An introduction to reinforcement learning and stochastic optimization

Francisco J. Silva

francisco.silva@unilim.fr

Master 2 ACSYON

September 2023

Markov decision processes

◇ [Basic definitions about Markov chains] Let $\mathcal{X}$ be a finite set. Recall that a stochastic process $(X_t)_{t\in\mathbb{N}}$ is a Markov chain with state space $\mathcal{X}$ if

$$\begin{aligned}&(\forall\, t \in \mathbb{N})(\forall\, x_0, \ldots, x_t, x_{t+1} \in \mathcal{X})\\ &\mathbb{P}(X_{t+1} = x_{t+1} \,|\, X_t = x_t, \ldots, X_0 = x_0) = \mathbb{P}(X_{t+1} = x_{t+1} \,|\, X_t = x_t).\end{aligned} \tag{1}$$

A Markov chain is completely determined by an initial distribution

$$\mu_0 \in \mathcal{P}(\mathcal{X}) := \{p \in [0,1]^{|X|} \mid \sum_{x\in\mathcal{X}} p_x = 1\}$$

and a family of transition probabilities

$$\{p_t(y|x) \,|\, t \in \mathbb{N},\ x,\ y \in \mathcal{X}\}$$

satisfying that

$$(\forall\, t \in \mathbb{N})(\forall\, x, y \in \mathcal{X}) \quad p_t(y|x) \geq 0 \quad \text{and, for all } t \in \mathbb{N},\ x \in \mathbb{N},\ \sum_{y \in \mathcal{X}} p_t(y|x) = 1,$$

i.e. $p_t(\cdot|x) \in \mathcal{P}(\mathcal{X})$. In this framework, the equality (1) can be written as

$$(\forall\, t \in \mathbb{N})(\forall\, x_0, \ldots, x_t, x_{t+1} \in \mathcal{X})$$
$$\mathbb{P}(X_{t+1} = x_{t+1} \,|\, X_t = x_t, \ldots, X_0 = x_0) = p_t(x_{t+1}|x_t).$$

Remark 1. If p_t does not depend on t, then the Markov chain is said to be homogeneous.

◇ [Markov decision processes] A Markov Decision Process (MDP) is defined by

1. A state space $\mathcal{X}$ (which is assumed to be finite).
2. An action (or control) space A (which is assumed to be finite).
3. A family of controlled transition probabilities

$$\{p_t(y|x,a) \,|\, t \in \mathbb{N},\ x,\ y \in \mathcal{X},\ a \in A\}$$

satisfying that, for every $(t, x, a) \in \mathbb{N} \times \mathcal{X}$, $p_t(\cdot|x,a) \in \mathcal{P}(\mathcal{X})$.

4. A family $\{r_t(x,a) \mid t \in \mathbb{N},\ x \in \mathcal{X},\ a \in A\} \subset \mathbb{R}$.

For every $t \in \mathbb{N}$, x, y and $a \in A$, $p_t(y|x,a)$ is interpreted as the probability to move from x, at time step t, to y, at time step $t+1$, under the decision a, and $r_t(x,a)$ is the reward of being at point x at time t and to have adopted decision a.
An agent will chose an action at each time step that will depend on his current state. Thus, this type of strategy is given in the so-called *feedback form*.

Definition 1. *We say that $(\pi_t)_{t\in\mathbb{N}}$ is a policy if, for every $t \in \mathbb{N}$, $\pi_t \colon \mathcal{X} \to A$. The policy $(\pi_t)_{t\in\mathbb{N}}$ is stationary if it is independent of t, i.e. $\pi_s = \pi_t$ for all s, $t \in \mathbb{N}$. The set of policies is denoted by Π.*

If a policy $(\pi_t)_{t\in\mathbb{N}}$ is stationary, will simply write $\pi(x) = \pi_t(x)$ for all $t \in \mathbb{N}$ and $x \in \mathcal{X}$.
Given $\mu \in \mathcal{P}(\mathcal{X})$ and a policy $\pi = (\pi_t)_{t\in\mathbb{N}}$, we set $X^{\mu,\pi}$ for the Markov chain whose initial law is given by μ and its transition probabilities are given by

$$(\forall\, t \in \mathbb{N})(\forall\, x,\ y \in \mathcal{X}) \quad p_t^{\pi}(y|x) = p_t(y|x, \pi_t(x)).$$

Notice that if p_t does not depend on t and $(\pi_t)_{t\in\mathbb{N}}$ is stationary, then the resulting

Markov chain is homogeneous.

Associated with an initial distribution μ and a policy π, several reward functionals can be considered, which defines the corresponding optimization problem.

1. Finite horizon case. Let $T \in \mathbb{N}^*$. The optimization problem solved by the decision maker is given by

$$\sup_{\pi \in \Pi} \mathbb{E}\left(\sum_{t=0}^{T-1} r_t\big(X_t^{\mu,\pi}, \pi_t(X_t^{\mu,\pi})\big) + \Psi(X_T^{\mu,\pi}) \right),$$

   where $\Psi : \mathcal{X} \to \mathbb{R}$ represents the final reward.

2. Discounted infinite horizon case. Here we assume that $r_t(x,a)$ and $p_t(y|x,a)$ do not depend on t and that Π denotes the set of stationary policies $\Pi = \{\pi : \mathcal{X} \to A\}$. Given $\gamma \in]0,1[$, the optimization problem solved by the decision maker is given by

$$\sup_{\pi \in \Pi} \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t r\big(X_t^{\mu,\pi}, \pi(X_t^{\mu,\pi})\big) \right)$$

   This type of problem is adequate when the time horizon is unknown but suspected to

be large. One also uses this formalism when one does not want that the final horizon and final reward affect the main properties of the solution. The interpretation of the discount factor is that future costs matter to us less than the same cost incurred at the present time.

3. Infinite horizon with terminal state (or stochastic shortest path problem). We assume the existence of an absorbing state $\hat{x} \in \mathcal{X}$, i.e.

$$(\forall\, t \in \mathbb{N})(\forall\, a \in A) \quad p_t(y|\hat{x}, a) = \begin{cases} 1 & \text{if } y = \hat{x}, \\ 0 & \text{otherwise.} \end{cases}$$

Consider the stopping time

$$\tau^{\mu,\pi} = \inf\{t \in \mathbb{N} \,|\, X_t^{\mu,\pi} = \hat{x}\} \in \mathbb{N} \cup \{+\infty\}. \tag{2}$$

The optimization problem solved by the decision maker is given by

$$\sup_{\pi \in \Pi} \mathbb{E}\left(\sum_{t=0}^{\tau^{\mu,\pi}} r_t(X_t^{\mu,\pi}, \pi_t(X_t^{\mu,\pi}) \right)$$

This type of problem is interesting when reaching the termination state $\hat{x}$ is inevitable, i.e. when $\tau^{\mu,\pi} < +\infty$ holds almost surely. When we will work with this problem, we will assume that this property holds and that the final state $\hat{x}$ is cost-free, i.e. $r(\hat{x}, \pi(\hat{x})) = 0$ for every $\pi \in \Pi$.

4. Average cost problems. Here the optimization problem to be solved is given by

$$\inf_{\pi \in \Pi} \limsup_{T \to +\infty} \frac{1}{T} \mathbb{E} \left(\sum_{t=0}^{T-1} r_t(X_t^{\mu,\pi}, \pi_t(X_t^{\mu,\pi}) \right)$$

and one usually considers the case where π is stationary and $r_t(x, a)$ does not depend on t.
This model is adapted to problems where discounting is inappropriate and there is no natural cost-free absorbing state.

◇ [Examples]

- [The inventory problem][1] We have $T > 1$ periods of time where, at the beginning of each period $t \in \{0, \ldots, T-1\}$, a manager of an inventory orders a quantity

[1] In this example the state space $\mathcal{X}$ and the action space A are continuous.

a_t of stock in order to meet a stochastic demand D_t. We assume that $(D_t)_{t=0}^T$ is a sequence of i.i.d. random variables and that the excess demand is backlogged and served as soon as new stock is available.
We denote by

- X_t the stock at the beginning of period t. Notice that X_t can be negative as it takes into account the eventual backlogged demand from the previous time. The cost at period t of stocking X_t is given by $r(X_t)$, where $r\colon \mathbb{R} \to \mathbb{R}$. Notice that it is meaninful to ask that r satisfies $r(0) = 0$ and $r(x) > 0$ for all $x \neq 0$.
- a_t the stock ordered (and inmediately delivered) at the beginning of period t. Each unit of stock ordered at period t has a cost given by $c > 0$.

The problem of minimizing the operational cost in this situation is modelled by

$$\inf_{\pi \in \Pi} \mathbb{E}\Big(\sum_{t=0}^{T-1} \big[r(X_t^{\mu,\pi}) + c\pi_t(X_t) \big] + r(X_T^{\mu,\pi}) \Big),$$

where the policies take values in the action space $A = [0, +\infty[$ and the transition

probabilities of $(X_t^{\mu,\pi})_{t=0}^T$ are given by: for every $B \in \mathcal{B}(\mathbb{R})$,

$$p_t^\pi(B|x) = \mathbb{P}(x + \pi_t(x) - D_t \in B) = \mathbb{P}(D_0 \in x + \pi_t(x) - B)$$

for all $t = 0, \ldots, T-1,\ x,\ y \in \mathbb{R}$.

◇ [Dynamic programming] Let us quote the following explanation from Bellman (1957) on the principle of dynamic programming:

"An optimal policy has the property that, whatever the initial state and the initial decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision".

Given $x \in \mathcal{X}$ and a family of controlled transition probabilities $\{p_s(y|x,a) \mid s \in \mathbb{N},\ x,\ y \in \mathcal{X},\ a \in A\}$ and a time-dependent policy $\pi \in \Pi_t$, we define $\{X_s^{t,x,\pi} \mid s \in \mathbb{N},\ s \geq t\}$ as the Markov chain with values in $\mathcal{X}$ satisfying that

$$X_t^{t,x,\pi} = x \quad \text{and} \quad \mathbb{P}(X_{s+1}^{t,x,\pi} = y | X_s^{t,x,\pi} = x) = p_s(y|x, \pi_s(x)),$$

for all $s \in \mathbb{N}$, $s \geq t$, and $y \in \mathcal{X}$.

◇ [The finite horizon case] Given a policy $\pi \in \Pi$, set

$$V_T^\pi(x) = \Psi(x) \quad \text{for all } x \in \mathcal{X}$$

and, for all $t = 0, \ldots, T-1$, $x \in \mathcal{X}$, define

$$V_t^\pi(x) = \mathbb{E}\left(\sum_{s=t}^{T-1} r_s\left(X_s^{x,t,\pi}, \pi_s(X_s^{x,t,\pi})\right) + \Psi(X_T^{x,t,\pi})\right). \tag{3}$$

Notice that

$$\begin{aligned} V_t^\pi(x) &= r_t\left(x, \pi_t(x)\right) \\ &\quad + \mathbb{E}\left(\sum_{s=t+1}^{T-1} r_s\left(X_s^{x,t,\pi}, \pi_s(X_s^{x,t,\pi})\right) + \Psi(X_T^{x,t,\pi})\right) \\ &= r_t\left(x, \pi_t(x)\right) \\ &\quad + \mathbb{E}\left(\mathbb{E}\left(\textstyle\sum_{s=t+1}^{T-1} r_s\left(X_s^{x,t,\pi}, \pi_s(X_s^{x,t,\pi})\right) + \Psi(X_T^{x,t,\pi}) | X_{t+1}^{x,t,\pi}\right)\right), \end{aligned}$$

from which we deduce the identity

$$V_t^\pi(x) = r_t\left(x, \pi_t(x)\right) + \mathbb{E}\left(V_{t+1}^\pi(X_{t+1}^{x,t,\pi})\right), \tag{4}$$

or, equivalently,

$$V_t^\pi(x) = r_t\left(x, \pi_t(x)\right) + \sum_{y\in\mathcal{X}} p_t(y|x, \pi_t(x)) V_{t+1}^\pi(y). \tag{5}$$

Let us define the value function $V^*\colon \{0, \dots, T-1\} \times \mathcal{X} \to \mathbb{R}$ as $V_T^* = \Psi(x)$ for all $x \in \mathcal{X}$ and

$$V_t^*(x) = \sup_{\pi\in\Pi} V_t^\pi(x) \quad \text{for all } t = 0, \dots, T-1,\ x \in \mathcal{X}. \tag{6}$$

Remark 2. Since $|\mathcal{X}| < \infty$ and $|A| < \infty$, the set of possible policies between t and T, i.e. $\{\pi_s\colon \mathcal{X} \to A \,|\, s = t, \dots, T-1\}$ is also finite and hence there exists at least one policy $\pi^* \in \Pi$ such that

$$V_s^*(x) = V_s^{\pi^*}(x) \quad \text{for all } s = t, \dots, T-1. \tag{7}$$

Note that such policy π^* satisfies

$$\pi_s^*(x) \in \text{argmax}_{a \in A} V_s^{(\pi^*)^{s,a}}(x) \quad \text{for all } s = t, \ldots, T-1, \tag{8}$$

where, given a policy $\pi \in \Pi$, the policy $\pi^{t,a} = (a, a, \ldots, a, \pi_{t+1}, \pi_{t+2}, \ldots)$ satisfies that, for every $s \in \mathbb{N}$, $s \leq t$, π_s^a is the constant a policy and, for every $s \in \mathbb{N}$, $s > t$, we have $\pi_s^a = \pi_s$.

