Processus de décisions markoviens.

Un PDM est défini par (S, A, T, R, γ) où

- S est un ensemble fini d'états
  on notera $(s_1, \dots, s_n) = S$
- A est un ensemble fini d'actions
  on notera $(a_1, \dots, a_m) = A$
- On notera S_t l'état à l'instant $t \in \mathbb{N}$
  A_t l'action prise à l'instant $t \in \mathbb{N}$
- T est la matrice de transition

$$T_{ss'}^{a} = p\left(S_{t+1} = s \mid S_t = s, A_t = a\right)$$

probabité de se trouver en s' sachant qu'on a pris l'action a dans l'état s.

- R est la matrice de récompense

$$R_s^a = \mathbb{E}\left(r_{t+1} \mid S_t = s, A_t = a\right)$$

espérance d'obtenir la récompense $r_{t+1} \in \mathbb{R}$ après avoir pris l'action a dans l'état s.

- γ est le taux d'escompte
  Quelle est la valeur aujourd'hui d'obtenir 1 unité demain.

N.B. On trouve d'autres formulation pour modéliser la récompense :
- R_s : récompense quand on arrive dans l'état s
- $R_{ss'}^a$: récompense obtenue après avoir pris l'action a dans l'état s et être arrivé en s'

→ on peut ensuite reformuler et écrire des résultats équivalents.

def : La fonction de valeur pour la politique π est

$$v_\pi(s) = \mathbb{E}\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \;\middle|\; S_t = s, \pi\right)$$

def : La fonction de valeur optimale v^* est $\max_\pi v_\pi$.

prop : • v_π satisfait l'équation de Bellman

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) R_s^a + \gamma \sum_{s' \in S} \sum_{a \in A} \pi(a|s) T_{ss'}^a v_\pi(s')$$

• v^* satisfait l'équation de programmation dynamique

$$v^*(s) = \max_{a \in A} R_s^a + \gamma \sum_{s' \in S} T_{ss'}^a v^*(s')$$

$$\begin{aligned}
v_\pi(s) &= \mathbb{E}\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \;\middle|\; S_t = s, \pi\right) \\
&= \mathbb{E}\left(r_{t+1} + \sum_{k=1}^{\infty} \gamma^k r_{t+k+1} \;\middle|\; S_t = s, \pi\right) \\
&= \mathbb{E}\left(r_{t+1} \mid S_t = s, \pi\right) + \mathbb{E}\left(\sum_{k=1}^{\infty} \gamma^k r_{t+k+1} \;\middle|\; S_t = s, \pi\right) \\
&= \sum_{a \in A} \pi(a|s) R_s^a + \gamma \sum_{a \in A} \pi(a|s) \sum_{s' \in S} T_{ss'}^a \sum_{k=1}^{\infty} \underbrace{\mathbb{E}\left(\gamma^{k-1} r_{t+k+1} \mid S_{t+1} = s', S_t = s, A_{t+1} = a\right)}_{\overset{\text{markov}}{=} \mathbb{E}\left(\gamma^{k-1} r_{t+k+1} \mid S_{t+1} = s', A_{t+1} = a\right)} \\
&= \sum_{a \in A} \pi(a|s) R_s^a + \gamma \sum_{a \in A} \sum_{s' \in S} T_{ss'}^a \underbrace{\sum_{k=0}^{\infty} \mathbb{E}\left(\gamma^k r_{t+1+k+1} \mid S_{t+1} = s', A_{t+1} = a\right)}_{v_\pi(s')} \\
&= \sum_{a \in A} \pi(a|s) R_s^a + \gamma \sum_{s' \in S} \sum_{a \in A} T_{ss'}^a v_\pi(s')
\end{aligned}$$

$\square$

$$\begin{aligned}
v^*(s) &= \max_\pi \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \;\middle|\; S_t = s, \pi\right) \\
&= \max_{a, \pi'} R_s^a + \gamma \sum_{s' \in S} T_{ss'}^a v_{\pi'}(s') \\
&= \max_a \left[R_s^a + \max_{\pi'} \gamma \sum_{s' \in S} T_{ss'}^a v_{\pi'}(s')\right] \\
&= \max_a \left[R_s^a + \gamma \sum_{s' \in S} T_{ss'}^a \max_{\pi'} v_{\pi'}(s')\right] \\
&= \max_a \left[R_s^a + \gamma \sum_{s' \in S} T_{ss'}^a v^*(s')\right]
\end{aligned}$$

$\square$

On peut ainsi définir une politique optimale $\pi^* = \underset{\pi}{\operatorname{argmax}}\, v_\pi$

On est maintenant prêt à définir deux opérateurs.

- $T^\pi : \mathbb{R}^n \to \mathbb{R}^n$
  $W \mapsto T^\pi W(s) = \sum_{s' \in S} \pi(a|s) T_{ss'}^a W(s')$
- $T : \mathbb{R}^n \to \mathbb{R}^n$
  $W \mapsto T W(s) = \max_{a \in A} \left[R_s^a + \gamma \sum_{s' \in S} T_{ss'}^a W(s')\right]$

Ceci nous permet de revisiter v_π et v^* sous forme matricielle

- $V^\pi = \left(I - \gamma T^\pi\right)^{-1} r^\pi$ où $\begin{cases} T^\pi(s, s') = \sum_{a \in A} \pi(a|s) T_{ss'}^a \\ r^\pi(s) = \sum_{a \in A} \pi(a|s) R_s^a \end{cases}$
- V^π est l'unique point fixe de T^π
- V^* est l'unique point fixe de T^*
- $\pi^* = \underset{a \in A}{\arg\max} \left[R_s^a + \gamma \sum_{s' \in S} T_{ss'}^a V^*(s')\right]$ est optimale
- $\forall W \in \mathbb{R}^n \quad \lim_{k \to +\infty} (T^\pi)^k W = V^\pi$
  $\lim_{k \to \infty} T^k W = V^*$
- Opérateurs contractant en $\|\cdot\|_\infty$

$$\forall W_1, W_2 \in \mathbb{R}^n \quad \|T^\pi W_1 - T^\pi W_2\|_\infty \leq \gamma \|W_1 - W_2\|_\infty$$

$$\|T W_1 - T W_2\|_\infty \leq \gamma \|W_1 - W_2\|_\infty$$

$$\begin{aligned}
\left|TW_1(s) - TW_2(s)\right| &= \left|\max_{a \in A}\left[R_s^a + \gamma \sum_{s' \in S} T_{ss'}^a W_1(s')\right] - \max_a \left[R_s^a + \gamma \sum_{s' \in S} T_{ss'}^a W_2(s')\right]\right| \\
&\leq \gamma \max_{a \in A} \sum_{s' \in S} T_{ss'} \left|W_1(s') - W_2(s')\right| \\
&\leq \gamma \|W_1 - W_2\|_\infty
\end{aligned}$$

Démonstrations

1. On a $V^\pi = r^\pi + \gamma T^\pi V^\pi \Leftrightarrow (I - \gamma T^\pi) V^\pi = r^\pi$
   La matrice T^π est stochastique, donc ses valeurs propres sont de modules ≤ 1
   Donc les valeurs propres de $I - \gamma T^\pi$ sont de module $\geq 1 - \gamma$
   et donc $I - \gamma T^\pi$ est inversible.

2 et 3, 5 ⇒ T et T^π sont contractants.