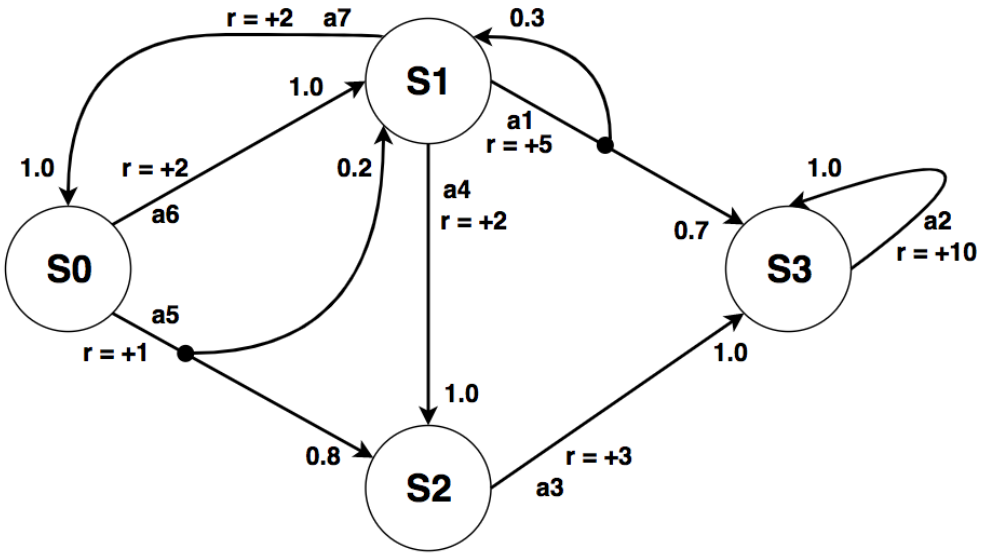
# 1 Markov Decision Processes (20 points)

Consider the MDP given in the figure below. Assume the discount factor $\gamma = 0.9$. The r-values are rewards, while the numbers next to the arrows are outcome probabilities. Note that:

- State $S_0$ has two actions: $a5$ and $a6$;
- state $S_1$ has three actions: $a1$, $a4$, and $a7$;
- State $S_2$ has only one action: $a3$;
- state $S_3$ has only one action: $a2$.



1. *(8 points)* Write down the numerical value of $V_{S1}$ after the first two iterations of Value Iteration. No partial credit.

   Use the initial value function: $V^0_{S_0} = 0, V^0_{S_1} = 0, V^0_{S_2} = 0, V^0_{S_3} = 0$;

   **[Answer:]:**
   $V^1_{S_1} =$

   $V^2_{S_1} =$

   **Solution:** $V^1_{S_1} = 5$

   After the first iteration, $V^1_{S_0} = 2, V^1_{S_1} = 5, V^1_{S_2} = 3, V^1_{S_3} = 10$;

   For action $a_1$: $5 + 0.9 * (0.3 * 5 + 0.7 * 10) = 12.65$
   For action $a_4$: $2 + 0.9 * 3 = 4.7$
   For action $a_7$: $2 + 0.9 * 2 = 3.8$
   $V^2_{S_1} = 12.65$

2. *(12 points)* Suppose you run a "simplified" Policy Iteration. During each iteration, you compute the exact value of each state under the current policy *only ONCE* and then update the policy based on those values. Write down the numerical value of each state, each corresponding policy, and its Expected Cumulative Reward (ECR) after the first two iterations of this "simplified" Policy Iteration. No partial credit.

   The distribution of the initial states is $P(S_0) = 0.2, P(S_1) = 0.3, P(S_2) = 0.3, P(S_3) = 0.2$.

Initial value function: $V_{S_0}^0 = 0, V_{S_1}^0 = 0, V_{S_2}^0 = 0, V_{S_3}^0 = 0$;
Initial Policy $\pi^0$:

(a) At $S_0$, take action $a_5$;

(b) At $S_1$, take action $a_7$;

(c) At $S_2$, take action $a_3$;

(d) At $S_3$, take action $a_2$;

**[Answer:]**

**1st Iteration:**
$V_{S_0}^{\pi^0} =$ , $V_{S_1}^{\pi^0} =$ , $V_{S_2}^{\pi^0} =$ , $V_{S_3}^{\pi^0} =$

Policy $\pi^1$:

ECR of policy $\pi^1$:

**2nd Iteration:**
$V_{S_0}^{\pi^1} =$ , $V_{S_1}^{\pi^1} =$ , $V_{S_2}^{\pi^1} =$ , $V_{S_3}^{\pi^1} =$

Policy $\pi^2$:

ECR of policy $\pi^2$:

**Solution:**

$V^1_{S_0} = 1$, $V^1_{S_1} = 2$, $V^1_{S_2} = 3$, $V^1_{S_3} = 10$

$Q(S_0, a_6) = 2 + 0.9 * 2 = 3.8$
$Q(S_0, a_5) = 1 + 0.9 * (2 * 0.2 + 3 * 0.8) = 3.52$
$Q(S_1, a_7) = 2 + 0.9 * 1 = 2.9$
$Q(S_1, a_4) = 2 + 0.9 * 3 = 4.7$
$Q(S_1, a_1) = 5 + 0.9 * (0.7 * 10 + 0.3 * 2) = 11.84$

Policy $\pi^1$: $S_0 \rightarrow a_6$, $S_1 \rightarrow a_1$, $S_2 \rightarrow a_3$, $S_3 \rightarrow a_2$

ECR of $\pi^1$: $0.2 * 1 + 0.3 * 2 + 0.3 * 3 + 0.2 * 10 = 3.7$

After the second iteration:
$V(S_0) = Q(S_0, a_6) = 2 + 0.9 * 2 = 3.8$
$V(S_1) = Q(S_1, a_1) = 5 + 0.9 * (0.7 * 10 + 0.3 * 2) = 11.84$
$V(S_2) = Q(S_2, a_3) = 3 + 0.9 * 10 = 12$
$V(S_3) = Q(S_3, a_2) = 10 + 0.9 * 10 = 19$

$V^2_{S_0} = 3.8$, $V^2_{S_1} = 11.84$, $V^2_{S_2} = 12$, $V^2_{S_3} = 19$

Policy $\pi^2$: No change. $\pi^2$: $S_0 \rightarrow a_6$, $S_1 \rightarrow a_1$, $S_2 \rightarrow a_3$, $S_3 \rightarrow a_2$

$Q(S_0, a_6) = 2 + 0.9 * 11.84 = 12.656$
$Q(S_0, a_5) = 1 + 0.9 * (11.84 * 0.2 + 12 * 0.8) = 11.7712$
$Q(S_1, a_7) = 2 + 0.9 * 3.8 = 5.42$
$Q(S_1, a_4) = 2 + 0.9 * 12 = 12.8$
$Q(S_1, a_1) = 5 + 0.9 * (0.7 * 19 + 0.3 * 11.84) = 20.1668$

ECR of $\pi^2$: $0.2 * 3.8 + 0.3 * 11.84 + 0.3 * 12 + 0.2 * 19 = 11.712$