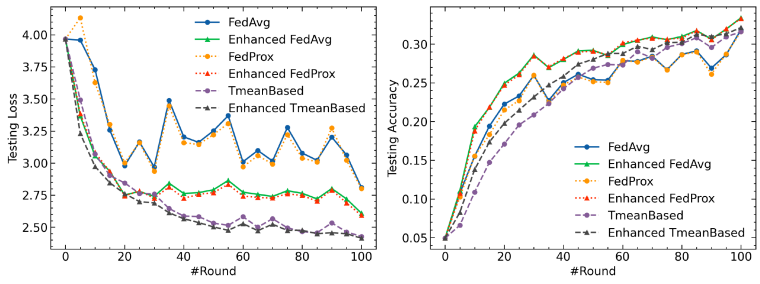
The experiments explore our method's capability to deal with non-IID data by learning from different degrees of non-IID distribution on the CIFAR10 dataset. As Figure 3 shows, the distribution of the data becomes more IID while the γ increases. We vary γ from 1 to 2, 4, 8, and 16 to instantiate different degrees of non-IID distribution. The metric to evaluate this capability is the communication rounds required to get a certain accuracy. As Figure 6 shows, as γ decreases, the



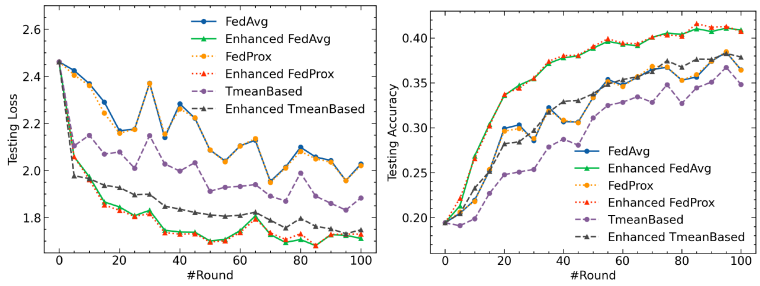
(a) MNIST



(b) CIFAR10



(c) PASCAL VOC



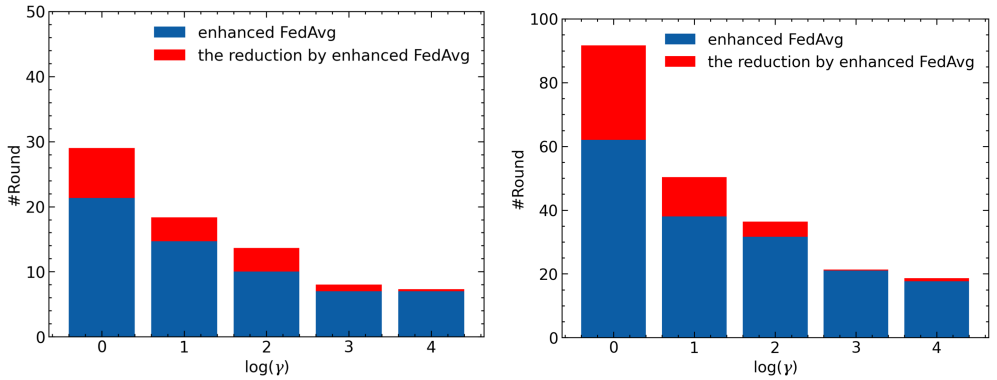
(d) FER

Fig. 5. The performance of FL with/without enhancing process on MNIST, CIFAR10, PASCAL VOC, and FER. Each point's vertical ordinate indicates the mean value of metric in the last five rounds before current round. All experiments here are repeated three times with different random seeds.

Table 1. The Communication Rounds to Achieve a Certain Accuracy of Several Federated Frameworks with/without Enhancing Process, Specifically, 85% on MNIST, 50% on CIFAR10, and 30% on PASCAL VOC and 40% on FER

	MNIST	PASCAL VOC	CIFAR10	FER
FedAvg	**11±2**	41±14	91±8	59±5
Enhanced FedAvg	**11±2**	**23±5**	**62±14**	**48±8**
FedProx	**37±5**	42±13	98±8	57±6
Enhanced FedProx	**37±5**	**29±9**	**73±17**	**46±6**
tr-mean-based	10±2	**54±8**	108±15	**102±10**
Enhanced tr-mean-based	**9±1**	57±9	**107±10**	109±15

This table shows the means and standard deviations of experiments by repeating three times. "Enhanced *" indicates a certain federated framework "*" with the enhancing method AggEnhance. The best results are marked in **bold**.



