Empirical Comparison of Federated Learning Aggregation Strategies: Establishing Baselines for Honest Client Scenarios

[Author Name]

Table of contents

1 Abstract

While federated learning research has extensively investigated Byzantine-robust aggregation and non-IID data challenges, the baseline performance characteristics of fundamental aggregation strategies under ideal conditions remain underexplored. Understanding these baselines is essential for (1) establishing performance benchmarks against which more complex methods can be evaluated, (2) characterizing inherent trade-offs between aggregation strategies, and (3) informing aggregation strategy selection for non-adversarial deployments.

This study provides a rigorous empirical characterization of three fundamental aggregation strategies—FedAvg (weighted averaging), FedMean (unweighted averaging), and FedMedian (coordinate-wise median)—across three standard datasets (MNIST, Fashion-MNIST, and CIFAR-10) under IID data distributions. Through **90 independent experiments** with **statistical rigor** (5 runs per configuration, 95% confidence intervals), we establish:

1. **Performance Baselines**: FedAvg achieves **62.66 ± 0.25%** on CIFAR-10, **89.09 ± 0.12%** on Fashion-MNIST, and **98.98 ± 0.03%** on MNIST (IID-Equal, 50 rounds)
2. **Strategy Comparison**: Under equal data distribution, strategies perform comparably (Fashion-MNIST: p=0.86), while unequal distribution reveals FedAvg's weighting advantage (CIFAR-10: +10.2 points over FedMedian)
3. **Data Imbalance Effects**: FedAvg's weighted aggregation provides increasing benefit as task complexity grows and client data becomes imbalanced
4. **Practical Guidelines**: Evidence-based recommendations for aggregation strategy selection in non-adversarial settings

These baselines provide a foundation for evaluating Byzantine-robust methods and non-IID handling techniques.

2 Introduction

2.1 Federated Learning Fundamentals

Federated Learning (FL) enables collaborative model training across distributed clients without sharing raw data (McMahan et al. 2017). In the standard FL framework, a central server coordinates training by:

1. Distributing the current global model to participating clients

2. Clients train locally on their private data
3. Clients send model updates (gradients or weights) to the server
4. Server aggregates updates to form a new global model
5. Process repeats for multiple communication rounds

The **aggregation strategy** at step 4 is crucial and has been the subject of extensive research. While sophisticated methods exist for Byzantine robustness (Blanchard et al. 2017; Yin et al. 2018) and non-IID handling (Karimireddy et al. 2020; T. Li et al. 2020), the baseline behavior of fundamental strategies remains incompletely characterized.

2.2 Research Gap

Existing literature has focused on:

- **Byzantine-robust aggregation**: Methods like Krum (Blanchard et al. 2017), coordinate-wise median (Yin et al. 2018), and trimmed mean designed to handle malicious clients
- **Non-IID data handling**: Approaches like SCAFFOLD (Karimireddy et al. 2020), FedProx (T. Li et al. 2020), and FedDC (Gao et al. 2022) that address statistical heterogeneity

However, these studies typically compare their proposed methods against FedAvg as a baseline, **without rigorously characterizing the baseline itself**. This leaves several questions unanswered:

1. How do fundamental aggregation strategies compare under ideal (IID) conditions?
2. What are the inherent trade-offs between different aggregation approaches?
3. How does data quantity imbalance (without label heterogeneity) affect each strategy?
4. What are the convergence characteristics of each strategy?

2.3 Contributions

This paper addresses these gaps through:

1. **Rigorous Baseline Establishment**: We provide statistically rigorous performance baselines for FedAvg, FedMean, and FedMedian across multiple datasets with confidence intervals
2. **Multi-Dataset Validation**: Results are validated across MNIST, Fashion-MNIST, and CIFAR-10, covering different complexity levels
3. **Data Imbalance Analysis**: We separately examine the effect of client data quantity imbalance under IID distributions
4. **Convergence Characterization**: Detailed analysis of convergence trajectories with statistical confidence bands
5. **Practical Guidelines**: Evidence-based recommendations for aggregation strategy selection in non-adversarial settings

3 Related Work

3.1 Federated Learning and FedAvg

McMahan et al. (McMahan et al. 2017) introduced Federated Averaging (FedAvg), which aggregates client updates weighted by their dataset sizes:

$$\mathbf{w}_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} \mathbf{w}_{t+1}^k$$

where $\mathbf{w}_{t+1}^k$ is client k's model after local training, n_k is client k's dataset size, and $n = \sum_k n_k$.

