Ex 2 Aadhar Bansal CS5180

Wednesday, September 28, 2022 6:31 PM

Q.1

a) State space = { set of all possible state from (0, 0) to (10, 10) except the Wall cells }

(x, y) ∀ all x, y (0, 10) except where (x, y) is wall

Action Space = The action choosen by the State { UP, DOWN, LEFT, RIGHT }

b) dynamics: $p(s', r / s, a)$ As in the Ex(0) the 0.2 Error to Go in ⊥ direction

{ (0,0) , DOWN } = $P((0,0) / (0,0), \text{DOWN}) = 0.8$
$P((0,0) / (0,0), \text{LEFT}) = 0.1$
$P((0,1) / (0,0), \text{RIGHT}) = 0.1$

{ (1,5) , UP } = $P((1,5) / (1,5), \text{UP}) = 0.8$
$P((1,4) / (1,5), \text{Left}) = 0.1$
$P((1,6) / (1,5), \text{RIGHT}) = 0.1$

{ (9,10), RIGHT } = $P((10,10) / (9,10), \text{RIGHT}) = 0.8$
$P((9,10) / (9,10), \text{UP}) = 0.1$
$P((9,9) / (9,10)\ \text{DOWN}) = 0.1$

②

(a) Suppose you treated pole-balancing as an episodic task but also used discounting, with all rewards zero except for −1 upon failure. What then would the return be at each time? How does this return differ from that in the discounted, continuing formulation of this task? (Derive expressions for both the episodic and continuing cases.)

Episodic Case

$R_{\text{terminal state}} = -1$

Cumulative Reward

$$G_t \doteq R_{t+1} + R_{t+2} \dots R_T \quad - (3.7)$$

with Discounting

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

$$= \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1} = \boxed{-\gamma^{T-t-1}}$$

for Continuous Case

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

$$G_t \doteq -\gamma^K$$

The difference b/w the episodic with discounting & Continuing case is that
episodic with discounting will have one failure & return the G_t. whereas with Continuous Case it will have multiple failure. & will be keep on updated.

b) Imagine that you are designing a robot to run a maze. You decide to give it a reward of +1 for escaping from the maze and a reward of zero at all other times. The task seems to break down naturally into episodes – the successive runs through the maze – so you decide to treat it as an episodic task, where the goal is to maximize expected total reward (Equation 3.7). There is no discounting, i.e. γ = 1. After running the learning agent for a while, you find that it is showing no improvement in escaping from the maze. What is going wrong? Have you effectively communicated to the agent what you want it to achieve?

The problem here is that no matter the time agent spend in the maze, the Reward maximum will be +1 for escaping the maze, as there is no discount factor γ.

Q.3 (a) In the gridworld example (Figure 3.2 in RL2e, see below), rewards are positive for goals, negative for running into the edge of the world, and zero the rest of the time. Are the signs of these rewards important, or only the intervals between them? Prove, using Equation 3.8, that adding a constant c to all the rewards adds a constant, v_c, to the values of all states, and thus does not affect the relative values of any states under any policies. What is v_c in terms of c and γ?

$$G_t \doteq R_{t+1} + \gamma R_{t+2} \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad - (3.8)$$

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

Adding Constant C

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

$$= E_\pi\left[\sum_{i=0}^{\infty} \gamma^i R_{t+k+1} \;\middle|\; S_t = s\right] + E_\pi\left[\sum_{i=0}^{\infty} \gamma^i c \;\middle|\; S_t = s\right]$$

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

(b) Now consider adding a constant c to all the rewards in an episodic task, such as maze running. Would this have any effect, or would it leave the task unchanged as in the continuing task above? Why or why not? Give an example.

In the episodic task, such as maze running, the sign of reward is important, as the -ve reward makes the agent to leave the maze fast.
Therefore adding constant C to the reward. if it keep negative reward remains intact the agent will find the exit path to the maze but the urgency to find it may vary & if the reward turns +ve after adding constant, the agent might not leave the maze, as be in the maze would be

most beneficial.

Q.4

(a) The Bellman equation (Equation 3.14) must hold for each state for the value function v_π shown in Figure 3.2 (right) (see above) of Example 3.5. Show numerically that this equation holds for the center state, valued at +0.7, with respect to its four neighboring states, valued at +2.3, +0.4, −0.4, +0.7. (These numbers are accurate only to one decimal place.) Note that Figure 3.2 (right) (see above) is the value function for the equiprobable random policy. γ = 0.9

$$V_\pi(s) \doteq \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma V_\pi(s')\right] \quad \forall \text{ all } s \in S \text{ — (3.14)}$$

$$= 1 \cdot \Big(\tfrac{1}{4}\big(0 + 0.9\,(2.3)\big) + \tfrac{1}{4}\big(0 + 0.9 \times 0.4\big) + \tfrac{1}{4}\big(0 + 0.9 \times (-0.4)\big)$$

$$+ \tfrac{1}{4}\big(0 + 0.9 \times 0.7\big)\Big)$$

$$= \frac{1}{4} \times 0.9\,(2.3 + 0.4 - 0.4 + 0.7)$$

$$V_\pi(s) = \frac{2.7}{4} = \underline{0.675}$$

$$= 0.7 \quad \text{(for one decimal pt)}$$

(b) The Bellman equation holds for all policies, including optimal policies. Consider v_* and π_* shown in Figure 3.5 (middle, right respectively) (See below). Similar to the previous part, show numerically that the Bellman equation holds for the center state, valued at +17.8, with respect to its four neighboring states, for the optimal policy π_* shown in Figure 3.5 (right) (see below).

