Research Statement

Mengzhe RUAN
Department of Computer Science, City University of Hong Kong

January 30, 2025

My research centers on advancing the efficiency, scalability, and robustness of distributed machine learning systems, with a focus on federated learning, reinforcement learning, and optimization-driven algorithm design.

1 Research Overview & Philosophy

1.1 Part 1: Bridging Theory and Practice in Distributed Optimization

My research prioritizes *structure-aware algorithm design* to close the gap between theoretical optimization and real-world machine learning systems. By identifying latent structures in problems—such as gradient sparsity in federated learning or task relationships in reinforcement learning—I develop tailored frameworks that exploit these properties for efficiency. For example, gradient decoupling separates invariant and spurious gradient dimensions to accelerate convergence in heterogeneous federated learning, while adaptive Top-K SGD dynamically balances communication costs and convergence error. These works rigorously analyze theoretical guarantees (e.g., non-convex convergence rates) while demonstrating empirical impact, ensuring algorithms are both principled and deployable.

1.2 Part 2: Toward Scalable and Robust Learning Systems

I focus on building scalable and robust systems under resource constraints (e.g., limited bandwidth) and data scarcity. My approach combines *efficiency-driven optimization* (e.g., power control strategies for over-the-air federated learning) with *structured knowledge transfer* (e.g., linear MDP-based frameworks for offline RL). This dual strategy minimizes resource consumption while generalizing across tasks, supported by theoretical guarantees like suboptimality bounds.

2 Summary of Past and Ongoing Research

2.1 Federated Learning: Efficiency and Communication Co-Design

My work bridges optimization theory and practical challenges in federated learning (FL):

2.1.1 Gradient Decoupling for Heterogeneous FL (Ruan et al. [2025])

- **Unified Framework for Generalization and Optimization**: Proposes a gradient decoupling mechanism that jointly addresses convergence efficiency and generalization in federated learning. By separating invariant and spurious gradient dimensions through variance analysis, the framework adaptively assigns higher learning rates to spurious components while preserving invariant patterns, achieving faster convergence than FedAvg/ FeProx / FedOpt on several heterogeneous datasets.
- **Theoretical-Practical Synergy**: Establishes formal convergence guarantees for non-convex objectives. The plug-and-play design enables seamless integration with existing FL algorithms (FedProx, FedOpt), maintaining privacy constraints without requiring client data sharing.

Integration Perspective: The work bridges optimization dynamics with generalization theory by treating gradient dimensions as carriers of domain-invariant knowledge, fundamentally rethinking how FL algorithms process heterogeneous signals. This dual focus enables both accelerated training and robust out-of-distribution performance.

2.1.2 Communication-Efficient FL Systems (Ruan et al. [2023, 2024a,b])

- Co-designed power control and gradient compression: We proposed one novel power control strategy. Derived optimality gap bounds for over-the-air FL with gradient compression to improve the training performance.
- Adaptive Top-K sparsification: Lower the upper bound of convergence rate under general Top-K sparsification, to adjust the sparsity ratio in training process to reach faster convergence and better accuracy with and without error compensation.

These efforts highlight my focus on *resource-aware FL systems* that balance theory, efficiency, and scalability.

2.2 2. Knowledge Transfer in Data-Scarce Reinforcement Learning

In recent work Ruan and Gan [2025], we proposed the **KT-RL framework** for offline RL to address data scarcity:

1. **Framework Innovation**: We proposed novel flexible knowledge transfer framework under beyond conventional similarity assumptions between tasks, enabling transfer even when source and target relationships are loosely aligned. And we establishes a suboptimality upper bound for the learned policy and proves the algorithm's minimax optimality, demonstrating rigorous performance guarantees.
2. **Algorithm Design**: Introduces the method that aggregates statistical quantitiesfrom source tasks to achieve knowledge transfer instead of raw data, enhancing efficiency and scalability under privacy-preserving.

These efforts highlight my focus on *resource-efficient RL* to learning better policies with limited data.

3 Future Research Agenda

My recent research topic includes fairness-aware experimentation, demand estimation and best arm identification. I also will expand FL robustness (both algorithm and theory) to heterogeneous data even to combining LLMs.

Through unifying distributed optimization and RL-driven decision science, I seek to address societal challenges like healthcare sustainability and efficient LLM, taking advantage of an interdisciplinary environment to progress both theory and real-world AI deployment.

More specific, I aim to develop frameworks integrating LLMs with distributed optimization for enhanced efficiency and generalization in heterogeneous AI systems, focusing on LLM-driven adaptive optimization, synthetic data generation for federated learning, and theoretical guarantees for decentralized LLM fine-tuning. Additionally, I will bridge RL and operations research to solve dynamic problems in supply chains and healthcare, such as several interesting topics like RL-based inventory management, RL with fairness constraints, and causal RL for treatment optimization.

References

Mengzhe Ruan and Guangyan Gan. Bridging the data scarcity gap via knowledge transfer in offline reinforcement learning. ***SUBMIT To*** *The 42nd International Conference on Machine Learning (ICML 2025)*, 2025.

Mengzhe Ruan, Guangfeng Yan, Yuanzhang Xiao, Linqi Song, and Weitao Xu. Adaptive top-k in sgd for communication-efficient distributed learning. In *GLOBECOM 2023-2023 IEEE Global Communications Conference*, pages 5280–5285. IEEE, 2023.

Mengzhe Ruan, Yunhe Li, Weizhou Zhang, Linqi Song, and Weitao Xu. Optimal power control for over-the-air federated learning with gradient compression. In *2024 IEEE 30th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 326–333. IEEE, 2024a.

Mengzhe Ruan, Guangfeng Yan, Yuanzhang Xiao, Linqi Song, and Weitao Xu. Adaptive top-k in sgd for communication-efficient distributed learning in multi-robot collaboration. *IEEE Journal of Selected Topics in Signal Processing*, 2024b.

Mengzhe Ruan, Yunhe Li, Hao Shi, Hanxu Hou, Jianping Wang, Weitao Xu, and Linqi Song. Gradient decoupling: A plug-and-play framework for accelerating general federated learning. ***SUBMIT To*** *The 42nd International Conference on Machine Learning (ICML 2025)*, 2025.