EN.520.637: Foundations of Reinforcement Learning Homework 4

Ha Manh Bui (CS Department)
hbui13@jhu.edu

Fall 2023

1 Problem 1

Given a finite MDP with optimal policy π^* and the corresponding optimal action value function $q^*(s,a)$. Let $q(s,a)$ be another state-action value function with the greedy policy of q is given by $\pi_q(s) = \arg\max_a q(s,a)$. Let $0 < \gamma < 1$ be the discounting factor of the reward. Here we assume π^* and π_q are deterministic polices. We let $||q||_\infty := \max_{(s,a)} |q(s,a)|$ and $||v||_\infty := \max_s |v(s)|$.

(a) Suppose for the reward, we have $|r| \le r_{\max}, \forall r \in \mathbb{R}$, show that for any policy π,

$$||q_\pi||_\infty \le \frac{1}{1-\gamma} r_{\max}.$$

Proof. By definition, for any policy π, we have

$$q_\pi(s,a) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{k+1} | S_0 = s, A_0 = a \right].$$

Due to $|r| \le r_{\max}, \forall r \in \mathbb{R}$, we get

$$q_\pi(s,a) \le \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{\max} \right] = r_{\max} \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k \right].$$

Since $0 < \gamma < 1$, we achieve

$$q_\pi(s,a) \le \lim_{k \to \infty} \frac{1-\gamma^k}{1-\gamma} r_{\max} = \frac{1}{1-\gamma} r_{\max}. \tag{1}$$

On the other hand, since $||q_\pi||_\infty := \max_{(s,a)} |q_\pi(s,a)|$, combining with the result from Eq. 1, $\forall \pi(a|s)$, we obtain

$$||q_\pi||_\infty \le \frac{1}{1-\gamma} r_{\max}. \tag{2}$$

□

(b) Let $v^*(s)$ be the optimal value function and $v_{\pi_q}(s)$ be the state value function under policy π_q. (i) Show that $\forall s \in \mathcal{S}$,

$$v^*(s) - v_{\pi_q}(s) = q^*(s, \pi^*(s)) - q^*(s, \pi_q(s)) + \gamma \sum_{s'} p(s'|s, \pi_q(s)) \left[v^*(s') - v_{\pi_q}(s') \right]$$

Proof. Since π^* and $\pi_q(s) = \arg\max_a q(s,a)$ are deterministic policies, $\forall s \in \mathcal{S}$, we have

$$\begin{aligned} v^*(s) - v_{\pi_q}(s) &= \max_a q^*(s,a) - \max_a q(s,a) \\ &= q^*(s, \pi^*(s)) - q(s, \pi_q(s)) \\ &= q^*(s, \pi^*(s)) - q^*(s, \pi_q(s)) + q^*(s, \pi_q(s)) - q(s, \pi_q(s)). \end{aligned} \tag{3}$$

Consider the term $q^*(s, \pi_q(s)) - q(s, \pi_q(s))$, by definition of the Q-function, we have

$$\begin{aligned} q^*(s, \pi_q(s)) - q(s, \pi_q(s)) &= \mathbb{E}_\pi \left[R_{t+1} + \gamma v^*(S_{t+1}) | S_t = s, A_t = \pi_q(s)\right] - \mathbb{E}_\pi \left[R_{t+1} + \gamma v_{\pi_q}(S_{t+1}) | S_t = s, A_t = \pi_q(s)\right] \\ &= r(s, \pi_q(s)) + \gamma \sum_{s'} p(s'|s, \pi_q(s)) v^*(s') - r(s, \pi_q(s)) - \gamma \sum_{s'} p(s'|s, \pi_q(s)) v_{\pi_q}(s') \\ &= \gamma \sum_{s'} p(s'|s, \pi_q(s)) \left[v^*(s') - v_{\pi_q}(s')\right]. \end{aligned}$$

Replace this result to Eq. 3, $\forall s \in \mathcal{S}$, we obtain

$$v^*(s) - v_{\pi_q}(s) = q^*(s, \pi^*(s)) - q^*(s, \pi_q(s)) + \gamma \sum_{s'} p(s'|s, \pi_q(s)) \left[v^*(s') - v_{\pi_q}(s')\right]. \tag{4}$$

