

CME 242

16 janvier 2020

# Table of Contents

I	Introduction . . . . .	2
II	Markov Decision Processes . . . . .	2
II.1	Markov Processes . . . . .	2
II.2	Markov Reward Processes . . . . .	2
II.3	Markov Decision Process . . . . .	3

# 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 :

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

- 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  $S_1, S_2, \dots$  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
- $\gamma$  is a discount factor

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

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

*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  $s$

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

### 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 :

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

And we can vectorize this equation by writing :

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

- This equation is linear.
- The direct solution is  $\mathbf{v} = (I - \gamma \mathcal{P})^{-1} \mathcal{R}$ .
- We want to solve the value function analytically (rather than simulations).
- Computational 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  
 $\mathcal{P}_{ss'}^a = \mathcal{P}(S_{t+1} = s' | S_t = s, A_t = a)$
- $\mathcal{R}$  is a reward function
- $\gamma$  is a discount factor

**Definition** : A policy  $\pi$  is a distribution over action given states :

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

**Definition** : The state-value function  $v_\pi(s)$  of an MRP is the expected return starting from state  $s$ , and then following policy  $\pi$ .

**Definition** The action value-function

The Bellman expectation equation can be expressed concisely

**Definition [Optimal Value Function]**

A MDP is solved when we know the optimal value function.

**Theorem [Optimal Policy]** : For any Markov Decision Processes

---