An Introduction to Reinforcement Learning

From theory to algorithms

Joon Kwon

December 5, 2023

Contents

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, robotics, 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

The framework for reinforcement learning is the Markov Decision process, which is a repeated interaction between an agent and a dynamic environment, which can be informally described as follows.

We are given three finite nonempty sets $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{R} \subset \mathbb{R}$. The environment chooses an initial *state* $S_0 \in \mathcal{S}$ and reveals it to the agent. The agent then chooses an *action* $A_0 \in \mathcal{A}$, possibly at random. The environment then draws $(R_1, S_1) \in \mathcal{R} \times \mathcal{S}$ according to a probability distribution that depends on S_0 and A_0. The *reward* R_1 and the new state S_1 are revealed to the agent. The agent then chooses $A_2 \in \mathcal{A}$, possibly at random. The environement then draws $(R_2, S_2) \in \mathcal{R} \times \mathcal{S}$ according to a probability distribution which depends on S_0 and A_0, and so on.

The total reward of the agent $\sum_{t=1}^{+\infty} \gamma^{t-1} R_t$, where $0 < \gamma < 1$ is a given *discount factor*. The goal is to find the decision rule for the agent that yields the highest expected total reward.

Note that at stage $t \geqslant 1$, the choice of actions A_t by the agent may depend on all previously observed information, meaning $(S_0, A_0, R_1, \ldots, R_t, S_t)$.

Depending on the problem, the dynamics of the environement (which maps a state-action pair to a probability distribution over reward-state pairs) may be known or not.

This chapter presents basic notions regading MDPs, in a formal fashion.

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 Formal 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

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

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 $t \geqslant 0$, $\pi_t = \pi_0$; 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 proposition below.

We first introduce the following notation. For $T \geqslant 0$ and $h^{(T)} = (s_0, a_0, r_1, \ldots r_T, s_T)$, we consider the following associated subset of $\mathcal{H}^\infty$:

$$\operatorname{Cyl} h^{(T)} = \{s_0\} \times \{a_0\} \times \{r_1\} \times \cdots \times \{r_T\} \times \{s_T\} \times (\mathcal{A} \times \mathcal{R} \times \mathcal{S})^{\mathbb{N}}.$$

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 σ-algebra) such that for all $T \geqslant 0$, and all $h^{(T)} = (s_0, a_0, r_1, \ldots, r_T, s_T) \in \mathcal{H}^{(T)}$,*

$$\mathbb{P}_{\mu,\pi}\left(\operatorname{Cyl} h^{(T)}\right) = \mu(s_0) \prod_{t=0}^{T-1} \pi_t(a_t|h^{(t)}) p(r_{t+1}, s_{t+1}|s_t, a_t).$$

where for each $0 \leqslant t \leqslant T$, $h^{(t)} = (s_0, a_0, r_1, \ldots, s_{t-1}, a_{t-1}, r_t, s_t)$.

Sketch of proof. The above expression defines a value for each set of the form $\operatorname{Cyl} h^{(T)}$ for $T \geqslant 0$ and $h^{(T)} \in \mathcal{H}^{(T)}$. 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:

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

$\mathbb{P}_{s,a,\pi}$ 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 $\pi = (\pi_t)_{t \geqslant 0}$ be a policy and $s \in \mathcal{S}$. Then,*

$$\mathbb{P}_{s,\pi} = \sum_{a \in \mathcal{A}} \pi_0(a|s) \cdot \mathbb{P}_{s,a,\pi}.$$

Proof. It is sufficient to prove the identity between those two measures on the sets that appear in the statement of Proposition 1.3.1, because they would then uniquely extend to all measurable subsets of $\mathcal{H}^\infty$.

Let $T \geqslant 0$ and $h^{(T)} = (s_0, a_0, r_1, \ldots, r_T, s_T) \in \mathcal{H}^{(T)}$, and denote $h^{(t)} := (s_0, a_0, r_1, \ldots, r_r, s_t)$ for $0 \leqslant t \leqslant T$. If $s_0 \neq s$, then the measures of the identity are zero when evaluated at $\operatorname{Cyl} h^{(T)}$. We now assume $s_0 = s$.

Fix $a \in \mathcal{A}$ and consider π' defined as in Definition 1.3.2. Then,

$$\begin{aligned}
\pi_0(a|s) \cdot \mathbb{P}_{s,a,\pi}\left(\operatorname{Cyl} h^{(T)}\right) &= \pi_0(a|s) \prod_{t=0}^{T-1} \pi'_t(a_t|h^{(t)}) p(r_{t+1}, s_{t+1}|s_t, a_t) \\
&= \mathbb{1}\left\{s_0 = s\right\} \prod_{t=0}^{T-1} \pi_t(a_t|h^{(t)}) p(r_{t+1}, s_{t+1}|s_t, a_t) \\
&= \mathbb{1}\left\{a_0 = a\right\} \cdot \mathbb{P}_{s,\pi}\left(\operatorname{Cyl} h^{(T)}\right).
\end{aligned}$$

Summing over $a \in \mathcal{A}$ then gives

$$\sum_{a \in \mathcal{A}} \pi_0(a|s) \cdot \mathbb{P}_{s,a,\pi}\left(\operatorname{Cyl} h^{(T)}\right) = \mathbb{P}_{s,\pi}\left(\operatorname{Cyl} h^{(T)}\right).$$

□

Proposition 1.3.4. *Let $s \in \mathcal{S}$, $a \in \mathcal{A}$, π a stationary policy, $f : \mathcal{H}^\infty \to \mathbb{R}$ a bounded measurable function (with respect to the product σ-algebra) and random variables $(S'_0, A'_0, R'_1, S'_2, A'_2, R'_2, \ldots)$ with distribution $\mathbb{P}_{s,\pi}$ or $\mathbb{P}_{s,a,\pi}$. Then, almost-surely,*

(i) For all $t \geqslant 0$,

$$\mathbb{E}_{S'_t,\pi}\left[f(S_0, A_0, R_1, \ldots)\right] = \mathbb{E}\left[f(S'_t, A'_t, R'_{t+1}, \ldots) \,\middle|\, S'_t\right],$$

(ii) and for all $t \geqslant 1$,

$$\mathbb{E}_{S'_t,A'_t,\pi}\left[f(S_0, A_0, R_1, \ldots)\right] = \mathbb{E}\left[f(S'_t, A_t,' R'_{t+1}, \ldots) \,\middle|\, S'_t, A'_t\right].$$

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 $q = (q(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$.

We equip both spaces with the ℓ^∞ norm:

$$\|v\|_\infty = \max_{s\in\mathcal{S}} |v(s)|, \qquad \|q\|_\infty = \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |q(s,a)|,$$

and with component-wise inequalities:

$$v \leqslant v' \quad \Longleftrightarrow \quad \forall s \in \mathcal{S},\ v(s) \leqslant v'(s),$$

$$q \leqslant q' \quad \Longleftrightarrow \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A},\ q(s,a) \leqslant q'(s,a).$$

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.

Remark 1.4.4. $v_\pi(s)$ corresponds to the expected total reward starting from state s and following policy π.

