RL A2

1) arms = 1, 2, ... 10

$$P(A=a) = \begin{cases} 2p, & a = 2,4,6,8,10 \\ p & a = 1,3,5,7,9 \end{cases} = \begin{cases} \frac{2}{15} & a = 2,4,6,8,10 \\ 1/15 & a = 1,3,5,7,9 \end{cases}$$

$5p + (2p)\times 5 = 1 \Rightarrow p = \frac{1}{15}$

Arm i $R_i \sim N(\mu = i, \sigma = 1)$

No of steps = 10

Expected sum reward $E[R^1 + R^2 + \ldots R^{10}]$

$R^j \rightarrow$ reward at time step j $j = 1, \ldots 10$

$E[R_i] = \mu = i$

$$E\left[\sum_{j=1}^{10} R^j\right] = \sum_{j=1}^{10} E[R^j]$$

$$E[R^j] = \sum_{x} x\, p(x/\text{arm}) \qquad r = E[R]$$

$$= \frac{1}{15}(1+3+5+7+9) + \frac{2}{15}(2+4+6+8+10) = 1.66 + 4$$

$$= 5.66$$

$$E\left[\sum_{j=1}^{10} R^j\right] = 10 \times 5.66 = 56.66$$

(2.) K=10 (no of arms)

arms - 1, 2, 4, 5, 7, 9, 10 reward = 0
probability 0.5

~~P(Reward=0/arm={1,2,4,5,7,9,10}~~

arm = 3, 6, 8 reward = 0 P = 0.3
R = 0.2 P = 0.3
R = 1 p = 0.4

$$P(r/a_1) = \begin{cases} 0.3 & r=0,\ a=3,6,8 \\ 0.3 & r=0.2,\ a=3,6,8 \\ 0.4 & r=1,\ a=3,6,8 \end{cases}$$

$a_1 = \{3, 6, 8\}$

$$P(r/a_2) = \begin{cases} 0.5 & r=0 \\ 0.5 & r=1 \end{cases}$$

$a_2 = \{1, 2, 4, 5, 7, 9, 10\}$

Maximise E[Reward]

$$\max_{\{a_1, a_2 \ldots a_m\}} E\left[\sum_{i=1}^{m} R_i\right]$$

E[R] for arm {3,6,8} = 0×0.3 + 0.2×0.3 + 0.4×1
= 0.06 + 0.4
= 0.46

E[R] for arm {1,2,4,5,7,9,10} = 0.5×0 + 0.5×1
= 0.5

Optimal stochastic policies

① $P(A=a) = \begin{cases} \frac{1}{7} & \text{arm} = 1,2,4,5,7,9,10 \\ 0 & \text{arm} = 3,6,8 \end{cases}$

② $P(A=a) = \begin{cases} \frac{1}{11} & a = 1,2,4 \\ \frac{2}{11} & a = 5,7,9,10 \\ 0 & a = 3,6,8 \end{cases}$

$3p + 4(2p) = 1$

$3p + 8p = 1$

$11p = 1$

③ $P(A=a) = \begin{cases} \frac{1}{11} & a = 1,5,7 \\ \frac{2}{11} & a = 2,4,9,10 \\ 0 & a = 3,6,8 \end{cases}$

④ $P(A=a) = \begin{cases} \frac{1}{11} & a = 9,10,2 \\ \frac{2}{11} & a = 1,5,7,4 \\ 0 & a = (3,6,8) \end{cases}$

⑤ $P(A=a) = \begin{cases} 3p = \frac{3}{17} & a = 9,10 \\ 4p = \frac{4}{17} & a = 5,7 \\ p = 1/17 & a = 2,4,10 \\ 0 & a = 3,6,8 \end{cases}$

$6p + 8p + 3p = 1$

$p = \frac{1}{17}$

(6) $P(A=a) = \begin{cases} 3/17 & a = 5, 7 \\ 4/17 & a = 9, 1 \\ 1/17 & a = 2, 4, 10 \\ 0 & a = 3, 6, 8 \end{cases}$

Q3.) ϵ-greedy explore only non greedy options

$k=3$ $R=\{0,1\}$

$Q_1(a)$ $a \in \{1,2,3\}$

$Q_1(1) = 1$
$Q_1(2) = 4$
$Q_1(3) = 9$

$Q_t(a), A_t, R_t$ for $t = 1,2,3,4,5,6$

explore - odd time
$t = 1, 3, 5$
exploit - even time
$t = 2, 4, 6$

Example seq:

