Finite MDP Solutions

3.1

Answer

1. A robot that is attempting to reach a object, the actions can be the angle selected, states are the current place, rewards is +1 if the object is touch, else 0.
2. Pacman game, actions can be the movement directions, states are the places and the bullet places, rewards are positive scores getted by eating bullets.
3. Puzzle game, actions can be a mapping from patches to places, states are the whole current map, rewards are the positive scores if the place is right.

3.2

Answer

No, MDP framework is not adequate to the problems that the decision depends not only on current but also the previous state, such as the prediction of weather.

3.3

Answer

1. It depends on the tasks.
2. The evaluation can focus on how it is close to the decision part of the problem.
3. Actions have to be things that the agent can actually control.

3.4

Answer

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

3.5

Answer

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

3.6

Answer

- Episodic case:

  $$

  ```
  G_t = -{\gamma}^{T-t}
  ```

  $$
- Continous case:

  $$

  ```
  G_t = \sum_{k \in K} {\gamma}^{k-t}
  ```

  $$

3.7

Answer

- Firstly, under most strategies the reward shown in **3.7** will be 1, which encourages the agent to move randomly.
- Secondly, in some cases, the agent will not escape from the maze.
- Thirdly, the real aim is exactly to make the agent escape from the maze as quickly as possible.

3.8

Answer

$$\begin{aligned} G_0 = 2G_1 = 6 \\ G_2 = 8 \\ G_3 = 4 \\ G_4 = 2 \\ G_5 = 0 \end{aligned}$$

3.9

Answer

$$G_1 = \sum_{t=2}^{\infty} \gamma^{(t-2)} \times 7 = \frac{7\gamma}{1-\gamma}$$

$$G_0 = R_1 + \sum_{t=2}^{\infty} \gamma^{(t-2)} \times 7 = 2 + \frac{7\gamma}{1-\gamma}$$

3.10

Answer

By the formulation of the sum of geometric sequence, $G_t = \sum_{k=0}^{\infty} y^k = \lim_{N \to \infty} \frac{1-\gamma^N}{1-\gamma} = \frac{1}{1-\gamma}$

3.11

Answer

$$E_\pi[R_{t+1}|S_t=s]=\sum_a \pi(a|s)\sum_{s',r} rp(s',r|s,a)$$

3.12

Answer

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

3.13

Answer

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

3.14

Answer

$$0.7 \approx \frac{(2.3+0.4-0.4+0.7)*0.9}{4}=0.675$$

3.15

Answer

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

Since v_c is added to every state's value, the relative values are unchanged.

3.16

Answer

$$v_c=c\frac{1-\gamma^T}{1-\gamma}$$

So the change depends on the termination time of every episode.

3.17

Answer

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

3.18

Answer

$$\begin{aligned} v_\pi(s) &= E[q_\pi(S_t,A_t)|S_t=s,A_t=a] \\ &= \sum_a \pi(a|s)q_\pi(s,a) \end{aligned}$$

3.19

Answer

$$
\begin{aligned}
q_\pi(s,a) &= E[R_{t+1} + v_\pi(s') | S_t = s, A_t = a] \\
&= \sum_{s',r} p(s', r | s, a)[r + v_\pi(s')]
\end{aligned}
$$

3.20

Answer

It is apparent that the optimal policy is to putt in the green area, and to drive off the green area.

So the optimal state-value function is using the function of putt when in the green area, and the function of driver when off the green area.

3.21

Answer

As described in **3.20**.

3.22

Answer

- γ is 0: π_{left} is optimal.
- γ is 0.9: π_{right} is optimal.
- γ is 0.5: both are optimal.

3.23

Answer

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

3.24

Answer

$$\gamma = \frac{16.0}{17.8} = 0.9$$
$$24.4 = 10 + 16\gamma$$

3.25

Answer

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

3.26

Answer

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

3.27

Answer

$$\pi_*(a|s) = \mathtt{1}\{a = argmax_{a'} q_*(s,a')\}$$

3.28

Answer

$$\pi_*(a|s) = \mathtt{1}\{a = argmax_{a'} \sum_{s',r} p(s', r|s, a')[r + \gamma v_*(s')]\}$$

3.29

Answer

$$\begin{aligned}
v_\pi(s) &= \sum_a \pi(a|s) \sum_{s'} p(s'|s, a)[r(s, a) + \gamma v_\pi(s')] \\
v_*(s) &= \max_a \sum_{s'} p(s'|s, a)[r(s, a) + \gamma v_*(s')] \\
q_\pi(s, a) &= \sum_{s'} p(s'|s, a)[r(s, a) + \gamma \sum_{a'} \pi(a'|s') q_\pi(s', a')] \\
q_*(s, a) &= \sum_{s'} p(s'|s, a)[r(s, a) + \gamma \max_{a'} q_*(s', a')]
\end{aligned}$$