Date / /

Page no.

RL Assignment 2

Vishaal Udandarao
2016119

1) We know that all the states as given in set S ⇒

$S = \{high, low\}$

We also know that the action sets for each of the individual states ⇒

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

Now, we know ⇒

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

Also, expected reward is ⇒

$$r(s, a, s') = \sum_{r} r\, p(r | s, a, s') \quad (2)$$

Now, rewards are binary i.e. 1 if an empty can is found or 0 if no can is found.

Now, from equations (1) and (2) ⇒

We can find the value of $p(s', r | s, a)$ for every expected reward i.e. as given in the example.

Date / /

Page no.

Therefore, the final table will be constructed as ⇒

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

3) Exercise 3.15

The signs of the reward are not important at an absolute scale but are important relatively. That is, we can change the value of the reward by adding a large positive constant value to all rewards This would not affect the value function. However, if the signs of the rewards are inverted, then the value function will change because then negative rewards which were positive rewards initially would mean that in the particular state, that action will not be preferred, when in fact it should be preferred.
Therefore, the signs of the rewards will only matter if they are not inverted for the value-function computation.

We know, state-value function ⇒

$$V_\pi(s) = E_\pi[G_t | S_t = s]$$

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

Now, if add a constant c to every reward ⇒

$$V'_\pi(s) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) \,\middle|\, S_t = s\right]$$

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

$$= E_\pi[G_t | S_t = s] + E_\pi\left[c \sum_{k=0}^{\infty} \gamma^k \,\middle|\, S_t = s\right]$$

Now, the second term is clearly a constant and hence can be written as ⇒

$$V_c = E_\pi\left[c \sum_{k=0}^{\infty} \gamma^k \,\middle|\, S_t = s\right]$$

$$= c \sum_{k=0}^{\infty} \gamma^k$$

Now, we know $0 \le \gamma \le 1 \Rightarrow \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$

$$\Rightarrow \boxed{V_c = c\left(\frac{1}{1-\gamma}\right)}$$

Page No.
Date:
MIRAJ
PREMIUM

Therefore, we can write $V'_\pi(s) \Rightarrow$

$$V'_\pi(s) = E_\pi[G_t | S_t = s] + V_c$$

$$\boxed{V'_\pi(s) = V_\pi(s) + V_c}$$

Therefore, $V'_\pi(s)$ is a sum of $V_\pi(s)$ and a constant V_c. Thus, the relative values of all states under any policy remain unaffected.

Exercise 3.16

For the case of an episodic task, we know state-value function as $\Rightarrow$

$$V_\pi(s) = E_\pi[G_t | S_t = s]$$

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

Here, T denotes no of steps to terminal state.

Following a similar procedure as in exercise 3.15 $\Rightarrow$ on adding constant c to all rewards $\Rightarrow$

$$V'_\pi(s) = E_\pi[G_t | S_t = s] + E_\pi\left[c \sum_{k=0}^{T} \gamma^k \,\middle|\, S_t = s\right]$$

$\therefore$ If we write it in the form as in exercise 3.15 $\Rightarrow$

$$V'_\pi(s) = V_\pi(s) + V_c \text{ where} \Rightarrow$$

$$V_c = E_\pi\left[c \sum_{k=0}^{T} \gamma^k \,\middle|\, S_t = s\right]$$

$$= c \sum_{k=0}^{T} \gamma^k$$

$$= c\left[\frac{1 - \gamma^{T+1}}{1 - \gamma}\right] \quad (\text{sum of a G.P})$$

$$\therefore \boxed{V_c = C\left[\frac{1 - r^{T+1}}{1 - r}\right]}$$

We clearly see that V_c is stochastic as it is a random variable dependent on T.

However, for one particular episode, T remains the same and hence, relatively for all the states, the term V_c is a constant since T is a constant for all states, in a given episode.

Therefore, within a given episode, the state-value functions will remain unaffected relatively.

5) We know,

$$v_*(s) = \max_{\pi} v_\pi(s)$$

$$= \max_{a \in A(s)} q_{\pi_*}(s, a)$$

That is $\Rightarrow v_*(s)$ is the maximum value of optimal action value functions over all actions.

$$\Rightarrow v_*(s) = \max_{a \in A(s)} q_{\pi_*}(s, a)$$

$$\Rightarrow v_*(s) = \max_{a \in A(s)} E[G_t | S_t = s, A_t = a]$$

$$\Rightarrow v_*(s) = \max_{a \in A(s)} \sum_{s', r} p(s', r \mid s, a) \left[r + \gamma v_*(s') \right]$$

Now, we also know $\Rightarrow$

$$v_*(s') = \max_{a' \in A(s')} q_{\pi_*}(s', a')$$

$$\Rightarrow \boxed{v_*(s) = \max_{a \in A(s)} \sum_{s', r} p(s', r \mid s, a) \left[r + \gamma \max_{a' \in A(s')} q_{\pi_*}(s', a') \right]}$$