Proceedings of the 4th International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/2026.33488

Adaptive Gradient-Aligned Aggregation for Federated Learning under Non-IID Data

Yichen Liu

School of Intelligent Systems Science and Engineering/JNU-Industry School of Artificial Intelligence, Jinan University, Guangzhou, China
lyc3407518322@stu2023.jnu.edu.cn

Abstract. Federated learning (FL) is a distributed training paradigm with privacy-preserving capabilities and has attracted considerable attention in recent years. But the heterogeneity of client statistics still limits its performance. In a non-IID environment, inconsistent local optimization directions can lead to client drift, unstable convergence, and a decrease in the accuracy of the global model. So this paper proposes a multi-factor adaptive aggregation strategy for non-IID scenarios, which comprehensively evaluates clients from three aspects: data volume, consistency of gradient directions, and local training quality, and allocates adaptive aggregation weights. A gradient direction filtering mechanism is also introduced to alleviate the impact of conflicting local updates before global aggregation. Experiments are set based on the MNIST and CIFAR-10 datasets to construct two scenarios, compared with FedAvg, FedProx, and SCAFFOLD as baselines. The results show the method of this paper has stable performance under IID conditions and better performance under non-IID conditions. Subsequent ablation experiments further validate the effectiveness of gradient consistency modeling, training quality weighting, and the update filtering mechanism.The main contribution of this work is to improve convergence stability and enhance the robustness of federated optimization in heterogeneous environments.

Keywords: Federated learning, Non-IID data, Adaptive aggregation, Gradient alignment

1. Introduction

FL is an efficient distributed learning method. It allows multiple clients to train a global model jointly without sharing their raw local data [1,2]. Among existing FL methods, Federated Averaging (FedAvg) is widely used because it is practical and communication-efficient [1].But the real performance of FL is often limited by statistical heterogeneity among clients. In real settings, client data usually has clear Non-IID characteristics. This can lead to large differences in local optimization directions, which may further cause client drift, slower model convergence, and lower model performance [3-5]. Standard FedAvg mainly uses data size to assign aggregation weights, but it does not fully consider the optimization quality of client updates or their direction contribution. As a result, when low-quality updates or even updates that conflict with the global optimization direction are included in aggregation, training instability in heterogeneous data settings can become more serious [5-7].To solve these problems, SCAFFOLD introduces control variates to address

DOI: 10.54254/2755-2721/2026.33488

client drift [5]. FedProx improves training robustness by adding a proximal regularization term to the local objective function [6]. FedNova reduces the objective inconsistency caused by heterogeneous local updates through normalized aggregation [7]. FedBN keeps local batch normalization statistics [8]. FedDyn aligns local and global objectives through dynamic regularization [9]. Adaptive federated optimization methods improve convergence in heterogeneous non-convex settings by using adaptive optimization on the server side [10]. Although these methods have made good progress, most of them focus on only one correction mechanism, such as regularization, normalization, or server-side adaptive updates. Lightweight aggregation strategies that consider data size, update direction consistency, and local training quality at the same time are still lacking.To fill this gap, this paper proposes an adaptive weighted aggregation method based on gradient alignment. The method evaluates client contribution from three aspects: data size, gradient direction consistency, and local training quality. Adaptive aggregation weights are assigned to different clients based on this.

This paper also designs a gradient-direction-based filtering mechanism to suppress local updates that go against the global optimization trend, so as to reduce client drift in Non-IID settings. In this paper, a multi-factor adaptive aggregation strategy is shown by combining gradient alignment, data size, and local training quality, a gradient-direction-based filtering mechanism is designed to reduce the negative effect of harmful client updates on global training, a lightweight and easy-to-implement federated aggregation framework is built to improve convergence stability and final model performance under heterogeneous data distributions.

2. Method

2.1. Dataset preparation

To evaluate aggregation strategies in heterogeneous federated settings, two benchmark image classification datasets were used. Representative samples are shown in Figure 1, where MNIST contains grayscale handwritten digits and CIFAR-10 includes more complex RGB natural images shown in Figure 2.

