6.390 Introduction to Machine Learning

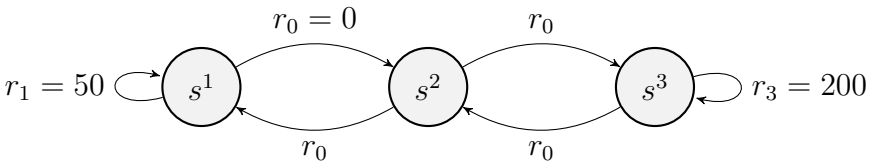
Recitation Week #11

Issued November 21, 2022

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.

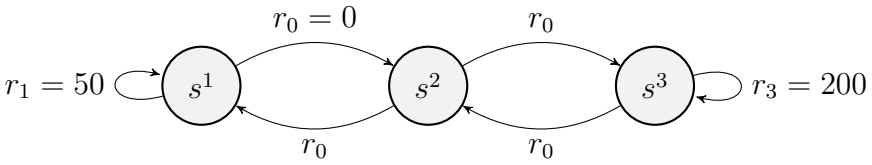


   (a) For this part, assume that the MDP is *deterministic* with a *finite-horizon*, and that $r_0 = 0$. There are non-zero immediate rewards associated with only two state-action pairs: $R(s^1, a_{\text{left}}) = 50$ and $R(s^3, a_{\text{right}}) = 200$ (denoted by r_1 and r_3 in the diagram, respectively.)

      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, $V^h_{\pi^*}(s)$.

	$h=0$	$h=1$	$h=2$	$h=3$	$h=4$
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 $V^{\infty}_{\pi^*}(s)$ with discount factor $\gamma = 0.8$. Recall the expanded form of the geometric series:

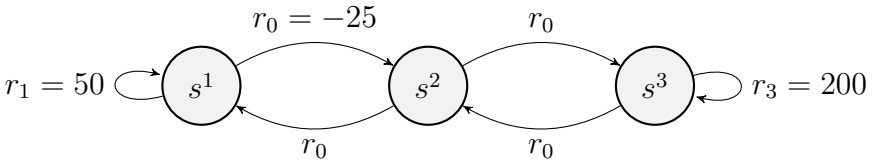
$$\sum_{k=0}^{\infty} a\rho^k = \frac{a}{1-\rho}, \quad \text{for } |\rho| < 1.$$

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

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

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

iii. Suppose that your initial state is given to be s^1. Give a value for γ (constrained by $0 < \gamma < 1$) that results in an optimal policy which results in Wobot taking actions in the opposite direction than the optimal policy from the previous subpart, from state s^1.



(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 $V^{\infty}_{\pi_{\text{left}}}(s)$ for starting in each state $s \in \{s^1, s^2, s^3\}$?

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

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

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

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

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

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

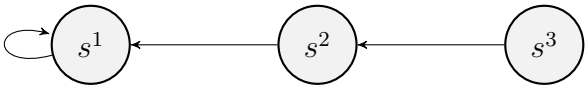
$V^{\infty}_{\pi_{\text{right}}}(s^1) =$

ii. The policies π_{left} and π_{right} are for Wobot to always take the same action irrespective of state. Now we want to consider the state-dependent optimal policy, $\pi^*(s)$. That is, for each state $s \in \{s^1, s^2, s^3\}$, Wobot can decide if it is best to take action a_{left} or a_{right} for that state s.
Take $\gamma = 0.8$ and $r_0 = -25$. For each state $s \in \{s^1, s^2, s^3\}$, what is $\pi^*(s)$?

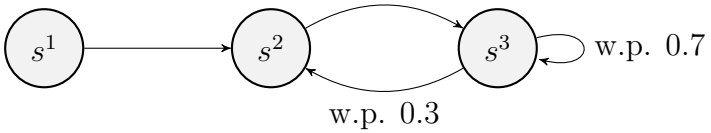
iii. Suppose that your initial state is given to be s^1. Give a value for r_0 that results in an optimal policy which results in Wobot taking actions in the opposite direction than the optimal policy from the part (c) ii.

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.

   (a) With a finite horizon of $h = 1$, what is the expected undiscounted reward obtained when in state s^3 and taking action a_{right} ?

   (b) With a finite horizon of $h = 1$, what is the expected undiscounted reward obtained when in state s^2 and taking action a_{right} ?

   (c) With a finite horizon of $h = 2$, what is the expected horizon-2 undiscounted value when starting in state s^3 obtained by the policy π_{right}? That is, what is $V^2_{\pi_{\text{right}}}(s^3)$?

   $V^2_{\pi_{\text{right}}}(s^3) =$