A Note on Double Q-Learning

Author Po-Wei Huang
Department of Computer Science
National Yang Ming Chiao Tung University
a0716084.cs07@nycu.edu.tw

1 Introduction

The main research challenges tackled by this paper is Q-learning's poor performance in stochastic MDPs due to large overestimations of the action values. This paper shows that why overestimation in Q-learning occurs and propose another algorithm called Double Q-learning to avoid this. Instead of using single estimator like Q-learning, double Q-learning uses an alternative method, double estimator, to find the maximum expected value. This paper also shows that double Q-learning converges to the optimal value and can handle the overestimation problem, but may have underestimation problem, so the double Q-learning method is not always better than the Q-learning method.

2 Problem Formulation

2.1 Notations

- MDP: $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$
  - $\mathcal{S}$: state space
  - $\mathcal{A}$: action space
  - P: fixed state transition distribution, $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$
  - R: fixed reward distribution, $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathcal{R}$
  - $\gamma \in [0, 1)$: discout factor
- Q-learning:
  - $Q_t(s, a)$: the value of the action a in state s at time t.
  - r_t: the reward drawn from R, where $E[r_t|(s, a, s') = (s_t, a_t, s_{t+1})] = R^{s'}_{sa}$
  - $\alpha_t(s, a) \in [0, 1]$: learning rate
- Estimating the maximum expected value
  - $X = (X_1, \ldots, X_M)$: a set of M random variables
  - $\max_i E[X_i]$: the maximum expected value of the variables
  - $S = \cup_{i=1}^{M} S_i$: a set of iid samples, S_i is the subset containing samples for X_i
  - μ_i: an estimator for X_i
  - f_i: the probability density function of X_i
  - F_i: the cumulative distribution function of f_i, $F_i = \int_{-\infty}^{x} f_i(x)\mathrm{d}x$

It is impossible to exactly determine $\max_i E[X_i]$ for $X = (X_1, \ldots, X_M)$ in most cases. How can we estimate the maximum expected value? This section shows two kinds of estimator: single estimator and double estimator.
The sample average from S_i for each X_i is an unbiased estimator for the expected values:

$$E[X_i] = E[\mu_i] \approx \mu_i(S) \stackrel{def}{=} \frac{1}{|s_i|}\Sigma_{s \in S_i} s \tag{1}$$

The maximum expected value:

$$\max_i E[X_i] = \max_i \int_{-\infty}^{x} x f_i(x) \mathrm{d}x \tag{2}$$

2.2 The Single Estimator

Q-learning approximates the value $\max_i E[X_i] = \max_i E[\mu_i] \approx \max_i \mu_i(S)$ by maximizing over the estimated action values in that state. $\max_i \mu_i$ is distributed according to some PDF $f^{\mu}_{\max}$ that is dependent on the PDFs of the estimators f^{μ}_i. To determine this PDF, consider the CDF $F^{\mu}_{\max}(x)$:

$$F^{\mu}_{\max} \stackrel{def}{=} P(\max_i \mu_i \leq x) = \Pi^M_{i=1} P(\mu_i \leq x) \stackrel{def}{=} \Pi^M_{i=1} F^{\mu}_i(x) \tag{3}$$

$\max_i \mu_i(S)$ is an unbiased estimate for:

$$E[\max_j \mu_j] = \int_{-\infty}^{\infty} x f^{\mu}_{\max}(x) \mathrm{d}x = \int_{-\infty}^{\infty} x \frac{\mathrm{d}}{\mathrm{d}x} \Pi^M_{i=1} F^{\mu}_i(x) \mathrm{d}x = \Sigma^M_j \int_{-\infty}^{\infty} x f^{\mu}_j(s) \Pi^M_{i \neq j} F^{\mu}_i(x) \mathrm{d}x \tag{4}$$

However, it is not an unbiased estimate for $\max_j E[\mu_j]$.

