Department of Computer Science

Why is RLHF Data-Efficient in Policy Optimization?

Yihan Du
University of Illinois at Urbana-Champaign

Hosted by Jing Li

Date: Tuesday, December 10, 2024
Coffee: 2:15 PM – 2:30 PM
Time: 2:30 PM – 3:30 PM
Location: GITC 2121(2nd floor Seminar Lecture Hall)

Zoom Link: https://njit-edu.zoom.us/j/92739973145?pwd=NmmZuyJawyJsQ4Il9yaDqGpkq6M22z.1

Abstract:
Reinforcement Learning from Human Feedback (RLHF) has achieved a huge empirical success while using a small amount of human preference data. However, there is limited theoretical understanding for this phenomenon. In this work, we consider and analyze the policy optimization-based RLHF algorithmic framework, which has been successfully applied in ChatGPT. In this algorithm, the knowledge of the reward function is not assumed as standard RL, and instead, we use human preference feedback on trajectories to infer the reward function. We provide performance guarantees for this algorithm with low query complexity, and exhibit an insight into why a small amount of human feedback may be sufficient to achieve good performance in RLHF. We design algorithms and provide results for two important settings, i.e., linear and neural function approximation, respectively. In addition, in this talk, Dr. Du will also briefly introduce her research on other popular topics of decision making, including risk-sensitive and safe decision making, and multi-agent (multi-task) decision making.

Bio:
Dr. Yihan Du is currently a postdoctoral researcher at the University of Illinois at Urbana-Champaign, working with Prof. R. Srikant. Her research interests focus on machine learning, with emphases on reinforcement learning and online learning. Dr. Du obtained her Ph.D. degree from the Institute for Interdisciplinary Information Sciences (headed by Turing award winner Prof. Andrew Chi-Chih Yao) at Tsinghua University in 2023. She has published multiple papers on the top conferences in machine learning, including ICML, NeurIPS, ICLR and AAAI. Dr. Du also received several honors, including the China Computer Federation (CCF) Agent and Multi-Agent System Doctoral Dissertation Award, and the Tsinghua Outstanding Doctoral Dissertation Award.