An Introduction to Reinforcement Learning

From theory to algorithms

Joon Kwon

November 19, 2025

Contents

Foreword

As of Fall 2024, this document contains lecture notes from a course given in Master 2 in *Université Paris–Saclay* since Fall 2023. 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 $\mathcal{R}$, the latter being a subset of $\mathbb{R}$. 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 action $A_2 \in \mathcal{A}$, possibly at random. The environment then draws $(R_2, S_2) \in \mathcal{R} \times \mathcal{S}$ according to a probability distribution which depends on S_1 and A_1, and so on.

The total reward of the agent is defined as $\sum_{t=1}^{+\infty} \gamma^{t-1} R_t$, 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 environment (which maps a state-action pair to a probability distribution over reward-state pairs) may be known or not.

This chapter presents basic notions regarding 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 $\mathcal{R} \subset \mathbb{R}$, and $p : \mathcal{S} \times \mathcal{A} \times \mathcal{R} \times \mathcal{S} \to [0, 1]$ 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.

Remark 1.2.2. When using policy π, $\pi^{(t)}(a|h^{(t)})$ is interpreted as the probability of the agent choosing action a at time t after having observed history $h^{(t)}$.

Definition 1.2.3. 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 $t \geqslant 0$, $\pi^{(t)}$ 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; $\pi^{(t)}$ 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 environment.

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 $S_0, A_0, R_1, S_1, A_1, R_2, \dots$. This is formalized by the proposition below.

We first introduce the following notation. For $T \geqslant 0$ and $h^{(T)} = (s_0, a_0, r_1, \dots r_T, s_T)$, 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 $h^{(T)} = (s_0, a_0, r_1, \dots, r_T, s_T) \in \mathcal{H}^{(T)}$,*

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

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

Sketch of proof. The above expression defines associates 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:

$$\begin{aligned}
&\mathbb{P}_{\mu,\pi}\left(\prod_{t=0}^{T}(\mathcal{S}_t \times \mathcal{A}_t \times \mathcal{R}_{t+1}) \times \mathcal{S}_{T+1} \times (\mathcal{A} \times \mathcal{R} \times \mathcal{S})^{\mathbb{N}}\right) \\
&\quad = \sum_{\substack{s_0 \in \mathcal{S}_0 \\ \vdots \\ s_{T+1} \in \mathcal{S}_{T+1}}} \sum_{\substack{a_0 \in \mathcal{A}_0 \\ \vdots \\ a_T \in \mathcal{A}_T}} \sum_{\substack{r_1 \in \mathcal{R}_1 \\ \vdots \\ r_{T+1} \in \mathcal{R}_{T+1}}} \mu(s_0) \prod_{t=0}^{T} \pi^{(t)}(a_t|h^{(t)})\, p(s_{t+1}, r_{t+1}|s_t, a_t).
\end{aligned}$$

$\mathbb{P}_{\mu,\pi}$ 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 $\mathcal{H}^{\infty}$. □

Remark 1.3.2. In particular, Proposition 1.3.1 implies that a measure on $\mathcal{H}^{\infty}$ coincide with $\mathbb{P}_{\mu,\pi}$ as soon as they coincide on sets of the form $\operatorname{Cyl} h^{(T)}$. This will be used in the proofs of Propositions 1.3.4 and 1.3.5 below.

Definition 1.3.3. Let $\mu \in \Delta(\mathcal{S})$, $\pi \in \Pi$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

(i) $\mathbb{P}_{\mu,\pi}$ from Proposition 1.3.1 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 $\tilde{\pi} = (\tilde{\pi}^{(t)})_{t \geqslant 0}$ be defined as

$$\begin{aligned}
\tilde{\pi}^{(0)}(s) &= \delta_a, \\
\tilde{\pi}^{(0)}(s') &= \pi^{(0)}(s') \quad \text{for } s' \neq s \\
\tilde{\pi}^{(t)} &= \pi^{(t)} \quad \text{for } t \geqslant 1.
\end{aligned}$$

$\mathbb{P}_{s,\tilde{\pi}}$ 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 by π only for $t \geqslant 1$. In general, it cannot be defined as $\mathbb{P}_{s,a}$ conditioned on the event $\{A_0 = a\}$ because the probability $\pi(a|s)$ of this event may be zero.

Proposition 1.3.4. *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 $\operatorname{Cyl} h^{(T)}$ 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 $h^{(T)} = (s_0, a_0, r_1, \dots, r_T, s_T) \in \mathcal{H}^{(T)}$, and denote $h^{(t)} := (s_0, a_0, r_1, \dots, r_r, s_t)$ 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 $\tilde{\pi}$ defined as in Definition 1.3.3. Then,

$$\begin{aligned}
\pi^{(0)}(a|s) \cdot \mathbb{P}_{s,a,\pi}\left(\operatorname{Cyl} h^{(T)}\right) &= \pi^{(0)}(a|s) \prod_{t=0}^{T-1} \tilde{\pi}^{(t)}(a_t|h^{(t)})\, p(r_{t+1}, s_{t+1}|s_t, a_t) \\
&= \mathbb{1}\{s_0 = s\} \prod_{t=0}^{T-1} \pi^{(t)}(a_t|h^{(t)})\, p(r_{t+1}, s_{t+1}|s_t, a_t) \\
&= \mathbb{1}\{a_0 = a\} \cdot \mathbb{P}_{s,\pi}\left(\operatorname{Cyl} h^{(T)}\right).
\end{aligned}$$

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

$$\sum_{a\in\mathcal{A}} \pi^{(0)}(a|s) \cdot \mathbb{P}_{s,a,\pi}\left(\operatorname{Cyl} h^{(T)}\right) = \mathbb{P}_{s,\pi}\left(\operatorname{Cyl} h^{(T)}\right).$$

□

The following proposition demonstrates that a given stationary policy induces a distribution over histories that has a Markov property in the following sense: for all $t \geqslant 0$, the distribution of $(S_t, A_t, R_{t+1}, \dots)$ conditionnally on $\{S_t = s\}$ has the same is the same as the distribution of $(S_0, A_0, R_1, \dots)$ when the latter has initial state s.

Proposition 1.3.5 (Markov property)**.** *Let $s, s' \in \mathcal{S}$, $a, a' \in \mathcal{A}$, π a stationary policy, $f : \mathcal{H}^\infty \to \mathbb{R}$ a bounded measurable function (with respect to the product σ-algebra), random variables $(S'_0, A'_0, R'_1, S'_2, A'_2, R'_2, \dots)$ with distribution $\mathbb{P}_{s,\pi}$ or $\mathbb{P}_{s,a,\pi}$, and $t \geqslant 0$.*

(i) If $\mathbb{P}[S'_t = s'] > 0$, the distribution of $(S'_t, A'_t, R'_{t+1}, S'_{t+1}, \dots)$ conditionnaly on $\{S'_t = s'\}$ is $\mathbb{P}_{s',\pi}$.

(ii) Almost-surely,

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

(iii) If $\mathbb{P}\left[S'_t = s',\ A'_t = a'\right] > 0$*, the distribution of* $(S'_t, A'_t, R_{t+1}, S'_{t+1}, \dots)$ *conditionnaly on* $\{S'_t = s',\ A'_t = a'\}$ *is* $\mathbb{P}_{s',a',\pi}$*.*

(iv) Almost-surely,

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

Proof. Let us assume $\mathbb{P}\left[S'_t = s'\right] > 0$. To prove (i), thanks to Proposition 1.3.1, it is enough to prove that $\mathbb{P}\left[\,\cdot \mid S'_t = s'\right]$ and $\mathbb{P}_{s',\pi}$ coincide on sets on the form $\operatorname{Cyl} h^{(T)}$. Let $T \geqslant t$ and $(s_t, a_t, r_{t+1}, \dots, r_T, s_T) \in \mathcal{H}^{(T-t)}$. Using the expression from the statement of Proposition 1.3.1,

$$\begin{aligned}
&\mathbb{P}\left[S'_t = s_t,\ A'_t = a_t,\ R'_{t+1} = r_{t+1}, \dots, R_T = r_T,\ S_T = s_T \,\middle|\, S'_t = s'\right] \\
&\quad= \frac{\mathbb{P}\left[S'_t = s',\ S'_t = s_t,\ A'_t = a_t,\ R'_{t+1} = r_{t+1}, \dots, R_T = r_T,\ S_T = s_T\right]}{\mathbb{P}\left[S'_t = s'\right]} \\
&\quad= \frac{\mathbb{1}_{\{s_0 = s'\}} \times \displaystyle\sum_{(s_0,\dots,s_{t-1}) \in \mathcal{S}^t} \delta_s(s_0) \prod_{\tau=0}^{T-1} \pi(a_\tau)\, p(r_{\tau+1}, s_{\tau+1} | s_\tau, a_\tau)}{\displaystyle\sum_{(s_0,\dots,s_{t-1}) \in \mathcal{S}^t} \delta_s(s_0) \prod_{t=0}^{T-1} \pi(a_\tau)\, p(r_{\tau+1}, s_{\tau+1} | s_\tau, a_\tau)} \\
&\quad= \mathbb{1}_{\{s_0 = s'\}} \times \prod_{\tau=t}^{T-1} \pi(a_\tau)\, p(r_{\tau+1}, s_{\tau+1} | s_\tau, a_\tau) \\
&\quad= \mathbb{P}_{s',\pi}\left[S_0 = s_t,\ A_0 = a_t,\ R_1 = r_{t+1}, \dots, R_{T-t} = r_T,\ S_{T-t} = s_T\right],
\end{aligned}$$

and (i) follows.

We now turn to (ii). By definition of the conditionnal expectation, $\mathbb{E}\left[f(S'_t, A'_t, R'_{t+1}, \dots) \,\middle|\, S'_t\right]$ designates any random variable, measurable with respect to S'_t and with expectation equal to $\mathbb{E}\left[f(S'_t, A'_t, R'_{t+1}, \dots)\right]$. Let us prove that $\mathbb{E}_{S'_t,\pi}\left[f(S_0, A_0, R_1, \dots)\right]$ indeed satisfy those properties. It is obviously measurable with respect to S'_t, as a deterministic function of the value of S'_t. Regarding the expectation, we can write

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}_{S_t,\pi}\left[f(S_0, A_0, R_1, \dots)\right]\right] &= \sum_{s' \in \mathcal{S}} \mathbb{P}\left[S'_t = s'\right] \times \mathbb{E}_{s',\pi}\left[f(S_0, A_0, R_1, \dots)\right] \\
&= \sum_{\substack{s' \in \mathcal{S} \\ \mathbb{P}[S'_t = s'] > 0}} \mathbb{P}\left[S'_t = s'\right] \times \mathbb{E}\left[f(S'_t, A'_t, R'_{t+1}, \dots) \,\middle|\, S'_t = s'\right] \\
&= \mathbb{E}\left[f(S'_t, A'_t, R'_{t+1}, \dots)\right].
\end{aligned}$$

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

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:

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

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 from 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

This chapter introduces Bellman operators, which are the fundamental tools for solving MDPs. 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

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

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

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

(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) $B_\pi^{(V,\gamma)} = E_\pi \circ D^{(\gamma)}$ (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.

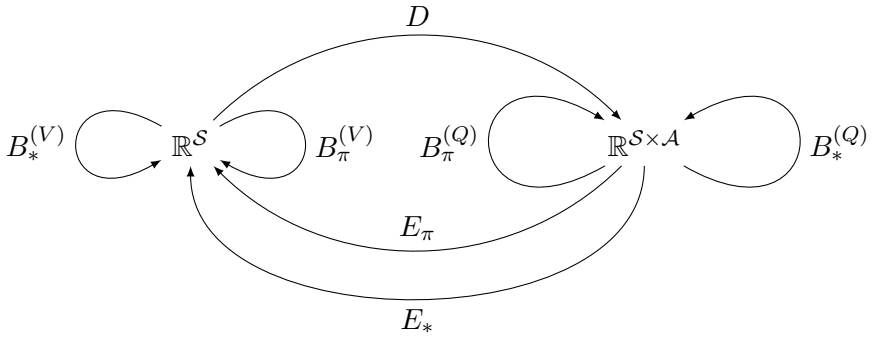

Figure 2.1: Operators $D, E_\pi, E_*, B_\pi^{(V)}, B_*^{(V)}, B_\pi^{(Q)}$ and $B_*^{(Q)}$.

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.*

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

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

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

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

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 stationary policy, $s \in \mathcal{S}$, $a \in \mathcal{A}$. Then,*

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

(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 $\tilde{\pi}$ the policy associated with (s,a) used in Definition 1.3.3 to define $\mathbb{P}_{s,a,\pi}$. Using the definition of the probability measure $\mathbb{P}_{s,\pi}$ (see Proposition 1.3.1),

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

We now turn to (ii).

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

We now deduce (iii) using Proposition 1.3.4:

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

For (iv), we combine (i) and (ii) with the help of the Markov property from Proposition 1.3.5; let $(S'_0, A'_0, R'_1, \dots) \sim \mathbb{P}_{s,a,\pi}$, then

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

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

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

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

Proof. (i) Let $s \in \mathcal{S}$ and $\pi \in \Pi_0$.

$$\begin{aligned}(E_\pi q)(s) &= \sum_{a\in\mathcal{A}} \pi(a|s)\, q(s,a) \leqslant \sum_{a\in\mathcal{A}} \pi(a|s) \max_{a'\in\mathcal{A}} q(s,a') \\ &= (E_* q)(s) \sum_{a\in\mathcal{A}} \pi(a|s) = (E_* q)(s).\end{aligned}$$

Taking the supremum over $\pi \in \Pi_0$ yields

$$\sup_{\pi\in\Pi_0} (E_\pi q)(s) \leqslant (E_* q)(s).$$

Besides, for each $s \in \mathcal{S}$, there exists a maximizer of $q(s, \cdot\,)$ (because the number of values is finite). Let $\pi_{0,d}(\,\cdot\,|s)$ be a Dirac at one of the maximizers. This defines a stationary and deterministic policy $\pi_{0,d}$, which satisfies $(E_{\pi_0} q)(s) = \max_{a\in\mathcal{A}} q(s,a)$ for all $s \in \mathcal{S}$. We then write for $s \in \mathcal{S}$,

$$\begin{aligned}\sup_{\pi\in\Pi_0} (E_\pi q)(s) \leqslant (E_* q)(s) &= \max_{a\in\mathcal{A}} q(s,a) = (E_{\pi_{0,d}} q)(s) \\ &\leqslant \sup_{\pi\in\Pi_{0,d}} (E_\pi q)(s) \leqslant \sup_{\pi\in\Pi_0} (E_\pi q)(s).\end{aligned}$$

The above lowest and highest quantities are the same. Therefore, all inequalities are equalities, and the supremums are maximums because they are attained by $\pi_{0,d}$.

Then (ii) and (iii) follow from 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:

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

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:

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

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.

Definition 2.2.2. Let $\eta > 0$, (X, d) and (Y, d') be metric spaces. A map $F : X \to Y$ is a η-contraction if $\gamma \in [0, 1)$ and F is η-Lipschitz continuous.

