Impact of Data and System Heterogeneity on Federated Learning and Biased Client Selection

Abdirashid Chorshanbiyev
Politecnico di Torino
Torino, Italy
s314800@studenti.polito.it

Mirkomil Egamberdiev
Politecnico di Torino
Torino, Italy
s315875@studenti.polito.it

Abstract

Federated learning (FL) is a machine learning setting in which many clients (e.g. mobile devices or entire organizations) collaboratively train a model under the orchestration of a central server (e.g. a service provider) while keeping the training data decentralized. FL embodies the principles of focused data collection and minimization, mitigating many of the systemic privacy risks and costs associated with traditional, centralized machine learning and data science approaches. In this report, we examine how data and system heterogeneity influence model training in the context of FL. Data heterogeneity is simulated through non-IID data splits, while system heterogeneity is modeled by varying client participation levels. Our study demonstrates that selecting clients based on local loss can accelerate convergence and improve test accuracy. Using this insight, we explore an adaptive client selection framework that balances convergence speed and model performance. Our experimental results show that strategic client selection can lead to significantly faster convergence and improved accuracy compared to traditional random selection methods.

1. Introduction

In recent years, the exponential growth of data and the need for privacy-preserving machine learning solutions have led to the emergence of Federated Learning (FL) as a transformative paradigm. Unlike traditional centralized machine learning, where data is collected and processed in a central server, FL enables model training directly on distributed devices or local nodes. This decentralized approach ensures that sensitive data remains on user devices, aligning with privacy regulations such as GDPR, while also reducing communication overhead and latency.

Despite its advantages, FL introduces several challenges that can impact model performance and convergence. One of the primary challenges is data heterogeneity, as client data distributions are often non-IID (non-independent and identically distributed). This results in models that generalize poorly across all clients. Additionally, system heterogeneity arises due to varying computational resources, network conditions, and participation frequencies among clients. FL also suffers from communication bottlenecks, as frequent model updates between clients and the central server can lead to significant network overhead. Furthermore, client participation is often biased or unpredictable, as not all clients are available or equally involved in each training round. This imbalance can slow down convergence and degrade model accuracy. To address these challenges, particularly those related to slow convergence and suboptimal performance caused by random client participation, this project explores biased client selection based on high local loss. In standard FL frameworks, clients are typically selected at random or proportionally based on data availability. However, Some previous studies [pow-d] suggest that prioritizing clients with higher local loss—those whose models currently exhibit the most error—can accelerate convergence and improve global model accuracy. Inspired by this insight, this project implements and evaluates a biased client selection strategy, where clients with higher local loss are given preference in training rounds. The effectiveness of this approach is assessed using the CIFAR-100 and Shakespeare datasets under various IID and non-IID scenarios. The code for all the experiments in the project is here [4].

2. Related Work

The Power-of-Choice (PoW-d) strategy [2] introduces an adaptive client selection method that prioritizes clients with higher local loss. POWER-OF-CHOICE is designed to incur minimal communication and computation overhead, enhancing resource efficiency in federated learning.This approach improves convergence speed and communication efficiency while maintaining model performance. Our results demonstrate that PoW-d achieves accuracy comparable to

uniform selection, making it a practical and efficient strategy for real-world FL applications.

3. Methods

3.1. Datasets

The CIFAR-100 dataset is a popular benchmark for evaluating computer vision models. It consists of 60,000 color images, each of size 32×32 pixels, split into 50,000 training images and 10,000 test images. There are 100 fine-grained classes. Each fine-grained class contains 500 training images and 100 test images.

The Shakespeare dataset is designed for the next-character prediction task, presenting a challenge in natural language processing. It contains sentence-character pairs, where each sentence consists of 80 characters, and the predicted character is the 81st in the sequence. The dataset is derived from the complete works of William Shakespeare, encompassing over 4.2 million words. This dataset is provided by [1].

