Generalized Optimistic Q-Learning with Provable Efficiency*

Grigory Neustroev[1][0000−0002−7706−7778] and
Mathijs M. de Weerdt[1][0000−0002−0470−6241]

Delft University of Technology, EEMCS, Algorithmics,
P.O. Box 5031, 2600 GA, Delft, the Netherlands
{g.neustroev, m.m.deweerdt}@tudelft.nl

This is an abstract of a paper published at the 19th International Conference on Autonomous Agents and Multiagent Systems, Auckland, New Zealand [2].

1 Introduction

When a learning algorithm requires as few data samples as possible, it is called sample efficient. Recently, Jin *et al.* introduced the first provably efficient model-free reinforcement learning (RL) algorithm [1]. Later, a few other sample-efficient model-free algorithms were developed [4, 3]. The key factor that allows these algorithms to achieve sample efficiency is their use of the principle of *optimism in the face of uncertainty*.

The paper studies the effect of optimism on sample efficiency of RL. It presents a generalized theory on optimistic model-free RL, unifying the existing algorithms. Using this theory, we establish sample efficiency of optimistic Q-learning by showing that its regret grows sub-linearly with respect to the number of samples. Moreover, we show that the regret of optimistic Q-learning can be explained by three distinct factors.

2 Generalized Optimistic Q-Learning

In learning, optimism is used in two ways: optimistic initialization and optimistic exploration. We look at the existing optimistic model-free RL methods [1, 4, 3] to see how they incorporate these aspects of optimism.

In initialization, large values are assigned to all state-action combinations. This guarantees that actions never chosen before seem especially lucrative. When such initialization is not possible (e.g., in deep RL), the Q-values of unvisited states are augmented with a bonus term that we call a *bonus for optimism*.

Optimistic exploration is done by using upper confidence bounds (UCBs) on state-action values. In each interaction, the action with the highest UCB is chosen. This happens either if the true optimal value is high, or if there is not enough confidence in it yet. In the former case, the agent essentially performs

* This research received funding from the Netherlands Organization for Scientific Research (NWO).

exploitation, as the chosen action is the best one. The latter case represents exploration, because an action with high uncertainty in its outcome is chosen. Thus, UCBs help to automatically balance exploration and exploitation. To maintain the UCBs, a *confidence bonus* is added to the Q-values during learning.

We incorporate these bonuses in an algorithm that we call *generalized optimistic Q-learning* and perform a theoretical analysis of its sample efficiency. Unlike previous results, our analysis does not rely on the particular form of the bonuses to determine whether the resulting algorithm is sample efficient or not.

The general form we use allows us to show that the total regret R of optimistic Q-learning is asymptotically bounded by the sum of three different terms:

$$R = \mathrm{O}\big(\mu(X + B + E)\big). \tag{1}$$

The state-action space size X represents the effect of the *optimistic initialization*, as the number of initial values is equal to X. The bonus effect B depends on the bonuses for optimism and for confidence. The last term E represents the required number of interactions with the environment to ensure that all of the possible outcomes are experienced with high probability. The magnitude μ depends on the reward range and the discounting factor and represents the scale of Q-value.

The formal proof of this regret bound relies on some mild necessary conditions. They can be found in the paper along with the formal definitions of the terms μ, X, B, and E and the proof itself. The paper also gives an example of a new algorithm designed within the generalized optimistic Q-learning framework. This algorithm, called UCB-H$^+$, is similar to UCB-H [1], but uses a different learning rate. Using the theoretical framework of the paper, we prove that it is sample-efficient. Then we evaluate UCB-H$^+$ in two experiments, which demonstrate a regret reduction of 13% and 43% compared to UCB-H.

3 Conclusions

Generalized optimistic Q-learning incorporates existing optimistic model-free reinforcement learning, and our proof does not rely on a particular form of learning rate or bonuses, allowing transfer of these results to new algorithms.

References

1. Jin, C., Allen-Zhu, Z., Bubeck, S., Jordan, M.I.: Is Q-learning provably efficient? In: Advances in Neural Information Processing Systems 31, pp. 4863–4873. Curran Associates, Inc. (2018)
2. Neustroev, G., de Weerdt, M.M.: Generalized optimistic Q-learning with provable efficiency. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems. pp. 913–921 (2020)
3. Rashid, T., Peng, B., Boehmer, W., Whiteson, S.: Optimistic exploration even with a pessimistic initialisation. In: International Conference on Learning Representations. Addis Ababa, Ethiopia (Apr 2020)
4. Wang, Y., Dong, K., Chen, X., Wang, L.: Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. In: International Conference on Learning Representations. Addis Ababa, Ethiopia (Apr 2020)