Ex - 02

1) Formulating an MDP:

a) The State Space S in 4 rooms. Consists of 4 rooms themselves.

i.e State : (t, y) $t, y \in [0, 10]$
$(t, y) \neq (5, i), (j, 5), (k, 4)$.

$i = 0, 2, 3, 4, 5, 6, 7, 8, 9, 10$.
$j = 0, 2, 3, 4,$
$k = 6, 7, 9, 10$.

$x \in [0, 10], y \in [0, 10] - [(0,5), (2,5), (3,5), (4,5), (5,5), (6,4), (7,4), (9,4), (10,4), (5,0), (5,2), (5,3), (5,4), (5,6), (5,7), (5,9), (5,10)]$.

Where as, action Space A consists of 4 possible actions move.

Action = {North, South, East, West}.

b) In 4 room domain has 17 walls.
So State $S_n = 121 - 17$
$= 104$.

Action is 'East' or 'West' when not blocked by walls (S, a) pair there are 3 Non zero $P(s', r \mid s, a)$ value.

No. of non zero when non bloced by walls $= 104 \times 4 \times 3$
$= 1248$.

when blocked by walls $= 104 \times 4 \times 1 = 416$.

Appox. the no. of non zero rows can be $\underline{832}$.

Q1 C)

PSEUDOCODE

```
Initialize an empty table for transition probabilities

For each cell in the grid (excluding the boundary):
  If the cell is not a wall:
    For each action in [UP, DOWN, LEFT, RIGHT]:
      Calculate the resulting cell after the action
      If the resulting cell is within the grid and not a wall:
        If the action does not take the agent out of bounds:
          Add an entry to the table with:
            - The current cell as the current state s
            - The resulting cell as the next state s'
            - The action a
            - The transition probability (0.8 for the intended direction, 0.1 for the perpendicular
directions)
            - The reward r (which is 0 unless transitioning into the goal state, which gives a reward of 1)
      If the current cell is the goal state (10, 10):
        Add an entry to the table with:
          - The current cell as the current state s
          - The start state (0, 0) as the next state s' for all actions
          - The action a
          - A transition probability of 1
          - The reward r of 0
```

2) 1 point RL objective.

a) Treated pole balancing as an episodic task but also used discounting, with all rewards zero except -1 upon failure.

Denoting failure time as T

Sum of all failure discounted rewards, $t < T$ will be 0

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

$\therefore G_t = 0$ for $t < T$ as $R_{t+1} = R_{t+2} = 0$.

As the angent receives a reward (-1) is a terminal time step G_T is simply imidiate reward

$G_T = -1$

$\therefore$ we can formulate

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

While continuing formulation. (continuous task)

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

$$= \gamma^k (-1)$$

$G_t = -\gamma^k$ Same as expression above.

$\therefore G_t = \gamma^k (-1) + \gamma^{2k} (-1) + \gamma^{3k} (-1) \dots$

$\therefore \quad G_t = -\gamma^K[1+\gamma^K+\gamma^{2K} \dots]$

$$\therefore \quad \boxed{G_t = -\gamma^K \cdot \frac{1}{1-\gamma^K}}$$

Thus the episodic case agent focuses on single outcome of episode end, while in continuing case.

b) For episodic task while discounting $\gamma=1$, where reward is 0 at every time step T is a terminal state. and $R_T = 1$ for the final time step.

The expected reward

$$G_T = R_1 + \gamma R_2 + \gamma^2 R_3 + \gamma^3 R_4 \dots \gamma^{T-1} R_n$$

$$R_1 = R_2 = R_3 = 0.$$

$$\therefore \quad G_t = 0 + 1(0) + 1^2(0) \dots + 1 \cdot 1$$

$$\therefore \quad \boxed{G_t = 1}$$

$\therefore$ G_t always be 1, so it will never have changes,

⇒ Thus, we need to know to provide intermediate reward to let agent know and learn to achieve better.

3) 1 point : Discounted return.

a)

$\gamma = 0.5$, $T = 5$

$R_1 = -1$ $\quad$ $R_4 = 3$

$R_2 = 2$ $\quad$ $R_5 = 2$ $\quad$ } (given)

$R_3 = 6$

To find:

$G_0, G_1, G_2, G_3, G_4, G_5 = ?$

Sol^n:

$G_t = R_{t+1} + \gamma G_{t+1}$ — (eq^n 3.9)

$\therefore$ $t = 5$

$G_5 = R_6 + \gamma G_6 = 0$

$\therefore \boxed{G_5 = 0}$

$t = 4$

$G_4 = R_5 + \gamma \cdot G_5$

$= 2 + 0.5(0)$

$\therefore \boxed{G_4 = 2}$

$t = 3$

$G_3 = R_4 + \gamma G_4$

$= 3 + (0.5) 2$

$\boxed{G_3 = 4}$

$t = 2$

$G_2 = R_3 + \gamma G_3$

$= 6 + (0.5) 4$

$\boxed{G_2 = 8}$

$t=1$.

