# A Note on Maxmin Q-Learning

**Yu-Heng Hung**
Department of Computer Science
National Chiao Tung University
j5464654.cs08g@nctu.edu.tw

## 1 Introduction

Q-learning suffers from the overestimation problem and may lead to divergence. There are several methods have been introduced to reduce the overestimation bias like Double Q-learning. But none of them give a control of trading overestimation and underestimation. This paper provide a algorithm called Maxmin Q-learning which is a variant of Q-learning. This algorithm uses a set of action-value functions to control estimation bias and the estimation variance. Moreover this paper give a theoretically proof of how Maxmin Q-learning can lead to unbiased estimation with lower variance and the convergence of this algorithm.

Maxmin Q-learning is a simple variant of Q-learning, it is designed to control the estimation bias and can also reduces the estimation variance of action values. The main idea of Maxmin Q-learning is to create $N$ different action-value functions, and use the minimum of these $N$ action-value functions to be the Q-learning target, for example, if $N = 1$ the update role is the same as Q-learning which is suffered from overestimation problem; As $N$ increase, the overestimation decreases and switches to underestimate for some $N \geq 1$. On the whole, they provide a algorithm which can easily control the estimation bias and is also easily implemented.

## 2 Problem Formulation

The optimization problem is modeled as a Markov decision process (MDP) which can be represented by $(S, A, P, r, \gamma)$, Under the MDP setting, an agent and the environment interact over a sequence of discrete time steps $t$. At every time step $t$, the agent observes a state $S_t \in S$, where $S$ is the state space of this MDP. After taking a actine $A_t$ from the action space $A$, the agent moves to next state $S_{t+1}$ with probability $P(S_{t+1}|S_t, A_t)$ then receive a reward $R_{t+1} = r(S_t, A_t, S_{t+1})$ and $\gamma \in [0, 1]$ is the discount factor to the reward. The goal of the agent is to find a policy $\pi(a|s)$ to maximizes the expected cumulative reward $E[G_t] = E\left[\sum_{k=0}^{T-t-1} \gamma^k R_{t+1+k}\right]$ starting from some initial state $S_0$.

The Q-learning is a off-policy algorithm which want to learn the optimal policy which $\pi^*(a|s) = \arg\max_{a \in A} Q^*(s, a)$ and $Q^*(s, a)$ is the optimal action value function. In Maxmin Q-learning, We assume that the approximation error of each action-value function is $e^i_{sa}$

$$Q^i(s, a) = Q^*(s, a) + e^i_{sa} \tag{1}$$

where $Q^i(s, a)$ is i-th estimator of $Q^*$ and $e_{sa}$ is a uniform random variable follows a uniform distribution $U(-\tau, \tau)$ for some $\tau > 0$. Using $M$ to donate the number of actions applicable at state $s'$, we can define the estimation bias $Z_{MN}$ for transition $s, a, r, s'$ to be

$$Z_{MN} = (r + \gamma \max_{a'} Q^{min}(s', a')) - (r + \gamma \max_{a'} Q^*(s', a')) \tag{2}$$

$$= \gamma(\max_{a'} Q^{min}(s', a')) - \max_{a'} Q^*(s', a')) \tag{3}$$

**Question** : I am still confused on the depiction of **Theorem 2** in Lan et al. [2020], how to prove the convergence when $\gamma = 1$?

Preprint. Under review.

**Question** : Why it is need to assume the estimation bias is follow a uniform distribution. Is there any distribution which can base on the same proof sketch?

## 3 Theoretical Analysis

In this section, I summarize the two main theorems provide by the paper, the first one is to give the relation of $N$ between estimation bias and the estimation variance; The second one is to proof the convergence of Maxmin Q-learning.

**Theorem 1** For the Maxmin Q-Learning, it can control the estimation bias and variance of target action value by $N$ which is the number of action-value functions $\{Q^1, ..., Q^N\}$ used in the Maxmin Q-Learning algorithm. Such that the expected estimation bias is

$$E[Z_{MN}] = \gamma\tau\left(1 - 2\frac{M!\frac{1}{N}!}{(M+\frac{1}{N})!}\right) \tag{4}$$

and the estimation variance of target action value is

$$Var[Q^{min}_{sa}] = \frac{4N\tau^2}{(N+1)^2(N+2)} \tag{5}$$

From (4) and (5), we can easily check that $E[Z_{MN}]$ and $Var[Q^{min}_{sa}]$ both decrease as N increases. Notice that $E[Z_{MN}] = \gamma\tau\frac{M-1}{M+1}$ for $N = 1$ and $E[Z_{MN}] = -\gamma\tau$ for $N = \infty$; $Var[Q^{min}_{sa}] = \frac{\tau^2}{3}$ for $N = 1$ and $Var[Q^{min}_{sa}] = 0$ for $N = \infty$.

