Lec-14, IT567, 24-25

$E[X|Y] \overset{?}{=} f(y)$; $E_2[E_1[X|Y]]$, the first expectation is over x, for a given y. the second one is over y.

$\int_x x p(x|y) dx$

or $\sum x p(x|y)$

Substituting $E_{\pi'}[R_{t+2} + \gamma V_\pi(S_{t+2}) \mid S_{t+1}, A_{t+1} = \pi'(S_{t+1})]$

as $E_{\pi'}[R_{t+2} + \gamma V_\pi(S_{t+2}) \mid S_{t+1}]$ $\rightarrow$ follows from top of slide no. 2 of lec 13.

Going back to the last inequality on slide 3 of lec13, we get

$$V_\pi(s) \le E_{\pi'}\left[R_{t+1} + \gamma E_{\pi'}\left[R_{t+2} + \gamma V_\pi(S_{t+2}) \mid S_{t+1}\right] \mid S_t = s\right]$$

$$= E_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 V_\pi(S_{t+2}) \mid S_t = s\right]$$

To show $\gamma E_{\pi'}\left[E_{\pi'}\left[\overbrace{R_{t+2} + \gamma V_\pi(S_t}^{x}{}_{+2}) \mid S_{t+1}\right] \mid S_t = s\right]$ — ①

$$= \gamma E_{\pi'}\left[R_{t+2} + \gamma V_\pi(S_{t+2}) \mid S_t = s\right] \quad \text{— ②}$$

① $\rightarrow$ $E_{\pi'}\left\{\left[\sum_{r', s''} x \, p(r', s'' \mid s')\right] \mid S_t = s\right\}$

$$= \sum_{s'} \sum_{r', s''} x \, p(r', s'' \mid s') \cdot p(s' \mid s)$$

$\hookrightarrow \sum_{r', s''} x \, p(r', s'' \mid s)$

$\quad \parallel$

$\sum_{s'} p(r', s'' \mid s', s) \, p(s' \mid s)$

$p(r', s'' \mid s', s) = p(r', s'' \mid s')$ ∵ MDP

In all states, given $v_\pi(s)$, i.e., expected return follow-ing policy π.

$q_\pi(s, a) \quad \forall a \in A(s)$

& over all s.

$\hookrightarrow E[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a]$

s \ a	a_1	a_2	a_3	a_4	a_5
s_1	–	–	✓	✓	✓
⋮					
s_n	✓	✓	✓	–	–

$s_1 \xrightarrow{a_3, r_1} s_2 \quad v_\pi(s_2)$

$s_1 \xrightarrow{a_4, r_2} s_1 \quad v_\pi(s_1)$

$s_1 \xrightarrow{a_5, r_3} s_5 \quad v_\pi(s_5)$

So far, we have seen how given a policy & its value function we can easily evaluate a change in the policy at a single state to a particular action.

We can apply changes at all states & to all possible actions, selecting at each state the action that appears best according to $q_\pi(s,a)$.

Consider the new 'greedy' policy π', given by

$$\pi'(s) \triangleq \arg\max_a q_\pi(s,a)$$

$$= \arg\max_a E\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, A_t = a\right]$$

$A = \{1, -1, 3, 9\}$

$\arg\max A = 4$

$\max A = 9$

$$= \arg\max_a \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V_\pi(s') \right]$$

The process of making a new policy that improve on an original policy, by making it greedy w.r.t the value function of the original policy is called policy improvement (PI).

Suppose the new greedy policy π' is as good as, but not better than the old policy π. Then

$V_\pi = V_{\pi'} \quad \forall s \in \mathcal{S}^+$