Proceedings of Machine Learning Research vol 125:1–4, 2020

$\sqrt{n}$-Regret for Learning in Markov Decision Processes with Function Approximation and Low Bellman Rank

Kefan Dong* KEFANDONG@GMAIL.COM
Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China

Jian Peng JIANPENG@ILLINOIS.EDU
Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL 61820, USA

Yining Wang YNWANG.YINING@GMAIL.COM
Warrington College of Business, University of Florida, Gainesville, FL 32611, USA

Yuan Zhou YUANZ@ILLINOIS.EDU
Department of ISE, University of Illinois Urbana-Champaign, Urbana, IL 61820, USA

Abstract

In this paper, we consider the problem of *online* learning of Markov decision processes (MDPs) with very large state spaces. Under the assumptions of realizable function approximation and low Bellman ranks, we develop an online learning algorithm that learns the optimal value function while at the same time achieving very low cumulative *regret* during the learning process. Our learning algorithm, Adaptive Value-function Elimination (AVE), is inspired by the policy elimination algorithm proposed in Jiang et al. (2017), known as OLIVE. One of our key technical contributions in AVE is to formulate the elimination steps in OLIVE as *contextual bandit* problems. This technique enables us to apply the active elimination and expert weighting methods from Dudik et al. (2011), instead of the random action exploration scheme used in the original OLIVE algorithm, for more efficient exploration and better control of the regret incurred in each policy elimination step. To the best of our knowledge, this is the first $\sqrt{n}$-regret result for reinforcement learning in stochastic MDPs with general value function approximation.

1. Introduction

Consider a Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{X}, \mathcal{A}, H, p, r)$ with state space $\mathcal{X}$, action space $\mathcal{A}$, horizon H, transition probabilities $p : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$[1] and reward function $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. For notational simplicity, we assume that $\mathcal{X}$ can be partitioned into disjoint subsets as $\mathcal{X} = \mathcal{X}_1 \cup \cdots \mathcal{X}_H$, such that $\mathcal{X}_h \cap \mathcal{X}_{h'} = \emptyset$ if $h \neq h'$. A *policy* $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ is a function that maps a state $x \in \mathcal{X}$ to a distribution over actions $a \in \mathcal{A}$. The objective of policy learning is usually formulated as an optimization of finding π that achieves as large the *expected reward* as possible under $\mathcal{M}$, which is defined as

$$R(\pi) := \mathbb{E}\big[\sum_{h=1}^{H} r_h | r_h \sim r(x_h, a_h), a_h \sim \pi(x_h), x_h \sim p(x_{h-1}, a_{h-1})\big]. \tag{1}$$

The *optimal policy* π that maximizes Eq. (1) is denoted as π^*. Without further confusion, for *deterministic* policies (i.e., policies whose $\pi(\cdot)$ is a singleton for all states) we abuse the notation $\pi(x) \in \mathcal{A}$ for the action the policy takes at state x. We remark that the optimal policy π^* can always be made deterministic.

* Extended abstract. Full version appears as https://arxiv.org/abs/1909.02506. Author names are listed in alphabetical order. Correspondence to: yuanz@illinois.edu. Work done while KD was a visiting student at UIUC. KD and YZ were supported by a Ye Grant.

1. $\Delta(\mathcal{X})$ denotes all probability distributions over $\mathcal{X}$.

© 2020 K. Dong, J. Peng, Y. Wang & Y. Zhou.

In this paper, we consider the problem of learning near-optimal policy π with unknown p and r from two perspectives: the *sample complexity* perspective, which seeks for the smallest number of realized trajectories (possibly obtained using different exploration policies) in order to obtain a good policy with high probability, and the *online learning* perspective which characterizes how the exploration policies themselves evolve and improve over time.

