Proceedings of Machine Learning Research vol 75:1–4, 2018

Open Problem: The Dependence of Sample Complexity Lower Bounds on Planning Horizon

Nan Jiang NANJIANG@ILLINOIS.EDU

Alekh Agarwal ALEKHA@MICROSOFT.COM

Microsoft Research, New York

Abstract

In reinforcement learning (RL), problems with long planning horizons are perceived as very challenging. The recent advances in PAC RL, however, show that the sample complexity of RL does not depend on planning horizon except at a superficial level. How can we explain such a difference? Noting that the technical assumptions in these upper bounds might have hidden away the challenges of long horizons, we ask the question: *can we prove a lower bound with a horizon dependence when such assumptions are removed?* We also provide a few observations on the desired characteristics of the lower bound construction.

Keywords: reinforcement learning, sample complexity, planning horizon

1. Introduction

Reinforcement learning (RL) is a machine learning paradigm for sequential decision making, where the agent takes actions to maximize cumulative rewards over multiple time steps (Sutton and Barto, 1998). Compared to its simpler form, known as contextual bandits, the full RL problem is more challenging due to the notion of *planning horizon*: error propagates over time steps, and distribution drifts as time elapses. In general, exploration and temporal credit assignment, which are core challenges of RL, both become more difficult when the planning horizon is longer. Many approaches, including hierarchical RL (Sutton et al., 1999) and reward shaping (Ng, 1999), are proposed to effectively shrink planning horizons using side information.

While such intuitions are agreed upon widely, recent advances in RL theory seem to tell a very different story. When the state and action spaces are finite and no function approximation is used ("tabular RL"), recent works have closed the gap between some lower and upper bounds (e.g., Dann and Brunskill, 2015; Azar et al., 2017). The achieved sample complexity / regret rates, however, have effectively *no dependence* on planning horizon, apart from the superficial dependence due to unnormalized total reward, non-stationary dynamics, counting steps instead of episodes, etc. In addition, constructions in matching lower bounds clearly exhibit bandit structure, reflecting no challenges from long planning horizons.

We believe explaining such a difference between theoretical results and practical intuitions is very important. As we will argue, while the matching upper and lower bounds have no horizon dependence, some technical assumptions in the upper bound might have hidden away the challenges of long horizons. Once they are removed, a horizon-dependent lower bound is no longer excluded. Therefore, the question we ask here is whether we can actually prove such a lower bound. Besides resolving the aforementioned difference, the lower bound will also serve as a foundation for theoretical analyses of e.g., hierarchical RL and reward shaping: We will only be able to theoretically show the benefit of reducing planning horizon if vanilla RL has a nontrivial dependence on it.

2. Definitions, Problems, and Related Results

For concreteness we adopt the PAC framework for episodic RL for most of this document. The question can also be meaningfully asked under regret minimization in episodic and average-reward settings, which we will briefly touch on in Section 2.4. We welcome resolutions of our problem in any reasonable formulations as long as they shed light on the conceptual question raised earlier.

Let $M = (\mathcal{S}, \mathcal{A}, P, R, H, \mu)$ be an episodic Markov Decision Process (MDP), where $\mathcal{S}$ is a finite state space of size S, $\mathcal{A}$ is a finite action space of size A, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function, $R : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ is the reward function, $H \in \mathbb{N}$ is the horizon (episode length), and $\mu \in \Delta(\mathcal{S})$ is the initial state distribution. An episode $s_1, a_1, r_1, s_2, \ldots, s_H, a_H, r_H$ is rolled out as $s_1 \sim \mu$, $r_1 \sim R(s_1, a_1), s_2 \sim P(s_1, a_1), \ldots, s_H \sim P(s_{H-1}, a_{H-1}), r_H \sim R(s_H, a_H)$, where $a_1, \ldots, a_H$ are chosen by the agent. Given any non-stationary policy $\pi : \mathcal{S} \times [H] \to \mathcal{A}$ which takes action a_h based on the current state $s_h \in \mathcal{S}$ and time step $h \in [H]$, we measure its performance by the expected total rewards $v^\pi := \mathbb{E}_{a_{1:H} \sim \pi}[\sum_{h=1}^H r_h]$. Let $v^\star := \max_{\pi : \mathcal{S} \times [H] \to \mathcal{A}} v^\pi$.

