

RESEARCH INTEREST

• Machine Learning, Optimization Theory, Operations Research

EDUCATION

Peking University

Beijing, China

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

Jun. 2023–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).

Peking University

Beijing, China

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

Sep. 2022–Jun. 2023

SELECTED 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

• The prize is awarded to the top 150 high school students in chemical 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 Lean theorem proving, RAG search) 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

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

Oct. 2024 – May. 2025

◦ Optimizing LLM Inference: Fluid-Guided Online Scheduling with Memory Constraints

(α - β) [Ruicheng Ao](#)*, [Gan Luo](#)*, [David Simchi-Levi](#), [Xinshang Wang](#) [SSRN] [Arxiv] [Code1] [Code2]

- * 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

• Advisor: Prof. Jing Dong, Columbia University

Apr. 2025 – Present

◦ Work in progress, Analysis of Continuous Batching Algorithm in LLM Inference

- * In this work, we first formulated the continuous batching algorithm as a discrete-time model. Then we analyzed its steady state and its dynamics behavior under overloaded conditions. Next we will further analyze its dynamics under admission control.

LLM Agent and Workflow Generation

DAMO Academy

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

Apr. 2025 – Present

- **MetaFlow: A Meta Approach of Training LLMs into Generalizable Workflow Generators**
(α - β) [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

Convergence and Speedup Analysis of Distributed Optimization Algorithms

Peking University

Advisor: Kun Yuan, Peking University

Nov. 2023 – Jun. 2025

- **Push-Pull Algorithm Provably Achieves Linear Speedup Over Arbitrary Network Topologies**
[Liyuan Liang*](#), [Gan Luo*](#), [Kun Yuan](#) [Arxiv] [Code: Linear Speedup] [Notes]
 - * **Submitted to SIAM Journal on Optimization.**
 - * Proposed a novel multi-step descent analysis framework and first proved that the [Push-Pull algorithm](#) achieves linear speedup over arbitrary strongly connected digraphs. Our multi-step analysis resolved the non-vanishing noise issue inherent in [traditional single-step approaches](#). Also see the [\[notes\]](#).
 - * Conducted all numerical experiments to validate the linear speedup property we proved.
- **Achieving Linear Speedup and Optimal Complexity for Decentralized Optimization over Row-stochastic Networks**
[Liyuan Liang*](#), [Xinyi Chen*](#), [Gan Luo*](#), [Kun Yuan](#) [Arxiv] [Code]
 - * **Accepted to ICML 2025, Spotlight**
 - * 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.
 - * Developed a new analysis framework proving that [PULL-DIAG](#) achieves linear speedup and proposed a multi-gossip protocol that resolves instability issues and attains the lower bound with near-optimal complexity.
 - * Conducted all numerical experiments to validate the theoretical results on convergence lower bound, linear speedup, and near-optimal complexity.

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