CS 5180 Exercise-2

1) a) State space :- $\{x, y\}$

where, $0 \leq x \leq 10$; $0 \leq y \leq 10$

Action space :- {UP, DOWN, LEFT, RIGHT}

b) For $p(s', r | s, a)$

The total no. of possible states are :- $x \cdot y$ in which 17 states are not possible. Therefore,

no. of possible states :- $(11 \times 11) - 17 = 104$

Total no. of possible actions in each state = 4

The no. of possible next states for a given action = 3

∴ Total no. of possible non-zero rows = $(104)(4)(3)$

$= 1248.$

2) a) Episodic task :-

We know, total return when the Episode is completed in T steps, then,

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots + \gamma^{T-1} R_T$$

$$= 0 + \gamma(0) + \gamma^2(0) + \cdots + \gamma^{T-1}(-1)$$

$$\boxed{G_t = -\gamma^{T-1}}$$

Continuing task:-

Total return $G_t = R_t + \gamma R_{t+1} + \cdots + \gamma^k(-1) + \cdots \gamma^{k+n} R_{t+k+1}$

$$\Rightarrow G_t = 0 + \gamma(0) + \cdots + \gamma^k(-1) + \cdots \gamma^{k+n}(-1)$$

$$G_t = -(\gamma^k + \gamma^{k+n} + \cdots)$$

This cannot be simplified further

⇒ For episodic task there is fixed return value for each episode and for the continuing task there will be no ~~finite~~ fixed value for the return.

3) a) we know,

$$G_t = R_{t+1} + \gamma G_{t+1}$$

$$G_5 = 0$$

$$G_4 = R_5 + \gamma G_5 \Rightarrow 2.$$

$$G_3 = R_4 + \gamma G_4 \Rightarrow 3 + (0.5)(2) = 4$$

$$G_2 = R_3 + \gamma G_3 \Rightarrow 6 + (0.5)(4) = 8$$

$$G_1 = R_2 + \gamma G_2 \Rightarrow 2 + (0.5)(8) = 6$$

$$G_0 = R_1 + \gamma G_1 \Rightarrow (-1) + (0.5)(6) = 2$$

b) Given, $\gamma = 0.9 \quad R_1 = 2 \; ; \; R_2 = R_3 = R_4 = \cdots = 7$

$$G_0 = 2 + \gamma G_1$$

$$= 2 + 0.9\left[R_2\right] + (0.9)^2 (R_3) + \cdots - (0.9)^{\infty} (R_\infty)$$

$$= 2 + 0.9(7) + (0.9)^2 (07) + \cdots - (0.9)^{\infty} (7)$$

$$= 2 + 7\left[0.9 + (0.9)^2 + (0.9)^3 + \text{----}\right]$$

$$= 2 + 7\left[\frac{0.9}{1-0.9}\right] \quad \text{(Geometric progression)}$$

$$= 65$$

$$G_8 = R_2 + \gamma R_3 + \gamma^2 R_4 + \text{----}$$

$$= 7 + \gamma(7) + \gamma^2(7) + \text{----}$$

$$= 7\left[1 + \gamma + \gamma^2 + \gamma^3 + \text{----}\right]$$

$$= 7\left[\frac{1}{1-\gamma}\right] = 7\left(\frac{1}{1-0.9}\right)$$

$$= 70$$

4) For 'up' action,

$$V(UP) = 50 + \gamma(-1) + \gamma^2(-1) + \text{----} \gamma^{100}(-1)$$

$$= 50 - \left[\gamma + \gamma^2 + \gamma^3 + \text{----} + \gamma^{100}\right]$$

$$= 50 - \left[\frac{\gamma(1-\gamma^{100})}{1-\gamma}\right]$$

For 'DOWN' action,

$$V(\text{DOWN}) = -50 + \gamma(+1) + \gamma^2(1) + \gamma^3(+1) + \cdots - \gamma^{100}(1)$$

$$= -50 + \left[\frac{\gamma(1-\gamma^{100})}{1-\gamma}\right]$$

The agent will choose the action with highest value function.

For choosing 'UP' :- $V(UP) > V(DOWN)$

For choosing 'DOWN' :- $V(UP) < V(DOWN)$

5) a) The signs of rewards are not important, only the interval between them is important.

Eq :- 3.8, $G_t = R_{t+1} + \gamma R_{t+2} + \cdots$

Let G_t' be the new return after adding 'c' to all the rewards.

