Ravina Lad.

# Ex2: Markov Decision Process.

10/04/23

Q.1 Formulating an MDP.

Let's consider 4-rooms domain.

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

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

Where

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

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

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

(except: walls)

Actions: {left, right, up, down}

b) Consider, dynamics fn $p(s', r | s, a)$.

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

→ As shown in grid image —

There 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) 4 states - in middle grid world = 2 direction.

(orange) 1 goad 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 time step is sum of all discounted rewards that agent expects to receive from that t until end of episode

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

$G_t$ → episodic return at t
$R_t$ → reward at t
$\gamma$ → discount factor.

episodic fn: $G_t = R_{t+1} + R_{t+2} + R_{t+3} + R_{t4} \cdots R_T$

with discounting →

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

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

Continuing fn → $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = \gamma^k$ T ... 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, whin in continuing task is to maximize expected return over entire task

b) We know that the agent's reward $f^n$ is very sparse. agent only receives +1 when it escapes & 0 at all other times. it has very little feedback to learn from.

- Reward $f^n$ 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 $f^n$ 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 it's performance.

Q.3 a) Discounted Return.

suppose $\gamma = 0.5$
$R_1 = -1$ $R_2 = 2$ $R_3 = 6$ $R_4 = 3$ $R_5 = 2$ with $T = 5$
What are $G_0, G_1 \ldots G_5$? Hint: work backwards.

→

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

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

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

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

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

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

) suppose $\gamma = 0.9$

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

$G_1$ ? $G_0$?

→

① $G_1 = R_2 + \gamma R_3 + \gamma^2 R_4 + \dots \gamma^n R_{n+2}$

$= \sum_{i=0}^{\infty} \gamma^i R_{n+2}$ ( $R_1 = 2$, $R_3 = R_3 = \dots R_n = 7$)

$= \frac{1}{1-\gamma} \times 7 = \frac{1}{1-0.9} \times 7 = \frac{1}{0.1} \times 7 = \underline{70}$

② $G_0 = R_1 + 0.9 \times 70$

$= R_1 + 63 = \underline{65}$.

8.4

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

$$\sum_{i=1}^{100} \gamma^i = \gamma^1 + \gamma^2 \ldots + \gamma^{100} = \frac{1-\gamma^{100}}{1-\gamma}$$

$G_t$ – discounted return at $t$

$$G_{up} = R_1 + \sum_{i=1}^{100} \gamma^i R_{10+} = 50 + \frac{(-1)(1-\gamma^{100})}{1-\gamma}$$

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

if $G_{up} > G_{down} \longrightarrow$ { select UP / Down otherwise

Let's find $\gamma$,

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

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

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

Thus, while $0 < \gamma < 0.9844$ (-1.047) $\rightarrow$ it will go UP, otherwise down.

Q5 a) Modifying reward fn.

→ signs of the rewards are important in gridworld.
→ signs tell us about the agent's progess.
→ 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 → Vc to all values of all states.
it doesn't affect the relative values of any states under any policy.

$$G_t = \sum_{i=0}^{\infty} \gamma^i R_{t+k+1}$$

$$V_\pi(s) = E_\pi[G_t \mid s_t = s] \quad \ldots$$

$$= E_\pi\left[\sum_{i=0}^{\infty} \gamma^i R_{t+k+1} \mid s_t = s\right]$$

$V_\pi(s)$ = ~~optimal~~ value of state $s$
$\pi \rightarrow$ policy.
$G_t$ = discounted return at time step $t$
$s_t \rightarrow$ state at $t$.

$G'_t = G_t + c$ ... if we add const $c$ to all rewards.

$$= E_\pi\left[\sum_{i=0}^{\infty} \gamma^i R_{t+k+1} + \sum_{i=0}^{\infty} \gamma^i c \mid s_t = s\right]$$

$$= E_\pi\left[\sum_{i=0}^{\infty} \gamma^i R_{t+k+1} \mid s_t = s\right] + E_\pi\left[\sum_{i=0}^{\infty} \gamma^i c \mid s_t = s\right]$$

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

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

b) Yes, it would have an affect, when we will add const c to all the rewards in an episodic task → maze running.

