Lecture 2 Markov Decision Processes

MDP
- formally describe an environment for reinforcement learning
- Where the environment is fully observable
- The current state completely characterises the process

Markov Property :
- "The future is independent of the past given the present"
- The state captures all relevant information from the history
- A state S_t is Markov if and only if.......

Markov Process :
- A Markov process or Markov chain is a tuple $< S, P >$
- S : a finite set of states
- P : transition probability matrix

Markov Reward Process :
- In this process, the only stochastic process is transition to next state.
- A Markov Reward Process is a tuple $< S, P, R, \gamma >$
- R is a reward function, $R_s := E[\mathrm{R}_{t+1} | S_t = s]$
- Important : R_t depends on a state. It is deterministic. Therefore, R_{t+1} means reward of S_{t+1}.
- γ is a discount factor, $\gamma \in [0,1]$
- Return $G_t = \mathrm{R}_{t+1} + \gamma \mathrm{R}_{t+2} + ... = \sum_{k=0}^{\infty} \gamma^k \mathrm{R}_{t+k+1}$
- Therefore, $G_t = R_{t+1} + \gamma G_{t+1}$
- Question : Why G_t doesn't include R_t?
- The value of receiving reward R after $k+1$ time-steps is $\gamma^k \mathrm{R}$

Value Function
- The value function $v(s)$ gives the long-term value of state s.
- The state value function v(s) of an MRP is the expected return
- starting from state s
- $v(s) = E[G_t | S_t = s]$
- Therefore, for transition matrix $P := P_{ss'} = P[S_{t+1} = s' | S_t = s]|_{P \in M_{N \times N}}$, reward vector $r := r_k = R_{s_k}$, and state probability vector $s := s_k = P[S_t = s_k]$,

$$v(s) = E[G_t | S_t = s] = \left(\sum_{k=0}^{\infty} \gamma^k P^k \right) (sr)$$

Bellman Equation for MRP

- $v(s) = E[G_t|S_t = s]$
  $v(s) = E[R_{t+1} + \gamma G_{t+1}|S_t = s]$
  $v(s) = E[R_{t+1}|S_t = s] + E[\gamma G_{t+1}|S_t = s]$
  $v(s) = R_s + E[\gamma G_{t+1}|S_t = s]$
- Question : According to ppt, $G_{t+1} = v(S_{t+1})$. At least $E[G_{t+1}|S_t = s] = E[v(S_{t+1})|S_t = s]$. Why?
- According to my calculation, $v(s) = \left(\sum_{k=0}^{\infty}\gamma^k P^k\right)(sr)$

$$= \left(\left(\sum_{k=0}^{0}\gamma^k P^k\right) + \left(\sum_{k=1}^{\infty}\gamma^k P^k\right)\right)(sr)$$
$$= \left(I + \gamma P\left(\sum_{k=0}^{\infty}\gamma^k P^k\right)\right)(sr)$$
$$= sr + \gamma P\left(\sum_{k=0}^{\infty}\gamma^k P^k\right)(sr)$$
$$= sr + \gamma P v(s)$$

- Therefore, $v(s) = sr + \gamma P v(s)$
  $\therefore v(s)(1-\gamma P) = sr$
  $\therefore v(s) = sr(1-rP)^{-1}$

Markov Decision Process

- A Markov decision process (MDP) is a Markov reward process with decisions. It is an environment in which all states are Markov
- A Markov Decision Process is a tuple $< S, A, P, R, \gamma >$
- A is a finite set of actions
- P is a state transition probability matrix, $P_{ss'}^a = P[S_{t+1} = s'|S_t = s, A_t = a]$
- Therefore, P can be written in $M \times N \times N$ matrix that $M = n(A)$ and $N = n(S)$
- R is a reward function, $R_s^a = E[R_{t+1}|S_t = s, A_t = a]$

Policy

- A policy π is a distribution over actions given states, $\pi(a,s) = P[A_t = a|S_t = s]$ [1)]
- Therefore, π can be written in $M \times N$ matrix that $M = n(S)$ and $N = n(A)$.
- MDP policies depend on the current state
- Policies are stationary(time-independent), $A_t \sim \pi(\bullet, S_t), \forall t > 0$
- Then, my question is What is P^π? By definition, $P^\pi = P_{ss'}^\pi = \sum_{a \in A}\pi(a,s)P_{ss'}^a$.
- $P_{ss'}^\pi = \sum_{a \in A}\pi(a,s)P_{ss'}^a$
  $\therefore P_{ss'}^\pi = \sum_{a \in A}P[a|s]P_{ss'}^a$
  $\therefore P_{ss'}^\pi = E[P_{ss'}^A]$
- It means that, P^π means ultimate probability considering action.

1) Note that I used $\pi(a,s)$ instead $\pi(a|s)$