As a regularity assumption, let $r_h \geq 0$ and $\sum_{h=1}^H r_h \in [0, 1]$ hold almost surely. While this may seem stronger than the usual assumption in literature ($r_h \in [0, 1]$ and hence $\sum_{h=1}^H r_h \in [0, H]$), it is actually substantially *weaker*, as we will explain later. We say an algorithm PAC-learns an MDP if w.p. at least $1 - \delta$, the algorithm can identify a policy $\hat{\pi}$ such that $v^{\hat{\pi}} \geq v^\star - \epsilon$ after collecting $poly(S, A, H, 1/\epsilon, 1/\delta)$ episodes.[1] This polynomial is the *sample complexity* of the algorithm.

Problem Statement Can we prove a lower bound that depends polynomially in H? We conjecture that $\Omega(\frac{SAH}{\epsilon^2})$ is the right scaling, as opposed to the known $\Omega(\frac{SA}{\epsilon^2})$ lower bound. Showing an $\Omega(\frac{SAH^\alpha}{\epsilon^2})$ lower bound for any $\alpha \in (0, 1)$ would also be interesting and positive progress.

2.1. Issues with the known upper bounds

Dann and Brunskill (2015) proposed UCFH, which has a sample complexity of $\tilde{O}(\frac{S^2 A}{\epsilon^2} \log \frac{1}{\delta})$ with no dependence on H.[2,3] However, the following technical assumptions of the upper bound might have hidden the challenges of long horizons:

Reward Uniformity The regularity assumption of Dann and Brunskill (2015) is the standard $r_h \in [0, 1]$ (and hence $\sum_{h=1}^H r_h \in [0, H]$). To remove the dependence on H due to reward scaling, we should normalize their cumulative reward to $[0, 1]$ by dividing reward by H. Now compare their assumption (after normalization) to ours:

Standard assumption (e.g., Dann and Brunskill, 2015): $r_h \in [0, \frac{1}{H}]$, and hence $\sum_{h=1}^H r_h \in [0, 1]$.
Our assumption (e.g., Krishnamurthy et al., 2016): $r_h \geq 0$, and $\sum_{h=1}^H r_h \in [0, 1]$.

It is clear that our assumption is strictly weaker, despite that it might seem more restrictive at the first glance. (A key subtlety here is on the interpretation of ϵ: only after normalization does ϵ

[1] There is a slightly different notion of PAC for RL, which counts the number of episodes where the agent's policy is more than ϵ suboptimal (e.g., Dann and Brunskill, 2015). In fact, the lower bound given by Dann and Brunskill (2015) also applies to our PAC model, and in general we believe such a difference is minor to our discussion here.

[2] $\tilde{O}$ suppresses poly-logarithmic dependence on S, A, H, $1/\epsilon$, $\log \frac{1}{\delta}$. Such dependence is also ignored in our discussion (e.g., we want to show $\Omega(H^\alpha)$ dependence as opposed to $\Omega(\log H)$).

[3] The original bound has H^2 dependence; see the next paragraph for an explanation of the difference.

represent the *relative* suboptimality gap (Kakade, 2003, Chapter 2.2.3).) In fact, requiring $r_h \in [0, \frac{1}{H}]$ effectively imposes a uniformity requirement on rewards, and cannot model environments with sparse rewards—for which we believe long horizons are most challenging—in a tight manner.

Asymptotics The other assumption they have is $\epsilon \in [0, \frac{1}{H}]$ (after normalization). For some of our motivating scenarios, such an asymptotic situation is uninteresting: for example, the horizon of a control task can be, say, $H \sim 10^6$, when we control motors that respond in millisecond intervals ("flat RL"), but the horizon may reduce significantly if pre-defined macro actions are available ("hierarchical RL"). In this case, learning a policy 10^{-6} close to optimal is unnecessary, and to show the advantage of hierarchical RL we are interested in the regime of $\epsilon \gg 1/H$.

