CME 242

22 janvier 2020

Table of Contents

I Introduction

Markov decision processes formally describe an environment for reinforcement learning when the environment is fully observable. Almost all RL problems can be formalised as MDPs.

II Markov Decision Processes

II.1 Markov Processes

Definition [Markov Property] : A state S_t is Markov iif :

$$\mathbb{P}(S_{t+1}|S_t) = \mathbb{P}(S_{t+1}|S_1, \ldots, S_t)$$

— The state captures all relevant information from the history.
— The state is a sufficient statistic of the future.

Definition [State Transition Matrix] : For a Markov state s and successor state $s^{'}$, the state transition probability is defined by

$$(P)_{ss'} = \mathbb{P}(S_{t+1} = s^{'}|S_t = s)$$

Definition [Markov Process] : A Markov process is a memoryless random process, i.e. a sequence of random states S_1, S_2, ... with the Markov property. A Markov Process (Markov Chain) is a tuple $(\mathcal{S}, \mathcal{P})$ such that :
— $\mathcal{S}$ is a (finite) set of states.
— $\mathcal{P}$ is a state transition probability matrix.

II.2 Markov Reward Processes

Definition [Markov Reward Process] : A Markov Reward process is a tuple $(\mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma)$:

— $\mathcal{S}$ is a finite set of states
— $\mathcal{P}$ is a state transition probability matrix
— $\mathcal{R}$ is a reward function
— γ is a discount factor

Definition [Return] : The return G_t is the total discounted reward from time-step t.

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

Q : Most Markov reward and decision processes are discounted, why ?
— Avoids infinite returns in cyclic Markov processes

— Mathematically convenient

Definition [state value function] : The state value function $v(s)$ of an MRP is the expected return starting from state s

$$\boxed{v(s) = \mathbb{E}(G_t|S_t = s)} \tag{1}$$

Bellman Equation for MRPs

The value function can be decomposed into two parts :
— immediate reward : R_{t+1}
— discounted value of successor state : $\gamma v(S_{t+1})$

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

So we have :

$$\boxed{v(s) = \mathcal{R}_s + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'} v(s')} \tag{2}$$

And we can vectorize this equation by writing :

$$\boldsymbol{v} = \mathcal{R} + \gamma \mathcal{P} \boldsymbol{v}$$

— This equation is linear.
— The direct solution is $\boldsymbol{v} = (I - \gamma\mathcal{P})^{-1}\mathcal{R}$.
— We want to solve the value function analytically (rather than simulations).
— Computional complexity is $O(n^3)$ for n states.
— Direct solution only possible for direct MRPs.
— Otherwise : iterative methods such as DP, Monte-Carlo evaluation.

II.3 Markov Decision Process

Definition [Markov Decision Process] : A Markov Decision process is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$:

— $\mathcal{S}$ is a finite set of states
— $\mathcal{A}$ is a finite set of actions
— $\mathcal{P}$ is a state transition probability matrix
$\mathcal{P}^a_{ss'} = \mathcal{P}(S_{t+1} = s'|S_t = s, A_t = a)$
— $\mathcal{R}$ is a reward function, $\mathcal{R}^a_s = \mathbb{E}(R_{t+1}|S_t = s, A_t = a)$
— γ is a discount factor

Definition [Policy] : A policy π is a distribution over action given states :

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

NB : Polices are time-independent (stationary).

Definition [State-Value Function] : The state-value function $v_\pi(s)$ of an MDP is the expected return starting from state s, and then following policy π :

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

Definition [Action-Value Function] The action value-function $q_\pi(s, a)$ is the expected return starting from state s, taking action a, and then following policy π :

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

The Bellman expectation equations can be expressed concisely :

Bellman Expection Equation for $\mathcal{V}^\pi$

$$\boxed{v_\pi(s) = \sum_{a\in\mathcal{A}} \pi(s|a)q_\pi(s, a)} \tag{3}$$

Bellman Expection Equation for $\mathcal{Q}^\pi$

$$\boxed{q_\pi(s, a) = \mathcal{R}_s^a + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s')} \tag{4}$$

Definition [Optimal Value Function]

The optimal state-value function $v_*(s)$ is the maximum value function over all policies :

$$v_*(s) = max_\pi v_\pi(s)$$

The optimal action-value function $q_*(s, a)$ is the maximum action-value function over all policies :

$$q_*(s, a) = max_\pi q(s, a)$$

MDP is solved when we know the optimal value function.

Theorem [Optimal Policy] : For any Markov Decision Process

— There exists an optimal policy π_* that is better than or equal to all other policies, $\pi_* \geq \pi, \forall \pi$
— All optimal policies achieve the optimal value function : $v_{\pi_*}(s) = v_*(s)$
— All optimal policies achieve the optimal action-value function : $q_{\pi_*}(s,a) = q_*(s,a)$

There is always a deterministic optimal policy for any MDP. If we know $q_*(s,a)$, we immediately have the optimal policy :

$$\pi_*(a|s) = \delta_{a,argmax_{a \in \mathcal{A}} q_*(s,a)}$$

The Bellman Expectation Equations are the same with π_*.