Chapter 2

Bellman operators & optimality

Bellman operators are the fundamental tool for solving MDPs. This chapter introduces their definitions and properties. We then define optimal value functions and policies, and characterize them with the help of the Bellman operators.

We assume that $\gamma \in (0, 1)$ in given. The image of an element $x \in X$ by a map $F : X \to Y$ will often be denoted Fx instead of $F(x)$.

2.1 Bellman operators

Definition 2.1.1. Let π be a stationary policy. We define the following operators.

(i) $D^{(\gamma)} : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ as

$$(D^{(\gamma)}v)(s, a) = \sum_{(r,s')\in\mathcal{S}\times\mathcal{R}} p(r, s'|s, a)(r + \gamma v(s')), \quad s \in \mathcal{S},\ a \in \mathcal{A}.$$

(ii) $E_\pi : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}}$ as

$$(E_\pi q)(s) = \sum_{a\in\mathcal{A}} \pi(s|a)q(s, a), \quad s \in \mathcal{S}.$$

(iii) $E_* : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}}$ as

$$(E_* q)(s) = \max_{a\in\mathcal{A}} q(s, a), \quad s \in \mathcal{S}.$$

(iv) $B^{(V,\gamma)}_\pi = E_\pi \circ D^{(\gamma)}$ (Bellman expectation operator for state-value functions)

(v) $B_*^{(V,\gamma)} = E_* \circ D^{(\gamma)}$ (Bellman optimality operator for state-value functions)

(vi) $B_\pi^{(Q,\gamma)} = D^{(\gamma)} \circ E_\pi$ (Bellman expectation operator for action-value functions)

(vii) $B_*^{(Q,\gamma)} = D^{(\gamma)} \circ E_*$ (Bellman optimality operator for action-value functions)

We will use lighter notation $D, E_\pi, E_*, B_\pi, B_*$ as soon as context prevents confusion. The following expressions follow from the definitions.

Proposition 2.1.2 (Explicit expression of Bellman operators)**.** *Let $v \in \mathbb{R}^{\mathcal{S}}$, $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, and π a stationary policy. Then, the following expressions hold.*

$$(B_\pi v)(s) = \sum_{(a,r,s')\in\mathcal{A}\times\mathcal{S}\times\mathcal{R}} \pi(a|s)p(r,s'|s,a)\left(r + \gamma v(s')\right), \quad s \in \mathcal{S},$$

$$(B_* v)(s) = \max_{a\in\mathcal{A}} \sum_{(r,s')\in\mathcal{S}\times\mathcal{R}} p(r,s'|s,a)(r + \gamma v(s')), \quad s \in \mathcal{S},$$

$$(B_\pi q)(s,a) = \sum_{(r,s',a')\in\mathcal{S}\times\mathcal{R}\times\mathcal{A}} p(r,s'|s,a)\left(r + \gamma\pi(a'|s')q(s',a')\right), \quad (s,a) \in \mathcal{S}\times\mathcal{A},$$

$$(B_* q)(s,a) = \sum_{(r,s')\in\mathcal{S}\times\mathcal{R}} p(r,s'|s,a)\left(r + \gamma \max_{a'\in\mathcal{A}} q(s',a')\right), \quad (s,a) \in \mathcal{S}\times\mathcal{A}.$$

Proof. Immediate from the definitions. □

Proposition 2.1.3 (Bellman operators as expectations)**.** *Let $v \in \mathbb{R}^{\mathcal{S}}$, $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, π a policy, $s \in \mathcal{S}$, $a \in \mathcal{A}$. Then,*

(i) $(Dv)(s,a) = \mathbb{E}_{s,a,\pi}\left[R_1 + \gamma v(S_1)\right]$,

and if π is stationnary,

(ii) $(E_\pi q)(s) = \mathbb{E}_{s,\pi}\left[q(s,A_0)\right]$,

(iii) $(B_\pi v)(s) = \mathbb{E}_{s,\pi}\left[R_1 + \gamma v(S_1)\right]$,

(iv) $(B_\pi q)(s,a) = \mathbb{E}_{s,a,\pi}\left[R_1 + \gamma q(S_1,A_1)\right]$.

(v) $(B_* v)(s) = \max_{a\in\mathcal{A}} \mathbb{E}_{s,a,\pi}\left[R_1 + \gamma v(S_1)\right]$,

(vi) $(B_* q)(s,a) = \mathbb{E}_{s,a,\pi}\left[R_1 + \gamma \max_{a'\in\mathcal{A}} q(S_1,a')\right]$.

Proof. Let us prove (i). Let π' the policy associated with (s,a) used in Definition 1.3.2 to define $\mathbb{P}_{s,a,\pi}$. Using the definition of the probability measure $\mathbb{P}_{s,\pi}$ (see Proposition 1.3.1),

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

We now turn to (ii).

$$\begin{aligned}
E_{s,\pi}\mathbb{E}\left[q(s,A_0)\right] &= \sum_{a\in\mathcal{A}} q(s,a)\times\mathbb{P}_{s,a}\left(\mathcal{S}\times\{a\}\times(\mathcal{R}\times\mathcal{S}\times\mathcal{A})^{\mathbb{N}}\right) \\
&= \sum_{a\in\mathcal{A}} q(s,a)\pi(a|s) = (E_\pi q)(s).
\end{aligned}$$

We now deduce (iii) using Proposition 1.3.3:

$$\begin{aligned}
(B_\pi v)(s) = (E_\pi(Dv))(s) &= \sum_{a\in\mathcal{A}} \pi(a|s)\mathbb{E}_{s,a,\pi}\left[R_1+\gamma v(S_1)\right] \\
&= \mathbb{E}_{s,\pi}\left[R_1+\gamma v(S_1)\right].
\end{aligned}$$

For (iv), we combine (i) and (ii) with the help of the Markov property from Proposition 1.3.4; let $(S'_0, A'_0, R'_1, \dots) \sim \mathbb{P}_{s,a,\pi}$, then

$$\begin{aligned}
(B_\pi q)(s,a) = (D(E_\pi q))(s,a) &= \mathbb{E}\left[R'_1+\gamma(E_\pi q)(S'_1)\right] \\
&= \mathbb{E}\left[R'_1+\gamma\cdot\mathbb{E}_{S'_1,\pi}\left[q(S_0,A_0)\right]\right] \\
&= \mathbb{E}\left[R'_1+\gamma\cdot\mathbb{E}\left[q(S'_1,A'_1)\mid S'_1\right]\right] \\
&= \mathbb{E}\left[R'_1+\gamma\cdot q(S'_1,A'_1)\right].
\end{aligned}$$

Finally, (v) and (vi) follow by composition. □

Remark 2.1.4. If for each $s\in\mathcal{S}$, $v(s)$ is interpreted as an estimate of the total reward obtained starting from state s and using policy π, $(B_\pi v)(s)$ is then an alternative estimate, as it is the expectation, when starting from state s of the actual first reward R_1, *plus* $\lambda v(S_1)$ which is an estimate of remaining discounted rewards, as estimated by v. A similar interpretation holds for $B_\pi q$. We will see that the latter estimate is in some sense better: the Bellman operators will thus be used to iteratively *update* the estimates.