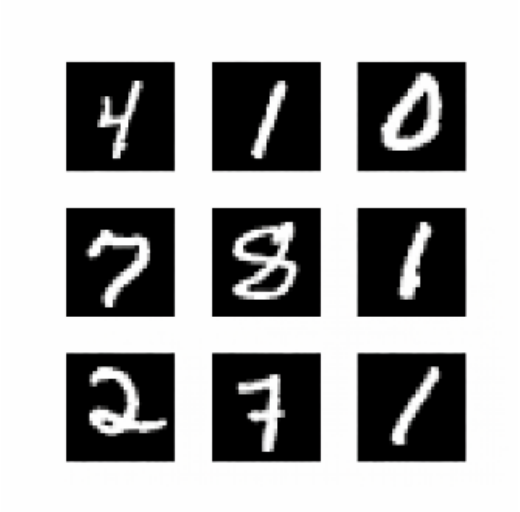

Figure 1. Representative samples from MNIST (picture credit: original)

DOI: 10.54254/2755-2721/2026.33488

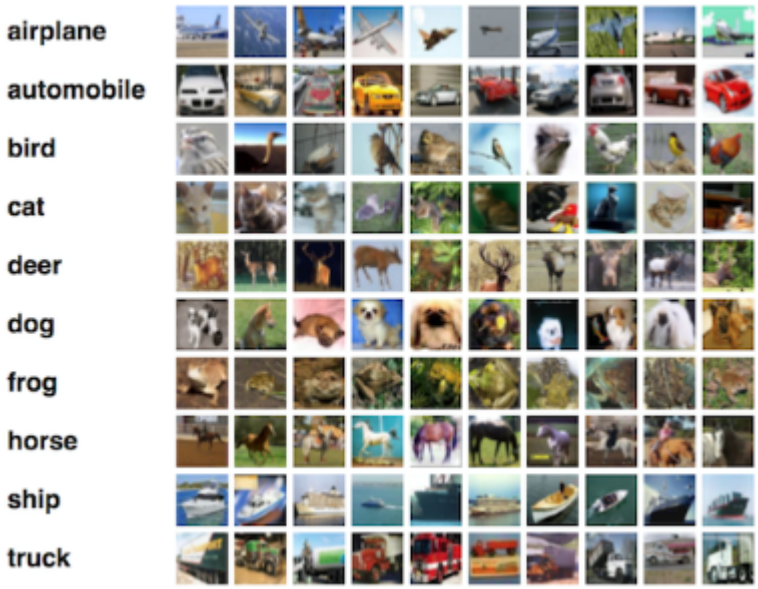

Figure 2. Representative samples from CIFAR-10 (picture credit: original)

Before federated partitioning, the raw data needs to undergo standard preprocessing to improve training stability and data consistency. For MNIST, the pixel values are converted to tensors and normalized. For CIFAR-10, the three RGB channels are normalized separately. Considering that CIFAR-10 images have more complex content, data augmentation operations are further applied during the local training stage, including random horizontal flipping and random cropping with padding, to increase the model's generalization ability and alleviate local small-sample overfitting issues. Since the MNIST data structure is relatively simple, only normalization is applied.

Both datasets retain the official train/test set splits. The global test set is kept on the server side and is used to evaluate the performance of the model only. The original training set is divided among multiple clients to construct a FL environment. To simulate realistic statistical heterogeneity, the training set is partitioned using a Dirichlet-distribution-based non-IID method, so that different clients have different class proportion distributions. This setup is consistent with the issue raised in the introduction, namely, that statistical heterogeneity causes client drift and inconsistent local optimization directions [3,4]. This design allows for a comparison of the performance differences between traditional aggregation methods and the proposed adaptive aggregation method in heterogeneous scenarios.

2.2. Federated learning

The global objective can be formulated as

$$\min_{w} F\left(w\right) = \sum_{k=1}^{K} \frac{n_k}{n} F_k\left(w\right) \tag{1}$$

where K denotes the total number of clients, n_k is the number of samples owned by client k, and $F_k(w)$ denotes the local empirical loss on client k. In standard FedAvg, the server aggregates local models according to the proportion of client data volume [1]. This mechanism is simple and effective in IID settings, but its performance often declines in non-IID environments because different clients may follow quite different optimization directions [3-7]. To better examine the

DOI: 10.54254/2755-2721/2026.33488

aggregation strategy, this paper uses a lightweight convolutional neural network as the base classifier.

2.3. Multi-factor adaptive aggregation strategy

The proposed method aims to address the limitation of conventional data-size-based aggregation by jointly considering data volume, gradient direction consistency, and local training quality. A client update should receive a high aggregation weight only when it is statistically meaningful, directionally aligned with the overall optimization trend, and associated with satisfactory local training behavior.

