An Introduction to Reinforcement Learning

From theory to algorithms

Joon Kwon

November 28, 2023

Contents

1 Markov decision processes **4**
1.1 Definition 4
1.2 Policies 5
1.3 Induced probability distributions over histories 6
1.4 Value functions 8

Foreword

As of Fall 2023, this document contains lecture notes from a course given in Master 2 in *Université Paris–Saclay*. These are highly incomplete and constantly updated as the lectures are given.

Acknowledgements

These notes highly benefited from discussions with Sylvain Sorin, Erwan Le Pennec, the expertise of Jaouad Mourtada, and the encouragements from Liliane Bel.

Introduction

Reinforcement learning deals with problems where an agent sequentially interacts with a dynamic environement, which yields a sequence of rewards. We aim at finding the decision rule for the agent which yields the highest cumulative reward. We first study the case where characteristics of the environements are known, and then turn to techniques for dealing with unknown environements, which must then be progressively learnt through repeated interaction.

Reinforcement learning achives great success in various applications: super-human algorithm for Go, robitics, finance, protein structure prediction, to name a few. Because it is so sucessful in practice, many resources are practice-oriented.

In these lectures, we first aim at a very rigorous presentation of the basic notions and tools. These building blocks will then be used to define algorithms, and establish theoretical guarantees for some of them.

Chapter 1

Markov decision processes

This chapter presents some basic notations regading MDPs.

For a finite set I, we denote $\Delta(I)$ the corresponding unit simplex in $\mathbb{R}^I$:

$$\Delta(I) = \left\{ x \in \mathbb{R}^I_+,\ \sum_{i \in I} x_i = 1 \right\}$$

and interpret it as set the probability distributions over I. For $i \in I$, the corresponding Dirac measure is denoted δ_i.

1.1 Definition

Definition 1.1.1. A *finite Markov Decision Process* (MDP) is a 4-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$ where $\mathcal{S}, \mathcal{A}, \mathcal{R}$ are nonempty finite sets and $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{R} \to [0,1]$ is such that for all $s, a \in \mathcal{S} \times \mathcal{A}$,

$$\sum_{(r,s') \in \mathcal{R} \times \mathcal{S}} p(s, a, r, s') = 1.$$

The elements of $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{S}$ are respectively called *states*, *actions* and *rewards*. The following notation will be used:

$$p(r, s'|s, a) = p(s, a, r, s'), \quad (s, a, r, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{R} \times \mathcal{S}.$$

The knowledge of $\mathcal{S}$ and $\mathcal{A}$ is always assumed, but $\mathcal{R}$ and p may not be known, depending on the context.

From now on, we assume that a finite MDP is given.

Remark 1.1.2. For fixed values $(s, a) \in \mathcal{S} \times \mathcal{A}$, $p(s, a, \cdot\,)$ defines a probability distribution on $\mathcal{R} \times \mathcal{S}$, which justifies notation $p(\,\cdot\,|s, a)$.

Definition 1.1.3. Let $t \geqslant 1$. A *history of length t* is a finite sequence of the form

$$(s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, \ldots, s_{t-1}, a_{t-1}, r_t, s_t) \in (\mathcal{S} \times \mathcal{A} \times \mathcal{R})^t \times \mathcal{S}.$$

By convention, a history of length 0 is an element $s_0 \in \mathcal{S}$. $\mathcal{H}^{(t)}$ denotes the set of histories of length t and $\mathcal{H}^\infty = (\mathcal{S} \times \mathcal{A} \times \mathcal{R})^{\mathbb{N}}$ the set of infinite histories.

Remark 1.1.4. Histories of length t correspond to the information observed by the agent at step t before choosing its action.

1.2 Policies

We now define policies, which are the formalization of decision rules for the agent. We first consider general policies, which allow for random decisions, as well as decision rules that depend on all available information (from the beginning of the interaction to the present state).

