ECE 239AS Project Proposal: 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 Motivation and Background

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 Course Relevance

Analyzing this paper [1] builds upon our knowledge of Q-learning in model-free environments and provides additional exposure to the concept of Upper Confidence Bounds (UCB), which it relies upon.

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

3 Planned Theoretical Analysis

We explore in our own words wherever possible- the main results, theorems, and lemmas, respectively ensuring we elaborate on underlying assumptions and fill the gaps left by the authors for the readers to figure out. The main theoretical contribution of the paper is to arrive at competitive sample complexities for Q-learning for the episodic MDP setting with the incorporation of upper confidence bounds (UCB). In contrast to the commonly used Q-learning approach with ε-greedy, which has proven to be sample inefficient for learning episodes (i.e. $\Omega(\min\{T, A^{H/2}\})$ in regret with no optimism), we dive deeper into other exploration/exploitation algorithms, namely, the UCB-Hoeffding and UCB-Bernstein algorithms. The intuition being that better strategies for balancing exploration and exploitation could make the learning more effective and thereby improve the sample efficiency.

We will take a look at the results and proofs used for Q-learning with UCB-Hoeffding- with the more challenging UCB-Berstein analysis envisaged as a stretch goal. The goal is to arrive at a sharp bound for sample complexity for the regret — which the authors hope is better than model-free Q-learning with ε-greedy and comparable to the efficiency of model-based approaches. We'll employ a step-by-step analysis using simpler lemmas to prove key ideas and finally show that the regret sample complexity is of the form $\mathcal{O}(\sqrt{H^4SAT})$. Proving Theorem 2 (Hoeffding) is thus the meat of the paper. It states the existence of a strictly positive constant c such that for any probability p, if an upper confidence bonus b^t of $c\sqrt{H^3}$ is chosen, then the total regret of the UCB-Hoeffding method is at most $\mathcal{O}(\sqrt{H^4SAT\iota})$ with a probability of $1-p$, where $\iota := \log(SAT/p)$. In comparison to model-based approaches, the Hoeffding theorem shows similar efficiency in terms of dependency on the number of states S, actions A, and discrete steps T, with only a slightly higher dependency on the number of steps per episode H. However, the primary advantage of UCB-Hoeffding over model-based systems comes from the fact that it learns through current observation (online) as opposed to learning over groups of patterns (batch) and without storing any representation of the model. Consequently, it serves better than model-based approaches in time and space complexities. If possible we'll look into UCB-Berstein as it further improves the regret complexity by a factor of $\sqrt{H}$ at the cost of a complicated exploration design with a significantly harder analysis.

4 Measure of Success

Baseline: We plan to summarize the related work and detail all the main theorems / lemmas by the end of the quarter. This includes:

- Discusses pros and cons of model-free vs. model-based [2, 5]
- Discussing empirical evidence that model-free algorithms generally require higher sample complexity [2, 3].
- Discuss attempts to improve sample efficiency by combining key elements of model-free and model-based approaches [6, 7].
- Discuss *model-based* approaches that have achieved asymptotically optimal sample efficiency [8, 9, 10, 11, 12]
- Discuss *model-free* theoretical results for episodic MDP [4], which is the most related work to the present paper [1].
- Rigorously explicate the sample complexity analysis described in the paper.

In order to accomplish this, we will need to familiarize ourselves with a swath of related research on the methods used to characterize the theoretical guarantees of both model-based and model-free RL algorithms. We will also likely need to investigate unwritten assumptions relied upon within the lemmas and their application to the main theorem proofs.

Stretch Goal: Apply the theoretical approach used in this paper to analyze the pairing of Q-learning with another kind of exploration strategy, such as optimistic initial values. NOTE: Q-Learning with ε-greedy exploration was already demonstrated in Appendix A of [1]. Alternatively, we could also attempt to arrive at the same conclusion via the application of other mathematical inequalities [13] as a means of verifying the correctness of the original proofs in [1].

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] Anusha Nagabandi, Gregory Kahn, Ronald Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning, 05 2018.

[7] Vitchyr Pong, Shixiang Gu, Murtaza Dalal, and Sergey Levine. Temporal difference models: Model-free deep rl for model-based control. *ArXiv*, abs/1802.09081, 2018.

[8] Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *ArXiv*, abs/1705.07041, 2017.

[9] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *ICML*, 2017.

[10] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, August 2010.

[11] Sham Kakade, Mengdi Wang, and Lin Yang. Variance reduction methods for sublinear reinforcement learning, 02 2018.

[12] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions, 2016.

[13] https://en.wikipedia.org/wiki/List_of_inequalities. List of inequalities — wikipedia, the free encyclopedia, 2020.