Assignment 2

Neetu R.R
PhD 19108.

(1) States = {High, low}

Actions = {search, wait, Recharge}

$p(s' = high \mid s = high, a = search) = \alpha$

$p(s' = low \mid s = high, a = search) = 1-\alpha$

$p(s' = high \mid s = low, a = search) = 1-\beta$.

$\vdots$

The table can be filled as.

s	a	s'	r	$p(s', r \mid s, a)$
high	search	high	r_{search}	α
high	search	low	r_{search}	$1-\alpha$
low	search	high	-3	$1-\beta$.
low	search	low	r_{search}	β
high	wait	high	r_{wait}	1
high	wait	low	–	0
low	wait	high	–	0
low	wait	low	r_{wait}	1
low	recharge	high	0	1
low	recharge	low	–	0

Exercise 3.15

(3) Eqn (3.8)

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

where γ is a parameter, $0 \le \gamma \le 1$ called the discount rate.

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

$$G_t^c = \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + C_{t+k+1})$$

$$= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + C_{t+k+1} \sum_{k=0}^{\infty} \gamma^k$$

$C_{t+k+1} = C$

$$= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \frac{c}{1-\gamma}$$

$$= G_t + \frac{c}{1-\gamma}$$

$$V_c = \mathbb{E}[G_t^c \mid S_t = s]$$

$$= \mathbb{E}\left[G_t + \frac{c}{1-\gamma} \,\middle|\, S_t = s\right]$$

$$= \mathbb{E}[G_t \mid S_t = s] + \mathbb{E}\left[\frac{c}{1-\gamma}\right]$$

$$= \mathbb{E}[G_t \mid S_t = s] + \frac{c}{1-\gamma}$$

$\mathbb{E}[\text{constant}] = \text{constant}$

$\therefore$ V^c does not affect the relative values of any states.

Exercise 3.16.

Episodic task :- $G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$.

The rewards, we will sum over all the rewards to find the best action to be picked.

When adding a constant,

$$G_t^c = \sum_{k=t+1}^{T} \gamma^{k-t-1} (R_k + c)$$

The rewards obtained in the episodic task is very important. Adding a constant can affect the rewards obtained in a state, and affect the action to be picked in the state.

eg: If $R = -5$ & $c = 2$.

The reward at a state $s = -5+2$
$= \underline{-3}$

The reward value increased, and chances of picking a non-best arm is high.

$$q_\pi(s,a) = \mathbb{E}[G_t \mid S_t = s, A_t = a]$$

$$= \mathbb{E}\left[\sum_{k=t+1}^{T} \gamma^{k-t-1} (R_k + c) \mid S_t = s, A_t = a\right]$$

∴ The action value function will also change.

⑤ Optimal state value function $v*$

A policy π is defined to be better than or equal to a policy π' if its expected return is greater than or equal to that of π' for all states. $\pi \geq \pi'$ if and only if $v_\pi(s) \geq v_{\pi'}(s)$. The optimal policy is that policy which is better than or equal to all other policies. denoted by $\pi*$. ~~They~~ It share a ~~same~~ same state value function denoted by $v*$.

$$v*(s) = \max_{\pi} v_\pi(s) \quad \forall s \in S.$$

Optimal action value function $q*$ gives the expected return for taking action a in state s and thereafter following an optimal policy.

$$q*(s,a) = \max_{\pi} q_\pi(s,a)$$

$$\boxed{v*(s) = \max_{a \in A(s)} q_{\pi*}(s,a)}$$

It is taking the best action by taking best policy $\pi*$ into account and giving the best expected return.

(8) Yes, we can express R_{t+2} in terms of S_t & A_t. That means R_{t+2} is dependent on S_t & A_t. This happens when the previous state of S_{t+2} ie S_{t+1} doesn't produce any reward or didn't take any action.

$$\therefore p(S_{t+2}=s'', R_{t+2}=r'' \mid S_t=s, A_t=a) =$$

$$Pr\{S_{t+2}=s'', R_{t+2}=r'' \mid S_t=s, A_t=a\}$$

$$r(s, a, s'') = \mathbb{E}[R_{t+2}=r'' \mid S_t=s, A_t=a, S_{t+2}=s'']$$

$$= \sum_{r \in R} r'' \frac{p(s'', r'' \mid s, a)}{p(s'' \mid s, a)}$$

This happens when the previous state of R_{t+2} ie S_{t+1} does not happen.

