Solution of Reinforcement Learning: An Introduction by Sutton and Barto

Chapter-3 Finite Markov Decision Process

Exercise 3.1

① Mars Rover, it can have multiple objective but for the time being, say it has objective to go from point A to point B.

Rewards

Small +ve reward for each step in the correct direction and
large +ve " " reaching point B.
Small -ve " " moving in wrong direction.

States

The charge of its battery.
The other type of states may be the type of terrain it is on. (like sandy or rocky).

Action

Whether it has to stand and charge itself.
The direction in which it has to go.

Note: S, A, R are not ~~finite~~ discrete here.

② A system that could sense the presence of a person in a room and then adjust the voltage of light and fans, so as to minimize the total power consumption.

Reward

~~-ve~~ -1 for high electricity consumed.
1 " low "
0 " normal "

States

(A, B); A is the prob. of person in room.
B " current power supplied

Action

Whether to inc. or dec. power supplied.

Scanned with CamScanner

Ex 3.2

No, there are situations where $p(s', r \mid s, a)$ can not be defined ~~or~~. is not discrete probability distribution. Example is the Mars Rover in Ex 3.1 1 .

Ex 3.3

The general rule is that anything that can not be changed arbitrarily by the agent is considered a part of its environment.
So, the correct choice for the line b/w agent and env. is accelerator, steering wheel and brakes.

Ex 3.4

$$E_{\pi}[R_{t+1} \mid S_t] = \sum_{a \in A} \pi(a \mid s_t) \cdot \sum_{s' \in S} r \cdot p(s', r \mid s_t, a)$$

Ex 3.6

If $s \neq S_T$

$$\sum_{s' \in S} \sum_{r \in R} p(s', r \mid s, a) = 1$$

If $s = S_T$

$$\sum_{s' \in S} \sum_{r \in R} p(s', r \mid s, a) = 0$$ (As the state S_T is final state and no transcition can be made from it).

Ex 3.7

$R_t = -\gamma^K$ [K is number of time steps until failure]

Assuming that once it has failed, it will be in failure state without external intervention and the task to be non episodic the return would be

$$-\gamma^{K} - \gamma^{K+1} - \dots$$

$$= -\gamma^{K}(1 + \gamma + \gamma^2 + \dots) = \frac{-\gamma^{K}}{1-\gamma} \quad (|\gamma| < 1)$$

Scanned with CamScanner

Ex 3.8. While formulating a MDP for practical purpose one of the things to keep in mind is that the agent should receive reward in regular intervals for effective learning.

A much nicer way to formulate this problem is to give a reward -1 for each time step the agent is inside the maze and a big +ve reward when the agent has escaped the maze.

Ex 3.9 Assuming R_6 & G_6 to be 0.

$$G_5 = R_6 + \gamma G_6$$
$$= R_6 + \frac{G_6}{2} = 0$$

$$G_4 = R_5 + \gamma G_5$$
$$G_4 = 2 + \frac{1}{2} G_5 = 2$$
$$G_3 = 3 + \frac{G_4}{2} = 4$$
$$G_2 = 6 + \frac{G_3}{2} = 8$$
$$G_1 = 2 + \frac{G_2}{2} = 6$$
$$G_0 = -1 + \frac{G_1}{2} = 2.$$

Ex 3.10

$$G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + \dots$$
$$= 2 + \gamma [7 + \gamma 7 + \gamma^2 7 \dots]$$
$$= 2 + \frac{0.9 \times 7}{1 - 0.9}$$
$$= 2 + 63$$
$$= 65$$

$$G_1 = 7 + \gamma 7 + \gamma^2 7 + \dots$$
$$= \frac{63}{0.9} = 70$$

Scanned with CamScanner

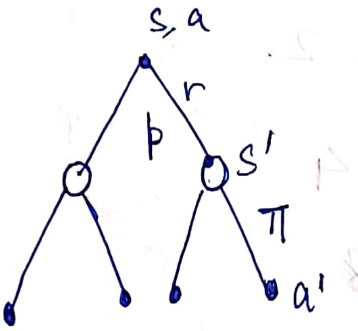
Ex 3.11

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$
$$= 1 + \gamma + \gamma^2 + \gamma^3 + \cdots$$
$$= \frac{1}{1-\gamma} \quad (|\gamma| < 1)$$

Ex 3.12

$$0.7 = \frac{1}{4}\left[0 + 0.9 \times 0.7\right] + \frac{1}{4}\left[0 + 0.9 \times 2.3\right]$$
$$+ \frac{1}{4}\left[0 + \cancel{0.9 \times 0.4}\right] + \frac{1}{4}\left[0 - \cancel{0.9 \times 0.4}\right]$$
$$= \frac{0.9}{4}\left[3\right] = \frac{2.7}{4} = 0.68 \approx 0.7$$

Ex 3.13



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

Ex 3.14

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

$$G_t' = (R_{t+1} + c) + \gamma(R_{t+2} + c) + \gamma^2(R_{t+2} + c) + \cdots$$
$$= G_t + c + c\gamma + c\gamma^2 + \cdots$$

$$G_t' = G_t + \frac{c}{1-\gamma}$$
$$V_c = \frac{c}{1-\gamma}$$

Scanned with CamScanner

Ex 3.15 Yes, it would have effect as we get $v_c = \frac{c}{1-\gamma}$ for of states only if the task is on episodic, when the task is episodic v_c will be different for all the states.

Ex 3.16

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

Ex 3.17

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

or

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

Ex 3.20

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

$$q_*(\text{high},\text{wait}) = r_{\text{wait}} + \gamma \max(q_*(\text{high},\text{wait}), q_*(\text{high},\text{search}))$$

$$q_*(\text{high},\text{search}) = \alpha\left[r_{\text{search}} + \gamma \max(q_*(\text{high},\text{wait}), q_*(\text{high},\text{search}))\right] + (1-\alpha)\left[r_{\text{search}} + \gamma \max(q_*(\text{low},\text{wait}), q_*(\text{low},\text{recharge}), q_*(\text{low},\text{search}))\right]$$

Scanned with CamScanner

Ex 3.22

$\gamma = 0$

$q(q_0, \text{left}) = 1 + 0 = 1$

$q(q_0, \text{right}) = 0.$

$1 > 0$

Correct choice left.

$\gamma = 0.9$

$q(q_0, \text{left}) = 1 \times (1 + 0.9(0)) = \phi$

$q(q_0, \text{right}) = 0 + 0.9(2) = 1.8$

$1.8 > 1 \Rightarrow$ right.

$\gamma = 0.5$

$q(q_0, \text{left}) = 1 \times (1 + 0.5 \times 0) = 1$

$q(q_0, \text{right}) = 1 \times (0 + 0.5 \times 2) = 1$

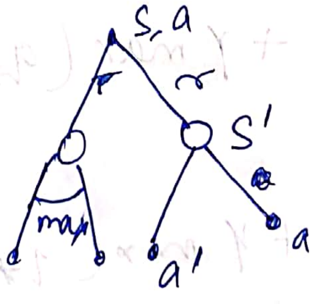
$1 == 1$ (any direction)

Ex 3.23

$$v_*(s) = \max_a (q_*(s,a))$$

Ex 3.24

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



Ex 3.25

$$a_* \doteq \arg\max_a q_*(s,a)$$

Ex 3.26 $a_* = \arg\max_a \sum_{s',r} p(s',r|s,a)[r + v_*(s')]$