Markov Decission Processes

Reinforcement Learning

Daniel Parra

24-Dic-2020

Markov Property

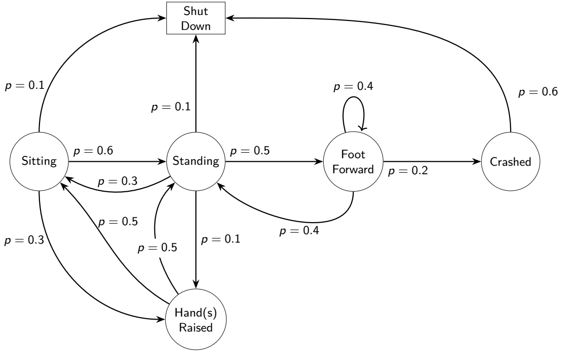
A state S_t, is Markov if and only if:

$$\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, \ldots, S_t] \tag{1}$$

Definition

Formally, for a state S_t to be **Markov** the probability of the next state S_{t+1} being s' should only be dependent on the current state $S_t = s_t$, and not on the rest of the past states.

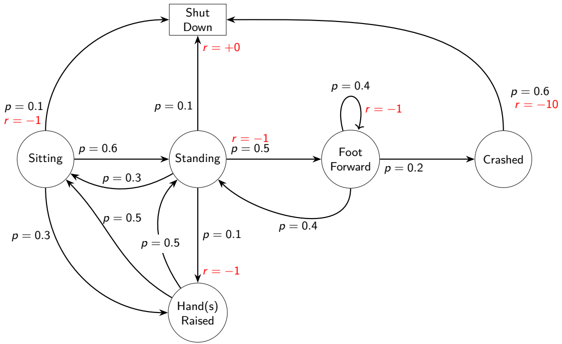
Markov Process



A **Markov Process** or **Markov Chains** is composed by states and probabilities of transition $p(s'|s)$ defined as:

$$p(s'|s) = \mathbb{P}\left\{S_t = s'|S_{t-1} = s\right\} \tag{2}$$

Markov Reward Process (MRP)

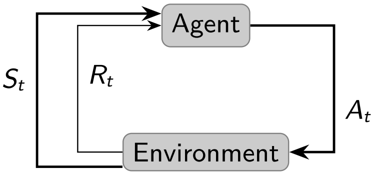


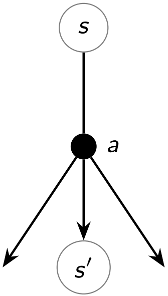
Markov Reward Process (MRP)
An MRP is defined by $(s, p(s'|s), r(s), \gamma)$, where s are states, $p(s'|s)$ is the state-transition probability, $r(s)$ is the reward given s, and γ is the discount factor.

$$p(s'|s) = \mathbb{P}\left\{S_t = s' | S_{t-1} = s\right\} \tag{3}$$

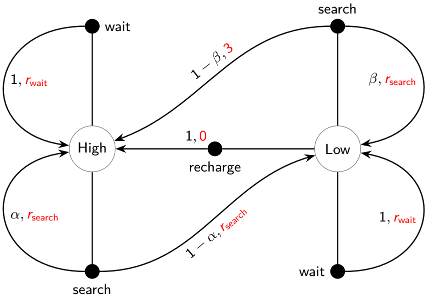
$$r(s) = \mathbb{E}\left[R_t | S_{t-1} = s\right] \tag{4}$$

Markov Decision Process





Mathematical Formulation



Dynamics Function

$$p(s', r|s, a) \doteq \mathbb{P}\left\{S_t = s', R_t = r | S_{t-1} = s, A_{t-1} = a\right\} \tag{5}$$

$$\sum_{s' \in S} \sum_{r \in R} p(s', r|s, a) = 1 \tag{6}$$

Some probabilities

State Transition Probabilities

$$
\begin{aligned}
p(s'|s,a) &\doteq \mathbb{P}\left\{ S_t = s' \,\middle|\, S_{t-1} = s, A_{t-1} = a \right\} \\
&= \sum_{r \in R} p(s', r | s, a)
\end{aligned} \tag{7}
$$

Some probabilities

Expected rewards for State-Action Pairs

$$
\begin{aligned}
r(s,a) &\doteq \mathbb{E}\left[R_t | S_{t-1}=s, A_{t-1}=a\right] \\
&= \sum_{r \in R} r \cdot \mathbb{P}\left\{R_t = r | S_{t-1}=s, A_{t-1}=a\right\} \\
&= \sum_{r \in R} r \cdot \sum_{s' \in S} \mathbb{P}\left\{R_t = r, S_t = s' | S_{t-1}=s, A_{t-1}=a\right\} \\
&= \sum_{r \in R} r \cdot \sum_{s' \in S} p(s', r | s, a)
\end{aligned} \tag{8}
$$

