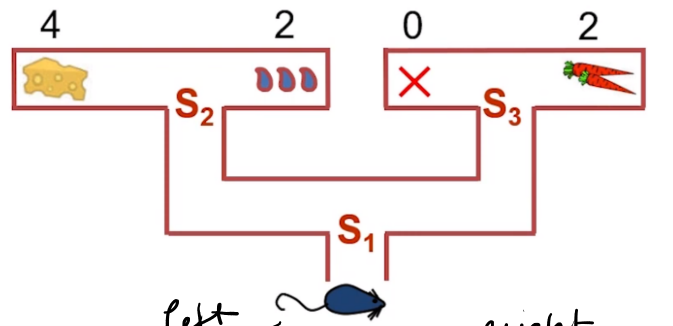
RL-Week-2

Practice Assignment - 2

1) Consider the following problem: **1 point**



left ← → right

Find the undiscounted return for each of the following trajectories:

(1) left-left s_1 s_2 cheese → 4

(2) left-right s_1 s_2 blue → 2

(3) right-left s_1 s_3 nothing → 0

(4) right-right s_1 s_3 carrot → 2

- [x] $4, 2, 0, 2$
- [ ] $4, 4, 4, 4$
- [ ] $2, 2, 2, 2$
- [ ] $0, 2, 4, 2$

2) Consider a robot that can take four actions: (left, right, front, back). An observer starts watching the actions performed by the robot. (~~Every time the robot takes~~ an action, the observer measures the time elapsed on his watch. The timer on the watch is set to 0 when the robot makes its first action. The following table summarizes the information recorded by the observer over 60 seconds: **1 point**

S.No.	Time (s)	Action
1	0	left
2	3	right
3	10	front
4	35	front
5	60	right

front ↑ left ← • → right ↓ back

$a(0) \rightarrow$ left

instead of $a(3)$ we write it as $a(1)$

In terms of the notation a_t used for actions, how would you represent these five actions?

- [x] a_0, a_1, a_2, a_3, a_4
- [ ] $a_0, a_3, a_{10}, a_{35}, a_{60}$

why? because currently we are focussing on discrete time steps (which means at any instant when an action is taken), and not particularly following the wall-clock.

Consider the following sequence of rewards received in a continuing task with $\gamma = 0.9$

$R_1 = -2, \quad R_2 = 5, \quad R_3 = 4, \quad R_4 = 2, \quad R_5 = 1$

Assume that $R_t = 0, t \geq 6$. Assume that states and actions follow zero-indexing. A typical trajectory looks like this: $s_0, a_0, r_1, s_1, a_1, r_2, \cdots$

3. Find the value of G_5.

$G_t = r_{t+1} + r_{t+2} - - -$

$G_5 = \overset{0}{r_6} + \overset{0}{r_7} - -$

given, $R_t = 0$, $\forall\ t \geq 6$

$\underline{G_5 = 0}$

4. Find the value of G_0 contin.

$G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + \gamma^3 R_4 + \gamma^4 R_5$

$= -2 + 0.9 \times 5 + 0.81 \times 4 + 0.729 \times 2 + 0.6561 \times 1$

$= -2 + 4.5 + 3.24 + 1.458 + 0.6561 = \underline{7.8541}$

5) If $q_\pi(s,a) = 1.5$ for some state s in an MDP, which of the following statements is always true? **1 point**

$q_\pi(s,a) = 1.5$ → state s

- ○ The expected return starting from state s and following policy π is equal to 1.5.
- ◉ The expected return starting from state s, taking action a and then following policy π is equal to 1.5
- ○ The return starting from state s is equal to 1.5 in some episode.
- ○ The return starting from state s, taking action a is equal to 1.5 in some episode.

$\left.\begin{array}{l} (s) \rightarrow \text{take ac}^n\ a \\ (\text{following policy } 1.5) \end{array}\right\}$ 1.5

$$q_\pi(s,a) = E\left[G_t \mid S_t = s, A_t = a\right]$$

6) Which of the following statements are true regarding the state and action value functions for a stochastic policy π? **1 point**

