Homework 1

Submitted by:

Sayan Chakraborty

Roll No: EE18MTECH11030

Subject: Introduction to Modern AI

Ans 1:

Derive the Bellman's equation for:

(i) State-value function $V_\pi(s)$.

We know,

$$v_\pi(s) = E_\pi\left[R_t \mid S_t = s\right], \quad R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid S_t = s\right] \quad \text{— (1)}$$

We use the law of total expectation to derive the Bellman's equation.

The law of total expectation is given as,

$$E[X] = \sum_i E[X \mid A_i]\, \mathbb{P}(A_i),$$ $\{A_i\}_i$ is a countable partition of the sample space.

Now,

conditioning (1) over each chosen action at time step t, i.e, a_t

$$v_\pi(s) = \sum_a E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid S_t = s, a_t = a\right] \mathbb{P}(a_t = a \mid S_t = s) \quad \text{— (2)}$$

Now, $\mathbb{P}(a_t = a \mid S_t = s)$ is the policy $\pi(s,a)$

(2) $\Rightarrow$

$$v_\pi(s) = \sum_a \pi(s,a)\, E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid S_t = s, a_t = a\right] \quad \text{— (3)}$$

Now, conditiong (3) on the future state s'

we get,

(3) $\Rightarrow v_\pi(s) = \sum_a \pi(s,a) \sum_{s'} E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\middle|\, s_t = s, a_t = a, s_{t+1} = s'\right] \cdot$
$$\mathbb{P}[s_{t+1} = s' \mid s_t = s, a_t = a]$$

denote, $\mathbb{P}_{(s,a)}^{s'} = \mathbb{P}[s_{t+1} = s' \mid s_t = s, a_t = a]$

$$\Rightarrow v_\pi(s) = \sum_a \pi(s,a) \sum_{s'} E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\middle|\, s_t = s, a_t = a, s_{t+1} = s'\right] \mathbb{P}_{(s,a)}^{s'}$$

$$\Rightarrow v_\pi(s) = \sum_a \pi(s,a) \sum_{s'} \mathbb{P}_{(s,a)}^{s'} E_\pi[r_{t+1} \mid s, a, s']$$
$$+ \sum_a \pi(s,a) \sum_{s'} \mathbb{P}_{(s,a)}^{s'} E_\pi\left[\sum_{k=0}^{\infty} \gamma^{k+1} r_{t+k+2} \,\middle|\, s, a, s'\right]$$

denote, $\mathbb{R}_{s,a,s'}^{r_{t+1}} = E_\pi[r_{t+1} \mid s, a, s']$

Also by Markov property we have that

$$E_\pi\left[\sum_{k=0}^{\infty} \gamma^{k+1} r_{t+k+2} \,\middle|\, s, a, s'\right] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^{k+1} r_{t+k+2} \,\middle|\, s'\right]$$

$\therefore$ we have,

$$\Rightarrow v_\pi(s) = \sum_a \pi(s,a) \sum_{s'} \mathbb{P}_{(s,a)}^{s'} \mathbb{R}_{s,a,s'}^{r_{t+1}}$$
$$+ \gamma \sum_a \pi(s,a) \sum_{s'} \mathbb{P}_{(s,a)}^{s'} E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \,\middle|\, s'\right]$$

$$\Rightarrow v_\pi(s) = \sum_a \pi(s,a) \sum_{s'} \mathbb{P}_{(s,a)}^{s'} \mathbb{R}_{(s,a,s')}^{r_{t+1}}$$
$$+ \gamma \sum_a \pi(s,a) \sum_{s'} \mathbb{P}_{(s,a)}^{s'} v_\pi(s')$$

$$\Rightarrow \boxed{v_\pi(s) = \sum_a \pi(s,a) \sum_{s'} \mathbb{P}_{(s,a)}^{s'} \left[\mathbb{R}_{(s,a,s')}^{r_{t+1}} + \gamma v_\pi(s')\right]}$$

(ii) Action-value function:

We know,

$$q_\pi(s,a) = E_\pi[R_t \mid s_t = s, a_t = a], \quad R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\Big|\, s, a\right] \quad (4)$$