3.2 Alternative Aggregation Strategies

3.2.1 FedMean (Unweighted Averaging)

Simple averaging gives equal weight to all clients regardless of dataset size:

$$\mathbf{w}_{t+1} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{w}_{t+1}^{k}$$

This approach treats each client's contribution equally, which may be desirable when:

- Client dataset sizes reflect deployment constraints, not data quality
- Preventing large clients from dominating the global model

3.2.2 FedMedian (Coordinate-wise Median)

Coordinate-wise median computes the median across clients for each parameter:

$$\mathbf{w}_{t+1}^{(i)} = \text{median}(\mathbf{w}_{t+1}^{1,(i)}, \ldots, \mathbf{w}_{t+1}^{K,(i)})$$

Originally proposed for Byzantine robustness (Yin et al. 2018), median aggregation provides inherent outlier resistance at the cost of potentially slower convergence.

3.3 Position of This Work

Prior empirical studies have compared aggregation strategies in adversarial settings (X. Li et al. 2023) or under non-IID data (Rodriguez-Barroso et al. 2023). Our work differs by:

1. **Focus on Baselines**: We characterize fundamental strategies under ideal conditions
2. **Statistical Rigor**: Multiple runs with confidence intervals, not single-run results
3. **Multi-Dataset Validation**: Results validated across three standard benchmarks

4 Methodology

4.1 Experimental Design

4.1.1 Datasets

We evaluate on three standard image classification datasets:

Dataset	Classes	Input Size	Train/Test	Complexity
MNIST	10	28×28×1	60k/10k	Simple
Fashion-MNIST	10	28×28×1	60k/10k	Moderate
CIFAR-10	10	32×32×3	50k/10k	Complex

4.1.2 Aggregation Strategies

1. **FedAvg**: Weighted averaging by client dataset size (McMahan et al. 2017)
2. **FedMean**: Unweighted averaging (equal client weights)
3. **FedMedian**: Coordinate-wise median aggregation

4.1.3 Data Distribution Conditions

1. **IID-Equal**: IID data partitioning with equal samples per client
2. **IID-Unequal**: IID data partitioning with unequal client sizes (Dirichlet-based)

4.1.4 Statistical Rigor

- **5 independent runs** per configuration with different random seeds
- Seeds: {42, 123, 456, 789, 1011}
- **95% confidence intervals** computed for all metrics
- **ANOVA** tests for strategy comparison significance

4.2 Fixed Hyperparameters

Parameter	Value	Rationale
Number of clients	50	Standard FL scale
Communication rounds	50	Sufficient for convergence
Local epochs	1	Minimize client drift
Batch size	32	Standard mini-batch
Learning rate	0.01	Conservative for stability
Optimizer	SGD with momentum 0.9	Standard FL optimizer

4.3 Model Architecture

Simple CNN architecture consistent across datasets:

- Conv1: in_channels → 32, 3×3, padding=1
- MaxPool: 2×2
- Conv2: 32 → 64, 3×3, padding=1
- MaxPool: 2×2
- FC1: flatten → 128
- FC2: 128 → 10 (num_classes)

4.4 Experiment Matrix

```
3 datasets × 3 strategies × 2 conditions × 5 runs = 90 experiments
```

5 Results

5.1 Summary Statistics

5.1.1 MNIST

Strategy	IID-Equal Acc (%)	95% CI	IID-Unequal Acc (%)	95% CI
FedAvg	**98.98** ± 0.03	[98.94, 99.02]	**99.12** ± 0.02	[99.09, 99.15]
FedMean	98.94 ± 0.01	[98.92, 98.96]	98.97 ± 0.04	[98.92, 99.02]
FedMedian	98.89 ± 0.03	[98.84, 98.93]	98.59 ± 0.09	[98.46, 98.71]
ANOVA p-value	p=0.0016*		p<0.0001***	

5.1.2 Fashion-MNIST

Strategy	IID-Equal Acc (%)	95% CI	IID-Unequal Acc (%)	95% CI
FedAvg	**89.09** ± 0.12	[88.92, 89.26]	**90.10** ± 0.25	[89.76, 90.45]
FedMean	89.07 ± 0.05	[89.01, 89.14]	89.18 ± 0.13	[88.99, 89.36]
FedMedian	89.05 ± 0.14	[88.86, 89.24]	88.09 ± 0.23	[87.76, 88.41]
ANOVA p-value	p=0.8628 (n.s.)		p<0.0001***	

