MARKOV DECISION PROCESSES: ON THE CONVERGENCE OF THE MONTE-CARLO FIRST VISIT ALGORITHM

SYLVAIN DELATTRE AND NICOLAS FOURNIER

Abstract. We consider the Monte-Carlo first visit algorithm, of which the goal is to find the optimal control in a Markov decision process with finite state space and finite number of possible actions. We show its convergence when the discount factor is smaller than 1/2.

1. Introduction

This paper deals with some algorithms of which the goal is to find the optimal policy in a Markov decision process (MDP) with finite state/action set. A MDP is a model where an agent interacts with its environment: at each step, the agent lies in some state x, chooses some action a, and the environment chooses at random, with some law depending on (x, a), the next state to which the agent goes.

In this context, Monte-Carlo algorithms are based on simulating a sequence of episodes, to find the optimal policy. We consider their optimistic versions, in that we modify, after each episode, the policy used to produce the next episode. Such algorithms are model free, in that we do not suppose that we know the transition probabilities: we only assume we can interact with the MDP by producing some episodes, given a policy.

There are two main available Monte Carlo algorithms: the *first visit algorithm* (FVA) and the *every visit algorithm* (EVA), see Sutton-Barto [7, Section 5.1]. In the FVA (resp. EVA), we update the policy by estimating the value function at each pair (x, a) of state/action using only the first visit at (x, a) (resp. all the visits at (x, a)) of each episode.

Let us quote Sutton-Barto [7, page 99], discussing about the convergence of the FVA/EVA: "In our opinion, this is one of the most fundamental open theoretical questions in reinforcement learning (for a partial solution, see Tsitsiklis, 2002)."

We study here the FVA, which is easier than the EVA from a theoretical point of view. Assuming that the value to be maximized is the expectation of $\sum_{t\in\mathbb{N}} \gamma^t R_{t+1}$, where R_t is the reward at step $t \in \mathbb{N}_*$, we are able to prove that when $\gamma \in [0, 1/2)$, the FVA produces some policy of which the value function converges to that of the optimal policy. Actually, we consider a more general family of (abstract) algorithms and show with counter example (strongly inspired by Example 5.12 in Bertsekas-Tsitsiklis [1]) that the convergence cannot hold true when $\gamma \geq 1/2$ for this general family. However, the counter example is far from concerning the FVA: it shows that our proof breaks down, not at all that the FVA does not converge.

Tsitsiklis [8] studies, among other algorithms, a *synchronous optimistic policy iteration algorithm*, which is a simplified synchronous version of the FVA: each iteration consists of simulating

2020 *Mathematics Subject Classification.* 90C40, 60J20.

Key words and phrases. Markov decision processes, First-visit algorithm.

many trajectories, one per couple state/action. More or less, this implies that the step $\alpha_k(x,a)$ is deterministic and does not depend on (x,a) in the general algorithm, see Subsection 2.3 below. He shows the convergence of his algorithm for any $\gamma \in [0,1)$.

Liu [2] extends the result of [8] to the case $\gamma = 1$, assuming that all the policies are proper: some final state space is reached with probability 1, starting from any state and using any policy.

Wang-Yuan-Shao-Ross [9] show the convergence of the FVA, with $\gamma = 1$ (but, as they mention, this extends to any $\gamma \in [0,1]$), assuming a structural condition on the MDP: it has to be *optimal policy feed forward*, meaning that a state cannot be visited twice under any optimal policy.

Winnicki-Srikant [10] study some related algorithm that uses lookahead policies. The computational cost of this algorithm seems larger. They show the convergence of the state values under a smallness condition on γ, depending on two parameters H and m used in their algorithm.

2. Notation and results

2.1. **The model.** The following objects are fixed in the whole paper.

Setting 1. *Let $\mathcal{X}$ be a non-empty finite state space and, for each $x \in \mathcal{X}$, let $\mathcal{A}_x$ be a non-empty finite set of possible actions. We also set $\mathcal{Z} = \{(x,a) : x \in \mathcal{X}, a \in \mathcal{A}_x\}$. For each $(x,a) \in \mathcal{Z}$, we consider a probability measure $P(x,a,\cdot)$ on $\mathcal{X}$. We also consider, for each $(x,a) \in \mathcal{Z}$ and $y \in \mathcal{X}$, a probability measure $S(x,a,y,\cdot)$ on $\mathbb{R}$ satisfying $\int_{\mathbb{R}} z^2 S(x,a,y,\mathrm{d}z) < \infty$ and we set*

$$g(x,a,y) = \int_{\mathbb{R}} zS(x,a,y,\mathrm{d}z). \tag{1}$$

Finally, we consider a real number $\gamma \in [0,1)$.

For $t \in \mathbb{N}$, we denote by $\mathcal{B}_t = \{(x_0,a_0,\dots,x_{t-1},a_{t-1},x_t) : x_i \in \mathcal{X}, a_i \in \mathcal{A}_{x_i}\}$. A *policy* is a family $\Pi = (\Pi_t)_{t\geq 0}$, where for each $t \geq 0$, for each $h = (x_0,a_0,\dots,x_{t-1},a_{t-1},x_t) \in \mathcal{B}_t$, $\Pi_t(h,\cdot)$ is a probability measure on $\mathcal{A}_{x_t}$.

A policy $\Pi = (\Pi_t)_{t\geq 0}$ is said to be SM (for stationary Markov) if there is a family $(\pi(x,\cdot))_{x\in\mathcal{X}}$, with $\pi(x,\cdot)$ a probability measure on $\mathcal{A}_x$, such that for all $t \geq 0$, all $h = (x_0,a_0,\dots,x_{t-1},a_{t-1},x_t) \in \mathcal{B}_t$, $\Pi_t(h,\cdot) = \pi(x_t,\cdot)$. In such a case, we simply say that $\pi = (\pi(x,\cdot))_{x\in\mathcal{X}}$ is a SM policy.

Given a starting point $x \in \mathcal{X}$ and a policy Π, we build recursively the stochastic process $(X_t,A_t)_{t\in\mathbb{N}}$ as follows: we set $X_0 = x$, we build A_0 with law $\Pi_0(X_0,\cdot)$ and, assuming that we have built $(X_s,A_s)_{s\in[\![0,t-1]\!]}$ for some $t \geq 1$, we first build X_t with conditional law $P(X_{t-1},A_{t-1},\cdot)$ knowing $(X_s,A_s)_{s\in[\![0,t-1]\!]}$, we set $H_t = (X_0,A_0,\dots,X_{t-1},A_{t-1},X_t)$ and we build A_t with conditional law $\Pi_t(H_t,\cdot)$ knowing H_t.

At each step, there is a reward: conditionally on the whole process $(X_s,A_s)_{s\in\mathbb{N}}$ the family of rewards $(R_s)_{s\in\mathbb{N}_*}$ is independent and for each $t \geq 1$, the reward R_t is $S(X_{t-1},A_{t-1},X_t,\cdot)$-distributed. The total reward is then given by

$$G = \sum_{t\in\mathbb{N}} \gamma^t R_{t+1} \qquad \text{(convention: } 0^0 = 1\text{).} \tag{2}$$

We indicate in subscript of the probability $\mathbb{P}_{x,\Pi}$ and expectation $\mathbb{E}_{x,\Pi}$ the starting point $x \in \mathcal{X}$ and the policy Π used to build the above random variable G, and we consider the value function

$$V_\Pi(x) = \mathbb{E}_{x,\Pi}[G]. \tag{3}$$

For π a SM policy, we simply denote by $\mathbb{P}_{x,\pi}$, $\mathbb{E}_{x,\pi}$ and $V_\pi(x)$ the corresponding objects. Finally, we set, for $x \in \mathcal{X}$,

$$V^*(x) = \sup\{V_\Pi(x),\ \Pi \ \text{ policy}\}.$$

2.2. **Optimal policy.** The existence of an optimal policy, which is SM and does not depend on the starting point, is well known, see Puterman [3] and Sutton-Barto [7]. The proofs are handled in the Appendix A for the sake of completeness. Things can be summarized as follows. For $(x, a) \in \mathcal{Z}$, we set

$$Q^*(x,a) = r(x,a) + \gamma PV^*(x,a), \qquad \text{where} \qquad r(x,a) = \sum_{y\in\mathcal{X}} P(x,a,y)g(x,a,y) \tag{4}$$

and where $PV^*(x,a) = \sum_{y\in\mathcal{X}} P(x,a,y)V^*(y)$. Observe that $r(x,a)$ stands for the mean (instantaneous) reward, when the process lies in state x and when one chooses the action a.

For π a SM policy, for $x, y \in \mathcal{X}$ and $a \in \mathcal{A}_x$, we set

$$r_\pi(x) = \sum_{a\in\mathcal{A}_x} r(x,a)\pi(x,a), \qquad P_\pi(x,y) = \sum_{a\in\mathcal{A}_x} P(x,a,y)\pi(x,a), \tag{5}$$

$$\text{and} \quad Q_\pi(x,a) = r(x,a) + \gamma PV_\pi(x,a), \tag{6}$$

where $PV_\pi(x,a) = \sum_{y\in\mathcal{X}} P(x,a,y)V_\pi(y)$. Observe that $r_\pi(x)$ represents the mean (instantaneous) reward, when the process lies in state x and when using the policy π, while $Q_\pi(x,a)$ stands for the mean total reward, when starting from the state x, when choosing a as first action, and when using the policy π for the rest of the process.

Theorem 2. *(i) Consider a SM policy π^* such that*

$$\pi^*\Big(x, \arg\max Q^*(x,\cdot)\Big) = 1 \quad \textit{for all } x \in \mathcal{X}. \tag{7}$$

Then $V_{\pi^*}(x) = V^*(x)$ for all $x \in \mathcal{X}$.

(ii) For π a SM policy, we have $Q_\pi(x,a) \le Q^*(x,a)$ for all $(x,a) \in \mathcal{Z}$ and

$$V_\pi(x) = r_\pi(x) + \gamma P_\pi V_\pi(x) = \sum_{a\in\mathcal{A}_x} Q_\pi(x,a)\pi(x,a) \qquad \textit{for all } x \in \mathcal{X},$$

$$Q_\pi(x,a) = r(x,a) + \gamma \sum_{(y,b)\in\mathcal{Z}} P(x,a,y)\pi(y,b)Q_\pi(y,b) \qquad \textit{for all } (x,a) \in \mathcal{Z}.$$

(iii) If π^* satisfies (7)*, then $Q_{\pi^*} = Q^*$. Moreover,*

$$V^*(x) = \max_{a\in\mathcal{A}_x} Q^*(x,a) \qquad \textit{for all } x \in \mathcal{X},$$

$$Q^*(x,a) = r(x,a) + \gamma \sum_{y\in\mathcal{X}} P(x,a,y) \max_{b\in\mathcal{A}_y} Q^*(y,b) \qquad \textit{for all } (x,a) \in \mathcal{Z}.$$

As a consequence, we may and will, from now on, consider only some SM policies.

2.3. **A general class of algorithms.** We consider the following family of algorithms.

General algorithm

Start with some deterministic function $\hat{Q}^0 : \mathcal{Z} \to \mathbb{R}$. For $k \ge 0$, assume that k first episodes

$$(X_0^i, A_0^i, X_1^i, R_1^i, A_1^i, X_2^i, R_2^i, A_2^i, \dots), \quad i = 1, \dots, k,$$

have already been built, and consider the sigma-field $\mathcal{F}^k$ generated by all these random variables (with $\mathcal{F}^0 = \{\emptyset, \Omega\}$). We also introduce

$$\mathcal{G}^i_n = \sigma(X^i_0, A^i_0, X^i_1, R^i_1, A^i_1, \dots, X^i_n, R^i_n, A^i_n),$$

so that $\mathcal{F}^k = \vee_{i=1}^k \mathcal{G}^i_\infty$. Assume also that the $\mathcal{F}^k$-measurable (random) function $\hat{Q}^k : \mathcal{Z} \to \mathbb{R}$ has been built. For some $\mathcal{F}^k$-measurable (random) function $\varepsilon_k : \mathcal{X} \to [0,1]$, we define the policy

$$\hat{\pi}^k(x,a) = \frac{\varepsilon_k(x)}{|\mathcal{A}_x|} + \mathbf{1}_{\{a \in \arg\max \hat{Q}^k(x,\cdot)\}} \frac{1-\varepsilon_k(x)}{|\arg\max \hat{Q}^k(x,\cdot)|}. \tag{8}$$

For some $\mathcal{F}^k$-measurable (random) probability ν_k on $\mathcal{Z}$, we build the $(k+1)$-th episode

$$(X^{k+1}_0, A^{k+1}_0, X^{k+1}_1, R^{k+1}_1, A^{k+1}_1, X^{k+1}_2, R^{k+1}_2, A^{k+1}_2, \dots)$$

using $(X^{k+1}_0, A^{k+1}_0) \sim \nu_k$ and then the SM policy $\hat{\pi}^k$, as in Subsection 2.1. For each $(x,a) \in \mathcal{Z}$, we consider the $(\mathcal{G}^{k+1}_n)_{n\geq 0}$-stopping-time

$$\tau^{k+1}_{x,a} = \inf\{t \geq 0 : (X^{k+1}_t, A^{k+1}_t) = (x,a)\} \qquad \text{(convention: } \inf\emptyset = \infty).$$

For all $(x,a) \in \mathcal{Z}$, for some $\mathcal{F}^k \vee \mathcal{G}^{k+1}_{\tau^{k+1}_{x,a}}$-measurable random variable $\alpha_k(x,a) \in [0,1]$, we set

$$\hat{Q}^{k+1}(x,a) = (1 - \alpha_k(x,a)\mathbf{1}_{\{\tau^{k+1}_{x,a}<\infty\}})\hat{Q}^k(x,a) + \alpha_k(x,a)\mathbf{1}_{\{\tau^{k+1}_{x,a}<\infty\}} G^{k+1}_{x,a}, \tag{9}$$

$$\text{where} \qquad G^{k+1}_{x,a} = \sum_{t\in\mathbb{N}} \gamma^t R^{k+1}_{\tau^{k+1}_{x,a}+t+1} \qquad \text{on } \{\tau^{k+1}_{x,a} < \infty\}. \tag{10}$$

This algorithm depends on the choices of the initial function $\hat{Q}^0$, the family of functions $(\varepsilon_k)_{k\geq 0}$ and $(\alpha_k)_{k\geq 0}$ and the family of probability measure $(\nu_k)_{k\geq 0}$. As we will see in the next subsection, for some specific choice of these parameters, this gives the so-called first-visit algorithm.

