TP3: Understanding Federated learning Attacks and Counteract Schemes - Report

Mahmoud Mayaleh

June 18, 2025

1 Introduction

In this report, I will analyze the robustness of three aggregation schemes: **FedAvg**, **FedMedian**, and **Krum** under varying data heterogeneity (Dirichlet $\alpha \in \{10, 1, 0.1\}$) and a fixed 25% malicious client ratio using data poisoning. We also compare their performance under different attack intensities.

2 Experimental Setup

- **Number of clients:** 10
- **Malicious ratio:** 0%, 25%, 50%
- **Attack type:** Data poisoning and model poisoning
- **Dirichlet α:** 10 (low heterogeneity), 1 (medium), 0.1 (high)
- **Aggregation schemes:** FedAvg, FedMedian, Krum
- **Other hyperparameters:** Fixed (see codebase)

3 Results Overview

3.1 Effect of Malicious Client Ratio (from `TP_attack_results`)

Figures 3–12 show the effect of increasing the percentage of malicious clients (0%, 25%, 50%) on each aggregation scheme under data poisoning.

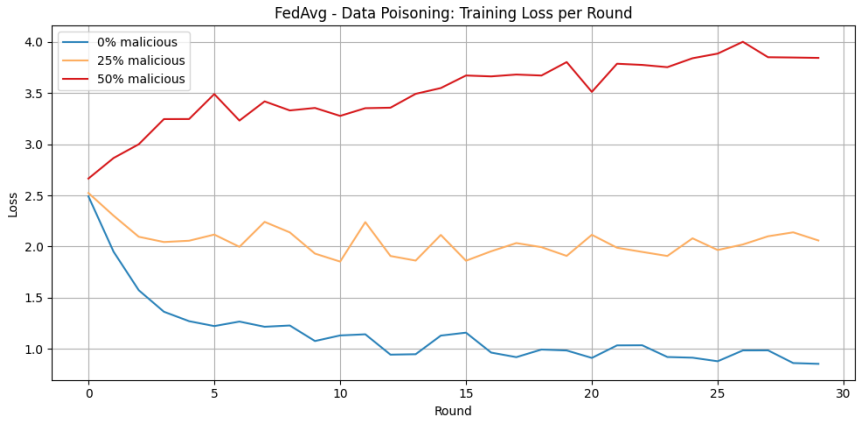

Figure 1: FedAvg: Training loss for different malicious client ratios (Data).

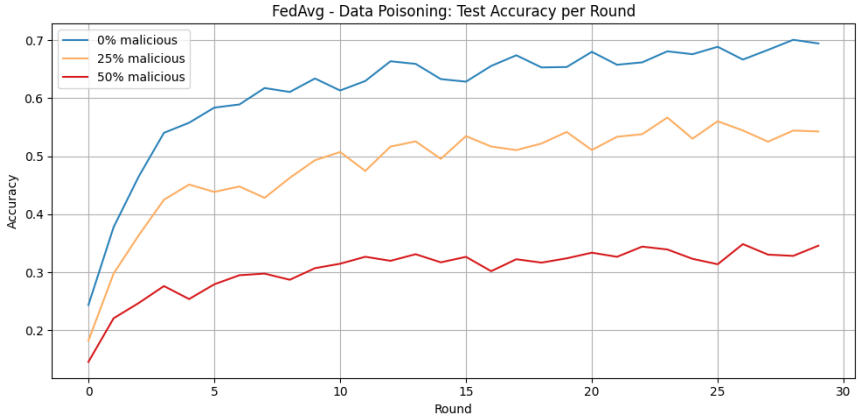

Figure 2: FedAvg: Test accuracy for different malicious client ratios (Data).

