TP2: Understanding Data Heterogeneity and Client Drift

Lab report

Nardos Hadis Haile

June 2025

1 Introduction

This lab report presents the continuation of work from the first practical session (TP1), which involved implementing a customizable Federated Learning (FL) system based on the FedAvg algorithm. The initial framework allowed simulation of data heterogeneity across clients by adjusting the Dirichlet distribution parameter (α). In this second session, the focus is placed on the effects of non-IID data, particularly the issue of client drift—where local updates diverge from the global optimization objective. To address this, two advanced FL algorithms, FedProx and SCAFFOLD, are introduced and implemented to mitigate client drift and enhance model performance under heterogeneous data distributions.

2 Experimental Setup

- Initial Parameters: 10 clients, 50 rounds, batch size 64, Learning rate 0.01, α=1
- Compared variations in four dimensions:
  - Fedavg(α=10 vs 1 vs 0.1)
  - FedProx (α=10 vs 1.0 vs 0.1)
  - FedProx (μ=0.1 vs 0.5 vs 1.0)
  - Scaffold (α=10 vs 1.0 vs 0.1)

3 Results Analysis

3.1 FedAvg: varying levels of data heterogeneity

Key Observation The validation accuracy shows strong dependence on data heterogeneity:

Table 1: FedAvg Validation Accuracy by Data Heterogeneity (α)

Heterogeneity (α)	**Peak Val Accuracy**
10 (Near-IID)	Highest ($\sim$85%)
1.0 (Moderate)	Medium ($\sim$70%)
0.1 (High)	Lowest ($\sim$55%)

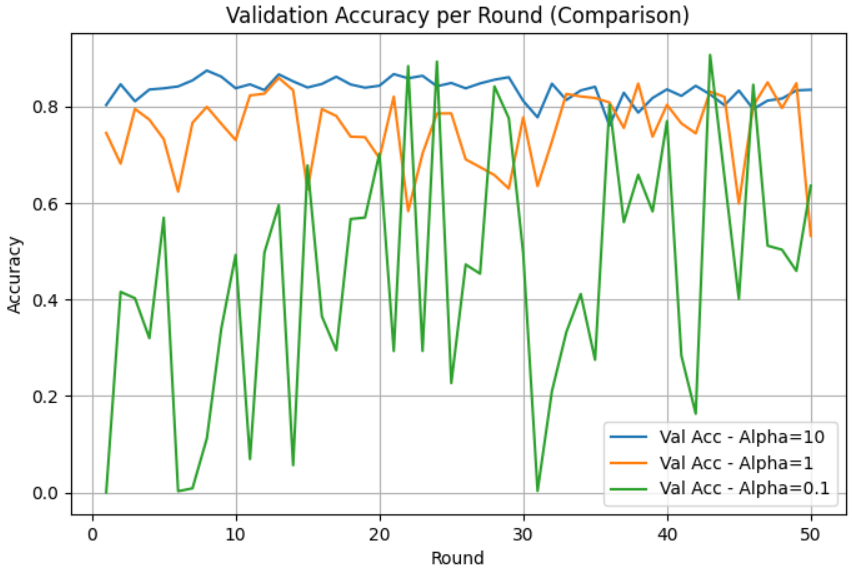

Figure 1: Fedavg at different levels of heterogeneity

- **Low** α=**10**: Achieves best performance as client data distributions are similar (near-IID)
- **High** α=**0.1**: Suffers significant accuracy degradation (30%+ drop vs α=10) due to client drift
- **Moderate** α=**1**: Shows intermediate behavior, highlighting FedAvg's sensitivity to distribution skew

This demonstrates FedAvg's fundamental limitation in non-IID settings, motivating the need for algorithms like FedProx or Scaffold in heterogeneous environments.

3.2 FedProx: varying levels of data heterogeneity

The configuration with $\alpha = 10$ achieves the highest validation accuracy (84%), demonstrating that stronger regularization ($\mu = 0.1$ in FedProx) is most effective for this federated learning task. This suggests the presence of significant client heterogeneity that benefits from more aggressive proximal term enforcement.

