Bastion: Adversarial-Aware Federated Learning with Privacy-Preserving Game Theory

Abstract: Federated Learning (FL) faces critical challenges in balancing privacy preservation with adversarial robustness, particularly under evolving threats of gradient inversion and model poisoning attacks. This paper introduces **Adversarial-Aware Federated Learning with Privacy-Preserving Game Theory (AAFL-PGT)**, a novel two-stage framework that synergizes differential privacy (DP), adversarial training, and hardware acceleration to address these challenges. In the first stage, participants engage in a cooperative game-theoretic negotiation to dynamically allocate DP noise budgets, optimizing the trade-off between privacy guarantees and model utility. The second stage employs adversarial training to harden the global model against evasion and membership inference attacks. To enhance efficiency, FPGA-based enclaves enforce DP constraints during gradient aggregation, ensuring real-time privacy compliance while minimizing computational overhead.

Literature Review:

1. Introduction to Federated Learning and Its Challenges

Federated Learning (FL) has emerged as a distributed machine learning paradigm designed to collaboratively train models on decentralized data without requiring data sharing. This approach not only addresses data volume and communication challenges but also offers a promising framework for preserving data privacy. However, as FL systems have evolved, they have revealed vulnerabilities in both privacy and robustness. These issues manifest in several forms, including inference attacks, gradient leakage, poisoning attacks, and adversarial examples. The literature reveals a strong research focus on integrating privacy-preserving techniques such as differential privacy (DP) with mechanisms to defend against adversarial behavior, while simultaneously considering the efficiency and incentive aspects of large-scale collaborative learning.

2. Major Themes in the Literature

A. Privacy and Robustness Trade-offs in FL

- **Differential Privacy Integration:**
  Researchers have incorporated various forms of DP—ranging from local to joint differential privacy—to mitigate privacy risks in FL systems. For example, the work by Lyu et al. discusses how adversaries can exploit gradient information

and proposes defenses that combine privacy guarantees with robustness against poisoning and inference attacks .

- **Adversarial Robustness and Defense Mechanisms:**
  Several documents focus on the vulnerability of FL models to adversarial examples. Approaches like federated adversarial training and decision boundary adjustments have been explored to mitigate such attacks, as discussed in the work by Jie Zhang et al. which evaluates the performance degradation when applying adversarial training in non-IID settings .

B. Incentive Mechanisms and Game-Theoretic Approaches

- **Incentive-Driven FL Schemes:**
  Recognizing that clients incur costs (e.g., computation, communication, privacy risks) by participating in FL, multiple studies have integrated incentive mechanisms into FL protocols. For instance, the IMFL framework leverages personalized differential privacy settings along with game-theoretic models (such as Stackelberg games) to balance privacy, data utility, and participation incentives .
- **Game-Theoretical Client Selection and Reward Allocation:**
  Other works employ game theory to analyze and design collaboration strategies. By modeling the interaction between a central server and heterogeneous clients, these approaches aim to determine optimal strategies for client selection, resource allocation, and reward payment. The Stackelberg game analysis for FL with DP is a notable example that offers insights into client-server dynamics and .

C. Hardware Acceleration and Efficiency Improvements

- **FPGA and ASIC-based Acceleration:**
  To address the computational overhead inherent in privacy-preserving and robust FL, several studies have explored hardware acceleration. The PipeFL framework, for example, presents a co-design of hardware/software solutions to accelerate FL aggregation and differential privacy enforcement on FPGAs .
- **Edge and FPGA-Accelerated FL Systems:**
  Other research, such as the FLAIRS project, leverages FPGA-based Trusted Execution Environments (TEEs) to secure and accelerate FL while mitigating backdoor and inference attacks. This line of work demonstrates substantial speed-ups over software-only implementations and highlights the potential of specialized hardware in enabling real-time privacy defenses and .