$\square$

(ii) Show that $\forall s \in \mathcal{S}$

$$v^*(s) - v_{\pi_q}(s) \leq 2||q - q^*||_\infty + \gamma ||v_{\pi_q} - v^*||_\infty$$

Proof. Firstly, let consider the first term $q^*(s, \pi^*(s)) - q^*(s, \pi_q(s))$ in Eq. 4 we have

$$q^*(s, \pi^*(s)) - q^*(s, \pi_q(s)) = q^*(s, \pi^*(s)) - q(s, \pi^*(s)) + q(s, \pi^*(s)) - q^*(s, \pi_q(s)).$$

Since $\pi_q(s) = \arg\max_a q(s,a)$, we get

$$\begin{aligned} & q(s, \pi_q(s)) \geq q(s,a), \forall a \in \mathcal{A}, s \in \mathcal{S} \\ \Rightarrow \quad & q(s, \pi_q(s)) \geq q(s, \pi^*(s)), \forall s \in \mathcal{S}. \end{aligned}$$

This yields

$$\begin{aligned} q^*(s, \pi^*(s)) - q^*(s, \pi_q(s)) &= q^*(s, \pi^*(s)) - q(s, \pi^*(s)) + q(s, \pi^*(s)) - q^*(s, \pi_q(s)) \\ &\leq q^*(s, \pi^*(s)) - q(s, \pi^*(s)) + q(s, \pi_q(s)) - q^*(s, \pi_q(s)) \\ &\leq \max_s |q^*(s, \pi^*(s)) - q(s, \pi^*(s))| + \max_s |q(s, \pi_q(s)) - q^*(s, \pi_q(s))|. \end{aligned}$$

By $||q||_\infty = \max_{(s,a)} |q(s,a)|$, we obtain

$$q^*(s, \pi^*(s)) - q^*(s, \pi_q(s)) \leq 2||q - q^*||_\infty. \tag{5}$$

On the other hand, consider the second term $\gamma \sum_{s'} p(s'|s, \pi_q(s)) \left[v^*(s') - v_{\pi_q}(s')\right]$ in Eq. 4, since $||v||_\infty = \max_s |v(s)|$, we have

$$\begin{aligned} \gamma \sum_{s'} p(s'|s, \pi_q(s)) \left[v^*(s') - v_{\pi_q}(s')\right] &\leq \gamma \sum_{s'} p(s'|s, \pi_q(s)) \left[\max_{s'} |v^*(s') - v_{\pi_q}(s')|\right] \\ &\leq \gamma ||v_{\pi_q} - v^*||_\infty. \end{aligned} \tag{6}$$

Replace results in Eq. 5 and Eq. 6 to Eq. 4, $\forall s \in \mathcal{S}$, we obtain

$$v^*(s) - v_{\pi_q}(s) \leq 2||q - q^*||_\infty + \gamma ||v_{\pi_q} - v^*||_\infty. \tag{7}$$

$\square$

(iii) Show that

$$||v^* - v_{\pi_q}||_\infty \leq \frac{2||q - q^*||_\infty}{1-\gamma}.$$

In other word, when we obtain a q that is "close" to the optimal q^*, the value function of its greedy policy π_q is also "close" to the optimal value function.

Proof. From the result in Eq. 7, we have

$$v^*(s) - v_{\pi_q}(s) \leq 2||q - q^*||_\infty + \gamma||v_{\pi_q} - v^*||_\infty,$$

$\forall s \in \mathcal{S}$, this yields

$$\max_{s'} \left[v^*(s') - v_{\pi_q}(s')\right] \leq 2||q - q^*||_\infty + \gamma||v_{\pi_q} - v^*||_\infty.$$