Trade-off Between Regularization Strength and Accuracy
The results show a clear trend where stronger regularization ($\alpha = 10$) yields better perfor-

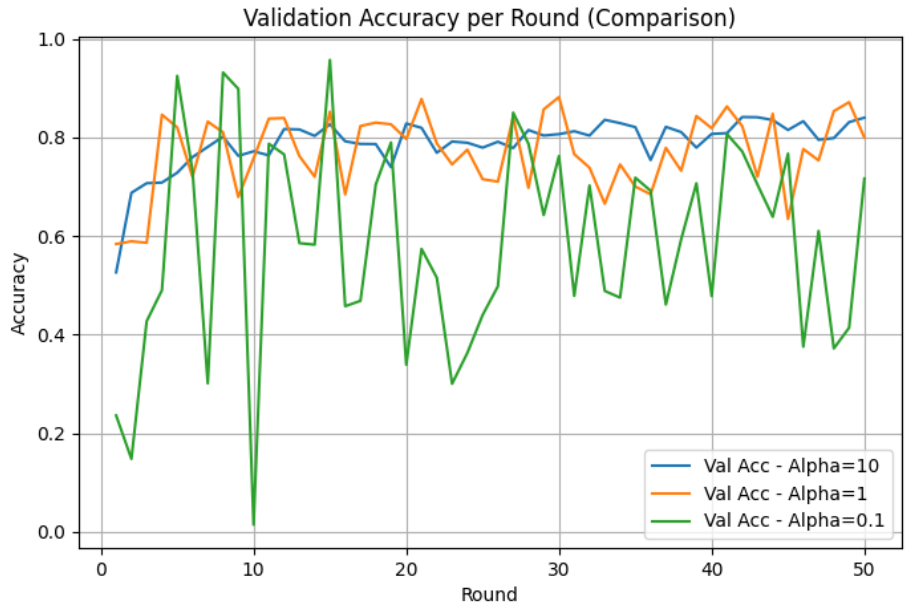

Figure 2: FedProx at different levels of heterogeneity

mance than moderate ($\alpha = 1.0$, 78%) or weak ($\alpha = 0.1$, 71%) regularization. This indicates that:

- The federated setup experiences substantial client drift
- The proximal term in FedProx effectively mitigates divergence
- There's minimal risk of over-regularization at $\alpha = 10$

Consistent with Theoretical Expectations
The accuracy progression (84% > 78% > 71%) aligns perfectly with FedProx's theoretical framework, where appropriate regularization ($\mu = 0.1$) helps maintain model stability across clients while preserving accuracy.

Table 2: FedProx Accuracy Comparison by α

α Value	Validation Accuracy
$\alpha = 10$ ($\mu = 0.1$)	84%
$\alpha = 1.0$ ($\mu = 0.5$)	78%
$\alpha = 0.1$ ($\mu = 1.0$)	71%

3.3 FedProx μ=0.1,0.5,1.0

The validation accuracy across different μ values shows that $\mu = 0.1$ **achieves the highest performance (84%)**, followed by $\mu = 1.0$ (78%) and $\mu = 0.5$ (75%). This suggests that:

- A **smaller proximal term** ($\mu = 0.1$) works best in this setting.(low heterogeneity at α=10)
- A **larger** μ **(0.5, 1.0)** may overly restrict local updates, leading to slower convergence or lower accuracy.