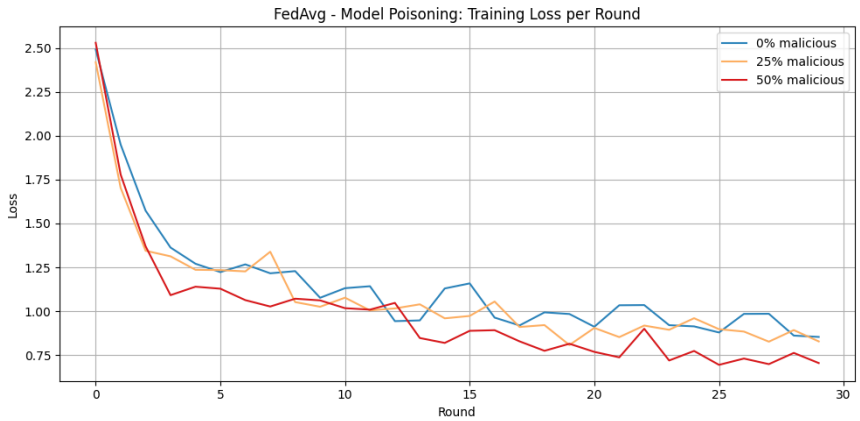

Figure 3: FedAvg: Training loss for different malicious client ratios (Model).

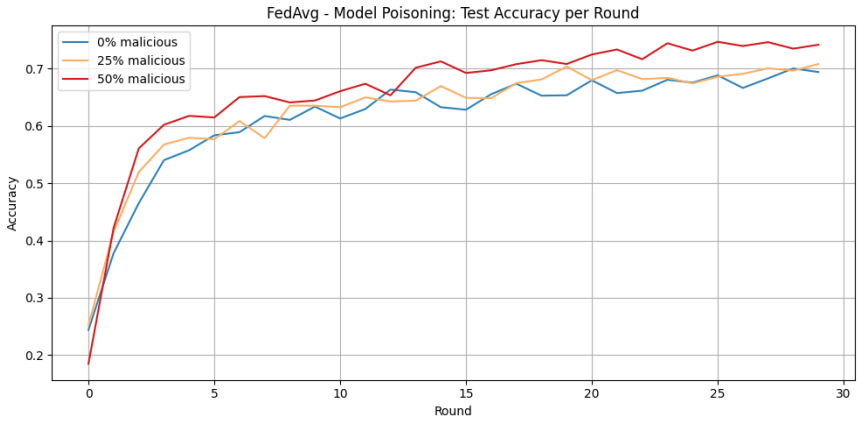

Figure 4: FedAvg: Test accuracy for different malicious client ratios (Model).

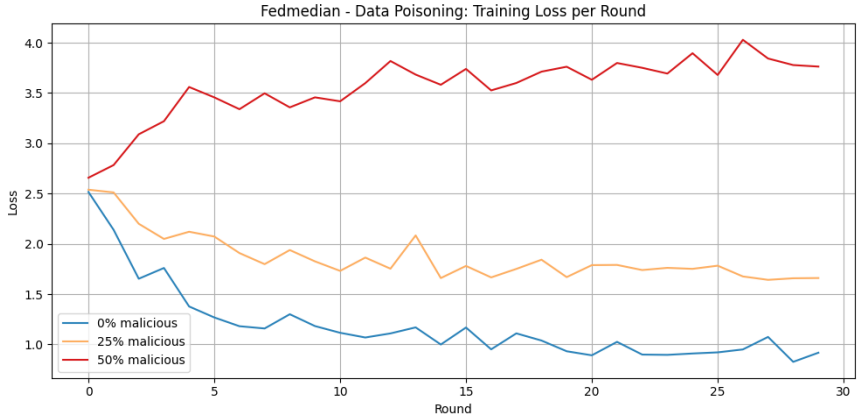

Figure 5: FedMedian: Training loss for different malicious client ratios (Data).