Theorem 3. *Assume that $\gamma \in [0,1/2)$. Consider the algorithm defined above and assume that*

$$\begin{cases} \textit{a.s., for all } (x,a) \in \mathcal{Z}, \quad \lim_{k\to\infty} \varepsilon_k(x) = 0, \quad \sum_{k\geq 1} \alpha_k(x,a)\mathbf{1}_{\{\tau^{k+1}_{x,a}<\infty\}} = \infty, \\ \qquad\qquad \textit{and} \quad \sum_{k\geq 1}(\alpha_k(x,a))^2 \mathbf{1}_{\{\tau^{k+1}_{x,a}<\infty\}} < \infty. \end{cases} \tag{11}$$

Then a.s., $\lim_{k\to\infty} V_{\hat{\pi}^k}(x) = V^*(x)$ for all $x \in \mathcal{X}$.

Although this is not very convincing as far as the first-visit algorithm is concerned, the condition that $\gamma \in [0,1/2)$, or at least that $\gamma \in [0,1/2]$, is necessary in the above theorem. This is strongly inspired by Example 5.12 in Bertsekas-Tsitsiklis [1].

Proposition 4. *Consider the simplest possible Markov decision process: take a state space $\mathcal{X} = \{e\}$ with one element, take $\mathcal{A}_0 = \{0,1\}$, so that necessarily $P(e,0,e) = P(e,1,e) = 1$, and take $S(e,0,e,\cdot) = \delta_0$ and $S(e,1,e,\cdot) = \delta_1$. If $\gamma \in (1/2,1) \cap \mathbb{Q}$, we can design the parameters of the above general algorithm (with in particular $\varepsilon_k(e) = 0$ for all $k \geq 0$) such that* (11) *holds true but $\lim_{k\to\infty} V_{\hat{\pi}^k}(e)$ does a.s. not exist.*

We assume that $\gamma \in (1/2,1) \cap \mathbb{Q}$ for simplicity. It is likely that this result extends to any $\gamma \in (1/2,1)$ with a little more work.

2.4. **The first-visit algorithm.** The algorithm presented in this section is very classical in reinforcement learning, see Singh-Sutton [6] and Sutton-Barto [7].

First visit algorithm

We consider $\theta > 0$ and a law μ_0 on $\mathcal{X}$. We set $\bar{Q}^0(x,a) = 0$ for all $(x,a) \in \mathcal{Z}$. For $k \geq 0$, assume that k first episodes

$$(X_0^i, A_0^i, X_1^i, R_1^i, A_1^i, X_2^i, R_2^i, A_2^i, \dots), \quad i = 1, \dots, k,$$

have been built, as well as the function $\bar{Q}^k : \mathcal{Z} \to \mathbb{R}$. For $(x,a) \in \mathcal{Z}$, set

$$N_k(x,a) = \sum_{i=1}^{k} \mathbf{1}_{\{\tau^i_{x,a} < \infty\}} \qquad \text{and} \qquad N_k(x) = \sum_{a \in \mathcal{A}_x} N_k(x,a) \qquad (\text{convention: } \textstyle\sum_{i=1}^{0} = 0), \tag{12}$$

$$\text{where} \qquad \tau^i_{x,a} = \inf\{t \geq 0 : (X_t^i, A_t^i) = (x,a)\} \qquad (\text{convention: } \inf \emptyset = \infty).$$

Consider the SM policy defined, for $(x,a) \in \mathcal{Z}$, by

$$\bar{\pi}^k(x,a) = \frac{\varepsilon_k(x)}{|\mathcal{A}_x|} + \mathbf{1}_{\{a \in \arg\max \bar{Q}^k(x,\cdot)\}} \frac{1 - \varepsilon_k(x)}{|\arg\max \bar{Q}^k(x,\cdot)|}, \quad \text{where} \quad \varepsilon_k(x) = \frac{1}{(1 + N_k(x))^\theta}. \tag{13}$$

We then build the $(k+1)$-th episode

$$(X_0^{k+1}, A_0^{k+1}, X_1^{k+1}, R_1^{k+1}, A_1^{k+1}, X_2^{k+1}, R_2^{k+1}, A_2^{k+1}, \dots)$$

using $X_0^{k+1} \sim \mu_0$ and the policy $\bar{\pi}^k$, as in Subsection 2.1. We then compute, for each $(x,a) \in \mathcal{Z}$,

$$\bar{Q}^{k+1}(x,a) = \frac{\sum_{i=1}^{k+1} \mathbf{1}_{\{\tau^i_{x,a} < \infty\}} G^i_{x,a}}{N_{k+1}(x,a)} \qquad (\text{convention: } \tfrac{0}{0} = 0), \tag{14}$$

$$\text{where} \qquad G^i_{x,a} = \sum_{t \in \mathbb{N}} \gamma^t R^i_{\tau^i_{x,a}+t+1} \qquad (\text{convention : } G^i_{x,a} = 0 \text{ when } \tau^i_{x,a} = \infty). \tag{15}$$

Theorem 5. *Assume that $\gamma \in [0, 1/2)$. If $\theta \in (0,1]$ and $\mu_0(x) > 0$ for all $x \in \mathcal{X}$, then a.s., $\lim_k V_{\bar{\pi}^k}(x) = V^*(x)$ for all $x \in \mathcal{X}$.*

2.5. **A more realistic situation.** In the previous subsection, we were using some episodes with infinite length. This is of course not realistic. We now consider the following common situation. For $x, y \in \mathcal{X}$, we set $Q(x,y) = \max_{a \in \mathcal{A}_x} P(x,a,y)$.

Assumption 6. *There is a subset $\triangle$ of $\mathcal{X}$ such that $S(x,a,y,\mathrm{d}z) = \delta_0(\mathrm{d}z)$ for all $x \in \Delta$, all $a \in \mathcal{A}_x$, all $y \in \mathcal{X}$ and such that $P(x,a,\triangle) = 1$ for all $x \in \triangle$, all $a \in \mathcal{A}_x$. Moreover, for all $x \in \mathcal{X}$, there are $y \in \triangle$ and $n \geq 0$ such that $Q^n(x,y) > 0$.*

In other words, the reward is identically equal to 0 when the process lies in $\triangle$, whatever the action. Moreover, when the process reaches $\triangle$, it remains stuck in $\triangle$ forever. Finally, it is possible to reach $\triangle$ from anywhere, at least if choosing suitably the actions. We will check the following.

Remark 7. *Grant Assumption 6*

(i) Consider two SM policies π, π' such that $\pi(x,a) = \pi'(x,a)$ for all $x \in \mathcal{X} \setminus \{\triangle\}$, all $a \in \mathcal{A}_x$. Then $V_\pi(x) = V_{\pi'}(x)$ for all $x \in \mathcal{X}$.

(ii) Consider any positive *SM policy π, i.e. such that $\pi(x,a) > 0$ for all $(x,a) \in \mathcal{Z}$. Then $\mathbb{P}_{x,\pi}(T_\triangle < \infty) = 1$ for all $x \in \mathcal{X}$, where $T_\triangle = \inf\{t \geq 0 : X_t \in \triangle\}$.*

Remark 8. *Grant Assumption 6. Consider the first-visit algorithm with finite episodes below, using the same random elements as in the first-visit algorithm above. It holds that* $\tilde\pi^k(x,a) = \bar\pi^k(x,a)$ *for all* $k \geq 0$*, all* $x \in \mathcal{X} \setminus \{\triangle\}$*, all* $a \in \mathcal{A}_x$*. Hence* $V_{\tilde\pi^k}(x) = V_{\bar\pi^k}(x)$ *for all* $x \in \mathcal{X}$ *by Remark 7-(i).*

First visit algorithm with finite episodes

We consider $\theta > 0$ and a law μ_0 on $\mathcal{X}$. We set $\tilde Q^0(x,a) = 0$ for all $(x,a) \in \mathcal{Z}$. For $k \geq 0$, assume that k first finite episodes

$$(X_0^i, A_0^i, X_1^i, R_1^i, A_1^i, \dots, X^i_{T^i_{\triangle-1}}, R^i_{T^i_{\triangle-1}}, A^i_{T^i_{\triangle-1}}, X^i_{T^i_{\triangle}}), \quad i = 1, \dots, k,$$

have been built, with $T^i_\triangle = \inf\{t \geq 0 : X^i_t \in \triangle\}$, as well as the function $\tilde Q^k : \mathcal{Z} \to \mathbb{R}$. For $(x,a) \in \mathcal{Z}$, we set

$$(16) \quad \tilde N_k(x,a) = \sum_{i=1}^k \mathbf{1}_{\{\tilde\tau^i_{x,a} < \infty\}} \qquad \text{and} \qquad \tilde N_k(x) = \sum_{a \in \mathcal{A}_x} \tilde N_k(x,a) \qquad \text{(convention: } \textstyle\sum_{i=1}^0 = 0\text{)},$$

$$\text{where} \qquad \tilde\tau^i_{x,a} = \inf\{t \in \llbracket 0, T^i_\triangle \rrbracket : (X^i_t, A^i_t) = (x,a)\} \qquad \text{(convention: } \inf\emptyset = \infty\text{)}.$$

Consider the SM policy defined, for $(x,a) \in \mathcal{Z}$, by

$$(17) \quad \tilde\pi^k(x,a) = \frac{\tilde\varepsilon_k(x)}{|\mathcal{A}_x|} + \mathbf{1}_{\{a \in \arg\max \tilde Q^k(x,\cdot)\}} \frac{1 - \tilde\varepsilon_k(x)}{|\arg\max \tilde Q^k(x,\cdot)|}, \quad \text{where} \quad \tilde\varepsilon_k(x) = \frac{1}{(1 + \tilde N_k(x))^\theta}.$$

We then build the $(k+1)$-th episode

$$(X_0^{k+1}, A_0^{k+1}, X_1^{k+1}, R_1^{k+1}, A_1^{k+1}, \dots, X^{k+1}_{T^{k+1}_{\triangle-1}}, R^{k+1}_{T^{k+1}_{\triangle-1}}, A^{k+1}_{T^{k+1}_{\triangle-1}}, X^{k+1}_{T^{k+1}_{\triangle}})$$

using $X_0^{k+1} \sim \mu_0$ and the policy $\tilde\pi^k$, as in Subsection 2.1. We then compute, for each $(x,a) \in \mathcal{Z}$,

$$(18) \qquad \tilde Q^{k+1}(x,a) = \frac{\sum_{i=1}^{k+1} \mathbf{1}_{\{\tilde\tau^i_{x,a} < \infty\}} \tilde G^i_{x,a}}{\tilde N_{k+1}(x,a)} \qquad \text{(convention: } \tfrac{0}{0} = 0\text{)},$$

$$(19) \qquad \text{where} \qquad \tilde G^i_{x,a} = \sum_{t=0}^{T^i_\triangle} \gamma^t R^i_{\tilde\tau^i_{x,a}+t+1} \qquad \text{(convention : } \tilde G^i_{x,a} = 0 \text{ when } \tilde\tau^i_{x,a} = \infty\text{)}.$$

2.6. **Plan of the paper.** In Section 3, we prove a crucial contraction result. Section 4 is devoted to the proof of an abstract convergence result. This abstract result is applied to show Theorem 3, that concerns the general algorithm, in Section 5. We show in Section 6 that Theorem 5, that concerns the first-visit algorithm, can be deduced from Theorem 3. We check the results of Subsection 2.5, concerning the case where the episodes are finite, in Section 7. In Section 8, we prove Proposition 4 through a counter-example, which shows that our strategy cannot be extended to the case where $\gamma \in (1/2, 1)$. Finally, we quickly recall in Appendix A the proofs of some known results, namely Theorem 2 and a simple version of the Robbins-Monro lemma, see Lemma 12.

