

## RESEARCH INTEREST

• Machine learning, Optimization, Operations management

## EDUCATION

### Peking University

Beijing, China

• B.A. in Mathematics, School of Mathematical Sciences

Sep. 2022–Jul. 2026 (Expected)

- **Mathematics Courses:** Mathematical Analysis, Advanced Algebra, Complex Analysis, Probability Theory, Mathematical Statistics, Stochastic Processes, Stochastic Analysis, Basic Numerical Method, Basic Optimization Method.
- **Computer Science Courses:** Basic Artificial Intelligence and Deep Learning, Parallel and Distributed Computing, Introduction to Computer Vision, Introduction to Multi-model(auditor).
- **Honors and Awards:** [Applied Mathematics Elite Program](#) (The program accepted only 15 people this year)

## RESEARCH EXPERIENCE

### Convergence and Speedup Analysis of Distributed Optimization Algorithms

Peking University

• Undergraduate Research Assistant

Nov. 2023 – Present

- **Advisor:** [Kun Yuan](#)
- **Project 1: Analysis of Push-Pull Algorithm**
  - \* **Push-Pull Algorithm Provably Achieves Linear Speedup Over Arbitrary Network Topologies**  
[Liyuan Liang\\*](#), [Gan Luo\\*](#), [Kun Yuan](#) (\*Equal contribution) [[Arxiv](#)]
  - \* **This paper is submitted to SIOPT.**
  - \* Conducted research on the [Push-Pull Algorithm](#), focusing on convergence and linear speedup properties in non-convex and stochastic settings on arbitrary topology.
  - \* First to prove convergence and linear speedup properties of the [Push-Pull Algorithm](#) under non-convex and stochastic settings on arbitrary topology.
  - \* Validated the proposed theoretical results by conducting distributed optimization numerical experiments on the MNIST and CIFAR10 datasets.
- **Project 2: Analysis on Decentralized Optimization over Row-stochastic Networks**
  - \* **Achieving Linear Speedup and Optimal Complexity for Decentralized Optimization over Row-stochastic Networks**  
[Liyuan Liang\\*](#), [Xinyi Chen\\*](#), [Gan Luo\\*](#), [Kun Yuan](#) (\*Equal contribution) [[Arxiv](#)]
  - \* **Accepted to ICML2025, Spotlight**
  - \* Introduced effective metrics to capture the influence of row-stochastic mixing matrices
  - \* Established the first convergence lower bound for decentralized learning over row-stochastic networks
  - \* Incorporated a multi-step gossip (MG) protocol, to attain the lower bound, achieving optimal complexity.
  - \* Proposed a novel analysis framework demonstrating that [PULL-DIAG-GT](#) achieves linear speedup, which is the first such result for row-stochastic decentralized optimization.
  - \* Conducted numerical experiments to validate theoretical results.

### Online Scheduling on LLM Inference

Massachusetts Institute of Technology

• Undergraduate Research Assistant

Oct. 2024 – Present

- **Advisor:** [David Simchi-Levi](#)
- **Project 1: Online Batching Algorithm on LLM Inference**
  - \* **Optimizing LLM Inference: Fluid-Guided Online Scheduling with Memory Constraints**  
( $\alpha$ - $\beta$ ) [Ruicheng Ao\\*](#), [Gan Luo\\*](#), [David Simchi-Levi](#), [Xinshang Wang](#), available on [[SSRN](#)] and [[Arxiv](#)]
  - \* **This paper is submitted to Operations Research.**
  - \* Formulated the LLM inference as a multi-stage online scheduling task with stochastic queueable requests.
  - \* Proposed a novel online batching algorithm for LLM inference.
  - \* Proved that the algorithm achieves near-optimal throughput while controlling latency and Time to First Token (TTFT).
  - \* Conducted numerical experiments on synthetic and real-world datasets with Llama-7B on A100 GPU to validate theoretical results.