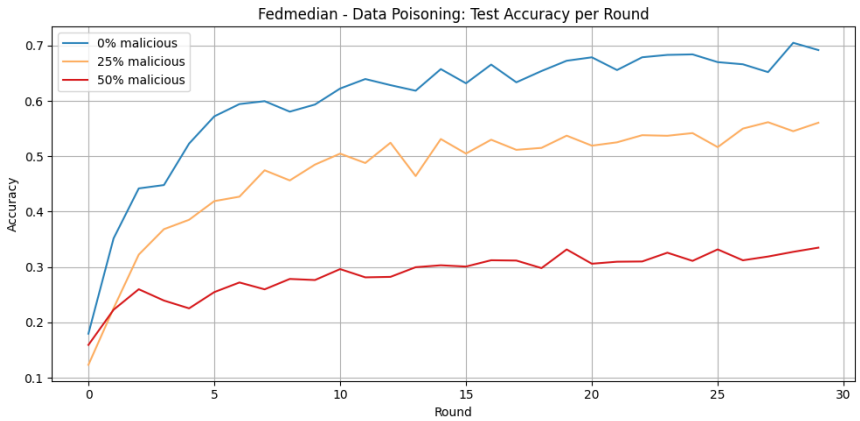

Figure 6: FedMedian: Test accuracy for different malicious client ratios (Data).

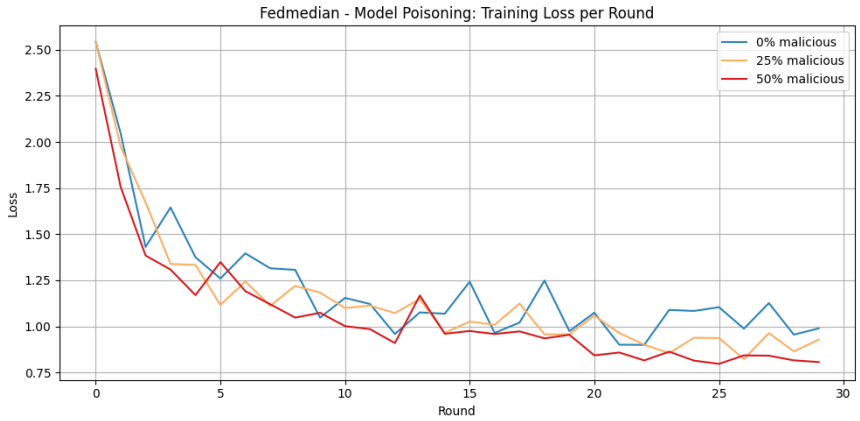

Figure 7: FedMedian: Training loss for different malicious client ratios (Model).

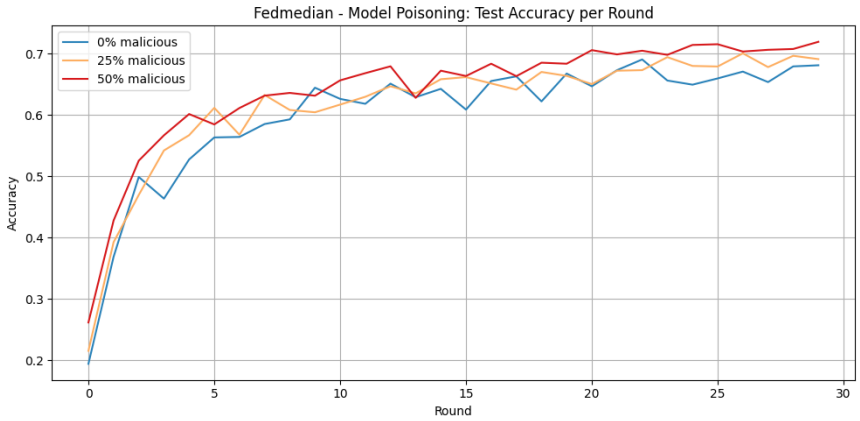

Figure 8: FedMedian: Test accuracy for different malicious client ratios (Model).