Theorem 2.2.3 (Banach's fixed point theorem)**.** *Let $0 \leqslant \eta < 1$, (X, d) a complete metric space, and $F : X \to X$ a η-contraction. 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*

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

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

Proof. For all $k \geqslant 1$, using the Lipschitz continuity of F,

$$d(x_{k+1}, x_k) = d(Fx_k, Fx_{k-1}) \leqslant \eta\, d(x_k, x_{k-1}),$$

which by a simple induction implies

$$d(x_{k+1}, x_k) \leqslant \eta^k \, d(x_1, x_0),$$

from which we deduce that $(x_k)_{k \geqslant 0}$ is a Cauchy sequence and thus admits a limit $x_* \in X$. Map F is continuous because of its Lipschitz property, and taking the limit in the identity $x_{k+1} = Fx_k$ yields $x_* = Fx_*$, in other words, x_* is indeed of fixed point of F. If $x_{**} \in X$ is also a fixed point, it holds that

$$d(x_*, x_{**}) = d(Fx_*, Fx_{**}) \leqslant \eta \, d(x_*, x_{**}),$$

which, because $0 \leqslant \eta < 1$, is only possible when $x_* = x_{**}$. The fixed point is therefore unique.

Besides, for all $k \geqslant 0$,

$$d(x_{k+1}, x_*) = d(Fx_k, Fx_*) \leqslant \eta \, d(x_k, x_*),$$

which by a simple induction yields

$$d(x_k, x_*) \leqslant \eta^k \, d(x_0, x_*).$$

□

Remark 2.2.4. The above convergence is guaranteed *regardless* of the initial point x_0.

Proposition 2.2.5. *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 a γ-contraction,

(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-contractions 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 2.2.6. *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* $B_\pi^{(Q)}$.

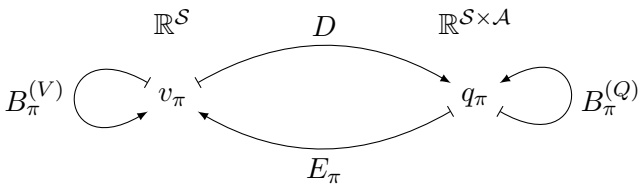

Figure 2.2: Relations between v_π, q_π, D, E_π $B_\pi^{(V)}$ and $B_\pi^{(Q)}$.

Proof. Let $s \in \mathcal{S}$. We prove (i) using Proposition 1.3.4:

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

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 operator D 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 1.3.5.

Combining (i) and (ii) together with Banach's fixed point theorem from Theorem (2.2.3) yields (iii) and (iv). □

Remark 2.2.7. 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 a state-value function $v \in \mathbb{R}^{\mathcal{S}}$ if it is greedy with respect to Dv.

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

Remark 2.3.2. $\pi_g[q]$ corresponds to a policy which selects actions by simply comparing values of the action-value function q. In the case of $\pi_g[v]$, 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 $\pi_g \in \Pi_g[q]$. By definition of a greedy policy,

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

Then,

$$B_*^{(Q)} q = D(E_* q) = D(E_{\pi_g} q) = B_{\pi_g} q$$

and similarly

$$B_*^{(V)} v = E_*(Dv) = E_{\pi_g}(Dv) = B_\pi v.$$

□

2.4 Optimal value functions and policies

Definition 2.4.1. Let $\gamma \in (0,1)$. The *optimal state-value* and *actions-value functions* for 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) $q_* = D v_*$.

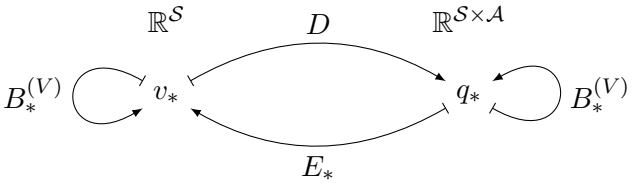

Figure 2.3: Relations between v_*, q_*, D, E_* $B_*^{(V)}$ and $B_*^{(Q)}$.

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 2.2.6.

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 $(S_0, A_0, R_1, S_1, A_1, R_2, \dots) \sim \mathbb{P}_{s,\pi}$. Let $t \geqslant 0$,

$$v_0(S_t) = (B_* v_0)(S_t) = \max_{a \in \mathcal{A}} (Dv_0)(S_t, a) \geqslant (Dv_0)(S_t, A_t).$$

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 $\mathbb{P}_{s,\pi}$ (meaning the expression from Proposition 1.3.1),

$$\begin{aligned}
(Dv_0)(s_0, a_0) &= \sum_{(r,s') \in \mathcal{R} \times \mathcal{S}} p(r, s' | s_0, a_0)(r + \gamma\, v_0(s')) \\
&= \sum_{(r,s') \in \mathcal{R} \times \mathcal{S}} \frac{\mathbb{P}[R_{t+1} = r,\ S_{t+1} = s',\ S_t = s_0,\ A_t = a_0]}{\mathbb{P}[S_t = s_0,\ A_t = a_0]} (r + \gamma\, v_0(s')) \\
&= \mathbb{E}[R_{t+1} + \gamma\, v_0(S_{t+1}) \,|\, S_t = s_0, A_t = a_0].
\end{aligned}$$

We deduce

$$(Dv_0)(S_t, A_t) = \mathbb{E}[R_{t+1} + \gamma\, v_0(S_{t+1}) \,|\, S_t, A_t],$$

and

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

Then using the expresion of B_* from Proposition 2.1.7 and the expression

of B_π from Proposition 2.1.3, applying the above recursively, we get

$$\begin{aligned} v_0(s) = (B_* v_0)(s) &= \sup_{\pi' \in \Pi} (B_{\pi'} v_0)(s) \\ &\geqslant (B_\pi v_0)(s) = \mathbb{E}\left[R_1 + \gamma\, v_0(S_1)\right] \\ &\geqslant \mathbb{E}\left[R_1 + \gamma \mathbb{E}\left[R_2 + \gamma\, v_0(S_2) \,|\, S_1, A_1\right]\right] \\ &= \mathbb{E}\left[R_1 + \gamma R_2 + \gamma^2 v_0(S_2)\right] \\ &\geqslant \cdots \geqslant \mathbb{E}_{s,\pi}\left[\sum_{t=1}^{T} \gamma^{t-1} R_t + \gamma^T v_0(S_T)\right]. \end{aligned}$$

Using the dominated convergence theorem, we can take the limit in the above last expectation, as for all $T \geqslant 1$, almost-surely,

$$\begin{aligned} \left|\sum_{t=1}^{T} \gamma^{t-1} R_t + \gamma^\top v_0(S_T)\right| &\leqslant \sum_{t=1}^{T} \gamma^{t-1} |\mathcal{R}| + \gamma^T \|v_0\|_\infty \\ &\leqslant |\mathcal{R}| \sum_{t=1}^{+\infty} \gamma^{t-1} + \|v_0\|_\infty, \end{aligned}$$

which provides an integrable upper bound, where $|\mathcal{R}| := \max_{r \in \mathcal{R}} |r|$. Hence,

$$v_0(s) \geqslant \mathbb{E}\left[\sum_{t=1}^{+\infty} \gamma^{t-1} R_t\right] = v_\pi(s).$$

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 inequalities are equalities, and the supremums are maximums because they are attained for $\pi_g \in \Pi_{0,d} \subset \Pi$.

Then, we write

$$\begin{aligned} q_* = \sup_{\pi \in \Pi} q_\pi &\geqslant \max_{\pi \in \Pi_{0,d}} q_\pi \geqslant q_{\pi_g} = q_0 = D v_0 \\ &= D\left(\max_{\pi \in \Pi} v_\pi\right) \geqslant \sup_{\pi \in \Pi} D v_\pi = \sup_{\pi \in \Pi} q_\pi = q_*, \end{aligned}$$

where the last inequality holds by monotonicity of D from Proposition 2.1.6 (by writing for $\pi \in \Pi$, $D(\max_{\pi \in \Pi} v_\pi) \geqslant D v_\pi$ 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 construction 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[1] 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 stationary 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 optimal policy. The (synchronous) value iteration for control approximately computes v_* (resp. q_*) by iterating the Bellman expectation operator $B_*^{(V)}$ (resp. $B_*^{(Q)}$) and then considers a greedy policy.

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

$$v_{k+1} = B_\pi v_k, \qquad (\text{resp.} \quad q_{k+1} = B_\pi q_k)$$

[1] as in *sampling*

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

$$v_{k+1} = B_* v_k, \qquad (\text{resp.} \quad q_{k+1} = B_* q_k)\,.$$

Algorithm 1: Synchronous value iteration for state-value functions for policy evaluation.

Input: Stationary policy π, initial value function $v \in \mathbb{R}^{\mathcal{S}}$, number of iterations $n \geqslant 1$.
for $k = 1, \ldots, n$ **do**
  $v' \leftarrow v$
  for $s \in \mathcal{S}$ **do**
    $v(s) \leftarrow \sum_{(a,r,s')\in\mathcal{A}\times\mathcal{S}\times\mathcal{R}} \pi(a|s)p(r,s'|s,a)(r+\gamma\, v'(s'))$
return $v, \pi_g\,[v]$

Algorithm 2: Synchronous value iteration for state-value functions for control.

Input: Initial value function $v \in \mathbb{R}^{\mathcal{S}}$, number of iterations $n \geqslant 1$.
for $k = 1, \ldots, n$ **do**
  $v' \leftarrow v$
  for $s \in \mathcal{S}$ **do**
    $v(s) \leftarrow \max_{a\in\mathcal{A}} \sum_{(r,s')\in\mathcal{S}\times\mathcal{R}} p(r,s'|s,a)(r+\gamma\, v'(s'))$
return $v, \pi_g\,[v]$

Proposition 3.1.2 (Equivalence between synchronous state-value and action-value iterations)**.** *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. 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,

Algorithm 3: Synchronous value iteration for action-value functions policy evaluation.

Input: Stationary policy π, initial value function $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, number of iterations $n \geqslant 1$.

for $k = 1, \ldots, n$ **do**

  $q' \leftarrow q$

  for $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**

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

return q

Algorithm 4: Synchronous value iteration for action-value functions for control.

Input: Initial value function $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, number of iterations $n \geqslant 1$.

for $k = 1, \ldots, n$ **do**

  $q' \leftarrow q$

  for $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**

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

return $q, \pi_g\,[q]$

(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 $k \geqslant 0$,

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

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

Proposition 3.1.3 (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 2.2.6 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 a γ-contraction with respect to $\|\cdot\|_\infty$ according to Proposition 2.2.5. The Banach's fixed point theorem (Theorem 2.2.3) then applies and gives the result. □

Remark 3.1.4 (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, 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 2.2.6), 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 3.1.3, taking the limit as $k \to +\infty$ gives (i). Besides, using the monotonicity of D from Proposition 2.1.6, together with Proposition 2.2.6 gives (ii):

$$q_\pi = D v_\pi \leqslant D v_{\pi_g} = q_{\pi_g}.$$

We now turn to (iii) and assume $v_{\pi_g} = v_\pi$. Using Propositions 2.2.6 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 stationary policies is a *policy iteration* if $\pi_{k+1} \in \Pi_g [v_{\pi_k}]$ 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 presented in Algorithm 5 and discussed in Remark 3.2.6 below.

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

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

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

$$\Pi_g [v_{\pi_k}] = \Pi_g [D v_{\pi_k}] = \Pi_g [q_{\pi_k}],$$

where we used Proposition 2.2.6 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 $\Pi_g [v_{\pi_k}]$ requires by definition the computation of $D v_{\pi_k} = q_{\pi_k}$.

Algorithm 5: An approximate policy iteration

Input: Initial stationary and deterministic policy π, initial value function $v \in \mathbb{R}^{\mathcal{S}}$, number of inner iterations for approximate policy evaluation $m \geqslant 1$, number of iterations $n \geqslant 1$.

for $k = 1, \dots, n$ **do**
  for $\ell = 1, \dots, m$ **do**
    $v' \leftarrow v$
    for $s \in \mathcal{S}$ **do**
      $v(s) \leftarrow \sum_{(r,s') \in \mathcal{R} \times \mathcal{S}} p(r, s' | s, \pi(a))(r + \gamma\, v'(s'))$
  for $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
    $q(s, a) \leftarrow \sum_{(r,s') \in \mathcal{R} \times \mathcal{S}} p(r, s' | s, a) \left(r + \gamma\, v(s')\right)$
  $\pi(s) \leftarrow \arg\min_{a \in \mathcal{A}} q(s, a)$
return π

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 2.2.5,

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

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 for all $k\geqslant 0$,

$$\begin{aligned} \pi_k &\in \Pi_g\left[v_k\right], \qquad (\text{resp.} \quad \pi_k \in \Pi_g\left[q_k\right]) \\ v_{k+1} &= B_{\pi_k}^{m_k} v_k, \qquad \left(\text{resp.} \quad q_{k+1} = B_{\pi_k}^{m_k} q_k\right), \end{aligned}$$

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 $k\geqslant 0$). An approximate policy iteration where m_k may be large but not infinite, as in Algorithm 5, belongs to this family.

3.3 Asynchronous fixed point iterations

We consider generalized fixed point iterations where a given operator is used to only update, at each step, a subset of the components. This is called *asynchronous* updates, which corresponds to the idea that an updated component need not wait for all other components to be updated before being used. Using updated information more quickly may lead to faster convergence in practice, which is sometimes called the *Gauss-Seidel effect*. This is especially true when the number of components is large. We give a sufficient condition for convergence in the case of a contraction for $\|\cdot\|_\infty$. This is then applied in Section 3.4 to obtain asynchronous variants of value iterations.

Theorem 3.3.1 (A generalized fixed point theorem)**.** *Let (X,d) be a metric space, $(\gamma_k)_{k\geqslant 0}$ a nonnegative sequence 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 for each $k\geqslant 0$. If $(x_k)_{k\geqslant 0}$ satisfies $x_{k+1}=F_k x_k$ for all $k\geqslant 0$, then*

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

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

Proof. Let $k\geqslant 0$.

