Markov Decision Process

The 1st step to apply RL is to formulate the problem as a **MDP**

Starting from a Markov Process:

- if we add R we get a MARKOV REWARD PROCESS
- if we add A we get a MARKOV DECISION PROCESS

Why MDP?

1. Actions influence **present** and **future** rewards
2. MDP involves **delayed** rewards
3. With MDP we introduce the concept of **STATE** (not present in bandit)

MDPs describe an environment for RL <u>FULLY OBSERVABLE</u>

<u>STATE</u>: A state S_t is Markov $\Leftrightarrow$ $\mathbb{P}[S_{t+1} | S_t] = \mathbb{P}[S_{t+1} | S_1 \dots S_t]$

<u>STATE TRANSITION PROBABILITY</u>: $P_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s]$

↓

the sum of each row is 1

1

MARKOV PROCESS: y_t is a tuple $\langle S, P \rangle$

- S : finite set of states
- P : state transition probability matrix

2

MARKOV REWARD PROCESS: tuple $\langle S, P, R, \gamma \rangle$

- R : reward function $R_s = E[R_{t+1} \mid S_t = s]$
- γ : discount factor $[0, 1]$

γ → present value of future rewards

GOAL: MAXIMIZE TOTAL REWARD

RETURN: $G_t = R_{t+1} + \gamma R_{t+2} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+1+k}$

VALUE FUNCTION: $V_s = E[G_t \mid S_t = s]$

↳ expected return, starting from s

BELLMAN EQUATION FOR MRP

$$v(s) = R_s + \gamma \sum_{s' \in S} P_{ss'} v(s')$$

This value func. can be devided in 2 parts:
- immediate reward
- discounted value of successor state

- Bellman Equation expres a relationship between the value of a state and the values of succesor states
- BE averages over all the possibilities (sl. 24)
- CC: $O(n^3)$ → very expensive

(3)

MARKOV DECISION PROCESS is a tuple $\langle S, A, P, R, \gamma \rangle$

- A: set of action
- $P^a_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a]$ ↑ matrix ∀ action
- $R^a_s = E[R_{t+1} \mid A_t = a, S_t = s]$

How do we make decisions? POLICIES

POLICY: $\pi(a|s) = \mathbb{P}[A_t = a \mid S_t = s]$

↳ distribution over actions given states

↳ defines the behaviour of an agent

VALUE FUNC FOR MDP: $V_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s]$

↳ Same as value func but it follows the policy π

ACTION-VALUE FUNC: $q_\pi(a|s) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$

BELLMAN EXPECTATION EQUATION

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

$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')$$

The probability of taking one action or another one depends on the policy.

DIFFERENCE BETWEEN B.E. AND B.E.E.