

Yushun Zhang

✉ yushunzhang@link.cuhk.edu.cn; yushunz@stanford.edu

🌐 Personal website: zyushun.github.io

📖 Google scholar: ≥ 900 citations (Research papers only; technical reports excluded)

🐙 GitHub: <https://github.com/zyushun>

Professional Experience

- 09/2026 - 📌 **Postdoc.**, Stanford University, MS & E Department
Mentor: Madeleine Udell
- 03/2026 - 📌 **Intern**, DeepSeek AI, Backbone Team
Mentor: Huazuo Gao
- 07/2025 - 02/2026 📌 **Intern**, Moonshot AI (Kimi), Scaling Team
Mentor: Jianlin Su

Education

- 09/2019 - 03/2026 📌 **Ph.D.**, The Chinese University of Hong Kong
School of Data Science
Thesis: *On the Principles Behind Neural Network Optimizers*
Advisor: Zhi-Quan (Tom) Luo
GPA: 3.936 / 4. Ranking: unknown
- 09/2015 - 06/2019 📌 **B.S.**, Southern University of Science and Technology
Department of Mathematics
Magna Cum Laude
Outstanding Graduation Thesis Award
Scholarship Award for Excellence (Top 10 Students) in the Department of Mathematics
GPA: 3.81 / 4. Ranking: 2 / 40
- 03/2019 - 06/2019 📌 **Visiting student**, University of California, San Diego
Department of Mathematics
GPA: 4 / 4. Ranking: unknown
- 09/2012 - 06/2015 📌 Shenzhen Foreign Language School

Research Areas: an Overview

I recently obtained the Ph.D. degree in Data Science from The Chinese University of Hong Kong, working with Prof. Zhi-Quan (Tom) Luo at CUHK-Shenzhen. I am also deeply fortunate to have a long-standing and close collaboration with Prof. Ruoyu Sun at CUHK-Shenzhen during my Ph.D. study.

My research interests lie in optimization theory & algorithms, and their applications in large-scale AI model training. My Ph.D. thesis is titled "*On the Principles Behind Neural Network Optimizers*". The thesis develops theoretical foundations for modern NN optimizers and uses these insights to design better methods.

Below, I highlight key contributions from my previous research (by topics):

- **Converge or diverge? A story of Adam.** Adam is the de facto optimizer for training GenAI models including LLMs, with over 230k citations. An influential prior work ([[ICLR 2018 Best Paper](#)]) showed Adam can diverge, raising concerns about its reliability in multi-million-dollar training runs for GenAI models. We identify a key mismatch between the divergence example and practical settings, and prove that Adam converges without modification under problem-dependent hyperparameters. This result turns Adam from a fragile heuristic into a more reliable and predictable method.

Math tools: new concentration results for stochastic non-linear dynamics.

Impact: This work is mentioned at the [\[ICLR 2025 Test of Time Award Speech\]](#) as the convergence guarantee of Adam optimizer. I am also extremely honored to be invited by the inventor of Adam (Dr. Durk Kingma) to give a [\[1-hour talk\]](#) of this work at Google Brain.

Our theory also provides guidance for tuning Adam’s hyperparams β_1, β_2 . These suggestions are adopted by recent empirical studies, which observe improved LLM pretraining performance when applying them. (e.g., [\[work by Yair Carmon team\]](#), and [\[work by Harvard & Amazon team\]](#)).

1. **Zhang**, Chen, Shi, Luo, and Sun. "Adam Can Converge Without Any Modification On Update Rules." NeurIPS 2022 **Spotlight** (5% of 10k+ submissions).
2. **Zhang**, Li, Chen, Sun, and Luo. "Adam Converges Without Any Modification On Update Rules." This is the extended journal version. Under Review at Mathematical Programming.

- **Hessian of LLMs and Adam-mini.** To deepen our understanding of Adam and LLM training, we explore the special Hessian structure & spectrum of LLMs and general NNs. Through the lens of Hessian, we explain why Transformers need Adam (NeurIPS 2024), and we further theoretically quantify when and why the special Hessian structure occurs (under review at TIT).

