Is Q-learning Provably Efficient?

Ding-Neng Liang
Department of Computer Science
National Chiao Tung University
asdsdlgl12@gmail.com

1 Introduction

It is believed that model-free algorithms suffer from a higher sample complexity compared to model-based approaches. This has been evidenced empirically in the previous research. However, the basic theoretical questions still remain that " Can we design model-free algorithms that are sample efficient? " in particular, " Is Q-learning provably efficient? ". There are no absolute answers to this problems, and the answers still remain elusive. In this paper, they try to prove or find a new method that make the model-free algorithms sample complexity as efficient as model-based approaches.

Generally, the key to achieving good sample efficiency generally lies in managing the trade-off between exploration and exploitation. In this paper, they answer the two aforementioned questions affirmatively. They show that Q-learning, which is the classical model-free paradigm, when equipped with a upper confidence bounds (UCB) exploration policy that incorporates estimates of the confidence of Q values and assign exploration bonuses, and the result achieves the total regret $\widetilde{O}(\sqrt{H^4SAT})$, where S, A, H and T represent the numbers of states, the numbers of actions, the number of steps per episode and the total number of steps respectively. The regret result of Q-learning with UCB catches up the regret result of model-based algorithms. Since their algorithm is the Q-learning method, it does not store additional data besides the table of Q values, and at the same time, it also have the advantage of model-based algorithms in terms of time and space complexities.

One previous research tried the method that the standard Q-learning heuristic of incorporating $\epsilon - greedy$ exploration and the regret result appeared to take exponentially many episodes to learn, the others like the only existing theoretical result on model-free RL that applies to the episodic setting is for delayed Q-learning, and this algorithm is quite sample-inefficient compared to model-based approaches. Most of prior works of model-free algorithm are still very limited compared to model-based approaches. So, the method,Q-learning with UCB algorithm, provided from the paper is truly close to model-based approaches although it has a relatively higher$\sqrt{H}$ compared to model-based approaches, but the main factor$\sqrt{T}$ is catching up the model-based approaches.

From my point of view, this method is incredible for me. It successfully combines the model-free approaches and UCB exploration policy. Moreover, the assumption for bonuses and the choice of α that i will mention later are just right. Under all the perfect assumptions they made, it successfully create the state-of-the-art methods and brand-new thinking to show the potential efficiency of model-free approaches.

2 Problem Formulation

First of all, I will introduce the following equations provided from this paper, and slightly explain what the notations are, and the detail analysis will explain in the next section.

Notation. We denote by (x_h^k, a_h^k) the actual state-action pair observed and chosen at step h of episode k. We also denote by Q_h^k, V_h^k, N_h^k respectively the Q_h, V_h, N_h functions at the beginning of

Preprint. Under review.

	Algorithm	Regret	Time	Space
Model-based	UCRL2 [10] [1]	at least $\tilde{\mathcal{O}}(\sqrt{H^4S^2AT})$	$\Omega(TS^2A)$	$\mathcal{O}(S^2AH)$
	Agrawal and Jia [1] [1]	at least $\tilde{\mathcal{O}}(\sqrt{H^3S^2AT})$		
	UCBVI [5] [2]	$\tilde{\mathcal{O}}(\sqrt{H^2SAT})$	$\tilde{\mathcal{O}}(TS^2A)$	
	vUCQ [12] [2]	$\tilde{\mathcal{O}}(\sqrt{H^2SAT})$		
Model-free	Q-learning (ε-greedy) [14] (if 0 initialized)	$\Omega(\min\{T, A^{H/2}\})$	$\mathcal{O}(T)$	$\mathcal{O}(SAH)$
	Delayed Q-learning [25] [3]	$\tilde{\mathcal{O}}_{S,A,H}(T^{4/5})$		
	Q-learning (UCB-H)	$\tilde{\mathcal{O}}(\sqrt{H^4SAT})$		
	Q-learning (UCB-B)	$\tilde{\mathcal{O}}(\sqrt{H^3SAT})$		
	lower bound	$\Omega(\sqrt{H^2SAT})$	-	-

Table 1: Regret comparisons for RL algorithms on episodic MDP. $T = KH$ is totally number of steps, H is the number of steps per episode, S is the number of states, and A is the number of actions. For clarity, this table is presented for $T \geq \text{poly}(S, A, H)$, omitting low order terms.

episode k. Using this notation, the update equation at episode k can be rewritten as follows, for every $h \in [H]$:

$$Q_h^{k+1}(x,a) = \begin{cases} (1-\alpha_t)Q_h^k(x,a) + \alpha_t[r_h(x,a) + V_{h+1}^k(x_{h+1}^k) + b_t] & \text{if (x, a) = } (x_h^k, a_h^k) \\ Q_h^k(x,a) & \text{otherwise} \end{cases} \tag{1}$$