Theorem 1. *For every $t = 0, \ldots, T-1$ and $x \in \mathcal{X}$, we have*

$$V_t^*(x) = \sup_{a \in A} \left\{ r_t(x, a) + \sum_{y \in \mathcal{X}} p_t(y|x, a) V_{t+1}^*(y) \right\}. \tag{9}$$

Proof. Let us fix $t \in \{0, \dots, T-1\}$ and $x \in \mathcal{X}$. For every $\pi \in \Pi$ we have

$$\begin{aligned} V_t^{\pi}(x) &= r_t(x, \pi_t(x)) + \sum_{y \in \mathcal{X}} p_t(y|x, \pi_t(x)) V_{t+1}^{\pi}(y) \\ &\leq r_t(x, \pi_t(x)) + \sum_{y \in \mathcal{X}} p_t(y|x, \pi_t(x)) V_{t+1}^{*}(y) \\ &\leq \sup_{a \in A} \left\{ r_t(x, a) + \sum_{y \in \mathcal{X}} p_t(y|x, a) V_{t+1}^{*}(y) \right\}. \end{aligned} \tag{10}$$

Conversely, let $\pi^* \in \Pi$ such that (7) holds. For every $a \in A$, we have that

$$r_t(x, a) + \sum_{y \in \mathcal{X}} p(y|x, a) V_{t+1}^{\pi^*}(y) = V_t^{(\pi^*)^{t,a}}(x) \leq V_t^{*}(x)$$

and, since $V_{t+1}^{\pi^*} = V_{t+1}^{*}$ and $a \in A$ is arbitrary, we deduce that

$$\sup_{a \in A} \left\{ r_t(x, a) + \sum_{y \in \mathcal{X}} p_t(y|x, a) V_{t+1}^{*}(y) \right\} \leq V_t^{*}(x) \tag{11}$$

and (9) follows from (10) and (11). □

Corollary 1. *Denote by $\pi^* \in \Pi$ a policy satisfying*

$$V_t^*(x) = r_t(x, \pi_t^*(x)) + \sum_{y \in \mathcal{X}} p_t(y|x, \pi_t^*(x)) V_{t+1}^*(y), \tag{12}$$

for all $t = 0, \ldots, N-1$ and $x \in \mathcal{X}$. Then π^* is optimal, i.e. $V_t^{\pi^*}(x) = V_t^*(x)$ for all $t = 0, \ldots, T$.

Proof. By definition, we have $V_T^* = V_T^{\pi^*}$, which, by (12), implies that $V_{T-1}^* = V_{T-1}^{\pi^*}$. Arguing in this manner, one obtains the result. □

◇ [The infinite horizon discounted case] We suppose that we deal with autonomous data, i.e. we assume that the rewards $r_t(x, a)$ and the transition probabilities $p_t(y|x, a)$ do not depend on t. We also assume that the rewards $\{r_t \mid t \in \mathbb{N}\}$ are bounded. Fix $\gamma \in]0, 1[$, called discounting factor, and, given $\pi \in \Pi$, define

$$V^\pi(x) = \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t r\left(X_t^{x,0,\pi}, \pi_t(X_t^{x,0,\pi})\right)\right) \quad \text{for all } x \in \mathcal{X}. \tag{13}$$

Let us define the value function

$$V^*(x) = \sup_{\pi \in \Pi} V^\pi(x) \quad \text{for all } x \in \mathcal{X}. \tag{14}$$

Since $\gamma \in]0,1[$, our boundedness assumption on the rewards implies that, for all $x \in \mathcal{X}$, $V^\pi(x) \in \mathbb{R}$, for all $\pi \in \Pi$, and $V^*(x) \in \mathbb{R}$.

Theorem 2. *The value function V^* satisfies*

$$V^*(x) = \sup_{a \in A} \left\{ r(x,a) + \gamma \sum_{y \in \mathcal{X}} p(y|x,a) V^*(y) \right\} \quad \textit{for all } x \in \mathcal{X}. \tag{15}$$

Proof. Fix a policy $\pi \in \Pi$. We have

$$\begin{aligned} V^{\pi}(x) &= r\left(x, \pi_0(x)\right) + \mathbb{E}\left(\sum_{t=1}^{\infty} \gamma^t r\left(X_t^{x,0,\pi}, \pi_t(X_t^{x,0,\pi})\right)\right) \\ &= r\left(x, \pi_0(x)\right) + \gamma\mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t r\left(X_{t+1}^{x,0,\pi}, \pi_{t+1}(X_{t+1}^{x,0,\pi})\right)\right) \\ &= r\left(x, \pi_0(x)\right) + \gamma\mathbb{E}\left(\mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t r\left(X_{t+1}^{x,0,\pi}, \pi_{t+1}(X_{t+1}^{x,0,\pi})\right) | X_1^{x,0,\pi}\right)\right) \\ &\leq r\left(x, \pi_0(x)\right) + \gamma \sum_{y \in \mathcal{X}} p(y|x, \pi_0(x)) V^*(y), \end{aligned} \tag{16}$$

from which we get

$$V^*(x) \leq \sup_{a \in A} \left\{ r(x,a) + \sum_{y \in \mathcal{X}} p(y|x,a) V^*(y) \right\} \quad \text{for all } x \in \mathcal{X}. \tag{17}$$

Conversely, let $\varepsilon > 0$ and for all $x \in \mathcal{X}$ and $t \in \mathbb{N}$, $t \geq 1$ let $\pi^{t,x} \in \Pi$ be such that

$$V^*(x) \leq V^{\pi^{t,x}} + \frac{\varepsilon}{(2\gamma)^t}. \tag{18}$$

We define the policy $\pi^{\varepsilon} \in \Pi$ as follows: let $\pi_0^{\varepsilon}(x)$ be such that

$$\sup_{a\in A}\left\{ r(x,a) + \gamma \sum_{y\in\mathcal{X}} p(y|x,a)V^*(y) \right\} = r(x,\pi_0^{\varepsilon}(x)) + \gamma \sum_{y\in\mathcal{X}} p(y|x,\pi_0^{\varepsilon}(x))V^*(y) \tag{19}$$

and, for $t \in \mathbb{N}$, $t \geq 1$, and $y \in \mathcal{X}$, define $\pi_t^{\varepsilon}(y) = \pi_0^{t,y}$. Thus, by (18) with $t = 1$ and (19), we have

$$\sup_{a\in A}\left\{ r(x,a) + \gamma \sum_{y\in\mathcal{X}} p(y|x,a)V^*(y) \right\} \leq r(x,\pi_0^{\varepsilon}(x)) + \gamma \sum_{y\in\mathcal{X}} p(y|x,\pi_0^{\varepsilon}(x))V^{\pi^{1,y}}(y) + \frac{\varepsilon}{2}.$$

Applying consecutively (16) and (18) for $t \in \mathbb{N}$, $t \geq 2$, we obtain

$$\sup_{a\in A}\left\{ r(x,a) + \sum_{y\in\mathcal{X}} p(y|x,a)V^*(y) \right\} \leq V^{\pi^{\varepsilon}} + \varepsilon \leq V^* + \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we deduce that

$$\sup_{a\in A}\left\{ r(x,a) + \sum_{y\in\mathcal{X}} p(y|x,a)V^*(y) \right\} \leq V^* \tag{20}$$

and the result follows from (17) and (20).

□

Denote by E the space of functions defined on $\mathcal{X}$ with values in $\mathbb{R}$ endowed with the maximum norm

$$\|f_1 - f_2\|_\infty = \max\{|f_1(x) - f_2(x)| \mid x \in \mathcal{X}\} \quad \text{for all } f_1,\ f_2 \in E.$$

As the computation in (16) shows, for every stationary policy $\pi \in \Pi_{\text{stat}}$, we have

$$V^\pi(x) = r(x, \pi(x)) + \gamma \sum_{y \in \mathcal{X}} p(y|x, \pi(x)) V^\pi(y).$$

Thus, $V^\pi \in E$ is a fixed point of the operator $T^\pi \colon E \to E$ defined by

$$\left(T^\pi f\right)(x) = r(x, \pi(x)) + \gamma \sum_{y \in \mathcal{X}} p(y|x, \pi(x)) f(y) \quad \text{for all } f \in E,\ x \in \mathcal{X}. \tag{21}$$

Similarly, V^* is a fixed point of the operator $T^*\colon E \to E$ defined by

$$\left(T^* f\right)(x) = \max_{a \in A} \left\{ r(x,a) + \gamma \sum_{y \in \mathcal{X}} p(y|x,a) f(y) \right\} \quad \text{for all } f \in E,\ x \in \mathcal{X}. \tag{22}$$

Recall the following fundamental result[2].

Theorem 3 (Banach fixed-point theorem (1922) [1]). *Let (X,d) be a complete metric space, let $M \subset X$ be closed, let $T\colon M \to M$, and let $\gamma \in [0,1[$. Assume that T is γ-contractive, i.e.*

$$d(T(x),T(y)) \leq \gamma d(x,y) \quad \textit{for all } x,\ y \in M.$$

Then the following hold:

(i) *T admits a unique fixed-point x^*.*

(ii) *The sequence of Picard (or successive) iterations $x_0 \in M$ and, for every $k \geq 0$, $x_{k+1} = T(x_k)$ converges to x^*.*

[2] For a proof see [6, Theorem 1.A].

(iii) *The following error estimates hold:*

$$d(x_k, x^*) \leq \frac{\gamma^k}{1-\gamma} d(x_1, x_0) \quad \textit{for all } k \in \mathbb{N}.$$

(iv) *The sequence $(x_k)_{k\in\mathbb{N}}$ linearly converges towards x^* with rate of convergence equal to γ. More precisely, we have*

$$d(x_{k+1}, x^*) \leq \gamma d(x_k, x^*) \quad \textit{for all } k \in \mathbb{N}.$$

The next result is a consequence of Theorem 3.

Theorem 4. *The following assertions hold:*

(i) *For every $\pi \in \Pi_{\text{stat}}$ the map T^π, defined in (21), is a contraction. Its unique fixed point $V^\pi \in E$ can be approximated as follows:*

$$V^0 \in E \textit{ arbitrary } \ (\forall\, k \in \mathbb{N}) \quad V^{k+1} = T^\pi V^k. \tag{23}$$

Moreover, $(V^k)_{\in\mathbb{N}}$ linearly converges towards V^π and the following estimate

holds:

$$\|V^k - V^\pi\|_\infty \leq \frac{\gamma^k}{1-\gamma}\|V^1 - V_0\|_\infty.$$

(ii) *The map T^*, defined in* (22)*, is a contraction. Its unique fixed point $V^* \in E$ can be approximated as follows:*

$$V^0 \in E \text{ arbitrary } \ (\forall\, k \in \mathbb{N}) \quad V^{k+1} = T^* V^k. \tag{24}$$

Moreover, the following estimate holds:

$$\|V^k - V^*\|_\infty \leq \frac{\gamma^k}{1-\gamma}\|V^1 - V_0\|_\infty.$$

Proof. In view of Banach fixed point theorem, it suffices to check that T^π and T^* are contractions. Indeed, for a fixed $\pi \in \Pi_{\text{stat}}$, and $f_1,\ f_2 \in E$, the triangular inequality implies

$$\|T^\pi f_1 - T^\pi f_2\|_\infty = \gamma \sup_{x\in\mathcal{X}} \left| \sum_{y\in\mathcal{X}} p(y|x,\pi(x))\,(f_1(x) - f_2(x)) \right| \leq \gamma \|f_1 - f_2\|_\infty.$$

Similarly[3], we have

$$\|T^* f_1 - T^* f_2\|_\infty \le \sup_{x\in\mathcal{X}} \sup_{a\in A} \left| \sum_{y\in\mathcal{X}} p(y|x,a)\,(f_1(x) - f_2(x)) \right| \le \gamma \|f_1 - f_2\|_\infty.$$

□

Remark 3. (i) The approximation (24) is called the value iteration method to approximate V^*.

(ii) V^* can also be computed by solving a linear programming problem. Consider the problem

$$\begin{aligned} &\min \sum_{x\in\mathcal{X}} V(x) \\ &\text{s.t.} \quad V(x) \ge r(x,a) + \sum_{y\in X} p(y|x,a) V(y) \quad \text{for all } x \in \mathcal{X},\ a \in A. \end{aligned} \tag{25}$$

[3]Recall that $|\sup_{a\in A} p(a) - \sup_{a\in A} q(a)| \le \sup_{a\in A} |p(a) - q(a)|$ for any two functions $p,\ q : A \to \mathbb{R}$.

Let us show that V^* is the unique solution to (25). Indeed, (2) implies that V^* is feasible for (25) and, by definition of the feasible set, for any other feasible V, we have $V \geq T^*V$ and hence, by Lemma 1(ii) below, we have $V \geq (T^*)^k V$ for any $k \in \mathbb{N}$. Letting $k \to \infty$ and using Theorem 4(ii), with $V^0 = V$, we obtain that $V \geq V^*$, which implies that V^* solves (25). The uniqueness of the solution V^* follows from the fact that if $\hat{V}$ solves (25), then $\sum_{x \in \mathcal{X}} \left(\hat{V}(x) - V^*(x) \right) = 0$. Since the feasibility of V implies that $\hat{V} \geq V^*$, we deduce that $\hat{V}(x) = V^*(x)$ for all $x \in \mathcal{X}$.

Corollary 2. *The following assertions hold:*

(i) *Let* $\pi^* \in \Pi_{\text{stat}}$ *and suppose that*

$$V^*(x) = r(x, \pi^*(x)) + \gamma \sum_{y \in Y} p(y|x, \pi^*(x)) V^*(y) \quad \textit{for all } x \in \mathcal{X}. \quad (26)$$

Then π^* *is optimal, i.e.* $V^{\pi^*} = V^*$.

(ii) *There exists at least one optimal stationary policy.*

Proof. (i) If (26) holds, then V^{π^*} and V^* are both fixed points of T^{π^*} and, since T^{π^*} and T^* have unique fixed points, one has $V^{\pi^*} = V^*$.

(ii) Since A is finite, a policy $\pi^* \in \Pi_{\text{stat}}$ satisfying (26) always exists. □

Remark 4. As a consequence of Corollary 2, we have

$$V^*(x) = \sup_{\pi \in \Pi} V^\pi(x) = \sup_{\pi \in \Pi_{\text{stat}}} V^\pi(x).$$

The following useful properties of operators T^π ($\pi \in \Pi_{\text{stat}}$) and T^* will be useful.

Lemma 1. *The following assertions hold:*

(i) *Let $\pi \in \Pi_{\text{stat}}$. Then T^π is monotone:*

$$T^\pi f_1 \le T^\pi f_2 \quad \textit{for all} \quad f_1,\ f_2 \in E,\ f_1 \le f_2.$$

(ii) *T^* is monotone:*

$$T^* f_1 \le T^* f_2 \quad \textit{for all} \quad f_1,\ f_2 \in E,\ f_1 \le f_2.$$

Proof. Both assertions follow directly from the definitions taking into account that $p(y|x,a) \geq 0$ for all x, $y \in \mathcal{X}$, and $a \in A$. □

Consider the following method, called policy iterations method:

- $\pi^0 \in \Pi_{\text{stat}}$ arbitrary,
- $\pi^{k+1}(x) \in \operatorname{argmax}_{a \in A} \left\{ r(x,a) + \gamma \sum_{y \in \mathcal{X}} p(y|x,a) V^{\pi^k}(y) \right\}$, (27)
  for all $k \in \mathbb{N}$, $x \in \mathcal{X}$.