Now, conditioning (4) on the future state s', we obtain,

$$\Rightarrow q_\pi(s,a) = \sum_{s'} E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\Big|\, s,a,s'\right] \mathbb{P}(s' \mid s,a)$$

denote, $\mathbb{P}^{s'}_{(s,a)} = \mathbb{P}(s' \mid s,a)$

Then,

$$\Rightarrow q_\pi(s,a) = \sum_{s'} E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\Big|\, s,a,s'\right] \mathbb{P}^{s'}_{(s,a)}$$

$$= \sum_{s'} \mathbb{P}^{s'}_{(s,a)} E_\pi[r_{t+1} \mid s,a,s'] +$$

$$\gamma \sum_{s'} \mathbb{P}^{s'}_{(s,a)} E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \,\Big|\, s,a,s'\right]$$

denote, $\mathbb{R}^{r_{t+1}}_{(s,a,s')} = E_\pi[r_{t+1} \mid s,a,s']$

and observe that, by Markov property,

$$E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \,\Big|\, s,a,s'\right] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \,\Big|\, s'\right]$$

$$\Rightarrow q_\pi(s,a) = \sum_{s'} \mathbb{P}^{s'}_{(s,a)} \mathbb{R}^{r_{t+1}}_{(s,a,s')} + \gamma \sum_{s'} \mathbb{P}^{s'}_{(s,a)} E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \,\Big|\, s'\right] \quad (5)$$

Now,

conditioning (5) on a' (possible actions when on state s'),

we have,

$$(5) \Rightarrow q_\pi(s,a) = \sum_{s'} \mathbb{P}^{s'}_{(s,a)} \mathbb{R}^{t+1}_{(s,a,s')} + \gamma \sum_{s'} \mathbb{P}^{s'}_{(s,a)} E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \,\middle|\, s'\right]$$

$$= \sum_{s'} \mathbb{P}^{s'}_{(s,a)} \mathbb{R}^{t+1}_{(s,a,s')} + \gamma \sum_{s'} \mathbb{P}^{s'}_{(s,a)} \sum_{a'} E_\pi\left[R_{t+1} \mid s', a'\right] \mathbb{P}(a' \mid s')$$

$$= \sum_{s'} \mathbb{P}^{s'}_{(s,a)} \mathbb{R}^{t+1}_{(s,a,s')} + \gamma \sum_{s'} \mathbb{P}^{s'}_{(s,a)} \sum_{a'} q_\pi(s',a')\, \pi(s',a')$$

$$\boxed{q_\pi(s,a) = \sum_{s'} \mathbb{P}'_{(s,a)} \left[\mathbb{R}^{t+1}_{(s,a,s')} + \gamma \sum_{a'} q_\pi(s',a')\, \pi(s',a')\right]}$$

Ans: 2

(a)

(i) State transition table for example 1

Description: A mobile robot has the job of picking cans and placing them on bin. The bot runs on battery. The decision on how to search for a can is made by reinforcement learning agent based on the bot's battery. The charge level of the bot's battery. The agent should decide whether the bot should (i) actively search for cans, (ii) wait for someone to bring the cans to it (iii) recharge the battery.

The states are:

$S = \{low, high\}$ (charge levels of bot's battery)

The actions corresponding to states are:

$A(low) = \{search, wait, recharge\}$

$A(high) = \{search, wait\}$

The corresponding rewards are assumed to be:

r_{search} : average reward during search.

r_{wait} : average reward during wait.

-3 : when recharge after running out of battery.

0 : during recharge

State transition table:

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

(ii) State transition table for example 2

The states are:

$$\mathcal{S} = \{ G, B \}, \quad \begin{array}{l} G = \text{good state} \\ B = \text{bad state} \end{array}$$

The actions corresponding to states are:

$$\mathcal{A}(G) = \{\text{stay}, \text{move}\}$$

$$\mathcal{A}(B) = \{\text{stay}, \text{move}\}$$

The corresponding rewards are assumed to be:

$r_{GSG} = 3$ (for staying in state G)

$r_{GSB} = -1$ (for going to state B but attempting to stay at state G)

$r_{BSB} = -1$ (for staying in state B)

