Chapter 3 - Finite Markov Decision Processes

Stéphane Liem NGUYEN

August 15, 2021

Some keywords, citations and formulas.

What are value functions ? Value functions tell us *how good* it is for the agent to be in a given state or how good it is for the agent to take an action from a given state. "How good" is related to the expected return, the expected cumulative future reward when the agent follows a particular behaviour, a policy.

For example, the optimal action-value function $q_*(s, a)$ is telling us what is the expected return if we start in state s, take action a then follows the optimal policy/strategy π_* afterwards.

Here are the formal definitions of the value functions under policy π; the first one is for the state-value function and the second is the action-value function.

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t | S_t = s] \tag{1}$$

$$q_\pi(s, a) \doteq \mathbb{E}_\pi[G_t | S_t = s, A_t = a] \tag{2}$$

with $G_t = R_{t+1} + \gamma G_{t+1}, \forall t < T$ and $G_T = 0$. The discount factor/rate γ is a scalar value in $[0, 1]$ that determines how farsighted is the agent. Is the agent taking into account more than the immediate reward for deciding which action to pick in a given state ?

The undiscounted case ($\gamma = 1$) is used most of the time for episodic tasks where there's a notion of a terminal state or final time step T (random variable). It can be used for continuing tasks but it's not covered in the Chapter so we can just keep in mind that the discounted case ($0 \leq \gamma < 1$) is used for continuing tasks to keep the return finite.

The value of the *terminal state* (or *absorbing state*. There's only one single terminal state with different possible rewards for different outcomes.) for the *episodic tasks* is 0.

Action-value function and Model-free When we do not want to select actions based on the knowledge of the environment dynamics, action-value functions can be used because they "cache the results of all one-step-ahead searches." (Upper part of the page 65). In other words, action-value functions are used in the *model-free* case where methods select actions without creating a model of the environment–without estimating the transition probabilities as well as the expected rewards based on real trajectories.

Bellman (expectation) equations for the four value functions Bellman equation for the state-value function for policy π

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma v_\pi(s')\right] \tag{3}$$

Bellman equation for the optimal state-value function (Bellman optimal equation for v_*)

$$v_*(s) = \max_a \sum_{s',r} p(s', r|s, a) \left[r + \gamma v_*(s')\right] \tag{4}$$

Bellman equation for the action-value function for policy π

$$q_\pi(s, a) = \sum_{s',r} p(s', r|s, a) \left[r + \gamma \sum_{a'} \pi(a'|s') q_\pi(s', a')\right] \tag{5}$$

Bellman equation for the optimal action-value function (Bellman optimal equation for q_*)

$$q_*(s, a) = \sum_{s',r} p(s', r|s, a) \left[r + \gamma \max_{a'} q_*(s', a')\right] \tag{6}$$

where $a \in \mathcal{A}(s)$, s' and $s \in \mathcal{S}$ and $r \in \mathcal{R}$

1 Citations of some parts of the book

1. p. 49: "In a Markov decision process, the probabilities given by p completely characterize the **environment's dynamics**. [...] The state must include *information about all aspects of the past* agent–environment interaction that make a *difference for the future.* If it does, then the state is said to have the **Markov property**. We will assume the Markov property throughout this book, though starting in Part II we will consider **approximation methods** that do not rely on it, and in Chapter 17 we consider how a **Markov state** can be efficiently **learned** and **constructed from non-Markov observations**."

2. p. 50: "In general, *actions can be any decisions we want to learn how to make*, and states can be anything we can know that might be useful in making them."

3. p. 50: "The general rule we follow is that **anything that cannot be changed arbitrarily** by the agent is considered to be outside of it and thus **part of its environment**."

4. p. 50: "The agent-environment **boundary** represents the limit of the agent's **absolute control, not of its knowledge**."

5. p. 53, *reward hypothesis* : "That all of what we mean by **goals** and purposes can be well thought of as the **maximization of the expected value** of the cumulative **sum** of a received scalar signal (called **reward**)."

6. p. 68: "The **online nature** of reinforcement learning makes it possible to **approximate optimal policies** in ways that put more effort into learning to make *good decisions* for **frequently encountered states**, at the expense of less effort for infrequently encountered states. This is one key property that distinguishes reinforcement learning from other approaches to approximately solving MDPs."