Gap-Dependent Bounds for Q-Learning using Reference-Advantage Decomposition

Zhong Zheng, Haochen Zhang and Lingzhou Xue

Department of Statistics,
the Pennsylvania State University

April 26, 2025

Motivation

In model-free tabular episodic Markov Decision Processes, several algorithms have been developed, such as UCB-Hoeffding, UCB-Advantage, and Q-EarlySettled-Advantage. While the latter two algorithms successfully use upper confidence bounds and reference-advantage decomposition to achieve near-optimal regret bounds, gap-dependent results have only been established for UCB-Hoeffding.

Motivation

In model-free tabular episodic Markov Decision Processes, several algorithms have been developed, such as UCB-Hoeffding, UCB-Advantage, and Q-EarlySettled-Advantage. While the latter two algorithms successfully use upper confidence bounds and reference-advantage decomposition to achieve near-optimal regret bounds, gap-dependent results have only been established for UCB-Hoeffding.

Question: Can we improve the gap-dependent regret bound for Q-learning by incorporating variance estimators in the bonuses and leveraging reference-advantage decomposition?

Problem Formulation

- **Tabular episodic Markov Decision Process (MDP)** In a tabular episodic MDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$:
  - $\mathcal{S}$: state space, $\mathcal{A}$: action space, H: number of steps.
  - $\mathbb{P} := \{\mathbb{P}_h\}_{h=1}^{H}$ is the time-inhomogeneous transition kernel.
  - $r := \{r_h\}_{h=1}^{H}$ is the collection of reward functions.

Problem Formulation

- **Tabular episodic Markov Decision Process (MDP)** In a tabular episodic MDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$:
  - $\mathcal{S}$: state space, $\mathcal{A}$: action space, H: number of steps.
  - $\mathbb{P} := \{\mathbb{P}_h\}_{h=1}^H$ is the time-inhomogeneous transition kernel.
  - $r := \{r_h\}_{h=1}^H$ is the collection of reward functions.
- **Policy and Value Functions**
  - A policy π is a collection of H functions $\left\{\pi_h : \mathcal{S} \to \Delta^{\mathcal{A}}\right\}_{h \in [H]}$, where $\Delta^{\mathcal{A}}$ is the set of probability distributions over $\mathcal{A}$.

Problem Formulation

- **Tabular episodic Markov Decision Process (MDP)** In a tabular episodic MDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$:
  - $\mathcal{S}$: state space, $\mathcal{A}$: action space, H: number of steps.
  - $\mathbb{P} := \{\mathbb{P}_h\}_{h=1}^H$ is the time-inhomogeneous transition kernel.
  - $r := \{r_h\}_{h=1}^H$ is the collection of reward functions.
- **Policy and Value Functions**
  - A policy π is a collection of H functions $\left\{\pi_h : \mathcal{S} \to \Delta^{\mathcal{A}}\right\}_{h\in[H]}$, where $\Delta^{\mathcal{A}}$ is the set of probability distributions over $\mathcal{A}$.
  - We use $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $V_h^\pi : \mathcal{S} \to \mathbb{R}$ to denote the state-action value function and the state value function at step h under policy π.

$$Q_h^\pi(s, a) := r_h(s, a) + \sum_{h'=h+1}^{H} \mathbb{E}_{(s_{h'}, a_{h'})\sim(\mathbb{P},\pi)}\left[r_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a\right].$$

$$V_h^\pi(s) := \sum_{h'=h}^{H} \mathbb{E}_{(s_{h'}, a_{h'})\sim(\mathbb{P},\pi)}\left[r_{h'}(s_{h'}, a_{h'}) \mid s_h = s\right].$$

Problem Formulation

There always exists an optimal policy $\pi^\star$ for all states and steps. In detail, it achieves the optimal value function $V_h^\star(s) = V_h^{\pi^*}(s) = \sup_\pi V_h^\pi(s)$ and $Q_h^\star(s,a) = Q_h^{\pi^*}(s,a) = \sup_\pi Q_h^\pi(s,a)$ for all $s \in \mathcal{S}$ and $h \in [H]$.

Problem Formulation

There always exists an optimal policy $\pi^\star$ for all states and steps. In detail, it achieves the optimal value function $V_h^\star(s) = V_h^{\pi^*}(s) = \sup_\pi V_h^\pi(s)$ and $Q_h^\star(s,a) = Q_h^{\pi^*}(s,a) = \sup_\pi Q_h^\pi(s,a)$ for all $s \in \mathcal{S}$ and $h \in [H]$.

