Zijun, Wu   1488834

Question 1

(a) X, Left, 0, $\overbrace{\text{X, Left, 0}}^{\text{loop}}$, ...

(b) X, right, +1, X, right, -1, Y, right, +3, Absorbing state

(c) $G_0 = R_1 + \gamma R_2 + \gamma^2 R_3$
$= 1 + 0.5 \times (-1) + 0.5^2 \times 3$
$= 1 - 0.5 + 0.75$
$= 1.25$

(d) $V_{\pi_1}(Y) = E_{\pi_1}[G_t \mid S_t = Y] = 3$

(e) $q_{\pi_1}(X, \text{left}) = E_{\pi_1}[0 + 0.5 \times 0 + 0.5^2 \times 0 + \cdots \mid S_t = X, A_t = \text{Left}]$
$= 0$

(f) $V_{\pi_2}(X) = \sum_a \pi_2(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma V_\pi(s')]$
$= \frac{2}{3}[1 + 0.5 V_{\pi_2}(X)] + \frac{1}{3}[-1 + 0.5 V_{\pi_2}(Y)]$
$= \frac{2}{3}[1 + 0.5 V_{\pi_2}(X)] + \frac{1}{3}(-1 + 0.5 \times 3)$
$= \frac{2}{3} + \frac{1}{3} V_{\pi_2}(X) + \frac{1}{6}$

So $V_{\pi_2}(X) = \frac{5}{4}$

2(a)

Example 1. Using reinforcement learning to playing football. The agent's position in the court is the states. The actions is the agent's direction of carring the ball. If the ball is stolen, the reward is -1, otherwise, the reward is 0, if the agent carries the ball to the goal, then the reward is +1.

Example 2. A robot is on a bicycle learns how to ride it. It can control the speed and change its center of gravity to make itself not fall over. It need to go as far as possible, if it falls over, getting -1 reward, othermise, is 0. So the states are its speed and its relative gravity center position. But the limits is the agent cannot learn with the uncertain forward, for example, a rock, a pitting, which makes it fall over.

Example 3. An auto drive system learn how to drive safely, the state is the latest data by sensors like radar and cameras. It takes action to control the direction and speed. If it drives safely without any collision, then gets 0 reward, else, gets -1.

(b) We set the reward wrong, because if we only give reward of +1 for escaping from maze, the agent can get it no matter how many steps it takes. No reward for escaping with less time step. $G_t = 0+0+\cdots+1 = 1$ for all time steps. What we need to correct is to set reward of -0.5 at all other time. Then the agent can get the feedback and learn to escape out of the maze as quickly as possible.

(c). $G_4 = R_5 + r G_5$

Since $T = 5$, $G_5 = R_6 + rR_7 + \cdots = 0 + 0 + 0 + \cdots = 0$

$G_4 = 2 + 0.5 \times 0 = 2$.

$G_3 = R_4 + rG_4 = 3 + 0.5 \times 2 = 4$

$G_2 = R_3 + rG_3 = 6 + 0.5 \times 4 = 8$

$G_1 = R_2 + rG_2 = 2 + 0.5 \times 8 = 6$

$G_0 = R_1 + rG_1 = -1 + 0.5 \times 6 = 2$

(d) $G_1 = R_2 + rR_3 + r^2 R_4 + \cdots r^n R_{n+2} + \cdots$

$= 7 + 0.9 \times 7 + (0.9)^2 \times 7 + \cdots + (0.9)^n \times 7 + \cdots$

$= \frac{7}{1 - 0.9}$

$= 70$.

$G_0 = R_1 + rG_1$

$= 2 + 0.9 \times 70$

$= 65$

(e) $V_\pi(\text{center}) = \frac{1}{4}[0 + 0.9\, V_\pi(\text{Left})] + \frac{1}{4}[0 + 0.9\, V_\pi(\text{right})]$

$+ \frac{1}{4}[0 + 0.9\, V_\pi(\text{up})] + \frac{1}{4}[0 + 0.9\, V_\pi(\text{down})]$

$= \frac{1}{4}(0 + 0.9 \times 0.7) + \frac{1}{4}[0 + 0.9 \times 0.4]$

$+ \frac{1}{4}(0 + 0.9 \times 2.3) + \frac{1}{4}[0 + 0.9 \times (-0.4)]$

$= 0.1575 + 0.5175$

$= 0.675 \approx 0.7$

(f) $G_t' = (R_{t+1} + C) + \gamma(R_{t+2} + C) + \gamma^2(R_{t+3} + c) + \cdots$

$= \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c)$

$= \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1}) + \sum_{k=0}^{\infty} \gamma^k c$, $G_t = R_{t+1} + \gamma R_{t+2} + \cdots$

$V_c = G_t' - G_t = \sum_{k=0}^{\infty} \gamma^k c = \frac{c}{1-\gamma}$

Only the intervals between the rewards are important.

(g) $V_c = \sum_{k=0}^{T} \gamma^k c$

In each episode, the sum doesn't go into infinity. For example, the maze problem. In each episode the agent learn to escape, the number of steps in each episode is determined by the actions agent takes, so the V_c could alter for G_t, become a changing factor for G_t, which may cause the learning time be longer.

(h) $q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = a]$

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

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

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

(i) $V_\pi(s) = E_\pi[G_t \mid S_t = s]$

$= \sum_a \pi(a \mid s) \cdot q_\pi(s, a)$

(j) $V_*(A) = \max_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma V_*(s')]$

$$= \max \begin{cases} p(S_{left}, r \mid A, left)[0 + 0.9 \times V_*(S_{left})], \\ p(S_{right}, r \mid A, right)[0 + 0.9 \times V_*(S_{right})], \\ p(A, r \mid A, up)[-1 + 0.9 \times V_*(A)], \\ p(A', r \mid A, down)[10 + 0.9 \times V_*(A')] \end{cases}$$

$$= \max \begin{Bmatrix} 0 + 0.9 \times 22.0, \\ 0 + 0.9 \times 22.0, \\ -1 + 0.9 \times V_*(A), \\ 10 + 0.9 \times 16 \end{Bmatrix}$$

$= \max\{19.8, 19.8, -10, 24.4\}$

$= 24.4$

Bonus, Question 3.

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

$$= 0.5 \times [0.75 \times [3+0.8\times 2] + 0.25 \times [-6+0.8\times 7]] +$$
$$0.5 \times [0.2 \times [-3+0.8\times(-1)] + 0.8 \times [4 + 0.8 \times V_\pi(top)]]$$

$$= 0.5 \times 3.35 + 0.5[-0.76 + 3.2 + 0.64 V_\pi(top)]$$

$$= 1.675 + 1.22 + 0.32 V_\pi(top)$$

$$V_\pi = 4.26$$

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

$$= \max \begin{Bmatrix} 0.75 \times (3+0.8\times 2) + 0.25 \times (-6+0.8\times 7), \\ 0.2 \times (-3+0.8\times(-1)) + 0.8 \times (4 + 0.8 V_*) \end{Bmatrix}$$

$$= \max \{ 3.35, \ 2.44 + 0.64 V_* \}$$

$$= 6.78$$

Question 4.

The return at each time is the negative probability the pole will fall. In discounted, continuing task, the return at each time is related to $-\gamma^k$, where k are the time step of falling.