$t=1$ (explore)
$A_1 = 1, R_1 = 0$
$Q_2(1) = 0$
$Q_2(2) = 4$
$Q_2(3) = 9$

$t=2$ (exploit)
$A_2 = 3$ $R_2 = 0$
$Q_3(1) = 0$
$Q_3(2) = 4$
$Q_3(3) = 0$

$t=3$ (explore)
$A_3 = 1$, $R_3 = 1$
$Q_4(1) = 1$
$Q_4(2) = 4$
$Q_4(3) = 0$

$t=4$ (exploit)
$A_4 = 2$, $R_4 = 0$
$Q_5(1) = 1$
$Q_5(2) = 0$
$Q_5(3) = 0$

$t=5$ (explore)
$A_5 = (3)$ $R_5 = 1$
$Q_6(1) = 1$
$Q_6(2) = 0$
$Q_6(3) = 1$

$t=6$ (exploit)
$A_6 = 1$ $R_6 = 1$
$Q_7(1) = 1$
$Q_7(2) = 0$
$Q_7(3) = 1$

(6) $P(A=a) = \begin{cases} 3/17 & a = 5, 7 \\ 4/17 & a = 9, 1 \\ 1/17 & a = 2, 4, 10 \\ 0 & a = 3, 6, 8 \end{cases}$

Q3.) ϵ-greedy explore only non greedy options

$k=3$ $R=\{0,1\}$

$Q_1(a)$ $a \in \{1,2,3\}$

$Q_1(1) = 1$
$Q_1(2) = 4$
$Q_1(3) = 9$

$Q_t(a), A_t, R_t$ for $t = 1, 2, 3, 4, 5, 6$

explore - odd time
$t = 1, 3, 5$

exploit - even time
$t = 2, 4, 6$

Example seq:

$t=1$ (explore)
$A_1 = 1, R_1 = 0$
$Q_2(1) = 0$
$Q_2(2) = 4$
$Q_2(3) = 9$

$t=2$ (exploit)
$A_2 = 3$ $R_2 = 0$
$Q_3(1) = 0$
$Q_3(2) = 4$
$Q_3(3) = 0$

$t=3$ (explore)
$A_3 = 1$, $R_3 = 1$
$Q_4(1) = 1$
$Q_4(2) = 4$
$Q_4(3) = 0$

$t=4$ (exploit)
$A_4 = 2$, $R_4 = 0$
$Q_5(1) = 1$
$Q_5(2) = 0$
$Q_5(3) = 0$

$t=5$ (explore)
$A_5 = (3)$ $R_5 = 1$
$Q_6(1) = 1$
$Q_6(2) = 0$
$Q_6(3) = 1$

$t=6$ (exploit)
$A_6 = 1$ $R_6 = 1$
$Q_7(1) = 1$
$Q_7(2) = 0$
$Q_7(3) = 1$

(6) $P(A=a) = \begin{cases} 3/17 & a = 5, 7 \\ 4/17 & a = 9, 1 \\ 1/17 & a = 2, 4, 10 \\ 0 & a = 3, 6, 8 \end{cases}$

(3.) ϵ-greedy explore only non greedy options

$k=3$ $R=\{0,1\}$

$Q_1(a)$ $a \in \{1,2,3\}$

$Q_1(1) = 1$
$Q_1(2) = 4$
$Q_1(3) = 9$

$Q_t(a), A_t, R_t$ for $t = 1, 2, 3, 4, 5, 6$

explore - odd time
$t = 1, 3, 5$

exploit - even time
$t = 2, 4, 6$

Example seq:

$t=1$ (explore)
$A_1 = 1, R_1 = 0$
$Q_2(1) = 0$
$Q_2(2) = 4$
$Q_2(3) = 9$

$t=2$ (exploit)
$A_2 = 3$ $R_2 = 0$
$Q_3(1) = 0$
$Q_3(2) = 4$
$Q_3(3) = 0$

$t=3$ (explore)
$A_3 = 1$, $R_3 = 1$
$Q_4(1) = 1$
$Q_4(2) = 4$
$Q_4(3) = 0$

$t=4$ (exploit)
$A_4 = 2$, $R_4 = 0$
$Q_5(1) = 1$
$Q_5(2) = 0$
$Q_5(3) = 0$

$t=5$ (explore)
$A_5 = (3)$ $R_5 = 1$
$Q_6(1) = 1$
$Q_6(2) = 0$
$Q_6(3) = 1$

$t=6$ (exploit)
$A_6 = 1$ $R_6 = 1$
$Q_7(1) = 1$
$Q_7(2) = 0$
$Q_7(3) = 0$

Q4) episodic → repeated attempts to balance the pole
reward − +1 failure did not occur
+ discounting
−1 upon failure
~~continuous~~
return at each time

Q4) 3.4 $p(s', r \mid s, a)$ s, a, s', r & $p(s', r/s, a)$
and a row for every 4-tuple for which
$p(s', r/s, a) > 0$

Since the rewards do not have a probability distribution, they are constant depending on the state

$\therefore p(s', r/s, a) = p(s'/s, a)$
hence table will be same.