3. Main contraction result

The following contraction estimate is the key of our study.

Lemma 9. *Assume that $\gamma \in [0,1)$. For a function $\varepsilon : \mathcal{X} \to [0,1]$, for a function $q : \mathcal{Z} \to \mathbb{R}$ and for $x \in \mathcal{X}$, we define the probability measure $\pi_q^\varepsilon(x,\cdot)$ on $\mathcal{A}_x$ by*

$$\pi_q^\varepsilon(x,a) = \frac{\varepsilon(x)}{|\mathcal{A}_x|} + \mathbf{1}_{\{a \in \arg\max q(x,\cdot)\}} \frac{1-\varepsilon(x)}{|\arg\max q(x,\cdot)|}.$$

We also introduce the function $\mathcal{H}(\varepsilon,q) : \mathcal{Z} \to \mathbb{R}$ defined by $\mathcal{H}(\varepsilon,q)(x,a) = Q_{\pi_q^\varepsilon}(x,a)$, recall (6). It holds that $\mathcal{H}(\varepsilon,q)(x,a) \le Q^*(x,a)$ for all $(x,a) \in \mathcal{Z}$ and

$$||\mathcal{H}(\varepsilon,q) - Q^*||_\infty \le \frac{\gamma}{1-\gamma}\Big(||Q^* - q||_\infty + ||(q-Q^*)_+||_\infty + 2||Q^*||_\infty ||\varepsilon||_\infty\Big),$$

where $(q-Q^*)_+(x,a) = \max\{q(x,a) - Q^*(x,a), 0\}$ and where $||\cdot||_\infty$ stand for $\max$ norms.

Proof. First, know from Theorem 2-(ii) that $\mathcal{H}(\varepsilon,q)(x,a) = Q_{\pi_q^\varepsilon}(x,a) \le Q^*(x,a)$ for all $(x,a) \in \mathcal{Z}$. Next, by Theorem 2-(ii)-(iii), we have, for $(x,a) \in \mathcal{Z}$,

$$\begin{aligned}
0 \le Q^*(x,a) - Q_{\pi_q^\varepsilon}(x,a) =& \gamma \sum_{y \in \mathcal{X}} P(x,a,y)\Big[\max_{b \in \mathcal{A}_y} Q^*(y,b) - \sum_{b \in \mathcal{A}_y} \pi_q^\varepsilon(y,b) Q_{\pi_q^\varepsilon}(y,b)\Big] \\
=& \gamma \sum_{y \in \mathcal{X}} P(x,a,y)\Big[\max_{b \in \mathcal{A}_y} Q^*(y,b) - \sum_{b \in \mathcal{A}_y} \pi_q^\varepsilon(y,b) Q^*(y,b)\Big] \\
&+ \gamma \sum_{y \in \mathcal{X}} P(x,a,y) \sum_{b \in \mathcal{A}_y} \pi_q^\varepsilon(y,b)\Big[Q^*(y,b) - Q_{\pi_q^\varepsilon}(y,b)\Big] \\
\le& \gamma \sum_{y \in \mathcal{X}} P(x,a,y)\Delta(y) + \gamma ||Q^* - Q_{\pi_q^\varepsilon}||_\infty,
\end{aligned}$$

where

$$\Delta(y) := \max_{b \in \mathcal{A}_y} Q^*(y,b) - \sum_{b \in \mathcal{A}_y} \pi_q^\varepsilon(y,b) Q^*(y,b).$$

We write $\Delta(y) = \Delta_1(y) + \Delta_2(y) + \Delta_3(y) + \Delta_4(y)$, with

$$\begin{aligned}
\Delta_1(y) &= \max_{b \in \mathcal{A}_y} Q^*(y,b) - \max_{b \in \mathcal{A}_y} q(y,b), & \Delta_2(y) &= \max_{b \in \mathcal{A}_y} q(y,b) - \sum_{b \in \mathcal{A}_x} q(y,b)\pi_q(y,b), \\
\Delta_3(y) &= \sum_{b \in \mathcal{A}_y} [q(y,b) - Q^*(y,b)]\pi_q(y,b), & \Delta_4(y) &= \sum_{b \in \mathcal{A}_y} Q^*(y,b)[\pi_q(y,b) - \pi_q^\varepsilon(y,b)],
\end{aligned}$$

where we have set

$$\pi_q(x,a) = \mathbf{1}_{\{a \in \arg\max q(x,\cdot)\}} \frac{1}{|\arg\max q(x,\cdot)|} \qquad \text{for all } (x,a) \in \mathcal{Z}.$$

We clearly have $\Delta_1(y) \le ||Q^* - q||_\infty$, as well as $\Delta_3(y) \le ||(q - Q^*)_+||_\infty$. By definition of π_q, it holds that $\Delta_2(y) = 0$. Finally,

$$\Delta_4(y) \le ||Q^*||_\infty ||\pi_q - \pi_q^\varepsilon|| \ \le 2||Q^*||_\infty ||\varepsilon||_\infty.$$

Thus

$$\Delta(y) \le ||Q^* - q||_\infty + ||(q - Q^*)_+||_\infty + 2||Q^*||_\infty ||\varepsilon||_\infty,$$

whence

$$\sum_{y \in \mathcal{X}} P(x,a,y)\Delta(y) \le ||Q^* - q||_\infty + ||(q - Q^*)_+||_\infty + 2||Q^*||_\infty ||\varepsilon||_\infty.$$

All in all, we have proved that

$$0 \le Q^*(x,a) - Q_{\pi_q^\varepsilon}(x,a) \le \gamma\Big(||Q^* - q||_\infty + ||(q - Q^*)_+||_\infty + 2||Q^*||_\infty ||\varepsilon||_\infty\Big) + \gamma ||Q^* - Q_{\pi_q^\varepsilon}||_\infty.$$

We end with

$$||Q^* - Q_{\pi_q^\varepsilon}||_\infty \le \gamma\Big(||Q^* - q||_\infty + ||(q - Q^*)_+||_\infty + 2||Q^*||_\infty||\varepsilon||_\infty\Big) + \gamma||Q^* - Q_{\pi_q^\varepsilon}||_\infty,$$

from which the result readily follows, recalling that $\mathcal{H}(\varepsilon, q) = Q_{\pi_q^\varepsilon}$. □

4. A general convergence result

We consider a finite set $\mathcal{Y}$ and denote by $\mathbf{F}(\mathcal{Y}, [0,1])$ the set of all functions from $\mathcal{Y}$ to $[0,1]$ and by $\mathbf{F}(\mathcal{Y}, \mathbb{R})$ the set of all functions from $\mathcal{Y}$ to $\mathbb{R}$.

Assumption 10. *There exists $f_* \in \mathbf{F}(\mathcal{Y}, \mathbb{R})$, $\rho \in (0,1)$ and $\beta > 0$ such that the function*

$$\mathcal{M} : \mathbf{F}(\mathcal{Y}, [0,1]) \times \mathbf{F}(\mathcal{Y}, \mathbb{R}) \to \mathbf{F}(\mathcal{Y}, \mathbb{R})$$

satisfies, for all $f \in \mathbf{F}(\mathcal{Y}, \mathbb{R})$, all $\eta \in \mathbf{F}(\mathcal{Y}, [0,1])$,

$$\mathcal{M}(\eta, f)(y) \le f_*(y) \qquad \text{for all } y \in \mathcal{Y},$$
$$||f_* - \mathcal{M}(\eta, f)||_\infty \le \rho||f_* - f||_\infty + \beta||(f - f_*)_+||_\infty + \beta||\eta||_\infty.$$

Our main results will be deduced from the following general abstract result. A similar, with less complex filtrations, is stated by Singh-Jaakkola-Littman-Szepesvári [5].

Proposition 11. *Assume that $\mathcal{M}$ satisfies Assumption 10 and consider*

- *a family $(\mathcal{I}_y^k)_{k\ge 0, y\in\mathcal{Y}}$ of σ-fields such that $\mathcal{I}_y^k \subset \mathcal{I}_{y'}^{k+1}$ for all $k \ge 0$, all $y, y' \in \mathcal{Y}$,*
- *some random $\hat f_0 \in \mathbf{F}(\mathcal{Y}, \mathbb{R})$ such that $\hat f_0(y)$ is $\mathcal{I}_y^0$-measurable for all $y \in \mathcal{Y}$,*
- *for each $k \ge 0$, some random $\eta_k \in \mathbf{F}(\mathcal{Y}, [0,1])$ which is $\mathcal{I}_y^k$-measurable for all $y \in \mathcal{Y}$.*
- *for each $k \ge 1$, some random $\xi_k \in \mathbf{F}(\mathcal{Y}, \mathbb{R})$ such that $\xi_k(y)$ is $\mathcal{I}_y^k$-measurable for all $y \in \mathcal{Y}$,*
- *for each $k \ge 0$, some random $\lambda_k \in \mathbf{F}(\mathcal{Y}, [0,1])$ such that $\lambda_k(y)$ is $\mathcal{I}_y^k$-measurable for all $y \in \mathcal{Y}$.*

Assume moreover that a.s., for all $y \in \mathcal{Y}$, we have

$$\lim_k \eta_k(y) = 0, \qquad \sum_{k\ge 0} \lambda_k(y) = \infty, \qquad \sum_{k\ge 0} (\lambda_k(y))^2 < \infty$$

and that there is a constant $C > 0$ such that for all $k \ge 0$, all $y \in \mathcal{Y}$,

$$\mathbb{E}[\xi_{k+1}(y)|\mathcal{I}_y^k] = 0 \qquad \text{and} \qquad \mathbb{E}[(\xi_{k+1}(y))^2|\mathcal{I}_y^k] \le C.$$

Define recursively, for $k \ge 0$,

$$\hat f_{k+1}(y) = \hat f_k(y) + \lambda_k(y)\Big(\mathcal{M}(\eta_k, \hat f_k)(y) - \hat f_k(y) + \xi_{k+1}(y)\Big) \qquad \text{for all } y \in \mathcal{Y}.$$

Then a.s., for all $y \in \mathcal{Y}$, $\lim_{k\to\infty} \hat f_k(y) = f_*(y)$.

The proof of this result relies on the following simple version of the Robbins-Monro theorem [4]. See also its (short) proof in Appendix A.

Lemma 12. *Consider a filtration $(\mathcal{G}_k)_{k\ge 0}$, a $\mathcal{G}_0$-measurable real random variable Z_0, as well as some $(\mathcal{G}_k)_{k\ge 0}$-adapted sequences $(\theta_k)_{k\ge 0}$ and $(\zeta_k)_{k\ge 1}$ of real random variables, with θ_k valued in $[0,1]$. Assume that a.s., $\sum_{k\ge 0} \theta_k = \infty$ and $\sum_{k\ge 0} \theta_k^2 < \infty$ and that there is a constant $C > 0$ such that for all $k \ge 0$, $\mathbb{E}[\zeta_{k+1}|\mathcal{G}_k] = 0$ and $\mathbb{E}[(\zeta_{k+1})^2|\mathcal{G}_k] \le C$. Define recursively, for $k \ge 0$,*

$$Z_{k+1} = (1 - \theta_k)Z_k + \theta_k\zeta_{k+1}.$$

Almost surely, $\lim_k Z_k = 0$.

Proof of Proposition 11. First, one easily checks by induction that for all $k \geq 0$, all $y \in \mathcal{Y}$, $\hat{f}^k(y)$ is $\mathcal{I}_y^k$-measurable. Our goal is to show that a.s., for all $y \in \mathcal{Y}$, $\lim_k \hat{\varphi}_k(y) = 0$, where $\hat{\varphi}_k(y) = \hat{f}_k(y) - f_*(y)$. For all $k \geq 0$, all $y \in \mathcal{Y}$, we have

$$\begin{aligned}\hat{\varphi}_{k+1}(y) =&\hat{\varphi}_k(y) + \lambda_k(y)[\mathcal{M}(\eta_k, \hat{f}_k)(y) - f_*(y) - \hat{\varphi}_k(y) + \xi_{k+1}(y)]\\ =&(1-\lambda_k(y))\hat{\varphi}_k(y) + \lambda_k(y)[\mathcal{M}(\eta_k, \hat{f}_k)(y) - f_*(y)] + \lambda_k(y)\xi_{k+1}(y). \quad (20)\end{aligned}$$

We now divide the proof into several steps.

Step 1. We define $(W_k(y))_{k\geq 0, y \in \mathcal{Y}}$ by $W_0(y) = \hat{\varphi}_0(y)$ and, for $k \geq 0$,

$$W_{k+1}(y) = (1-\lambda_k(y))W_k(y) + \lambda_k(y)\xi_{k+1}(y).$$

For each fixed $y \in \mathcal{Y}$, we can apply Lemma 12 with $\mathcal{G}_k = \mathcal{I}_y^k$, $Z_0 = W_0(y)$, $\theta_k = \lambda_k(y)$, $\zeta_k = \xi_k(y)$, so that $Z_k = W_k(y)$. We get that a.s., for each $y \in \mathcal{Y}$, a.s., $\lim_k W_k(y) = 0$.