(a) The rounds to achieve 40% accuracy. (b) The rounds to achieve 50% accuracy.

Fig. 6. The communication rounds needed to obtain a certain accuracy with different non-IID levels on CIFAR10. The heights of blue bins indicate the FedAvg pipeline with the enhancing process, while the red parts indicate the reductions by our aggregation enhancement method. The more non-IID the data are, the more communication federated learning needs and the more rounds our enhancing method reduces.

reduction of communication costs by the enhanced FedAvg is more noticeable. Even if the data distribution is approximate to IID, the proposed method still maintains its advantages. To sum up, our enhancing method is flexible to handle various non-IID scenarios.

This work also tests the algorithms on CIFAR10 with the simple CNN mentioned in Section 4.2, ResNet18 [7], and VGG16 [22]. Demonstrated as Figure 7, our AggEnhance method achieves less communication than the original FedAvg, no matter what the model structure is, which means that AggEnhance has wide applicability to different neural networks.

5.2 The Validity of CIPs

We verify the validity of the class interior points, taking the MNIST dataset as an example. Shown in Figure 8, the model only trained on the class interior points from scratch achieves up to 78.2% accuracy on the testing set. Besides, the distance from class interior points' feature set to the one of training set is closer than that from random noises. As the training round increases, the distance is closer to that from testing set, which implies that the class interior points are storing more and more meaningful class-specific information about data distribution.

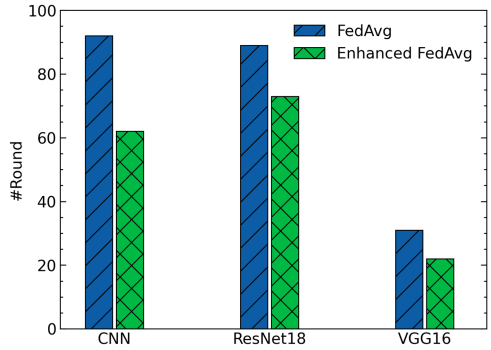
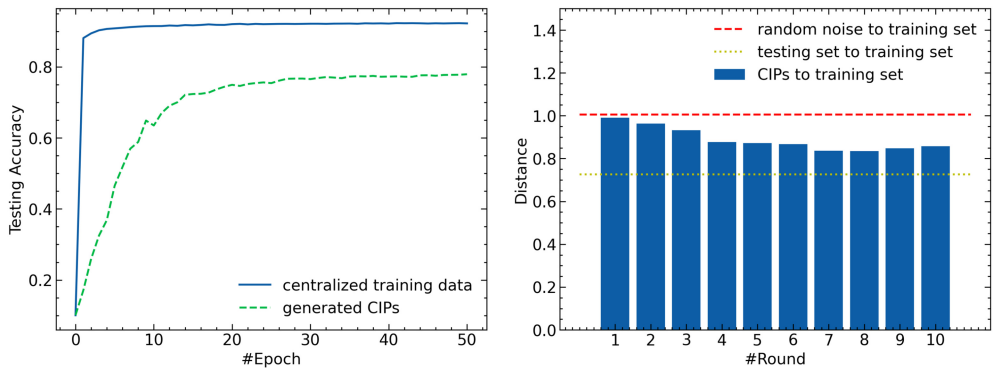


Fig. 7. The communication rounds needed to achieve 50% accuracy on CIFAR10 by different deep models.



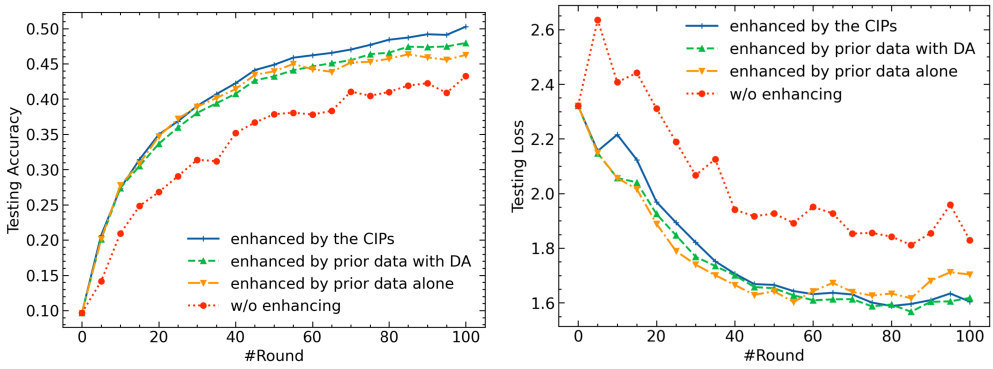
(a) MNIST's training set v.s. class interior points. (b) The similarity of features.