Some probabilities

Expected rewards for State-Action-Next-State Triple

$$\begin{aligned} r(s, a, s') &\doteq \mathbb{E}\left[R_t | S_{t-1} = s, A_{t-1} = a, S_t = s'\right] \\ &= \sum_{r \in R} r \cdot \mathbb{P}\left\{R_t = r | S_{t-1} = s, A_{t-1} = a, S_t = s'\right\} \\ &= \sum_{r \in R} r \cdot p(r | s, a, s') \end{aligned} \tag{9}$$

Product Rule applied on Dynamics Function

$$p(s', r | s, a) = p(s' | s, a) \cdot p(r | s, a, s') \tag{10}$$

$$r(s, a, s') = \sum_{r \in R} r \cdot p(r | s, a, s') = \sum_{r \in R} r \cdot \frac{p(s', r | s, a)}{p(s' | s, a)} \tag{11}$$

Expected Return

Formal Definition

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T \tag{12}$$

Discount

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^K R_{t+k+1} \tag{13}$$

Episodic Tasks

$$G_T \doteq \sum_{K=t+1}^{T} \gamma^{K-t-1} R_K \tag{14}$$

Policies

State-Value Function for Policy π

$$v_\pi \doteq \mathbb{E}_\pi \left[G_t | S_t = s \right] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^K R_{t+k+1} \middle| S_t = s \right], \quad \forall s \in S \tag{15}$$

Action-Value Function for Policy π

$$q_\pi \doteq \mathbb{E}_\pi \left[G_t | S_t = s, A_t = a \right] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^K R_{t+k+1} \middle| S_t = s, A_t = a \right] \tag{16}$$

Policies

Give an equation for v_π in terms of q_π and π

$$\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi\left[G_t | S_t = s\right] \\
&= \sum_{g_t} p(g_t | S_t = s) g_t \\
&= \sum_{g_t} \sum_{a} p(g_t, a | S_t = s) g_t \\
&= \sum_{a} p(a | S_t = s) \times \sum_{g_t} p(g_t | S_t = s, A_t = a) g_t \\
&= \sum_{a} p(a | S_t = s) \times \mathbb{E}_\pi\left[G_t | S_t = s, A_t = a\right] \\
&= \sum_{a} \underbrace{\pi(a|s)}_{\substack{\text{Prob. to take} \\ \text{action i.e. policy}}} \times \underbrace{q_\pi(s, a)}_{\substack{\text{Action-Value} \\ \text{Function}}}
\end{aligned} \tag{17}$$

Policies

Give an equation for q_π in terms of v_π and the four argument p

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi\left[G_t | S_t = s, A_t = a\right] \\
&= \sum_{s',r,g_t} p(s',r,g_t|s,a)\cdot g_t \\
&= \sum_{s',r,g_t} p(s',r,g_t|s,a)\cdot [r+\gamma g_{t+1}] \\
&= \sum_{s',r} p(s',r|s,a) \sum_{g_t} p(g_t|s,a)\cdot [r+\gamma g_{t+1}] \\
&= \sum_{s',r} p(s',r|s,a)\left[r+\gamma\sum_{g_t} p(g_t|s')g_t\right] \\
&= \sum_{s',r} p(s',r|s,a)\left[r+\gamma\mathbb{E}_\pi\left[G_t|S_t=s'\right]\right] \\
&= \sum p(s',r|s,a)\left[r+\gamma v_\pi(s')\right]
\end{aligned}
\tag{18}
$$