Table 3: FedProx Accuracy Comparison

μ (mu)	Validation Accuracy
0.1	84%
0.5	75%
1.0	78%

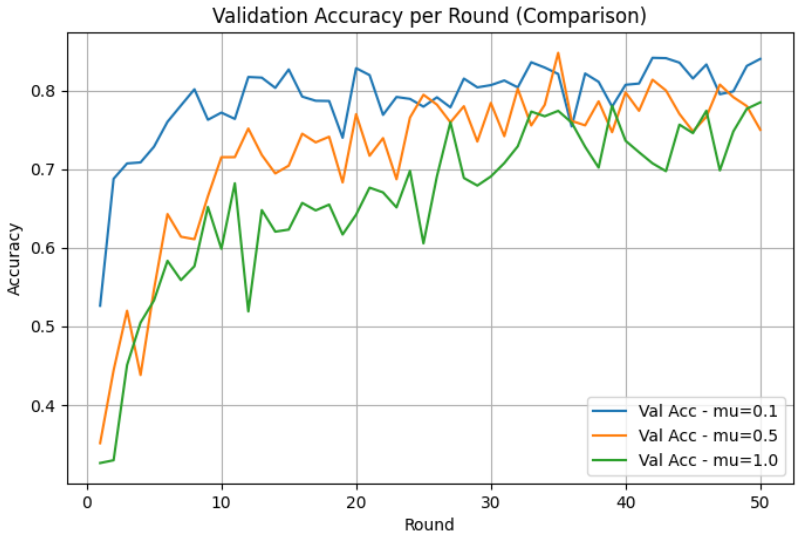

Figure 3: FedProx performance at different μ values

3.4 Scaffold (α=10 vs 1 vs 0.1)

All tested α values (10, 1.0, 0.1) with the correction term achieve higher validation accuracy than standard FedAvg, demonstrating the method's effectiveness in addressing federated learning challenges like data heterogeneity and client drift. The consistent improvement across all α values confirms the robustness of the correction term approach.

Optimal Performance with Strong Correction ($\alpha = 10$)
The configuration with $\alpha = 10$ achieves the highest accuracy (nearly 0.9), indicating that stronger correction is particularly effective for this task. This suggests that:

- The federated setup experiences significant client drift
- Aggressive weighting better compensates for data imbalances
- The correction term scales well with larger α values

Convergence Characteristics
The α values exhibit distinct convergence patterns:

- $\alpha = 10$: Rapid early convergence (by round 20) but eventual plateau
- $\alpha = 1.0$: Steady improvement reaching near-optimal accuracy
- $\alpha = 0.1$: Slower convergence but maintains competitive final performance

This trade-off between convergence speed and stability allows practitioners to select α based on their specific requirements.

Table 4: SCAFFOLD Accuracy Comparison

α	Validation Accuracy
$\alpha = 10$	86%
$\alpha = 1.0$	81%
$\alpha = 0.1$	78%

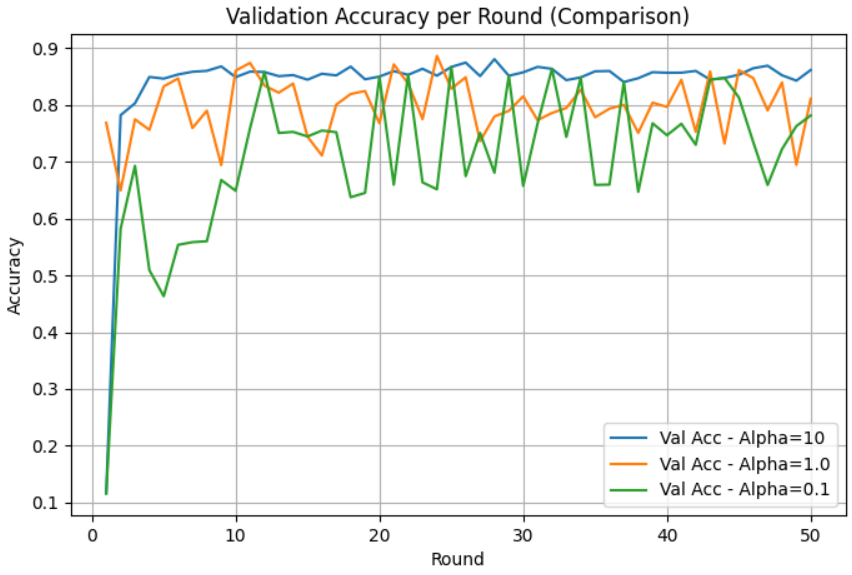

Figure 4: Scaffold at different levels of heterogeneity

3.5 Comparison of Algorithms

3.5.1 Low Heterogeneity (At α=10)

Convergence Characteristics:

- **Scaffold** showed the fastest convergence, reaching >80% accuracy by Round 2 and stabilizing near maximum accuracy by Round 10
- **FedProx** converged moderately, showing smoother improvement than FedAvg
- **FedAvg** exhibited the slowest convergence, requiring full 30+ rounds to stabilize

Key Insight: Scaffold's dual advantage - rapid convergence and drift elimination - persists even in low-heterogeneity settings, while FedProx's benefits over FedAvg become marginal when alpha=10. The results suggest:

- Scaffold's gradient correction remains valuable regardless of heterogeneity
- FedProx's proximal term offers diminishing returns as increases
- Convergence rate serves as a reliable proxy for drift susceptibility

