Proceedings of the 4th International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/2026.33840

Quantitative Analysis of the Impact of Data Heterogeneity on the Convergence Speed of Federated Learning

Zifeng Wang

Faculty of Science, the Hong Kong Polytechnic University, Hong Kong, China
wzf9697@gmail.com

Abstract. Data heterogeneity presents a significant challenge in federated learning, often severely impeding training efficiency. This study quantitatively measures its impact on convergence speed. A controlled experimental framework is established, utilizing the Dirichlet distribution to generate client data with tunable heterogeneity levels, and systematic tests are conducted on the MNIST dataset. The investigation further examines how the scale of the client population and the selection of the aggregation algorithm influence the outcomes. The key findings are: 1) Increased heterogeneity necessitates substantially more communication rounds to reach a target accuracy. 2) A larger client population significantly amplifies the negative effects of data heterogeneity. 3) Algorithm robustness varies greatly; SCAFFOLD demonstrates superior performance, followed by FedProx, while the standard FedAvg is only effective with nearly uniform data distributions. This work provides clear, quantitative guidance for designing efficient federated learning systems capable of handling real-world, non-uniform data. These insights offer practical guidance for scalable and robust deployments.

Keywords: Federated Learning, Data Heterogeneity, Non-IID, Convergence Speed, Communication Efficiency

1. Introduction

Federated learning is a type of distributed machine learning. It enables model training on different devices or organizations collaboratively without sharing the data, which protects its privacy. Several areas like healthcare, finance, and the internal of things shows a great potential [1, 2]. The training process takes place over several rounds. During each round, clients will carry out training on their end and send updates of their model to the server. Nonetheless, a big challenge in real-world applications is that the data of different clients is often not independent and identically distributed (Non-IID) [3]. The difference in data types slows down the training, causes instability in model prediction and also the failure to converge [4].

According to different studies, data heterogeneity aggravates the issue of client drift. The phenomenon known as client drift causes the local model updates on clients to take a path increasingly distant from the global optimum. Therefore, client drift causes convergence to slow. Most previous work describes this problem in a more general sense and/or works on specific algorithms [5, 6]. A fundamental question that has not been rigorously studied is how the degree of

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.33840

data heterogeneity changes the convergence speed of federated learning quantitatively. In particular, it is unclear if the number of communication rounds to reach a target accuracy displays a particular pattern under increasing levels of heterogeneity. Investigating this question may serve as solid evidence for the widely held belief that "heterogeneity slows training." It may also help in resource planning, algorithm selection, and parameter tuning for real federated learning systems.

Existing studies quantifying the impact of data heterogeneity in federated learning shares some important shortcomings. In the first place, most experimental studies are limited to a single or only a few heterogeneity datasets. This prevents a systematic exploration of how performance varies over different grades of heterogeneity intensity [7]. It becomes hard to envisage a situation where product heterogeneity would not affect behavior. Evaluation metrics often focus on model accuracy by the end of convergence; however, this could lead to a learning outcome that doesn't reflect the speed at which things learn, or when it stops learning. It neglects important practical issues like the costs of communication and time to reach accurate conclusions. There is still no standardized and reproducible analytical framework and benchmark to study the convergence behaviour of fundamental algorithms, such as Federated Averaging (FedAvg) under controlled and quantifiable heterogeneity conditions. The gap in the understanding of this issue makes any assessment impractical [5].

To fill in these gaps, this paper offers a controlled experimental study to analyze the quantitative correlation between data heterogeneity and the convergence speed of federated learning. A quantifiable framework using the Dirichlet distribution for generating client data with adjustable heterogeneity (parameter α) is proposed. Experiments that were systematic are conducted on MNIST to isolate the effect of distribution. Basically, it is possible to see a clear quantitative pattern in the data showing that greater heterogeneity requires more communication rounds to achieve a target accuracy. The aim of this work is to develops an extensible benchmark to evaluate reliable federated learning algorithms performance and behaviour in Non-IID environments.

2. Method

2.1. Controllable heterogeneity study based on dirichlet distribution

The purpose of this paper is to set up a continuous heterogeneity calibration research framework in order to study the quantitative relationship between the strength of heterogeneity and convergence performance.

