Lec-13, IT527, 24-25

Policy improvement theorem :-

Let π & π' be pair of deterministic policies s.t. $\forall s \in S$

$$q_\pi(s, \pi'(s)) \geq V_\pi(s) \quad -①$$

then π' must be as good as or better than π, i.e, it must obtain $\geq$ expected return from all states $s \in S$

$$V_{\pi'}(s) \geq V_\pi(s) \quad -②$$

$\pi(s) = a_1$

$a_2 \neq \pi(s)$

If $q_\pi(s, a_2) > V_\pi(s)$

then following an 'a' is always beneficial

Should be

now, if there is strict inequality at any state in ① $\Rightarrow$ " in ② for that state.

We know that $\forall s \in S$

$$V_\pi(s) \leq q_\pi(s, \pi'(s))$$

$$= E\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, A_t = \pi'(s) \right]$$

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

R.H.S $\sum_{s'} \sum_{r} \times p(s', r \mid s)$

$\sum \sum \sum \pi'(a|s) p(s', r \mid s, a) \times$
$s' \; r \; a$

because π' is det.

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

$E[X \mid Y, Z]$

$= \sum_x p(x \mid y, z)$

$p(z \mid y)$

$p(x \mid y, z)$

$X, Y, Z = X + Y$

$E[Z]$

$Z_1 = X_1 + Y_1$

$\sum_x \sum_y (x+y) p(x, y)$

$p(x, z \mid y) =$

$p(z \mid y) \; p(x \mid y, z)$

$p(x \mid y) = \sum_z p(x, z \mid y)$

$= \sum_z p(z \mid y) \; p(x \mid y, z)$

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

$\because \quad V_{\pi}(s) \leq q_{\pi}(s, \pi'(s)) \quad \forall s \in S$

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