Fig. 8. The generated class interior points on MNIST contain valid information about the training set.

Particularly, the real data are a special type of class interior points. If there are real samples on the server, then we can enhance the aggregated result by tuning it on these real data rather than the generated ones, as shown in Figure 9. We fetch one real sample per class and put them into a prior dataset in advance and compare the effect of our generated points with that of the real data in FedAvg framework. The real data may contain more information about the whole distribution than the generated points if the prior real dataset is enough large. However, while the prior dataset is small, our generated class interior points achieve better performance than the prior real data. Besides, it is insecure for transmitting real data to the server and saving on it for the privacy issue. When the server stores real data, it is easy for the server to infer the related data by attacking the local models. Therefore, to generate class interior points is a more safe alternative.

5.3 Discussions about the Privacy Security

In our AggEnhance algorithm, we generate the feasible class interior points of the local optimal solutions instead of the real data. Therefore, it is hard to guess the real data due to the lack of information. In Figure 10, the generated data might follow certain patterns but is entirely different from the real one. We cannot infer the training data from the set of class interior points. Since the major difference between the enhanced framework and the non-enhanced one is that the former inserts an enhancing procedure with our AggEnhance algorithm after the aggregation, they still have the same privacy security level.



(a) Loss curves. (b) Accuracy curves.

Fig. 9. The loss and accuracy curves of different types of class interior points on CIFAR10 dataset. The curve named "enhanced by prior data with DA" indicates the FedAvg enhanced by the real data with data augmentation techniques, while the "enhanced by prior data alone" indicates that without data augmentation.

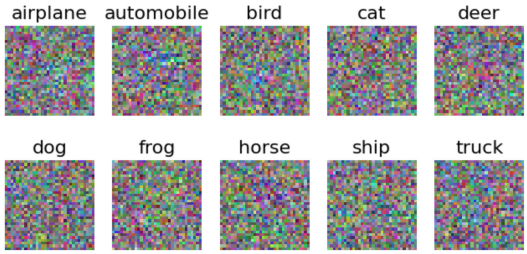


Fig. 10. Instances of the class interior points about CIFAR10. We cannot recognize the generated data by the naked eye without any prior knowledge.

Besides, it is easy to apply other privacy-preserving techniques, e.g., **differential privacy (DP)**, to our enhanced frameworks. This work explores the influence of DP on our AggEnhance method. The experiment applies Laplace mechanism to achieve differential privacy, which adds Laplace noises to the model weights. Equation (20) is the probability density function of Laplace Distribution, where μ is the location of the peak value and $b > 0$ is the scale of the distribution. Zhu [28] demonstrated that DP prevents the **deep leakage from gradient (DLG)** well, while the scale of Laplacian noise is higher than 10^{-2}. As Figure 11, though the performance of FL degrades after applying DP, the AggEnhance method still achieves better convergence than the original FedAvg. Therefore, our method still takes effect after using differential privacy to protect the model weights.

$$Pr(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right) \tag{20}$$

It is also applicable for secure aggregation-based FL, where the local models are not accessible to the server. It only requires the CIP data to be generated on local devices. The global CIP data are reachable by broadcasting and scoring all the local ones via the server (with encryption if necessary), as Figure 12 shows. Finally, fine-tune the results of secure aggregation on the global CIP data to attain a better result.

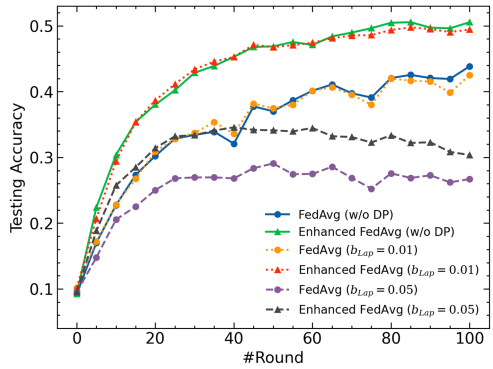


Fig. 11. The results after applying differential privacy on CIFAR10 dataset.

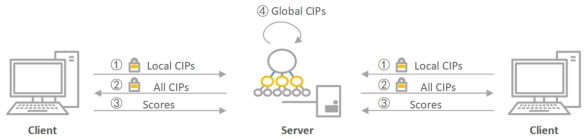


Fig. 12. The process to generate global CIP data in secure aggregation-based FL.

