Double Q-Learning

Hao-Chen Lin
Department of Computer Science
National Yang Ming Chiao Tung University
f964966828.cs08@nycu.edu.tw

1 Introduction

A well known reinforcement learning value-based algorithm Q-learning always suffer from overestimations, and these overestimations result from the single estimator used in Q-learning, to solve this problem this paper use two estimators in Q-learning, and construct a new off-policy value based reinforcement learning algorithm called Double Q-learning.

In this paper, it's first show why single estimator might lead to overestimations and using similar concept to derive formula in two estimators, then show that two estimators method might meet underestimations problem instead.

Then using the idea of two estimator, it derived the pseudocode of Double Q-learning and show the convergence of Double Q-learning using concept of stochastic approximation and γ contraction operator. In the end, it shows some experiment results in testing environments.

2 Problem Formulation

When prove the convergence of Double Q-learning, there's two Q-value function notation as Q^A and Q^B, we want to show both two function will converge to the optimal value function Q^*, here we use the lemma of convergence in stochastic process, then give some constraint so that our notation will satisfy the limit of convergence of stochastic process, and here is constraints given in theorem.

- The MDP is finite, i.e. $|S \times A| < \infty$
- The discount factor $\gamma \in$ [0,1)
- The Q values are stored in a lookup table
- Both Q^A and Q^B receive an infinite number of updates
- The update rate $\alpha_t(s,a) \in$ [0,1), $\sum_t \alpha_t(s,a) = \infty$, $\sum_t (\alpha_t(s,a))^2 < \infty$

3 Theoretical Analysis

In this paper, they implement Double-Q learning algorithm in two testing environments called Roulette and Grid World, and in both environments does Double Q-learning have the better performance, avoid overestimation problem and need fewer epoch than Q-learning to get good performance.

Figure 1 shows state action values of Q-learning and Double Q-learning on the specific 'walk-away' action which worth \$0, we can see that Double Q-learning can always predict the true value while Q-learning always overestimate the value.

Figure 2 shows result in the grid world, first row is average reward per time step and second row is the maximal action value in the starting state, we can see Double Q-learning can get higher average reward in the same amount of time step, and as the same discover as in Roulette, Double Q-learning can get the true value while Q-learning always overestimate the value.

-

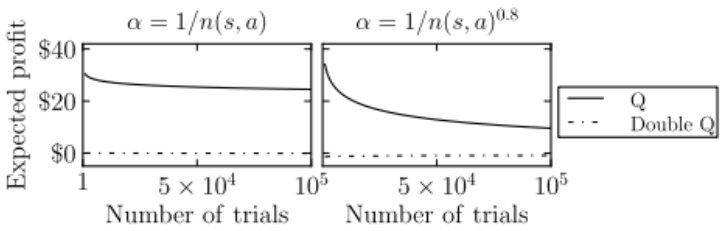

Figure 1: Results in the roulette for Q-learning and Double Q-learning.

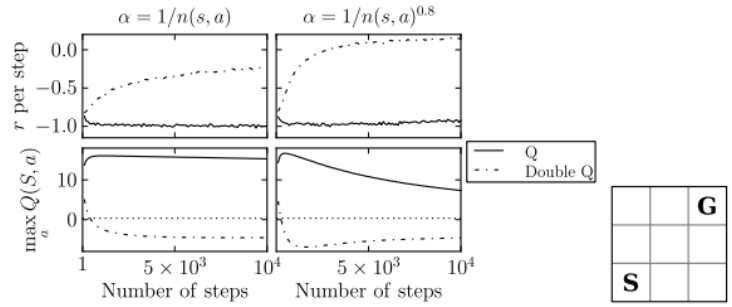

Figure 2: Results in the grid world for Q-learning and Double Q-learning.

4 Conclusion

This is the first off-policy value based reinforcement learning algorithm that does not have a positive bias in estimating the action value in stochastic environment at that time. It provide a more robust and efficient algorithm than Q-learning.

There still have some limitation according to this paper, as problem formulation above, the proof of convergence of Double Q-learning is under the constraint that MDP should be finite, it means if our state and action be continuous there might not guarantee convergence. Second point, even though Double Q-learning will not suffer from overestimation, it might sometimes meet underestimation instead, which might be a potential problem while training.

References

Hado van Hasselt, Double Q-learning, NIPS 2010