Efficient and Robust Reinforcement Learning from Human Feedback

Huazheng Wang
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR, USA
huazheng.wang@oregonstate.edu

Reinforcement Learning (RL) has emerged as a powerful paradigm for sequential decision-making with numerous real-world applications. However, in practical environments such as recommender systems, search engines, and LLMs, RL algorithms must efficiently learn from biased human feedback that may be subject to corruption. In this talk, I will present our recent efforts in developing robust RL algorithms that can provably effectively handle such challenging scenarios. First, I will introduce our works on reinforcement learning from biased click feedback in ranking. In addition to controlling bias with online learning to rank methods (Wang et al. 2018, 2019; Jia et al. 2021), off-policy methods received high attention where previous approaches typically relied on strong assumptions about human click behavior (formalized as click models) and required specialized debiasing methods for different models. We propose a novel unified framework that formulates the ranking process under general click models as a Markov Decision Process, enabling the development of a click model-agnostic RL algorithm (Zhang et al. 2024). Second, I will introduce the fundamental vulnerability of bandits and reinforcement learning under corrupted feedback (Wang, Xu, and Wang 2022; Balasubramanian et al. 2024; Wang, Wang, and Wang 2024; Wang et al. 2024). Our theoretical analysis provides complete necessity and sufficiency characterizations of the attackability of linear bandits and linear RL, revealing their intrinsic robustness and limitations. Lastly, I will discuss our recent works on improving RL finetuning for LLMs, including sample efficient off-policy RLHF and solving the gradient entanglement issue in margin-based alignment methods (Yuan et al. 2025).

References

Balasubramanian, R.; Li, J.; Tadepalli, P.; Wang, H.; Wu, Q.; and Zhao, H. 2024. Adversarial Attacks on Combinatorial Multi-Armed Bandits. In *Forty-first International Conference on Machine Learning*.

Jia, Y.; Wang, H.; Guo, S.; and Wang, H. 2021. PairRank: Online Pairwise Learning to Rank by Divide-and-Conquer. In *Proceedings of the Web Conference 2021*, 146–157.

Wang, H.; Kim, S.; McCord-Snook, E.; Wu, Q.; and Wang, H. 2019. Variance Reduction in Gradient Exploration for Online Learning to Rank. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, 835–844. ACM.

Wang, H.; Langley, R.; Kim, S.; McCord-Snook, E.; and Wang, H. 2018. Efficient exploration of gradient space for online learning to rank. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 145–154. ACM.

Wang, H.; Xu, H.; and Wang, H. 2022. When Are Linear Stochastic Bandits Attackable? In *International Conference on Machine Learning*, 23254–23273. PMLR.

Wang, Z.; Balasubramanian, R.; Yuan, H.; chenyu song; Wang, M.; and Wang, H. 2024. Adversarial Attacks on Online Learning to Rank with Stochastic Click Models. *Transactions on Machine Learning Research*.

Wang, Z.; Wang, H.; and Wang, H. 2024. Stealthy Adversarial Attacks on Stochastic Multi-Armed Bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15770–15777.

Yuan, H.; Zeng, Y.; Wu, Y.; Wang, H.; Wang, M.; and Leqi, L. 2025. Common Pitfalls of Margin-based Preference Optimization in Language Model Alignment. In *The Thirteenth International Conference on Learning Representations*.

Zhang, Z.; Su, Y.; Yuan, H.; Wu, Y.; Balasubramanian, R.; Wu, Q.; Wang, H.; and Wang, M. 2024. Unified off-policy learning to rank: a reinforcement learning perspective. *Advances in Neural Information Processing Systems*, 36.