By collecting these insights, we design Adam-mini, a memory-efficient optimizer that matches or improves Adam’s performance with **50%** less memory footprint (ICLR 2025). Adam-mini not only can bring savings in training cost and energy, but also democratizes access to AI research and encourages participation from more diverse researchers, especially those with limited GPU resources.

Math tools: Random matrix theory; Quadrature methods.

Reproducibility and open-source software: The effectiveness of Adam-mini is independently reproduced by [\[Recent benchmark by Stanford\]](#) with the conclusion that Adam-mini “*closely track the performance of AdamW*” and “*sometimes even perform better than AdamW*”.

We open sourced Adam-mini on [\[GitHub\]](#) and have received 400 + stars, along with sustained downloads (3000+ monthly) via [\[Pypi\]](#). Adam-mini is also used by diverse researchers, including [\[particle physicists\]](#) who find that AdamW is “*memory-intensive*”, and all experiments are run by Adam-mini. Researchers beyond our labs are also using the code-base of Adam-mini to conduct other research, such as [\[architecture design\]](#). All these show that Adam-mini enables scientists from diverse fields to carry out research, especially under limited GPU resources.

Finally, Adam-mini is acknowledged at the [\[ICLR 2025 Test of Time Award Speech\]](#) for its contribution to reducing the memory footprint of Adam.

1. **Zhang**, Chen, Ding, Li, Sun, and Luo. "Why Transformers Need Adam: A Hessian Perspective." NeurIPS 2024.
2. Dong*, **Zhang***, Yao, Sun. "Towards Quantifying the Hessian Structure of Neural Networks." (*: Equal contribution. Alphabetically ordered.) Under review at Transaction on Information Theory.
3. **Zhang**, Chen, Li, Ding, Wu, Kingma, Ye, Luo, and Sun. "Adam-mini: Use Fewer Learning Rates To Gain More." ICLR 2025.

- **Optimization under finite budget.** In modern engineering scenarios, there is often a strict time limit for the optimization solvers. This raises the question of optimal algorithmic configuration for a fixed and finite time budget. To meet this need, we re-design the classic stepsize rules of the primal-dual methods for Linear Programming (LP). We also discover faster algorithms for special matrix multiplication in the form of $XX^\top$, reducing the number of operations by 5% for both small and large-scaled matrices.

Tools: The algorithmic discovery is assisted by RL + optimization solvers.

1. **Zhang**, Rybin, and Luo. "Finite Horizon Optimization: Framework and Applications." Preprint.
2. Rybin, **Zhang**, and Luo. " XX^t Can Be Faster" Preprint.

Industrial Technical Reports

- 1 [DeepSeek V4: Towards Highly Efficient Million-Token Context Intelligence](#)
DeepSeek AI
- 2 [Kimi K2.5: Visual Agentic Intelligence](#)
Moonshot AI

Selected Publication

- 3 [Adam Converges Without Any Modification On Update Rules](#)
Yushun Zhang, Bingran Li, Congliang Chen, Zhi-Quan Luo, Ruoyu Sun
Under Review at Mathematical Programming
- 4 [Towards Quantifying the Hessian Structure of Neural Networks](#)
Zhaorui Dong*, **Yushun Zhang***, Jianfeng Yao, Ruoyu Sun
(*: Equal contribution. Alphabetically ordered.)
Under review at Transaction on Information Theory
- 5 [XX^t Can Be Faster](#)
Dmitry Rybin, **Yushun Zhang**, Zhi-Quan Luo
Preprint
- 6 [Finite Horizon Optimization: Framework and Applications](#)
Yushun Zhang, Dmitry Rybin, Zhi-Quan Luo
Preprint
- 7 [Adam-mini: Use Fewer Learning Rates To Gain More](#)
Yushun Zhang*, Congliang Chen*, Ziniu Li, Tian Ding, Chenwei Wu, Diederik P. Kingma, Yinyu Ye, Zhi-Quan Luo, Ruoyu Sun
(*: Equal contribution.)
(This work is acknowledged at the [\[2025 Test of Time Award Speech\]](#) for its contribution to reducing the memory footprint of Adam.)
ICLR 2025
- 8 [Why Transformers Need Adam: A Hessian Perspective](#)
Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, Zhi-Quan Luo
NeurIPS 2024
- 9 [Adam Can Converge Without Any Modification on Update Rules](#)
Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, Zhi-Quan Luo
(This work is acknowledged at the [\[2025 Test of Time Award Speech\]](#) as the convergence guarantee of Adam optimizer.)
NeurIPS 2022 **Spotlight** (5% of 10k+ submissions)
- 10 [Does Adam Converge and When?](#)
Yushun Zhang, Congliang Chen, Zhi-Quan Luo
ICLR Blog Track 2022
- 11 [When Expressivity Meets Trainability: Fewer than n Neurons Can Work](#)
Jiawei Zhang*, **Yushun Zhang***, Mingyi Hong, Ruoyu Sun, Zhi-Quan Luo
(*: Equal contribution. Alphabetically ordered.)
NeurIPS 2021

