Q(3) <b>

$p(Q|S)$ — probability of querying in state Stale
$p(Q|F)$ " " " " " fresh

We will find

$$V_{t+i}(\text{state}=S) = E[R_t + R_{t+i}\gamma^i$$

$$V_{t+i}(\text{state}=S) = E[R_{t+i} + \gamma^2 R_{t+2+3}$$

$$V_{t+1}(\text{states}=S) = E[R_{t+i} + \gamma R_{t+i+k} \quad | S_t = S]$$

$$+ p(S_{t+3}=F | S_t=S)[10] + p(S_{t+3}=S|S_t=S)[-10]$$

$$E[R_{t+i} + \ldots | S_t = S] = \sum_{k=0}^{2} \gamma^k E[R_{t+2+k} | \text{state}=S]$$

I have calculated $V_3(F)$.

policy will be dynamic because $V_t(S,F)$ changes with time-step (i)

Q13) <b>

$V_3(F)$ = Reward after 3 time steps

$= E[R_{t+1} | F] + \gamma E[R_{t+2} | S_t = F] + \gamma^2 [R_{t+3} | S_t = F]$

$[1 + p(S_{t+3} = F | S_t = F)][-10] +$

$p(S_{t+3} = F | S_t = F)[10]$

Last two terms Represent additional penalty or reward

$E[R_{t+1} | F] = p(Q|F)(-4) + (1 - p(Q|F))(+4)$

$= 4 - 8p(Q|F)$ — (i)

$E[R_{t+2} | F] = p(S_{t+1} = F | S_t = F) E[R_{t+2} | S_{t+1} = F]$

$+ [1 - p(S_{t+1} = F | S_t = F)] E[R_{t+2} | S_{t+2} = S]$

$= p(S_{t+1} = F | S_t = F) \{ E[R|F] + E[R|S] \} + E[R|S]$

$\rightarrow p(S_{t+1} = F | S_t = F) = p(Q|F)(0.9) + (1 - p(Q|F))(0.5)$

$p(S_{t+1} = F | S_t = F) = 0.5 - 0.4 p(Q|F)$

$E[R|S] = p(Q|S)(-8) + [1 - p(Q|S)] 4$

$= 4 - 12 p(Q|S)$ — (ii)

$[R_{t+2} | S_t = F] = [0.5 - 0.4 p(Q|F)][1 - 8p(Q|F) - 4 + 12p(Q|S)]$

$+ 4 - 12 p(Q|S)$

$= -0.15 + p(Q|F)[-4.0 - 0.4 + 3.2 p(Q|F)] + 4$

$+ p(Q|S)[-6] - 3.6 p(Q|S) p($

$E[R_{t+3} \mid S_t = F] =$

$p[S_{t+2} = F]\, E[R \mid F] + (1 - p[S_{t+2} = F \mid S_t = F])\, E[R \mid S]$ ← (iii)

$\Rightarrow p[S_{t+2} = F \mid S_t = F] = p(Q \mid S_t = F)(0.9)\, p(Q \mid S_t = F)(0.9)$

$+ p(Q \mid F)(0.5)\, p(Q \mid S)(0.8) +$

$[1 - p(Q \mid F)](0.5)\, p(Q \mid F)(0.9) +$

$[1 - p(Q \mid F)](0.5)[1 - p(Q \mid F)](0.5)$

~~$E[R_{t+3} \mid S_t = F]$~~

$\Rightarrow p(Q \mid F)[0.81\, p(Q \mid F) - 0.45\, p(Q \mid F) + 0.45]$

$+ p(Q \mid S)\, p(Q \mid F)(4) +$

$+ p(Q \mid S)[4 - 4\, p(Q \mid F)]\, 0.25[p(Q \mid F)^2 - 2p(Q \mid F) + 1]$

$\Rightarrow p(Q \mid F)[-1.19\, p(Q \mid F)$

~~$E[R_{t+3} \mid S_t = F]$~~ $= p(Q \mid F)$

$= p(Q \mid F)[1.36\, p(Q \mid F) - 2] +$

$p[S_{t+2} = F \mid S_t = F] =$

$= p(Q \mid F)[1.36\, p(Q \mid F) - 1.55] + 0.25 + p(Q \mid S)\, p(Q \mid F)^2$ ← (iv)

