

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

Chenghao Wang

Action space: $A = \{A_{up}, A_{down}, A_{left}, A_{right}\}$

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

$$P\{(0,0), 0 | (0,0), DOWN\} = 0.9$$

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

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

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

$$P\{(1,5), 0 | (1,5), UP\} = 0.2$$

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

$$P\{(10,10), 1 | (9,10), RIGHT\} = 0.8$$

$$P\{(9,10), 0 | (9,10), RIGHT\} = 0.1$$

$$P\{(9,9), 0 | (9,10), RIGHT\} = 0.1$$

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

where T is terminal time step

In continue task, having a discount factor could be helpful because it prevents blowing up. But in episodic task there is no need to having discount factor.

(b) Based on the reward setting, the goal for agent here is TO ESCAPING THE MAZE, rather than 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.

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

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

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

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

$$V_a(s) = E_a[G_t | S_t = s] = E_a\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right]$$

$$V_{ac}(s) = E_a\left[G_t + \sum_{k=0}^{\infty} \gamma^k C \mid S_t = s\right] = V_a(s) + \sum_{k=0}^{\infty} \gamma^k C$$

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

(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 b to all the rewards, the agent will never escape the maze.

4. (a)
$$\begin{aligned}
 V_{\pi}(\text{center}) &= P((\text{centerUP}), 2.3 | \text{center}, \text{UP}) \times [0 + 0.9 \times 2.3] \\
 &\quad + P((\text{centerLEFT}), 0.7 | \text{center}, \text{LEFT}) \times [0 + 0.9 \times 0.7] \\
 &\quad + P((\text{centerRIGHT}), 0.4 | \text{center}, \text{RIGHT}) \times [0 + 0.9 \times 0.4] \\
 &\quad + P((\text{centerDOWN}), -0.4 | \text{center}, \text{DOWN}) \times [0 + 0.9 \times -0.4] \\
 &= 0.25 \times 2.07 + 0.25 \times 0.63 + 0.25 \times 0.36 - 0.25 \times 0.36 \\
 &= 0.6125 + 0.1575 \\
 &\approx 0.7
 \end{aligned}$$

(b)
$$\begin{aligned}
 V_{\pi}(\text{center}) &= P((\text{centerUP}), 19.8 | \text{center}, \text{UP}) \times [0 + 0.9 \times 19.8] \\
 &\quad + P((\text{centerLEFT}), 19.8 | \text{center}, \text{LEFT}) \times [0 + 0.9 \times 19.8] \\
 &= 0.5 \times 0.9 \times 19.8 + 0.5 \times 0.9 \times 19.8 \\
 &= 17.8
 \end{aligned}$$

5 (a)

$$\begin{aligned}
 U_{\bar{a}}(\text{high}) &= \pi(\text{search}|\text{high}) \cdot P(\text{high}, r_{\text{search}}|\text{high}, \text{search}) \times [r_{\text{search}} + \gamma U_{\bar{a}}(\text{high})] \\
 &\quad + \pi(\text{search}|\text{high}) \cdot P(\text{low}, r_{\text{search}}|\text{high}, \text{search}) \times [r_{\text{search}} + \gamma U_{\bar{a}}(\text{low})] \\
 &\quad + \pi(\text{wait}|\text{high}) \cdot P(\text{high}, r_{\text{wait}}|\text{high}, \text{wait}) \times [r_{\text{wait}} + \gamma U_{\bar{a}}(\text{high})] \\
 &= \pi(\text{search}|\text{high}) \cdot \alpha \times [r_{\text{search}} + \gamma U_{\bar{a}}(\text{high})] \\
 &\quad + \pi(\text{search}|\text{high}) \cdot (1-\alpha) \cdot [r_{\text{search}} + \gamma U_{\bar{a}}(\text{low})] \\
 &\quad + \pi(\text{wait}|\text{high}) \cdot [r_{\text{wait}} + \gamma U_{\bar{a}}(\text{high})]
 \end{aligned}$$

$$\begin{aligned}
 U_{\bar{a}}(\text{low}) &= \pi(\text{wait}|\text{low}) \cdot P(\text{low}, r_{\text{wait}}|\text{low}, \text{wait}) \times [r_{\text{wait}} + \gamma U_{\bar{a}}(\text{low})] \\
 &\quad + \pi(\text{search}|\text{low}) \cdot P(\text{low}, r_{\text{search}}|\text{low}, \text{search}) \times [r_{\text{search}} + \gamma U_{\bar{a}}(\text{low})] \\
 &\quad + \pi(\text{recharge}|\text{low}) \cdot P(\text{low}, 0|\text{low}, \text{recharge}) \times [0 + \gamma U_{\bar{a}}(\text{high})] \\
 &\quad + \pi(\text{search}|\text{low}) \cdot P(\text{high}, r_{\text{search}}|\text{low}, \text{search}) \times [-3 + \gamma U_{\bar{a}}(\text{low})] \\
 &= \pi(\text{wait}|\text{low}) \cdot [r_{\text{wait}} + \gamma U_{\bar{a}}(\text{low})] \\
 &\quad + \pi(\text{search}|\text{low}) \cdot \beta [r_{\text{search}} + \gamma U_{\bar{a}}(\text{low})] \\
 &\quad + \pi(\text{recharge}|\text{low}) \cdot \gamma U_{\bar{a}}(\text{high}) \\
 &\quad + \pi(\text{search}|\text{low}) \cdot (1-\beta) \cdot [-3 + \gamma U_{\bar{a}}(\text{low})]
 \end{aligned}$$

