Action-gap in Reinforcement Learning

Ziniu Li

October 11, 2022

The Chinese University of Hong Kong, Shenzhen

Table of contents

Introduction

Markov Decision Processes

- Infinite-horizon MDPs with time-independent dynamics $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \gamma, P, R)$.
- Bellman Optimality Equation:

$$Q^\star(x, a) = R(x, a) + \gamma \mathbb{E}_{x' \sim P(\cdot|x,a)} \left[\max_{a' \in \mathcal{A}} Q^\star(x', a') \right], \quad \forall (x, a) \in \mathcal{X} \times \mathcal{A}.$$

- Bellman operator $\mathcal{T}$:

$$\mathcal{T}(Q)(x, a) = R(x, a) + \gamma \mathbb{E}_{x' \sim P(\cdot|x,a)} \left[\max_{a'} Q(x', a') \right].$$

However, in practice, we do not know P so that $\mathcal{T}$ is not applicable.

- γ-contractility ($0 < \gamma < 1$):

$$\max_{(x,a)} |\mathcal{T}(Q_1)(x, a) - \mathcal{T}(Q_2)(x, a)| \leq \gamma \max_{(x,a)} |Q_1(x, a) - Q_2(x, a)|.$$

Guarantee of Approximate Value Function

(Proposition 1) Worst-case Guarantee of Approximate Value Function

Suppose a state-value function $\widehat{V}$ satisfies $\|\widehat{V} - V^\star\|_\infty \leq \varepsilon$ for some $\varepsilon \geq 0$. If $\widehat{\pi}$ is a greedy policy based on $\widehat{V}$, then

$$\left\| V^{\widehat{\pi}} - V^\star \right\|_\infty \leq \frac{2\gamma\varepsilon}{1-\gamma}.$$

Remark: even though the value function $\widehat{V}$ is close to $V^\star$, the induced greedy policy $\widehat{\pi}$ may suffer compounding errors in the worst-case.

Proof of Proposition 1

We use the Bellman operator $\mathcal{T}$ and $\mathcal{T}_{\widehat{\pi}}$ defined as

$$(\mathcal{T}V)(x) = \max_a \sum_{x'} P(x'|x,a)\left(r(x,a) + \gamma V(x')\right).$$

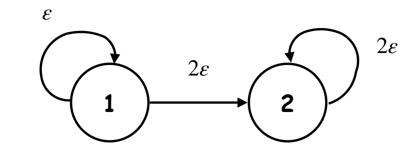
$$(\mathcal{T}^{\widehat{\pi}}V)(x) = \sum_{x'} P(x'|x,\widehat{\pi}(x))\left(r(x,\widehat{\pi}(x)) + \gamma V(x')\right).$$

Then, we have

$$\begin{aligned}
\left\|V^{\widehat{\pi}} - V^\star\right\|_\infty = \left\|\mathcal{T}^{\widehat{\pi}}V^{\widehat{\pi}} - \mathcal{T}V^\star\right\|_\infty &\leq \left\|\mathcal{T}^{\widehat{\pi}}V^{\widehat{\pi}} - \mathcal{T}^{\widehat{\pi}}\widehat{V}\right\|_\infty + \left\|\mathcal{T}^{\widehat{\pi}}\widehat{V} - \mathcal{T}V^\star\right\|_\infty \\
&\leq \gamma\left\|V^{\widehat{\pi}} - \widehat{V}\right\|_\infty + \left\|\mathcal{T}^{\widehat{\pi}}\widehat{V} - \mathcal{T}V^\star\right\|_\infty \\
&= \gamma\left\|V^{\widehat{\pi}} - \widehat{V}\right\|_\infty + \left\|\mathcal{T}\widehat{V} - \mathcal{T}V^\star\right\|_\infty \\
&\leq \gamma\left\|V^{\widehat{\pi}} - \widehat{V}\right\|_\infty + \gamma\left\|\widehat{V} - V^\star\right\|_\infty \\
&\leq \left[\gamma\left\|V^{\widehat{\pi}} - V^\star\right\|_\infty + \gamma\left\|V^\star - \widehat{V}\right\|_\infty\right] + \gamma\left\|\widehat{V} - V^\star\right\|_\infty \\
&\leq \gamma\left\|V^{\widehat{\pi}} - V^\star\right\|_\infty + 2\gamma\left\|\widehat{V} - V^\star\right\|_\infty.
\end{aligned}$$

Rearranging yields the desired result.