(9) $$\mathbb{E}[R_{t+2} \mid S_t=s, A_t=a] = \mathbb{E}_{s''}\left[\mathbb{E}[R_{t+2} \mid S_{t+2}=s'', s, a] \mid s, a\right]$$

$$= \mathbb{E}_{s''}\left[\sum_{r''} r'' \frac{p(s'', r'' \mid s, a)}{p(s'' \mid s, a)} \middle| s, a\right]$$

$$= \mathbb{E}_{s''}\left[r''(s,a,s'') \mid s,a \right]$$

$s'' = S_{t+2}$

$r'' = R_{t+2}.$

$$= \cancel{\sum_{s'' \in S} s'' \, r(s,a,s'')}$$

$$=$$

$$\therefore \mathbb{E}\left[R_{t+2} \mid S_t = s, A_t = a\right] = \mathbb{E}_{S_{t+2} = s''}\left[r''(s,a,s'') \mid s,a \right]$$

(10) The State value function

$$V_\pi(s) = \mathbb{E}_\pi\left[G_t \mid S_t = s \right]$$

As we know, $G_t = R_{t+1} + \gamma \underbrace{G_{t+1}}_{\text{Reward to go function}}$.

If $\gamma \approx 1$, the agent is far-sighted of the future.

$$\therefore V_\pi(s) = \mathbb{E}_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s \right]$$

$$= \sum_r \mathbb{E}\left[R_{t+1} \mid S_t = s \right] + \gamma \mathbb{E}_\pi\left[G_{t+1} \mid S_{t+1} = s' \right]$$

$$= \sum_{s'} \sum_r r + \gamma \mathbb{E}_\pi\left[G_{t+1} \mid S_{t+1} = s' \right]$$

Summing over all actions, to get the state value function.

$$= \sum_{a} \pi(a|s) \sum_{s', r} p(s', r|s, a) \left[r + \gamma v_\pi(s')\right] \quad \forall s \in S.$$

This is the sum of all values of three variables a, s', r.

For each value of $a, s', \& r$, we compute the probability ie ~~$\pi(a)$~~ $\pi(a|s)\, p(s', r|s, a)$ & multiply with the current reward & reward to go term.

Then we sum over all possibilities to get an expected value.

(11) An agent receives a sequence of rewards $R_1 = 2$, $R_2 = -1$, $R_3 = 10$ & $R_4 = -3$.

$\gamma = 0.5$

The infinite horizon discounted return is

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

A:- $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

$G_t = R_{t+1} + \gamma G_{t+1}$

$G_4 = 0.$

$G_3 = R_{t+4} = R_4 = -3.$

$G_2 = R_3 + \gamma G_3 = -3 \times 0.5 + 10.$

$= 8.5$

$G_1 = R_2 + \gamma G_2 = -1 + 0.5 \times 8.5$

$= \underline{\underline{3.25}}$

$G_0 = R_1 + \gamma G_2 = 2 + 0.5 \times 3.25$

$= \underline{\underline{3.625}}.$

The γ discounted reward/return for each time step.

$G_0 = 3.625$

$G_1 = 3.25$

$G_2 = 8.5$

$G_3 = -3$

$G_4 = 0.$

(12) We know, the optimal state value function

$$v*(s) = \mathbb{E}_{\pi*}\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s\right]$$

By Bellman's optimality principle,

$$v*(s) = \max_a q*(s,a)$$

As we have given $v*(s)$,

$$v*(s) = \max_a q*(s,a) = \max_a \left[R_{t+1} + \gamma v*(S_{t+1}) \mid S_t = s, A_t = a\right]$$

$$\therefore \pi*(s) = \underset{a \in A(s)}{\operatorname{argmax}} \mathbb{E}\left[R_{t+1} + \gamma v*(S_{t+1}) \mid S_t = s, A_t = a\right]$$

