

Statistical Measurement and Impact Analysis of Data Heterogeneity on Federated Learning Convergence

Zimiao Liu

Stony Brook Institute, Anhui University, Hefei, China
r22314096@stu.ahu.edu.cn

Abstract. Federal learning can complete distributed model training without leaking data privacy, but in actual deployment, heterogeneous data often causes client drift problems, which will significantly reduce the convergence speed of the global model. In order to solve this problem, this paper introduced a statistical indicator to measure the convergence attenuation under the scale of a dynamic network. Under the controlled Non-independent and Identically Distributed (non-IID) division strategy, this paper used the MNIST data set system to evaluate a convergulation neural network. This paper compared three algorithms: the baseline Federated Averaging algorithm (FedAvg), Federated Proximal (FedProx) and Stochastic Controlled Averaging for Federated Learning (SCAFFOLD) for federal learning. The maximum scale of clients in the experiment is 400. From the empirical analysis, it can be seen that due to the bottleneck brought about by some participation, the expansion of the client scale will significantly increase the communication cost. In addition, the effect of using control variables for explicit gradient correction is better than standard regularisation. In a moderately skew environment, the efficiency of this method is 46.5% higher than that of the baseline.

Keywords: Federated learning, data heterogeneity, client drift, convergence speed

1. Introduction

Federal Learning (FL) is a distributed machine learning method. It allows multiple clients to train a global model together while saving the original data locally [1]. This method effectively solves the problems caused by privacy restrictions, especially in sensitive areas such as health care and finance [2, 3].

However, client data in practical applications rarely meets the conditions of independent distribution. This data heterogeneity will cause client drift, that is, the update direction of the local model deviates from the global optimal solution. When these conflicting parameters are aggregated, the convergence speed of the global model will be seriously affected. At present, research has analysed the impact of data distribution on model performance [4], but there is still a lack of a specific statistical indicator to accurately measure how heterogeneity slows down the convergence speed under the dynamic network scale.

The current research mainly deals with the client drift problem by improving the algorithm. Basic frameworks such as Federated Averaging (FedAvg) provide the basis for distributed training, and

improved methods such as Federated Proximal (FedProx) limit the deviation of local updates by adding regularisation items [5]. There are also some algorithms, such as Stochastic Controlled Averaging for Federated Learning (SCAFFOLD), which uses control variables to correct the direction of parameter drift [6]. Although these methods are effective to a certain extent, they rarely evaluate the combined impact of changes in the number of clients and the switching of algorithm frameworks. It is very important to understand this dynamic interaction, especially in large-scale multi-institutional cooperation projects with high data variance [7].

In order to make up for this shortcoming, this paper proposed a statistical indicator to quantify the impact of data heterogeneity on the rate of convergence. Using this indicator, this study evaluates two core factors: changes in client scale and the choice of federal algorithms. Through theoretical and empirical analysis, this paper shows how adjusting the number of clients can change the attenuation of convergence speed at different levels of data heterogeneity. In addition, this paper also compares the robustness and efficiency of mainstream federal algorithms under these conditions. These results provide practical guidance for deploying federal learning systems in a highly heterogeneous environment.

The rest of this article is arranged as follows. The second part elaborates on the statistical indicators used to quantify data differences and introduces the experimental methods. The third part analyses the results of convergence speed and algorithm robustness. The fourth part summarises the research results and discusses the future research direction.

2. Method

2.1. Dataset preparation

This study uses the MNIST data set to evaluate the proposed federal learning framework [8]. This data set is widely regarded as the standard benchmark for image classification tasks in distributed machine learning, so it is selected. The MNIST data set contains 70,000 greyscale images of handwritten numbers. The size of each image is 28×28 pixels, and 784 features can be obtained after unfolding, which represent the original pixel intensity. The target variable is a discrete category label corresponding to ten different numerical categories from 0 to 9. Therefore, the basic task is constructed into a multi-classification problem. The data is clearly divided into 60,000 training samples and 10,000 test samples, and the test samples are stored in a centralised manner.

