Is Q-learning Provably Efficient?

Kushagra Rastogi, Jonathan Lee, Aditya Joglekar, Fabrice Harel-Canada

University of California, Los Angeles

Abstract

We propose to analyze the theoretical results presented within the paper *Is Q-Learning Provably Efficient?* by Jin *et al.* [1]. This analysis shall include a survey of related research to contextualize the need for strengthening the theoretical guarantees related to perhaps the most important threads of model-free reinforcement learning. We also hope to expound upon the reasoning used in the proofs to highlight the critical leading steps to the main results: Q-learning with UCB exploration achieves a sample efficiency that matches the optimal regret that can be achieved by any model-based approach.

Related Work

- *Sample efficiency* measures the number of inputs an agent requires in order to achieve a given level of performance on a particular task.
- *Sample complexity* measures the minimum number of inputs required to guarantee a probably approximately correct (PAC) estimator.

	Algorithm	Regret	Time	Space
MB	UCRL2 [25]	$\geq \mathcal{O}(\sqrt{H^4S^2AT})$	$\Omega(TS^2A)$	$\mathcal{O}(S^2AH)$
	Agrawal & Jia [23]	$\geq \mathcal{O}(\sqrt{H^3S^2AT})$		
	UCBVI [24]	$\mathcal{O}(\sqrt{H^2SAT})$	$\mathcal{O}(TS^2A)$	
	vUCQ [26]	$\mathcal{O}(\sqrt{H^2SAT})$		
MF	Delayed Q-learning [30]	$\mathcal{O}_{S,A,H}(T^{4/5})$	$\mathcal{O}(T)$	$\mathcal{O}(SAH)$
	Q-learning (UCB-H) [1]	$\mathcal{O}(\sqrt{H^4SAT})$		
	Q-learning (UCB-B) [1]	$\mathcal{O}(\sqrt{H^3SAT})$		
	information theoretic lower bound	$\Omega(\sqrt{H^2SAT})$	—	—

Regret Comparisons for other RL methods in model-based (MB) and model-free (MF) episodic MDPs: $T = KH$ is the total number of steps, H is the steps per episode, S is the number of states, and A is the number of actions.

Setting & Notation

- Consider a tabular episodic Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$.
- $\mathcal{S}$ is the set of states. $\mathcal{A}$ is the set of actions. H is the number of steps in each episode. $\mathbb{P}$ is the transition matrix. $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ is a deterministic reward function at step h.
- Denote $V_h^\pi : S \to \mathbb{R}$ as the value function at step h under policy π. Denote $Q_h^\pi(x,a) : S \times \mathcal{A} \to \mathbb{R}$ as the Q-value function at step h under policy π. In symbols:

$$V_h^\pi(x) = \mathbb{E}[\sum_{h'=h}^{H} r_{h'}(x_{h'}, \pi_{h'}(x_{h'}))|x_h = x]$$

$$Q_h^\pi(x,a) = r_h(x,a) + \mathbb{E}[\sum_{h'=h+1}^{H} r_{h'}(x_{h'}, \pi_{h'}(x_{h'}))|x_h = x, a_h = a]$$

- Define $[\mathbb{P}_h V_{h+1}](x,a) = \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)} V_{h+1}(x')$ and its empirical counterpart $[\hat{\mathbb{P}}_h^k V_{h+1}](x,a) = V_{h+1}(x_{h+1}^k)$ which is only defined for $(x,a) = (x_h^k, a_h^k)$.
- The agent plays the game for K episodes $k = 1, 2, ..., K$ and we let the adversary pick a starting state x_1^k for each episode k and let the agent choose a policy π_k before starting the k-episode.
- The total expected regret is $Regret(K) = \sum_{k=1}^K [V_1^*(x_1^k) - V_1^{\pi_k}(x_1^k)]$
- Define $\delta_h^k := (V_h^k - V_h^{\pi_k})(x_h^k)$ and $\phi_h^k := (V_h^k - V_h^*)(x_h^k)$. ϕ is associated with the optimal action while δ pertains to the regret associated by choosing a certain action at step k not necessarily the optimal one.