The equation (1) implement the line 7 in Algorithm1, and it's the key thought through the full method. In this equation, t is the counter for how many times the algorithm has visited the state-action pair (x, a) at step h, α_t is the learning rate, r_h is the reward at step h and b_t is the confidence bonus indicating how certain the algorithm is about current state-action pair.
Next, this paper makes the assumptions that define the learning rate as follows:

$$\alpha_t := \frac{H+1}{H+t} \tag{2}$$

where t is the counter for how many times the algorithm has visited the state-action and H is the number of steps per episode. For notational convenience, we can introduce the following related quantities:

$$\alpha_t^0 := \prod_{j=1}^{t}(1-\alpha_j), \quad \alpha_t^i := \alpha_i \prod_{j=i+1}^{t}(1-\alpha_j) \tag{3}$$

From equation(3), we can easily to get the two result one is $\sum_{j=1}^{t} \alpha_j = 1$ and $\alpha_t^0 = 1$ when $t \geqslant 1$, another is $\alpha_t^0 = 0$ if t = 0. According to equation(1) and (3), we can revise the Q-function and the result is:

$$Q_h^k(x,a) = \alpha_t^0 H + \sum_{j=1}^{t} \alpha_j^i [r_h(x,a) + V_{h+1}^k(x_{h+1}^k) + b_t] \tag{4}$$

we can make assumption that t is the numbers of (x, a) we observed in every episode, According to equation (4), the Q value at episode k equals a weighted average of the V values of the next state. So we can draw the weighted graph according to the choice of different learning rate function, and we can set H = 10, and now, we can found something result from the figure. If we choose 1/t, the step length is uniform distribution, if we choose $1/\sqrt{t}$, and the result is that 15% sample data have the

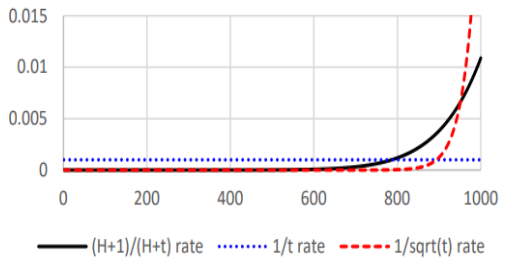

Figure 1: Illustration of $\{\alpha^i_{1000}\}_{i=1}^{1000}$ for learning rates $\alpha_t = \frac{H+1}{H+t}, \frac{1}{t}$ and $\frac{1}{\sqrt{t}}$ when $H = 10$.

Algorithm 1 Q-learning with UCB-Hoeffding

initialize $Q_h(x,a) \leftarrow H$ and $N_h(x,a) \leftarrow 0$ for all $(x,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.
for episode $k = 1, \ldots, K$ **do**
  receive x_1.
  for step $h = 1, \ldots, H$ **do**
    Take action $a_h \leftarrow \text{argmax}_{a'} Q_h(x_h, a')$, and observe x_{h+1}.
    $t = N_h(x_h, a_h) \leftarrow N_h(x_h, a_h) + 1$; $b_t \leftarrow c\sqrt{H^3\iota/t}$.
    $Q_h(x_h, a_h) \leftarrow (1-\alpha_t)Q_h(x_h, a_h) + \alpha_t[r_h(x_h, a_h) + V_{h+1}(x_{h+1}) + b_t]$.
    $V_h(x_h) \leftarrow \min\{H, \max_{a' \in \mathcal{A}} Q_h(x_h, a')\}$.

most of the weights, it will cause the higher variance and indirectly lead to much higher regret. And the choice of α from this paper look like just being moderate.

Next assumption is the confidence bonus b_t, in this paper, it assumes $b_t = O(\sqrt{H^3\iota/t})$, where $\iota := log(SAT/p)$. The reason why make this assumption is that one is Q-values are upper-bounded by H, another is Hoeffding-type martingale concentration inequalities imply that if we have visited (x, a) for t times, then a confidence bound for the Q value.

Now, there something we are interesting in, under the assumption we made,what the total regret of Q-learning with UCB-Hoeffding is if we choose $b_t = c\sqrt{H^3\iota/t}$ with the probability 1-p and c > 0.

3 Theoretical Analysis

First of all, we can show the properties of α_t^i, and i will not show the prove of lemma In general, we can use lemma(3.2) to prove lemma(3.3), and lemma(3.1) and lemma(3.3) could use to prove the complexity.

