Is Q-learning Provably Efficient?

Lee Xin
Department of Computer Science
National Chiao Tung University
nene.cs08g@nctu.edu.tw

1 Introduction

Please provide a clear overview of the selected paper. You may want to discuss the following aspects:

- The main research challenges tackled by the paper
  Because the state space and action space is finite, so there exist an optimal policy. If we know the distribution of r_h, P_h, we can find the Q^* V^*, and then we can find the optimal policy. But we don't know the distribution of Q^* V^*. What we can do is to collect many samples (state, action, rewards, etc.) from the environment. After we get these samples, we can use these samples to estimate the Q_value function, and find an optimal policy to maximize the estimated Q_value function. So the main problem is how to estimate the Q_value function. In the prior works, they use ε-greedy exploration, which means to use the sample mean. But if we want to use UCB, we need to find the confidence interval of Q_value function. This is more complicated then finding the confidence interval in the bandit problem because the Q_value function is a random variable related to MDP. The following is algorithm that add the UCB exploration into Q-learning.

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 $a_h \leftarrow \operatorname{argmax}_{a'} Q_h(x_h, a')$, and observe x_{h+1}.
        $t = N_h(x_h, a_h) \leftarrow N_h(x_h, a_h) + 1;\ \ b_t \leftarrow c\sqrt{H^3\iota/t}$.
        $Q_h(x_h, a_h) \leftarrow (1-\alpha_t)Q_h(x_h, a_h) + \alpha_t[r_h(x_h, a_h) + V_{h+1}(x_{h+1}) + b_t]$.
        $V_h(x_h) \leftarrow \min\{H, \max_{a' \in \mathcal{A}} Q_h(x_h, a')\}$.

Figure 1: Algorithm 1 Q-learning with UCB-Hoeffding

- The high-level technical insights into the problem of interest
  There are two kinds of Q-learning with UCB, they named it UCB-H and UCB-B. The difference is that different concentration inequality need different upper confidence bound. This paper focus on the proof of UCB-H. In the algorithm, $Q_h(x_h, a_h) \leftarrow (1-\alpha_t)Q_h(x_h, a_h) + \alpha_t[r_h(x_h, a_h) + V_{h+1}(x_{h+1}) + b_t]$ means how to estimate the Q_value function. There are two terms, the first term $(1-\alpha_t)Q_h(x_h, a_h)$ can be seen as the momentum term in optimize algorithm. The second term $\alpha_t[r_h(x_h, a_h) + V_{h+1}(x_{h+1}) + b_t]$ is the UCB bound iteration. Here, they found that if use learning rate $\alpha_t = O(H/t)$, instead of the prior works' setting $\alpha_t = 1/t$, will get a better result. This means that we can't use an uniform weight and we need to give the near term a higher weight.
- The main contributions of the paper (compared to the prior works)

Preprint. Under review.

The main contributions of the paper is they prove that in an episodic MDP setting, Q-learning with UCB exploration achieves regret $\widetilde{O}(\sqrt{H^3SAT})$, where S is the number of states, A is the number of actions, H is the number of steps per episode, T is the total number of steps. This sample efficiency matches the optimal regret that can be achieved by any model-based approach, up to a single $\sqrt{H}$ factor. This is the first analysis in the model-free setting that establishes $\sqrt{T}$ regret without requiring access to a simulator.

- Your personal perspective on the proposed method

  In this paper, they use UCB exploration instead of ϵ-greedy exploration. UCB exploration use confidence bound instead of ϵ-greedy using the expected sample mean. Before I know the UCB exploration method, I think the ϵ-greedy is the only method to deal with the exploration-exploitation problem. Maybe in the future, I can study more to learn others exploration method.

2 Problem Formulation

Please present the formulation in this section. You may want to cover the following aspects:

- Your notations (e.g. MDPs, value functions, function approximators,...etc)

  Here is the notations of the proof.

  First, give the definition of regret :

  $$regret(K) = \sum_{k=1}^{K}[V_1^*(x_1^k) - V_1^{\pi_k}(x_1^k)]$$

  Here, K is the number of episode, π_k is the policy we use in episode k, x_1^k is the initial state of every episode k.

  We denote by (x_1^k, a_1^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. By these notations, the code for update Q function:

  $$Q_h(x_h, a_h) \leftarrow (1-\alpha_t)Q_h(x_h, a_h) + \alpha_t[r_h(x_h, a_h) + V_{h+1}(x_{h+1}) + b_t] \tag{1}$$

  can be rewritten as follows, for every $h \in [H]$:

  $$Q_h^{k+1}(x,a) = \begin{cases} (1-\alpha_t)Q_h^k + \alpha_t[r_h(x,a) + V_{h+1}^k(x_{h+1}^k) + b_t] & if\ (x,a) = (x_h^k, a_h^k) \\ Q_h^k(x,a) \quad otherwise. \end{cases} \tag{2}$$

- The technical assumptions

  They have chosen the learning rate as $\alpha_t := \frac{H+1}{H+t}$, where t is the counter for how many times the algorithm has visited the state-action pair (x, a).

  For notational convenience, introduce the follow related quantities :

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

  It can be verified that :

  $$\sum_{i=1}^{t}(\alpha_t^i) = 1\ and\ \alpha_t^0 = 0\ for\ t \geq 1$$
  $$\sum_{i=1}^{t}(\alpha_t^i) = 0\ and\ \alpha_t^0 = 1\ for\ t = 0$$

  With (2) and (3), we have :

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

  Here, we discover that α_t^i reflect the UCB bound weight allocating of our Q-learning algorithm.

  In Figure 2, $1/t$ is uniform, and $1/\sqrt{t}$ gives the weight to the nearest samples, and this will cause high variance. Compare to the learning rate $1/t$ and $1/\sqrt{t}$, set learning rate $\alpha_t = \frac{H+1}{H+t}$ will has a stable performance.

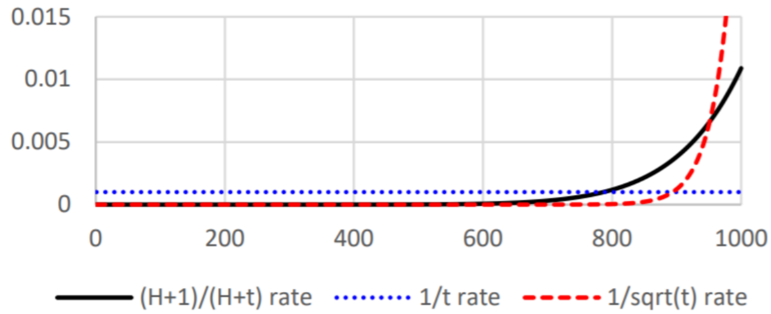

Figure 2: Illustration of $\{\alpha^i_{1000}\}_{i=1}^{1000}$ for learning rates $\alpha_t = \frac{H+1}{H+t}$, $\frac{1}{t}$ and $\frac{1}{\sqrt{t}}$ when $H = 10$

- The optimization problem of interest
  Q-learning with Hoeffding-style bonus (UCB-H)

  In this method, their choice of b_t is $b_t = O(\sqrt{H^3\iota/t})$, where $\iota := log(SAT/p)$, denote a log factor. This choice can make Q-values upper-bounded by H. Also, Hoeffding-type martingale concentration inequalities imply that if we have visited (x, a) for t times, then a confidence bound for the Q value scales as $1/\sqrt(t)$. According to this choice, they give the theorem 1 :

  Theorem 1 : 1 (Hoeffding). There exists an absolute constant $c > 0$ such that, for any $p \in (0, 1)$, if we choose $b_t = c\sqrt{H^3\iota/t}$, then with probability $1 - p$, the total regret of Q-learning with UCB-Hoeffding is at most $O(\sqrt{H^4SAT\iota})$, where $\iota := log(SAT/p)$.

  Q-learning with Bernstein-style bonus (UCB-B)

  Here, they set b_t by making use of a Bernsteinstyle upper confidence bound. Because the length may be too long, so I didn't write the proof UCB-B in this theory project and only give the theorem here.

  Theorem 2 (Bernstein). For any $p \in (0, 1)$, one can specify b_t so that with probability $1 - p$, the total regret of Q-learning with UCB-Bernstein is at most $O\left(\sqrt{H^3SAT\iota} + \sqrt{H^9S^3A^3 \cdot \iota^2}\right)$.

3 Theoretical Analysis

Please present the theoretical analysis in this section. Moreover, please formally state the major theoretical results using theorem/proposition/corollary/lemma environments. Also, please clearly highlight your new proofs or extensions (if any).

In the proof, they give three Lemmas based on the definition of α_t first :

Lemma 4.1 The following properties hold for α^i_t :

(a) $\frac{1}{\sqrt{t}} \le \sum_{i=1}^{t} \frac{\alpha^i_t}{\sqrt{i}} \le \frac{2}{\sqrt{t}}$ for every $t \ge 1$

(b) $\max_{i\in[t]} \alpha^i_t \le \frac{2H}{t}$ and $\sum_{i=1}^{t} \left(\alpha^i_t\right)^2 \le \frac{2H}{t}$ for every $t \ge 1$

(c) $\sum_{t=i}^{\infty} \alpha^i_t = 1 + \frac{1}{H}$ for every $i \ge 1$

Lemma 4.2 (recursion on Q). For any $(x, a, h) \in S \times A \times [H]$ and episode $k \in [K]$, let $t = N^k_h(x, a)$ and suppose (x, a) was previously taken at step h of episodes $k_1, ..., k_t < k$. Then :

$$\left(Q^k_h - Q^\star_h\right)(x, a) =$$
$$\alpha^0_t \left(H - Q^\star_h(x, a)\right) + \sum_{i=1}^{t} \alpha^i_t \left[\left(V^{k_i}_{h+1} - V^\star_{h+1}\right)\left(x^{k_i}_{h+1}\right) + \left[\left(\hat{\mathbb{P}}^{k_i}_h - \mathbb{P}_h\right) V^\star_{h+1}\right](x, a) + b_i\right]$$

Lemma 4.3 (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 $\beta_t = 2\sum_{i=1}^{t}(\alpha_t^i b_i) \le 4c\sqrt{H^3\iota/t}$ and, with probability at least $1-p$, the following holds simultaneously for all $(x,a,h,k) \in S \times A \times [H] \times [K]$:

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

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.

The proof of these three Lemmas are at the end of this section. Now, after we get the three Lemma, we can start to prove Theorem 1.

proof of Theorem 1

Denote by

$$\delta_h^k := \left(V_h^k - V_h^{\pi_k}\right)\left(x_h^k\right) \quad \text{ and } \quad \phi_h^k := \left(V_h^k - V_h^\star\right)\left(x_h^k\right)$$

By Lemma 4.3, $Q_h^k \ge Q_h^*$ with probability $1-p$ and thus $V_h^k \ge V_h^*$ with probability $1-p$. Thus, the total regret can be upper bounded :

$$\operatorname{Regret}(K) = \textstyle\sum_{k=1}^{K}\left(V_1^\star - V_1^{\pi_k}\right)\left(x_1^k\right) \le \sum_{k=1}^{K}\left(V_1^k - V_1^{\pi_k}\right)\left(x_1^k\right) = \sum_{k=1}^{K}\delta_1^k$$

What we want to do next is to use the next step $\sum_{k=1}^{K}\delta_{h+1}^k$ to upper bound $\sum_{k=1}^{K}\delta_h^k$, and we can get a recursive formula to calculate total regret. We can obtain such a recursive formula by relating $\sum_{k=1}^{K}\delta_h^k$ to $\sum_{k=1}^{K}\phi_h^k$.

Let $t = N_h^k(x_h^k, a_h^k)$ for any fixed $(x,h) \in [K] \times [H]$, and suppose (x_h^k, a_h^k) were previously taken at step h of episodes $k_1, k_2, ..., k_t < k$. Then we have :

$$\delta_h^k = \left(V_h^k - V_h^{\pi_k}\right)\left(x_h^k\right) \tag{5}$$

Because $V_h^k\left(x_h^k\right) \le \max_{a' \in \mathcal{A}} Q_h^k\left(x_h^k, a'\right) = Q_h^k\left(x_h^k, a_h^k\right)$, so we have :

$$\left(V_h^k - V_h^{\pi_k}\right)\left(x_h^k\right) \le \left(Q_h^k - Q_h^{\pi_k}\right)\left(x_h^k, a_h^k\right) = \left(Q_h^k - Q_h^\star\right)\left(x_h^k, a_h^k\right) + \left(Q_h^\star - Q_h^{\pi_k}\right)\left(x_h^k, a_h^k\right) \tag{6}$$

By Lemma 4.3 and Bellman equation, we have :

$$\left(Q_h^k - Q_h^\star\right)\left(x_h^k, a_h^k\right) + \left(Q_h^\star - Q_h^{\pi_k}\right)\left(x_h^k, a_h^k\right) \le \alpha_t^0 H + \sum_{i=1}^{t}\alpha_t^i \phi_{h+1}^{k_i} + \beta_t + \left[\mathbb{P}_h\left(V_{h+1}^\star - V_{h+1}^{\pi_k}\right)\right]\left(x_h^k, a_h^k\right) \tag{7}$$

By definition $\delta_{h+1}^k - \phi_{h+1}^k = \left(V_{h+1}^\star - V_{h+1}^{\pi_k}\right)\left(x_{h+1}^k\right)$, we have :

$$\alpha_t^0 H + \sum_{i=1}^{t}\alpha_t^i \phi_{h+1}^{k_i} + \beta_t + \left[\mathbb{P}_h\left(V_{h+1}^\star - V_{h+1}^{\pi_k}\right)\right]\left(x_h^k, a_h^k\right) = \alpha_t^0 H + \sum_{i=1}^{t}\alpha_t^i \phi_{h+1}^{k_i} + \beta_t - \phi_{h+1}^k + \delta_{h+1}^k + \xi_{h+1}^k \tag{8}$$

The $\beta_t = 2\sum \alpha_t^i b_i \le O(1)\sqrt{H^3\iota/t}$ and $\xi_{h+1}^k := \left[\left(\mathbb{P}_h - \hat{\mathbb{P}}_h^k\right)\left(V_{h+1}^\star - V_{h+1}^k\right)\right]\left(x_h^k, a_h^k\right)$ is a martingale difference sequence.

So here we have :

$$\begin{aligned}
\delta_h^k &= \left(V_h^k - V_h^{\pi_k}\right)\left(x_h^k\right) \le \left(Q_h^k - Q_h^{\pi_k}\right)\left(x_h^k, a_h^k\right) \\
&= \left(Q_h^k - Q_h^\star\right)\left(x_h^k, a_h^k\right) + \left(Q_h^\star - Q_h^{\pi_k}\right)\left(x_h^k, a_h^k\right) \\
&\le \alpha_t^0 H + \sum_{i=1}^{t}\alpha_t^i \phi_{h+1}^{k_i} + \beta_t + \left[\mathbb{P}_h\left(V_{h+1}^\star - V_{h+1}^{\pi_k}\right)\right]\left(x_h^k, a_h^k\right) \\
&= \alpha_t^0 H + \sum_{i=1}^{t}\alpha_t^i \phi_{h+1}^{k_i} + \beta_t - \phi_{h+1}^k + \delta_{h+1}^k + \xi_{h+1}^k
\end{aligned} \tag{9}$$

