JAVAD MOEIN

Q1)

a) $S = \{(x, y) \in N \times N \mid 0 \le x \le 10, 0 \le y \le 10\}$

we also consider walls states, which could not be accessible

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

b) we have 11x11 states which for each state we have 4 actions (possible action) and for each action we usually have 3 possible transfer. of course we could not to transfer to wall states and never have transformation from wall states so approximately we will have

$[11 \times 11 - num\_of\_wall\_states] \times 4 \times 3 - num\_of\_wall\_states$

$= [11 \times 11 - 17] \times 4 \times 3 - 17 = 1231$ non-zero Rows

Q2)-

a-) episodic task with discount

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots \gamma^{T-t-1} R_T = \sum_{K=0}^{T-t-1} \gamma^K R_{t+K+1}$$

If $R_{t+1} = \cdots = R_{T-1} = 0$, $R_T = -1$

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

while for continuing task, the return is $-\gamma^K$ where K is the number of timesteps before failure.

So for episodic $G_t = -\gamma^{T-t-1} = -\gamma^{K-1}$, considering t is K steps before T, failure.

b) Because there is no discount, we do not force the robot to exit the maze as soon as possible. In this case when the robot exit get rewards 1 otherwise it gets zero which does not for it to exit quickly. So we have not communicated effectively to the robot. For improving it'd better to have discount or to consider negative the rewards for non-goal (states

So we have not communicated effectively to the robot. For improving it'd better to have discount or to consider negative rewards for non-goal states

3)

a) In continuing task, the sign of rewards is not important and the intervals beetwen them is important, as shown in following if we add a positive value to alle rewards finally the relative values between states will not change

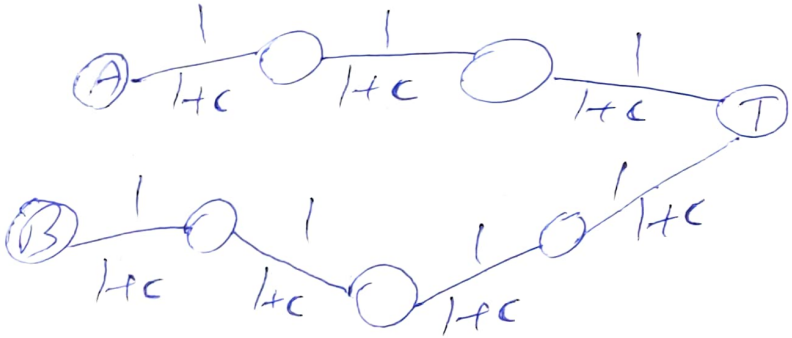
$$G_t(s) = R_{t+1} + \gamma R_{t+2} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad \forall s$$

$$G_t'(s) = (R_{t+1} + c) + \gamma[R_{t+2} + c] + \dots = c\sum_{k=0}^{\infty} \gamma^k + \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad \forall s$$

$$G_t'(s) = c\sum_{k}^{\infty} \gamma^k + G_t(s) \quad \forall s \Rightarrow V_c = c\sum_{k}^{\infty} \gamma^k \quad \forall s$$

If $|\gamma| < 1 \Rightarrow V_c = \frac{c}{1-\gamma} \quad \forall s$

b) In the episodic task, adding a constant would be problematic because at first it will change the relative values between states and second it cause some negative rewards to be positive that change our problem.



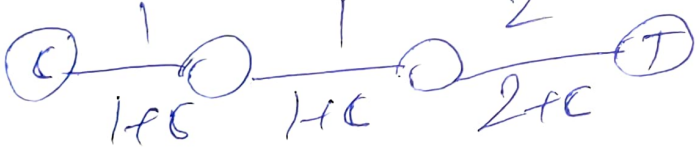
$G_t(A) = 3$

$G'_t(A) = 3 + 3\varepsilon$

$G_t(B) = 4$

$G'_t(B) = 4 + 4\varepsilon$

T: is terminal state



$G_t(C) = 4$

$G'_t(C) = 4 + 3\varepsilon$

$G_t(A) - G_t(B) = -1$

$G_t(A) - G_t(C) = -1$

$G'_t(A) - G'_t(B) = -(1+\varepsilon)$

$G'_t(A) - G'_t(C) = -(1)$

Before adding ε to each reward the relative value between A-B and A-C was same but after adding ε this relative value changes. which makes our problem distinct from the first one

