Runlong Zhou (周润龙)

July 7, 2025

Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA 98195, USA
`vectorzh@cs.washington.edu`, `zhourunlongvector@gmail.com`
vectorzhou.com

Research Interests

- Reinforcement learning (RL) + large language models (LLMs)
- RL theory

Education

- **University of Washington** — Seattle, USA
  PhD, Paul G. Allen School of Computer Science & Engineering, advised by Prof. Simon S. Du — 2022.9 - Now (Est. 2027.8)
- **Tsinghua University** — Beijing, China
  BEng, Yao Class, Institute for Interdisciplinary Information Sciences (IIIS) — 2018.8 - 2022.6

Academic Experience

- **Apple AIML** — Seattle, USA
  Research intern with Dr. Lefan Zhang and Xuan Kelvin Zou — 2025.6 - Now
- **Microsoft Research** — Redmond, USA
  Research intern with Dr. Baolin Peng and Dr. Hao Cheng — 2025.2 - 2025.6
- **Microsoft Research** — Redmond, USA
  Research intern with Dr. Yi Zhang — 2024.6 - 2024.8
- **Microsoft Research** — Redmond, USA
  Research intern with Dr. Beibin Li — 2023.6 - 2023.9
- **University of Washington** — Remote
  Research intern with Prof. Simon S. Du — 2020.9 - 2022.9
- **Facebook AI Research** — Remote
  Research intern with Dr. Alessandro Lazaric and Dr. Matteo Pirotta — 2021.3 - 2021.5

Publications

* denotes equal contribution or alphabetical ordering.

1. **CASCADE Your Datasets for Cross-Mode Knowledge Retrieval of Language Models** [Link]
   Runlong Zhou, Yi Zhang
   COLM 2025 — Poster
   We qualitatively identify and quantitatively formulate the problem of cross-mode knowledge retrieval of LLMs, and propose CASCADE algorithm to improve the performance by allowing the LLMs to capture knowledge with arbitrary lengths and occurrence locations.
2. **Extragradient Preference Optimization (EGPO): Beyond Last-Iterate Convergence for Nash Learning from Human Feedback** [Link]
   Runlong Zhou, Maryam Fazel, Simon S. Du
   COLM 2025 — Poster
   We propose EGPO algorithm for Nash learning from human feedback, achieving a last-iterate linear convergence and a simple online IPO implementation.

3. **Transformers are Efficient Compilers, Provably** [Link]
   Xiyu Zhai, **Runlong Zhou**, Liao Zhang, Simon S. Du
   COLM 2025 Poster
   We propose Cybertron as a proof vehicle for transformers' expressive ability and show that for a compilation task, transformers need only a logarithm number of parameters while any recurrent neural network needs at least a linear number of parameters.
4. **The Crucial Role of Samplers in Online Direct Preference Optimization** [Link]
   Ruizhe Shi*, **Runlong Zhou***, Simon S. Du
   ICLR 2025 Poster
   We prove that online DPO with a mixture of samplers achieves quadratic convergence with exact gradients and linear convergence with estimations.
5. **Reflect-RL: Two-Player Online RL Fine-Tuning for LMs** [Link]
   Runlong Zhou, Simon S. Du, Beibin Li
   ACL 2024 Poster
   We developed Reflect-RL, a two-player system to align language models with interactive decision-making tasks. Techniques included are reflection, negative example generation, single-prompt action enumeration, and curriculum learning.
6. **Free from Bellman Completeness: Trajectory Stitching via Model-based Return-conditioned Supervised Learning** [Link]
   Zhaoyi Zhou, Chuning Zhu, **Runlong Zhou**, Qiwen Cui, Abhishek Gupta, Simon S. Du
   ICLR 2024 Poster
   We study the strengths and weaknesses of return-conditioned supervised learning, and propose an empirically improved algorithm.
7. **Sharp Variance-Dependent Bounds in Reinforcement Learning: Best of Both Worlds in Stochastic and Deterministic Environments** [Link]
   Runlong Zhou, Zihan Zhang, Simon S. Du
   ICML 2023 Poster
   We provide a systematic study of variance-dependent regret bounds of model-based and model-free reinforcement learning for tabular MDPs. The proposed model-based algorithm is both optimal for stochastic and deterministic MDPs.
8. **Variance-Dependent and Horizon-Free Reinforcement Learning for Latent Markov Decision Processes** [Link]
   Runlong Zhou, Ruosong Wang, Simon S. Du
   ICML 2023 Poster
   We provide an algorithm framework for Latent MDPs (with context in hindsight), achieving the first horizon-free minimax regret. We complement the study by giving a novel regret lower bound for LMDPs using the symmetrization technique.
9. **Understanding Curriculum Learning in Policy Optimization for Online Combinatorial Optimization** [Link]
   Runlong Zhou, Zelin He, Yuandong Tian, Yi Wu, Simon S. Du
   TMLR
   We formulate of canonical online Combinatorial Optimization problems as Latent MDPs and give convergence guarantee of Natural Policy Gradient on LMDPs. We show effectiveness of Curriculum Learning through the perspective of relative conditional number.
10. **Stochastic Shortest Path: Minimax, Parameter-Free and Towards Horizon-Free Regret** [Link]
    Jean Tarbouriech*, **Runlong Zhou***, Simon S. Du, Matteo Pirotta, Michal Valko, Alessandro Lazaric
    NeurIPS 2021 Spotlight, 3% acceptance rate
    We propose an algorithm (EB-SSP) for SSP problems, which is the first to achieve minimax optimal regret while being parameter-free.