Now we want to find the summation $\sum_{k=1}^{K}\delta_h^k$. Denoting by $n_h^k = N_h^k(x_h^k, a_h^k)$, we have :

$$\sum_{k=1}^{K} \alpha_{n_h^k}^0 H = \sum_{k=1}^{K} H \cdot \mathbb{I}\left[n_h^k = 0\right] \leq SAH$$

To upper bound the second term in (9) :

$$\sum_{k=1}^{K} \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \phi_{h+1}^{k_i\left(x_h^k, a_h^k\right)}, \tag{10}$$

where $k_i(x_h^k, a_h^k)$ is the episode in which (x_h^k, a_h^k) was taken at step h for the ith time.

For every $k' \in [K]$, the term $\phi_{h+1}^{k'}$ appears in the summand with $k > k'$ if and only if $(x_h^k, a_h^k) = (x_h^{k'}, a_h^{k'})$. Because we have $n_h^k =_h^{k'} +1$ when it appears the first time and $n_h^k =_h^{k'} +2$ when it appears the first time, so we have:

$$\sum_{k=1}^{K} \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \phi_{h+1}^{k_i\left(x_h^k, a_h^k\right)} \leq \sum_{k'=1}^{K} \phi_{h+1}^{k'} \sum_{t=n_h^{k'}+1}^{\infty} \alpha_t^{n_h^{k'}}$$

By $\sum_{t=i}^{\infty} \alpha_t^i = 1 + \frac{1}{H}$ in Lemma 4.1 (c), we have :

$$\sum_{k'=1}^{K} \phi_{h+1}^{k'} \sum_{t=n_h^{k'}+1}^{\infty} \alpha_t^{n_h^{k'}} \leq \left(1 + \frac{1}{H}\right) \sum_{k=1}^{K} \phi_{h+1}^k$$

Plug back into (9) :

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

Uses $\phi_{h+1}^k \leq \delta_{h+1}^k$ (owing to the fact that $V^* \geq V * \pi_k$), we have :

$$\begin{aligned} SAH + \left(1 + \frac{1}{H}\right) \sum_{k=1}^{K} \phi_{h+1}^k - \sum_{k=1}^{K} \phi_{h+1}^k + \sum_{k=1}^{K} \delta_{h+1}^k + \sum_{k=1}^{K} \left(\beta_{n_h^k} + \xi_{h+1}^k\right) \\ \leq SAH + \left(1 + \frac{1}{H}\right) \sum_{k=1}^{K} \delta_{h+1}^k + \sum_{k=1}^{K} \left(\beta_{n_h^k} + \xi_{h+1}^k\right) \end{aligned} \tag{11}$$

Recursing the result for $h = 1, 2, ..., H$, and using the fact $\delta_{H+1}^K \equiv 0$, we have :

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

Here, Use the pigeonhole principle, for any $h \in [H]$:

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

Because $\sum_{x,a} N_h^k(x,a) = K$ and $O(1) \cdot \sum_{x,a} \sum_{n=1}^{N_h^K(x,a)} \sqrt{\frac{H^3 \iota}{n}}$ is maximized when $N_h^k(x,a) = K/SA$ for all x, a, we have :

$$O(1) \cdot \sum_{x,a} \sum_{n=1}^{N_h^K(x,a)} \sqrt{\frac{H^3 \iota}{n}} \leq O(\sqrt{H^3 SAK_l}) = O(\sqrt{H^2 SAT\iota}) \tag{12}$$

By the AzumaHoeffding inequality, with probability $1 - p$, we have :

$$\left|\sum_{h=1}^{H} \sum_{k=1}^{K} \xi_{h+1}^k\right| = \left|\sum_{h=1}^{H} \sum_{k=1}^{K} \left[\left(\mathbb{P}_h - \hat{\mathbb{P}}_h^k\right) \left(V_{h+1}^\star - V_{h+1}^k\right)\right] \left(x_h^k, a_h^k\right)\right| \leq cH\sqrt{T_l} \tag{13}$$