3. Identified Gaps in the Literature

Despite significant progress, the literature reveals several areas where further research is needed:

- **Interplay Between Differential Privacy and Adversarial Robustness:**
  Although multiple studies examine DP's role in FL and its impact on privacy, there is less clarity on how different levels of local differential privacy (e.g., varying privacy budgets) affect adversarial robustness. For instance, while some works indicate that increased noise can improve robustness against specific attacks, a systematic theoretical and empirical analysis remains underdeveloped ; .
- **Client Heterogeneity and Personalized Protection:**
  Many frameworks assume a relatively homogeneous client population or consider only limited dimensions (e.g., privacy budgets or data quality). More work is required to address the full spectrum of client heterogeneity—including varying computational capabilities, non-IID data distributions, and different risk tolerances—and to design mechanisms that provide personalized privacy protection without compromising global model performance .
- **Integration of Robustness, Privacy, and Efficiency:**
  While individual aspects (robustness, privacy, efficiency) are often tackled in isolation, integrated solutions that jointly optimize these dimensions are rare. For example, hardware acceleration schemes frequently focus on computational performance without explicitly addressing adversarial robustness or dynamic privacy requirements .
- **Dynamic and Multi-dimensional Incentive Models:**
  Existing game-theoretic models provide static or simplified views of client-server interactions. There is a need for dynamic models that account for evolving client behavior, real-time performance metrics, and the adversarial potential of participants, which could further enhance the incentive mechanisms ; .

4. Potential Areas for Future Research

Based on the literature synthesis, several promising directions emerge:

- **Holistic Frameworks for FL:**
  Future research could focus on developing integrated frameworks that simultaneously optimize privacy, robustness, computational efficiency, and economic incentives. These frameworks should allow for adaptive privacy mechanisms (e.g., dynamic noise allocation) while maintaining high model accuracy even under adversarial conditions.

- **Theoretical Analysis of Privacy-Robustness Trade-offs:**
  A deeper theoretical investigation into the trade-offs between differential privacy (especially under local settings) and adversarial robustness could help design more effective training protocols. This would involve deriving formal bounds and empirical validations to guide the selection of privacy parameters.
- **Advanced Incentive Mechanism Design:**
  Further exploration of multi-dimensional, dynamic game-theoretic models could lead to more nuanced incentive mechanisms that account for the temporal evolution of client participation, data quality variations, and adversarial strategies. Incorporating machine learning into these incentive models (e.g., using reinforcement learning to adaptively adjust rewards) is another promising avenue.
- **Leveraging Heterogeneous Hardware:**
  Extending hardware acceleration research beyond FPGAs to include emerging accelerators (e.g., ASICs or edge TPUs) while integrating privacy and robustness considerations could yield scalable and efficient FL systems. Additionally, combining hardware acceleration with secure enclaves or TEEs on various platforms might provide a unified solution to performance and security challenges.
- **Real-world Deployment and Benchmarking:**
  Finally, as many proposed methods have been evaluated in controlled experimental setups, more research is needed to validate these approaches in real-world scenarios. Developing standardized benchmarks for adversarial robustness, privacy leakage, and efficiency in FL, and conducting large-scale field studies, would help bridge the gap between theoretical proposals and practical implementations.

Problem Statement:

Federated Learning (FL) struggles to harmonize three critical requirements: (1) *privacy preservation* against gradient inversion and membership inference attacks, (2) *adversarial robustness* to poisoning and evasion attacks, and (3) *computational efficiency* in large-scale, heterogeneous environments. Existing approaches address these challenges in isolation, leading to suboptimal trade-offs (e.g., excessive noise degrades model utility, static privacy budgets ignore client heterogeneity, and hardware acceleration lacks adversarial awareness).

Core Hypothesis

The **AAFL-PGT framework** will achieve superior privacy-robustness-efficiency trade-offs by:

1. **Dynamically optimizing local differential privacy (DP) noise budgets** through a cooperative game-theoretic mechanism, adapting to client heterogeneity (e.g., data quality, risk tolerance).
2. **Integrating adversarial training** into global model aggregation to harden against evasion and poisoning attacks while preserving utility under non-IID data.
3. **Leveraging FPGA-based enclaves** to enforce DP constraints and accelerate adversarial training, ensuring real-time compliance with privacy guarantees.