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

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

$$(F^{|J}x)_j = \begin{cases} (Fx)_j & \text{if } j \in J \\ x_j & \text{if } j \notin J, \end{cases} \qquad 1 \leqslant j \leqslant d.$$

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 $F : \mathbb{R}^d \to \mathbb{R}^d$ 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 3.3.5. *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 3.3.6. *Let $\eta \in [0, 1]$ and $F : \mathbb{R}^d \to \mathbb{R}^d$ a η-Lipschitz continuous map for $\|\cdot\|_\infty$ and $M \geqslant 1$ an integer. 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 η-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,

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

where we used the η-contraction property of F and for the last inequality the 1-Lipschitz continuity of each map $F_1, F_2, \ldots, F_{m-1}$ from Proposition 3.3.5. Taking the maximum over j yields

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

□

Theorem 3.3.7. *Let $\eta \in (0,1)$, $F : \mathbb{R}^d \to \mathbb{R}^d$ a η-contraction 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 an increasing sequence of integers $(k_\ell)_{\ell\geqslant 0}$ as follows. Let $k_0 = 0$ and k_1 be the smallest integer such that

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

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

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

Denote $F_k = F^{|J_k}$ for all $k \geqslant 0$ and $G_\ell = F_{k_{\ell+1}-1} \circ \cdots \circ F_{k_\ell+1} \circ F_{k_\ell}$ for all $\ell \geqslant 0$. Then we can apply Proposition 3.3.6 which gives that each map G_ℓ is a η-contraction 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

$$\|x_{k_\ell} - x_*\| \leqslant \eta^\ell \|x_0 - x_*\|, \qquad \ell \geqslant 0,$$

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 3.3.5 and 3.3.4, we can write for $k > k_\ell$,

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

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

Example 3.3.8. In practice, an important special case is where only one component is updated at each iteration. If moreover, the components are updated in a cyclic fashion, the iteration is sometimes called nonlinear Gauss-Seidel.

3.4 Asynchronous value iterations

We consider asynchronous variants of 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 stationary 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 □

Algorithm 6: Asynchronous value iteration for state-value functions for policy evaluation.

Input: Stationary policy π, initial value function $v \in \mathbb{R}^{\mathcal{S}}$, number of iterations $n \geqslant 1$.

for $k = 1, \ldots, n$ **do**
- $v' \leftarrow v$
- Choose $\mathcal{S}_0 \subset \mathcal{S}$
- **for** $s \in \mathcal{S}_0$ **do**
  - $v(s) \leftarrow \sum_{(a,r,s') \in \mathcal{A} \times \mathcal{S} \times \mathcal{R}} \pi(a|s) p(r, s'|s, a)(r + \gamma\, v'(s'))$

return v

Algorithm 7: Asynchronous value iteration for state-value functions for control.

Input: Initial value function $v \in \mathbb{R}^{\mathcal{S}}$, number of iterations $n \geqslant 1$.

for $k = 1, \ldots, n$ **do**
- $v' \leftarrow v$
- Choose $\mathcal{S}_0 \subset \mathcal{S}$
- **for** $s \in \mathcal{S}_0$ **do**
  - $v(s) \leftarrow \max_{a \in \mathcal{A}} \sum_{(r,s') \in \mathcal{S} \times \mathcal{R}} p(r, s'|s, a)(r + \gamma\, v'(s'))$

return $v, \pi_g\,[v]$

Algorithm 8: Asynchronous value iteration for action-value functions for policy evaluation.

Input: Stationary policy π, initial value function $q \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$, number of iterations $n \geqslant 1$.

for $k = 1, \ldots, n$ **do**
- $q' \leftarrow q$
- Choose $\mathcal{Q} \subset \mathcal{S} \times \mathcal{A}$
- **for** $(s, a) \in \mathcal{Q}$ **do**
  - $q(s, a) \leftarrow \sum_{(r,s',a') \in \mathcal{R} \times \mathcal{S} \times \mathcal{A}} p(r, s'|s, a) \left(r + \gamma \pi(a'|s') q'(s', a')\right)$

return q

Algorithm 9: Asynchronous value iteration for action-value functions for control.

Input: Initial value function $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, number of iterations $n \geqslant 1$.
for $k = 1, \ldots, n$ **do**
  $q' \leftarrow q$
  Choose $\mathcal{Q} \subset \mathcal{S} \times \mathcal{A}$
  for $(s, a) \in \mathcal{Q}$ **do**
    $$q(s,a) \leftarrow \sum_{(r,s')\in\mathcal{R}\times\mathcal{S}} p(r, s'|s, a) \left(r + \gamma \max_{a'\in\mathcal{A}} q'(s', a') \right)$$
return $q, \pi_g\,[q]$

Chapter 4

Tabular reinforcement learning

Starting from this chapter, we consider MDPs with unknown dynamics, in the sense that the algorithms we are allowed to consider may interact with the environement but do not have access to the transition dynamic p in an explicit form. In particular, the operator D, and therefore the Bellman operators cannot be computed exactly. The image of a value function by a Bellman operator will then be replaced by a stochastic estimator, and the replacement-based updates from the deterministic fixed point iterations will be generalized into averaging-based stochastic ones.

Another consequence of operator D being unavailable is that we cannot determine a greedy policy with respect to a state-value function $v \in \mathbb{R}^{\mathcal{S}}$. For that reason, we will rely on action-value functions instead. An alternative, which will be dealt with later, is *model-based methods* which estimate the transition dynamics through interaction, and which can then derive an approximatively greedy policy with respect to a state-value function.

This chapter focuses on *tabular* methods, which work by manipulating whole action-value functions $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and whole stationnary policies $\pi \in \Delta(\mathcal{A})^{\mathcal{S}}$. In the following chapters, we will study methods that approximate action-value functions and/or policies with parametric families for better scalability.

4.1 Stochastic asynchronous fixed point iterations

The asynchronous fixed-point iterations in $\mathbb{R}^d$ ($d \geqslant 1$) from Section 3.3 can be written as

$$x_{t+1} = (1 - \mathbb{1}_{J_k}) \otimes x_k + \mathbb{1}_{J_k} \otimes F x_k, \quad k \geqslant 0,$$

where $\otimes$ denotes component-wise multiplication, J_k is the set of components that are updated at iteration k, and $\mathbb{1}_{J_k}$ the correpsonding indicator vec-

tor, meaning for all $1 \leqslant j \leqslant d$, $(\mathbb{1}_{J_k})_j = \mathbb{1}_{\{j \in J_k\}}$. This expression easily generalises into an averaging procedure of the form

$$x_{t+1} = (1 - \alpha_k) \otimes x_t + \alpha_k \otimes \hat{f}_k, \quad k \geqslant 0,$$

where α_k is a vector in $[0,1]^d$, which is sometimes called a *stochastic approximation* procedure and presented as a method which computes an approximate *zero* of operator $F - \mathrm{Id}$, if $\hat{f}_k$, conditionally on x_k is an unbiased estimator of Fx_k. In addition to being a stochastic generalization of fixed-point iterations, it can also be seen as an extension of basic mean estimation. If $(Z_k)_{k \geqslant 0}$ are i.i.d. random vectors with common mean $\mu \in \mathbb{R}^d$, then μ is the unique fixed point of operator $x \mapsto \mu$, and procedure $x_{k+1} = (1 - \frac{1}{k+1})x_k + \frac{1}{k+1} Z_k$ is equivalent to simple averaging $x_{k+1} = \frac{1}{k+1} \sum_{\ell=0}^{k} Z_\ell$.

We give without proof the following convergence guarantee.

Theorem 4.1.1 (Tsitsiklis, 1994)**.** *Let $d \geqslant 1$, $\gamma \in (0,1)$, $F : \mathbb{R}^d \to \mathbb{R}^d$ a γ-contraction with respect to $\|\cdot\|_\infty$, $(x_k)_{k \geqslant 0}$, $(\alpha_k)_{k \geqslant 0}$ and $(\hat{f}_k)_{k \geqslant 0}$ sequences of random vectors in $\mathbb{R}^d$ and $(\mathcal{F}_k)_{k \geqslant 0}$ a filtration. We assume that*

(i) $x_{k+1} = (1 - \alpha_k) \otimes x_k + \alpha_k \otimes \hat{f}_k$, *for all* $k \geqslant 0$,

(ii) x_k *and* α_k *are* $\mathcal{F}_k$*-measurable, for all* $k \geqslant 0$,

(iii) there exists $c_1, c_2 \geqslant 0$ *such that for all* $k \geqslant 0$ *and* $1 \leqslant j \leqslant d$,

$$\alpha_{k,j} \neq 0 \implies \begin{cases} \mathbb{E}\left[\hat{f}_{k,j} \middle| \mathcal{F}_k\right] = (Fx_k)_j, \\ \mathrm{Var}\left(\hat{f}_{k,j} \middle| \mathcal{F}_k\right) \leqslant c_1 + c_2 \|x_k\|_\infty^2, \end{cases}$$

(iv) for all $1 \leqslant j \leqslant d$, *almost-surely,* $(\alpha_{k,j})_{k \geqslant 0}$ *is a nonnegative sequence and*

$$\sum_{k=0}^{+\infty} \alpha_{k,j} = +\infty \quad \text{and} \quad \sum_{k=0}^{+\infty} \alpha_{k,j}^2 < +\infty.$$

Then, $(x_k)_{k \geqslant 0}$ *converges almost-surely to the unique fixed point of* F.

4.2 Stochastic estimators of Bellman operators

We know from Proposition 2.1.3 that Bellman expectation operators can be written as expectations. We derive similar expressions for $(B_\pi^{(V)})^T$ and $(B_\pi^{(Q)})^T$ (for $T \geqslant 1$) that involve an expectation of the discounted sum of rewards up to time T *plus* an approximation of the remaining rewards. We then also derive a similar expression for $B_*^{(Q)}$. Those will be used to construct stochastic estimators of Bellman operators.

Proposition 4.2.1. *Let $T \geqslant 1$, π a stationary policy, $v \in \mathbb{R}^{\mathcal{S}}$, $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$.*

(i) $$(B_\pi^T v)(s) = \mathbb{E}_{s,\pi}\left[\sum_{t=1}^{T} \gamma^{t-1} R_t + \gamma^T v(S_T)\right].$$

(ii) $$(B_\pi^T q)(s,a) = \mathbb{E}_{s,a,\pi}\left[\sum_{t=1}^{T} \gamma^{t-1} R_t + \gamma^T q(S_T, A_T)\right].$$

(iii) $$(B_*^T v)(s) = \max_{\pi' \in \Pi_{0,d}} \mathbb{E}_{s,\pi'}\left[\sum_{t=1}^{T} \gamma^{t-1} R_t + \gamma^T v(S_T)\right].$$

(iv) $$(B_*^T q)(s,a) = \max_{\pi' \in \Pi_{0,d}} \mathbb{E}_{s,a,\pi'}\left[\sum_{t=1}^{T} \gamma^{t-1} R_t + \gamma^T q(S_T, A_T)\right].$$

Proof. We proceed by induction. Property (i) is true for $T = 1$ by Proposition 2.1.3. For $T \geqslant 2$, let $(S'_0, A'_0, R'_0, \dots) \sim \mathbb{P}_{s,\pi}$ and we write

$$\begin{aligned}
(B_\pi^T)(s) = (B_\pi^{T-1} B_\pi v)(s) &= \mathbb{E}\left[\sum_{t=1}^{T-1} \gamma^{t-1} R'_t + \gamma^{T-1} (B_\pi v)(S'_{T-1})\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T-1} \gamma^{t-1} R'_t + \gamma^{T-1} \mathbb{E}_{S'_{T-1},\pi}\left[R_1 + \gamma v(S_1)\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T-1} \gamma^{T-1} R'_t + \gamma^{T-1} \mathbb{E}\left[R'_T + \gamma v(S'_T) \,\middle|\, S'_{T-1}\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} R_t + \gamma^T v(S'_T)\right],
\end{aligned}$$

where we used the Markov property from Proposition 1.3.5 to get the third line. This proves property (i). Property (ii) is proved similarly. Then, properties (iii) and (iv) immediatly follow from Proposition 2.1.7. □

Regarding Bellman optimality operators, the above expressions for $(B_*^{(V)})^T$ and $(B_*^{(V)})^T$ are not *a priori* written as expectations, but as a maximums of expectations, and therefore do not yield straightforward constructions for unbiaised stochastic estimators. In the special case of $B_*^{(Q)}$ however (with $T = 1$), the following corollary does give such an expression.

Corollary 4.2.2. *For $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$(B_* q)(s,a) = \mathbb{E}_{(R,S') \sim p(\,\cdot\,|s,a)}\left[R + \gamma \max_{a' \in \mathcal{A}} q(S', a')\right].$$

Proof. Proposition 4.2.1 gives

$$(B_* q)(s,a) = \max_{\pi \in \Pi_{0,d}} \mathbb{E}_{s,a,\pi} \left[R_1 + \gamma q(S_1, A_1) \right].$$

Let $\pi \in \Pi_{0,d}$ and $(S'_0, A'_0, R'_1, \dots) \sim \mathbb{P}_{s,a,\pi}$. Then,

$$\begin{aligned}
\mathbb{E}_{s,a,\pi} \left[R_1 + \gamma q(S_1, A_1) \right] &= \mathbb{E} \left[R'_1 + \gamma q(S'_1, A'_1) \right] \\
&\leqslant \mathbb{E} \left[R'_1 + \gamma \max_{a' \in \mathcal{A}} q(S'_1, a') \right] \\
&= \mathbb{E} \left[R'_1 + \gamma q(S'_1, \pi_g \left[q \right] (S'_1)) \right] \\
&= \mathbb{E}_{s,a,\pi_g} \left[R_1 + \gamma q(S_1, A_1) \right],
\end{aligned}$$

where $\pi_g \in \Pi_g \left[q \right]$. Therefore, taing the maximum over $\pi \in \Pi_{0,d}$ and because $\pi_g \in \Pi_{0,d}$ by definition of greedy policies, π_g attains the maximum and

$$\begin{aligned}
(B_* q)(s,a) = \max_{\pi \in \Pi_{0,d}} \mathbb{E}_{s,a,\pi} \left[R_1 + \gamma q(S_1, A_1) \right] \\
\mathbb{E}_{s,a,\pi_g} \left[R_1 + \gamma q(S_1, A_1) \right] \\
= \mathbb{E} \left[R'_1 + \gamma \max_{a \in \mathcal{A}} q(S'_1, a') \right] \\
= \mathbb{E}_{(R,S') \sim p(\,\cdot\,|s,a)} \left[R + \gamma \max_{a \in \mathcal{A}} q(S', a') \right],
\end{aligned}$$

where the last equality stands because $(R'_1, S'_1) \sim p(\,\cdot\,|s,a)$ as an immediate consequence of the definition of $\mathbb{P}_{s,a,\pi}$ (meaning the expression from Proposition 1.3.1). □

Definition 4.2.3. For $T \geqslant 1$, we call *history of length* T^+ a tuple of the form:

$$(S_0, A_0, R_1, \dots, S_{T-1}, A_{T-1}, R_T, S_T, A_T) \in (\mathcal{S} \times \mathcal{A} \times \mathcal{R})^{T-1} \times \mathcal{S} \times \mathcal{A}.$$

We denote $\mathcal{H}^{(T^+)}$ the set of histories of length T^+.

Definition 4.2.4. For $T \geqslant 1$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$, and $\pi \in \Pi_0$, we denote $\mathbb{P}_{s,a,\pi}^{(T^+)}$ the probability distribution on $\mathcal{H}^{(T^+)}$ induced by $\mathbb{P}_{s,a,\pi}$. By convention and for consistency, we denote $\mathcal{H}^{(\infty^+)} = \mathcal{H}^{(\infty)}$ and $\mathbb{P}_{s,a,\pi}^{(\infty^+)} = \mathbb{P}_{s,a,\pi}$.