Step 2. We now show by induction that for all $k \geq 0$, all $y \in \mathcal{Y}$, $\hat{\varphi}_k(y) \leq W_k(y)$. This will imply, thanks to Step 1, that $\lim_k ||(\hat{\varphi}^k)_+||_\infty = 0$ a.s.

First, $\hat{\varphi}_0(y) = W_0(y)$ for all $y \in \mathcal{Y}$. If next $\hat{\varphi}_k(y) \leq W_k(y)$ for some $k \geq 0$, then using (20) and Assumption 10,

$$\hat{\varphi}_{k+1}(y) \leq (1-\lambda_k(y))\hat{\varphi}_k(y) + \lambda_k(y)\xi_{k+1}(y) \leq (1-\lambda_k(y))W_k(y) + \lambda_k(y)\xi_{k+1}(y) = W_{k+1}(y).$$

Step 3. We prove here that a.s., $\sup_{k\geq 0} ||\hat{\varphi}_k||_\infty < \infty$.

To this end, it suffices to prove that $\sup_{k\geq 0} ||\Delta_k||_\infty < \infty$, where $\Delta_k(y) = W_k(y) - \hat{\varphi}_k(y)$, because $\sup_{k\geq 0} ||W_k||_\infty < \infty$ a.s., since $\lim_k ||W_k||_\infty = 0$ a.s. by Step 1. We recall that $\Delta_k(y) \geq 0$ by Step 2 and write

$$\begin{aligned}\Delta_{k+1}(y) =&(1-\lambda_k(y))\Delta_k(y) + \lambda_k(y)|f_*(y) - \mathcal{M}(\eta_k, \hat{f}_k)(y)|\\ \leq&(1-\lambda_k(y))\Delta_k(y) + \lambda_k(y)[\rho||\hat{\varphi}_k||_\infty + \beta||(\hat{\varphi}_k)_+||_\infty + \beta||\eta_k||_\infty] \quad (21)\end{aligned}$$

by Assumption 10. Since $||\hat{\varphi}_k||_\infty \leq ||\Delta_k||_\infty + ||W_k||_\infty$, this gives

$$\Delta_{k+1}(y) \leq (1-\lambda_k(y))||\Delta_k||_\infty + \lambda_k(y)\rho||\Delta_k||_\infty + \lambda_k(y)H_k$$

where $H_k = \rho||W_k||_\infty + \beta||(\hat{\varphi}_k)_+||_\infty + \beta||\eta_k||_\infty$. Setting $K = \sup_{k\geq 0} H_k$, which is a.s. finite by Steps 1 and 2, we end with

$$\Delta_{k+1}(y) \leq (1-(1-\rho)\lambda_k(y))||\Delta_k||_\infty + \lambda_k(y)K.$$

If $||\Delta_k||_\infty \geq K/(1-\rho)$, we conclude that $|\Delta_{k+1}(y)| \leq ||\Delta_k||_\infty$.

If $||\Delta_k||_\infty < K/(1-\rho)$, we conclude that $|\Delta_{k+1}(y)| \leq K/(1-\rho)$.

We thus always have $||\Delta_{k+1}||_\infty \leq \max\{||\Delta_k||_\infty, K/(1-\rho)\}$, from which we easily conclude that for all $k \geq 0$, $||\Delta_k||_\infty \leq \max\{||\Delta_0||_\infty, K/(1-\rho)\}$.

Step 4. We fix $k_0 \geq 1$ and set

$$D_{k_0} = ||W_{k_0}||_\infty + \sup_{k\geq k_0}\Big[||\hat{\varphi}_k||_\infty + \frac{\beta}{\rho}\Big(||(\hat{\varphi}_k)_+||_\infty + ||\eta_k||_\infty\Big)\Big]. \quad (22)$$

For each $y \in \mathcal{Y}$ fixed, we define $(Y_k^{k_0}(y))_{k\geq k_0}$ by $Y_{k_0}^{k_0}(y) = D_{k_0}$ and by induction, for $k \geq k_0$,

$$Y_{k+1}^{k_0}(y) = (1-\lambda_k(y))Y_k^{k_0}(y) + \rho\lambda_k(y)D_{k_0}.$$

We show here that $\lim_k Y_k^{k_0}(y) = \rho D_{k_0}$ a.s.

It suffices to note that for all $k \geq k_0$, $(Y^{k_0}_{k+1}(y) - \rho D_{k_0}) = (1 - \lambda_k(y))(Y^{k_0}_k(y) - \rho D_{k_0})$, whence $Y^{k_0}_{k+1}(y) - \rho D_{k_0} = (1-\rho)D_{k_0}\prod_{\ell=k_0}^k(1-\lambda_\ell(y))$. This last quantity a.s. tends to 0 as $k \to \infty$ because by assumption, $\sum_{\ell \geq k_0}\lambda_\ell(y) = \infty$ a.s.

Step 5. Here we check that $\Delta_k(y) \leq Y^{k_0}_k(y)$ for all $k \geq k_0 \geq 0$ and all $y \in \mathcal{Y}$.

We obviously have $\Delta_{k_0}(y) \leq ||W_{k_0}||_\infty + ||\hat{\varphi}_{k_0}||_\infty \leq D_{k_0} = Y^{k_0}_{k_0}(y)$ and, assuming by induction that $\Delta_k(y) \leq Y^{k_0}_k(y)$ for some $k \geq k_0$, we write, recalling (21),

$$\begin{aligned}\Delta_{k+1}(y) \leq&(1-\lambda_k(y))Y^{k_0}_k(y) + \lambda_k(y)[\rho||\hat{\varphi}_k||_\infty + \beta||(\hat{\varphi}_k)_+||_\infty + \beta||\eta_k||_\infty]\\ =&(1-\lambda_k(y))Y^{k_0}_k(y) + \rho\lambda_k(y)\Big[||\hat{\varphi}_k||_\infty + \frac{\beta}{\rho}||(\hat{\varphi}_k)_+||_\infty + \beta||\eta_k||_\infty\Big]\\ \leq&(1-\lambda_k(y))Y^{k_0}_k(y) + \rho\lambda_k(y)D_{k_0}\\ =&Y^{k_0}_{k+1}(y).\end{aligned}$$

Step 6. Since $\lim_k ||W_k||_\infty = 0$ a.s. by Step 1, we conclude that a.s., $\limsup_k ||\hat{\varphi}_k||_\infty = \limsup_k ||\Delta_k||_\infty$. Hence by Steps 5 and 4 (and since $\Delta_k(y) \geq 0$, see Step 3), for any $k_0 \geq 0$, a.s.,

$$\limsup_k ||\hat{\varphi}_k||_\infty \leq \limsup_k ||Y^{k_0}_k||_\infty = \rho D_{k_0}.$$

As a consequence, we a.s. have

$$\limsup_k ||\hat{\varphi}_k||_\infty \leq \rho \limsup_{k_0} D_{k_0}.$$

Recalling the definition (22) of D_{k_0} and that $\lim_k ||W_k||_\infty = 0$ by Step 1, that $\lim_k ||(\hat{\varphi}_k)_+||_\infty = 0$ by Step 2 and that $\lim_k ||\eta_k||_\infty = 0$ by assumption, we end with

$$\limsup_k ||\hat{\varphi}_k||_\infty \leq \rho \limsup_k ||\hat{\varphi}_k||_\infty.$$

Since finally $\limsup_k ||\hat{\varphi}_k||_\infty < \infty$ by Step 3 and since $\rho \in (0,1)$, we find that $\limsup_k ||\hat{\varphi}_k||_\infty = 0$ a.s., which was our goal. □

5. Convergence of the general algorithm

Here we give the

Proof of Theorem 3. We divide the proof in two steps.

Step 1. We claim that it suffices to show that $\lim_k \hat{Q}^k(x,a) = Q^*(x,a)$ a.s. for all $(x,a) \in \mathcal{Z}$.

Indeed, assume this is the case. Observe that by construction, see (8) and Lemma 9,

$$\hat{\pi}^k = \pi^{\varepsilon_k}_{\hat{Q}^k}, \qquad \text{so that} \qquad Q_{\hat{\pi}^k} = \mathcal{H}(\varepsilon_k, \hat{Q}^k).$$

By Lemma 9, the facts that $\lim_k \varepsilon_k(x) = 0$ by assumption and that $\lim_k \hat{Q}^k(x,a) = Q^*(x,a)$ for all $(x,a) \in \mathcal{Z}$ imply that $\lim_k Q_{\hat{\pi}^k}(x,a) = Q^*(x,a)$ for all $(x,a) \in \mathcal{Z}$. But Theorem 2-(ii) tells us that for all $x \in \mathcal{X}$,

$$V_{\hat{\pi}^k}(x) = \sum_{a\in\mathcal{A}_x} Q_{\hat{\pi}^k}(x,a)\hat{\pi}^k(x,a) = \frac{\varepsilon_k(x)}{|\mathcal{A}_x|}\sum_{a\in\mathcal{A}_x} Q_{\hat{\pi}^k}(x,a) + (1-\varepsilon_k(x))\max_{a\in\mathcal{A}_x} Q_{\hat{\pi}^k}(x,a)$$

by definition of $\hat{\pi}^k$, see (8) again. Using that $\lim_k \varepsilon_k(x) = 0$ and $\lim_k Q_{\hat{\pi}^k}(x,a) = Q^*(x,a)$, we conclude that

$$\lim_k V_{\hat{\pi}^k}(x) = \max_{a\in\mathcal{A}_x} Q^*(x,a),$$

which equals $V^*(x)$ by Theorem 2-(iii).

Step 2. To prove that $\lim_k \hat{Q}^k(x,a) = Q^*(x,a)$ a.s. for all $(x,a) \in \mathcal{Z}$, we aim to apply Proposition 11 with $\mathcal{Y} = \mathcal{Z}$, with $\hat{f}^k = \hat{Q}^k$ for all $k \geq 0$, with the map, $\mathcal{M}(\varepsilon, q) = \mathcal{H}(\varepsilon, q) = Q_{\pi_q^\varepsilon}$ for all $\varepsilon \in \mathbf{F}(\mathcal{Z}, [0,1]))$ and all $q \in \mathbf{F}(\mathcal{Z}, \mathbb{R})$, with $\eta_k(z) = \varepsilon_k(x)$ and $\lambda_k(z) = \mathbf{1}_{\{\tau_{x,a}^{k+1}<\infty\}}\alpha_k(x,a)$ for all $k \geq 0$, all $z = (x,a) \in \mathcal{Z}$, with

$$\xi_k(z) = \mathbf{1}_{\{\tau_{x,a}^{k}<\infty\}}\Big[G_{x,a}^k - \mathcal{H}(\varepsilon_{k-1}, \hat{Q}^{k-1})(x,a)\Big] \qquad \text{for all } k \geq 1, \text{ all } z = (x,a) \in \mathcal{Z},$$

and with the family of filtrations $(\mathcal{I}_z^k)_{k\geq 0, z\in\mathcal{Z}}$ defined by (for $z = (x,a)$)

$$\mathcal{I}_z^k = \mathcal{F}^k \vee \mathcal{G}^{k+1}_{\tau_{x,a}^{k+1}},$$

where $\mathcal{G}_\ell^k = \sigma(X_0^k, A_0^k, R_1^k, X_1^k, A_1^k, \dots, R_\ell^k, X_\ell^k, A_\ell^k)$ and $\mathcal{F}^k = \vee_{i=1}^k \mathcal{G}_\infty^i$ (and $\mathcal{F}^0 = \{\emptyset, \Omega\}$).

First, Lemma 9 tells us that $\mathcal{H}$ satisfies Assumption 10 with $f_* = Q^*$, with $\rho = \gamma/(1-\gamma) \in (0,1)$ (recall that $\gamma \in (0,1/2)$) and $\beta = [\gamma/(1-\gamma)]\max\{1, 2||Q^*||_\infty\}$.

We obviously have $\mathcal{I}_z^k \subset \mathcal{I}_{z'}^{k+1}$ for all $k \geq 0$, all $z, z' \in \mathcal{Z}$. Moreover, $\hat{Q}^0$ is deterministic by assumption and thus $\hat{Q}^0(z)$ is $\mathcal{I}_z^0$-measurable for all $z \in \mathcal{Z}$. For each $k \geq 0$, each $z = (x,a) \in \mathcal{Z}$, $\eta_k(z) = \varepsilon_k(x)$ is $\mathcal{F}^k$-measurable by assumption and thus $\vee_{z'\in\mathcal{Z}}\mathcal{I}_{z'}^k$-measurable. For each $k \geq 1$, each $z = (x,a) \in \mathcal{Z}$, $\xi_k(z) = \mathbf{1}_{\{\tau_{x,a}^k<\infty\}}[G_{x,a}^k - \mathcal{H}(\varepsilon_{k-1}, \hat{Q}^{k-1})(x,a)]$ is $\mathcal{F}^k$-measurable and thus $\mathcal{I}_z^k$-measurable. Finally, for each $k \geq 0$, each $z = (x,a) \in \mathcal{Z}$, $\lambda_k(z) = \mathbf{1}_{\{\tau_{x,a}^{k+1}<\infty\}}\alpha_k(x,a)$ is $\mathcal{I}_z^k$-measurable, since $\alpha_k(x,a)$ is $\mathcal{F}^k \vee \mathcal{G}^{k+1}_{\tau_{x,a}^{k+1}}$-measurable by assumption.