Since $||v||_\infty = \max_s |v(s)|$, we obtain

$$\begin{aligned} & ||v^* - v_{\pi_q}|| \leq 2||q - q^*||_\infty + \gamma||v_{\pi_q} - v^*||_\infty \\ \Leftrightarrow \quad & ||v^* - v_{\pi_q}|| - \gamma||v_{\pi_q} - v^*||_\infty \leq 2||q - q^*||_\infty \\ \Leftrightarrow \quad & ||v^* - v_{\pi_q}||_\infty \leq \frac{2||q - q^*||_\infty}{1-\gamma}. \end{aligned} \tag{8}$$

□

(c) The Value Iteration on state-action value function is given by

$$q^{(k+1)} = \mathcal{T}^q_{\max} q^{(k)}, \quad q^{(0)} = 0,$$

where $\mathcal{T}^q_{\max}$ is the Bellman optimality operator for q, and $q^* = \mathcal{T}^q_{\max} q^*$. Show that when $k \geq \left(\log \frac{1}{\gamma}\right)^{-1} \log \frac{2r_{\max}}{(1-\gamma)^2\epsilon}$, we have

$$||v_{\pi_{q(k)}} - v^*||_\infty \leq \epsilon.$$

Proof. We have

$$\begin{aligned} & k \geq \left(\log \frac{1}{\gamma}\right)^{-1} \log \frac{2r_{\max}}{(1-\gamma)^2\epsilon} \Leftrightarrow k \log \frac{1}{\gamma} \geq \log \frac{2r_{\max}}{(1-\gamma)^2\epsilon} \Leftrightarrow \frac{1}{\gamma^k} \geq \frac{2r_{\max}}{(1-\gamma)^2\epsilon} \\ \Leftrightarrow \quad & \gamma^k \leq \frac{(1-\gamma)^2\epsilon}{2r_{\max}} \Leftrightarrow \gamma^k 2r_{\max} \leq (1-\gamma)^2\epsilon \Leftrightarrow \frac{\gamma^k 2r_{\max}}{(1-\gamma)^2} \leq \epsilon. \end{aligned}$$

Using the result from Eq. 2, i.e., $||q_\pi||_\infty \leq \frac{1}{1-\gamma} r_{\max}$, we get

$$\epsilon \geq \frac{\gamma^k 2r_{\max}}{(1-\gamma)^2} \geq \frac{2\gamma^k}{1-\gamma}||q_\pi - 0||_\infty \geq \frac{2\gamma^k}{1-\gamma}||q^* - q^{(0)}||_\infty.$$

Apply the fact that $\mathcal{T}^q_{\max}$ is γ-contracting w.r.t. $||\cdot||_\infty$, i.e., $||\mathcal{T}^q_{\max} q^* - \mathcal{T}^q_{\max} q^{(0)}||_\infty \leq ||q^* - q^{(0)}||_\infty$, we obtain

$$\epsilon \geq \frac{2\gamma^k}{1-\gamma}||q^* - q^{(0)}||_\infty \geq \frac{2\gamma^k}{1-\gamma}||\mathcal{T}^q_{\max} q^* - \mathcal{T}^q_{\max} q^{(0)}||_\infty.$$

Using the result from Eq. 8, i.e., $||v^* - v_{\pi_q}||_\infty \leq \frac{2||q - q^*||_\infty}{1-\gamma}$, we obtain

$$\epsilon \geq \frac{2\gamma^k}{1-\gamma}||\mathcal{T}^q_{\max} q^* - \mathcal{T}^q_{\max} q^{(0)}||_\infty \geq ||v_{\pi_{q(k)}} - v^*||_\infty.$$

□

2 Problem 2

Given an MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, p, \gamma)$, where $0 < \gamma < 1$ is the discounting factor. Consider a modified/approximate MDP $\hat{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \hat{p}, \gamma)$. The $\hat{p}(s', r|s, a)$ is chosen such that

$$\hat{p}(s'|s,a) = p(s'|s,a), \forall s', s, a,$$

and

$$|\hat{r}(s,a) - r(s,a)| \leq \epsilon, \text{i.e., } \left| \sum_r r\hat{p}(r|s,a) - \sum_r rp(r|s,a) \right| \leq \epsilon, \forall s, a.$$

