北京理工大學

数学与统计学院学术报告

Safe Alignment of LLMs under Reward Over-Optimization: Mitigating Catastrophic Goodhart Effect via Rényi Regularization

报告人：张慧铭 北京航空航天大学

时　间：2026年11月05日16：00-17：00

地　点：文萃楼E 218

摘　要： This talk introduces a novel information-theoretic framework for analyzing generalization in machine learning algorithms, specifically addressing scenarios involving exponentially decaying heavy-tailed data where traditional KL-based mutual information bounds, reliant on moment generating functions (MGFs), become ineffective. The core issue is that many modern pipelines, such as Robust Learning, Reinforcement Learning from Human Feedback (RLHF), and Stochastic Optimization, exhibit heavy-tailed losses or rewards, for which MGFs may not exist, rendering classical tools vacuous. Even when the KL mutual information is small, rare extreme events can dominate generalization performance. We show that under heavy-tailed rewards, KL divergence-based trust regions in RLHF can suffer from the "Catastrophic Goodhart effect" - where the proxy reward can blow up without bound even when the KL constraint is small. In contrast, RLHF with Rényi regularization effectively controls reward inflation under heavy-tailed sub-Weibull rewards. Furthermore, for the Best-of-n sampling strategy, we show that Rényi-regularized alignment provides finite reward guarantees and ensures that best-of-N policies remain well-controlled. Finally, we successfully apply Rényi regularization to Direct Preference Optimization (DPO) under the Bradley-Terry preference framework for LLMs.

个人简介：

张慧铭，北航人工智能研究院的副教授(准聘)、硕士生导师；北航数学科学学院兼职博导。曾在澳门大学数学系担任濠江学者博后研究员(2020-2022)；曾就读于北京大学(2016-2020)获得统计学博士学位。研究方向：稳健机器学习、泛化误差理论、LLM对齐理论、高维概率统计、集中不等式、子抽样算法、Lévy过程等。发表SCI/顶会论文30+篇(包括AI与自动化领域顶刊JMLR, IEEE-TAC, 顶会NeurIPS；统计顶刊JASA, Biometrika、精算顶刊IME)，谷歌学术引用1100+(其中5篇(曾)为高被引论文)。曾担任美国《数学评论》评论员；概率统计、AI与机器学习领域顶刊(AOS,AOAP,JASA,JMLR,IEEE-TSP)的审稿人。

主办单位：北京理工大学数学与统计学院

School of Mathematics and Statistics, Beijing Institute of Technology