N-fold Q-learning

James Lue
Department of Computer Science
National Yang Ming Chiao Tung University
james2323123.cs10@nycu.edu.tw

1 Introduction

In this paper(Hasselt [2010]), the authors designed a new approximator for $\max_i E[X_i]$ by using two separated estimator and a new RL algorithm by applying the new approximator on Q-learning. In the original Q-learning, the overestimation bias of the approximator lead to the bad performance of the Q-learning. In this paper, the authors proved that the double estimator they proposed does not have overestimation bias. Instead, the double estimator may have underestimation bias.

The approximator used in Q-learning for a set random variables $X = \{X_1, \ldots, X_M\}$ is:

$$\max_i E[X_i] = \max_i E[\mu_i] \approx \max_i \mu_i(S) \tag{1}$$

Where $S = \bigcup_{i=1}^{M} S_i$ is the set of samples, S_i is the set of samples of X_i, and $\mu_i(S) \stackrel{\text{def}}{=} \frac{1}{|S_i|} \sum_{s \in S_i} s$ is the estimator of X_i

However, (1) is actually an unbiased estimate for $E[\max_j \mu_j]$:

$$E[\max_j \mu_j] = \int_{-\infty}^{\infty} x \frac{d}{dx} F^{\mu}_{max}(x) dx = \int_{-\infty}^{\infty} x \frac{d}{dx} \prod_{i=1}^{M} F^{\mu}_i(x) dx = \sum_{j=1}^{M} \int_{-\infty}^{\infty} x f^{\mu}_j(x) \prod_{i \neq j}^{M} F^{\mu}_i(x) dx \tag{2}$$

Where $F^{\mu}_{max}(x) \stackrel{\text{def}}{=} P(\max_i \mu_i \leq x) = \prod_{i=1}^{M} P(\mu_i \leq x) = \prod_{i=1}^{M} F^{\mu}_i(x)$, F^{μ}_i is the CDF of μ_i and f^{μ}_j is the PDF of μ_i

Since $\max_j \mu_j$ is a convex function of $\mu_1, \ldots, \mu_M$, by Jensen's Inequality, $E[\max_j \mu_j] \geq \max_i E[X_i]$, and thus (1) has a overestimation bias.

The double estimator proposed in this paper uses two independent sets of estimators μ^A and μ^B(i.e. the samples used S^A and S^B has no intersection). The approximator uses one of the set of estimators to find which random variable have the maximum expected value, then uses the other one to find the actual value. That is:

$$a^* = \arg\max_i \mu^A_i(S) \tag{3}$$

$$\max_i E[X_i] = \max_i E[\mu^B_i] \approx \mu^B_{a^*}(S) \tag{4}$$

Where $\mu^A_i(S) = \frac{1}{|S^A_i|} \sum_{s \in S^A_i} s$ and S^A_i is the set of samples of X_i in S^A

The expected value of the double estimator can be given by:

$$\sum_{j=1}^{m} P(j = a^*) E[\mu^B_j] = \sum_{j=1}^{M} E[\mu^B_j] \int_{-\infty}^{\infty} f^A_j(x) \prod_{i \neq j}^{M} F^A_i(x) dx \tag{5}$$

Where F^A_i and f^A_i are CDF and PDF of μ^A_i Which instead has underestimation bias since $P(j = a^*)$ sums up to 1 and (5) is therefore a weighted average of unbiased expected values. This can be shown by Lemma 1 in the paper:

$$
\begin{aligned}
E[\mu^B_{a^*}] &= P(a^* \in \mathcal{M})E[\mu^B_{a^*}|a^* \in \mathcal{M}] + P(a^* \notin \mathcal{M})E[\mu^B_{a^*}|a^* \notin \mathcal{M}] \\
&= P(a^* \in \mathcal{M})max_i E[X_i] + P(a^* \notin \mathcal{M})E[\mu^B_{a^*}|a^* \notin \mathcal{M}] \\
&\leq P(a^* \in \mathcal{M})max_i E[X_i] + P(a^* \notin \mathcal{M})max_i E[X_i] \\
&= max_i E[X_i]
\end{aligned} \tag{6}
$$

Where $\mathcal{M} \stackrel{\text{def}}{=} \{j|E[X_j] = \max_i E[X_i]\}$ is the set of elements that maximize the expected values. It can also be shown that when X_i's all have the same expected value, the approximator is unbiased since $P(a^* \in \mathcal{M}) = 1$

Using this approximator, the authors propsed the double Q-learning algorithm:

Algorithm 1 Double Q-learning

Initialize Q^A, Q^B, s
repeat
  Choose a, based on $Q^A(s, \cdot)$ and $Q^B(s, \cdot)$, observe r, s'
  Choose either UPDATE(A) or UPDATE(B) randomly
  if UPDATE(A) **then**
    Define $a^* = \arg\max_a Q^A(s', a)$
    $Q^A(s, a) \leftarrow Q^A(s, a) + \alpha(s, a)(r + \gamma Q^B(s', a^*) - Q^A(s, a))$
  else if UPDATE(B) **then**
    Define $b^* = \arg\max_a Q^B(s', a)$
    $Q^B(s, a) \leftarrow Q^B(s, a) + \alpha(s, a)(r + \gamma Q^A(s', a^*) - Q^B(s, a))$
  end if
  $s \leftarrow s'$
until end