3.2. Models

For CIFAR-100 classification tasks, the EnhancedLeNet architecture, inspired by LeNet cmahan2017communication, was employed. This architecture consists of two convolutional layers, each employing 5x5 kernels with 64 channels, followed by 2x2 max-pooling layers. Batch Normalization is applied after each convolutional layer to stabilize learning and accelerate convergence. Subsequently, two fully connected layers with 384 and 192 channels, respectively, were utilized, culminating in a softmax classifier for final classification. Dropout layers are incorporated after the fully connected layers to mitigate overfitting.

For Shakespearean text analysis, the foundational model utilized was a stacked LSTM structure. The CharLSTM model incorporates an embedding layer that transforms character indices into dense vectors of dimension 8. These embeddings are processed through a stacked LSTM with 2 layers, each comprising 100 hidden units. The LSTM layers sequentially process the input sequences, and the final hidden state from the last time step is extracted. This hidden state undergoes further processing through a fully connected layer that outputs 80 classes, ultimately utilized in a softmax classifier for classification tasks.

3.3. Data Splitting

To evaluate model performance under varying data distribution scenarios, the CIFAR-100 dataset was partitioned into subsets for simulated clients using IID (Independent and Identically Distributed) and non-IID (non-uniform) splitting strategies which are represented in Fig. 1 and Fig. 2 respectively. IID Splitting: The dataset was randomly divided into K equally sized subsets, ensuring that each client receives a uniform and balanced distribution of all class labels. This configuration reflects idealized learning conditions.

Non-IID Splitting: The dataset was sorted by class labels and partitioned to simulate more realistic, heterogeneous client data distributions. Each client was randomly assigned a subset of n labels, from which an equal number of samples per label were drawn.

In the Shakespeare dataset, the data splitting follows a similar logic, with the primary difference being that the non-IID split (its native partitioning) is structured based on individual clients. Specifically, each client is treated as a character from one of Shakespeare's works and is assigned sentences attributed to that character. To create the dataset, 100 clients were randomly selected from a total of 1,128 characters, ensuring that each chosen client received at least 2,000 sentences.

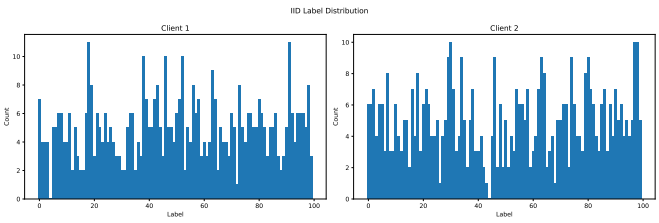

Figure 1. IID Label Distribution

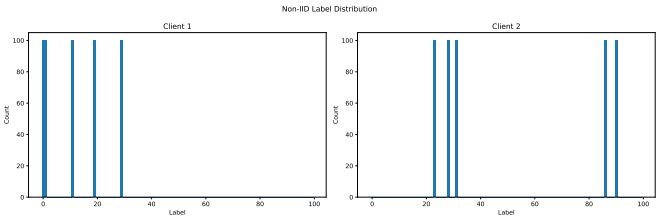

Figure 2. Non-IID Label Distribution

3.4. Client Selection

Uniform Participation: Each client is selected with equal probability:

$$p = \left(\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}\right). \tag{1}$$

Skewed Participation: Selection probabilities p follow a K-dimensional Dirichlet distribution:

$$p \sim \mathcal{D}(\alpha), \quad \text{where } \gamma = \frac{1}{\alpha}. \tag{2}$$

- Smaller γ (larger α) produces near-uniform participation.

- Larger γ (smaller α) increases heterogeneity, favoring frequent participation by a subset of clients.

4. Results

4.1. Centrilized Baseline

The model for the image classification task was trained in 150 epochs using stochastic gradient descent (SGD) with the following hyperparameters. Learning rate $\{\eta : 10^{-2}, 10^{-3}\}$, Weight decay: $\{10^{-4}, 4 \times 10^{-4}\}$

A grid search over hyperparameter combinations revealed the optimal configuration: Learning rate: $\{\eta = 10^{-2}\}$, Weight decay: $\{4 \times 10^{-4}\}$

The following Cosine annealing, Exponential decay and StepLR (step learning rate) learning rate schedulers were evaluated

The cosine annealing scheduler achieved the best performance, obtaining a test accuracy of **57.76%**. The accuracy and loss curves for the training epochs are illustrated in Fig. 3 and Fig. 4, respectively.

