FedProx: Federated Optimization with Proximal Terms

Abhishek Sushil, Alhad Sethi

December 10, 2024

Abstract

FedProx is a robust optimization framework for federated learning, designed to address challenges such as heterogeneity in data distributions and computational resources among clients. This paper provides an overview of the FedProx algorithm, its theoretical properties, and empirical evaluations on the MNIST dataset under IID and Non-IID settings. Results demonstrate the effectiveness of FedProx in achieving convergence and improving federated model performance.

1 Introduction

Federated Learning (FL) enables collaborative model training across distributed clients without requiring raw data to be shared. However, FL faces challenges such as data heterogeneity (Non-IID data distributions) and system heterogeneity (variable computational capabilities). To address these, FedProx introduces a proximal term in the optimization objective, ensuring robust convergence and efficient training in federated environments.

In this paper, we explore the theoretical foundations of FedProx and its empirical performance on the MNIST dataset under IID and Non-IID conditions. The proximal term added by FedProx mitigates the adverse effects of heterogeneity, making it a versatile tool for real-world FL applications.

2 Theory

FedProx modifies the standard FL optimization problem by introducing a proximal term that penalizes deviations from the global model. The objective for client k is:

$$\min_{w_k} f_k(w_k) + \frac{\mu}{2}\|w_k - w\|^2, \tag{1}$$

where w_k is the local model, w is the global model, $f_k(w_k)$ is the local objective (e.g., empirical loss), and $\mu > 0$ is the regularization parameter.

2.1 Convergence for Strongly Convex Functions

For strongly convex local objectives, FedProx guarantees convergence to a global optimum. Let L be the smoothness constant, μ the regularization parameter, and κ the condition number. The convergence rate is:

$$\|w^{(t+1)} - w^*\|^2 \leq \rho \|w^{(t)} - w^*\|^2, \tag{2}$$

where $\rho < 1$ is a function of L and κ, ensuring linear convergence.

2.2 Convergence for Non-Convex Smooth Functions

For non-convex objectives, FedProx converges to a stationary point. Let η_t be the learning rate. After T iterations, the expected gradient norm satisfies:

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\nabla f(w^{(t)})\|^2] \leq O\left(\frac{1}{T}\right). \tag{3}$$

This guarantees that FedProx minimizes the local variance and stabilizes updates, even under heterogeneity.

3 Experimental Setup

3.1 Dataset and Preprocessing

We used the MNIST dataset for empirical evaluation. The dataset contains 60,000 training and 10,000 testing images of handwritten digits (0-9), each resized to 28×28 pixels and normalized to [0, 1].

3.2 Synchronization Function

To implement the Non-IID data split from scratch, we designed a function that assigns a subset of labels to each client, ensuring heterogeneous data distributions. First, the synchronization function was implemented to align the local models with the global model. This function loads the global model's parameters into each client's local model before their training begins. During training, clients independently optimize their local objectives while using the FedProx proximal term to constrain their updates. After training, the updated local parameters are sent back to the server, where they are aggregated using an averaging scheme. This iterative process ensures that the global model incorporates contributions from all clients while mitigating the impact of Non-IID data heterogeneity.

3.3 Experimental Design

FedProx was evaluated under two conditions:

- **IID Setting**: Data uniformly distributed among clients, ensuring similar label distributions. Thus representing ideal data-sharing conditions.

- **Non-IID Setting**: Created imbalanced data distributions by assigning specific subsets of labels to each client. This models real-world scenarios where clients have heterogeneous data distributions, such as hospitals sharing medical records or personalized smartphone applications.

We used a simple neural network with three hidden layers as the model architecture. Three clients participated in training, and the proximal term parameter μ was varied to analyze its impact.

3.4 Training Configuration

- Learning Rate: 0.1
- Local Epochs: 10
- Batch Size: 25
- Number of Rounds: 10

4 Results and Discussion

4.1 IID Results

Under the IID setting, FedProx achieved stable convergence, matching or exceeding the performance of standard FL algorithms. The inclusion of the proximal term reduced client-to-client variability, as illustrated in Figure 1.

4.2 Non-IID Results

In the Non-IID setting, FedProx demonstrated significant improvements over traditional FL approaches. The proximal term mitigated the adverse effects of data heterogeneity, achieving higher accuracy and faster convergence (Figure 2).