Theorem 5 (Convergence of the policy iterations method). *Let $(\pi^k)_{k \in \mathbb{N}} \subset \Pi_{\text{stat}}$ be the sequence obtained by* (27)*. Then the following hold:*

(i) $V^{\pi^k} \leq V^{\pi^{k+1}}$ *for all* $k \in \mathbb{N}$.

(ii) *There exists* $k^* \in \mathbb{N}$ *such that* $V^{\pi^k} = V^{\pi^{k^*}} = V^*$ *for all* $k \in \mathbb{N}$, $k \geq k^*$.

Proof. (i) Notice that, for all $k \in \mathbb{N}$ and $x \in \mathcal{X}$,

$$\begin{aligned} V^{\pi^k}(x) &= r(x, \pi^k(x)) + \gamma \sum_{y \in \mathcal{X}} p(y|x, \pi^k(x)) V^{\pi^k}(y) \\ &\leq r(x, \pi^{k+1}(x)) + \gamma \sum_{y \in \mathcal{X}} p(y|x, \pi^{k+1}(x)) V^{\pi^k}(y) \\ &= T^{\pi^{k+1}} V^{\pi^k}(y). \end{aligned}$$

Thus, Lemma 1 implies that

$$V^{\pi^k}(x) \leq \underbrace{\left(T^{\pi^{k+1}} \circ \ldots \circ T^{\pi^{k+1}} \right)}_{n \text{ times}} V^{\pi^k}(y) \quad \text{for all } n \in \mathbb{N}.$$

Letting $n \to \infty$ and using Theorem 4(i), we deduce that $V^{\pi^k} \leq V^{\pi^{k+1}}$.

(ii) Suppose that there exists $k \in \mathbb{N}$ such that $V^{\pi^k} = V^{\pi^{k+1}}$. Then (27) implies that for all

$x \in \mathcal{X}$, we have that

$$\begin{aligned} V^{\pi^{k+1}}(x) &= r(x, \pi^{k+1}(x)) + \gamma \sum_{y \in \mathcal{X}} p(y|x, \pi^{k+1}(x)) V^{\pi^{k+1}}(x) \\ &= r(x, \pi^{k+1}(x)) + \gamma \sum_{y \in \mathcal{X}} p(y|x, \pi^{k+1}(x)) V^{\pi^{k}}(x) \\ &= \sup_{a \in A} \left\{ r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y|x, a) V^{\pi^{k}}(x) \right\}. \end{aligned}$$

Therefore, $V^{\pi^k} = V^{\pi^{k+1}}$ solves $V^{\pi^k} = T^* V^{\pi^k}$ and hence, by the uniqueness of the fixed point of T^*, we deduce that $V^{\pi^k} = V^*$. The result follows from the fact that $|\Pi_{\text{stat}}| < \infty$. □

In the implementation of (27), given π^k, for $k \in \mathbb{N}$, the computation of V^{π^k} is needed. Typically, this can be done by the following two methods:

- enumerating the elements in $\mathcal{X} = \{x_1, \ldots, x_N\}$, defining the matrix $P^k \in \mathbb{R}^{N \times N}$ as $(P^k)_{i,j} = p(x_j|x_i, \pi^k(x_i))$, for all $i, j \in \{1, \ldots, N\}$, setting $r^k = (r(x_1, \pi^k(x_1)), \ldots, r(x_N, \pi^k(x_N))) \in \mathbb{R}^N$, and solving the linear

equation

$$\left(I_N - \gamma P^k\right) V^{\pi^k} = r^k, \tag{28}$$

to find $V^{\pi^k} = \left(V^{\pi^k}(x_1), \ldots, V^{\pi^k}(x_N)\right) \in \mathbb{R}^N$. In (28), I_N denotes the $N \times N$ identity matrix and, as the following exercise shows, $\left(I_N - \gamma P^k\right)$ is invertible.

Exercise 1.

(i) Show that $\rho(P^k) := \max\{|\lambda| \mid \lambda \text{ is an eigenvalue of } P^k\} = 1$.

(ii) Conclude that $\left(I_N - \gamma P^k\right)$ is invertible.

- Using iterates (23). This method is efficient if $\gamma \ll 1$.

◇ [The undiscounted infinite horizon case with terminal state] As in the previous case, we suppose that data is autonomous, i.e. $r_t(x, a)$ and $p_t(y|x, a)$ are independent of $t \in \mathbb{N}$. We will also work only with stationary policies. Given $x \in \mathcal{X}$ and

$\pi \in \Pi_{\text{stat}}$, in what follows we set $X_t^{x,\pi} = X_t^{0,x,\pi}$, for all $t \in \mathbb{N}$, and

$$V^{\pi}(x) = \mathbb{E}\left(\sum_{t=0}^{\tau^{x,\pi}} r\left(X_t^{x,\pi}, \pi\left(X_t^{x,\pi}\right)\right)\right),$$

where we recall that given a terminal state $\hat{x} \in \mathcal{X}$, the stopping time $\tau^{x,\pi}$ is defined in (2) and we assume that $r(\hat{x}, a) = 0$ for all $a \in A$. In particular, we have

$$V^{\pi}(x) = \mathbb{E}\left(\sum_{t=0}^{\infty} r\left(X_t^{x,\pi}, \pi\left(X_t^{x,\pi}\right)\right)\right),$$

Definition 2. *We say that $\pi \in \Pi_{\text{stat}}$ is proper if there exists $\hat{t} \in \mathbb{N}$ such that*

$$\min_{x \in \mathcal{X}} \mathbb{P}\left(X_{\hat{t}}^{x,\pi} = \hat{x}\right) > 0.$$

Remark 5. Since $|\mathcal{X}| < \infty$, a policy $\pi \in \Pi$ is proper iff for every $x \in \mathcal{X}$ implementing the policy π there exists a path, with positive probability, connecting

x and $\hat{x}$ in $\hat{t}$ steps. A sufficient condition for the properness of π is that

$$p(\hat{x}|x, \pi(x)) > 0 \quad \text{for all } x \in \mathcal{X}$$

and, in this case, one can take $\hat{t} = 1$.

Proposition 1. *If $\pi \in \Pi_{\text{stat}}$ is proper then, for every $x \in \mathcal{X}$, with probability one the Markov chain $(X_t^{x,\pi})_{t\in\mathbb{N}}$ reaches the terminal state $\hat{x}$ in finite time. Moreover, $\|V^\pi\|_\infty < \infty$.*

Proof. Let $\pi \in \Pi_{\text{stat}}$ be proper, set $\delta = 1 - \min_{x\in\mathcal{X}} \mathbb{P}\left(X_{\hat{t}}^{x,\pi} = \hat{x}\right) < 1$, and let $x \in \mathcal{X}$. By assumption, $\mathbb{P}(X_{\hat{t}}^{y,\pi} \neq \hat{x}) \leq \delta$ for every $y \in \mathcal{X}$, and, using that $\hat{x}$ is absorbing and the Markov

property, for every $k \in \mathbb{N}^*$, $k \geq 2$, we have

$$\begin{aligned}
\mathbb{P}(X^{x,\pi}_{k\hat{t}} \neq \hat{x}) &= \sum_{y_{k-1} \in \mathcal{X}, y_{k-1} \neq \hat{x}} \mathbb{P}(X^{x,\pi}_{k\hat{t}} \neq \hat{x} | X^{x,\pi}_{(k-1)\hat{t}} = y_{k-1}) \mathbb{P}(X^{x,\pi}_{(k-1)\hat{t}} = y_{k-1}) \\
&\leq \delta \sum_{y_{k-1} \in \mathcal{X}, y_{k-1} \neq \hat{x}} \mathbb{P}(X^{x,\pi}_{(k-1)\hat{t}} = y_{k-1}) \\
&= \delta \mathbb{P}(X^{x,\pi}_{(k-1)\hat{t}} \neq \hat{x}) \\
&\leq \delta^k.
\end{aligned}$$

Thus, by the Borel-Cantelli Lemma[4], the event $\{X^{x,\pi}_{k\hat{t}} \neq \hat{x}\}$ occurs finitely often with probability

[4]Consider a sequence $(A_n)_{n\in\mathbb{N}} \subseteq \mathcal{F}$ such that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$. Then $\mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\right) = 0$.

1. In addition, recalling that $r(\widehat{x}, a) = 0$ for all $a \in A$, we have

$$\begin{aligned}
|V^{\pi}(x)| &\leq \left| \mathbb{E}\left(\sum_{t=0}^{\infty} r\left(X_t^{x,\pi}, \pi\left(X_t^{x,\pi}\right)\right) \right) \right| \\
&\leq \sum_{k=0}^{\infty} \mathbb{E}\left(\mathbb{I}_{\left\{X_{k\hat{t}}^{x,\pi} \neq \hat{t}\right\}} \sum_{t=k\hat{t}}^{(k+1)\hat{t}-1} \left| r\left(X_t^{x,\pi}, \pi\left(X_t^{x,\pi}\right)\right)\right| \right) \\
&\leq \|r\|_{\infty} \hat{t} \sum_{k=0}^{\infty} \mathbb{P}(X_{k\hat{t}}^{x,\pi} \neq \hat{t}), \\
&\leq \|r\|_{\infty} \hat{t} \sum_{k=0}^{\infty} \delta^k \\
&= \frac{\hat{t}\|r\|_{\infty}}{1-\delta} < +\infty.
\end{aligned}$$

□

Set $V^*(x) = \sup_{\pi \in \Pi_{\text{stat}}} V^{\pi}(x)$ for all $x \in \mathcal{X}$. As in the infinite horizon discounted case, one can show a dynamic programming principle in infinite horizon undiscounted case. We will admit the following results and we refer the reader to [3, Chapter 7] for the proofs.

Theorem 6. *Assume that every policy $\pi \in \Pi_{\text{stat}}$ is proper. Then the following hold:*

(i) [DPP] *For every $x \in \mathcal{X}$ one has*

$$V^*(x) = \sup_{a \in A} \left\{ r(x,a) + \sum_{y \in \mathcal{X}} p(y|x,a) V^*(y) \right\}.$$

(ii) *There exists at least one optimal stationary policy.*

(iii) [Value iterations] *Let $V^0 \colon \mathcal{X} \to \mathbb{R}$ be arbitrary and, for every $k \in \mathbb{N}$, define $V^{k+1} \colon \mathcal{X} \to \mathbb{R}$ by*

$$V^{k+1}(x) = \sup_{a \in A} \left\{ r(x,a) + \sum_{y \in \mathcal{X}} p(y|x,a) V^k(y) \right\} \quad \text{for all } x \in \mathcal{X}.$$

Then the sequence of functions $(V^k)_{k \in \mathbb{N}}$ converges towards V^*.

(iv) [Policy iterations] *Let $\pi^0 \colon \mathcal{X} \to A$ be arbitrary and, for every $k \in \mathbb{N}$, define*

$\pi^{k+1}\colon \mathcal{X} \to \mathbb{R}$ *by*

$$\pi_{k+1}(x) \in \operatorname*{argmax}_{a\in A} \left\{ r(x,a) + \sum_{y\in\mathcal{X}} p(y|x,a) V^{\pi^k}(y) \right\} \quad \textit{for all } x \in \mathcal{X}.$$

Then $V^{\pi^k} \leq V^{\pi^{k+1}}$ *and there exists* $k^* \in \mathbb{N}$ *such that* $V^{\pi^k} = V^{\pi^{k^*}} = V^*$ *for all* $k \in \mathbb{N}$, $k \geq k^*$.

Remark 6.

(i) If V^* is defined by (14) one can show that there exists at least one optimal stationary policy and, hence, $V^*(x) = \sup_{\pi\in\Pi_{\text{stat}}} V^\pi(x)$ holds for all $x \in \mathcal{X}$.

(ii) If all the policies are proper, one can show that, for every $\pi \in \Pi_{\text{stat}}$, V^π is the limit of the sequence

$$V^0 \text{ arbitrary}, \ (\forall\, k \geq 0) \quad V^{k+1}(x) = r(x,\pi(x)) + \sum_{y\in\mathcal{X}} p(y|x,\pi(x)) V^k(y).$$

An introduction to stochastic algorithms

◇ [Introduction and main convergence result] Let $H : \mathbb{R}^d \to \mathbb{R}^d$. We are interested in solving the fixed-point equation

$$H(r) = r. \tag{29}$$

If H is a contraction, then Theorem 3 ensures that the sequence

$$\mathbb{R}^d \ni r_0 \text{ arbitrary}, \quad (\forall\, k \geq 0) \;\; r_{k+1} = H(r_k) \tag{30}$$

converges to the unique fixed point r^* of H.
We can also modify the previous sequence by taking $\gamma \in (0, 1]$ and letting

$$\mathbb{R}^d \ni r_0 \text{ arbitrary}, \quad (\forall\, k \geq 0) \;\; r_{k+1} = (1 - \gamma) r_k + \gamma H(r_k), \tag{31}$$

or, equivalently,

$$\mathbb{R}^d \ni r_0 \text{ arbitrary}, \quad (\forall\, k \geq 0) \;\; r_{k+1} = r_k + \gamma (H(r_k) - r_k), \tag{32}$$

Since $\mathbb{R}^d \ni r \mapsto (1-\gamma)r + \gamma H(r) \in \mathbb{R}^d$ is also a contraction, the previous sequence also converges to r^*.

If H is not a contraction, then, in some cases, one can still hope to show that the sequence in (31) converges or, at least, to show that $H(r_k) - r_k$ converges to 0. In the former stronger case, if H is continuous at the limit point, the latter is a fixed point of H.

◇ [Deterministic gradient methods]

Given $f : \mathbb{R}^d \to \mathbb{R}$ of class C^1, in unconstrained minimization problems one often tries to solve the equation

$$\nabla f(r) = 0 \quad \text{or, equivalently,} \quad r - \nabla f(r) = r. \tag{33}$$

Setting $H = \mathrm{id} - \nabla f$, the sequence in (32) reads

$$\mathbb{R}^d \ni r_0 \text{ arbitrary}, \quad (\forall\, k \geq 0)\ \ r_{k+1} = r_k - \gamma \nabla f(r_k), \tag{34}$$

i.e. one recovers the standard steepest descent gradient method with constant stepsize. Notice that H is not a contraction. However, let us recall the following

result.