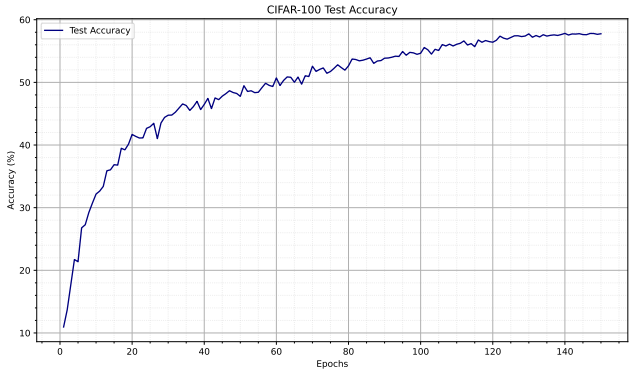

Figure 3. CIFAR-100 Centralized Test Accuracy

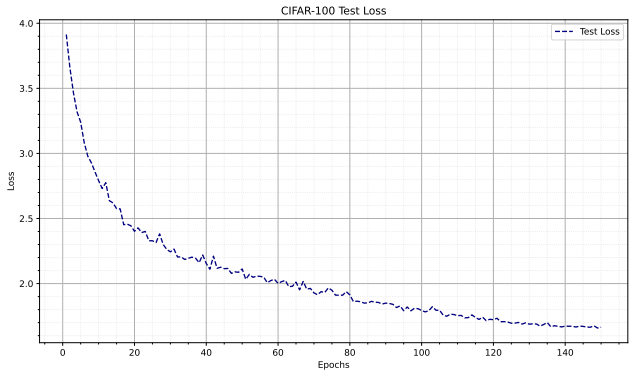

Figure 4. CIFAR-100 Centralized Test Loss

The text processing model was trained for 20 epochs using stochastic gradient descent (SGD) also with the following hyperparameters: Learning rate $\{\eta : 10^{-1}, 10^{-2}, 10^{-3}\}$, Weight decay: $\{10^{-2}, 10^{-3}, 10^{-4}\}$

A grid search over hyperparameter combinations identified the optimal configuration with learning rate: $\eta = 10^{-2}$ and weight decay: 10^{-4}. Using them and cosine annealing scheduler we achieved **57.06%** of test accuracy. The accuracy and loss curves for the training epochs are illustrated in Fig. 5 and Fig. 6, respectively.

