Lec 2

MDP ⇒ environment
(FULLY OBSERVABLE)

Any partially observable problem can be completely converted to MDP

Recall — Markov Property where current state categorizes all relevant information

- State transition probability

$$P_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s]$$

now we can create a matrix from all s to all s'

$$P = \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ & \text{current} \quad \text{next} & \\ P_{n1} & \cdots & P_{nm} \end{bmatrix}$$

- Markov Process / Chain → sequence of states with Markov Property

Tuple (S, P)

terminal state (self loop) like a rest state

Sample episodes just mean sequence of states

Markov Reward Process (of a sample episode)

Chain with value judgements

(S, P, R, γ)

- P: matrix
- R: $R_s = \mathbb{E}[R_{t+1} \mid S_t = s]$ — next time step pe you get reward of current step
- $\gamma \in [0,1]$: farsighted / shortsighted/myopic

Return G_t

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

the more close it is to 0: we prefer immediate rewards

" " to 1: indifferent to when rewards arrive

γ — mathematically convenient
- avoid infinite returns in cyclic processes
- if financial, we may want immediate rewards coz interest

State-value Function

long term value of being in state s: expected return starting from state s

$$v(s) = \mathbb{E}[G_t \mid S_t = s]$$

(s: state)

Take one start state, for many episodes figure value & average (one way)

Bellman Equation for MRP

$$v(s) = \mathbb{E}[G_t \mid S_t = s]$$

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

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

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

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

↑ immediate reward

↑ value function of next state

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

all states

reward from exiting state

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

$v = R + \gamma P v$

$(1 - \gamma P) v = R$

$v = (1 - \gamma P)^{-1} R \longrightarrow O(n^3)$ not practicable

only works for small MDPs

for large MDP — DP, Monte Carlo, Temporal Difference

MDP are MRP with decisions / actions

tuple (S, A, P, R, γ)

$R_s^a = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$

finite set of states & actions

state transition matrix

[it now depends on action so new matrix for each action]

Policy $\pi(a|s) = \mathbb{P}[A_t = a | S_t = s]$

Markov property holds that only current policy matters

i.e. time independent (policy independent) i.e. same policy in each time step

New definitions

$v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s]$

policy

$\mathbb{E}_\pi$ — when we sample all actions according to π

how good it is to be in state s if i am following policy π

Action value function — how good it is to take action 'a' from state 's'

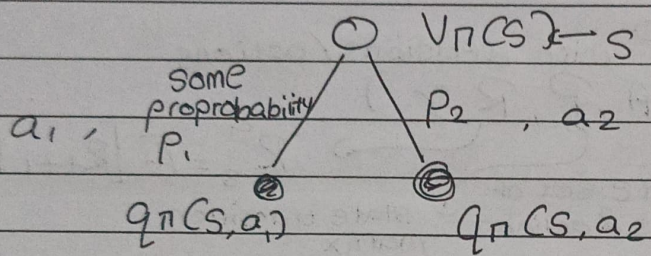
$q_\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$

Bellman Expectation Eqn

$$v_\pi(s) = \mathbb{E}_\pi[\underbrace{R_{t+1}}_{\text{immediate reward}} + \gamma\, v_\pi(s_{t+1}) \mid s_t = s]$$

$$q_\pi(s,a) = \mathbb{E}[R_{t+1} + \gamma q_\pi(s_{t+1}, A_{t+1}) \mid s_t = s, A_t = a]$$

v & q relate to each other

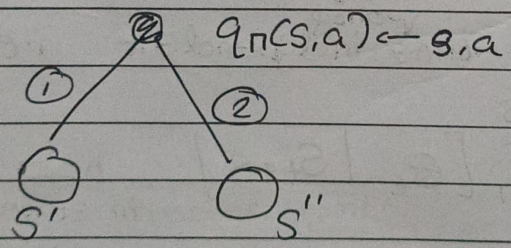


$\therefore v_\pi(s) = p_1 q_\pi(s,a_1) + p_2 q_\pi(s,a_2)$

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

Converse relation

Environment might blow you over ① or ②
So v(s) gives how good ① & ② are

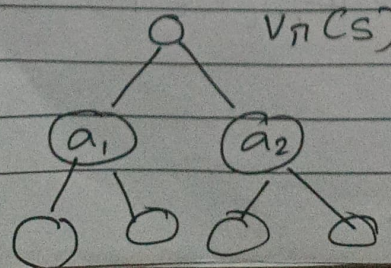


consider specific action we take from state

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

combining both

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) \left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a\, v_\pi(s') \right)$$



MDP can be converted to MRP

__/__/___

other

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

Basic idea

value function = immediate reward + value function of next state

Optimal value function (best behaviour/ in MDP) solution

to pick best policy, see which policy gives max value

$$V_*(s) = \max V_\pi(s)$$

↳ max possible reward to extract from system

Similarly optimal action value max reward you get when you fix state & action

$$q_*(s,a) = \max q_\pi(s,a)$$

↓

sufficient for RL model

MDP is "solved" when you know q^*

~~Optimal Policy~~

__/__/____

Optimal Policy

$$\pi \geq \pi' \text{ if } v_\pi(s) \geq v_{\pi'}(s)$$
$$\forall \text{ all states}$$

All MDP have ~~one~~ unique optimal policy (could be more than 1)

$$\pi_* \geq \pi \quad \forall \pi$$

$$v_{\pi_*}(s) = v_*(s)$$

optimal value function

$$q_{\pi_*}(s,a) = q_*(s,a)$$

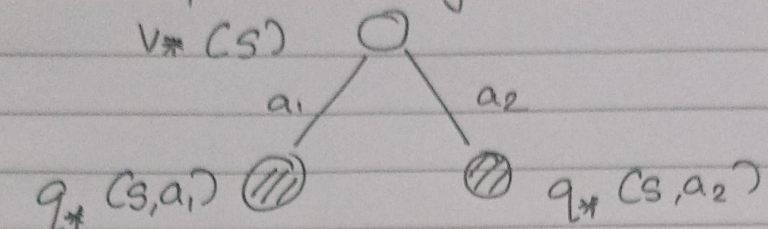
$$\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \underset{a \in A}{\text{argmax}}\ q_*(s,a) \\ 0 & \text{o/w} \end{cases}$$

Pick action with more q

how to arrive at q_* : Bellman Optimality

$R_{t+1} = f(S_t)$

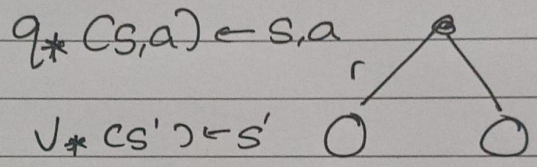
Bellman optimal Policy (how to figure out q^*)



how good is it to be in state s, is = max q_* value of actions you can perform

$$V_*(S) = \max_a q_*(S,a) \quad \text{— (i)}$$

Opp



WE DONT KNOW FOR CERTAIN WHERE ENVIRONMENT MIGHT TAKE US

$$q_*(S,a) = R_s^a + \gamma \sum_{S' \in S} P_{SS'}^a V_*(S') \quad \text{(ii)}$$

Combining (i) & (ii) we get (2 step look ahead)

$$\# \quad V_*(S) = \max_a R_s^a + \gamma \sum_{S' \in S} P_{SS'}^a V_*(S')$$

$$\# \quad q_*(S,a) = R_s^a + \gamma \sum_{S' \in S} P_{SS'}^a \max_{a'} q_*(S',a')$$

This is not linear (max coz of), so methods to solve

- → value iteration
- → policy iteration
- → Q-learning
- → sarsa