MARKOV DECISION PROCESSES

Consider an agent $\mathfrak{a}$, in an environment $\mathcal{E}$, restricted to the state-space $\mathfrak{X}$, with an available action-space $\mathcal{A}$, and securing rewards from the reward-space $\mathcal{R}$.

Mathematically, we have:

$$\mathfrak{X}, \mathcal{A} \in \mathrm{Ob}(\mathcal{S}et), \; \mathcal{E} = \mathfrak{X} \times \mathcal{A}, \quad \mathcal{R} \in \mathbb{R}^{\dim(\mathcal{E})}$$

Let the transition-probabilities be written as:

$$P_{x,y}[a] := \mathbb{P}[Y_n = y \mid X_n = x, A_n = a]$$

Def^n: A Markov Decision Process $\mathcal{M}$, is described by the 4-tuple $(\mathfrak{X}, \mathcal{A}, P, \mathcal{R})$.

- At $t=n$, let the agent register state $x_n \in \mathfrak{X}$, choose an action $a \in \mathcal{A}$, and recieve a probabilistic reward $r_n \in \mathbb{R}$, s.t.:

$$\mathbb{E}[r_n](x_n, a_n) =: \rho_{x_n}(a_n)$$

Let the discounting factor be $\gamma \in]0,1[$.

- If $(R_t)_{t \in \mathbb{N} \cup \{0\}}$ describes the reward process and $(G_t)_{t \in \mathbb{N} \cup \{0\}}$ " " cumulative " ", then,

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \infty \\ &= R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \cdots \infty) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$

<u>Def^n</u>: Let $x \in \mathfrak{X}$. A policy function $\pi : \mathfrak{X} \to \mathcal{A}$ is defined as the solution for $\mathcal{M}$.

Def^n:

(i) For a given policy π, the value of state x, under π, written $V^\pi : \mathfrak{X} \to \mathbb{R}$, is defined as:

$$V^\pi(x) = \mathbb{E}[G_t \mid S_t = x]$$
$$= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = x\right]$$

- $V^\pi(x)$ is the same as the expected reward at x, ~~plus~~ plus the discounted V^π at next step:

$$V^\pi(x) = \rho_x(\pi(x)) + \gamma \sum_{y \in \mathfrak{X}(x)} P_{x,y}[\pi(x)]\, V^\pi(y)$$

(ii) For a given π, Q-value of $(x,a) \in \mathfrak{X}$, is defined as:

$$Q^\pi(x,a) = \rho_x(a) + \gamma \sum_{y \in \mathfrak{X}(x)} P_{x,y}[\pi(x)]\, V^\pi(y)$$

TASK (Optimization problem):

Theorem [Bellman & Dreyfus, '62]: For a M.D.P. $\mathcal{M}$, described with, as above, $\exists\, \pi^*$ (optimal policy), s.t.:

$$V^* := V^{\pi^*}(x) = \max_{a \in \mathcal{A}(x)} \left\{ \rho_x(a) + \gamma \sum_{y \in \mathcal{X}(x)} P_{x,y}[a]\, V^{\pi^*}(y) \right\} \quad \text{---} (\mathcal{B}\mathcal{E})$$

Objective: Determine Q-values for an optimal policy:

$$Q^*(x,a) := Q^{\pi^*}(x,a), \qquad \forall x \in \mathcal{X},\ \forall a \in \mathcal{A}$$

Then, we have: $$V^*(x) = \max_{a \in \mathcal{A}(x)} \{ Q^*(x,a) \},$$

corresponding to an optimal policy $\pi^*(x)$.

Q-learning ALGORITHM :

Def^n: Consider a M.D.P. described as above. We define the Q-learning, as the sequence $(Q_n(\cdot,\cdot) : \mathcal{X} \times \mathcal{A} \to \mathbb{R})_{n \in \mathbb{N} \cup \{0\}}$:

$$Q_n(x,a) := \begin{cases} (1-\alpha_n) Q_{n-1}(x,a) + \alpha_n (r_n + \gamma V_{n-1}(y_n)), & \text{if } (x,a) = (x_n, a_n) \\ Q_{n-1}(x,a), & \text{otherwise} \end{cases} \quad \text{--- (1)}$$

