RL Mid sem :-

Markov Decision Process :-

A Markov Decision Process (MDP) model contains :

(i) a set of possible states 'S'

(ii) a set of possible actions A

(iii) a real valued reward function R(s,a)

(iv) a transition T of each action's effects in each state which followes the Markov Property: The effects of an action taken in a state depend only on that state and not on the prior history.

Model- Based RL :- Model generated objective of reinforcement learning is to –

(i) learn an optimal policy π that maximizes the expected total reward.

i.e $$\max_{\pi} \mathop{E}_{\substack{a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim p(\cdot|s_t,a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$$

(ii) Maximize the expected cumulative discounted rewards $r(s_t, a_t)$ from according to a policy π in an environment that is governed by system dynamics p.

Issues of model based RL

(i) In Model based RL we assume a model of the environment and learn it from the interactions with the environment.

(ii) This methods learn with significantly lower sample the model-free RL methods.

(iii) learning an accurate model of the environment has proven to be a challenging problem in certain domains. i.e model Bias.

Outline :-

i) A process in observed at time points : $t = 0, 1, 2, \dots, n$

ii) at stage 't' the process is in a state 's_t' with probability $p(s_t)$ where $s_t \in S$ $(|S| < \infty_{\text{count}})$

(iii) After observing the state of the process at stage/time 't' an action 'a_t' must be choosen, where $a_t \in A$ $(|A| < \infty_{\text{count}})$

$$s_0 \xrightarrow{a_0} s_1 \xrightarrow{a_1} s_2 \xrightarrow{a_2} s_3 \xrightarrow{a_3} \dots$$

→ A state 's_t' is Markov iff : $P(s_{t+1} | s_t) = P(s_{t+1} | s_1, \dots, s_t)$

(iv) After the action a_t has been taken when state of the process was at state 's_t' the process goes to state 's_{t+1}' with probability $P(s_{t+1} | a_t ; s_t)$

$$\begin{array}{ccccccccc} \dots \xrightarrow{A_{t-2}} & s_{t-1} & \xrightarrow{A_{t-1}} & s_t & \xrightarrow{A_t} & s_{t+1} & \xrightarrow{A_{t+1}} \dots \\ & \downarrow & & \downarrow & & \downarrow & \\ & R_{t-1} & & R_t & & R_{t+1} & \end{array}$$

(v) The reward earned is $R(a_t, s_t)$ or $R(s_t)$ is earned.

(vi) Both reward and transition probabilities are functions only of the last state and the last action. (Markov Property)

$$P(s_{t+1} = s_{t+1} | (a_0, s_0), (a_1, s_1), \dots, (a_t, s_t)) = P(s_{t+1} = s_{t+1} | a_t, s_t)$$

$$R(s_t, a_t | (a_1, s_1), (a_2, s_2), \dots, (a_t, s_t)) = R(s_t, a_t)$$

(vii) For a Markov Process having present state s and successor state s', the state transition probability is defined by

$$P_{ss'} = \text{Prob}\,[s_{t+1} = s' \,|\, s_t = s, a_t]$$

(viii) PTM defines the transition probabilities from all present states to all successor states

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & & & \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$$

[each row sums to 1]

(*) A Markov Reward Process is a tuple $\langle S, P, R, \gamma \rangle$

- S is a finite set of states
- P is a state PTM, $P_{ss'} = P[S_{t+1} = s' \mid S_t = s]$
- R is a reward function, $R_s = E[R_{t+1} \mid S_t = s]$
- γ is a discount factor, $\gamma \in [0, 1]$

(*) upon visiting sequence of states $s_0, s_1, \ldots$ with actions $a_0, a_1, \ldots$, $\boxed{\text{Total Payoff}} = R(s_0, a_0) + \gamma R(s_1, a_1) + \gamma^2 R(s_2, a_2) + \cdots$

where $\gamma \in [0, 1]$

(*) Goal : maximize expected total discounted reward

$$E(R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots)$$

(*) $\boxed{\text{Undiscounted Markov reward process} \quad \gamma = 1}$

Policy:- is any function mapping from the states to the actions ; $a = \pi(s)$ where $\pi : S \to A$

Stationary Policy :- is one which is followed at every stage

Value function :-

The state value function $v(s)$ of an MDP is the expected return starting from the state 's'

$$v(s) = E(G_t \mid S_t = s)$$

$$= E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid S_t = s]$$

$$= E[R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \cdots) \mid S_t = s]$$

$$= E[R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

$$= E[R_{t+1} \mid S_t = s] + \gamma v(S_{t+1})$$

$$\therefore v^{\pi}(s) = E[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots \mid s_0 = s, \pi]$$

$$= R(s) + \gamma \sum_{s' \in S} P_{ss'} v^{\pi}(s')$$

Bellman Equation :-

$$v^{\pi}(s) = R(s) + \gamma \sum_{s' \in S} P_{ss'} v^{\pi}(s')$$

i.e $V = R + \gamma P V$

$$\begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix} = \begin{bmatrix} R_1 \\ \vdots \\ R_n \end{bmatrix} + \gamma \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & & \\ p_{n1} & \cdots & p_{nn} \end{bmatrix} \begin{bmatrix} v(1) \\ v(2) \\ \vdots \\ v(n) \end{bmatrix}$$

$\therefore$ **Direct solution** : $V = (I - \gamma P)^{-1} R$

Complexity: $O(n^3)$ for n states.