Equation (13) establishes $\sum_{k=1}^{K} \delta_1^k \leq O\left(H^2 SA + \sqrt{H^4 SAT\iota}\right)$. Because

(1) when $T \geq \sqrt{H^4 SAT\iota}$, $\sqrt{H^4 SAT\iota} \geq H^2 SA$
(2) when $T \leq \sqrt{H^4 SAT\iota}$, $\sum_{k=1}^{K} \delta_1^k \leq HK = T \leq \sqrt{H^4 SAT\iota}$

, we can remove the H^2SA term in the regret upper bound.

In sum, we have $\sum_{k=1}^{K} \delta_1^k \leq O\left(H^2SA + \sqrt{H^4SAT\iota}\right)$, with probability at least $1-2p$. Finally, change the p to $p/2$ and the proof is over.

Below are the proof of Lemma 4.1, 4.2, 4.3

derive the properties implied by the choice of the learning rate. Recall the notation :

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

$Lemma 4.1$ The following properties hold for α_t^i :
(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 $i \geq 1$

proof of Lemma 4.1

(a) Here, we use induction on t to prove.

If $t = 1$:

$$\sum_{i=1}^{t} \frac{\alpha_t^i}{\sqrt{i}} = \alpha_1^1 = 1$$

hold.

If $t \geq 2$, $\alpha_t^i = (1-\alpha_t)\,\alpha_{t-1}^i$ for $i = 1, 2, ..., t-1$, we have :

$$\sum_{i=1}^{t} \frac{\alpha_t^i}{\sqrt{i}} = \frac{\alpha_t}{\sqrt{t}} + (1-\alpha_t)\sum_{i=1}^{t-1} \frac{\alpha_{t-1}^i}{\sqrt{i}}$$

On the one hand, by induction :

$$\frac{\alpha_t}{\sqrt{t}} + (1-\alpha_t)\sum_{i=1}^{t-1} \frac{\alpha_{t-1}^i}{\sqrt{i}} \geq \frac{\alpha_t}{\sqrt{t}} + \frac{1-\alpha_t}{\sqrt{t-1}} \geq \frac{\alpha_t}{\sqrt{t}} + \frac{1-\alpha_t}{\sqrt{t}} = \frac{1}{\sqrt{t}}$$

On the other hand, by induction :

$$\begin{aligned} \frac{\alpha_t}{\sqrt{t}} + (1-\alpha_t)\sum_{i=1}^{t-1} \frac{\alpha_{t-1}^i}{\sqrt{i}} &\leq \frac{\alpha_t}{\sqrt{t}} + \frac{2(1-\alpha_t)}{\sqrt{t-1}} = \frac{H+1}{\sqrt{t}(H+t)} + \frac{2\sqrt{t-1}}{H+t} \\ &\leq \frac{H+1}{\sqrt{t}(H+t)} + \frac{2\sqrt{t}}{H+t} = \frac{2}{\sqrt{t}} + \frac{1}{\sqrt{t}} \cdot \frac{1-H}{t+H} \leq \frac{2}{\sqrt{t}} \end{aligned}$$

Because $H \geq 1$, the final inequality holds.

(b) We have :

$$\begin{aligned} \alpha_t^i &= \frac{H+1}{i+H} \cdot \left(\frac{i}{i+1+H}\frac{i+1}{i+2+H}\cdots\frac{t-1}{t+H}\right) \\ &= \frac{H+1}{t+H} \cdot \left(\frac{i}{i+H}\frac{i+1}{i+1+H}\cdots\frac{t-1}{t-1+H}\right) \leq \frac{H+1}{t+H} \leq \frac{2H}{t} \end{aligned}$$

Because $\sum_{i=1}^{t}\left(\alpha_t^i\right)^2 \leq \left[\max_{i\in[t]} \alpha_t^i\right] \cdot \sum_{i=1}^{t} \alpha_t^i$ and $\sum_{i=1}^{t} \alpha_t^i = 1$, we have :

$$\frac{H+1}{t+H} \leq \frac{2H}{t}$$

Therefore, we have proved $\max_{i\in[t]} \alpha_t^i \leq 2H/t$.