For $T \geqslant 1$, $H = (S_0, A_0, R_1, \dots S_T, A_T) \in \mathcal{H}^{(T^+)}$ and $q \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$, we denote

$$(\hat{B}^T q)(H) = \sum_{t=1}^{T} \gamma^{t-1} R_t + \gamma^T q(S_T, A_T),$$

and if $H \sim \mathbb{P}_{s,a,\pi}^{(T^+)}$ for some $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $\pi \in \Pi_0$, according to Proposition 4.2.1, $(\hat{B}^T q)(H)$ is an unbiased estimator of $(B_\pi^T q)(s,a)$.

For consistency, we also denote for $H = (S_0, A_0, R_1, \dots) \in \mathcal{H}^{(\infty)}$

$$(\hat{B}^{\infty} q)(H) = \sum_{t=1}^{+\infty} \gamma^{t-1} R_t,$$

which does not depend on q, and which is an unbiased estimator of $q_\pi = \lim_{T \to +\infty} B_\pi^T q$ by definition of the latter (as soon as $H \sim \mathbb{P}_{s,a,\pi}$).

For $(R, S') \in \mathcal{R} \times \mathcal{S}$, we denote

$$(\hat{B}_* q)(R, S') = R + \gamma \max_{a \in \mathcal{A}} q(S', a),$$

and if $(R, S') \sim p(\cdot\,|s, a)$ for some $(s, a) \in \mathcal{S} \times \mathcal{A}$, then $(\hat{B}_* q)(R, S')$ is an unbiased estimator of $(B_* q)(s, a)$ thanks to Corollary 4.2.2.

The following proposition establishes upper bounds on the variance of those estimators.

Proposition 4.2.5. *Let $T \geqslant 0$, π a stationary policy, $q \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$. Denote $M_{\mathcal{R}} = \sup_{r \in \mathcal{R}} |r|$.*

(i) If $H \sim \mathbb{P}_{s,a,\pi}^{(T^+)}$, then

$$\operatorname{Var}\Big((\hat{B}^T q)(H)\Big) \leqslant \frac{4M_{\mathcal{R}}^2}{(1-\gamma)^2} + \frac{2}{1-\gamma} \|q\|_\infty^2 .$$

(ii) If $(R, S') \sim p(\cdot\,|s, a)$, then

$$\operatorname{Var}\Big((\hat{B}_* q)(R, S')\Big) \leqslant 2M_{\mathcal{R}}^2 + 2\gamma^2 \|q\|_\infty^2 .$$

Proof. Denote $H = (S_0, A_0, R_1, \dots, S_T, A_T)$. Note that for all $1 \leqslant t \leqslant T$,

$$\operatorname{Var}(R_t) \leqslant \mathbb{E}\left[R_t^2\right] \leqslant M_{\mathcal{R}}^2,$$

and

$$\operatorname{Var}(q(S_T, A_T)) \leqslant \mathbb{E}\left[q(S_T, A_T)^2\right] \leqslant \|q\|_\infty^2 ,$$

and similarly $\operatorname{Var}(\max_{a \in \mathcal{A}} q(S, a)) \leqslant \|q\|_\infty^2$. Then, using formula

$$\operatorname{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \operatorname{Var} X_i + 2 \sum_{i<j} \operatorname{Cov}(X_i, X_j)$$

and Cauchy-Schwarz inequality which yields

$$\operatorname{Cov}(X, Y) \leqslant \sqrt{(\operatorname{Var} X)(\operatorname{Var} Y)},$$

we write

$$\begin{aligned}
\operatorname{Var}\left((\hat{B}^T q)(H)\right) &= \operatorname{Var}\left(\sum_{t=1}^{T} \gamma^{t-1} R_t + \gamma^T q(S_T, A_T)\right) \\
&= \sum_{t=1}^{T} \operatorname{Var}\left(\gamma^{t-1} R_t\right) + \operatorname{Var}\left(\gamma^T q(S_T, A_T)\right) + \\
&\quad + 2 \sum_{1 \leqslant t < t' \leqslant T} \operatorname{Cov}\left(\gamma^{t-1} R_t, \gamma^{t'-1} R_{t'}\right) \\
&\quad + 2 \sum_{t=1}^{T} \operatorname{Cov}(\gamma^{t-1} R_t, \gamma^T q(S_T, A_T)) \\
&\leqslant M_{\mathcal{R}}^2 \sum_{t=1}^{T} \gamma^{2t-2} + \gamma^{2T} \|q\|_\infty^2 + 2M_{\mathcal{R}}^2 \sum_{1 \leqslant t < t' \leqslant T} \gamma^{t+t'-2} \\
&\quad + \sum_{t=1}^{T} \gamma^{T+t-1} \left(M_{\mathcal{R}}^2 + \|q\|_\infty^2\right).
\end{aligned}$$

(i) follows by simplifying. We now turn to (ii):

$$\begin{aligned}
\operatorname{Var}\left((\hat{B}_* q)(R, S')\right) &= \operatorname{Var}\left(R + \gamma \max_{a \in \mathcal{A}} q(S', a)\right) \\
&\leqslant \operatorname{Var} R + \gamma^2 \operatorname{Var}\left(\max_{a \in \mathcal{A}} q(S', a)\right) \\
&\qquad + 2 \operatorname{Cov}(R,\ \gamma \max_{a \in \mathcal{A}} q(S', a)) \\
&\leqslant M_{\mathcal{R}}^2 + \gamma^2 \|q\|_\infty^2 + 2\sqrt{(\operatorname{Var} R)\left(\operatorname{Var}\left(\gamma \max_{a \in \mathcal{A}} q(S', a)\right)\right)} \\
&\leqslant M_{\mathcal{R}}^2 + \gamma^2 \|q\|_\infty^2 + \operatorname{Var} R + \operatorname{Var}\left(\gamma \max_{a \in \mathcal{A}} (S', a)\right) \\
&\leqslant 2M_{\mathcal{R}}^2 + 2\gamma^2 \|q\|_\infty^2 .
\end{aligned}$$

□

4.3 Policy evaluation

In this section, we study methods that compute the action-value function q_π of a stationary policy π by combining stochastic synchronous fixed point iteration from Section 4.1 and estimators of Bellman expectation operator $B_\pi^{(Q)}$ from Section 4.2.

Let us first give an informal description. Let $T \in \{1, 2, \dots\} \cup \{+\infty\}$. For each $k \geqslant 0$, we choose a state-action pair $(S_{0,k}, A_{0,k})$, possibly at random

as a function of previous observations. Starting from initial state $S_{0,k}$ and action $A_{0,k}$, we generate a history of length T^+ by using policy π:

$$H_k = (S_{0,k}, A_{0,k}, R_{1,k}, \dots, S_{T,k}, A_{T,k}) \,|\, S_{0,k}, A_{0,k} \sim \mathbb{P}^{(T^+)}_{S_{0,k},A_{0,k},\pi},$$

and compute the corresponding estimator T-step Bellman expectation operator:

$$(\hat{B}^T q_k)(H_k) = \sum_{t=1}^{T} \gamma^{t-1} R_{t,k} + \gamma^T q_k(S_{T,k}, A_{T,k}).$$

We asynchronously update the component $(S_{0,k}, A_{0,k})$ of the action-value function so that the new value is the Cesarò average of the corresponding estimators computed so far:

$$q_{k+1}(S_{0,k}, A_{0,k}) = \frac{\displaystyle\sum_{\ell=0}^{k} \mathbb{1}\left\{(S_{0,\ell}, A_{0,\ell}) = (S_{0,k}, A_{0,k})\right\} \times (\hat{B}^T q_k)(H_\ell)}{\displaystyle\sum_{\ell=0}^{k} \mathbb{1}\left\{(S_{0,\ell}, A_{0,\ell}) = (S_{0,k}, A_{0,k})\right\}},$$

and $q_{k+1}(s,a) = q_k(s,a)$ for $(s,a) \neq (S_{0,k}, A_{0,k})$, which can be equivalently written as

$$q_{k+1} = (1 - \alpha_k) \otimes q_k + (\hat{B}^T q_k)(H_k)\, \alpha_k,$$

where

$$\alpha_k = \left(\frac{\mathbb{1}\left\{S_{0,k} = s,\ A_{0,k} = a\right\}}{\sum_{\ell=0}^{k} \mathbb{1}\left\{S_{0,\ell} = s,\ A_{0,\ell} = a\right\}} \right)_{(s,a) \in \mathcal{S} \times \mathcal{A}}.$$

- The case $T = 1$ is called *Temporal Difference (TD).*
- The case $1 < T < +\infty$ is called *T-step Temporal Difference.*
- The case $T = +\infty$ is called *on-policy Monte-Carlo policy evaluation*, and cannot be implemented in general, because they require to generate histories of infinite length. They may however be implemented in cases where either the policy at hand and/or the assumptions on the MDP force the rewards to be nonzero only for a finite number of steps.

Proposition 4.3.1. *Let $T \in \{1, 2, \dots\} \cup \{+\infty\}$, π a stationary policy, $(H_k)_{k \geqslant 0}$ a family of random histories of length T^+ denoted*

$$H_k = (S_{0,k}, A_{0,k}, R_{1,k}, \dots, S_{T,k}, A_{T,k}), \quad k \geqslant 0.$$

Consider filtration $\mathcal{F} = (\mathcal{F}_k)_{k \geqslant 0}$ where

$$\mathcal{F}_k = \sigma(H_0, \dots, H_{k-1}, S_{0,k}, A_{0,k}), \quad k \geqslant 0.$$

We assume that:

(i) *for all $k \geqslant 0$, the law of random history H_k conditionally on $\mathcal{F}_k$ is $\mathbb{P}^{(T+)}_{S_{0,k},A_{0,k},\pi}$,*

(ii) *$(q_k)_{k\geqslant 0}$ is a random sequence in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ satisfying for all $k \geqslant 0$, almost-surely,*

$$q_{k+1} = (1-\alpha_k)\otimes q_k + (\hat{B}^T q_k)(H_k)\,\alpha_k,$$

where,

$$\alpha_k = \left(\frac{\mathbb{1}\{S_{0,k}=s,\ A_{0,k}=a\}}{\sum_{\ell=0}^{k}\mathbb{1}\{S_{0,\ell}=s,\ A_{0,\ell}=a\}}\right)_{(s,a)\in\mathcal{S}\times\mathcal{A}},$$

(iii) *almost-surely, for all $(s,a)\in\mathcal{S}\times\mathcal{A}$,*

$$\sum_{k=0}^{+\infty}\mathbb{1}\{S_{0,k}=s,\ A_{0,k}=a\} = +\infty.$$

Then, q_k converges almost-surely to q_π as $k\to+\infty$.

Proof. We can see that above satisfy the assumptions of Theorem 4.1.1 with operator $(B_\pi^{(Q)})^T$ (where in the case $T=+\infty$, $(B_\pi^{(Q)})^\infty$ corresponds to the constant map $q\mapsto q_\pi$), which is indeed a γ-contraction thanks to Proposition 2.2.5 and which gives the almost-sure convergence of q_k to its unique fixed point q_π.

Indeed, the assumptions on the stochastic estimators are given by Propositions 4.2.1 & 4.2.5 and Corollary 4.2.2.

Regarding measurability, for $k\geqslant 0$, it follows from the definition of α_k that it is $\mathcal{F}_k$-measurable. We then prove by an immediate induction that q_k is $\mathcal{F}_k$-measurable.

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

$$\sum_{k=0}^{+\infty}\alpha_k(s,a) = \sum_{k=0}^{+\infty}\frac{\mathbb{1}\{S_{0,k}=s,\ A_{0,k}=a\}}{\sum_{\ell=0}^{k}\mathbb{1}\{S_{0,\ell}=s,\ A_{0,\ell}=a\}} = \sum_{m=1}^{+\infty}\frac{1}{m} = +\infty,$$

and

$$\sum_{k=0}^{+\infty}\alpha_k(s,a)^2 = \sum_{k=0}^{+\infty}\frac{\mathbb{1}\{S_{0,k}=s,\ A_{0,k}=a\}}{(\sum_{\ell=0}^{k}\mathbb{1}\{S_{0,\ell}=s,\ A_{0,\ell}=a\})^2} = \sum_{m=1}^{+\infty}\frac{1}{m^2} < +\infty.$$

□

Lemma 4.3.2. *Let $\pi\in\Pi_0$, $(S_0,A_0,R_1,\dots)$ a random infinite history which, conditionnaly on (S_0,A_0) is distributed according to $\mathbb{P}_{S_0,A_0,\pi}$. Then, for all integers $T\geqslant 1$ and $m\geqslant 0$, conditionally on*

$$(S_0,A_0,R_1,\dots,S_{mT},A_{mT}),$$

the law of $(S_{mT},A_{mT},R_{mT+1},\dots,S_{(m+1)T},A_{(m+1)T})$ is $\mathbb{P}^{(T+)}_{S_{mT},A_{mt},\pi}$.

Proof. □

Algorithm 10: Temporal difference learning for policy evaluation.

Input: Stationary policy π, number of episodes $N \geqslant 1$
for $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
  $q(s, a) \leftarrow 0$
  $N(s, a) \leftarrow 0$
for $n = 1, \dots, N$ **do**
  Observe S_0
  Choose A_0
  $t \leftarrow 0$
  while *episode is not terminated nor truncated* **do**
    Observe (R_{t+1}, S_{t+1})
    Draw $A_{t+1} \sim \pi(S_{t+1})$
    $H \leftarrow (S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}) \quad \in \mathcal{H}^{(1+)}$
    $\hat{B} \leftarrow (\hat{B}^1 q)(H) \quad = R_{t+1} + \gamma\, q(S_{t+1}, A_{t+1})$
    $N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$
    $q(S_t, A_t) \leftarrow \left(1 - \dfrac{1}{N(S_t, A_t)}\right) q(S_t, A_t) + \dfrac{1}{N(S_t, A_t)} \hat{B}$
    $t \leftarrow t + 1$
return q

4.4 Q-learning

We now turn to Q-learning, which is one of the core reinforcement learning algorithms for control. It combines the stochastic fixed point iterations from Tsitsiklis' Theorem 4.1.1 with the stochastic estimator of Bellman optimality operator from Corollary 4.2.2.

It can be informally described as follows. For each $k \geqslant 0$, we choose a state-action pair (S_k, A_k), possibly at random as a function of previous observations. Starting from initial state S_k and action A_k, we generate a history of length 1, meaning a reward and a next state given by the MDPs dynamic (which needs not be known explicitely):

$$H_k = (S_k, A_k, R_k, S'_k) \,|\, S_k, A_k \sim p(\,\cdot\,|S_k, A_k).$$

We asynchronously update the component of the action-value function corresponding to (S_k, A_k) so that the new value is the Cesarò average of the corresponding estimators computed so far:

$$q_{k+1}(S_k, A_k) = \frac{\displaystyle\sum_{\ell=0}^{k} \mathbb{1}\left\{(S_\ell, A_\ell) = (S_k, A_k)\right\} \times (\hat{B}_* q_k)(R_k, S'_k)}{\displaystyle\sum_{\ell=0}^{k} \mathbb{1}\left\{(S_\ell, A_\ell) = (S_k, A_k)\right\}},$$

Algorithm 11: T-step temporal difference learning for policy evaluation.

