Reinforcement Learning Markov Decision Process Notes

Lu Hong [1]

[1] *Nanjing University of Aeronautics and Astronautics*

September 21, 2019

1 Markov Process

"The future is independent of the past given the present". A state S_t is *Markov* iif.

$$P[S_{t+1}|S_t] = P[S_{t+1}|S_1, S_2, ..., S_t]$$

Markov Process (or Markov Chain) is a *memoryless random process*, represented by a tuple $< S, P >$

2 Markov Reward Process

Markov Reward Process is Markov Process with values, represented by a tuple $< S, P, R, \gamma >$, where $R = E[R_{t+1}|S_t = s]$ and γ is a discount factor.

Return G_t is the *total discounted reward* at time-step t presented by

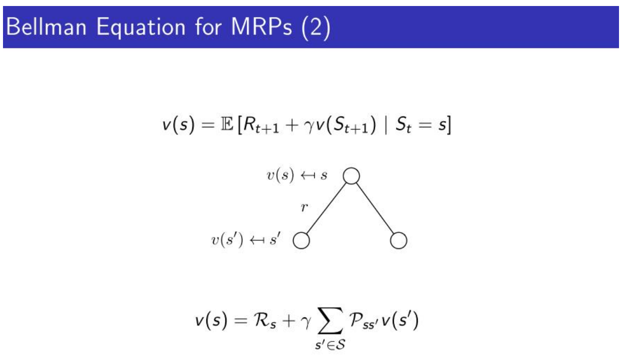
$$G_t = \sum_{i=0}^{\infty} R_{t+i+1}$$

Also, we define Value Function to indicate the long-term value of the state s.

$$V(s) = E[G_t|S = s]$$

* Bellman Equation for MRPs, it demonstrate that MRPs can be presented in recursive format.

$$\begin{aligned} v(s) &= E[G_t|S_t = s] \\ &= E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ...|S_t = s] \\ &= E[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + ...)|S_t = s] \\ &= E[R_{t+1} + \gamma G_{t+1}|S_t = s] \\ &= E[R_{t+1} + \gamma v(S_{t+1})|S_t = s] \end{aligned}$$



To express Bellman Equation in matrix form as

$$v = R + \gamma P v,$$

we can solve Bellman Equation easily as linear equation.

$$\begin{aligned} v &= R + \gamma P v \\ (I - \gamma P)v &= R \\ v &= (I - \gamma P)^{-1} R \end{aligned}$$

•

3 Markov Decision Process

A MDP is a MRP with decisions, represented by a tuple $< S, A, P, R, \gamma >$, to be specific, some params have been changed after actions added. $P^a_{ss'} = P[S_{t+1} = s'|S_t = s, A_t = a]$ and $R^a_s = E[R_{t+1}|S_t = s, A_t = a]$

*Policies: A policy π is a distribution over actions given states,

$$\pi(a|s) = P[A_t = a|S_t = s]$$

Policies (2)

- Given an MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ and a policy π
- The state sequence $S_1, S_2, ...$ is a Markov process $\langle \mathcal{S}, \mathcal{P}^{\pi} \rangle$
- The state and reward sequence $S_1, R_2, S_2, ...$ is a Markov reward process $\langle \mathcal{S}, \mathcal{P}^{\pi}, \mathcal{R}^{\pi}, \gamma \rangle$
- where

$$\mathcal{P}^{\pi}_{s,s'} = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{P}^{a}_{ss'}$$
$$\mathcal{R}^{\pi}_{s} = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{R}^{a}_{s}$$