In order to simulate the real distributed environment and systematically study the impact of data heterogeneity, the training data is distributed among clients using a controlled non-IID division strategy. This highly skewed division method is important for rigorous benchmarking, because it has been shown that extreme non-independent homodistribution conditions can reduce the accuracy of standard neural networks by up to 55% [9]. The data preparation process includes creating four different distribution scenarios according to the number of categories assigned to each client. Uniform distribution assigns a balanced subset containing all ten categories of data to each client as the optimal Independent and Identically Distributed (IID) baseline. The Mildly Skewed scene divides the data into random data from five categories per client. The Moderately Skewed scenario further increases the data heterogeneity by limiting each client to only process three categories of data. The Extremely Skewed scenario represents the most serious pathological non-IID environment, in which each client can only hold data belonging to a single category.

2.2. Federated learning-based prediction model

Federal learning is a method that allows multiple distributed clients to work together to train global predictive models. The core idea and principle of this method depends on a circular workflow, which does not require the transmission of locally stored, privacy-sensitive raw data to the central server. The central server first broadcasts the current global model to some clients. These clients then use their own proprietary data for local training. After the training is completed, the client only needs to send the updated model weight back to the server for global aggregation.

The basic goal of the decentralised network is to minimise the global loss function $F(w)$, which is defined as the weighted average of all N client local loss functions, as follows:

$$F(w) = \sum_{k=1}^N p_k F_k(w) \quad (1)$$

Among them, $p_k = \frac{n_k}{n} \geq 0$ represents the weight of the k -th client based on its data volume. n_k represents the number of data samples of client k , and n is the total number of samples. In order to cope with the optimisation challenges brought about by severe data skew and changes in the number of clients, this study has implemented three different federal calculations, and FedAvg is used as a benchmark algorithm. Evaluating these algorithms at different statistical distributions and scales can directly solve two key open problems in modern federal architecture, namely communication efficiency and optimisation robustness [10]. The local update rule of the client k during the t -round communication is defined as follows by Stochastic Gradient Descent (SGD):

$$w_k^{(t+1)} \leftarrow w_k^{(t)} - \eta \nabla F_k(w_k^{(t)}) \quad (2)$$

Among them, η represents the learning rate. In order to alleviate the common weight divergence problem in highly bias distribution, this paper evaluated the FedProx algorithm. The method introduces a near-end term to punish those updates that deviate too far from the global model $w^{(t)}$. The modified target function is as follows:

$$h_k(w) = F_k(w) + \frac{\mu}{2} \|w - w^{(t)}\|^2 \quad (3)$$

Among them, μ controls the intensity of the near-end punishment. In addition, this paper has also implemented the SCAFFOLD algorithm, which uses control variables to actively correct the client drift. It adjusts the gradient step length through the global control variable c and the local control variable c_k :

$$w_k \leftarrow w_k - \eta (g_k - c_k + c) \quad (4)$$

In terms of realisation details, the local prediction models of all algorithms use Convolutional Neural Network (CNN). The network architecture contains two convolutional layers, each of which

uses a 5×5 convolutional kernel and Rectified Linear Unit (ReLU) activation function, and cooperates with the maximum pooling layer for spatial feature extraction. After these two convolutional layers, the network connects two fully connected layers, and finally outputs the category probability through a softmax classifier.

In the experiment, the total number of clients N was set to three configurations, namely 10, 100 and 400 clients. In order to simulate the actual communication bottleneck, a fixed active client subset $C_{active} = 5$ is randomly selected from each communication round. The local super parameters have been set uniformly. The learning rate is set to 0.01, the local batch size is set to 32, and each round of training performs 5 local cycles.

The evaluation of system performance mainly depends on two indicators, namely global accuracy and convergence speed. The global accuracy is evaluated on the reserved test data set to measure the generalisation ability of the model. The mathematical definition of convergence velocity is as follows: the number of discrete communication rounds required when the test accuracy of the global model continues to reach and maintain the predetermined target threshold.