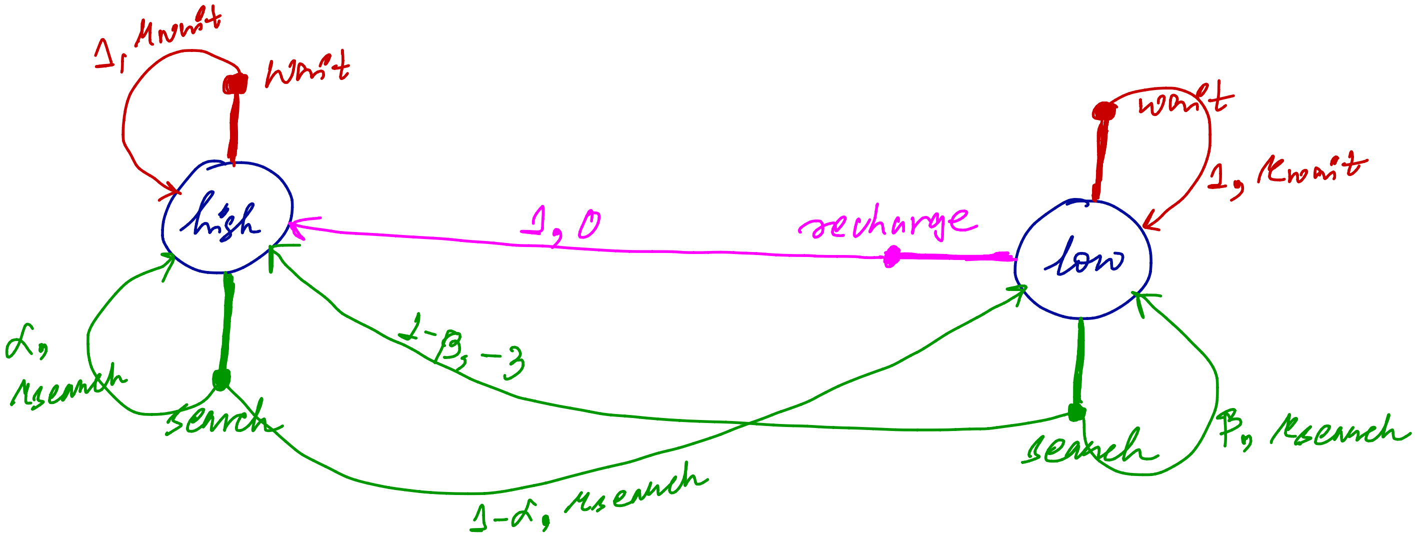
$r_{BMG} = 3$ (for moving to state G from state B)

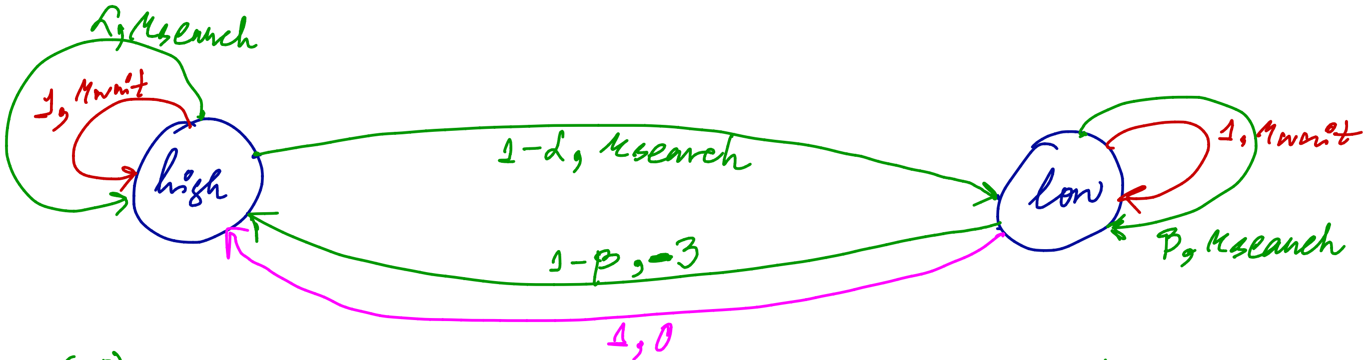
$r_{GMB} = -1$ (for moving to state B from state G)

State transition table:

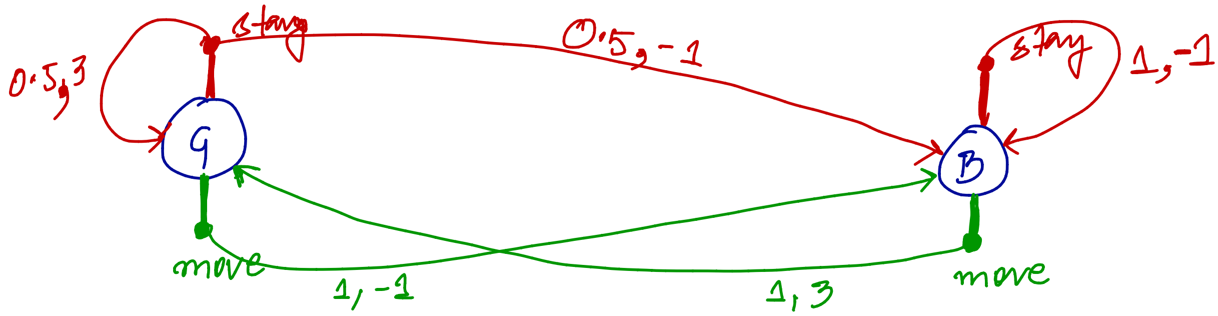
s	s'	a	$p(s' \mid s, a)$	$r(a, s, s')$
G	G	stay	0.5	3
G	B	stay	0.5	-1
G	G	move	0	0
G	B	move	1	-1
B	B	stay	1	-1
B	G	stay	0	0
B	B	move	0	0
B	G	move	1	3

