# Finite Markov Decision Processes

In this chapter we formalize the reinforcement learning problem, framed as a Markov Decision Process. Key aspects are

- Evaluative feedback
- Associative aspect (choosing different actions in different situations)
- Sequential decision making, with actions influencing future rewards and subsequent states

## 1.1 The agent-environment interface

The learner and decision maker is called the agent, and the thing that interacts with it, everything outside the agent, is the environment.

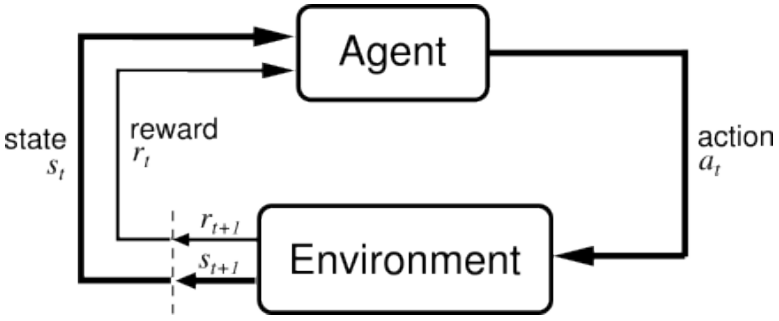


**Figure 1.** From book.

For simplicity[1], the agent and environment interact with each other at discrete timesteps. At each timestep, the agent receives some representation of the state $S_t \in \mathcal{S}$ and on that basis selects an action $A_t \in \mathcal{A}$. One step later, the agent gets a reward $R_{t+1} \in \mathbb{R}$. Thus we get the *trajectory*

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, S_3, A_3 \ldots.$$

1. Can be generalized to continuous case

**Definition 1.1.** In a **Finite MDP**, the set of possible states, actions, and rewards $\mathcal{A}, \mathcal{S},$ and $\mathcal{R}$ are all finite. Moreover, $R_t$ and $S_t$ have well defined probability distributions based on only the preceding state and action.

The probability of an action $s'$ and a reward $r$ occuring at a time $t$ is given by

$$p(s', r \mid s, a) \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}. \tag{1.1}$$

This is defined for all $s', s \in \mathcal{S}, r \in \mathcal{R}$ and $a \in \mathcal{A}$. Function $p$ defines the *dynamics* of the MDP, and

$$p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0, 1],$$

is an ordinary deterministic function.

*Remark* 1.1. In a *Markov* decision process, the probabilities given by $p$ completely characterize the environment's dynamics. That is, each possible value for $S_t$ and $R_t$ are uniquely determined by preceding state and action $S_{t-1}$ and $A_{t-1}$.

This is viewed as a restriction not on the decision process, by on the *state*. It must include all possible information of the past agent-environment interaction that impact the future.

**Definition 1.2.** A process is said to have the **Markov Property** if the state include all past agent-environment information that impacts the future.

## 1.1.1. Consequences of the dynamics function

From $p$ we can compute anything else we would like to know about the problem. Such as *transition-state probabilities* $p' : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$

$$p'(s' \mid s, a) \doteq \Pr\{S_t = s' \mid S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in \mathcal{R}} p(s', r \mid s, a). \tag{1.2}$$

Or the expected rewards for *state-action* pairs as a function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

$$r(s, a) \doteq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a] = \sum_{r \in \mathcal{R}} r \sum_{s \in \mathcal{S}} p(s', r \mid s, a). \tag{1.3}$$

Or the expected rewards for *state-action-next-state* triples as a function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

$$r(s, a, s') \doteq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in \mathcal{R}} r \frac{p(s', r \mid s, a)}{p(s' \mid s, a)}. \tag{1.4}$$

*Remark* 1.2.

- As a general rule anything that cannot be changed arbitrarily by an agent is considered part of the external environment.
- The reward computation is always considered external to the agent.
- The agent-environment boundary represents the limit of the agent's *absolute control*.