4.3 Conclusion

FedProx provides a robust framework for federated optimization, particularly under challenging Non-IID scenarios. By penalizing large deviations from the global model, it ensures stable and efficient convergence, making it a valuable tool for real-world federated learning applications.

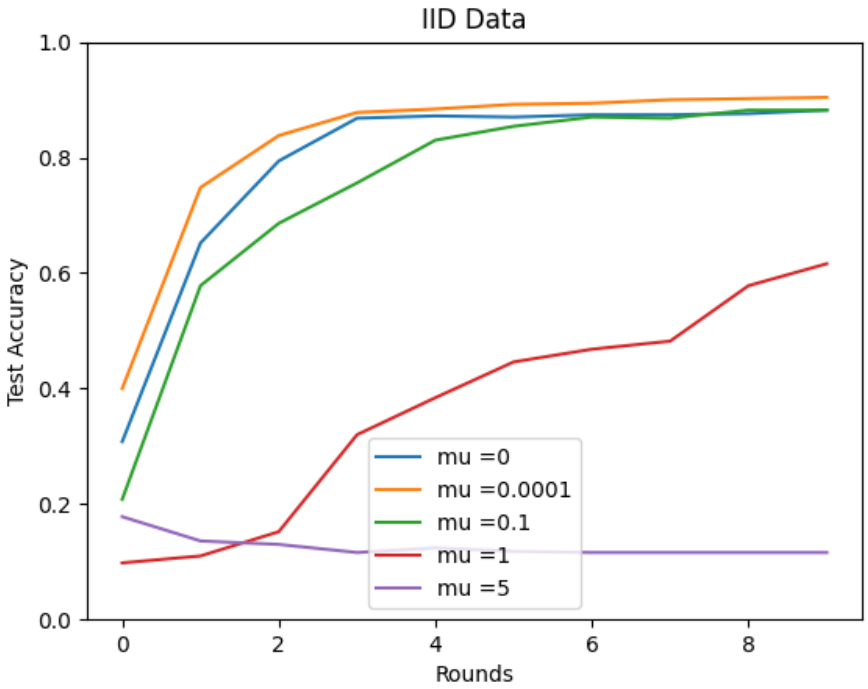

Figure 1: Training accuracy under IID conditions.

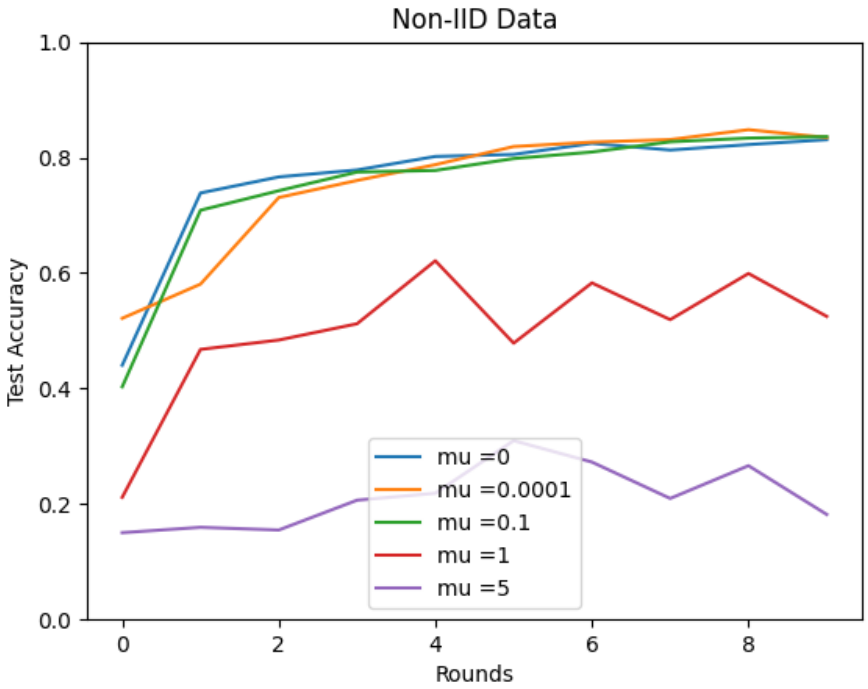

Figure 2: Training accuracy under Non-IID conditions.