Existence :-

we have to show that $(I - \gamma P)$ is invertible

Now, P is a Stochastic matrix, $P\mathbb{1} = \underline{1} \Rightarrow 1$ is an eigen value of P.

let $\exists\ \lambda > 1$ and non-zero $\underline{x}$ s.t. $P\underline{x} = \lambda \underline{x}$.

as P has non-negative row values and it sum to 1 for each row then each element of $P\underline{x}$ is a convex combination of the components of $\underline{x}$:

as a convex combination can't greater that x_{max} (the largest comp- of $\underline{x}$) $\Rightarrow$ our assumption is wrong [our assumptions $\Rightarrow$ atleast one element λx_{max} in the RHS (i.e in $\lambda\underline{x}$) is greater than x_{max}]

$\Rightarrow \lambda > 1$ is not possible.

ie largest eigen value of P is 1.

$\therefore$ the smallest eigen value of $(I - \gamma P)$ is $(1-\gamma)$ for $\gamma < 1 \Rightarrow$

$(I - \gamma P)$ is invertible [as $(1-\gamma) > 0$]

[side proof: For all eig. val. λ_i of A and corresponding eign vec v_i

s.t. $Av_i = \lambda_i v_i$ then

$eig(I + \gamma A) = 1 + \gamma\lambda_i$ [γ is a scaler]

$\rightarrow\ Av_i = \lambda_i v_i$

$\gamma A v_i = \gamma\lambda_i v_i$

$v_i + \gamma A v_i = v_i + \gamma\lambda_i v_i$

$\Rightarrow (I + \gamma A) v_i = (1 + \gamma\lambda_i) v_i$]

Value iteration :-

Consider only MDPs with finite state and action spaces. The value iteration algorithm —

(i) For each state 's', initialize $V(S) = 0$

(ii) Repeat untill convergence :

$$V(S) : R(S) + \max_{\pi} \gamma \sum_{s' \in S} P_{ss'} V(s')$$

(iii) Value iteration will cause v to converge to v^*.

(iv) Having found v^* we can find π^* as

$$\pi^*(s) = \arg\max_{\pi} \gamma \sum_{s' \in S} P_{ss'} V^*(s')$$

Convergence Proof :-

Value iteration converges to optimal value $\hat{V} \to v^*$.

Proof : For any estimate of v ; $\hat{v}$ we define the Bellman Backup operator $B : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ such that.

$$B\hat{v}(s) = R(s) + \gamma \max_{a \in A} \sum_{s' \in S} P(s'|s,a)\, \hat{v}(s')$$

In order to prove that $\hat{v} \to v^*$; we've to simply show that B is a contraction map.

i.e $$\max_{s \in S} |Bv_1(s) - Bv_2(s)| \leq \gamma \max_{s \in S} |v_1(s) - v_2(s)|.$$

$$|Bv_1(s) - Bv_2(s)| = \gamma \left| \max_{a \in A} \sum_{s' \in S} P(s'|s,a) v_1(s') - \max_{a \in A} \sum_{s' \in S} P(s'|s,a) v_2(s') \right|$$

$$\leq \gamma \max_{a \in A} \left| \sum_{s' \in S} P(s'|s,a) v_1(s') - \sum_{s' \in S} P(s'|s,a) v_2(s') \right|$$

$$= \gamma \max_{a \in A} \sum_{s' \in S} P(s'|s,a) \left| v_1(s') - v_2(s') \right| = \gamma \max_{a \in A} |v_1(s') - v_2(s')|$$

$$\leq \gamma \max_{s \in S} |v_1(s) - v_2(s)|$$

$$\therefore \max_{s \in S} |B v_1(s) - B v_2(s)| < \gamma \max_{s \in S} |v_1(s) - v_2(s)|$$