5.4 Discussions about the Computation

At the same time, the additional computation of our method is done only on the global server. We spend extra time inferring the class interior points from models, and it costs extra space to store them. In normal scenarios, the extra time is far less than the local machines' response time due to the global server's enormous computing resources. Besides, we can store the class interior points in a length-limited queue to save the space resources. Overall, the extra computation costs are negligible.

6 CONCLUSIONS

In this work, we provide a novel idea to accelerate the convergence of federated learning frameworks with non-IID data. Specifically, we design an independent module for aggregation enhancement in existing federated learning pipelines. We find that a type of data, defined as the class interior points, is helpful to refine decision boundaries. This work provides an alternative method to synthesize these data from the collected models. Since our method requires no additional data transmitted from the clients, it still has the security level of existing federated learning frameworks. Our results provide empirical evidence to show that the enhanced frameworks outperform the frameworks without the enhancing process. Our method still performs well even though differential privacy is applied to prevent the deep leakage from gradient. It is easy to apply the proposed method to the prevalent frameworks, including the secure aggregation-based FL.

In our future research, we will explore the advantage of the class interior points in the local training step, which might help generalize the local models. Besides, we will explore the applications of this method to other types of data, e.g., text data. We will also try other methods to generate the class interior points that maintain knowledge of the population distribution.

A APPENDIX

A.1 Proof of Theorem 3.8

To prove the Theorem 3.8, we give some additional notations and lemmas as follows: In the local update phase between the two server aggregation phases, let $w_{k,i}, i = 1, 2, \ldots, I$ be the model weights maintained in the kth device at the ith step, where I denotes the iterations of SGD for each client in each local update phase. Note that all local models have the same initialization $w_{k,0} = w_0$.

We assume that the learning rate η is constant throughout the local update phase of a certain round, then we have $w_{k,i+1} = w_{k,i} - \eta \nabla_w \ell_k(w_{k,i}, \mathcal{B}_{k,i})$. We define that $\bar{w}_i = \sum_k p_k w_{k,i}$, $g_i = \sum_k p_k \nabla_w \ell_k(w_{k,i}, \mathcal{B}_{k,i})$ and $\bar{g}_i = \sum_k p_k \nabla_w \mathcal{L}_k(w_{k,i})$. Therefore, $\bar{w}_{i+1} = \bar{w}_i - \eta g_i$ and $\mathbb{E} g_i = \bar{g}_i$.

According to the proof of Reference [17], there is a lemma about the results of one step SGD as follows:

Lemma A.1. *Assume Assumption 1 holds. If $\eta \le \frac{1}{4L}$, then we have*

$$\mathbb{E}\left\|\bar{w}_{i+1} - w^*\right\|^2 \le (1 - \eta\mu)\,\mathbb{E}\left\|\bar{w}_i - w^*\right\|^2 + \eta^2 \mathbb{E}\left\|g_i - \bar{g}_i\right\|^2 + 6L\eta^2\Gamma + 2\mathbb{E}\sum_k p_k \left\|\bar{w}_i - w_{k,i}\right\|^2, \quad (21)$$

where $\Gamma = \mathcal{L}^* - \sum_k p_k \mathcal{L}_k^* \ge 0$.

Moreover, assume Assumption 2 holds, we have

$$\begin{aligned}
\mathbb{E}\left\|g_i - \bar{g}_i\right\|^2 &= \mathbb{E}\left\|\sum_k p_k \left(\nabla \ell_k\left(w_{k,i}, \mathcal{B}_{k,i}\right) - \nabla \mathcal{L}_k\left(w_{k,i}\right)\right)\right\|^2 \\
&= \sum_k p_k^2 \mathbb{E}\left\|\nabla \ell_k\left(w_{k,i}, \mathcal{B}_{k,i}\right) - \nabla \mathcal{L}_k\left(w_{k,i}\right)\right\|^2 \\
&\le \sum_k p_k^2 \sigma^2,
\end{aligned} \quad (22)$$