Definition 2.1.5. Let $d, n \geqslant 1$ integers. A map $F : \mathbb{R}^d \to \mathbb{R}^n$ is *monotone* if for all $x, x' \in \mathbb{R}^d$, $x \leqslant x'$ implies $Fx \leqslant Fx'$, where the inequalities are to be understood component-wise.

Proposition 2.1.6. *Let π be a stationary policy. Then, operators D, E_π, $B_\pi^{(V)}$ and $B_\pi^{(Q)}$ are affine with nonnegative coefficients. E_π is moreover linear. In particular, they are monotone.*

Proof. Immediate from the definitions. □

Proposition 2.1.7. *Let $v \in \mathbb{R}^{\mathcal{S}}$, $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Then,*

(i) $(E_* q)(s) = \sup_{\pi\in\Pi_0} (E_\pi q)(s) = \sup_{\pi\in\Pi_{0,d}} (E_\pi q)(s)$,

(ii) $(B_* v)(s) = \sup_{\pi\in\Pi_0} (B_\pi v)(s) = \sup_{\pi\in\Pi_{0,d}} (B_\pi v)(s)$,

(iii) $(B_* q)(s,a) = \sup_{\pi\in\Pi_0} (B_\pi q)(s,a) = \sup_{\pi\in\Pi_{0,d}} (B_\pi q)(s,a)$.

Proof. (i) is an easy consequence from the definition of E_*. Then (ii) and (iii) follow using the monotonicity from Proposition 2.1.6. □

2.2 Bellman equations

Definition 2.2.1. Let X be a set and $F : X \to X$. An element $x \in X$ is a *fixed point* of F is $Fx = x$.

The fixed points of Bellman operators will be of particular interest. They are often written in the form of the so-called Bellman equations: for a given stationary policy π, a state-value function $v \in \mathbb{R}^{\mathcal{S}}$ is a fixed point of $B_\pi^{(V)}$ if, and only if:

$$v(s) = \sum_{(a,r,s')\in\mathcal{A}\times\mathcal{S}\times\mathcal{R}} \pi(a|s)p(r,s'|s,a)\left(r + \gamma v(s')\right), \quad s \in \mathcal{S}.$$

The above is called the *Bellman expectation equation* for state-value functions. Similarly, v is the fixed point of $B_*^{(V)}$ if, and only if:

$$v(s) = \max_{a\in\mathcal{A}} \sum_{(r,s')\in\mathcal{S}\times\mathcal{R}} p(r,s'|s,a)(r + \gamma v(s')), \quad s \in \mathcal{S},$$

which is called the Bellman *optimality equation*. The corresponding equations for action-value functions are similarly defined. We establish below that these equations have unique solutions and that they correspond respectively to v_π and v_*, where v_* is the value function associated with an optimal policy.

Theorem 2.2.2 (Banach's fixed point theorem)**.** *Let $0 \leqslant \gamma < 1$, (X, d) a complete metric space, and $F : X \to X$ a γ-Lipschitz map (with respect to distance d). Then, F has a unique fixed point $x_* \in X$ and for all sequence $(x_k)_{k \geqslant 0}$ satisfying $x_{k+1} = Fx_k$ ($k \geqslant 0$), it holds that*

$$d(x_k, x_*) \leqslant \gamma^k d(x_0, x_*), \quad k \geqslant 0,$$

and thus $x_k \longrightarrow x_*$ as $k \to +\infty$.

Remark 2.2.3. The above convergence is guaranteed *regardless* of the initial point x_0.

Proposition 2.2.4. *Let π be a stationary policy. With respect to the norms $\|\cdot\|_\infty$ in $\mathbb{R}^{\mathcal{S}}$ and $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,*

(i) $D^{(\gamma)}$ is γ-Lipschitz continuous,

(ii) E_π is 1-Lipschitz continuous,

(iii) E_* is 1-Lipschitz continuous,

(iv) $B_\pi^{(V,\gamma)}$, $B_*^{(V,\gamma)}$, $B_\pi^{Q,\gamma}$ and $B_*^{(Q,\gamma)}$ are γ-Lipschitz continuous and admit unique fixed points.

Proof. Let $v, v' \in \mathbb{R}^{\mathcal{S}}$.

$$\begin{aligned}
\left\|Dv' - Dv\right\|_\infty &= \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|Dv'(s,a) - Dv(s,a)\right| \\
&= \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| \sum_{(r,s')\in\mathcal{R}\times\mathcal{S}} p(r,s'|s,a)\gamma(v'(s') - v(s)) \right| \\
&\leqslant \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \gamma \left\|v' - v\right\|_\infty \sum_{(r,s')\in\mathcal{R}\times\mathcal{S}} p(r,s'|s,a) \\
&= \gamma \left\|v' - v\right\|_\infty,
\end{aligned}$$

where the last inequality follows from $p(\,\cdot\,|s,a)$ being a probability distribution over $\mathcal{R}\times\mathcal{S}$, which proves (i).

Let $q, q' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and π a stationary policy.

$$\begin{aligned}
\left\|E_\pi q' - E_\pi q\right\|_\infty &= \max_{s\in\mathcal{A}} \left| \sum_{a\in\mathcal{A}} \pi(a|s) \left|q'(s,a) - q(s,a)\right| \right| \\
&\leqslant \max_{s\in\mathcal{A}} \sum_{a\in\mathcal{A}} \pi(a|s) \left\|q' - q\right\|_\infty \\
&= \left\|q' - q\right\|_\infty,
\end{aligned}$$

where the last inequality follows from $\pi(\,\cdot\,|s)$ being a probability distribution over $\mathcal{A}$.

Let $s \in \mathcal{S}$. If $(E_* q')(s) \geqslant (E_* q)(s)$, then

$$\begin{aligned}
\left|(E_* q')(s) - (E_* q)(s)\right| &= (E_* q')(s) - (E_* q)(s) \\
&= \max_{a' \in \mathcal{A}} q'(s, a') - \max_{a \in \mathcal{A}} q(s, a) \\
&\leqslant \max_{a' \in \mathcal{A}} \left\{q'(s, a') - q(s, a')\right\} \\
&\leqslant \max_{a' \in \mathcal{A}} \left|q'(s, a') - q(s, a')\right| \\
&\leqslant \left\|q' - q\right\|_\infty .
\end{aligned}$$

Similarly, if $(E_* q')(s) \leqslant (E_* q)(s)$, then

$$\left|E_* q'(s) - E_* q(s)\right| \leqslant \left\|q' - q\right\|_\infty .$$

Taking the maximum over $s \in \mathcal{S}$ yields (iii):

$$\left\|E_* q' - E_* q\right\|_\infty \leqslant \left\|q' - q\right\|_\infty .$$

The Lipschitz property (iv) of Bellman operators then follow by composition. □

Proposition 2.2.5. *Let π be a stationary policy. Then,*

(i) $v_\pi = E_\pi q_\pi$,

(ii) $q_\pi = D v_\pi$,

(iii) v_π *is the unique fixed point of* $B_\pi^{(V)}$,

(iv) q_π *is the unique fixed point of* $B_\pi^{(Q)}$.