Lemma 3.1.
(a) $\frac{1}{\sqrt{t}} \leq \sum_{i=1}^t \frac{\alpha_t^i}{\sqrt{i}} \leq \frac{2}{\sqrt{t}}$ *for every* $t \geq 1$
(b) $\max_{i\in[t]} \alpha_t^i \leq \frac{2H}{t}$ *and* $\sum_{i=1}^t \left(\alpha_t^i\right)^2 \leq \frac{2H}{t}$ *for every* $t \geq 1$
(c) $\sum_{t=i}^{\infty} \alpha_t^i = 1 + \frac{1}{H}$ *for every* $i \geq 1$

Lemma 3.2. *(recursion on Q). For any* $(x,a,h) \in S \times A \times [H]$ *and episode* $k \in [K]$, *let* $t = N_h^k(x,a)$ *and suppose* (x,a) *was previously taken at step* h *of episodes* $k_1, ..., k_t < k$. *Then :*

$$\left(Q_h^k - Q_h^\star\right)(x,a) =$$
$$\alpha_t^0 \left(H - Q_h^\star(x,a)\right) + \sum_{i=1}^t \alpha_t^i \left[\left(V_{h+1}^{k_i} - V_{h+1}^\star\right)\left(x_{h+1}^{k_i}\right) + \left[\left(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h\right) V_{h+1}^\star\right](x,a) + b_i\right]$$

Lemma 3.3. *(bound on* $Q^k - Q^*$*). There exists an absolute constant* $c > 0$ *such that, for any* $p \in (0,1)$, *letting* $b_t = c\sqrt{H^3\iota/t}$, *we have* $\beta_t = 2\sum_{i=1}^t (\alpha_t^i b_i) \leq 4c\sqrt{H^3\iota/t}$ *and, with probability at least* $1-p$, *the following holds simultaneously for all* $(x,a,h,k) \in S \times A \times [H] \times [K]$*:*

$$0 \leq \left(Q_h^k - Q_h^\star\right)(x,a) \leq \alpha_t^0 H + \sum_{i=1}^c \alpha_t^i \left(V_{h+1}^{k_i} - V_{h+1}^\star\right)\left(x_{h+1}^{k_i}\right) + \beta_t,$$

where $t = N_h^k(x,a)$ and $k_1, k_2, ..., k_t < k$ are the episodes where (x,a) was taken at step h.

Now we want to show the regret of Q-learning with UCB.Denote by

$$\delta_h^k := (V_h^k - V_h^{\pi_k})(x_h^k) \quad and \quad \phi_h^k := (V_h^k - V_h^*)(x_h^k)$$

By Lemma(3.3), we have the probability 1 - p, that $Q_h^k \geqslant Q_h^*$ and $V_h^k \geqslant V_h^*$. So the total regret can be bounded:

$$Regret(K) = \sum_{k=1}^K (V_1^* - V_1^{\pi_k})(x_1^k) \leqslant \sum_{k=1}^K (V_1^k - V_1^{\pi_k})(x_1^k) = \sum_{k=1}^K \delta_1^k$$

Now, we can use $V_h^k - V_h^{\pi_k}$ to bound $Q_h - Q_h^*$, and at the same time, we can use dynamic programming recursive to get $V_h^k - V_h^{\pi_k}$, or the bound of regret. According to the thought above, we have the the following induction:

$$\begin{aligned}
\delta_h^k &= (V_h^k - V_h^{\pi_k})(x_h^k) \leqslant (Q_h^k - Q_h^{\pi_k})(x_h^k, a_h^k) \\
&= (Q_h^k - Q_h^*)(x_h^k, a_h^k) + (Q_h^* - Q_h^{\pi_k})(x_h^k, a_h^k) \\
&\leqslant \alpha_t^0 H + \sum_{i=1}^t (\phi_t^i \delta_{h+1}^{k_i} + \beta_t + [(P_h - P_h^k)(V_{h+1}^* - V_{h+1}^k)](x_h^k, a_h^k)) \qquad (5) \\
&= \alpha_t^0 H + \sum_{i=1}^t (\phi_t^i \delta_{h+1}^{k_i} + \beta_t - \phi_{h+1}^k + \delta_{h+1}^k + \xi_{h+1}^k)
\end{aligned}$$

where $\beta_t = 2\sum \alpha_t^i b_i \leqslant O(1)\sqrt{H^3\iota/t}$ and $\phi_{h+1}^k := [(P_h - P_h^k)(V_{h+1}^* - V_{h+1}^k)](x_h^k, a_h^k)$ is a martingale difference sequence.