- dependent of the value of c → +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)

| | | | | |
|---|---|---|---|---|
| 3.3 | 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 |

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

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

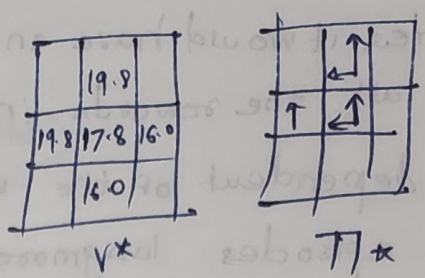
$= \frac{1}{4} \times 0.9 (2.3 + 0.7) = \frac{0.9 \times 3}{4} = \frac{2.7}{4} = 0.675$

$\simeq 0.7$ (approximately)

As you can see in the grid.

b) show - for optimal policy —

Gridworld

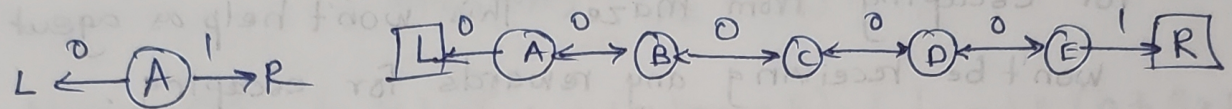



$V_\pi^*$ (centre state) = 17.8

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

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

Q.7) a) Guessing & varifying value fn:-




3 step MDP. All episodes start in the centre st. A. then proceed L/R. with equal probability. When episode terminates on R → Reward +1 all other rewards 0. undiscounted MDP $\gamma = 1$.

Guess: value fn.

→ I guess it will be →

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

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

$V(L) = 0$
$V(R) = +1$

Verify →

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

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

$$= 1/2 \times 1 + 1/2 \times 0 = 1/2.$$

Verified

b) extended MDP → 7 states (right)

Guess & verify value fn.

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

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

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

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

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

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

Similarly,

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

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

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

$V_A = 1/6$ $\quad V_B = 2V_A = 1/3$

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

$V_E = 5V_A = 5/6$

c) arbitary no. states n,

$$V_{\pi(s)}(i) = \frac{i}{n-1} \quad (i = 1, 2, 3 \ldots (n-2))$$

Q.8. solving for value $s^n$.

a) Expand Bellman's eq. for 2 states. in recycling robot, for an arbitrary policy $\pi(a/s)$
discount factor $\gamma$, domain paras $\alpha, \beta$, & $r_{search}, r_{wait}$
$S = \{high, low\}$ ... assume that only 2 charge levels are distinguished.

action sets are then - recharging is always foolish

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

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

$$V(high) = \sum_a \pi(a/s) \sum_{s', r} p(s', r \mid s, a)[r + \gamma V_{\pi(s')}]$$

$$= \pi(search \mid high)[r_{search} + \alpha \cdot \gamma \times V(high) + (1-\alpha) \times \gamma \times V(low)]$$
$$+ \pi(wait \mid high)[1 \times (r_{wait} + \gamma \cdot V(high))]$$

$$V(low) = \pi(search \mid low)[r_{search} + \beta \times \gamma \times V(low) + (1-\beta) \times \gamma \times V(high)] +$$
$$\pi(wait \mid low)[1 \times (r_{wait} + \gamma V(low))]$$
$$+ \pi(recharge \mid low)[0 + 1 \times \gamma V(high)]$$

b) $\alpha = 0.7$ $\quad \gamma = 0.9$ $\quad \pi(\text{search} \mid \text{high}) = 1$

$\beta = 0.6$ $\quad r_{\text{search}} = 10$ $\quad \pi(\text{wait} \mid \text{low}) = 0.5$

$r_{\text{wait}} = 3$ $\quad \pi(\text{recharge} \mid \text{low}) = 0.5$

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

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

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 $\pi(\text{search} \mid \text{Low}) = 0$.

Hence,

$$v(\text{high}) = 1 \times [10 + 0.7 \times 0.9 \times v(\text{high}) + 0.3 \times 0.9\, v(\text{low})]$$

$$= 10 + 0.63\, v(\text{high}) + 0.27\, v(\text{low})$$

$$v(\text{low}) = 0 + 0.5(3 + 0.9\, v(\text{low})) + 0.5(0.9\, v(\text{high}))$$

$$= 1.5 + 0.45\, v(\text{low}) + 0.45\, v(\text{high})$$

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

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

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

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

Solved it using online tool

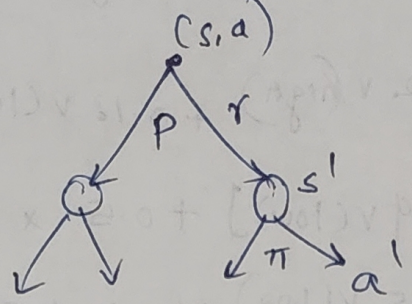
$$v(\text{high}) = \frac{5905}{82}$$

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

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

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

q.9) Action-value fn:




$q_\pi$ back diagram.

a) $V_\pi(s) = E_\pi(G_t \mid s_t = s)$

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

$q_\pi(s,a) = E_\pi(G_t \mid s_t = s, A_t = a)$

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

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

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

$= \sum_{s',r} p(s', r \mid s, a)\,(r + \gamma V_\pi(s'))$

c) $V_\pi(s') = \sum_{a'} \pi(a'|s')\, q(s', a')$

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