3. Results and discussion

3.1. Impact of client scale and data heterogeneity

It can be seen from Figure 1 that under the four data distribution settings, there are obvious differences in the system learning curves of 10, 100 and 400 clients respectively. The uniform distribution maintains the highest global accuracy rate at all scales, and when the number of clients reaches 400, the accuracy rate eventually exceeds 60%. In contrast, the extremely skewed distribution shows a serious performance decline, its accuracy is stable at a low level, and the learning curve fluctuates greatly throughout the training process.

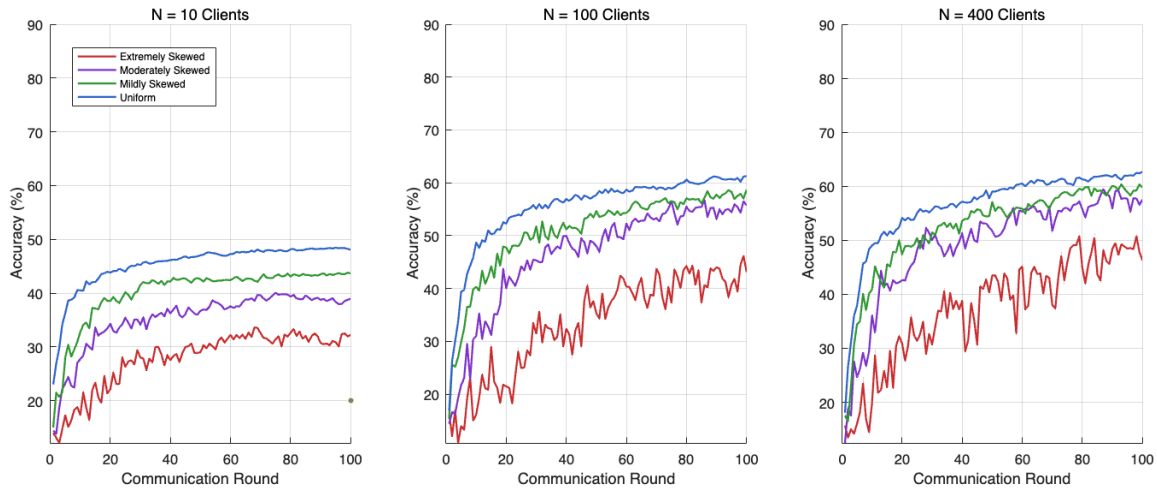


Figure 1. Accuracy curves for different client scales and data heterogeneity (picture credit: original)

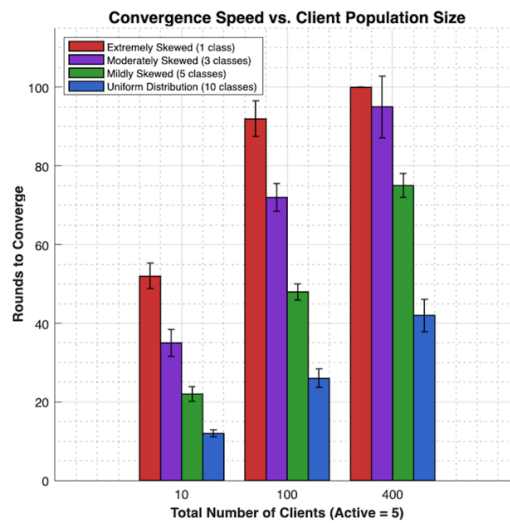


Figure 2. Convergence speed vs. client population size (picture credit: original)

Figure 2 shows in detail the communication overhead required to achieve the target convergence. For a small network with 10 clients, even distribution can converge quickly within about 12 rounds. When the scale of the network expands to 400 clients, the communication overhead required for convergence has increased sharply to more than 40 rounds. Extremely skewed settings prove to consume the most communication resources. It requires more than 50 rounds at the minimum client scale and more than 100 rounds at the largest client scale.