2.3 The Double Estimator

An alternative method to approximate $\max_i E[X_i]$, uses two sets of estimators: $\mu^A = \{\mu^A_1, \ldots, \mu^A_M\}$ and $\mu^B = \{\mu^B_1, \ldots, \mu^B_M\}$.Both sets of estimators are updated with a subset of the samples we draw, $S = S^A \cup S^B, S^A \cap S^B = \phi, \mu^A_i(S) = \frac{1}{|s^A_i|} \Sigma_{s \in S^A_i} s, and \mu^B_i(S) = \frac{1}{|s^B_i|} \Sigma_{s \in S^B_i} s$, assume that the samples are split in a proper manner, both μ^A and μ^B are unbiased. Let the set of maximal estimates in $\mu^A(S) : Max^A(S) \stackrel{def}{=} \{j | \mu^A_j(S) = \max_i \mu^A_i(S)\}$ Since μ^B is an independent, unbiased set of estimators. $E[\mu^B_j] = E[X_j]\ \forall j$, including all $j \in Max^A$. Pick an estimator a^* that maximizes $\mu^A : \mu^A_{a^*}(S) \stackrel{def}{=} \max_i \mu^A_i(S)$, then we can use $\mu^B_{a^*}(S)$ as an estimate for $\max_i E\{\mu^B_i\}$, then obtain the approximation: $\max_i E\{X_i\} = \max_i E\{\mu^B_i\} \approx \mu^B_{a^*}$. Assume that the underlying PDFs are continuous, f^A_j and F^A_j are the PDF and CDF of μ^A_i:

$$P(j = a^*) = \int_{-\infty}^{\infty} P(\mu^A_j = x) \Pi^M_{i \neq j} P(\mu^A_i < x) \mathrm{d}x \stackrel{def}{=} \int_{-infty}^{\infty} f^A_j(x) \Pi^M_{i \neq j} F^A_i(x) \mathrm{d}x \tag{5}$$

The expected value of the double estimator's approximation is:

$$\Sigma^M_j P(j = a^*) E[\mu^B_j] = \Sigma^M_j E[\mu^B_j] \int_{-infty}^{\infty} f^A_j(x) \Pi^M_{ij} F^A_i(x) \mathrm{d}x \tag{6}$$

For discrete PDFs, use sums to replace the integrals.
Comparing the double estimator $\Sigma^M_j E[\mu^B_j] \int_{-\infty}^{\infty} f^A_j(x) \Pi^M_{i \neq j} F^A_i(x) \mathrm{d}x$ to single estimator $\Sigma^M_j \int_{-\infty}^{\infty} x f^{\mu}_j(s) \Pi^M_{i \neq j} F^{\mu}_i(x) \mathrm{d}x$, the double estimator uses $E[\mu^B_j]$ in place of x within integral.

- The single estimator overestimates because x within integral and the CDF product $\Pi^M_{i \neq j} F^{\mu}_i(x)$ is monotonically increasing.
- The double estimator underestimates because $\Sigma^M_j P(j = a^*) = 1$ and the approximation $\Sigma^M_j P(j = a^*) E[\mu^B_j]$ can be viewed as a weighted estimate of unbiased expected values.

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

Proof.

1. $E[\mu^B_{a^*}] \leq \max_i E[X_i]$:

if $a^* \in M$, then $E[X_{a^*}] = \max_i E[X_i]$, otherwise $E[X_{a^*}] < \max_i E[X_i]$
so $E[X_{a^*}] \leq \max_i E[X_i]$

2. $E[\mu^B_{a^*}] < \max_i E[X_i] \Leftrightarrow P(a^* \notin M) > 0$:

$E[\mu^B_{a^*}] = P(a^* \in M)E[X_{a^*}|a^* \in M] + P(a^* \notin M)E[X_{a^*}|a^* \notin M]$
$E[X_{a^*}|a^* \in M] = \max_i E[X_i]$
$(\Rightarrow) E[\mu^B_{a^*}] < \max_i E[X_i] \Rightarrow P(a^* \notin M) < 0$:

if $P(a^* \notin M) = 0$

then

$$\begin{aligned} E[\mu^B_{a^*}] &= P(a^* \in M)E[X_{a^*}|a^* \in M] + P(a^* \notin M)E[X_{a^*}|a^* \notin M] \\ &= 1 \cdot E[X_{a^*}|a^* \in M] + 0 \cdot E[X_{a^*}|a^* \notin M] \\ &= \max_i E[X_i] \\ &\text{contradiction with } E[\mu^B_{a^*}] < \max_i E[X_i], \text{so } P(a^* \notin M) > 0 \end{aligned} \tag{7}$$