- **Suboptimality Gap**
  - **Suboptimality gap** $\Delta_h(s,a) := V_h^\star(s) - Q_h^\star(s,a), \forall(s,a,h)$.
  - **Minimum gap** $\Delta_{\min} := \inf\{\Delta_h(s,a) : \Delta_h(s,a) > 0, \forall(s,a,h)\}$.
- **Maximum conditional variance** $\mathbb{Q}^\star := \max_{s,a,h}\{\mathbb{V}_{s,a,h}(V_{h+1}^\star)\}$.

Key techniques: surrogate reference functions

Key techniques: surrogate reference functions

Key idea of the reference-advantage decomposition:
For any (s, h), we expect to maintain a collection of non-increasing reference values $\{V_h^{\mathrm{R},k}(s)\}_{s,k,h}$, which form reasonable estimates of $\{V_h^\star(s)\}_{s,h}$. Our goal is to ensure that, for the final value of the reference function $V_h^{\mathrm{R},K+1}(s)$ and some predefined parameter β, it holds

$$|V_h^{\mathrm{R},K+1}(s) - V_h^\star(s)| \leq \beta.$$

Key techniques: surrogate reference functions

Key idea of the reference-advantage decomposition:
For any (s, h), we expect to maintain a collection of non-increasing reference values $\{V_h^{\mathrm{R},k}(s)\}_{s,k,h}$, which form reasonable estimates of $\{V_h^\star(s)\}_{s,h}$. Our goal is to ensure that, for the final value of the reference function $V_h^{\mathrm{R},K+1}(s)$ and some predefined parameter β, it holds

$$|V_h^{\mathrm{R},K+1}(s) - V_h^\star(s)| \leq \beta.$$

Problem:
The sum of differences involving $V_h^{\mathrm{R},K+1}(s)$ has a non-martingale issue and cannot be bounded directly by concentration inequalities.

Key techniques: surrogate reference functions

Solution: We propose our surrogate reference functions $\hat{V}_h^{\mathrm{R},k}(s)$. They are defined as follows:

$$\hat{V}_h^{\mathrm{R},k}(s) := \max\left\{V_h^{\star}(s), \min\{V_h^{\star}(s)+\beta, V_h^{\mathrm{R},k}(s)\}\right\}, \forall (s,h,k).$$

Key techniques: surrogate reference functions

Solution: We propose our surrogate reference functions $\hat{V}_h^{\mathrm{R},k}(s)$. They are defined as follows:

$$\hat{V}_h^{\mathrm{R},k}(s) := \max\left\{V_h^\star(s), \min\{V_h^\star(s)+\beta, V_h^{\mathrm{R},k}(s)\}\right\}, \forall(s,h,k).$$

It is adaptive to the learning process and has the same property as $V_h^{\mathrm{R},K+1}(s)$:

$$|\hat{V}_h^{\mathrm{R},k}(s) - V_h^\star(s)| \leq \beta.$$

Main Results

Main Results

Define

$$\text{Regret} = \sum_{\text{all episodes } e} \left(V_1^\star(s_{1,e}) - V_1^{\pi_e}(s_{1,e})\right),$$

where $s_{1,e}$ is the initial state for the episode e.

Theorem (Regret of UCB-Advantage)

For UCB-Advantage algorithm with $\beta \in (0, H]$, we have

$$\mathbb{E}[\text{Regret}(T)] \le O\left(\frac{(\mathbb{Q}^\star + \beta^2 H)\, H^3 SA \log(SAT)}{\Delta_{\min}} + \frac{H^8 S^2 A \log(SAT) \log(T)}{\beta^2}\right).$$

Theorem (Regret of Q-EarlySettled-Advantage)

For Q-EarlySettled-Advantage algorithm with $\beta \in (0, H]$, we have

$$\mathbb{E}[\text{Regret}(T)] \le O\left(\frac{(\mathbb{Q}^\star + \beta^2 H)\, H^3 SA \log(SAT)}{\Delta_{\min}} + \frac{H^7 SA \log^2(SAT)}{\beta^2}\right).$$

Main Results

Our gap-dependent bounds are better than it for UCB-Hoeffding:

- Under the worst-case $\mathbb{Q}^* = \Theta(H^2)$ and setting $\beta = O(1/\sqrt{H})$ or $\beta = O(1)$ as in UCB-Advantage and Q-EarlySettled-Advantage algorithms, the upper bounds becomes $\tilde{O}(H^5 SA/\Delta_{\min})$, which is a factor of H better.