2.1.1. Data partitioning and heterogeneity generation

A Dirichlet distribution is used to partition the training data [8, 9], thereby simulating Non-IID label distribution across different clients. For a dataset with K classes, the data distribution of each client k is characterized by $p_k \sim Dir(\alpha \cdot 1)$, where α represents the concentration parameter. Lowering the value of α yields a sparser p_k, meaning the client's data is more likely to be concentrated on just a few classes, which leads to stronger heterogeneity. This experiment uses five gradient values of $\alpha = \{0.1, 0.3, 0.5, 1.0, 5.0\}$ to cover a broad range of data distributions from extremely heterogeneous to almost homogeneous (IID).

2.1.2. Model and training configuration

• Dataset and Task: The MNIST handwritten digit classification dataset is used.

DOI: 10.54254/2755-2721/2026.33840

• Client Settings: A total of 50 clients is simulated. Every round of communication randomly selects 10 users for training.

• Model Architecture: The organization uses a vanilla CNN which can be very simple but effective.

• Federated Learning Algorithm: The classic Federated Averaging (FedAvg) algorithm is employed as the baseline.

• Training Parameters: To prevent the influencing of the other testing results, all the local training epochs are set to 5, learning rate to 0.01, communication rounds to 100 and target test set accuracy to 0.85.

The key feature of this experiment is to keep all training hyperparameters constant and vary only the data partitioning parameter α. Thus, all performance differences in this experiment can be attributed to data heterogeneity.

2.2. Discrete heterogeneity configuration study based on class distribution

In order to test the conclusions with different patterns of qualitative heterogeneity and to study the impact of outliers, the experiment designs four client data configurations with clear explanations.

2.2.1. Heterogeneity configuration design

There are four configurations (the total number of samples per client is fixed at 50 so that the effect of sample size differences can be eliminated) shown in Figure 1.

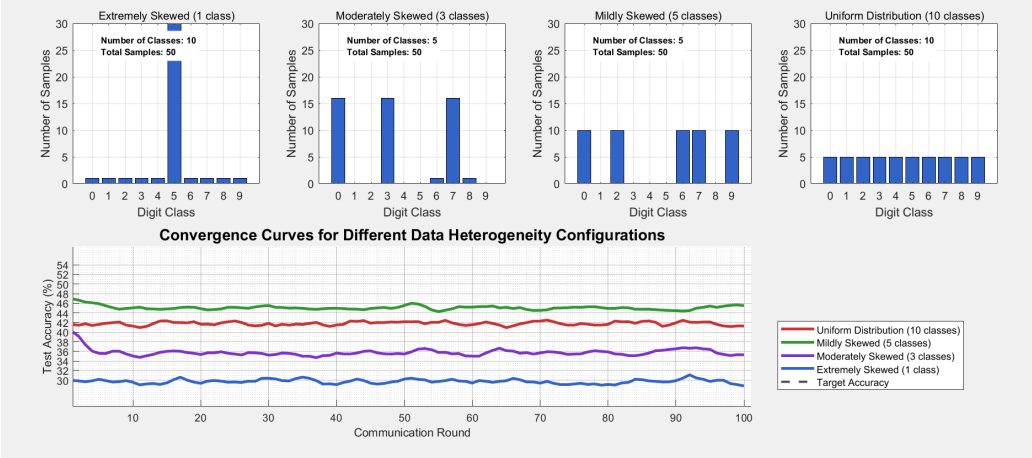


Figure 1. The example of different client settings (picture credit: original)

• Extremely Skewed: The client contains samples from only 1 class.

• Moderately Skewed: Client samples are concentrated in 3 classes.

• Mildly Skewed: Client samples are distributed across 5 classes.

• Uniform Distribution: Client samples are evenly distributed across all 10 classes (5 samples per class).

2.2.2. Model and training configuration

• Dataset: The MATLAB Digit Dataset (8k training, 2k test images) is used.

• Federated Settings: The use of a template matching classifier simplifies the computation and emphasizes the effect of the data distribution itself.

DOI: 10.54254/2755-2721/2026.33840

• Model: There are 20 clients in total, randomly selected 5 clients from that pool on per round basis. There are 100 communication rounds. Every configuration is performed 8 times for statistical stability.