After reading this paper, a very simple insight may come to one's mind: *If using two estimators is better than one, what about using three or more estimators?* In this project, I will:

- Propose a new estimator that uses multiple estimators to reduce the bias
- Use the proposed estimator to develop a new RL algorithm
- Discuss the possible caveats to this new estimator

2 Problem Formulation

In this project, we use $\mu^i = \{\mu^i_1, \ldots, \mu^i_M\}$ to denote the i-th set of estimators for random variables $X_1, \ldots, X_M$. First, let us consider the case where there is 3 sets of estimators($i \in [1, 3]$), in this case, we can construct a new approximator:

$$
a^* = \arg\max_i \mu^1_i(S) \tag{7}
$$

$$
\max_i E[X_i] \approx \max\{\mu^2_{a^*}(S), \mu^3_{a^*}(S)\} \tag{8}
$$

Note that there is another possible way to construct a new approximator using 3 sets of estimators:

$$
a^* = \arg\max_i \max \mu^1_i(S), \mu^2_i(S) \tag{9}
$$

$$
\max_i E[X_i] \approx \mu^3_{a^*}(S) \tag{10}
$$

But, we will focus on the first way in this project, where only one set of the estimator is used to find which random variable have the maximum expected value and the rest are used to find the actual value. Using N estimators, we can construct an approximator with the form:

$$a^* = \arg\max_i \mu_i^1(S) \tag{11}$$

$$\max_i E[X_i] \approx \max\{\mu_{a^*}^2(S), \ldots, \mu_{a^*}^N(S)\} \tag{12}$$

Where $\mu_i^j(S) = \frac{1}{|S_i^j|} \sum_s \in S_i^j$ and S_i^j is the samples of X_i in S^j, the set of samples for the j-th estimator.

3 Theoretical Analysis

By writing an equation for the expected value of the new estimator similar to (5), we can find that:

$$\sum_{j=1}^{m} P(j = a^*) E[\max\{\mu_j^2, \ldots, \mu_j^N\}] \geq \sum_{j=1}^{m} P(j = a^*) E[\mu_j^2] \tag{13}$$

However, if we try to use a proof similar to Lemma 1 in the paper:

$$\begin{aligned} &E[\max\{\mu_{a^*}^2(S), \ldots, \mu_{a^*}^N(S)\}] \\ &= P(a^* \in \mathcal{M}) E[\max\{\mu_{a^*}^2, \ldots, \mu_{a^*}^N\} | a^* \in \mathcal{M}] \\ &+ P(a^* \notin \mathcal{M}) E[\max\{\mu_{a^*}^2, \ldots, \mu_{a^*}^N\} | a^* \notin \mathcal{M}] \end{aligned} \tag{14}$$

We can find that $E[\max\{\mu_{a^*}^2, \ldots, \mu_{a^*}^N\}] = \max_i E[\mu_{a^*}^i] = \max_i E[X_i]$ is not necessarily true when $a^* \in \mathcal{M}$. Furthermore, by Jensen's inequality, we can see that $E[\max\{\mu_{a^*}^2, \ldots, \mu_{a^*}^N\}] \geq \max_i E[\mu_{a^*}^i]$. Therefore, even though the constructed approximator may have less underestimation bias, it might suffer from overestimation bias instead. Aside from that, more sets of estimators means that the samples are separated into more(and consequently smaller) disjoint sets of samples for each estimators. This leads to each estimator having a larger variance with the same total samples, and thus takes more samples to converge.

With this new approximator of $max_i E[X_i]$, we can design a RL algorithm similar to double Q-learning:

Algorithm 2 N-fold Q-learning

Initialize $Q^1, \ldots, Q^N, s$
repeat
  Choose a, based on $Q^1(s, \cdot), \ldots, Q^N(s, \cdot)$, observe r, s'
  Choose $j \in [1, N]$ randomly
  Define $a^* = \arg\max_a Q^j(s', a)$
  $Q^j(s, a) \leftarrow Q^j(s, a) + \alpha(s, a)(r + \gamma \max_{i \neq j}\{Q^i(s', a^*)\} - Q^j(s, a))$
  $s \leftarrow s'$
until end

4 Conclusion

In this project, we discussed about the feasibility of using more than 2 estimators to construct an approximator of $\max_i E[X_i]$, in contrast of using just 2 in the paper. We found that although such approximator suffers less from underestimation, it might still suffer from overestimation just like using only one set of estimator.
Potential future work of this project may include:

- Using the second way mentioned in (9) and (10) to construct an approximator, since (10) does not have $\max\{\ldots\}$ in the estimated value, it would not overestimate the actual value.
- The use of average instead of $\max\{\ldots\}$. Similar to above, we avoided $\max\{\ldots\}$ so that the approximator does not overestimate.
- Comparison of different values of N based on real-world experiments.

References

Hado Hasselt. Double q-learning. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/091d584fced301b442654dd8c23b3fc9-Paper.pdf.