Definition 1.2.1. A *policy* is a sequence of maps $\pi = (\pi_t)_{t \geqslant 0}$ where $\pi_t : \mathcal{H}^{(t)} \to \Delta(\mathcal{A})$. For each $t \geqslant 0$ and $h^{(t)} \in \mathcal{H}^{(t)}$, denote

$$\pi_t(a|h^{(t)}) := \pi_t(h^{(t)})_a.$$

Π denotes the set of all policies.

Definition 1.2.2. A policy $\pi = (\pi_t)_{t \geqslant 0}$ is

- *deterministic* if for each $t \geqslant 0$ and $h^{(t)} \in \mathcal{H}^{(t)}$, there exists $a \in \mathcal{A}$ such that $\pi_t(h^{(t)})$ is the Dirac distribution in a;
- *Markovian* if for each $t \geqslant 0$, π_t is constant in all its variables but the last: in other words for a fixed value $s_t \in \mathcal{S}$, the map $\pi_t(\,\cdot\,, s_t)$ is constant; π_t can then be represented as $\pi_t : \mathcal{S} \to \Delta(\mathcal{A})$;
- *stationary* if it is Markovian and if for all $t \geqslant 0$ and $(s_0, a_0, r_1, \dots, r_t, s_t) \in \mathcal{H}^{(t)}$,

  $$\pi_t(s_0, a_0, r_1, \dots, r_t, s_t) = \pi_0(s_t);$$

  π can then be represented as $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ and denoted $\pi(a|s) = \pi(s)_a$ for $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Denote Π_0 (resp. $\Pi_{0,d}$) the set of stationary policies (resp. stationary and deterministic policies). A stationary and deterministic policy can be represented as $\pi : \mathcal{S} \to \mathcal{A}$.

In the next chapter, we will establish that there exists a stationary and deterministic optimal policy, and focus on stationary policies. We will however continue working with non-deterministic strategies, as they will later prove handy for *exploring* an unknown environement.

1.3 Induced probability distributions over histories

As soon as an MDP, a policy π, and an initial state distribution μ are given, the interaction produces random variables $S_0, A_0, R_1, S_1, A_0, R_2, \ldots$. This is formalized by the following proposition.

Proposition 1.3.1. *Let $\mu \in \Delta(\mathcal{S})$ and a policy π. There exists a unique probability measure $\mathbb{P}_{\mu,\pi}$ on $\mathcal{H}^\infty = (\mathcal{S} \times \mathcal{A} \times \mathcal{R})^{\mathbb{N}}$ (equipped with the product sigma-algebra) such that for all $T \geqslant 0$, $a_0, \ldots, a_T \in \mathcal{A}$, $s_0, \ldots, s_{T+1} \in \mathcal{S}$, and $r_1, \ldots, r_{T+1} \in \mathcal{R}$,*

$$\mathbb{P}_{\mu,\pi}\left(\prod_{t=0}^{T}(\{s_t\} \times \{a_t\} \times \{r_{t+1}\}) \times \{s_{T+1}\} \times (\mathcal{S} \times \mathcal{A} \times \mathcal{R})^{\mathbb{N}}\right)$$

$$= \mu(s_0)\prod_{t=0}^{T}\pi_t(a_t|h^{(t)})p(r_{t+1}, s_{t+1}|s_t, a_t).$$

where for each $1 \leqslant t \leqslant T$, $h^{(t)} = (s_0, a_0, r_1, \ldots, s_{t-1}, a_{t-1}, r_t, s_t)$.

