An Introduction of Double Q-Learning

Bing-Jing Hsieh
Department of Computer Science
National Chiao Tung University
bingjing2000.cs08@nctu.edu.tw

1 Introduction

In this report, first I introduce the paper about Double Q-Learning and what they want to solve. Second, I show some problem formulation in Double Q-Learning paper they give and the problem of single estimator. Third, I show the theoretical analysis of the Double Q-Learning paper and the proof of convergence. Finally, I give some conclusion of this report and the recent research.

In value-based reinforcement learning methods always faced a problem that action value function approximate an error and caused overestimating the estimation. In 2010, van Hasselt et al. introduce a well-known method to solve Q-Learning overestimating action values under some conditions named Double Q-Learning(van Hasselt, 2010). The main reason that Q-Learning algorithm overestimate action value function is Q-Learning using the maximum action value to approximate the maximum expected action value. Therefore, they apply double estimator to Q-Learning that can underestimate action value rather than overestimating. They give a proof of convergence and some experiment to show that the double estimator of Q-Learning can solve the overestimating problem. Although that paper has been 10 years since it publish, and even now using deep neural network to approximate action value function, most of us still using this method to solve action value overestimating. It shows that this is really powerful method and easy to use.

2 Problem Formulation

In this section I show you some notations and assumptions in Double Q-Learning paper and the problem of overestimate action value in single estimator.

We first consider a set of M random variables $X = X_1, ..., X_M$, and we are interested in the maximum expected value of the variables in such a set:

$$\max_i E\{X_i\}. \tag{1}$$

Since we have no knowledge of the the variables in X, it is hard to determine max value of the set exactly. This value is approximated by constructing approximations for $E\{X_i\}$ for all i. They assume that there is a set of samples $S = \bigcup_{i=1}^{M} S_i$, and the samples in S_i are independent and identically distributed (iid)(if it is not iid, the following theory will be hard to prove). Then we can obtain the unbiased expected value of X for each variable by average the sample $E\{X_i\} = E\{\mu_i\} \approx \mu_i(S) \equiv \frac{1}{|S_i|}\sum_{s\in S_i} s$, where μ_i is an estimator for variable X_i. Since the sample $s \in S_i$ is unbiased estimate for the value of expected X_i, the approximation would be unbiased.

Now define the probability density function (PDF) f_i be the ith variable X_i and $F_i(x) = \int_{-\infty}^{x} f_i(x)dx$ is the cumulative distribution function (CDF) of the PDF. And the f_i^{μ} and F_i^{μ} be the PDF and CDF of the ith estimator. Now the maximum expected value can be expressed in terms of the underlying PDF as $\max_i E\{X_i\} = \max_i \int_{-\infty}^{\infty} x f_i(x)dx$.

By notations above in Double Q-Learning paper, we now can approximate the value in (1) with :

$$\max_i E\{X_i\} = \max_i E\{\mu_i\} \approx \max_i \mu_i(S). \tag{2}$$

Preprint. Under review.

According to PDF $f^{\mu}_{\max}$ that is dependent on the PDF of the estimators f^{μ}_i, the maximal estimator $\max_i \mu_i$ is distributed. To determine the PDF, we consider that the CDF $F^{\mu}_{\max}(x)$ gives the probability that the maximum estimate is lower or equal to x. This probability is equal to the probability that all the estimates are lower or equal to $x : F^{\mu}_{max}(x) \equiv P(\max_i \mu_i \leq x) = \prod_{i=1}^{M} P(\mu_i \leq x) \equiv \prod_{i=1}^{M} F^{\mu}_i(x)$. The value $\max_i \mu_i(S)$is an unbiased estimate for $E\{\max_j \mu_j\} = \int_{-\infty}^{\infty} x f^{\mu}_{\max}(x)dx$, which can be written by :

$$E\{\max_j \mu_j\} = \int_{-\infty}^{\infty} x \frac{d}{dx} \prod_{i=1}^{M} F^{\mu}_i(x)dx = \sum_j^M \int_{-\infty}^{\infty} x f^{\mu}_j(s) \prod_{i \neq j}^{M} F^{\mu}_i(x)dx. \tag{3}$$

The single estimator overestimates, because x is within integral and therefore correlates with the monotonically increasing product $\prod_{i \neq j} F^{\mu}_i(x)$. Therefore they use double estimator to handle this problem.