Q6) ~~3.15~~ 3.14
Bellman eq :

$$V_\pi(s) = E_\pi[G_t / S_t = s]$$

$$= \sum_a \pi(a/s) \sum_{s', r} p(s', r/s, a)\,[r + \gamma V_\pi(s')]$$

$\forall s \in S$

Centre state of +0.7
2.3 0.7
0.4
−0.4

~~for s' 2.3 Action ↑~~

~~$V_\pi(s) =$~~ $\sum_{s',r} p(s', r/s, a)[r + \gamma V_\pi(s')$

Calculating for all actions

Assuming $P(A = a) = \begin{cases} \frac{1}{4} \end{cases}$

1 state s' corresponding to 1 action

reward -1 off the grid
or 0 others
from A → reward +10
B → +5

Given: $\gamma = 0.9$

To find $V_\pi(s)$ of centre state

$a = \uparrow$ $\frac{1}{4}(0 + 0.9(2.3))$

+

$a = \downarrow$ $\frac{1}{4}(0 + 0.9 \times (-0.4)) +$

+

$a = \leftarrow$ $\frac{1}{4}(0 + 0.9 \times 0.7)$

+

$a = \rightarrow$ $\frac{1}{4}(0 + 0.9 \times 0.4)$

$0.9 \times 0.25 [2.3 - 0.4 + 0.7 + 0.4)$

$= 0.9 \times 0.25 \times 3.0$

$= 0.675 \approx 0.7$

3.15 reward +ve goal
-ve running into edge
0 rest

→ adding constant c to reward
V_c to all value of state
does not affect the relative value of any state

→ V_c in terms of c & γ

3.8 $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$

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

$G_t' = (R_{t+1} + c) + \gamma(R_{t+2} + c) + \gamma^2(R_{t+3} + c) + \ldots$

(adding constant c to all reward)

$G_t' = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \ldots +$
$c + \gamma c + \gamma^2 c + \ldots$

c - constant ⇒ $c + \gamma c + \gamma^2 c \ldots$ is a GP
γ - constant

Sum of GP (infinite) $= \frac{a}{1-r} = \frac{c}{1-\gamma}$ $\quad a = c,\ r = \gamma$

$\frac{c}{1-\gamma}$ — constant

$$V_c = \frac{c}{1-\gamma} \qquad \gamma < 1$$

3.16

For episodic task

$$G_t = R_{t+1} + R_{t+2} + \dots R_T$$

$$G_t' = (R_{t+1} + c) + (R_{t+2} + c) \dots + (R_T + c)$$

adding constant c to all rewards

$$G_t' = R_{t+1} + R_{t+2} + \dots R_T + [T - (t+1)]\, c$$

There is a constant added for the time interval of the episodic task

Since now the value depends on the time interval, the task is changed as the return is incentivized to add additional delay in completing the task.

Q8.) equation for V_* in terms of q_*

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

$$V_*(s) = \max_{a \in A} \sum_{s',r} p(s', r / s, a) \left[r + \gamma V_*(s') \right]$$

Q11) MDP → PMF

R_{t+1} depends on state S_t & Action A_t

Is R_{t+2} dependent on S_t & A_t. Using PMFs used to define MDP. → No

Conditional prob of R_{t+2}

According to MDP,

$$p(s', r / s, a) = Pr\{S_{t+1} = s', R_{t+1} = r / S_t = s, A_t = a\}$$

$$P(R_{t+1} / S_t, A_t) = P(R_{t+1} / S_t = s, A_t = a, S_{t+1})$$

For R_{t+2}

$$p(s', r / s, a) = Pr\{S_{t+2} = s', R_{t+2} = r / S_{t+1} = s, A_{t+1} = a\}$$

In case on MDP, the probability of each possible values for S_t & R_t depends only on the preceding state & action. ∴ R_{t+2} depends only on S_{t+1} & A_{t+1} & and not S_t & A_t.

Q12) $$E[R_{t+2} / S_t = s, A_t = a] = \sum_{r \in R_{t+2}} r \, p(r/s, a)$$

$$P(R_{t+1} / S_t, A_t) = P(R_{t+1} / S_t = s, A_t = a, S_{t+1})$$

$$\hookrightarrow \sum_{s' \in S_{t+1}} P(r', s' / s, a) \quad \not\in S_{t+1}$$

$$P(R_{t+2} / S_t, A_t) =$$

$$P(R_{t+2} / S_{t+1} = s, A_t = a, S_{t+2} = s')$$

$$\sum_{s'' \in S_{t+2}} P(R_{t+2} = r, S_{t+2} = s'' / S_{t+1} = s', A_t = a)$$

$$\sum_{s'' \in S_{t+2}} p(r, s''/s, a) \times p(S_{t+1}=s'/S_t=s, A_t=a) \; P(A_{t+1}=a/s) \; \pi(a/s)$$