We next observe that for all $k \geq 0$, all $z = (x,a) \in \mathcal{Z}$, recalling (9),

$$\begin{aligned}
\hat{Q}^{k+1}(z) =& (1-\lambda_k(z))\hat{Q}^k(z) + \lambda_k(z)G_{x,a}^{k+1} \\
=& \hat{Q}^k(z) + \lambda_k(z)\Big(\mathcal{H}(\varepsilon_k, \hat{Q}^k)(z) - \hat{Q}^k(z) + [G_{x,a}^{k+1} - \mathcal{H}(\varepsilon_k, \hat{Q}^k)(z)]\Big) \\
=& \hat{Q}^k(z) + \lambda_k(z)\Big(\mathcal{H}(\varepsilon_k, \hat{Q}^k)(z) - \hat{Q}^k(z) + \mathbf{1}_{\{\tau_{x,a}^{k+1}<\infty\}}[G_{x,a}^{k+1} - \mathcal{H}(\varepsilon_k, \hat{Q}^k)(z)]\Big)
\end{aligned}$$

because $\lambda_k(z) = 0$ when $\tau_{x,a}^{k+1} = \infty$. Hence

$$\hat{Q}^{k+1}(z) = \hat{Q}^k(z) + \lambda_k(z)\Big(\mathcal{M}(\eta_k, \hat{Q}^k)(z) - \hat{Q}^k(z) + \xi_{k+1}(z)\Big)$$

as desired.

For $k \geq 0$ and $z = (x,a) \in \mathcal{Z}$, it holds that

$$\mathbb{E}[\xi_{k+1}(z)|\mathcal{I}_z^k] = \mathbf{1}_{\{\tau_{x,a}^{k+1}<\infty\}}\Big(\mathbb{E}[G_{x,a}^{k+1}|\mathcal{I}_z^k] - \mathcal{H}(\varepsilon_k, \hat{Q}^k)(z)\Big) = 0.$$

Indeed, on $\{\tau_{x,a}^{k+1}<\infty\}$, we have $\mathbb{E}[G_{x,a}^{k+1}|\mathcal{I}_z^k]=Q_{\hat\pi^k}(x,a)$ (i.e. $\mathbb{E}[G_{x,a}^{k+1}|\mathcal{I}_z^k]=\mathcal{H}(\varepsilon_k,\hat Q^k)(x,a)$, see Step 1): recalling (10),

$$\begin{aligned}\mathbb{E}[G_{x,a}^{k+1}|\mathcal{I}_z^k]=&\mathbb{E}\Big[R^{k+1}_{\tau_{x,a}^{k+1}+1}\Big|\mathcal{I}^k_{x,a}\Big]+\gamma\mathbb{E}\Big[\sum_{t\in\mathbb{N}}\gamma^tR^{k+1}_{\tau^{k+1}_{x,a}+2}\Big|\mathcal{I}^k_{x,a}\Big]\\=&\sum_{y\in\mathcal{X}}P(x,a,y)g(x,a,y)+\gamma\mathbb{E}\Big[\mathbb{E}_{X_{\tau^{k+1}_{x,a}+1},\hat\pi^k}[G]\Big|\mathcal{I}^k_{x,a}\Big]\end{aligned}$$

by the strong Markov property, recall (1) and (2). Recalling that $V_\pi(x)=E_{x,\pi}[G]$, (4) and (6), this gives

$$\mathbb{E}[G_{x,a}^{k+1}|\mathcal{I}_z^k]=r(x,a)+\gamma\sum_{y\in\mathcal{X}}P(x,a,y)V_{\hat\pi^k}(y)=Q_{\hat\pi^k}(x,a)$$

as desired.

Moreover,

$$\mathbb{E}[(\xi_{k+1}(z))^2|\mathcal{I}_z^k]=\mathbf{1}_{\{\tau^{k+1}_{x,a}<\infty\}}\mathrm{Var}[G^{k+1}_{x,a}|\mathcal{I}^k_z]\le\mathbf{1}_{\{\tau^{k+1}_{x,a}<\infty\}}\mathbb{E}[(G^{k+1}_{x,a})^2|\mathcal{I}^k_z].$$

As previously, using that $(u+v)^2\le 2u^2+2v^2$, we find that on $\{\tau^{k+1}_{x,a}<\infty\}$,

$$\mathbb{E}[(\xi_{k+1}(z))^2|\mathcal{I}_z^k]\le 2\sum_{y\in\mathcal{X}}P(x,a,y)\int_{\mathbb{R}}z^2S(x,a,y,\mathrm{d}z)+2\gamma^2\mathbb{E}\Big[\mathbb{E}_{X_{\tau^{k+1}_{x,a}+1},\hat\pi^k}[G^2]\Big|\mathcal{I}^k_{x,a}\Big].$$

The first term is bounded by $2K$, where $K=\sup_{(x,a)\in\mathcal{Z},y\in\mathcal{X}}\int_{\mathbb{R}}z^2S(x,a,y,\mathrm{d}z)$ (see Setting 1). The second term is bounded, because $\sup_{x\in\mathcal{X},\pi}\mathbb{E}_{x,\pi}[G^2]<\infty$: thanks to the Minkowski inequality,

$$\mathbb{E}_{x,\pi}[G^2]\le\Big(\sum_{t\in\mathbb{N}}\gamma^t\mathbb{E}_{x,\pi}[R^2_{t+1}]^{1/2}\Big)^2\le\frac{K}{(1-\gamma)^2}.$$

All in all, we have proved that

$$\mathbb{E}[(\xi_{k+1}(z))^2|\mathcal{I}_z^k]\le 2K+\frac{2\gamma^2K}{(1-\gamma)^2}.$$

By assumption, see (11), the three conditions $\lim_k\eta_k(z)=\lim_k\varepsilon_k(x)=0$, $\sum_{k\ge0}\lambda_k(z)=\sum_{k\ge0}\mathbf{1}_{\{\tau^{k+1}_{x,a}<\infty\}}\alpha_k(x,a)=\infty$ and $\sum_{k\ge0}(\lambda_k(z))^2=\sum_{k\ge0}\mathbf{1}_{\{\tau^{k+1}_{x,a}<\infty\}}(\alpha_k(x,a))^2<\infty$ are a.s. fulfilled for all $z=(x,a)\in\mathcal{Z}$.

By Proposition 11, we deduce that a.s., for all $z=(x,a)\in\mathcal{Z}$, $\lim_k\hat Q^k(x,a)=Q^*(x,a)$. □

6. Convergence of the first-visit algorithm

We now handle the

Proof of Theorem 5. The first-visit algorithm is a particular case of the general algorithm, namely when choosing $\hat Q^0=0$ (so that $\hat\pi^0(x,a)=|\mathcal{A}_x|^{-1}$ for all $(x,a)\in\mathcal{Z}$) and, for all $k\ge0$ and $(x,a)\in\mathcal{Z}$, $\nu_k(x,a)=\mu_0(x)\hat\pi^k(x,a)$, $\varepsilon_k(x)=(1+N_k(x))^{-\theta}$ and $\alpha_k(x,a)=(N_{k+1}(x,a))^{-1}$, where we recall that (with the convention that $\sum_{i=1}^0=0$)

$$N_k(x,a)=\sum_{i=1}^k\mathbf{1}_{\{\tau^i_{x,a}<\infty\}}\qquad\text{and}\qquad N_k(x)=\sum_{a\in\mathcal{A}_x}N_k(x,a).$$

Observe that $\alpha_k(x,a)$ is indeed $\mathcal{F}^k \vee \mathcal{G}^{k+1}_{\tau^{k+1}_{x,a}}$-measurable. Indeed, the only issue is to check that for all $k \geq 0$, all $(x,a) \in \mathcal{Z}$, we have

$$\bar{Q}^{k+1}(x,a) = (1 - \alpha_k(x,a)\mathbf{1}_{\{\tau^{k+1}_{x,a}<\infty\}})\bar{Q}^k(x,a) + \alpha_k(x,a)\mathbf{1}_{\{\tau^{k+1}_{x,a}<\infty\}} G^{k+1}_{x,a},$$

but this immediately follows from the definition (14) of $\bar{Q}^{k+1}$ and the facts that $N_{k+1}(x,a) = N_k(x,a) + \mathbf{1}_{\{\tau^{k+1}_{x,a}<\infty\}}$ (when $k=0$, this also uses that $\bar{Q}^0(x,a) = 0$).

To show that Theorem 3 applies, we only have to verify (11). Assume for a moment that for all $(x,a) \in \mathcal{Z}$, $\lim_k N_k(x,a) = \infty$ a.s. Then of course $\lim_k N_k(x) = \infty$, whence $\lim_k \varepsilon_k(x) = 0$ a.s. Moreover,

$$\sum_{k\geq 0} \alpha_k(x,a)\mathbf{1}_{\{\tau^{k+1}_{x,a}<\infty\}} = \sum_{k\geq 1} \frac{\mathbf{1}_{\{\tau^k_{x,a}<\infty\}}}{N_k(x,a)} = \sum_{\ell\geq 1}\frac{1}{\ell} = \infty,$$

and

$$\sum_{k\geq 0} (\alpha_k(x,a))^2\mathbf{1}_{\{\tau^{k+1}_{x,a}<\infty\}} = \sum_{k\geq 1} \frac{\mathbf{1}_{\{\tau^k_{x,a}<\infty\}}}{(N_k(x,a))^2} = \sum_{\ell\geq 1}\frac{1}{\ell^2} < \infty.$$

It thus only remains to show that for each $(x,a) \in \mathcal{Z}$, that we now fix, $\lim_k N_k(x,a) = \infty$ a.s. First, since the family $(X^i_0)_{i\geq 1}$ is i.i.d. and μ_0 distributed with $\mu_0(x) > 0$, we conclude from the law of large numbers that a.s.,

$$N_k(x) \geq \sum_{i=1}^k \mathbf{1}_{\{X^i_0 = x\}} \sim \mu_0(x)k \qquad \text{as } k \to \infty. \tag{23}$$

Next, since

$$N_k(x,a) \geq \sum_{i=1}^k \mathbf{1}_{\{X^i_0 = x, A^i_0 = a\}} =: S_k,$$

it suffices to show that $\lim_k S_k = \infty$. We introduce $\Lambda_k = \sum_{i=1}^k \mathbb{P}(X^i_0 = x, A^i_0 = a|\mathcal{F}^{i-1})$, so that $(M_k = S_k - \Lambda_k)_{k\geq 0}$ is a $(\mathcal{F}^k)_{k\geq 0}$-martingale (with $S_0 = \Lambda_0 = 0$). We recall that by construction, see (13), for all $i \geq 1$,

$$\mathbb{P}(X^i_0 = x, A^i_0 = a|\mathcal{F}^{i-1}) = \mu_0(x)\bar{\pi}^{i-1}(x,a) \geq \mu_0(x)\frac{\varepsilon_{i-1}(x)}{|\mathcal{A}_x|} = \frac{\mu_0(x)}{|\mathcal{A}_x|(1+N_{i-1}(x))^\theta}.$$

Hence (23) implies that $\lim_k \Lambda_k = \infty$ a.s., because $\theta \in (0,1]$. For $A > 0$, we introduce the stopping time $\sigma_A = \inf\{k \geq 0 : S_k \geq A\}$. For all $k \geq 1$, We have $\mathbb{E}[M_{\sigma_A\wedge k}] = 0$, i.e. $\mathbb{E}[S_{\sigma_A\wedge k}] = \mathbb{E}[\Lambda_{\sigma_A\wedge k}]$. Hence $\mathbb{E}[\Lambda_{\sigma_A\wedge k}] \leq A$ whence, by monotone convergence, $\mathbb{E}[\Lambda_{\sigma_A}] \leq A$. This implies that $\Lambda_{\sigma_A} < \infty$ a.s., whence $\sigma_A < \infty$ a.s. because $\lim_k \Lambda_k = \infty$. Since this holds true for all $A > 0$, we conclude that $\lim_k S_k = \infty$ as desired. □

7. The case where the episodes are finite

In this section, we grant Assumption 6. We set $\mathcal{Z}_\triangle = \{(x,a) : x \in \mathcal{X} \setminus \triangle, a \in \mathcal{A}_x\}$.