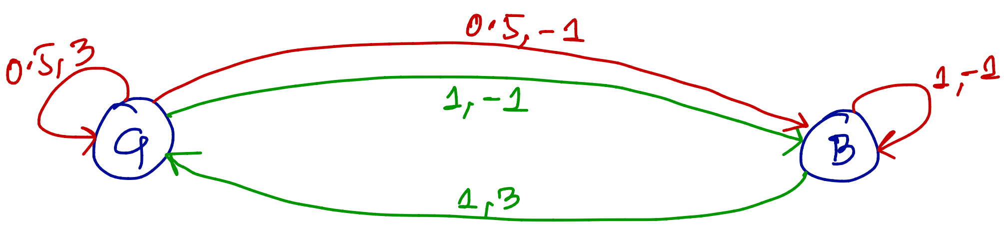
(6) (i) State-space diagram for example 1





(ii) State-space diagram for example 2





(c) After implementing the value iteration algorithm for the can collecting robot example, the following results are obtained:

Considering $\Delta = 100$, $\varepsilon = 1\times10^{-9}$,

optimal values

$v^*(S=\text{High}) = 3.17647058$7

$v^*(S=\text{Low}) = 1.9999999$

Iteration needed = 31

policy

Policy: State = High, Action = search

Policy: State = Low, Action = wait

Python file: EE18MTECH11030_Question_2_C_final.ipynb

(d) After implementing the value iteration algorithm for the second MDP example, the following results are obtained:

Considering $\Delta = 100$, $\varepsilon = 5\times10^{-9}$,

optimal values

$v^*(S=\text{High}) = 2.7999999$

$v^*(S=\text{Low}) = 4.3999999$

Iteration needed = 30

policy

Policy: State = Good, Action = Stay

Policy: State = Bad, Action = Move

Python file: EE18MTECH11030_Question_2_d_final.ipynb

Ans: 1)

Deriving the Bellman's equation for state value function as derived in class.

Now,

$$v_\pi(s) = E_\pi[G_t \mid S_t]$$

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

$$= E_\pi\left[R_{t+1} + \sum_{k=0}^{\infty} \gamma^{k+1} R_{t+k+2} \mid S_t\right]$$

$$= E_\pi\left[r + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \mid S_t\right]$$

$$= E_\pi[r + \gamma G_{t+1} \mid S_t]$$

$$= E_\pi[r \mid S_t] + \gamma E_\pi[G_{t+1} \mid S_t] \quad \text{—} (1)$$

Now,

$$E_\pi[E_\pi[G_{t+1} \mid S_{t+1}] \mid S_t] = E_\pi[G_{t+1} \mid S_t]$$

Proof: Let $G_{t+1} = g'$, $S_{t+1} = s'$, $G_t = g$, $S_t = s$

then,

$$E_\pi[E_\pi[g' \mid s'] \mid s] = E_\pi[E_\pi[g' \mid s', s]]$$

$$= E_\pi\left[\sum_{g'} g' \, p(g' \mid s', s)\right]$$

$$= \sum_{s'} \sum_{g'} g' \, p(g' \mid s', s) \, p(s' \mid s)$$

$$= \sum_{s'} \sum_{g} g' \times \frac{p(g', s', s)}{\cancel{p(s', s)}} \cdot \frac{\cancel{p(s', s)}}{p(s)}$$

$$= \sum_{s'} \sum_{g} g' \, \frac{p(g', s', s)}{p(s)}$$

$$= \sum_{s'} \sum_{g'} g' \frac{p(g', s', s)}{p(s)}$$

$$= \sum_{g'} g' \frac{p(g', s)}{p(s)}$$

$$= \sum_{g'} g' \, p(g' \mid s)$$

$$= E_\pi [g' \mid s]$$

$$= E_\pi [G_{t+1} \mid S_t]$$

$\therefore$ (1) $\Rightarrow$

$$v_\pi(s) = E_\pi [r \mid s_t] + \gamma \, E_\pi [G_{t+1} \mid s_t]$$

$$= E_\pi [r \mid s_t] + \gamma \, E_\pi [E_\pi [G_{t+1} \mid S_{t+1}] \mid s_t]$$

$$= E_\pi [r + \gamma \, E_\pi [G_{t+1} \mid S_{t+1}] \mid s_t]$$

$$\boxed{v_\pi(s) = E_\pi [r + \gamma \, v_\pi(s') \mid s_t]}$$