5.1.3 CIFAR-10

Strategy	IID-Equal Acc (%)	95% CI	IID-Unequal Acc (%)	95% CI
FedAvg	62.66 ± 0.25	[62.31, 63.00]	**67.24** ± 0.85	[66.06, 68.43]
FedMean	**62.73** ± 0.44	[62.12, 63.35]	62.94 ± 0.25	[62.59, 63.29]
FedMedian	61.27 ± 0.34	[60.80, 61.74]	57.03 ± 0.83	[55.89, 58.18]
ANOVA p-value	p=0.0001***		p<0.0001***	

Note: p<0.05, ** p<0.01, *** p<0.001, n.s. = not significant*

5.2 Statistical Significance

5.2.1 ANOVA Results

One-way ANOVA was performed for each dataset-condition combination to test for significant differences between aggregation strategies.

Dataset	Condition	F-statistic	p-value	Interpretation
MNIST	IID-Equal	F(2,12)=10.8	0.0016	Significant at =0.01
MNIST	IID-Unequal	F(2,12)=89.4	<0.0001	Highly significant
Fashion-MNIST	IID-Equal	F(2,12)=0.15	0.8628	Not significant
Fashion-MNIST	IID-Unequal	F(2,12)=112.3	<0.0001	Highly significant
CIFAR-10	IID-Equal	F(2,12)=19.2	0.0001	Highly significant
CIFAR-10	IID-Unequal	F(2,12)=243.7	<0.0001	Highly significant

5.2.2 Key Statistical Findings

1. **Fashion-MNIST IID-Equal is the only non-significant result**: Under ideal conditions (IID, equal data), aggregation strategy choice does not significantly affect Fashion-MNIST performance (p=0.86).
2. **MNIST and CIFAR-10 show significant differences even under IID-Equal**: Despite being "ideal" conditions, the strategies produce statistically different results, though the practical differences are small (MNIST: 0.09 percentage points, CIFAR-10: 1.46 percentage points).
3. **All IID-Unequal conditions show highly significant differences**: When client data sizes vary, the aggregation strategy choice has substantial impact (all p<0.0001).