Now Let $v_K = B^{K-1} v_0$ $\left[\text{as } v_{K+1}(s) = B\, v_K(s) \right]$

$$\Rightarrow \max_{s \in S} |v_K(s) - v^*(s)| = \max_{s \in S} |B v_{K-1}(s) - B v^*(s)| < \gamma \max_{s \in S} |v_{K-1}(s) - v^*(s)|$$

$$\leq \dots \leq \gamma^K \max_{s \in S} |v_0(s) - v^*(s)|$$

Now as $K \to \infty$ we have

$$\max_{s \in S} |v_K(s) - v^*(s)| \to 0$$

$$\Rightarrow \lim_{K \to \infty} v_K = v^*. \quad (\text{Proved})$$

Note:- ① $v^*(s) \leq \sum_{t=1}^{\infty} \gamma^t R_{max} = \frac{R_{max}}{1-\gamma}$ $\left[\because v^{\pi}(s) = E\left[\sum \gamma^t R(s_t)\right] \right]$

② $\max_{s \in S} |v^K(s) - v^*(s)| \leq \frac{\gamma^K R_{max}}{1-\gamma}$

Policy Iteration :-

a) given policy π, Calculate $v := v^{\pi}$ (utility of each state if π were to be executed)

b) Calculate a new policy using:

$$\pi^*(s) := \arg\max_{a \in A} \gamma \sum_{s' \in S} P(s'|s,a)\, v^*(s')$$

$$\left[\pi_0 \to v^{\pi_0} \to \pi_1 \to v^{\pi_1} \to \pi_2 \to v^{\pi_2} \to \dots \to \pi^* \to v^{\pi^*} \right]$$

Action Value Function :-

$Q^{\pi}(s,a)$ where a is an action and s is a state,

$Q^{\pi}(s,a)$ is the expected value of doing 'a' in state 's', then following policy 'π'.

$$Q^{\pi}(s,a) = \sum_{s'} P(s'|a,s)\left(r(s,a,s') + \gamma V^{\pi}(s')\right)$$

$$V^{\pi}(s,a) = Q^{\pi}(s,\pi(s))$$

For V^{π}

$$V_{\pi}(s) = \sum_{a \in A} \pi(a|s)\, q_{\pi}(s,a)$$

For Q^{π}

$$q_{\pi}(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_{\pi}(s')$$

Understanding V and Q Functions :-

Value function $V^{\pi}(s) = E\left(R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots \mid s_0 = s, \pi\right)$

$$= \underbrace{R(s)}_{\text{immediate Reward}} + \underbrace{\gamma \sum_{s' \in S} P_{ss'} V^{\pi}(s')}_{\text{Expected sum of future discounted Reward.}}$$

Optimal Value function :-

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

$$= R(s) + \max_{\pi} \gamma \sum_{s' \in S} P_{ss'} V^{\pi}(s')$$

Optimal policy :- $\pi^*(s) = \arg\max_{\pi} \gamma \sum_{s' \in S} P_{ss'} V^{\pi}(s')$.

Q Function :-

The optimal value function gives the expected return if we start in state 's' and always acts according to the optimal policy in the environment.

$$V^*(s) = \max_{\pi} \underset{\tau \sim \pi}{E} \left[R(\tau) \mid s_0 = s \right]$$

The optimal value action function $Q^{*\pi}(s,A)$ gives the optimal expected reward if we start 's', take an arbitary action a (may not come from policy) and then forever after act according to optimal policy π.

$$Q^{\pi}(s,a) = \underset{\tau \sim \pi}{E} \left[R(\tau) \mid s_0 = s, a_0 = a \right]$$

$$Q^*(s,a) = \max_{\pi} E \left[R(\tau) \mid s_0 = s, a_0 = a \right]$$

$$a^*(s) = \pi^* = \arg\max_{a} Q^*(s,a)$$

Relation between $V^{\pi}(s)$ and $Q^{\pi}(s,a)$:-

$$V^{\pi}(s) = \underset{\tau \sim \pi}{E} \left[R(\tau) \mid s_0 = s \right]$$

$$= \underset{a \sim \pi}{E} \left[R(\tau) \mid s_0 = s, a_0 = a \right]$$

$$= \underset{a \sim \pi}{E} \left[E (R(\tau) \mid s_0 = s, a_0 = a) \right] = \underset{a \sim \pi}{E} \left(Q^{\pi}(s,a) \right).$$

and we can have

$$V^{\pi}(s,a) = Q^{\pi}(s, \pi(s)).$$ [value function and Q- function are eaual when $a \sim \pi$]

Compact Bellman equations :-

$$V^{\pi}(s) = \underset{\substack{a \sim \pi \\ s' \sim P}}{E} \left[r(s,a) + \gamma V^{\pi}(s') \right]$$

$$\therefore Q^{\pi}(s,a) = \underset{s' \sim P}{E} \left[r(s,a) + \gamma \underset{a' \sim \pi}{E} \left[Q^{\pi}(s',a') \right] \right]$$

$$\therefore V^{*}(s) = \max_{a} \underset{s' \sim P}{E} \left[r(s,a) + \gamma V^{*}(s') \right]$$

$$Q^{*}(s,a) = \underset{s' \sim P}{E} \left[r(s,a) + \gamma \max_{a} Q^{*}(s,a) \right]$$