Proof. Let $s \in \mathcal{S}$. We prove (i) using Proposition 1.3.3:

$$\begin{aligned}
(E_\pi q_\pi)(s) &= \sum_{a \in \mathcal{A}} \pi(a|s) \cdot \mathbb{E}_{s,a,\pi}\left[\sum_{t=1}^{+\infty} \gamma^{t-1} R_t\right] \\
&= \mathbb{E}_{s,\pi}\left[\sum_{t=1}^{+\infty} \gamma^{t-1} R_t\right] \\
&= v_\pi .
\end{aligned}$$

We now turn to (ii). Let $a \in \mathcal{A}$. Let $(S_0', A_0', R_1', \dots) \sim \mathbb{P}_{s,a,\pi}$. Then, using the expression of the Bellman operator as an expectation (from Proposition 2.1.3), we write

$$\begin{aligned}
(Dv_\pi)(s,a) &= \mathbb{E}_{s,a,\pi}\left[R_1 + \gamma v_\pi(S_1)\right] \\
&= \mathbb{E}\left[R_1' + \gamma v_\pi(S_1')\right] \\
&= \mathbb{E}\left[R_1' + \gamma \cdot \mathbb{E}_{S_1',\pi}\left[\sum_{t=1}^{+\infty} \gamma^{t-1} R_t\right]\right] \\
&= \mathbb{E}\left[R_1' + \gamma \cdot \mathbb{E}\left[\sum_{t=1}^{+\infty} \gamma^{t-1} R_{t+1}' \,\middle|\, S_1'\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{+\infty} \gamma^{t-1} R_t\right] = v_\pi,
\end{aligned}$$

where for the fourth equality we used the Markov property for $\mathbb{P}_{s,a,\pi}$ from Proposition 1.3.4.

Combining (i) and (ii) together with Banach's fixed point theorem from Theorem (2.2.2) yields (iv) and (iv). □

Remark 2.2.6. In other words, v_π (resp. q_π) is the unique solution of the Bellman expectation equation for state-value function (resp. action-value functions).

2.3 Greedy policies

Definition 2.3.1. A stationary and deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ is

(i) a *greedy policy* with respect to an action-value function $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ if for all $s \in \mathcal{S}$,

$$\pi(s) \in \operatorname*{Arg\,max}_{a\in\mathcal{A}} q(s,a),$$

where Arg max denotes the set of maximizers.

(ii) a *greedy policy* with respect to an state-value function $v \in \mathbb{R}^{\mathcal{S}}$ if $\pi \in \Pi_g\left[Dv\right]$.

$\Pi_g\left[q\right]$ denotes the set of greedy policies with respect to q and $\Pi_g\left[v\right]$ is a shorthand for $\Pi_g\left[Dv\right]$. Notation $\pi_g\left[q\right]$ (resp. $\pi_g\left[v\right]$) denotes any element from $\Pi_g\left[q\right]$ (resp. $\Pi_g\left[v\right]$).

Remark 2.3.2. $\pi_g\left[q\right]$ corresponds to a policy which selects actions by simply comparing values of the action-value function q. In the case of $\pi_g\left[v\right]$, the action selection is based on a *one-step look-ahead*, as it rewrites as follows using Proposition 2.1.3:

$$\pi_g(s) \in \operatorname*{Arg\,max}_{a\in\mathcal{A}} \mathbb{E}_{s,a,\pi}\left[R_1 + \gamma v(S_1)\right].$$

Proposition 2.3.3. *For $v \in \mathbb{R}^{\mathcal{S}}$ (resp. $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$), $\Pi_g[v]$ (resp. $\Pi_g[q]$) is nonempty.*

Proof. The set of actions $\mathcal{A}$ being finite (and nonempty), $\operatorname{Arg\,max}_{a\in\mathcal{A}} q(s,a)$ is nonempty, and the result follows. ☐

Proposition 2.3.4. *Let $v \in \mathbb{R}^{\mathcal{S}}$ and $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. Then,*

(i) $E_* q = E_{\pi_g[q]} q$,

(ii) $B_* q = B_{\pi_g[q]} q$.

(iii) $B_* v = B_{\pi_g[v]} v$,

Proof. Let $s \in \mathcal{S}$ and $\pi \in \Pi_g[q]$. By definition of a greedy policy,

$$(E_* q)(s) = \max_{a\in\mathcal{A}} q(s,a) = q(s,\pi(s)) = \sum_{a\in\mathcal{A}} \pi(s|a) q(s,a) = (E_\pi q)(s).$$

Then, $B_*^{(Q)} q = D \circ E_* = D \circ E_\pi = B_\pi$ and $B_*^{(V)} q = E_* \circ D = E_\pi \circ D = B_\pi$. ☐

2.4 Optimal value functions & policies

Definition 2.4.1. Let $\gamma \in (0,1)$. The *optimal state-value* and *actions-value functions* with respect to discount factor γ are respectively defined as

$$v_*^{(\gamma)}(s) = \sup_{\pi\in\Pi} v_\pi^{(\gamma)}(s), \quad s \in \mathcal{S},$$
$$q_*^{(\gamma)}(s,a) = \sup_{\pi\in\Pi} q_\pi^{(\gamma)}(s,a), \quad (s,a) \in \mathcal{S}\times\mathcal{A}.$$

As soon as discount factor γ is clear from the context, we may simply use notation v_* and q_*.

Remark 2.4.2. v_* and q_* are well-defined because v_π and q_π can be easily seen to by bounded by $(1-\gamma)^{-1} \max_{r\in\mathbb{R}} |r|$.

Definition 2.4.3. A policy π_* is *optimal* if $v_{\pi_*} = v_*$.

Theorem 2.4.4. *Let v_0 and q_0 the unique fixed points of $B_*^{(V)}$ and $B_*^{(Q)}$ respectively. Then, $\Pi_g[v_0] = \Pi_g[q_0]$ and for π_g in the latter set,*

(i) $v_* = v_0 = v_{\pi_g}$,

(ii) $q_* = q_0 = q_{\pi_g}$,

(iii) $v_* = E_* q_*$,

(iv) $q_* = D v_*$.

Remark 2.4.5. Some important takeaways from the above theorem are the following:

- v_* (resp. q_*) is the unique fixed point of $B_*^{(V)}$ (resp. $B_*^{(Q)}$), meaning the unique solution to the Bellman optimality equation for state-value function (resp. action-value function);
- there exists a stationary and deterministic optimal policy.

Proof. Let us first prove that $q_0 = Dv_0$ and $v_0 = E_* q_0$. Indeed,

$$Dv_0 = DB_* v_0 = DE_* Dv_0 = B_*(Dv_0),$$

therefore, Dv_0 is the unique fixed point of B_*, in other words $q_0 = Dv_0$. Then,

$$E_* q_0 = E_* Dv_0 = B_* v_0 = v_0.$$

Therefore, $\Pi_g[v_0] = \Pi_g[Dv_0] = \Pi_g[q_0]$. We recall that a set of greedy policies is never empty, as stated in Proposition 2.3.3.