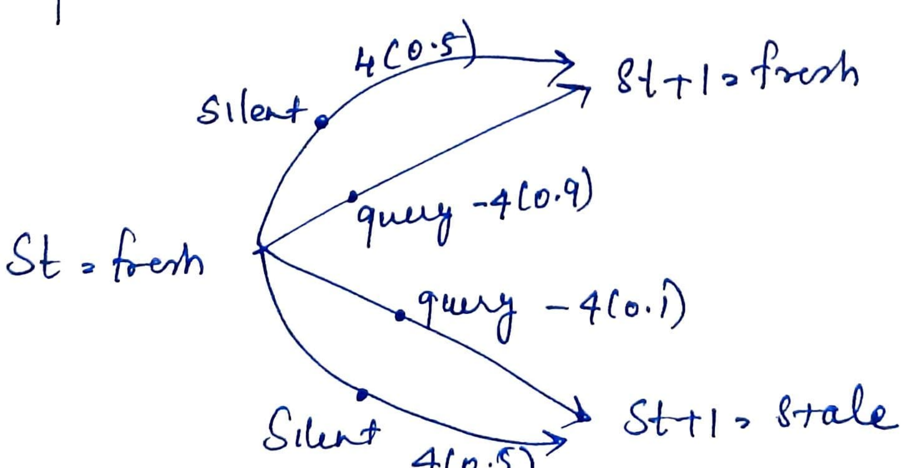
$$\pi*(s) = \underset{a \in A(s)}{\operatorname{argmax}} \sum_{s',r} p(s',r \mid s,a)\left[r + v*(s')\right]$$

(13)

(i) MDP.

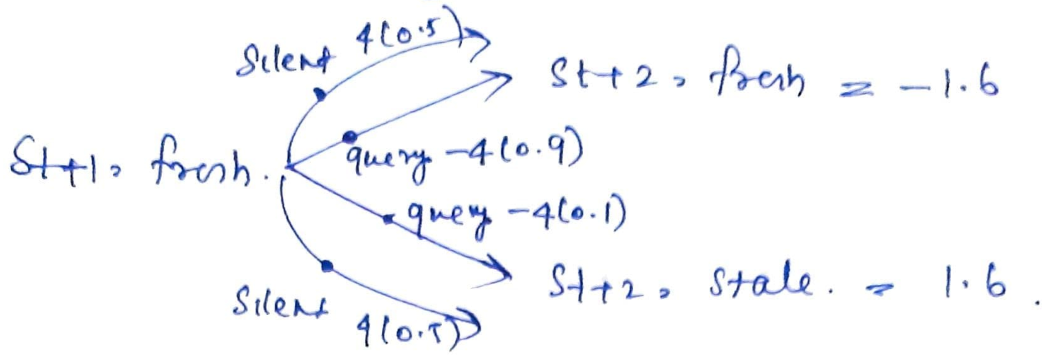
S_t	A_t	S_{t+1}	r_{t+1}	p
fresh	query	fresh	-4	0.9
fresh	query	stale	-4	0.1
Stale	query	fresh	-8	0.8
stale	query	stale	-8	0.2
fresh	Silent	fresh	+4	0.5
fresh	silent	Stale	+4	0.5
stale	silent	~~stale~~ fresh	+4	0
stale	silent	stale	+4	1

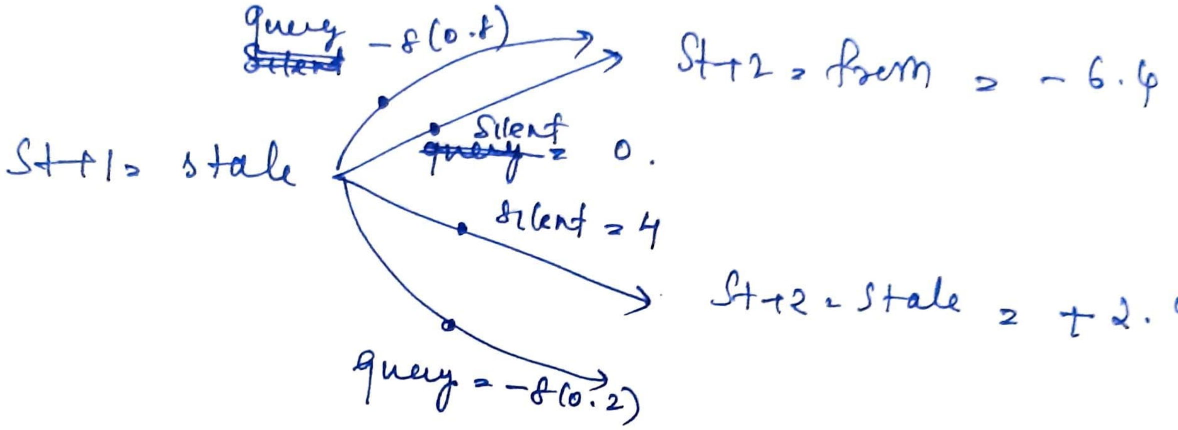
(ii) Starting with state = fresh.

