A Note on Maxmin Q-Learning

Shang-Hsuan Yang
Department of Computer Science
National Chiao Tung University
sandy861003.eo05@nctu.edu.tw

1 Introduction

This paper proposed a new variant of Q-Learning, called *Maxmin Q-Learning*, in order to address the overestimation bias issues in Q-Learning. The key idea is to maintain N estimates of the action values, and use the minimum of these estimates in the Q-learning target. [1]

The overestimation bias may not always be detrimental. In some cases, overestimation bias can help encourage exploration for overestimated actions, which might be beneficial in highly stochastic areas if they correspond to high-value regions. On the other hand, if highly stochastic areas also have low values, overestimation bias might cause an agent to over-explore, and an underestimation bias may help prevent the situation.

Therefore, instead of fully moving towards underestimation, the method proposed in this paper tries to figure out how strongly we should correct for overestimation bias, and how to determine—or control—the level of bias .

The main contributions of this paper are: 1) It provides an easily-implemented method to determine overestimation or underestimation by simply tuning the parameters (i.e. the number of action-value estimates). 2) It introduces a new Generalized Q-learning framework to prove the convergence of *Maxmin Q-learning* as well as other Q-learning methods that use same number of action-value estimates as Maxmin Q-learning does.

Overall, the paper is interesting and the idea is novel. The theoretical analyses seem to be solid as well. And many other methods are listed to compare with Maxmin Q-learning, which makes the insight of this paper much clearer. I'm looking forward to a future work based on this paper which doesn't rely on parameter-tuning technique.

2 Problem Formulation

The paper formalizes the problem as a *Markov Decision Process (MDP)*, $(\mathcal{S}, \mathcal{A}, \mathbf{P}, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathrm{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probabilities, $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward mapping, and $\gamma \in [0, 1]$ is the discount factor. At each time step t, the agent observes a state $S_t \in \mathcal{S}$ and takes an action $A_t \in \mathcal{A}$ and then transitions to a new state $S_{t+1} \in \mathcal{S}$ according to the transition probabilities $\mathbf{P}$ and receives a scalar reward $R_{t+1} = r(S_t, A_t, S_{t+1}) \in \mathbb{R}$. The goal of the agent is to find a policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ that maximizes the expected return starting from some initial state.

Q-learning is an off-policy algorithm which attempts to learn the state-action values $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for the optimal policy. It tries to solve for

$$Q^*(s, a) = \mathbb{E}\left[R_{t+1} + \max_{a' \in \mathcal{A}} Q^*(S_{t+1}, a') \mid S_t = s, A_t = a\right]$$

Preprint. Under review.

The optimal policy is to act greedily with respect to these action values: from each s select a from

$$\arg\max_{a\in\mathcal{A}} Q^*(s,a)$$

The update rule for an approximation Q for a sampled transition $(s_t, a_t, r_{t+1}, s_{t+1})$ is:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left(Y_t^Q - Q(s_t,a_t)\right) \quad \text{for } Y_t^Q \stackrel{\text{def}}{=} r_{t+1} + \gamma \max_{a'\in\mathcal{A}} Q(s_{t+1},a')$$

where α is the step-size.

The overestimations result from a positive bias that is introduced because Q-learning uses the maximum action value $\max_{a'\in\mathcal{A}} Q(s_{t+1},a')$ as an approximation for the maximum expected action value [2]. For compactness, we write Q^i_{sa} instead of $Q^i(s,a)$. Each Q^i_{sa} has random approximation error e^i_{sa}.

$$Q^i_{sa} = Q^*_{sa} + e^i_{sa}$$

Assume that e^i_{sa} is a uniform random variable $U(-\tau,\tau)$ for some $\tau > 0$. The uniform random assumption was used by [3]. If e^i_{sa} is a constant instead of a ramdom variable, then it has no influence on the optimal policy. Moreover, due to stochasticity, it must be a random variable. Thus, the uniform assumption is required. However, the tuning of τ is still remained to be concerned.