Proof of Remark 7. We start with (i). For a given starting point $x \in \mathcal{X}$ and two SM policies π, π' coinciding on $\mathcal{Z}_\triangle$, we can build the corresponding processes $(X_0, A_0, X_1, R_1, A_1, \dots, X_t, R_t, A_t, \dots)$ and $(X'_0, A'_0, X'_1, R'_1, A'_1, \dots, X'_t, R'_t, A'_t, \dots)$ such that $T_\triangle = T'_\triangle$ and

$$\begin{aligned}(X_0, A_0, X_1, R_1, A_1, \dots, X_{T_\triangle-1}, &R_{T_\triangle-1}, A_{T_\triangle-1}, X_{T_\triangle})\\ &=(X'_0, A'_0, X'_1, R'_1, A'_1, \dots, X'_{T'_\triangle-1}, R'_{T'_\triangle-1}, A'_{T'_\triangle-1}, X'_{T'_\triangle}).\end{aligned}$$

where $T_\triangle, T'_\triangle$ are the entrance times of $(X_t)_{t\geq 0}$ and $(X'_t)_{t\geq 0}$ in $\triangle$. Moreover, $R_{t+1} = R'_{t+1} = 0$ for all $t \geq T_\triangle = T'_\triangle$, since the set $\triangle$ is absorbing and since $S(x,a,y,\mathrm{d}z) = \delta_0(\mathrm{d}z)$ as soon as $x \in \triangle$. Thus $G = G'$, where

$$G = \sum_{t\geq 0} \gamma^t R_{t+1} = \sum_{t=0}^{T_\triangle - 1} \gamma^t R_{t+1} \qquad \text{and} \qquad G' = \sum_{t\geq 0} \gamma^t R'_{t+1} = \sum_{t=0}^{T'_\triangle - 1} \gamma^t R'_{t+1},$$

with the convention that $\sum_{t=0}^{-1} = 0$. Recalling (3), we conclude that indeed, $V_\pi(x) = V_{\pi'}(x)$ for all $x \in \mathcal{X} \setminus \triangle$. Finally, $V_\pi(x) = V_{\pi'}(x) = 0$ when $x \in \triangle$.

For (ii), consider a *positive* SM policy π. Recall that when using π, the process $(X_t)_{t\geq 0}$ is a Markov chain with transition matrix $P_\pi(x,y) = \sum_{a\in\mathcal{A}_x} \pi(x,a)P(x,a,y)$. Clearly, $P_\pi(x,y) > 0$ if and only if $Q(x,y) = \max_{a_i n \mathcal{A}_x} P(x,a,y) > 0$. Hence we deduce from Assumption 6 that for all $x \in \mathcal{X}$, there are $y \in \triangle$ and $n \geq 0$ such that $P^n_\pi(x,y) > 0$. The state space $\mathcal{X}$ being finite, we classically conclude that indeed, $T_\triangle = \inf\{t \geq 0 : X_t \in \triangle\} < \infty$ a.s. under $\mathbb{P}_{x,\pi}$ for all $x \in \mathcal{X}$. □

We next explain the

Proof of Remark 8. Here we use simultaneously the *first-visit algorithm* and the *first-visit algorithm with finite episodes.*

For all $(x,a) \in \mathcal{Z}$, we have $\tilde{Q}^0(x,a) = \bar{Q}^0(x,a) = \tilde{N}_0(x,a) = N_0(x,a) = \tilde{N}_0(x) = N_0(x) = 0$.

Hence $\tilde{\pi}^0(x,a) = \bar{\pi}^0(x,a)$. Thus we can build the same first episode in both algorithms (stopped when reaching $\triangle$, at time $T^1_\triangle$, for the second one). Since $\triangle$ is absorbing whatever the policy and since $R^1_t = 0$ for all $t \geq T^1_\triangle$, we conclude that for all $(x,a) \in \mathcal{Z}_\triangle$, $\tilde{\tau}^1_{x,a} = \tau^1_{x,a}$ and $\tilde{G}^1_{x,a} = G^1_{x,a}$, whence $\tilde{N}_1(x,a) = N_1(x,a)$, $\tilde{N}_1(x) = N_1(x)$ and $\tilde{Q}^1(x,a) = \bar{Q}^1(x,a)$.

Thus $\tilde{\pi}^1(x,a) = \bar{\pi}^1(x,a)$ for all $(x,a) \in \mathcal{Z}_\triangle$, and we can build the same second episode in both algorithms (stopped when reaching $\triangle$, at time $T^1_\triangle$, for the second one). Observe that this does not require that $\tilde{\pi}^1(x,a) = \bar{\pi}^1(x,a)$ when $x \in \triangle$, because such values are never used in the second algorithm. Since $\triangle$ is absorbing whatever the policy and since $R^2_t = 0$ for all $t \geq T^2_\triangle$, we conclude that for all $(x,a) \in \mathcal{Z}_\triangle$, $\tilde{\tau}^2_{x,a} = \tau^2_{x,a}$ and $\tilde{G}^2_{x,a} = G^2_{x,a}$, whence $\tilde{N}_2(x,a) = N_2(x,a)$, $\tilde{N}_2(x) = N_2(x)$ and $\tilde{Q}^2(x,a) = \bar{Q}^2(x,a)$.

Iterating this argument, we find that $\tilde{\pi}^k(x,a) = \bar{\pi}^k(x,a)$ for all $(x,a) \in \mathcal{Z}_\triangle$, all $k \geq 0$. □

8. Counter-example

We assume here that $\gamma \in (1/2, 1) \cap \mathbb{Q}$ and we give the

Proof of Proposition 4. We recall that $\mathcal{X} = \{e\}$, that $\mathcal{A}_0 = \{0,1\}$, that $P(e,0,e) = P(e,1,e) = 1$, and that $S(e,0,e,\cdot) = \delta_0$ and $S(e,1,e,\cdot) = \delta_1$. In other words, there is only one possible state and two actions. When one chooses the action 0, the (deterministic) reward is 0, and when one chooses the action 1, the (deterministic) reward is 1. There are of course two extremal SM policies, that we denote by π_0 and π_1, and that are given by

$$\pi_0(e,0) = 1, \quad \pi_0(e,1) = 0 \qquad \text{and} \qquad \pi_1(e,0) = 0, \quad \pi_1(e,1) = 1.$$

Let us mention, although we will not use it, that using Theorem 2 and (6), with $r(e,0) = 0$, $r(e,1) = 1$ and $P_{\pi_0}(e,e) = P_{\pi_1}(e,e) = 1$, one finds that $Q^*(e,0) = \gamma/(1-\gamma)$, $Q^*(e,1) = 1/(1-\gamma)$,

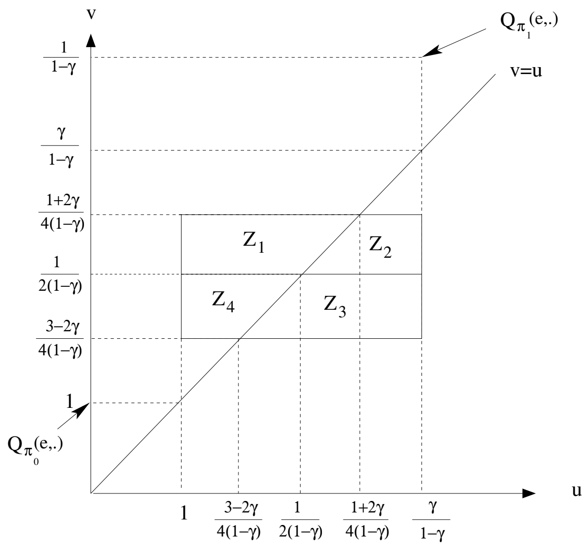
v
$Q_{\pi_1}(e,.)$
$\frac{1}{1-\gamma}$
v=u
$\frac{\gamma}{1-\gamma}$
$\frac{1+2\gamma}{4(1-\gamma)}$
Z_1
Z_2
$\frac{1}{2(1-\gamma)}$
Z_4
Z_3
$\frac{3-2\gamma}{4(1-\gamma)}$
1
$Q_{\pi_0}(e,.)$
u
1
$\frac{3-2\gamma}{4(1-\gamma)}$
$\frac{1}{2(1-\gamma)}$
$\frac{1+2\gamma}{4(1-\gamma)}$
$\frac{\gamma}{1-\gamma}$

FIGURE 1.

so that the optimal SM strategy is $\pi^* = \pi_1$ and that

$$V_{\pi_0}(e) = 0, \quad Q_{\pi_0}(e,0) = 0, \quad Q_{\pi_0}(e,1) = 1,$$
$$V_{\pi_1}(e) = \frac{1}{1-\gamma}, \quad Q_{\pi_1}(e,0) = \frac{\gamma}{1-\gamma}, \quad Q_{\pi_0}(e,1) = \frac{1}{1-\gamma}.$$

For $Q : \mathcal{Z} \to \mathbb{R}$, we will use the notation $Q = (u, v)$, where $u = Q(e, 0)$ and $v = Q(e, 1)$.

We now design $\hat{Q}^0 = (u_0, v_0)$ and the families $(\nu_k)_{k\geq 0}$ and $(\alpha_k)_{k\geq 0}$ such that when applying the general algorithm with $\varepsilon_k(e) = 0$ for all $k \geq 0$, the limit $\lim_k V_{\hat{\pi}^k}(e)$ does a.s. not exist.

Step 1. We introduce the zones, see Figure 1,

$$\begin{aligned}
Z_1 =& \Big\{(u,v) \in \mathbb{R}^2 : v > u > 1 \quad \text{and} \quad v \in \Big(\frac{1}{2(1-\gamma)}, \frac{1+2\gamma}{4(1-\gamma)}\Big)\Big\},\\
Z_2 =& \Big\{(u,v) \in \mathbb{R}^2 : v < u < \frac{\gamma}{1-\gamma} \quad \text{and} \quad v \in \Big(\frac{1}{2(1-\gamma)}, \frac{1+2\gamma}{4(1-\gamma)}\Big)\Big\},\\
Z_3 =& \Big\{(u,v) \in \mathbb{R}^2 : v < u < \frac{\gamma}{1-\gamma} \quad \text{and} \quad v \in \Big(\frac{3-2\gamma}{4(1-\gamma)}, \frac{1}{2(1-\gamma)}\Big)\Big\},\\
Z_4 =& \Big\{(u,v) \in \mathbb{R}^2 : v > u > 1 \quad \text{and} \quad v \in \Big(\frac{3-2\gamma}{4(1-\gamma)}, \frac{1}{2(1-\gamma)}\Big)\Big\}.
\end{aligned}$$

We choose (u_0, v_0) in Z_1 such that $u_0 \in \mathbb{Q}$ and $v_0 \in \mathbb{R} \setminus \mathbb{Q}$ and set $\hat{Q}^0 = (u_0, v_0)$. We also fix

$$q \in \mathbb{Q} \cap \Big(0, \frac{\gamma}{2} - \frac{1}{4}\Big), \quad \text{whence} \quad q < \frac{1}{4} \wedge \Big(\gamma - \frac{1}{2}\Big) \wedge \Big(\frac{\gamma}{2} - \frac{1}{4}\Big) \wedge \frac{2\gamma - 1}{3 - 2\gamma}. \tag{24}$$

For all $k \geq 0$ and for $a \in \{0,1\}$, we define (once the $k-1$ first episodes have been built)

$$\alpha_k(e,a) = \frac{q\rho_k(a)}{1+L_k(a)}, \quad \text{where} \quad L_k(a) = \sum_{i=0}^{k} \rho_i(a) \quad \text{and} \quad \rho_i(a) = \begin{cases} \mathbf{1}_{\{a=0\}} & \text{if } \hat{Q}^i \in Z_1 \cup Z_3, \\ \mathbf{1}_{\{a=1\}} & \text{if } \hat{Q}^i \in Z_2 \cup Z_4, \end{cases}$$

and we set

$$\nu_k(e,a) = \begin{cases} \mathbf{1}_{\{a=0\}} & \text{if } \hat{Q}^k \in Z_1 \cup Z_3, \\ \mathbf{1}_{\{a=1\}} & \text{if } \hat{Q}^k \in Z_2 \cup Z_4. \end{cases}$$

Step 2. We now check that for all $k \geq 0$, all $i \in \{1,2,3,4\}$, if $\hat{Q}^k = (u_k, v_k) \in Z_i$ with $u_k \in \mathbb{Q}$ and $v_k \in \mathbb{R} \setminus \mathbb{Q}$, then, with the convention that $Z_5 = Z_1$,

- $\hat{Q}^{k+1} = (u_{k+1}, v_{k+1}) \in Z_i \cup Z_{i+1}$ with also $u_{k+1} \in \mathbb{Q}$ and $v_{k+1} \in \mathbb{R} \setminus \mathbb{Q}$;
- there is $\ell > k$ such that $\hat{Q}^\ell \in Z_{i+1}$.

This in particular implies that $\hat{Q}^k$ belongs to $Z_1 \cup Z_2 \cup Z_3 \cup Z_4$ for all $k \geq 0$, so that the definitions of α_k and ν_k are sufficient to produce the whole algorithm.