Let $\pi_g \in \Pi_g[v_0]$. Then using the property of greedy policies from Proposition 2.3.4, $v_0 = B_* v_0 = B_{\pi_g} v_0$ and $q_0 = B_* q_0 = B_{\pi_g} q_0$. Value functions v_0 and q_0 are therefore the unique fixed points of $B_{\pi_g}^{(V)}$ and $B_{\pi_g}^{(Q)}$, respectively. In other words $v_0 = v_{\pi_g}$ and $q_0 = q_{\pi_g}$, by Proposition 2.2.5.

Therefore, $v_0 = v_{\pi_g} \leqslant \sup_{\pi \in \Pi_{0,d}} v_\pi$ because $\pi_g \in \Pi_{0,d}$ by definition, and similarly $q_0 \leqslant \sup_{\pi \in \Pi_{0,d}} q_\pi$.

Let us now prove that $v_0 \geqslant \sup_{\pi \in \Pi} v_\pi$. Let $\pi = (\pi_t)_{t \geqslant 0}$ be any policy, $s \in \mathcal{S}$, and consider random variables $(S_0, A_0, R_1, S_2, A_2, R_3, \dots) \sim \mathbb{P}_{s,\pi}$. Let $t \geqslant 0$,

$$v_0(S_t) = (B_* v_0)(S_t) = \max_{a \in \mathcal{A}} (Dv)(S_t, a) \geqslant (Dv)(S_t, A_t).$$

Let us rewrite this last quantity. Let $(s_0, a_0) \in \mathcal{S}$ such that $\mathbb{P}[S_t = s_0,\ A_t = a_0] > 0$. Then, using the definition of $\mathbb{P}_{s,\pi}$,

$$\begin{aligned}
(Dv)(s_0, a_0) &= \sum_{(r,s') \in \mathcal{R} \times \mathcal{S}} p(r, s' | s_0, a_0)(r + \gamma v(s')) \\
&= \sum_{(r,s') \in \mathcal{R} \times \mathcal{S}} \frac{\mathbb{P}[R_{t+1} = r,\ S_{t+1} = s',\ S_t = s_0,\ A_t = a_0]}{\mathbb{P}[S_t = s_0,\ A_t = a_0]}(r + \gamma v(s')) \\
&= \mathbb{E}[R_{t+1} + \gamma v(S_{t+1}) \mid S_t, A_t].
\end{aligned}$$

Therefore,

$$v_0(S_t) \geqslant \mathbb{E}[R_{t+1} + \gamma v(S_{t+1}) \mid S_t, A_t].$$

Then using the expression of $(Bv_0)(s)$ from Proposition 2.1.3, applying the above recursively, we get

$$\begin{aligned} v_0(s) = (Bv_0)(s) &= \mathbb{E}\left[R_1 + \gamma v_0(S_1)\right] \\ &\geqslant \mathbb{E}\left[R_1 + \gamma \mathbb{E}\left[R_2 + \gamma v(S_2) \,|\, S_1, A_1\right]\right] \\ &\geqslant \cdots \geqslant \mathbb{E}_{s,\pi}\left[\sum_{t=1}^{T} \gamma^{t-1} R_t + \gamma^T v(S_T)\right] \\ &\geqslant \mathbb{E}\left[\sum_{t=1}^{+\infty} \gamma^{t-1} R_t\right] = v_\pi(s). \end{aligned}$$

Therefore,

$$v_* = \sup_{\pi \in \Pi} v_\pi \leqslant v_0 = v_{\pi_g} \leqslant \sup_{\pi \in \Pi_{0,d}} v_\pi \leqslant \sup_{\pi \in \Pi} v_\pi = v_*,$$

and the lower and upper bounds being equal, all inequalies are equalities, and the supremums are maximums because they are attained for $\pi_g \in \Pi_{0,d} \subset \Pi$.

Then, we write

$$\begin{aligned} q_* = \sup_{\pi \in \Pi} q_\pi \geqslant \max_{\pi \in \Pi_{0,d}} q_\pi \geqslant q_{\pi_g} = q_0 = Dv_0 \\ = D\left(\max_{\pi \in \Pi} v_\pi\right) \geqslant \sup_{\pi \in \Pi} Dv_\pi = \sup_{\pi \in \Pi} q_\pi = q_*, \end{aligned}$$

where the last inequality holds by monotonicity of D from Proposition 2.1.6 (by writing for $\pi \in \Pi$, $D \max_{\pi \in \Pi} v_\pi \geqslant Dv_\pi$ and then taking the supremum over $\pi \in \Pi$). Therefore, all inequalities are equalities and all supremums are maximums. □

Chapter 3

Dynamic programming

The properties of the Bellman operators established in the previous chapter allow the contruction and analysis of dynamic programming algorithms (DP), meaning algorithms that solve MDPs with known dynamics. Starting from Chapter 4, we will study reinforcement learning, which is solving MDPs with either unknown dynamics, and/or by approximating the problem in some way. Most reinforcement learning methods (RL) are sample variants of dynamic programming algorithms.

3.1 Value iteration

Policy evaluation is the computation of the value function v_π or q_π of a policy π. Many dynamic programming and reinforcement learning algorithms use policy evaluation as an intermediate step in finding the optimal policy. The (synchronous) value iteration for policy evaluation computes v_π (or q_π), in the case of a stationnary policy, by iterating the Bellman expectation operator $B_\pi^{(V)}$ (resp. $B_\pi^{(Q)}$). *Synchronous* means that all values (for each state, or each state-action pair) are updated simultaneously using the values from the current iterate.

In the context of MDPs, *control* is the computation of an optimal value function v_* or q_* and an optimal policy. The (synchronous) value iteration for control computes v_* (resp. q_*) by iterating the Bellman expectation operator $B_*^{(V)}$ (resp. $B_*^{(Q)}$).

Definition 3.1.1 (Synchronous value iteration)**.** Let π be a stationary policy, $(v_k)_{k\geqslant 0}$ and $(q_k)_{k\geqslant 0}$ two sequences in $\mathbb{R}^{\mathcal{S}}$ and $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ respectively.

