RL

Homework - 2

Ex. 3.15

We know that, for a policy π & state s,

$$v_\pi(s) = \mathbb{E}[G_t \mid S_t = s]$$

$$\Rightarrow v_\pi(s) = \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s]$$

using eq 3.8 in the chapter, ie,

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

Adding constant 'c' to all the rewards,

$$\Rightarrow v_\pi'(s) = \mathbb{E}[(R_{t+1} + c) + \gamma(R_{t+2} + c) + \gamma^2(R_{t+3} + c) + \dots \mid S_t = s]$$

$$\Rightarrow v_\pi'(s) = \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s] + \mathbb{E}[c\gamma + c\gamma + c\gamma^2 + \dots \mid S_t = s]$$

$$\Rightarrow v_\pi'(s) = v_\pi(s) + c\,\mathbb{E}\Big[\underbrace{\sum_{k=0}^{\infty} \gamma^k}_{\text{constant}} \mid S_t = s\Big]$$

$$\Rightarrow v_\pi'(s) = v_\pi(s) + \frac{c}{1-\gamma}$$ (since, $\gamma < 0$ & it is an infinite sum of G.P)

Hence, by adding 'c' to all the rewards changes the value ~~&~~ of each state by $\frac{c}{1-\gamma}$. ~~The~~ Thus,

$$v_c = \frac{c}{1-\gamma}$$

Ex. 3.16

Given : an episodic Task & adding 'c' to all the rewards

$$v_\pi(s) = \mathbb{E}[G_t \mid S_t = s]$$

$$v_\pi(s) = \mathbb{E}\left[\sum_{k=0}^{T-t-1} \gamma^k R_{t+1+k} \mid S_t = s\right]$$

Now, adding 'c' to all the rewards,

$$v'_\pi(s) = \mathbb{E}\left[\sum_{k=0}^{T-t-1} \gamma^k (R_{t+1+k} + c) \mid S_t = s\right]$$

$$= \mathbb{E}\left[\sum_{k=0}^{T-t-1} \gamma^k R_{t+1+k} + c\sum_{k=0}^{T-t-1} \gamma^k \mid S_t = s\right]$$

$$= \mathbb{E}[G_t \mid S_t = s] + \frac{c(\gamma^{T-t} - 1)}{\gamma - 1}$$

$$v'_\pi(s) = v_\pi(s) + \frac{c(1-\gamma^{T-t})}{1-\gamma}$$

For a given t, the increase is smaller, for a smaller value of $(T-t)$.

So states that are at a shorter distance to terminate will end up having a relatively smaller change.

Ex. 3.4

s	a	s'	r	$p(s', r \mid s, a)$
high	search	high	1	$\alpha \gamma_{search}$
high	search	high	0	$\alpha - \alpha \gamma_{search}$
high	search	low	1	$(1-\alpha)\gamma_{search}$
high	search	low	0	$(1-\alpha)(1-\gamma_{search})$
low	search	high	-3	$1-\beta$
low	search	low	1	βr_{search}
~~low~~	~~search~~			
high	wait	high	1	γ_{wait}
high	wait	high	0	$1-\gamma_{wait}$
low	wait	low	1	γ_{wait}
low	wait	low	0	$1-\gamma_{wait}$
low	recharge	high	0	1

The above table has been found out using the below formulas & the given conditions: ~~such as~~

$$p(s' \mid s, a) = \sum_{r \in R} p(s', r \mid s, a)$$

$$r(s, a, s') = \sum_{r \in R} r \, \frac{p(s', r \mid s, a)}{p(s' \mid s, a)}$$

Some of the conditions are:

→ If the energy level is high, then a period of active search can always be completed without risk of depleting the battery.

→ A reward of -3 results whenever the robot has to be rescued.

→ No cans can be collected during a run home for recharging.

1st row has been explained below:

$$p(high,1 \mid high, search) + p(high, 0 \mid high, search) = p(high \mid high, search)$$
$$= \alpha$$

& $r(high, search, high) = r_{search}$

$$r_{search} = 1 \cdot \frac{p(high,1 \mid high, search)}{p(high \mid high, search)} + 0 \cdot \frac{p(high,0 \mid high, search)}{p(high \mid high, search)}$$

$$\Rightarrow \quad p(high,1 \mid high, search) = \alpha\, r_{search}$$

$$\Rightarrow \quad p(high,0 \mid high, search) = \alpha - \alpha\, r_{search}$$

Q.5 $\quad v_*(s) = \max_a q_*(s,a)$