Is Reinforcement Learning More Difficult Than Bandits?
A Near-optimal Algorithm Escaping the Curse of Horizon

Zihan Zhang
Tsinghua University
ZIHAN-ZH17@MAILS.TSINGHUA.EDU.CN

Xiangyang Ji
Tsinghua University
XYJI@TSINGHUA.EDU.CN

Simon S. Du
University of Washington
SSDU@CS.WASHINGTON.EDU

Abstract
1 Episodic reinforcement learning and contextual bandits are two widely studied sequential decision-making problems. Episodic reinforcement learning generalizes contextual bandits and is often perceived to be more difficult due to long planning horizon and unknown state-dependent transitions. The current paper shows that the long planning horizon and the unknown state-dependent transitions (at most) pose little additional difficulty on sample complexity.

We consider the episodic reinforcement learning with $S$ states, $A$ actions, planning horizon $H$, total reward bounded by 1, and the agent plays for $K$ episodes. We propose a new algorithm, Monotonic Value Propagation (MVP), which relies on a new Bernstein-type bonus. Compared to existing bonus constructions, the new bonus is tighter since it is based on a well-designed monotonic value function. In particular, the constants in the bonus should be subtly setting to ensure optimism and monotonicity.

We show MVP enjoys an $O\left(\sqrt{SAK} + S^2 A \right) \log(SAHK)$ regret, approaching the $\Omega\left(\sqrt{SAK}\right)$ lower bound of contextual bandits up to logarithmic terms. Notably, this result 1) exponentially improves the state-of-the-art polynomial-time algorithms by Dann et al. [2019] and Zanette et al. [2019] in terms of the dependency on $H$, and 2) exponentially improves the running time in [Wang et al. 2020] and significantly improves the dependency on $S$, $A$ and $K$ in sample complexity.

References