Worst-case for Proposition 1



We have

$$V^\star(1) = \frac{2\varepsilon}{1-\gamma} \quad \text{and} \quad V^\star(2) = \frac{2\varepsilon}{1-\gamma}$$

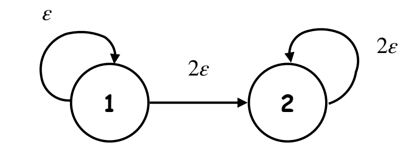
$$\widehat{V}(1) = \frac{2\varepsilon}{1-\gamma} + \varepsilon \quad \text{and} \quad \widehat{V}(2) = \frac{2\varepsilon}{1-\gamma} - \varepsilon$$

The agent always picks the sub-optimal action on state 1 because

$$r(1, a_\varepsilon) + \gamma \widehat{V}(1) = \frac{2}{1-\gamma}\varepsilon$$

$$r(1, a_{2\varepsilon}) + \gamma \widehat{V}(2) = \frac{2-\gamma(1-\gamma)}{1-\gamma}\varepsilon.$$

Worst-case for Proposition 1



- Summary of intuition: in the worst-case, the greedy policy fail to identify the optimal action due to a small gap between two actions.
- However, this worst-case is ε-dependent. Real applications have fixed (and potentially large) action gaps.

Action-gap Theory

Action-gap Theory

Define the action gap function $g_{Q^\star} : \mathcal{X} \to \mathbb{R}$ as

$$g_{Q^\star}(x) \triangleq |Q^\star(x,1) - Q^\star(x,2)|\,.$$

(Assumption 1)

For a fixed MDP $(\mathcal{X}, \mathcal{A}, P, \mathcal{R}, \gamma)$ with $|\mathcal{A}| = 2$, there exist constants $c_g > 0$ and $\zeta \geq 0$ such that for all $t > 0$, we have

$$\mathbb{P}_{\rho^\star}\left(0 < g_{Q^\star}(X) \leq t\right) = \int_{\mathcal{X}} \mathbf{1}\left\{0 < g_{Q^\star}(x) \leq t\right\} d^{\rho^\star(x)} \leq c_g t^\zeta.$$

(Definition 1) Concentrability of the Future-State Distribution

Given $\rho, \rho^\star \in \mathcal{M}(\mathcal{X})$, a policy π, and an integer $m \geq 0$, let $\rho(P^\pi)^m \in \mathcal{M}(\mathcal{X})$ denote the future-state distribution obtained when the first state is distributed according to ρ and we follow the policy π for m steps. Denote the supremum of the Radon-Nikodym derivative of $\rho(P^\pi)^m$ w.r.t. $\rho^\star$ by $c(m, \pi)$, i.e.,

$$c(m;\pi) \triangleq \left\| \frac{d(\rho(P^\pi)^m)}{d\rho^\star} \right\|_\infty .$$

If $\rho(P^\pi)^m$ is not absolutely continuous w.r.t. $\rho^\star$, we set $c(m;\pi) = \infty$. The concentrability of the future-state distribution coefficient is defined as

$$C(\rho, \rho^\star) \triangleq \sup_\pi \sum_{m \geq 0} \gamma^m c(m;\pi).$$

Action-gap Theory

(Theorem 1) Action-gap dependent bound

Consider an MDP $(\mathcal{X}, \mathcal{A}, P, \mathcal{R}, \gamma)$ with $|\mathcal{A}| = 2$ and an estimate $\widehat{Q}$ of the optimal action-value function. Let Assumption 1 hold and $C(\rho, \rho^\star) < \infty$. Denote $\widehat{\pi}$ as the greedy policy w.r.t. $\widehat{Q}$. We then have

$$\left\|V^\star - V(\widehat{\pi})\right\|_\rho \leq \begin{cases} 2^{1+\zeta} c_g C(\rho, \rho^\star) \left\|\widehat{Q} - Q^\star\right\|_\infty^{1+\zeta} \\ 2^{1+\frac{p(1+\zeta)}{p+\zeta}} c_g^{\frac{p-1}{p+\zeta}} C(\rho, \rho^\star) \left\|\widehat{Q} - Q^\star\right\|_{p,\rho^\star}^{\frac{p(1+\zeta)}{p+\zeta}} & (1 \leq p < \infty) \end{cases}$$

Proof of Theorem 1

Let function $F : \mathcal{X} \to \mathbb{R}$ be defined as

$$F(x) = V^\star(x) - V^{\widehat{\pi}}(x) = Q^{\pi^\star}(x, \pi^\star(x)) - Q^{\widehat{\pi}}(x, \widehat{\pi}(x)).$$

Note that $\|V^\star - V(\widehat{\pi})\|_\rho = \rho F$ (i.e., the inner production between two vectors). Decompose $F(x)$ as

$$F(x) = \underbrace{\left(Q^{\pi^\star}(x, \pi^\star(x)) - Q^{\pi^\star}(x, \widehat{\pi}(x))\right)}_{F_1(x)} + \underbrace{\left(Q^{\pi^\star}(x, \widehat{\pi}(x)) - Q^{\widehat{\pi}}(x, \widehat{\pi}(x))\right)}_{F_2(x)}.$$

