Exercise 2: Markov Decision Processes (MDPs) Hongyan Yang

1. Formulating an MDP

(a). Considering the four-rooms from Ex 0:

state spaces $S = \{$ coordinates $(x, y) : x \in [0, 10], y \in [0, 10], x \in \mathbb{Z}, y \in \mathbb{Z} \}$

action spaces $A = \{$ UP, DOWN, LEFT, RIGHT $\}$

(b). According to the information given,

For $\{(0,0), \text{DOWN}\}$: $\begin{cases} p((1,0), 0 \mid s, a) = 0.1 \\ p((0,0), 0 \mid s, a) = 0.9 \end{cases}$

For $\{(1,5), \text{UP}\}$: $\begin{cases} P((1,5), 0 \mid s, a) = 0.2 \\ P((1,6), 0 \mid s, a) = 0.8 \end{cases}$

For $\{(9,10), \text{RIGHT}\}$: $\begin{cases} P((9,10), 0 \mid s, a) = 0.1 \\ P((9,9), 0 \mid s, a) = 0.1 \\ P((10,10), 1 \mid s, a) = 0.8 \end{cases}$

2. The RL objective:

(a) Based on the information given,

For an episodic task: $G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots + \gamma^{T-t-1} R_T$

$\therefore$ with all rewards zero except for -1 upon failure,

$G_t = -\gamma^{T-t-1}$ at each time t

It's the same from that in the discounted, continuing formulation

$G_t = -\gamma^{K-1}$, where K is the number of timesteps before failure

(b) Using expected total reward without a discount factor γ will yield same result for episodes with different lengths, thus can't inspire the robot to escape the maze faster. It's better to add discount factor γ to the total return or add a small negative reward for every step the robot takes before escape the maze.

3. Modifying the reward function

(a) According to equation 3.8 $G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

After adding a constant C to all the rewards

$$G_t' = \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + C) = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \sum_{k=0}^{\infty} \gamma^k C = G_t + \sum_{k=0}^{\infty} \gamma^k C$$

$$= G_t + \frac{1}{1-\gamma} C$$

$\therefore$ Adding a constant C to all the rewards adds a constant V_c to all states, $V_c = \frac{1}{1-\gamma} C$

(b) Adding a constant C to all the rewards in an episodic task will change the task in the continuing task above. Because $G_t' \doteq \sum_{k=t+1}^{T} \gamma^{k-t-1}(R_k + C)$

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

~~Give maze running as an Example,~~

$\therefore$ Earlier steps will be compensated by a larger add-on reward,

A constant C large enough could lead the agent to collect more reward in the earlier steps, thus escape the maze slower.

4. Bellman Equation

(a) According to the information given.

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

$$= \frac{1}{4} \times (0 + 2.3 \times 0.9) + \frac{1}{4}(0 + 0.4 \times 0.9) + \frac{1}{4}\left[0 + (-0.4) \times 0.9\right] + \frac{1}{4}(0 + 0.7 \times 0.9)$$

$$= \frac{1}{4}(2.3 + 0.7) \times 0.9 = \frac{3}{4} \times \frac{9}{10} = \frac{7.5 - 0.75}{10} = 0.675 \approx 0.7$$

4(b) Based on the ~~information~~ given,

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

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

$$= 19.8\times 0.9 = 17.82 \approx 17.8$$

5. Solving for the value function

(a) Based on the information given,

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

$$= \pi(\text{wait}|\text{high})[r_{\text{wait}} + \gamma V_\pi(\text{high})] + \pi(\text{search}|\text{high})\cdot\Big[\alpha(r_{\text{search}} + \gamma V_\pi(\text{high})) + (1-\alpha)(r_{\text{search}} + \gamma V_\pi(\text{low}))\Big]$$

$$V_\pi(\text{low}) = \pi(\text{recharge}|\text{low})\cdot[0+\gamma V_\pi(\text{high})] + \pi(\text{search}|\text{low})\cdot[\beta(r_{\text{search}}+\gamma V_\pi(\text{low})) + (1-\beta)(-3+\gamma V_\pi(\text{high}))] + \pi(\text{wait}|\text{low})\cdot[r_{\text{wait}}+\gamma V_\pi(\text{low})]$$

(b) Based on the information given,

$$\begin{cases} V_\pi(\text{high}) = 0.8[10+0.9V_\pi(\text{high})] + 0.2[10+0.9V_\pi(\text{low})] & ① \\ V_\pi(\text{low}) = 0.5[3+0.9V_\pi(\text{low})] + 0.5[0+0.9V_\pi(\text{high})] & ② \end{cases}$$

$$\therefore \begin{cases} V_\pi(\text{high}) \approx 79.041 \\ V_\pi(\text{low}) \approx 67.397 \end{cases}$$

Check with Bellman equation:

$$V_\pi(\text{high}) = \sum_a \pi(a|s) \sum_{s',r} (s',r|s,a)[r+\gamma V_\pi(s')] = 0.8[10+0.9\times 79.041] + 0.2[10+0.9\times 67.397] \approx 79.041$$

$$V_\pi(\text{low}) = 0.5[3+0.9\times 67.397] + 0.5[0+0.9\times 79.041] \approx 67.397$$

$\therefore$ It satisfies ~~with~~ the Bellman Equation

6. Action-value function

(a) V_π in terms of q_π and π: $V_\pi(s) = \sum_a \pi(a|s) \cdot q_\pi(s,a)$

(b) q_π in terms of V_π and the four-argument p:

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

(c) $\because$ $\begin{cases} q_\pi(s,a) = \sum_{s',r} p(s',r|s,a) \cdot [r + \gamma V_\pi(s')] & ① \\ V_\pi(s') = \sum_{a'} \pi(a'|s') q_\pi(s',a') & ② \end{cases}$

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