AMINA TABASSUM
NUID: 002190127



Reinforcement Learning & Sequential Decision Making

HW #02

Sep 28, 2022

---

Question #01

(a) In this case, state space is coordinate points $(x, y)$ such that:

$$S = \{(x, y)\} \qquad x \in [0, 10] \quad y \in [0, 10]$$

There are four possible actions, so,

$$A = \{\text{'left'}, \text{'right'}, \text{'up'}, \text{'Down'}\}$$

(b) Since, dynamics function is given by:

$$p(s', r \mid s, a) = \{Pr(S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a)\}$$

so, for given state action pairs:

| $s$ | $a$ | $r$ | $s'$ | $p(s', r \mid s, a)$ |
|---|---|---|---|---|
| (0,0) | Down | 0 | (0,0) due to noise | $\alpha$ where $\alpha = 0.9$ |
| (0,0) | Down | 0 | (0,1) | $1-\alpha$ so, $(1-\alpha) = 0.1$ |
| (1,5) | up | 0 | (1,6) | 1 |
| ... | ... | ... | ... | ... |
| (9,10) | Right | 0 | (9,9) | $1-\alpha = 0.1$ |
| (9,10) | Right | 1 | (10,10) | $\alpha$ $(\because \alpha = 0.9)$ |

## Question #02

(a) Since, unified episodic discounting return is given by:

$$G_t = \sum_{K=t+1}^{T} \gamma^{K-t-1} R_K$$

So, episodic discounted return will be:

$$G_t = \gamma^0 R_{t+1} + \gamma^1 R_{t+2} + \cdots \gamma^{T+1-t-1} R_T$$

Here reward = -1 for failure and zero otherwise,

So,

$$G_t = 0 + \gamma^{T-t}(-1)$$

$$\boxed{G_t = -\gamma^{T-t}}$$

where T is timestep when task is failed and hence episode ends.

Similarly,
For continuing case:

$$G_t = -\sum_{K \in K} \gamma^{K-1}$$

where K is set of timesteps where task is failed. Note that this reward will increase in long run, irrespective of improved performance. So, designing it as continuous task does not make sense here.

---

## Question # 02 (b)

We have designed it such that reward = +1 only when it has excited the maze, and are therefore not incentivising it to learn how to exit the maze faster. To do so, we will need to provide a negative reward proportional to time in maze e.g. -1 per timestep.

---

## Question #03 (a)

↳ The sign of reward is not important. Infact, it is interval between each reward that drives behaviour.

↳ Since, eq 3.8 is given by:

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

when we add a constant to all rewards, expected return will just receive constant additive term i.e,

$$G_t = \sum_{K=0}^{\infty} \gamma^K [R_{t+K+1} + c]$$

$$= \sum_{K=0}^{\infty} \gamma^K R_{t+K+1} + \sum_{K=0}^{\infty} \gamma^K c$$

$$G_t = \frac{R_{t+k+1}}{1-\gamma} + \frac{c}{1-\gamma}$$

Since,

$$V_\pi(s) \doteq E[G_t | S_t = s]$$

So,

$$V_c = E[G_t | S_t = s] \qquad \because R_k = c$$

$$= E\left[\sum_{k=0}^{\infty} \gamma^k R_t\right]$$

$$= E\left[\sum_{k=0}^{\infty} \gamma^k c\right]$$

$$V_c = \frac{c}{1-\gamma}$$

---

Question # 03 (b)

In episodic task adding a constant to all rewards does not affect the agent. Here,

$$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$$

cumulative reward depends on length of episode. Timesteps that incur positive rewards act to lengthen the episode and vice versa. In maze running example, we may have chosen to give agent

-1 reward at each timestep to ensure it completes task quickly.

So, adding c=2 at every reward such that reward at each timestep is now positive, the agent is now incentivised to not find exit and continue collecting intermediate rewards indefinitely.

---

Question # 04 (a)

Since, Bellman equation is:

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

Here, probaulty for all 4 actions is equal,

So, $\pi(a|s)\, p(s',r|s,a) = \frac{1}{4} = 0.25$

where $s \in \{$ north, south, east, west $\}$ so,

$r = 0$ except for state 'A' & B

So,

$$V_\pi(s) = (0.25)(0.9 \times 2.3) + (0.25)(0.9 \times 0.7) + (0.25)(0.9 \times 0.4) + (0.25)(0.9 \times -0.4)$$

$$= 0.68 \approx 0.7$$

Hence, it satisfies bellman equation.