(i) $(v_k)_{k\geqslant 0}$ is a *synchronous state-value iteration for the evaluation of* π if for all $k \geqslant 0$,
$$v_{k+1} = B_\pi v_k.$$

(ii) $(q_k)_{k\geqslant 0}$ is a *synchronous action-value iteration for the evaluation of* π if for all $k \geqslant 0$,

$$q_{k+1} = B_\pi q_k.$$

(iii) $(v_k)_{k\geqslant 0}$ is a *synchronous state-value iteration for control* if for all $k \geqslant 0$,

$$v_{k+1} = B_* v_k.$$

(iv) $(q_k)_{k\geqslant 0}$ is a *synchronous action-value iteration for control* if for all $k \geqslant 0$,

$$q_{k+1} = B_* q_k.$$

Remark 3.1.2. Value iterations for state-value functions explicitely write as

$$v_{k+1}(s) = \sum_{(a,r,s')\in\mathcal{A}\times\mathcal{S}\times\mathcal{R}} \pi(a|s)p(r,s'|s,a)(r+\gamma v_k(s')), \qquad s\in\mathcal{S}, \quad k\geqslant 0,$$

for the evaluation of π, and as

$$v_{k+1}(s) = \max_{a\in\mathcal{A}} \sum_{(r,s')\in\mathcal{S}\times\mathcal{R}} p(r,s'|s,a)(r+\gamma v_k(s')), \qquad s\in\mathcal{S}, \quad k\geqslant 0,$$

for control. Similar expression hold for action-value functions.

Proposition 3.1.3 (Equivalence between synchronous state-value and action-value iterations)**.** *Let π be a stationnary policy, $(v_k)_{k\geqslant 0}$ and $(q_k)_{k\geqslant 0}$ two sequences in $\mathbb{R}^{\mathcal{S}}$ and $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ respectively. Consider the following assertions.*

(a) $\forall k\geqslant 0, \quad v_{k+1} = B_\pi v_k$;

(b) $\forall k\geqslant 0, \quad q_{k+1} = B_\pi q_k$;

(c) $\forall k\geqslant 0, \quad v_{k+1} = B_* v_k$;

(d) $\forall k\geqslant 0, \quad q_{k+1} = B_* q_k$;

(e) $\forall k\geqslant 0, \quad q_k = D v_k$;

(f) $\forall k\geqslant 0, \quad v_k = E_\pi q_k$;

(g) $\forall k\geqslant 0, \quad v_k = E_* q_k$.

Then,

(i) (a) *and* (e) *imply* (b),

(ii) (b) *and* (f) *imply* (a),

(iii) (c) *and* (e) *imply* (d),

(iv) (d) *and* (g) *imply* (c).

Proof. Assume (a) and (e). Then for all $k\geqslant 0$,

$$B_\pi q_k = DE_\pi D v_k = DB_\pi v_k = D v_{k+1} = q_{k+1},$$

and (b) holds. The other implications are proved similarly. □

Proposition 3.1.4 (Linear convergence of synchronous value iteration)**.** *Let π be a stationary policy.*

- *If $(v_k)_{k \geqslant 0}$ and $(q_k)_{k \geqslant 0}$ are synchronous state-value (resp. action-value) iterations for the evaluation of policy π, then for all $k \geqslant 0$,*

$$\|v_k - v_\pi\|_\infty \leqslant \gamma^k \|v_0 - v_\pi\|_\infty ,$$
$$\|q_k - q_\pi\|_\infty \leqslant \gamma^k \|q_0 - q_\pi\|_\infty .$$

- *If $(v_k)_{k \geqslant 0}$ and $(q_k)_{k \geqslant 0}$ are synchronous state-value (resp. action-value) iterations for control, then for all $k \geqslant 0$,*

$$\|v_k - v_*\|_\infty \leqslant \gamma^k \|v_0 - v_*\|_\infty ,$$
$$\|q_k - q_*\|_\infty \leqslant \gamma^k \|q_0 - q_*\|_\infty .$$

Proof. We know from Proposition 2.2.5 and Theorem 2.4.4 that v_π (resp. q_π, v_*, q_*) is the unique fixed point of Bellman operator $B_\pi^{(V)}$ (resp. $B_\pi^{(Q)}$, $B_*^{(V)}$, $B_*^{(Q)}$). The latter is γ-Lipschitz continuous with respect to $\|\cdot\|_\infty$ according to Proposition 2.2.4. The Banach's fixed point theorem (Theorem 2.2.2) then applies and gives the result. □

Remark 3.1.5 (Computational complexity and memory requirements). The above results demonstrate that both algorithms for policy evaluation (resp. control) are equivalent in terms of output solutions. However, computational complexity and memory requirements of the state-value counterpart are lower by a factor $|\mathcal{A}|$. There is therefore no reason to choose action-value iteration in the context of dynamic programming. In reinforcement learning however, the additional stored values of the latter will be of great help.

3.2 Policy iteration

Proposition 3.2.1 (Greedy policy improvement)**.** *Let π be a stationary policy and $\pi_g \in \Pi_g [v_\pi]$. Then,*

(i) $v_{\pi_g} \geqslant v_\pi$,

(ii) $q_{\pi_g} \geqslant q_\pi$,

(iii) $v_{\pi_g} = v_\pi$ *implies* $v_\pi = v_*$,

(iv) $q_{\pi_g} = q_\pi$ *implies* $q_\pi = q_*$.

Proof. Using the fact that v_π is a fixed point of B_π (Proposition 2.2.5), the property $B_* = \sup_{\pi_0 \in \Pi_0} B_{\pi_0}$ from Proposition 2.1.7 and the property of greedy policies from Proposition 2.3.4,

$$v_\pi = B_\pi v_\pi \leqslant B_* v_\pi = B_{\pi_g} v_\pi .$$

Then, applying on both sides operator B_{π_g}, which is monotone thanks to Proposition 2.1.6, we get $B_{\pi_g} v_\pi \leqslant B_{\pi_g}^2 v_\pi$. Therefore, $v_\pi \leqslant B_{\pi_g}^k v_\pi$ for all $k \geqslant 1$, and by Proposition 3.1.4, taking the limit as $k \to +\infty$ gives (i).

Besides, using the monotonicity of D from Proposition 2.1.6, together with Proposition 2.2.5 gives (ii):

$$q_\pi = Dv_\pi \leqslant Dv_{\pi_g} \leqslant q_{\pi_g}.$$

Using Propositions 2.2.5 and 2.3.4, we write $v_\pi = v_{\pi_g} = B_{\pi_g} v_{\pi_g} = B_{\pi_g} v_\pi = B_* v_\pi$. Thus, v_π is a fixed point of B_*, and $v_\pi = v_*$ by Theorem 2.4.4, which proves (iii). (iv) is proved similarly. □

Definition 3.2.2 (Policy iteration)**.** A sequence $(\pi_k)_{k\geqslant 0}$ of stationnary policies is a *policy iteration* if $\pi_{k+1} \in \Pi_g\left[v_{\pi_k}\right]$ for all $k \geqslant 0$.

Remark 3.2.3 (Policy iteration is an idealized algorithm)*.* Except in situations where v_{π_k} can be computed exactly, policy iteration is only an idealized algorithm because each step would involve the computation of v_{π_k} by iterating B_{π_k} infinitely. A practical variant, where B_{π_k} is only iterated a finite number of times is discussed below.

Remark 3.2.4 (Equivalent definition from action-value functions)*.* Policy iteration can be written with action-value functions:

$$\pi_{k+1} \in \Pi_g\left[q_{\pi_k}\right],$$

which is equivalent to the above, because by definition of greedy policies for state-value functions:

$$\Pi_g\left[v_{\pi_k}\right] = \Pi_g\left[Dv_{\pi_k}\right] = \Pi_g\left[q_{\pi_k}\right],$$

where we used Proposition 2.2.5 for the last equality. Unlike value iterations, the corresponding algorithm is exactly the same even regarding the computational and memory requirements, because determining a greedy policy in $\Pi_g\left[v_{\pi_k}\right]$ requires by definition the computation of $Dv_{\pi_k} = q_{\pi_k}$.