(a) If first $\hat{Q}^k \in Z_1$, then $\hat{\pi}^k = \pi_1$ (recall (8) and that $\varepsilon_k \equiv 0$). Moreover, we have $\nu_k(e,a) = \mathbf{1}_{\{a=0\}}$. Hence $\tau_{e,0}^{k+1} = 0$ and $\tau_{e,1}^{k+1} = 1$, and $G_{e,0}^{k+1} = \gamma/(1-\gamma)$ and $G_{e,1}^{k+1} = 1/(1-\gamma)$. By (9) and by definition of α_k, we find

$$u_{k+1} = (1-\alpha_k(e,0))u_k + \alpha_k(e,0)\frac{\gamma}{1-\gamma} = u_k + \frac{q}{1+L_k(0)}\Big(\frac{\gamma}{1-\gamma} - u_k\Big) \quad \text{and} \quad v_{k+1} = v_k. \tag{25}$$

Hence $u_{k+1} \in \mathbb{Q}$ and $v_{k+1} \in \mathbb{R} \setminus \mathbb{Q}$, recall that $\gamma \in \mathbb{Q}$ by assumption. Moreover, $\hat{Q}^{k+1} \in Z_1 \cup Z_2$, because $v_{k+1} \neq u_{k+1}$ and, using that $u_k \in (1, (1+2\gamma)/[4(1-\gamma)])$ and that $q < 1/4$, see (24),

$$v_{k+1} = v_k \in \Big(\frac{1}{2(1-\gamma)}, \frac{1+2\gamma}{4(1-\gamma)}\Big),$$

$$1 < u_k < u_{k+1} < u_k + q\Big(\frac{\gamma}{1-\gamma} - u_k\Big) < \frac{1+2\gamma}{4(1-\gamma)} + q\frac{2\gamma-1}{1-\gamma} < \frac{\gamma}{1-\gamma}.$$

Assume next that $\hat{Q}^\ell \in Z_1$ for all $\ell \geq k$, then for all $n \geq k$, we would have $L_n(0) = L_k(0) + \ell - k$ and, by (25) and since $u_n < (1+2\gamma)/[4(1-\gamma)]$ for all $n \geq k$, for all $\ell \geq k$,

$$u_\ell = u_k + \sum_{n=k}^{\ell-1} \Big(\frac{\gamma}{1-\gamma} - u_n\Big)\frac{q}{1+L_n(0)} \geq u_k + \frac{(2\gamma-1)q}{4(1-\gamma)} \sum_{n=k}^{\ell-1} \frac{1}{1+L_k(0)+n-k}.$$

This would imply that $u_\ell \to \infty$ as $\ell \to \infty$ and contradict the fact that $\hat{Q}^\ell \in Z_1$ for all $\ell \geq k$.

(b) If next $\hat{Q}^k \in Z_2$, then $\hat{\pi}^k = \pi_0$ and $\nu_k(e,a) = \mathbf{1}_{\{a=1\}}$. Hence $\tau_{e,0}^{k+1} = 1$ and $\tau_{e,1}^{k+1} = 0$, while $G_{e,0}^{k+1} = 0$ and $G_{e,1}^{k+1} = 1$. By (9) and by definition of α_k, we find

$$u_{k+1} = u_k \qquad \text{and} \qquad v_{k+1} = (1-\alpha_k(e,1))v_k + \alpha_k(e,1) \times 1 = v_k - \frac{q}{1+L_k(1)}(v_k - 1). \tag{26}$$

Thus $u_{k+1} \in \mathbb{Q}$ and $v_{k+1} \in \mathbb{R} \setminus \mathbb{Q}$. Moreover, $\hat{Q}^{k+1} \in Z_2 \cup Z_3$, because $u_{k+1} \neq v_{k+1}$ and, since $v_k > 1/[2(1-\gamma)]$ and since $q < \gamma - 1/2$, see (24),

$$u_{k+1} = u_k \in \Big(\frac{1}{2(1-\gamma)}, \frac{\gamma}{1-\gamma}\Big),$$

$$u_{k+1} = u_k > v_k > v_{k+1} > v_k - q(v_k - 1) > v_k(1-q) > \frac{1-q}{2(1-\gamma)} > \frac{3-2\gamma}{4(1-\gamma)}.$$

If $\hat{Q}^\ell \in Z_2$ for all $\ell \geq k$, then for all $n \geq k$, we would have $L_n(1) = L_k(1) + n - k$ and, by (26) and since $v_n > 1/[2(1-\gamma)]$ for all $n \geq k$, for all $\ell \geq k$,

$$v_\ell = v_k - \sum_{n=k}^{\ell-1} \frac{q(v_n - 1)}{1 + L_n(1)} \leq v_k - \frac{(2\gamma-1)q}{2(1-\gamma)} \sum_{n=k}^{\ell-1} \frac{1}{1 + L_k(1) + n - k}.$$

This would imply that $v_\ell \to -\infty$ as $\ell \to \infty$ and contradict the fact that $\hat{Q}^\ell \in Z_2$ for all $\ell \geq k$.

(c) If now $\hat{Q}^k \in Z_3$, then $\hat{\pi}^k = \pi_0$ and $\nu_k(e,a) = \mathbf{1}_{\{a=0\}}$. Hence $\tau_{e,0}^{k+1} = 0$ and $\tau_{e,1}^{k+1} = \infty$, and $G_{e,0}^{k+1} = 0$. By (9) and by definition of α_k, we find

$$u_{k+1} = (1 - \alpha_k(e,0))u_k + \alpha_k(e,0) \times 0 = u_k - \frac{q}{1 + L_k(0)} u_k \qquad \text{and} \qquad v_{k+1} = v_k. \tag{27}$$

Thus $u_{k+1} \in \mathbb{Q}$ and $v_{k+1} \in \mathbb{R} \setminus \mathbb{Q}$. Moreover, $\hat{Q}^{k+1} \in Z_3 \cup Z_4$, because $v_{k+1} \neq u_{k+1}$ and, since $u_k > (3-2\gamma)/[4(1-\gamma)]$ and $q < (2\gamma-1)/(3-2\gamma)$, see (24),

$$v_{k+1} = v_k \in \Big(\frac{3-2\gamma}{4(1-\gamma)}, \frac{1}{2(1-\gamma)}\Big),$$

$$\frac{\gamma}{1-\gamma} > u_k > u_{k+1} > u_k(1-q) > \frac{(1-q)(3-2\gamma)}{4(1-\gamma)} > 1.$$

If $\hat{Q}^\ell \in Z_3$ for all $\ell \geq k$, then for all $n \geq k$, we would have $L_n(0) = L_k(0) + n - k$ and, by (27) and since $u_n > 1$ for all $n \geq k$, for all $\ell \geq k$,

$$u_\ell = u_k - \sum_{n=k}^{\ell-1} \frac{q u_n}{1 + L_n(0)} \leq u_k - q \sum_{n=k}^{\ell-1} \frac{1}{1 + L_k(0) + n - k}.$$

This would imply that $u_\ell \to -\infty$ as $\ell \to \infty$ and contradict the fact that $\hat{Q}^\ell \in Z_3$ for all $\ell \geq k$.

(d) If finally $\hat{Q}^k \in Z_4$, then $\hat{\pi}^k = \pi_1$ and $\nu_k(e,a) = \mathbf{1}_{\{a=1\}}$. Hence $\tau_{e,0}^{k+1} = \infty$ and $\tau_{e,1}^{k+1} = 0$, and $G_{e,1}^{k+1} = \frac{1}{1-\gamma}$. By (9) and by definition of α_k, we find

$$u_{k+1} = u_k \qquad \text{and} \qquad v_{k+1} = (1 - \alpha_k(e,1))v_k + \frac{\alpha_k(e,1)}{1-\gamma} = v_k + \frac{q}{1 + L_k(1)} \Big(\frac{1}{1-\gamma} - v_k\Big). \tag{28}$$

Thus $u_{k+1} \in \mathbb{Q}$ and $v_{k+1} \in \mathbb{R} \setminus \mathbb{Q}$. Moreover, $\hat{Q}^{k+1} \in Z_4 \cup Z_1$, because $v_{k+1} \neq u_{k+1}$ and, since $v_k < 1/[2(1-\gamma)]$ and since $q < \gamma/2 - 1/4$, see (24),

$$u_{k+1} = u_k \in \Big(1, \frac{1}{2(1-\gamma)}\Big),$$

$$1 < u_{k+1} = u_k < v_k < v_{k+1} < v_k + q\Big(\frac{1}{1-\gamma} - v_k\Big) < \frac{1}{2(1-\gamma)} + \frac{q}{1-\gamma} < \frac{1+2\gamma}{4(1-\gamma)}.$$

If $\hat{Q}^\ell \in Z_4$ for all $\ell \geq k$, then for all $n \geq k$, we would have $L_n(1) = L_k(1) + n - k$ and, by (28) and since $v_n < 1/[2(1-\gamma)]$ for all $n \geq k$, for all $\ell \geq k$,

$$v_\ell = v_k + \sum_{n=k}^{\ell-1} \frac{q}{1 + L_n(1)} \Big(\frac{1}{1-\gamma} - v_n\Big) \geq v_k + \frac{q}{2(1-\gamma)} \sum_{n=k}^{\ell-1} \frac{1}{1 + L_k(1) + n - k}.$$

This would imply that $v_\ell \to \infty$ as $\ell \to \infty$ and contradict the fact that $\hat{Q}^\ell \in Z_4$ for all $\ell \geq k$.

Step 3. We conclude that $V_{\hat{\pi}^k}(e)$ does not converge as $k \to \infty$, because $V_{\hat{\pi}^k}(e) = V_{\pi_1}(e) = 1/(1-\gamma)$ for those k's such that $\hat{Q}^k \in Z_1 \cup Z_4$ (there are infinitely many such k's by Step 2), while

$V_{\hat{\pi}^k}(e) = V_{\pi_0}(e) = 0$ for those k's such that $\hat{Q}^k \in Z_2 \cup Z_3$ (there are infinitely many such k's by Step 2). It only remains to show (11). But, recalling the definition of α_k and that $\tau_{e,0}^{k+1} < \infty$ when $\hat{Q}^k \in Z_1 \cup Z_3$, see Steps 2-(a)-(c), we find

$$\sum_{k\geq 0} \alpha_k(e,0)\mathbf{1}_{\{\tau_{e,0}^{k+1}<\infty\}} = \sum_{k\geq 0} \frac{q\mathbf{1}_{\{\hat{Q}^k\in Z_1\cup Z_3\}}}{1+\sum_{i=0}^k \mathbf{1}_{\{\hat{Q}^i\in Z_1\cup Z_3\}}} = \sum_{\ell\geq 1} \frac{q}{1+\ell} = \infty.$$

We used that $\sum_{i=0}^k \mathbf{1}_{\{\hat{Q}^i\in Z_1\cup Z_3\}} = \infty$ by Step 2. Next,

$$\sum_{k\geq 0} (\alpha_k(e,0))^2\mathbf{1}_{\{\tau_{e,0}^{k+1}<\infty\}} = \sum_{k\geq 0} \frac{q^2\mathbf{1}_{\{\hat{Q}^k\in Z_1\cup Z_3\}}}{(1+\sum_{i=0}^k \mathbf{1}_{\{\hat{Q}^i\in Z_1\cup Z_3\}})^2} = \sum_{\ell\geq 1} \frac{q^2}{(1+\ell)^2} < \infty.$$

Similarly, using that $\tau_{e,1}^{k+1} < \infty$ when $\hat{Q}^k \in Z_2 \cup Z_4$, see Steps 2-(b)-(d),

$$\sum_{k\geq 0} \alpha_k(e,1)\mathbf{1}_{\{\tau_{e,1}^{k+1}<\infty\}} = \sum_{k\geq 0} \frac{q\mathbf{1}_{\{\hat{Q}^k\in Z_2\cup Z_4\}}}{1+\sum_{i=0}^k \mathbf{1}_{\{\hat{Q}^i\in Z_2\cup Z_4\}}} = \sum_{\ell\geq 1} \frac{q}{1+\ell} = \infty,$$

because $\sum_{i=0}^k \mathbf{1}_{\{\hat{Q}^i\in Z_2\cup Z_4\}} = \infty$ by Step 2, and

$$\sum_{k\geq 0} (\alpha_k(e,1))^2\mathbf{1}_{\{\tau_{e,1}^{k+1}<\infty\}} = \sum_{k\geq 0} \frac{q^2\mathbf{1}_{\{\hat{Q}^k\in Z_2\cup Z_4\}}}{(1+\sum_{i=0}^k \mathbf{1}_{\{\hat{Q}^i\in Z_2\cup Z_4\}})^2} = \sum_{\ell\geq 1} \frac{q^2}{(1+\ell)^2} < \infty.$$

The proof is complete. □

Appendix A. Quick proofs of know results

Here we recall, for the sake of completeness, the

Proof of Theorem 2. We adopt the notation introduced in Subsections 2.1 and 2.2.

Step 1. For $f : \mathcal{X} \to \mathbb{R}$ and $x \in \mathcal{X}$, we set $\mathcal{T}(f)(x) = \max_{a\in\mathcal{A}_x}[r(x,a) + \gamma Pf(x,a)]$, where $Pf(x,a) = \sum_{y\in\mathcal{X}} P(x,a,y)f(y)$. For $f, g : \mathcal{X} \to \mathbb{R}$, it holds that

$$||\mathcal{T}(f) - \mathcal{T}(g)||_\infty = \gamma \max_{x\in\mathcal{X}, a\in\mathcal{A}_x} |Pf(x,a) - Pg(x,a)| \leq \gamma||f-g||_\infty.$$

Since $\gamma \in [0,1)$, we conclude that there exists a unique function $\tilde{V} : \mathcal{X} \to \mathbb{R}$ such that $\mathcal{T}(\tilde{V}) = \tilde{V}$.