## Question # 04 (b)

Since, $V_*(s) = \max_{\pi} V_\pi(s)$

So, for bellman equation

$$V_*(s) = \max_a \left( \sum_{s',r} P(s', r | s, a) [r + \gamma V_*(s') \right)$$

Here, there are two possible actions for center state, north and west. Both of these action result in same state-value, so, we can choose any of these. Hence, probability of each action is 1/2.

$$V_*(s) = 0.5 \times (0.9 \times 19.8) + 0.5 (0.9 \times 19.8)$$

$$= 17.8$$

Hence, it satisfies bellman equation.

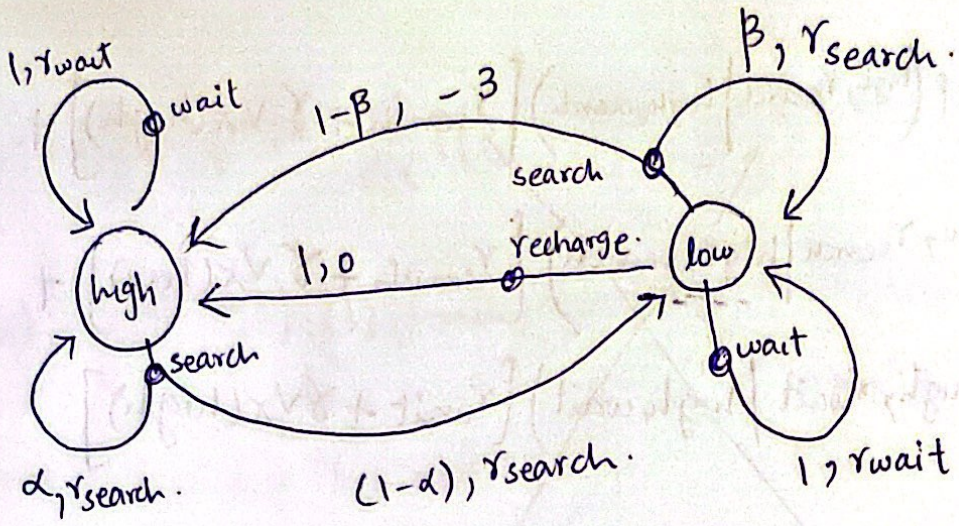
---

## Question (5) (a)

Since, Bellman equation is:

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

Now, for 2-state recycling robot,




In tabular form

| s | a | s' | r | $p(s', r \mid s, a)$ |
|---|---|---|---|---|
| high | search | high | $r_{search}$ | $\alpha$ |
| high | search | low | $r_{search}$ | $1-\alpha$ |
| low | search | high | $-3$ | $1-\beta$ |
| low | search | low | $r_{search}$ | $\beta$ |
| high | wait | high | $r_{wait}$ | 1 |
| high | wait | low | $r_{wait}$ | 1 |
| low | recharge | high | 0 | 1 |

putting values of all these states in bellman equation gives:

So,

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

$$V_\pi(high) = \sum_a \pi(a|s)\Big\{\Big((1)\left(r_{wait} + \gamma V_\pi(high)\right) + \Big(\alpha\left(r_{search} + \gamma V_\pi(high)\right) +$$

$$\Big((1-\alpha)\left(r_{search} + \gamma V_\pi(low)\right)\Big)$$

$$= \sum_a \pi(a|s)\Big\{r_{wait} + \gamma V_\pi(high) + \alpha r_{search} + \alpha\gamma V_\pi(high) + r_{search} +$$

$$\gamma V_\pi(low) - \alpha r_{search} - \alpha\gamma V_\pi(low)\Big]$$

$$V_\pi(high) = \sum_a \pi(a|s)\Big\{r_{wait} + r_{search} + (1+\alpha)\gamma V_\pi(high) + (1-\alpha)\gamma V_\pi(low)\Big\}$$