Full Publication List (By Time)

- 1 [Adam Converges Without Any Modification On Update Rules](#)
Yushun Zhang, Bingran Li, Congliang Chen, Zhi-Quan Luo, Ruoyu Sun
Under Review at Mathematical Programming
- 2 [Towards Quantifying the Hessian Structure of Neural Networks](#)
Zhaorui Dong*, **Yushun Zhang***, Jianfeng Yao, Ruoyu Sun
(*: Equal contribution. Alphabetically ordered.)
Under review at Transaction on Information Theory
- 3 [XX^t Can Be Faster](#)
Dmitry Rybin, **Yushun Zhang**, Zhi-Quan Luo
Preprint
- 4 [MoFO: Momentum-Filtered Optimizer for Mitigating Forgetting in LLM Fine-Tuning](#)
Yupeng Chen, Senmiao Wang, **Yushun Zhang**, Tian Ding, Ruoyu Sun
TMLR 2025
- 5 [Exploring and Improving Initialization for Deep Graph Neural Networks: A Signal Propagation Perspective](#)
Senmiao Wang, Yupeng Chen, **Yushun Zhang**, Ruoyu Sun, Tian Ding
TMLR 2025
- 6 [Finite Horizon Optimization: Framework and Applications](#)
Yushun Zhang, Dmitry Rybin, Zhi-Quan Luo
ArXiv Preprint
- 7 [Adam-mini: Use Fewer Learning Rates To Gain More](#)
Yushun Zhang*, Congliang Chen*, Ziniu Li, Tian Ding, Chenwei Wu, Diederik P. Kingma, Yinyu Ye, Zhi-Quan Luo, Ruoyu Sun
(*: Equal contribution.)
ICLR 2025
- 8 [Why Transformers Need Adam: A Hessian Perspective](#)
Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, Zhi-Quan Luo
NeurIPS 2024
- 9 [ReMax: A Simple, Effective, and Efficient Reinforcement Learning Method for Aligning Large Language Models](#)
Ziniu Li, Tian Xu, **Yushun Zhang**, Zhihang Lin, Yang Yu, Ruoyu Sun, Zhi-Quan Luo
ICML 2024
- 10 [Provable Adaptivity of Adam under Non-Uniform Smoothness](#)
Bohan Wang*, **Yushun Zhang***, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, Zhi-Quan Luo, Tie-Yan Liu, Wei Chen
(*: Equal contribution.)
KDD 2024
- 11 [HyperDQN: A Randomized Exploration Method for Deep Reinforcement Learning](#)
Ziniu Li, Yingru Li, **Yushun Zhang**, Tong Zhang, Zhi-Quan Luo
ICLR 2022
- 12 [Adam Can Converge Without Any Modification on Update Rules](#)
Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, Zhi-Quan Luo
NeurIPS 2022 **Spotlight** (5% of 10k+ submissions)
- 13 [Does Adam Converge and When?](#)
Yushun Zhang, Congliang Chen, Zhi-Quan Luo
ICLR Blog Track 2022

- 14 **When Expressivity Meets Trainability: Fewer than n Neurons Can Work**
Jiawei Zhang*, **Yushun Zhang***, Mingyi Hong, Ruoyu Sun, Zhi-Quan Luo
(*: Equal contribution. Alphabetically ordered.)
NeurIPS 2021