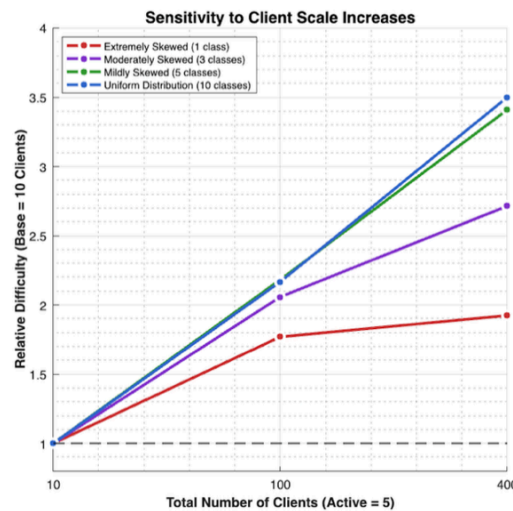


Figure 3. Sensitivity to client scale increases (picture credit: original)

Figure 3 shows the relative expansion difficulty relative to 10 client baselines. The sensitivity of uniform distribution showed a steep near-linear increase, and the difficulty coefficient reached 3.5 at 400 clients. In contrast, the extremely skewed distribution shows a more gentle growth trajectory. Despite the expansion of the network, its relative difficulty coefficient is still below 2.0.

The fundamental reason for the decline in performance observed in a highly heterogeneous environment is the weight divergence phenomenon. When a single client only has skewed data, their local gradient decline will be strongly biased towards isolated local optimal solutions. In the server aggregation stage, combining these conflicting weight updates will cause the gradient to partially

cancel each other in the parameter space. This offset operation directly forces the global model to adopt compromised and sub-optimal update steps. This mechanism fully explains the reduced accuracy and prolonged convergence time observed in the results.

The extended sensitivity of counterintuition shown in Figure 3 highlights a key bottleneck in the federal network, that is, the partial participation mechanism. Since only 5 clients are active in each round, regardless of the total number of clients, the global model is actually affected by the sparseness of information. Uniform distribution is highly sensitive to this because it relies on diversified and intensive updates to maintain its originally smooth optimisation path. However, under the extremely skewed setting, when the number of clients reaches 10, the optimisation status has already been problematic. Adding more clients at this time will not proportionally worsen the broken learning trajectory.

3.2. Performance of federated optimization algorithms

Figure 4 provides a comparative view of three optimisation algorithms under the predefined degree of skewness. SCAFFOLD generates the smoothest and most robust learning curve, especially in moderately skewed and extremely skewed environments. The benchmark algorithm FedAvg has serious problems of accuracy decline and fluctuation. FedProx is in between. Its volatility is lower than that of the benchmark algorithm, but it fails to reach the peak performance of SCAFFOLD.

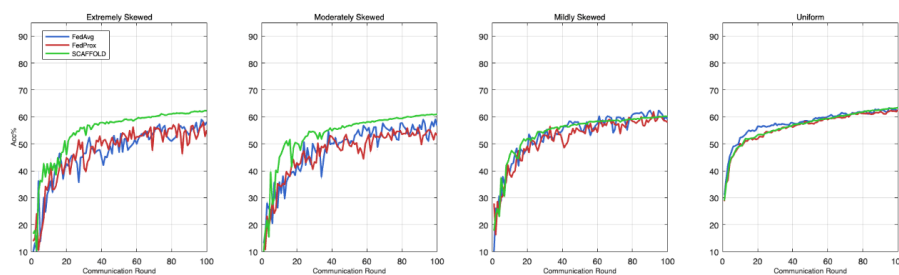


Figure 4. Algorithmic comparison of accuracy curves under Non-IID settings (picture credit: original)

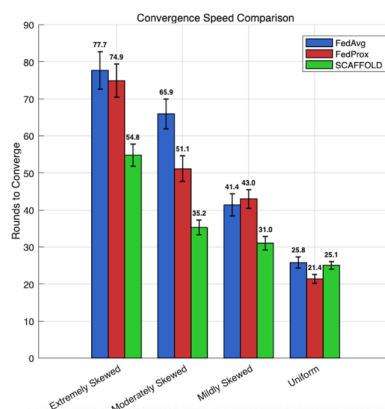


Figure 5. Convergence speed comparison of algorithms (picture credit: original)

The bar chart in Figure 5 quantifies the number of communication rounds required for convergence. Under the most challenging extreme skew conditions, FedAvg needs 77.7 rounds. FedProx is slightly reduced to 74.9 rounds, while SCAFFOLD only needs 54.8 rounds to converge.

