Instance-Dependent Complexity of Contextual Bandits and Reinforcement Learning: A Disagreement-Based Perspective*

Dylan J. Foster   DYLANFOSTER@MICROSOFT.COM
Microsoft Research, New England

Alexander Rakhlin   RAKHLIN@MIT.EDU
Massachusetts Institute of Technology

David Simchi-Levi   DSLEVI@MIT.EDU
Massachusetts Institute of Technology

Yunzong Xu   YXU@MIT.EDU
Massachusetts Institute of Technology

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

In the classical multi-armed bandit problem, *instance-dependent* algorithms attain improved performance on "easy" problems with a gap between the best and second-best arm. Are similar guarantees possible for contextual bandits? While positive results are known for certain special cases, there is no general theory characterizing when and how instance-dependent regret bounds for contextual bandits can be achieved for rich, general classes of policies. We introduce a family of complexity measures that are both sufficient and necessary to obtain instance-dependent regret bounds. We then introduce new oracle-efficient algorithms that adapt to the gap whenever possible, while also attaining the minimax rate in the worst case. Finally, we provide structural results that tie together a number of complexity measures previously proposed throughout contextual bandits, reinforcement learning, and active learning and elucidate their role in determining the optimal instance-dependent regret. In a large-scale empirical evaluation, we find that our approach often gives superior results for challenging exploration problems.

Turning our focus to reinforcement learning with function approximation, we develop new oracle-efficient algorithms for reinforcement learning with rich observations that obtain optimal gap-dependent sample complexity.

Keywords: contextual bandits, reinforcement learning, function approximation, instance dependence, disagreement coefficient, star number, eluder dimension

---