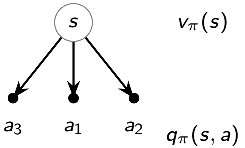
Bellman Equation Derivation

$$
\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi\left[G_t | S_t = s\right] \\
&= \mathbb{E}_\pi\left[R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{(t+1)+k+1} \middle| S_t = s \right] \\
&= \mathbb{E}_\pi\left[R_{t+1} + \gamma G_{t+1} | \, S_t = s\right] \\
&= \mathbb{E}_\pi\left[R_{t+1} + \gamma \mathbb{E}_\pi\left[G_{t+1} | S_{t+1}\right] | \, S_t = s\right] \\
&= \mathbb{E}_\pi\left[R_{t+1} + \gamma v_\pi(S_{t+1}) | \, S_t = s\right] \\
&= \sum_{s',a,r} p(a, r, s' | s) \left[r + \gamma v_\pi(s') \right] = \sum_a \pi(a|s) \sum_{r,s'} p(r|s,a) p(s'|s,a) \left[r + \gamma v_\pi(s') \right] \\
&= \sum_a \pi(a|s) \sum_r p(r|s,a) r + \gamma \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) v_\pi(s') \\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(r, s'|s,a) r + \gamma \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s,a) v_\pi(s') \\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(r, s'|s,a) \left[r + \gamma v_\pi(s') \right]
\end{aligned} \tag{19}
$$

Tomado de: `https://jmichaux.github.io/_notebook/2018-10-14-bellman/`

Bellman Equation

Exercise 3.18 The value of a state depends on the values of the actions possible in that state and on how likely each action is to be taken under the current policy. We can think of this in terms of a small backup diagram rooted at the state and considering each possible action:



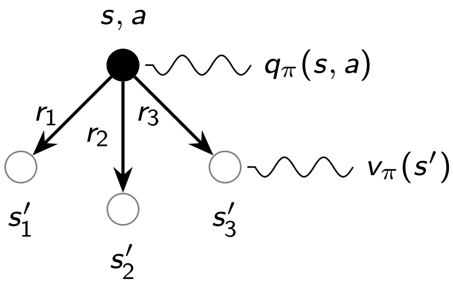
Give the equation corresponding to this intuition and diagram for the value at the root node, $v_\pi(s)$, in terms of the value at the expected leaf node, $q_\pi(s,a)$, given $S_t = s$.

$$\begin{aligned} v_\pi(s) &= \sum_a \pi(a|s) q_\pi(s,a) \\ &= \mathbb{E}_\pi \left[q_\pi(s,a) | s \right] \end{aligned} \tag{20}$$

Then give a second equation in which the expected value is written out explicitly in terms of $\pi(a|s)$ such that no expected value notation appears in the equation.

Bellman Equation

Exercise 3.19: The value of an action, $q_\pi(s, a)$, depends on the expected next reward and the expected sum of the remaining rewards. Again we can think of this in terms of a small backup diagram, this one rooted at an action (state–action pair) and branching to the possible next states:



Optimal Policies and Optimal Value Functions

There is always at least one policy that is better than or equal to all other policies, *optimal policy* (π_*). There may be more than one optimal policy, and they share the same *state-value function* (v_*).

$$v_*(s) \doteq \max_\pi v_\pi(s) \quad \forall s \in S \tag{21}$$

Optimal policies also share the same *optimal action-value function*, denoted q_*, and defined as:

$$\begin{aligned} q_*(s,a) &\doteq \max_\pi q_\pi(s,a) \;\; \forall s \in S \text{ and } a \in A(s) \\ &= \mathbb{E}\left[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a\right] \end{aligned} \tag{22}$$

Optimal Value Function

$$
\begin{aligned}
v_*(s) &= \max_{a \in A(s)} q_{\pi*}(s,a) \\
&= \max_a \mathbb{E}_{\pi*}\left[G_t | S_t = s, A_t = a\right] \\
&= \max_a \mathbb{E}_{\pi*}\left[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a\right] \\
&= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] \\
&= \max_a \sum_{s',r} p(s', r | s, a) \left[r + \gamma v_*(s')\right]
\end{aligned} \tag{23}
$$

$$
\begin{aligned}
q_*(s,a) &= \mathbb{E}\Big[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') | S_t = s, A_t = a\Big] \\
&= \sum_{s',r} p(s', r | s, a) \left[r + \gamma \max_{a'} q_*(s', a')\right]
\end{aligned} \tag{24}
$$

Practical Quizz Value Functions and Bellman Equations

Q1: A policy is a function which maps States to probability distributions over actions.
Q2: The term "backup" most closely resembles the term Update in meaning.
Q3: At least one deterministic optimal policy exists in every Markov decision process. True Let's say there is a policy π_1 which does well in some states, while policy π_2 does well in others. We could combine these policies into a third policy π_3, which always chooses actions according to whichever of policy π_1 and π_2 has the highest value in the current state. π_3 will necessarily have a value greater than or equal to both π_1 and π_2 in every state! So we will never have a situation where doing well in one state requires sacrificing value in another. Because of this, there always exists some policy which is best in every state. This is of course only an informal argument, but there is in fact a rigorous proof showing that there must always exist at least one optimal deterministic policy.

