

## RESEARCH INTEREST

- Operations Research, Machine Learning, Optimization Theory

## EDUCATION

### Peking University

Beijing, China

Jun. 2023–Jul. 2026 (Expected)

- B.A. in Mathematics, School of Mathematical Sciences
  - 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).

Beijing, China

Sep. 2022–Jun. 2023

### Peking University

- B.A. in Chemistry, College of Environmental Sciences and Engineering

## SLECTED HONORS AND AWARDS

### Finalist in INFORMS Applied Probability Society Best Student Paper Prize, 2025

- Annual award recognizing outstanding student research in applied probability [\[Link\]](#) [\[Award\]](#)

Oct. 2025

### Applied Mathematics Elite Program

- The program lead by Prof. Weinan E accepted only 15 people this year

Jun. 2024

### Silver Medal, 35th Chinese Chemical Olympiad (National Final)

- The prize is awarded to the top 150 high school students in chemistry throughout China

Nov. 2021

## WORKING EXPERIENCE

### Decision Intelligence Lab, DAMO Academy

Hangzhou, China

- Research Intern

Jul. 2025–Sep. 2025

- Developed the complete codebase for MetaFlow and built execution sandboxes (including tool calling like RAG search and Lean4 theorem proving) for online generated workflow and verification, enabling real-time execution feedback for reinforcement learning with verifiable rewards.

## RESEARCH EXPERIENCE ON LARGE LANGUAGE MODELS AND OPERATIONS RESEARCH

### Online Scheduling on LLM Inference

Massachusetts Institute of Technology

Oct. 2024 – May. 2025

- Advisor: Prof. David Simchi-Levi, Massachusetts Institute of Technology

- Optimizing LLM Inference: Fluid-Guided Online Scheduling with Memory Constraints  
 $(\alpha\beta)$  [Ruicheng Ao\\*](#), [Gan Luo\\*](#), David Simchi-Levi, Xinshang Wang [\[SSRN\]](#) [\[Arxiv\]](#) [\[Code1\]](#) [\[Code2\]](#) [\[Demo\]](#)
  - Submitted to Operations Research.
  - Preliminary version accepted to NeurIPS 2025 MLxOR Workshop.
  - Finalist in INFORMS Applied Probability Society Best Student Paper Prize, 2025. [\[Link\]](#) [\[Award\]](#)
  - Formulated the LLM inference as a multi-stage online scheduling task with stochastic queueable requests, proposed a novel online batching algorithm for LLM inference and 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.

### Analysis of batching and scheduling algorithms in LLM inference

Columbia University

Apr. 2025 – Present

- Advisor: Prof. Jing Dong, Columbia University

- Work in progress, Analysis of Continuous Batching Algorithm in LLM Inference
  - In this work, we first formulated the continuos batching algorithm as a discrete-time model. Then we analyzed its steady state and its dynamics behavior under overloaded conditions. Next we will futher analyze its dynamics under admission control.

### LLM Agent and Workflow Generation

DAMO Academy

Apr. 2025 – Present

- Advisor: Prof. Wotao Yin, DAMO Academy & Prof. Bin Dong, Peking University

- **MetaFlow: A Meta Approach of Training LLMs into Generalizable Workflow Generators**  
 $(\alpha\text{-}\beta)$  [Gan Luo\\*](#), [Zihan Qin\\*](#), [Bin Dong](#), [Wotao Yin](#)
  - \* Submitted, Under Review.
  - \* Formulated workflow generation as a meta-learning problem where LLMs learn to compose task-level solution strategies from operators, producing reusable workflows that generalize across problem instances rather than instance-specific solutions.
  - \* Developed a two-stage training approach combining supervised fine-tuning on synthetic workflow data with reinforcement learning with verifiable rewards (RLVR), using execution feedback across instances to improve end-to-end success rates.
  - \* Demonstrated strong zero-shot generalization to untrained tasks and novel operator sets, achieving performance comparable to state-of-the-art baselines on in-domain tasks across benchmarks in question answering, code generation, and mathematical reasoning.

## RESEARCH EXPERIENCE ON OPTIMIZATION THEORY

---

- |  |   |
|--|---|
| <p><b>Convergence and Speedup Analysis of Distributed Optimization Algorithms</b></p> <p><i>Advisor: Prof. Kun Yuan, Peking University</i></p> <ul style="list-style-type: none"> <li>◦ <b>Push-Pull Algorithm Provably Achieves Linear Speedup Over Arbitrary Network Topologies</b><br/> <a href="#">Liyuan Liang*</a>, <a href="#">Gan Luo*</a>, <a href="#">Kun Yuan</a> [Arxiv] [Code: Linear Speedup] [Notes]           <ul style="list-style-type: none"> <li>* Submitted to SIAM Journal on Optimization.</li> <li>* Proposed a novel multi-step descent analysis framework and first proved that the <a href="#">Push-Pull algorithm</a> achieves linear speedup over arbitrary strongly connected digraphs. Our multi-step analysis resolved the non-vanishing noise issue inherent in <a href="#">traditional single-step approaches</a>. Also see the <a href="#">[notes]</a>.</li> <li>* Conducted all numerical experiments to validate the linear speedup property we proved.</li> </ul> </li> <li>◦ <b>Achieving Linear Speedup and Optimal Complexity for Decentralized Optimization over Row-stochastic Networks</b><br/> <a href="#">Liyuan Liang*</a>, <a href="#">Xinyi Chen*</a>, <a href="#">Gan Luo*</a>, <a href="#">Kun Yuan</a> [Arxiv] [Code]           <ul style="list-style-type: none"> <li>* Accepted to ICML 2025, Spotlight</li> <li>* Introduced novel metrics to characterize the influence of row-stochastic mixing matrices and established the first convergence lower bound for decentralized optimization over row-stochastic networks.</li> <li>* Developed a new analysis framework proving that <a href="#">PULL-DIAG</a> achieves linear speedup and proposed a multi-gossip protocol that resolves instability issues and attains the lower bound with near-optimal complexity.</li> <li>* Conducted all numerical experiments to validate the theoretical results on convergence lower bound, linear speedup, and near-optimal complexity.</li> </ul> </li> </ul> | <p><b>Peking University</b></p> <p><i>Nov. 2023 – Jun. 2025</i></p> |
|--|---|

## SKILLS & OTHERS

---

- **Computer Skills:** Python (&PyTorch), Cuda, Lean4, MATLAB, C++, LaTeX
- **Language:** [Sichuanese dialects](#) (Native), Mandarin Chinese (Native), English (Fluent)
- **The Test of English as a Foreign Language (TOEFL-IBT):** Total 101 with 27(R)+28(L)+22(S)+24(W)
- **Conference Reviewing:** ICLR 2026