Markov Decision Process

Farnaz Adib Yaghmaie

Linkoping University, *Sweden*
farnaz.adib.yaghmaie@liu.se

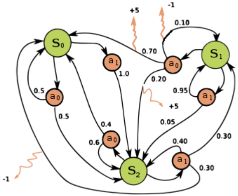
March 12, 2021

Markov Decision Processes

- describe environment in RL framework
- describe dynamical systems
- In optimal control problems MDPs are continuous

A Markov Decision Process (MDP) is a tuple $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$

- $\mathcal{S}$: The set of states.
- $\mathcal{A}$: The set of actions.
- $\mathcal{P}$: The set of transition probability.
- $\mathcal{R}$: The set of immediate rewards associated with the state-action pairs.
- $0 \leq \gamma \leq 1$: Discount factor.



Modified version of @ https://en.wikipedia.org/wiki/Markov_decision_process

States: Describe internal status of MDP

Actions: Possible choices to make in each state of MDP

The state and action space can be finite or infinite and it is extremely important!

Transitions probability: $\mathcal{P}$ is the set of transition probability with n_a matrices each of dimension $n_s \times n_s$ where $s,\ s'$ entry reads

$$[\mathcal{P}^a]_{ss'} = p[s_{t+1} = s' | s_t = s,\ a_t = a] \tag{1}$$

Reward:

$$r_t = r(s,\ a) \tag{2}$$

Total reward:

$$R(T) = \sum_{t=1}^{T} \gamma^t r_t \tag{3}$$

Average reward:

$$R(T) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} r_t \tag{4}$$

Do you care about future as much as now (and past)?

- $\gamma \to 0$: We only care about the current reward not what we'll receive in future
- $\gamma \to 1$: We care about all rewards equally

$$\mathcal{S} = \{s_0,\ s_1,\ s_2\},$$

$$\mathcal{A} = \{a_0,\ a_1\},$$

$$\mathcal{P}^{a_0} = \begin{bmatrix} 0.5 & 0 & 0.5 \\ 0.7 & 0.1 & 0.2 \\ 0.4 & 0 & 0.6 \end{bmatrix},$$

$$\mathcal{P}^{a_1} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0.95 & 0.05 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}.$$

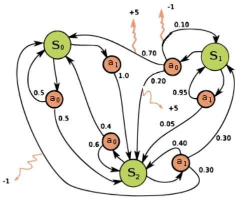


Photo Credit: © https://en.wikipedia.org/wiki/Markov_decision_process

- Policy: The agent's decision
  - Deterministic policy $a = \pi(s)$
  - stochastic policy $\pi(a|s) = P\left[a_t = a|s_t = s\right]$
- Value function: how good the agent does in a state

$$V(s) = \mathbf{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ...|s_t = s\right]$$

- Model: The agent's interpretation of the environment

Not all components are necessary!

Email your questions to

farnaz.adib.yaghmaie@liu.se