**Remark 1** From **Corollary 1** in Lan et al. [2020], $\tau$ will be proportional to some function of $\frac{n_{sa}}{N}$, where $n_{sa}$ is the total samples for updating $Q_{sa}$ because of that $N$ estimators share among $n_{sa}$ samples. Assume that the $n_{sa}$ samples are evenly distributed to $N$ estimators, then $\tau = \sqrt{\frac{3\sigma^2 N}{n_{sa}}}$ where $\sigma^2$ is the variance of a single estimator $Q_{sa}$ that uses all $n_{sa}$ samples for updating. Therefore

$$Var[Q^{min}_{sa}] < Var[Q_{sa}] \text{ ,when } N \geq 8\,. \tag{6}$$

**Question** : Why we can assume that $\tau = \sqrt{\frac{3\sigma^2 N}{n_{sa}}}$ ?

We start from the three lemma provided in this paper. Let $X_1, ..., X_N$ be N i,i,d random variables with probability density function $f(x)$ and cumulative distribution function $F(x)$. Define $\mu = E(X_i)$ and $\sigma^2 = Var[X_i] < \infty$. Denote $X^m = \min\{X_1, ..., X_N\}$, $X^M = \max\{X_1, ..., X_N\}$. The **PDF** and **CDF** of $X^m$ are $f^m(x)$ and $F^m(x)$ respectively. Similarly, **PDF** and **CDF** of $X^M$ are $f^M(x)$ and $F^M(x)$ respectively.

**Lemma 1**

$$f^m(x) = Nf(x)(1 - F(x))^{N-1} \tag{7}$$

$$F^m(x) = 1 - (1 - F(x))^N \tag{8}$$

**Proof** We can start from the **CDF** of $X^m$

$$\begin{aligned}
F^m(x) &= P(X^m \leq x) && (9)\\
&= 1 - P(X^m > x) && (10)\\
&= 1 - P(X_1 > x, ..., X_N > x) && (11)\\
&= 1 - P(X_1 > x)\cdots P(X_N > x) && (12)\\
&= 1 - (1 - F(x))^N && (13)
\end{aligned}$$

and the **PDF** of $X^m$ is $f^m(x) = \frac{dF^m(x)}{dx} = Nf(x)(1 - F(x))^{N-1}$.

**Lemma 2**

$$f^M(x) = Nf(x)(F(x))^{N-1} \tag{14}$$

$$F^M(x) = (F(x))^N \tag{15}$$

**Proof** Similar to the proof of **Lemma 1**.

**Lemma 3** If $X_1, ..., X_N$ are sampled from a uniform distribution called $U(-\tau, \tau)$. The variance of $X^m$ is

$$Var(X^m) = \frac{4N\tau^2}{(N+1)^2(N+2)} \tag{16}$$

**Proof** Because of the uniform distribution $U(-\tau, \tau)$, we have

$$f(x) = \frac{1}{2\tau} \tag{17}$$

$$F(x) = \frac{1}{2} + \frac{x}{2\tau} \tag{18}$$

and then go through the definition of variance

$$\begin{aligned} Var(X^m) &= E((X^m)^2) - E(X^m)^2 & (19)\\ &= \frac{8\tau^2}{(N+1)(N+2)} - \frac{4\tau^2}{(N+1)^2} & (20)\\ &= \frac{4N\tau^2}{(N+1)^2(N+2)} & (21) \end{aligned}$$

We can easily check that $Var(X^m)$ decreases as $N$ increasing and equals to $\sigma^2$ when $N = 1$. **Question** : I try to get the term with color-red by starting with $\int_{-\tau}^{\tau} x^2 \frac{N}{2\tau}(\frac{1}{2} + \frac{x}{2\tau})^{N-1} dx$. But I can't get the result as same as the paper gives.