$$\sum_{s'' \in S_{t+2}} P(R_{t+2}=r, S_{t+2}=s'/S_{t+1}=s, A_{t+1}=a)$$

PMF R_{t+2} | $S_t=s$, $A_t=a$

$$\sum_{\substack{s'' \in S_{t+2}, \\ r' \in R_{t+1}, \\ s' \in S_{t+1} \\ a' \in A_{t+1}}} P(R_{t+2}=r', S_{t+2}=s''/S_{t+1}=s', A_{t+1}=a') \; P(R_{t+1}=r', S_{t+1}=s'/S_t=s, A_t=a) \; \pi(a'/s')$$

$$E[R_{t+2}/A_t=a, S_t=s] = \sum_{r'' \in R_{t+2}} r'' \, P[R_{t+2}=r''/S_t=s, A_t=a]$$

$$= \sum_{\substack{r'' \in R_{t+2} \\ s'' \in S_{t+2} \\ r' \in R_{t+1} \\ s' \in S_{t+1} \\ a' \in A_{t+1}}} r'' \; p(r', s''/s', a') \; p(r, s'/s, a) \; \pi(a'/s')$$

Q13.) $V_\pi(s) = E[G_t/S_t=s]$

Bellman eq for $V_\pi(s)$ $\forall$ $s \in S$

↳ in terms of MDP

policy π

PMF $p(s', r/s, a)$

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

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$$

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

$$V_\pi(s) = E_\pi\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \mid S_t = s\right]$$

$$= E_\pi\left[R_{t+1} \mid S_t = s\right] + \gamma E\left[R_{t+2} + \gamma R_{t+3} + \ldots \mid S_t = s\right]$$

$$\left[V_\pi(S_{t+1}) = E_\pi\left[R_{t+2} + \gamma R_{t+3} + \ldots \mid S_{t+1} = s'\right]\right]$$

$$V_\pi(s) = E_\pi\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s\right]$$

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

$$= \sum_{r,s'} \left(r + \gamma V_\pi(s')\right) \sum_a P\left[R_{t+1} = r, S_{t+1} = s' \mid S_t = s, A_t = a\right] \pi(a/s)$$

$$V_\pi(s) = \sum_{a,r,s'} \left(r + \gamma V_\pi(s')\right) p(r, s' \mid s, a)\, \pi(a/s)$$

Q14) $R_1 = 2$, $R_2 = -1$, $R_3 = 10$, $R_4 = -3$

$\gamma = 0.5$ $\quad G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4}$

At time step 0 : $[t = 0]$

$$G_t = 2 + 0.5(-1) + (0.5)^2 10 + (0.5)^3 (-3)$$

$$(t=0) = 1.5 + 2.5 - 0.375$$

At $t=1$ $G_1 = ?$

At $t=2$ $G_2 = R_2 + \gamma G_1$

At time $t=0$

$$G_{t=0} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4}$$
$$= 3.625$$

At time $t=1$ $\quad \gamma(G_{t=2})$

$$G_{t=1} = R_2 + \gamma R_3 + \gamma^2 R_4 = -1 + 0.5(8.5)$$
$$= 3.25$$

$t=2$

$$G_{t=2} = R_3 + \gamma R_4 = 10 + 0.5(-3)$$
$$= 8.5$$

$t=3$

$$G_{t=3} = R_4 = -3$$

$$G_t' = (R_{t+1} + C) + \gamma(R_{t+2} + C) + \gamma^2(R_{t+3} + C) + \ldots$$
$$= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$$
$$+ \underbrace{C + \gamma C + \gamma^2 C + \ldots}$$

$C + \gamma C + \gamma^2 C + \ldots \rightarrow$ GP Sum of GP infinite $= \frac{C}{1-\gamma}$

Infinite horizon discounted return

$$G_t' = G_t + \left[\frac{C}{1-\gamma}\right]$$

Q15) $V_*(s) \quad \forall \; s \in S$

Optimal policy can be found using policy iteration

When $V_{K+1}(s) = V_K(s)$ and policy is optimal at this condition

However since we already have $V_*(s)$

$$\pi_*(s) = \underset{a \in A(s)}{\operatorname{argmax}} \underbrace{E\left[R_{t+1} + \gamma V_*(S_{t+1}) \,\middle|\, S_t = s, A_t = a\right]}_{\forall s}$$

