Hongpei Li

ishongpeili@gmail.com | Github | Homepage | Google Scholar

Education

Shanghai University of Finance and Economics 2021 - 2025
Bachelor of Engineering, Pilot Class of Interdisciplinary Sciences Shanghai, China
GPA 3.77/4.0 Average Score 89.55/100 Major GPA 3.87 Major Average Score 91.48/100
Selected Coursework: Advanced Operation Research, Linear and Nonlinear Programming, Probability, Numerical Optimization, Mathematical Analysis, Linear Algebra, Data Structures & Algorithms, Machine Learning & Deep Learning.
Honors & Awards:

- Honor Bachelor Degree, 2025. **(Top 5% in Major)**
- Honor of Excellence Graduation Thesis, 2025. **(Top 3% in Major)**
- Shanghai Municipal Bronze Award in the Internet+ University Student Innovation and Entrepreneurship Competition. **(0.13%)**

Work Experience

LLMs Technical Researcher Feb.2025 - Present
QingTian Lab, Cardinal Operations Shanghai, China

- `Research`: Researched acceleration algorithms and deployed high-performance inference engines using vLLM.
- `Infrastructure Management`: Maintained GPU clusters for large-scale model training and serving.

Research Interests

Mathematical Optimization, Large-Scale Optimization, AI for Optimization, Language Model Architectures, ML System

Publications

Large-Scale Optimization

- [P1] A Restarted Primal-Dual Hybrid Conjugate Gradient Method for Large-Scale Quadratic Programming (**Accepted by INFORMS Journal on Computing**)
  Y. Huang, W. Zhang, **H. Li**, D. Ge, H. Liu, and Y. Ye. (2024). *arXiv preprint* ([Paper])([Code])
- [P2] A Scalable First-Order Method for Large-Scale Competitive Market Equilibrium Computation (**Submitted to INFORMS Journal on Computing**)
  H. Liu, Y. Huang, **H. Li**, D. Ge and Y. Ye (2025). *arXiv preprint* ([Paper])([Code])

Deep Learning Application

- [P3] A Learning and Searching Method for Integrated Processing Planning and Scheduling (**Submitted to Production and Operations Management**)
  H. Li, H. Zhang, Z. He, Y. Jia, B. Jiang, X. Huang, and D. Ge. (2024). *arXiv preprint* ([Paper])([Code])
- [P4] BenLOC: A Benchmark for Learning to Configure MIP Optimizers
  H. Li, Z. He, Y. Wang, S. Pu, Q. Deng, W. Tu, and D. Ge. (2025). *arXiv preprint* ([Paper])([Code])
- [P5] FMIP: Joint Continuous-Integer Flow for Mixed-Integer Linear Programming (**Submitted to ICLR 2026, Rank 20%**)
  H. Li, H. Yuan, H. Zhang, J. Lin, D. Ge, M. Wang and Y. Ye (2025). *arXiv preprint* ([Paper])([Code])

ML System

- [P6] OptPipe: Memory- and Scheduling-Optimized Pipeline Parallelism for LLM Training
  H. Li, H. Zhang, H. Liu, D. Ge and Y. Ye (2025). *arXiv preprint* ([Paper])

Selected Research Experience

Large-Scale Optimization

Improvement of Primal-Dual Hybrid Gradient Method For Large-Scale Linear Programming Oct.2025 - Present
Adviser: Prof. Haihao Lu Massachusetts Institute of Technology
Further improvement the efficiency and design post-processing phase of cuPDLPx, a GPU-accelerated primal-dual hybrid gradient method for large-scale linear programming problems.

- Develop fused GPU kernels, specified for the features of problem instances, to further improve the efficiency of cuPDLPx.
- Design a post-processing phase to further improve the solution quality of cuPDLPx after the major PDHG iterations.

