MS&E 346 Assignment 4

Junting Duan

January 23, 2022

1

1.1

The action-value function for $k = 1$ is $q_1(s,a) = R(s,a) + \gamma \cdot \sum_{s' \in \mathcal{N}} P(s,a,s') \cdot v_0(s')$. Plugging the values into it, we have

$$q_1(s_1,a_1) = 10.6, \quad q_1(s_1,a_2) = 11.2,$$
$$q_1(s_2,a_1) = 4.3, \quad q_1(s_2,a_2) = 4.3.$$

As a result, we have $v_1(s_1) = B^*(v_0)(s_1) = \max_{a\in\mathcal{A}} q_1(s_1,a) = 11.2$ and $v_1(s_2) = B^*(v_0)(s_2) = \max_{a\in\mathcal{A}} q_1(s_2,a) = 4.3$. The greedy policy $\pi_1(s_1) = \arg\max_{a\in\mathcal{A}} q_1(s_1,a) = a_2$ and $\pi_1(s_2) = \arg\max_{a\in\mathcal{A}} q_1(s_2,a) = a_1$.

For $k = 2$, we update the action-value function as

$$q_2(s_1,a_1) = 12.82, \quad q_2(s_1,a_2) = 11.98,$$
$$q_2(s_2,a_1) = 5.65, \quad q_2(s_2,a_2) = 5.89.$$

Thus, there is $v_2(s_1) = B^*(v_1)(s_1) = \max_{a\in\mathcal{A}} q_2(s_1,a) = 12.82$ and $v_2(s_2) = B^*(v_1)(s_2) = \max_{a\in\mathcal{A}} q_2(s_2,a) = 5.89$. The greedy policy $\pi_2(s_1) = \arg\max_{a\in\mathcal{A}} q_2(s_1,a) = a_1$ and $\pi_2(s_2) = \arg\max_{a\in\mathcal{A}} q_2(s_2,a) = a_2$.

1.2

We just need to prove that $q_k(s_1,a_1) > q_k(s_1,a_2)$ and $q_k(s_2,a_1) < q_k(s_2,a_2)$ for $k > 2$. Observe that

$$\begin{aligned}
&q_k(s_1,a_1) - q_k(s_1,a_2) \\
=&R(s_1,a_1) - R(s_1,a_2) + (P(s_1,a_1,s_1) - P(s_1,a_2,s_1)) \cdot v_{k-1}(s_1) + (P(s_1,a_1,s_2) - P(s_1,a_2,s_2)) \cdot v_{k-1}(s_2) \\
=&-2 + 0.1 * v_{k-1}(s_1) + 0.4 * v_{k-1}(s_2).
\end{aligned}$$

Since $v_k(s) \geq v_{k-1}(s)$, we have $q_k(s_1,a_1) - q_k(s_1,a_2) \geq 1.638 > 0$. Furthermore, we there is

$$\begin{aligned}
&q_k(s_2,a_2) - q_k(s_2,a_1) \\
=&R(s_2,a_2) - R(s_2,a_1) + (P(s_2,a_2,s_1) - P(s_2,a_1,s_1)) \cdot v_{k-1}(s_1) + (P(s_2,a_2,s_2) - P(s_2,a_1,s_2)) \cdot v_{k-1}(s_2) \\
=&-2 + 0.2 * v_{k-1}(s_1) \geq 0.564 > 0.
\end{aligned}$$

Thus we complete our proof.