from Bellman equation

from Centre State +17.8, Can take two State by these action from policy π^*

$$V_\pi(s)$$

$$1 \cdot \Big(\tfrac{1}{2}\big(0 + 0.9 \times 19.8\big) + \tfrac{1}{2}\big(0 + 0.9 \times 19.8\big)\Big)$$

$$= 0.9 \times 19.8 = 17.82$$

$$\boxed{V_\pi(s) = 17.8} \quad \text{(one decimal place}$$

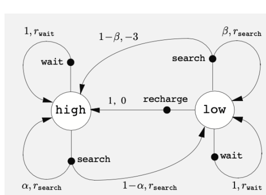
Q.5

a) 2 State recycling robot from Example 3.3

Charge level $S = \{\text{high}, \text{low}\}$

$$A(\text{high}) = \{\text{search}, \text{wait}\}$$

$$A(\text{low}) = \{\text{search}, \text{weight}, \text{recharge}\}$$



Bellman eqn

$$V_\pi(s) \doteq \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma V_\pi(s')\right] \quad \forall \text{ all } s \in S$$

$$V(\text{high}) = \pi(\text{search}/\text{high})\Big[\alpha\big[\overset{s'=\text{high}}{r_{search} + \gamma V(\text{high})}\big] + (1-\alpha)\big[\overset{s'=\text{low}}{r_{search} + \gamma V(\text{low})}\big]$$

$$+ \pi(\text{wait}/\text{high})\Big[1 \cdot \big[\overset{s'=\text{high}}{r_{wait} + \gamma V(\text{high})}\big] + \overset{s'=\text{low}}{0}\Big]$$

$$v(low) = \pi(search/low)\Big[1-\beta\big[-3+\gamma(v(high))\big)+\beta\big(r_{search}+\gamma(v(low))\big)\Big]$$
$$+\ \pi(wait/low)\Big[0+1\big(r_{wait}+\gamma\, v(low)\big)\Big]$$
$$+\ \pi(recharge/low)\Big[\big(\gamma\, v(high)\big)\Big]$$

b)

(b) You should now have two linear equations involving two unknowns, v(high) and v(low), as well as involving the policy π(a|s), γ, and the domain parameters. Let α = 0.8, β = 0.6, γ = 0.9, r_{search} = 10, r_{wait} = 3. Consider the policy π(search | high) = 1, π(wait | low) = 0.5, and π(recharge | low) = 0.5.
Find the value function for this policy, i.e., solve the equations for the values of v(high) and v(low). Check that your solution satisfies the Bellman equation.

$$V(high) = 1\cdot\big[0.8\big[10+0.9\,v(high)\big) + 0.2\big[10+0.9\,V(low)\big]$$

Let $v(high) = x$, $v(low) = y$

$$x = 8 + 0.72x + 2 + 0.18y$$

$$0.28x - 0.18y = 10$$
$$28x - 18y = 1000$$
$$14x - 9y = 500 \quad -\ ①$$

$$v(low) = 0.5\big[3+0.9\,v(low)\big)\big] + 0.5\big[0.9\,v(high)\big]$$

$$y = 1.5 + 0.45y + 0.45x$$

$$0.45x - 0.55y + 1.5 = 0$$
$$45x - 55y + 150 = 0$$
$$9x - 11y + 30 = 0 \quad -\ ②$$

① × 11 − ② × 9

$$(14\times 11)x - 99y = 500\times 11$$
$$-(9\times 9)x + 99y = +30\times 9$$
$$(154-81)x = 5500 + 270$$
$$73x = 5770$$

$$\boxed{\begin{array}{l} V(high) = 79.04 \\ V(low) = 67.39 \end{array}}$$

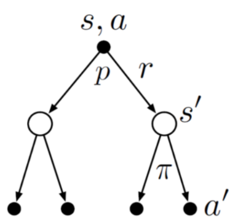
Verify by substituting them back in using Bellman equation

$$V(high) = 0.8\times(10+0.9\times 79.04) + 0.2\,(10+0.9\times 67.39)$$
$$= 64.90 + 14.13$$

$= 79.03$

$$v(tow) = 0.5 \times 1 \times (3 \times 0.9 \times 67.39) + 0.5(0.9 \times 79.04)$$
$$= 31.83 + 35.57$$
$$v(tow') = 67.40$$

Q.6



q_π backup diagram

a)
$$v_\pi(s) = \mathbb{E}_\pi[G_t / S_t = s]$$
$$= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} / S_t = s\right]$$
$$q_\pi(s,a) = \mathbb{E}_\pi[G_t / S_t = s, A_t = a]$$
$$= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} / S_t = s, A_t = a\right]$$
$$\therefore \quad v_\pi(s) = \sum_a \pi(a/s)\, q(s,a)$$

b)
$$q_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} / S_t = s, A_t = a\right]$$
$$= \sum_{s',r} p(s',r/s,a)\left[r + \gamma\, v_\pi(s')\right] \quad \text{— eq (b)}$$

c)
$$v_\pi(s') = \sum_{a} \pi(a'/s')\, q(s',a') \quad \text{— Put this in eq (b)}$$
$$q_\pi(s,a) = \sum_{s',r} p(s',r/s,a)\left[r + \gamma \sum_{a} \pi(a'/s')\, q(s',a')\right]$$