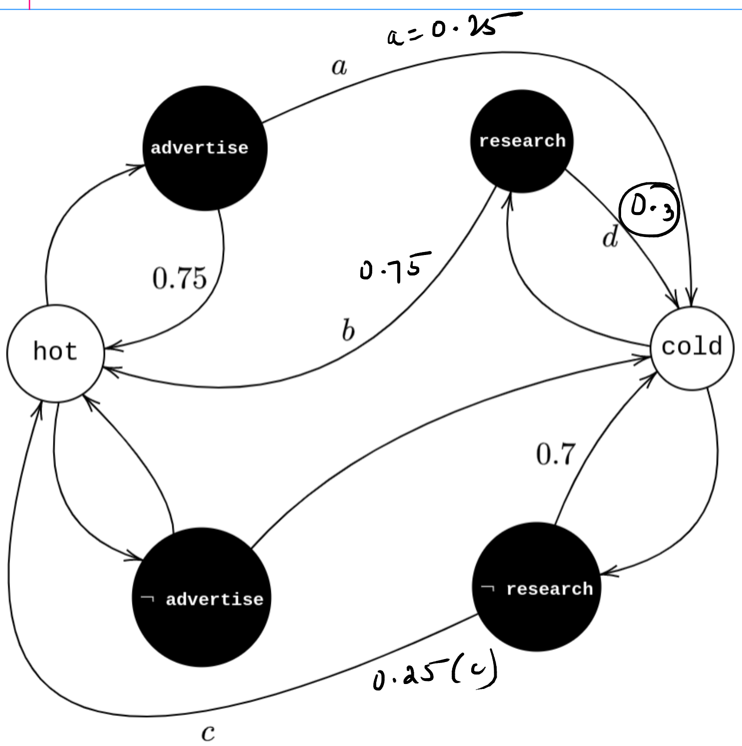
- ☑ $v_\pi(s) = \sum_a \pi(a \mid s) \cdot q_\pi(s,a)$
- ☐ $q_\pi(s,a) = \sum_a \pi(a \mid s) \cdot v_\pi(s)$
- ☑ $v_\pi(s) = E_\pi[q_\pi(S_t, A_t) \mid S_t = s]$
- ☐ $v_\pi(s) = q_\pi(s, \pi(s))$

$$v_\pi(s) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right] \rightsquigarrow \text{state value func}^n$$

ac^n value func^n

$$q_\pi(s|a) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

$$v_\pi(s) = \sum_a \pi(a|s)\, q_\pi(s|a)$$



$0.25 + 0.75 + 0.3 + 0.25$

$= 1.55$

7. What is the value of $a + b + c + d = 1.55$

8) Which of the following is true for any MDP? **1 point**

- ○ $Pr(s_{t+1}, r_{t+1} \mid s_t, a_t) = Pr(s_{t+1}, a_{t+1})$
- ◉ $Pr(s_{t+1}, r_{t+1} \mid s_t, a_t, s_{t-1}, a_{t-1}, \cdots, s_0, a_0) = Pr(s_{t+1}, a_{t+1} \mid s_t, a_t)$
- ○ $Pr(s_{t+1}, r_{t+1} \mid s_t, a_t) = Pr(s_{t+1}, a_{t+1} \mid s_0, a_0)$
- ○ $Pr(s_{t+1}, r_{t+1} \mid s_t, a_t) = Pr(s_{t-1}, a_{t-1})$

This is known as the Markov Property, which says that the history or the older state - ac^n value pairs are not considered.

9) Consider a continuing task for which the reward at any time step is either zero or 1. Let us call this task T_1. Consider another continuing task, T_2, that is identical to T_1 in all respects but for the rewards. The rewards in T_2 at every time step is c more than the reward in T_1 for the corresponding situation. If $v_\pi(s)$ is the state value function in T_1 for some policy π, then what is the state value function for T_2 for the same policy? Assume that $0 < \gamma < 1$ and is the same for both tasks. **1 point**

- ○ $v_\pi(s)$
- ○ $v_\pi(s) \cdot c$
- ○ $v_\pi(s) + c$
- ○ $v_\pi(s) - c$
- ✓ $v_\pi(s) + \frac{c}{1-\gamma}$
- ○ The value function of T_2 cannot be expressed in terms of the one for T_1.

$T_1 \rightarrow$ continuing task (reward = 0 or 1)

$T_2 \xrightarrow{\text{con.}} c + r_{T_1}$ (at that time step)

$v_\pi(s) \rightarrow T_1$

T_2 (state value func) ?

$$G_{T_1} = R_{t+1} + \gamma R_{t+2} \cdots$$

$$G_{T_2} = (c + R_{t+1}) + \gamma(c + R_{t+2}) \cdots$$

$$= (c + \gamma c \cdots)\ \underbrace{(R_{t+1} + \gamma R_{t+2})}_{G_{T_1}} = G_{T_1} + c(1 + \gamma + \gamma^2 \cdots)$$

