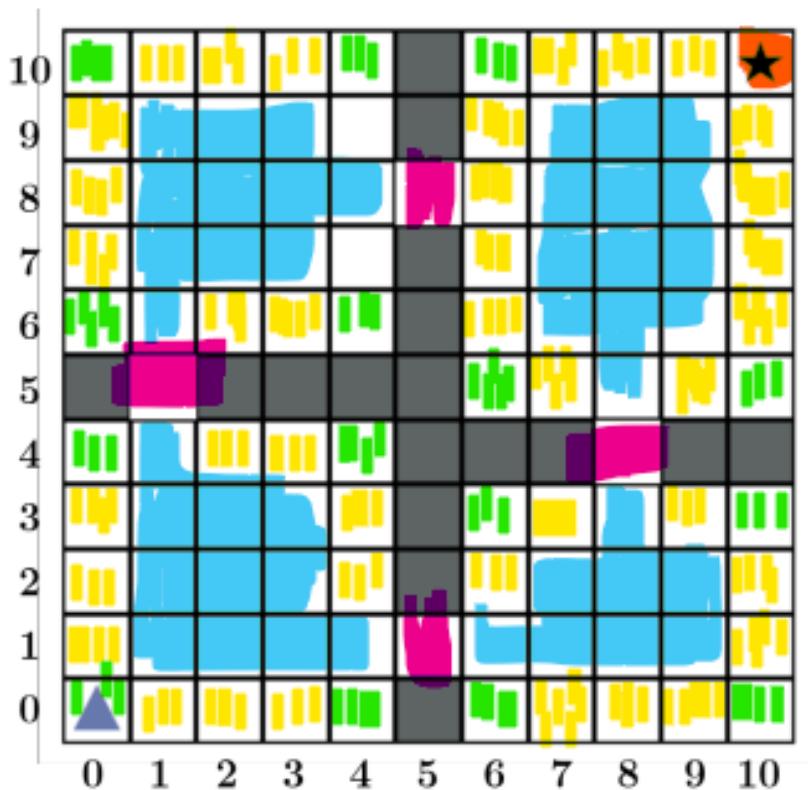


Exercise 2: Markov Decision Processes

Ravina Lad

Q1 b)



I have discussed the q1b and 7b, c with Utkarsh Ghime and Samruddhi Raut.

Ravina Lad.

Ex 2: Markov Decision Process.

10/04/23

Q.1 Formulating an MDP.

Let's consider 4-rooms domain.

a) States: $s \rightarrow (x, y); x, y \in [0, 10]$

$k = 0, 2, 3, 4, 5, 6, 7, 9, 10.$

$(x, y) \neq (i, 5), (j, 4), (5, k)$

$j = 5, 7, 9, 10.$

Actions: $\{ \text{left, right, up, down} \}$

$i = 0, 2, 3, 4.$

(except walls)

b) consider, dynamics $f^n p(s', r | s, a)$.

How many non-zero rows are in this conditional probability table?

→ As shown in grid image —

These are 44 states → that can move in 4 direction (Blue)

40 in 3 direction (Yellow)

(green) 15 states (starting + close adjacent to wall) → 2 directions

(purple/pink) 1 states - in middle gridworld = 2 direction.

(orange) 1 good state that return to start state.

Hence,

$$44 \times 4 \times 3 + 40 \times 4 \times 3 + 15 \times 3 \times 2$$

$$+ 15 \times 2 \times 2$$

$$+ 4 \times 3 \times 2$$

$$+ 4 \times 2 \times 2 + 1$$

$$1199$$

I guess there are 1199 non-zero rows in this conditional probability.

Q2. RL objective.

a) Episodic discounted pole-balancing,

agent receives reward of 0 → each step

- 1 → failure. Episodic return at each timestep is sum of all discounted rewards that agent expects to receive from that t until end of episode

$$G_t = R_t + \gamma G_{t+1}$$

$G_t \rightarrow$ episodic return at t

$R_t \rightarrow$ reward at t

$\gamma \rightarrow$ discount factor

$$\text{episodic fn: } G_t = R_{t+1} + R_{t+2} + R_{t+3} + R_{t+4} + \dots R_T$$

(with discounting) →

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots + \gamma^{T-t-1} R_T$$

$$= \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1} = \gamma^{T-t-1} R_T$$

$$\text{(continuing fn) } G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = \gamma^T R_T \quad T \dots \text{No of steps in the task}$$

→ episodic has only 1 failure, it returns last G_t .