Now we can start to proof **Theroem 1**. Notice that we use the same notion as above in the following proof. Such that the estimation bias $e_{sa}^i = X_i$.

**Proof** The expectation of $Z_{MN}$ is

$$\begin{aligned} E[Z_{MN}] &= \gamma E[\max_{a'} Q_{s'a'}^{min} - \max_{a'} Q_{s'a'}^{*}] & (22)\\ &= \gamma E[\max_{a'} \min_{i \in \{1,...,N\}} e_{sa}^i] & (23)\\ &= \gamma \int_{-\tau}^{\tau} xMf^M(x)(F^M(x))^{M-1} dx & (24) \end{aligned}$$

where the last equation is based on $Q_{sa}^i = Q_{sa}^* + e_{sa}^i$. And then we plug in the **PDF** and **CDF** of uniform distribution $U(-\tau, \tau)$ into $f^m(x)$ and $F^m(x)$ which have the formula we give in the result of **Lemma** 2. Such that we have

$$E[Z_{MN}] = \gamma \int_{-\tau}^{\tau} x \underbrace{MN\frac{1}{2\tau}\left(\frac{1}{2} - \frac{x}{2\tau}\right)^{N-1}\left[1 - (\frac{1}{2} - \frac{x}{2\tau})^N\right]^{M-1}}_{g(x)} dx \tag{25}$$

And the term $g(x)$ can be writen as $\frac{dh(x)}{dx}$ where

$$h(x) = \left[1 - (\frac{1}{2} - \frac{x}{2\tau})^N\right]^M = (1 - y^N)^M \text{ , } y = \frac{1}{2} - \frac{x}{2\tau} \text{ , } dy = -\frac{1}{2\tau}dx \tag{26}$$

Therefore

$$\begin{aligned} E[Z_{MN}] &= \gamma \int_{-\tau}^{\tau} xg(x)dx & (27)\\ &= \gamma \int_{-\tau}^{\tau} xdh(x) & (28)\\ &= \gamma(\tau h(\tau) + \tau h(-\tau)) - \gamma \int_{-\tau}^{\tau} h(x)dx & (29)\\ &= \gamma\tau - \gamma \int_{-\tau}^{\tau} h(x)dx & (30)\\ &= \gamma\tau - \gamma \int_{0}^{1} (1 - y^N)^M dy & (31) \end{aligned}$$

We can solve $\int_0^1 (1-y^N)^M dy$ by transform it to the form of beta function. Define $t = y^N$ we have

$$\int_0^1 (1-y^N)^M dy = \frac{1}{N}\int_0^1 t^{\frac{1}{N}-1}(1-t)^M dt \tag{32}$$
$$= \frac{1}{N}\frac{\Gamma(M+1)\Gamma(\frac{1}{N})}{\Gamma(M+\frac{1}{N}+1)} \tag{33}$$
$$= \frac{\Gamma(M+1)\Gamma(\frac{1}{N}+1)}{\Gamma(M+\frac{1}{N}+1)} \tag{34}$$
$$= \frac{M!\frac{1}{N}!}{(M+\frac{1}{N})!}\text{ , }\Gamma(n) = (n-1)!\text{ is the gamma function.} \tag{35}$$

The following theorem give the convergence of Maxmin Q Learning. In this paper, they propose a more general result called Generalized Q-learning: Q-learning where the bootstrap target uses a function G of N action values, and then apply Maxmin Q-learning to the proof. I start from checking the update rule of Maxmin Q-learning is a $\gamma$-contraction operator, and then use the similar way of proving the convergence of traditional Q learning.

**Theorem 2** Let $Q_s = (Q_s^1, ..., Q_s^N)$ and $G(Q_s) = \max_{a\in A}\min_{i\in\{1,...,N\}} Q^i(s,a)$ , we have the update operator $H_Q(s,a) = \sum_{s'\in S} P(s,a,s')\left[r(s,a,s') + \gamma G(Q_{s'})\right]$ . Maxmin Q-learning which uses the operator $H$ to update will converge to the optimal action-value function.

