366 Assignment 2 Jiaqi11 Jiaqi Liu.

Q1 a(1) × left 0 × left 0 ... (repeat)

b(2) × right 1 × right 1 × right -1 Y

right 3.

c(3) $G_0 = R_0 + \frac{1}{2}R_1 + \frac{1}{4}R_2 + \frac{1}{8}R_3 = 1 + 0.5 + (-0.25) + 0.375$

$= 1.625.$

d(4) $V_{\pi_1}(Y) = 3$ directly go to the terminal state

e(5) $q_{\pi 1}(x, \text{left}) = 0$ × left 0 repeat

f(6) $V_{\pi_2}(x) = \frac{2}{3}(1 + \frac{1}{2}V_{\pi_2}(x)) + \frac{1}{3}(-1 + \frac{1}{2}V_{\pi_2}(y))$

$V_{\pi_2}(y) = V_{\pi_1}(y) = 3.$

$\Rightarrow V_{\pi_2}(x) = \frac{5}{4}$

Q2 c a) ① FIFA soccer football Bot AI: if successfully passing the ball gets +1, a goal gets +10, If failed to pass the ball get -1, if lose a goal gets -10. By giving value to different tatics., the aim is to get the maximum reward out of each game. Agent should find out the biggest reward in 90 minutes of the game.

② Android Phone assistant: kill one of the app which is not being used currently, get more space in RAM +1, run the test programme, if the phone get better benchmarch +5, if the user open that app in 5 mins -5, the aim is that the agent will help the phone to work more officiently in the long term Based on different user behavior to make a custome plan for everyday using.

③ Bio - medical producer: choose right temperature and amount of water to produce ~~chemical~~ medicine from plants The sensor detect the amount of production of the chemical materials Get an above average amount get +1, per hour yet lower than average get a -1, the agent will ~~know~~ finally know the right choice the maximize production

2 (b) Exercise 3.7.

① Since we want to let the agent to get out of the maze as soon as possible, the reward for each step should be zero. agent will always find a way out when it makes huge amount of choices. Despite the fact that the agent will always win the reward no matter how many steps it takes, it will never learn.

② If the reward for each step is 0, it is not communicating efficient. We need to set the reward for each step to -1, to tell the agent if it waste ~~few~~ 0+ steps on useless choice it will get less reward. In this way, it can learn properly.

3 (c). $G_t = R_{t+1} + \gamma G_{t+1}$

$G_5 = 0$

$G_4 = R_5 + \gamma G_5 = 2 + \frac{1}{2} \cdot 0 = 2$

$G_3 = R_4 + \gamma G_4 = 3 + \frac{1}{2} \times 2 = 4$

$G_2 = R_3 + \gamma G_3 = 6 + \frac{1}{2} \times 4 = 8$

$G_1 = R_2 + \gamma G_2 = 2 + \frac{1}{2} \times 8 = 6$

$G_0 = R_1 + \gamma G_1 = -1 + \frac{1}{2} \times 6 = 2$

4 (d). 2 7 7 7 ...

$G_0 = R_1 + 0.9 G_1$

$G_1 = R_2 + 0.9 \cdot G_2$

$G_2 = \lim_{n\to\infty} \frac{7(1-r^n)}{1-r} = \lim_{n\to\infty} \frac{7(1-r^n)}{1-r} = \frac{7}{1-r} = \frac{7}{1-0.9} = 70$

$G_1 = 7 + 0.9 \times 70 = 70$

$G_0 = R_1 + 0.9 G_1$

$= 2 + 0.9 \times 70 = 65$

$G_1 = 70, \ G_0 = 65$

(5) (e). 3.14

$V = \frac{1}{4}(0 + 0.9 \times 2.3) + \frac{1}{4}(0 + 0.9 \times 0.4) + \frac{1}{4}(0 + 0.9 \times -0.4)$

$+ \frac{1}{4}(0 + 0.9 \times 0.7).$

$= 0.675$

(6) (f) (3.8) $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \quad \gamma^n (R_{t+n})$

3.15

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

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

$\gamma^n (R_{t+n} + C)$

(geometric series).

$$= \frac{R_{t+1}(1-\gamma^n)}{1-\gamma} + \boxed{\frac{C(1-\gamma^n)}{1-\gamma}}$$

$$\lim_{n \to \infty} \frac{C(1-\gamma^n)}{1-\gamma}$$

$\because 0 < \gamma < 1$

$$\therefore \lim_{n \to \infty} \frac{C(1-\gamma^n)}{1-\gamma} = \frac{C}{1-\gamma}$$ which is a constant V_c.

