Bonus 2

1 Non–iid's Influence on FL Model Performance

In the case of non-iid data, data partitioning in federated learning may have a negative impact on the model. This is because in this case, there may be large differences between parts of the dataset, and these differences may affect the training of the model.

Specifically, if the data is not properly segmented, it may result in a trained model that performs well in some regions but not in others. In this way, the model does not fit the entire data set well. In addition, data segmentation may also cause the model to over-fit some specific subsets of data, thus affecting the generalization ability of the model. Specifically, data partitioning can lead to two problems:

Reduced generalization ability: In the training phase, a federation learning model needs to collect data from multiple clients with different but homogeneous distributions and train these data jointly to improve the generalization ability of the model. However, if the data is partitioned in this process so that only part of the data from each client is involved in the training, the model trained by each client will also be affected by its data distribution due to its different data distribution, which leads to the degradation of the generalization ability of the model.

Contribution imbalance: In the case of non-independent and homogeneous distribution of the dataset, some clients' data may be richer or more representative, while other clients' data may be less or more difficult to train. Therefore, when performing data partitioning, some clients may contribute more to the training of the model, while others may contribute less to the training of the model, resulting in an imbalance in contribution and affecting the overall performance of the model.

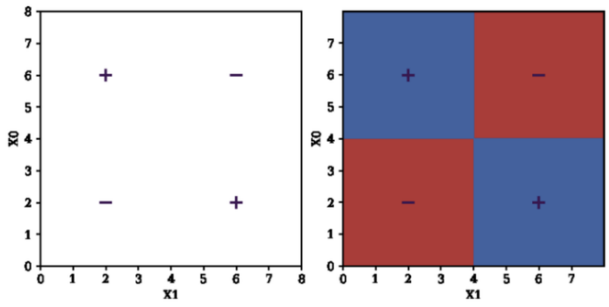
2 Analysis

We specifically analyze the impact of the Non-IID dataset on the FedAVG algorithm from two perspectives, quantitative and qualitative, respectively.

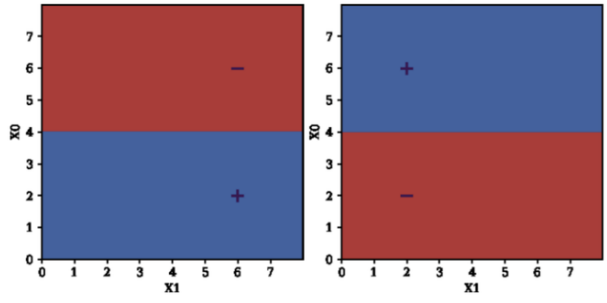
2.1 Qualitative Analysis

First we analyze the impact of Non-IID on model training by a simple classification task. As shown in the left panel of Figure 1(a), we have four samples, each containing two features and a label (x0,x1,label): sample 1 (6,2,+), sample 2 (6,6,-), sample 3 (2,2,-), and sample 4 (2,6,+). In centralized training, classification can be easily performed according to the range of values of features x0,x1, as shown in the right panel of Figure 1(a).

Suppose in federal learning, sample 2 and sample 4 are in client1, sample 1 and sample 3 are in client2, and the data distribution is Non-IID. when local training is performed in client, the classification is performed according to the range of values of features x0,x1, and the corresponding classification model is obtained. As shown in Figure 1(b), the classification models of client1 and client2 become the exact opposite prediction results, so the federally trained aggregation models cancel each other in the corresponding regions, and the model performance is necessarily poor.



(a) Prediction in centralized setting



(b) Hypothesis conflict between two sites in distributed setting

Fig. 1 Non–IID classification simplified model

2.2 Quantitative Analysis

For easy understanding, as shown in Figure 2, the model is assumed to be a two-dimensional vector, and the gradient is the arrow in the figure. For the

data distribution of IID, the model gradient direction is almost the same between clients when FedAVG training is performed, and after mT times of federation training, the federated aggregation model

$$W_{mT}^{f}$$

The model obtained with central training (assuming that all data are trained on the server) is almost identical.

$$W_{mT}^{c}$$

For the Non-IID data distribution, there is a large difference between the client model gradient direction

$$W_{mT}^{k}$$

The federal aggregation model differs significantly from the centralized model in both size and direction

$$W_{mT}^{c}$$

Therefore, the comparison in Figure 1 reveals that the data set of Non-IID causes the federal model to fail to converge to the central training model. In addition, in the federation learning process using algorithms such as SGD, even if the training members are IID data, when the batch_size is small, the data distribution of each batch is not the same, so there is still a large performance loss when performing model aggregation.

