Kellogg School of Management

VASISHT DUDDU
2015137

Q5. Give an equation for v^* in terms of q^*.

The optimal state-value function v^* is maximum over all policies,

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

where $v_\pi(s) = E_\pi[G_t \mid S_t = s]$

The optimal action-value function $q^*(s,a)$ is maximum over all $q_\pi(s,a)$

$$q^*(s,a) = \max_{\pi} q_\pi(s,a)$$

where $q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = a]$

$$q^*(s,a) = R_{t+1} + \gamma \sum_{s'} \pi(a|s')\, v^*(s')$$

R_{t+1} : $E[R_t \mid S_t = s, A_t = a]$

$$\boxed{v^*(s) = \max_{a} q^*(s,a)}$$

Q3. Exercise 3.15

$$v_\pi^*(s) = E_\pi[G_t \mid S_t = s]$$

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

Add a constant to R_{t+k+1} : $\hat{R}_{t+k+1} = R_{t+k+1} + c$

$$\hat{V}_\pi(s) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k \hat{R}_{t+k+1} \mid S_t = s\right]$$

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

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

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

$$= V_\pi(s) + C \cdot \underbrace{\frac{1}{1-\gamma}}_{V_C}$$

V_C is added to all states & hence relative value does not change.

$$V_C = \frac{C}{1-\gamma}$$

Exercise 3.16 In episodic task, we have fixed no. of episodes, say K.

$$V_\pi(s) = E[R_{t+1} + \cdots \gamma^{k-1} R_{t+k} \mid S_t = s]$$

$$\bar{V}_\pi(s) = E\left[(R_{t+1} + C) + \gamma(R_{t+2} + C) + \cdots + \gamma^{k-1} R_{t+k+1}\right] \mid S_t = s$$

$$= V_\pi(s) + C\left[\frac{1-\gamma^k}{1-\gamma}\right]$$

Since, constant value is added to all states, the relative importance of state value does not change.

Kellogg School of Management

Q1.

$$R_{t+1} = E[r_{t+1} \mid s_t = s, a_t = a, s_{t+1} = s']$$

↑ Expected reward at $t+1$

$$= \sum_{r'} r' \, P\{r_{t+1} = r' \mid s_t = s, a_t = a, s_{t+1} = s'\}$$

$$P(r_{t+1} = r' \mid s_t = s, a_t = a, s_{t+1} = s') = \frac{P(s_{t+1} = s', r_{t+1} = r' \mid s_t = s, a_t = a)}{P(s_{t+1} = s' \mid s_t = s, a_t = a)}$$

$$P(s_{t+1} = s', r_{t+1} = r' \mid s_t = s, a = a_t) = P(r_{t+1} \mid s_t, a_t, s_{t+1}) \times P(s_{t+1} \mid s_t, a_t)$$

We are given $P(r \mid s, a, s')$ & $P(s' \mid s, a)$ in the table.

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