Preprints

* denotes equal contribution or alphabetical ordering.

1. **Sharp Gap-Dependent Variance-Aware Regret Bounds for Tabular MDPs** [Link]
   Shulun Chen, **Runlong Zhou**, Zihan Zhang, Maryam Fazel, Simon S. Du
   We propose a novel analysis of gap-dependent regrets by introducing a necessary term named maximum conditional total variance. The proposed model-based algorithm is tight on both the horizon and the variance dependencies.
2. **Understanding the Performance Gap in Preference Learning: A Dichotomy of RLHF and DPO** [Link]
   Ruizhe Shi*, Minhak Song*, **Runlong Zhou**, Zihan Zhang, Maryam Fazel, Simon S. Du
   We theoretically study the separation between RLHF and DPO when the optimization step is exact while the policy model and reward model are differently mis-specified, and when only finite samples are accessible.

3. **Multi-Agent Reinforcement Learning from Human Feedback: Data Coverage and Algorithmic Techniques** [Link]
   Natalia Zhang*, Xinqi Wang*, Qiwen Cui*, **Runlong Zhou**, Sham M. Kakade, Simon S. Du
   We study Multi-Agent Reinforcement Learning from Human Feedback (MARLHF) by exploring both theoretical foundations and empirical validations. Included in our proposed methods are Mean Squared Error (MSE) regularization and imitation learning.

Awards, Grants & Honors

- **Graduate school:**

Institute for Foundations of Data Science (IFDS) Research Assistant	2025

- **Undergraduate school:**

IIIS Outstanding Graduate	2022
The 2021 China Collegiate Programming Contest, Guilin Site (Gold Medal)	2021
IIIS Research Innovation Scholarship	2021
IIIS Academic Performance Scholarship	2021
Tsinghua University Air Rifle Competition (First Place)	2021
The 2019 ACM-ICPC Asia Regional Contest, Xuzhou Site (Gold Medal)	2019
The 2018 ACM-ICPC Asia Regional Contest, Beijing Site (Gold Medal)	2018

- **Secondary school:**

The 34th National Olympiad in Informatics (Silver Medal)	2017
China Team Selection Competition 2017 (Gold Medal)	2017
The 2016 ACM-ICPC Asia CHINA-Final Contest (Gold Medal)	2016
The 2016 China Collegiate Programming Contest Finals (Silver Medal)	2016
The 33rd National Olympiad in Informatics (Silver Medal)	2016

Other Projects

1. **Ray Tracing Renderer** [Link]
   Runlong Zhou
   Advanced Computer Graphics course project C++
   Optimized path tracing framework supporting Mitsuba configurations, many textures and sampling methods

Miscellanea

- **Professional skills:** Algorithm design, Data structures, Deep (Reinforcement) learning

- **Programming skills:** C++ / C, python, LaTeX, HTML, JavaScript (Node.js), CUDA, Java, MATLAB, Rust

- **Hobbies:** Air rifle / pistol shooting, Archery

- **Teaching:** Teaching Olympiad in Informatics to secondary school students between 2018 and 2021. Teaching assistant of UW CSE 543 Deep Learning, CSE 446 Machine Learning.

- **Reviewing:** TMLR, JMLR, ICML 2022, NeurIPS 2022, ICML 2024, COLT 2024, ICLR 2025, ICML 2025, NeurIPS 2025.