Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/160/2025.TJ23539

Analysis on Optimizing Federated Proximal Algorithm for Heterogeneous and Secure Collaborative Learning

Donglin Yu

The Grainger College of Engineering, University of Illinois Urbana Champaign, Champaign, USA
donglin5@illinois.edu

Abstract: Federated Learning (FL) has to decentralize the model training but maintains users' data privacy, hence it is potentially essential in critical applications such as healthcare, finance, etc. For FL, the main obstacles remain the client heterogeneity and the sensitivity to any security attacks, which severely hinder its application to real scenarios. In this paper, the thesis studies the edge cases of the Federated Proximal (FedProx) algorithm that incur this phenomenon and suggests six ways for mitigating them. More precisely, the thesis considers adaptive regularization, knowledge distillation and transfer, optimization on efficiency, security defenses, client selection strategies, and approaches dealing with behavioural heterogeneity. Experiments conducted on benchmark datasets such as Canadian Institute for Advanced Research (CIFAR)-10 and Federated Extended Modified National Institute of Standards and Technology (FEMNIST) demonstrate that these strategies can improve FedProx accuracy by up to 7.2% and reduce communication rounds by up to 30%. The thesis's findings enhance the robustness, scalability, and personalization of FedProx in heterogeneous and adversarial settings. Such enhancements have a practical benefit for implementing FL systems over various real-world settings.

Keywords: Federated Learning, FedProx, Client Heterogeneity, Security Defense.

1. Introduction

Federated Learning (FL) allows collaborative model training across decentralized devices while preserving data privacy, making it ideal for domains like healthcare, finance, mobile applications, and IoT. Introduced by McMahan et al. in 2016, FL avoids raw data exchange, aligning well with privacy regulations and data localization needs [1]. Despite its benefits, FL faces two core challenges: client heterogeneity and security threats. Heterogeneity appears in three forms: statistical (non-IID data distributions), system (variations in client hardware or bandwidth), and behavioral (irregular participation or unreliable updates). These factors often degrade model convergence and performance. Meanwhile, security threats, particularly Byzantine attacks, where clients maliciously submit harmful updates, undermine model integrity and trust. Various algorithms have extended the basic Federated Averaging (FedAvg) approach to tackle these challenges. Among them, the Federated Proximal (FedProx) algorithm, proposed by Li et al., introduces a proximal term to constrain local updates, thus mitigating "client drift" under non-IID settings [2]. This makes FedProx more stable and better suited for heterogeneous systems.

However, FedProx is not without limitations. Key issues include: regularization tuning difficulty, limited robustness under extreme heterogeneity, increased computational cost, weak security

DOI: 10.54254/2755-2721/160/2025.TJ23539

guarantees, communication inefficiency, basic client selection, and neglect of behavioral heterogeneity. This survey analyzes how FedProx addresses these challenges and proposes six optimization strategies: adaptive regularization, knowledge distillation and transfer, efficiency improvements, security defenses, more brilliant client selection, and behavioral heterogeneity handling. This study aims to provide a focused view of FedProx's limitations and actionable paths for practical enhancement.

2. Methods

In this work, the thesis conducts a comprehensive study of the deficiencies of the FedProx algorithm and possible optimization efforts for reducing the impact of the FedProx algorithm for solving different challenging issues of federated learning across heterogeneous systems with possibly adversarial participation. This thesis investigates six optimization directions based on the deficiencies of the FedProx algorithm, which cover a wide range of considerations for solving challenging issues with this algorithm. They include adaptive regularization approaches, security-based optimization, heterogeneity of participants' behavior, etc. In Figure 1, the study is presented following a systematic way where the first step is to reveal FedProx's core deficits, in a subsequent step, consider optimized methods in the literature, and finally, in the third step, provide a comprehensive solution that is adoptable in an industrial FL implementation.

