## Question 3 : Exercise 3.15

In the given gridworld example, the signs aren't important but their magnitude difference is what matters.

We need to have higher values for goals as compared to edges and rest so that the agent learns to favor this goal state in order to maximize the overall reward.

Similarily, running to the edge should have the least value and stationary should be somewhere in between.

$$V_\pi(S) = E_\pi[G_t \mid S_t = S]$$

$$= E_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \cdots \mid S_t = S]$$

Adding a constant 'c' to every reward gives

$$V_\pi^{new}(S) = E_\pi\left[R_{t+1}+c+\gamma(R_{t+2}+c)+\gamma^2(R_{t+3}+c)+\dots\infty \;\middle|\; S_t=s\right]$$

$$= E_\pi\left[(R_{t+1}+\gamma R_{t+2}+\dots+c+\gamma c+\dots) \mid S_t=S\right]$$

$$= E_\pi\left[(R_{t+1}+\gamma R_{t+2}+\gamma^2 R_{t+3}+\dots) \mid S_t=S\right]$$

$$+ E_\pi\left[c+\gamma c+\gamma^2 c+\dots \mid S_t=S\right]$$

(Using property of expectation)

$$= E_\pi\left[G_t \mid S_t=S\right] + E_\pi\left[c(1+\gamma+\gamma^2+\dots) \mid S_t=S\right]$$

$$= v_\pi(S) + E_\pi\left[c(1+\gamma+\gamma^2+\dots) \mid S_t=S\right]$$

We know that 'c' and gamas are constant, so we can take it out of the expectation.

$$v_\pi^{new}(S) = v_\pi(S) + c\times\frac{1}{(1-\gamma)} \quad \text{[Sum to infinite G.P]}$$

$$\boxed{v_\pi^{new}(S) = v_\pi(S) + v_c} \quad \left[v_c = \frac{c}{1-\gamma}\right]$$

Hence, proved.

Scanned with CamScanner

## Question 3: Exercise 3.16

Solving like the previous question, we get

$$\cancel{v_\pi^{new}(s) = E_\pi\left[R_{t+1} + \gamma R_{t+2} \dots \gamma^k R_T \mid S_t = s\right]}$$

$$\cancel{= E_\pi} +$$

$$V_\pi^{new}(s) = E_\pi\left[R_{t+1} + \gamma R_{t+2} + \dots \gamma^k R_T \mid S_t = s\right]$$
$$+ E_\pi\left[c + \gamma c + \dots \gamma^k c \mid S_t = s\right]$$

$$V_\pi^{new}(s) = v_\pi(s) + E_\pi\left[c + \gamma c + \dots \gamma^k c \mid S_t = s\right]$$

Here the sum is not till infinite but only till T.

$$E_\pi\left[c + \gamma c + \dots \gamma^k \mid S_t = s\right] = c(1 + \gamma + \gamma^2 + \dots \gamma^k$$

There are $(T - t - 1)$ terms.

$$\therefore c\left(\frac{1 - \gamma^{T-t}}{1 - \gamma}\right)$$

$$\therefore v_\pi^{new}(s) = v_\pi(s) + c\left(\frac{1 - \gamma^{T-t}}{1 - \gamma}\right)$$

Here, we can see that $v_c$ depends on a random variable T.

However, this will not affect the $v_s$ in one particular episode because $T$ is constant for that episode.

$$\therefore v_c = c\left(\frac{1-\gamma^{T-t}}{1-\gamma}\right)$$

Question 5

We know that

$$v^*(s) = \max_a q^*(s,a) \quad -(1)$$

At any state, the optimal value function will be the optimal action value function.

$$v^*(s) = \max_a \sum_{s'} p(s'|s,a) \cdot \left[E[r|s,a,s'] + \gamma v^*(s')\right]$$

$x_1 + x_2 + x_3 = 0$

~~Best~~ From (1), we have

$$\boxed{v^*(s) = \max_a \sum_{s'} p(s'|s,a)\left[E[r|s,a,s'] + \gamma \max_{a'} q^*(s',a')\right]}$$

<u>Question 1: 3.4</u>

In the table, value of $p(s'|s,a)$ is mentioned. We need to find the value of $p(s',r|s,a)$.

We know that
$$p(r,s'|a,s) = p(s'|s,a) \times p(r|s',s,a)$$

$$E[r|s,a,s'] = \sum_r r \times p(r|s',a,s')$$

This can be done as follow, when $s$ = low, ~~and~~ $a$ = ~~action~~ search, and next ~~state is~~ $s' = l$, we can have two rewards, 1 and 0 so we will have to take expectation over both of these.

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