3 Theoretical Analysis

Maxmin Q-learning keeps N estimates of the action values, Q^i, and use the minimum of these estimates to update Q-learning target: $\max_{a'} \min_{i\in\{1,\dots,N\}} Q^i(s',a')$. The goal is to control overestimation by tuning N. It wants the overestimation decreases when N increases, and for some $N > 1$, this estimator switches from an overestimate, in expectation, to an underestimate. For $N = 1$, the update is simply Q-learning, and so likely has overestimation bias.

The full algorithm is summarized in Algorithm 1.

Algorithm 1: Maxmin Q-learning

Input : step size α, exploration parameter $\epsilon > 0$, number of action-value functions N
Initialize N action-value functions $\{Q^1,\dots,Q^N\}$ randomly
Initialize empty replay buffer D
Observe initial state s
while *Agent is interacting with the Environment* **do**
  $Q^{min}(s,a) \leftarrow \min_{k\in\{1,\dots,N\}} Q^k(s,a), \forall a \in \mathcal{A}$
  Choose action a by ϵ-greedy based on Q^{min}
  Take action a, observe r, s'
  Store transition (s,a,r,s') in D
  Select a subset S from $\{1,\dots,N\}$ (e.e., randomly select on i to update)
  for $i \in S$ **do**
    Sample random mini-batch of transitions (s_D, a_D, r_D, s'_D) from D
    Get update target: $Y^{MQ} \leftarrow r_D + \gamma \max_{a'\in A} Q^{min}(s'_D,a')$
    Update action-value Q^i : $Q^i(s_D,a_D) \leftarrow Q^i(s_D,a_D) + \alpha\left[Y^{MQ} - Q^i(s_D,a_D)\right]$
  end
  $s \leftarrow s'$
end

Let M denotes the number of applicable actions at state s'. Then the estimation bias Z_{MN} for transition s,a,r,s' is defined as:

$$\begin{aligned} Z_{MN} &\stackrel{\text{def}}{=} \left(r + \gamma \max_{a'} Q^{\min}_{s'a'}\right) - \left(r + \gamma \max_{a'} Q^*_{s'a'}\right) \\ &= \gamma\left(\max_{a'} Q^{\min}_{s'a'} - \max_{a'} Q^*_{s'a'}\right) \end{aligned}$$

where

$$Q^{\min}_{sa} \stackrel{\text{def}}{=} \min_{i\in\{1,\dots,N\}} Q^i_{sa} = Q^*_{sa} + \min_{i\in\{1,\dots,N\}} e^i_{sa}$$

3.1 The Expected Estimation Bias And The Variance of Q

First of all, in Theorem 1, the paper wants to show that by controlling N in Maxmin Q-learning, the expectation of estimation bias $E[Z_{MN}]$ and the variance of $Q_{sa}^{\min}$ will be affected.

Theorem 1. *Under the conditions stated above,*

(i) *the expected estimation bias is*

$$E[Z_{MN}] = \gamma\tau[1 - 2t_{MN}] \quad \textit{where } t_{MN} = \frac{M(M-1)\cdots 1}{\left(M+\frac{1}{N}\right)\left(M-1+\frac{1}{N}\right)\cdots\left(1+\frac{1}{N}\right)}$$

$E[Z_{MN}]$ *decreases as* N *increases:* $E[Z_{M,N=1}] = \gamma\tau\frac{M-1}{M+1}$ *and* $E[Z_{M,N\to\infty}] = -\gamma\tau$.

(ii)

$$\mathrm{Var}\left[Q_{sa}^{\min}\right] = \frac{4N\tau^2}{(N+1)^2(N+2)}$$

$\mathrm{Var}\left[Q_{sa}^{\min}\right]$ *decreases as* N *increases:* $\mathrm{Var}\left[Q_{sa}^{\min}\right] = \frac{\tau^2}{3}$ *for* $N = 1$ *and* $\mathrm{Var}\left[Q_{sa}^{\min}\right] = 0$ *for* $N \to \infty$.