$$q_*(s, a)$$

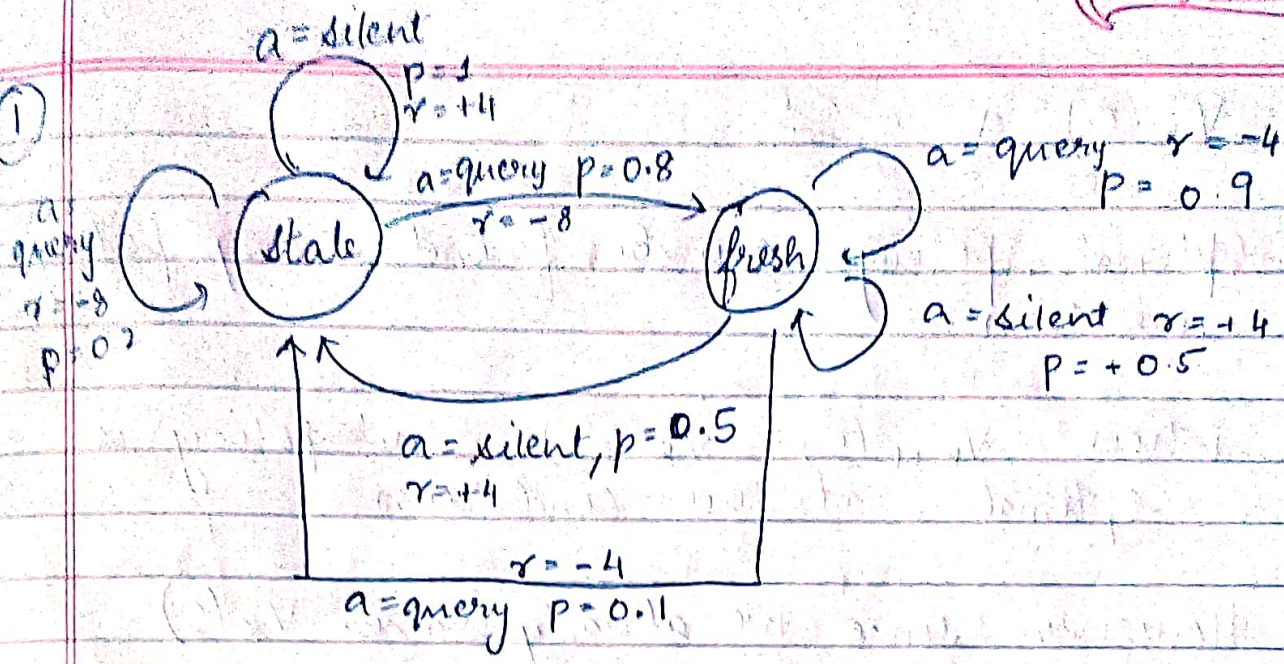
Q16) States: fresh or stale
Action: query or remain silent

$P(\text{fresh} / \text{stale}) = 0.8$ } given
$P(\text{stale} / \text{fresh}) = 0.1$ } query

$P(\text{state} / \text{stale}) = 1$ & $P(\text{stale} / \text{fresh}) = 0.5$ given silent

Query when current is stale = -8
" " " is fresh = -4
Silent reward = $+4$

(1)



s	a	s'	r	p(s', r/s, a)
fresh	silent	stale	+4	0.5
fresh	query	stale	-4	0.1
fresh	silent	fresh	+4	0.5
fresh	query	fresh	-4	0.9
stale	silent	stale	+4	1
stale	silent	fresh	-	- [NP]
stale	query	stale	-8	0.2
stale	query	fresh	-8	0.8

For $t = 1$

$$\pi(\text{fresh}) = \max \begin{cases} q = -4 + 0.5(V_2(\text{fresh})0.9 + V_2(\text{stale})0.1 \\ s = +4 + 0.5(V_2(\text{fresh})0.5 + V_2(\text{stale})0.5) \end{cases}$$

Action = silent

$$\pi(\text{silent}) = \max \begin{cases} -8 + 0.5(V_2(\text{fresh})0.8 + V_2(\text{Stale})0.2 = -6.5 \\ s \to 4 + 0.5(V_2(\text{fresh})0. + V_2(\text{stale}).1 = 3.5 \end{cases}$$

a = silent

$t = 0$

$$\pi(\text{fresh}) = \max \begin{cases} q \to -4 + 0.5(V_1(\text{fresh})0.9 + V_2(\text{stale})0.1 = -1.68 \\ s \to +4 + 0.5(V_1(\text{fresh})0.5 + V_1(\text{stale})0.5) = 6.0 \end{cases}$$

$\Rightarrow$ silent

$$\pi(\text{stale}) = \max \begin{cases} -8 + 0.5(V_1(\text{fresh})0.8 + V_1(\text{stale})0.2 = -5.75 \\ s \to 4 + 0.5(V_1(\text{fresh}))0 + V_1(\text{stale})1] - 5.7 \end{cases}$$

