How to make decisions?
- Markov Decision Processes -

November 20, 2024

1 MDP: theory

Definition 1.1 (MDP)**.** An MDP is specified by

- A set of states $\mathcal{S}$ (knonw as state space), e.g. $\{1, 2, \cdots, n\}$ or $\{x_1, x_2, \cdots\}$
- A set of actions $\mathcal{A}$ (known as action space), e.g. $\{1, 2, \cdots, m\}$ or $\{a_1, a_2, \cdots\}$
- - A time-indexed sequence of environment-generated random variables (r.v.s.) (state) S_t taking values in $\mathcal{S}$, $t = 0, 1, \cdots$,

  - a time-indexed sequence of environment-generated r.v.s. (reward) R_t, taking values in some space $\mathcal{R}$, $t = 0, 1, \cdots$, and

  - a time-indexed sequence of agent-controllable r.v.s. (action) A_t taking values in $\mathcal{A}$, $t = 0, 1, \cdots$
- Markov property:

$$\begin{aligned}&\mathbb{P}((R_{t+1}, S_{t+1}) \in B | S_t = s_t, A_t = a_t, S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \cdots, S_0 = s_0, A_0 = a_0)\\ =\ &\mathbb{P}((R_{t+1}, S_{t+1}) \in B | S_t = s_t, A_t = a_t).\end{aligned}$$

  With Markov property, the transition probabilities are defined by[1]

$$P(x'|x, a) = \mathbb{P}(\text{Next state } = x' | \text{Action } a \text{ in the current state } x)$$

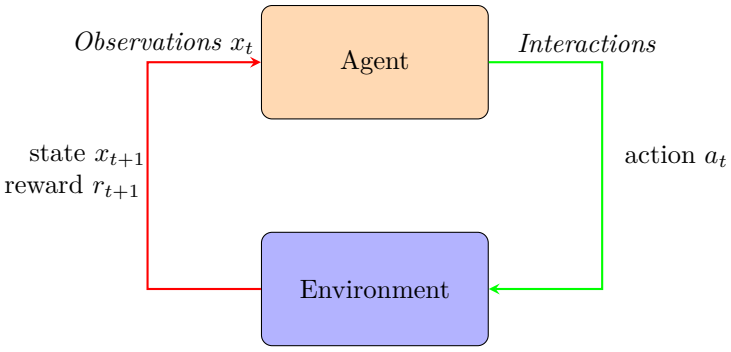
Remark 1.2.
- - The reward R_t is typically a function of state and action, $R_t = r(x, a)$ if $X_t = x$ and $A_t = a$.

  - Reward can be random with mean $r(x, a)$.

  - Reward can depend on x only or (x, a, x') as well.
- The mechanism for the agent to make decisions: at each time t, the agent observe the state x_t, after which the agent performs action a_t, after which the environment (upon seeing x_t and a_t) produces a random pair x_{t+1} and r_{t+1}, after which the agent observes his next state x_{t+1} and the cycle repeats.

  The agents wants to choose actions to maximize the expected accumulated reward.

[1] We concentrate on time-homogeneous MDP. That is, the transition probabilities are independent of t.



- For now, let us assume r and P are known to the decision maker/agent.

Example 1.3 (Example of MDP: inventory management)**.** Each state $s \in \mathcal{S}$ represents the an inventory level. For example, s_t could be the number of units available at time t.

Actions are the quantities of products ordered: $S_{t+1} = S_t + A_t + \epsilon_{t+1}$, where ϵ is a shock with independent components, and with distribution $\mathbb{P}(\epsilon_t = k) := p_t(k)$.

The reward function reflects the cost structure and is defined as:

$$r(s, a) = -C_o \cdot a - C_h \cdot \max(0, s + a - D) - C_p \cdot \max(0, D - (s + a))$$

where:

- C_o is the ordering cost per unit.
- C_h is the holding cost for excess inventory.
- C_p is the penalty cost for unmet demand.
- D is the demand.

Transition probabilities describe the likelihood of moving from state s to state s' after taking action a.

Exercise. Calculate the transition probability $P(x'|x, a)$.

The goal is to minimizes the expected total cost over a horizon T. This is done by maximizing the expected sum of rewards:

$$\max_A \sum_{t=0}^{T} \mathbb{E}[r(S_t, A_t)]$$

subject to the transition dynamics.

Definition 1.4 (Deterministic policy)**.** A deterministic policy is a mapping from the state space to the action space, i.e. $\pi : x \in \mathcal{S} \mapsto \pi(x) \in \mathcal{A}$, where $\pi(x)$ is the action recommended by the policy π for the state x. Denote by $\mathcal{A}(x)$ the set of all actions available at the state x, i.e. $\mathcal{A}(x) = \{\pi(x) : \pi \text{ is a deterministic policy}\}$.

Definition 1.5 (Randomized policy[2])**.** A *randomized* policy is a mapping $\pi : X \to \mathcal{P}(\mathcal{A})$, where $\mathcal{P}(\mathcal{A})$ is the set of probabilities on the action space $\mathcal{A}$.

Remark 1.6. Each deterministic policy corresponds to a randomized policy.