Proposition 2 (Convergence of the gradient method for constant step sizes)**.** *Let* $L > 0$*, assume that* f *has a* L*-Lipschitz gradient, i.e.*

$$|\nabla f(x) - \nabla f(y)| \leq L|x - y| \quad \textit{for all } x,\ y \in \mathbb{R}^d, \tag{35}$$

and consider the sequence $(r_k)_{k\in\mathbb{N}}$ *constructed in* (34) *with* $\gamma \in]0, 2/L[$*. Then, as* $k \to \infty$*, either* $f(r_k) \to -\infty$*, or* $(f(r_k))_{k\in\mathbb{N}}$ *is nonincreasing, convergent, and* $\lim_{k\to\infty} \nabla f(r_k) = 0$*.*

The result in Proposition 2 can be extended to more general *gradient methods*, i.e. when the iterates are given by

$$\mathbb{R}^d \ni r_0 \text{ arbitrary}, \quad (\forall\, k \geq 0)\ \ r_{k+1} = r_k + \gamma s_k, \tag{36}$$

where the sequence $(s_k)_{k\in\mathbb{N}}$ satisfies

$$(\exists\, c_1,\ c_2 \in]0, \infty[)(\forall\, k \in \mathbb{N}) \quad \begin{array}{l} c_1|\nabla f(r_k)|^2 \leq -\langle \nabla f(r_k), s_k\rangle, \quad \text{descent direction} \\ |s_k| \leq c_2|\nabla f(r_k)|, \quad \text{bounded} \end{array} \tag{37}$$

and a key point of its proof is the following consequence of (35).

Lemma 2 (Descent Lemma). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be of class C^1 and such that* (35) *holds. Then*

$$(\forall\, x,\, h \in \mathbb{R}^d) \quad f(x) + \langle \nabla f(x), h\rangle - \frac{L}{2}|h|^2 \leq f(x+h) \leq f(x) + \langle \nabla f(x), h\rangle + \frac{L}{2}|h|^2.$$

Proof. For all x, $h \in \mathbb{R}^d$, we have

$$\begin{aligned} f(x+h) &= f(x) + \int_0^1 \langle \nabla f(x+\tau h), h\rangle \mathrm{d}\tau \\ &= f(x) + \langle \nabla f(x), h\rangle + \int_0^1 \langle \nabla f(x+\tau h) - \nabla f(x), h\rangle \mathrm{d}\tau. \end{aligned}$$

Therefore, by (35), we have

$$|f(x+h) - [f(x) + \langle \nabla f(x), h\rangle]| \leq \frac{L}{2}|h|^2,$$

from which the result follows. □

In what follows we will mainly use the second inequality in the result of Lemma 2. We have the following extension of Proposition 2.

Theorem 7 (Convergence results for gradient methods with constant step size)**.** *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be of class C^1 and such that* (35) *holds. Let $(r_k)_{k\in\mathbb{N}}$ be the sequence defined by* (36)*, with*

$$\gamma \in \left]0, \frac{2c_1}{Lc_2^2}\right[, \tag{38}$$

and assume that (37) *holds. Then, as $k \to \infty$, either $f(r_k) \to -\infty$, or $(f(r_k))_{k\in\mathbb{N}}$ is nonincreasing, $f(r_{k+1}) < f(r_k)$ if $\nabla f(r_k) \neq 0$, convergent, and $\lim_{k\to\infty} \nabla f(r_k) = 0$.*

Proof. Taking $x = r_k$ and $h = \gamma s_k$ in Lemma 2, inequalities (37) yield

$$\begin{aligned} f(r_{k+1}) &\leq f(r_k) + \gamma\langle \nabla f(r_k), s_k\rangle + \tfrac{\gamma^2 L}{2}\|s_k\|^2 \\ &\leq f(r_k) + \gamma\left(\tfrac{\gamma L c_2^2}{2} - c_1\right)\|\nabla f(r_k)\|^2. \end{aligned} \tag{39}$$

Thus, (38) implies that if f is nonincreasing and $f(r_{k+1}) < f(r_k)$ if $\nabla f(r_k) \neq 0$. The first property implies that if $f(r_k) \to -\infty$ is false, as $k \to \infty$, then the sequence $(f(r_k))_{k\in\mathbb{N}}$ is convergent. In particular, as $k \to \infty$, $f(r_k) - f(r_{k+1}) \to 0$ and, by (39), $\nabla f(r_k) \to 0$. □

If $f\colon \mathbb{R}^d \to \mathbb{R}$ is of class C^1 and it satisfies (35), but L is not easy to estimate, then one can use a sequence of variable $(\gamma_k)_{k\in\mathbb{N}}$ step sizes and consider the sequence

$$\mathbb{R}^d \ni r_0 \text{ arbitrary}, \quad (\forall\, k \geq 0)\ \ r_{k+1} = r_k + \gamma_k s_k. \tag{40}$$

For this type of iterates, the following result holds:

Proposition 3. *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be of class C^1 and let $(\gamma_k)_{k\in\mathbb{N}} \subseteq]0, +\infty[$ be such that*

$$\gamma_k \underset{k\to\infty}{\longrightarrow} 0, \quad \textit{and} \quad \sum_{k=0}^{\infty} \gamma_k = +\infty. \tag{41}$$

Let $(r_k)_{k\in\mathbb{N}}$ be the sequence defined by (40), and assume that (37) holds. Then, as $k \to \infty$, either $f(r_k) \to -\infty$, or $(f(r_k))_{k\in\mathbb{N}}$ is convergent, $\lim_{k\to\infty} \nabla f(r_k) = 0$, and there exists $\overline{k}$ such that $(f(r_k))_{k\geq\overline{k}}$ is nonincreasing and, for all $k \geq \overline{k}$, we have $f(r_{k+1}) < f(r_k)$ if $\nabla f(r_k) \neq 0$.

Proof. See [2, Proposition 3.5]. □

Remark 7. Relation (40) implies that

$$|r_{k+1} - r_0| = \left| \sum_{t=0}^{k} \gamma_t s_t \right| \leq \left(\sup_{t \in \{0,\dots,k\}} |s_t| \right) \sum_{t=0}^{\infty} \gamma_t.$$

Therefore, if the sequence $(s_k)_{k\in\mathbb{N}}$ is bounded and $\sum_{k=0}^{\infty} \gamma_k < \infty$ then the sequence $(r_k)_{k\in\mathbb{N}}$ remains in a ball centered at r_0 and, hence, it cannot converge to a stationary point of ∇f if the latter is far enough from r_0. This motivates the second condition in (41) on the sequence $(\gamma_k)_{k\in\mathbb{N}}$.

In the general framework of (29), and iterates (31), if, at the iteration $k \in \mathbb{N}$, we do not have acces to $H(r_k)$ but we observe a noisy value $H(r_k) + \xi_k$, where ξ_k is a random variable, then it is reasonable to replace $H(r_k)$ in (32) by $H(r_k) + \xi_k$. This yields the random iterates

$$\mathbb{R}^d \ni r_0 \text{ arbitrary}, \quad (\forall\, k \geq 0) \;\; r_{k+1} = r_k + \gamma(H(r_k) + \xi_k - r_k). \tag{42}$$

Being constructed with a constant step size γ, it is easy to check that in this stochastic setting the previous iterates, in general, do not converge to a deterministic limit. Indeed, suppose, for instance, that the sequence of noises $(\xi_k)_{k\in\mathbb{N}}$ is i.i.d. with

variance $V(\xi_1) = \sigma^2 > 0$. Then, by independence,

$$(\forall\, k \geq 0)\ \ V(r_{k+1}) = V((1-\gamma)r_k + \gamma H(r_k)) + \gamma^2 V(\xi_k) \geq \gamma^2\sigma^2 > 0.$$

Therefore, it is natural to consider a variant of (42) with variable step size. More precisely, we consider the iterates

$$\mathbb{R}^d \ni r_0 \text{ arbitrary}, \quad (\forall\, k \geq 0)\ \ r_{k+1} = r_k + \gamma_k s_k, \tag{43}$$

where the sequences $(s_k)_{k\in\mathbb{N}} \subset \mathbb{R}^d$ and $(\gamma_k)_{k\in\mathbb{N}} \subset [0, +\infty[$ are random, and we assume that the step size γ_k $(k \in \mathbb{N})$ can be written as a deterministic function of the random variables in

$$\mathcal{F}_k = \{r_0, \ldots, r_k, \gamma_0, \ldots, \gamma_{k-1}, s_0, \ldots, s_{k-1}\}\,. \tag{44}$$

In this case, we say that the sequence $(\gamma_k)_{k\in\mathbb{N}}$ is adapted to the filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$.
In what follows, we focus on the case where $H = \mathrm{id} - \nabla f$, for some function $f : \mathbb{R}^d \to \mathbb{R}$ of class C^1, i.e. we are interesting in solving

$$\nabla f(r) = 0.$$

The main assumptions on f and the iterates (43) are the following Lyapunov conditions:[5]

$$\begin{cases} \text{(i)} \quad f \text{ is bounded from below.} \\ \text{(ii)} \ (\exists\, L > 0)(\forall\, x_1,\, x_2 \in \mathbb{R}^d) \quad |\nabla f(x_1) - \nabla f(x_2)| \le L|x_1 - x_2|. & \text{smoothness} \\ \text{(iii)} \ (\exists\, C \ge 0)(\forall\, k \in \mathbb{N}) \quad C|\nabla f(r_k)|^2 \le -\langle \nabla f(r_k), \mathbb{E}(s_k|\mathcal{F}_k)\rangle \text{ a.s.} & \text{descent direction} \\ \text{(iv)} \ (\exists\, K_1,\, K_2 > 0)(\forall\, k \in \mathbb{N}) \quad \mathbb{E}(|s_k|^2|\mathcal{F}_k) \le K_1 + K_2|\nabla f(r_k)|^2 \text{ a.s.} & \text{boundedness} \end{cases} \tag{45}$$

Theorem 8. *Assume that* (45) *holds and that the step sizes* $(\gamma_k)_{k\in\mathbb{N}}$ *in* (43) *satisfy*

$$\sum_{k=0}^{\infty} \gamma_k = +\infty \quad \textit{and} \quad \sum_{k=0}^{\infty} \gamma_k^2 < +\infty \quad \textit{a.s.} \tag{46}$$

Then the following holds:

[5] Recall that, given a σ-field $\mathcal{F}_k$, $\mathbb{E}(X|\mathcal{F}_k)$ is well defined if $\min\{\mathbb{E}(X^+|\mathcal{F}_k), \mathbb{E}(X^-|\mathcal{F}_k)\} < \infty$. The existence of $\mathbb{E}(X^\pm|\mathcal{F}_k)$ follows from $X^\pm \ge 0$, almost surely, and setting $\mathbb{E}(X^\pm|\mathcal{F}_k) = \mathrm{d}\mathbb{Q}^\pm/\mathrm{d}\mathbb{P}$, where $\mathbb{Q}^\pm \colon \mathcal{F}_k \to [0, +\infty[\colon A \mapsto \int_A X^\pm(\omega)\mathrm{d}\mathbb{P}(\omega)$ (which is a σ-finite measure).

(i) *The sequence* $(f(r_k))_{k\in\mathbb{N}}$ *converges almost surely.*

(ii) *It holds that*

$$\lim_{k\to\infty} \nabla f(r_k) = 0 \quad a.s.$$

Remark 8. An extension of the previous result to the case of a general function $H : \mathbb{R}^d \to \mathbb{R}^d$, i.e. when equation (29) and iterates (42) (with variable step sizes) are considered, can be found in [2, Chapter 4.3].

◇ [Two important applications] We apply Theorem 8 to two well-known stochastic algorithms.

◇ [The stochastic gradient descent] Let $f : \mathbb{R}^d \to \mathbb{R}$ be of class C^1. Consider the iterates

$$\mathbb{R}^d \ni r_0 \text{ arbitrary}, \quad (\forall\, k \geq 0)\ \ r_{k+1} = r_k - \gamma_k(\nabla f(r_k) + \xi_k), \tag{47}$$

where $(\gamma_k)_{k\in\mathbb{N}} \subset [0, +\infty[$ and $(\xi_k)_{k\in\mathbb{N}} \subset \mathbb{R}^d$ are sequences of random variables. For every $k \in \mathbb{N}$ set $s_k = -\nabla f(r_k) - \xi_k$ and let $\mathcal{F}_k$ be defined by (44). We

assume that $(\gamma_k)_{k\in\mathbb{N}}$ is adapted to $(\mathcal{F}_k)_{k\in\mathbb{N}}$ and that

$$\begin{cases} \text{(i)} \quad f \text{ is bounded from below.} \\ \text{(ii)}\ (\exists\, L>0)(\forall\, x_1,\, x_2 \in \mathbb{R}^d) \quad |\nabla f(x_1) - \nabla f(x_2)| \le L|x_1 - x_2|. \\ \text{(iii)}(\forall\, k \in \mathbb{N}) \quad \mathbb{E}(\xi_k|\mathcal{F}_k) = 0 \\ \text{(iv)}\ (\exists\, A,\, B>0)(\forall\, k \in \mathbb{N}) \quad \mathbb{E}(|\xi_k|^2|\mathcal{F}_k) \le A + B|\nabla f(r_k)|^2 \text{ a.s.} \end{cases} \tag{48}$$

For the iterates (43), let us check that (45)(iii) holds. For every $k \in \mathbb{N}$, (48)(iii) implies

$$-\langle \nabla f(r_k), \mathbb{E}(s_k|\mathcal{F}_k)\rangle = -\langle \nabla f(r_k), (-\nabla f(r_k))\rangle = |\nabla f(r_k)|^2,$$

and hence (45)(iii) is satisfied with $C = 1$. Let us check that (45)(iv) holds. For

every $k \in \mathbb{N}$, (48)(iii)(iv) yield

$$\begin{aligned}
\mathbb{E}(|s_k|^2|\mathcal{F}_k) &= |\nabla f(r_k)|^2 + 2\mathbb{E}\left(\langle \nabla f(r_k), \xi_k\rangle \big| \mathcal{F}_k\right) + \mathbb{E}(|\xi_k|^2|\mathcal{F}_k) \\
&= |\nabla f(r_k)|^2 + \mathbb{E}(|\xi_k|^2|\mathcal{F}_k) \\
&\leq A + (B+1)|\nabla f(r_k)|^2
\end{aligned}$$

and hence (45)(iv) is satisfied with $K_1 = A$ and $K_2 = B + 1$. Thus, we obtain the following result.

