Reinforecment Learning: Notes and Selected Exercises

Somnath Sikdar

November 12, 2020

Contents

Chapter 1

Markov Decision Processes

1.1 Introduction

Reinforcement learning consists in teaching an agent how to behave using observations from its environment with the goal of maximizing future rewards that it receives in response to its behavior. The set-up consists of an *agent*, which is the learner or decision maker and the *enviroment*, which is the thing that the agent interacts with, comprising everything outside the agent. The agent and the environment interact in discrete time steps. At a given time step t, the agent takes an *action* and the environment responds by presenting a new state and a scalar feedback signal called the *reward*.

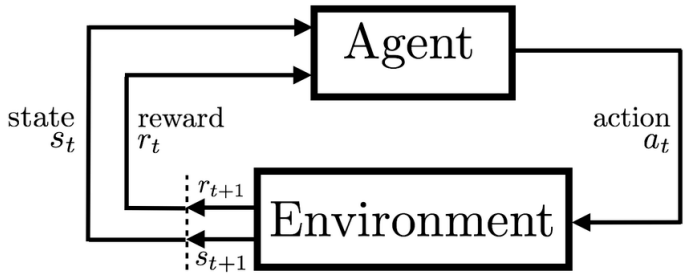

Figure 1.1: The agent and the environment

Problems in reinforcement learning are typically formulated in the formal setting of a Markov Decision Process (MDP). An MDP is a five-tuple $\langle \mathscr{S}, \mathscr{A}, p, \mathscr{R}, \gamma \rangle$

- a set $\mathscr{S}$ of states
- a set $\mathscr{A}$ of actions
- a set $\mathscr{R}$ of rewards
- a probability function $p\colon \mathscr{S} \times \mathscr{R} \times \mathscr{S} \times \mathscr{A} \to [0, 1]$
- a discount factor $\gamma \in [0, 1]$

The probablity function p defines the dynamics of the MDP. It tells us what the most likely next state and reward combination are given the current state and the action taken by the agent.

$$p(s', r \mid s, a) \coloneqq \mathbf{Pr}\left\{S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a\right\}. \tag{1.1}$$

In the above equation, we assume that the probability does *not* depend on the time step t. Such an MDP is called *stationary*. Since this is a probablity distribution, for all $s \in \mathscr{S}$ and for all $a \in \mathscr{A}$,

$$\sum_{s' \in \mathscr{S}} \sum_{r \in \mathscr{R}} p(s', r \mid s, a) = 1. \tag{1.2}$$

We also define a state-transition probability function $p\colon \mathscr{S} \times \mathscr{A} \times \mathscr{S} \to [0, 1]$ defined as:

$$p(s' \mid s, a) \coloneqq \sum_{r \in \mathscr{R}} p(s', r \mid s, a). \tag{1.3}$$

In response to the action taken by the agent, the environment provides it a scalar rewards at each time step. The total reward of the agent in time step t is defined as

$$G_t \coloneqq \sum_{k=0}^{\infty} R_{t+k+1}. \tag{1.4}$$

With infinite time horizons, this reward could go to infinity. To prevent this, one uses a discounting factor $\gamma \in [0, 1]$ and defines the total reward as

$$G_t \coloneqq R_{t+1} + \gamma \cdot R_{t+2} + \gamma^2 \cdot R_{t+3} + \cdots \tag{1.5}$$

1.1.1 Policy and Value Functions

A *policy* π is a mapping from states to probabilities of selecting each possible action.

$$\pi(a \mid s) \coloneqq \mathbf{Pr}\left\{A_t = a \mid S_t = s\right\}. \tag{1.6}$$

Note that this probability distribution does not depend on the time step t.

Exercise 1.1. If the current state is S_t, and actions are selected according to a stochastic policy π, then what is the expectation of R_{t+1} in terms of π and the four-argument function p?

