Qn 1. Given $u \in V$, $v \in V$. and $v \geq u$

Let there be s states denoted by $1 \ldots s$.

Now we know that

$$V_1 > U_1$$
$$V_2 > U_2$$
$$\vdots$$
$$V_s > U_s$$

where V_i represents value corresponding to value function 'v' and state 'i'.

Similar representation for u_i.

Now.., $BV = \max_a \sum P(s', r|s,a)[r + \gamma V(s')]$ ... ①

and $Bu = \max_a \sum P(s', r|s,a)[r + \gamma u(s')]$ ... ②

Since both u and $v \in V$.., probabilities and reward will be same.

Let say that ② is maximum when we goto state s_k

i.e. $\max_a \sum P(s', r|s,a)[r + \gamma u(s')]$ is maxm when s' is s_k.

(w.l.o.g we can say action taken is a_k)

~~But ...,~~

~~$\sum p(s', r|s,a)[r + \gamma V(s')] \geq \sum p(s', r|s,a)[r + \gamma u(s_k)]$~~

i.e. when we take action a_k and we are in state s and by taking action a_k we reach s_k then ② is maxm.

Also we know that for each state $V_i \geq U_i$

$\Rightarrow V_{s_k} \geq U_{s_k}$. Therefore for same action taken value of ① $\geq$ ② Hence $Bv \geq Bu$. Note that a_k action may not give maxm for Bv but we are sure that if for action a_k, ① $\geq$ ② then for $Bv \geq Bu$ will also hold.

Ques 2. Given $\{v_0, v_1, v_2 \ldots v_n \ldots\}$ successive iteration value function.

v^* ← optimal value function

$$v_1 = Bv_0$$

$$v^2 = B(v_1) = B(Bv_0) = B^2(v_0)$$

$$\vdots$$

$$v^n = B^n(v_0).$$

and

$$v^* = Bv_* = B^n v_* \quad \text{since } v^* \text{ is optimal.}$$

Now...

$$\|v^n - v^*\| = \|B^n(v_0) - B^n(v^*)\|$$

$$\leq \gamma^n \|v_0 - v^*\| \quad \ldots (1)$$

$$\|v_0 - v_*\| \leq \|v_0 - v_1\| + \|v_1 - v_2\| \ldots\ldots$$

$$\leq \|v_0 - v_1\| (1 + \gamma + \gamma^2 + \ldots)$$

$$\leq \frac{\|v_0 - v_1\|}{1 - \gamma} \quad \ldots (2)$$

combining (1) and (2)...,

$$\|v_n - v^*\| \leq \frac{\gamma^n}{1-\gamma} \|v_0 - v_1\|.$$

(Hence proved).

Ques 3.

(a)

loop:

$\Delta \leftarrow 0$

loop for each $s \in S$: } ← Runs for $|S|$ times

$v \leftarrow V(s)$

$V(s) \leftarrow \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$ } Runs for $|A| \times |S|$ times

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

until $\Delta < \theta$

Runs for $|A||S|^2$ times

Output a deterministic policy $\pi \approx \pi^*$, such that

$\pi(s) = \arg\max_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$.

$\rightarrow$ $O(|S|^2|A|)$

as $|A|$ actions and we can goto $|S|$ states (next states)

Ques 3. Given S states, A actions.

(b)

1. Initialization

2. Evaluation ← $|S|^3$ times

loop :

$\Delta \leftarrow 0$

S times ← loop for each $s \in S$:

$v \leftarrow V(s)$

at most S times as at most S next steps ← $V(s) \leftarrow \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

until $\Delta < \theta$.

} S times × S times

↳ Terminates atmost S times

So → $|S| \times |S| \times |S|$ times.

$|A| \times |S| \times |S|$

3. Improvement

policy stable ← true

for each $s \in S$ ← S times

old action ← $\pi(s)$

$\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$.

if old action ≠ $\pi(s)$, policy stable ← false

} $|A| \times |S|$ times for each action check ↓ we can go to next S atmost state

if policy stable then stop!

Overall time complexity : $|S|^3 + |A||S|^2$ times.

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

Ques 3.

(C) if policy evaluation step runs for k iteration then it will take $O(k|S|^2)$.

Policy improvement will be same i.e. $O(|S|^2|A|)$.

$\longrightarrow$ Overall : $O(k|S|^2 + |S|^2|A|)$.

Ques 4 given $q_\pi(s,a) > v_\pi(s)$ ...

Policy improvement theorem states that if $v_\pi(s) \leq q_\pi(s, \pi'(s))$

then $v_\pi(s) \leq v_{\pi'}(s)$ ... ①

here if the action a is taken by policy π' i.e. $\pi'(s) = a$

then ... using policy improvement theorem ...,

if $\pi'(s) = a$ ... ②

$q_\pi(s,a) > v_\pi(s)$ ... ③

Then $v_{\pi'}(s) > v_\pi(s)$ using ①, ② and ③ ..

hence we can say that there exists a better policy than π for given case i.e. π is not an optimal policy.

~~Proof of policy imp. theorem~~ ...

$$v_\pi(s) \leq q_\pi(s, \pi'(s)) \quad (\text{where } \pi'(s) = a)$$

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

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

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

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

$$\vdots$$

$$< E_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \dots \mid S_t = s]$$

$$< v_{\pi'}(s).$$

Solution: Earlier.., if we are in state s and we goto state s_1 to s_2

Ques 5.

$$s \xrightarrow{r_1} s_1 \xrightarrow{r_2} s_2 ----$$

expected return will ~~be~~ be $r_1 + \gamma r_2 + \gamma^2 r_3 \ldots$ for (s)

now..,

$$s \xrightarrow{r_1+c} s_1 \xrightarrow{r_2+c} s_2 ----$$

i.e. $(r_1 + \gamma r_2 + \ldots) + (c + \gamma c + \gamma^2 c + \ldots)$

$\frac{c}{1-\gamma}$

$$V_\pi^{new}(s) = U_\pi(s) + \frac{c}{1-\gamma}.$$

Since states are infinite and no terminal state this is valid for all s.

$$V_\pi^{new} = U_\pi + \frac{c}{1-\gamma}$$

Ques 6.

(a) $R_S = -1$.

$V(1) = 0$

$V(2) = +1$

$V(3) = +2$

$V(4) = +3$

$V(5) = -5$

$V(6) = +2$

$V(7) = +3$

$V(8) = +4$

$V(9) = +2$

$V(10) = +3$

$V(11) = +4$

$V(12) + 5$

$V(13) = +1.$

$V(15) = -1$

$V(16) = -2$

$V(14) = 0$

(b) If we are using previous policy, then we will still get the shortest path but value function will be different.

$V(1) = +12$
$V(2) = +11$
$V(3) = +10$
$V(4) = +9$
$V(5) = -3$

$V(6) = +10$
$V(7) = +9$
$V(8) = +8$
$V(9) = +10$
$V(10) = +9$
$V(11) = +8$

$V(12) = +7$
$V(13) = +11$
$V(14) = +12$
$V(15) = +13$
$V(16) = +14$.