A Note on Double Q-Learning

Pei-Chiun Peng
Department of Electronic Engineering and Computer Science
National Yang Ming Chiao Tung University
lollypeng100.ee07@nycu.edu.tw

1 Introduction

The main research challenge tackled by Double Q-learning is the Q-learning overestimation problem. The overestimation problem is that Q-learning uses the maximum action value as an approximation for the maximum expected action value. However, by consuming acceptable extra memory and computation resources, Double estimator to Q-learning shows a new off-policy TD control reinforcement learning algorithm which gets rid of the overestimation problem.

2 Problem Formulation

Technical assumptions for convergence of Double Q-learning

In a given ergodic MDP, bothQ^A, Q^B as updated by Double Q-learning as described in Algorithm 1 will converge to the optimal value function Q^* as given in the bellman optimality equation. with probability one if an infinite number of experiences in the form of rewards and state transitions for each sate action pair are given by a proper learning policy.
The assumptions are: 1) the MDP is finite, i.e. $|SxA| < \infty$. 2) $\gamma \in [0, 1)$. 3) The Q values are stored in lookup table. 4) Both $Q^A and Q^B$ receive an infinite number of updates. 5) $\alpha_t(s, a) = \infty, \Sigma_t \alpha_t(s, a) < \infty$ w.p.1, and $\forall (s, a) \neq (s_t, a_t) : \alpha(s, a) = 0.6$. 6) $\forall s, a, a^{'} : Var\{R^{s}_{sa}\} < \infty$

3 Theoretical Analysis

Lemma 1.
Let $X = \{X_1, ..., X_M\}$ be a set of random variables and let $\mu^A = \{\mu^A_1, ..., \mu^A_M\}$ and $\mu^B = \{\mu^B_1, ..., \mu^B_M\}$ be two sets of unbiased estimators such that $E\{\mu^A_i\} = E\{\mu^B_i\} = E\{X_i\}$, for all i. Let $M \equiv \{j | E\{X_j\} = max_i E\{X_i\}\}$ be a set of elements that maximize the expected values. Let a^* be an element that maximized $\mu^A : \mu^A_{a^*} = max_i \mu^A_i$.
Then $E\{\mu^B_{a^*}\} = E\{X_{a^*}\} \leq max_i E\{X_i\}$. Furthermore, the inequality is strict if and only if $P(a^* \notin M) > 0$.
Double Q-learning

4 Conclusion

This paper presented Double Q-learning that uses a double estimator approach to determine the value of the next state. Double Q-learning sometimes underestimates the action values, but does not have overestimate bias like Q-learning. Whether it is possible to construct an unbiased off-policy reinforcement-learning algorithm is still a question. Maybe this can be solved by deducting the overestimation size from the estimate. However, the size of overestimation depends on the number of actions and unknown distributions of the reward, which still becomes a complicated extension.

Delayed Double Q-learning [2] and other extensions of Q-learning will become noticeable with Double Q-learning applied.

References

[1] Hado van Hasselt et al., Double Q Learning, NIPS 2010

[2] A. L. Strehl, L. Li, E. Wiewiora, J. Langford, and M. L. Littman. PAC model-free reinforce- ment learning. In Proceedings of the 23rd international conference on Machine learning, pages 881–888. ACM, 2006.