Sharper Model-free Reinforcement Learning for Average-reward Markov Decision Processes

Zihan Zhang ZSUBFUNC@OUTLOOK.COM
Princeton University

Qiaomin Xie QIAOMIN.XIE@WISC.EDU
University of Wisconsin-Madison

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We study model-free reinforcement learning (RL) algorithms for infinite-horizon average-reward Markov decision process (MDP), which is more appropriate for applications that involve continuing operations not divided into episodes. In contrast to episodic/discounted MDPs, theoretical understanding of model-free RL algorithms is relatively inadequate for the average-reward setting. In this paper, we consider both the online setting and the setting with access to a simulator. We develop computationally efficient model-free algorithms that achieve sharper guarantees on regret/sample complexity compared with existing results.

In the online setting, we design an algorithm, $\texttt{UCB}-\texttt{AVG}$, based on an optimistic variant of variance-reduced Q-learning. We show that $\texttt{UCB}-\texttt{AVG}$ achieves a regret bound $\widetilde{O}(S^5A^2\,\mathrm{sp}(h^*)\sqrt{T})$ after T steps, where $S\times A$ is the size of state-action space, and $\mathrm{sp}(h^*)$ the span of the optimal bias function.[1] Our result provides the first computationally efficient model-free algorithm that achieves the optimal dependence in T (up to log factors) for weakly communicating MDPs, which is necessary for low regret (Bartlett and Tewari, 2009). In contrast, prior results either are suboptimal in T or require strong assumptions of ergodicity or uniformly mixing of MDPs.

In the simulator setting, we adapt the idea of $\texttt{UCB}-\texttt{AVG}$ to develop a model-free algorithm that finds an ϵ-optimal policy with sample complexity $\widetilde{O}\left(SA\,\mathrm{sp}^2(h^*)\epsilon^{-2}+S^2A\,\mathrm{sp}(h^*)\epsilon^{-1}\right)$. This sample complexity is near-optimal for weakly communicating MDPs, in view of the minimax lower bound $\Omega(SA\,\mathrm{sp}(^*)\epsilon^{-2})$ established in Wang et al. (2022). Existing work mainly focuses on ergodic MDPs and the results typically depend on t_{mix}, the *worst-case* mixing time induced by a policy. We remark that the diameter D and mixing time t_{mix} are both lower bounded by $\mathrm{sp}(h^*)$, and t_{mix} can be arbitrarily large for certain MDPs (Wang et al., 2022).

On the technical side, our approach integrates two key ideas: learning an γ-discounted MDP as an approximation, and leveraging reference-advantage decomposition for variance reduction (Zhang et al., 2020) in optimistic Q-learning. As recognized in prior work, a naive approximation by discounted MDPs results in suboptimal guarantees. A distinguishing feature of our method is maintaining estimates of *value-difference* between state pairs to provide a sharper bound on the variance of reference advantage. We also crucially use a careful choice of the discounted factor γ to balance approximation error due to discounting and the statistical learning error, and we are able to maintain a good-quality reference value function with $O(SA)$ space complexity. [2]

Keywords: Model-free reinforcement learning; average-reward markov decision processes; reference-advantage decomposition

1. We use the notation $\widetilde{O}(\cdot)$ to suppress constant and logarithmic factors.
2. Extended abstract. Full version appears as [arXiv:2306.16394, v1]

Acknowledgements

We thank Kaiqing Zhang for insightful discussions. We also acknowledge the anonymous reviewers for their valuable feedback. Q. Xie is partially supported by NSF grant CNS-1955997.

References

Peter L Bartlett and Ambuj Tewari. Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42, 2009.

Jinghan Wang, Mengdi Wang, and Lin F Yang. Near sample-optimal reduction-based policy learning for average reward mdp. *arXiv preprint arXiv:2212.00603*, 2022.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33: 15198–15207, 2020.