Date : / /
Page No.

Q2 Ex 3.15

$$G_t = \sum_{k=1}^{\infty} \gamma^{k-1} \cdot R_{t+k}$$

If we add c to each reward then the new G_t will be:

$$G_t' = \sum_{k=1}^{\infty} \gamma^{k-1} (R_{t+k} + c)$$

$$= \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k} + \sum_{k=1}^{\infty} \gamma^{k-1} \cdot c$$

$$= G_t + c\left(\frac{1}{1-\gamma}\right) \quad \text{(GP formula)}$$

Taking expectation on both sides conditioned on state s

$$E[G_t' | s] = E\left[G_t + \frac{c}{1-\gamma} \middle| s\right]$$

$$\Rightarrow V_\pi'(s) = E[G_t | s] + \frac{c}{1-\gamma} \quad \text{(as c \& } \gamma \text{ are constants)}$$

$$V_\pi'(s) = V_\pi(s) + \frac{c}{1-\gamma} \quad \forall\ s \text{ in states.}$$

Thus, adding constant c adds value of $v_c = \frac{c}{1-\gamma}$ to values of all states.

Q3. Ex 3.16 Consider that a state s ends in T steps sq!

$$V_\pi(s) = \sum_{k=1}^{T} \gamma^k R_{t+k}$$

$$V_\pi(s)' = \sum_{k=1}^{T} \gamma^k (R_{t+k} + c) = \sum_{k=1}^{T} \gamma^k \cdot c + V_\pi(s)$$

$$\boxed{V_\pi(s)' = V_\pi(s) + \frac{c(1-\gamma^T)}{(1-\gamma)}}$$

Date : / /
Page No.

Now here the constant term added for each state s depends on the no. of timesteps T after which the terminal state is reached. Since these timesteps can be diff for each state hence the value func^n of each state will increase by different amount and thus adding c to all rewards will have an effect & it will give preference to states whose episodes last longer (as $1-\gamma^T$ will be more for these states).

Q1. Ex. 3.4

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

The table is obtained using the formula:

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

Since we have been provided with no prob. distribution but only given expected values of rewards for each state so we'll assume that that reward$_s$ comes with prob = 1 & other with prob = 0

$$r(s,a,s') = E[r \mid s,a,s']$$
$$= \sum_r r \cdot p(r \mid s,a,s')$$

So for low, search, low, r_{search}:

$$r(low, search, low) = \sum_r r \cdot p(r \mid low, search, low)$$

$$r_{search} = \sum_r r \cdot \frac{p(r, low \mid low, search)}{p(low \mid low, search)}$$

$$r_{search} = r_{search} \cdot \frac{p(r, low \mid low, search)}{\beta}$$

$$p(r, low \mid low, search) = \beta$$

Q5. We know that by def^n V_* is the optimal state value func^n and q_* is the optimal state action value func^n, so from a given state, V_* is the max of all possible q_* possible from that state (based on all diff actions).

Hence $$V_*(s) = \max_a q_*(s,a)$$