2.3. Comparative experiments: analysis of client population size and algorithm robustness

In addition to those mentioned experiments, this study further extends experimental studies on system scale and more advanced algorithms in heterogeneous environments.

2.3.1. Client population size impact experiment

For the same four heterogeneity configurations, the total client population size was set to {20, 50, 80, 100}, while keeping 5 clients active per round. The experiment examines the amplifying effect of system scaling on heterogeneity impact by comparing the number of communication rounds needed to achieve target accuracy (75%) under different population sizes.

2.3.2. Aggregation algorithm comparison experiment

Using the same experimental setup, comparison of three following federated aggregation algorithms is done.

• FedAvg: The standard Federated Averaging algorithm, serving as the performance baseline.

• FedProx: The FedProx technique incorporates a proximal term to restrict local updates and curtail client drift. In this experiment, a value of $\mu=0.01$ is set for the proximal term coefficient.

• SCAFFOLD: The client drift error is corrected using the control variables. To make this comparison in the experiment, its client number is fixed to 50.

2.3.3. Evaluation and quantification metrics

• The convergence speed is directly measured by the number of communication rounds needed to achieve a predetermined target test accuracy.

• In comparative trials, the experiment measures a heterogeneity score (a heterogeneity measure based on the spread of the label distribution) and the average number of classes per client for all heterogeneity configurations for correlation analysis.

• To establish relationship between heterogeneity score and total rounds of communication, a linear regression is used. The correlation coefficient rand and the p-value are calculated to verify statistical correlation.

3. Results and discussion

3.1. Quantitative impact of heterogeneity strength on convergence speed

Through a controlled experiment utilizing Dirichlet-based data partitioning, it is shown that data heterogeneity significantly slows down convergence in federated learning. As shown in Figure 2, the average number of communication rounds needed for the model to achieve 85% target accuracy increases significantly as the Dirichlet parameter α decreases. Moreover, in the case of strong heterogeneity ($\alpha=0.1$), the final average accuracy of the model is also lower than in weaker heterogeneity. This indicates that heterogeneity can not only increase the time required for convergence but may also cause the resulting model to have lower performance.

DOI: 10.54254/2755-2721/2026.33840

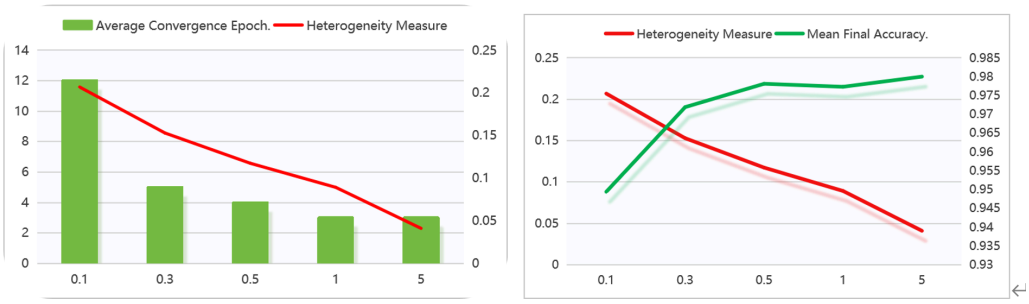


Figure 2. Impact of dirichlet parameter α on convergence speed and model accuracy (picture credit: original)

The test with four distinct heterogeneity configurations offers more intuitive support. Of the four configurations, the convergence curve for the "Extremely Skewed" is almost flat where a final accuracy of 40% reached shown in Figure 3. This is because clients can only alter the model parameters for a single digit class. It is important to point out that the "Mildly Skewed" configuration (i.e. samples distributed among 5 classes) performs similarly to or slightly better than the "Uniform Distribution" configuration. Such form of heterogeneity, provided it is concentrated in certain classes only, can help client become expert at their respective specialized classes. Such a phenomenon can help in aggregating global model. It provides a fresh angle on the use of heterogeneity.