$$\begin{aligned}
\mathbb{E}\sum_k p_k \left\|\bar{w}_i - w_{k,i}\right\|^2 &= \mathbb{E}\sum_k p_k \left\|\left(w_{k,i} - \bar{w}_0\right) - \left(\bar{w}_i - \bar{w}_0\right)\right\|^2 \\
&\overset{(a)}{\le} \mathbb{E}\sum_k p_k \left\|w_{k,i} - \bar{w}_0\right\|^2 \\
&= \mathbb{E}\sum_k p_k \left\|\sum_{\tau=0}^{i} \eta \nabla_w \ell_k(w_{k,i}, \mathcal{B}_{k,i})\right\|^2 \\
&\overset{(b)}{\le} \mathbb{E}\sum_k p_k i \sum_{\tau=0}^{i} \left\|\eta \nabla \ell_k(w_{k,i}, \mathcal{B}_{k,i})\right\|^2 \\
&\le \sum_k p_k i^2 \eta^2 G^2 \\
&\le \eta^2 I^2 G^2,
\end{aligned} \quad (23)$$

where (a) follows $\mathbb{E}\|X - \mathbb{E}X\|^2 \le \mathbb{E}\|X\|^2$ and (b) follows the Jensen inequality. Hence, from Equations (21), (22), and (23), it follows that

$$\mathbb{E}\left\|\bar{w}_{i+1} - w^*\right\|^2 \le (1 - \eta\mu)\,\mathbb{E}\left\|\bar{w}_i - w^*\right\|^2 + \eta^2 \sum_k p_k^2 \sigma^2 + 6L\eta^2\Gamma + 2\eta^2 I^2 G^2. \quad (24)$$

Let $B = \sum_k p_k^2 \sigma^2 + 6L\Gamma + 2I^2G^2$, then we have

$$\begin{aligned}\mathbb{E}\left\|\bar{w}_I - w^*\right\|^2 &\leq (1-\eta\mu)\,\mathbb{E}\left\|\bar{w}_{I-1} - w^*\right\|^2 + \eta^2 B \\ &\overset{(recurrence)}{\leq} (1-\eta\mu)^I\,\mathbb{E}\left\|\bar{w}_0 - w^*\right\|^2 + \frac{1-(1-\eta\mu)^I}{1-(1-\eta\mu)}\eta^2 B \\ &\leq (1-\eta\mu)^I\,\mathbb{E}\left\|\bar{w}_0 - w^*\right\|^2 + \eta^2 IB.\end{aligned} \tag{25}$$

Note that for the local update phase before the tth aggregation round in vanilla FedAvg algorithm, we have $\bar{w}_0 = \bar{w}^{t-1}$ and $\bar{w}_I = \bar{w}^t$, which means Equation (13) holds.

Then, according to the Algorithm 1, in our AggEnhance method, we have $\bar{w}_0 = w^{t-1}$,$\bar{w}_I = \bar{w}^t$ and $w^t = \bar{w}^t - \eta\nabla_w\mathcal{L}(\bar{w}^t; S)$. Assume $\cos(\bar{w}^t - w^*, \nabla_w\mathcal{L}(\bar{w}^t; S)) = \epsilon > 0$, we have

$$\langle(\bar{w}^t - w^*), \nabla_w\mathcal{L}(\bar{w}^t; S)\rangle = \epsilon\left\|\bar{w}^t - w^*\right\|\left\|\nabla_w\mathcal{L}(\bar{w}^t; S)\right\|. \tag{26}$$

Then, we have

$$\begin{aligned}\|w^t - w^*\|^2 &= \|\bar{w}^t - w^* - \eta_s\nabla_w\mathcal{L}(\bar{w}^t; S)\|^2 \\ &= \|\bar{w}^t - w^*\|^2 - 2\eta_s\langle(\bar{w}^t - w^*), \nabla_w\mathcal{L}(\bar{w}^t; S)\rangle + \eta_s^2\|\nabla_w\mathcal{L}(\bar{w}^t; S)\|^2 \\ &= \|\bar{w}^t - w^*\|^2 - 2\eta_s\epsilon\left\|\bar{w}^t - w^*\right\|\left\|\nabla_w\mathcal{L}(\bar{w}^t; S)\right\| + \eta_s^2\|\nabla_w\mathcal{L}(\bar{w}^t; S)\|^2 \\ &= (1-\epsilon^2)\|\bar{w}^t - w^*\|^2 + (\epsilon\|\bar{w}^t - w^*\| - \eta_s\|\nabla_w\mathcal{L}(\bar{w}^t; S)\|)^2.\end{aligned} \tag{27}$$