Solution. Let us assume that the current state $S_t = s$. We may write $\mathbf{E}_\pi\left[R_{t+1} \mid S_t = s\right]$ as:

$$\begin{aligned}
\mathbf{E}_\pi\left[R_{t+1} \mid S_t = s\right] &= \sum_{r \in \mathscr{R}} r \cdot \mathbf{Pr}_\pi\left\{R_{t+1} = r \mid S_t = s\right\} \\
&= \sum_{r \in \mathscr{R}} r \cdot \sum_{a \in \mathscr{A}} \mathbf{Pr}_\pi\left\{R_{t+1} = r \mid S_t = s, A_t = a\right\} \cdot \mathbf{Pr}_\pi\left\{A_t = a \mid S_t = s\right\} \\
&= \sum_{r \in \mathscr{R}} r \cdot \sum_{a \in \mathscr{A}} \sum_{s' \in \mathscr{S}} p(s', r \mid s, r) \cdot \pi(a \mid s)
\end{aligned}$$

□

The *value function* of a state s under a policy π, denoted $v_\pi(s)$, is the expected return when starting in state s and following π thereafter. For an MDP, we may write the value function as:

$$v_\pi(s) \coloneqq \mathbf{E}_\pi \left[G_t \mid S_t = s \right] = \mathbf{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k \cdot R_{t+k+1} \;\middle|\; S_t = s \right]. \tag{1.7}$$

The function $v_\pi(s)$ is called the *state-value function* of the policy π.

We also define the value of taking an action a in state s under a policy π, denoted $q_\pi(s, a)$, as the expected return when starting in state s, taking action a and following π thereafter.

$$q_\pi(s, a) \coloneqq \mathbf{E}_\pi \left[G_t \mid S_t = s, A_t = a \right] = \mathbf{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k \cdot R_{t+k+1} \;\middle|\; S_t = s, A_t = a \right]. \tag{1.8}$$

The function $q_\pi(s, a)$ is the *action-value function* of the policy π.

Exercise 1.2. Give an equation for v_π in terms of q_π and π.

Solution. We may write $v_\pi(s)$ as:

$$\begin{aligned} v_\pi(s) &\coloneqq \mathbf{E}_\pi \left[G_t | S_t = s \right] \\ &= \sum_{a \in \mathcal{A}} \mathbf{E}_\pi \left[G_t | S_t = s, A_t = a \right] \cdot \mathbf{Pr}_\pi \left\{ A_t = a | S_t = a \right\} \\ &= \sum_{a \in \mathcal{A}} q_\pi(s, a) \pi(a|s). \end{aligned}$$

□

Exercise 1.3. Give an equation for $q_\pi(s, a)$ in terms of v_π and the four parameter function p.

Solution.

$$\begin{aligned} q_\pi(s, a) &\coloneqq \mathbf{E}_\pi \left[R_{t+1} + \gamma \cdot R_{t+2} + \gamma^2 \cdot R_{t+3} + \cdots | S_t = s, A_t = a \right] \\ &= \mathbf{E}_\pi \left[R_{t+1} | S_t = s, A_t = a \right] + \gamma \cdot \mathbf{E}_\pi \left[G_{t+1} | S_t = s, A_t = a \right] \\ &= \sum_{r \in \mathcal{R}} r \cdot \mathbf{Pr}_\pi \left\{ R_{t+1} = r | S_t = s, A_t = a \right\} + \\ &\qquad \gamma \cdot \sum_{s' \in \mathcal{S}} \mathbf{E}_\pi \left[G_{t+1} | S_{t+1} = s', S_t = s, A_t = a \right] \cdot \mathbf{Pr}_\pi \left\{ S_{t+1} = s' | S_t = s, A_t = a \right\} \\ &= \sum_{r \in \mathcal{R}} r \cdot \sum_{s' \in \mathcal{S}} p(s', r|s, a) + \gamma \cdot \sum_{s' \in \mathcal{S}} \mathbf{E}_\pi \left[G_{t+1} | S_{t+1} = s' \right] \cdot \sum_{r \in \mathcal{R}} p(s', r|s, a) \\ &= \sum_{r \in \mathcal{R}} \sum_{s' \in \mathcal{S}} \left[r + \gamma \cdot v_\pi(s') \right] p(s', r|s, a). \end{aligned}$$

Based on the last two exercises, we may write:

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \cdot \sum_{r \in \mathcal{R}} \sum_{s' \in \mathcal{S}} \left[r + \gamma \cdot v_\pi(s') \right] p(s', r|s, a). \tag{1.9}$$

Exercise 1.4. Give an equation for $q_\pi(s,a)$ in terms of the value $q_\pi(s',a')$ of the next state-action pair (s',a').