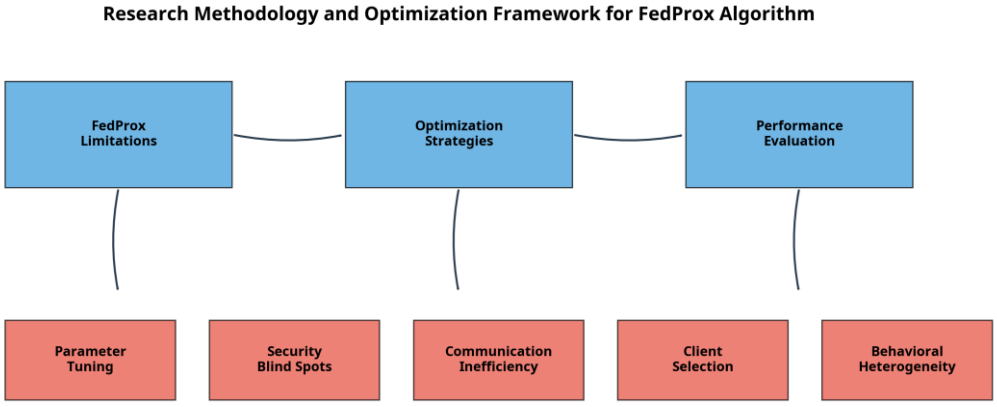

Figure 1: The research methodology and optimization framework for FedProx algorithm (photo credit: original)

FL trains models on distributed clients without viewing the local data. The server orchestrates this communication with three iterative updates: model broadcasting, regional training, and aggregation. FL mitigates the need to account for privacy as well as bandwidth limitations. However, it faces difficulties with client heterogeneity and adversarial clients. FedAvg is the baseline algorithm in FL [1]. Each client is selected by a subset each round. Multiple rounds of local training are utilized for model training before aggregating model updates. Yet, non-IID data distributions are vulnerable to client drift—divergent models trained by different clients, as the local data is different. FedProx enhances FedAvg by introducing a proximal term into the loss at the client side and penalizes divergence against the global model, so the local models update a less diverged model for better

DOI: 10.54254/2755-2721/160/2025.TJ23539

convergence under heterogeneity [2]. Although FedProx works well, it presents hyperparameter tuning and computational overhead, and is not robust to malicious clients.

3. Challenges in federated learning

FL encounters two fundamental challenges—client heterogeneity and security threats—that critically impact its efficiency and robustness.

3.1. Client heterogeneity

Heterogeneity of clients: In FL, the heterogeneity of clients includes data distribution, computing and communications capabilities, and behavioral patterns. It can be roughly classified into three types.

Non-IID data, also called statistical heterogeneity, is the heterogeneity in the data among clients, such as feature skew (e.g., distinct vocabularies in a Neuro-Linguistic Programming (NLP) task), label distribution skew (e.g., some clients have only a few class label types), and quantity skew (i.e., clients differ in dataset size). The disparity in such characteristics can cause local gradients to grow apart, resulting in divergence that can further exacerbate the global model and even collapse. Empirical studies have reported a 50% drop in FedAvg's accuracy on Modified National Institute of Standards and Technology (MNIST) when all clients are limited to only two-digit classes, indicating the scale of the problem [2].

System heterogeneity: System heterogeneity means variations in the computational capabilities (Central Processing Unit (CPU), Graphics Processing Unit (GPU)), Random Access Memory (RAM), and networking capacity of clients. Resource-limited devices can also become stragglers or withdraw from training altogether, thereby blocking training and disrupting the uniformity in client participation.

The behaviour heterogeneity includes clients' diverse participation rates and participation quality. There are some high-quality, frequent contributors, some occasional contributors, and some noisy gradient contributors. The behaviour heterogeneity disturbs the coordination in the training process and disrupts the model consistency.

Numerous approaches have been used to tackle these issues, such as data-sharing schemes, algorithmic tweaks (FedProx and SCAFFOLD), asynchronous training regimes, and meta-learning [3]. Hybrid techniques, e.g., combining proximal regularization with data augmentation, have generated impressive improvements, reducing the accuracy disparity between weak and strong nodes by 24% to 7% on Canadian Institute for Advanced Research (CIFAR)-10 [4].