Proposition 4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be of class C^1, let $(r_k)_{k\in\mathbb{N}}$, $(\gamma_k)_{k\in\mathbb{N}}$, and $(\xi_k)_{k\in\mathbb{N}}$ be three sequences of random variables satisfying* (47)*. Assume that* (48) *holds and that the step sizes $(\gamma_k)_{k\in\mathbb{N}}$ are adapted to $(\mathcal{F}_k)_{k\in\mathbb{N}}$ and satisfy* (46)*. Then the sequences $(f(r_k))_{k\in\mathbb{N}} \subset \mathbb{R}$ and $(\nabla f(r_k))_{k\in\mathbb{N}}$ a.s. converge and*

$$\lim_{k\to\infty} \nabla f(r_k) = 0 \quad \text{a.s.} \tag{49}$$

◇ [An incremental gradient method] Assume that we can write f as

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \quad \text{for all } x \in \mathbb{R}^d, \tag{50}$$

where $N \in \mathbb{N}^*$, for all $i = 1, \ldots, N$ the function $f_i \colon \mathbb{R}^d \to \mathbb{R}$ is of class C^1, and there exist $C_1,\ C_2 > 0$ such that

$$|\nabla f_i(x)|^2 \leq C_1 + C_2|\nabla f(x)|^2 \quad \text{for all } x \in \mathbb{R}^d. \tag{51}$$

In the incremental gradient method, instead of considering the steepest descent iterates

$$\mathbb{R}^d \ni r_0 \text{ arbitrary}, \quad (\forall\, k \geq 0)\ \ r_{k+1} = r_k - \frac{\gamma_k}{N}\sum_{i=1}^{N} \nabla f_i(r_k),$$

at each iteration k on selects, independently from the results in previous iterations,

an index I_k uniformly in $\{1, \ldots, N\}$ and constructs the iterates

$$\mathbb{R}^d \ni r_0 \text{ arbitrary}, \quad (\forall\, k \geq 0)\ \ r_{k+1} = r_k - \gamma_k \nabla f_{I_k}(r_k).$$

Notice that, for all $k \geq 0$,

$$r_{k+1} = r_k - \gamma_k \left(\nabla f(r_k) + \xi_k\right), \quad \text{with} \quad \xi_k = \nabla f_{I_k}(r_k) - \nabla f(r_k).$$

For every $k \geq 0$ we have

$$\mathbb{E}\left(\xi_k | \mathcal{F}_k\right) = \frac{1}{N} \sum_{i=1}^{N} f_i(r_k) - \nabla f(r_k) = 0$$

and, by (51),

$$\begin{aligned}
\mathbb{E}\left(|\xi_k|^2|\mathcal{F}_k\right) &= \mathbb{E}\left(|\nabla f_{I_k}(r_k)|^2|\mathcal{F}_k\right) - 2\langle \nabla f(r_k), \mathbb{E}(\nabla f_{I_k}(r_k)|\mathcal{F}_k)\rangle \\
&\quad +|\nabla f(r_k)|^2 \\
&= \tfrac{1}{N}\sum_{i=1}^N |\nabla f_i(r_k)|^2 - |\nabla f(r_k)|^2 \\
&\leq \tfrac{1}{N}\sum_{i=1}^N |\nabla f_i(r_k)|^2 \\
&\leq C_1 + C_2|\nabla f(r_k)|^2.
\end{aligned}$$

Thus, if (48)(i)(ii) hold and the step sizes $(\gamma_k)_{k\in\mathbb{N}}$ satisfy (46), we can apply the results on the stochastic gradient method to deduce that the sequences $(f(r_k))_{k\in\mathbb{N}} \subset \mathbb{R}$ and $(\nabla f(r_k))_{k\in\mathbb{N}}$ converge and

$$\lim_{k\to\infty} \nabla f(r_k) = 0.$$

Exercise 2. Let $N \in \mathbb{N}^*$. For $i = 1, \ldots, N$, let $Q_i \in \mathcal{M}_n(\mathbb{R})$ be symmetric and positive

semidefinite, let $b_i \in \mathbb{R}^d$, let $c_i \in \mathbb{R}$, and define $f_i : \mathbb{R}^d \to \mathbb{R}$ by

$$f_i(x) = \frac{1}{2}\langle Q_i x, x\rangle + \langle b, x\rangle + c \quad \text{for all } x \in \mathbb{R}^d.$$

Assume that $\sum_{i=1}^N Q_i$ is positive definite and define $f : \mathbb{R}^d \to \mathbb{R}$ by (50).

(i) Show that (51) holds.

(ii) Deduce that the sequence $(r_k)_{k\in\mathbb{N}}$ converges, almost surely, to the unique minimizer r^* of f.

◇ [Proof of Theorem 8] Let us first recall some useful results on discrete time martingales. Consider two stochastic process $(X_k)_{k\in\mathbb{N}}$ and $(Y_k)_{k\in\mathbb{N}}$ taking values in $\mathbb{R}$. Assume that

$$(\forall\, k \in \mathbb{N}) \quad \mathbb{E}(|X_k|) < +\infty.$$

We say that

- $(X_k)_{k\in\mathbb{N}}$ is a martingale with respect ot $(Y_k)_{k\in\mathbb{N}}$ if
  - $\mathbb{E}(X_{k+1}|Y_k, \ldots, Y_0) = X_k$ for all $k \in \mathbb{N}$.
- $(X_k)_{k\in\mathbb{N}}$ is a submartingale with respect ot $(Y_k)_{k\in\mathbb{N}}$ if
  - $\mathbb{E}(X_{k+1}|Y_k, \ldots, Y_0) \geq X_k$ for all $k \in \mathbb{N}$.
  - X_k is a function of $(Y_0, \ldots, Y_k)$.
- $(X_k)_{k\in\mathbb{N}}$ is a supermartingale with respect ot $(Y_k)_{k\in\mathbb{N}}$ if
  - $\mathbb{E}(X_{k+1}|Y_k, \ldots, Y_0) \leq X_k$ for all $k \in \mathbb{N}$.
  - X_k is a function of $(Y_0, \ldots, Y_k)$.

Remark 9. Notice that:

(i) If $(X_k)_{k\in\mathbb{N}}$ is a martingale with respect to $(Y_k)_{k\in\mathbb{N}}$, then for every $k \in \mathbb{N}$, the equality $X_k = \mathbb{E}(X_{k+1}|Y_k, \ldots, Y_0)$ implies that X_k is a function of $(Y_0, \ldots, Y_k)$.

(ii) $(X_k)_{k\in\mathbb{N}}$ is a supermartingale with respect to $(Y_k)_{k\in\mathbb{N}}$ iff $(-X_k)_{k\in\mathbb{N}}$ is submartingale with respect to Y_k.

(iii) $(X_k)_{k\in\mathbb{N}}$ is a martingale with respect to $(Y_k)_{k\in\mathbb{N}}$ iff it is a sub- and a supermartingale with respect to $(Y_k)_{k\in\mathbb{N}}$.

(iv) Suppose that $(X_k)_{k\in\mathbb{N}}$ is a martingale with respect to $(Y_k)_{k\in\mathbb{N}}$. Then, if $\varphi\colon \mathbb{R} \to \mathbb{R}$ is convex, Jensen's inequality implies that

$$(\forall\, k \in \mathbb{N}) \quad \mathbb{E}(\varphi(X_{k+1})|Y_k, \ldots, Y_0) \geq \varphi\left(\mathbb{E}(X_{k+1}|Y_k, \ldots, Y_0)\right) = \varphi(X_k)$$

and, hence, if for all $k \in \mathbb{N}$, $\mathbb{E}(|\varphi(X_k)|) < +\infty$, we have that $(\varphi(X_k))_{k\in\mathbb{N}}$ is a submartingale. For instance, if $(X_k)_{k\in\mathbb{N}}$ is a martingale with respect to $(Y_k)_{k\in\mathbb{N}}$ then $(|X_k|)_{k\in\mathbb{N}}$ is a submartingale with respect to $(Y_k)_{k\in\mathbb{N}}$. Moreover, if

$\mathbb{E}(X_k^2) < \infty$ for all $k \in \mathbb{N}$, then $(X_k^2)_{k\in\mathbb{N}}$ is also a submartingale with respect to $(Y_k)_{k\in\mathbb{N}}$.

Similarly, if $(X_k)_{k\in\mathbb{N}}$ is a submartingale with respect to $(Y_k)_{k\in\mathbb{N}}$ and $\varphi\colon \mathbb{R} \to \mathbb{R}$ is convex, non-decreasing and $\mathbb{E}(|\varphi(X_k)|) < \infty$ for all $k \in \mathbb{N}$, then $(\varphi(X_k))_{k\in\mathbb{N}}$ is a submartingale with respect to $(Y_k)_{k\in\mathbb{N}}$. For instance, if $(X_k)_{k\in\mathbb{N}}$ is a submartingale with respect to $(Y_k)_{k\in\mathbb{N}}$ and $c \in \mathbb{R}$ then $(\max\{X_k, c\})_{k\in\mathbb{N}}$ is a submartingale with respect to $(Y_k)_{k\in\mathbb{N}}$.

(v) Suppose that $(X_k)_{k\in\mathbb{N}}$ is such that $\mathbb{E}(|X_k|) < \infty$, for all $k \in \mathbb{N}$. Then the following hold:

- If $(X_k)_{k\in\mathbb{N}}$ a martingale then $(\mathbb{E}(X_k))_{k\in\mathbb{N}}$ is constant.
- If $(X_k)_{k\in\mathbb{N}}$ is a submartingale then $(\mathbb{E}(X_k))_{k\in\mathbb{N}}$ is non-decreasing.
- If $(X_k)_{k\in\mathbb{N}}$ is a supermartingale then $(\mathbb{E}(X_k))_{k\in\mathbb{N}}$ is non-increasing.

Example 1. Let $(X_k)_{k\in\mathbb{N}}$, $(Y_k)_{k\in\mathbb{N}}$, and $(Z_k)_{k\in\mathbb{N}}$ be three sequences of random variables. Assume that $(X_k)_{k\in\mathbb{N}}$ is a martingale with respect to $(Y_k)_{k\in\mathbb{N}}$ and that, for each $k \in \mathbb{N}$, Z_k is a bounded function of $Y_0, \ldots, Y_k$. Then the sequence $(R_k)_{k\in\mathbb{N}}$,

defined by

$$(\forall\, k \in \mathbb{N}) \quad R_k = X_0 + \sum_{t=0}^{k-1} Z_t(X_{t+1} - X_t), \tag{52}$$

is a martingale with respect to $(Y_k)_{k\in\mathbb{N}}$. Indeed, for every $k \in \mathbb{N}$, $\mathbb{E}(|X_k|) < +\infty$ and $|Z_k|$ is bounded, which, by (52), implies that $\mathbb{E}(|R_k|) < +\infty$. We also have

$$\begin{aligned}
\mathbb{E}(R_{k+1}|Y_k, \ldots, Y_0) &= X_0 + \sum_{t=0}^{k-1} Z_t(X_{t+1} - X_t) \\
&\quad + \mathbb{E}(Z_k(X_{k+1} - X_k)|Y_k, \ldots, Y_0) \\
&= R_k + Z_k\mathbb{E}(X_{k+1} - X_k|Y_k, \ldots, Y_0) \\
&= R_k.
\end{aligned}$$

Similarly, under the same assumptions on $(Z_k)_{k\in\mathbb{N}}$, if $(X_k)_{k\in\mathbb{N}}$ is a submartingale with respect to $(Y_k)_{k\in\mathbb{N}}$ (resp. supermartingale with respect to $(Y_k)_{k\in\mathbb{N}}$), then $(R_k)_{k\in\mathbb{N}}$ is a submartingale with respect to $(Y_k)_{k\in\mathbb{N}}$ (resp. supermartingale with respect to $(Y_k)_{k\in\mathbb{N}}$).

One of the main results in martingale theory is the following.

Theorem 9 (Martingale convergence theorem)**.** *Let $(X_k)_{k\in\mathbb{N}}$ and $(Y_k)_{k\in\mathbb{N}}$ be two sequences of real-valued random variables. Then the following holds:*

(i) *If $(X_k)_{k\in\mathbb{N}}$ is a submartingale with respect to $(Y_k)_{k\in\mathbb{N}}$ and $\sup_{k\in\mathbb{N}} \mathbb{E}(X_k^+) < \infty$, then X_k converges a.s.*

(ii) *If $(X_k)_{k\in\mathbb{N}}$ is a supermartingale with respect to $(Y_k)_{k\in\mathbb{N}}$ and $\sup_{k\in\mathbb{N}} \mathbb{E}(X_k^-) < \infty$, then X_k converges a.s.*

In particular, if $(X_k)_{k\in\mathbb{N}}$ is a martingale with respect to $(Y_k)_{k\in\mathbb{N}}$ and $\sup_{k\in\mathbb{N}} \mathbb{E}(|X_k|) < \infty$, then X_k converges a.s.

Proof. The proof of assertion (i) can be found, for instance, in [5, Theorem 9.23]. Using this result, assertion (ii) follows from Remark 9(ii). □

The following corollary to the previous result is often useful.