3 Theoretical Analysis

In this section first I will show you the the double estimator is unbiased and the convergence of Double Q-Learning in their paper.

3.1 Double Estimators

In order to use double estimators, first we need to define two set of estimator $\mu^A = \{\mu^A_1, ..., \mu^A_M\}$ and $\mu^B = \{\mu^B_1, ..., \mu^B_M\}$. Both set of estimators need to update with two set of samples S^A and S^B, such that $S^A \cup S^B = S$ and $S^A \cap S^B = \phi$. The expected values $E\{\mu^A_i\} \approx \mu^A_i(S) = \frac{1}{|S^A_i|} \sum_{s \in S^A_i} s$ and $E\{\mu^B_i\} \approx \mu^B_i(S) = \frac{1}{|S^B_i|} \sum_{s \in S^B_i} s$, and both μ^A_i and μ^B_i are unbiased. Let $Max^A(S) \equiv \{j | \mu^A_j(S) = \max_i \mu^A_i(S)\}$ be the set of maximal estimates in $\mu^A(S)$. Then we have $E\{\mu^B_j\} = E\{X_j\}$ for all j and $j \in Max^A$, because μ^B is a set of independent, unbiased estimators. Since we have two set of estimators, we can get a^* form μ^A and $\mu^A_{a^*} \equiv \max_i \mu^A_i(S)$ then use μ_{a^*} as an estimate for $\max_i E\{\mu^B_i\}$ and therefore also for $max_i E\{X_i\}$.Finally, we obtain the approximation $\max_i E\{X_i\} = \max_i E\{\mu^B_i\} \approx \mu^B_{a^*}$ and it will converges to the correct result as we gain more samples and $\mu^A_i(S) = \mu^B_i(S) = E\{X_i\}$ for all i.

The next step we need to do is to prove that the double estimator underestimate. Assume that the PDFs are continuous. The probability $P(j = a^*)$ for any j is then equal to the probability that all $i \neq j$ give lower estimates. Then with probability $\prod_{i \neq j}^{M} P(\mu^A_i < x)$ the $\mu^A_j = x$ is maximal for some x. We then integrate out x giving $P(j = a^*) = \int_{-\infty}^{\infty} P(\mu^A_j = x) \prod_{i \neq j}^{M} P(\mu^A_i < x)dx \equiv \int_{-\infty}^{\infty} f^A_j(x) \prod_{i \neq j}^{M} F^A_i(x)dx$, where f^A_j and F^A_i is the PDF and CDF of μ^A_i. Now we can write the expect value approximate by double estimators

$$\sum_j^M E\{\mu^B_j\} P(j = a^*) = \sum_j^M E\{\mu^B_j\} \int_{-\infty}^{\infty} f^A_j(x) \prod_{i \neq j}^{M} F^A_i(x)dx. \tag{4}$$

Comparing (3) and (4) we can see that the x is replace by $E\{\mu^B_j\}$. Therefore, there is no x in the integral, this will not overestimate the value. Then the double estimator underestimates because the probabilities $P(j = a^*)$ sum to one and the approximation is a weighted estimate of unbiased expected values, which must be lower or equal to the maximum expected value(the proof they give in that paper will show in lemma 1).

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

proof.

if $a^* \in M$, then $E\{\mu^B_{a^*}\} = E\{X_{a^*}\} \equiv \max_i E\{X_i\}$;

else $a^* \notin M$, then we choose $j \in M$, then $E\{\mu^B_{a^*}\} = E\{X_{a^*}\} < E\{X_j\} \equiv \max_i E\{X_i\}$

Therefore we have

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

where the inequality is strict if and only if $P(a^* \notin M) > 0$. This happens when the variables have different expected values, but their distributions overlap. In contrast with the single estimator, the double estimator is unbiased when the variables are iid, since then all expected values are equal and $P(a^* \in M) = 1$. □

