Jordan Rowley

# CME 241: Assignment 4

## 4.1

Using backwards induction, we find the state-action function $q(\cdot,\cdot)$ for the inputs $(s_1,a_1),(s_1,a_2),(s_2,a_1),(s_2,a_2)$. Starting with the initial value vector $v_0 = [v_0(s_1), v_0(s_2)] = [10,1]$, we have

$$q_0(s_1,a_1) = 8 + [0.2, 0.6]v_0^T = 10.6$$
$$q)_0(s_1,a_2) = 10 + [0.1, 0.2]v_0^T = 11.2$$
$$q_0(s_2,a_1) = 1 + [0.3, 0.3]v_0^T = 4.3$$
$$q_0(s_2,a_2) = -1 + [0.5, 0.3]v_0^T = 4.3$$

Therefore, we should choose action $a_2$ from $s_1$, and we can choose either action from $s_2$. This gives us $v_1(s_1) = 11.2$ and $v_1(s_2) = 4.3$. We will appeal to the general format of $q(\cdot,\cdot)$ to determine whether monotonicity persists.

$$q_i(s_2,a_2) - q_i(s_2,a_1) = -2 + [0.2, 0.0]v_i^T > 0 \Leftrightarrow 10 > v_i(s_1)$$

After the first iteration, we saw that $v_1(s_1) = 11.2$. Therefore, action $a_2$ will become the best action from state $s_2$. From the other state,

$$q_i(s_1,a_2) - q_i(s_1,a_1) = 2 - 0.1v_i(s_1) - 0.4v_i(s_2) > 0 \Leftrightarrow 2 - 0.4v_i(s_2) > 1 \Leftrightarrow v_i(s_2) < 2.5$$

Since we saw $v_1(s_2) = 4.3$, we have that $q_i(s_1,a_2) < q_i(s_1,a_1)$ indefinitely. Therefore, action $a_1$ ought to be chosen from state $s_1$.

The optimal policy $\pi^*$ is therefore defined by

$$\pi^*(s_1) = a_1$$
$$\pi^*(s_2) = a_2$$