Corollary 3. *Let $(X_k)_{k\in\mathbb{N}}$ and $(Y_k)_{k\in\mathbb{N}}$ be two sequences of real-valued random variables. Then the following hold:*

(i) *If* $(X_k)_{k\in\mathbb{N}}$ *is a submartingale with respect to* $(Y_k)_{k\in\mathbb{N}}$*, bounded from above, then it converges a.s.*

(ii) *If* $(X_k)_{k\in\mathbb{N}}$ *is a supermartingale with respect to* $(Y_k)_{k\in\mathbb{N}}$*, bounded from below, then it converges a.s.*

Theorem 10. *Let* $(W_k)_{k\in\mathbb{N}}$*,* $(X_k)_{k\in\mathbb{N}}$*,* $(Y_k)_{k\in\mathbb{N}}$*, and* $(Z_k)_{k\in\mathbb{N}}$ *be fourth sequences of random variables. Assume that, for all* $k \in \mathbb{N}$*,* W_k*,* X_k*, and* Z_k *are non-negative functions of* $(Y_0, \ldots, Y_k)$*, Moreover, suppose that*

$$\sum_{k=0}^{\infty} Z_k < +\infty$$

and that

$$(\forall\, k \in \mathbb{N}) \quad \mathbb{E}(X_{k+1}|Y_k, \ldots, Y_0) \leq X_k - W_k + Z_k. \tag{53}$$

Then, almost surely, $(X_k)_{k\in\mathbb{N}}$ *converges and* $\displaystyle\sum_{k=0}^{\infty} W_k < +\infty$*.*

Proof.

$$R_k = X_k + \sum_{t=0}^{k-1} W_t - \sum_{t=0}^{k-1} Z_t. \tag{54}$$

Given $N \in \mathbb{N}$, let us set

$$\tau_N = \inf \left\{ k \in \mathbb{N} \,\middle|\, \sum_{t=0}^{k} Z_t > N \right\}$$

and, for all $k \in \mathbb{N}$, let us define $R_k^N = R_{k \wedge \tau_N}$. Then we have

$$\mathbb{E}\left(R_{k+1}^N \middle| Y_k, \ldots, Y_0\right) - R_k^N = \begin{cases} \mathbb{E}\left(X_{k+1} \middle| Y_k, \ldots, Y_0\right) - X_k + W_k - Z_k & \text{if } k < \tau^N, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, by (53), we have that $(R_k^N)_{k\in\mathbb{N}}$ is a supermartingale bounded from below (by $-N$) and hence Corollary 3 implies that $(R_k^N)_{k\in\mathbb{N}}$ is convergent a.s. By countable additivity, we deduce that a.s. $(R_k^N)_{k\in\mathbb{N}}$ is convergent for every $N \in \mathbb{N}$. Since $\sum_{k=0}^{\infty} Z_k < \infty$ a.s., we obtain that $(R_k)_{k\in\mathbb{N}}$ is convergent a.s. Thus, since $\sum_{t=0}^{k-1} W_t$ is a.s. non-decreasing, and $X_k \geq 0$, by (54), we deduce that $\sum_{t=0}^{k-1} W_t$ and $(X_k)_{k\in\mathbb{N}}$ are convergent. The result follows. □

Let us now prove Theorem 8. By (45)(i), without loss of generality, that $f \geq 0$. For

every $k \in \mathbb{N}$, Lemma 2 implies that

$$\begin{aligned}
\mathbb{E}\left(f(r_{k+1})\middle|\mathcal{F}_k\right) &\leq \mathbb{E}\left(f(r_k) + \gamma_k\langle\nabla f(r_k), s_k\rangle + \tfrac{L}{2}\|s_k\|^2\middle|\mathcal{F}_k\right) \\
&= f(r_k) + \gamma_k\langle\nabla f(r_k), \mathbb{E}(s_k|\mathcal{F}_k)\rangle + \tfrac{L\gamma_k^2}{2}\mathbb{E}\left(|s_k|^2\middle|\mathcal{F}_k\right) \\
&\leq f(r_k) - \left(C\gamma_k - \tfrac{LK_2\gamma_k^2}{2}\right)|\nabla f(r_k)|^2 + \tfrac{LK_1\gamma_k^2}{2}
\end{aligned}$$

Since $\lim_{k\to\infty}\gamma_k = 0$, there exists $\overline{k} \in \mathbb{N}$ such that, if $k \geq \overline{k}$, we have

$$\mathbb{E}\left(f(r_{k+1})\middle|\mathcal{F}_k\right) \leq f(r_k) - \frac{C\gamma_k}{2}|\nabla f(r_k)|^2 + \frac{LK_1\gamma_k^2}{2} \tag{55}$$

Since $\sum_{k=0}^{\infty}\gamma_k^2 < +\infty$, we can apply Theorem 10 with

$$X_k = f(r_k), \quad Y_k = \frac{C\gamma_k}{2}|\nabla f(r_k)|^2, \quad \text{and} \quad Z_k = \frac{LK_1\gamma_k^2}{2}$$

to deduce that $(f(r_k))_{k\in\mathbb{N}}$ converges almost surely and

$$\sum_{k=0}^{\infty} \gamma_k |\nabla f(r_k)|^2 < +\infty \quad \text{a.s.}, \tag{56}$$

which, together with the condition $\sum_{k=0}^{\infty} \gamma_k = +\infty$, almost surely, implies that

$$\liminf_{k\to\infty} |\nabla f(r_k)| = 0 \quad \text{a.s.} \tag{57}$$

In order to conclude the proof, we must show that

$$\limsup_{k\to\infty} |\nabla f(r_k)| = 0 \quad \text{a.s.} \tag{58}$$

Let $\varepsilon > 0$. Given k, $\overline{k} \in \mathbb{N}$, with $\overline{k} > k$, we say that $\{k, k+1, \ldots, \overline{k}\}$ is an

upcrossing set from $\varepsilon/2$ to ε if

$$\begin{cases} \bullet \quad |\nabla f(r_k)| < \varepsilon/2, \\ \bullet \quad (\forall\, k' = k+1, \dots, \overline{k}-1) \quad \varepsilon/2 \le |\nabla f(r_{k'})| \le \varepsilon, \\ \bullet \quad |\nabla f(r_{\overline{k}})| > \varepsilon. \end{cases} \tag{59}$$

By (57), it suffices to show that for every $\varepsilon > 0$ the number UC_ε of upcrossing sets from $\varepsilon/2$ to ε is almost surely finite. Indeed, (57) implies that

$$\{\omega \in \Omega \mid \limsup_{k\to\infty} |\nabla f(r_k)| \neq 0\} \subset \bigcup_{\varepsilon\in\mathbb{Q}} \{UC_\varepsilon < +\infty\}.$$

We will show that if, with positive probability, $UC_\varepsilon = +\infty$, we get a contradiction with (56). Thus, suppose that $UC_\varepsilon = \infty$ and denote by

$$\{k_1, \dots, \overline{k}_1\}, \dots, \{k_t, \dots, \overline{k}_t\}, \dots, \quad \text{with } (k_t)_{t\in\mathbb{N}} \text{ increasing and } k_t \underset{t\longrightarrow+\infty}{\longrightarrow} \infty,$$

the sequence of upcrossing sets from $\varepsilon/2$ to ε. For all $k \in \mathbb{N}$ let us set

$\overline{s}_k = \mathbb{E}(s_k|\mathcal{F}_k)$, $\xi_k = s_k - \overline{s}_k$ and notice that $\mathbb{E}(\xi_k|\mathcal{F}_k) = 0$. By Jensen's inequality and (45)(iv) we have that, for all $t \in \mathbb{N}$ and $\ell = k_t, \ldots, \overline{k}_t - 1$,

$$|\overline{s}_\ell|^2 \leq \mathbb{E}(|s_\ell|^2|\mathcal{F}_\ell) \leq K_1 + K_2|\nabla f(r_\ell)|^2 \leq K_1 + K_2\varepsilon^2. \tag{60}$$

Let $(\chi_k)_{k\in\mathbb{N}}$ be the sequence defined as $\chi_k = 1$ if $|\nabla f(r_k)| \leq \varepsilon$ and $\chi_k = 0$, otherwise. Consider now the sequence of random variables $(u_k)_{k\in\mathbb{N}}$ defined by

$$u_0 = 0, \quad (\forall\, k \geq 1) \quad u_k = \sum_{t=0}^{k-1} \chi_t\gamma_t\xi_t.$$

Claim: The sequence $(u_k)_{k\in\mathbb{N}}$ is an a.s. convergent martingale.

Suppose that the claim is true. Then $\gamma_{k_t}\xi_{k_t} = \chi_{k_t}\gamma_{k_t}\xi_{k_t} \to 0$ a.s. and, hence, (60) implies that

$$\begin{aligned} \big||\nabla f(r_{k_t+1})| - |\nabla f(r_{k_t})|\big| &\leq \big|\nabla f(r_{k_t+1}) - \nabla f(r_{k_t})\big| \\ &\leq L\gamma_{k_t}\left(|\overline{s}_{k_t}| + |\xi_{k_t}|\right) \underset{t\to\infty}{\longrightarrow} 0. \end{aligned} \tag{61}$$

In particular, there exists $\overline{t} \in \mathbb{N}$ such that for all $t \in \mathbb{N}$, $t \geq \overline{t}$, we have $|\nabla f(r_{k_t})| \geq \varepsilon/4$ and hence

$$\sum_{k=0}^{\infty} \gamma_k |\nabla f(r_k)|^2 \geq \sum_{t=\overline{t}}^{\infty} \sum_{\ell=k_t}^{\overline{k}_t - 1} \gamma_\ell |\nabla f(r_\ell)|^2 \geq \frac{\varepsilon^2}{16} \sum_{t=\overline{t}}^{\infty} \sum_{\ell=k_t}^{\overline{k}_t - 1} \gamma_\ell. \tag{62}$$

On the other hand, for every $t \in \mathbb{N}$, (59) and (45)(ii) yield

$$\begin{aligned} \frac{\varepsilon}{2} &\leq \big| |\nabla f(r_{\overline{k}_t})| - |\nabla f(r_{k_t})| \big| \\ &\leq \big| \nabla f(r_{\overline{k}_t}) - \nabla f(r_{k_t}) \big| \\ &\leq L \sum_{\ell=k_t}^{\overline{k}_t - 1} \gamma_\ell |\overline{s}_\ell| + L \sum_{\ell=k_t}^{\overline{k}_t - 1} \gamma_\ell |\xi_\ell|, \end{aligned} \tag{63}$$

Notice that

$$\sum_{\ell=k_t}^{\overline{k}_t-1} \gamma_\ell \xi_\ell = \sum_{\ell=k_t}^{\overline{k}_t-1} \chi_\ell \gamma_\ell \xi_\ell \underset{t\to\infty}{\longrightarrow} 0$$

and hence (63), the inequality $a \leq a^2 + 1$, for all $a \in [0, +\infty[$, and (60) imply that

$$\frac{\varepsilon}{4} \leq L(1 + K_1 + K_2\varepsilon^2) \sum_{\ell=k_t}^{\overline{k}_t-1} \gamma_\ell,$$

provided that $\overline{t}$ is large enough. Therefore,

$$C_\varepsilon \leq \sum_{\ell=k_t}^{\overline{k}_t-1} \gamma_\ell \quad \text{for all } t \in \mathbb{N},\ t \geq \overline{t},$$

where $C_\varepsilon := \varepsilon/4L(1 + K_1 + K_2\varepsilon^2)$, which, in view of (62), is in contradiction with (56).

It remains to prove the claim. For all $k \in \mathbb{N}$, we have

$$\mathbb{E}(u_{k+1}|\mathcal{F}_k) = \mathbb{E}\left(\sum_{t=0}^{k-1} \chi_t \gamma_t \xi_t + \chi_k \gamma_k \xi_k \middle| \mathcal{F}_k\right) = u_k + \chi_k \gamma_k \mathbb{E}(\xi_k|\mathcal{F}_k) = u_k,$$

which shows that $(u_k)_{k\in\mathbb{N}}$ is a martingale. Suppose that there exists $N \in \mathbb{N}$ such that

$$\sum_{t=0}^{k} \gamma_k^2 \leq N \quad a.s. \tag{64}$$

Then, for all $k \in \mathbb{N}$, we have

$$\begin{aligned}
\mathbb{E}\left(|u_{k+1}|^2|\mathcal{F}_k\right) &= \mathbb{E}\left(|u_k + \chi_k \gamma_k \xi_k|^2|\mathcal{F}_k\right) \\
&= |u_k|^2 + \chi_k \gamma_k^2 \mathbb{E}(\xi_k^2|\mathcal{F}_k) \\
&\leq |u_k|^2 + 2\chi_k \gamma_k^2 (K_1 + K_2 |\nabla f(r_k)|^2) \\
&\leq |u_k|^2 + 2(K_1 + K_2 \varepsilon^2)\gamma_k^2,
\end{aligned}$$

which implies that

$$\mathbb{E}\left(|u_k|^2|\mathcal{F}_k\right) \leq |u_0|^2 + 2(K_1 + K_2\varepsilon^2)\sum_{t=0}^{\infty}\gamma_t^2 \leq |u_0|^2 + 2N(K_1 + K_2\varepsilon^2).$$

Thus, by the martingale convergence theorem (Theorem 9), we deduce that $(u_k)_{k\in\mathbb{N}}$ is a.s. convergent.
Finally, if (64) does not hold, for every $N \in \mathbb{N}$, define the sequence $(\gamma_k^N)_{k\in\mathbb{N}}$ as

$$(\forall\, k \in \mathbb{N}) \quad \gamma_k^N = \begin{cases} \gamma_k & \text{if } \sum_{t=0}^{k}\gamma_k^2 \leq N, \\ 0 & \text{otherwise.} \end{cases}$$

Then, considering the sequence $(u_k^N)_{k\in\mathbb{N}}$ defined by

$$u_0^N = 0, \quad (\forall\, k \geq 1) \quad u_k^N = \sum_{t=0}^{k-1}\chi_t\gamma_t^N\xi_t,$$

we have that, for every $N \in \mathbb{N}$, $(u_k^N)_{k\in\mathbb{N}}$ is a martingale and, since $\sum_k(\gamma_k^N)^2 \leq N$,

it also converges almost surely. In particular, $(u_k)_{k\in\mathbb{N}}$ is almost surely convergent (indeed, $\{(u_k)_{k\in\mathbb{N}} \text{ is convergent}\} = \bigcap_{N\in\mathbb{N}}\{(u_k^N)_{k\in\mathbb{N}} \text{ is convergent}\}$).

Some important reinforcement learning algorithms

◇ [Monte Carlo methods] Let X be a real-valued random variable such that $\mathbb{E}(|X|) < +\infty$. Assume that we are interested in approximating the mean $\mu := \mathbb{E}(X)$. For this purpose, we implement a *Monte Carlo approach*. Consider first an i.i.d. sampling $(X_k)_{k\in\mathbb{N}}$ of X and set

$$(\forall\, k \in \mathbb{N}) \quad S_k = \frac{1}{k}\sum_{t=1}^{k} X_t. \tag{65}$$

The main idea of Monte Carlo methods is to approximate μ by the sequence $(S_k)_{k\in\mathbb{N}}$. From the definition, for all $k \in \mathbb{N}$, we have that

$$\mathbb{E}(S_k) = \mu, \tag{66}$$

which means that the *sample mean* (or *empirical mean*) estimator S_k is *unbiased*. If

we assume, in addition, that

$$\sigma^2 := \mathbb{E}(|X - \mu|^2) < \infty,$$

then the variance of S_k is given by

$$\mathrm{Var}(S_k) = \frac{\sigma^2}{k^2} \tag{67}$$

Noticing that, for any couple of real-valued radom variables (Y, Z), we have

$$\begin{aligned}(Y - Z)^2 &= (Y - Z - \mathbb{E}(Y - Z))^2 \\ &\quad +2(Y - Z - \mathbb{E}(Y - Z))(\mathbb{E}(Y) - \mathbb{E}(Z)) \\ &\quad +(\mathbb{E}(Y) - \mathbb{E}(Z))^2,\end{aligned}$$

and hence

$$\mathbb{E}\left(|Y - Z|^2\right) = \mathrm{Var}(Y - Z) + (\mathbb{E}(Y) - \mathbb{E}(Z))^2,$$

we deduce from (66) and (67) that

$$\mathbb{E}(|S_k - \mu|^2) \underset{k\to\infty}{\longrightarrow} 0.$$

In particular, $(S_k)_{k\in\mathbb{N}}$ converges to μ in probability. This result is called the weak law of large numbers. The following sharper result[6], and more difficult to prove, is due to Kolmogorov.

Theorem 11 (Strong law of large numbers)**.** *Suppose that* $\mathbb{E}(|X|) < +\infty$*. Then it holds that*

$$S_k \underset{k\to\infty}{\longrightarrow} \mu \quad \text{a.s.} \tag{68}$$

Proof. See e.g. [4, Chapter 4, Theorem 3] for Kolmogorov's classical proof and [5, Theorem 9.40] for a proof based on martingale methods. If $\mathbb{E}(|X|^2) < \infty$, we can also derive the result from the convergence of the stochastic gradient method. Indeed, the sequence $(S_k)_{k\in\mathbb{N}}$ can be computed as

$$(\forall\, k \in \mathbb{N}) \quad S_{k+1} = S_k + \gamma_k(X_{k+1} - S_k), \tag{69}$$

[6] Recall that almost sure convergence of a sequence of random variables implies convergence in probability.

where $\gamma_k = 1/(k+1)$. Letting $f\colon \mathbb{R}^d \to \mathbb{R}\colon x \mapsto \frac{1}{2}|x-\mu|^2$, the iterates (69) correspond to the stochastic gradient method iterates (47) applied to f, with $r_k = S_k$ and $\xi_k = \mu - X_{k+1}$. Since $\mathbb{E}\left(\xi_k|\mathcal{F}_k\right) = \mathbb{E}(\xi_k) = 0$ and (48) holds, because, for all $k \in \mathbb{N}$, $\mathbb{E}\left(|\xi_k|^2|\mathcal{F}_k\right) = \mathrm{Var}(X)$, it follows from Proposition 4 that, as $k \to \infty$, $\nabla f(S_k) = S_k - \mu \to 0$ a.s. □

Remark 10. In particular, if $A \in \mathcal{B}(\mathbb{R}^d)$ is such that $p = \mathbb{P}(X \in A) > 0$, then, since

$$p = \lim_{k\to\infty} \frac{\sum_{t=1}^{k} \mathbb{I}_A(X_t)}{k} \quad \text{a.s.},$$

we see that, with probability one, X_k must visit A infinitely often.

If the sequence $(X_k)_{k\in\mathbb{N}}$ is not i.i.d., then (68) does not necessarily holds. However, there are instances where a small dependence between X_m and X_n, with $n \gg m$, yields to similar results. This is the case, for instance, of irreducible, positively recurrent Markov chains.[7]

[7] Recall that if $(X_k)_{k\in\mathbb{N}}$ is irreducible Markov chain then the state space $\mathcal{X}$ is the only communicating class. This class can be trasient, null recurrent and positive recurrent. Only in the latter case, a stationary distribution exists and, in this case, it is unique.

Theorem 12 (Ergodic theorem for Markov chains). *Let $(X_k)_{k\in\mathbb{N}}$ be an irreducible positive recurrent Markov chain with a countable state space $\mathcal{X}$ and denote by $\pi = \{\pi_x \,|\, x \in \mathcal{X}\}$ its unique stationary distribution. Suppose that $f : \mathcal{X} \to \mathbb{R}$ satisfies*

$$\sum_{x\in\mathcal{X}} |f(x)|\pi(x) < +\infty. \tag{70}$$

Then, it holds that

$$\lim_{k\to\infty} \frac{\sum_{t=1}^{k} f(X_t)}{k} = \sum_{x\in\mathcal{X}} f(x)\pi(x) \quad \text{a.s.}$$

Remark 11. If $\mathcal{X}$ is finite then (70) automatically holds for any function $f : \mathcal{X} \to \mathbb{R}$.

Let us comeback to the empirical estimator (65) and show, with the help of two simple examples, that adding dependent samples can improve or deteriorate the variance of the estimator S_k.

Example 2.

1. Suppose that we have only two i.i.d. samples X_1, X_2 of X, and let $X_3 = X_1$. Assume that $\mathrm{Var}(X) < \infty$. Then

$$\mathrm{Var}\left(\frac{1}{2}(X_1 + X_2)\right) = \frac{\mathrm{Var}(X)}{2} \quad \text{and} \quad \mathrm{Var}\left(\frac{1}{3}(X_1 + X_2 + X_3)\right) = \frac{5\mathrm{Var}(X)}{9}.$$

   Thus, in this case, adding the dependent variable $X_3 = X_1$ deteriorates the variance.

2. If we have an i.i.d. sample $X_1, \ldots, X_N$, we can consider a second sample $\tilde{X}_1, \ldots, \tilde{X}_N$ such that $(X_1, \tilde{X}_1), \ldots, (X_N, \tilde{X}_N)$ are i.i.d. In this case, setting $X_{N+k} = \tilde{X}_k$ for all $k = 1, \ldots, N$, we have

$$\mathrm{Var}\left(\frac{\sum_{k=1}^{2N} X_k}{2N}\right) = \frac{1}{4N^2} N \mathrm{Var}\left(X_1 + \tilde{X}_1\right) = \frac{1}{2N}\left(\mathrm{Var}(X_1) + \mathrm{Cov}(X_1, \tilde{X}_1)\right).$$

   Therefore, if $\mathrm{Cov}(X_1, \tilde{X}_1) \leq 0$, then

$$\mathrm{Var}\left(\frac{\sum_{k=1}^{2N} X_k}{2N}\right) \leq \frac{\mathrm{Var}(X_1)}{2N},$$

which corresponds to the variance of $\frac{\sum_{k=1}^{2N} X_k}{2N}$ if the entire sample $X_1, \ldots, X_{2N}$ was chosen i.i.d.
A typical application of this idea is to consider the approximation of

$$\int_0^1 f(x)\mathrm{d}x = \mathbb{E}(f(U)),$$

where $f\colon [0,1] \to \mathbb{R}$ integrable, i.e. f is measurable and $\int_0^1 |f(x)|\mathrm{d}x < \infty$, monotone, and U is a uniform random variable in $[0,1]$.

Exercise 3. Show that, under the previous assumptions,

$$\mathrm{Cov}(f(U), f(1-U)) \leq 0,$$

In view of the previous exercise, the estimator of $\int_0^1 f(x)\mathrm{d}x$ given by

$$\frac{1}{2N}\sum_{K=1}^{N} \left(f(U_k) + f(1-U_k)\right),$$

where $(U_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. uniform random variables in $[0,1]$, has a variance which is less or equal than the one of the standard estimator

$$\frac{1}{2N}\sum_{k=1}^{2N} f(U_k).$$

The following result, known as Wald's identity, is often useful in order to show the consistence of some estimators of μ.

Lemma 3. *Consider a sequence $(X_k)_{k\in\mathbb{N}^*}$ of real-valued random variables such that $\mathbb{E}(|X_k|) < +\infty$, for all $k \in \mathbb{N}^*$, and let N be a random variable taking values in $\mathbb{N}$. Assume that $\mathbb{E}(N) < \infty$ and the existence of $C > 0$ such that $\mathbb{E}(|X_k||N \geq k) \leq C$ for all $k \in \mathbb{N}^*$. Moreover, assume that $\mathbb{E}(X_k|N \geq k)$ is independent of*

$k \in \mathbb{N}^*$. *Then the following holds*[8]

$$\mathbb{E}\left(\left(\sum_{k=1}^{N} X_k\right)\mathbb{I}_{\{N\geq 1\}}\right) = \mathbb{E}(N)\mathbb{E}(X_1|N \geq 1). \tag{71}$$

Proof. First, recall that $\mathbb{E}(N) = \sum_{k=1}^{\infty} \mathbb{P}(N \geq k)$[9].

---

[8] We use the convention : $\left(\sum_{k=1}^{N(\omega)} X_k(\omega)\right)\mathbb{I}_{\{N\geq 1\}}(\omega) = 0$ if $N(\omega) = 0$.

[9] Indeed,

$$\mathbb{E}(N) = \sum_{\ell=0}^{\infty} \ell\mathbb{P}(X=\ell) = \sum_{\ell=1}^{\infty} \ell\mathbb{P}(X=\ell) = \sum_{\ell=1}^{\infty}\sum_{k=1}^{\ell} \mathbb{P}(X=\ell) = \sum_{k=1}^{\infty}\sum_{\ell=k}^{\infty} \mathbb{P}(X=\ell) = \sum_{k=1}^{\infty} \mathbb{P}(X\geq k).$$

We have

$$\begin{aligned}\mathbb{E}\left(\left(\sum_{k=1}^{N} X_k\right)\mathbb{I}_{\{N\geq 1\}}\right) &= \mathbb{E}\left(\sum_{k=1}^{\infty} X_k \mathbb{I}_{\{N\geq k\}}\right)\\ &= \sum_{k=1}^{\infty} \mathbb{E}(X_k|N\geq k)\mathbb{P}(N\geq k)\\ &= \Big(\sum_{k=1}^{\infty} \mathbb{P}(N\geq k)\Big)\mathbb{E}(X_1|N\geq 1)\\ &= \mathbb{E}(N)\mathbb{E}(X_1|N\geq 1),\end{aligned}$$

where the exchange of the expectation and summation in the second equality is justified by the assumptions $\mathbb{E}(N) < \infty$, $\mathbb{E}(|X_k||N \geq k) \leq C$, and Fubini's theorem. □

Remark 12. When N takes values in $\mathbb{N}^*$, (71) takes following form, which is the one presented in most of the textbooks,

$$\mathbb{E}\left(\sum_{k=1}^{N} X_k\right) = \mathbb{E}(N)\mathbb{E}(X_1). \tag{72}$$

◇ [Evaluation of the value of a policy by using Monte Carlo methods] We consider the framework of ???. Let π be a proper stationary deterministic policy. We suppose that the terminal state provides zero profit, i.e. $r(\hat{x}, a) = 0$ for all $a \in A$, which implies that $V^\pi(\hat{x}) = 0$. Recall that, for $x \in \mathcal{X} \setminus \{\hat{x}\}$,

$$V^\pi(x) = \mathbb{E}\left(\sum_{k=0}^{\tau^{x,\pi}-1} r\left(X_k^{x,\pi}, \pi(X_k^{x,\pi})\right) \right).$$

Based on the strong law of large numbers, a natural procedure to approximate $V^\pi(x)$ would be to consider N independent copies $\{X_k^{x,\pi,\ell} \mid k \in \mathbb{N},\ \ell = 1, \dots, N\}$, of $\{X_k^{x,\pi} \mid k \in \mathbb{N}\}$ and to consider the unbiased and consistent estimator

$$V_N(x) = \frac{\displaystyle\sum_{\ell=1}^{N} \sum_{k=0}^{\tau^{x,\pi,\ell}-1} r\left(X_k^{x,\pi,\ell}, \pi(X_k^{x,\pi,\ell})\right)}{N},$$

where

$$\tau^{x,\pi,\ell} = \inf\{k \in \mathbb{N} \mid X_k^{x,\pi,\ell} = \hat{x}\}.$$

However, from the practical point of view, it is better to consider several trajectories generated by π and update the values of the states visited by these trajectories. More precisely, let $(X_k^\pi)_{k\in\mathbb{N}}$ be a Markov chain with (time-homogeneous) transition probabilities given by $\{p(y|x,\pi(x)) \,|\, x,\ y \in \mathcal{X}\}$ and let

$$\tau^\pi = \inf\{k \in \mathbb{N} \,|\, X_k^\pi = \hat{x}\}.$$

Given $k = 0, \ldots, \tau^\pi - 1$, we set

$$J(X_k^\pi, \ldots, X_{\tau^\pi-1}^\pi) = \sum_{t=k}^{\tau^\pi-1} r\left(X_t^\pi, \pi\left(X_t^\pi\right)\right). \tag{73}$$

For $x \in \mathcal{X} \setminus \{\hat{x}\}$, define the visit times to x as

$$T_1^x = \inf\{k \in \mathbb{N} \,|\, X_k^\pi = x\},$$
$$(\forall\, k \geq 1) \quad T_{k+1}^x = \inf\{t \in \mathbb{N} \,|\, X_t^\pi = x,\ t > T_k^x\}.$$

The number of visits $\mathfrak{N}^x$ to state x is defined as

$$\mathfrak{N}^x = \#\{k \in \mathbb{N} \,|\, T_k^x < +\infty\}.$$

Remark 13. The policy π being proper, we have that $\mathfrak{N}^x < +\infty$ almost surely and $\mathbb{E}(\tau^{\pi,x}) < +\infty$.

Consider now N independent and identically distributed samples

$$(X_k^{\pi,\ell})_{k\in\mathbb{N}} \quad \ell = 1, \ldots, N, \tag{74}$$

of $(X_k^\pi)_{k\in\mathbb{N}}$ and, for $\ell = 1, \ldots, N$, denote by $\tau^{\ell,\pi}$, J^ℓ, $(T_k^{x,\ell})_{k\in\mathbb{N}}$, and $\mathfrak{N}^{x,\ell}$, the terminal time, the sum of the rewards (73), the sequence of visit times to x, and the number of visits to x, respectively, associated with $(X_k^{\pi,\ell})_{k\in\mathbb{N}}$. If $\mathbb{P}(\mathfrak{N}^x \geq 1) > 0$,

then Remark 10 implies that, almost surely,

$$V_N^\pi(x) = \frac{\displaystyle\sum_{\ell=1}^{N} \left(\sum_{k=1}^{\mathfrak{N}^{x,\ell}} J^\ell \left(X_{T_k^\ell}^{\pi,\ell}, \ldots, X_{\tau^{\pi,\ell}-1}^{\pi} \right) \right) \mathbb{I}_{\mathfrak{N}^{x,\ell} \geq 1}}{\sum_{\ell=1}^{N} \mathfrak{N}^{x,\ell}} \tag{75}$$

is well-defined for N large enough. The quantity above is called the every-visit Monte Carlo iterate to approximate $V^\pi(x)$. It can be written in the form

$$(\forall\, m \geq 1) \quad V_{m+1} = V_m + \gamma_m (J(x,m) - V_m), \tag{76}$$

where V_m denotes the m-th update of the approximation of $V^\pi(x)$, which is performed when x is visited for the m-th time by the sampled trajectories (74), $\gamma_m = 1/(m+1)$, and $J(x,m)$ is the cost (73), with k being replaced by the time where the m-th visit takes place.