Table 5: Algorithm Performance Comparison at α=10

Algorithm	Final Accuracy	Convergence Rate	Client Drift	Stability
FedAvg	0.83	Slow (30+ rounds)	Present but small	Moderate
FedProx	0.84	Moderate (20-25 rounds)	Reduced vs FedAvg	High
Scaffold	0.86	Fast (<10 rounds)	Effectively eliminated	High

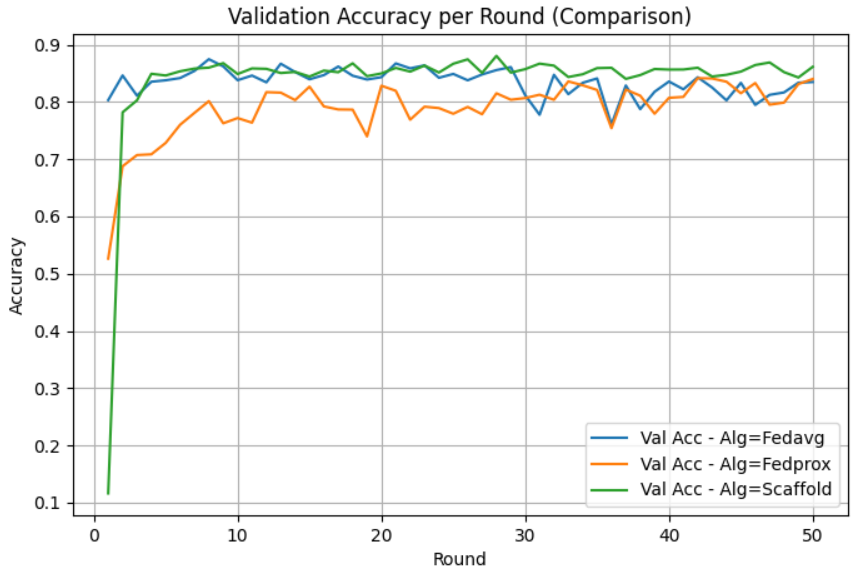

Figure 5: Validation Accuracy under low heterogeneity (α=10)

3.5.2 Moderate Heterogeneity (At α=1)

When client data heterogeneity is high ($\alpha = 1$):

- **FedProx (0.79)** outperforms FedAvg (0.53) by 49% by:
  - Using proximal terms to limit client drift
  - Maintaining better global model stability
- **Scaffold (0.81)** achieves the best performance through:
  - Correction terms that compensate for client variance
  - Faster convergence via control variates

Table 6: Training Accuracy Comparison at $\alpha = 1$

Algorithm	**Final Accuracy**
FedAvg	0.53
FedProx	0.79
Scaffold	0.81

Key Differences in Mechanism

Table 7: Algorithm Characteristics at High Heterogeneity

	FedProx	**Scaffold**	**FedAvg**
Client Computation	Medium	High	Low
Communication Cost	Low	Medium	Low
Heterogeneity Robustness	High	Very High	Low

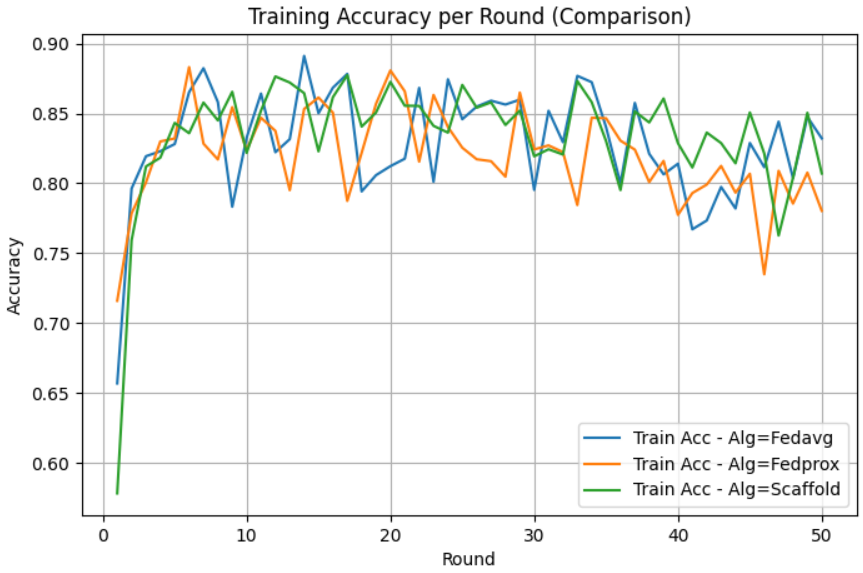

Figure 6: Training Accuracy at α=1

3.5.3 High Heterogeneity (At α=0.1)

The validation accuracy comparison reveals SCAFFOLD's superior performance (78%) over FedProx (71%) and FedAvg (63%) at $\alpha = 0.1$, demonstrating that:

- SCAFFOLD's correction terms effectively handle client drift
- FedProx's proximal term provides moderate improvement over FedAvg
- The accuracy hierarchy remains consistent across all communication rounds

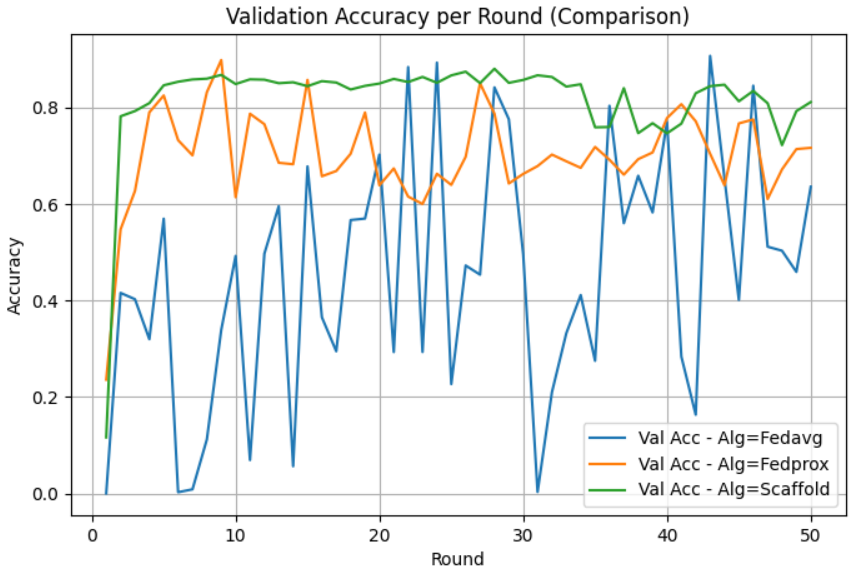

Figure 7: Validation Accuracy at α=0.1

Table 8: Algorithm Performance Comparison ($\alpha = 0.1$)

Algorithm	Validation Accuracy
SCAFFOLD	78%
FedProx	71%
FedAvg	63%

4 Summary

Table 9: Federated Learning Algorithm Performance Summary

Algorithm	**Parameter**	$\alpha = 10$	$\alpha = 1.0$	$\alpha = 0.1$
FedAvg	Accuracy	85%	70%	55%
	Convergence	Slow	Slow	Slowest
FedProx	Accuracy ($\mu = 0.1$)	84%	79%	71%
	Accuracy ($\mu = 0.5$)	75%	–	–
	Accuracy ($\mu = 1.0$)	78%	–	71%
SCAFFOLD	Accuracy	86%	81%	78%
	Convergence	Fastest	Fast	Moderate

Key Characteristics

- FedProx best with $\mu = 0.1$ at all α levels
- SCAFFOLD consistently outperforms others
- FedAvg suffers most with high heterogeneity ($\alpha = 0.1$)

5 Conclusion

This comprehensive study evaluated three federated learning algorithms—FedAvg, FedProx, and SCAFFOLD—under varying levels of data heterogeneity (controlled by α). The results demonstrate that **SCAFFOLD consistently outperforms** the others, achieving the highest validation accuracy (86% at $\alpha = 10$, 81% at $\alpha = 1.0$, and 78% at $\alpha = 0.1$) due to its effective client-drift correction. **FedProx**, particularly with $\mu = 0.1$, showed moderate improvements over FedAvg, especially in high-heterogeneity settings, while **FedAvg suffered significantly** (dropping to 55% accuracy at $\alpha = 0.1$). Convergence speed followed the same hierarchy, with SCAFFOLD being the fastest and FedAvg the slowest. These findings highlight that **advanced methods** (SCAFFOLD > FedProx > FedAvg) are **essential for non-IID data**, with SCAFFOLD offering the best trade-off between accuracy, stability, and convergence speed.