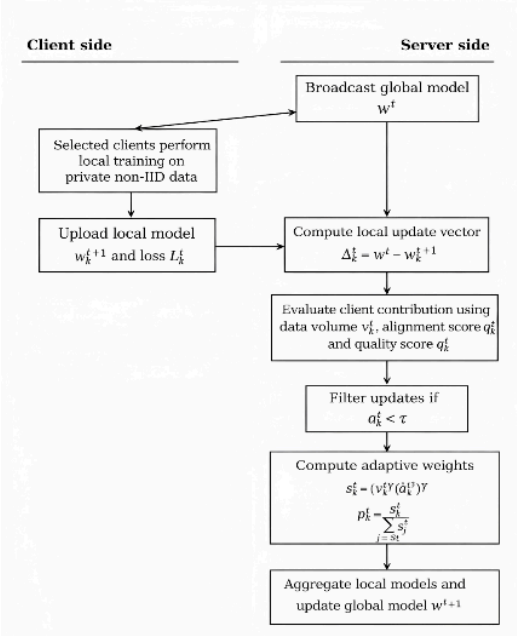

Figure 3. Workflow of the proposed multi-factor adaptive aggregation strategy (picture credit: original)

The proposed method first sends the global model to selected clients, collects local models and local losses after local training, evaluates client contributions from data volume, update alignment, and local training quality, filters conflicting updates, and then performs adaptive weighted aggregation to update the global model, shown in Figure 3.

The local update vector is defined as

$$\Delta_k^t = w^t - w_k^{t+1} \tag{2}$$

To characterize the overall optimization trend of the current round, an average update direction is computed as

$$\bar{\Delta}^t = \frac{1}{|S_t|} \sum_{k \in S_t} \Delta_k^t \tag{3}$$

S_t denotes the selected clients in round t. The directional consistency of client k is then measured by cosine similarity:

DOI: 10.54254/2755-2721/2026.33488

$$a_k^t = \frac{\Delta_k^t \cdot \bar{\Delta}^t}{\|\Delta_k^t\| \|\bar{\Delta}^t\|} \tag{4}$$

A larger value of a_k^t indicates that the client update is more consistent with the global optimization tendency. If the similarity falls below a predefined threshold τ, the corresponding update is regarded as harmful or weakly relevant. Such updates are filtered out or significantly down-weighted before aggregation, which helps suppress client drift in heterogeneous settings.In addition to directional consistency, client data volume is incorporated as

$$v_k^t = \frac{n_k}{\sum_{j \in S_t} n_j} \tag{5}$$

which preserves the statistical contribution of clients with more training samples. Local training quality is represented by the inverse of the final local loss:

$$q_k^t = \frac{1}{L_k^t + \varepsilon} \tag{6}$$

where L_k^t is the final local loss of client k, and ε is the small constant of numerical stability. After normalization, the three factors are fused into a composite score:

$$s_k^t = \left(v_k^t\right)^{\alpha} \left(\hat{a}_k^t\right)^{\beta} \left(\hat{q}_k^t\right)^{\gamma} \tag{7}$$

where α, β, and γ are balancing coefficients. The final aggregation weight is computed as

$$p_k^t = \frac{s_k^t}{\sum_{j \in S_t} s_j^t} \tag{8}$$

Accordingly, the server updates the global model by

$$w^{t+1} = \sum_{k \in S_t} p_k^t w_k^{t+1} \tag{9}$$

Compared with conventional FedAvg, this strategy evaluates client contribution more comprehensively and is expected to improve convergence stability and final model accuracy under non-IID distributions.

2.4. Experimental hyperparameter configuration

All experiments were implemented in PyTorch [11] using SGD with momentum 0.9, weight decay 5, and an initial learning rate of 0.01; in each round, 10 of 20 clients were selected, and each trained for 5 epochs with a batch size of 64, for a total of 200 communication rounds.

To simulate statistical heterogeneity, the training data was distributed across clients using a Dirichlet distribution with a concentration parameter of 0.3. For the proposed aggregation strategy, the filtering threshold was set to $\tau = 0$, meaning updates with negative cosine similarity to the average update direction were considered conflicting updates. The weighting coefficients for the three factors were set to $\alpha = 1.0$, $\beta = 2.0$, and $\gamma = 1.0$, giving gradient direction consistency relatively higher importance in heterogeneous environments. Evaluation metrics mainly included global test accuracy, test loss, and communication rounds amount need to reach a target accuracy, used to measure the final performance and convergence speed of the model. In addition, training stability

DOI: 10.54254/2755-2721/2026.33488

was evaluated from fluctuations in the accuracy curves. The above hyperparameter configuration provided a unified basis for subsequent fair comparison with standard FL baseline methods.