A variant, called the first-visit Monte Carlo method, is given by

$$V_N^\pi(x) = \frac{\sum_{\ell=1}^{N} J^\ell \left(X^{\pi,\ell}_{T_1^\ell}, \ldots, X^{\pi}_{\tau^{\pi,\ell}-1} \right) \mathbb{I}_{\{\mathfrak{N}^{x,\ell} \geq 1\}}}{K^{N,x}}, \tag{77}$$

where

$$K^{N,x} := \sum_{\ell=1}^{N} \mathbb{I}_{\{\mathfrak{N}^{x,\ell} \geq 1\}}.$$

Notice that (77), which is well defined in the event $\{K^{N,x} \geq 1\}$, can be put in the form (76), with V_m being updated every time the state x is visited for the first time in the trajectory.

Example 3. Suppose that $\mathcal{X} = \{0, 1\}$, $r(0, a) = 0$, $r(1, a) = 1$ for all $a \in A$. Let $\pi \in \Pi_{\text{stat}}$ and suppose that $p(1|1, \pi(1)) = 1 - p$, for some $p \in]0, 1[$, and $p(1|0, \pi(0)) = 1$. Notice that π is proper with absorbing state $\hat{x} = 0$. We have

$$V_\pi(1) = \mathbb{E}(\tau^\pi) \quad \text{where } \tau^\pi \text{ follows a geometric distribution of parameter } p$$

i.e. $V_\pi(1) = 1/p$. Consider a simulated trajectory of $(X_k^\pi)_{k\in\mathbb{N}}$ having the form $X_0^\pi = 1$, $X_1^\pi = 1, \ldots, X_{\tau^\pi-1}^\pi = 1$, $X_{\tau^\pi}^\pi = 0$. For this trajectory, the every-visit Monte Carlo estimate takes the form

$$V_1^\pi(1) = \frac{\sum_{k=0}^{\tau^\pi-1}(\tau^\pi - k)}{\tau^\pi} = \tau^\pi - (\tau^\pi - 1)/2 = \frac{\tau^\pi + 1}{2}$$

and hence

$$\mathbb{E}\left(V_1^\pi(1)\right) = \frac{1}{2} + \frac{1}{2p} \neq \frac{1}{p},$$

which shows that the every-visit estimator is biased. On the other hand, for the same trajectory, the first-visit estimator of $V_\pi(1)$ is given by

$$V_1^\pi(1) = \tau^\pi,$$

which is unbiased.

The following result shows that, in general, the first-visit estimator of $V^\pi(x)$ is unbiased.

Proposition 5. *Let $x \in \mathcal{X} \setminus \{\hat{x}\}$ and let π be a deterministic stationary policy. Assume that π is proper and that $\mathbb{P}(\mathfrak{N}^x \geq 1) > 0$. Then the first-visit Monte Carlo estimator (77) is an unbiased estimator of $V^\pi(x)$ in the following sense:*

$$\mathbb{E}\left(V_N^\pi(x) | K^{N,x} \geq 1\right) = V^\pi(x). \tag{78}$$

Proof. For notational convenience, in what follows we write K for $K^{N,x}$ and, for $\ell = 1, \ldots, N$, we write X^ℓ for $X^{\pi,\ell}$, $J^\ell(X_{T_1^\ell})$ for $J^\ell\left(X^{\pi,\ell}_{T_1^\ell}, \ldots, X^{\pi}_{\tau^{\pi,\ell}-1}\right)$, and $\mathfrak{N}^\ell$ for $\mathfrak{N}^{x,\ell}$. First, notice that

$$\begin{array}{rcl} \mathbb{E}\left(V_N^\pi(x)\mathbb{I}_{K\geq 1}\right) & = & \mathbb{E}\left(V_N^\pi(x)\sum_{k=1}^N \mathbb{I}_{\{K=k\}}\right) \\ & = & \sum_{k=1}^N \mathbb{E}\left(V_N^\pi(x)\mathbb{I}_{\{K=k\}}\right). \end{array} \tag{79}$$

For every $\ell = 1, \ldots, N$, let us define $K^{-\ell} = \displaystyle\sum_{m=1,\, m\neq\ell}^{N} \mathbb{I}_{\{\mathfrak{N}^m \geq 1\}}$. Since the trajectories $\{(X_k^{\pi,m})_{k\in\mathbb{N}} \,|\, m \in \{1, \ldots, N\}\}$ are i.i.d., we have that $\mu := \mathbb{P}(\mathfrak{N}^m \geq 1)$ is independent of

$m \in \{1, \ldots, N\}$ and, hence, $K^{-\ell} \sim B(N-1, \mu)$. Consequently, for any $k = 1, \ldots, N$,

$$
\begin{aligned}
\mathbb{E}\left(V_N^{\pi}(x)\mathbb{I}_{\{K=k\}}\right) &= \sum_{\ell=1}^{N} \tfrac{1}{k}\mathbb{E}\left(J^{\ell}\left(X^{\ell}_{T_1^{\ell}}\right)\mathbb{I}_{\{\mathfrak{N}^{\ell}\geq 1\}}\mathbb{I}_{\{K=k\}}\right) \\
&= \tfrac{1}{k}\sum_{\ell=1}^{N}\mathbb{E}\left(J^{\ell}\left(X^{\ell}_{T_1^{\ell}}\right)\mathbb{I}_{\{\mathfrak{N}^{\ell}\geq 1\}}\mathbb{I}_{\{K^{-\ell}=k-1\}}\right) \\
&= \tfrac{N}{k}\mathbb{E}\left(J^{1}\left(X^{1}_{T_1^{1}}\right)\mathbb{I}_{\{\mathfrak{N}^{1}\geq 1\}}\right)\mathbb{P}\left(K^{-1}=k-1\right) \\
&= \tfrac{N}{k}\mathbb{E}\left(\mathbb{E}\left(J^{1}\left(X^{1}_{T_1^{1}}\right)|\mathcal{F}_{T_1^{1}}\right)\mathbb{I}_{\{\mathfrak{N}^{1}\geq 1\}}\right)\mathbb{P}\left(K^{-1}=k-1\right) \\
&= V^{\pi}(x)\tfrac{N}{k}\mu\tbinom{N-1}{k-1}\mu^{k-1}(1-\mu)^{N-k} \\
&= V^{\pi}(x)\tbinom{N}{k}\mu^{k}(1-\mu)^{N-k},
\end{aligned}
$$

where, to deduce fifth inequality from the fourth one, we have used the strong Markov property to infer

that $\mathbb{E}\left(J^1\left(X_{T_1^1}\right)|\mathcal{F}_{T_1^1}\right)\mathbb{I}_{\{\mathfrak{N}^1\geq 1\}} = V^\pi(x)\mathbb{I}_{\{\mathfrak{N}^1\geq 1\}}$. Finally, (79) implies that

$$\mathbb{E}\left(V_N^\pi(x)\mathbb{I}_{K\geq 1}\right) = V^\pi(x)\sum_{k=1}^{N}\binom{N}{k}\mu^k(1-\mu)^{N-k} = V^\pi(x)(1-(1-\mu)^N) = V^\pi(x)\mathbb{P}(K\geq 1),$$

which yields (78). □

The next result shows that both the every- and first-visit Monte Carlo are consistent estimators of V^π.

Proposition 6. *Let $x \in \mathcal{X}\setminus\{\hat{x}\}$ and let π be a deterministic stationary policy. Assume that π is proper and that $\mathbb{P}(\mathfrak{N}^x \geq 1) > 0$. Then both* (75) *and* (77) *are consistent estimators of $V^\pi(x)$.*

Proof. We begin with estimator (77) and, in addition to the notations in the proof of Proposition 5, we write X for X^π, $\mathfrak{N}$ for $\mathfrak{N}^x$, and τ for $\tau^{x,\pi}$. First, from Proposition 1 we get

$$\mathbb{E}\left(\left|J\left(X_{T_1^1}\right)\mathbb{I}_{\mathfrak{N}\geq 1}\right|\right) \leq \|r\|_\infty\mathbb{E}(|\tau|) < \infty$$

and hence, by (77), the strong law of large numbers, and the strong Markov property, we have

$$\begin{aligned} V_N^\pi(x) &\underset{N\to\infty}{\longrightarrow} \frac{\mathbb{E}\Big(J\Big(X_{T_1}\Big)\mathbb{I}_{\mathfrak{N}\geq 1}\Big)}{\mathbb{P}(\mathfrak{N}\geq 1)} \\ &= \frac{\mathbb{E}\left(\mathbb{E}\left(J\Big(X_{T_1}\Big)\middle|\mathcal{F}_{T_1}\right)\mathbb{I}_{\mathfrak{N}\geq 1}\right)}{\mathbb{P}(\mathfrak{N}\geq 1)} \\ &= \frac{\mathbb{E}\Big(V^\pi(x)\mathbb{I}_{\mathfrak{N}\geq 1}\Big)}{\mathbb{P}(\mathfrak{N}\geq 1)} \\ &= V^\pi(x), \end{aligned}$$

which shows that (77) is consistent with $V^\pi(x)$. Suppose now that the estimator $V_N^\pi(x)$ is given by (75). Notice that the strong Markov property implies that

$$\begin{aligned}
\mathbb{E}\left(\left|J\left(X_{T_k}\right)\right|\left|\mathfrak{N}\geq k\right.\right) &= \frac{\mathbb{E}\left(\left|J\left(X_{T_k}\right)\right|\mathbb{I}_{\mathfrak{N}\geq k}\right)}{\mathbb{P}(\mathfrak{N}\geq k)} \\
&= \frac{\mathbb{E}\left(\mathbb{E}\left(\left|J\left(X_{T_k}\right)\right|\mathbb{I}_{\mathfrak{N}\geq k}\middle|\mathcal{F}_{T_k}\right)\right)}{\mathbb{P}(\mathfrak{N}\geq k)} \\
&= \frac{\mathbb{E}\left(\mathbb{E}\left(\left|J\left(X_{T_k}\right)\right|\middle|\mathcal{F}_{T_k}\right)\mathbb{I}_{\mathfrak{N}\geq k}\right)}{\mathbb{P}(\mathfrak{N}\geq k)} \\
&= \mathbb{E}\left(\left|J\left(x\right)\right|\right).
\end{aligned}$$

Therefore, since $N \leq \tau^{\pi}$, we have $\mathbb{E}(\mathfrak{N}) \leq \mathbb{E}(\tau^{\pi}) < +\infty$, and hence Lemma 3 yields

$$\begin{aligned}
\mathbb{E}\left(\left|\left(\sum_{k=1}^{\mathfrak{N}} J\left(X_{T_k}\right)\right)\mathbb{I}_{\mathfrak{N}\geq 1}\right|\right) &\leq \mathbb{E}\left(\left(\sum_{k=1}^{\mathfrak{N}}\left|J\left(X_{T_k}\right)\right|\right)\mathbb{I}_{\mathfrak{N}\geq 1}\right) \\
&= \mathbb{E}(\mathfrak{N})\mathbb{E}\left(\left|J\left(x\right)\right|\right) \\
&\leq \mathbb{E}(\mathfrak{N})\|r\|_{\infty}\mathbb{E}(\tau^{\pi}) \\
&< \infty.
\end{aligned}$$

Therefore, we can apply the strong law of large numbers (Theorem 11) to obtain that

$$V_N^\pi \underset{N\to\infty}{\longrightarrow} \frac{\mathbb{E}\left(\left(\sum_{k=1}^{\mathfrak{N}} J\left(X_{T_k}\right)\right)\mathbb{I}_{\mathfrak{N}\geq 1}\right)}{\mathbb{E}\left(\mathfrak{N}\right)}. \tag{80}$$

Arguing as in (80), for every $k \in \mathbb{N}$ we have that

$$\mathbb{E}\left(J\left(X_{T_k}\right)\middle|\mathfrak{N}\geq k\right) = V^\pi(x).$$

Thus, it follows from Lemma 3 that

$$\mathbb{E}\left(\left(\sum_{k=1}^{\mathfrak{N}} J\left(X_{T_k}\right)\right)\mathbb{I}_{\mathfrak{N}\geq 1}\right) = \mathbb{E}(\mathfrak{N})V^\pi(x),$$

which, combined with (80), yields the consistency of V_N^π. □

Remark 14. Regarding the comparison of both estimators, even if the every-visit estimator is, in general, biased, it has the positive feature of having a mean quadratic error smaller than the one of the first-visit estimator (which is unbiased) when one

trajectory is considered. Let us, for instance, consider the model in Example 3. In this case, the mean quadratic error of the every-visit estimator is given by

$$\mathbb{E}\left(\left(V_1^\pi(1)-\frac{1}{p}\right)^2\right)=\mathbb{E}\left(V_1^\pi(1)^2\right)-\frac{2}{p}\mathbb{E}\left(V_1^\pi(1)\right)+\frac{1}{p^2}.$$

Since

$$\mathbb{E}\left(V_1^\pi(1)^2\right)=\frac{1}{4}\left(\mathbb{E}((\tau^\pi)^2)+2\mathbb{E}(\tau^\pi)+1\right)=\frac{1}{4p^2}\left(p^2+p+2\right),$$

we get

$$\begin{aligned}\mathbb{E}\left(\left(V_1^\pi(1)-\tfrac{1}{p}\right)^2\right) &= \tfrac{1}{4p^2}\left(p^2+p+2\right)-\tfrac{2}{p}\tfrac{(p+1)}{2p}+\tfrac{1}{p^2}\\ &= \tfrac{1}{2p^2}-\tfrac{3}{4p}+\tfrac{1}{4}.\end{aligned}$$

On the other hand, the mean quadratic error of the first-visit estimator is given by

$$\mathbb{E}\left(\left(V_1^\pi(1) - \frac{1}{p}\right)^2\right) = \mathsf{Var}(\tau^\pi) = \frac{1-p}{p^2}.$$

Since $p(p+1) \leq 2$, the mean quadratic error of the every-visit Monte Carlo estimator of $V^\pi(1)$ is less or equal than the corresponding error for the first-visit Monte Carlo estimator.

On the other hand, for large, but finite, state space, most of the states, if they are visited, they are only visited once. In this case, the every- and first-visit method yield to similar performances.

References

[1] S. Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. Fund. Math., 3:133–181, 1922.

[2] D. Bertsekas and J.N. Tsitsiklis. Neuro-Dynamic Programming. Athena Scientific, 1996.

[3] D. P. Bertsekas. Dynamic programming and optimal control. Vol. I. Athena Scientific, Belmont, MA, third edition, 2005.

[4] A.N. Shiryaev. Probability. 2, volume 95 of Graduate Texts in Mathematics. Springer, New York, 2019.

[5] J.B. Walsh. Knowing the odds, volume 139 of Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2012. An introduction to probability.

[6] E. Zeidler. Nonlinear functional analysis and its applications. I. Springer-Verlag, New York, 1986. Fixed-point theorems, Translated from the German by Peter R. Wadsack.