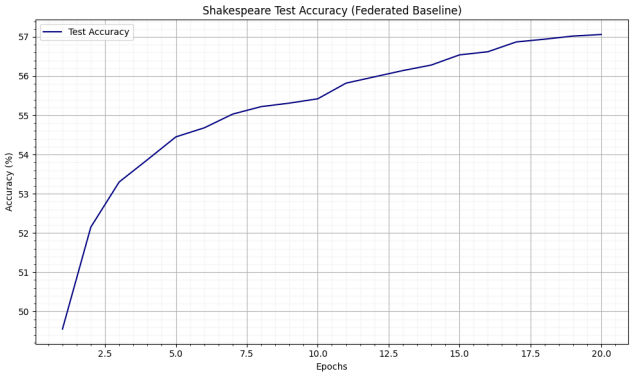

Figure 5. Shakespeare Centralized Test Accuracy

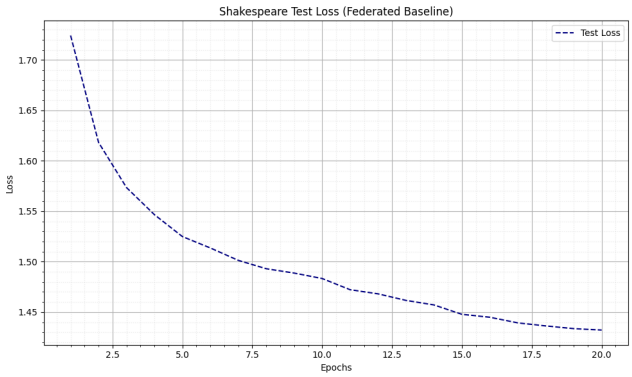

Figure 6. Shakespeare Centralized Test Loss

4.2. Federated Baseline

The federated learning baseline experiments for CIFAR-100 dataset utilized the **Federated Averaging (FedAvg)** [3] algorithm with $K = 100$ clients, each performing $J = 4$ local training epochs per communication round. In each round, a fraction $C = 0.1$ of the clients (10 clients) was randomly selected to participate in the global model update. The model was trained in 1000 global rounds with a fixed learning rate of $\eta = 10^{-2}$. This configuration achieved a final **test accuracy of 56.14%** and a **test loss of 1.6146**, demonstrating stable convergence under heterogeneous client participation. The progression of test accuracy and loss between training rounds are visualized in Fig. 7 and Fig. 8, respectively.

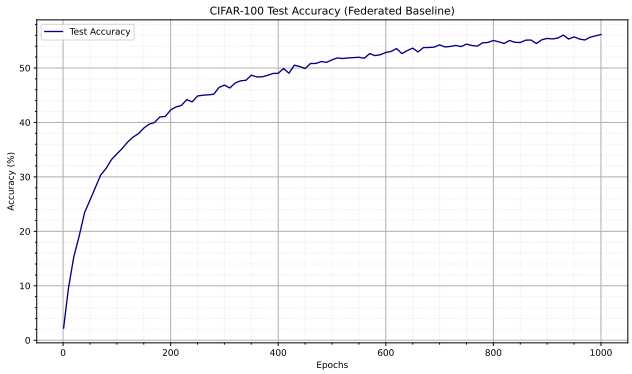

Figure 7. CIFAR-100 Test Accuracy of Federated Baseline

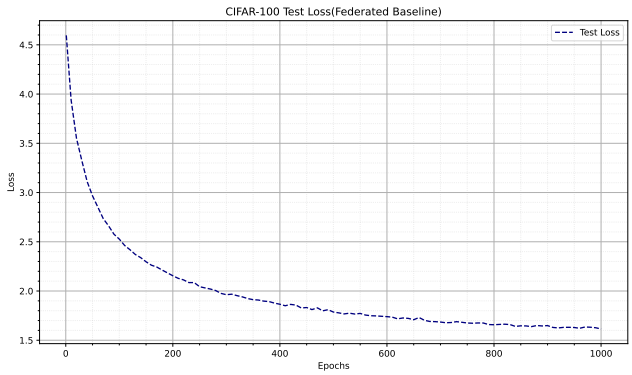

Figure 8. CIFAR-100 Test Loss of Federated Baseline

The same training process, using the same parameters, was applied to the Shakespeare dataset but with a reduced number of 200 global rounds. This configuration achieved a final **test accuracy of 54.97%** and a **test loss of 1.5064**, demonstrating stable convergence despite heterogeneous client participation. The progression of test accuracy and loss across training rounds is illustrated in Fig. 9 and Fig. 10, respectively.