GPU-Accelerated First-Order Method for the Robotic Control via SOS Relaxation Jun.2025 - Present
Adviser: Prof. Yinyu Ye, Prof. Huikang Liu Stanford University
Developing a high-performance first-order solver for large-scale Semidefinite Programming (SDP) arising from Sum-of-Squares (SOS) relaxations in robotic control and trajectory planning.
- Implement whole algorithm framework and solver components.
- Design efficient multi-GPU PSD projection algorithm and other fused GPU kernels.

GPU-Accelerated First-Order Method for Large-Scale Fisher Equilibrium Problems [P2] Nov.2024 - Apr.2025
Adviser: Prof. Yinyu Ye, Prof. Huikang Liu Stanford University
An efficient and GPU-accelerated first order method for large-scale Fisher equilibrium problems.
- Implement efficient GPU kernels for efficiency improvement.
- Develop a CUDA-C based solver for large-scale Fisher equilibrium problems.

Primal-Dual Hybrid Conjugate Gradient Method (PDHCG) [P1] Apr.2024 - Oct.2024
Adviser: Prof. Yinyu Ye, Prof. Dongdong Ge, Prof. Huikang Liu Stanford University, Shanghai Jiao Tong University
An efficient and GPU-accelerated first order method for large-scale quadratic programming problems.
- Implement GPU version and low-rank acceleration of PDHCG solver.
- Implement computational techniques on GPU, including asynchronous computation, fused kernels, memory access optimization.

Deep Learning Application

Generative Models for Linear Programming&Mixed Integer Programming [P5] Sep.2024 - May.2025
Adviser: Prof. Mengdi Wang, Prof. Yinyu Ye Princeton University, Stanford University
Use generative models to accelerate the solving of linear programming (LP) and mixed-integer programming (MIP) problems.
- Build a framework for generating high qulified solution of LPs and MILPs using flow matching.
- Design a joint continuous-integer flow for MILP problems and a optimality-aware training-free guidance strategy.

Deep Reinforcement Learning (DRL) for Scheduling Problems [P3] Mar.2024 - Aug.2024
Adviser: Prof. Dongdong Ge, Prof. Bo Jiang Shanghai Jiao Tong University, SUFE, Cardinal Optimizer
Use DRL to solve the integrated process planning and scheduling problem, a kind of realistic and difficult scheduling problem.
- Implement a GPU-based simulation environment for training and Design a novel graph representation of the problem, a dense reward function and a real-time action space reduction method.
- Design a pruning strategy to improve learning-guided searching method.

Machine Learning for MIP Optimizer Configuration [P4] Dec.2023 - Oct.2024
Adviser: Prof. Dongdong Ge Shanghai Jiao Tong University
Use machine learning to help configure MIP optimizers for better performance.
- Build a whole framework from dataset generation, feature engineering, model training to deployment.
- Design GNNs to predict the best configuration of MIP optimizers.

ML System

Improve Pipeline Parallelsim using Mathematic Programming [P6] Mar.2025 - Present
Adviser: Prof. Yinyu Ye Stanford University
Use mathematic programming to optimize the memory and scheduling of pipeline parallelism in LLM training.
- Implement a framework based on Megatron-LM, which including building and solving a mixed-integer programming model.
- Design practical techniques to reduce required solving time and improve the solution quality.

Volunteer Experience

Shanghai University of Finance and Economics Shanghai, China
- `Teaching Assistant`: Advanced Operations Research (Spring '24), Object-Oriented Analysis & Design (Fall '23), Programming Peer Tutor (Fall '22).
- `Service & Volunteering`: Member of Students' Union (2021-2022); Campus Return Recruitment Volunteer (Winter 2021).

Referees

Yinyu Ye
Professor Emeritus
Department of Management Science & Engineering
Stanford University
yyye@stanford.edu

Haihao Lu
Assistant Professor, Operations Research and Statistics
Sloan School of Management
Massachusetts Institute of Technology
haihao@mit.edu

Dongdong Ge
Distinguished Professor, Antai Collegee of Economics and Management
Dean, Institute of Intelligent Computing
Shanghai Jiao Tong University
ddge@sjtu.edu.cn