3.2. Security threats

FL faces a wide range of security threats, particularly from Byzantine attacks. Byzantine attacks happen when a malicious client submits a poisoned update that may disrupt global performance by perturbing local updates, such as a gradient inversion attack, a label flipping attack, or a noise injection attack. Gradient inversion attempts to reverse the learning path, and label flipping and noise injection insert semantic or statistical perturbations, leading to corrupted model parameters over time.

Traditional FL algorithms, including FedAvg and FedProx, use aggregations as a core component of the updates. Thus, all these aggregations are susceptible to these manipulations. Xia et al. state that FL schemes have no natural defenses against cheating [4]. In the federated learning literature, robust aggregation methods, e.g., median, trimmed mean, and Krum, were proposed for resisting these attacks, assuming all but a few clients are honest [5]. However, they are passive defenses and may become ineffective as the attacker ratio increases.

In contrast, trust-based algorithms such as FLTrust rate clients according to their similarity to an untrusted reference update and update the aggregation weights based on their scores [6]. Such

DOI: 10.54254/2755-2721/160/2025.TJ23539

techniques offer proactive protection based on similarities to an untrusted reference update instead of majority consensus.

Nonetheless, defense impacts performance across cost, dimension, and attack type. Fu et al. find that even light Gaussian noise attacks adversely affect FedAvg performance for non-IID datasets, implying the need for adaptive security depending on the use case [7].

4. FedProx limitations and optimization directions

While FedProx significantly enhances FedAvg under non-IID settings by regularizing local updates with a proximal term, some limitations still hamper its adoption in practice.

Parameter tuning. Selecting the appropriate value for the proximal term coefficient (μ) remains a nontrivial problem. One that is too small does not sufficiently suppress client drift, and one that is too large does not adequately facilitate local optimization. The proposed algorithm performs very differently depending on the task and the type of data skew.

Scalability. Another issue of FedProx is its increasing overhead at the client side for the extra computations of regularization, which can put the computation-intensive requirement on low-power devices and restrict deployments in such environments. Another major limitation is security blind spots. As noted by Xia et al. FedProx, as well as other conventional federated learning systems, cannot resist malicious activities intrinsically. And aggregation processes are sensitive to Byzantine attacks from malicious clients that strongly influence model performance [8-10]. Fu et al. It has been proven that aggregation-based methods are especially prone to attack [11]. However, communication inefficiency is still an issue of proximal-constraint stabilized FedProx. Since FedProx can only converge fast in several communication rounds for canonical benchmarks such as MNIST, it increases the usage of communication bandwidth and the total training time.

Random sampling can't consider clients with low value, low data utility, and unstable clients, preventing the model from reaching optimal performance and causing ineffective resource utilization. Without considering clients' behavioral diversity, the proposed FedProx cannot provide a mechanism to adapt to changing participation trends of clients (and their reliability), which is necessary for FedProx to work in real deployments where they vary greatly.

4.1. Adaptive regularization

In an alternative approach, dynamic μ strategies adjust the proximal strength over time or per client, offering more flexible and context-aware regularization. These approaches include phase-aware μ implementation, which starts with strong constraints (e.g., μ=1) and then gradually decays (e.g., to μ=0.01) as training progresses [12]. This temporal adaptation allows for stronger regularization during early, unstable phases and more freedom for local optimization as the model matures.

Divergence-based μ adjustment represents another promising direction, using Kullback-Leibler divergence to guide μ calibration. Experiments on the CIFAR-10 dataset have demonstrated that this approach can significantly improve FedProx accuracy compared to fixed μ implementations [2].

Client-specific μ tuning employs reinforcement learning-based models to adjust regularization strength based on client data statistics. This personalized approach has shown promising results on the FEMNIST dataset, achieving faster convergence than standard implementations [9].

4.2. Knowledge transfer and distillation

Improving generalization performance without compromising personalization is another direction of optimization for FedProx. Bidirectional distillation allows the global and local models to transfer soft logits to each other [7]. Their approach attains a significant increase in performance for benchmarks such as CIFAR-10. These knowledge distillation techniques can be introduced into FedProx to

DOI: 10.54254/2755-2721/160/2025.TJ23539

counteract the drop in performance with the non-IID assumption by fostering the exchange of knowledge more uniformly across all heterogeneous clients.