Here, (α_n), (γ_n) are, resp., the learning-rate sequence and the reward-sequence, and $V_{n-1}(y)$ is :

$$V_{n-1}(y) = \max_{b \in \mathcal{A}} \{ Q_{n-1}(y,b) \} \quad \text{--- (2)}$$

- Not^n: Define $n^i(x,a)$: index of i^{th} time, a is tried in x.
  Define $I := \{ n^i : i \in \mathbb{N} \}$

Theorem ($\mathcal{C}_e$) [convergence theorem]:

Given bdd rewards, $|r_n| \leq R$, and learning rates $0 \leq \alpha_n < 1$, with $(\alpha_n)_{n\in I} \notin \ell^1(I)$, and $(\alpha_n)_{n\in I} \in \ell^2(I)$.

$$\Downarrow \qquad\qquad\qquad\qquad \Downarrow$$

$$\left[\sum_{i=1}^{\infty} \alpha_{n^i(x,a)} = +\infty\right] \qquad \left[\sum_{i=1}^{\infty} \alpha^2_{n^i(x,a)} < +\infty\right]$$

Then, we have:

$$\boxed{Q_n(x,a) \xrightarrow[n\to+\infty]{} Q^*(x,a)}, \quad \text{w.P. } 1.$$

- Strategy of Proof-$\mathcal{C}_e$:

  Step 1: Action-Replay-Process, $\mathbb{A}$, description

  Step 2: • lemma A: Q_n optimal for $\mathbb{A}$

  • lemma B1–B3: $\left\{\begin{array}{l}\text{preparatory lemmas, and,}\\ \mathbb{A} \longrightarrow \text{real process}\end{array}\right.$

  Step 3: Use lemmas A–B3, to prove $\mathcal{C}_e$.

Action-Replay-Process [A]

Consider the state-space $\{<x, n> \mid x \in \mathcal{X}\}$, action-space $\mathcal{A}$.
Let the discount factor $\gamma \in]0,1[$ for A, be the same as well.
Recall n^i:

$$n^i := n^i(x,a) : \begin{cases} \text{index of the } i^{th}\text{-time, action a} \\ \text{was tried at space } x \end{cases}$$

Also, define i_* as:

$$i_* := \begin{cases} \text{argmax}_i \{n^i < n\}, & \text{if } (x,a) \text{ expected before 'n'} \\ 0, & \text{otherwise} \end{cases}$$

$\Rightarrow n^{i_*}$ is the last time before episode 'n', that (x,a) was expected

Now if $i_* = 0$, then the reward is set as $Q_0(x, a)$

$\Rightarrow$ $\mathcal{A}$ absorbs

Otherwise,

let i_e be the index of the episode replayed as :

$$i_e = \begin{cases} i_* , & \text{with prob. } \alpha_{n^{i_*}} \\ i_* - 1 , & \text{with prob. } (1-\alpha_{n^{i_*}})\, \alpha_{n^{i_*-1}} \\ i_* - 2 , & \text{with prob. } (1-\alpha_{n^{i_*}})(1-\alpha_{n^{i_*-1}})\, \alpha_{n^{i_*-2}} \\ \vdots & \\ 0 , & \text{with prob. } \prod_{i=1}^{i_*} (1-\alpha_{n^i}) \end{cases}$$

Lemma A : [Q_n optimal for $\mathbb{A}$]

$Q_n(x,a)$ are the optimal action-values for $\mathbb{A}$, with state $\langle x,n \rangle$ and action a :

$$Q_n(x,a) = Q^*_{\mathbb{A}}(\langle x,n \rangle, a), \quad \forall (a,x) \in \mathcal{E}, \ \forall n \geq 0$$

Proof. From the construction of $\mathbb{A}$: $Q_0(x,a) = Q^*_{\mathbb{A}}(\langle x, n \rangle, a)$

Hence, the theorem holds for $n=0$.
Suppose Q_{n-1} - values (by Q-learning) are optimal for $\mathbb{A}$:

$$Q_{n-1}(x,a) = Q^*_{\mathbb{A}}(\langle x, n-1 \rangle, a), \qquad \forall (a,x) \in \mathcal{A} \times \mathcal{E}$$

