

CSE 150 Assignment 5 - Policy Improvement

Josh Anthony - A14281769

Consider the Markov decision process (MDP) with two states $s \in \{0, 1\}$, two actions $a \in \{0, 1\}$, discount factor $\gamma = \frac{2}{3}$, and rewards and transition matrices as shown below:

s	R(s)
0	-2
1	4

s	s'	P(s' s, a=0)
0	0	3/4
0	1	1/4
1	0	1/4
1	1	3/4

s	s'	P(s' s, a=1)
0	0	1/2
0	1	1/2
1	0	1/2
1	1	1/2

- a) (2.5 points) Consider the policy π that chooses the action $a = 0$ in each state. For this policy, solve the linear system of Bellman equations to compute the state-value function $V^\pi(s)$ for $s \in \{0, 1\}$. Your answers should complete the following table:

s	$\pi(s)$	$V^\pi(s)$
0	0	
1	0	

$$\begin{aligned}
 V(s) &= R(s) + \gamma \max_a \sum_{s'} T(s, a, s') V(s') \\
 V^\pi(s) &= R(s) + \frac{2}{3} \left[T(s, 0, s') V^\pi(s') \right. \\
 &\quad \left. + T(s, 1, s') V^\pi(s') \right] \\
 V^\pi(0) &= -2 + \frac{2}{3} \left[\frac{3}{4} V^\pi(0) + \frac{1}{4} V^\pi(1) \right] \\
 V^\pi(1) &= 4 + \frac{2}{3} \left[\frac{1}{4} V^\pi(0) + \frac{3}{4} V^\pi(1) \right]
 \end{aligned}$$

$$\textcircled{1} 2 = -\frac{1}{2} V^{\pi}(0) + \frac{1}{6} V^{\pi}(1)$$

$$\textcircled{2} -4 = \frac{1}{6} V^{\pi}(0) - \frac{1}{2} V^{\pi}(1)$$

$$\textcircled{1} \frac{2}{3} = -\frac{1}{6} V^{\pi}(0) + \frac{1}{18} V^{\pi}(1)$$

$$\textcircled{2} + \textcircled{1} \quad \frac{-10}{3} = -\frac{4}{9} V^{\pi}(1)$$

$$\hookrightarrow \underline{V^{\pi}(1) = 7.5}$$

$$\underline{V^{\pi}(0)} = -2 \left[2 - \frac{1}{6} V^{\pi}(1) \right]$$

$$= -2 [2 - 7.5] = 11$$

So $V^{\pi}(0) = -1.5$ and $V^{\pi}(1) = 7.5$

b) (2.5 points) Compute the greedy policy $\pi'(s)$ with respect to the state-value function $V^\pi(s)$ from part (a). Your answers should complete the following table:

s	$\pi(s)$	$\pi'(s)$
0	0	
1	0	

$$\pi'(s) = \underset{a}{\operatorname{argmax}} \sum_{s'} P(s'|s, a) V^\pi(s')$$

So for $s=0$

$$\pi'(0) = \underset{a}{\operatorname{argmax}} \sum_{s'} P(s'|s=0, a) V^\pi(s')$$

for $a=0 \rightarrow \frac{3}{4} \cdot -1.5 + \frac{1}{4} \cdot 7.5 = \underline{\underline{3/4}}$

$a=1 \rightarrow \frac{1}{2} \cdot -1.5 + \frac{1}{2} \cdot 7.5 = \underline{\underline{3}}$

And for $s=1$

$$\pi(1) = \underset{a}{\operatorname{argmax}} \sum_{s'} P(s'|s=1, a) V^\pi(s')$$

for $a=0 \rightarrow \frac{1}{4} \cdot -1.5 + \frac{3}{4} \cdot 7.5 = \underline{\underline{5.25}}$

$a=1 \rightarrow \frac{1}{2} \cdot -1.5 + \frac{1}{2} \cdot 7.5 = \underline{\underline{3}}$

So $\pi'(0) = 1$ and $\pi'(1) = 0$