**Proof** We can easily proof the convergence by checking the Bellman optimality backup operator $H$ is a $\gamma$-contraction operator. For any $Q_s = (Q_s^1, ..., Q_s^N)$ and $Q'_s = (Q'^1_s, ..., Q'^N_s)$, we have

$$|G(Q_s) - G(Q'_s)| \le |\max_a \min_i Q^i(s,a) - \max_{a'}\min_{i'} Q'^{i'}(s',a')| \le \max_{a,i} |Q^i(s,a) - Q'^i(s,a)| \tag{36}$$

To proof $H$ is a $\gamma$-contraction operator, we write

$$\|H_Q - H_{Q'}\|_\infty = \max_{s,a}\left|\sum_{s'\in S} P(s,a,s')\left[r(s,a,s') + \gamma G(Q_{s'}) - r(s,a,s') - \gamma G(Q'_{s'})\right]\right| \tag{37}$$
$$= \max_s \gamma \left|\sum_{s'\in S} P(s,a,s')\left(G(Q_{s'}) - G(Q'_{s'})\right)\right| \tag{38}$$
$$\le \max_s \gamma \sum_{s'\in S} P(s,a,s')\left|G(Q_{s'}) - G(Q'_{s'})\right| \tag{39}$$
$$\le \max_s \gamma \sum_{s'\in S} P(s,a,s')\max_{a,i}\left|Q^i(s',a) - Q'^i(s',a)\right| \tag{40}$$
$$= \max_{s,a,i} \gamma \sum_{s'\in S} P(s,a,s')\left|Q^i(s',a) - Q'^i(s',a)\right| \tag{41}$$
$$= \gamma\|Q - Q'\|_\infty \tag{42}$$

and then we use the following lemma to finish the proof.

**Lemma 4** The random process $\Delta_t$ taking values in $R^n$ and defined as

$$\Delta_{t+1}(x) = (1-\alpha_t(x))\Delta_t(x) + \alpha_t(x)F_t(x) \tag{43}$$

converges to zero $w.p.1$ under the following assumptions:

- $0 \le \alpha_t \le 1$ , $\sum_t \alpha_t(x) = \infty$ and $\sum_t \alpha_t^2(x) < \infty$ ;
- $\|E\left[F_t(s,a)\right]\|_\infty \le \gamma\|\Delta_t\|_\infty$ with $\gamma \le 1$;
- $Var\left[F_t(s,a)\right] \le C(1+\|\Delta_t\|_\infty^2)$, for $C \ge 0$;

**Proof** Provided by Melo [2001] and Jaakkola et al. [1994].

Define

$$F_t(s,a) = r(s,a,s') + \gamma G(Q_{s'}) - Q^*(s,a) \tag{44}$$

Because $Q^* = H_{Q^*}$ and the reward is bounded, we have

$$\|E\big[F_t(s,a)\big]\|_\infty = \|H_Q(s,a) - H_{Q^*}(s,a)\|_\infty \leq \gamma\|Q - Q'\|_\infty = \gamma\|\Delta_t\|_\infty \tag{45}$$

and

$$Var\big[F_t(s,a)\big] = E\big[F_t(s,a) - (H_Q(s,a) - Q^*(s,a))\big] \tag{46}$$

$$= E\big[r(s,a,s') + \gamma G(Q_{s'}) - H_Q(s,a)\big] \tag{47}$$

$$= Var[r(s,a,s') + \gamma G(Q_{s'}] \tag{48}$$

$$\leq C(1 + \|\Delta_t\|_\infty^2), \text{ for some constant } C. \tag{49}$$

Then, by Lemma 4, $\Delta_t$ converges to zero $w.p.1$, i.e., $Q^i$ converges to $Q^*$ for all $i$.

## 4 Conclusion

In this paper, they assume that the estimation bias follows some uniform distribution, I consider we can extend the proof to another distribution, it is useful for applying the Maxmin Q-learning to the model which generates reward with some noise. On the other hand, the $N$ is a hyperparemeter used in Maxmin Q-learning, and it can control the estimation bias between overestimated and underestimated. I think if we can use the dynamic $N$ to auto control the bias property will make the algorithm converge easily and more stable.

## References

Qingfeng Lan, Yangchen Pan, Alona Fyshe, and Martha White. Maxmin q-learning: Controlling the estimation bias of q-learning. *arXiv preprint arXiv:2002.06487*, 2020.

Francisco S Melo. Convergence of q-learning: A simple proof. *Institute Of Systems and Robotics, Tech. Rep*, pages 1–4, 2001.

Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems*, pages 703–710, 1994.