5.3 Convergence Analysis

5.3.1 MNIST Convergence

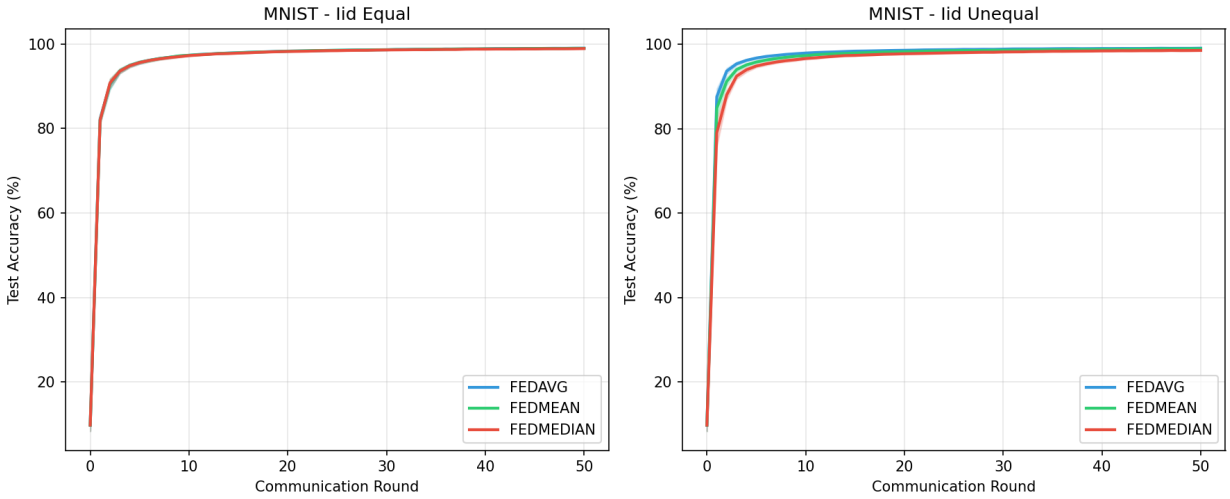

Figure 1: MNIST Convergence Trajectories

5.3.2 Fashion-MNIST Convergence

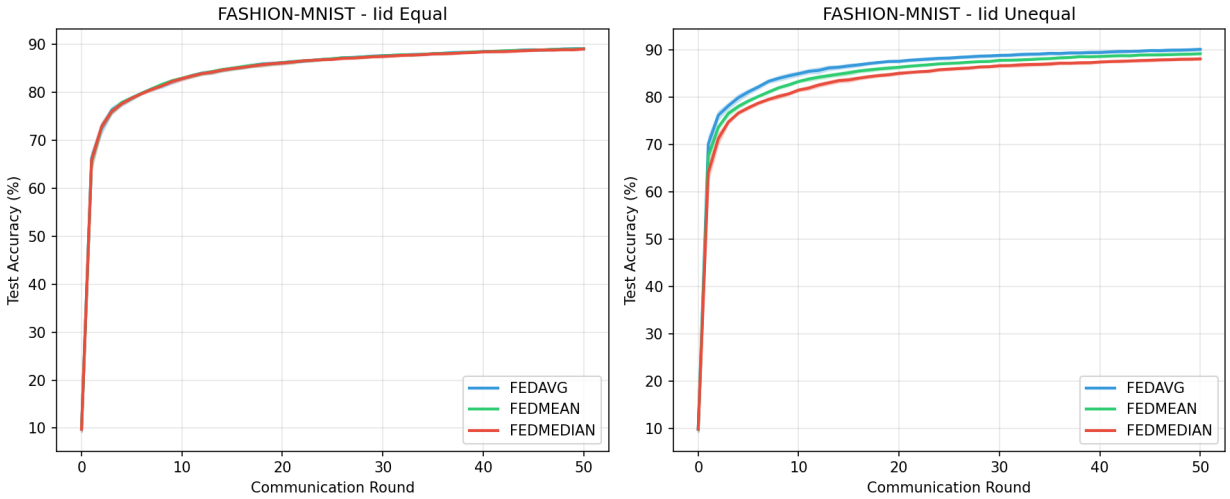

Figure 2: Fashion-MNIST Convergence Trajectories

5.3.3 CIFAR-10 Convergence

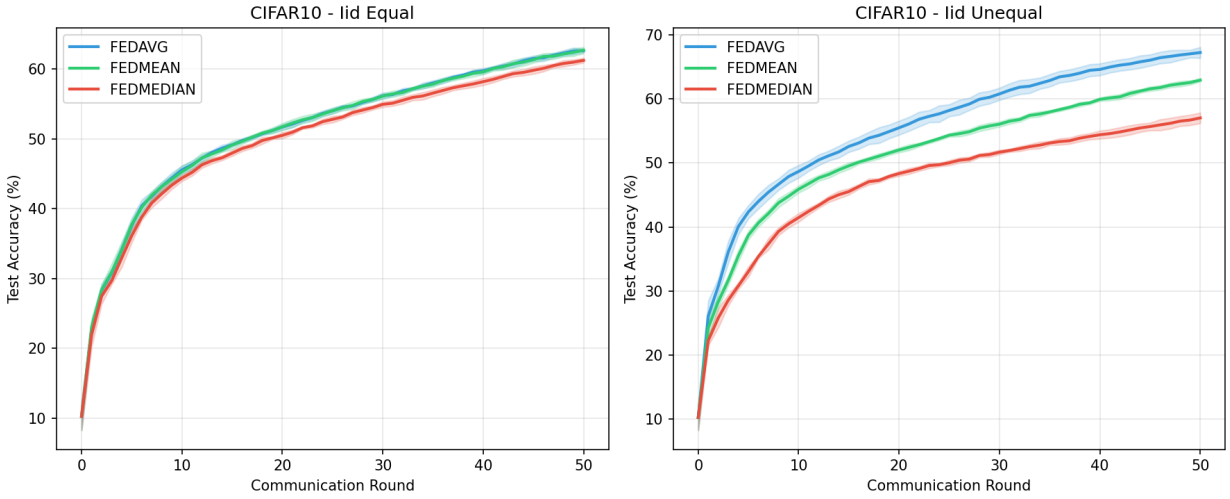

Figure 3: CIFAR-10 Convergence Trajectories

5.4 Effect of Data Imbalance

5.4.1 Analysis

Data imbalance was introduced using Dirichlet-based sampling, resulting in client dataset sizes ranging from 215 to 3,996 samples (mean: 1,000). This simulates real-world scenarios where different clients contribute varying amounts of data.