3.2 Double Q-Learning

To connect the relationship of single estimator and Q-Learning is to interpret Q-learning as using the single estimator to estimate the action value function $Q(s, a)$ and $\max_a Q(s, a)$ is an estimate for $E\{\max_a Q(s, a)\}$, which in turn approximates $\max_a E\{Q(s, a)\}$. With the connection of single estimator and Q-Learning, we can use double estimator to deal with the overestimating action value for Q-Learning. First, we need two Q function as two estimators: Q^A and Q^B. In Q-Learning we use $Q(s', a^*) = \max_a Q(s', a)$ to update Q. In Double Q-Learning we use $Q^B(s', a^*)$ to update Q^A, where a^* is the maximal valued action in state s', according to the value function Q. Since Q^B is updated by the different set of experience samples and can be considered as an unbiased estimate for the value of this action. The update of Q^B is similar to Q^A that is updated by Q^A and b^*, where b^* is the maximal valued action in state s', according to the value function Q^B. Although single and double estimator both converge to the same answer in the limit, it does not transfer immediately to bootstrapping action values. Then we need to show the convergence of Double Q-Learning in limit. In lemma 2 I will give you their proof in Double Q-Learning paper.

Lemma 2 *Consider a stochastic process $(\zeta_t, \Delta_t, F_t), t \geq 0$, where $\zeta_t, \Delta_t, F_t : X \to R$ satisfy the equation:*

$$\Delta_{t+1}(x_t) = (1 - \zeta_t(x_t))\Delta_t(x_t) + \zeta_t(x_t)F_t(x_t), \tag{5}$$

where $x_t \in X$, and $t = 0, 1, 2, ...$ Let Pt be a sequence of increasing σ-fields such that ζ_0 and δ_0 are P_0-measurable and ζ_t, δ_t and F_{t-1} are P_t-measurable, $t = 1, 2, ...$ Assume that the following hold:

- *The set X is finite*
- *$\zeta_t(x_t) \in [0, 1], \sum_t \zeta_t(x_t) = \infty, \sum_t (\zeta_t(x_t))^2 < \infty$ with probability 1 and $\forall x \neq x_t$: $\zeta(x) = 0$.*
- *$||E\{F_t|P_t\}|| \leq \kappa||\Delta_t|| + c_t$ where $\kappa \in [0, 1)$ and c_t converges to 0 with probability 1.*
- *$Var[F_t(x_t)|P_t] \leq K(1 + \kappa||\Delta_t||)^2$, where K is some constant*

Where $|| \cdot ||$ denotes the maximum norm. Then Δ_t converges to 0 with probability 1.

Theorem 1 *With assumption above then given the following conditions:*

- *The MDP is finite.*
- *Q values are stored in a lookup table.*
- *Each state action pair is sampled an infinite number of times.*
- *Both Q^A and Q^B receive an infinite number of updates.*
- *$\gamma \in [0, 1)$*
- *The learning rates α_t satisfy $\alpha_t(s, a) \in [0, 1], \sum_t \alpha_t(s, a) = \infty, \sum_t (\alpha_t(s, a))^2 < \infty$ with probability 1 and $\forall (s, a) \neq (s_t, a_t) : \alpha_t(s, a) = 0$*
- *$Var[R(s, a)] < \infty, \forall s, a.$*

proof. We first apply Lemma 2 with $P_t = \{Q_0^A, Q_0^B, s_0, a_0, \alpha_0, r_1, s_1, ..., s_t, a_t\}, X = S \times A, \Delta_t = Q_t^A - Q^*, \zeta_t = \alpha_t$.

Defining $a^* = argmax_a Q^A(s_{t+1}, a)$. *and* $F_t(s_t, a_t) = r_t + \gamma Q_t^B(s_{t+1}, a^*) - Q_t^*(s_t, a_t)$. *we can write*

$$F_t(s_t, a_t) = F_t^Q(s_t, a_t) + \gamma(Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, a^*)).$$

where $F_t^Q(s_t, a_t) = r_t + \gamma Q_t^A(s_{t+1}, a^*) - Q_t^*(s_t, a_t)$ *is the value of* F_t *in standard Q-learning and* $c_t = \gamma(Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, a^*))$. *As* $E\{F_t^Q|P_t\} \leq \gamma||\Delta_t||$ *is a well-known result, then apply the lemma if* c_t *converge to zero,* $\Delta_t^{BA} = Q_t^B - Q_t^A$ *converge to zero. Depending on whether* Q^A *or* QB *is updated, the update of* Δ_t^{BA} *at time t is either*