→ continuing will have many failures & return value will always be updated.

→ The episode ends when the pole falls. Hence, return at each time step is equal to the discounted expected number of steps until failure.

→ in continuing discounted pole-balancing, the task never ends. Hence, return at each time step is ∞ .

→ in case of episodic, agent's goal is to maximize the expected return over course of episode, while in continuing task is to maximize expected return over entire task

- b) We know that the agent's reward r^n is very sparse.
 agent only receives +1 when it escapes & 0 at all other times.
 it has very little feedback to learn from.
- Reward r^n doesn't provide any information about how close it is to escaping from the maze. even if its close to escaping, it wouldn't know & take proper action towards the goal & it might go away also.
 - To improve agent's action selection, need to give reward r^n more information and less sparse.
 eg - we can give agent reward of +0.1 for every steps that it takes towards escaping. & -0.1 for going away from escaping the maze.
 - $R_1 = R_2 = R_{t+k} = 0 ; R_{t+k+1} = 1$
 then G_t will be always 1.
 Hence, it will never improve its performance.

Q.3 a) Discounted Return.

suppose $\gamma = 0.5$

$R_1 = -1 \quad R_2 = 2 \quad R_3 = 6 \quad R_4 = 3 \quad R_5 = 2$ with $T=5$

What are G_0, G_1, \dots, G_5 ? Hint: work backwards.

$$\rightarrow G_5 = R_6 + \gamma G_6 = 0$$

$$G_4 = R_5 + \gamma G_5 = 2 + 0.5(0) = 2$$

$$G_3 = R_4 + \gamma G_4 = 3 + 0.5(2) = 4$$

$$G_2 = R_3 + \gamma G_3 = 6 + 0.5(4) = 8$$

$$G_1 = R_2 + \gamma G_2 = 2 + (0.5)8 = 6$$

$$G_0 = R_1 + \gamma G_1 = -1 + (0.5)6 = 2$$

Suppose $V = 0.9$

$R_1 = 2$ followed by infinite sequence of 7's.

$G_1 ? G_0 ?$

$$\begin{aligned} \text{① } G_1 &= R_2 + V R_3 + V^2 R_4 + \dots + V^n R_{n+2} \\ &= \sum_{i=0}^{\infty} V^i R_{n+2} \quad (R_1 = 2, R_3 = R_5 = \dots = R_n = 7) \\ &= \frac{1}{1-V} \times 7 = \frac{1}{1-0.9} \times 7 = \frac{1 \times 7}{0.1} = 70 \end{aligned}$$

$$\begin{aligned} \text{② } G_0 &= R_1 + 0.9 \times 70 \\ &= R_1 + 63 = 65. \end{aligned}$$

Q4

Q4(a) Discounted return.

$$A = 0.8$$

$$R_1 = -1$$

$$A = 0.8 + 0.8 \times R_2 + 0.8^2 R_3 + \dots$$

$$A = R_1 + A \times R_2 + A^2 R_3 + \dots$$

$$A = (1) 2 \cdot 0 + 2 = 2A + A^2 R_2 + A^3 R_3 + \dots$$

$$A = (1) 2 \cdot 0 + 2 = 2A + A^2 R_2 + A^3 R_3 + \dots$$

$$2 = (1) 2 \cdot 0 + 2 = 2A + A^2 R_2 + A^3 R_3 + \dots$$

$$2 = 2(2 \cdot 0) + 2 = 2A + A^2 R_2 + A^3 R_3 + \dots$$

Q-4 Discount factor.

if $\gamma = 1 \rightarrow$ agent will select action that maximizes the immediate reward.

$\gamma < 1 \rightarrow$ agent will discount future rewards.

agent will select action that maximizes the sum of discounted rewards. & agent's decision (U/D) will depend on relative values of future rewards.