(b)

$$\begin{aligned}U_a(\text{high}) &= \pi(\text{search}|\text{high}) \cdot \alpha \cdot [r_{\text{search}} + \gamma U_a(\text{high})] \\&\quad + \pi(\text{search}|\text{high}) \cdot (1-\alpha) \cdot [r_{\text{search}} + \gamma U_a(\text{low})] \\&\quad + \pi(\text{wait}|\text{high}) \cdot [r_{\text{wait}} + \gamma U_a(\text{high})]\end{aligned}$$

$$= 1 \times 0.8 \times [10 + 0.9 U_a(\text{high})]$$

$$+ 1 \times 0.2 \times [10 + 0.9 U_a(\text{low})]$$

$$U_a(\text{high}) = 8 + 0.72 U_a(\text{high}) + 2 + 0.18 U_a(\text{low})$$

$$0.28 U_a(\text{high}) = 10 + 0.18 U_a(\text{low}) \quad (1)$$

$$\begin{aligned}U_a(\text{low}) &= \pi(\text{wait}|\text{low}) \cdot [r_{\text{wait}} + \gamma U_a(\text{low})] \\&\quad + \pi(\text{search}|\text{low}) \cdot \beta \cdot [r_{\text{search}} + \gamma U_a(\text{low})] \\&\quad + \pi(\text{recharge}|\text{low}) \cdot \gamma U_a(\text{high}) \\&\quad + \pi(\text{search}|\text{low}) \cdot (1-\beta) \cdot [-3 + \gamma U_a(\text{high})] \\&= 0.5 \times [3 + 0.9 \times U_a(\text{low})] \\&\quad + 0.5 \times 0.9 \times U_a(\text{high})\end{aligned}$$

$$U_a(\text{low}) = 1.5 + 0.45 U_a(\text{low}) + 0.45 U_a(\text{high})$$

$$0.55 U_a(\text{low}) = 1.5 + 0.45 U_a(\text{high})$$

$$11 U_a(\text{low}) = 30 + 9 U_a(\text{high}) \quad (2) \Rightarrow 77 U_{\text{low}} = 210 + 63 U_{\text{high}}$$

Combine (1) and (2)

$$\Rightarrow 14 U_a(\text{high}) = 500 + 9 U_a(\text{low}) \Rightarrow 63 U_{\text{high}} = 2250 + 40.5 U_{\text{low}}$$

$$\Downarrow \\ 36.5 U_{\text{low}} = 2460$$

$$\Rightarrow U_{low} = 67$$

$$\Rightarrow U_{high} = 79$$

$$14 U_{high} = 500 + 603$$

$$\begin{aligned} 6 \text{ (a)} \quad U_{\bar{a}} &= E_{\bar{a}}[G_t | S_t = s] \\ &= \sum_a E_{\bar{a}}[G_t | S_t = s, A_t = a] P[A_t = a | S_t = s] \\ &= \sum_a q_{\bar{a}}(s, a) \pi(a|s) \end{aligned}$$

$$\begin{aligned} \text{(b)} \quad q(s, a) &= E_{\bar{a}}[G_t | S_t = s, A_t = a] \\ &= \sum_{s'} [G_t | S_t = s, A_t = a, S_{t+1} = s'] P[S_{t+1} = s' | S_t = s, A_t = a] \\ &= \sum_{s'} [G_t | S_{t+1} = s'] P[S_{t+1} = s' | S_t = s, A_t = a] \\ &= \sum_{s', r} \left[r + \gamma E_{\bar{a}}[G_{t+1} | S_{t+1} = s'] \right] P(s', r | s, a) \\ &= \sum_{s', r} [r + \gamma U_{\bar{a}}(s')] P(s', r | s, a) \end{aligned}$$

$$\begin{aligned} \text{(c)} \quad q_{\bar{a}}(s, a) &= \sum_{s', r} P(s', r | s, a) [r + \gamma U_{\bar{a}}(s')] \\ &= \sum_{s', r} P(s', r | s, a) \left[r + \gamma \sum_{a' \in A} \pi(a' | s) \cdot \gamma q_{\bar{a}}(s', a') \right] \end{aligned}$$