Practical Quizz Value Functions and Bellman Equations

Q4: The optimal state-value function: Is unique in every finite Markov decision process. The Bellman optimality equation is actually a system of equations, one for each state, so if there are N states, then there are N equations in N unknowns. If the dynamics of the environment are known, then in principle one can solve this system of equations for the optimal value function using any one of a variety of methods for solving systems of nonlinear equations. All optimal policies share the same optimal state-value function.

Q5: Does adding a constant to all rewards change the set of optimal policies in episodic tasks? Yes, adding a constant to all rewards changes the set of optimal policies. Adding a constant to the reward signal can make longer episodes more or less advantageous (depending on whether the constant is positive or negative).

Q6: Does adding a constant to all rewards change the set of optimal policies in continuing tasks? No, as long as the relative differences between rewards remain the same, the set of optimal policies is the same. Since the task is continuing, the agent will accumulate the same amount of extra reward independent of its behavior.

Practical Quizz Value Functions and Bellman Equations

Q7: Select the equation that correctly relates v_* to q_*. Assume π is the uniform random policy.

$$v_* = \max_a q_*(s, a) \tag{25}$$

Q8: Select the equation that correctly relates q_* to v_*, using four argument function p.

$$q_*(s, a) = \sum_{s', r} p(s', r|a, s) \left[r + \gamma v_*(s') \right] \tag{26}$$

Q9: Write a policy π_* in terms of q_*

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \max_{a'} q_*(s, a') \\ 0 & \text{else} \end{cases} \tag{27}$$

The probability of taking an action is constrained between 0 and 1. The value of an action can be arbitrary.

Q10: Give an equation for some π_* in terms of v_* and the four argument p.

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } v_*(s) = \sum_{r,s'} p(s', r|s, a) \left[r + \gamma v_*(s') \right] \\ 0 & \text{else} \end{cases} \tag{28}$$

Second Quiz

Q1: A function which maps ___ to ___ is a value function.

- State-action pairs to expected returns.
- States to expected returns.

Q2: Consider the continuing Markov decision process shown below. The only decision to be made is in the top state, where two actions are available, left and right. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, π_{left} and π_{right}

Q3: Every finite Markov decision process has:

- A deterministic optimal policy
- A unique optimal value function

Q4: The ___ of the reward for each state-action pair, the dynamics function p, and the policy π is ___ to characterize the value function v_π. Mean; sufficient

Q5: The Bellman equation for a given a policy π:

- Expresses state values $v(s)$ in terms of state values of successor states.

Second Quiz

Q6: An optimal policy: Is not guaranteed to be unique, even in finite Markov decision processes.

Q7: The Bellman optimality equation for v_π:

- Holds for the optimal state value function.
- Expresses state values $v_*(s)$ in terms of state values of successor states.
- ~~Holds when $v_* = v_\pi$ for a give policy π~~
- ~~Expresses the improved policy in terms of the existing policy~~
- ~~Holds when the policy is greedy with respect to the value function.~~

Q8: Give an equation for v_π in terms of q_π and π

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s, a) \tag{29}$$

Q9: Give an equation for q_π, in terms of v_π and the four argument p

$$q_\pi(s, a) = \sum_{s', r} p(s', r|s, a) \left[r + \gamma v_\pi(s') \right] \tag{30}$$

Second Quizz

Q10: Let $r(s, a)$ be the expected reward for taking action a in state s. Which of the following are valid ways to re-express the Bellman equations, using the expected reward function?

$$
\begin{aligned}
q_\pi(s, a) &= r(s, a) + \gamma \sum_{s', a} p(s'|s, a)\pi(a'|s')q_\pi(s', a') \\
q_*(s, a) &= r(s, a) + \gamma \sum_{s'} p(s'|s, a) \max_{a'} q_*(s', a') \\
v_\pi(s) &= \sum_a \pi(a|s) \left[r(s, a) + \gamma \sum_{s'} p(s'|s, a)v_\pi(s') \right] \\
v_*(s) &= \max_a \left[r(s, a) + \gamma \sum_{s'} p(s'|s, a)v_*(s') \right]
\end{aligned}
\tag{31}
$$

Q11: Consider an episodic MDP with one state and two actions (left and right). The left action has stochastic reward 1 with probability p and 3 with probability $1 - p$. The right action has stochastic reward 0 with probability q and 10 with probability $1 - q$. What relationship between p and q makes the actions equally optimal?

$$7 + 2p = 10q \tag{32}$$