To prove it, the paper leverages the first lemma in [3].

Lemma 1. *Let* $X_1, \ldots, X_N$ *be* N *i.i.d. random variables from an absolutely continuous distribution with probability density function(PDF)* $f(x)$ *and cumulative distribution function (CDF)* $F(x)$ *Denote* $\mu \overset{def}{=} E[X_i]$ *and* $\sigma^2 \overset{def}{=} \mathrm{Var}[X_i] < +\infty$. *Set* $X_{1:N} \overset{def}{=} \min_{i\in\{1,\ldots,N\}} X_i$ *and* $X_{N:N} \overset{def}{=} max_{i\in\{1,\ldots,N\}} X_i$. *Denote the PDF and CDF of* $X_{1:N}$ *as* $f_{1:N}(x)$ *and* $F_{1:N}(x)$, *respectively. Similarly, denote the PDF and CDF of* $X_{N:N}$ *as* $f_{N:N}(x)$ *and* $F_{N:N}(x)$, *respectively. We then have*

i. $\mu - \frac{(N-1)\sigma}{\sqrt{2n-1}} \le E[X_{1:N}] \le \mu$ *and* $E[X_{1:N+1}] \le E[X_{1:N}]$

ii. $F_{1:N}(x) = 1 - (1 - F(x))^N . f_{1:N}(x) = Nf(x)(1 - F(x))^{N-1}$

iii. $F_{N:N}(x) = (F(x))^N . f_{N:N}(x) = Nf(x)(F(x))^{N-1}$

iv. If $X_1, \ldots, X_N \sim U(-\tau, \tau)$, *we have* $\mathrm{Var}(X_{1:N}) = \frac{4N\tau^2}{(N+1)^2(N+2)}$ *and* $\mathrm{Var}(X_{1:N+1}) < \mathrm{Var}(X_{1:N}) \le \mathrm{Var}(X_{1:1}) = \sigma^2$ *for any positive integer* N

Next, Theorem 1 is proved.

Proof. Let $f(x)$ and $F(x)$ be the cdf and pdf of e_{sa}, respectively. Similarly, Let $f_N(x)$ and $F_N(x)$ be the cdf and pdf of $\min_{i\in\{1,\ldots,N\}} e_{sa}^i$. since e_{sa} is sampled from Uniform$(-\tau, \tau)$, it is easy to get $f(x) = \frac{1}{2\tau}$ and $F(x) = \frac{1}{2} + \frac{x}{2\tau}$. By Lemma 1, we have $f_N(x) = Nf(x)[1 - F(x)]^{N-1} =$

$\frac{N}{2\tau}\left(\frac{1}{2}-\frac{x}{2\tau}\right)^{N-1}$ and $F_N(x) = 1-(1-F(x))^N = 1-\left(\frac{1}{2}-\frac{x}{2\tau}\right)^N$. The expectation of Z_{MN} is

$$\begin{aligned}
E\left[Z_{MN}\right] &= \gamma E\left[\left(\max_{a'} Q^{\min}_{s'a'} - \max_{a'} Q^{*}_{s'a'}\right)\right] \\
&= \gamma E\left[\max_{a'} \min_{i\in\{1,\ldots,N\}} e^{i}_{sa'}\right] \\
&= \gamma \int_{-\tau}^{\tau} Mx f_N(x) F_N(x)^{M-1} dx \\
&= \gamma \int_{-\tau}^{\tau} MN \frac{x}{2\tau}\left(\frac{1}{2}-\frac{x}{2\tau}\right)^{N-1}\left[1-\left(\frac{1}{2}-\frac{x}{2\tau}\right)^N\right]^{M-1} dx \\
&= \gamma \int_{-\tau}^{\tau} x d\left[1-\left(\frac{1}{2}-\frac{x}{2\tau}\right)^N\right]^M \\
&= \gamma\tau - \gamma\int_{-\tau}^{\tau}\left[1-\left(\frac{1}{2}-\frac{x}{2\tau}\right)^N\right]^M dx \\
&= \gamma\tau\left[1-2\int_0^1 \left(1-y^N\right)^M dy\right] \quad \left(y \overset{\text{def}}{=} \frac{1}{2}-\frac{x}{2\tau}\right)
\end{aligned}$$