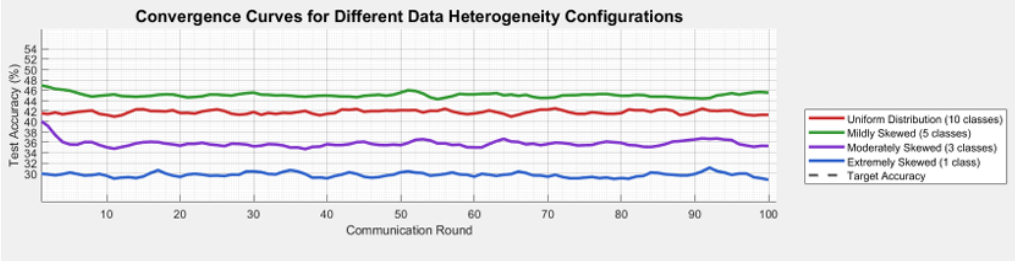


Figure 3. Convergence behavior under four discrete data heterogeneity configurations (picture credit: original)

3.2. Amplifying effect of client population size on heterogeneity impact

The findings of the comparative experiment indicate an important interaction between system scale and data heterogeneity. The heterogeneity score showed a strong positive correlation ($r > 0.83$, $p < 0.001$) with the number of rounds to converge to the target accuracy according to the statistical analysis. According to the quantitative model, for every increase of 0.1 units in the heterogeneity score, an average of 4.2 more communication rounds is required for economic convergence shown in Figure 4.

DOI: 10.54254/2755-2721/2026.33840

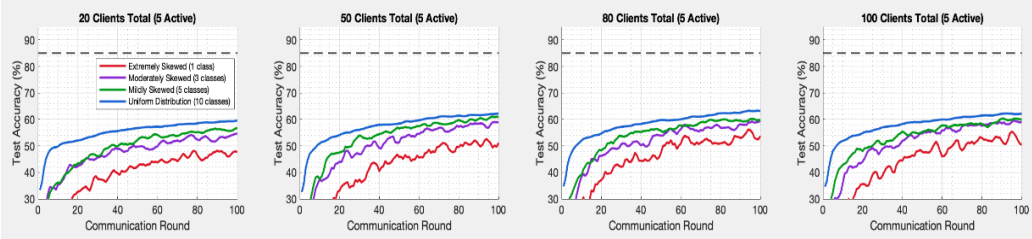


Figure 4. Interaction between system scale (total clients) and data heterogeneity on convergence (picture credit: original)

Above all, the greater the total number of clients, the greater the negative effect of heterogeneity. When the client population scale is increased from 20 to 100 under the "Extremely Skewed" configuration, there is a sharp rise in the rounds required for convergence. Meanwhile, there is little effect of this increase under the "Uniform Distribution" configuration. It suggests that additional communication costs may need to be taken into account when designing a large-scale federated learning system with high data heterogeneity.

3.3. Comparison of robustness of different aggregation algorithms in heterogeneous environments

The performance of the three algorithms differs significantly when facing heterogeneity, and this difference widens as heterogeneity increases.

Performance Ranking: Across all heterogeneity configurations and client population sizes, the ranking for robustness is: SCAFFOLD > FedProx > FedAvg. Under the most challenging conditions of highest heterogeneity ("Extremely Skewed") and largest scale (100 clients), FedAvg requires 84.6 rounds to converge, while SCAFFOLD requires only 42.3 rounds, achieving roughly double the efficiency shown in Figure 5.