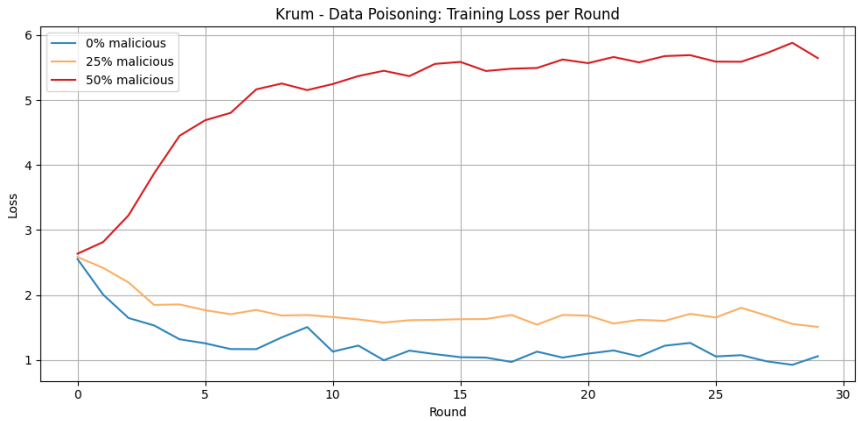

Figure 9: Krum: Training loss for different malicious client ratios (Data).

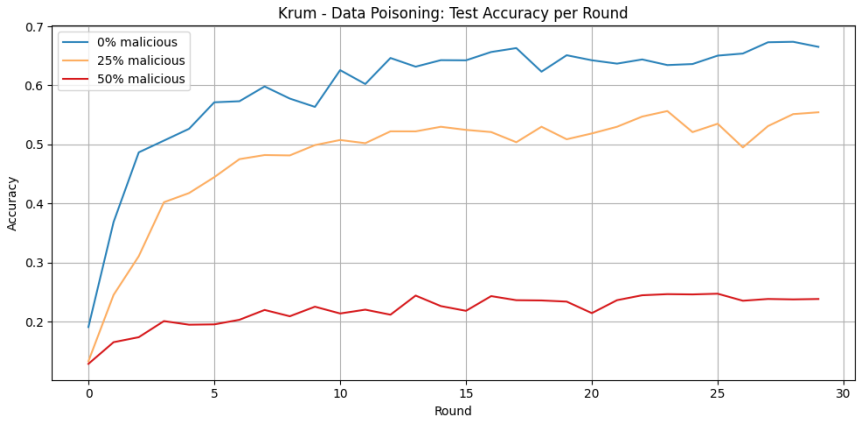

Figure 10: Krum: Test accuracy for different malicious client ratios (Data).

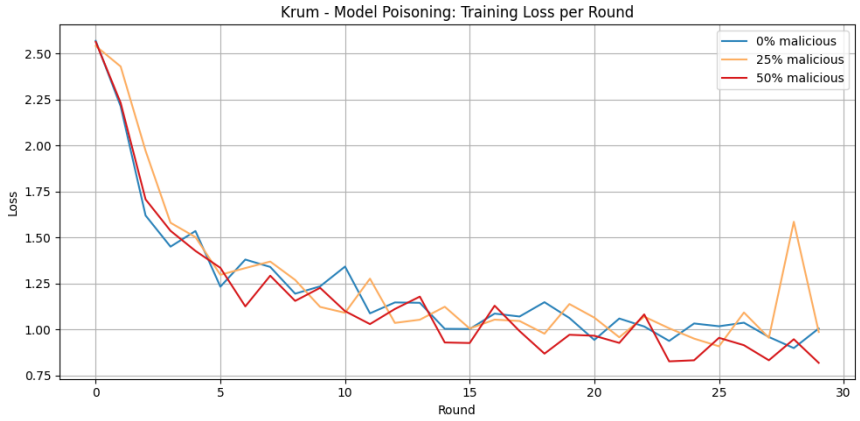

Figure 11: Krum: Training loss for different malicious client ratios (Model).

