Université Claude Bernard Lyon 1
Département Informatique

4A, 2019–2020

IA

TD1. MDPs et Planification sous incertitudes.(correction)

Laëtitia Matignon

---

1 Exercice 1

Nous disposons d'un robot qui, aux instants où il doit prendre une décision, a le choix entre 3 comportements :

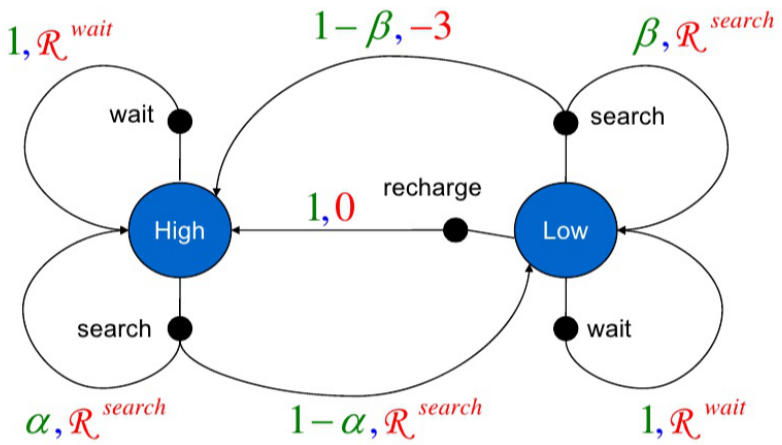
- Chercher activement une cannette à recycler.
- Attendre sur place qu'on lui amène une cannette à recycler.
- Aller à sa station de rechargement pour recharger ses batteries

Evidemment, il y a plus de chance de trouver une cannette en cherchant qu'en attendant qu'on lui amène. Cependant, chercher une cannette coute de l'énergie au robot (vide sa batterie), ce qui n'est pas le cas quand il attend.

Nous allons considérer que **l'état de la batterie du robot peut prendre deux valeurs** qui sont `à-bloc`, `presque-vide`. Quand la batterie est `presque-vide`, dépenser de l'énergie en cherchant une canette peut complètement vider la batterie, il faut alors qu'un opérateur viennent manuellement recharger la batterie, ce qui est ennuyeux et coûteux, mais **immédiat**. Enfin, si le robot décide d'aller se recharger, il y parvient sans problème et se retrouve avec une batterie gonflée `à-bloc` (immédiat, moins coûteux que l'intervention d'un opérateur).

Nous voulons que le robot recycle le plus de cannettes possibles sans faire intervenir l'opérateur de maintenance. Sa prise de décision est fonction de l'état de sa batterie.

Question 1 *Modéliser ce problème à l'aide d'un MDP représenté sous forme d'un graph. Pour toute variable introduite non précisée dans l'énoncé, vous préciserez leurs significations.*



$\alpha < 1$, $\beta < 1$, $R^{search} > R^{wait} > 0$

2 Exercice 2

Question 1 *Calculer la fonction de valeur V aux itérations 1 et 2 pour l'exemple fil rouge (labyrinthe) du cours dans le cas stochastique (cf. fin de section 4 du CM1, transparent 59).*

Voir la version corrigée du CM1.

3 Exercice 3

3.1 Rappels

L'algorithme *Value iteration* [Bellman,1957] calcule itérativement la fonction de valeur optimale V^* à partir du modèle MDP.

- Initialisation arbitraire de $V_0(s) \forall s$
- Mise à jour de $V_k(s)$ en utilisant les valeurs estimées à $k-1$ des états voisins s' :

$$\forall s \in S \quad V_k(s) \leftarrow \max_{a \in A} \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_{k-1}(s')]$$

- Répète jusqu'à convergence :
  critère d'arrêt $\max_{s \in S} |V_k(s) - V_{k+1}(s)| < \epsilon$

La politique gloutonne, notée π^g, est extraite de la fonction de valeur V en choisissant, dans chaque état s, l'action qui maximise le retour espéré :