Input: Stationary policy π, number of episodes $N \geqslant 1$, $T \geqslant 1$
for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
  $q(s,a) \leftarrow 0$
  $N(s,a) \leftarrow 0$
for $n = 1, \ldots, N$ **do**
  Observe S_0
  Choose A_0
  $t \leftarrow 0$
  while *episode is not terminated nor truncated* **do**
    $s \leftarrow 1$
    while $s \leqslant T$ *and episode is not terminated nor truncated* **do**
      Observe (R_{t+s}, S_{t+s})
      Draw $A_{t+s} \sim \pi(S_{t+s})$
      $s \leftarrow s + 1$
    if $s = T + 1$ **then**
      $H \leftarrow (S_t, A_t, R_{t+1}, \ldots, S_{t+T}, A_{t+T}) \quad \in \mathcal{H}^{(T+)}$
      $\hat{B} \leftarrow (\hat{B}^T q)(H) \quad = \sum_{s=1}^{T} \gamma^{s-1} R_{t+s} + \gamma^T \, q(S_{t+T}, A_{t+T})$
      $N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$
      $q(S_t, A_t) \leftarrow \left(1 - \dfrac{1}{N(S_t, A_t)}\right) q(S_t, A_t) + \dfrac{1}{N(S_t, A_t)} \hat{B}$
      $t \leftarrow t + T$
return q

and $q_{k+1}(s,a) = q_k(s,a)$ for $(s,a) \neq (S_k, A_k)$, which can be equivalently written as

$$q_{k+1} = (1 - \alpha_k) \otimes q_k + (\hat{B}_* q_k)(R_k, S'_k)\, \alpha_k,$$

where

$$\alpha_k = \left(\frac{\mathbb{1}\{S_{0,k} = s,\ A_{0,k} = a\}}{\sum_{\ell=0}^{k} \mathbb{1}\{S_{0,\ell} = s,\ A_{0,\ell} = a\}} \right)_{(s,a) \in \mathcal{S} \times \mathcal{A}}.$$

Q-learning is said to be an *off-policy* method, because although a stationary policy π may be used to generate histories (so that $A_k | S_k \sim \pi(S_k)$ and $S_{k+1} = S'_k$ for all $k \geqslant 0$), that policy may very well be quite different from the one that is being learnt.

Proposition 4.4.1. *Let $(H_k)_{k \geqslant 0}$ be a family of random histories of length 1 denoted $H_k = (S_k, A_k, R_k, S'_k)$. Consider filtration $\mathcal{F} = (\mathcal{F}_k)_{k \geqslant 0}$ where*

$$\mathcal{F}_k = \sigma(H_0, \ldots, H_{k-1}, S_k, A_k), \quad k \geqslant 0.$$

We assume that

(i) for all $k \geqslant 0$, the law of (R_k, S'_k) conditionally on $\mathcal{F}_k$ is $p(\cdot|S_k, A_k)$,

(ii) $(q_k)_{k\geqslant 0}$ is a random sequence in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ satisfying, for all $k \geqslant 0$, almost-surely,

$$q_{k+1} = (1-\alpha_k) \otimes q_k + \hat{B}_*(R_k, S'_k)\,\alpha_k, \quad k \geqslant 0,$$

where

$$\alpha_k = \left(\frac{\mathbb{1}\{S_k = s,\ A_k = a\}}{\sum_{\ell=0}^{k} \mathbb{1}\{S_\ell = s,\ A_\ell = a\}} \right)_{(s,a)\in\mathcal{S}\times\mathcal{A}},$$

(iii) almost-surely, for all $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\sum_{k=0}^{+\infty} \mathbb{1}\{S_k = s,\ A_k = a\} = +\infty.$$

Then, q_k converges almost-surely to q_* as $k \to +\infty$.

Proof. Similar to Proposition 4.3.1. □

Remark 4.4.2 (On exploration). The last assumption of the above proposition is extremely important and they are several ways to ensure that all state-action pairs are updated infinitely often. One possibility is to draw each pair (S_k, A_k) e.g. uniformly from $\mathcal{S}\times\mathcal{A}$. This is not always possible, as the state of the environnement may not be freely chosen: states may be accessible only through interaction starting from some initial state. In that case, S_k can only be chosen as given by the last interaction, meaning $S_k = S'_{k-1}$. Another possibility is to follow a policy which at all states ensures that each action is selected with positive probability, and then hope that this would yield an infinite exploration of all state-action pairs.

4.5 Policy iteration

Definition 4.5.1 (Exploring ε-greedy policies)**.** Let $\varepsilon \geqslant 0$. The *exploring ε-greedy policy* with respect to an action-value function $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is the stationary policy denoted $\pi_{g,\varepsilon}[q]$ defined as:

$$\pi_{g,\varepsilon}[q](a|s) = \begin{cases} \dfrac{1-\varepsilon}{|\operatorname{Arg\,max}_{a'\in\mathcal{A}} q(s,a')|} + \dfrac{\varepsilon}{|\mathcal{A}|}, & \text{if } a \in \operatorname*{Arg\,max}_{a'\in\mathcal{A}} q(s,a') \\ \dfrac{\varepsilon}{|\mathcal{A}|} & \text{otherwise,} \end{cases}$$

which can be equivalently written as the following identity in the simplex $\Delta(\mathcal{A})$:

$$\pi_{g,\varepsilon}[q](\cdot|s) = \varepsilon\,\mathcal{U}(\mathcal{A}) + (1-\varepsilon)\,\mathcal{U}\left(\operatorname*{Arg\,max}_{a'\in\mathcal{A}} q(s,a')\right), \quad s \in \mathcal{S},$$

where $\mathcal{U}(\cdot)$ denotes the uniform distribution over a finite set.

Algorithm 12: Q-learning for control.

Input: Number of episodes $N \geqslant 1$
for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
  $q(s,a) \leftarrow 0$
  $N(s,a) \leftarrow 0$
for $n = 1, \dots, N$ **do**
  $t \leftarrow 0$
  while *episode is not terminated nor truncated* **do**
    Observe S_t
    Choose A_t
    Observe (R_{t+1}, S_{t+1})
    $\hat{B} \leftarrow (\hat{B}_* q)(R_{t+1}, S_{t+1}) \quad = R_{t+1} + \gamma \max_{a \in \mathcal{A}} q(S_{t+1}, a)$
    $N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$
    $q(S_t, A_t) \leftarrow \left(1 - \dfrac{1}{N(S_t, A_t)}\right) q(S_t, A_t) + \dfrac{1}{N(S_t, A_t)} \hat{B}$
    $t \leftarrow t + 1$
return $q,\ \pi_g[q]$

Remark 4.5.2. The above (restrictive) definition uniquely defines a policy. In the case $\varepsilon = 0$, the above may not be a greedy policiy, as the latter are deterministic by definition.

We now describe a family of stochastic generalized policy iteration methods, which like their deterministic counterpart, alternate between an approximate policy evaluation step and a policy improvement step.

Let $T \in \{1, 2, \dots\} \cup \{+\infty\}$ and $(\varepsilon_k)_{k \geqslant 0}$ be a positive sequence. For each $k \geqslant 0$, we first compute $\pi_k = \pi_{g,\varepsilon_k}[q_k]$. Then, we choose a state-action pair $(S_{0,k}, A_{0,k})$, possibly at random and as a function of previous observations. Starting from initial state and $S_{0,k}$ and action $A_{0,k}$, we generate a history of length T^+ using policy π_k:

$$H_k = (S_{0,k}, A_{0,k}, R_{1,k}, \dots, S_{T,k}, A_{T,k}) \,|\, S_{0,k}, A_{0,k} \sim \mathbb{P}_{S_{0,k}, A_{0,k}, \pi_k}.$$

The action-value function is then asynchronously updated as:

$$q_{k+1} = (1 - \alpha_k) \otimes q_k + (\hat{B}^T q_k)(H_k)\, \alpha_k,$$

$$\alpha_k = \left(\frac{\mathbb{1}\{S_{0,k} = s,\ A_{0,k} = a\}}{\sum_{\ell=0}^{k} \mathbb{1}\{S_{0,\ell} = s,\ A_{0,\ell} = a\}} \right)_{(s,a) \in \mathcal{S} \times \mathcal{A}}.$$

- In the case $T = 1$, the corresponding algorithm is called *SARSA* (where the letters correspond to the use of variables $S_{0,k}, A_{0,k}, R_{1,k}, S_{1,k}, A_{1,k}$ when computing estimator $(\hat{B}^1 q_k)(H_k)$).
- In the case $1 < T < +\infty$, the algorithm is called *T-step SARSA*.

- In the case $T = +\infty$, the algorithm is called *on-policy Monte-Carlo control*.

The above algorithms are said to be *on-policy* because the policy used to generate the histories are close to the policy that is being learnt.

Algorithm 13: SARSA for control with ε-greedy exploration.

Input: Number of episodes $N \geqslant 1$
for $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
  $q(s, a) \leftarrow 0$
  $N(s, a) \leftarrow 0$
for $n = 1, \ldots, N$ **do**
  Observe S_0
  Choose A_0
  $t \leftarrow 0$
  while *episode is not terminated nor truncated* **do**
    Observe (R_{t+1}, S_{t+1})
    Choose ε
    Draw $A_{t+1} \sim \pi_{g,\varepsilon}[q](S_{t+1})$
    $H \leftarrow (S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}) \quad \in \mathcal{H}^{(1+)}$
    $\hat{B} \leftarrow (\hat{B}^1 q)(H) \quad = R_{t+1} + \gamma\, q(S_{t+1}, A_{t+1})$
    $N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$
    $q(S_t, A_t) \leftarrow \left(1 - \dfrac{1}{N(S_t, A_t)}\right) q(S_t, A_t) + \dfrac{1}{N(S_t, A_t)} \hat{B}$
    $t \leftarrow t + 1$
return $q,\ \pi_g[q]$

Algorithm 14: T-step SARSA for control with ε-greedy exploration.

Input: $1 \leqslant T \leqslant \infty$, number of episodes $N \geqslant 1$
for $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
  $q(s, a) \leftarrow 0$
  $N(s, a) \leftarrow 0$
for $n = 1, \ldots, N$ **do**
  Observe S_0
  Choose A_0
  $t \leftarrow 0$
  while *episode is not terminated nor truncated* **do**
    Choose ε
    $s \leftarrow 1$
    while $s \leqslant T$ *and episode is not terminated nor truncated* **do**
      Observe (R_{t+s}, S_{t+s})
      Draw $A_{t+s} \sim \pi_{g,\varepsilon}[q](S_{t+s})$
      $s \leftarrow s + 1$
    if $s = T + 1$ **then**
      $H \leftarrow (S_t, A_t, R_{t+1}, \ldots, S_{t+T}, A_{t+T}) \quad \in \mathcal{H}^{(T+)}$
      $\hat{B} \leftarrow (\hat{B}^T q)(H) \quad = \sum_{s=1}^{T} \gamma^{s-1} R_{t+s} + \gamma^T q(S_{t+T}, A_{t+T})$
      $N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$
      $q(S_t, A_t) \leftarrow \left(1 - \frac{1}{N(S_t, A_t)}\right) q(S_t, A_t) + \frac{1}{N(S_t, A_t)} \hat{B}$
      $t \leftarrow t + T$
return $q,\ \pi_g[q]$

Chapter 5

Value function approximation

Previous chapters presented *tabular* methods, meaning that they handled value functions in $\mathbb{R}^{\mathcal{S}}$ or $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. In the case where the set of states $\mathcal{S}$ is huge, which happens e.g. in cases where the problem has a combinatorial character, this may be impractical. Then, one possible approach is to solve an approximate problem by restricting to a parametric class of value functions $\{v_w\}_{w\in\mathbb{R}^p}$ where the dimensionality p of the parameter space is much lower than the number of states.

The Bellman equations that defined the solutions to policy evaluation and control in tabular settings are here turned into approximate fixed point problems, where we seek for e.g. value functions v_w within the parametric class, such that $v_w \approx B_\pi v_w$ or $v_w \approx B_* v_w$.

An important kind of parametrization is linear parametrization where for a given feature function $\phi : \mathcal{S} \to \mathbb{R}^p$, we restrict to value functions of the form $v_w(s) = \phi(s)^\top w$, and for which we are able to establish a few theoretical guarantees. Another important type is neural networks which have led to many practical successes.

5.1 Projected Bellman equations

Let $p \geqslant 1$ and $\mathcal{V} = \{v_w\}_{w\in\mathbb{R}^p}$ be a parametric class of state-value functions such that $w \mapsto v_w$ is differentiable. Denote $\operatorname{Jac} v_w$ the Jacobian matrix with respect to parameter $w \in \mathbb{R}^p$:

$$\operatorname{Jac} v_w = \left(\frac{\partial v_w(s)}{\partial w_j}\right)_{\substack{s\in\mathcal{S}\\ 1\leqslant j\leqslant p}}.$$

Let $\mu \in \Delta(\mathcal{S})$ such that $\mu(s) > 0$ for all $s \in \mathcal{S}$, denote $M_\mu = \operatorname{diag}_{s\in\mathcal{S}}(\mu(s))$

and consider norm $\ell_{2,\mu}$ defined on $\mathbb{R}^{\mathcal{S}}$ as

$$\|v\|_{2,\mu} = \sqrt{v^\top M_\mu v} = \sqrt{\sum_{s\in\mathcal{S}} \mu(s) v(s)^2}.$$

For $v_0 \in \mathbb{R}^{\mathcal{S}}$, denote $\operatorname{Proj}_{\mathcal{V},\mu} v_0$ the set of minimizers on $\mathcal{V}$ of the $\ell_{2,\mu}$ distance to v_0:

$$\operatorname{Proj}_{\mathcal{V},\mu} v_0 = \operatorname*{Arg\,min}_{v\in\mathcal{V}} \|v - v_0\|_{2,\mu}.$$

Proposition 5.1.1. *Let $v_0 \in \mathbb{R}^{\mathcal{S}}$. If $w \in \mathbb{R}^p$ is such that*

$$v_w \in \operatorname{Proj}_{\mathcal{V},\mu} v_0,$$

then

$$(\operatorname{Jac} v_w)^\top M_\mu (v_0 - v_w) = 0.$$

Proof. By definition of the projection, w is a minimizer of $w \mapsto \|v_w - v_0\|_{2,\mu}^2$ which is differentiable by composition. Simple computation gives that the gradient of this function is

$$(\operatorname{Jac} v_w)^\top M_\mu (v_0 - v_w).$$

Because the domain $\mathbb{R}^p$ is an open set, the gradient must vanish at a minimizer, hence the result. □

Corollary 5.1.2. *Let $T \geqslant 1$, $w \in \mathbb{R}^p$ and π a stationary policy. If*

$$v_w \in \operatorname{Proj}_{\mathcal{V},\mu} B_\pi^T v_w \qquad \left(\textit{resp.} \quad v_w \in \operatorname{Proj}_{\mathcal{V},\mu} B_* v_w\right), \tag{5.1}$$

then

$$(\operatorname{Jac} v_w)^\top M_\mu (B_\pi^T v_w - v_w) = 0, \qquad \left(\textit{resp.} \quad (\operatorname{Jac} v_w)^\top M_\mu (B_* v_w - v_w) = 0\right).$$

The above fixed point problems (5.1) are difficult to tackle directly for several reasons. First, existence of a solution is not guaranteed. Even if a solution exists, operators $\operatorname{Proj}_{\mathcal{V},\mu} B_\pi^T$ and $\operatorname{Proj}_{\mathcal{V},\mu} B_*$ may not be contractions, and even if they were, they are difficult to compute, because the projection is itself an optimization problem to be solved.