Let $t_{MN} = \int_0^1 \left(1-y^N\right)^M dy$, so that $E\left[Z_{MN}\right] = \gamma\tau\left[1-2t_{MN}\right]$. Substitute y by t where $t = y^N$ then

$$\begin{aligned}
t_{MN} &= \frac{1}{N}\int_0^1 t^{\frac{1}{N}-1}(1-t)^M dt \\
&= \frac{1}{N} B\left(\frac{1}{N}, M+1\right) \\
&= \frac{1}{N}\frac{\Gamma(M+1)\Gamma\left(\frac{1}{N}\right)}{\Gamma\left(M+\frac{1}{N}+1\right)} \\
&= \frac{\Gamma(M+1)\Gamma\left(1+\frac{1}{N}\right)}{\Gamma\left(M+\frac{1}{N}+1\right)} \\
&= \frac{M(M-1)\cdots 1}{\left(M+\frac{1}{N}\right)\left(M-1+\frac{1}{N}\right)\cdots\left(1+\frac{1}{N}\right)}
\end{aligned}$$

Each term in the denominator decreases as N increases, because $1/N$ gets smaller. Therefore, $t_{M,N=1} = \frac{1}{M+1}$ and $t_{M,N\to\infty} = 1$. Using this, we conclude that $E\left[Z_{MN}\right]$ decreases as N increases and $E\left[Z_{M,N=1}\right] = \gamma\tau\frac{M-1}{M+1}$ and $E\left[Z_{M,N\to\infty}\right] = -\gamma\tau$.

By Lemma 1, the variance of $Q^{\min}_{sa}$ is

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

$\mathrm{Var}\left[Q^{\min}_{sa}\right]$ decreases as N increases. In particular, $\mathrm{Var}\left[Q^{\min}_{sa}\right] = \frac{\tau^2}{3}$ for $N=1$ and $\mathrm{Var}\left[Q^{\min}_{sa}\right] = 0$ for $N\to\infty$ □

Corollary 1. *Assuming the n_{sa} samples are evenly allocated amongst the N estimators, then $\tau = \sqrt{3\sigma^2 N/n_{sa}}$ where σ^2 is the variance of samples for (s,a) and, for Q_{sa} the estimator that uses all n_{sa} samples for a single estimate,*

$$\mathrm{Var}\left[Q^{\min}_{sa}\right] = \frac{12N^2}{(N+1)^2(N+2)}\mathrm{Var}\left[Q_{sa}\right]$$

Under this uniform random noise assumption, for $N \geq 8$, $\mathrm{Var}\left[Q^{\min}_{sa}\right] < \mathrm{Var}\left[Q_{sa}\right]$

Proof. Because Q^i_{sa} is a sample mean, its variance is $\sigma^2 N/n_{sa}$ where σ^2 is the variance of samples for (s,a) and its mean is Q^*_{sa} (because it is an unbiased sample average). Consequently, e_{sa} has mean zero and variance $\sigma^2 N/n_{sa}$. Because e_{sa} is a uniform random variable which has variance $\frac{1}{3}\tau^2$, we know that $\tau = \sqrt{3\sigma^2 N/n_{sa}}$. Plugging this value into the variance formula in Theorem 1 we get that

$$\begin{aligned}\text{Var}\left[Q^{\min}_{sa}\right] &= \frac{4N\tau^2}{(N+1)^2(N+2)} \\ &= \frac{12N^2\sigma^2/n_{sa}}{(N+1)^2(N+2)} \\ &= \frac{12N^2}{(N+1)^2(N+2)}\,\text{Var}\left[Q_{sa}\right]\end{aligned}$$

because $\text{Var}\left[Q_{sa}\right] = \sigma^2/n_{sa}$ for the sample average Q_{sa} that uses all the samples for one estimator. Easy to verify that for $N \geq 8, \text{Var}\left[Q^{\min}_{sa}\right] < \text{Var}\left[Q_{sa}\right]$ □