Step 2. For $N \geq 1$, set $G_N = \sum_{t=0}^{N-1} \gamma^t R_{t+1}$. For any policy Π, any $x \in \mathcal{X}$, set $V_\Pi^N(x) = \mathbb{E}_{x,\Pi}[G_N]$. Let $\mathbf{0} : \mathcal{X} \to \mathbb{R}$ be the null function. For all $x \in \mathcal{X}$, it holds that

$$V_\Pi^N(x) \leq \mathcal{T}^{\circ N}(\mathbf{0})(x).$$

It suffices to show that for all $k \in [\![1, N]\!]$,

$$\mathbb{E}_{x,\Pi}\Big[\sum_{t=N-k}^{N-1} \gamma^t R_{t+1}\Big| H_{N-k}, A_{N-k}\Big] \leq \gamma^{N-k}\mathcal{T}^{\circ k}(\mathbf{0})(X_{N-k}),$$

recall that H_t was defined in Subsection 2.1. With $k = N$, this gives the result. First, when $k = 1$,

$$\mathbb{E}_{x,\Pi}\Big[\gamma^{N-1}R_N\Big|H_{N-1}, A_{N-1}\Big] = \gamma^{N-1}r(X_{N-1}, A_{N-1}) \leq \gamma^{N-1}\mathcal{T}(\mathbf{0})(X_{N-1}).$$

Next, assuming that the inequality holds true for some $k \in [\![1, N-1]\!]$,

$$\begin{aligned}
&\mathbb{E}_{x,\Pi}\Big[\sum_{t=N-k-1}^{N-1} \gamma^t R_{t+1}\Big| H_{N-k-1}, A_{N-k-1}\Big] \\
\leq&\mathbb{E}_{x,\Pi}[\gamma^{N-k-1} R_{N-k} | H_{N-k-1}, A_{N-k-1}] + \mathbb{E}_{x,\Pi}[\gamma^{N-k}\mathcal{T}^{\circ k}(\mathbf{0})(X_{N-k})|H_{N-k-1}, A_{N-k-1}] \\
=&\gamma^{N-k-1}\Big(r(X_{N-k-1}, A_{N-k-1}) + \gamma P\mathcal{T}^{\circ k}(\mathbf{0})(X_{N-k-1}, A_{N-k-1})\Big) \\
\leq&\gamma^{N-k-1}\mathcal{T}^{\circ(k+1)}(\mathbf{0})(X_{N-k-1}).
\end{aligned}$$

Step 3. We show here that for all policy Π, all $x \in \mathcal{X}$, we have $V_\Pi(x) \leq \tilde{V}(x)$. This of course implies that $V^*(x) \leq \tilde{V}(x)$ for all $x \in \mathcal{X}$.

We have $V_\Pi(x) = \lim_{N\to\infty} V_\Pi^N(x)$, because, recalling Setting 1,

$$|V_\Pi(x) - V_\Pi^N(x)| \leq \sum_{t\geq N} \gamma^t |\mathbb{E}_{x,\Pi}[R_N]| \leq ||g||_\infty \sum_{t\geq N} \gamma^t \to 0.$$

Hence by Step 2, $V_\Pi(x) \leq \lim_{N\to\infty} \mathcal{T}^{\circ N}(\mathbf{0})(x)$. This last quantity equals $\tilde{V}(x)$ by Step 1.

Step 4. For any SM policy π, for $f : \mathcal{X} \to \mathbb{R}$ and for $x \in \mathcal{X}$, we set $T_\pi(f)(x) = r_\pi(x) + \gamma P_\pi f(x)$, recall (5). We check in this step that V_π is the unique fixed point of T_π.

First, T_π is a contraction, since $||T_\pi f - T_\pi g||_\infty = \gamma||P_\pi f - P_\pi g||_\infty \leq \gamma||f - g||_\infty$. Next,

$$V_\pi(x) = \mathbb{E}_{x,\pi}\Big[R_1 + \sum_{t\geq 1}\gamma^t R_{t+1}\Big] = r_\pi(x) + \gamma\mathbb{E}_{x,\pi}\Big[\mathbb{E}_{x,\pi}\Big[\sum_{t\geq 1}\gamma^{t-1}R_{t+1}\Big| H_1\Big]\Big] = r_\pi(x) + \gamma\mathbb{E}_{x,\pi}[V_\pi(X_1)].$$

Since $\mathbb{E}_{x,\pi}[V_\pi(X_1)] = P_\pi V_\pi(x)$, we conclude that $V_\pi(x) = r_\pi(x) + \gamma P_\pi V_\pi(x) = T_\pi(V_\pi)(x)$.

Step 5. Let π^* be a SM policy satisfying

$$\pi^*\Big(x, \arg\max \tilde{Q}(x,\cdot)\Big) = 1 \quad \text{for all } x \in \mathcal{X}, \text{ where} \quad \tilde{Q}(x,a) = r(x,a) + \gamma P\tilde{V}(x,a). \tag{29}$$

Here we prove that then, $\tilde{V} = T_{\pi^*}(\tilde{V})$. By Step 4, we will conclude that $\tilde{V} = V_{\pi^*}$, whence $V_{\pi^*} \geq V^*$ by Step 3. By definition of V^*, this implies that $V_{\pi^*} = V^* = \tilde{V}$.

By definition, see Step 1, we have, for all $x \in \mathcal{X}$,

$$\tilde{V}(x) = \max_{a\in\mathcal{A}_x}[r(x,a) + \gamma P\tilde{V}(x,a)] = \sum_{a\in\mathcal{A}_x}[r(x,a) + \gamma P\tilde{V}(x,a)]\pi^*(x,a)$$

by (29). Hence

$$\tilde{V}(x) = r_{\pi^*}(x) + \gamma\sum_{a\in\mathcal{A}_x}\sum_{y\in\mathcal{X}} P(x,a,y)\tilde{V}(y)\pi^*(x,a) = r_{\pi^*}(x) + \gamma\sum_{y\in\mathcal{X}}\Big(\sum_{a\in\mathcal{A}_x} P(x,a,y)\pi^*(x,a)\Big)\tilde{V}(y),$$

i.e. $\tilde{V}(x) = r_{\pi^*}(x) + \gamma P_{\pi^*}\tilde{V}(x) = T_{\pi^*}(\tilde{V})(x)$.

Conclusion. By Step 5, we know that $\tilde{V} = V^*$. Hence $\tilde{Q} = Q^*$ (see (4) and (29)), so that π^* satisfies (7) if and only if it satisfies (29). Such a SM policy π^* satisfies $V_{\pi^*}(x) = V^*(x)$ for all $x \in \mathcal{X}$ by Step 5, which shows (i).

Let now π by any SM policy. Recalling (6), we have $Q_\pi = r + \gamma P V_\pi \le r + \gamma P V^* = Q^*$, see (4). We have seen that $V_\pi = T_\pi(V_\pi) = r_\pi + \gamma P_\pi V_\pi$ in Step 4. Moreover, we have

$$\sum_{a\in\mathcal{A}_x} Q_\pi(x,a)\pi(x,a) = \sum_{a\in\mathcal{A}_x} r(x,a)\pi(x,a) + \gamma \sum_{a\in\mathcal{A}_x} PV_\pi(x,a)\pi(x,a) = r_\pi(x) + \gamma P_\pi V_\pi(x) = V_\pi(x).$$

Finally,

$$Q_\pi(x,a) = r(x,a) + \gamma \sum_{y\in\mathcal{X}} P(x,a,y)V_\pi(y) = r(x,a) + \gamma \sum_{y\in\mathcal{X}} P(x,a,y) \sum_{b\in\mathcal{A}_y} Q_\pi(y,b)\pi(y,b),$$

and we have checked (ii).

If π^* satisfies (7) (or equivalently (29)), then $Q_{\pi^*} = r + \gamma P V_{\pi^*} = r + \gamma P V^* = Q^*$ because $V_{\pi^*} = V^*$. Moreover, recalling Steps 1 and 5,

$$V^*(x) = \tilde{V}(x) = \max_{a\in\mathcal{A}_x}[r(x,a) + \gamma P\tilde{V}(x,a)] = \max_{a\in\mathcal{A}_x}[r(x,a) + \gamma PV^*(x,a)] = \max_{a\in\mathcal{A}_x} Q^*(x,a),$$

and, by point (ii) applied to π^*,

$$Q^*(x,a) = r(x,a) + \gamma \sum_{(y,b)\in\mathcal{Z}} P(x,a,y)\pi^*(y,b)Q^*(y,b) = r(x,a) + \gamma \sum_{y\in\mathcal{X}} P(x,a,y) \max_{b\in\mathcal{A}_y} Q^*(y,b)$$

by (7). This proves (iii). □

Finally, we recall the

Proof of Lemma 12. A simple computation shows that for all $k \ge 0$,

$$\mathbb{E}[Z_{k+1}^2|\mathcal{G}_k] \le (1-\theta_k)^2 Z_k^2 + C\theta_k^2 = (1+\theta_k^2)Z_k^2 + C\theta_k^2 - 2\theta_k Z_k^2. \tag{30}$$

We set $\gamma_0 = 1$ and $M_0 = Z_0$, $N_0 = Z_0$. For $k \ge 1$, we set $\gamma_k = [\prod_{\ell=0}^{k-1}(1+\theta_\ell^2)]^{-1}$ and introduce

$$M_k = \gamma_k Z_k^2 - C\sum_{\ell=0}^{k-1}\gamma_{\ell+1}\theta_\ell^2 \qquad \text{and} \qquad N_k = M_k + 2\sum_{\ell=0}^{k-1}\gamma_{\ell+1}\theta_\ell Z_\ell^2.$$

One easily checks, using (30) and that γ_{k+1} is $\mathcal{G}_k$-measurable, that $(M_k)_{k\ge0}$ and $(N_k)_{k\ge0}$ are two supermartingales. Let now $L := \sum_{k=0}^\infty \gamma_{k+1}\theta_k^2 \le \sum_{k\ge0}\theta_k^2$, which is a.s. finite by assumption. Since $N_k \ge M_k \ge -CL$ for all $k \ge 0$, both $M_\infty = \lim_k M_k \in \mathbb{R}$ and $N_\infty = \lim_k N_k \in \mathbb{R}$ a.s. exist. Using again that $\sum_{k\ge0}\theta_k^2 < \infty$, we deduce that $\gamma_\infty = \lim_k \gamma_k > 0$ a.s. We first conclude that

$$\ell := \lim_k Z_k^2 = \frac{M_\infty + CL}{\gamma_\infty} \in \mathbb{R}_+$$

a.s. exists, and next that

$$2\sum_{k\ge0}\gamma_{k+1}\theta_k Z_k^2 = N_\infty - M_\infty \in \mathbb{R}_+ \quad \text{a.s.}$$

This tells us that necessarily $\ell = 0$ a.s. because $\gamma_\infty = \lim_k \gamma_k > 0$ a.s. and $\sum_{k\ge0}\theta_k = \infty$ a.s. by assumption. □

References

[1] D.P. Bertsekas, J.N. Tsitsiklis. Neuro-Dynamic Programming. Athena Scientific, Belmont, Massachusetts, 1996.

[2] J. Liu. *On the convergence of reinforcement learning with Monte Carlo Exploring Starts.* Automatica 129 (2021), 109693.

[3] M.L. Puterman. Markov decision processes: discrete stochastic dynamic programming. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1994.

[4] H. Robbins, S. Monro. *A stochastic approximation method.* Ann. Math. Statistics 22 (1951), 400–407.

[5] S. Singh, T. Jaakkola, M.L. Littman, C. Szepesvári. *Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms.* Machine Learning 39 (2000), 287–308.

[6] S.P. Singh, R.S. Sutton. *Reinforcement learning with replacing eligibility traces.* Machine Learning, 22(1-3):123–158.

[7] R.S. Sutton, A.G. Barto. Reinforcement learning: an introduction. Second edition. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2018.

[8] J.N. Tsitsiklis. *On the convergence of optimistic policy iteration.* J. Mach. Learn. Res. 3 (2003), no. 1, 59–72.

[9] C. Wang, S. Yuan, K. Shao, K. Ross. *On the Convergence of the Monte Carlo Exploring Starts Algorithm for Reinforcement Learning.* ICLR 2022 Conference.

[10] A. Winnicki, R. Srikant. *On The Convergence Of Policy Iteration-Based Reinforcement Learning With Monte Carlo Policy Evaluation.* Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023.

Sylvain Delattre, Université Paris Cité and Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75013 Paris, France.

Email address: sylvain.delattre@lpsm.paris

Nicolas Fournier, Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France.

Email address: nicolas.fournier@sorbonne-universite.fr