$G_1 = R_2 + \gamma G_2$

$= 2 + (0.5)\,8$

$\boxed{G_1 = 6}$.

$t=0$

$G_0 = R_1 + \gamma G_1$

$= -1 + (0.5)\,6$.

$\boxed{G_0 = 2}$

$G_0 = 2$, $G_1 = 6$, $G_2 = 8$, $G_3 = 4$, $G_4 = 2$, $G_5 = 0$ //.

b) $\gamma = 0.9$

$R_1 = 2$

$R_2 = R_3 = R_4 \ldots R_n = 7$.

To find: $G_0, G_1 = ?$

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

$\therefore$ $t=1$

$G_1 = R_2 + \gamma G_2$

$G_2 = R_3 + \gamma G_3$

$= 7 + \gamma(7) + \gamma^2(7)$

$= 7(1 + \gamma + \gamma^2 + \ldots)$

$= 7\left(\frac{1}{1-\gamma}\right) = 7\left(\frac{1}{1.09}\right)$

$\boxed{G_2 = 70}$

$G_1 = 7 + (0.9)(70)$

$\boxed{G_1 = 70}$

$t=0$,

$G_0 = R_1 + \gamma G_1$

$= 2 + (0.9)(70)$

$\boxed{G_0 = 65}$

$\therefore$ $\boxed{G_0 = 65 \quad G_1 = 70}$

4) Case I: Up motion

$$G_{up} = 50 + \sum_{n=2}^{101} \gamma^{n-1} R_n$$

$$= 50 + (-\gamma)\frac{(1-\gamma^{100})}{1-\gamma} = \frac{50 - \gamma(1-\gamma^{100})}{1-\gamma}.$$

Case II: Down motion.

$$G_{down} = -50 + \sum_{n=2}^{101} \gamma^{n-1} R_n$$

$$= -50 + \frac{\gamma(1-\gamma^{100})}{1-\gamma}.$$

So $G_{up} - G_{down}$.

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

$$= 100 - 2\gamma\frac{(1-\gamma^{100})}{1-\gamma}.$$

Using Wolfram alpha.

$$-1.047 < \gamma < 0.9843$$

And, $G_{down} - G_{up}$.

$$= -50 + \frac{\gamma(1-\gamma^{100})}{1-\gamma} - \frac{50 - \gamma(1-\gamma^{100})}{1-\gamma}$$

$$= -100 + 2\gamma\frac{(1-\gamma^{100})}{1-\gamma}.$$

For discount factor bet^n.

$0.9843 < \gamma < 1$ is better to take down motion

5) <u>1 point</u> : <u>Modifying the reward function.</u>

a) $G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

... (eqⁿ 3.8)

<u>Adding constant to all rewards.</u>

$G'_t = (R_{t+1} + c) + \gamma (R_{t+2} + c) + \gamma^2 (R_{t+3} + c) \dots$

<u>Factoring out c to see the effect:</u>

$G'_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \dots$
$c(1 + \gamma + \gamma^2 \dots)$

The sum of series

$G_t = \frac{1}{1-\gamma}$ if $\gamma < 1$

So, adding a constant across all states.

$$\boxed{G'_t = \frac{c}{1-\gamma}}$$

This added value V_c is constant across all states and does not depend on the policy, so it does not affect the relativvalues.

This means policy that maximizes the original return G_t will also maximize return G'_t.

$V_{\pi c}(s) = V_\pi(s) + V_c.$

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

(8)

b) For an episodic task.

$$G'_t = (R_1 + c) + \gamma(R_2 + c) + \gamma^2(R_3 + c) \ldots + \gamma^{n-1}(R_n + c)$$

$$G'_t = R_1 + \gamma R_2 + \gamma^2 R_3 + \ldots \gamma^{n-1} R_n + c(1 + \gamma + \gamma^2 \ldots + \gamma^{n-1})$$

The sum of series

$$\frac{1-\gamma^n}{1-\gamma}$$

$$V_{\pi_c}(s) = E_\pi[G_t \mid S_t = s]$$

$$= E_\pi[G_t \mid S_t = s] + E_\pi[cT \mid S_t = s]$$

$$= V_\pi(s) + \underbrace{c\, E_\pi[T \mid S_t = s]}$$

V_c now depends on n, length of the episode

∴ The shifts ~~with~~ in the return will not be constant.

∵ This will have a effect.

6) I point Bellman eq^n:

a)

$$V\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma V_\pi(s')\right] \quad (3.14)$$

for all $s \in S$

Since the policy is equiprobable action has a probability of 0.25.

$$\begin{array}{ccc} & 2.3 \leftarrow V_n & \\ V_w \rightarrow 0.7 & 0.7 & 0.4 \leftarrow V_e \\ & -0.4 \leftarrow V_s & \end{array}$$

∴ γ is less than 1

∴ $\gamma = 0.93$

∴ Equation for centre state V_c, reward is 0.

$$V_c = 0.25(r + \gamma V_n) + 0.25(r + \gamma V_s) + 0.25(r + \gamma V_e) + 0.25(r + \gamma V_w).$$

$$V_c = 0.25(0 + (0.93)(2.3)) + 0.25(0 + (0.93)(-0.4)) + 0.25(0 + (0.93)(0.4)) + 0.25(0 + (0.93)(0.7)).$$

∴ $\boxed{V_c = 0.7}$

This matches the value of centre state in the figure.

10

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

$\gamma = 0.9$

~~Val~~ centre value $= 17.8$

	19.8	
19.8	17.8	16
	16	

$\therefore$

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

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

$\therefore$

$V_c \approx 17.82$

$\boxed{V_c = 17.8}$

(11)

7) Guessing and verifying value funⁿ:

a) For 3-state MDP is,

$V = 0.5$

To verify the value funⁿ

$V\pi(S) = \sum \pi(a|s) \sum P(s', r|s, a) [r + \gamma V\pi(s')]$

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

$= \frac{1}{2} + \frac{1}{2} \times 0 = \frac{1}{2}$

$\therefore \boxed{V\pi(S) = 0.5}$ $\therefore$

$\because$ LHS = RHS.

b) Value funⁿ for 7 state MDP.

$\therefore V\pi(S) = \frac{1}{2} \times 1 \times (0 + 1 \times 0) + \frac{1}{2} \times 1 \times (0 + 1 \times V\pi(S)(B)) = \frac{1}{2} V\pi(S)(B)$

$\therefore V\pi(S)(B) = \frac{1}{2} \times 1 \times (0 + V\pi(S)(A)) + \frac{1}{2} \times 1 \times (0 + V\pi(S)(C))$

$\therefore V\pi(S)(C) = \frac{1}{2} \times 1 \times (0 + V\pi(S)(B)) + \frac{1}{2} \times 1 \times (0 + V\pi(S)(D))$

$\therefore V\pi(S)(D) = \frac{1}{2} \times 1 \times (0 + V\pi(S)(C)) + \frac{1}{2} \times 1 \times (0 + V\pi(S)(E))$

$\therefore V\pi(S)(E) = \frac{1}{2} \times 1 \times (0 + V\pi(S)(D)) + \frac{1}{2} \times 1 \times (0 + V\pi(S)(D))$

$\therefore \boxed{V(A) = \frac{1}{6},\ V(B) = 2 \cdot V_A = \frac{1}{3},\ V_C = 3 V_A = \frac{1}{2},\ V_D = 4 \cdot V_A = \frac{2}{3},\ V_E = \frac{5}{6}}$

c) For this value function for orbitrary is,

$V\pi(S) = \frac{i}{n-1}$ $\therefore$ where $[i = 1, 2, 3, \ldots (n-2)]$.

8) Solving for Value Function.

a) Equation for the 2 States in the recycling robot.
arbitrary policy $\pi(a|s)$

$$V_\pi = \sum_a \pi(a|s) \sum_{s',r} P(s',r|s,a)\,(r + \gamma V(s'))$$

So,

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

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

b) $\pi(\text{search}|\text{high}) = 1$, $\pi(\text{wait}|\text{low}) = 0.5$
$\pi(\text{recharge}|\text{low}) = 0.5$, $\alpha = 0.7$, $\beta = 0.8$, $\gamma = 0.9$
$V_{search} = 10$, $V_{wait} = 3$

$$V_{low} = 0.5 \times 1 \times (3 + 0.9 V_{low}) + 0.5 \times 1 \times (0.9 + V_{high}).$$

~~$V_{high} = 0.55 V_{low} = 1.5 + 0.45 V_{high}$ — (i)~~

$$V_{high} = 0.8 \times (10 \times 0.9 \times V_{high}) + 0.2 \times (10 + 0.9 V_L)$$

$\therefore\ 0.55 V_{low} = 1.5 + 0.45 V_{high}$ — (i)

$\therefore\ 0.28 V_{high} = 10 + 0.18 V_{low}$ — (ii)

Solving (i) & (ii) we get,

$$\boxed{\begin{aligned} V_{high} &= 79.03 \\ V_{low} &= 67.40 \end{aligned}}$$

c) So,

$\pi(\text{wait}|\text{low}) = \theta$

$\pi(\text{recharge}|\text{low}) = 1-\theta$

$V_{low} = \theta(3 + 0.9\,V_{low}) + (1-\theta)\,0.9\,V_{high}$

$= 3\theta + 0.9\theta\,V_{low} + 0.9 - 0.9\theta\,V_{high}$

$(1-0.9\theta)\,V_{low} = 3\theta + 0.9(1-\theta\,V_{high})$

from (b)

$0.28\,V_{high} = 10 + 0.18\,V_{low}$

Solving we get

$$V_{low} = \frac{9 - 8.16\theta}{0.118 - 0.90} = \frac{90.67 - .6986}{0.118 - 0.090}$$

$\therefore$ $V_{low} = 76.28$, $V_{high} = 84.75$ } when $\theta = 0$