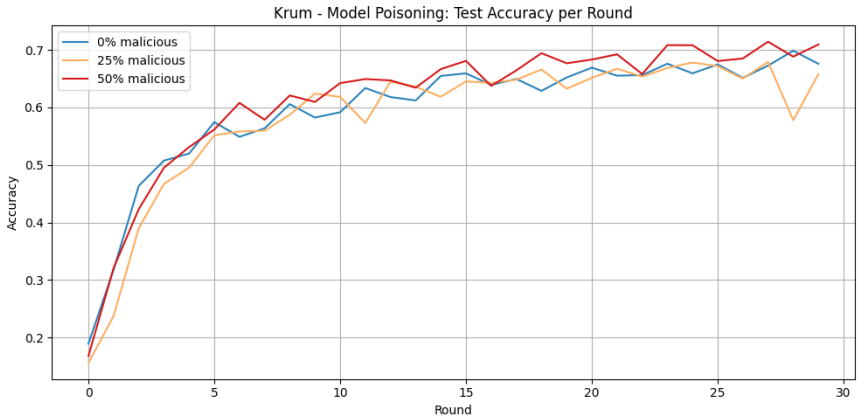

Figure 12: Krum: Test accuracy for different malicious client ratios (Model).

Analysis:

As the malicious client ratio increases, FedAvg's performance degrades rapidly, especially at

50%. FedMedian and Krum are more robust, but even they show reduced accuracy at high attack rates. Krum maintains the highest robustness, but may still be affected if too many clients are malicious.

3.2 Effect of Data Heterogeneity (from `TP3_results`)

Figures 13–18 show the effect of increasing heterogeneity (decreasing α) for each aggregation scheme, with 25% malicious clients.

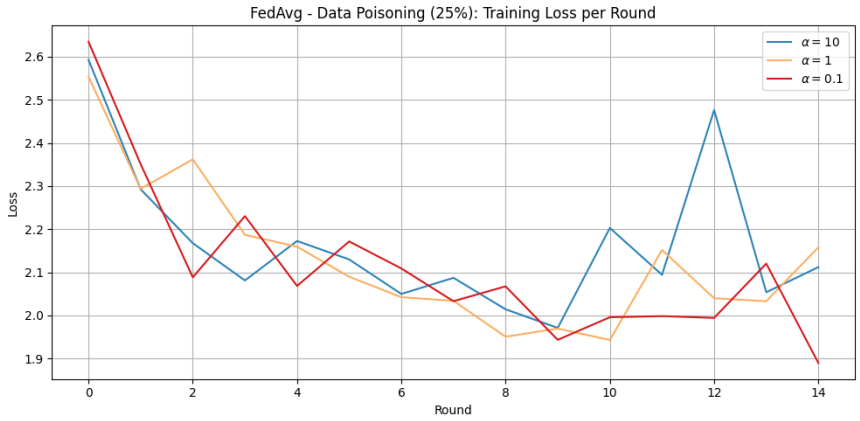

Figure 13: FedAvg: Training loss for $\alpha \in \{10, 1, 0.1\}$ (25% data poisoning).

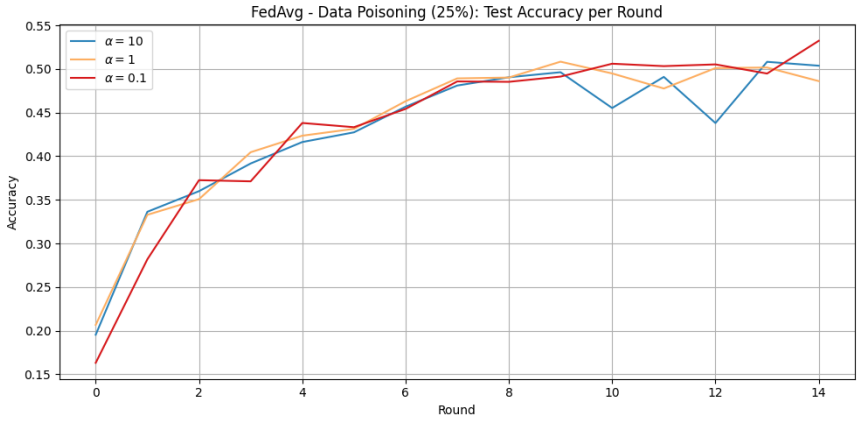

Figure 14: FedAvg: Test accuracy for $\alpha \in \{10, 1, 0.1\}$ (25% data poisoning).

