

Rappel : équations d'optimalité
de Bellman

$s \in \mathcal{S}$

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

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

$$q^*(s, a) = \sum_{s', r} p(s', r | s, a) (r + \max_{a'} q^*(s', a'))$$
$$= \sum_{s'} p(s' | s, a) (R(s, a, s') + \max_{a'} q^*(s', a'))$$

A la recherche de la politique optimale d'une MDP

Deux stratégies:

→ Itération de valeurs (value iteration)

→ Itérations de politiques
(Policy iteration)

VI: $J_0 \rightarrow J_1 \rightarrow J_2 \rightarrow \dots$

Principe.

- Initialiser $V(s) \leftarrow 0 \quad \forall s \in \mathcal{S}, k \leftarrow 0$

- Répéter

$$V_{k+1}(s) = \sum_{a \in \mathcal{A}} \pi(s, a) \sum_{s' \in \mathcal{S}} P(s' | s, a)$$

$$[R(s, a, s') + \gamma V_k(s')]$$

$k \leftarrow k+1$

jusqu'à ce que l'écart entre V_k et V_{k+1} soit faible

- l'algorithme converge vers V^* asymptotiquement.

- On mesure l'écart entre V_k et V_{k-1} par $\|V_k - V_{k-1}\|_\infty$ où $\|\cdot\|_\infty$ est

définie par $\max_{s \in S} |V_k(s) - V_{k-1}(s)|$.

PI: Principe: $\pi_0 \rightarrow \pi_1 \rightarrow \dots$

- 1 - démarrer avec une politique quelconque
- 2 - calculer sa valeur
- 3 - construire une politique meilleure que la précédente (comment)
- 4 - retourner à l'étape 2 tant que l'on arrive à produire une politique meilleure (strictement)

Comment faire 3.?

Propriété: soit π une politique et v^π la fonction valeur associée,

$$\begin{aligned}\pi'(s) &= \underset{a \in \mathcal{A}}{\text{arg max}} q^\pi(s, a) \\ &= \underset{a \in \mathcal{A}}{\text{arg max}} \left[p(s' | s, a) \left[R(s, a, s') + \gamma v^\pi(s') \right] \right]\end{aligned}$$

$\pi' \geq \pi$ et si $\pi' = \pi$ alors c'est une politique optimale.

Algorithme d'itération sur les politiques

Input: $PDM = (S, A, P, R)$

γ

ϵ

Output: Politique

initialiser π_0

$k \leftarrow 0$

répéter

initialiser $V_0^{\pi_0}$

$i \leftarrow 0$

répéter

pour tout $s \in S$ faire

$$V_{i+1}^{\pi^k}(s) \leftarrow \sum_{s' \in S} P(s'|s, \pi^k(s)) [r(s, \pi^k(s), s') + \gamma V_i^{\pi^k}(s')]$$

$i \leftarrow i+1$

jusqu'à $\|V_i^{\pi^k} - V_{i-1}^{\pi^k}\|_{\infty} \leq \epsilon \frac{1-\gamma}{2\gamma}$

pour tout $s \in S$

$$\pi_{k+1}(s) = \arg \max_{a \in A} \sum_{s' \in S} P(s'|s, a) [r(s, a, s') + \gamma V^{\pi^k}(s')]$$

$k \leftarrow k+1$

jusqu'à $\pi_k = \pi_{k-1}$

Algorithme d'itération sur la valeur :

Input : $MDP = (\mathcal{S}, \mathcal{A}, P, r)$, γ , ϵ

Initialiser $V_0 \leftarrow 0$

$k \leftarrow 0$

répéter

pour tout $s \in \mathcal{S}$ faire

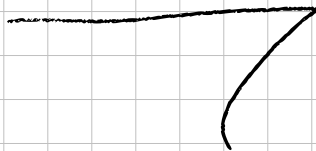
$$V_{k+1}(s) \leftarrow \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s'|s,a) \left[r(s,a,s') + \gamma V_k(s') \right]$$

$k \leftarrow k+1$

jusqu'à $\|V_k - V_{k-1}\|_\infty \leq \frac{\epsilon(1-\gamma)}{\gamma}$

pour tout $s \in \mathcal{S}$ faire

$$\pi(s) = \operatorname{argmax}_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s'|s,a) \left[r(s,a,s') + \gamma V(s') \right]$$



$$u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}; \quad v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

$$\|u-v\|_{\infty} = \max_{i=1}^n |u_i - v_i|$$

$$\|u-v\|_{\infty} = 0 \iff \max_{i=1}^n |u_i - v_i| = 0$$
$$\iff |u_i - v_i| = 0 \quad \forall i$$