Sketch of proof. The above expression defines a value for each set of the form

$$\prod_{t=0}^{T}(\{s_t\} \times \{a_t\} \times \{r_{t+1}\}) \times \{s_{T+1}\} \times (\mathcal{A} \times \mathcal{R} \times \mathcal{S})^{\mathbb{N}}.$$

The map $\mathbb{P}_{\mu,\pi}$ can then be extended to so-called cylinder sets of the form

$$\prod_{t=0}^{T}(\mathcal{S}_t \times \mathcal{A}_t \times \mathcal{R}_{t+1}) \times \mathcal{S}_{T+1} \times (\mathcal{A} \times \mathcal{R} \times \mathcal{S})^{\mathbb{N}},$$

where $\mathcal{S}_0, \ldots, \mathcal{S}_{T+1} \subset \mathcal{S}$, $\mathcal{A}_0, \ldots, \mathcal{A}_T \subset \mathcal{A}$ and $\mathcal{R}_1, \ldots, \mathcal{R}_{T+1} \subset \mathcal{R}$ by summing as follows:

$$\mathbb{P}_{\mu,\pi}\left(\prod_{t=0}^{T}(\mathcal{S}_t \times \mathcal{A}_t \times \mathcal{R}_{t+1}) \times \mathcal{S}_{T+1} \times (\mathcal{A} \times \mathcal{R} \times \mathcal{S})^{\mathbb{N}}\right)$$

$$= \sum_{\substack{s_0 \in \mathcal{S}_0 \\ \vdots \\ s_{T+1} \in \mathcal{S}_{T+1}}} \sum_{\substack{a_0 \in \mathcal{A}_0 \\ \vdots \\ a_T \in \mathcal{A}_T}} \sum_{\substack{r_1 \in \mathcal{R}_1 \\ \vdots \\ r_{T+1} \in \mathcal{R}_{T+1}}} \mu(s_0)\prod_{t=0}^{T}\pi_t(a_t|h^{(t)})p(s_{t+1}, r_{t+1}|s_t, a_t).$$

$\mathbb{P}_{\mu,\pi}$ can then be seen to satisfy the assumptions of Kolmogorov's extension theorem which assures that $\mathbb{P}_{\mu,\pi}$ can be extended to a unique probability measure on $\mathcal{H}^\infty$. $\square$

Definition 1.3.2. Let $\mu \in \Delta(\mathcal{S})$, $\pi \in \Pi$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

(i) $\mathbb{P}_{\mu,\pi}$ is called the *probability distribution over histories* induced by initial state distribution μ and policy π.

(ii) We write $\mathbb{P}_{s,\pi}$ instead of $\mathbb{P}_{\delta_s,\pi}$, which is called the probability distribution over histories induced by initial state s and policy π.

(iii) Let $\pi' = (\pi'_t)_{t \geqslant 0}$ defined as

$$\begin{aligned}\pi'_0(s) &= \delta_a, \\ \pi'_0(s') &= \pi_0(s') \quad \text{for } s' \neq s \\ \pi'_t &= \pi_t \quad \text{for } t \geqslant 1.\end{aligned}$$

$\mathbb{P}_{s,\pi'}$ is then called the probability distribution induced by initial state s, initial action a, and policy π, and is denoted $\mathbb{P}_{s,a,\pi}$.

The following shorthands will be used:

$$\begin{aligned}\mathbb{E}_{\mu,\pi}[\,\cdot\,] &= \mathbb{E}_{(S_0,A_0,R_1,\dots)\sim\mathbb{P}_{\mu,\pi}}[\,\cdot\,] \\ \mathbb{E}_{s,\pi}[\,\cdot\,] &= \mathbb{E}_{(S_0,A_0,R_1,\dots)\sim\mathbb{P}_{s,\pi}}[\,\cdot\,] \\ \mathbb{E}_{s,a,\pi}[\,\cdot\,] &= \mathbb{E}_{(S_0,A_0,R_1,\dots)\sim\mathbb{P}_{s,a,\pi}}[\,\cdot\,].\end{aligned}$$

$\mathbb{P}_{s,a,\pi}$ corresponds to the interaction where the initial state is s, initial action is a (deterministically), and decision rule is given π only for $t \geqslant 1$. It cannot be defined as $\mathbb{P}_{s,a}$ conditionned on the event $A_0 = a$ because the probability $\pi(a|s)$ of this event may be zero.

