Rappel : équations d'optimalité de Bellman

$\forall s \in \mathcal{S}$,

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

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

$$q^*(s, a) = \sum_{s', r} p(s', r \mid s, a) \left(r + \max_{a'} q^*(s', a') \right)$$

$$= \sum_{s'} p(s' \mid s, a) \left(R(s, a, s') + \max_{a'} q^*(s', a') \right)$$

A la recherche de la politique optimale d'une MDP

Deux stratégies :

→ Itération de valeurs (value iteration)

→ Itérations de politiques (Policy itération)

VI : $v_0 \to v_1 \to v_2 \to \dots$

Principe.

- Initialiser $V(s)$ à 0 $\forall s \in S$, $k \leftarrow 0$
- Répéter

$$V_{k+1}(s) = \sum_{a \in A} \pi(s,a) \sum_{s' \in S} p(s'|s,a) \left[R(s,a,s') + \gamma V_k(s') \right]$$

$k \leftarrow k+1$

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 : $\pi_0 \to \pi_1 \to \dots$

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.

$$\pi'(s) = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, q^\pi(s,a)$$

$$= \underset{a \in \mathcal{A}}{\operatorname{argmax}} \left[p(s'|s,a) \left[R(s,a,s') + \gamma V^\pi(s') \right] \right]$$

$\pi' \geq \pi$ 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

      $V_{i+1}^{\pi_k}(s) \leftarrow \sum_{s' \in S} p(s' | s, \pi_k(s)) \left[r(s, \pi_k(s), s') + \gamma V_i^{\pi_k}(s') \right]$

    $i \leftarrow i+1$

  jusqu'à $\| V_i^{\pi_k} - V_{i-1}^{\pi_k} \|_\infty \leq \varepsilon \left(\frac{1-\gamma}{2\gamma} \right)$

  pour tout $s \in S$

    $\pi_{k+1}(s) = \operatorname{argmax}_{a \in A} \sum_{s' \in S} p(s' | s, a) \left[r(s, a, s') + \gamma V^{\pi_k}(s') \right]$

  $k \leftarrow k+1$

jusqu'à $\pi_k = \pi_{k-1}$

Algorithme d'itération sur la valeur :

Input : $MDP = (S, A, P, r)$, γ, ε

Initialiser $V_0 \leftarrow 0$

$k \leftarrow 0$

répéter

  pour tout $s \in S$ faire

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

  $k \leftarrow k+1$

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

pour tout $s \in S$ faire

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

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

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

$\|u-v\|_\infty = 0$ ssi $\max |u_i - v_i| = 0$

ssi $|u_i - v_i| = 0 \quad \forall i$