$$\Delta_{t+1}^{BA}(s_t, a_t) = \Delta_t^{BA}(s_t, a_t) + \alpha(s_t, a_t)F_t^B(s_t, a_t), or$$
$$\Delta_{t+1}^{BA}(s_t, a_t) = \Delta_t^{BA}(s_t, a_t) - \alpha(s_t, a_t)F_t^A(s_t, a_t),$$

where $F_t^A(s_t, a_t) = r_t + \gamma Q_t^B(s_{t+1}, a^*) - Q_t^A(s_t, a_t)$ *and* $F_t^B(s_t, a_t) = r_t + \gamma Q_t^A(s_{t+1}, b^*) - Q_t^B(s_t, a_t)$. *Then we define* $\zeta_t^{BA} = \frac{1}{2}\alpha_t$,

$$\begin{aligned} E\{\Delta_{t+1}^{BA}(s_t, a_t)|P_t\} &= \Delta_t^{BA}(s_t, a_t) + E\{\alpha(s_t, a_t)F_t^B(s_t, a_t) - \alpha(s_t, a_t)F_t^A(s_t, a_t)|P_t\} \\ &= (1 - \zeta_t^{BA}(s_t, a_t))\Delta_t^{BA}(s_t, a_t) + \zeta_t^{BA}(s_t, a_t)E\{F_t^{BA}(s_t, a_t)|P_t\}, \end{aligned}$$

where $E\{F_t^{BA}(s_t, a_t)|P_t\} = \gamma E\{Q_t^A(s_{t+1}, b^*) - Q_t^B(s_{t+1}, a^*)|P_t\}$. *And they randomly select whether to update* Q^A *or* Q^B.

Assume $E\{Q_t^A(s_{t+1}, b^*)|P_t\} \geq E\{Q_t^B(s_{t+1}, a^*)|P_t\}$. *By definition of* a^* *we have* $Q_t^A(s_{t+1}, a^*) = \max_a Q_t^A(s_{t+1}, a) \geq Q_t^A(s_{t+1}, b^*)$, *then*

$$\begin{aligned} |E\{F_t^{BA}(s_t, a_t)|P_t\}| &= \gamma E\{Q_t^A(s_{t+1}, b^*) - Q_t^B(s_{t+1}, a^*)|P_t\} \\ &\leq \gamma E\{Q_t^A(s_{t+1}, a^*) - Q_t^B(s_{t+1}, a^*)|P_t\} \\ &\leq \gamma||\Delta_t^{BA}||. \end{aligned}$$

Now assume $E\{Q_t^A(s_{t+1}, a^*)|P_t\} > E\{Q_t^A(s_{t+1}, b^*)|P_t\}$ *and also that by definition of* b^* *we have*$Q_t^B(s_{t+1}, b^*) = \max_a Q_t^B(s_{t+1}, a) \geq Q_t^B(s_{t+1}, a^*)$.*Then*

$$\begin{aligned} |E\{F_t^{BA}(s_t, a_t)|P_t\}| &= \gamma E\{Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, b^*)|P_t\} \\ &\leq \gamma E\{Q_t^B(s_{t+1}, b^*) - Q_t^A(s_{t+1}, b^*)|P_t\} \\ &\leq \gamma||\Delta_t^{BA}||. \end{aligned}$$

One of two of assumptions must hold at each time step and in both cases we obtain the result that $E\{F_t^{BA}|P_t\} \leq \gamma||\Delta_t^{BA}||$. *By the Lemma 2 the* Δ_t^{BA} *converge to zero , which ensures that the original process also converges in the limit.* □

4 Conclusion

Double Q-learning is not a full solution to the problem of finding the maximum of the expected values of the actions. The action a^* may not be the action that maximizes the expected Q function $\max_a E\{Q^A(s', a)\}$. In general $E\{Q^B(s', a^*)\} \leq \max_a EQ^A(s', a^*)$, and underestimations of the action values can occur. There are still an underestimate bias in action value function. Since the underestimations bias, current research want to more precisely predict true and reduce the bias. Like this year Qingfeng Lan et al. proposed an method call max min Q-learning (MAXMIN Q-LEARNING: CONTROLLING THE ESTIMATION BIAS OF Q-LEARNING Qingfeng Lan et al. ICLR 2020) to flexible control the estimation bias of Q-Learning. In future work of this estimation problem we need to decrease the estimation bias and the estimation variance, so that we can approximate the true value and get the better result of many problems.