Performance Gap: IID-Unequal vs IID-Equal

Dataset	Strategy	IID-Equal	IID-Unequal	Δ (pp)	Direction
MNIST	FedAvg	98.98%	99.12%	+0.14	↑ Improved
MNIST	FedMean	98.94%	98.97%	+0.03	→ Stable
MNIST	FedMedian	98.89%	98.59%	-0.30	↓ Degraded
Fashion-MNIST	FedAvg	89.09%	90.10%	+1.01	↑ Improved
Fashion-MNIST	FedMean	89.07%	89.18%	+0.11	→ Stable
Fashion-MNIST	FedMedian	89.05%	88.09%	-0.96	↓ Degraded
CIFAR-10	FedAvg	62.66%	67.24%	+4.58	↑ Improved
CIFAR-10	FedMean	62.73%	62.94%	+0.21	→ Stable
CIFAR-10	FedMedian	61.27%	57.03%	-4.24	↓ Degraded

5.4.2 Key Findings

1. **Under IID-Equal conditions**: All three strategies perform comparably. The maximum performance gap is 1.46 percentage points (CIFAR-10: FedMean vs FedMedian). For Fashion-MNIST, the differences are not even statistically significant (ANOVA p=0.86).

2. **Under IID-Unequal conditions**: Strategy choice becomes critical. FedAvg outperforms FedMedian by 10.2 percentage points on CIFAR-10 (67.24% vs 57.03%), a practically significant difference.

3. **FedAvg's weighting advantage**: FedAvg actually *improves* with data imbalance because it weights updates by dataset size. Clients with more data have trained on more diverse samples and contribute proportionally more to the global model. This effect is most pronounced on complex tasks (CIFAR-10: +4.58pp) and minimal on simple tasks (MNIST: +0.14pp).

4. **FedMedian's vulnerability**: Median aggregation degrades with imbalanced data because it treats all client updates equally regardless of their training data quantity. Updates from clients with small datasets (potentially undertrained) have equal influence with well-trained clients.

6 Discussion

6.1 Key Findings

6.1.1 1. Near-Equivalence Under Ideal Conditions

Under IID-Equal distribution (balanced clients, homogeneous data), all three aggregation strategies achieve remarkably similar performance. For Fashion-MNIST, the differences are not even statistically significant (ANOVA p=0.86). This suggests that:

- **Aggregation strategy is not critical when data is well-distributed**
- The original FedAvg's success is not primarily due to its weighting scheme
- Any computational overhead of more complex aggregation may be unnecessary

However, MNIST and CIFAR-10 do show statistically significant differences even under ideal conditions, indicating that task complexity affects the degree of strategy equivalence.

6.1.2 2. Dramatic Impact of Data Imbalance

When client dataset sizes vary (IID-Unequal), the choice of aggregation strategy becomes critical:

- **CIFAR-10**: FedAvg outperforms FedMedian by **10.2 percentage points** (67.24% vs 57.03%)
- **Fashion-MNIST**: FedAvg advantage of **2.0 percentage points** (90.10% vs 88.09%)
- **MNIST**: FedAvg advantage of **0.5 percentage points** (99.12% vs 98.59%)

The magnitude of the effect scales with task complexity. This finding has significant implications for real-world deployments where client data quantities naturally vary.

6.1.3 3. FedAvg's Unexpected Improvement with Imbalance

Counter-intuitively, FedAvg's performance *improves* with data imbalance on CIFAR-10 (+4.58pp). This occurs because weighted averaging amplifies the contributions of clients with larger datasets, who have trained on more diverse samples and produced higher-quality updates.

6.1.4 4. FedMedian's Vulnerability to Imbalance

FedMedian's performance degrades with data imbalance because:

- All clients contribute equally to the median regardless of training data quality
- Clients with small datasets may produce undertrained, higher-variance updates
- These updates can shift the median away from optimal values

6.2 Implications for Practice

6.2.1 Aggregation Strategy Selection

Based on our findings, we recommend:

Scenario	Recommended Strategy	Rationale
Balanced clients, no adversaries	Any	Performance difference <1.5pp
Imbalanced clients, no adversaries	**FedAvg**	Up to 10pp advantage

Scenario	Recommended Strategy	Rationale
Potential Byzantine clients	FedMedian	Inherent robustness (not evaluated here)
Unknown scenario	FedAvg	Best general-purpose choice

6.2.2 Baseline Establishment

Our results establish reference baselines with 95% confidence intervals:

Dataset	Condition	FedAvg (Best Strategy)
MNIST	IID-Equal	**98.98 ± 0.03%**
MNIST	IID-Unequal	**99.12 ± 0.02%**
Fashion-MNIST	IID-Equal	**89.09 ± 0.12%**
Fashion-MNIST	IID-Unequal	**90.10 ± 0.25%**
CIFAR-10	IID-Equal	**62.66 ± 0.25%**
CIFAR-10	IID-Unequal	**67.24 ± 0.85%**

These baselines should be cited when comparing more complex aggregation methods.