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

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

$$= c\left[1 + \gamma + \gamma^2 + \gamma^3 + \cdots\right] + \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots\right]$$

$$G_t' = c\left[\frac{1}{1-\gamma}\right] + G_t$$

$$\boxed{G_t' = v_c + G_t}$$

$$\Rightarrow v_c = c\left[\frac{1}{1-\gamma}\right]$$

b) For tasks such as maze running, adding a constant for all the rewards may impact the outcome of the optimal policy.

Let's say hitting the walls result in '-1' reward before adding a constant c. As the agent tries to increase the return this will start avoiding hitting the walls as it will reduce the final return.

Let's say, $c = 2$ is added to all the rewards. then hitting the wall will result in '+1' reward The policy will try to increase the final return. therefore it may stuck in hitting the walls for shortterm rewards instead of finding the exit for the maze.

6) a) We know,

$$V_\pi(s) = E_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s\right]$$

$$\Rightarrow V_\pi(c) = 0 + 0.9\left[(0.25)(2.3) + (0.25)(0.4) + (0.25)(-0.4) + (0.25)(0.7)\right]$$

$$V_\pi(c) = 0.9\left[0.75\right] = 0.675 \approx 0.7$$

b) We know,

$$V_*(s) = \max_a E_{\pi_*}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right]$$

$$V_*(c) = 0 + 0.9\left[\max(19.8, 19.8, 16, 16)\right]$$

$$= 0.9(19.8) = 17.82 \approx 17.8$$

7) a) As the probability of ~~the~~ choosing an action is equal we can guess the V(A) as 0.5.

We know, $v(A) = \pi(L|A)\left[0 + v(L)\right] + \pi(R|A)\left[0 + v(R)\right]$

$$= (0.5)(0+0) + (0.5)(0+1)$$

$$V(A) = 0.5(0) + 0.5)$$
$$= 0.5$$

b) The value function drops as they move away from the right terminal. So, $V(E) = 0.5$, $V(D) = (0.5)^2$, $V(C) = 0.5^3$, $V(B) = 0.5^4$, $V(A) = 0.5^6$

$$V(E) = 0.5(V(D) + V(R)) = 0.5(0+1) = 0.5$$

$$V(D) = 0.5(0 + 0.5^2) = 0.5^2$$

$$V(C) = 0.5(0 + 0.5^2) = 0.5^3$$

$$V(B) = 0.5(0 + 0.5^3) = 0.5^4$$

$$V(A) = 0.5(0 + 0.5^4) = 0.5^5$$

c) $V(S_n) = 0.5^n$, where S_n is the n^{th} state from R.

8) a) we know,

$$V(\text{high}) = \pi(\text{search}|\text{high})\left[\alpha(r_s + \gamma \cdot V(\text{high}) + (1-\alpha)(r_s + V(\text{low}\right.$$

$$+ \pi(\text{wait}|\text{high})\left[r_w + \gamma \cdot V(\text{high})\right]$$

$$V(\text{low}) = \pi(\text{search}|\text{low})\left[(1-\beta)(-3 + \gamma V(\text{high}) + \beta \cdot (r_s + \gamma V(\text{low}))\right]$$

$$+ \pi(\text{wait}|\text{low})\left[r_w + \gamma V(\text{low})\right] + \pi(\text{recharge}|\text{low})\left[0 + \gamma V(\text{hig}\right.$$

b) by substituting, we get.

$$(0.685)\, V(\text{high}) = 0.135 \cdot V(\text{low}) + 5$$

$$(0.55)\, V(\text{low}) = 0.45 \cdot V(\text{high}) + 1.5$$

by solving we get, $V(\text{high}) = 9.33$, $V(\text{low}) = 10.36$

c) $V(\text{high}) = (0.135 \cdot V(\text{low}) + 5) / 0.685$

$$V(\text{low}) = \theta \cdot (3 + 0.9)(V(\text{low})) + (1-\theta)(0.9 \times V(\text{high}))$$

$$V(\text{low}) = (6.55 - 3.55(\theta)) / (0.83 + 0.73 \cdot \theta)$$

Choosing the value of θ which maximizes the value of $v(low)$ maximizes $v(high)$ as well.

At $\theta = 0$, $v(low) = 7.89$, $v(high) = 8.85$.

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

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

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