RL HW - 1

Priyash Shukla
2018MEB1022

Q1.

$v \geq u$. P.t. $Bv > Bu$.

B is the optimality bellman operator.

$$\therefore \quad B(v(s)) = \max_a \left[\sum_{s',r} p(s', r|s,a) \, [r + \gamma v(s')] \right].$$

value inside the max operator

$$\sum_{s',r} p(s', r|s,a) \, [r + \gamma v(s')]$$

$$\geq \sum_{s,r} p(s', r|s,a) \, [r + \gamma u(s')]$$

$\because \; v(s') \geq u(s')$.

$\therefore$ For any action the value of LHS (with v) > RHS (with u). But this guarantees that the best action for v > best action for u.

$\therefore$ if the value of RHS with the best action 'a' is determined, it is certain the LHS will have a value $\geq$ at that point. (or maybe somewhere else too then that will be the max.

$$\therefore \quad \max \sum_{s'r} p(s', r|s,a) \, [r + \gamma v(s')]$$
$$\geq$$
$$\max \sum_{s'r} p(s', r|s,a) \, [r + \gamma u(s')]$$

$\Rightarrow \; Bv \geq Bu$

Q2. $[v_0, v^1, v^2 \dots v^n \dots v^*]$ = iterations of value itr algorithm.

$\therefore \|Bv^{n-1} - Bv^*\| \leq \|v^{n-1} - v^*\| \gamma$

$\Rightarrow \|v^n - v^*\| \leq \gamma \|v^{n-1} - v^*\|$

$\because \|v^{n-1} - v^*\| \leq \gamma \|v^{n-2} - v^*\|$ (or) $v^k - v^* \leq \gamma \|v^{k-1} - v^*\|$

$\therefore \boxed{\|v^n - v^*\| \leq \gamma^n \|v^0 - v^*\|}$

Also, from Δ inequality,

$\|v^0 - v^*\| - \|v^1 - v^*\| \leq \|v^0 - v^1\|$

$\|v^0 - v^*\| \leq \|v^0 - v^1\| + \|v^1 - v^*\|$

on extending,

$\|v^0 - v^*\| \leq \|v^0 - v^1\| + \|v^1 - v^2\| + \|v^2 - v^3\| \dots \cancel{\|v^k - v^*\|} \to 0$

$\boxed{\|v^n - v^*\| \leq \gamma^n (\|v^0 - v^1\| + \|v^1 - v^2\| + \|v^3 - v^2\| \dots)}$

Also, $\|Bv^k - Bv^{k-1}\| \leq \gamma \|v^k - v^{k-1}\|$

$\|v^{k+1} - v^k\| \leq \gamma \|v^k - v^{k-1}\|$

Telescopically, $\|v^{k+1} - v^k\| \leq \gamma^k \|v^1 - v^0\|$

$\hookrightarrow \|v^2 - v^1\| \leq \gamma \|v^1 - v^0\|$

$\hookrightarrow \|v^3 - v^2\| \leq \gamma^2 \|v^1 - v^0\| \dots$

$$\|v^n - v^*\| \leq \gamma^n (\|v^0 - v^1\| + \gamma \|v^0 - v^1\| + \gamma^2 \|v^0 - v^1\| + \dots)$$

$$\leq \gamma^n \|v^0 - v^1\| (1 + \gamma + \gamma^2 + \dots)$$

$$\leq \frac{\gamma^n}{1-\gamma} \|v^0 - v^1\|$$

Q3

1. T.C of value iteration

$\Rightarrow$ $O(|S|^2 |A|)$ for each iteration, but there will be many iterations. [linear convergence]

For every state the no. of new steps it can reach by taking any action $\Rightarrow$ $|S| * |A|$ (maximum)

There are $|S|$ such states

$|S| \equiv$ no of states

$|A| \equiv$ No of actions possible at any state

$TC = \therefore O(|S| * |S| * |A|) = O(|S|^2 |A|)$

2. Policy Iteration

$\Rightarrow$ $O(|S|^3 + |S|^2 |A|)$ $\rightarrow$ Lower # of iterations.

The $|S|^2 |A|$ comes from the value iteration part, but the policy evaluation part adds to the overall TC.

Normal policy evaluation $\Rightarrow$ $O(|S|^3)$

$\therefore$ overall TC of policy itr $\Rightarrow$ $O(|S|^3) + O(S^2 A)$

$= O(|S|^3 + |S|^2 |A|)$

3. Modified Policy iteration $\rightarrow$ Lower # of iterations.

$\Rightarrow$ $O(|S|^2 k + |S|^2 |A|)$

$\rightarrow$ same as policy itr except the evaluation runs for k steps only. $\Rightarrow$ this step $= O(|S| k)$

There are $|S|$ such states $\Rightarrow$ $O(|S|^2 k)$

overall $\Rightarrow$ $O(|S|^2 (k + |A|))$.