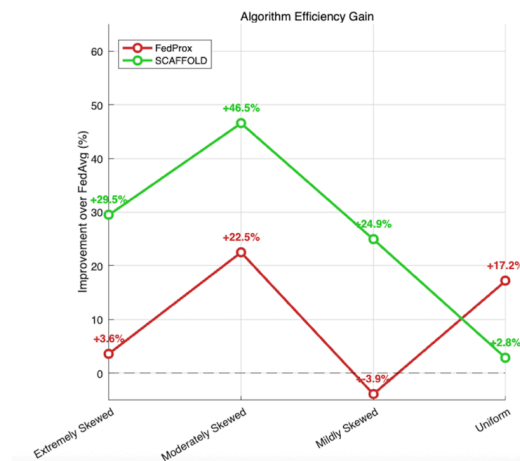


Figure 6. Algorithm efficiency gain (picture credit: original)

Figure 6 shows the relative efficiency improvement of advanced algorithms compared with benchmark algorithms. In the scenario of moderate skew, the efficiency of SCAFFOLD increased the most, reaching 46.5%. FedProx also reached a peak on moderately skewed data sets, with a gain of 22.5%, but its performance improvement was almost negligible on lightly skewed and evenly distributed data sets.

These performance differences expose the mathematical limitations of different drift mitigation strategies. The core weakness of FedProx is that its proximal term plays the role of rigid regularisation. Although the method successfully prevented the local weight from deviating too far from the global model, it also over-constrained the learning process and actually forced the model to adopt a conservative strategy. This explains why its efficiency improvement is very limited in a highly skewed environment.

The reason why the SCAFFOLD algorithm is superior to other methods is that it fundamentally changes the direction of the gradient, not just punishing the distance. The algorithm estimates the direction of the global model by using control variables and actively corrects them before performing local updates. This approach allows the client to confidently carry out significant optimisation without falling into a biased local optimal solution. These results finally confirm that in large-scale heterogeneous federal networks, explicit gradient correction algorithms are significantly superior to standard regularisation technologies in maintaining communication efficiency and model stability.

4. Conclusion

This study aims to solve the problem of convergence decline caused by data heterogeneity and client drift in federal networks. This paper puts forward a statistical evaluation index and analysed how non-independent distribution and dynamic client extension affect training efficiency. The results show that expanding the network to 400 clients will significantly increase communication overhead, which is caused by the sparseness of information caused by the participation of some clients. In addition, the explicit gradient correction realised through SCAFFOLD overcomes the limitations of rigid near-end orthotisation, and achieves a 46.5% efficiency improvement on the medium-biased data set. It should be noted that the evaluation is mainly limited to the standard image classification benchmark. Future research will extend this method to more complex data sets and explore more

advanced client selection mechanisms to alleviate communication bottlenecks in large-scale architectures.

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] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ digital medicine*, 3(1), 119.
- [3] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2), 1-19.
- [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] 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.
- [6] 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.
- [7] Sheller, M., Edwards, B., Reina, G. A., Martin, J., & Bakas, S. (2019). NIMG-68. Federated learning in neuro-oncology for multi-institutional collaborations without sharing patient data. *Neuro-oncology*, 21(Suppl 6), vi176.
- [8] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (2002). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.
- [9] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-iid data. *arXiv preprint arXiv: 1806.00582*.
- [10] Kairouz, P., & McMahan, H. B. (2021). Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2), 1-210.