Hence, when we choose an enough small η_s satisfies $0 < \eta_s < \frac{2\epsilon\|\bar{w}^t - w^*\|}{\|\nabla_w\mathcal{L}(\bar{w}^t; S)\|}$, might as well set $\frac{\eta_s\|\nabla_w\mathcal{L}(\bar{w}^t; S)\|}{\|\bar{w}^t - w^*\|} = k$, it follows that

$$\begin{aligned}\|w^t - w^*\|^2 &= (1-\epsilon^2)\|\bar{w}^t - w^*\|^2 + (\epsilon - k)^2\|\bar{w}^t - w^*\|^2 \\ &= (1 - k(2\epsilon - k))\|\bar{w}^t - w^*\|^2.\end{aligned} \tag{28}$$

Then, exists $\tau = k(2\epsilon - k) \in (0, 1)$, because $0 < k < 2\epsilon$ and $k(2\epsilon - k) \leq \epsilon^2 \leq 1$. Therefore, from Equation (25), we can derive that

$$\begin{aligned}\mathbb{E}\|w^t - w^*\|^2 &= (1-\tau)\mathbb{E}\|\bar{w}^t - w^*\|^2 \\ &\leq (1-\tau)\,(1-\eta\mu)^I\,\mathbb{E}\left\|w^{t-1} - w^*\right\|^2 + (1-\tau)\eta^2 IB.\end{aligned} \tag{29}$$

□

A.2 Proof of Theorem 3.9

We use the same notations as the proof of Theorem 3.8 in Appendix A.1. For a diminishing learning rate, $\eta_t = \frac{\beta}{\gamma + t}$ for some $\beta > \frac{1}{\mu}$ and $\gamma > 0$ such that $\eta_t < \min\{\frac{1}{\mu}, \frac{1}{4L}\}$, we can derive that $\gamma > \max\{\mu\beta, 4L\beta\}$ from above. Let $\upsilon = \max\{\gamma\Delta_0, \frac{\beta^2 IB}{\mu\beta - 1}\}$ where $\Delta_0 = \|w^0 - w^*\|^2$, we will prove that $\Delta_t^{Avg} \leq \frac{\upsilon}{\gamma + t}$ by induction first.

First, the definition of v ensures that it holds for $t = 0$. Assume the conclusion holds for some t, from Theorem 3.8, it follows that

$$\begin{aligned}
\Delta_{t+1}^{Avg} &\leq (1-\eta_t\mu)^I \Delta_t^{Avg} + \eta_t^2 IB \\
&\leq \left(1-\frac{\mu\beta}{\gamma+t}\right)^I \frac{v}{\gamma+t} + \frac{\beta^2 IB}{(\gamma+t)^2} \\
&= \frac{(\gamma+t-\mu\beta)^I + (\mu\beta-1)(\gamma+t)^{I-1}}{(\gamma+t)^{I+1}} v + \frac{\beta^2 IB}{(\gamma+t)^2} - \frac{(\mu\beta-1)}{(\gamma+t)^2} v \\
&\leq \frac{(\gamma+t-\mu\beta)(\gamma+t)^{I-1} + (\mu\beta-1)(\gamma+t)^{I-1}}{(\gamma+t)^{I+1}} v + 0 \\
&\leq \frac{\gamma+t-1}{(\gamma+t)^2-1} v \\
&= \frac{v}{\gamma+t+1}.
\end{aligned} \tag{30}$$

Specifically, when $\beta = \frac{2}{\mu}$ and $\gamma = \max\{\mu\beta, 4L\beta\} = \max\{2, \frac{8L}{\mu}\}$, we have

$$v = \max\left\{\gamma\Delta_0, \frac{\beta^2 IB}{\mu\beta-1}\right\} \leq \gamma\Delta_0 + \frac{\beta^2 IB}{\mu\beta-1} = \gamma\Delta_0 + \frac{4IB}{\mu^2} \tag{31}$$

$$\Delta_t^{Avg} = \mathbb{E}\left\|\bar{w}^t - w^*\right\|^2 \leq \frac{v}{\gamma+t} \leq \frac{1}{\gamma+t}\left(\gamma\Delta_0 + \frac{4IB}{\mu^2}\right) = \frac{1}{\gamma+t}\left(\gamma\|w^0 - w^*\|^2 + \frac{4IB}{\mu^2}\right). \tag{32}$$

Therefore, from Equations (28) and (32), we can derive that

$$\Delta_t^{Enh} = \mathbb{E}\|w^t - w^*\|^2 \leq \frac{1-\tau}{\gamma+t}\left(\gamma\|w^0 - w^*\|^2 + \frac{4IB}{\mu^2}\right). \tag{33}$$

□

A.3 Experimental Results with Different Hyper-parameters

This work trains the models on CIFAR10, PASCAL VOC, and FER datasets by FL with different settings of hyper-parameters. We set the hyper-parameters mentioned in Section 4.2 as the default settings and adjust a certain hyper-parameter while other hyper-parameters remain unchanged. Shown as Figure 13, in all experiments of different hyper-parameter combinations, the FedAvg method enhanced by AggEnhance outperforms the original one without aggregation enhancement.