$(\Leftarrow) P(a^* \notin M) < 0 \Rightarrow E[\mu^B_{a^*}] < \max_i E[X_i]$:

$$\begin{aligned} \max_i E[X_i] - E[\mu^B_{a^*}] &= E[X_{a^*}|a^* \in M] - (P(a^* \in M)E[X_{a^*}|a^* \in M] + P(a^* \notin M)E[X_{a^*}|a^* \notin M]) \\ &= P(a^* \notin M)(E[X_{a^*}|a^* \in M] - E[X_{a^*}|a^* \notin M]) > 0 \\ &(\because P(a^* \notin M) > 0 \text{ and } E[X_{a^*}|a^* \in M] > E[X_{a^*}|a^* \notin M]) \\ &\text{so } E[\mu^B_{a^*}] < \max_i E[X_i] \end{aligned} \tag{8}$$

$P(a^* \notin \mathcal{M})$ happens when the variables have different expected values but their distributions overlap. The double estimator is unbiased when the variables are iid since then all expected values are equal and $P(a^* \in \mathcal{M}) = 1$.

Double Q-laerning

Q-learning can be interpreted as using the single estimator to estimate the value of the next state:

$$\max_a Q_t(s_{t+1}, a) \approx E[\max_a Q_t(s_{t+1}, a)] \approx \max_a E[Q_t(s_{t+1}, a)] \tag{9}$$

$\max_a Q_t(s_{t+1}, a)$ is an unbiased sample, drawn from an iid distribution with mean $E[\max_a Q_t(s_{t+1}, a)]$. Empirically, Q-learning can suffer from large overestimations.The next section will present Double Q-learning Algorithm to avoid these overestimation issues and will show that it can converge to the optimal policy.

3 Theoretical Analysis

3.1 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 (e.g. random) either UPDATE(A) or UPDATE(B)
  if UPDATE(A) **then**
    Define $a^* = \arg\max_a Q^A(s',a)$
    $Q^A(s,a) \leftarrow Q^A(s,a) + \alpha(s,a)\left(r + \gamma Q^B(s',a^*) - Q^A(s,a)\right)$
  else if UPDATE(B) **then**
    Define $b^* = \arg\max_a Q^B(s',a)$
    $Q^B(s,a) \leftarrow Q^B(s,a) + \alpha(s,a)(r + \gamma Q^A(s',b^*) - Q^B(s,a))$
  end if
  $s \leftarrow s'$
until end

Double Q-learning stores two Q functions: Q^A and Q^B. Each Q function is updated with a value from the other Q function for the next state.a^* is he maximal valued action in state s' according to Q^A. However, instead of using $Q^A(s',a^*)$ to update Q^A like Q-learning, this algorithm uses $Q^B(s',a^*)$. Since Q^A and Q^B have the same update method but with a different set of experience samples,this can be considered an unbiased estimate for the value of this action.
Like the double estimator, double Q-learning may underestimate because action a^* may not be the action that maximizes the expected Q function $\max_a E[Q^A(s',a)]$. In general, $\max_a E[Q^B(s',a^*)] \leq \max_a E[Q^A(s',a^*)]$.

3.2 Convergence in the Limit

Theorem 1 *In the limit Double Q-learning converges to the optimal policy.*

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 (2) 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) $\forall s,a,s' : Var\{R^{s'}_{sa}\} < \infty$.*

Conditions

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. $\alpha_t(s,a) \in [0,1]$, $\Sigma_t \alpha_t(s,a) = \infty$, $\Sigma_t \alpha_t(s,a)^2 < \infty w.p.1$ *and* $\forall(s,a) \neq (s_t,a_t) : \alpha_t(s,a) = 0$
6. $\forall s,a,s' : Var[R^{s'}_{sa}] < \infty$

Proof This paper proves Theorem 1 by making use of lemma 2, which was also used to prove convergence of Sarsa:

Lemma 2

Lemma 2. *Consider a stochastic process (ζ_t, Δ_t, F_t), $t \geq 0$, where $\zeta_t, \Delta_t, F_t : X \to \mathbb{R}$ satisfy the equations:*

$$\Delta_{t+1}(x_t) = (1 - \zeta_t(x_t))\Delta_t(x_t) + \zeta_t(x_t)F_t(x_t) \ , \tag{8}$$

where $x_t \in X$ and $t = 0, 1, 2, \ldots$. Let P_t 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, $t = 1, 2, \ldots$. Assume that the following hold: 1) The set X is finite. 2) $\zeta_t(x_t) \in [0, 1]$, $\sum_t \zeta_t(x_t) = \infty$, $\sum_t (\zeta_t(x_t))^2 < \infty$ w.p.1 and $\forall x \neq x_t : \zeta_t(x) = 0$. 3) $||E\{F_t|P_t\}|| \leq \kappa||\Delta_t|| + c_t$, where $\kappa \in [0, 1)$ and c_t converges to zero w.p. 1. 4) $Var\{F_t(x_t)|P_t\} \leq K(1 + \kappa||\Delta_t||)^2$, where K is some constant. Here $||\cdot||$ denotes a maximum norm. Then Δ_t converges to zero with probability one.

Conditions

1. The set X is finite.
2. $\zeta_t(x_t) \in [0, 1],\ \Sigma_t\zeta_t(x_t) = \infty,\ \Sigma_t\zeta_t(x_t)^2 < \infty\ w.p.1,\ \forall x \neq x_t, \zeta_t(x) = 0$
3. $||E[F_tP_t]|| \leq \kappa||\Delta_t|| + c_t, where\ \kappa \in [0, 1)\ and\ c_t \to 0\ w.p.1$
4. $Var[F_t(x_t)P_t] \leq K(1 + \kappa||\Delta_t||)^2$,where K is some constant.

I will prove that in the algorithm, Q^A will converge to Q^*, the optimal value, and by the symmetry of A and B, Q^B will converge to Q^*, too.
Let $P_t = Q_0^A, Q_0^B, s_0, a_0, r_1, s_1, \ldots, s_t, a_t, X = S \times A, \Delta_t = Q_t^A - Q^*$

$$\begin{aligned} Q_{t+1}^A(s_t, a_t) &= Q_t^A(s_t, a_t) + \alpha(s_t, a_t)(r_t + \gamma Q_t^B(s_{t+1}, a^*) - Q^A(s_t, a_t)) \\ \Rightarrow Q_{t+1}^A(s_t, a_t) &= (1 - \alpha(s_t, a_t))Q_t^A(s_t, a_t) + \alpha(s_t, a_t)(r_t + \gamma Q_t^B(s_{t+1}, a^*)) \end{aligned}$$

subtract $Q^*(s_t, a_t)$ from both sides of the equation, we get:

$$\Delta_{t+1}(s_t, a_t) = (1 - \alpha(s_t, a_t))\Delta_t(s_t, a_t) + \alpha(s_t, a_t)(r_t + \gamma Q_t^B(s_{t+1}, a^*) - Q^*(s_t, a_t))$$

Let $\zeta_t = \alpha_t, F_t(s_t, a_t) = r_t + \gamma Q_t^B(s_{t+1}, a^*) - Q_t^*(s_t, a_t), a^* = argmax_a Q^A(s_{t+1}, a^*), \gamma = \kappa$
Check the four conditions in lemma 2 is satisfied:

1. Satisfied by the condition 1 of Theorem 1.
2. Satisfied by the condition 5 of Theorem 1.
3. Let $F_t^Q(s_t, a_t) = r_t + \gamma Q_t^A(s_{t+1}, a^*) - Q_t^*(s_t, a_t)$, the value of F_t in Q-learning:

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

Let $\Delta^{BA} = Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, a^*)$
Show that $||E[F_t^Q(s_t, a_t)P_t]||_\infty \leq \gamma||\Delta_t||_\infty$ first:
The update for Q^* is: $Q_{t+1}^*(s, a) = (1 - \alpha)Q_t^*(s, a) + \alpha(r + \gamma \cdot \max_a Q_t^*(s', a))$,
$\because Q_{t+1}^* = Q_t^* \ \therefore Q^*(s, a) = r + \gamma \cdot \max_a Q^*(s', a)$