Q4) $q_\pi(s,a) > v_\pi(s)$

$$v_\pi(s) = \sum_{a'} \pi(a'|s)\, q_\pi(s,a') \quad , \quad a' \in A(s).$$

$\therefore$ The value func. is the expected value of $q_\pi(s,a')$, or a weighted avg of it, the weights are the distribution of $\pi(a'|s)$

$\pi(a'|s)$ depends on the policy.

It is given that for some $s \in S$ & $a \in A(s)$,

$$q_\pi(s,a) > v_\pi(s).$$

$\Rightarrow$ This means that there exists a state s where an action a can be made that results in a higher expected reward than the one given by the policy π.

$\Rightarrow$ This means there exists another policy π^{new} where the action taken will be more exploitory

~~great~~ $\pi^{new}(a|s) > \pi(a|s)$

& will increase ~~$v_{\pi^{new}}$~~ $v_{\pi^{new}}(s)$.

$\therefore$ $v_\pi \neq$ an optimal policy for sure.

Q5. $0 < \gamma < 1$ & $\gamma \neq 1$ $\because$ (Horizon is inf)

Now, $V_\pi(S) = E_\pi[G_t | S_t = s]$, $\forall s \in S$

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

$\therefore$ if v_π^{new}, $R_{k+1}^{new} = c + R_{k+1}$

$$\Rightarrow v_\pi^{new}(S) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k (c + R_{k+1}) \middle| S_t = s\right]$$

$$\Rightarrow v_\pi^{new}(S) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{k+1} \middle| S_t = s\right] + E_\pi\left[\sum_{k=0}^{\infty} \gamma^k c\right]$$

$$= v_\pi(S) + \underbrace{E_\pi\left[\sum_{k=0}^{\infty} \gamma^k c\right]}_{\text{constant}}$$

General.

$$\boxed{v_\pi^{new}(S) = v_\pi(S) + \sum_{k=0}^{n} \gamma^k c}$$

for this question,
$0 < \gamma < 1$ & $n \to \infty$

$$\therefore \quad v_\pi^{new}(S) = v_\pi(S) + \sum_{k=0}^{\infty} \gamma^k c$$

$$= v_\pi(S) + [c + c\gamma + c\gamma^2 + \cdots \infty]$$

$$= v_\pi(S) + \frac{c}{1-\gamma}$$

Q6

a.) shortest path $\rightarrow$ least amount of steps.

$\therefore$ if $R_s = -1$, when the agent will try to maximize returns, it will find the shortest path.

$$V_*(S) = \max_a \sum_{s', r} P(s', r | s, a) [r + V_*(s')].$$

$\because$ Actions are deterministic, $P'(s', r | s, a) = 1$.

Given $\rightarrow$ $R_G = +5$, $\gamma = 1$.
$R_D = -5$

$R_{12} = +5$ $\quad R_5 = -5$

$$V_*(S) = \max_a \sum_{s', r} [r + V_*(s')]$$

$$\therefore V_*(8) = V_*(11) = \max(-1+5, 0, 0) = \max(4, 0, 0) = 4.$$

For $V_*(7)$, it would be $\max(-1+4, 0, 0) = 3$.

$\therefore$ It can be observed that the value of any state $\Rightarrow$
5 − (Its distance from G=12).

$$\left.\begin{matrix} V_1 = 0 & V_6 = 2 & V_{10} = 3 & V_{15} = -1 \\ V_2 = 1 & V_7 = 3 & V_{11} = 4 & V_{16} = -2 \\ V_3 = 2 & V_8 = 4 & V_{13} = 1 & V_5 = 0 - 5 \\ V_4 = 3 & V_9 = 2 & V_{14} = 0 & V_{12} = 0 + 5 \end{matrix}\right\}$$ for optimum value function.

---

b.) From Q5. except for states = ∞, it is the same question.
[the general solⁿ]

$\gamma = 1.$
$c = +2$

$$V_\pi^{new} = V_\pi + \sum_{K=0}^{n} c\,1^K$$

$$V_\pi^{new} = V_\pi + 2 \sum_{K=0}^{K=n} (1)$$

$$V_\pi^{new} = V_\pi + 2(n+1)$$

[$n \equiv$ no. of steps from current state to target $S = 12$.

$\therefore V_\pi = +5 - n$.

$$\Rightarrow V_\pi^{new} = V_\pi + 2(6 - V_\pi)$$
$$= V_\pi + 12 - 2V_\pi$$
$$= 12 - V_\pi$$

$V_1^{new} = 10$	$6 = 8$	$10 = 7$	$15 = 11$
$2 = 9$	$7 = 7$	$11 = 6$	$16 = 12$
$3 = 8$	$8 = 6$	$13 = 9$	$V_5^{new} = -3$
$4 = 7$	$9 = 8$	$14 = 10$	$V_{12}^{new} = 7$

$\longrightarrow$ [these values + 2.]