Solution. The probability that the next state-action pair is (s',a') given that the current state-action pair is (s,a) is:

$$\mathbf{Pr}_\pi\left\{S_{t+1}=s', A_{t+1}=a' | S_t=s, A_t=a\right\} = p(s'|s,a)\cdot\pi(a'|s').$$

Now,

$$\begin{aligned}
q_\pi(s,a) &= \mathbf{E}_\pi\left[G_t|S_t=s,A_t=a\right] \\
&= \mathbf{E}_\pi\left[R_{t+1}|S_t=s,A_t=a\right] + \gamma\cdot\mathbf{E}_\pi\left[G_{t+1}|S_t=s,A_t=a\right] \\
&= \sum_{s'\in\mathscr{S}}\sum_{r\in\mathscr{R}} r\cdot p(s',r|s,a) + \\
&\qquad \gamma\cdot\sum_{s'\in\mathscr{S}}\sum_{a'\in\mathscr{A}}\mathbf{E}_\pi\left[G_{t+1}|S_{t+1}=s',A_{t+1}=a',S_t=s,A_t=a\right]\cdot p(s'|s,a)\cdot\pi(a'|s') \\
&= \sum_{s'\in\mathscr{S}}\sum_{r\in\mathscr{R}} r\cdot p(s',r|s,a) + \gamma\cdot\sum_{s'\in\mathscr{S}}\sum_{a'\in\mathscr{A}} q_\pi(s',a')\cdot p(s'|s,a)\cdot\pi(a'|s')
\end{aligned}$$

1.1.2 Optimal Policies and Optimal Value Functions

A policy π is said to be better than or equal to a policy $\pi^\star$ if for all states $s\in\mathscr{S}$, $v_\pi(s)\geq v_{\pi^\star}(s)$. An optimal policy $\pi^\star$ is one which satisfies the condition:

$$v_{\pi^\star}(s) = \max_\pi v_\pi(s) \quad \forall\ s\in\mathscr{S} \tag{1.10}$$

In an MDP, one is guaranteed to have an optimal policy. Define $v_\star \coloneqq v_{\pi^\star}$. The following claim is easy to show.

Lemma 1.1. *Let $\pi^\star$ be an optimal policy. Then $q_{\pi^\star}(s,a)=\max_\pi q_\pi(s,a)$ for all $s\in\mathscr{S}$ and all $a\in\mathscr{A}$.*

Proof. As shown in Exercise 1.3, $\forall s\in\mathscr{S}$ and $\forall a\in\mathscr{A}(s)$

$$\begin{aligned}
q_{\pi^\star}(s,a) &= \sum_{r\in\mathscr{R}}\sum_{s'\in\mathscr{S}}\left[r+\gamma\cdot v_\star(s')\right]p(s',r|s,a) \\
&= \sum_{r\in\mathscr{R}}\sum_{s'\in\mathscr{S}}\left[r+\gamma\cdot\max_\pi v_\pi(s')\right]p(s',r|s,a) \\
&= \max_\pi\sum_{r\in\mathscr{R}}\sum_{s'\in\mathscr{S}}\left[r+\gamma\cdot v_\pi(s')\right]p(s',r|s,a) \\
&= \max_\pi q_\pi(s,a).
\end{aligned}$$

□

Since $v_\star$ is the optimal state value function, it must equal the expected return for the best action from any given state. Hence,

$$\begin{aligned} v_\star(s) &= \max_{a \in \mathscr{A}(s)} q_\star(s,a) \\ &= \max_{a \in \mathscr{A}(s)} \sum_{r \in \mathscr{R}} \sum_{s' \in \mathscr{S}} [r + \gamma \cdot v_\star(s')]\, p(s', r|s, a) \quad \text{(Exercise 1.3).} \end{aligned} \tag{1.11}$$

Similarly, we may start with the expression for $q_\star$ as in Exercise 1.3 and write:

$$\begin{aligned} q_\star(s,a) &= \sum_{r \in \mathscr{R}} \sum_{s' \in \mathscr{S}} [r + \gamma \cdot v_\star(s')]\, p(s', r|s, a) \\ &= \sum_{r \in \mathscr{R}} \sum_{s' \in \mathscr{S}} \left[r + \gamma \cdot \max_{a' \in \mathscr{A}(s')} q_\star(s', a') \right] p(s', r|s, a). \end{aligned} \tag{1.12}$$

Equations 1.11 and 1.12 are called the Bellman optimality equations for $v_\star$ and $q_\star$, respectively.