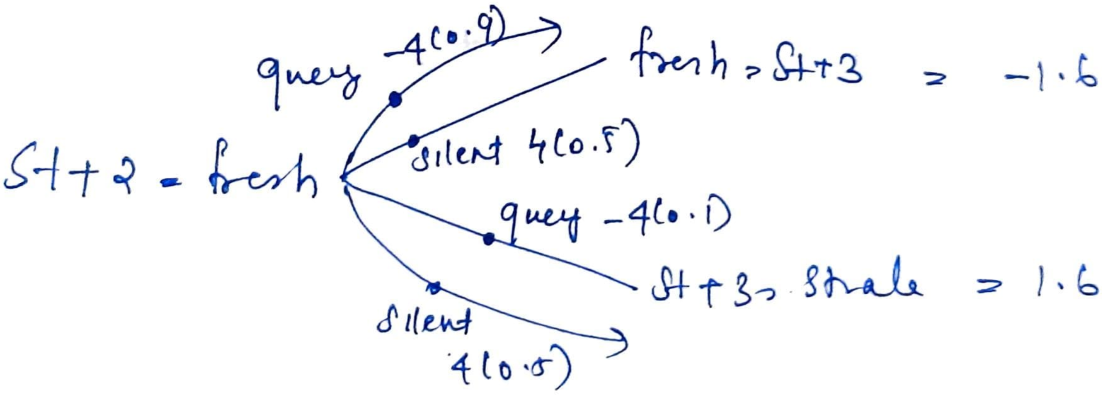


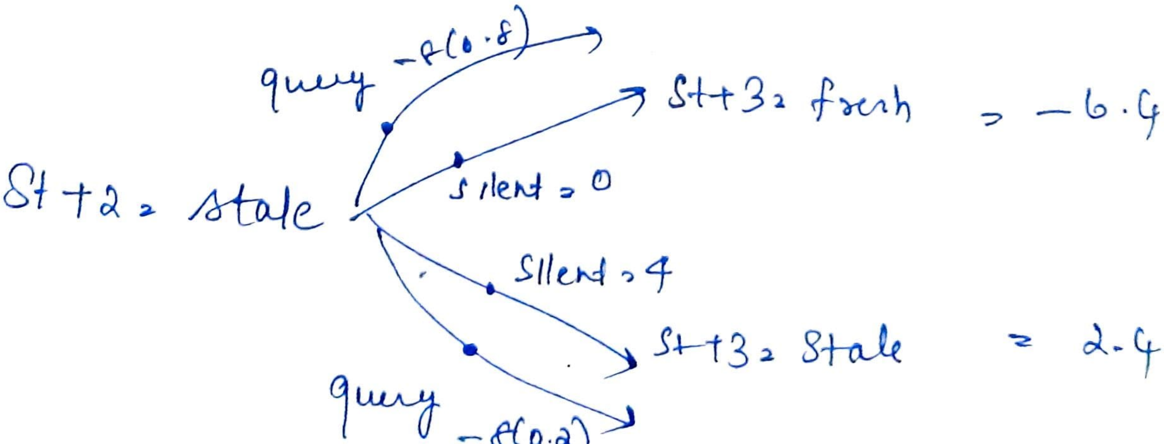
S_{t+1} = fresh = -1.6.

S_{t+1} = stale = 1.6.


Silent 4(0.5)
St+2 = fresh = −1.6
St+1 = fresh.
query −4(0.9)
query −4(0.1)
St+2 = stale. = 1.6.
Silent 4(0.5)


query −8(0.8)
St+2 = fresh = −6.4
St+1 = stale
Silent = 0.
silent = 4
St+2 = Stale = +2.4
query = −8(0.2)


query −4(0.9)
fresh = St+3 = −1.6
St+2 = fresh
silent 4(0.5)
query −4(0.1)
St+3 = Stale = 1.6
silent 4(0.5)


query −8(0.8)
St+3 = fresh = −6.4
St+2 = stale
silent = 0
Silent = 4
St+3 = Stale = 2.4
query −8(0.2)

