Master 2 Computer Science
RL Course M2 AI

Markov Decision Processes

Akka Zemmari

RI : Agent interacting with an environment

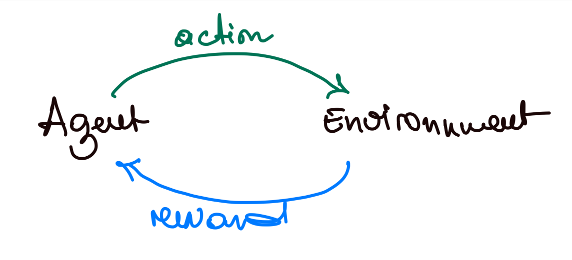


Figure 1: Agent interacting with an environment

Introduction to MDPs

A Markov Decision Process (MDP) is defined as a tuple:

$$M = (S, A, P, R, \gamma)$$

- S: set of states
- A: set of actions
- P: transition probability function $P(s'|s, a)$:

$$\begin{array}{rcccl} P & : & S \times A \times S & \rightarrow & [0, 1] \\ & & (s, a, s') & \mapsto & P(s', s, a) = \mathbb{P}r(s_{t+1} = s' \mid s_t = s, a_t = a) \end{array}$$

- R: reward function $R(s, a)$

$$R : S \times A \rightarrow \mathbb{R}$$

- γ: discount factor

Introduction to MDPs

Toy Example

The grid world is a simple MDP with a 2D grid of states.

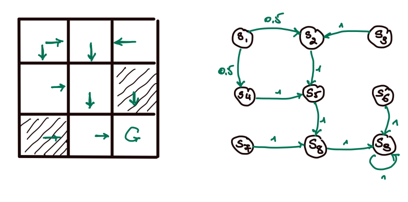


Figure 2: Grid world

Introduction to MDPs

Dynamic of the MDP

- The Dynamic of the MDP is defined by the transition probability function $P(s'|s, a)$ and the reward function $R(s, a)$.
- It can also be caracterized by:

$$p(s', r \mid s, a) = \mathbb{P}r(s_{t+1} = s', r_{t+1} = r \mid s_t = s, a_t = a)$$

Policy and Value Functions

- A policy π is a mapping from states to actions:

$$\pi : S \to A$$

- More generally, a policy can be stochastic. $\pi(a, s)$ (or $\pi(a|s)$) is the probability of taking action a in state s:

$$\begin{aligned} \pi : \quad S \times A &\to [0, 1] \\ (s, a) &\mapsto \pi(a, s) = \mathbb{P}r(a_t = a \mid s_t = s) \end{aligned}$$

Policy and Value Functions

The ultimate goal of an agent is to find a policy π that maximizes the expected sum of rewards:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]$$

Policy and Value Functions

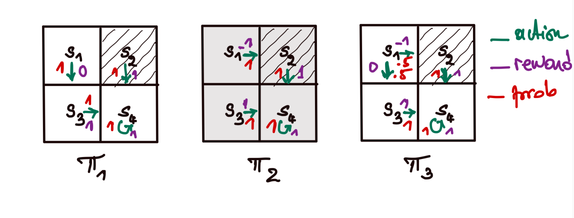
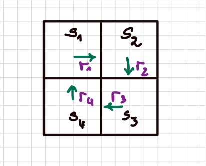


Figure 3: Question: Starting from state s_1 wich policy is best? (See the blackboard)

Policy and Value Functions

How to evaluate a policy?

Let v_i be the value of state s_i under policy π.



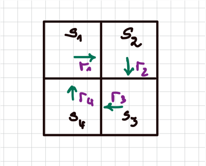
First method :

$$\begin{aligned} v_1 &= r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots \\ v_2 &= r_2 + \gamma r_3 + \gamma^2 r_4 + \cdots \\ v_3 &= r_3 + \gamma r_4 + \gamma^2 r_1 + \cdots \\ v_4 &= r_4 + \gamma r_1 + \gamma^2 r_2 + \cdots \end{aligned}$$

Policy and Value Functions

How to evaluate a policy?

Let v_i be the value of state s_i under policy π.



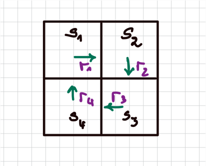
Rewriting the equations:

$$
\begin{aligned}
v_1 &= r_1 + \gamma\left(r_2 + \gamma r_3 + \cdots\right) &= r_1 + \gamma v_2 \\
v_2 &= r_2 + \gamma\left(r_3 + \gamma r_4 + \cdots\right) &= r_2 + \gamma v_3 \\
v_3 &= r_3 + \gamma\left(r_4 + \gamma r_1 + \cdots\right) &= r_3 + \gamma v_4 \\
v_4 &= r_4 + \gamma\left(r_1 + \gamma r_2 + \cdots\right) &= r_4 + \gamma v_1
\end{aligned}
$$

Policy and Value Functions

How to evaluate a policy?

Let v_i be the value of state s_i under policy π.



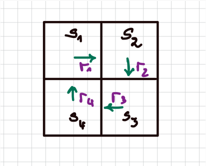
Rewriting the equations in a matrix form:

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} + \begin{bmatrix} \gamma v_2 \\ \gamma v_3 \\ \gamma v_4 \\ \gamma v_1 \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} + \gamma \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}$$

Policy and Value Functions

How to evaluate a policy?

Let v_i be the value of state s_i under policy π.