[Induction Hypothesis]

$$\Rightarrow \quad V^*(\langle x, n-1 \rangle) = V_{n-1}(x)$$
$$= \max_{a \in \mathcal{A}(x)} \{ Q_{n-1}(x,a) \}$$

Case (i): $(x,a) \neq (x_n, a_n)$

$$\Rightarrow \quad Q_n(x,a) = Q_{n-1}(x,a)$$
$$= Q^*_{\mathbb{A}}(\langle x, n-1\rangle, a)$$
$$= Q^*_{\mathbb{A}}(\langle x, n\rangle, a)$$

Case ii: $(x,a) = (x_n, a_n)$

$$Q^*_{\mathbb{A}}(\langle x_n, n\rangle, a_n) = (1-\alpha_n)Q^*_{\mathbb{A}}(\langle x_n, n-1\rangle, a_n) + \alpha_n\left(r_n + \gamma V^*(\langle y_n, n-1\rangle)\right)$$
$$= (1-\alpha_n)\, Q_{n-1}(x_n, a_n) + \alpha_n(r_n + \gamma V_{n-1}(y_n))$$
$$= Q_n(x_n, a_n)$$

~~Hence~~, from the induction hypothesis and Q_n-iteration formula (1).

Hence, $\quad Q_n(x,a) = Q^*_{\mathbb{A}}(\langle x, n\rangle, a), \quad \forall (a,x) \in \mathcal{A} \times \mathcal{S}$

$\blacksquare$

Lemma B.1: [Discounting infinite sequence]

Consider a finite Markov Process, with discounting factor γ, bdd-rewards ($|r_n| < R$), transition-probabilities $P_{x,y}[a]$. Let $\mathbb{X}_s = (x_0, x_1, \dots, x_s)$ be the fixed s-steps. Then,

$$\forall \pi : \quad \left| V^{\pi}(\mathbb{X}_s) - V^{\pi}(x_0) \right| \xrightarrow[s \to +\infty]{} 0$$

Proof.

Ignoring the value of $(s+1)^{th}$-state, incurs the penalty :

$$\delta := \gamma^s \sum_{y_{s+1} \in \xi(y_s)} P_{y_s, y_{s+1}} [a_s] \, V^{\pi}(y_{s+1})$$

Since $|r_n| < R, \quad \forall n \in \mathbb{N} \cup \{0\}$:

$$V^{\pi}(y_{s+1}) < R + \gamma R + \gamma^2 R + \cdots \infty$$
$$= R\left(\frac{1}{1-\gamma}\right) = \frac{R}{1-\gamma}$$

$$\Rightarrow \quad V^{\pi}(y_{s+1}) < \frac{R}{1-\gamma}$$

$$\therefore \quad |\delta| < \gamma^s \sum_{y_{s+1} \in \mathcal{S}(y_s)} P_{y_s, y_{s+1}}[a_s] \left(\frac{R}{1-\gamma}\right)$$
$$= \gamma^s \left(\frac{R}{1-\gamma}\right)(1) = \frac{\gamma^s R}{1-\gamma}$$

$$\Rightarrow |\delta| < \frac{\gamma^s R}{1-\gamma} \xrightarrow[s \to +\infty]{} 0$$

$\blacksquare$

Lemma B.2 : [Rewards & trans^n probabilities converge]

w.P. 1, $\quad P^{(n)}_{A(x,y)}[a] \xrightarrow[n \to +\infty]{} P^{(n)}_{(x,y)}[a]$

and $\quad \rho^{(n)}_{A_x}[a] \xrightarrow[n \to +\infty]{} \rho^{(n)}_{x}[a]$, $\quad \forall a \in \mathcal{A}$

Proof.

By the theorem of Kushner & Clark, 1978, "if $(X_n)_{n \in \mathbb{N} \cup \{0\}}$:

$$X_{n+1} = X_n + \beta_n (\xi_n - X_n),$$

with $0 \leq \beta_n < 1$, $(\beta_n) \notin \ell^1$, $(\beta_n) \in \ell^2$, ξ_n bdd, $\mathbb{E}[\xi_n] = \Xi$,
then,

$$X_n \xrightarrow[n \to +\infty]{} \Xi, \quad w.\mathbb{P}.1.$$

In our case:

$$\rho_{\langle x, n^{i+1}\rangle}(a) = \rho_{\langle x, n^i\rangle}(a) + \alpha_{n^{i+1}}\left(r_{n^{i+1}} - \rho_{\langle x, n^i\rangle}(a)\right),$$

and,

$$P_{A\langle x,y\rangle}^{(n^{i+1})}[a] = P_{A\langle x,y\rangle}^{(n^i)}[a] + \alpha_{n^{i+1}}\left(\mathbb{1}_{\{y_n = y\}} - P_{A\langle x,y\rangle}^{(n^i)}[a]\right),$$

such that,

$$\mathbb{E}[r_n] = \rho_x(a), \quad \mathbb{E}[\mathbb{1}_{y_x = y}] = P_{x,y}[a]$$

Hence, the theorem applies to both, and we have proved our lemma.

$\blacksquare$

Lemma B.3 : [close reward & probabilities $\Rightarrow$ close values]

Consider an s-step Markov Chain, formed according to the probabilities $\left(P^{(n)}_{(x_n, x_{n+1})}[a_n]\right)_{n\in\{1,2,\dots s\}}$. Let $\eta>0$ be given.

Let $\overline{Q}(x,a_1,\dots,a_s)$: expected reward for the real process
$\overline{Q}'(x,a_1,\dots,a_s)$: " " " " Markov Chain

If we have: $\left|P^j[a] - P_{x,y}[a]\right| < \frac{\eta}{R}$, $\forall a,x,y$, $\forall i\in\{1,\dots,s\}$

and, $\left|\int^{(i)}_x(a) - \int_x(a)\right| < \eta$, $\forall a,x$, $\forall i\in\{1,\dots,s\}$

then, $\left|\overline{Q}'(x,a_1,\dots,a_s) - \overline{Q}(x,a_1,\dots,a_s)\right| < \frac{s(s+1)}{2}\eta$

Proof. We have: $\bar{Q}(x, a_1, a_2) = \rho_x(a_1) + \gamma \sum_{y \in \mathcal{L}(x)} P_{x,y} \rho_y(a_2)$

and, $\bar{Q}'(x, a_1, a_2) = \rho_x^{(1)}(a_1) + \gamma \sum_{y \in \mathcal{L}(x)} P_{x,y}^{(1)} \rho_y^{(2)}(a_2)$

$$
\begin{aligned}
\Rightarrow |\bar{Q}'(x, a_1, a_2) - \bar{Q}(x, a_1, a_2)| &\leq |\rho_x^{(1)}(a_1) - \rho_x(a_1)| \\
&\quad + \Big|\gamma \sum_{y \in \mathcal{L}(x)} P_{x,y}^{(1)} (\rho_y^{(2)} - \rho_y)(a_2)\Big| \\
&\quad + \Big|\gamma \sum_{y \in \mathcal{L}(x)} \rho_y(a_2)(P_{x,y}^{(1)} - P_{x,y})\Big| \\
&\leq \eta + \gamma \cdot \eta \cdot 1 + \gamma \cdot R \cdot \frac{\eta}{R} \\
&\underset{(\because \gamma \leq 1)}{\leq} \eta + \eta + \eta = 3\eta
\end{aligned}
$$

Similarly, for the s-step chain, we get $\frac{s(s+1)}{2}$ terms and we get the result. ∎

Proof of $\mathcal{C}_e^{\infty}$:

Putting the above proved lemmas together, we get the following:

From lemma B.2: $$P^{(n)}_{A_{(x,y)}}[a] \xrightarrow[n\to+\infty]{} P^{(n)}_{(x,y)}[a]$$

$$\rho^{(n)}_{A_x}(a) \xrightarrow[n\to+\infty]{} \rho^{(n)}_{x}(a), \quad \forall a \in \mathcal{A}$$

From lemma B.3:

$$Q^*_A(x, a_1, \dots, a_s) \longrightarrow Q^*(x, a_1, \dots, a_s)$$

$$\Rightarrow \quad Q^*_A(x,a) \longrightarrow Q^*(x,a)$$

($\because$ lemma B.1)

From lemma A: $$Q_n(x,a) \longrightarrow Q^*_A(x,a)$$

$$\therefore \quad \boxed{Q_n(x,a) \longrightarrow Q^*(x,a)} \quad \blacksquare$$