Definition 1.7 (Policy value or value function)**.** The value of a randomized policy π at state $x \in \mathcal{S}$ is defined as the expected reward returned when starting at x and following the policy π:

- finite horizon: $V^\pi(x) = \mathbb{E}_{a_t \sim \pi(x_t)} \left[\sum_{t=0}^{T} r(x_t, a_t) | x_0 = x \right]$,
- infinite horizon: $V^\pi(x) = \mathbb{E}_{a_t \sim \pi(x_t)} [\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) | x_0 = x]$,

where the discount rate $0 < \gamma < 1$, and the expectations are taken over the random selection of an action a_t according to the distribution $\pi(x_t)$, and over the random states x_t reached and the reward values $r(x_t, a_t)$.[3]

Remark 1.8.
- How to understand $\mathbb{E}_{a_t \sim \pi(x_t)}$?
- What if π is a deterministic policy?

Question 1: How to get policy value?

Lemma 1.9 (Bellman equations)**.** *The (expected) value for a randomized policy π satisfies*

$$V^\pi(x) = \mathbb{E}_{a \sim \pi(x)} [r(x, a)] + \gamma \sum_{x'} \mathbb{E}_{a \sim \pi(x)} [P(x'|x, a)] V^\pi(x').$$

Proof.

□

Remark 1.10. If π is a deterministic policy, then

$$V^\pi(x) = r(x, \pi(x)) + \gamma \sum_{x'} P(x'|x, \pi(x)) V^\pi(x').$$

[2]It is also called relaxed control is the area of stochastic control, and mixed strategy in game theory. It is a popular strategy in economics.

[3]More generally, the randomization of the reward function and the next state will not be considered for simplicity.

Proof. The proof is similar and left as an exercise. □

With the above recursive formula, we can solve the value of a policy in a *finite state MDP* as follows: let

$$\mathbf{V}^\pi = \begin{pmatrix} V^\pi(1) \\ \vdots \\ V^\pi(n) \end{pmatrix}, \quad \mathbf{R}^\pi = \begin{pmatrix} \mathbb{E}_{a\sim\pi(1)}[r(1,a)] \\ \vdots \\ \mathbb{E}_{a\sim\pi(n)}[r(n,a)] \end{pmatrix}, \quad \mathbf{\Gamma}^\pi = \begin{pmatrix} \mathbb{E}_{a\sim\pi(1)}[P(1|1,a)] & \cdots & \mathbb{E}_{a\sim\pi(1)}[P(n|1,a)] \\ \vdots & \ddots & \vdots \\ \mathbb{E}_{a\sim\pi(n)}[P(1|n,a)] & \cdots & \mathbb{E}_{a\sim\pi(n)}[P(n|n,a)]. \end{pmatrix}$$

Then,

$$\mathbf{V}^\pi = \mathbf{R}^\pi + \gamma\mathbf{\Gamma}^\pi\mathbf{V}^\pi \Rightarrow (I - \gamma\mathbf{\Gamma}^\pi)\mathbf{V}^\pi = \mathbf{R}^\pi \Rightarrow \mathbf{V}^\pi = (I - \gamma\mathbf{\Gamma}^\pi)^{-1}\mathbf{R}^\pi,$$

where the matrix $I - \gamma\mathbf{\Gamma}^\pi$ is invertible (why?)

Example 1.11 (Calculation of value functions)**.** Transition rule/probability of the agent in the 3×4 world: the probability of moving to the correct direction is 0.8; the probability of moving at the right angle to the correct direction is 0.2; never going back; a collision with a wall results in no movement. For example, consider the policy $\uparrow$ at the position/state $(1, 1)$, then

Rewarding rule: The two terminal states have reward $+1$ and -1, respectively, and all other states have a reward of -0.04.

Table 1: One policy in the 3×4 world

3	$\rightarrow$	$\rightarrow$	$\rightarrow$	$\boxed{+1}$
2	$\uparrow$		$\uparrow$	$\boxed{-1}$
1	$\uparrow$	$\leftarrow$	$\leftarrow$	$\leftarrow$
	1	2	3	4

Let $\pi = \uparrow$. Given $V^\pi(1, 2) = 0.762$ and $V^\pi(2, 1) = 0.655$, compute $V^\pi(1, 1)$.

Table 2: Value functions of the policy in the above table

1	0.812	0.868	0.918	+1
2	0.762		0.660	-1
3	0.705	0.655	0.611	0.388
	1	2	3	4

The goal of the agent is to find a policy with a higher return based on his/her observations.

- Planning problem: the agent knows the environment, by given a fully specified MDP, e.g. playing chess. ✔
- Learning problem: the environment is unknown to the agent (agent interacts with an environment with unknown dynamics), e.g. driving a car.

Example 1.12. A robot navigates a 2×2 world. It can move *up*, *down*, *left*, or *right*. The transition rule follows Example 1.11. If it reaches the top-right corner (goal), it receives a reward of +10 and the process ends, and receives nothing at other states.

1. What are the states ($\mathcal{S}$) and actions ($\mathcal{A}$) in this MDP?
2. Define the reward function $r(s, a)$ for each state-action pair.
3. What will be the transition probabilities $P(s'|s, a)$?

Solution: