A Note on Double Q-Learning

JiJia Wu
Department of Computer Science
National Yang Ming Chiao Tung University
jia.cs07@nycu.edu.tw

1 Introduction

Q-learning performs poorly in some stochastic environments. The poor performance is usually caused by large overestimations of action values. To deal with this problem, Hasselt [2010] proposed Double Q-learning which uses a double estimator to tackle this issue.

In this note, the proofs in the paper will be discussed, some part of them are considered not rigorous enough and having some issues. I will provide my proof for the convergence of Double Q-learning to correct the issues. The estimators will be analyzed as well, which shows that Q-learning overestimates the maximum expected action value while Double Q-learning underestimates it.

I think this paper proposed an alternative way to estimate the maximum expected action value. Despite the experimental results in the paper show that Double Q-learning can have lower bias and better performance than Q-learning, however, there's no guarantee that Double Q-learning can always lead to lower bias. Besides, the effect of positive bias and negative bias can still be discussed, as there might be some different effects on training.

2 Problem Formulation

2.1 Q-learning

Algorithm 1 shows the algorithm of Q-learning. Q-learning stores single Q function: Q. $Q_t(s, a)$ gives the value of the action a in state s at time t, r represents the immediate reward in which r_t give the immediate reward at time t. The learning rate $\alpha(s, a) \in [0, 1]$ ensures that the update averages over possible randomness in the rewards and transitions in order to converge in the limit to the optimal action value function.

2.2 Double Q-learning

Algorithm 2 shows the algorithm of Double Q-learning. Double Q-learning stores two functions: Q^A and Q^B. Each Q function is updated with a value from the other Q function for the next state. The action a^* and b^* are the maximal valued action in state s', according to the value function Q^A and Q^B.

Both Q-learning and Double Q-learning are algorithms to solve Markov Decision Processes (MDPs) with finite state and action spaces, bounded rewards. In the limit, Double Q-learning converges to the optimal policy with probability one (w.p.1). The proofs will be provided in the next section.

In Double Q-learning, it replaces single Q function with two Q functions. The difference between these two methods will be analyzed.

Algorithm 1 Q-learning

```
Initialize $Q$, s
repeat
    Choose $\alpha$, based on $Q(s,\cdot)$, observe $r$, $s'$
    Define $a^* = argmax_{a'}Q(s',a')$
    $Q(s,a) \leftarrow Q(s,a) + \alpha(s,a)(r + \gamma Q(s',a^*) - Q^A(s,a))$
    $s \leftarrow s'$
until end
```

Algorithm 2 Double Q-learning

```
Initialize $Q^A$, $Q^B$, s
repeat
    Choose $\alpha$, based on $Q^A(s,\cdot)$ and $Q^B(s,\cdot)$, observe $r$, $s'$
    Choose (e.g. random) either UPDATE(A) or UPDATE(B)
    if UPDATE(A) then
        Define $a^* = argmax_{a'}Q^A(s',a')$
        $Q^A(s,a) \leftarrow Q^A(s,a) + \alpha(s,a)(r + \gamma Q^B(s',a^*) - Q^A(s,a))$
    else if  UPDATE(A) then
        Define $b^* = argmax_{a'}Q^B(s',a')$
        $Q^B(s,a) \leftarrow Q^B(s,a) + \alpha(s,a)(r + \gamma Q^A(s',b^*) - Q^B(s,a))$
    end if
    $s \leftarrow s'$
until end
```

3 Theoretical Analysis

In this section, I'll analysis the following properties. Overestimation of Q-learning, underestimation of Double Q-learning, and the convergence of Double Q-learning.

3.1 Single Estimator

Consider a set of M random variables $X = \{X_1, ..., X_M\}$. $\mu = \{\mu_1, ..., \mu_M\}$ is a set of unbiased estimators such that $E[\mu_i] = E[X_i]$, for all i. Let $\mathcal{M} \overset{\text{def}}{=} \{j|E[X_j] = max_i E[X_i]\}$ be the set of elements that maximize the expected values. Let a^* be an element that maximizes $\mu : \mu^A_{a^*} = max_i \mu^A_i$. Then

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

Proof. Assume $a^* \in \mathcal{M}$. Then $E[\mu_{a^*}] = E[X_{a^*}] \overset{\text{def}}{=} max_i E[X_i]$. Now assume $a^* \notin \mathcal{M}$ and $j \in \mathcal{M}$. Then $E[\mu_{a^*}] = E[X_{a^*}] < E[X_j] \overset{\text{def}}{=} max_i E[X_i]$. These two possibilities are mutually exclusive, so the combined expectation can be expressed as

$$\begin{aligned} E[max_i(\mu_i)] &= P(a^* \in \mathcal{M})E[\mu_{a^*}|a^* \in \mathcal{M}] + P(a^* \notin \mathcal{M})E[\mu_{a^*}|a^* \notin \mathcal{M}] \\ &= P(a^* \in \mathcal{M})max_i E[X_i] + P(a^* \notin \mathcal{M})E[\mu_{a^*}|a^* \notin \mathcal{M}] \\ &\geq P(a^* \in \mathcal{M})max_i E[X_i] + P(a^* \notin \mathcal{M})max_i E[X_i] \\ &= max_i E[X_i] \end{aligned} \tag{2}$$

where the inequality is strict if and only if $P(a^* \notin \mathcal{M}) > 0$. This happens when the variables' distributions overlap. □

3.2 Double Estimator

Consider a set of M random variables $X = \{X_1, ..., X_M\}$. Let $\mu^A = \{\mu^A_1, ..., \mu^A_M\}$ and $\mu^B = \{\mu^B_1, ..., \mu^B_M\}$ be a set of unbiased estimators such that $E[\mu^A_i] = E[\mu^B_i] = E[X_i]$, for all i. Let $M \overset{\text{def}}{=} \{j|E[X_j] = max_i E[X_i]\}$ be the set of elements that maximize the expected values. Let a^* be an element that maximizes μ^A: $\mu^A_{a^*} = max_i \mu^A_i$. Then

$$E[\mu^B_{a^*}] = E[X_{a^*}] \leq max_i E[X_i] \tag{3}$$

Proof. Assume $a^* \in \mathcal{M}$. Then $E[\mu_{a^*}] = E[X_{a^*}] \overset{\text{def}}{=} max_i E[X_i]$. Now assume $a^* \notin \mathcal{M}$ and $j \in \mathcal{M}$. Then $E[\mu_{a^*}] = E[X_{a^*}] < E[X_j] \overset{\text{def}}{=} max_i E[X_i]$. These two possibilities are mutually exclusive, so the combined expectation can be expressed as

$$\begin{aligned} E[\mu^B_{a^*}] &= P(a^* \in \mathcal{M}) E[\mu^B_{a^*} | a^* \in \mathcal{M}] + P(a^* \notin \mathcal{M}) E[\mu^B_{a^*} | a^* \notin \mathcal{M}] \\ &= P(a^* \in \mathcal{M}) max_i E[X_i] + P(a^* \notin \mathcal{M}) E[\mu^B_{a^*} | a^* \notin \mathcal{M}] \\ &\leq P(a^* \in \mathcal{M}) max_i E[X_i] + P(a^* \notin \mathcal{M}) max_i E[X_i] \\ &= max_i E[X_i] \end{aligned} \tag{4}$$

where the inequality is strict if and only if $P(a^* \notin \mathcal{M}) > 0$. This happens when the variables' 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 $P(a^* \in \mathcal{M}) = 0$. □

3.3 Convergence of Double Q-learning

In the part, the proof in the paper will be shown here. I'll also add some brief comments on the problems or typos founded in the proof. I'll discuss about whether those problems will cause some issues in Section 3.4. Also, I'll provide my proof for the parts I think that is not rigorous enough in Section 3.6. Before we start proving the convergence of Double Q-learning, let's consider Lemma 1 proposed by Jaakkola et al. [1994] first.

Lemma 1. *A random iterative process $\Delta_{t+1}(x) = (1 - \zeta_t(x))\Delta_t(x) + \beta_t(x)F_t(x)$ converges to zero w.p.1 under the following assumptions:*

1. *The state space is finite.*
2. $\sum_t \zeta_t(x) = \infty$, $\sum_t \zeta_t^2(x) < \infty$, $\sum_t \beta_t(x) = \infty$, $\sum_t \beta_t^2(x) < \infty$, *and* $E[\beta_t(x)|P_t] < E[\zeta_t(x)|P_t]$ *uniformly w.p.1.*
3. $\|E[F_t(x)|P_t]\| \leq \gamma \|\Delta_t\|$, *where* $\gamma \in [0, 1)$.
4. $Var[F_t(x)|P_t] \leq C(1 + \|\Delta_t\|)^2$, *where C is some constant.*

Here $P_t = \{\Delta_t, \Delta_{t-1}, ..., F_{t-1}, ..., \alpha_{t-1}, ..., \beta_{t-1}, ...\}$ *stands for the past at step t.* $F_t(x)$, $\alpha_t(x)$ *and* $\beta_t(x)$ *are allowed to depend on the past insofar as the above conditions remain valid. The notation* $\|\cdot\|$ *refers to maximum norm.*

With Lemma 1, they prove the convergence of Double Q-learning under similar conditions as Q-learning. Their theorem is as follows:

Theorem 1. *Assume the conditions below are fulfilled. Then, in a given ergodic MDP, both Q^A and 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 state action pair are given by a proper learning policy. The additional conditions are:*

1. *The MDP is finite, i.e.* $|S \times A| < \infty$.
2. $\gamma \in [0, 1)$.
3. *The Q values are stored in a lookup table.*
4. *Both Q^A and Q^B receive an infinite number of updates.*
5. $\alpha_t(s, a) \in [0, 1]$, $\sum_t \alpha_t(s, a) = \infty$, $\sum_t (\alpha_t(s, a))^2 < \infty$ *w.p.1, and* $\forall (s, a) \neq (s_t, a_t) :$ $\alpha_t(s, a) = 0$.
6. $\forall s, a, s' : Var[R^{s'}_{s\alpha}] < \infty$.

Proof. They sketch how to apply Lemma 1 to prove theorem 1. Because the updates on the functions Q^A and Q^B is symmetry, it suffices to show convergence for either of these. They apply Lemma 1 with the following definitions:

1. $P_t = \{Q_t^A, Q_t^B, s_0, a_0, \alpha_0, r_1, s_1, a_1, \alpha_1, ..., s_t, a_t\}$
2. $X = S \times A$
3. $\Delta_t = Q_t^A - Q_t^*$
4. $\zeta = \beta = \alpha$
5. $F_t(s_t, a_t) = F_t^Q(s_t, a_t) + \gamma(Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, a^*))$,
   where $F_t^Q(s_t, a_t) = r_t + \gamma Q_t^A(s_{t+1}, a^*) - Q_t^*(s_t, a_t)$

The first two conditions of the lemma hold. The fourth condition of the lemma holds as a consequence of the boundedness condition on the variance of the rewards in the theorem.

This leaves to show that the third condition on the expected contraction of F_t holds. We can write

$$F_t(s_t, a_t) = F_t^Q(s_t, a_t) + \gamma(Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, a^*)) \tag{5}$$

where $F_t^Q = r_t + \gamma Q_t^A(s_{t+1}, a^*) - Q_t^*(s_t, a_t)$ is the value of F_t if normal Q-learning would be under consideration. It is well-known that $E[F_t^Q|P_t] \leq \gamma \|\Delta_t\|$, so to apply the lemma, we identify $c_t = \gamma Q_t^B(s_{t+1}, a^*) - \gamma Q_t^A(s_{t+1}, a^*)$, converges to zero, and it suffices to show that $\Delta_t^{BA} = Q_t^B - Q_t^A$ converges to zero. Depending on whether Q^A or Q^B is updated, the update of δ_t^{BA} is either:

$$\begin{aligned} \Delta_{t+1}^{BA}(s_t, a_t) &= \Delta_t^{BA}(s_t, a_t) + \alpha_t(s_t, a_t)F_t^B(s_t, a_t), or \\ \Delta_{t+1}^{BA}(s_t, a_t) &= \Delta_t^{BA}(s_t, a_t) - \alpha_t(s_t, a_t)F_t^A(s_t, a_t) \end{aligned} \tag{6}$$

where

$$\begin{aligned} F_t^A(s_t, a_t) &= r_t + \gamma Q_t^B(s_{t+1}, a^*) - Q_t^A(s_t, a_t) \\ F_t^B(s_t, a_t) &= r_t + \gamma Q_t^A(s_{t+1}, b^*) - Q_t^B(s_t, a_t) \\ E[F_t^{BA}(s_t, a_t)|P_t] &= \gamma E[Q_t^A(s_{t+1}, b^*) - Q_t^B(s_{t+1}, a^*)] \\ \zeta &= \frac{1}{2}\alpha \end{aligned} \tag{7}$$

Then

$$\begin{aligned} E[\Delta_{t+1}^{BA}(s_t, a_t)|P_t] &= \Delta_t^{BA}(s_t, a_t) + E[\alpha_t(s_t, a_t)F_t^B(s_t, a_t) - \alpha_t(s_t, a_t)F_t^A(s_t, a_t)|P_t] \\ &= (1 - \zeta_t(s_t, a_t))\Delta_t^{BA}(a_t, a_t) + \zeta_t(s_t, a_t)E[F_t^{BA}(s_t, a_t)|P_t] \end{aligned} \tag{8}$$

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

$$\begin{aligned} |E[F_t^{BA}(s_t, a_t)|P_t]| &= \gamma E[Q_t^A(s_{t+1}, b^*) - Q_t^B(s_{t+1}, a^*)] \\ &\leq \gamma E[Q_t^A(s_{t+1}, a^*) - Q_t^B(s_{t+1}, a^*)] \leq \gamma \left\|\Delta_t^{BA}\right\| \end{aligned} \tag{9}$$

Now assume $Q_t^B(s_{t+1}, a^*)] \geq E[Q_t^A(s_{t+1}, b^*)]$. By definition of Algorithm **??** we have $Q_t^B(s_{t+1}, b^*) = max_a Q_t^B(s_{t+1}, a) \geq Q_t^B(s_{t+1}, a^*)$

$$\begin{aligned} |E[F_t^{BA}(s_t, a_t)|P_t]| &= \gamma E[Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, b^*)] \\ &\leq \gamma E[Q_t^B(s_{t+1}, b^*) - Q_t^A(s_{t+1}, b^*)] \leq \gamma \left\|\Delta_t^{BA}\right\| \end{aligned} \tag{10}$$

Clearly, one of the two assumptions must hold at each time step and in both cases we obtain the desired result that $|E[F_t^{BA}(s_t, a_t)|P_t]| \leq \gamma \left\|\Delta_t^{BA}\right\|$. Applying the lemma yields convergence of Δ_t^{BA} to zero, which in turn ensures that the original process also converges in the limit. □

3.4 My Opinion on the Proof of Double D-learning

In this part, I wrote two problems I noticed in the proof.

Problem 1 In the paper of Double Q-learning, it is written that it is well-known that $E[F_t^Q(s_t, a_t)|Pt] \leq \gamma \|\Delta_t\|$, but we can notice that it cannot give any support for $E[F_t(s_t, a_t)|Pt] \leq$

$\gamma \left\| \Delta_t \right\|$, which is the third condition of Lemma 1. In other words, there might exist some $F_t(s_t, a_t)$ such that $\forall \gamma \in [0, 1)$,

$$\begin{aligned} |E[F_t(s_t, a_t)|P_t]| &= |E[F_t^Q(s_t, a_t) + \gamma(Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, a^*))]| \\ &> \gamma \left\| \Delta_t \right\| \end{aligned} \tag{11}$$

Therefore, we cannot apply Lemma 1 to Theorem 1 since the third condition is not hold.

Problem 2 Let's take a look at Equation 8, Equation 9, and Equation 10 again, we can notice that they miss some $E[\cdot]$ and $| \cdot |$. Equation 8 should be as the following,

$$\begin{aligned} E[\Delta_{t+1}^{BA}(s_t, a_t)|P_t] &= E[\Delta_t^{BA}(s_t, a_t)] + E[\alpha_t(s_t, a_t)F_t^B(s_t, a_t) - \alpha_t(s_t, a_t)F_t^A(s_t, a_t)|P_t] \\ &= (1 - \zeta_t(s_t, a_t))E[\Delta_t^{BA}(a_t, a_t)] + \zeta_t(s_t, a_t)E[F_t^{BA}(s_t, a_t)|P_t] \end{aligned} \tag{12}$$

And the Equation 9 and Equation 10 should be:

$$\begin{aligned} |E[F_t^{BA}(s_t, a_t)|P_t]| &= |E[E[F_t^{BA}(s_t, a_t)|P_t]]| \\ &= \gamma |E[Q_t^A(s_{t+1}, b^*) - Q_t^B(s_{t+1}, a^*)]| \\ &\leq \gamma |E[Q_t^A(s_{t+1}, a^*) - Q_t^B(s_{t+1}, a^*)]| \leq \gamma \left\| E[\Delta_t^{BA}] \right\|, or \\ |E[F_t^{BA}(s_t, a_t)|P_t]| &= |E[E[F_t^{BA}(s_t, a_t)|P_t]]| \\ &= \gamma |E[Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, b^*)]| \\ &\leq \gamma |E[Q_t^B(s_{t+1}, b^*) - Q_t^A(s_{t+1}, b^*)]| \leq \gamma \left\| E[\Delta_t^{BA}] \right\| \end{aligned} \tag{13}$$

Does it really matter? The answer is "Yes". Let's apply this equation to Lemma 1 again, we can only go as the following way

- $\zeta = \beta = \alpha$
- $\Delta_t = E[\Delta_t^{BA}|P_t]$, for all t
- $F_t = E[F_t^{BA}|P_t]$, for all t

Therefore we can only have the result that $|E[E[F_t^{BA}|P_t]]| = |E[F_t]| \leq \gamma \left\| E[\Delta_t^{BA}] \right\| = \gamma \left\| \Delta_t \right\|$. which can only lead to the conclusion that $E[\Delta_t^{BA}]$ converges to zero. We have no guarantee of Δ_t^{BA} converges to zero since the variance of Δ_t^{BA} is still an issue.

3.5 Some Preparation for My Proof

In this part, the three lemmas to prove theorom 1 proposed by Jaakkola et al. [1994] will be mentioned. Besides, theorem 1 will be extended so that we can apply it to prove the convergence of Double Q-learning. The following are the three lemmas:

Lemma 2. *A random process*

$$w_{t+1}(x) = (1 - \alpha_t(x))w_t(x) + \beta_t(x)r_t(x) \tag{14}$$

converges to zero w.p.1 if the following conditions are satisfied:

1. $\sum_t \zeta_t(x) = \infty$, $\sum_t \zeta_t^2(x) < \infty$, $\sum_t \beta_t(x) = \infty$, $\sum_t \beta_t^2(x) < \infty$, *and* $E[\beta_t(x)|P_t] < E[\zeta_t(x)|P_t]$ *uniformly w.p.1*
2. $E[r_t(x)|P_t] = 0$ *and* $E[r_t^2(x)|P_t] \leq C$, *where C is a constant and* $P_t = \{w_t, w_{t-1}, ..., r_{t-1}, ..., \alpha_{t-1}, ..., \beta_{t-1}, ...\}$

Lemma 3. *Consider a random process $X_{n+1}(x) = G_n(X_n, x)$, where*

$$G_n(\beta X_n, x) = \beta G_n(X_n, x) \tag{15}$$

Let us suppose that if we kept $\|X_n\|$ bounded by scaling, then X_n would converge to zero w.p.1. This assumption is sufficient to guarantee that the original process converges to zero w.p.1.

Lemma 4. *A stochastic process $X_{n+1}(x) = (1 - \alpha(x))X_n(x) + \gamma\beta_n(x) \left\| X_n \right\|$ converges to zero w.p.1 provided*

1. *$x \in S$, S is a finite set.*
2. *$\sum_t \zeta_t(x) = \infty$, $\sum_t \zeta_t^2(x) < \infty$, $\sum_t \beta_t(x) = \infty$, $\sum_t \beta_t^2(x) < \infty$, and $E[\beta_t(x)|P_t] < E[\zeta_t(x)|P_t]$ uniformly w.p.1*

Lemma 5. *A stochastic process $X_{n+1}(x) = (1 - \alpha(x))X_n(x) + \gamma\beta_n(x)(\|X_n\| + \lambda_n)$ converges to zero w.p.1 provided*

1. *$x \in S$, S is a finite set.*
2. *$\sum_t \zeta_t(x) = \infty$, $\sum_t \zeta_t^2(x) < \infty$, $\sum_t \beta_t(x) = \infty$, $\sum_t \beta_t^2(x) < \infty$, and $E[\beta_t(x)|P_t] < E[\zeta_t(x)|P_t]$ uniformly w.p.1*
3. *λ converges to zero m.p.1.*

Proof. Similar to the proof of Lemma 4, which is shown in Jaakkola et al. [1994], this proof is an application of Lemma 3. Assume that we keep $\|X_n\| + |\lambda_n| \le M_1 + L_1 = C_1$ by scaling which allows the iterative process to be bounded by

$$|X_{n+1}(x)| \le (1 - \alpha_n(x))|X_n(x)| + \gamma\beta_n(x)C_1 \tag{16}$$

This is linear in $|X_n(x)|$ and can be easily shown to converge w.p.1 to some $X^*(x)$, where $\|X^*\| \le \gamma C_1$. Hence, for small enough ϵ, there exists $M_1(\epsilon)$ such that $\|X_n\| + |\lambda_n| \le C_1/(1+\epsilon)$ for all $n > M_1(\epsilon)$ with probability at least $p_1(\epsilon)$. With probability $p_1(\epsilon)$ the procedure can be repeated for $C_2 = C_1/(1+\epsilon)$. Continuing in this manner and choosing $p_k(\epsilon)$ so that $\prod_k p_k(\epsilon)$ goes to one as $\epsilon \to 0$ we obtain the w.p.1 convergence of the bounded iteration and Lemma 2 can be applied. □

Theorem 2. *A random iterative process $\Delta_{t+1}(x) = (1 - \zeta_t(x))\Delta_t(x) + \beta_t(x)F_t(x)$ converges to zero w.p.1 under the following assumptions:*

1. *The state space is finite.*
2. *$\sum_t \zeta_t(x) = \infty$, $\sum_t \zeta_t^2(x) < \infty$, $\sum_t \beta_t(x) = \infty$, $\sum_t \beta_t^2(x) < \infty$, and $E[\beta_t(x)|P_t] < E[\zeta_t(x)|P_t]$.*
3. *$\|E[F_t(x)|P_t]\| < \gamma\,\|\Delta_t\| + \|\Lambda\|$, where $\gamma \in [0,1)$, $\|\Lambda\| < \infty$ and $\|\Lambda\|$ converge to zero w.p.1.*
4. *$Var[F_t(x)|P_t] \le C(1 + \|\Delta_t\|)^2$, where C is some constant.*

Here $P_t = \{X_t.X_{t-1}, ..., F_{t-1}, .., \zeta_{t-1}, ..., \beta_{t-1}, ...\}$ stands for the past at step t. $F_n(x)$, $\alpha_n(x)$ and $\beta_n(x)$ are allowed to depend on the past insofar as the above conditions remain valid.

Proof. By defining $r_t(x) = F_t(x) - E[F_t(x)|P_t]$ we can show two parallel iterative processes

$$\begin{aligned}\delta_{t+1}(x) &= (1 - \zeta_t(x))\delta(x) + \beta_t(x)E[F_t(x)|P_t] \\ \omega_{t+1}(x) &= (1 - \zeta_t(x))\delta(x) + \beta_t(x)r_t(x)\end{aligned} \tag{17}$$

by using the absolute value and maximum norm operator, we can show that

$$\begin{aligned}|\delta_{t+1}(x)| &= |(1 - \zeta_t(x))\delta(x) + \beta_t(x)E[F_t(x)|P_t]| \\ &\le |(1 - \zeta_t(x))\delta(x)| + |\beta_t(x)E[F_t(x)|P_t]| \\ &\le (1 - \alpha_t(x))|\delta(x)| + \gamma\beta_t(x)(\|\delta + \omega\| + \|X\|) \\ &\le (1 - \alpha_t(x))|\delta(x)| + \gamma\beta_t(x)(\|\delta\| + \|\omega\| + \|X\|)\end{aligned} \tag{18}$$

As the variance of $r_t(x)$ bounded by some constant C and thereby ω_t converges to zero w.p.1 according to Lemma 1. Also, as the assumption of $|X|$ converges to zero w.p.1, there exists M such that for all $n > M$, $\|\omega\| < \epsilon/2$ and $\|X\| < \epsilon/2$ with probability at least 1 - ϵ. This implies that the δ_t process can be further bounded by

$$\begin{aligned}|\delta_{t+1}(x)| &\le (1 - \alpha_t(x))|\delta(x)| + \gamma\beta_t(x)(\|\delta\| + \epsilon/2 + \epsilon/2) \\ &= (1 - \alpha_t(x))|\delta(x)| + \gamma\beta_t(x)(\|\delta\| + \epsilon)\end{aligned} \tag{19}$$

with probability over $1 - \epsilon$. If we choose C such that $\gamma(C+1)/C < 1$ then for $\|\delta_t\| > C\epsilon$,

$$\gamma(\|\delta_t\| + \epsilon) \le \gamma(C+1)/C\,\|\delta_t\| \tag{20}$$

Therefore we can show that δ converges w.p.1 to some value bounded by $C\epsilon$. It shows that the convergence of Δ bounded by δ and ω. □

3.6 My Proof for the Convergence of Double Q-learning

In this part, I will proof Theorem 1 again and fix the problem mentioned in Section 3.4. To rewrite the proof, I want to start from the proof for problem 2.

Lemma 6. *In the same condition of Theorem 1, $Q^A - Q^B$ converges to zero w.p.1,*

Proof. Let's apply Lemma 1 with the following definitions:

1. $P_t = \{Q_t^A, Q_t^B, s_0, a_0, \alpha_0, r_1, s_1, a_1, \alpha_1, ..., s_t, a_t\}$
2. $X = S \times A$
3. $\Delta_t = Q_t^A - Q_t^B$
4. $\zeta = \beta = \alpha$
5. $F_t = F_t^A$ when updating Q^A, $F_t = F_t^B$ when updating Q^B, where
   $F_t^A(s_t, a_t) = -(r_t + \gamma Q_t^B(s_{t+1}, a^*) - Q_t^B(s_t, a_t))$
   $F_t^B(s_t, a_t) = r_t + \gamma Q_t^A(s_{t+1}, b^*) - Q_t^A(s_t, a_t)$

Remind the equation of random iterative process

$$\Delta_{t+1}(x) = (1 - \zeta_t(x))\Delta_t(x) + \beta_t(x)F_t(x) \tag{21}$$

We can show that when updating Q^A , by applying the definitions, and $Q_{t+1}^B = Q_t^B$ as when updating Q^A, Q^B is freezed, the equation of random iterative process is the same as the updating of Q^A such that

$$Q_{t+1}^A(s_t, a_t) = (1 - \alpha_t(s_t, a_t))Q_t^A(s_t, a_t) + \alpha_t(s_t, a_t)(r_t + \gamma Q_t^B(s_{t+1}, a^*)) \tag{22}$$

We can also show that when updating Q^B , by applying the definitions, and $Q_{t+1}^A = Q_t^A$ as when updating Q^B, Q^A is freezed, the equation of random iterative process is the same as the updating of Q^B such that

$$Q_{t+1}^B(s_t, a_t) = (1 - \alpha_t(s_t, a_t))Q_t^B(s_t, a_t) + \alpha_t(s_t, a_t)(r_t + \gamma Q_t^A(s_{t+1}, b^*)) \tag{23}$$

It is straightforward to show the first two conditions of the lemma hold. The fourth condition of the lemma holds as a consequence of the boundedness condition on the variance of the rewards in the theorem.

This leaves to show that the third condition on the expected contraction F_t holds. As $\gamma E[Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, b^*)] \leq \left\| Q_t^A - Q_t^B \right\| \overset{\text{def}}{=} \gamma \left\| \Delta_t^{BA} \right\|$ can be shown as the following:

$$\begin{aligned}
|E[F_t^{BA}(s_t, a_t)|P_t]| &= \gamma E[Q_t^A(s_{t+1}, b^*) - Q_t^B(s_{t+1}, a^*)] \\
&\leq \gamma E[Q_t^A(s_{t+1}, a^*) - Q_t^B(s_{t+1}, a^*)] \leq \gamma \left\| \Delta_t^{BA} \right\|, \text{or} \\
|E[F_t^{BA}(s_t, a_t)|P_t]| &= \gamma E[Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, b^*)] \\
&\leq \gamma E[Q_t^B(s_{t+1}, b^*) - Q_t^A(s_{t+1}, b^*)] \leq \gamma \left\| \Delta_t^{BA} \right\|
\end{aligned} \tag{24}$$

The absolute value of the expected value of F^t has the property that,

$$\begin{aligned}
|E[F_t(s_t, a_t)]| &= |\frac{1}{2}E[F_t^A(s_t, a_t)] + \frac{1}{2}E[F_t^B(s_t, a_t)]| \\
&= |\frac{1}{2}E[\gamma(Q_t^A(s_{t+1}, b^*) - Q_t^B(s_{t+1}, a^*)) + (Q_t^B(s_t, a_t) - Q_t^A(s_t, a_t))]| \\
&\leq |\frac{1}{2}\gamma E[Q_t^A(s_{t+1}, b^*) - Q_t^B(s_{t+1}, a^*)]| + |\frac{1}{2}E[Q_t^B(s_t, a_t) - Q_t^A(s_t, a_t)]| \\
&\leq \frac{1}{2}\gamma|E[Q_t^A(s_{t+1}, b^*) - Q_t^B(s_{t+1}, a^*)]| + \frac{1}{2}\left\| \Delta_t^{BA} \right\| \\
&\leq \frac{1}{2}\gamma \left\| \Delta_t^{BA} \right\| + \frac{1}{2}\left\| \Delta_t^{BA} \right\| \\
&= \frac{1}{2}(\gamma + 1)\left\| \Delta_t^{BA} \right\|
\end{aligned} \tag{25}$$

Therefore the third condition is also fulfilled since $\frac{1}{2}(\gamma + 1) \in [0, 1)$. So we can know that Δ^{AB} converges to zero by applying Lemma 1. □

Finally, we can show that Double Q-learning converges to the optimal policy by the following proof.

Proof. Let's apply following definitions to Theorem 2:

1. $P_t = \{Q_t^A, Q_t^B, s_0, a_0, \alpha_0, r_1, s_1, a_1, \alpha_1, ..., s_t, a_t\}$
2. $X = S \times A$
3. $\Delta_t = Q_t^A - Q_t^*$
4. $\zeta = \beta = \alpha$
5. $F_t(s_t, a_t) = F_t^Q(s_t, a_t) + \gamma(Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, a^*))$,
   where $F_t^Q(s_t, a_t) = r_t + \gamma Q_t^A(s_{t+1}, a^*) - Q_t^*(s_t, a_t)$
6. $\Lambda = Q_t^A - Q_t^B$

It is straightforward to show the first two conditions of the lemma hold. The fourth condition of the lemma holds as a consequence of the boundedness condition on the variance of the rewards in the theorem. As it is well-known that $|E[r_t + \gamma Q_t^A(s_{t+1}, a^*) - Q_t^*(s_t, a_t)]| \leq \left\|Q_t^A - Q_t^*\right\|$ can be prove by Bellman optimality equation, we have $E[F_t^Q|P_t] \leq \gamma \left\|\Delta_t\right\|$. Also, as we have proved that $Q_t^A - Q_t^B$ converges to zero w.p.1, the third condition is also satisfied. Therefore we can make the conclusion that Double Q-learning converge to the optimal solution in the limit.

□

4 Conclusion

In this note, We show that the double estimator underestimates the maximum expected action value while the single estimator overestimates it. And show that Double Q-learning converges to the optimal policy in the limit with my proof.

Double Q-learning gives us an alternative way to solve the MDPs with finite state and action spaces, and with bounded rewards. But this is not the full answer for this problem, because we have no guarantee that the bias in Double Q-learning will be lower than Q-learning's, and that underestimating is always better than overestimating.

To reduce the bias of the estimator, Lan et al. [2020] proposed Maxmin Q-learning to control the estimation bias varying from positive to negative.

Many deep reinforcement learning algorithms are proposed these years, which can apply the algorithms to MDPs with infinite state action space. The same author of Double Q-learning proposed Double DQN (Van Hasselt et al. [2016]). Averaged-DQN (Anschel et al. [2017]) is also proposed to tackle this problem by directly reduce the variance of approximation error to soothe the problem of overestimation, but it will require more time when estimating and more memory usage for neuron networks.

There are also some papers are exploring whether can the rewards in MDP be unbounded. Ma et al. [2020] proposed a transformation for Q-learning to tackle this problem.

For the stability of training and to converge faster, decreasing estimation bias and estimation variance is a way to achieve it. Therefore in future research, it is important to explore estimators which can do better approximation.

References

Hado Hasselt. Double q-learning. 2010.

Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. 1994.

Qingfeng Lan, Yangchen Pan, Alona Fyshe, and Martha White. Maxmin q-learning: Controlling the estimation bias of q-learning. 2020.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. 2016.

Oron Anschel, Nir Baram, and Nahum Shimkin. Averaged-dqn: Variance reduction and stabilization for deep reinforcement learning. 2017.

Qingyin Ma, John Stachurski, and Alexis Akira Toda. Unbounded dynamic programming via the q-learning transform. 2020.