Main Results

Our gap-dependent bounds are better than it for UCB-Hoeffding:

- Under the worst-case $\mathbb{Q}^* = \Theta(H^2)$ and setting $\beta = O(1/\sqrt{H})$ or $\beta = O(1)$ as in UCB-Advantage and Q-EarlySettled-Advantage algorithms, the upper bounds becomes $\tilde{O}(H^5 SA/\Delta_{\min})$, which is a factor of H better.
- Under the best variance $\mathbb{Q}^* = 0$ which will happen when the MDP is deterministic, our regret bound can linearly depend on $\tilde{O}(\Delta_{\min}^{-1/3})$, which is intrinsically better than the dependency on $\Delta_{\min}^{-1}$.

Main Results

Our gap-dependent bounds are better than it for UCB-Hoeffding:

- Under the worst-case $\mathbb{Q}^* = \Theta(H^2)$ and setting $\beta = O(1/\sqrt{H})$ or $\beta = O(1)$ as in UCB-Advantage and Q-EarlySettled-Advantage algorithms, the upper bounds becomes $\tilde{O}(H^5 SA/\Delta_{\text{min}})$, which is a factor of H better.
- Under the best variance $\mathbb{Q}^* = 0$ which will happen when the MDP is deterministic, our regret bound can linearly depend on $\tilde{O}(\Delta_{\text{min}}^{-1/3})$, which is intrinsically better than the dependency on Δ_{min}^{-1}.

Our results also provide new guidance on setting the hyper-parameter β when we have prior knowledge about the minimum gap Δ_{min}.

Main Results

The Policy Switching Cost for K episodes is defined as:

$$N_{\text{switch}} = \sum_{k=1}^{K-1} \tilde{N}_{\text{switch}}(\pi^{k+1}, \pi^k).$$

Here, $\tilde{N}_{\text{switch}}(\pi^{k+1}, \pi^k) := \sum_{s \in \mathcal{S}} \sum_{h=1}^{H} \mathbb{I}[\pi_h^{k+1}(s) \neq \pi_h^k(s)]$.

Theorem (Policy switching cost of UCB-Advantage)

For UCB-Advantage with $\beta \in (0, H]$ and any $\delta \in (0,1)$, with probability at least $1-\delta$, the policy switching cost is upper bounded by

$$O\left(H|D_{\text{opt}}| \log\left(\frac{T}{H|D_{\text{opt}}|} + 1\right) + H|D_{\text{opt}}^c| \log\left(\frac{H^4 S A^{\frac{1}{2}} \log(\frac{SAT}{\delta})}{\beta\sqrt{|D_{\text{opt}}^c|}\Delta_{\min}}\right)\right).$$

Here, $D_{\text{opt}} = \{(s,a,h) \mid a \in \mathcal{A}_h^\star(s)\}$, where $\mathcal{A}_h^*(s) = \{a \mid a = \arg\max_{a'} Q_h^*(s,a')\}$.

The Policy Switching Cost for K episodes is defined as:

$$N_{\text{switch}} = \sum_{k=1}^{K-1} \tilde{N}_{\text{switch}}(\pi^{k+1}, \pi^{k}).$$

Here, $\tilde{N}_{\text{switch}}(\pi^{k+1}, \pi^{k}) := \sum_{s \in \mathcal{S}} \sum_{h=1}^{H} \mathbb{I}[\pi_h^{k+1}(s) \neq \pi_h^{k}(s)]$.

Theorem (Policy switching cost of UCB-Advantage)

For UCB-Advantage with $\beta \in (0, H]$ and any $\delta \in (0,1)$, with probability at least $1-\delta$, the policy switching cost is upper bounded by

$$O\left(H|D_{\text{opt}}| \log\left(\frac{T}{H|D_{\text{opt}}|} + 1\right) + H|D_{\text{opt}}^c| \log\left(\frac{H^4 S A^{\frac{1}{2}} \log(\frac{SAT}{\delta})}{\beta\sqrt{|D_{\text{opt}}^c|}\Delta_{\min}}\right)\right).$$

Here, $D_{\text{opt}} = \{(s,a,h) \mid a \in \mathcal{A}_h^{\star}(s)\}$, where $\mathcal{A}_h^{*}(s) = \{a \mid a = \arg\max_{a'} Q_h^{*}(s,a')\}$.

It is the first gap-dependent upper bound for policy switching cost.

Thank you for listening!

Scan for more details about our paper