Chenghao Wang

Q1. (a). State space: $S = \{ S_{blank} \}$

Actionspace: $A = \{ a_{up}, a_{Down}, a_{left}, a_{right} \}$

(b) Dynamic of $\{(0,0), DOWN\}$

$$P\{(0,0), 0 \mid (0,0), DOWN\} = 0.9$$
$$P\{(0,1), 0 \mid (0,0), DOWN\} = 0.1$$

Dynamic of $\{(1,5), UP\}$

$$P\{(1,6), 0 \mid (1,5), UP\} = 0.8$$
$$P\{(1,5), 0 \mid (1,5), UP\} = 0.2$$

Dynamic of $\{(9,10), RIGHT\}$

$$P\{(10,10), 1 \mid (9,10), RIGHT\} = 0.8$$
$$P\{(9,10), 0 \mid (9,10), RIGHT\} = 0.1$$
$$P\{(9,9), 0 \mid (9,10), RIGHT\} = 0.1$$

Q2 (a) $G = R_0 + \gamma R_1 + \gamma^2 R_2 + \cdots + \gamma^T R_T = -\gamma^T$

where T is terminal time step

In countinue task, having a discountinue factor could be helpful because it prevents blowing up. But in epsodic task there is no need to having discoutinue factor.

(b) Based on the reward setting, the goal for agent here is TO ESCAPING THE MAZE, rather then ESCAPING THE MAZE AS FAST AS POSSIBLE. So there shows no improvement for the training process. If a small negative reward is added to each non terminal step, the effectively communtation could be achieved.

Q3 (a) The sign of those rewards is important when the task is episodic, and the sign become less important for a continuing task.

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

$$\Rightarrow G_{tc} = (R_{t+1}+C) + \gamma(R_{t+2}+c) + \gamma^2(R_{t+3}+C) + \cdots$$

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

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

$$V_\pi(s) = E_\pi[G_t | S_t = s] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s\right]$$

$$V_{\pi c}(s) = E_\pi\left[G_t + \sum_{k=0}^{\infty} \gamma^k C \middle| S_t = s\right] = V_\pi(s) + \sum_{k=0}^{\infty} \gamma^k C$$

$$= V_\pi(s) + \frac{C}{1-\gamma}$$

(b) This would have effect. For example. a maze. the reward for all non-exit grids is -0.1, the reward for exit is 5. if we add a constant 6 to all the rewards. the agent will never escape the maze.

4\. (a) $U_\pi(\text{center}) = P((\text{center UP}), 2.3 \mid \text{center}, UP) \times [0 + 0.9 \times 2.3]$

$+ P((\text{center LEFT}), 0.7 \mid \text{center}, LEFT) \times [0 + 0.9 \times 0.7]$

$+ P((\text{center RIGHT}), 0.4 \mid \text{center}, RIGHT) \times [0 + 0.9 \times 0.4]$

$+ P((\text{center Down}), -0.4 \mid \text{center}, DOWN) \times [0 + 0.9 \times -0.4]$

$= 0.25 \times 2.07 + 0.25 \times 0.63 + 0.25 \times 0.36 - 0.25 \times 0.36$

$= 0.5175 + 0.1575$

≈ 0.7

(b) $V_{\pi}(\text{center}) = P((\text{center UP}), 19.8 \mid \text{center}, UP) \times [0 + 0.9 \times 19.8]$

$+ P(\text{center LEFT}, 19.8 \mid \text{center}, LEFT) \times [0 + 0.9 \times 19.8$

$= 0.5 \times 0.9 \times 19.8 + 0.5 \times 0.9 \times 19.8$

$= 17.8$

5 (a)

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

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

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

(b)

$$V_{\pi}(high) = \pi(search|high) \cdot \alpha \times [r_{search} + \gamma V_{\pi}(high)]$$
$$+ \pi(search|high) \cdot (1-\alpha) \cdot [r_{search} + \gamma V_{\pi}(low)]$$
$$+ \pi(wait|high) \cdot [r_{wait} + \gamma V_{\pi}(high)]$$
$$= 1 \times 0.8 \times [10 + 0.9 V_{\pi}(high)]$$
$$+ 1 \times 0.2 \times [10 + 0.9 V_{\pi}(low)]$$

$$V_{\pi}(high) = 8 + 0.72 V_{\pi}(high) + 2 + 0.18 V_{\pi}(low)$$

$$0.28 V_{\pi}(high) = 10 + 0.18 V_{\pi}(low) \quad ①$$

$$V_{\pi}(low) = \pi(wait|low) \cdot [r_{wait} + \gamma V_{\pi}(low)]$$
$$+ \pi(search|low) \cdot \beta [r_{search} + \gamma V_{\pi}(low)]$$
$$+ \pi(recharge|low) \cdot \gamma V_{\pi}(high)$$
$$+ \pi(search|low) \cdot (1-\beta) \cdot [-3 + \gamma V_{\pi}(high)]$$
$$= 0.5 \times [3 + 0.9 \times V_{\pi}(low)]$$
$$+ 0.5 \times 0.9 \times V_{\pi}(high)$$

$$V_{\pi}(low) = 1.5 + 0.45 V_{\pi}(low) + 0.45 V_{\pi}(high)$$

$$0.55 V_{\pi}(low) = 1.5 + 0.45 V_{\pi}(high)$$

$$11 V_{\pi}(low) = 30 + 9 V_{\pi}(high) \quad ② \quad \Rightarrow 77 V_{low} = 210 + 63 V_{high}$$

Combine ① and ②

$$\Rightarrow 14 V_{\pi}(high) = 500 + 9 V_{\pi}(low) \quad \Rightarrow 63 V_{high} = 2250 + 40.5 V_{low}$$

$$\Downarrow$$

$$36.5 V_{low} = 2460$$

$\Rightarrow V_{low} = 67$

$\Rightarrow V_{high} = 79$

$14 V_{high} = 500 + 603$

6 (a) $v_\pi = E_\pi[G_t \mid S_t = s]$

$$= \sum_a E_\pi[G_t \mid S_t = s, A_t = a] \, P[A_t = a \mid S_t = s]$$

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

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

$$= \sum_{s'} [G_t \mid S_t = s, A_t = a, S_{t+1} = s'] \, \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a]$$

$$= \sum_{s'} [G_t \mid S_{t+1} = s'] \, \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a]$$

$$= \sum_{s', r} \{ r + \gamma E_\pi[G_{t+1} \mid S_{t+1} = s'] \} \, p(s', r \mid s, a)$$

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

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

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