Silent

Q17) $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$ → Selecting action

TP: policy improvement either improves current policy or current policy is optimal

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

π – original
π' – changed

TP:
$$\begin{cases} V_{\pi_{k+1}}(s) \geq V_{\pi_k}(s) & \forall s \\ \text{When } V_{\pi_k}(s) = V_{\pi_k}(s) \Rightarrow \pi_k = \text{optimal policy} & \forall s \end{cases}$$

$\therefore V_{\pi'}(s) \geq V_\pi(s) \ \forall s$ or $q_\pi(s, \underbrace{\pi'(s)}_{a}) \geq V_\pi(s)$

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

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

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

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

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

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

$\leq V_{\pi'}(s)$: Policy improvement for all states $\forall s \in S$ have value $V_{\pi'}(s) \geq V_\pi(s)$

In case $V_\pi = V_{\pi'} \quad \forall s \in S$

$$V_{\pi'}(s) = \max_a E[R_{t+1} + \gamma V_{\pi'}(S_{t+1}) / S_t = s, A_t = a]$$

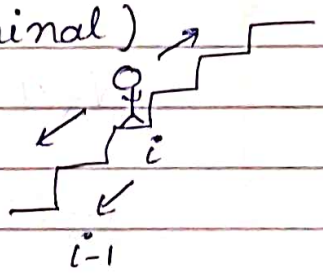
$$= \max_a \sum_{s', r} p(s', r / s, a) [r + \gamma V_{\pi'}(s)]$$

This same as Bellman optimality

$\therefore V_{\pi'}$ must be V_*

$\therefore$ both π & π' are optimal

Q18.) GF F - States (terminal)



n steps : 1, 2 ... n → F - states

G

reward 2 $p = 0.5$
0 $p = 1 - 0.5 = 0.5$
} i to i-1 ↓

$1 < i < n$

reward 1 → G reward = 1

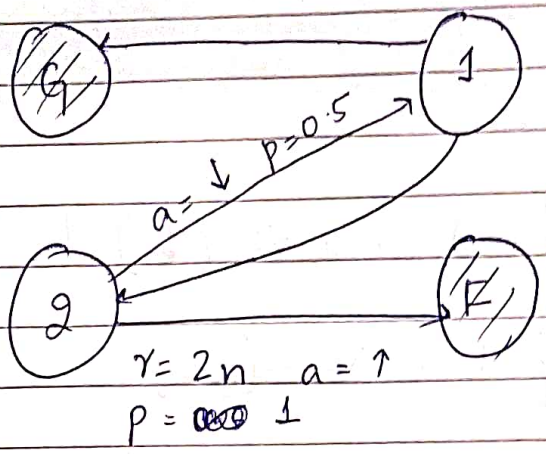
reward -1 ↕ i to i+1

but 2n n → F

PMF $P(A = a) = \begin{cases} 0.5 & a = \uparrow \\ 0.5 & a = \downarrow \end{cases}$

MDP for $n = 2$ $\gamma = 1$

States : G 1 2 F



s	a	s'	r	$p(s', r \mid a, s)$
1	↓	G	1	1
1	↑	2	-1	1
2	↑	F	$2n = 4$	1
2	↓	1	2	0.5
2	↓	1	0	0.5

Using current policy

$$\pi(a \mid s) = \begin{cases} 0.5 & a = \uparrow \\ 0.5 & a = \downarrow \end{cases}$$

Iteration 1 : Policy evaluation

$$V_\pi(s) = \sum_a \pi(a/s) \sum_{s',r} p(s', r/s, a)\left[r + \gamma V_\pi(s')\right]$$

$V_\pi(G) = 0$

$V_\pi(F) = 0$

$V_\pi(1) =$ ~~0.8~~

$a = \uparrow \rightarrow 0.5 \times 1\left[-1 + 1\, V_\pi(2)\right]$

$+$

$a = \downarrow \rightarrow 0.5 \times 1\left[1 + V_\pi(G)\right]$

$V_\pi(1) = 0.5\, V_\pi(2) \quad -(1)$

$V_\pi(2) =$

$a = \uparrow \rightarrow 0.5 \times \left[1 \times 4\right] = 2$

$+$

$a = \downarrow \rightarrow 0.5\left[0.5\left[2 + V_\pi(1)\right] + 0.5\left[0 + V_\pi(1)\right]\right]$

$V_\pi(2) = 2.5 + 0.5\, V_\pi(1) \quad -(2)$

Solving eq 1 & 2, $V_\pi(1) = 1.66$

$V_\pi(2) = 3.34$

Iteration 1: policy ~~iteration~~ improvement

$$\pi(a/s) = \underset{a \in A(s)}{\text{argmax}} \; q_\pi(s,a)$$

$$\underset{a \in A(s)}{\text{argmax}} \; E[R_{t+1} + \gamma V(S_{t+1}) / S_t = s, A_t = a]$$