Meta-learning approaches such as Per-FedAvg utilize Model-Agnostic Meta-Learning (MAML) to generate a shared initialization, followed by local adaptation using FedAvg [11]. While initially designed for FedAvg, this strategy could inspire similar personalization enhancements in FedProx.

Another strategy involves stronger regularization on shared layers and weaker constraints on personalized layers, enabling a flexible trade-off between global and local objectives [12].

4.3. Efficiency improvements

Several optimization strategies are proposed to reduce the computational and communications costs incurred by FedProx. Sparse proximal updates selectively regularize only the most impactful weights, substantially reducing the computational overhead introduced by the proximal term, particularly in large models such as ResNet-50, while maintaining performance [13]. This strategy helps address one of FedProx's key limitations: increased client-side computation.

Low-rank approximation methods have also compressed the proximal term in dense layers, which reduces the model's memory cost and computation complexity without damaging the model's quality. This makes FedProx more practical for deployment in low-powered devices such as mobile phones and IoT nodes. Sketching methods like those employed in FetchSGD use Count Sketches to compress gradient updates [13]. Integrating sketching into FedProx's update procedure can dramatically reduce communication overhead, making it more suitable for bandwidth-limited environments.

In asynchronous protocols, clients can arbitrarily update instead, which captures different computational power and network conditions. Xia et al. reported that asynchronous methods can be beneficial when device heterogeneity is high [4]. In addition, the convergence speed benefits from the stability advantages offered by FedProx. The asynchronous variants likely give both speedup and improve general hardware applicability, but the thesis needs to be cautious not to break convergence guarantees.

4.4. Security enhancements

It is possible to compose FedProx with many different defense layers to tackle its security shortcomings. For instance, like FLTrust, FedProx introduces trust-based μ scaling, where low-trust clients are updated with more regularization to keep them accurate at high attack rates [6].

Robust aggregation via prox-constrained robust aggregation merges geometric median computation and distance constraints for robust aggregation [10]. Geometric median provides stability when using prox terms, and the distance constraint helps aggregate outlier-free samples, rendering them more robust.

Gradient-based anomaly detection works on the same principle by flagging those clients that exhibit anomalous gradient updates compared to those of another client by their gradient direction and magnitude [5]. These detection methods do not require intricate trust infrastructure and are computationally inexpensive to detect and stop malicious contributions. For example, cosine similarity or Euclidean distance metrics might be employed to identify outliers that substantially differ from most updates for stronger robustness concerning adversarial scenarios.

Differential privacy, which applies calibrated noise to be resilient against gradient inversion attacks, offers formal privacy guarantees and decent model performance and can be considered a formal way to preserve privacy [14, 15].

DOI: 10.54254/2755-2721/160/2025.TJ23539

4.5. Advanced client selection

Intelligent client selection can be applied instead of naive client selection, which samples clients randomly, to enhance federated optimization's convergence speed and fairness in client selection. Utility-based client selection focuses on the most useful clients (e.g., those with larger data, more varied samples, or larger gradient magnitude). It encourages the heavy lifting on useful clients by speeding up the model convergence rate [9].

Resource-aware scheduling considers clients' memory, compute capacity, or network bandwidth to optimize participation. This strategy ensures smoother training and better load balancing across heterogeneous environments by avoiding low-resource bottlenecks. Several studies have demonstrated that adaptive scheduling can reduce round failures and idle times in practical FL deployments.

Trust selection mechanisms record clients' past performance and rank them based on their stability (not fluctuation) and accuracy of updates. They help the system adapt to a longer-term participation pattern rather than being influenced by a flaky or noisy worker. Thus, they serve longer-term FL training processes where trustworthiness cannot be assumed for all clients.

These selection mechanisms can be integrated with FedProx to improve robustness and personalization, particularly by allocating regularization effort to clients with lower reliability or greater heterogeneity.

4.6. Behavioral heterogeneity handling

A second group of Behavior-aware FedProx variants balances for participation frequency and the participation quality of the clients (trust). Availability models learn to compensate for irregular participation frequencies of clients.

Weighting by reliability allows authors to scale down the contribution of unreliable nodes to the global model rather than ignore them, thus regulating the compromise between the generic and the specific.