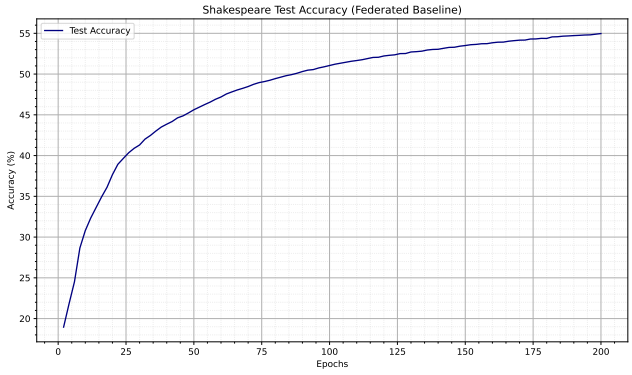

Figure 9. Shakespeare Test Accuracy of Federated Baseline

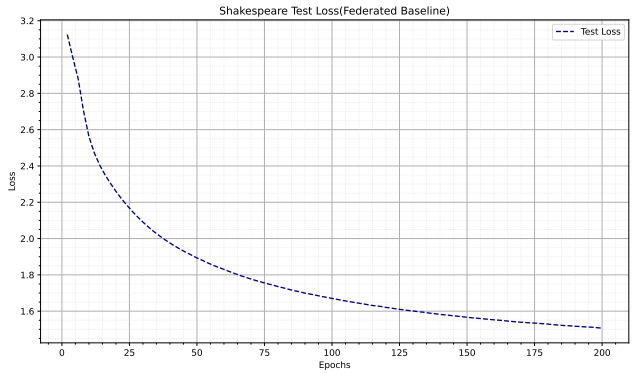

Figure 10. Shakespeare Test Loss of Federated Baseline

4.3. Client Participation

As demonstrated in Figure 11, our comparative evaluation of uniform and skewed client participation reveals critical performance differences. In the initial experiment, we systematically varied the Dirichlet distribution hyperparameter $\gamma \in \{0.1, 1, 3, 10\}$, which governs the degree of heterogeneity of the data between clients. Lower γ values induce stronger non-IID characteristics by amplifying data imbalance, while higher values approximate IID conditions. Uniform client participation consistently outperforms skewed sampling across all γ configurations. The test accuracy exhibits an inverse relationship with γ, with the accuracy decreasing as γ increases.

For the Shakespeare dataset, uniform client participation achieves the highest test accuracy (52.97%), while increasing γ values lead to slight variations. At $\gamma = 0.1$, the precision drops to 51. 12%, and for $\gamma = 10$, it also decreases to 50. 84%. This suggests that while the Shakespeare dataset remains relatively stable under non-IID conditions, uniform participation still results in the best performance. These metrics are demonstrated in the Table 1.

Type	**γ Value**	**CIFAR (%)**	**Shakespeare (%)**
Uniform	–	56.21	52.97
Skewed	0.1	55.03	51.12
Skewed	1	53.29	50.80
Skewed	3	52.12	51.10
Skewed	10	46.32	50.84

Table 1. Test Accuracy Comparison Between Uniform and Skewed Client Participation for Both Datasets

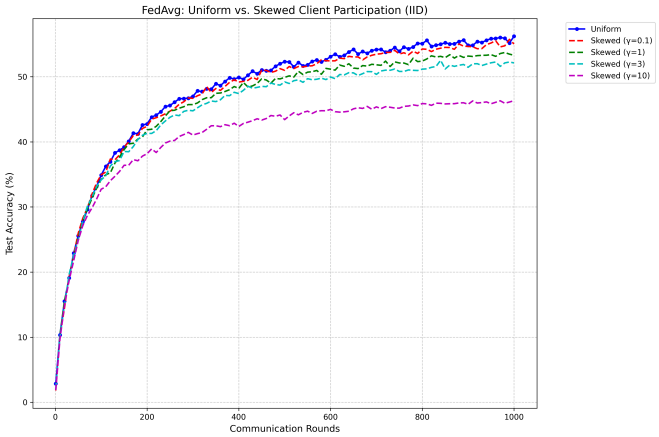

Figure 11. Test Accuracy for client participation schemes on CIFAR-100

4.4. Heterogeneous Distributions

Heterogeneous data distributions were simulated by controlling the number of class labels per client ($N_c \in \{1, 5, 10, 50\}$). We compared model performance against IID baselines while varying local update steps ($J = \{4, 8, 16\}$), with communication rounds adjusted to maintain constant global computation. Test accuracy evolution across configurations is shown in Figures 12–**??**.

