Theoretical Portion of Assignment 3

1. For this problem, we have $\mathcal{S} = \{s_1, s_2, s_3\}$, $\mathcal{T} = \{s_3\}$, and $\mathcal{A} = \{a_1, a_2\}$. Let $v_0(s_1) = 10.0$, $v_0(s_2) = 1.0$, and $v_0(s_3) = 0.0$. Then:

   (a) 1. Iteration 1:
       - $q_1(s_1, a_1) = 8 + 0.2 \times 10 + 0.6 \times 1 = 10.6$
       - $q_1(s_1, a_2) = 10 + 0.1 \times 10 + 0.2 \times 1 = 11.2$
       - $q_1(s_2, a_1) = 1 + 0.3 \times 10 + 0.3 \times 1 = 4.3$
       - $q_1(s_2, a_2) = -1 + 0.5 \times 10 + 0.3 \times 1 = 4.3$
       - $v_1(s_1) = 11.2$
       - $v_1(s_2) = 4.3$
       - $\pi_1(s_1) = a_1$
       - $\pi_1(s_2) = \{a_1, a_2\}$

       2. Iteration 2:
       - $q_2(s_1, a_1) = 8 + 0.2 \times 11.2 + 0.6 \times 4.3 = 12.82$
       - $q_2(s_1, a_2) = 10 + 0.1 \times 11.2 + 0.2 \times 4.3 = 11.98$
       - $q_2(s_2, a_1) = 1 + 0.3 \times 11.2 + 0.3 \times 4.3 = 5.65$
       - $q_2(s_2, a_2) = -1 + 0.5 \times 11.2 + 0.3 \times 4.3 = 5.89$
       - $v_2(s_1) = 12.82$
       - $v_2(s_2) = 5.89$
       - $\pi_2(s_1) = a_1$
       - $\pi_2(s_2) = a_2$

   (b) Since we have only two actions $\{a_1, a_2\}$ to choose from, we only need to compare their differences:
       - For s_1, we evaluate: $q_k(s_1, a_2) - q_k(s_1, a_1) = (10-8) + (0.1-0.2)v_{k-1}(s_1) + (0.2-0.6)v_{k-1}(s_2)$, which is $2 - 0.1v_{k-1}(s_1) - 0.4v_{k-1}(s_2)$.
       - For s_2, we evaluate: $q_k(s_2, a_2) - q_k(s_2, a_1) = (-1-1) + (0.5-0.3)v_{k-1}(s_1) + (0.3-0.3)v_{k-1}(s_2)$, which is $-2 + 0.2v_{k-1}(s_1)$.

       From the above, we can get the following:
       - $q_k(s_2, a_2) > q_k(s_2, a_1)$ when $-2 + 0.2v_{k-1}(s_1) > 0$, or $v_{k-1}(s_1) > 10$;
       - $q_k(s_1, a_2) < q_k(s_1, a_1)$ when $2 - 0.1v_{k-1}(s_1) - 0.4v_{k-1}(s_2) < 0$. If $v_{k-1}(s_1) > 10$, we can have $0.1v_{k-1}(s_1) > 1$. Then, $2 - 0.1v_{k-1}(s_1) - 0.4v_{k-1}(s_2) < 0$ when $0.4v_{k-1}(s_2) > 1$, or $v_{k-1}(s_2) > 2.5$.

       Therefore, we conclude that if $v_{k-1}(s_1) > 10$ and $v_{k-1}(s_2) > 2.5$, a_1 is always a better action than a_2 when we are in s_1, and a_2 is always a better action than a_1 when we are in s_2. By property of Value Iteration, our value function for both states can only get better. This means as soon as we reach a k that satisfies both conditions, the $\pi_k(s_1) = a_1$ and $\pi_k(s_2) = a_2$ will be our Optimal Deterministic Policy, and the deterministic policy for all future k will be the same as current $\pi_k(s_1)$ and $\pi_k(s_2)$. Since $v_1(s_1) = 11.2 > 10$, $v_1(s_2) = 4.3 > 2.5$, we have already satisfies both conditions during the second iteration. Therefore, we can have the following:

- $\pi_k(s_1) = \pi_2(s_1), \pi_k(s_2) = \pi_2(s_2), \forall k > 2;$
- $\pi^*(s_1) = a_1$ and $\pi^*(s_2) = a_2$.