# Aff par renforcement

## Exo 4

1) Processus markovien de décision

$S = \{s_1, s_2, s_3, s_4\}$

$A = \{\rightarrow, \leftarrow, \uparrow, \downarrow\}$

T et R

| | $\rightarrow$ | $\leftarrow$ | $\uparrow$ | $\downarrow$ |
|---|---|---|---|---|
| $s_1$ | $s_2$ 0 | $s_1$ -1 | $s_1$ -1 | $s_3$ 0 |
| $s_2$ | $s_2$ -1 | $s_1$ 0 | $s_2$ -1 | $s_4$ -0.5 |
| $s_3$ | $s_4$ 0 | $s_3$ -1 | $s_1$ 0 | $s_3$ -1 |
| $s_4$ | $s_1$ 1 | $s_1$ 1 | $s_1$ 1 | $s_1$ 1 |

2) Une stratégie est une fonction

$$\Pi : S \longrightarrow A$$

il s'agit donc d'associer une action à chaque état.

Considérons par exemple la fonction suivante

$\Pi(s_1) = \rightarrow$, $\Pi(s_3) = \leftarrow$, $\Pi(s_2) = \downarrow$ et $\Pi(s_4) = \leftarrow$

Nous souhaitons calculer $V_\pi(s)$ pour chaque état

$V_\pi(s_1) = ?$

Lorsque nous commençons en $s_1$ nous avons la suite états-actions suivante

$$s_1 \xrightarrow[-\frac{1}{2}]{0} s_2 \xrightarrow[+1]{} s_4 \xleftarrow[0]{} s_1 \xrightarrow[-\frac{1}{2}]{} s_2 \xrightarrow[+1]{} s_4 \leftarrow s_1 \cdots$$

Nous avons donc

$$V_\pi(s_1) = 0 + \gamma \times (-\tfrac{1}{2}) + \gamma^2 . 1 + \gamma^3 . 0 + \gamma^4 \times (-\tfrac{1}{2}) + \cdots$$

$$= \sum_{t=0}^{+\infty} \gamma^{3t+1} \times (-\tfrac{1}{2}) + \sum_{t=0}^{+\infty} \gamma^{3t+2} \times 1$$

$$= -\frac{\gamma}{2} \sum_{t=0}^{+\infty} (\gamma^3)^t + \gamma^2 \sum_{t=0}^{+\infty} (\gamma^3)^t$$

$$= -\frac{\gamma}{2} \frac{1}{1-\gamma^3} + \frac{\gamma^2}{1-\gamma^3}$$

$$= \frac{2\gamma^2 - \gamma}{2(1-\gamma^3)}$$

Les valeurs de $V_\pi(s_2)$ et $V_\pi(s_4)$ se calculent de la même manière

Calcul de $V_\pi(s)$

Nous avons la suite

$$s_3 \overset{-1}{\leftarrow} s_1 \overset{-1}{\leftarrow} s_3 \overset{-1}{\leftarrow} s_3 \dots$$

d'où $V_\pi(s_3) = -1 * \sum_{t=0}^{+\infty} \gamma^t = -\frac{1}{1-\gamma}$

3) Faisons tourner l'algorithme Value Iteration à la main

- Initialisation $V(s) = 0$ pour tous les états

- Etat $s_1$

action = $\rightarrow$

$Q(s_1, \rightarrow) = R(s_1, \rightarrow) + \gamma V(s_2)$

$= 0 + 0 = 0$

action = $\leftarrow$

$Q(s_1, \leftarrow) = R(s_1, \leftarrow) + \gamma V(s_1)$

$= -1 + 0 = -1$

action = $\uparrow$

$Q(s_1, \uparrow) = R(s_1, \uparrow) + \gamma V(s_1)$

$= -1 + 0 = -1$

action = $\downarrow$

$Q(s_1, \downarrow) = R(s_1, \downarrow) + \gamma V(s_3)$

$= 0 + 0$

D'où $V(s_1) = \max(0, -1, -1, 0) = 0$

...

Suite de la question : voir code