For policy iteration, the above corollary however suggests an alternative approach: one can look for zeros of operator $w \mapsto (\operatorname{Jac} v_w)^\top M_\mu (B_\pi^T v_w - v_w)$, or equivalently, fixed points of operator

$$w \mapsto w + \alpha (\operatorname{Jac} v_w)^\top M_\mu (B_\pi^T v_w - v_w),$$

for some $\alpha \neq 0$. Although it is unknown *a priori* for which value of α this operator is a contraction (if any), it is at least given in explicit form and is therefore easily computable.

Therefore, a possible synchronous dynamic programming algorithm for policy evaluation is given by:

$$w_{k+1} = w_k + \alpha(\operatorname{Jac} v_{w_k})^\top M_\mu (B_\pi^T v_{w_k} - v_{w_k}), \quad k \geqslant 0, \tag{5.2}$$

and similarly a synchronous dynamic programming algorithm for control is given by

$$w_{k+1} = w_k + \alpha(\operatorname{Jac} v_{w_k})^\top M_\mu (B_* v_{w_k} - v_{w_k}), \quad k \geqslant 0. \tag{5.3}$$

Similar iterations can be defined for action-value functions.

Although the above iterations (5.2) and (5.3) still require the computation of vectors in $\mathbb{R}^{\mathcal{S}}$, this will no longer be the case with sample-based reinforcement learning variants in Section 5.3.

Remark 5.1.3 (Analogy with gradient-based optimization). This approach is somewhat analogous to gradient-based optimization, where a gradient descent iterates operator $x \mapsto x - \alpha \nabla f(x)$ in the hope of finding a zero of the gradient, even though being a zero of the gradient is only a necessary condition for optimality. In the convex case however, this approach is better justified by the fact that zeros of the gradient exactly correspond to global minimizers.

Example 5.1.4 (Tabular case)**.** The tabular case, meaning when value functions are not approximated, is a special case corresponding to $\{v_w\}_{w \in \mathbb{R}^{\mathcal{S}}}$ and $v_w = w$ for all $w \in \mathbb{R}^{\mathcal{S}}$. Then, the Jacobian matrix is the identity and choosing μ as the uniform distribution over $\mathcal{S}$ and $\alpha = |\mathcal{S}|$ makes algorithms (5.2) and (5.3) boil down to synchronous value iterations from Section 3.1.

5.2 Linear parametrization

Let $p \geqslant 1$ and $\phi : \mathcal{S} \to \mathbb{R}^p$. In this section, we consider $\mathcal{V} = \{v_w\}_{w \in \mathbb{R}^{\mathcal{S}}}$ the class of linearly parameterized state-value functions associated with feature function ϕ, where $v_w : s \mapsto \phi(s)^\top w$ ($w \in \mathbb{R}^{\mathcal{S}}$), which is a subspace of $\mathbb{R}^{\mathcal{S}}$. For $w \in \mathbb{R}^p$, we alternatively denote $\Phi w = v_w \in \mathbb{R}^{\mathcal{S}}$, which can be interpreted as a matrix-vector product with $\Phi = (\phi(s)_j)_{\substack{s \in \mathcal{S} \\ 1 \leqslant j \leqslant p}} \in \mathbb{R}^{\mathcal{S} \times p}$.

Proposition 5.2.1. *The projection onto $\mathcal{V}$ with respect to $\ell_{2,\mu}$ exists and is unique, in other words,* $\operatorname{Proj}_{\mathcal{V},\mu} v$ *is a singleton for all $v \in \mathbb{R}^{\mathcal{S}}$.*

Proof. Because the norm derives from an inner product, and because $\mathcal{V}$ is closed and convex, the projection exists and is unique. □

Proposition 5.2.2. $\operatorname{Proj}_{\mathcal{V},\mu}$ *is* 1*-Lipschitz continuous for $\ell_{2,\mu}$.*

Then for a given stationary policy π, the fixed point problem (5.2) can be equivalently written, in the case $T = 1$, as

$$\Phi w = \mathrm{Proj}_{\mathcal{V},\mu} B_\pi \Phi w. \tag{5.4}$$

The following proposition proves that the necessary condition from Proposition 5.1.1 is in fact a characterization in the linear case, thanks to a convexity argument.

Proposition 5.2.3. *Let π be a stationary policy and $w \in \mathbb{R}^p$. Then, $\Phi w = \mathrm{Proj}_{\mathcal{V},\mu} B_\pi \Phi w$ if, and only if*

$$\Phi^\top M_\mu \left(B_\pi \Phi w - \Phi w\right) = 0.$$

Proof. Consider function $f(w') = \|B_\pi \Phi w - \Phi w'\|_{2,\mu}^2$, which is convex and differentiable on $\mathbb{R}^p$. By definition of the projection, $\Phi w = \mathrm{Proj}_{\mathcal{V},\mu} B_\pi \Phi w$ is equivalent to having

$$w \in \arg\min_{w' \in \mathbb{R}^p} f(w'),$$

which by convexity is equivalent to having $\nabla f(w') = 0$. The result then follows by deriving an explicit expression for ∇f. □

Following a given stationary policy defines a Markov chain on the set of states $\mathcal{S}$, which is finite. Therefore, there exists a stationary measure. We are able to establish a convergence guarantee for (5.2) in the case where μ is such a measure.

Definition 5.2.4. Let π be a stationary policy. A *stationary distribution associated with* π is a probability distribution $\mu_\pi \in \Delta(\mathcal{S})$ that satisfies

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

We then denote $M_\pi = \mathrm{diag}_{s\in\mathcal{S}}(\mu_\pi(s))$ and $\ell_{2,\pi} = \ell_{2,\mu_\pi}$

Proposition 5.2.5. *Let π be a stationary policy and μ_π an associated stationary distribution such that $\mu_\pi(s) > 0$ for all $s \in \mathcal{S}$. Operator $B_\pi^{(V,\gamma)}$ is then γ-Lipschitz continuous with respect to $\ell_{2,\pi}$.*

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

$$\begin{aligned}
\left\|B_\pi v' - B_\pi v\right\|_{2,\pi}^2 &= \sum_{s\in\mathcal{S}} \mu_\pi(s) \left((B_\pi v')(s) - (B_\pi v)(s)\right)^2 \\
&= \sum_{s\in\mathcal{S}} \mu_\pi(s) \left(\sum_{(a,r',s')\in\mathcal{A}\times\mathcal{R}\times\mathcal{S}} \pi(a|s)p(r,s'|s,a)\gamma(v'(s') - v(s')) \right)^2 \\
&\leqslant \sum_{s\in\mathcal{S}} \mu_\pi(s) \sum_{(a,r',s')\in\mathcal{A}\times\mathcal{R}\times\mathcal{S}} \pi(a|s)p(r,s'|s,a)\gamma^2(v'(s') - v(s'))^2 \\
&= \gamma^2 \sum_{s'\in\mathcal{S}} (v'(s') - v(s'))^2 \sum_{(s,a,r)\in\mathcal{S}\times\mathcal{A}\times\mathcal{R}} \pi(a|s)p(r,s'|s,a)\mu_\pi(s) \\
&= \gamma^2 \sum_{s'\in\mathcal{S}} (v'(s') - v(s'))^2 \mu_\pi(s') \\
&= \gamma^2 \left\|v' - v\right\|_{2,\pi}^2,
\end{aligned}$$

where we used the definition of a stationary distribution to get the fifth line, and for the third line Jensen's inequality together with the fact that for a given $s \in \mathcal{S}$,

$$\begin{aligned}
\sum_{(a,r',s')\in\mathcal{A}\times\mathcal{R}\times\mathcal{S}} \pi(a|s)p(r,s'|s,a) &= \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{(r,s')\in\mathcal{R}\times\mathcal{S}} p(r,s'|s,a) \\
&= \sum_{a\in\mathcal{A}} \pi(a|s) = 1.
\end{aligned}$$

□

Proposition 5.2.6. *Let π be a stationary policy and μ_π an associated stationary distribution such that $\mu_\pi(s) > 0$ for all $s \in \mathcal{S}$. Then, there exists a unique state-value function $v \in \mathcal{V}$ such that $v = \mathrm{Proj}_{\mathcal{V},\mu_\pi} B_\pi v$.*

Proof. Combining Propositions 5.2.2 and 5.2.5 by composition yields that $\mathrm{Proj}_{\phi,\mu} \circ B_\pi^{(V)}$ is a γ-Lipschitz continuous operator on $\mathcal{V}$ for $\ell_{2,\pi}$. By Theorem 2.2.3, the fixed point problem admits a unique solution. □

With the assumptions Proposition 5.2.6, of although operator $\mathrm{Proj}_{\mathcal{V},\mu_\pi} \circ B_\pi^{(V)}$ which defines the fixed point problem (5.4) is a contraction, it is inefficient to iterate it, as computing $\mathrm{Proj}_{\mathcal{V},\mu}$ is itself an optimization problem.

Proposition 5.2.7. *Let π be a stationary policy and μ_π an associated stationary distribution such that $\mu_\pi(s) > 0$ for all $s \in \mathcal{S}$. Then, if $\operatorname{Ker}\Phi = 0$, there exists $\alpha_0 > 0$ such that for all $0 < \alpha \leqslant \alpha_0$, operator*

$$I + \alpha\Phi^\top M_\pi (B_\pi - I)\Phi$$

is β-Lipschitz continuous for some $0 \leqslant \beta < 1$.

Remark 5.2.8. Under the assumptions of the above proposition, iterations (5.2) therefore converge to the unique solution of (5.4) for small enough step-size $\alpha > 0$.

5.3 Semi-gradient algorithms

We now define asynchronous reinforcement learning algorithms with parametric value-functions that interact with the MDP without prior knowledge of its dynamic p. The construction is based on action-value counterparts of the methods defined in Section 5.1 combined with stochastic estimators of Bellman operators and a new kind of asynchronous updates.

Analogously to Section 5.1, we consider a parametric class of action-value functions. Let $p \geqslant 1$ and $\mathcal{Q} = \{q_w\}_{w \in \mathbb{R}^p}$ a subset of $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ such that $w \mapsto q_w$ is differentiable. For $w \in \mathbb{R}^p$, denote

$$\operatorname{Jac} q_w = \left(\frac{\partial q_w(s,a)}{\partial w_j}\right)_{\substack{(s,a)\in\mathcal{S}\times\mathcal{A} \\ 1\leqslant j\leqslant p}}$$

and

$$\nabla q_w(s,a) = \left(\frac{\partial q_w(s,a)}{\partial w_j}\right)_{1\leqslant j\leqslant p}, \quad (s,a) \in \mathcal{S}\times\mathcal{A}.$$

Let $T \geqslant 1$, $\mu \in \Delta(\mathcal{S}\times\mathcal{A})$ and $M_\mu = \operatorname{diag}_{(s,a)\in\mathcal{S}\times\mathcal{A}}(\mu(s,a))$. The following iterations are the action-value counterparts of (5.2) and (5.3):

$$w_{k+1} = w_k + \alpha(\operatorname{Jac} q_{w_k})^\top M_\mu (B_{\pi_k}^T q_{w_k} - q_{w_k}), \quad k \geqslant 1, \tag{5.5}$$

where $(\pi_k)_{k\geqslant 0}$ is sequence of stationary strategies and

$$w_{k+1} = w_k + \alpha(\operatorname{Jac} q_{w_k})^\top M_\mu (B_* q_{w_k} - q_{w_k}), \quad k \geqslant 1. \tag{5.6}$$

For simplicity, we only consider below the case where μ is the uniform distribution over $\mathcal{S}\times\mathcal{A}$, which corresponds to $M_\mu = (|\mathcal{S}|\cdot|\mathcal{A}|)^{-1} I$. We then remove the factor $(|\mathcal{S}|\cdot|\mathcal{A}|)^{-1}$ by incorporating it into step-size α.

In Section 4, the definition of tabular methods could be theoretically seen as first constructing a stochastic estimator of the operator being iterated (evaluated at the current iterate), and then considering an asynchronous update based on this estimator. Although this approach is possible for the operators being iterated above in (5.6) and (5.6), this would not avoid manipulating vectors in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, which was the initial motivation for value function approximation. Instead, we first consider in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ a stochastic estimator of $B_{\pi_k}^T q_{w_k} - q_{w_k}$ (resp. $B_* q_{w_k} - q_{w_k}$) *restricted to a single component*, which reduces the computational burden from $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ to a scalar, and then apply matrix multiplication by $(\operatorname{Jac} q_{w_k})^\top$, which reduces to computing a single vector in $\mathbb{R}^p$ because the estimator has a single nonzero component. This construction is made precise in the following statement.

Lemma 5.3.1. *Let $w \in \mathbb{R}^p$.*

(i) Let $T \in \{1, 2, \dots\} \cup \{+\infty\}$ and π a stationary policy. Let $H = (S_0, A_0, R_1, \dots, S_T, A_T)$ be a random history of length T^+. If the law of H conditionally on (S_0, A_0) is $\mathbb{P}^{(T+)}_{S_0,A_0,\pi}$, then

$$\mathbb{E}\left[\Big((\hat{B}^T q_w)(H) - q_w(S_0, A_0)\Big)\nabla q_w(S_0, A_0)\Big|S_0, A_0\right] = (\operatorname{Jac} q_w)^\top \big((B^T_\pi q_w)(S_0, A_0) - q_w(S_0, A_0)\big)\,\mathbb{1}_{(S_0,A_0)},$$

where $\mathbb{1}_{(S_0,A_0)} = (\mathbb{1}\{(s, a) = (S_0, A_0)\})_{(s,a)\in\mathcal{S}\times\mathcal{A}} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$.

(ii) If (S, A, R, S') is a random history of length 1^+ such that $(R, S')|S, A \sim p(\cdot\,|S, A)$, then

$$\mathbb{E}\left[\Big((\hat{B}_* q_w)(R, S') - q_w(S, A)\Big)\nabla q_w(S, A)\Big|S, A\right] = (\operatorname{Jac} q_w)^\top ((B_* q_w)(S, A) - q_w(S, A))\,\mathbb{1}_{(S,A)}.$$

An important feature of the above estimators

$$\Big((\hat{B}^T q_w)(H) - q_w(S_0, A_0)\Big)\nabla q_w(S_0, A_0)$$

and

$$\Big((\hat{B}_* q_w)(R, S') - q_w(S, A)\Big)\nabla q_w(S, A)$$

is that they only involve a vector in $\mathbb{R}^p$ and no vector in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$.