1.1. Function approximation

When the state space $\mathcal{X}$ is finite with small cardinality $|\mathcal{X}|$, all states x can be enumerated in learning. This is known as the *tabular* MDP setting, which has been extensively studied (Jin et al., 2018; Zanette and Brunskill, 2019; Azar et al., 2017; Strehl et al., 2006; Azar et al., 2011; Even-Dar and Mansour, 2003; Sidford et al., 2018). In many real-world problems, however, $|\mathcal{X}|$ can be very large or even infinite. For example, in the Go game, the total number of states could be as large as 2×10^{170}, clearly infeasible for any approach that attempts to enumerate them.

It is clear that, in order to handle MDPs with very large state spaces, aggressive compression of the state space is required for practical purposes. In the literature, such compression is most naturally accomplished by the idea of *function approximation*, which considers a finite class[2] of functions $\mathcal{F} = \{f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}\}$ and restricts ourselves to policies $\Pi = \{\pi_f : f \in \mathcal{F}\}$ "induced" by certain function approximates, defined as

$$\pi_f(x) = \arg\max_{a \in \mathcal{A}} f(x, a). \tag{2}$$

In essence, the complexity of the function class $\mathcal{F}$ captures all inherent structures in the MDP $\mathcal{M}$ with a very large state space. In practice, the approximation function classes range from linear or low-degree polynomials in revenue management problems (Talluri and Van Ryzin, 1998; Adelman, 2007) to very complicated convolutional or recurrent neural networks for complex games (Mnih et al., 2015, 2016).

To ensure a considered function approximation is appropriate, we impose the following *realizability* assumption which guarantees the correspondence between the optimal policy π^* and a function $f^* \in \mathcal{F}$:

Assumption 1 (Realizability) *For the optimal policy π^*, there exists $f^* \in \mathcal{F}$ such that for all $h \in [H]$, $x_h \in \mathcal{X}_h$ and $a_h \in \mathcal{A}$,*

$$Q_h^{\pi^*}(x_h, a_h) := \mathbb{E}\big[\textstyle\sum_{h' \geq h} r_{h'} \big| r_{h'} = r(x_{h'}, a_{h'}), a_{h'} = \pi^*(x_{h'}), s_{h'} \sim p(x_{h'-1}, a_{h'-1})\big] = f^*(x_h, a_h).$$

We remark that Assumption 1 is a *monotonic* assumption, meaning that if it holds for function class $\mathcal{F}$ then it also holds for all $\mathcal{F}' \supseteq \mathcal{F}$. While such monotonicity property is desirable, allowing us to use slightly more than necessary function approximators, such property does *not* hold for many "completeness" type conditions in the literature, as we remark in more details in the full version of the paper.

1.2. From PAC-learning to online learning

Suppose the learning algorithm has access to n sequentially collected trajectories, and an *adaptive* policy $\pi^{(i)}$ can be used to generate the ith trajectory, which might depend on the algorithm's observations from the previous $(i-1)$ realized trajectories. Under the "Probably Approximately Correct (PAC)" framework, after observing data from n trajectories with n depending *polynomially* on the problem size, the algorithm is asked to output a policy $\widehat{\pi}$ which is near-optimal with high probability. The work in Jiang et al. (2017) provided the first PAC-learning result under the assumption of realizability and low bellman rank:

2. The requirement that $\mathcal{F}$ is finite could be removed, as shown in the full version of the paper.

Theorem 2 (Jiang et al. (2017)) *For a Markov Decision Process with bellman rank M, there exists an algorithm and a model-dependent constant $C_{\mathcal{M}}$ that is a polynomial of $H, |\mathcal{A}|, M$ and $\log |\mathcal{F}|$ such that, for any $\varepsilon \in (0, 1/2]$, with $n = \widetilde{O}(C_{\mathcal{M}}/\varepsilon^2)$ sample trajectories, the algorithm outputs a policy $\widehat{\pi}$ that satisfies $R(\widehat{\pi}) \geq R(\pi^*) - \varepsilon$ with probability at least 0.9.*

While PAC-learning results such as the one in Theorem 2 is very much desirable, the framework overlooks the aspect of exploration policy *improvement*, which expects the quality of the exploration policy to continuously improve as more data are collected. Such exploration policy improvement is important in applications where bad policies maybe lead to significant loss or even the cost of human lives, such as learning for self-driving cars. In these applications, an evaluation criterion of the "cumulative" gap of sub-optimality between the committed exploration policies and the optimal policy, known commonly as the cumulative *regret* in the online/bandit learning literature, is more suitable to measure the quality of policy improvement.