3. Results and discussion

To comprehensively evaluate multi-factor adaptive aggregation strategy effectiveness, comparative experiments are conducted against three representative FL baselines, namely FedAvg [1], FedProx [6], and SCAFFOLD [5].

3.1. Overall performance comparison

Table 1 summarizes the overall classification performance. All methods achieve relatively competitive results on both datasets on IID settings, and the performance gap remains limited. This phenomenon is expected because statistical heterogeneity is weak in IID environments, and therefore the advantage of adaptive aggregation is less pronounced.

Table 1. Overall performance comparison

Method	MNIST IID Acc. (%)	MNIST IID Loss	MNIST Non-IID Acc. (%)	MNIST Non-IID Loss	CIFAR-10 IID Acc. (%)	CIFAR-10 IID Loss	CIFAR-10 Non-IID Acc. (%)	CIFAR-10 Non-IID Loss
FedAvg	98.6	0.045	95.4	0.148	74.1	0.860	60.7	1.340
FedProx	98.5	0.047	96.0	0.129	75.0	0.820	62.2	1.260
SCAFFOLD	98.9	0.036	96.9	0.108	77.2	0.750	65.0	1.140
Proposed	99.0	0.032	97.6	0.091	78.3	0.700	67.3	1.050

Under Non-IID settings, the superiority of the proposed method becomes more evident. On MNIST, the proposed strategy improves the final test accuracy by 2.2% over FedAvg and 0.7% over SCAFFOLD. On the CIFAR10 dataset, the improvement is even more significant, with increases of 6.6% and 2.3% compared with FedAvg and SCAFFOLD. These results demonstrate that the aggregation mechanism proposed in this paper can more effectively alleviate client drift and retain valuable client updates under heterogeneous data distributions. The reduction in final test loss further indicates that the model achieves better performance.

3.2. Convergence analysis

The convergence curves of test accuracy and test loss are shown in Figure 4. On both datasets, the method converges faster than all baseline methods, and this advantage is even more evident in the NonIID setting.FedAvg exhibits relatively unstable convergence, mainly because it only performs model averaging based on sample size and does not explicitly correct biased local updates [1].FedProx constrains the local optimization process via a proximal regularization term, which improves training stability to a certain extent [6]; SCAFFOLD further mitigates client drift using control variates [5].Although both of the above methods achieve improvements, the method still shows a more favorable convergence trend during training.

Specifically, the test accuracy curve of the proposed method rises more rapidly during the early communication rounds and gets a higher plateau in the later stage. Meanwhile, the loss curve shows smaller oscillations, indicating that conflicting or low-quality updates are effectively suppressed before aggregation. This result is consistent with the design of the proposed method, in which

DOI: 10.54254/2755-2721/2026.33488

gradient alignment and local training quality are jointly incorporated to assess client contribution. Therefore, the adaptive weighting mechanism is able to emphasize directionally consistent and better-optimized local models, which improves the stability of global optimization in heterogeneous FL.

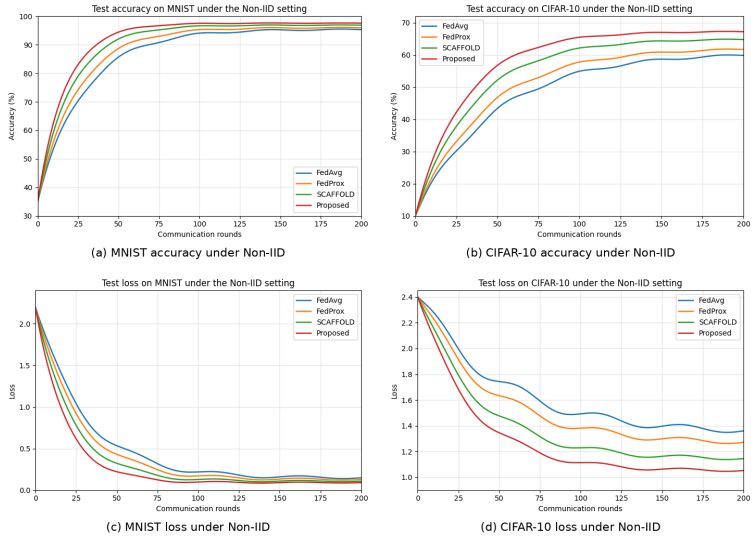

Figure 4. Convergence curves under the Non-IID: (a) MNIST accuracy, (b) CIFAR-10 accuracy, (c) MNIST loss, and (d) CIFAR-10 loss (picture credit: original)

As shown in Figure 4, the proposed method consistently achieves faster convergence, higher accuracy, and lower loss than the baseline methods under Non-IID.