——— (A)

and

$$V_\pi(low) = \sum_a \pi(a|s)\Big[(1)\left(r_{wait} + \gamma V_\pi(low)\right) + \beta\left(r_{search} + \gamma V_\pi(low)\right) +$$

$$(1-\beta)\left(r_{search} + \gamma V_\pi(high)\right) + (1)\left(0 + \gamma V_\pi(high)\right)\Big]$$

So,

$$V_\pi(low) = \sum_a \pi(a|s)\Big\{r_{wait} + \gamma V_\pi(low) + \beta r_{search} + \beta\gamma V_\pi(low) +$$

$$r_{search} + \gamma V_\pi(high) - \beta r_{search} - \beta\gamma V_\pi(high) +$$

$$\gamma V_\pi(high)\Big\}$$

$$V_\pi(low) = \sum_a \pi(a/s)\Big\{r_{wait} + r_{search} + (1+\beta)\gamma V_\pi(low) + (2-\beta)\gamma V_\pi(high)\Big\}$$

——(B)

Question #05 (b)

putting all values in (A)

$$V_\pi(high) = \sum_a \pi(a/s)\Big[r_{wait} + r_{search} + (1+\alpha)\gamma V_\pi(high) + (1-\alpha)\gamma V_\pi(low)\Big]$$

$$= (1)\Big[3 + 10 + (1+0.8)(0.9)V_\pi(high) + (1-0.8)(0.9)V_\pi(low)\Big]$$

$$= (1)\Big[13 + 1.62 V_\pi(high) + 0.18 V_\pi(low)\Big)$$

$$V_\pi(high) - 1.62 V_\pi(high) - 0.18 V_\pi(low) = 13$$

$$-0.62 V_\pi(high) - 0.18 V_\pi(low) = 13$$

$$0.62 V_\pi(high) + 0.18 V_\pi(low) = -13 \quad \text{——} (1)$$

and putting in eq (B):

$$V_\pi(low) = \sum_a \pi(a/s)\Big\{3 + 10 + (1+0.6)(0.9)V_\pi(low) + (2-0.6)(0.9)V_\pi(high)\Big]$$

$$= \sum_a \pi(a/s)\Big\{13 + 1.44 V_\pi(low) + 1.26 V_\pi(high)\Big\}$$

$$V_\pi(\text{low}) = \pi(\text{wait/low})\left[13 + 1.44\,V_\pi(\text{low}) + 1.26\,V_\pi(\text{high})\right] +$$

$$\pi(\text{recharge/low})\left[13 + 1.44\,V_\pi(\text{low}) + 1.26\,V_\pi(\text{high})\right]$$

$$= 0.5\left[13 + 1.44\,V_\pi(\text{low}) + 1.26\,V_\pi(\text{high})\right] +$$

$$0.5\left[13 + 1.44\,V_\pi(\text{low}) + 1.26\,V_\pi(\text{high})\right]$$

$$V_\pi(\text{low}) = 13 + 1.44\,V_\pi(\text{low}) + 1.26\,V_\pi(\text{high})$$

$$-0.44\,V_\pi(\text{low}) - 1.26\,V_\pi(\text{high}) = 13$$

$$1.26\,V_\pi(\text{high}) + 0.44\,V_\pi(\text{low}) = -13 \quad \text{——— (2)}$$

Solving eq (1) and (2) gives:

$$V_\pi(\text{high}) = -73.47$$

$$V_\pi(\text{low}) = 180.86.$$

# Question #06

(a) The state value function $V_x$ is equal to the expected cumulative return from that state given a distribution of actions. The state-action value function $q_\pi$ is the value of being in a state and taking a deterministic action. Therefore, the state value function is the weighted sum of the state action value function, with weights equal to probabilities of selecting each action:

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

---

(b) Given action $a$, the state-action value function is the probability distributions over possible next states & rewards from action times the one-step reward and distributed state value function at next timestep:

$$q_\pi = \sum_{s' \in S} \sum_{r \in R} p(s', r|s, a)\left[r + \gamma V_x(s_{t+1})\right]$$

(c)

$$q_{\pi}(s,a) = E_{\pi}\left[ G_t | S_t = s, A_t = a \right]$$

$$= E_{\pi}\left[ R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a \right]$$

$$= \sum_{s',r} p(s', r | s, a) \left[ r + \gamma E\left[ G_{t+1} | s', a' \right] \right]$$

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

---