For $F_2(x)$, we further have

$$\begin{aligned} F_2(x) &= \left[r(x, \widehat{\pi}(x)) + \gamma \int_{\mathcal{X}} P(dy|x, \widehat{\pi}(x)) Q^{\pi^\star}(y, \pi^\star(y))\right] \\ &\quad - \left[r(x, \widehat{\pi}(x)) + \gamma \int_{\mathcal{X}} P(dy|x, \widehat{\pi}(x)) Q^{\widehat{\pi}}(y, \pi^\star(y))\right] \\ &= \gamma P^{\widehat{\pi}}(\cdot|x) F(\cdot). \end{aligned}$$

Therefore, we obtain

$$F = (I - \gamma P^{\widehat{\pi}})^{-1} F_1 = \sum_{m \geq 0} (\gamma P^{\widehat{\pi}})^m F_1.$$

Thus,

$$\begin{aligned}
\rho F = \sum_{m \geq 0} \rho(\gamma P^{\widehat{\pi}})^m F_1 &= \sum_{m \geq 0} \gamma^m \int_{\mathcal{X}} \left(\rho(P^{\widehat{\pi}})^m \right)(dy) F_1(y) \\
&= \sum_{m \geq 0} \gamma^m \int_{\mathcal{X}} \frac{d(\rho(P^{\widehat{\pi}})^m)}{d\rho^\star}(y) d\rho^\star(y) F_1(y) \\
&\leq \sum_{m \geq 0} \gamma^m c(m; \widehat{\pi}) \rho^\star F_1 \leq C(\rho, \rho^\star) \rho^\star F_1.
\end{aligned}$$

Proof of Theorem 1

Claim: Note that for any given $x \in \mathcal{X}$, if for some value $\varepsilon > 0$, we have $\widehat{\pi}(x) \neq \pi^\star(x)$ and $|Q^{\pi^\star}(x, a) - \widehat{Q}(x, a)| \leq \varepsilon$ (for both $a = 1, 2$), then it holds that $g_{Q^\star}(x) = |Q^{\pi^\star}(x, 1) - Q^{\pi^\star}(x, 2)| \leq 2\varepsilon$.

Proof of Claim: suppose that instead $g_{Q^\star}(x) = |Q^{\pi^\star}(x, 1) - Q^{\pi^\star}(x, 2)| > 2\varepsilon$. Then because of the assumption $\widehat{\pi}(x) \neq \pi^\star(x)$ and $|Q^{\pi^\star}(x, a) - \widehat{Q}(x, a)| \leq \varepsilon$ (for both $a = 1, 2$), the ordering of $\widehat{Q}(x, 1)$ and $\widehat{Q}(x, 2)$ is the same as the ordering of $Q^\star(x, 1)$ and $Q^\star(x, 2)$, which contradicts the assumption that $\widehat{\pi}(x) = \pi^\star(x)$.

Proof of Theorem 1

Denote $\varepsilon_0 = \|Q^{\pi^\star} - \widehat{Q}\|_\infty$. Whenever $\widehat{\pi} = \pi^\star(x)$, the value of $F_1(x)$ is zero, so we get

$$\begin{aligned} F_1(x) &= \left[Q^{\pi^\star}(x, \pi^\star(x)) - Q^{\pi^\star}(x, \widehat{\pi}(x))\right] \left[\mathbf{1}\{\widehat{\pi}(x) = \pi^\star(x)\} + \mathbf{1}\{\widehat{\pi}(x) \neq \pi^\star(x)\}\right] \\ &= \left[Q^{\pi^\star}(x, \pi^\star(x)) - Q^{\pi^\star}(x, 1 - \pi^\star(x))\right] \mathbf{1}\{\widehat{\pi}(x) \neq \pi^\star(x)\} \\ &\quad \times \left[\mathbf{1}\{g_{Q^\star}(x) = 0\} + \mathbf{1}\{0 < g_{Q^\star}(x) \leq 2\varepsilon_0\} + \mathbf{1}\{g_{Q^\star}(x) > 2\varepsilon_0\}\right] \\ &\leq 0 + 2\varepsilon_0 \mathbf{1}\{0 < g_{Q^\star}(x) \leq 2\varepsilon_0\} + 0. \end{aligned} \tag{1}$$

This result together with Assumption 1 shows that $\rho^\star F_1 \leq 2\varepsilon_0 \mathbb{P}_{\rho^\star}(0 < g_{Q^\star}(x) \leq 2\varepsilon_0) \leq 2\varepsilon_0 c_g (2\varepsilon_0)^\xi$.

Frame Title

References