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

$\forall s \in S$

$\gamma = 0.9$

$p(s',r|s,a) = 1 \quad \forall s, s', a, r$

$r = 0$

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

$+ \frac{1}{4}\left[0 - 0.9 \times 0.4\right] = 0.675 \sim 0.7$

b) $17.8 = \frac{1}{2}\left[0 + 0.9 \times 19.8\right) + \frac{1}{2}\left[0 + 0.9 \times 19.8\right] = 17.82$

5)

a) Guess: $V_\pi(A) = \frac{1}{2}$ $\quad V(L) = 0 \quad V(R) = 1$

$$V_\pi(s) = \sum \pi(a|s) \sum p(s', r | s, a)\left[r + \gamma V_\pi(s')\right]$$

$\gamma = 1$

$p(s', r | s, a) = 1$

$V_\pi(s' = R) = 1$

$V_\pi(s' = L) = 0$

$r = 0$

$$V_\pi(s = A) = \pi(\text{Right} | s = A)\left[0 + V_\pi(s' = R)\right] + \pi(\text{Left} | s = A)\left[0 + V_\pi(s' = L)\right]$$

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

$V_\pi(s = R) = 1$ $\quad V_\pi(s = L) = 0$

b) because there is not discount ⇒ Guess:

$V(L)=0$ $V(A)=\frac{1}{6}$ $V(B)=\frac{2}{6}$ $V(C)=\frac{3}{6}$

$V(D)=\frac{4}{6}$ $V(E)=\frac{5}{6}$ $V(R)=1$

Ex: $V_\pi(S=E)=\pi(Right|S=E)[0+V_\pi(S'=R)]+$
$\pi(Left|S=E)[0+V_\pi(S'=D)]=$

$\frac{1}{2}\times 1+\frac{1}{2}\times\frac{4}{6}=\frac{5}{6}$

c) based on Part c we could guess that for n states the value function of states would be

$V_\pi(S=S_1)=\frac{1}{n}$ $V_\pi(S=S_2)=\frac{2}{n}$ ... $V_\pi(S=S_{n-2})=\frac{n-1}{n}$

$V_\pi(S=L)=0$ $V_\pi(S=R)=1$

6) H: high L: low

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

$V_\pi(S = High) = \pi(Search|H) \times \left(\alpha[r_{search} + \gamma V_\pi(H)] + (1-\alpha)[r_{search} + \gamma V_\pi(low)] \right)$

$+ \pi(wait|H)\left(1[r_{wait} + \gamma V_\pi(H)]\right)$

$V_\pi(S = L) = \pi(search|L)\left(\beta[r_{search} + \gamma V_\pi(L)] + (1-\beta)[-3 + \gamma V_\pi(H)]\right)$

$+ \pi(wait|L)\left(1[r_{wait} + \gamma V_\pi(L)]\right)$

$+ \pi(recharge|L)\left(1[0 + \gamma V_\pi(H)]\right)$

b) $\alpha = 0.8$ $\beta = 0.6$ $\gamma = 0.9$ $r_{search} = 10$

$r_{wait} = 3$ $\Pi(search|H) = 1$ $\Pi(wait|L) = 0.5$

$\Pi(recharge|L) = 0.5$

$$V_\Pi(H) = 10 + 0.72\, V_\Pi(H) + 0.18\, V_\Pi(L)$$

$$V_\Pi(L) = 1.5 + 0.45\, V_\Pi(L) + 0.45\, V_\Pi(H)$$

$$\Rightarrow V_\Pi(L) = 87.7 \qquad V_\Pi(H) = 89.0$$

7)

a) $V_\pi(s) = \sum_a \pi(a|s)\, q_\pi(a,s)$

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

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