The results in Table 2 reveal three principal trends in federated learning under heterogeneous data conditions. First, extreme non-IID configurations ($N_c = 1$) result in catastrophic training failure, with test accuracy stagnating near chance levels ($\sim$1%), while performance improves progressively with increased label diversity ($N_c = 50$ achieves 52.71% at $J = 8$).

Configuration	J=4	J=8	J=16
IID	55.33	**56.28**	55.28
$N_c = 1$	1.13	1.00	1.00
$N_c = 5$	33.20	30.84	25.68
$N_c = 10$	40.97	41.05	37.19
$N_c = 50$	52.30	52.71	51.68

Table 2. Test Accuracy (%) Across Heterogeneous Configurations

Second, the number of local steps (J) exhibits a dual role: optimal IID performance peaks at $J = 8$ (56.28%), but larger J values degrade accuracy across non-IID settings, with $N_c = 10$ showing a 3.86% drop from $J = 8$ to $J = 16$. Finally, moderate heterogeneity ($N_c = 50$) demonstrates a narrow optimization window, where $J = 8$ yields marginal gains (+0.41% over $J = 4$) before declining at $J = 16$, suggesting diminishing returns from increased local computation. These patterns highlight the critical balance required between data heterogeneity and optimization granularity in federated systems.

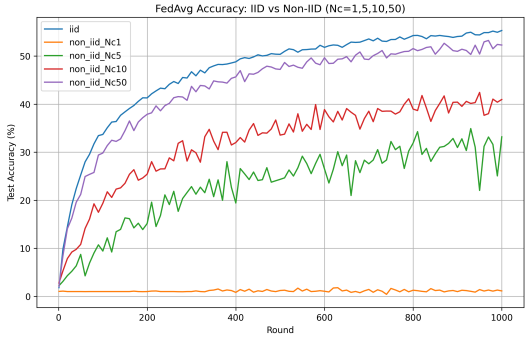

Figure 12. Test Loss for CIFAR-100 J=4

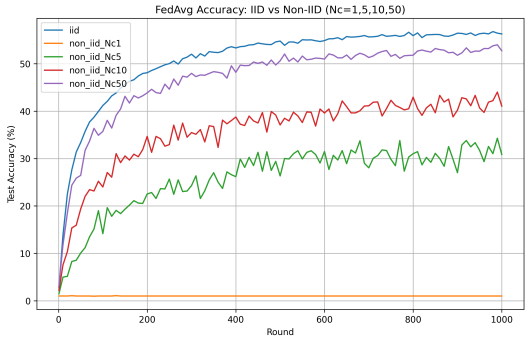

Figure 13. Test Loss for CIFAR-100 J=8

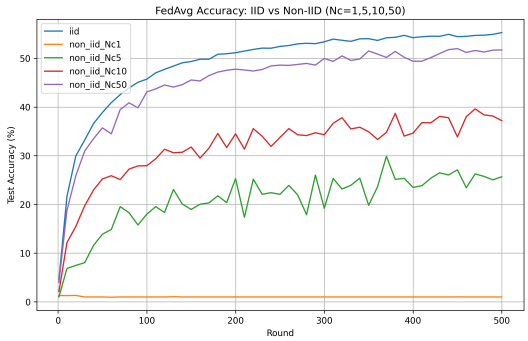

Figure 14. Test Loss for CIFAR-100 J=16

For the task of next-character prediction, both the IID and non-IID splits were already provided. The results of using different values of J in both settings show very similar performance curves. This is because the non-IID split of the Shakespeare dataset is far less extreme than that of CIFAR-100. This is further confirmed by the fact that repeating the experiment on the IID split of Shakespeare produces almost identical results. If the IID split were significantly different

from the non-IID one, a higher performance would have been expected.