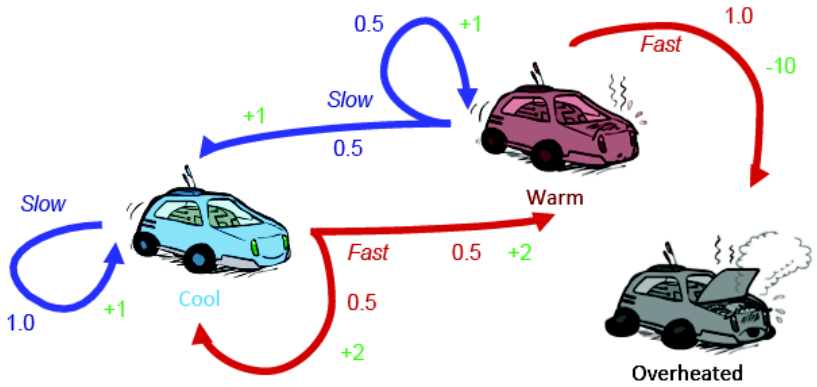
$$\forall s \in S \quad \pi^g(s) = \arg\max_{a \in A} \sum_{s' \in S} T(s,a,s')[R(s,a,s') + \gamma V(s')]$$

Ainsi, à partir la fonction de valeur optimale V^* obtenue par *Value iteration*, on peut calculer la politique optimale π^* comme la politique gloutonne sur V^* :

$$\forall s \in S \quad \pi^*(s) = \arg\max_{a \in A} \sum_{s' \in S} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$$

3.2 Exercice

Soit le MDP suivant :



Question 1 *Donnez la politique gloutonne obtenue après 2 itérations de value-iteration, en supposant qu'on utilise un facteur d'atténuation γ de 0.9 et en partant initialement (itération 0) avec des valeurs toutes égales à 0.*

Pour connaitre la politique gloutonne après 2 itérations, il faut tout d'abord calculer la fonction de valeur pour l'itération 2 en appliquant :

$$V_k(s) \leftarrow \max_{a \in A} \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_{k-1}(s')]$$

Ce qui donne :

$$\begin{aligned}
V_0(cool) &= 0 \\
V_0(warm) &= 0 \\
V_0(overheated) &= V_1(overheated) = V_2(overheated) = 0 \\
V_1(cool) &= MAX\{1 \times (1 + \gamma \times V_0(cool)), 0.5 \times (2 + \gamma \times V_0(cool)) + 0.5 \times (2 + \gamma \times V_0(warm)) \\
V_1(cool) &= 2 \\
V_1(warm) &= MAX\{0.5 \times (1 + \gamma \times V_0(warm)) + 0.5 \times (1 + \gamma \times V_0(cool)), \\
&\qquad 1 \times (-10 + \gamma \times V_0(overheated))\} \\
V_1(warm) &= 1 \\
V_2(cool) &= MAX\{1 \times (1 + \gamma \times V_1(cool)), 0.5 \times (2 + \gamma \times V_1(cool)) + 0.5 \times (2 + \gamma \times V_1(warm)) \\
V_2(cool) &= 3.35 \\
V_2(warm) &= MAX\{0.5 \times (1 + \gamma \times V_1(warm)) + 0.5 \times (1 + \gamma \times V_1(cool)), \\
&\qquad 1 \times (-10 + \gamma \times V_1(overheated))\} \\
V_2(warm) &= 2.35
\end{aligned}$$

On en déduit la politique gloutonne après 2 itérations en appliquant :

$$\pi_k(s) = \arg\max_{a \in A} \sum_{s' \in S} T(s,a,s')[R(s,a,s') + \gamma V_k(s')]$$

Ce qui donne :

$$\begin{aligned}
\pi_2(cool) &= ARGMAX\{1 \times (1 + \gamma \times V_2(cool)), 0.5 \times (2 + \gamma \times V_2(cool)) + 0.5 \times (2 + \gamma \times V_2(warm)) \\
\pi_2(cool) &= FAST \\
\pi_2(warm) &= ARGMAX\{0.5 \times (1 + \gamma \times V_2(warm)) + 0.5 \times (1 + \gamma \times V_2(cool)), \\
&\qquad 1 \times (-10 + \gamma \times V_2(overheated))\} \\
\pi_2(warm) &= SLOW
\end{aligned}$$