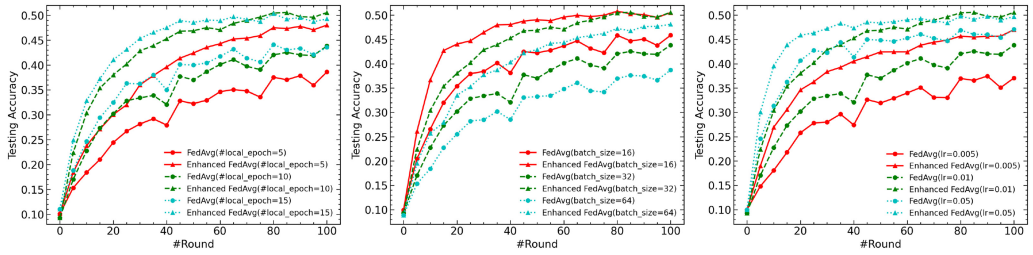
A.4 Instances of CIPs in Other Settings

Figure 10 shows the CIPs generated in the settings that the number of local training epochs is 10. Here show the CIP instances in other settings. Specifically, the CIPs in Figure 14 are generated when the model is trained for only 1 epoch on local devices per round. We cannot still recognize the CIPs by the naked eye even if the local epoch is less.

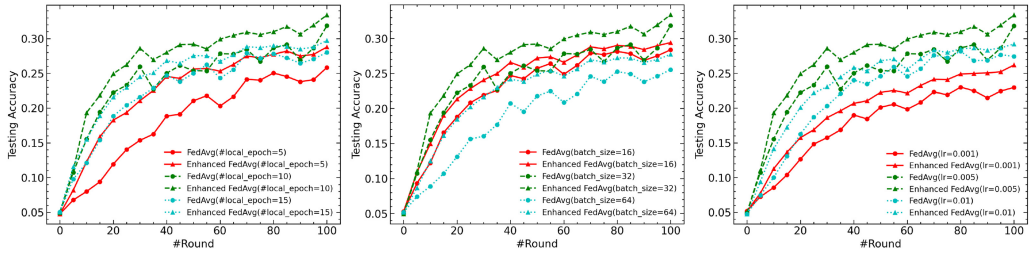
While adding a 0.05 scale of Laplace noise into the gradients, the generated CIPs are demonstrated in Figure 15. Compared with those generated in no-DP settings, more pixels of CIPs reach their extremum values. It is still impossible to distinguish the CIPs after applying differential privacy.

A.5 The Number of CIPs for Aggregation Enhancement

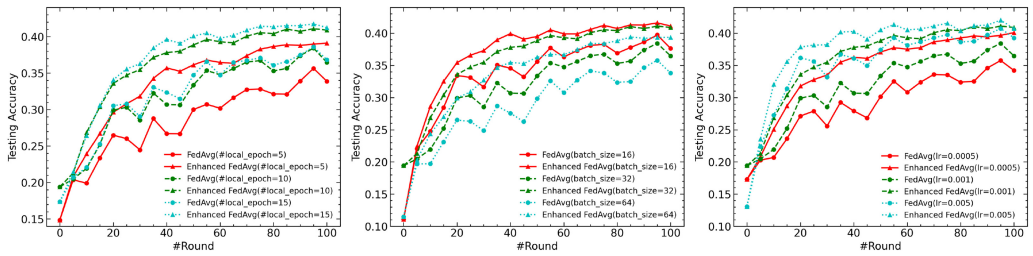
The 21st line of Algorithm 2 applies all CIPs generated in history instead of the new ones to fine-tune the global model, which reduces the biases of the last CIPs and makes the results smoother.



(a) CIFAR10



(b) PASCAL VOC



(c) FER

Fig. 13. The accuracy curves of experiments with different training hyper-parameters.

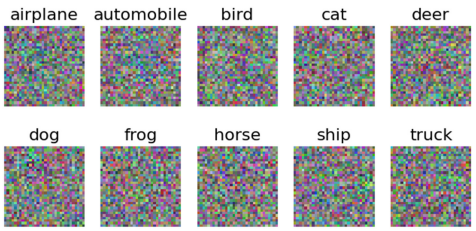


Fig. 14. Instances of the class interior points about CIFAR10 while *#local_epoch* = 1.