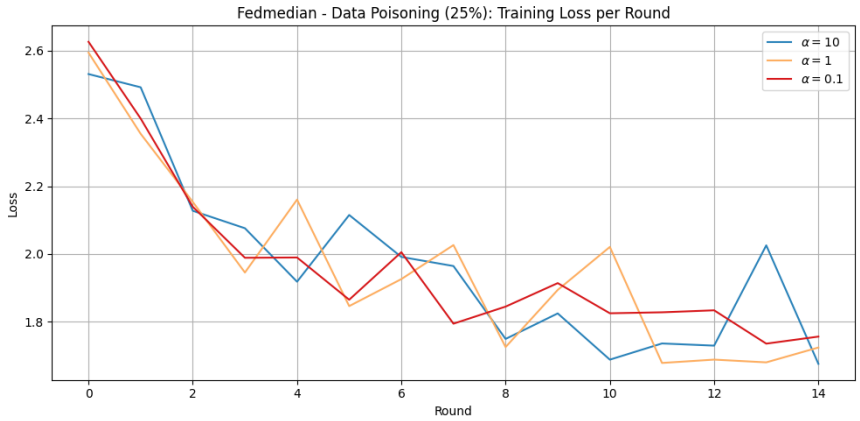

Figure 15: FedMedian: Training loss for $\alpha \in \{10, 1, 0.1\}$ (25% data poisoning).

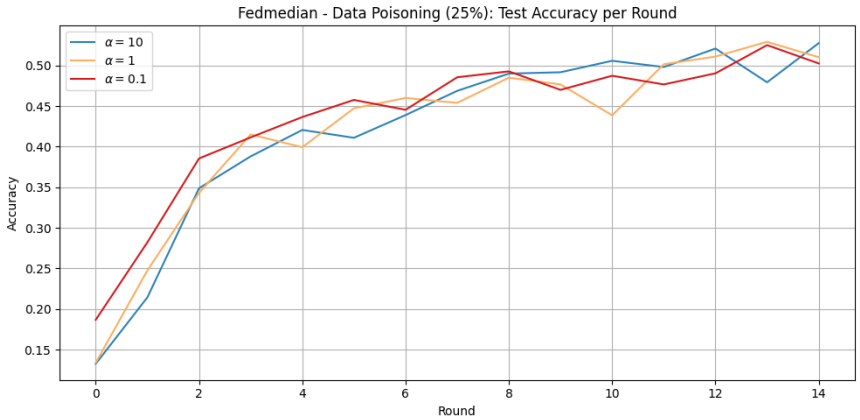

Figure 16: FedMedian: Test accuracy for $\alpha \in \{10, 1, 0.1\}$ (25% data poisoning).