$$G_t = R_{t+1} + \gamma G_{t+1}$$

S_t	S_{t+1}	S_{t+2}	S_{t+3}	
Fresh(F)	Fresh(F)	Fresh(F)	Fresh(F)	$-1.6 + \frac{1}{2}(-1.6 - 1.6) + 10 = \underline{6.8}$
F	F	F	Stale (S)	$-1.6 + \frac{1}{2}(-1.6 + 1.6) + 10 = \underline{8.4}$
F	F	S	S	$-1.6 + \frac{1}{2}(+1.6 + 2.4) + 10 = \underline{10.4}$
F	F	S	F	$-1.6 + \frac{1}{2}(1.6 + 6.4) + 10 = \underline{6}$
F	S	F	F	$1.6 + \frac{1}{2}(-6.4 - 1.6) + 10 = \underline{7.6}$
F	S	S	F	$1.6 + \frac{1}{2}(2.4 + 6.4) + 10 = \underline{9.6}$
F	S	F	S	$1.6 + \frac{1}{2}(-6.4 + 1.6) + 10 = \underline{9.2}$
F	S	S	S	$1.6 + \frac{1}{2}(2.4 + 2.4) - 10 = \underline{-6}$

The optimal policy is that giving maximum reward

$\therefore \pi_* =$ S_t = Fresh, S_{t+1} = Fresh, S_{t+2} = Stale, S_{t+3} = Stale.

(14) Policy improvement step:-

(i) Either improves policy or leaves it unchanged.

(ii) If it leaves unchanged, the policy is optimal policy.

The operator $T_\pi f(s) = \mathbb{E}[R_{t+1} + \gamma f(s_{t+1}) | s_t]$

$T_\pi V_{\pi k}(s) = V_{\pi k}(s) \quad s \in S.$

$\therefore T[V_{\pi k}] = T_{\pi k+1} V_{\pi k}.$

ie $V_{\pi k+1}(s) \geq V_{\pi k}(s)$

Proof:-

$T_{\pi k+1} V_{\pi k}(s) = T V_{\pi k}(s) \geq T_{\pi k} V_{\pi k}(s) = V_{\pi k}(s)$

$T_{\pi k+1} V_{\pi k}(s) \geq V_{\pi k}(s)$

$\therefore T_{\pi k+1}(T_{\pi k+1}(V_{\pi k}(s))) \geq V_{\pi k}(s)$

$\therefore T^N_{\pi k+1} V_{\pi k}(s) \geq V_{\pi k}(s) \quad \text{——} (1)$

Applying limit

$\lim_{N \to \infty} T^N_{\pi k+1} V_{\pi k}(s) = V_{\pi k+1}(s) \quad \text{——} (2)$

From (1), (2) can be written as

$\lim_{N \to \infty} T^N_{\pi k+1} V_{\pi k}(s) \geq V_{\pi k}(s)$

$$\therefore \boxed{V_{\pi_{k+1}}(s) \geq V_{\pi_k}(s)} \quad \forall\, s \in S$$

If $V_{\pi_{k+1}}(s) = V_{\pi_k}(s)$.

$\therefore$ $T_{\pi_{k+1}}$ has a unique fixed point $V_{\pi_{k+1}}$

$$\therefore \boxed{V_{\pi_k} = V_{\pi_{k+1}}} \quad \forall\, s \in S.$$

15) $n = 2$

S_t	a_t	S_{t+1}	r_{t+1}	P	S_{t+2}	r_{t+2}	P
		① Iteration			2nd iteration		
Step 1	up	2 (step)	-1	0.5	F	4	0.5
step 1	down	G	1	0.5	-	-	-
Step 2	up	F	4	0.5	-	-	-
Step 2	down	1 (step)	2	0.5	G	2	0.5

① Iteration $\Rightarrow$ Reward $\rightarrow R_1$

$$(-1 \times 0.5 + 1 \times 0.5) + 1(4 \times 0.5)$$

$$= \underline{\underline{2}}.$$

② Iteration $\Rightarrow$ Reward R_2.

$$(4 \times 0.5 + 2 \times 0.5) + 1(2 \times 0.5)$$

$$= \underline{\underline{4}}.$$

So as we are increasing the iterations, or increasing the value of n, the iterations increase proportional.

If we are considering 'n' steps, there will be 'n' iterations.

The optimal policy is given by.

$$\text{step } 2 \longrightarrow \text{step } 1 \longrightarrow \text{Ground.}$$