ECE 239AS Milestone Report: Is Q-Learning Provably Efficient?

Kushagra Rastogi
UID: 304640248
University of California, Los Angeles
Los Angeles, CA 90095
krastogi@g.ucla.edu

Jonathan Lee
UID: 104840173
University of California, Los Angeles
Los Angeles, CA 90095
jlee916@g.ucla.edu

Aditya Joglekar
UID: 405222904
University of California, Los Angeles
Los Angeles, CA 90095
adivj123@gmail.com

Fabrice Harel-Canada
UID: 705221880
University of California, Los Angeles
Los Angeles, CA 90095
fabricehc@cs.ucla.edu

Abstract

We propose to analyze the theoretical results presented within the paper *Is Q-Learning Provably Efficient?* by Jin *et al.* [1]. This analysis shall include a survey of related research to contextualize the need for strengthening the theoretical guarantees related to perhaps the most important threads of model-free reinforcement learning. We also hope to expound upon the reasoning used in the proofs to highlight the critical leading steps to the main results: Q-learning with UCB exploration achieves a sample efficiency that matches the optimal regret that can be achieved by any model-based approach.

1 Introduction

State-of-the-art reinforcement learning has been dominated by model-free algorithms (like Q-learning) because they are online, more expressive and need less space. However, empirical work has shown that model-free algorithms have a higher sample complexity [2, 3], meaning that they require many more samples in order to perform well on a given task. Can we make model-free algorithms sample-efficient? This is one of the most fundamental questions in the reinforcement learning community that has yet to be answered definitely. As seen in the setting of multi-armed bandits, good sample efficiency is the result of aptly managing the exploration-exploitation tradeoff. In our project, we aim to prove that Q-learning with Upper Confidence Bound (UCB) exploration, in an episodic MDP setting and without access to a "simulator", matches the information-theoretic regret optimum, up to a single $\sqrt{H}$ where H is the number of steps per episode. To do this, we will leverage our current understanding of Q-learning, examine literature on *delayed Q-learning* [4], and investigate ways to combine model-free algorithms with model-based approaches.

2 Preliminary Results

2.1 Related Work

Regarding the related work, we've collected and begun summarizing / synthesizing all of the papers referenced in the original work, as well as incorporating other published works we've found independently. We'll omit the subsections related to sample complexity and efficiency, saving it for the final report so that we can fit an abridged version of the proof evaluation progress in the next subsection.

34rd Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

2.1.1 Model-free vs. Model-based RL

- Model-free (MF) Trial-and-Error Learners
  - Pros
    - Computationally less complex than MB methods, requiring no model of the environment to be effective (which can be a bottleneck for MB methods) [5].
  - Cons
    - Requires (repeated) "personal experience" with many state-action pairs in order to train, makes exploration more costly [5].
    - Tend to be less sample efficient [5, 2, 6].
- Model-based (MB) Planners
  - Pros
    - Tend to be more sample efficient [2, 6].
    - More efficient handling of changing goals because it does not need "personal experience" with every state-action pair [6, 5].
  - Cons
    - Suffer from model bias, i.e., they inherently assume that the learned dynamics model sufficiently accurately resembles the real environment [2, 7, 8, 6].
    - Computationally more complex than MF methods - can be difficult to learn a good model of state transitions and rewards [5].

2.2 Proofs

We present the main theorem and supporting lemmas that will help us prove that Q-learning is efficient. In our paper, we only prove the efficiency of Q-learning with UCB-Hoeffding bonus.

Algorithm 1 Q-learning with UCB-Hoeffding

Initialize $Q_h(x,a) \leftarrow H$, $N_h(x,a) \leftarrow 0 \; \forall (x,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$
for episode $k = 1$ to K **do**
  get x_1
  for step $h = 1$ to H **do**
    $a_h \leftarrow argmax_{a'} Q_h(x_h, a')$
    $t = N_h(x,a) \leftarrow N_h(x,a) + 1$
    $b_t \leftarrow c\sqrt{H^3\iota/t}$ where $c > 0$ is a constant and $\iota = \log(SAT/p)$
    $Q_h(x_h,a_h) \leftarrow (1-\alpha_t)Q_h(x_h,a_h) + \alpha_t[r_h(x_h,a_h) + V_{h+1}(x_{h+1}) + b_t]$
    $V_h(x_h) \leftarrow \min(H, \max_{a' \in \mathcal{A}} Q_h(x_h,a'))$
  end for
end for

Theorem 1 (Hoeffding). *If $b_t = c\sqrt{H^3\iota/t}$, then with probability $1-p$, $\forall p \in (0,1)$, the total regret of Algorithm 1 is at most $O(\sqrt{H^4SAT\iota})$ where $c > 0$ is a constant and $\iota = \log(SAT/p)$.*

Lemma 4.1. *The learning rate is $\alpha_t = \frac{H+1}{H+t}$ and some related quantities are $\alpha_t^0 = \prod_{j=1}^t (1-\alpha_j)$ and $\alpha_t^i = \alpha_i \prod_{j=i+1}^t (1-\alpha_j)$. Properties of α_t^i:*

(a) *For every $t \geq 1$, $\frac{1}{\sqrt{t}} \leq \sum_{i=1}^t \frac{\alpha_t^i}{\sqrt{i}} \leq \frac{2}{\sqrt{t}}$.*

(b) *For every $t \geq 1$, $\max_{i \in [t]} \alpha_t^i \leq \frac{2H}{t}$ and $\sum_{i=1}^t (\alpha_t^i)^2 \leq \frac{2H}{t}$.*

(c) *For every $i \geq 1$, $\sum_{t=i}^\infty \alpha_t^i = 1 + \frac{1}{H}$.*

Proof of Lemma 4.1. Our choice of the learning rate is crucial for Q-learning to be efficient. Property (c) is particularly important to bound the regret by a constant factor of $(1+\frac{1}{H})^H$. The proofs of the properties are done using induction and simple manipulations of products and sums.