Proposition 3.2.5 (Linear convergence of policy iteration)**.** *Let* $(\pi_k)_{k\geqslant 0}$ *be a policy iteration. Then for all* $k \geqslant 0$*,*

$$\begin{aligned} \|v_{\pi_k} - v_*\|_\infty &\leqslant \gamma^k \|v_{\pi_0} - v_*\|_\infty, \\ \|q_{\pi_k} - q_*\|_\infty &\leqslant \gamma^k \|q_{\pi_0} - q_*\|_\infty. \end{aligned}$$

Proof. Denote $v_k = v_{\pi_k}$ for $k \geqslant 0$.

$$\begin{aligned} v_* - v_{k+1} &= B_* v_* - B_* v_k + (B_* - B_{\pi_{k+1}}) v_k + B_{\pi_{k+1}}(v_k - v_{k+1}) \\ &\leqslant B_* v_* - B_* v_k, \end{aligned}$$

where the inequality holds because the second term is zero:

$$B_{\pi_{k+1}} v_k = B_{\pi_g[v_k]} v_k = B_* v_k$$

and the last term is nonpositive because $B_{\pi_{k+1}}$ is monotone according to Proposition 2.1.6, and $v_k \leqslant v_{k+1}$ by property of greedy policy improvement from Proposition 3.2.1. Moreover, by definition of v_*, $v_* \geqslant v_{\pi_{k+1}} = v_{k+1}$. Therefore,

$$0 \leqslant v_* - v_{k+1} \leqslant B_* v_* - B_* v_k$$

and using the Lipschitz continuity of B_* from Proposition 2.2.4,

$$\|v_* - v_{k+1}\|_\infty \leqslant \|B_* v_* - B_* v_k\|_\infty \leqslant \gamma \|v_* - v_k\| .$$

The result for action-value functions is proved similarly. □

Remark 3.2.6 (Generalized iteration). It is possible to define a family of iterations, which generalizes both value iteration and policy iteration. It is sometimes called *generalized policy iteration* or *optimistic policy iteration.* A sequence $(v_k)_{k\geqslant 0}$ in $\mathbb{R}^{\mathcal{S}}$ (resp. $(q_k)_{k\geqslant 0}$ in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$) is a *generalized iteration* for state-value functions (resp. action-value functions) if there exists a sequence $(m_k)_{k\geqslant 0}$ in $\{1, 2, \dots\} \cup \{\infty\}$ such that

$$v_{k+1} = B_{\pi_g[v_k]}^{m_k} v_k, \qquad \left(\text{resp.} \quad q_{k+1} = B_{\pi_g[v_k]}^{m_k} q_k\right),$$

where by convention, $B_\pi^\infty v = v_\pi$ (for all $\pi \in \Pi_0$ and $v \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$). Then, value iteration corresponds to $m_k = 1$ (for all $k \geqslant 0$) and policy iteration corresponds to $m_k = \infty$ (for all $k \geqslant 0$). A practical implementation of policy iteration where m_k may be large but not infinite then belongs to this family.

3.3 Asynchronous fixed point iterations

Theorem 3.3.1 (A generalized fixed point theorem)**.** *Let (X, d) a complete metric space, $(\gamma_k)_{k\geqslant 0}$ nonnegative sequence in $(0, 1)$ and $(F_k)_{k\geqslant 0}$ a sequence of operators in X that share a common fixed point $x_* \in X$ and so that F_k is γ_k-Lipschitz continuous. If $(x_k)_{k\geqslant 0}$ satisfies $x_{k+1} = F_k x_k$ for all $k \geqslant 0$, then*

$$d(x_k, x_*) \leqslant d(x_0, x_*) \left(\prod_{\ell=0}^{k-1} \gamma_\ell\right).$$

If the above product converges to zero, then $x_k \longrightarrow x_*$ as $k \to +\infty$.

Proof. Let $k \geqslant 0$.

$$d(x_{k+1}, x_*) = d(F_k x_k, F_k x_*) \leqslant \gamma_k d(x_k, x_*),$$

hence the result. □

For the remaining of this section, $d \geqslant 1$ will be a given integer.

Definition 3.3.2. Let $F : \mathbb{R}^d \to \mathbb{R}^d$. For $J \subset \{1, \dots, d\}$ and denote $F^{|J} : \mathbb{R}^d \to \mathbb{R}^d$ the operator defined as

$$(F^{|J}x)_j = (Fx)_j$$

for $j \in J$ and and $(F^{|J}x)_{j'} = x_{j'}$ for $j' \notin J$. If $J = \{j\}$ for some $j \in \{1, \dots, d\}$, we denote $F^{|j} = F^{|\{j\}}$.

Remark 3.3.3. $F^{|J}$ can be written as

$$F^{|J} = I + \mathbb{1}_J \otimes (F - I) = (1 - \mathbb{1}_J) \otimes I + \mathbb{1}_J \otimes F$$

where $\mathbb{1}_J$ denotes the vector with value 1 for components in J and value 0 for the other components, and $\otimes$ denotes component-wise multiplication. This expression will be easier to generalize.

Proposition 3.3.4. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a 1-Lipschitz continuous map for $\|\cdot\|_\infty$. Then for all $J \subset \{1, \dots, d\}$, $F^{|J}$ is 1-Lipschitz continuous for $\|\cdot\|_\infty$.*

Proof. For $x, x' \in \mathbb{R}^d$,

$$\begin{aligned}
\left\|F^{|J}x - F^{|J}x'\right\|_\infty &= \max_{1 \leqslant j \leqslant d} \left|(F^{|J}x)_j - (F^{|J}x')_j\right| \\
&= \max\left\{\max_{j \in J}\left|(F^{|J}x)_j - (F^{|J}x')_j\right|,\ \max_{j \notin J}\left|(F^{|J}x)_j - (F^{|J}x')_j\right|\right\} \\
&\leqslant \max\left\{\max_{j \in J}\left|(Fx)_j - (Fx')_j\right|,\ \max_{j \notin J}\left|x_j - x'_j\right|\right\} \\
&\leqslant \max\left\{\left\|Fx - Fx'\right\|_\infty,\ \left\|x - x'\right\|_\infty\right\} \\
&\leqslant \left\|x - x'\right\|_\infty .
\end{aligned}$$

□

Proposition 3.3.5. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ and $x \in \mathbb{R}^d$. The following propositions are equivalent:*

(i) x a fixed point of F,

(ii) x a is fixed point of $F^{|j}$ for all $j \in \{1, \dots, d\}$,

(iii) x a is fixed point of $F^{|J}$ for all $J \subset \{1, \dots, d\}$.