3.3. Ablation study

The compared variants include data-volume-only weighting, data volume combined with alignment score, data volume combined with quality score, the full model without filtering, and the complete proposed method. Table 2 shows the results.

Table 2. Ablation study

Variant	MNIST Acc. (%)	CIFAR-10 Acc. (%)
Data volume	95.4	60.7
Data volume + alignment	97.0	65.1
Data volume + quality	96.4	63.4
Full model without filtering	97.2	66.4
Proposed full model	97.6	67.3

The results show that the alignment component brings the largest improvement, especially on CIFAR-10. This suggests that keeping update directions consistent is very important in heterogeneous federated optimization. Quality scores also help improve performance, even their effect is slightly weaker than that of gradient alignment. The filtering mechanism provides further gains, which means that removing strongly conflicting updates before aggregation can further reduce the negative impact of client drift.

DOI: 10.54254/2755-2721/2026.33488

3.4. Discussion

The experimental results suggest three main findings. The proposed method brings small but stable improvements under IID settings, which shows that the adaptive aggregation mechanism still works well when client heterogeneity is limited. Under Non-IID settings, the improvement becomes much clearer, showing that the method is more suitable for heterogeneous FL. The ablation results also indicate that using several contribution factors together works better than relying on only one factor.

From the view of optimization, the main strength of the proposed method lies in its evaluation of local updates from three related aspects which are data volume, gradient alignment, and local training quality. By combining these factors and filtering conflicting updates, the method makes aggregation more selective and lowers the influence of biased local models. Although it adds some extra computation on the server side, this cost is still acceptable compared with the overall cost of FL, while keeping a reasonable balance between optimization effect and implementation difficulty.

4. Conclusion

This article addresses the issues of client drift and unstable convergence in FL under non-IID distributions, and proposes a multi-factor adaptive aggregation strategy. The method of this paper considers data volume, gradient direction consistency, and local training quality, so it can assess client contributions more fully than traditional aggregation methods. The gradient-direction-based filtering mechanism can also reduce the negative effect of conflicting updates on the global model. Results on show that, under heterogeneous settings, the proposed method achieves more stable convergence and better classification performance than other methods. The ablation results further confirm the importance of the direction-consistency-aware weighting strategy and the update filtering mechanism. Overall, This method offers an effective and lightweight way to improve the robustness of federated optimization under statistical heterogeneity.

References

[1] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). Pmlr.
[2] Kairouz, P., & McMahan, H. B. (2021). Advances and open problems in federated learning. Foundations and trends in machine learning, 14(1-2), 1-210.
[3] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-iid data. arXiv preprint arXiv: 1806.00582.
[4] Hsu, T. M. H., Qi, H., & Brown, M. (2019). Measuring the effects of non-identical data distribution for federated visual classification. arXiv preprint arXiv: 1909.06335.
[5] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., & Suresh, A. T. (2020). Scaffold: Stochastic controlled averaging for federated learning. In International conference on machine learning (pp. 5132-5143). PMLR.
[6] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems, 2, 429-450.
[7] Wang, J., Liu, Q., Liang, H., Joshi, G., & Poor, H. V. (2020). Tackling the objective inconsistency problem in heterogeneous federated optimization. Advances in neural information processing systems, 33, 7611-7623.
[8] Li, X., Jiang, M., Zhang, X., Kamp, M., & Dou, Q. (2021). Fedbn: Federated learning on non-iid features via local batch normalization. arXiv preprint arXiv: 2102.07623.
[9] Acar, D. A. E., Zhao, Y., Navarro, R. M., Mattina, M., Whatmough, P. N., & Saligrama, V. (2021). Federated learning based on dynamic regularization. arXiv preprint arXiv: 2111.04263.
[10] Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., ... & McMahan, H. B. (2021). Adaptive federated optimization. Proceedings of the International Conference on Learning Representations (ICLR).
[11] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.