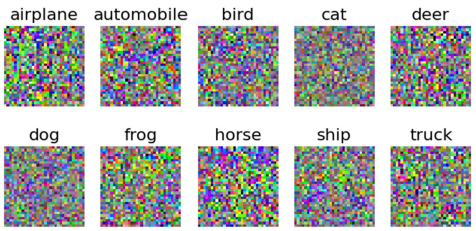


Fig. 15. Instances of the class interior points about CIFAR10 while applying DP.

Figure 16 shows the experimental result using different sizes of CIPs for aggregation enhancement, where "0 CIPs" indicates the original FedAvg without aggregation enhancement, "The Last 10 CIPs" indicates using the new 10 CIPs to enhance the results, and so on. "All CIPs(Weighted)" means using all points in CIP set for AggEnhance with decayed weights, whose decay rate is

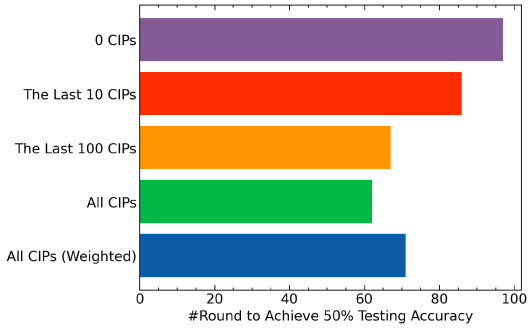


Fig. 16. The rounds required to achieve 50% accuracy on CIFAR10 in different CIPs settings.

0.01 by time. In Figure 16, the larger the CIPs queue, the less the communication rounds required to achieve the same accuracy. Compared with unweighted settings, the weighted CIPs perform poorly, since the model overfits the newest points.

REFERENCES

[1] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2010. The PASCAL visual object classes (VOC) challenge. *Int. J. Comput. Vis.* 88, 2 (2010), 303–338.

[2] Avishek Ghosh, Justin Hong, Dong Yin, and Kannan Ramchandran. 2019. Robust federated learning in a heterogeneous environment. *arXiv preprint arXiv:1906.06629* (2019).

[3] Jack Goetz, Kshitiz Malik, Duc Bui, Seungwhan Moon, Honglei Liu, and Anuj Kumar. 2019. Active federated learning. *arXiv preprint arXiv:1909.12641* (2019).

[4] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*. Springer, 117–124.

[5] Neel Guha, Ameet Talwalkar, and Virginia Smith. 2019. One-shot federated learning. *arXiv preprint arXiv:1902.11175* (2019).

[6] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. 2018. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604* (2018).

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

[8] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. 2019. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335* (2019).

[9] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*. PMLR, 448–456.

[10] Georgios A. Kaissis, Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* 2, 6 (2020), 305–311.

[11] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. 2020. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning*. PMLR, 5381–5393.

[12] Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. Ph. D. Dissertation. University of Toronto.

[13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[14] David Leroy, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau. 2019. Federated learning for keyword spotting. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6341–6345.

[15] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Proc. Mach. Learn. Syst.* 2 (2020), 429–450.

[16] Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M. Jorge Cardoso, et al. 2019. Privacy-preserving federated brain tumour segmentation. In *International Workshop on Machine Learning in Medical Imaging*. Springer, 133–141.

[17] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. 2020. On the convergence of FedAvg on non-IID data. In *International Conference on Learning Representations*.

[18] Paul Pu Liang, Terrance Liu, Liu Ziyin, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523* (2020).

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

[20] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, et al. 2020. The future of digital health with federated learning. *NPJ Digit. Med.* 3, 1 (2020), 1–7.

[21] Micah J. Sheller, G. Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. 2018. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *International MICCAI Brainlesion Workshop*. Springer, 92–104.

[22] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[23] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. 2020. Federated learning with matched averaging. In *International Conference on Learning Representations*.

[24] Yuxin Wu and Kaiming He. 2018. Group normalization. In *European Conference on Computer Vision (ECCV)*. 3–19.

[25] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*. PMLR, 5650–5659.

[26] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. 2018. Federated learning with non-IID data. *arXiv preprint arXiv:1806.00582* (2018).

[27] Yanlin Zhou, George Pu, Xiyao Ma, Xiaolin Li, and Dapeng Wu. 2020. Distilled one-shot federated learning. *arXiv preprint arXiv:2009.07999* (2020).

[28] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, Vol. 32.

Received July 2021; revised May 2022; accepted June 2022