State $s=1$, $a = \uparrow$ $\rightarrow$ $-1 + 1 \cdot V_\pi(2)$

$\rightarrow$ 2.334

$s=1$, $a = \downarrow$ $\rightarrow$ $1 + 1 \cdot V_\pi(G) = 1$

for $\pi(a/s=1) \rightarrow \uparrow$

State $s=2$, $a = \uparrow$ $\rightarrow$ $+4 + V_\pi(F) = 4$

$a = \downarrow$ $\rightarrow$ $0.5(2 + V_\pi(1)) + 0.5(0 + V_\pi(1))$

$= 2.667$

$\pi(a/s=2) = \uparrow$

Iteration 2: policy evaluation

$V_\pi(1)$ $a = \uparrow$

$= 1 \times 1 [-1 + V_\pi(2)]$

$= -1 + V_\pi(2)$ $V_\pi(1) = -1 + 4$

$= 3$

$V_\pi(2)$ $a = \uparrow$ $= 4$

policy improvement

$\pi(a/s)$ $s = 1$ $a = \uparrow$ $= -1 + V_\pi(2) = 3$

$s = 1$ $a = \downarrow$ $= 1$

$\pi(a/s=1) = \uparrow$

for $S=2$ $a=\uparrow$: $4 + V_\pi(F) = 4$

$a=\downarrow$ $[2 + V_\pi(1)]0.5 + 0.5[0 + V_\pi(1)]$

$= 4$

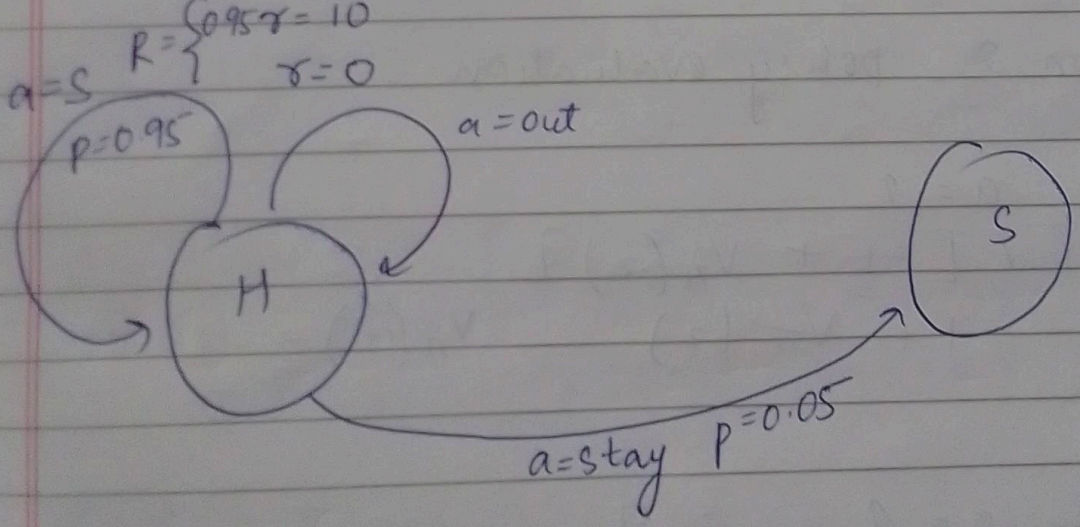
$\pi(a/s=2) = \uparrow$

Since the policy after iteration 1 & iteration are same, we have found the optimal policy