Proposition 1.3.3. *Let $f : \mathcal{S} \times \mathcal{R} \to \mathbb{R}$ and π a stationary policy. Then,*

(i) for all $s \in \mathcal{S}$,

$$\sum_{(a,r,s') \in \mathcal{A}\times\mathcal{R}\times\mathcal{S}} \pi(a|s) p(r, s'|s, a) f(r, s') = \mathbb{E}_{s,\pi}[f(S_1, R_1)],$$

(ii) for all $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\sum_{(r,s') \in \mathcal{S}\times\mathcal{A}} p(r, s'|s, a) f(r, s') = \mathbb{E}_{s,a,\pi}[f(S_1, R_1)].$$

Proof. Using the definition of $\mathbb{P}_{s,\pi}$, and more precisely the expression from

Proposition 1.3.1,

$$\begin{aligned}
&\mathbb{E}_{s,\pi}\left[f(S_1, R_1)\right] \\
&\quad = \sum_{(r,s')\in\mathcal{R}\times\mathcal{S}} f(r,s') \times \mathbb{P}_{s,\pi}\left(\mathcal{S}\times\mathcal{A}\times\{r\}\times\left\{s'\right\}\times(\mathcal{A}\times\mathcal{R}\times\mathcal{S})^{\mathbb{N}}\right) \\
&\quad = \sum_{(r,s')\in\mathcal{R}\times\mathcal{S}} f(r,s') \sum_{a\in\mathcal{A}} \pi(a|s)p(r,s'|s,a) \\
&\quad = \sum_{(a,r,s')\in\mathcal{A}\times\mathcal{R}\times\mathcal{S}} \pi(a|s)p(r,s'|s,a)f(r,s').
\end{aligned}$$

The other identity is proved similarly. □

1.4 Value functions

We now introduce value functions which are fundamental tools for solving MDPs. The *optimal* value function, defined in the next chapter, associates to each state the best possible average reward than can be obtained starting from that state. Almost all algorithms aim at getting close to the optimal value function through iterative updates.

Definition 1.4.1. (i) A *state-value function* (aka *V-function*) is a function $v : \mathcal{S} \to \mathbb{R}$ or equivalently a vector $v = (v(s))_{s\in\mathcal{S}} \in \mathbb{R}^{\mathcal{S}}$.

(ii) An *action-value function* (aka *Q-function*) is a function $q : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ or equivalently a vector $q = (q(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$.

Lemma 1.4.2. *Let $(R_t)_{t\geqslant 1}$ be a sequence of random variables with values in $\mathcal{R}$ and $\gamma \in (0,1)$. Then, the series $\sum_{t\geqslant 1}\gamma^{t-1}R_t$ converges almost-surely, and its sum is integrable.*

Proof. $\mathcal{R}$ being a finite subset of $\mathbb{R}$, it holds that $\max_{r\in\mathcal{R}}|r| < +\infty$. Then,

$$\left|\gamma^{t-1}R_t\right| \leqslant \gamma^{t-1}\max_{r\in\mathcal{R}}|r|\,, \quad \text{a.s.}$$

The result follows the dominated convergence theorem. □

Definition 1.4.3. Let $\pi \in \Pi$ and $\gamma \in (0,1)$.

(i) The *state-value function of policy* π with discount factor γ is defined as

$$v_\pi^{(\gamma)}(s) = \mathbb{E}_{s,\pi}\left[\sum_{t=1}^{+\infty}\gamma^{t-1}R_t\right], \quad s\in\mathcal{S}.$$

(ii) The *action-value function of policy* π with discount factor γ is defined as

$$q_\pi^{(\gamma)}(s,a) = \mathbb{E}_{s,a,\pi}\left[\sum_{t=1}^{+\infty}\gamma^{t-1}R_t\right], \quad (s,a) \in \mathcal{S} \times \mathcal{A}.$$

We may denote $v_\pi = v_\pi^{(\gamma)}$ and $q_\pi = q_\pi^{(\gamma)}$ when γ is clear from the context.