The following theorem is the main result we established in this paper:

Theorem 3 (Our results, informal) *For a Markov Decision Process with bellman rank M, there exists an algorithm and a model-dependent constant $C'_{\mathcal{M}}$ that is a polynomial of $H, |\mathcal{A}|, M$, $\log |\mathcal{F}|$ and $\log(1/\delta)$, such that, for sufficiently large n, the policies $\widehat{\pi}^{(1)}, \cdots, \widehat{\pi}^{(n)}$ the algorithm performs on the n trajectories satisfy with probability $(1-\delta)$ that*

$$\sum_{i=1}^{n} R(\pi^*) - R(\widehat{\pi}^{(i)}) = \widetilde{O}(C'_{\mathcal{M}} \times \sqrt{n}).$$

At a higher level, the result of Theorem 3 upper bounds the sub-optimalty gap of exploration policies $\{\widehat{\pi}^{(i)}\}$ for *every* trajectory $i = 1, 2, \cdots, n$ the algorithm obtains. Because the upper bound is on the order of $\widetilde{O}(\sqrt{n})$, the exploration policies have to constantly improve over themselves as otherwise a linear $O(n)$ regret will be incurred.

We make some additional remarks on Theorem 3, regarding its connection with the PAC-learning result in Theorem 2.

Remark 4 (online-to-batch conversion) *Because the expected reward function $R(\pi)$ is* linear *in policy π, by considering the "averaging policy" $\overline{\pi} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\pi}^{(i)}$ one has $R(\overline{\pi}) = \frac{1}{n}\sum_{i=1}^{n}[R(\pi^*) - R(\widehat{\pi}^{(i)})] \geq R(\pi^*) - \widetilde{O}(C'_{\mathcal{M}}/\sqrt{n})$ with high probability, matching the result in Theorem 2.*

Remark 5 (exploration and exploitation) *By running the PAC-learning algorithm implied by Theorem 2 on the first $n^{1/3}$ sample trajectories and then switching to the learnt policy π for the rest of the $n - n^{1/3}$ trajectories, one obtain a regret upper bound of $\widetilde{O}(\mathcal{C}_m \times n^{2/3})$, much worse than the $\widetilde{O}(\sqrt{n})$ upper bound in Theorem 3. The $\widetilde{O}(n^{2/3})$ regret bound cannot be improved by simply treating the PAC-learning algorithm as a black box.*

Acknowledgement

We thank Akshay Krishnamurthy and Zhizhou Ren for valuable discussions.

References

Daniel Adelman. Dynamic bid prices in revenue management. *Operations Research*, 55(4):647–661, 2007.

Mohammad Gheshlaghi Azar, Remi Munos, Mohammad Ghavamzadeh, and Hilbert Kappen. Speedy q-learning. In *Advances in neural information processing systems*, 2011.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 169–178, 2011.

Eyal Even-Dar and Yishay Mansour. Learning rates for q-learning. *Journal of Machine Learning Research*, 5(Dec):1–25, 2003.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1704–1713, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 4863–4873, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1928–1937, 2016.

Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787. Society for Industrial and Applied Mathematics, 2018.

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888. ACM, 2006.

Kalyan Talluri and Garrett Van Ryzin. An analysis of bid-price controls for network revenue management. *Management Science*, 44(11-part-1):1577–1593, 1998.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *arXiv preprint arXiv:1901.00210*, 2019.