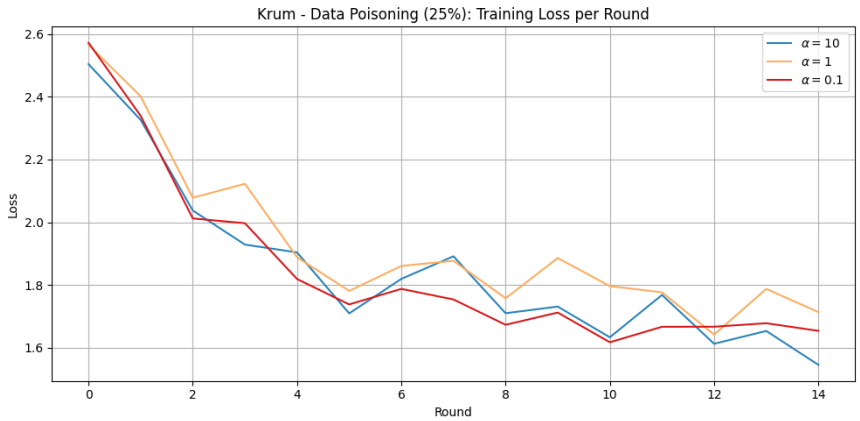

Figure 17: Krum: Training loss for $\alpha \in \{10, 1, 0.1\}$ (25% data poisoning).

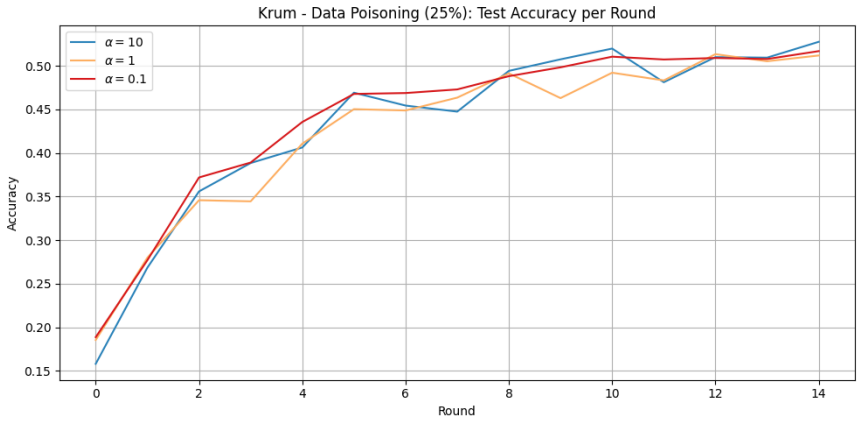

Figure 18: Krum: Test accuracy for $\alpha \in \{10, 1, 0.1\}$ (25% data poisoning).

Analysis:
As heterogeneity increases (lower α), all schemes experience a drop in accuracy and slower convergence. FedAvg is most affected, showing significant degradation under high heterogeneity. FedMedian and Krum are more robust, but their performance also declines as honest clients become more statistically different.

4 Discussion

How does increasing heterogeneity affect the performance of each scheme?

Increasing heterogeneity (lower α) makes the local data distributions more different across clients. This leads to:

- **FedAvg:** Severe performance drop, poor convergence, and high vulnerability to attacks.
- **FedMedian:** More robust than FedAvg, but accuracy drops as honest updates become outliers.
- **Krum:** Most robust, but can mistakenly reject honest clients with outlier updates, especially at high heterogeneity.

Do FedMedian or Krum mistakenly reject honest clients with outlier behavior?

Yes. Both schemes are designed to filter outliers, which helps against malicious updates. However, as heterogeneity increases, honest clients may produce updates that are far from the majority. This can cause:

- **FedMedian:** Honest but outlying updates may be ignored if they are far from the median.

- **Krum:** Honest clients with outlier updates may be scored as "suspicious" and excluded, especially when the number of honest clients is close to the number of malicious ones.

Under which conditions does each scheme work best?

- **FedAvg:** Best under low heterogeneity and no attacks.
- **FedMedian:** Robust to moderate attacks and moderate heterogeneity.
- **Krum:** Most robust to strong attacks, but can exclude honest clients under high heterogeneity or if f is set too high.

5 Conclusion

Robust aggregation schemes like **FedMedian** and **Krum** significantly improve federated learning security under attack, but their effectiveness is challenged by high data heterogeneity. The key challenge is distinguishing malicious updates from honest but statistically different ones. Future work should focus on adaptive or personalized aggregation strategies to address this issue.