Put value of $p[S_{t+2} = F \mid S_t = F]$ in eq (iii)

Since, We have Calculated

$E[R_{t+i} | S_t = F] \quad 1 \le i \le 3$

~~$p[S_{t+2} = F | S_t =$~~ in terms of

$p(Q|S)$ and $p(Q|F)$

Now calculate

$$p[S_{t+3} = F | S_t = F]$$

$$= p[S_{t+3} = F | S_{t+2} = F] \, p[S_{t+2} = F | S_t = F]$$

$\Rightarrow$

$$+ p[S_{t+3} = F | S_{t+2} = S] \left(1 - p(S_{t+2} = F | S_t = F)\right)$$

$= p(Q|F)(0.9) + p$ $p(S_{t+2} = F | S_t)$ can be taken from eq (iv)

$$p[S_{t+3} = F | S_{t+2} = F] = p(Q|F)\,0.9 + (1 - p(Q|F))(0.5)$$

$$p[S_t = F | S_{t+2} = S] = p(Q|S)(0.8) + (1 - p[Q|S])(0.$$

Now we can put all value's in equation of $V_3(F)$ to get state value, ~~same~~ approach for $V(S)$

follow same approach to calculate $V_2(F)$, $V_1(F)$
V_i (state = state) $1 \le i \le 3$

~~V_i (state_fi~~

$V_i(S_{t+i} = s)$ $s, s \in \{F, S\}$ $1 \le i \le 3$

$V_i(.)$ represent average ~~awar~~ Reward starting from ~~state~~ time_step $t+i$ to $t+3$ from starting state 's'.

optimal policy calculation :-

at time step $t+i$ and in ~~sta~~ state = s

~~arg max Reward(a|s) + γ~~
a

Take Action 'a' such that

$$\arg\max_{a \in \{Q, NQ\}} \sum_{r \in R} \sum_{s' \in S} p(r, s'|s, a)[r + \gamma V_{3-i}(s'$$

~~Q~~ Q → Query
NQ → not Query

R → set of reward's
S → set of states.

$p(r, s'|s, a)$ is Given in question

$V_{3-i}(s')$ can be calculated by above method

Q13) <c>

$$V_k(F) = p(Q|F)(0.9)[-4 + \gamma V_F] + p(Q|F)(0.1)[-4+\gamma V_S] + (1 - p(Q|F))(0.5)[4+\gamma V_F] + (1-p(Q|F))(0.5)[4+\gamma V_S]$$

$$V_F = V_F[\gamma 0.9\, p(Q|F) + 0.5\gamma - 0.5\, p(Q|F)] + V_S[0.1\gamma\, p(Q|F) + 0.5\gamma - 0.5\, p(Q|F)] + p(Q|F)[-0.45 - 0.4 \div 2 - 2] + 2 + 2$$

$$V_S = p(Q|S)[0.8)[-8 + \gamma V_F] + p(Q|S)(0.2)[-8 + \gamma V_S] + [1 - p(Q|S)](0)[4 + \gamma V_F] + [1-p(Q|S)](1.0)[4 + \gamma V_S]$$

$$V_S = V_F[-0.64\gamma\, p(Q|S)] + V_S[0.2\gamma\, p(Q|S) + \gamma - p(Q|S)] + [-0.64 + (-1.6) + 4]\, p(Q|S) + 4$$

Solve by value policy iteration

$p(Q|S)$ → probability of Query in stale state
$p(Q|F)$ → " " " " fresh "

γ → discounting factor

Ans 14) In policy improvement :-

for a Given state $s \in S$.

Calculate ~~action~~ state-action value for all action's expect deterministic action defind by current policy.

either $a \in A(s)$ such that $a \neq \pi(s)$

let this set of new state-action be SA.

if any action in SA give's more reward than by current policy.

either $q(s, a) > \pi(a|s) \sum\sum p(s',$

$q(s, a) > q(s, a')$

where $a' = \pi(a|s)$

then change $\pi(s) = a$

Otherwise do nothing.

Since, updating $\pi(s)$ [$s \in S$ and $s \neq s$] does not change other policies so, it only update current ~~state-value~~ policy.

So, it leads to policy improvement