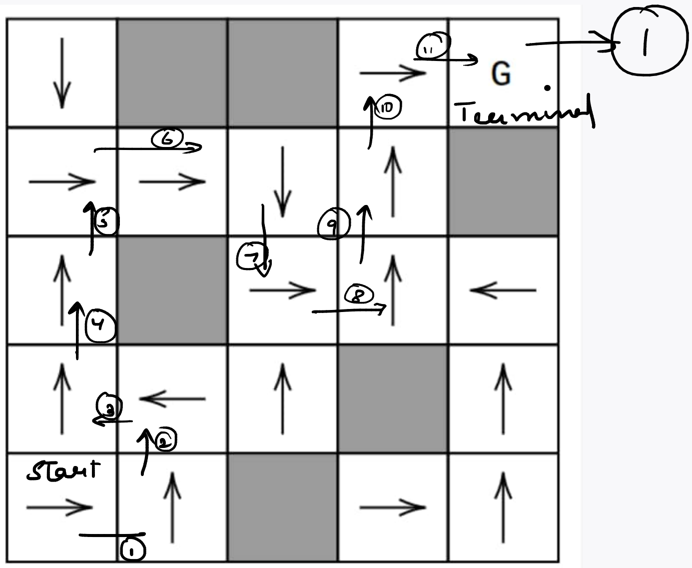
$$= G_{T_1} + \frac{c}{1-\gamma}$$

$$T_1 \rightarrow v_\pi(s) = E_\pi[G_{T_1} \mid S_t = s]$$

$$T_2 \rightarrow v'_\pi(s) = E_\pi[G_{T_2} \mid S_t = s]$$

$$= E_\pi\left[G_{T_1} + \frac{c}{1-\gamma} \mid S_t = s\right] \Rightarrow \boxed{v_\pi(s) + \frac{c}{1-\gamma}}$$

Consider a grid-world and a policy π corresponding to it. The agent starts at the bottom-left state. The goal (terminal state) is at the top right. The reward is 1 on reaching the goal and 0 for all other states. Use $\gamma = 0.9$. Assume that all transitions in the environment are deterministic.



10. What is the return starting from the bottom left state and following policy π?

11 steps are required to reach the goal state that means 10 steps have reward 0.

$$\therefore,\ G_0 = \gamma^n$$

$$= 0.9^{10} = 0.3486$$

11. What is $v_\pi(0)$?

$$v_\pi(0) = E_\pi[G_0 \mid S_t = s]$$

$$= 0.3486$$

Graded Assignment-2

1) Which of these statements is true regarding the rewards obtained in an MDP?

(1) The reward r_{t+1} obtained on choosing an action a_t depends only on a_t.

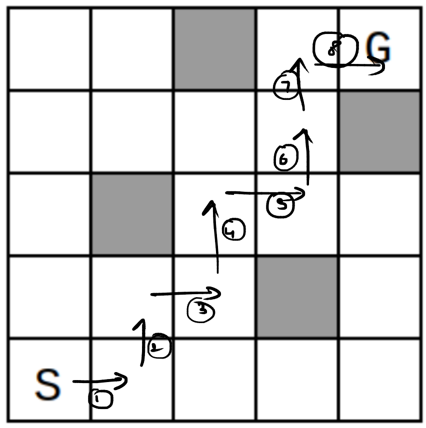
(2) The reward r_{t+1} obtained on choosing an action a_t depends only on s_t and a_t.

(3) The reward r_{t+1} obtained on choosing an action a_t depends on s_t, a_t and s_{t+1}.

$p_r(s', r \mid s, a)$

$r(s, a, s') \rightarrow$ the reward is a func^n that depends on current state s, takes ac^n a, and reaches new state s'.

2) Consider the following grid world. G is the goal and S is the cell where the agent starts. In each cell, the agent is allowed to choose one of the four actions: north, east, west and south. Actions that take the agent outside the grid or into an obstacle leave the state unchanged. The transitions are deterministic.



8 steps $\times$ -1 $=$ -8

4 ac^{ns} $\leftarrow \uparrow \rightarrow \downarrow$

If the reward at each time step is -1, find the maximum possible return starting from S. Assume that $\gamma = 1$.

-8

3) You work at a software company that is 15 km away from home. You would like to optimize the time taken to reach office from home in the morning on weekdays. The traffic density varies from hour to hour. You would like to choose from one of these 5 slots for starting from home: 7, 8, 9, 10 and 11. Assume that the traffic density for a given interval on all weekdays is roughly the same. Choose the most appropriate option. ***1 point***

- The best time can be found by treating this as a bandit problem.
- The best time can be found by treating this as a full-RL problem.
- This cannot be framed as an RL problem.