3.2 Convergence Analysis of Maxmin Q-Learning

Secondly, the convergence of Maxmin Q-Learning in the tabular setting will be shown. Here, the paper provides *Generalized Q-learning*: Q-learning where the bootstrap target uses a function G of N action values. The pros to use this framework to proof the convergence is that it can be used in many other variants of Q-learning. But the cons is that choosing G function may not be easy. G needs to be limited so that it maintains relative maximum values, as stated in Assumption 1.

Assumption 1. ***(Conditions on** G**)** *Let* $G : \mathbb{R}^{nNK} \mapsto \mathbb{R}$ *and* $G(Q) = q$ *where* $Q = \left(Q^{ij}_a\right) \in \mathbb{R}^{nNK}, a \in \mathcal{A}$ *and* $|\mathcal{A}| = n, i \in \{1,\ldots,N\}, j \in \{0,\ldots,K-1\}$ *and* $q \in \mathbb{R}$.

i. If $Q^{ij}_a = Q^{kl}_a, \forall i,k,\forall j,l,$ *and* $\forall a$, *then* $q = \max_a Q^{ij}_a$

ii. $\forall Q, Q' \in \mathbb{R}^{nNK}, |G(Q) - G\left(Q'\right)| \leq \max_{a,i,j} \left|Q^{ij}_a - Q'^{ij}_a\right|$

This assumption guarantees that even though information can be outdated, any old information is eventually discarded.

Proof. We can verify that Assumption 1 holds for Maxmin Q-learning. Set $K = 1$ and set N to be a positive integer. Let $Q_s = \left(Q^1_s,\ldots,Q^N_s\right)$ and define $G^{MQ}\left(Q_s\right) = \max_{a\in A} \min_{i\in\{1,\ldots,N\}} Q^i_{sa}$. It is easy to check that part i. of Assumption 1 is satisfied. Part ii. is also satisfied because

$$\left|G\left(Q_s\right) - G\left(Q'_s\right)\right| \leq \left|\max_a \min_i Q^i_{sa} - \max_{a'} \min_{i'} Q'^{i'}_{sa'}\right| \leq \max_{a,i} \left|Q^i_{sa} - Q'^i_{sa}\right|$$

□

The target action-value of Generalized Q-learning Y^{GQ} is defined based on action-value estimates from both dimensions:

$$Y^{GQ} = r + \gamma Q^{GQ}_{s'}(t-1)$$

where t is the current time step and the action-value function $Q^{GQ}_s(t)$ is a function of $Q^1_s(t-K),\ldots,Q^1_s(t-1),\ldots,Q^N_s(t-K),\ldots,Q^N_s(t-1)$:

$$Q^{GQ}_s(t) = G\begin{pmatrix} Q^1_s(t-K) & \ldots & Q^1_s(t-1) \\ Q^2_s(t-K) & \ldots & Q^2_s(t-1) \\ \vdots & \ddots & \vdots \\ Q^N_s(t-K) & \ldots & Q^N_s(t-1) \end{pmatrix}$$

For simplicity, the vector $\left(Q^{GQ}_{sa}(t)\right)_{a\in\mathcal{A}}$ is denoted as $Q^{GQ}_s(t)$, same for $Q^i_s(t)$. The corresponding update rule is

$$Q^i_{sa}(t) \leftarrow Q^i_{sa}(t-1) + \alpha^i_{sa}(t-1)\left(Y^{GQ} - Q^i_{sa}(t-1)\right)$$