The no of iteration will increase with $O(n)$

19) State: healthy or sick

Action: Stay home or go out → Healthy
med or no med → sick



for $S=2$ $a=\uparrow$: $4 + V_\pi(F) = 4$

$a=\downarrow$ $[2 + V_\pi(1)]0.5 + 0.5[0 + V_\pi(1)]$

$= 4$

$\pi(a/S=2) = \uparrow$

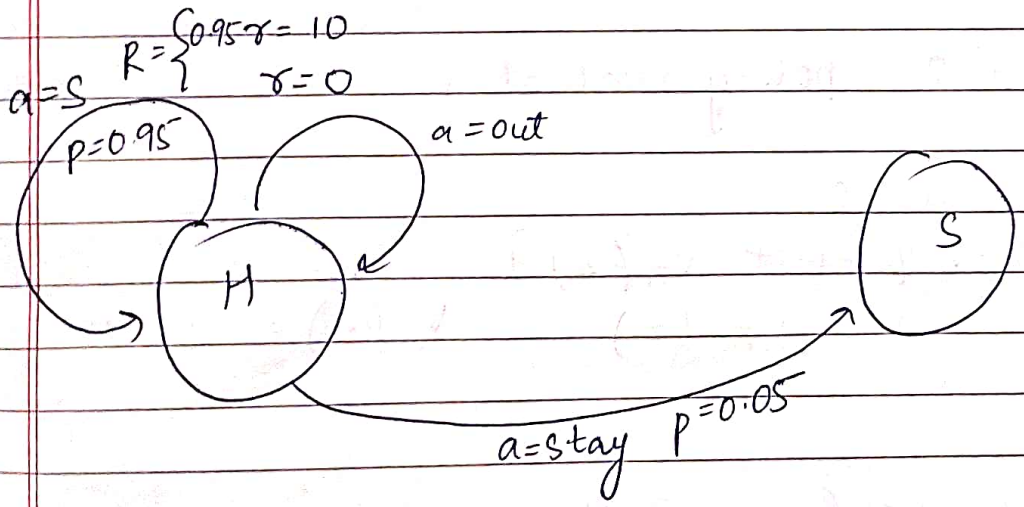
Since the policy after iteration 1 & iteration are same, we have found the policy optimal

The no of iteration will increase with $O(n)$

9.19) State: healthy or sick

Action: Stay home or go out → Healthy

med or no med → sick



s	a	s'	r	p(s', r / s, a)
H	stay	H	10	0.95 × 0.95 = 0.9025
H	stay	H	0	0.95 × 0.05 = 0.0475
H	stay	S	-10	0.05 × 0.9 = 0.045
H	stay	S	0	0.05 × 0.1 = 0.005
H	out	H	20	0.7 × 0.9 = 0.63
H	out	H	0	0.7 × 0.1 = 0.07
H	out	S	-10	0.3 × 0.9 = 0.27
H	out	S	0	0.3 × 0.1 = 0.03
S	med	S	-2	0.1
S	med	H	-1	0.9
S	no med	S	-1	0.4
S	no med	H	0	0.6

$\gamma = 0.9$

Q20) $\pi(a/s)$ - equal prob.

$V_\pi(s)\ \forall\ s$

$$V_\pi(s) = \sum_a \pi(a/s) \sum_{s'r} p(s', r/s, a)\left[r + \gamma V_\pi(s')\right]$$

State → H

$V_\pi(H)$

a = stay → $0.5\big[0.9025(10 + 0.9V_\pi(H)) +$
$0.0475(0 + 0.9V_\pi(H)) +$
$0.045(-10 + 0.9V_\pi(S)) +$
$0.005(0 + 0.9V_\pi(S))\big]$

+

a = out $0.5\big[0.63(20 + 0.9V_\pi(H)) +$
$0.07(0 + 0.9V_\pi(H)) +$
$0.27(-10 + 0.9V_\pi(S)) +$
$0.03(0 + 0.9V_\pi(S))\big]$

$V_\pi(H) = 0.7425\,V_\pi(H) + 0.15775\,V_\pi(S) +$
$9.2375 \quad -(1)$

State → S

$V_\pi(S)$

a = med → $0.5\big[0.1(-2 + 0.9V_\pi(S)) +$
$0.9(-1 + 0.9V_\pi(H))\big]$

+

a = no med → $0.5\big[0.6(0 + 0.9V_\pi(H)) +$
$0.4(-1 + 0.9V_\pi(S))$

$$V_\pi(s) = 0.675\, V_\pi(H) + 0.225\, V_\pi(s) - 0.75 \quad -(2)$$

Solving eq 1 & 2

$V_\pi(\text{healthy}) = 75.641$

$V_\pi(\text{sick}) = 64.913$

Q21.) t H $\quad$ $t+1$ S $\quad$ $\pi \to 0.5$ for all actions

$$E[G_t \mid S_t = \text{healthy}, S_{t+1} = \text{sick}]$$

$$G_t = R_{t+1} + \gamma V_\pi(S_{t+1})$$

$$\hookrightarrow E[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = \text{healthy}, S_{t+1} = \text{sick}]$$

~~$\sum_{a,r} r + \gamma V_\pi(S_{t+1})\, p(r \mid s, s', a)\, \pi(a \mid s)$~~

$$= \sum_{a,r} (r + \gamma V_\pi(S_{t+1}))\cdot p(r \mid s, s', a)\, \pi(a \mid s)$$

~~action~~ ~~reward~~

$a = \text{stay}$ $\quad 0.5\,[0.9\,(-10 + 0.9 \times 64.913) + 0.1\,(0 + 0.9 \times 64.913)]$

$+$

$a = \text{go out}$ $\quad 0.5\,[0.9\,(-10 + 0.9 \times 64.913) + 0.1\,(0 + 0.9 \times 64.913)]$

$= 2 \times 25 \left[43.57953 + 5.84217 \right]$

$= 49.4217$