4.5. Pow-d

The Federated Averaging (FedAvg) algorithm was evaluated on CIFAR-100 (non-IID) using three client selection strategies: **Uniform** (full participation), **Skewed** ($\gamma = 5$), and **Power-of-Choice (PoW-d,** $d = 3$**)**. The Uniform strategy achieved the highest test accuracy (46.75%) and lowest loss (2.0701), representing an idealized but impractical scenario. The Skewed strategy, with biased client participation, performed poorly (35.92% accuracy, 3.0899 loss), demonstrating the risks of participation imbalance. The proposed PoW-d strategy [2], which prioritizes clients with higher local losses, closely matched Uniform performance shown in Fig. 15 and Fig. 17, (45.90% accuracy, 2.1686 loss), validating its effectiveness in balancing convergence efficiency and practicality under resource constraints.

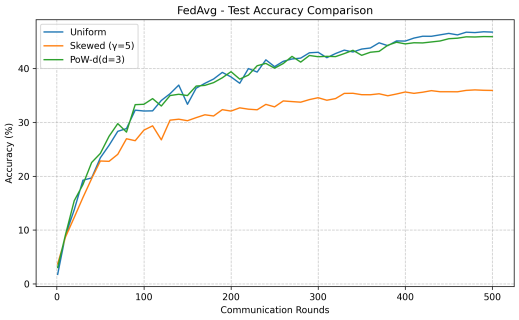

Figure 15. Test accuracy comparison of FedAvg on CIFAR-100 (Non-IID) across Uniform, Skewed ($\gamma = 0.1$), and PoW-d ($d = 3$) client selection strategies over 500 communication rounds

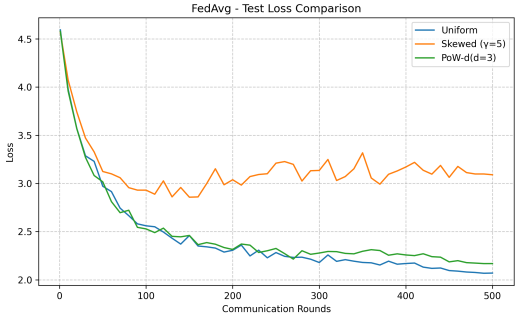

Figure 16. comparison$_{l}oss_{2}0250129_{1}10332.png$

Figure 17. Test loss comparison of FedAvg on CIFAR-100 (Non-IID) across Uniform, Skewed ($\gamma = 0.1$), and PoW-d ($d = 3$) client selection strategies over 500 communication rounds.

Increasing the candidate pool size d from 3 to 5 in the Power-of-Choice (PoW-d) strategy results in marginally reduced accuracy (46.03% vs. 45.90%) but improved loss (2.1502 vs. 2.1686), suggesting a trade-off between exploration (larger d) and exploitation (targeted high-loss clients). While a larger $d = 5$ allows the server to sample more clients and potentially select better candidates, the minimal performance gap from $d = 3$ indicates diminishing returns with increased computational overhead. Both $d = 3$ and $d = 5$ outperform Skewed ($\gamma = 5$) and nearly match Uniform selection, reinforcing that biased selection toward high-loss clients remains effective even with small candidate pools.

5. Conclusion

This study examined the impact of data and system heterogeneity on Federated Learning (FL) and explored biased client selection strategies to improve convergence and model performance. Our results demonstrate that non-IID data distributions and skewed client participation significantly affect FL efficiency. The Power-of-Choice (PoW-d) strategy effectively balances convergence speed and accuracy, closely matching uniform selection while reducing computational overhead. These findings highlight the importance of strategic client selection in FL, paving the way for more efficient and scalable decentralized learning systems.

References

[1] Sebastian Caldas et al. "Leaf: A Benchmark for Federated Settings". In: *Workshop on Federated Learning for Data Privacy and Confidentiality*. 2019. URL: https://github.com/TalwalkarLab/leaf/.

[2] Yae Jee Cho, Jianyu Wang, and Gauri Joshi. "Towards Understanding Biased Client Selection in Federated Learning". In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2022, pp. 10351–10375. URL: https://arxiv.org/pdf/2410.23131.

[3] Brendan McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)* (2017). URL: https://arxiv.org/abs/1602.05629.

[4] Mirkomil. *Federated Learning Project Repository*. 2024. URL: https://github.com/mirkomil579/fl_project/.