(a) Let $v^*(s)$ and $\hat{v}^*(s)$ be the optimal value functions of M and $\hat{M}$, respectively. Show that

$$||v^* - \hat{v}^*||_\infty \leq \frac{\epsilon}{1-\gamma}.$$

Proof. By definition of the Q-function, $\forall a \in \mathcal{A}, s \in \mathcal{S}$, we have

$$q^*(s,a) - \hat{q}^*(s,a) = \left[r^*(s,a) + \gamma \sum_{s'} v^*(s')p(s'|s,a) \right] - \left[\hat{r}^*(s,a) + \gamma \sum_{s'} \hat{v}^*(s')\hat{p}(s'|s,a) \right].$$

Using $\hat{p}(s'|s,a) = p(s'|s,a), \forall s', s, a$ and $|\hat{r}(s,a) - r(s,a)| \leq \epsilon, \forall s, a$, we get

$$q^*(s,a) - \hat{q}^*(s,a) \leq \epsilon + \gamma \sum_{s'} p(s'|s,a) \left[v^*(s') - \hat{v}^*(s') \right]. \tag{9}$$

Consider the term $v^*(s) - \hat{v}^*(s), \forall s \in \mathcal{S}$, by definition, we have

$$v^*(s) - \hat{v}^*(s) = q^*(s, \pi^*(s)) - \hat{q}^*(s, \hat{\pi}^*(s)),$$

since $\hat{\pi}^*$ is the estimator of the true optimal π^*, getting

$$v^*(s) - \hat{v}^*(s) \leq q^*(s, \pi^*(s)) - \hat{q}^*(s, \pi^*(s)),$$

applying the result in Eq. 9, yielding

$$\begin{aligned}
v^*(s) - \hat{v}^*(s) &\leq \epsilon + \gamma \sum_{s'} p(s'|s, \pi^*(s)) \left[v^*(s') - \hat{v}^*(s') \right] \\
&\leq \epsilon + \gamma \sum_{s'} p(s'|s, \pi^*(s)) ||v^* - \hat{v}^*||_\infty \\
&\leq \epsilon + \gamma ||v^* - \hat{v}^*||_\infty.
\end{aligned}$$

As a result, we obtain

$$||v^* - \hat{v}^*||_\infty \leq \epsilon + \gamma ||v^* - \hat{v}^*||_\infty,$$

i.e.,

$$||v^* - \hat{v}^*||_\infty \leq \frac{\epsilon}{1-\gamma}.$$

□

(b) Suppose that

$$\hat{r}(s,a) = r(s,a) + \epsilon, \forall s, a.$$

Show that

$$\hat{v}^*(s) = v^*(s) + \frac{\epsilon}{1-\gamma}, \forall s.$$

Proof. By definition, we have

$$\hat{v}^*(s) = \max_\pi \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k \hat{R}_{k+1} | S_0 = s \right],$$

since $\hat{r}(s,a) = r(s,a) + \epsilon, \forall s, a$ and $\hat{p}(s'|s,a) = p(s'|s,a), \forall s', s, a$, getting

$$\begin{aligned}
\hat{v}^*(s) &= \max_\pi \left\{ \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k \left(R_{k+1} + \epsilon \right) | S_0 = s \right] \right\} \\
&= \max_\pi \left\{ \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{k+1} | S_0 = s \right] + \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k \epsilon | S_0 = s \right] \right\},
\end{aligned}$$

combining with $0 < \gamma < 1$, we obtain

$$\begin{aligned}
\hat{v}^*(s) &= \max_\pi \left\{ \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{k+1} | S_0 = s \right] + \epsilon \lim_{k \to \infty} \frac{1 - \gamma^k}{1 - \gamma} \right\} \\
&= \max_\pi \left\{ \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{k+1} | S_0 = s \right] + \frac{\epsilon}{1 - \gamma} \right\} \\
&= \max_\pi \left\{ \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{k+1} | S_0 = s \right] \right\} + \frac{\epsilon}{1 - \gamma} \\
&= v^*(s) + \frac{\epsilon}{1 - \gamma}, \forall s.
\end{aligned}$$

$\square$