

6.390 Introduction to Machine Learning

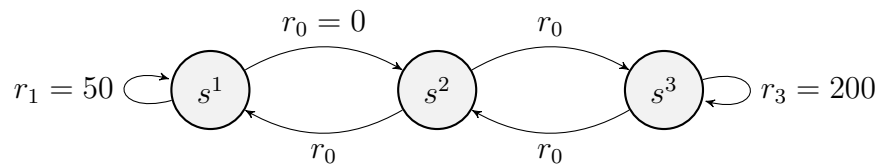
Recitation Week #10

Issued April 24, 2023

1. Wobot is learning how to navigate a 1-D grid world with three locations, l^1 , l^2 , l^3 . Location l^1 is directly to the left of l^2 , which is directly to the left of l^3 . Wobot is in state s^1 when it is in location l^1 , s^2 in l^2 , and s^3 in l^3 .

At each time step, Wobot takes an action $a \in \{a_{\text{left}}, a_{\text{right}}\}$: it will either attempt to move left, a_{left} , or to move right, a_{right} . At far left (from state s^1) there is a wall with a power outlet; when the Wobot takes action a_{left} from s^1 , it earns a reward $r_1 = 50$ and stays in state s^1 . Similarly, at far right is a wall with a stronger power output; when Wobot takes action a_{right} from s^3 , it earns $r_3 = 200$, and stays in s^3 . Other state-action pairs (s, a) in this world earn reward r_0 .

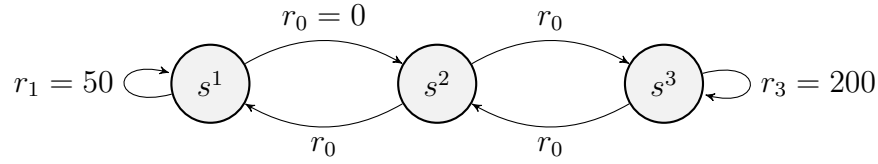
Note: Wobot does not earn a reward for transitioning or staying in a state. The reward comes from taking an action while in a state; e.g., the reward r_1 is a result of the state-action pair (s^1, a_{left}) , not whether or not a specific transition actually occurs.



- (a) For this part, assume that the MDP is *deterministic* with a *finite-horizon*, and that $r_0 = 0$. There are non-zero immediate rewards associated with only two state-action pairs: $R(s^1, a_{\text{left}}) = 50$ and $R(s^3, a_{\text{right}}) = 200$ (denoted by r_1 and r_3 in the diagram, respectively.)
 - i. A *policy* is a function $\pi : \mathcal{S} \mapsto \mathcal{A}$ that specifies what action to take in each state. Describe (in words) what the optimal policy π^* would be for the MDP in the diagram above, for the case of initializing in each of the three possible states. How does the horizon h affect the policy π ?

- ii. For each combination of initial state and horizon h , determine the horizon- h value obtained by the optimal policy, $V_{\pi^*}^h(s)$.

	$h = 0$	$h = 1$	$h = 2$	$h = 3$	$h = 4$
s^1	0				
s^2	0				
s^3	0				



(b) For this part, assume that the MDP is *deterministic* with an *infinite horizon*, and that $r_0 = 0$.

- i. Suppose that the discount factor $\gamma = 0.8$. Describe in words what actions the optimal policy π^* will take in each of the three states.

- ii. For each state $s \in \{s^1, s^2, s^3\}$, for the optimal policy π^* discovered in part (b) i. write the value for $V_{\pi^*}^\infty(s)$ with discount factor $\gamma = 0.8$. Recall the expanded form of the geometric series:

$$\sum_{k=0}^{\infty} a\rho^k = \frac{a}{1-\rho}, \text{ for } |\rho| < 1.$$

$V_{\pi^*}^\infty(s^3) =$

$V_{\pi^*}^\infty(s^2) =$

$V_{\pi^*}^\infty(s^1) =$

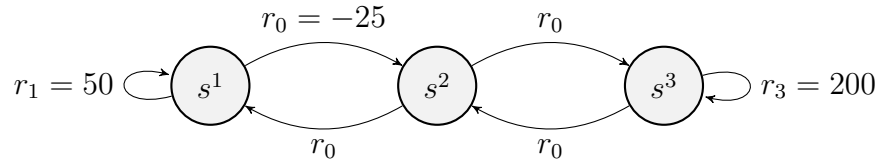
- iii. For each state-action pair, compute two iterations of infinite-horizon value iteration with $\gamma = 0.8$. Do not assume a particular policy. Assume that all value estimates start at 0 and that on each value iteration, you only use values that were computed on the previous iteration.