Sub-Hypotheses

1. **Dynamic DP Allocation via Game Theory**
    - *Hypothesis*: A cooperative game-theoretic negotiation for DP noise budgets will outperform static or centralized allocation strategies in balancing privacy-utility trade-offs, especially under client heterogeneity (e.g., non-IID data, varying computational resources).
    - *Basis*: Prior work highlights the effectiveness of Stackelberg games for FL incentives but lacks dynamic adaptation to evolving client conditions [3,4].
2. **Adversarial Training Enhances Robustness Under DP Noise**
    - *Hypothesis*: Jointly optimizing DP noise and adversarial training (via evasion attack simulations) will improve model robustness against gradient inversion and poisoning attacks without significant accuracy loss, even with non-IID data distributions.
    - *Basis*: Studies suggest DP noise can mask adversarial gradients, but its interaction with adversarial training remains underexplored [1,7].
3. **Hardware Acceleration Mitigates Efficiency Overheads**
    - *Hypothesis*: FPGA-based gradient aggregation with DP enforcement will reduce computation time by ≥30% compared to software-only implementations while maintaining provable privacy guarantees.
    - *Basis*: FPGA acceleration in FL shows promise [6,8], but existing solutions do not integrate adversarial training or dynamic DP allocation [9].
4. **Client Heterogeneity Drives Personalized Protection**
    - *Hypothesis*: Clients with higher data quality or lower risk tolerance will negotiate for lower DP noise budgets, improving global model accuracy without compromising their privacy.

    - *Basis*: Static DP budgets in IMFL [3] fail to account for client-specific constraints.

Expected Contributions:

1. **Framework Design**: A two-stage FL framework integrating game theory, DP, adversarial training, and hardware acceleration.
2. **Theoretical Insights**: Formal analysis of the interplay between DP noise levels and adversarial robustness, including bounds on model utility under adaptive attacks.
3. **Practical Acceleration**: Open-source FPGA modules for privacy-preserving adversarial training in FL.

Validation Plan

1. **Experiments**:
    - Compare AAFL-PGT against baselines (e.g., IMFL [3], PipeFL [9]) on benchmark datasets (CIFAR-10, MNIST) under varying attack scenarios (gradient inversion, label flipping).
    - Metrics: Privacy leakage (membership inference success rate), robustness (attack success rate drop), utility (test accuracy), and latency.
2. **Theoretical Analysis**:
    - Derive convergence guarantees for AAFL-PGT under dynamic DP noise and adversarial perturbations.
3. **Hardware Evaluation**:
    - Profile FPGA performance (throughput, power efficiency) against GPU/CPU implementations.

Potential Limitations & Mitigations

- **Limitation**: Game-theoretic negotiations may introduce communication overhead.
    - *Mitigation*: Implement lightweight negotiation protocols (e.g., batched client updates).
- **Limitation**: Adversarial training may amplify bias in non-IID settings.
    - *Mitigation*: Incorporate fairness-aware regularization during global aggregation.

Future Extensions

1. Extend the game-theoretic model to account for malicious clients attempting to manipulate negotiations.
2. Explore ASIC-based acceleration for energy-constrained edge devices.

References

[1] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, and P. S. Yu, “Privacy and robustness in federated learning: Attacks and defenses,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 7, pp. 8726–8741, Jul. 2024.

[2] Y. Han, Y. Cao, and M. Yoshikawa, “Understanding the interplay between privacy and robustness in federated learning,” in *ACM Symposium on Neural Gaze Detection*, Woodstock, NY, 2018.

[3] M. Li, Y. Tian, J. Zhang, Z. Zhou, D. Zhao, and J. Ma, “IMFL: An incentive mechanism for federated learning with personalized protection,” *IEEE Internet of Things Journal*, vol. 11, no. 13, pp. 23862–23875, Jul. 2024.

[4] G. Huang, Q. Wu, P. Sun, Q. Ma, and X. Chen, “Collaboration in federated learning with differential privacy: A Stackelberg game analysis,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 35, no. 3, pp. 455–469, Mar. 2024.

[5] L. Zhang, T. Zhu, P. Xiong, W. Zhou, and P. S. Yu, “A robust game-theoretical federated learning framework with joint differential privacy,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3333–3350, Apr. 2023.

[6] M. Mannino, A. Medaglini, B. Peccerillo, and S. Bartolini, “Accelerating differential privacy-based federated learning systems,” in *The Eighteenth International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP 2024)*, 2024.

[7] J. Zhang, B. Li, C. Chen, L. Lyu, S. Wu, S. Ding, and C. Wu, “Delving into the adversarial robustness of federated learning,” in *The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)*, 2023.

[8] H. Li, P. Rieger, S. Zeitouni, S. Picek, and A.-R. Sadeghi, “FLAIRS: FPGA-accelerated inference-resistant and secure federated learning,” *arXiv preprint arXiv:2308.00553*, Aug. 2023.

[9] Y. Han, “PipeFL: Hardware-software co-design of an FPGA accelerator for federated learning,” *Kyoto University Technical Report*, 2023.