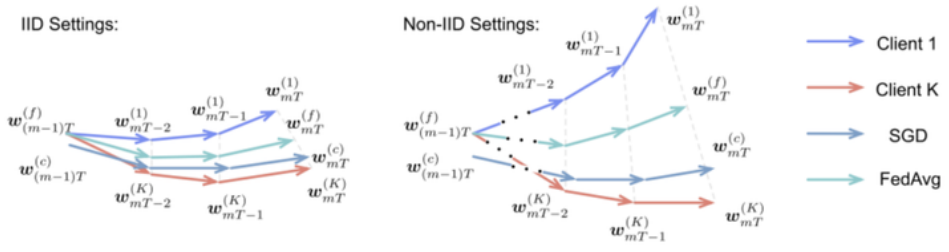


Fig. 2 Schematic diagram of convergence of IID and Non-IID data sets

The difference between the federal training model and the central training model can be represented by the following axiom:

$$\left\|\boldsymbol{w}_{mT}^{(f)}-\boldsymbol{w}_{mT}^{(c)}\right\| \leq \sum_{k=1}^{K} \frac{n^{(k)}}{\sum_{k=1}^{K} n^{(k)}}\left(a^{(k)}\right)^{T}\left\|\boldsymbol{w}_{(m-1)T}^{(f)}-\boldsymbol{w}_{(m-1)T}^{(c)}\right\| + \eta \sum_{k=1}^{K} \frac{n^{(k)}}{\sum_{k=1}^{K} n^{(k)}} \sum_{i=1}^{C}\left\|p^{(k)}(y=i)-p(y=i)\right\| \sum_{j=1}^{T-1}\left(a^{(k)}\right)^{j} g_{\max}\left(\boldsymbol{w}_{mT-1-k}^{(c)}\right),$$

Among them, ：

$$g_{\max}(\boldsymbol{w})=\max_{i=1}^{C}\left\|\nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x} \mid y=i} \log f_i(\boldsymbol{x}, \boldsymbol{w})\right\| \text{ and } a^{(k)}=1+\eta \sum_{i=1}^{C} p^{(k)}(y=i) \lambda_{\boldsymbol{x} \mid y=i}$$

- Corollary 1

The difference between the two is mainly composed of two parts: one part is the difference between the model at the last training, and the other part is caused by the difference between the distribution of clientk and the overall distribution.

- Corollary 2

The same model initialization for all clients helps reduce the differences between the federally trained model and the centrally trained model.

- Corollary 3

When the same initialized model is used, the difference between the federated training model and the central training model is mainly caused by the difference in data distribution, and the exact magnitude is also related to the learning rate η the number of training iterations T and the gradient gmax on the client side at each update of the federated training.

3 Potential Methods

There are several ways to reduce the impact of data partitioning for the case of non-independent homogeneous distribution of the data set as follows:

Optimization of aggregation strategy: In the federal learning framework, the aggregation strategy is the process of collecting model updates from all participants and averaging or weighting them. Optimizing the aggregation strategy can mitigate the contribution imbalance by increasing the weights of some clients or adaptively adjusting the weights of clients to improve the overall performance.

Local training and iteration: For some model layers that are capable of local training and iteration, these model layers can be left to clients for local training and iteration to better compensate for the different data distribution and reduce the dependency on the aggregation process.

Data augmentation: By performing manual data augmentation on client-side datasets, such as rotate, flip, and crop operations, the similarity between client-side data can be increased and the variability of data distribution can be reduced, thus increasing the generalization performance of federal learning.

Improvement of joint training strategy: In the joint training process, more effective joint training strategies, such as optimization algorithm design and gradient cropping techniques in federal learning, can be used to improve the generalization performance of the model and alleviate the problem of different data distributions.

Data filtering and selection: Data filtering and selection can be used to optimize the construction of data sets, such as removing noisy data, selecting representative data, or resampling data, so as to reduce the non-IID nature in the data set and improve the performance of the model under different data distributions.

4 Reference

[1] Liang, H., Hu, Y., Gan, C., & Liu, J. (2020). Towards federated learning at scale: System design. IEEE Transactions on Services Computing.

[2] Yin, W., Wang, Q., & Sun, Z. (2020). Federated learning with non-IID data: A convergence analysis. arXiv preprint arXiv:2006.05796.

[3] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., & Talwalkar, A. (2020). Fair resource allocation in federated learning. In Proceedings of the 37th International Conference on Machine Learning (ICML 2020) (pp. 6076-6086).