And next, we have the inequality of δ_h^k, the summation of δ_h^k is the upper bounds of Regret(K), so the thing we are going to do is the summation of δ_h^k to get the upper bounds of Regret(K). However, it's impossible for us to directly get the summation of δ_h^k, or $\sum_{k=1}^K \delta_1^k$, so we can divide δ_1^k into the format of above, equation(5). So, we have to compute the summation of respective term in the (5). For the first term, that is trivial we have, where $n_h^k = N_h^k(x_h^k, a_h^k)$:

$$\sum_{k=1}^K \alpha_{n_h^k}^0 H \leqslant SAH$$

Next, the second term in (5) and the summation of second term, which is:

$$\sum_{k=1}^K \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \phi_{h+1}^{k_i(x_h^k, a_h^k)} \leqslant (1 + \frac{1}{H}) \sum_{k=1}^K \phi_{h+1}^k$$

where the final inequality uses $\sum_{t=1}^\infty \alpha_t^i = 1 + \frac{1}{H}$ form the lemma Next, we can plugging the result back to (5), we have:

$$\sum_{k=1}^K \delta_h^k \leq SAH + \left(1 + \frac{1}{H}\right) \sum_{k=1}^K \phi_{h+1}^k - \sum_{k=1}^K \phi_{h+1}^k + \sum_{k=1}^K \delta_{h+1}^k + \sum_{k=1}^K \left(\beta_{n_h^k} + \xi_{h+1}^k\right)$$

this result implied that:

$$\sum_{k=1}^K \delta_1^k \leq O\left(H^2 SA + \sum_{h=1}^H \sum_{k=1}^K \left(\beta_{n_h^k} + \xi_{h+1}^k\right)\right)$$

Finally, by the pigeonhole principle, for any $h \in [H]$:

$$\textstyle\sum_{k=1}^K \beta_{n_h^k} \leq O(1) \cdot \sum_{k=1}^K \sqrt{\frac{H^3\iota}{n_h^k}} = O(1) \cdot \sum_{x,a} \sum_{n=1}^{N_h^K(x,a)} \sqrt{\frac{H^3\iota}{n}}$$

and this result is bound in $\widetilde{O}(\sqrt{H^2SAT})$
Also, by the AzumaHoeffding inequality with the probability 1 - p, we have:

$$\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\xi_{h+1}^{k}\right| = \left|\sum_{h=1}^{H}\sum_{k=1}^{K}\left[\left(P_h - P_h^k\right)\left(V_{h+1}^{\star} - V_{h+1}^{k}\right)\right]\left(x_h^k, a_h^k\right)\right| \leq cH\sqrt{T_l} \tag{6}$$

This establishes $\sum_{k=1}^{K}\delta_1^k \leq O(H^2SA + \sqrt{H^4SAT\iota})$, we can remove the H^2S term in the regret bound. So we can get the complexity of total regret $O(\sqrt{H^4SAT\iota})$

4 Conclusion

In my point of view, in this paper show that the regret complexity of model-free algorithm could be close to the model-based. It shows that Q-learning with UCB-Hoeffding could achieve $\widetilde{O}(\sqrt{H^4SAT})$, and from the table1 in page 2, we can find model-based approaches like UCBVI and vUCQ could achieve $\widetilde{O}(\sqrt{H^2SAT})$, the difference lies in the term $\sqrt{H}$. In fact, in addition to the Q-learning with UCB-Hoeffding, the author try another UCB algorithm, that is Q-learning with UCB-Bernstein(UCB-B) it achieved better result that is $O(\sqrt{H^3SAT})$. This result is much closer to the result of model-based approach. So, i mean there is no absolute answer to " Can we design model-free algorithms that are sample efficient? ". We have the same concept that using Q-learning with UCB, and it improve $\sqrt{H}$ in complexity, and the complexity of Q-learning with UCB algorithm itself is truly close to the model-based approaches. So, i think the potential future research directions that could combine the different model-free models like Sarsa and Policy Gradients, maybe it could get the better result. Moreover, the choice of exploration is truly importance, the choice of UCB and $\epsilon - greedy$ result in total different complexity, so the exploration algorithm is also a better research direction.
In the literature on bandits,including this paper , we are usually to make some special inequality to make the regret bound better, however, sometimes i think, it just played the bound because the perfect parameters are too far from the reality, we can not use in practical.
Recently research still have the debate on the efficiency between model-free and model-based, they try other method like Linear Quadratic Regulator, or someone talk about the worst-case regret bounds with value function. All the research are want to prove one thing "the effectiveness of model-based versus model-free methods", and still remain elusive, so this elusiveness is a charming work, right?

References

[1]Jin, C., Allen-Zhu, Z., Bubeck, S. and Jordan, MI, 2018. "Is Q-learning provably efficient?." Advances in Neural Information Processing Systems . 2018.