Definition 5.3.2 (Semi-gradient policy evaluation)**.** Let $T \in \{1, 2, \dots\} \cup \{+\infty\}$, π a stationary policy, $(H_k)_{k\geqslant 0}$ a family of random histories of length T^+:

$$H_k = (S_{0,k}, A_{0,k}, R_{1,k}, \dots, S_{T,k}, A_{T,k}), \quad k \geqslant 0.$$

Consider filtration $\mathcal{F} = (\mathcal{F}_k)_{k\geqslant 0}$ where

$$\mathcal{F}_k = \sigma(H_1, \dots, H_{k-1}, S_{0,k}, A_{0,k}), \quad k \geqslant 0.$$

A random sequence $(w_k)_{k\geqslant 0}$ in $\mathbb{R}^p$ is a *T-step semi-gradient value iteration for policy evaluation* if:

(i) for all $k \geqslant 0$, the law of random history H_k conditionally on $\mathcal{F}_k$ is $\mathbb{P}^{(T+)}_{S_{0,k},A_{0,k},\pi}$,

(ii) for all $k \geqslant 0$, almost-surely,

$$w_{k+1} = w_k + \alpha_k((\hat{B}^T q_{w_k})(H_k) - q_{w_k}(S_{0,k}, A_{0,k}))\nabla q_{w_k}(S_{0,k}, A_{0,k}),$$

where α_k is a $\mathcal{F}_k$-measurable positive random variable.

Algorithm 15: Semi-gradient policy evaluation.

Input: Stationary policy π, number of episodes $N \geqslant 1$, $T \geqslant 1$
$w \leftarrow 0_{\mathbb{R}^p}$
for $n = 1, \dots, N$ **do**
  Observe S_0
  Choose A_0
  $t \leftarrow 0$
  while *episode is not terminated nor truncated* **do**
    Observe (R_{t+1}, S_{t+1})
    Draw $A_{t+1} \sim \pi(S_{t+1})$
    $H \leftarrow (S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}) \quad \in \mathcal{H}^{(1+)}$
    $\hat{B} \leftarrow (\hat{B}^1 q_w)(H) \quad = R_{t+1} + \gamma\, q_w(S_{t+1}, A_{t+1})$
    Choose α
    $w \leftarrow w + \alpha(\hat{B} - q_w(S_t, A_t))\nabla q_w(S_t, A_t)$
    $t \leftarrow t + 1$
return w

Algorithm 16: T-step semi-gradient policy evaluation.

Input: Stationary policy π, $T \in \{1, 2, \dots\} \cup \{+\infty\}$, number of episodes $N \geqslant 1$
$w \leftarrow 0_{\mathbb{R}^p}$
for $n = 1, \dots, N$ **do**
  Observe S_0
  Choose A_0
  $t \leftarrow 0$
  while *episode is not terminated nor truncated* **do**
    $s \leftarrow 1$
    while $s \leqslant T$ *and episode is not terminated nor truncated* **do**
      Observe (R_{t+s}, S_{t+s})
      Draw $A_{t+s} \sim \pi(S_{t+s})$
      $s \leftarrow s + 1$
    if $s = T + 1$ **then**
      $H \leftarrow (S_t, A_t, R_{t+1}, \dots, S_{t+T}, A_{t+T}) \quad \in \mathcal{H}^{(T+)}$
      $\hat{B} \leftarrow (\hat{B}^T q_w)(H) \quad = \sum_{s=1}^{T} \gamma^{s-1} R_{t+s} + \gamma^T\, q_w(S_{t+T}, A_{t+T})$
      Choose α
      $w \leftarrow w + \alpha(\hat{B} - q_w(S_t, A_t))\nabla q_w(S_t, A_t)$
      $t \leftarrow t + T$
return w

Definition 5.3.3 (Semi-gradient Q-learning)**.** Let $(H_k)_{k \geqslant 0}$ be a family of

random histories of length 1^+:

$$H_k = (S_k, A_k, R_k, S'_k), \quad k \geqslant 0.$$

Consider filtration $\mathcal{F} = (\mathcal{F}_k)_{k\geqslant 0}$ where

$$\mathcal{F}_k = \sigma(H_1, \dots, H_{k-1}, S_k, A_k), \quad k \geqslant 0.$$

A random sequence $(w_k)_{k\geqslant 0}$ in $\mathbb{R}^p$ is a *semi-gradient Q-learning iteration* if:

(i) for all $k \geqslant 0$, the law of (R_k, S'_k) conditionally on $\mathcal{F}_k$ is $p(\,\cdot\,|S_k, A_k)$,

(ii) for all $k \geqslant 0$, almost-surely,

$$w_{k+1} = w_k + \alpha_k((\hat{B}_* q_{w_k})(R_k, S'_k) - q_{w_k}(S_{0,k}, A_{0,k}))\nabla q_{w_k}(S_{0,k}, A_{0,k}),$$

where α_k is a $\mathcal{F}_k$-measurable positive random variable.

Remark 5.3.4 (On exploration). There is no assumption on the choice of (S_k, A_k) in the above definition. However, similarly to tabular methods, *exploration* is an important issue, and must be addressed to obtain satisfactory results. Even if $\mathcal{S}$ and/or $\mathcal{A}$ are huge and all values $(s, a) \in \mathcal{S} \times \mathcal{A}$ can't then be explored, the method should ensure that sufficient *variety* of values are encountered. If the algorithm has the possibility to freely choose states, new state-action pairs can be e.g. drawn according to a given distribution. Otherwise, the method should merely follow the actual interaction with the environment, by choosing actions according to e.g. an ε-greedy policy with respect to the current estimate q_{w_k}.

Algorithm 17: Semi-gradient Q-learning for control.

Input: Number of episodes $N \geqslant 1$
$w \leftarrow 0_{\mathbb{R}^p}$
for $n = 1, \dots, N$ **do**
  $t \leftarrow 0$
  while *episode is not terminated nor truncated* **do**
    Observe S_t
    Choose A_t
    Observe (R_{t+1}, S_{t+1})
    $\hat{B} \leftarrow (\hat{B}_* q_w)(R_{t+1}, S_{t+1})$ $= R_{t+1} + \gamma \max_{a\in\mathcal{A}} q_w(S_{t+1}, a)$
    Choose α
    $w \leftarrow w + \alpha(\hat{B} - q_w(S_t, A_t))\nabla q_w(S_t, A_t)$
    $t \leftarrow t + 1$
return $w,\ \pi_g\,[q_w]$

Definition 5.3.5 (Semi-gradient (T-step) SARSA)**.** Let $T \in \{1, 2, \dots\} \cup \{+\infty\}$, $(\alpha_k)_{k\geqslant 0}$ and $(\varepsilon_k)_{k\geqslant 0}$ a positive random sequences, and $(H_k)_{k\geqslant 0}$ a family of random histories of length T^+:

$$H_k = (S_{0,k}, A_{0,k}, R_{1,k}, \dots, S_{T,k}, A_{T,k}), \quad k \geqslant 0.$$

Consider filtration $\mathcal{F} = (\mathcal{F}_k)_{k\geqslant 0}$ where

$$\mathcal{F}_k = \sigma(H_1, \dots, H_{k-1}, S_{0,k}, A_{0,k}), \quad k \geqslant 0.$$

A random sequence $(w_k)_{k\geqslant 0}$ in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is a *T-step semi-gradient SARSA iteration* if:

(i) for all $k \geqslant 0$, α_k and ε_k are $\mathcal{F}_k$-measurable,

(ii) for all $k \geqslant 0$, the law of random history H_k conditionally on $\mathcal{F}_k$ is $\mathbb{P}^{(T+)}_{S_{0,k},A_{0,k},\pi_k}$, where $\pi_k = \pi_{g,\varepsilon_k}[q_{w_k}]$,

(iii) for all $k \geqslant 0$, almost-surely,

$$w_{k+1} = w_k + \alpha_k((\hat{B}^T q_{w_k})(H_k) - q_{w_k}(S_{0,k}, A_{0,k}))\nabla q_{w_k}(S_{0,k}, A_{0,k}),$$

In the case $T = +\infty$, the method is instead called *Semi-gradient Monte-Carlo control*.

Remark 5.3.6 (Linear case). In the case of a linear parametric class presented in Section 5.2, $\nabla q_{w_k}(S_{0,k}, A_{0,k})$ in the above definition boils down to $\phi(S_{0,k}, A_{0,k})$.

Algorithm 18: Semi-gradient SARSA for control with ε-greedy exploration.

Input: Number of episodes $N \geqslant 1$
$w \leftarrow 0_{\mathbb{R}^p}$
for $n = 1, \ldots, N$ **do**
  Observe S_0
  Choose A_0
  $t \leftarrow 0$
  while *episode is not terminated nor truncated* **do**
    Observe (R_{t+1}, S_{t+1})
    Choose ε
    Draw $A_{t+1} \sim \pi_{g,\varepsilon}[q_w](S_{t+1})$
    $H \leftarrow (S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}) \quad \in \mathcal{H}^{(1+)}$
    $\hat{B} \leftarrow (\hat{B}^1 q_w)(H) \quad = R_{t+1} + \gamma\, q_w(S_{t+1}, A_{t+1})$
    Choose α
    $w \leftarrow w + \alpha(\hat{B} - q_w(S_t, A_t))\nabla q_w(S_t, A_t)$
    $t \leftarrow t + 1$
return $w,\ \pi_g[q_w]$

Algorithm 19: T-step semi-gradient SARSA for control with ε-greedy exploration.

Input: $1 \leqslant T \leqslant +\infty$, number of episodes $N \geqslant 1$
$w \leftarrow 0_{\mathbb{R}^p}$
for $n = 1, \ldots, N$ **do**
  Observe S_0
  Choose A_0
  $t \leftarrow 0$
  while *episode is not terminated nor truncated* **do**
    Choose ε
    $s \leftarrow 1$
    while $s \leqslant T$ *and episode is not terminated nor truncated* **do**
      Observe (R_{t+s}, S_{t+s})
      Draw $A_{t+s} \sim \pi_{g,\varepsilon}[q_w](S_{t+s})$
      $s \leftarrow s + 1$
    if $s = T + 1$ **then**
      $H \leftarrow (S_t, A_t, R_{t+1}, \ldots, S_{t+T}, A_{t+T}) \quad \in \mathcal{H}^{(T+)}$
      $\hat{B} \leftarrow (\hat{B}^T q_w)(H) \quad = \sum_{s=1}^{T} \gamma^{s-1} R_{t+s} + \gamma^T\, q_w(S_{t+T}, A_{t+T})$
      Choose α
      $w \leftarrow w + \alpha(\hat{B} - q_w(S_t, A_t))\nabla q_w(S_t, A_t)$
      $t \leftarrow t + T$
return $w,\ \pi_g[q_w]$

Chapter 6

Policy gradient methods

This chapter presents a different reinforcement learning approach where a hopefully good policy is obtained directly, without being derived from a value function. More specifically, we focus on methods that consider a parametric class of policies that are differentiable with respect to the parameter, in order to apply gradient-based optimization methods.

6.1 Policy parametrization

Let $d \geqslant 1$ and $\{\pi_\theta\}_{\theta \in \mathbb{R}^d}$ a parametric class of stationary policies such that $\theta \mapsto \pi_\theta$ is differentiable and for all $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\pi_\theta(a|s) > 0.$$

We denote ∇_θ the gradient with respect to the policy parameter $\theta \in \mathbb{R}^d$ (to avoid confusion with the parametrization $w \in \mathbb{R}^p$ of value functions, when both are considered as in Section 6.3 below).

Example 6.1.1 (Softmax parametrization)**.** A popular kind of parametric policies that ensure that all probabilities $\pi_\theta(a|s)$ are positive is given by

$$\pi_\theta(a|s) = \frac{\exp\left(h_\theta(s,a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(h_\theta(s,a')\right)}, \quad s \in \mathcal{S},\ a \in \mathcal{A},$$

where $h_\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is such that $\theta \mapsto h_\theta(s,a)$ is differentiable for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and can e.g. be a linear function, a neural network, etc.

Let $\mu \in \Delta(\mathcal{S})$ be given. The approach aims at maximizing the following objective function

$$\theta \mapsto \mathbb{E}_{S \sim \mu}\left[v_{\pi_\theta}(S)\right] \tag{6.1}$$

with stochastic gradient descent (SGD) (or rather *ascent*) or one of its variants e.g. RMSprop, Adam, etc. The following *policy gradient theorem* provides the expression of a stochastic estimator of the gradient of the above objective function, to be used in such an optimization algorithm.