(c) First note the following identity, which holds for all positive integers n and k with $n \geq k$:

$$\frac{n}{k} = 1 + \frac{n-k}{n+1} + \frac{n-k}{n+1}\frac{n-k+1}{n+2} + \frac{n-k}{n+1}\frac{n-k+1}{n+2}\frac{n-k+2}{n+3} + \cdots \quad (B.1)$$

To verify this, we write the terms of its right-hand side as $x_0 = 1, x_1 = \frac{n-k}{n+1}, \ldots$. It can be verify by induction that $\frac{n}{k} - \sum_{i=0}^{t} x_i = \frac{n-k}{k} \prod_{i=1}^{t} \frac{n-k+i}{n+i}$. This means $\lim_{t\to\infty} \frac{n}{k} - \sum_{i=0}^{t} x_i = 0$ and this proves (B.1).

Using (B.1) with $n = i + H$ and $k = H$, we have :

$$\sum_{t=i}^{\infty} \alpha_t^i = \frac{H+1}{i+H} \cdot \left(1 + \frac{i}{i+1+H} + \frac{i}{i+1+H}\frac{i+1}{i+2+H} + \cdots\right) = \frac{H+1}{i+H} \cdot \frac{i+H}{H} = \frac{H+1}{H}$$

proof of Lemma 4.2 (recursion on Q).

From the Bellman optimality equation :

$$Q_h^\star(x,a) = \left(r_h + \mathbb{P}_h V_{h+1}^\star\right)(x,a),$$

our notation :

$$\left[\hat{\mathbb{P}}_h^{k_i} V_{h+1}\right](x,a) := V_{h+1}\left(x_{h+1}^{k_i}\right),$$

and the fact that :

$$\sum_{i=0}^{t} \alpha_t^i = 1$$

we have :

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

Subtracting the formula (4) from this equation, we obtain Lemma 4.2.

proof of Lemma 4.3 (bound on $Q^k - Q^*$).
For each fixed $(x,a,h) \in S \times A \times H$, denote $k_0 = 0$, and denote

$$k_i = \min\left(\left\{k \in [K] \mid k > k_{i-1} \wedge \left(x_h^k, a_h^k\right) = (x,a)\right\} \cup \{K+1\}\right)$$

Here, k_i is the episode of which (x,a) was taken at step h for the ith time. Let $\mathcal{F}_i$ be the σ-field generated by all the random variables until episode k_i, step h. Then,

$$\left(\mathbb{I}\left[k_i \le K\right] \cdot \left[\left(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h\right) V_{h+1}^\star\right](x,a)\right)_{i=1}^{\tau}$$

is a martingale difference sequence w.r.t. the filtration $\{\mathcal{F}_i\}_{i\ge 0}$. By the Azuma-Hoeffding and a union bound, we have that with probability at least $1 - p/(SAH)$:

$$\forall \tau \in [K]: \quad \left|\sum_{i=1}^{\tau} \alpha_\tau^i \cdot \mathbb{I}\left[k_i \le K\right] \cdot \left[\left(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h\right) V_{h+1}^\star\right](x,a)\right| \le \frac{cH}{2}\sqrt{\sum_{i=1}^{\tau} \left(\alpha_\tau^i\right)^2 \cdot \iota} \le c\sqrt{\frac{H_l^3}{\tau}} \tag{14}$$