We have

$$\sum_{i=1}^{\infty} \gamma^i = \gamma^1 + \gamma^2 + \gamma^3 + \dots + \gamma^{100} = \frac{1 - \gamma^{100}}{1 - \gamma}$$

G_t - discounted return at t

$$G_{\text{up}} = R_1 + \sum_{i=1}^{100} \gamma^i R_{101} = 50 + (-1) \frac{(1 - \gamma^{100})}{1 - \gamma}$$

$$G_{\text{down}} = R_1 + \sum_{i=1}^{100} \gamma^i R_{101} = 1 - 50 + \frac{(1 - \gamma^{100})}{1 - \gamma}$$

if $G_{\text{up}} > G_{\text{down}} \rightarrow \begin{cases} \text{select UP} \\ \text{Down otherwise} \end{cases}$

Let's find γ ,

$$50 - \frac{(1 - \gamma^{100})}{1 - \gamma} = -50 + \frac{(1 - \gamma^{100})}{1 - \gamma}$$

$$2 \cdot 50 = \cancel{2} \frac{(1 - \gamma^{100})}{1 - \gamma}$$

$$\therefore -1.047 < \gamma < 0.9844.$$

Thus, while $0 < \gamma < 0.9844 \rightarrow$ it will go UP,
~~0.047~~ otherwise down.

Q5 a) Modifying reward function

- signs of the rewards are important in gridworld.
- signs tell us about the agent's progress.
- I think the interval rewards are also important, as they talk about how much the agent is rewarded for making progress towards goal.

- To prove adding const. c to all rewards adds a const $\rightarrow V_c$ to all values of all states.
it doesn't affect the relative values of any states under any policy.

$$G_t = \sum_{i=0}^{\infty} \gamma^i R_{t+k+i}$$

$$V_{\pi}(s) = E_{\pi}[G_t \mid s_t = s] \quad V_{\pi}(s) = \text{optimal value of state } s$$

$$= E_{\pi} \left[\sum_{i=0}^{\infty} \gamma^i R_{t+k+i} \mid s_t = s \right] \quad \pi \rightarrow \text{policy.} \\ G_t = \text{discounted return at time step } t$$

$$G'_t = G_t + c \quad \dots \text{if we add const } c \text{ to all rewards.} \\ s_t \rightarrow \text{state at } t.$$

$$= E_{\pi} \left[\sum_{i=0}^{\infty} \gamma^i R_{t+k+i} + \sum_{i=0}^{\infty} \gamma^i c \mid s_t = s \right]$$

$$= E_{\pi} \left[\sum_{i=0}^{\infty} \gamma^i R_{t+k+i} \mid s_t = s \right] + E_{\pi} \left[\sum_{i=0}^{\infty} \gamma^i c \mid s_t = s \right]$$

$$= V_{\pi}(s) + \sum_{i=0}^{\infty} \gamma^i c$$

$$= V_{\pi}(s) + \frac{c}{1-\gamma}$$

- b) Yes, it would have an effect, when we will add const c to all the rewards in an episodic task \rightarrow maze running.
- dependent of the value of c \rightarrow +ve / -ve. it will make longer episodes longmore advantageous or less.
 - e.g. if we add +1 for escaping from maze & 0 at all other time. Now, if we add a const +1 to all reward. then agent will receive +2 for escaping. this will be helpful as agent will receive higher reward.
 - if we add -1 to all rewards, agent will receive 0 for escaping from maze. This won't help as agent won't be receiving any rewards for escaping.

Q. 6

a)

3.2	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
1.9	-1.3	-1.2	-1.4	-2.0

$$V_{++}(s) = \frac{1}{4} \times 1 \times (0 + 0.9 \times 2.3) + \frac{1}{4} \times 1 \times (0 + 0.9 \times 0.4)$$

$$\text{Here, } s = \text{centrestate} + \frac{1}{4} \times 1 \times (0 + 0.9 + (-0.4)) + \frac{1}{4} \times 1 \times (0 + 0.9 \times 0.7)$$

$$= \frac{1}{4} \times 0.9 (2.3 + 0.7) = \frac{0.9 \times 3}{4} = \frac{2.7}{4} = 0.675$$

$$\approx 0.7 \text{ (approximately)}$$

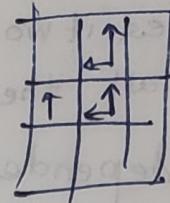
As you can see in the grid.

gridworld

b) show - for optimal policy -

$$V_{\pi^*}(\text{center state}) = +17.8$$

	19.8	
19.8	17.8	16.0
	16.0	

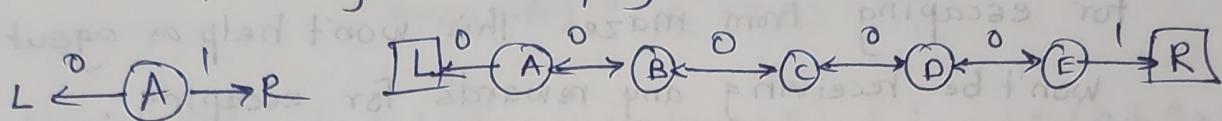


$$V_{\pi}(s) = \frac{1}{2} \times 1 \times (0 + 0.9 \times 19.8)$$

$$+ \frac{1}{2} \times 1 \times (0 + 0.9 \times 19.8)$$

$$V_{\pi}(s) = \frac{1}{2} \times 2 \times \underline{17.82}$$

Q.7) a) Guessing & verifying value f^n :



3 step MDP. All episodes start in the center st. A.

then proceed L/R. with equal probability.