First iteration, $Q^1(s, a)$:

	a_{left}	a_{right}
s^1		
s^2		
s^3		

Second iteration, $Q^2(s, a)$:

	a_{left}	a_{right}
s^1		
s^2		
s^3		



(c) For this part, assume that the MDP is *deterministic* with an *infinite horizon*, and that actions which transition between states will accumulate a *negative* reward, $r_0 = -25$.

- i. Suppose that the discount factor $\gamma = 0.8$. Imagine a policy, π_{left} , that has Wobot take action a_{left} for all states. What is $V_{\pi_{\text{left}}}^{\infty}(s)$ for starting in each state $s \in \{s^1, s^2, s^3\}$?

$$V_{\pi_{\text{left}}}^{\infty}(s^1) =$$

$$V_{\pi_{\text{left}}}^{\infty}(s^2) =$$

$$V_{\pi_{\text{left}}}^{\infty}(s^3) =$$

Now, imagine a policy, π_{right} , that has Wobot take action a_{right} for all states. What is $V_{\pi_{\text{right}}}^{\infty}(s)$ for starting in each state $s \in \{s^1, s^2, s^3\}$?

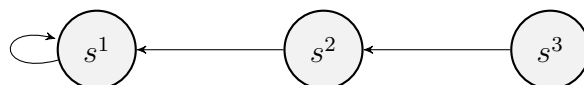
$$V_{\pi_{\text{right}}}^{\infty}(s^3) =$$

$$V_{\pi_{\text{right}}}^{\infty}(s^2) =$$

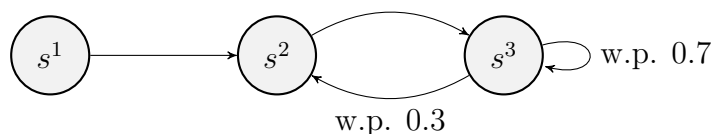
$$V_{\pi_{\text{right}}}^{\infty}(s^1) =$$

2. The season has changed in grid world, from summer to winter. Taking action a_{left} in s^1 will continue to yield reward $r_1 = 50$ with probability 1. However, now slippery, frictionless ice has formed around the stronger power supply and the third location l^3 . These new weather conditions change one of the state transitions to be *non-deterministic*. When Wobot is in s^3 , taking action a_{right} will always allow Wobot to access the stronger power supply and earn a reward $r_3 = R(s^3, a_{\text{right}}) = 200$, but now Wobot will only stay in s^3 with probability 0.7. With probability 0.3, Wobot will slip back to location l^2 and transition into state s^2 .

The diagram below defines state transitions under action a_{left} :



The diagram below defines state transitions under action a_{right} :



Unlabeled arcs correspond to probability 1 transitions. We will consider the infinite-horizon value iteration algorithm for states s^2, s^3 with a discount factor of $\gamma = 0.9$. Do not assume a particular policy. Assume that all value estimates start at 0 and that on each value iteration, you only use values that were computed on the previous iteration.

- (a) In the first value iteration, what are $Q^1(s^3, a_{\text{left}})$ and $Q^1(s^3, a_{\text{right}})$?

$$Q^1(s^3, a_{\text{left}}) =$$

$$Q^1(s^3, a_{\text{right}}) =$$

- (b) In the first value iteration, what are $Q^1(s^2, a_{\text{left}})$ and $Q^1(s^2, a_{\text{right}})$?

$$Q^1(s^2, a_{\text{left}}) =$$

$$Q^1(s^2, a_{\text{right}}) =$$

- (c) Now, at the second value iteration, what are $Q^2(s^3, a_{\text{left}})$ and $Q^2(s^3, a_{\text{right}})$?

$$Q^2(s^3, a_{\text{left}}) =$$

$$Q^2(s^3, a_{\text{right}}) =$$