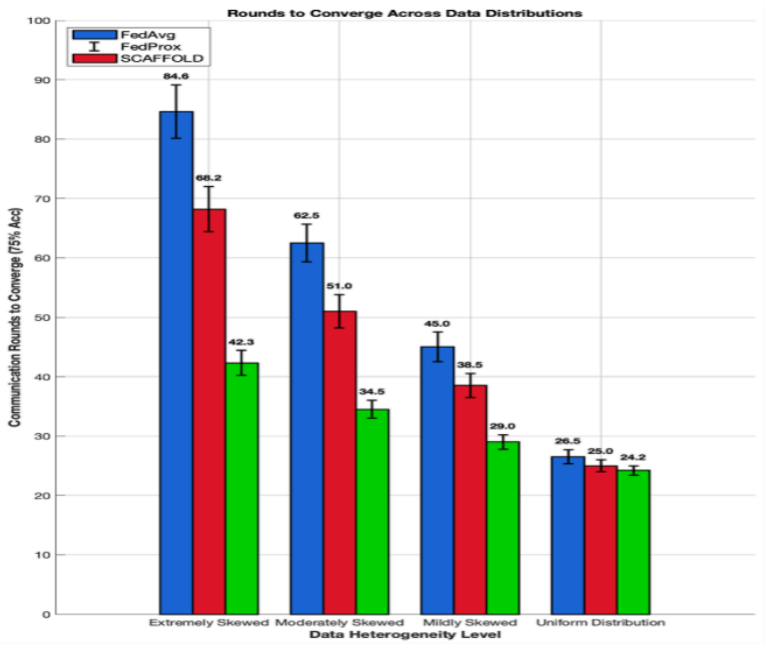


Figure 5. Communication rounds required for convergence by algorithm and heterogeneity level (picture credit: original)

DOI: 10.54254/2755-2721/2026.33840

Quantifying Robustness: Compared to the FedAvg baseline, SCAFFOLD reduces, on average, 43% of the negative impact caused by heterogeneity. A robustness metric (sensitivity to heterogeneity) is calculated for each algorithm; SCAFFOLD scored the best (0.15), followed by FedProx (0.24), with FedAvg being the worst (0.28) shown in Figure 6.



Figure 6. Efficiency gain of FedProx and scaffold over Fedavg baseline across heterogeneity levels (picture credit: original)

Applicable Scenarios: In low-heterogeneity environments, the performance of the three algorithms is similar, giving FedAvg an advantage due to its simplicity and efficiency shown in Figure 7. However, in medium-to-high heterogeneity environments, the advantages of FedProx and SCAFFOLD become prominent. SCAFFOLD demonstrates the greatest robustness when dealing with high heterogeneity, though the increased two-way communication overhead it introduces must be weighed in system design.

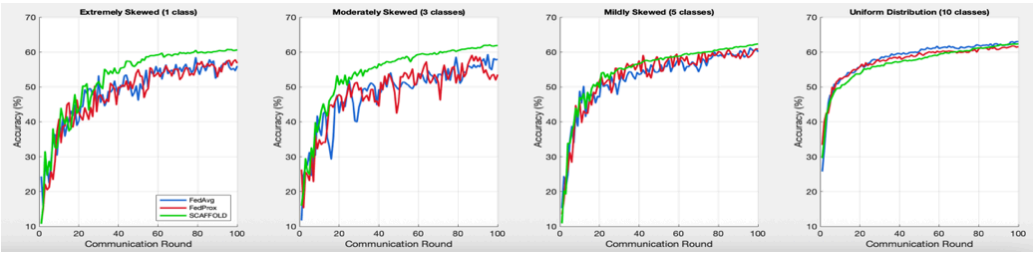


Figure 7. Performance comparison of federated aggregation algorithms (Fedavg, FedProx, Scaffold) under four discrete heterogeneity configurations (picture credit: original)

DOI: 10.54254/2755-2721/2026.33840

4. Conclusion

This work gives clear, measured answers on how data heterogeneity affects federated learning, which is a double-edged sword. Very uneven data stops learning, but a moderate skew can sometimes help clients learn specialized skills. It is also proved that adding more clients amplifies the slowdown from heterogeneity. Therefore, choosing the right algorithm is critical. For high heterogeneity, use SCAFFOLD. For medium cases, FedProx is a good choice. Use the simpler FedAvg only when data is nearly uniform. Before starting a federated project, teams should check their data's heterogeneity. This helps them pick the best system size and algorithm to save time and communication costs. Future work should test these ideas in more complex, real-world situations.

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 A. Singh & J. Zhu (Eds.), Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Vol. 54, pp. 1273–1282). PMLR.

[2] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D' Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., … Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 14(1–2), 1–210.

[3] Li, X., Huang, K., Yang, W., Wang, S., & Zhang, Z. (2020). On the convergence of FedAvg on non-IID data. In International Conference on Learning Representations. https: //openreview.net/forum?id=HJxNAnVtDS

[4] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-IID data. arXiv preprint arXiv: 1806.00582.

[5] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., & Suresh, A. T. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. In H. D. III & A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning (Vol. 119, 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] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)(pp. 1273-1282).

[8] Connor, R. J., & Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. Journal of the American Statistical Association, 64(325), 194-206.

[9] Lin, J. (2016). On the dirichlet distribution. Department of Mathematics and Statistics, Queens University, 40.