WEI XIONG

Computer Science, University of Illinois Urbana-Champaign
wx13@illinois.edu | Website | GitHub

RESEARCH INTERESTS

My research interests focus on reinforcement learning from human feedback (RLHF) for aligning large language model recently.

Previously, I have spent time on the mathematical foundation of RL, where I was fortunate to collaborate with many great senior mentors and talented peers. I also spent time on deep RL at Microsoft Research Asia.

EDUCATION

University of Illinois Urbana-Champaign *Urbana, USA*
PhD student, *Department of Computer Science* *2023.8 - present*
Advisor: Prof. Tong Zhang and Prof. Nan Jiang

The Hong Kong University of Science and Technology *Hong Kong, China*
Master of Philosophy, *Department of Mathematics* *2023.8*
Advisor: Prof. Tong Zhang

University of Science and Technology of China *Hefei, China*
Bachelor of Science, *Department of Mathematics* *2021.6*
Department of Electronic Engineering
Ranking: 1/72 in Statistics; 2/352 in EE.

Shanghai Jiao Tong University *Shanghai, China*
Exchange student at School of Electronic Information and Electrical Engineering *2018*

EXPERIENCE

University of Illinois Urbana-Champaign: *2024.2-present*
Advisor Prof. Tong Zhang
Worked as a core founding member of the RLHFlow project, which presents a full recipe for the worklow of online iterative RLHF, including SFT, reward/preference modeling, and iterative RLHF/DPO. The resulting reward/preference functions are the state-of-the-art open-source models, and the final LLM achieves comparable or even better performance compared to LLaMA3-8B-instruct.

The Hong Kong University of Science and Technology: *2023.1-present*
Advisor Prof. Tong Zhang
Worked as a core founding member of the LMFlow project, which is a framework that allows developing LLMs (fine-tuning, inference, RLHF...) with minimal cost and effort. The project received 7K+ star in github and ranked 2nd in the github trend. I am responsible for developing the RLHF part of the project.

Yale University: *2022 Spring*
Virtual visit with Prof. Zhuoran Yang
Worked on the mathematical foundation of reinforcement learning.

Microsoft Research Asia (MSRA): *Spring 2021*
Advisor: Dr. Wenxue Cheng and Dr. Li Zhao
Intern: Networking Research and Machine Learning Group

Worked on bandwidth estimation for real-time communications with reinforcement learning.

University of Virginia: *2019.8 - 2019.11*
Advisior: Prof. Cong Shen
Research Assistant: worked on multi-player multi-armed bandit (MPMAB).

SELECTED AWARDS AND FELLOWSHIPS

Hong Kong PhD Fellowship	*2021-2023*
Best Teaching Assistant Award at HKUST	*June 2022,2023*
Outstanding graduate (USTC and Anhui province)	*June 2021*
Nomination for Guo Moruo scholarship (1/72 in statistics, highest honor of USTC)	*November 2020*
Yuanqing Yang Scholarship	*November 2020*
Chinese Academy of Sciences Institute of Electronics Scholarship	*October 2018*
National Scholarship	*October 2017*
Honor Program in EE/AI at USTC	*2017 - 2019*
Zhuang Caifang Scholarship	*July 2016*

SELECTED PUBLICATIONS AND MANUSCRIPTS

See full list in Google Scholar.

(α, β) denotes random/alphabetical order and * denotes equal contribution

[1] (α, β) Hanze Dong*, Wei Xiong*, Bo Pang*, Haoxiang Wang*, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, Tong Zhang, "RLHF Workflow: From Reward Modeling to Online RLHF", We present the workflow of online iterative RLHF, which is widely reported to outperform its offline counterpart by a large margin in the recent LLM literature. However, existing open-source RLHF projects are still largely confined to the offline learning setting. In this repo, we aim to fill in this gap and provide a detailed recipe that is easy to be reproduced for online iterative RLHF. In particular, with our recipe, with only open-source data, we can achieve comparable or even better results than LLaMA3-8B-instruct [Code];

[2] Wei Xiong*, Hanze Dong*, Chenlu Ye*, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, Tong Zhang, "Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-Constraint", we formulate the real-world RLHF process as a reverse-KL regularized contextual bandits and establish the mathematical foundation of this process by proposing statistically efficient algorithms with finite-sample theoretical guarantee. With a reasonable approximation of some information-theoretical oracle, the results naturally lead to several new alignment algorithms, e.g., the iterative DPO and offline DPO with multi-step rejection sampling, which admit an impressive empirical performance and outperform existing strong baselines like DPO, and RSO in real-world LLM alignment experiments, short version is accepted by ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models as Oral presentation [ICML 2024].

[3] Haoxiang Wang*, Yong Lin*, Wei Xiong*, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, Tong Zhang, "Arithmetic Control of LLMs for Diverse User Preferences: Directional Preference Alignment with Multi-Objective Rewards" [Preprint].

[4] (α, β) Chenlu Ye*, Wei Xiong*, Yuheng Zhang*, Nan Jiang, Tong Zhang, "A Theoretical Analysis of Nash Learning from Human Feedback under General KL-Regularized Preference", we present an initial attempt to study the learnability of the KL-regularized NLHF framework, aiming to promote the development of reward-model-free preference learning under general preference oracle [Preprint].