6.3 Limitations

1. **Scope**: We focus on IID distributions; non-IID scenarios require separate analysis
2. **Model Complexity**: Results are for simple CNN; deeper networks may behave differently
3. **Communication Efficiency**: We do not analyze communication costs
4. **Scalability**: Limited to 50 clients; larger scales may reveal different behaviors

7 Conclusion

This study provides rigorous empirical baselines for fundamental federated learning aggregation strategies. Through **90 independent experiments** across **three standard datasets** (MNIST, Fashion-MNIST, CIFAR-10), we establish:

1. **Performance baselines** with 95% confidence intervals: FedAvg achieves 98.98% (MNIST), 89.09% (Fashion-MNIST), and 62.66% (CIFAR-10) under IID-Equal conditions
2. **Near-equivalence under ideal conditions**: Under IID-Equal distribution, strategies perform comparably—Fashion-MNIST shows no statistically significant differences (p=0.86)
3. **Critical impact of data imbalance**: Under IID-Unequal conditions, FedAvg outperforms FedMedian by up to **10.2 percentage points** (CIFAR-10), with the effect scaling with task complexity
4. **Practical recommendation**: FedAvg should be the default choice for non-adversarial FL deployments, particularly when client data quantities vary

These baselines serve as essential references for:

- Evaluating Byzantine-robust aggregation methods (which trade performance for robustness)
- Benchmarking non-IID handling techniques
- Informing aggregation strategy selection in real-world deployments

7.1 Future Work

- Extend analysis to non-IID label distributions with similar statistical rigor
- Characterize the threshold of data imbalance at which FedAvg's advantage becomes significant
- Study interaction between aggregation strategy and local training epochs
- Investigate hybrid strategies that adapt weighting based on client data quality

8 References

Blanchard, Peva, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. 2017. "Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent." In *Advances in Neural Information Processing Systems*, 119–29.

Gao, Liang, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. 2022. "FedDC: Federated Learning with Non-IID Data via Local Drift Decoupling and Correction." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10112–21. https://openaccess.thecvf.com/content/CVPR2022/html/Gao_FedDC_Federated_Learning_With_Non-IID_Data_via_Local_Drift_Decoupling_CVPR_2022_paper.html.

Karimireddy, Sai Praneeth, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. 2020. "SCAFFOLD: Stochastic Controlled Averaging for Federated Learning." In *International Conference on Machine Learning*, 5132–43. PMLR. https://proceedings.mlr.press/v119/karimireddy20a.html.

Li, Tian, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. "Federated Optimization in Heterogeneous Networks." In *Proceedings of Machine Learning and Systems*, 2:429–50.

Li, Xudong, Zhaohua Qu, Shengling Zhao, Bo Tang, Zeyue Lu, and Yifei Liu. 2023. "An Experimental Study of Byzantine-Robust Aggregation Schemes in Federated Learning." *arXiv Preprint arXiv:2302.07173*. https://arxiv.org/abs/2302.07173.

McMahan, H Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. "Communication-Efficient Learning of Deep Networks from Decentralized Data." In *Artificial Intelligence and Statistics*, 1273–82. PMLR. https://proceedings.mlr.press/v54/mcmahan17a.html.

Rodriguez-Barroso, Nuria, Daniel Jimenez-Lopez, M. Victoria Luzón, Francisco Herrera, and Eugenio Martinez-Camara. 2023. "Reviewing Federated Learning Aggregation Algorithms: Strategies, Contributions, Limitations and Future Perspectives." *Electronics* 12 (10): 2287. https://www.mdpi.com/2079-9292/12/10/2287.

Yin, Dong, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. 2018. "Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates." *arXiv Preprint arXiv:1803.01498*. https://arxiv.org/abs/1803.01498.

9 Appendix

9.1 A. Detailed Results Tables

[Full results for each configuration]

9.2 B. Convergence Trajectories

[All convergence plots with confidence bands]

9.3 C. Statistical Test Details

[Full ANOVA and t-test results]