for some absolute constant c. Inequality (14) holds for all fixed $\tau \in [K]$ uniformly, and holds for $\tau = t = N_h^k(x,a) \le K, k \in [K]$. Putting everything together and using a union bound, with least $1-p$ probability,

$$\left|\sum_{i=1}^{t} \alpha_t^i \left[\left(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h\right) V_{h+1}^\star\right](x,a)\right| \le c\sqrt{\frac{H^{3_L}}{t}} \quad \text{where} \quad t = N_h^k(x,a) \tag{15}$$

for all $(x,a,h,k) \in S \times A \times [H] \times [K]$.
If choose $b_t = c\sqrt{H^3\iota/t}$ for the same constant c in the equation (14), by Lemma 4.1.a, we have :

$$\beta_t/2 = \sum_{i=1}^{t} \alpha_t^i b_i \in [c\sqrt{H^3\iota/t}, 2c\sqrt{H^3 l/t}]$$

Then the right-hand side of Lemma 4.3 follows immediately from Lemma 4.2 and inequality (15). The left-hand side also follows from Lemma 4.2 and inequality (15) and induction on $h = H, H-1, ..., 1$.

4 Conclusion

Please provide succinct concluding remarks for your report. You may discuss the following aspects:

- The potential future research directions

  In an episodic MDP, the information-theoretic lower bound is $\widetilde{O}(\sqrt{H^2SAT})$. The method they give in this paper, which called UCB-H and UCB-B, tried to get close to the theoretically optimal regret. UCB-H got a regret $\widetilde{O}(\sqrt{H^4SAT})$, which has a $\sqrt{H^2}$ difference from the optimal regret. The other method they give is UCB-B, which got a regret $\widetilde{O}(\sqrt{H^3SAT})$, has a $\sqrt{H}$ difference from the optimal model.

  The method UCB-B has given a solution that only has a $\sqrt{H^2}$ difference from the optimal regret, which means that the method is very close to the optimal. But there is still a gap between UCB-B and optimal method. So, maybe how to get the optimal $\widetilde{O}(\sqrt{H^2SAT})$ regret in model-free setting will be the potential future research directions.

- Any technical limitations

  This paper want to find out whether a model-free method can receive the same sample efficient as model-based method or not. I think this sound impossible at first because doing sample in model-free method is harder then model-based method. But after reading this paper, I found that they discover a model-free method that can regret is close to model-based regret. Although the model-based method's sample efficient is still better than model-free method, I think maybe in the future, model-free method can achieve the same sample efficient as model-based method.

- Any latest results on the problem of interest

  This paper proposed a Q-learning algorithm with UCB exploration policy for finite-horizon MDP. In paper "Q-learning with UCB Exploration is Sample Efficient for Infinite-Horizon MDP", they adapt Q-learning with UCB-exploration bonus to infinite-horizon MDP with discounted rewards without accessing a generative model. They find that the sample complexity of exploration of our algorithm is bounded by $\tilde{O}\left(\frac{SA}{\epsilon^2(1-\gamma)^7}\right)$. This result improves the previously best known result of $\tilde{O}\left(\frac{SA}{\epsilon^4(1-\gamma)^8}\right)$ in this setting achieved by delayed Q-learning, and matches the lower bound in terms of ϵ as well as S and A up to logarithmic factors.

References

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is q-learning provably efficient?, 2018.

Chen-Yu Wei, Mehdi Jafarnia-Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes, 2019.

Kefan Dong, Yuanhao Wang, Xiaoyu Chen, and Liwei Wang. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp, 2019.

Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 881–888, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143955. URL `https://doi.org/10.1145/1143844.1143955`.

Marc Peter Deisenroth and Carl Edward Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 465–472, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Mohammad Gheshlaghi Azar, Remi Munos, and Bert Kappen. On the sample complexity of reinforcement learning with a generative model, 2012.

Jin et al. [2018] Wei et al. [2019] Dong et al. [2019] Strehl et al. [2006] Deisenroth and Rasmussen [2011] Azar et al. [2012]