[5] (α, β) Yong Lin*, Hangyu Lin*, Wei Xiong*, Shizhe Diao*, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Yuan Yao, Heng Ji, and Tong Zhang, "Mitigating the Alignment Tax of RLHF", we quantitatively study the alignment tax of popular RLHF algorithms including PPO, RAFT, and DPO and investigate various methods to alleviate the forgetting, including regularization, low-rank finetuning, data replay, reawrd regularization, and model averaging. We also propose adaptive model averaging, which is most competitive across different tasks of the benchmark [Preprint].

[6] (α, β) Hanze Dong*, Wei Xiong*, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum and Tong Zhang, "RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment", we develop a simple but effective alignment framework with minimal hyper-parameter configuration; the proposed framework is friendly in implementation and memory resources, and also interpretable with clear learning objectives. [TMLR] [Code].

[7] Shizhe Diao*, Rui Pan*, Hanze Dong*, KaShun Shen, Jipeng Zhang, Wei Xiong, and Tong Zhang, "LMFlow: An Extensible Toolkit for Finetuning and Inference of Large Foundation Models", we introduce an extensible and lightweight toolkit, LMFlow, which aims to simplify the development of general LLMs; the project received 7K+ GitHub stars and I was responsible for developing the RLHF part of the whole project. [NAACL 2024] [Code].

[8] (α, β) Han Zhong*, Wei Xiong*, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang, "GEC: A Unified Framework for Interactive Decision Making in MDP, POMDP, and Beyond", we measure the hardness of the sequential decision making problem as the coefficient to generalize in the online manner and show that the online problems in this framework can be reduced to an offline supervised learning in terms of in-sample error estimation; the proposed framework captures most of known trackable MAB, Contextual Banit, MDP and POMDP problems; a generalized posterior sampling framework is also provided. [Under Major Revision at Mathematical Operation Research (MOR)] [Slide].

[9] Wei Xiong, "A Sufficient Condition of Sample-Efficient Reinforcement Learning with General Function Approximation", *Master Thesis*, we develop GEC in this thesis with a thorough description of the motivation and application; we also develop a new optimization-based framework, as a counterpart of the sampling framework in original GEC paper.

[10] Wei Xiong*, Han Zhong*, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang,"Nearly Minimax Optimal Offline Reinforcement Learning with Linear Function Approximation: Single-Agent MDP and Markov Game", An application of weighted regression by using the variance information to achieve a sharper bound. , [ICLR 2023].

[11] Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, and Tong Zhang,"A Self-Play Posterior Sampling Algorithm for Zero-Sum Markov Game", this is a straightforward extension of the single-agent conditional posterior sampling, which also provides an extension of the eluder coefficient to the multi-agent case, [ICML 2022].

[12] Han Zhong*, Wei Xiong*, Jiyuan Tan*, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang,"Pessimistic Minimax Value Iteration: Provably Efficient Equilibrium Learning from Offline Datasets", [ICML 2022] [Slide].

[13] Chengshuai Shi, Wei Xiong, Cong Shen, and Jing Yang, "Heterogeneous Multi-player Multi-armed Bandits: Closing the Gap and Generalization", we developed a carefully-crafted exploration strategy in the heterogeneous MPMAB setting, as well as a delicate differential communication scheme; the proposed BEACON achieves the minimax optimal regret bound and also demonstrates an impressive empirical performance. [NeurIPS 2021] [Code].

TEACHING

The Hong Kong University of Science and Technology: *2021.9-2023.6*
Teaching Assistant: MATH 2421 - Probability, MATH 2121 - Linear Algebra, MATH 6913W - Reading Course: Statistical Learning Theory, MATH 2023 - Multivariable Calculus (**Best TA Award for all courses**).

University of Science and Technology of China: *2018-2021*
Teaching Assistant: Mathematical Statistics, Data Structures and Databases, Algorithms and Data Structures.

PROFESSIONAL ACTIVITY

Conference Reviewer:
ICLR 2024; Neurips 2022, 2023, 2024 (**Top Reviewer Award**); ICML 2022, 2023; AISTATS 2023, 2024.

Journal Reviewer:
Machine Learning, JMLR.

Talks:
"Reinforcement Learning from Human Feedback: From Theory to Algorithm", *Google Multi-turn RLHF Workshopt, MTV*, *2024.6*
"Reinforcement Learning from Human Feedback: From Theory to Algorithm", *Google Learning Theory Seminar, NYC*, *2024.5*
"Reinforcement Learning from Human Feedback: From Theory to Algorithm", *Center for Machine Learning Research, Peking University*, *2024.4*
"Reinforcement Learning from Human Feedback: From Theory to Algorithm", *University of Virginia*, host: Cong Shen, *2024.4*
"Alignment for Foundation Language Models: Mathematical Principle and Algorithmic Designs", *Yale University*, host: Zhuoran Yang, *2024.3*
"Alignment for Foundation Language Models: Mathematical Principle and Algorithmic Designs", *University of California, Los Angeles*, host: Quanquan Gu, *2024.2*
"Alignment for Foundation Language Models: Mathematical Principle and Algorithmic Designs", *Microsoft Research Asia*, host: Chuheng Zhang, *2024.1*
"Reinforcement Learning From Human Feedback with Rejection Sampling: A RL-free Approach", *Hong Kong University*, host: Qi Xiaojuan, *2023.8*
"Reinforcement Learning From Human Feedback with Rejection Sampling: A RL-free Approach", *University of Toronto*, host: Qiang Sun, *2023.6*
"Reinforcement Learning From Human Feedback with Rejection Sampling: A RL-free Approach", *Stanford University*, host: Mert Pilanci, *2023.5*