2.2. Existing lower bounds do not yield H dependence

The lower bound construction given by Dann and Brunskill (2015) is as follows: the agent chooses an action in the first step, transitions to either a good state or a bad state with action-dependent probabilities, and then loops in the good / bad state for the remaining time steps receiving either $+1$ or 0 reward per time step. Once we normalize total reward, the construction is exactly a multi-armed bandit with Bernoulli distributed rewards, which obviously will not yield any H dependence.

Another type of lower bound constructions in literature utilize lazy Markov chains (Jaksch et al., 2010): typically there are a good state and a bad state, and under all actions the agent will stay in its current state and only transition to the other state with small probabilities (see also Osband and Van Roy, 2016). Lattimore and Hutter (2012) considered batch-mode RL where the agent gets data from all state-action pairs in a synchronous manner. They used lazy-chain style constructions to show that the number of samples needed is $\Omega(H)$.[4] In the episodic setting, however, such a dependence is cancelled by the episode length. One fundamental reason that such a construction does not lead to $\Omega(H^\alpha)$ lower bound is the following: The small probabilities of switching states are set as $O(\frac{1}{H})$; As H increases, the MDP simply becomes lazier, and can be emulated by sampling episodes from an MDP with smaller H and adding uninformative "elapsing" time steps.

When $\mathcal{S}$ is very large or even continuous, there exists lower bounds that are exponential in H even with compact function approximation (Krishnamurthy et al., 2016). However, such results assume unbounded states so the exponential can be explained away by S.

2.3. Beyond existing constructions

We make a few more observations on the necessary characteristics of the desired lower bound construction. As a first step, to show an indisputable dependence, we would like to construct a family of MDPs where $S, A, 1/\epsilon \ll H$ and show that the sample complexity is $\Omega(H^\alpha)$ for $\alpha > 0$.

Our first observation is that if S and A are constant and optimal policies are stationary, then there exists an $O(\frac{1}{\epsilon^2})$ upper bound: Since there are $O(1)$ stationary policies, we can simply evaluate them one-by-one via Monte-Carlo and choose the best. This should not be surprising: It is known that policy search ignores the temporal structure and hence does not incur as much dependence on horizon as dynamic programming, and policy search is effective when the policy space is simple. Unfortunately, all the known constructions (see Section 2.2) yield stationary optimal policies.

To break this argument, one could try to make the optimal policies non-stationary. In particular, if an optimal policy switches from one stationary policy to another after $h_0 \in [H]$ steps, where h_0

[4] Their result is in the discounted setting with discount factor γ, and $\frac{1}{1-\gamma}$ corresponds to H in the episodic setting.

varies in the family of MDPs of interest, the upper bound argument will be broken as there are HA^S "1-switch" non-stationary policies. In fact, we find that such a situation can be created by adding a small amount of instantaneous reward to the lazy-chain style constructions. The difficulty is that the algorithm may not need to know the switching timing precisely. In the cases we have inspected, the algorithm can basically discretize $[H]$ into intervals of length $O(\epsilon H)$ and guarantee that one of those $O(1/\epsilon)$ switching timings is ϵ-optimal. Thus, the construction is still subject to a variant of the Monte-Carlo upper bound argument.

2.4. Alternative formulations

The results for regret minimization are in similar situations. The lower and upper bounds for the episodic setting are closed under reward uniformity and asymptotic assumptions (Azar et al., 2017). In the average-reward case, the notion of horizon is replaced by the MDP's diameter, D; here the lower and the upper bounds still have a gap of $\sqrt{D}$ (Jaksch et al., 2010; Agrawal and Jia, 2017).

We also welcome resolution of our problem in more realistic and challenging settings beyond tabular RL, such as rich observations and function approximation (Krishnamurthy et al., 2016; Jiang et al., 2017). While a richer setting enables more powerful lower bounds, existing work have not leveraged the power yet (e.g., Jiang et al., 2017, Thm 6 still uses a tabular construction).

Acknowledgements We thank Christoph Dann and John Langford for insightful discussions.

References

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, 2017.

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

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, 2015.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, 2017.

Sham Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University College London, 2003.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, 2016.

Tor Lattimore and Marcus Hutter. PAC bounds for discounted MDPs. In *Algorithmic Learning Theory (ALT)*, 2012.

Andrew Ng. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning*, 1999.

Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.