When episode terminates on R \rightarrow Reward +1

all other rewards 0. undiscounted MDP $V=1$.

Guess: Value f^n :

$$V_{\pi}(s) = \frac{1}{2} V(L) + \frac{1}{2} V(R)$$

\rightarrow I guess it will be \rightarrow

$$\frac{1}{2} \times 0 + \frac{1}{2} \times 1 = \frac{1}{2}$$

$$V(L) = 0$$

$$V(R) = +1$$

Verify \rightarrow

$$V_{\pi}(s) = \sum_a (a|s) \sum_{s', r} P(s', r | s, a) [r + V_{\pi}(s')]$$

$$= \frac{1}{2} \times 1 \times (1 + 1 \times 0) + \frac{1}{2} \times 1 (0 + 1 \times 0)$$

$$= \frac{1}{2} \times 1 + \frac{1}{2} \times 0 = \frac{1}{2}$$

Verified

b) extended MDP \rightarrow 7 states (right)

Guess & verify valuefn.

$$V_{\pi(s)}(A) = \frac{1}{2} \times 1 \times (0 + 1 \times 0) + \frac{1}{2} \times 1 \times (0 + 1 \times V_{\pi(s)}(B))$$

$$= \frac{1}{2} V_{\pi(s)}(B)$$

$$V_{\pi(s)}(B) = \frac{1}{2} \times 1 \times (0 + 1 \times V_{\pi(s)}(A)) + \frac{1}{2} \times 1 \times (0 + 1 \times V_{\pi(s)}(C))$$

$$= \frac{1}{2} V_{\pi(s)}(A) + \frac{1}{2} V_{\pi(s)}(C)$$

$$V_{\pi(s)}(C) = \frac{1}{2} \times 1 \times (0 + 1 \times V_{\pi(s)}(B)) + \frac{1}{2} \times 1 \times (0 + 0 \times V_{\pi(s)}(D))$$

$$= \frac{1}{2} V_{\pi(s)}(B) + \frac{1}{2} V_{\pi(s)}(D)$$

Similarly,

$$V_{\pi(s)}(D) = \frac{1}{2} V_{\pi(s)}(C) + \frac{1}{2} V_{\pi(s)}(E)$$

$$V_{\pi(s)}(E) = \frac{1}{2} \times 1 \times (0 + 1 \times V_{\pi(s)}(D)) + \frac{1}{2} \times 1 \times (0 + 1 \times 0)$$

$$= \frac{1}{2} V_{\pi(s)}(D) + \frac{1}{2}$$

$$V_A = \frac{1}{6} \times (V_B - 2VA) = \frac{1}{3}$$

$$V_C = 3VA = \frac{1}{2} \quad V_D = 4VA = \frac{2}{3}$$

$$+ (V_E - 5VA = \frac{5}{6})$$

$$[(VA - 1) \times 2 + (VA - 2)] (VA + 1) = (VA)^2$$

$$[(VA - 1) \times 3 + (VA - 2)] (VA + 1) = (VA)^2$$

c) arbitrary no. states n ,

$$V_{\pi}(s_t | i) = \frac{1}{n-1} \left(i = 1, 2, 3 \dots (n-2) \right)$$

Q.8. solving for value s^n .

g) expand Bellman's eq. for 2 states. in recycling robot,

for an arbitrary policy $\pi(a|s)$

discount factor γ , domain pars $\alpha, \beta, r_{\text{search}}, r_{\text{wait}}$

$s = \{\text{high}, \text{low}\}$... assume that only 2 charge levels are distinguished.

Action sets are then - recharging is always full

$$A(\text{high}) = \{\text{search, wait}\}$$

$$A(\text{low}) = \{\text{search, wait, recharge}\}$$

$$V(s_t) = E[R_{t+1} + \gamma V(s_{t+1}) | s_t = s]$$

$$V(\text{high}) = \sum_a \pi(a|s) \sum_{s', r} P(s', r | s, a) [r + \gamma V_{\pi}(s')]$$

$$= \pi(\text{search} | \text{high}) [r_{\text{search}} + \alpha \cdot r \cdot V(\text{high}) + (1-\alpha) \times r \times V(\text{low})]$$

$$+ \pi(\text{wait} | \text{high}) [1 \times (r_{\text{wait}} + \gamma \cdot V(\text{high}))]$$

$$V(\text{low}) = \pi(\text{search} | \text{low}) [r_{\text{search}} + \beta \times r \times V(\text{low}) + (1-\beta) \times r \times V(\text{high}) + \pi(\text{wait} | \text{low}) [1 \times (r_{\text{wait}} + \gamma V(\text{low}))]]$$

$$+ \pi(\text{recharge} | \text{low}) [0 + 1 \times r \times V(\text{high})]$$

$$\begin{array}{ll}
 b) \quad \alpha = 0.7 & \gamma = 0.9 \\
 \beta = 0.6 & r_{\text{search}} = 10 \\
 & r_{\text{wait}} = 3
 \end{array}
 \quad \begin{array}{l}
 \pi(\text{search} | \text{high}) = 1 \\
 \pi(\text{wait} | \text{low}) = 0.5 \\
 \pi(\text{recharge} | \text{low}) = 0.5
 \end{array}$$

$\pi(\text{wait} | \text{high}) \rightarrow$ probability of choosing the wait action when in the high state.

Here $\pi(\text{search} | \text{high}) = 1$ so $\pi(\text{wait} | \text{high}) = 0$.

As the policy always tells us to search when in high state.
So never wait in high state.

As the agent will always select to wait/recharge in low state
 $\pi(\text{search} | \text{low}) = 0$.

Hence,

$$\begin{aligned}
 v(\text{high}) &= 1 \times [10 + 0.7 \times 0.9 \times v(\text{high}) + 0.3 \times 0.9 \times v(\text{low})] \\
 &= 10 + 0.63v(\text{high}) + 0.27v(\text{low})
 \end{aligned}$$

$$\begin{aligned}
 v(\text{low}) &= 0 + 0.5(3 + 0.9v(\text{low})) + 0.5(0.9v(\text{high})) \\
 &= 1.5 + 0.45v(\text{low}) + 0.45v(\text{high})
 \end{aligned}$$

$$0.37v(\text{high}) = 10 + 0.27v(\text{low}) \quad \text{--- } ①$$

$$0.55v(\text{low}) = 1.5 + 0.45v(\text{high}) \quad \text{--- } ②$$

$$\begin{array}{rcl}
 + 0.37v(\text{high}) - 0.27v(\text{low}) & = 10 \\
 + 0.55v(\text{low}) - 0.45v(\text{high}) & = 1.5
 \end{array}
 \xrightarrow{\text{Solved it using online tool}}$$

$$v(\text{high}) = \frac{5905}{82}$$

$$v(\text{high}) = 72.01$$

$$v(\text{low}) = \frac{5055}{82}$$

$$v(\text{low}) = 61.64$$

Q π back diagram.

```

graph TD
    Root((s, a)) -- p --> P(( ))
    Root -- r --> R(( ))
    P -- "π" --> SPrime((s'))
    R -- "π" --> APrime((a'))
  
```

$$d) \quad V(T(s)) = E_{TT}(g+1 | st=s)$$

$$= E_T \left(\sum_{k=0}^{\infty} y^k R_{t+k+1} \mid s_t = s \right)$$

$$q_{\pi \pi}(s, a) = E_{\pi \pi} \left(g_t \mid s_t = s, A_t = a \right)$$

$$= E_T \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a \right]$$

$$V(s) = \sum_a \pi(a|s) q(s,a)$$

$$b) Q^{\pi\pi}(s,a) = E_{\pi\pi} \left(\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid s_t = s, A_t = a \right)$$

$$= \sum_{s', r} P(s', r | s, a) (r + \gamma V_{\pi^*}(s'))$$

$$c) VTT(s') = \sum_{a'} T(a'|s') q(s', a')$$

$$q_{\pi}(s, a) = \sum_{s', r} p(s', r | s, a) [r + \gamma \sum_{a'} q_{\pi}(s', a')]$$