Assumption 2. ***(Conditions on the step-sizes)*** *There exists some (deterministic) constant C such that for every $(s,a) \in \mathcal{S} \times \mathcal{A}, i \in \{1,\ldots,N\}, 0 \leq \alpha_{sa}^i(t) \leq 1$, and with probability 1,*

$$\sum_{t=0}^{\infty} \left(\alpha_{sa}^i(t)\right)^2 \leq C, \quad \sum_{t=0}^{\infty} \alpha_{sa}^i(t) = \infty$$

This assumption is needed on the step-sizes employed by the Generalized Q-learning algorithm which is standard for stochastic approximation algorithms. In particular, it requires that every state-action pair (s,a) is simulated an infinite number of times.

Theorem 2. *Assume a finite MDP $(\mathcal{S}, \mathcal{A}, \mathrm{P}, R)$ and that Assumption 1 and 2 hold. Then the actionvalue functions in Generalized Q-learning will converge to the optimal action-value function with probability 1, in either of the following cases: (i) $\gamma < 1$ or $(ii)\gamma = 1, \forall a \in \mathcal{A}, Q_{s_1 a}^i(t=0) = 0$ where s_1 is an absorbing state and all policies are proper.*

Proof. To proof Theorem 2, some assumptions and theorems proposed in [4] are required.

Assumption 3. *For any i and $j, \lim_{t\to\infty} \tau_j^i(t) = \infty$, with probability 1 where each $\tau_j^i(t)$ is an integer satisfying $0 \leq \tau_j^i(t) \leq t$.*
This assumption of parameter is involved in the following assumptions.

Assumption 4. *Let $\{\mathcal{F}(t)\}_{t=0}^{\infty}$ be an increasing sequence of subfields of $\mathcal{F}$*

i. $x(0)$ is $\mathcal{F}(0)$ -measurable.

ii. For every i and $t, w_i(t)$ is $\mathcal{F}(t+1)$ -measurable.

iii. For every i, j and $t, \alpha_i(t)$ and $\tau_j^i(t)$ are $\mathcal{F}(t)$ -measurable.

iv. For every i and t, we have $E\left[w_i(t) \mid \mathcal{F}(t)\right] = 0$

v. There exist (deterministic) constants A and B such that

$$E\left[w_i^2(t) \mid \mathcal{F}(t)\right] \leq A + B \max_j \max_{\tau \leq t} |x_j(\tau)|^2, \quad \forall i, t$$

The measure-theoretic terminology that "a random variable Z is $F(t)$-measurable" intuitively means that Z is completely determined by the history represented by $F(t)$.

Assumption 5. *There exists a vector $x^* \in \mathbb{R}^n$, a positive vector v, and a scalar $\beta \in [0,1)$, such that*

$$\|F(x) - x^*\|_v \leq \beta \|x - x^*\|_v, \quad \forall x \in \mathbb{R}^n$$

Theorem 3. *Let Assumptions 3, 4, 5, 2 hold. Then $x(t)$ converges to x^* with probability 1 Detailed proofs can be found in the work of [4].*

If Assumptions 3, 4, 5, 2 are all satisfied, then convergence is gauranteed according to Theorem 3. □

4 Conclusion

It is inspiring that this paper not only proposed a novel method to control overestimation bias, it also provides an elegant framework to help prove convergence with theoretically support. However, the tuning of parameters requires lots of efforts. Therefore, the future work should investigate how best to select N.

References

[1] Qingfeng Lan, Yangchen Pan, Alona Fyshe, and Martha White. Maxmin q-learning: Controlling the estimation bias of q-learning. 02 2020.

[2] Hado Van Hasselt. Double q-learning. pages 2613–2621, 01 2010.

[3] Sebastian Thrun and Anton Schwartz. Issues in using function approximation for reinforcement learning. 10 1993.

[4] Dimitri Bertsekas and John Tsitsiklis. Parallel and distributed computation : numerical methods / dimitri p. bertsekas, john n. tsitsiklis. *SERBIULA (sistema Librum 2.0)*, 23, 01 1989.