Regularizations that minimize per-client variance in fairness prevent performance improvement from being disproportionate to a specific subset of clients or a particular data distribution.

5. Future research directions

Balancing Heterogeneity and Security: Future FL frameworks should jointly model client diversity and adversarial threats. Adaptive defense mechanisms (e.g., heterogeneity-aware filtering thresholds or trust scores) are promising. Multi-objective optimization strategies can help navigate the accuracy, fairness, privacy, and robustness trade-offs. Theoretical Foundations: Stronger guarantees for adaptive μ convergence, client selection policies, and security bounds (considering heterogeneous clients and adapting to heterogeneous setups) are required theoretically; models unifying statistical and system heterogeneity will be significant. Cross-Domain Applications: Optimized FedProx variants can benefit domains like healthcare, finance, transportation, and edge IoT. For instance, hierarchical FL with trust-weighted μ is promising in multi-hospital networks or distributed smart city infrastructure.

6. Conclusion

FedProx greatly enhances FL's robustness against heterogeneity while lacking scalability, security, and personalization. Adaptive regularization, robust aggregation, and efficient communication enhance its feasibility. Blending the theoretical foundation into multifold cross-domain applications

DOI: 10.54254/2755-2721/160/2025.TJ23539

is the next direction. More innovative ideas, FedProx and its successors, will yield secure and inclusive collaborative intelligence.

References

[1] *McMahan, H. B., Moore, E., Ramage, D. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the International Conference on Artificial Intelligence and Statistics, 54, 1273–1282.*

[2] *Li, T., Sahu, A. K., Zaheer, M. (2020). Federated optimization in heterogeneous networks. Proceedings of Machine Learning and Systems, 2, 429–450.*

[3] *Li Q, Diao Y, Chen Q, et al. (2022). Federated learning on non-iid data silos: An experimental study. IEEE international conference on data engineering, 965-978.*

[4] *Xia, Y., Yu, W., & Li, Q. (2025). Byzantine-resilient federated learning via distributed optimization. arXiv preprint 10792.*

[5] *Feng, S., Zhang, Y., & Wang, X. (2024). A survey of security threats in federated learning. Complex & Intelligent Systems. 1664.*

[6] *Cao, X., Fang, M., Liu, J. (2021). FLTrust: Byzantine-robust federated learning via trust bootstrapping. Proceedings of the Network and Distributed System Security Symposium (NDSS).*

[7] *Jeong, E., & Kountouris, M. (2023). Personalized decentralized federated learning with knowledge distillation. arXiv preprint: 12156.*

[8] *Ma, X., Zhu, J., Lin, Z. (2022). A state-of-the-art survey on solving non-IID data in federated learning. Information Fusion, 89, 244–258.*

[9] *Wang, H., Kaplan, Z., Niu, D. (2020). Optimizing federated learning on non-IID data with reinforcement learning. In IEEE INFOCOM 2020 - IEEE Conference on Computer Communications, 1698–1707.*

[10] *Pillutla, K., Kakade, S. M., & Harchaoui, Z. (2022). Robust aggregation for federated learning. IEEE Transactions on Signal Processing, 70, 1142–1154.*

[11] *Yang L, Huang J, Lin W, et al. (2023). Personalized federated learning on non-IID data via group-based meta-learning[J]. ACM Transactions on Knowledge Discovery from Data, 17(4), 1-20.*

[12] *Yang X, Huang W, Ye M. (2023). Dynamic personalized federated learning with adaptive differential privacy. Advances in Neural Information Processing Systems, 36: 72181-72192.*

[13] *Rothchild, D., Panda, A., Ullah, E. (2020). FetchSGD: Communication-efficient federated learning with sketching. In Proceedings of the International Conference on Machine Learning, 8253–8265.*

[14] *Liu P, Xu X, Wang W. (2022). Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives. Cybersecurity, 5(1), 4.*

[15] *Li, S., Ngai, E. C.-H., & Voigt, T. (2023). An experimental study of Byzantine-robust aggregation schemes in federated learning. arXiv preprint arXiv:2302.07173.*