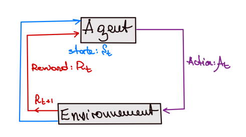
Wich can be written as:

$$v = r + \gamma P v$$

$\rightarrow$ this is the Bellman equation

Policy and Value Functions

More formally, back to the schema of RL:

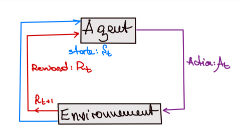


We have the following notations and random variables:

- t: time step
- S_t: state at time t
- A_t: action at time t at state S_t
- R_{t+1}: reward at time $t+1$ after taking action A_t at state S_t
- S_{t+1}: state at time $t+1$ after taking action A_t at state S_t

Policy and Value Functions

More formally, back to the schema of RL:



The steps are determined by the following distributions (we assume we know them):

- $S_t \rightarrow A_t$ by $\pi(A_t = a|S_t = s)$
- $S_t, A_t \rightarrow S_{t+1}$ by $P(S_{t+1} = s'|S_t = s, A_t = a)$
- $S_t, A_t \rightarrow R_{t+1}$ by $p(R_{t+1} = r|S_t = s, A_t = a)$

Policy and Value Functions

Consider a trajectory of states, actions and rewards (described by the r.v. above):

$$S_t \xrightarrow{A_t} S_{t+1}, R_{t+1} \xrightarrow{A_{t+1}} S_{t+2}, R_{t+2} \xrightarrow{A_{t+2}} \longrightarrow \cdots$$

The discounted return is:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

The value function is the expected return:

$$v_\pi(s) = \mathbb{E}_\pi \left[G_t \mid S_t = s \right]$$

Policy and Value Functions

Definition:

The value function or state-value function $v_\pi(s)$ is defined as:

$$v_\pi(s) = \mathbb{E}_\pi\left[G_t \mid S_t = s\right] = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s\right]$$

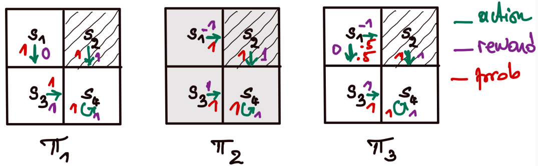
Policy and Value Functions

Remarks:

- It is a function of s. It is a conditional expectation with the condition that the state starts from s.
- It is based on the policy π. For a different policy, the state value may be different.
- If the policy, the transition function and the reward function are all deterministic, then the value function is simply the return, i.e., the sum of the rewards along the trajectory.

Policy and Value Functions

Back to our Example



Rewriting the equations:

$$
\begin{aligned}
v_{\pi_1}(s_1) &= 0 + \gamma + \gamma^2 + \cdots = \tfrac{\gamma}{1-\gamma} \\
v_{\pi_2}(s_1) &= -1 + \gamma + \gamma^2 + \cdots = -1 + \tfrac{\gamma}{1-\gamma} \\
v_{\pi_3}(s_1) &= 0.5\left(-1 + \tfrac{\gamma}{1-\gamma}\right) + 0.5\left(\tfrac{\gamma}{1-\gamma}\right) = -0.5 + \tfrac{\gamma}{1-\gamma}
\end{aligned}
$$

Policy and q-value functions

Intuition and Definition: Similar to the value function, the action-value function or q-value function caracterizes the value of taking an action in a state under a policy.
It is the expected return starting from state s, taking action a, and then following policy π:

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}_\pi\left[G_t \mid S_t = s, A_t = a\right] \\ &= \sum_r P(r \mid s,a) r + \gamma \sum_{s'} P(s' \mid s,a) v_\pi(s') \end{aligned}$$

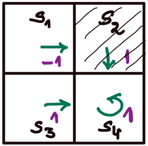
Policy state and q-value functions

Let's rewrite the equation for th value function, considering the action taken at time t:

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi \left[G_t \mid S_t = s \right] \\
&= \textstyle\sum_a \pi(a|s) \mathbb{E}_\pi \left[G_t \mid S_t = s, A_t = a \right] \\
&= \textstyle\sum_a \pi(a|s) q_\pi(s, a)
\end{aligned}
$$

Policy and q-value functions

Example:



$$
\begin{aligned}
q_\pi(s_1, a_1) &= -1 + \gamma v_\pi(s_1) \\
q_\pi(s_1, a_2) &= -1 + \gamma v_\pi(s_2) \\
q_\pi(s_1, a_3) &= 0 + \gamma v_\pi(s_3) \\
q_\pi(s_1, a_4) &= -1 + \gamma v_\pi(s_1) \\
q_\pi(s_1, a_5) &= 0 + \gamma v_\pi(s_1)
\end{aligned}
$$

Summary

- A Markov Decision Process (MDP) is defined as a tuple:

$$M = (S, A, P, R, \gamma)$$

- The value function

$$v_\pi(s) = \mathbb{E}_\pi \left[G_t \mid S_t = s \right]$$

  is the expected return starting from state s under policy π.

- The action-value function

$$q_\pi(s, a) = \mathbb{E}_\pi \left[G_t \mid S_t = s, A_t = a \right]$$

  is the expected return starting from state s, taking action a, and then following policy π.

Summary

- The Bellman equation is a recursive equation that caracterizes the value function:

$$
\begin{aligned}
v_\pi(s) &= \textstyle\sum_a \pi(a|s) q_\pi(s,a) \\
&= \textstyle\sum_a \pi(a|s) \left(\sum_r P(r \mid s,a) r + \gamma \sum_{s'} P(s' \mid s,a) v_\pi(s')\right)
\end{aligned}
$$

- The Bellman equation in matrix form is:

$$v_\pi = r\pi + \gamma P \pi v \pi$$

- How to solve the Bellman equation? See the next lecture.