Proof. Immediate □

Proposition 3.3.6. *Let $\gamma \in [0, 1]$ and $F : \mathbb{R}^d \to \mathbb{R}^d$ a γ-Lipschitz continuous map for $\|\cdot\|_\infty$. Let $J_1, \dots, J_M$ be such that $\bigcup_{m=1}^M J_m = \{1, \dots, d\}$. Then, $F^{|J_M} \circ \cdots \circ F^{|J_1}$ is γ-Lipschitz continuous for $\|\cdot\|_\infty$.*

Proof. For each $1 \leqslant m \leqslant M$, denote $F^{1:m} = F^{|J_m} \circ F^{|J_{m-1}} \circ \cdots \circ F^{|J_1}$.

Now fix $1 \leqslant j \leqslant d$ and let m be the largest integer such that $j \in J_m$. Then it follows that,

$$(F^{1:M}x)_j = (F^{|J_M}(F^{1:M-1}x))_j = (F^{1:M-1}x)_j = \cdots = (F^{1:m}x)_j = (F(F^{1:m-1}x))_j.$$

Similarly, $(F^{1:M}x')_j = (F(F^{1:m-1}x'))_j$. Then using the above,

$$\begin{aligned}\left|(F^{1:M}x)_j - (F^{1:M}x')_j\right| &= \left|(F(F^{1:m-1}x))_j - (F(F^{1:m-1}x'))_j\right| \\ &\leqslant \left\|F(F^{1:m-1}x) - F(F^{1:m-1}x')\right\|_\infty \\ &\leqslant \gamma \left\|F^{1:m-1}x - F^{1:m-1}x'\right\|_\infty \\ &\leqslant \gamma \left\|x - x'\right\|_\infty ,\end{aligned}$$

where used for the last inequality the 1-Lipschitz continuous of each map $F_1, F_2, \ldots, F_{m-1}$ from Proposition 3.3.4. Taking the maximum over j yields

$$\left\|F^{1:M}x - F^{1:M}x'\right\|_\infty \leqslant \gamma \left\|x - x'\right\|_\infty .$$

□

Theorem 3.3.7. *Let $\gamma \in (0,1)$, $F : \mathbb{R}^d \to \mathbb{R}^d$ a γ-Lipschitz continuous map for $\|\cdot\|_\infty$, and $(J_k)_{k\geqslant 0}$ a sequence of sets so that each $j \in \{1,\ldots,d\}$ belongs to infinitely many sets. If $(x_k)_{k\geqslant 0}$ satisfies*

$$x_{k+1} = F^{|J_k}x_k,$$

then it converges to the unique fixed point of F.

Proof. Recursively define a sequence $(k_\ell)_{\ell\geqslant 0}$ of integers as follows. Let $k_0 = 0$ and k_1 be the smallest integer such that

$$\bigcup_{k=0}^{k_1} J_k = \{1,\ldots,d\} ,$$

which exists by assumption. Similarly for $\ell \geqslant 2$, let k_ℓ the smallest integer larger than $k_{\ell-1}$ such that

$$\bigcup_{k=k_{\ell-1}+1}^{k_\ell} J_k = \{1,\ldots,d\} .$$

Denote $F_k = F^{|J_k}$ for all $k \geqslant 0$ and $G_\ell = F_{k_{\ell+1}} \circ \cdots \circ F_{k_\ell}$ for all $\ell \geqslant 0$. Then we can apply Proposition 3.3.6 which gives that each map $G_{\ell+1}$ is γ-Lipschitz continuous for $\|\cdot\|_\infty$. Because $x_{k_{\ell+1}} = G_\ell x_{k_\ell}$ for all $\ell \geqslant 0$, by Theorem 3.3.1, we can write

$$\left\|x_{k_{\ell+1}} - x_*\right\| \leqslant \gamma \left\|x_{k_{\ell+1}} - x_*\right\| , \qquad \ell \geqslant 0,$$

where x_* is the unique fixed point of F. Moreover, using the fact that each map F_k ($k \geqslant 0$) is 1-Lipschitz continuous for $\|\cdot\|_\infty$ and has x_* as fixed point thanks to Propositions 3.3.4 and 3.3.5, we can write for $k > k_{\ell+1}$,

$$\begin{aligned}\|x_k - x_*\|_\infty &= \left\|(F_{k-1} \circ \cdots \circ F_{k_{\ell+1}+1})x_{k_{\ell+1}} - (F_{k-1} \circ \cdots \circ F_{k_{\ell+1}+1})x_*\right\|_\infty \\ &\leqslant \left\|x_{k_{\ell+1}} - x_*\right\|_\infty \leqslant \gamma \left\|x_{k_\ell} - x_*\right\|_\infty \\ &\leqslant \cdots \leqslant \gamma^{k_{\ell+1}} \|x_0 - x_*\|_\infty .\end{aligned}$$

Hence the convergence of x_k to x_* as $k \to +\infty$. □

3.4 Asynchronous value iterations

Definition 3.4.1 (Asynchronous value iterations)**.** Let π be a stationary policy.

(i) A sequence $(v_k)_{k\geqslant 0}$ in $\mathbb{R}^{\mathcal{S}}$ is an asynchronous state-value iteration for the evaluation of policy π (resp. for control) if there exists a sequence $(\mathcal{S}_k)_{k\geqslant 0}$ of subsets of $\mathcal{S}$ such that

$$v_{k+1} = B_\pi^{|\mathcal{S}_k} v_k, \qquad \left(\text{resp.} \quad v_{k+1} = B_*^{|\mathcal{S}_k} v_k\right).$$

$(\mathcal{S}_k)_{k\geqslant 0}$ is then called the sequence of updated states.

(ii) A sequence $(q_k)_{k\geqslant 0}$ in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is an asynchronous state-value iteration for the evaluation of policy π (resp. for control) if there exists a sequence $(\mathcal{Q}_k)_{k\geqslant 0}$ of subsets of $\mathcal{S} \times \mathcal{A}$ such that

$$q_{k+1} = B_\pi^{|\mathcal{Q}_k} q_k, \qquad \left(\text{resp.} \quad q_{k+1} = B_*^{|\mathcal{Q}_k} q_k\right).$$

$(\mathcal{Q}_k)_{k\geqslant 0}$ is then called the sequence of updated state-action pairs.

Proposition 3.4.2 (Convergence of asynchronous value iterations)**.** *Let π be a stationnary policy.*

(i) Let $(v_k)_{k\geqslant 0}$ be a state-value iteration for the evaluation of policy π (resp. for control) where each state is updated infinitely. Then, v_k converges to v_π (resp. v_*) as $k \to +\infty$.

(ii) Let $(q_k)_{k\geqslant 0}$ be a action-value iteration for the evaluation of policy π (resp. for control) where each state-action pair is updated infinitely. Then, q_k converges to q_π (resp. q_*) as $k \to +\infty$.

Proof. Follows from Theorem 3.3.7. □

Remark 3.4.3 (Single-component updates).

Chapter 4

Tabular reinforcement learning

4.1 Asynchronous stochastic approximations

4.2 Stochastic estimators of Bellman equations

4.3 Policy evaluation

4.4 Control