Lec-7, IT 567, 24-25

$E_\pi[\cdot]$ denotes the expected value of a RV. given that the agent follows policy π & t is any time step.

policy is a mapping from states to probabilities of selecting each possible action. It is denoted as $\pi(a|s)$, which is basically the prob. that $A_t = a$ if $S_t = s$.

$\Rightarrow$ agent is following policy π at time t.

'|' in $\pi(a|s)$ reminds that it defines a prob. distribution over $a \in A(s)$ for each $s \in S$. $\Rightarrow$ i.e. $\sum_{a \in A(s)} \pi(a|s) = 1$

true for each $s \in S$

$V_\pi \rightarrow$ called as state-value function for policy π.

Another quantity :- the 'value' of taking action a in state s under a policy π $\triangleq$ $q_\pi(s,a)$ as the expect-ed return starting from s, taking the action a, & thereafter following policy π.

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

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

comment ?

$q_\pi(s,a) = v_\pi(s')$

(s, a, s')

q_π $\doteq$ called the action-value function for policy π.

As in return, value function (used in RL & DP) satisfy recursive relationship. We derive it, & it is the famous Bellman equation for V_π. It expresses a relationship b/w the value of a state & the values of its successor state

$$V_\pi(s) \doteq E_\pi[G_t \mid S_t = s] = E_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma E_\pi[G_{t+1} \mid S_{t+1} = s'] \right]$$

$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V_\pi(s') \right], \quad \forall s \in \mathcal{S}$$

$$E[X|Y] = \sum x\, p(x|y) \overset{a}{=} \sum_z \sum_x x\, p(x,z|y) \overset{b}{=}$$

$$\sum_z \sum_x x\, p(z|y)\, p(x|y,z) \overset{c}{=} \sum_z E[X|y,z]\, p(z|y)$$

(a) $\rightarrow$ $p(x|y) = \sum_z p(x,z|y)$ $\because$ marginal distribution

(b) $\rightarrow$ $p(a,b) = p(a)\,p(b|a)$ or $p(a,b|c) = p(a|c)\,p(b|a,c)$

(c) $E[X|y,z] = \sum_x x\, p(x|y,z)$

(see (c) above)

$$E[G_{t+1} | S_t = s] = \sum_{s'} p(s'|s)\, E_\pi[G_{t+1} | S_t = s, S_{t+1} = s']$$

$$V_\pi(s') = E_\pi[G_{t+1} | S_{t+1} = s']$$

$S_t = s$ $\longrightarrow$ $S_{t+1} = s'$