7, 8, 9, 10, 11

all five days $\rightarrow$ can start with a single state (that is starting from home. The ac^n to choose the slot. You get the reward or penalty just after the ac^n. (case of immediate reward $\rightarrow$ Bandit)

4) Suppose $\gamma = 0.8$ and the reward sequence is $R_1 = 1$ followed by an infinite sequence of 5s, that is $R_2 = R_3 = \cdots = 5$. What is the value of the return G_0?

$$G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + \gamma^3 R_4 + \gamma^4 R_5$$

$$= 1 + 0.8 \times 5 + 0.8^2 \times 5$$

$$= 1 + 0.8 \times 5 (1 + \gamma + \gamma^2) = 1 + \frac{\cancel{0.8}^{4} \times 5 \times 1}{0.\cancel{2}}$$

$$\left(\frac{c}{1-\lambda} \right) \quad c = 1, \ \lambda = 0.8$$

$$= 21$$

5) The table given below is a representation of the action value function for a policy π at a state s for all actions that are possible from that state.

a	$q_\pi(s,a)$
0	1.3
1	1.8
2	2.1
3	1.2

What is the value of $v_\pi(s)$ if $\pi(. \mid s) = [0.3, 0.2, 0.4, 0.1]$?

$$v_\pi(s) = \sum q_\pi(s|a)\, \pi(a|s)$$

$= 1.3 \times 0.3 = 0.39$
$+ 1.8 \times 0.2 = 0.36$
$+ 2.1 \times 0.4 = 0.84$
$+ 1.2 \times 0.1 = 0.12$

1.71

6) If the policy π is deterministic, then is the following statement true or false?

$v_\pi(s) = q_\pi(s, \pi(s))$

1 point

- True ✓
- False

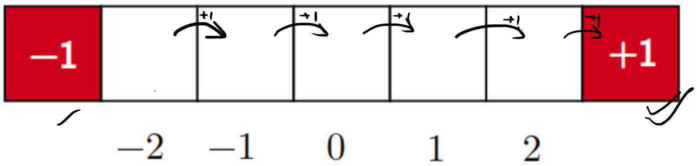
$v_\pi(s) = q_\pi(s|a)$

$a = \pi(s)$

$v_\pi(s) = q_\pi(s \mid \pi(s))$

eligible only for deterministic policies

Consider the following environment. Non-terminal states are numbered from -2 to 2. The terminal states are colored red with the corresponding rewards written inside them. Rewards for moving to non-terminal states are zero. All transitions in the environment are deterministic. Assume that $\gamma = 1$.



7) Consider a deterministic policy in which the agent always moves right. What is the value function corresponding to this policy? Express your answer as a vector of length 5, starting from state -2 to 2.

1 point

- $v_\pi(.) = [0, 0, 0, 0, 0]$
- $v_\pi(.) = [-1, -1, -1, -1, -1]$
- $v_\pi(.) = [1, 1, 1, 1, 1]$ ✓
- $v_\pi(.) = [-1, -1, 0, 1, 1]$

8) Consider a stochastic policy in which the agent moves right 50% of the times from every state. Consider any episode under this policy starting from state 0. What is the return for this episode?

1 point

50%

- 0
- 1
- -1
- 0.5
- -0.5
- Insufficient information. We need the complete trajectory to compute the return. ✓

−2 −1 0 +1 +2

in this scenario, our random variable is returns, since the policy becomes stochastic, the rewards are penalized on both sides either going to +1 or −1.

9) Consider a stochastic policy in which the agent moves right 50% of the times from every state. What is the value function for state 0 under this policy?

1 point

- 0 ✓
- 1
- -1
- 0.5
- -0.5
- Insufficient information. We need the complete trajectory to compute the return.

$v_\pi(0)$

-1 $(P=0.5)$ $+1$ $(P=0.5)$

-2 $(P=0.5)$ 0 $P(0.5)$ 0 $P(0.5)$ $+2$ $P(0.5)$

value function, $v_\pi(s) = 0$

10) Consider a modification to the environment. The transitions are now stochastic:

$p(s_L \mid s, \text{left}) = 0.8, \quad p(s_R \mid s, \text{left}) = 0.2$
$p(s_L \mid s, \text{right}) = 0.2, \quad p(s_R \mid s, \text{right}) = 0.8$

Here, s_L, s_R are the states immediately to the left and right of state s for $s \in \{-2, -1, 0, 1, 2\}$. Find the probability of seeing the following trajectory under the stochastic policy mentioned in the previous question given that the agent starts from state 0. The rewards are not mentioned here:

$0, \;\text{left}, \;1, \;\text{right}, \;0, \;\text{right}, \;1, \;\text{right}, \;2, \;\text{right}, \;\text{right-terminal}$

Enter your answer correct to five decimal places.

$P(s_L \mid s, \text{left}) = 0.8$

$P(s_L \mid s, \text{right}) = 0.2$

$P(s_R \mid s, \text{left}) = 0.2$

$P(s_R \mid s, \text{right}) = 0.8$

$s \to a$

$\underbrace{0 \; l \; -1}_{0.8} \;\; \underbrace{r \; 0}_{0.8} \;\; \underbrace{r \; 1}_{0.8} \;\; \underbrace{r \; 2}_{0.8} \;\; \underbrace{r \; \text{term}}_{1}$

$(0.8)^4 = 0.4096$