Assignment 4

Miao-Chin Yen

February 21, 2022

Problem 1 (Manual Value Iteration)

1. Initialize the Value Function for each state to be it's max (over actions) reward, i.e., we initialize the Value Function to be $v_0(s_1) = 10.0, v_0(s_2) = 1.0, v_0(s_3) = 0.0$. Then manually calculate $q_k(\cdot,\cdot)$ and $v_k(\cdot)$ from $v_{k-1}(\cdot)$ using the Value Iteration update, and then calculate the greedy policy $\pi_k(\cdot)$ from $q_k(\cdot,\cdot)$ for $k = 1$ and $k = 2$ (hence, 2 iterations).

$$q_1(s_1, a_1) = \mathcal{R}(s_1, a_1) + \mathcal{P}(s_1, a_1, s_1) \cdot v_0(s_1) + \mathcal{P}(s_1, a_1, s_2) \cdot v_0(s_2) = 10.6$$

$$q_1(s_1, a_2) = \mathcal{R}(s_1, a_2) + \mathcal{P}(s_1, a_2, s_1) \cdot v_0(s_1) + \mathcal{P}(s_1, a_2, s_2) \cdot v_0(s_2) = 11.2$$

$$v_1(s_1) = max(q_1(s_1, a_1), q_1(s_1, a_2)) = 11.2 \Rightarrow \pi_1(s_1) = a_2$$

$$q_1(s_2, a_1) = \mathcal{R}(s_2, a_1) + \mathcal{P}(s_2, a_1, s_1) \cdot v_0(s_1) + \mathcal{P}(s_2, a_1, s_2) \cdot v_0(s_2) = 4.3$$

$$q_1(s_2, a_2) = \mathcal{R}(s_2, a_2) + \mathcal{P}(s_2, a_2, s_1) \cdot v_0(s_1) + \mathcal{P}(s_2, a_2, s_2) \cdot v_0(s_2) = 4.3$$

$$v_1(s_2) = max(q_1(s_2, a_1), q_1(s_2, a_2)) = 4.3 \Rightarrow \pi_1(s_2) = a_1$$

$$q_2(s_1, a_1) = \mathcal{R}(s_1, a_1) + \mathcal{P}(s_1, a_1, s_1) \cdot v_1(s_1) + \mathcal{P}(s_1, a_1, s_2) \cdot v_1(s_2) = 12.82$$

$$q_2(s_1, a_2) = \mathcal{R}(s_1, a_2) + \mathcal{P}(s_1, a_2, s_1) \cdot v_1(s_1) + \mathcal{P}(s_1, a_2, s_2) \cdot v_1(s_2) = 11.98$$

$$v_1(s_1) = max(q_2(s_1, a_1), q_2(s_1, a_2)) = 12.82 \Rightarrow \pi_2(s_1) = a_1$$

$$q_2(s_2, a_1) = \mathcal{R}(s_2, a_1) + \mathcal{P}(s_2, a_1, s_1) \cdot v_1(s_1) + \mathcal{P}(s_2, a_1, s_2) \cdot v_1(s_2) = 5.65$$

$$q_2(s_2, a_2) = \mathcal{R}(s_2, a_2) + \mathcal{P}(s_2, a_2, s_1) \cdot v_1(s_1) + \mathcal{P}(s_2, a_2, s_2) \cdot v_1(s_2) = 5.89$$

$$v_2(s_2) = max(q_2(s_2, a_1), q_2(s_2, a_2)) = 5.89 \Rightarrow \pi_2(s_2) = a_2$$

2. Now argue that $\pi_k(\cdot)$ for $k > 2$ will be the same as $\pi_2(\cdot)$. Hint: You can make the argument by examining the structure of how you get $q_k(\cdot,\cdot)$ from $v_{k-1}(\cdot)$. With this argument, there is no need to go beyond the two iterations you performed above, and so you can establish $\pi_2(\cdot)$ as an Optimal Deterministic Policy for this MDP.

$$q_k(s_1, a_1) - q_k(s_1, a_2)$$
$$= \mathcal{R}(s_1, a_1) - \mathcal{R}(s_1, a_2) + (\mathcal{P}(s_1, a_1, s_1) - \mathcal{P}(s_1, a_2, s_1)) \cdot v_{k-1}(s_1) + (\mathcal{P}(s_1, a_1, s_2) - \mathcal{P}(s_1, a_2, s_2)) \cdot v_{k-1}(s_2)$$
$$= -2.0 + 0.1 \cdot v_{k-1}(s_1) + 0.4 \cdot v_{k-1}(s_2),$$
$$q_k(s_2, a_2) - q_k(s_2, a_1)$$
$$= \mathcal{R}(s_2, a_2) - \mathcal{R}(s_2, a_1) + (\mathcal{P}(s_2, a_2, s_1) - \mathcal{P}(s_2, a_1, s_1)) \cdot v_{k-1}(s_1) + (\mathcal{P}(s_2, a_2, s_2) - \mathcal{P}(s_2, a_1, s_2)) \cdot v_{k-1}(s_2)$$
$$= -2.0 + 0.2 \cdot v_{k-1}(s_1)$$

Because $v_{k-1}(s_1) \geq v_2(s_1)$ and $v_{k-1}(s_2) \geq v_2(s_2) \forall k \geq 3$,

$$q_k(s_1, a_1) - q_k(s_1, a_2) \geq -2.0 + 0.1 \cdot v_2(s_1) + 0.4 \cdot v_2(s_2) > 0 \ \forall k \geq 3$$

$$q_k(s_2, a_2) - q_k(s_2, a_1) \geq -2.0 + 0.2 \cdot v_2(s_1) > 0 \ \forall k \geq 3$$

Hence $q_k(s_1, a_1) > q_k(s_1, a_2)$ and $q_k(s_2, a_2) > q_k(s_2, a_1)$ $\forall k \geq 3 \Rightarrow \pi_k(s_1) = a_1$ and $\pi_k(s_2) = a_2$ $\forall k \geq 3$