$$\begin{aligned} ||E[F_t^Q(s_t, a_t)P_t]||_\infty &= ||E[r_t + \gamma Q_t^A(s_{t+1}, a^*) - Q^*(s_t, a_t)]||_\infty \\ &= \max_{a_t} |E[r_t + \gamma \max_a Q_t^A(s_{t+1}, a) - Q^*(s_t, a_t)]| \\ &= \max_{a_t} |E[r_t + \gamma \max_{a_t} Q_t^A(s_{t+1}, a_t) - (r_t + \gamma \cdot \max_{a_t} Q^*(s_{t+1}, a_t))]| \\ &= \gamma \max_{a_t} |E[\max_{a_t} Q_t^A(s_{t+1}, a_t) - \max_{a_t} Q^*(s_{t+1}, a_t)]| \\ &\leq \gamma \max_{a_t} |E[Q_t^A(s_{t+1}, a_t) - Q^*(s_{t+1}, a_t)]| \\ &\leq \gamma \max_{a_t} E[|Q_t^A(s_{t+1}, a_t) - Q^*(s_{t+1}, a_t)|](Jensen'sinequality) \\ &\leq \gamma||Q_t^A - Q^*||_\infty = \gamma||\Delta_t||_\infty \end{aligned}$$

Then show that $\Delta^{BA} = Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, a^*)$ converges to 0:
Let $F_t^B(s_t, a_t) = r_t + \gamma Q_t^A(s_{t+1}, b^*) - Q_t^B(s_t, a_t)$,
$F_t^A(s_t, a_t) = r_t + \gamma Q_t^B(s_{t+1}, a^*) - Q_t^A(s_t, a_t)$
For Update A or Update B,

$\Delta^{BA}_{t+1}(s_t, a_t) = \Delta^{BA}_t(s_t, a_t) + \alpha(s_t, a_t)F^B_t(s_t, a_t)$ or
$\Delta^{BA}_{t+1}(s_t, a_t) = \Delta^{BA}_t(s_t, a_t) - \alpha(s_t, a_t)F^A_t(s_t, a_t)$

$$E\Delta^{BA}_{t+1}(s_t, a_t)P_t = \Delta^{BA}_t(s_t, a_t) + E[\alpha(s_t, a_t)(F^B_t(s_t, a_t) - F^A_t(s_t, a_t))|P_t]$$

Let $\zeta^{BA} = 1/2\alpha_t, F^{BA}(s_t, a_t) = \gamma(Q^A_t(s_t, b^*) - Q^B_t(s_t, a^*)) \geq 0$ ($\because$ the symmetry of A and B)

$= \Delta^{BA}_t(s_t, a_t) + \zeta^{BA}(s_t, a_t)E[F^{BA}(s_t, a_t) - \Delta^{BA}_t(s_t, a_t)|P_t]$
$= (1 - \zeta^{BA}(s_t, a_t))\Delta^{BA}_t(s_t, a_t) + \zeta^{BA}(s_t, a_t)E[F^{BA}(s_t, a_t)|P_t]$
$\because E[F^{BA}(s_t, a_t)|P_t] = \gamma EQ^A_t(s_{t+1}, b^*) - Q^B_t(s_{t+1}, a^*)P_t \leq \gamma EQ^A_t(s_{t+1}, a^*) - Q^B_t(s_{t+1}, a^*)P_t \leq \gamma||\Delta^{BA}_t||$
$\therefore$ By lemma 2, $\Delta^{BA}_t \to 0$

Because $F_t(s_t, a_t) = F^Q_t(s_t, a_t) + \gamma\Delta^{BA}_t$,
$||E[F_t(s_t, a_t)P_t]||_\infty = ||E[F^Q_t(s_t, a_t) + \gamma\Delta^{BA}_t]||_\infty \leq \gamma||\Delta_t||_\infty$
This condition is satisfied.

4. From the condition 6 of theorem 1, $Var(R^{s'}_{sa})$ is bounded, so $Var(r_t)$, $Var(Q^B_t)$, $Var(Q^*_t)$ are bounded, too.
$F_t(s_t, a_t) = r_t + \gamma Q^B_t(s_{t+1}, a^*) - Q^*_t(s_t, a_t)$
$\Rightarrow Var(F_t(s_t, a_t)) = Var(r_t) + \gamma Var(Q^B_t) - Var(Q^*_t) < \infty$, this condition is satisfied.
By lemma 2, Q^A will converge toQ^*
Because of the symmetry of A and B, Q^B will converge to Q^*.
By theorem 1, both Q^A and Q^B will converge to the optimal value.

4 Conclusion

Q-learning has overestimation bias, the double Q-learning algorithm in this paper uses a double estimator approach to determine the value of the next state, and it sometimes underestimates the action values, but does not suffer from the overestimation bias.
Since double Q-learning is negatively biased and Q-learning is positively biased, it may be possible to construct an unbiased off-policy reinforcement-learning algorithm in the future.