Main Results

- Primary challenge: Choice of exploration policy is important. In episodic MDP, Q-learning with ε-greedy exploration can be inefficient.
- Contributions of the original paper: **(1)** Deriving an upper bound on the sample complexity of Q-learning with UCB exploration and an exploration bonus **(2)** Establishing an information-theoretic lower bound for the expected regret for the episodic MDP problem.
- Choice of learning rate $\alpha_t = \frac{H+1}{H+t}$ is very important to admit a regret that is sub-exponential in H.
- Q-learning with UCB exploration and Hoeffding-style bonus is presented below.

Algorithm 1 Q-learning with UCB-Hoeffding

```
initialize Q_h(x,a) ← H and N_h(x,a) ← 0 for all (x,a,h) ∈ S × A × [H].
for episode k = 1,...,K do
    receive x_1.
    for step h = 1,...,H do
        Take action a_h ← argmax_{a'} Q_h(x_h, a'), and observe x_{h+1}.
        t = N_h(x_h, a_h) ← N_h(x_h, a_h) + 1;  b_t ← c√(H³ι/t).
        Q_h(x_h, a_h) ← (1 − α_t)Q_h(x_h, a_h) + α_t[r_h(x_h, a_h) + V_{h+1}(x_{h+1}) + b_t].
        V_h(x_h) ← min{H, max_{a'∈A} Q_h(x_h, a')}.
```

- **Theorem 1** (Hoeffding). *If $b_t = c\sqrt{H^3\iota/t}$, then with probability $1-p \; \forall p \in (0,1)$, the total regret of Algorithm 1 is at most $\mathcal{O}(\sqrt{H^4SAT\iota})$ where $c > 0$ is a constant and $\iota = \log(SAT/p)$.*

  A list of some useful, related quantities: $\alpha_t^0 = \prod_{j=1}^t 1 - \alpha_j$ and $\alpha_t^i = \prod_{j=i+1}^t 1 - \alpha_j$.

  Lemma 1.1. *Properties of α_t^i:*

  (a) *For every $t \geq 1$, $\frac{1}{\sqrt{t}} \leq \sum_{i=1}^t \frac{\alpha_t^i}{\sqrt{i}} \leq \frac{2}{\sqrt{t}}$.*

  (b) *For every $t \geq 1$, $\max_{i\in[t]} \alpha_t^i \leq \frac{2H}{t}$ and $\sum_{i=1}^t (\alpha_t^i)^2 \leq \frac{2H}{t}$.*

  (c) *For every $i \geq 1$, $\sum_{t=i}^\infty \alpha_t^i = 1 + \frac{1}{H}$.*

  Lemma 1.2. *For any $(x,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ and episode $k \in [K]$ let $t = N_h^k(x,a)$ and suppose (x,a) was previously taken at step h of episodes $k_1, k_2, ..., k_t < k$. Then:*

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

  Lemma 1.3. *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 $\forall (x,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$:*

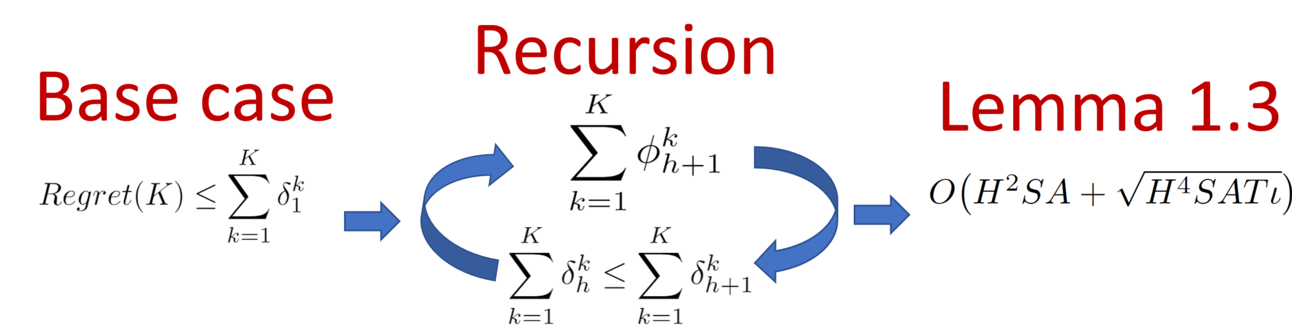
$$0 \leq (Q_h^k - Q_h^*)(x,a) \leq \alpha_t^0 H + \sum_{i=1}^{t} \alpha_t^i (V_{h+1}^{k_i} - V_{h+1}^*)(x_{h+1}^{k_i}) + \beta_t$$