Invited Talks

- Jun 2026  I gave a talk at OP 2026, hosted by [Prof. Xiao Li](#).
 - Topic: Towards Quantifying the Hessian Structure of Neural Networks
- Jan 2026  I gave two talks at SJTU and CUHK. The talk at SJTU is hosted by [Prof. Yinyu Ye](#).
 - Topic: Towards Quantifying the Hessian Structure of Neural Networks
- Jul 2025  I gave a talk at ICCOPT 2025, hosted by [Yiping Lu](#).
 - Topic: Towards Quantifying the Hessian Structure of Neural Networks
 - Slides can be seen [here](#)
- Jun 2025  I gave a talk at FAI seminar, hosted by [Jiaye Teng](#).
 - Topic: Towards Quantifying the Hessian Structure of Neural Networks
 - Slides can be seen [here](#)
- Dec 2024  I gave a talk at Tsinghua University, hosted by [Kaifeng Lyu](#).
 - Topic: Adam for Transformers: Why and Why Not
 - Slides can be seen [here](#)
- Oct 2024  I gave a talk at University of Minnesota, hosted by [Prof. Mingyi Hong](#).
 - Topic: Why Transformers Need Adam: A Hessian Perspective
 - Slides can be seen [here](#) (this is a >40 min version with Adam-mini included) I gave a talk at INFORMS Annual Meeting, Seattle, hosted by [Jianhao Ma](#).
 - Topic: Why Transformers Need Adam: A Hessian Perspective
 - Slides can be seen [here](#) (this is a 15 min version)
- Sep 2023  I gave a talk at Tsinghua University, hosted by [Prof. Jian Li](#).
 - Topic: Converge or Diverge? A Story of Adam
 - Slides can be seen [here](#)
- Jan 2023  I gave a talk at Google Brain, hosted by [Dr. Diederik P. Kingma](#).
 - Topic: Adam Can Converge Without Any Modification on Update Rules
 - Slides can be seen [here](#)

Awards

- Dec 2023  Duan Yongping Outstanding Research Award (1st place)
 Excellent TA Award (2nd place), School of Data Science
- Aug 2022  Best Paper Presentation Award (1st place), 2nd Doctoral and Postdoctoral Daoyuan Academic Forum
(Topic: Does Adam Converge and When?)
- Jul 2021  Best Paper Presentation Award (1st place), 3rd Tsinghua-Berkeley workshop on Learning Theory
(Topic: When Expressivity Meets Trainability: Width $< n$ Can Work)
- Jul 2019  Magna Cum Laude of SUSTech
- Jun 2019  Outstanding Graduation Thesis, SUSTech
- Sep 2018  Scholarship Award for Excellence (Top 10 Students), Department of Mathematics, SUSTech

Teaching

I serve as Teaching Assistant and for the following courses (listed by time). For each course, I design and deliver tutorials independently.

- DDA4300: Optimization for Machine Learning, by Prof. Yinyu Ye (2025 Spring)
- DDA4300: Optimization for Machine Learning, by Prof. Yinyu Ye (2023 Spring)
- DDA 6060: Machine Learning, by Prof. Hongyuan Zha & Prof. Shuang Li (2022 Spring)
- DDA 4002: Multivariate Statistics, by Prof. Zhaoyuan Li (2021 Autumn)
- DDA 4250: Mathematics for Deep Learning, by Prof. Arnulf Jentzen (2021 Spring)
- MFE 5100: Optimization, by Prof. Zizhuo Wang (2020 Autumn)
- STA 2002: Probability and Statistics, by Prof. Xinyun Chen (2020 Summer)
- CSC 4020: Fundamentals of Machine Learning, by Prof. Hongyuan Zha (2020 Spring)
- MAT 2040: Linear algebra, by Prof. Shenghao Yang (2019 Autumn)

Service

- I hosted a session [[Optimization Issues in Recent AI Models](#)] at INFORMS Annual Meeting, Oct, 2024.
- I serve as a reviewer for machine learning conferences including NeurIPS, ICLR, ICML, COLT, as well as journals including JMLR and TMLR.