Lemma 4.2. *For any $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and episode $k \in [K]$ let $t = N_h^k(x, a)$ and suppose (x, a) was previously taken at step h of episodes $k_1, k_2, ..., k_t < k$. Then:*

$$(Q_h^k - Q_h^*)(x, a) = \alpha_t^0(H - Q_h^*(x, a)) + \sum_{i=1}^{t} \alpha_t^i[(V_{h+1}^{k_i} - V_{h+1}^*)(x_{h+1}^{k_i}) + [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h)V_{h+1}^*](x, a) + b_i]$$

Proof of Lemma 4.2. The proof follows from manipulating the Bellman optimality equation $Q_h^*(x, a) = (r_h + \mathbb{P}_h V_{h+1}^*)(x, a)$ using the facts $\sum_{i=0}^{t} \alpha_t^i = 1$ and $[\hat{\mathbb{P}}_h^{k_i} V_{h+1}](x, a) = V_{h+1}(x_{h+1}^{k_i})$.

Lemma 4.3. *There exists an absolute constant $c > 0$ such that, for any $p \in (0, 1)$, letting $b_t = c\sqrt{H^3\iota/t}$, we have $\beta_t = 2\sum_{i=1}^{t} \alpha_t^i b_i \leq 4c\sqrt{H^3\iota/t}$ and, with probability at least $1 - p$, the following holds simultaneously $\forall (x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$:*

$$0 \leq (Q_h^k - Q_h^*)(x, a) \leq \alpha_t^0 H + \sum_{i=1}^{t} \alpha_t^i (V_{h+1}^{k_i} - V_{h+1}^*)(x_{h+1}^{k_i}) + \beta_t$$

Proof of Lemma 4.3. This lemma, which serves as the crux for the proof of Theorem 1, utilizes several concepts within measurement and probability theory, namely, σ-fields, filtration, Martingale difference sequences, and the Azuma-Hoeffding inequality. The main steps in the proof involve bringing the upper bound on the state-action value functions $Q_h^*(x, a)$ to an optimistic estimation $Q_h^k(x, a)$, particularly through the empirical difference $[(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h)V_{h+1}^*](x, a)$. The former half of the proof focuses on showing that $|\sum_{i=1}^{t} \alpha_t^i \cdot [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h)V_{h+1}^*](x, a)| \leq c\sqrt{\frac{H^3\iota}{t}}$, while the latter half takes the upper limit and applies properties from Lemma 4.1 to put everything together when bounding $Q_h^*(x, a)$ defined in Lemma 4.2.

Proof of Theorem 1. Having laid the foundations of the proof above- it's time to tie things up nicely. The big picture idea is to think of the 'episodic MDP with H steps per episode' as a 'contextual bandit of H layers'. The key challenge then is to control the way error and confidence propagate through the different "layers" in an online fashion, meaning that the exploration bonus and learning rate make the regret as sharp as possible. Lemma 4.3 and the Bellman equation are crucial for the derivation. Also, using smarter schemes like UCB exploration and a modified *weighted learning rate* ($\alpha_t = \frac{H+1}{H+t}$) are necessary conditions to achieve the competitive sample complexity.

The crux of the proof is relating the upper bound of the regret and the sample complexity linked by a recursive formula. The recursive inequality formula for the regret follows from the invocation of Lemma 4.3 and the Bellman equation. Their combination basically links the Q-value term from the Bellman equation to the sample complexity equations and the probability thresholds from Lemma 4.3. Denoting $(V_h^k - V_h^\pi)(x_h^k)$ as δ_h^k we get the following upper-bound for the total regret

$$Regret(K) = \sum_{k=1}^{K} (V_1^* - V_1^\pi)(x_1^k) <= \sum_{k=1}^{K} \delta_h^k$$

We now just need a recursive relation between the upper bounds at each step that is, $\sum_{k=1}^{K} \delta_h^k$ and $\sum_{k=1}^{K} \delta_h^k + 1$. This follows from manipulating the above equation and invoking Lemma 4.3 which will relate the recursive difference of the upper bounds for the regret with the probabilistic and complexity terms finally resulting in

$$\sum_{k=1}^{K} \delta_1^k \leq O(H^2SA + \sqrt{H^4SAT\iota})$$

which is the efficient sample complexity under the probabilistic constraints defined in Lemma 4.3 hence completing the proof. The details of relating the upper bound difference with the complexity terms will be explored in the main paper. It is just a short step of repeatedly invoking the Lemmas on the recursive equations defined above hence completing the proof.

References

[1] Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is q-learning provably efficient? In *NeurIPS*, 2018.

[2] Marc Deisenroth and Carl Rasmussen. Pilco: A model-based and data-efficient approach to policy search., 01 2011.

[3] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization, 2015.

[4] 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*, ICML '06, page 881–888, New York, NY, USA, 2006. Association for Computing Machinery.

[5] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[6] Christopher G. Atkeson and Juan Carlos Santamaria. A comparison of direct and model-based reinforcement learning. In *IN INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION*, pages 3557–3564. IEEE Press, 1997.

[7] Jeff G. Schneider. Exploiting model uncertainty estimates for safe dynamic control learning. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, NIPS'96, page 1047–1053, Cambridge, MA, USA, 1996. MIT Press.

[8] Stefan Schaal. Learning from demonstration. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, NIPS'96, page 1040–1046, Cambridge, MA, USA, 1996. MIT Press.