Theorem 6.1.2. *We assume that map $\theta \mapsto \pi_\theta$ is of class $\mathcal{C}^1$. Then, the objective function (6.1) is differentiable. Moreover, if $(b_t)_{t\geqslant 0}$ a bounded sequence of functions in $\mathbb{R}^{\mathcal{S}}$, then for all $\theta \in \mathbb{R}^d$,*

$$\nabla_\theta \mathbb{E}_{S\sim\mu}\left[v_{\pi_\theta}(S)\right] = \mathbb{E}_{\mu,\pi_\theta}\left[\sum_{t=0}^{+\infty}\gamma^t\left(q_{\pi_\theta}(S_t,A_t) - b_t(S_t)\right)\nabla_\theta \log \pi_\theta(A_t|S_t)\right].$$

Proof.

$$\begin{aligned}
\nabla_\theta \mathbb{E}_{S\sim\mu}\left[v_{\pi_\theta}(S)\right] &= \nabla_\theta \mathbb{E}_{\mu,\pi_\theta}\left[v_{\pi_\theta}(S_0)\right] = \nabla_\theta \mathbb{E}_{\mu,\pi_\theta}\left[(B_{\pi_\theta} v_{\pi_\theta})(S_0)\right] \\
&= \nabla_\theta \mathbb{E}_{\mu,\pi_\theta}\left[\sum_{(a,r,s')\in\mathcal{A}\times\mathcal{R}\times\mathcal{S}} \pi_\theta(a|S_0)p(r,s'|S_0,a)(r+\gamma\, v_{\pi_\theta}(s'))\right] \\
&= \mathbb{E}_{\mu,\pi_\theta}\left[\sum_{(a,r,s')\in\mathcal{A}\times\mathcal{R}\times\mathcal{S}} \nabla_\theta\pi_\theta(a|S_0)p(r,s'|S_0,a)(r+\gamma\, v_{\pi_\theta}(s'))\right] \\
&\quad + \mathbb{E}_{\mu,\pi_\theta}\left[\sum_{(a,r,s')\in\mathcal{A}\times\mathcal{R}\times\mathcal{S}} \pi_\theta(a|S_0)p(r,s'|S_0,a)\nabla_\theta v_{\pi_\theta}(s')\right] \\
&= \mathbb{E}_{\mu,\pi_\theta}\left[\sum_{a\in\mathcal{A}} \nabla_\theta\pi_\theta(a|S_0)\,(Dv_{\pi_\theta})(S_0,a)\right] + \gamma\,\mathbb{E}_{\mu,\pi_\theta}\left[\nabla_\theta v_{\pi_\theta}(S_1)\right] \\
&= \mathbb{E}_{\mu,\pi_\theta}\left[\sum_{a\in\mathcal{A}} \nabla_\theta\pi_\theta(a|S_0)q(S_0,a)\right] + \gamma\,\mathbb{E}_{\mu,\pi_\theta}\left[\mathbb{E}_{S_1,\pi_\theta}\left[\nabla_\theta v_{\pi_\theta}(S_0)\right]\right] \\
&= \mathbb{E}_{\mu,\pi_\theta}\left[\sum_{a\in\mathcal{A}} \nabla_\theta\pi_\theta(a|S_0)q(S_0,a)\right] \\
&\quad + \gamma\,\mathbb{E}_{\mu,\pi_\theta}\left[\mathbb{E}_{S_1,\pi_\theta}\left[\sum_{a\in\mathcal{A}} \nabla_\theta\pi_\theta(a|S_0)q(S_0,a)\right]\right] \\
&\quad + \gamma^2\,\mathbb{E}\left[\nabla_\theta v_{\pi_\theta}(S_2)\right] \\
&= \mathbb{E}_{\mu,\pi_\theta}\left[\sum_{a\in\mathcal{A}} \nabla_\theta\pi_\theta(a|S_0)q(S_0,a)\right] \\
&\quad + \gamma\,\mathbb{E}_{\mu,\pi_\theta}\left[\sum_{a\in\mathcal{A}} \nabla_\theta\pi_\theta(a|S_0)q(S_1,a)\right] + \gamma^2\,\mathbb{E}\left[\nabla_\theta v_{\pi_\theta}(S_2)\right] \\
&= \cdots = \mathbb{E}_{\mu,\pi_\theta}\left[\sum_{t=0}^{+\infty}\gamma^t\sum_{a\in\mathcal{A}} \nabla_\theta\pi_\theta(a|S_t)q_{\pi_\theta}(S_t,a)\right] \\
&= \mathbb{E}_{\mu,\pi_\theta}\left[\sum_{t=0}^{+\infty}\gamma^t\sum_{a\in\mathcal{A}} \pi_\theta(a|S_t)\frac{\nabla_\theta\pi_\theta(a|S_t)}{\pi_\theta(a|S_t)}q_{\pi_\theta}(S_t,a)\right]
\end{aligned}$$

$$= \mathbb{E}_{\mu,\pi_\theta}\left[\sum_{t=0}^{+\infty} \gamma^t \frac{\nabla_\theta \pi_\theta(A_t|S_t)}{\pi_\theta(A_t|S_t)} q_{\pi_\theta}(S_t, A_t)\right].$$

Besides,

$$\mathbb{E}_{\mu,\pi_\theta}\left[\sum_{t=0}^{+\infty} \gamma^t b(S_t) \frac{\nabla \pi_\theta(A_t|S_t)}{\pi(A_t|S_t)}\right] = \mathbb{E}_{\mu,\pi_\theta}\left[\sum_{t=0}^{+\infty} \gamma^t b(S_t) \nabla_\theta \sum_{a\in\mathcal{A}} \pi_\theta(a|S_t)\right]$$
$$= \mathbb{E}_{\mu,\pi_\theta}\left[\sum_{t=0}^{+\infty} \gamma^t b(S_t) \nabla_\theta(1)\right],$$

and the result follows because $\sum_{a\in\mathcal{A}} \pi_\theta(a|S_t) = 1$ and therefore its gradient is zero. □

Remark 6.1.3 (Baseline functions). Remarkably, the above identity holds for any functions b_t, which are called *baseline functions*. The choice of baseline functions affects the variance of the estimator, as will be further discussed below.

6.2 REINFORCE

The REINFORCE algorithm simply applies the stochastic estimator of the gradient suggested by the following corollary.

Corollary 6.2.1. *Under the assumptions of Theorem 6.1.2, for all* $\theta \in \mathbb{R}^d$,

$$\nabla_\theta \mathbb{E}_{S\sim\mu}\left[v_{\pi_\theta}(S)\right] = \mathbb{E}_{\mu,\pi_\theta}\left[\sum_{t=0}^{+\infty} \gamma^t \left(\sum_{t'=1}^{+\infty} \gamma^{t'-1} R_{t+t'} - b_t(S_t)\right) \nabla_\theta \log \pi_\theta(A_t|S_t)\right].$$

Proof. Let $(S'_0, A'_0, R'_1, \dots) \sim \mathbb{P}_{\mu,\pi_\theta}$.

$$\nabla_\theta \mathbb{E}_{S\sim\mu}\left[v_{\pi_\theta}(S)\right] = \mathbb{E}\left[\sum_{t=0}^{+\infty} \gamma^t \left(q_{\pi_\theta}(S'_t, A'_t) - b_t(S'_t)\right) \nabla_\theta \log \pi_\theta(A'_t|S'_t)\right]$$
$$= \mathbb{E}\left[\sum_{t=0}^{+\infty} \gamma^t \left(\mathbb{E}_{S'_t,A'_t,\pi_\theta}\left[\sum_{t'=1}^{+\infty} \gamma^{t'-1} R_{t'}\right] - b_t(S'_t)\right) \nabla_\theta \log \pi_\theta(A'_t|S'_t)\right]$$
$$= \mathbb{E}\left[\sum_{t=0}^{+\infty} \gamma^t \left(\mathbb{E}\left[\sum_{t'=1}^{+\infty} \gamma^{t'-1} R_{t'+t} \,\middle|\, S'_t, A'_t\right] - b_t(S'_t)\right) \nabla_\theta \log \pi_\theta(A'_t|S'_t)\right]$$
$$= \mathbb{E}_{\mu,\pi_\theta}\left[\sum_{t=0}^{+\infty} \gamma^t \left(\sum_{t'=1}^{+\infty} \gamma^{t'-1} R'_{t+t'} - b_t(S'_t)\right) \nabla_\theta \log \pi_\theta(A'_t|S'_t)\right],$$

where we used the definition of q_{π_θ} for the second equality and the Markov property from Proposition 1.3.5 for the third equality. □

We first present the basic version with no baseline. Let $(H_k)_{k\geqslant 0}$ be a family of random infinite histories:

$$H_k = (S_{0,k}, A_{0,k}, R_{1,k}, \dots), \qquad k \geqslant 0,$$

and $(\theta_k)_{k\geqslant 0}$ a random sequence of parameters in $\mathbb{R}^d$. We assume that for all $k \geqslant 0$, the law of H_k conditionally on $H_0, \dots, H_{k-1}$ is $\mathbb{P}_{\mu,\pi_{\theta_k}}$. Then, $(\theta_k)_{k\geqslant 0}$ is a REINFORCE iteration if:

$$\theta_{k+1} = \theta_k + \alpha \sum_{t=0}^{+\infty} \gamma^t \left(\sum_{t'=1}^{+\infty} \gamma^{t'-1} R_{t+t',k} \right) \nabla_\theta \log \pi_{\theta_k}(A_{t,k}|S_{t,k}), \qquad k \geqslant 0,$$

where $\alpha > 0$ is the step-size (aka learning rate).

Remark 6.2.2. The above method requires the computation of infinite discounted rewards, which in practice can only be computed exactly for *episodic* MDPs (i.e. where a terminal state with zero future rewards is necessarily reached). For general MDPs, the above iteration can be approximated by truncation.

Algorithm 20: REINFORCE without baseline

Input: Parametric family of stationary policies $\{\pi_\theta\}_\theta$, number of episodes $N \geqslant 1$

$\theta \leftarrow 0_{\mathbb{R}^d}$
for $k = 1, \dots, N$ **do**
  $t \leftarrow 0$
  while *episode is not terminated nor truncated* **do**
    Observe and record S_t
    Draw and record $A_t \sim \pi_\theta(S_t)$
    Record R_{t+1}
    $t \leftarrow t + 1$
  $T \leftarrow t$
  $G \leftarrow 0$
  Choose $\alpha > 0$
  for $s = 0, \dots, T-1$ **do**
    $G \leftarrow R_{T-s} + \gamma \cdot G$
    Update $\theta \leftarrow \theta + \alpha \cdot \gamma^{T-1-s} \cdot G \cdot \nabla_\theta \log \pi_\theta(A_{T-1-s}|S_{T-1-s})$
return θ

Theorem 6.1.2 also allows the use of a *baseline*, which can reduce the variance of the stochastic estimator of the gradient; ideally, the baseline $b_t(S_t)$ should be, conditionally on S_t, the expectation of $q_{\pi_\theta}(S_t, A_t)$, in other words $v_{\pi_\theta}(S_t)$. Since the latter is unknown, one possibility is to estimate it using a parametric class of state-value function and a semi-gradient iteration.

Let $p \geqslant 1$ and $\{v_w\}_{w \in \mathbb{R}^p}$ a parametric family of state-value functions such that $w \mapsto v_w$ is differentiable. For all $k \geqslant 0$, we can consider the following REINFORCE iteration with baseline:

$$G_{t,k} = \sum_{t'=1}^{+\infty} \gamma^{t'-1} R_{t+t',k}$$

$$\theta_{k+1} = \theta_k + \alpha \sum_{t=0}^{+\infty} \gamma^t \left(G_{t,k} - v_{w_k}(S_{t,k})\right) \nabla_\theta \log \pi_{\theta_k}(A_{t,k}|S_{t,k}),$$

$$w_{k+1} = w_k + \beta \sum_{t=0}^{+\infty} \left(G_{t,k} - v_{w_k}(S_{t,k})\right) \nabla_w v_{w_k}(S_{t,k}),$$

where $\alpha > 0$ and $\beta > 0$ are step-sizes.

Algorithm 21: REINFORCE with baseline

Input: Parametric family of stationary policies $\{\pi_\theta\}_\theta$, parametric family of state-value functions $\{v_w\}_{w \in \mathbb{R}^p}$, number of episodes $N \geqslant 1$

$\theta \leftarrow 0_{\mathbb{R}^d}$
$w \leftarrow 0_{\mathbb{R}^p}$
for $k = 1, \ldots, N$ **do**
  $t \leftarrow 0$
  while *episode is not terminated nor truncated* **do**
    Observe and record S_t
    Draw and record $A_t \sim \pi_\theta(S_t)$
    Record R_{t+1}
    $t \leftarrow t + 1$
  $T \leftarrow t$
  $G \leftarrow 0$
  Choose $\alpha, \beta > 0$
  for $s = 0, \ldots, T-1$ **do**
    $G \leftarrow R_{T-s} + \gamma \cdot G$
    $\delta \leftarrow G - v_w(S_{T-1-s})$
    Update $w \leftarrow w + \beta \cdot \delta \cdot \nabla_w v_w(S_{T-1-s})$
    Update $\theta \leftarrow \theta + \alpha \cdot \gamma^{T-1-s} \cdot \delta \cdot \nabla_\theta \log \pi_\theta(A_{T-1-s}|S_{T-1-s})$
return θ

6.3 An actor-critic algorithm

We present present a variant of the above REINFORCE algorithm with baseline, where the quantity $q_{\pi_\theta}(S_t, A_t)$ from Theorem 6.1.2 is no longer estimated with a corresponding discounted sum of rewards, but approximated

with $R_{t+1} + \gamma v_w(S_{t+1})$, where w is the current estimate parameter. Then, updates can be performed after each step (without waiting the end of the episode). We write it for *episodic* problems.

Definition 6.3.1. A state $s \in \mathcal{S}$ is *terminal* if for all $a \in \mathcal{A}$, $p(\cdot\,|s,a) = \delta_{(0,s)}$. We denote $\mathcal{S}^*$ the set of terminal states.

Let $(H_k)_{k\geqslant 0}$ be a family of random infinite histories:

$$H_k = (S_{0,k}, A_{0,k}, R_{1,k}, \dots), \qquad k \geqslant 0,$$

and $(\theta_{t,k})_{t,k\geqslant 0}$ and $(w_{t,k})_{t,k\geqslant 0}$ random families of parameters in $\mathbb{R}^d$ and $\mathbb{R}^p$ respectively. We assume that for all $k \geqslant 0$ and $t \geqslant 0$, the law of $A_{t,k}$ conditionally on $H_0, \dots, H_{k-1}, S_{0,k}, A_{0,k}, R_{1,k}, \dots, S_{t,k}$ is $\pi_{\theta_{t,k}}(\cdot\,|S_{t,k})$. We assume that there exists a random sequence of integers $(T(k))_{k\geqslant 0}$ such that $S_{T(k)+1,k} \in \mathcal{S}^*$ almost-surely for all $k \geqslant 0$.

For all $k \geqslant 0$ and $0 \leqslant t \leqslant T(k)$, we consider

$$\Delta_{t,k} = R_{t+1,k} + \gamma v_{w_{t,k}}(S_{t+1,k}) - v_{w_{t,k}}(S_{t,k})$$

$$\theta_{t+1,k} = \theta_{t,k} + \alpha\gamma^t \Delta_{t,k} \nabla_\theta \log \pi_{\theta_{t,k}}(A_{t,k}|S_{t,k}),$$

$$w_{t+1,k} = w_{t,k} + \beta \Delta_{t,k} \nabla_w v_{w_{t,k}}(S_{t,k}),$$

$$\theta_{0,k+1} = \theta_{T(k)+1,k} \qquad \text{and} \qquad w_{0,k+1} = w_{T(k)+1,k}.$$

where $\alpha > 0$ and $\beta > 0$ are step-sizes.

Algorithm 22: An actor-critic algorithm

Input: Parametric family of stationary policies $\{\pi_\theta\}_{\theta\in\mathbb{R}^d}$,
parametric family of state-value functions $\{v_w\}_{w\in\mathbb{R}^p}$,
number of episodes $N \geqslant 1$.

$\theta \leftarrow 0_{\mathbb{R}^d}$
$w \leftarrow 0_{\mathbb{R}^p}$
for $k = 1, \dots, N$ **do**
  $t \leftarrow 0$
  Observe S_0
  while *episode is not terminated nor truncated* **do**
    Draw $A_t \sim \pi_\theta(S_t)$
    Observe (R_{t+1}, S_{t+1})
    $\delta \leftarrow R_{t+1} + \gamma v_w(S_{t+1}) - v_w(S_t)$
    Choose $\alpha, \beta > 0$
    Update $w \leftarrow w + \beta \cdot \delta \cdot \nabla_w v_w(S_t)$
    Update $\theta \leftarrow \theta + \alpha \cdot \gamma^t \cdot \delta \cdot \nabla_\theta \log \pi_\theta(A_t|S_t)$
    $t \leftarrow t + 1$
return θ