- The upper bound in Theorem 1 can be improved by a factor of $\sqrt{H}$ if the Hoeffding-style upper confidence bound is replaced with Bernstein-style martingale concentration inequalities. However, this requires the bonus term b_t to be designed more carefully.
- **Theorem 2** (Bernstein). *For a specified b_t, with probability $1-p \; \forall p \in (0,1)$, the total regret of Q-learning with UCB-Bernstein exploration is at most $\mathcal{O}(\sqrt{H^3SAT\iota} + \sqrt{H^9S^3A^3\iota^4})$.*
- **Theorem 3** (Information-theoretic lower bound). *The total regret for any algorithm in an episodic MDP setting must be at least $\Omega(\sqrt{H^2SAT})$.*

Proofs

- This section only provides proofs for Theorem 1 and its lemmas.
- **Proof of Lemma 1.1.** The properties are based on simple manipulations of α_t and can be proved by induction.

  Note that property (c) is particularly important since it is used to show that the regret can only blow up by a constant factor of $(1 + 1/H)^H$ in each episode.
- **Proof of Lemma 1.2.** The main idea is to express $Q - Q*$ as a weighted average of previous updates of the Q-value function. This is achieved by manipulating the Bellman optimality equation $Q_h^*(x,a) = (r_h + \mathbb{P}_h V_{h+1}^*)(x,a)$ using $\sum_{i=1}^t \alpha_t^i = 1$ and $[\hat{\mathbb{P}}_h^{k_i} V_{h+1}](x,a) = V_{h+1}(x_{h+1}^{k_i})$.
- **Proof of Lemma 1.3.** The crux of the proof stems from the realization of the empirical error $[(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h)V_{h+1}^*](x,a)$ from Lemma 1.2 as a martingale difference sequence (MDS). Properties of Lemma 1.1 along with the invocation of the Azuma-Hoeffding inequality given the existence of an MDS are subsequently used to place an upper bound on the sum over the elements of the MDS with high confidence $1-p \in (0,1)$. Applying this upper confidence bound on $Q_h^*(x,a)$ defined in Lemma 1.2 completes the proof.
- **Proof of Theorem 1.** Recall the formulae for δ_h^k and ϕ_h^k. The proof relates the regret to these terms and manipulates them using the aforementioend lemmas. Note that $Regret(K) \leq \sum_{k=1}^K \delta_1^k$. The main idea is then to upper-bound the $\sum_{k=1}^K \delta_h^k$ by the next step $\sum_{k=1}^K \delta_{h+1}^k$, resulting in a recursive relation for the total regret. It is important to upper-bound $\sum_{k=1}^K \delta_h^k$ by $\sum_{k=1}^K \phi_h^k$ and $\sum_{k=1}^K \phi_{h+1}^k$ by $\sum_{k=1}^K \delta_{h+1}^k$. The former follows by Lemma 1.3 and the Bellman equation while the latter follows from the definition of δ and ϕ ($\phi_{h+1}^k \leq \delta_{h+1}^k$). ϕ is associated with the optimal value and δ pertains to the regret associated by choosing a certain action at step k which may not be the optimal one . Lemma 1.3 is used to introduce probability and complexity terms (p, H, ι) into the upper-bound δ equation for $Regret(K)$. Finally, applying recursion and the pigeon-hole principle proves the theorem and upper bounds the total regret by $\mathcal{O}(\sqrt{H^4SAT\iota})$.



Key Takeaways

1. Use UCB over ε-greedy to facilitate exploration in the model-free setting.
2. Use dynamic learning rates $\alpha_t = \mathcal{O}(H/t)$ such as $\frac{H+1}{H+t}$ instead of the commonly used $1/t$ for updates at time step t. This applies more weight to more recent updates and is critical for sample-efficiency guarantees.

References

[1] Chi Jin et al. "Is Q-learning Provably Efficient?" In: *NeurIPS*. 2018.

NOTE: we drop most of the citations due to space considerations.