The sign of the reward is important, because it can show the agent how to improve its decision base on the value it learns.

task above.

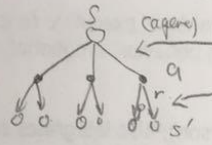
(g) 3.1b

If a constent c is add to all rewards including punishment reward in the "maze" problem.

If the -1 for punishment turns to 0 or a positive number, the agent will never learn. So it will have a effect is the constent turns the punishment reward into positive ones.

In "maze" case, each reward > -1, if all rewards are 0 it won't find the shortest way, if all rewards are positive, it will stay in the maze so that it gets the greatest rewards.

3.17 From book we have. $V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\,[r + \gamma V_\pi(s')]$



Now if we want to know $q(s,a)$ there for one of the branch

we have to go have reward first that do the policy. which is.

$$q(s,a) = 1 \times \sum_{s',r} p(s',r|s,a)\,[R + \gamma \sum_{a'} \pi(a'|s') \sum_{s',r'} p(s',r'|s',a')\, q(a'$$

3.18

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

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

$$V_\pi(s) = E_\pi\left[\sum_{a_i, i=1}^{a_i, i=3} \pi(a|s)\, q_\pi(s,a) \,\middle|\, S_t = s\right]$$

$$= \sum_{a_i, i=1}^{a_i, i=3} \pi(a|s)\, q_\pi(s, a_i)$$

$$= \pi(a_1|s)\, q_\pi(s,a_1) + \pi(a_2|s)\, q_\pi(s,a_2) + \pi(a_3|s)\, q_\pi(s,a_3)$$

3.24 Since the reward is 0 for each normal step.

$a_1 = a9 \times 24.4$

$a_2 = 0.9 \times a_1$

$a_3 = 0.9 \times a_2$

$a_4 = 0.9 \times a_3$

$(a9 \times 24.4) \times 0.9 \times 0.9 \times 0.9 = 16.00884$

Went back $16.00884 \times 0.9 + 10 = 24.407956$

equally 50% for 1 50% for 2

(a) $V_\pi = \frac{1}{2}(0 + V_1) + \frac{1}{2}(0 + V_2)$

$V_1 = \frac{3}{4}(3 + 0.8\times 2) + \frac{1}{4}(-6 + 0.8\times 7) = 3.35$

$V_2 = \frac{1}{5}[-3 + 0.8\times(-1)] + \frac{4}{5}\times(4 + 0.8V_\pi)$

$V_\pi = \frac{1}{2}\times 3.35 + \frac{1}{2}\times[-0.76 + 3.2 + 0.64V_\pi]$

$V_\pi = \frac{67}{40} + \frac{61}{50} + \frac{8}{25}V_\pi$

$V_\pi = \frac{\frac{579}{200}}{\frac{17}{25}} = \frac{579}{136} \approx 4.257\overline{352941176470588}$

(b) left

$V_* = 3 + 0.8\times 2$

$= 4.6$

right

$V_* = 0.2\times(-3 + 0.8\times -1) + 0.8\times(4 + 0.8V_*)$

$V_* = -0.6 - 0.16 + 3.2 + 0.64V_*$

$V_* = \frac{61}{4} = 6.\overline{777}$

$6.78 > 4.6$

right is the optimal choice

Exercise 3.6.

① The reward can be a small negative number which bigger than -1, each episode return > -0.5 which is unlikely to fail or < -0.5 if it is likely to fail in the next episode.

② In this case, the formulation will be a model that push the pole to the limit to stop. All kinds of details like angle or length or centroid will be taken into consideration.