UCBL1 Master 2 IA,
Département Informatique

Intelligent Agents

TD2. Apprentissage par renforcement.

Laëtitia Matignon

Rappels

L'algorithme ***Monte-Carlo prediction*** met à jour la valeur d'un état s comme la moyenne des n échantillons de la valeur de s obtenus **sur tous les épisodes** :

$$V(s) = \frac{1}{n}\sum_{i=1}^{n}\{v(s_i)|s_i = s\}$$

avec $v(s_i)$ un échantillon pour la valeur de l'état s dans l'épisode i :

$$v(s_i) = r_{1,i} + \gamma r_{2,i} + \gamma^2 r_{3,i} + ...$$

L'algorithme ***Incremental Monte-Carlo prediction*** met à jour **à la fin de chaque épisode**, les états visités pendant l'épisode selon une moyenne incrémentale :

$$\underbrace{V^\pi(s)}_{nouvelle\_estimation} = \underbrace{V^\pi(s)}_{ancienne\_estimation} + \alpha(\underbrace{v(s)}_{echantillon} - \underbrace{V^\pi(s)}_{ancienne\_estimation})$$

avec $v(s)$ échantillon pour l'épisode de la valeur de s : $v(s) = r_1 + \gamma r_2 + \gamma^2 r_3 + ...$ $\alpha \in [0; 1]$ est le coefficient d'apprentissage.

L'algorithme ***TD prediction*** met à jour **à chaque étape** $< s, a = \pi(s), s', r >$ la valeur de l'état s visité selon :

$$\underbrace{V^\pi(s)}_{nouvelle\_estimation} = \underbrace{V^\pi(s)}_{ancienne\_estimation} + \alpha(\underbrace{r + \gamma V^\pi(s')}_{échantillon} - \underbrace{V^\pi(s)}_{ancienne\_estimation})$$

Dans cette équation, l'agent atteint l'état s' et reçoit la récompense r après avoir effectué l'action a dans l'état s. $\alpha \in [0; 1]$ est le coefficient d'apprentissage.

L'algorithme du ***Q-learning*** met à jour **à chaque étape** $< s, a, s', r >$ la Q-valeur du couple (s, a) selon :

$$Q(s, a) = (1 - \alpha)Q(s, a) + \alpha(r + \gamma \max_{b \in A(s')} Q(s', b))$$

Dans cette équation, l'agent atteint l'état s' et reçoit la récompense r après avoir effectué l'action a dans l'état s.

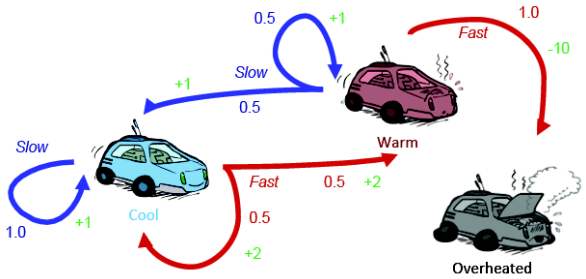
Une **politique gloutonne** est définie comme : $\forall s \in S, \pi(s) = \arg\max_{a \in A} Q(s,a)$

L'algorithme ***Approximate Q-learning*** met à jour **à chaque étape** $< s, a, s', r >$ les paramètres w_i :

$$\forall i \in [1,n] \quad w_i \leftarrow w_i + \alpha(r + \gamma \max_b Q_w(s',b) - Q_w(s,a)) f_i(s,a)$$

1 Exercice

Soit le MDP suivant :



On ne connaît ni la fonction de transition ni celle de récompense mais on observe les deux épisodes suivants :

— Episode 1 : Cool $\xrightarrow{Fast}$ (+2) Warm $\xrightarrow{Fast}$ (-10) Overheated.

— Episode 2 : Cool $\xrightarrow{Fast}$ (+2) Cool $\xrightarrow{Fast}$ (+2) Warm $\xrightarrow{Fast}$ (-10) Overheated.

où X $\xrightarrow{A}$ (R) Y signifie que l'agent a fait l'action A dans l'état X, et qu'il a fait une transition vers l'état Y et a obtenu la récompense R.

1.1 AR passif

On suppose que ces deux épisodes ont été réalisés en suivant la politique π suivante :

— $\pi(Cool) = Fast$

— $\pi(Warm) = Fast$

On utilise comme paramètres $\gamma = 1$ et $\alpha = 0.1$.

Question 1 *Calculez $V^\pi(Cool)$ et $V^\pi(Warm)$ avec l'algorithme Monte Carlo Prediction en version every visit. Vous détaillerez vos calculs.*

Question 2 *Calculez $V^\pi(Cool)$ et $V^\pi(Warm)$ avec l'algorithme Incremental Monte Carlo Prediction en version every visit. On suppose que la fonction de valeur V^π est initialisée à 0 $\forall s \in S$. Vous détaillerez vos calculs.*

Question 3 *Calculez $V^\pi(Cool)$ et $V^\pi(Warm)$ avec l'algorithme TD Prediction. On suppose que la fonction de valeur V^π est initialisée à 0 $\forall s \in S$. Vous détaillerez vos calculs.*

1.2 AR actif

On suppose que la fonction de Q-valeur est initialisée à 0 $\forall s \in S, a \in A$.

Question 4 *Quelle est la politique gloutonne initiale ?*

Question 5 *Quelle est la fonction de Q-valeur après les 2 épisodes ?*

Question 6 *Quelle est la politique gloutonne après les 2 épisodes ?*

2 Exercice Approximate QLearning

Soit le jeu de Pacman, dans lequel on utilise l'approximate Q-Learning pour que le Pacman apprenne à jouer au jeu.

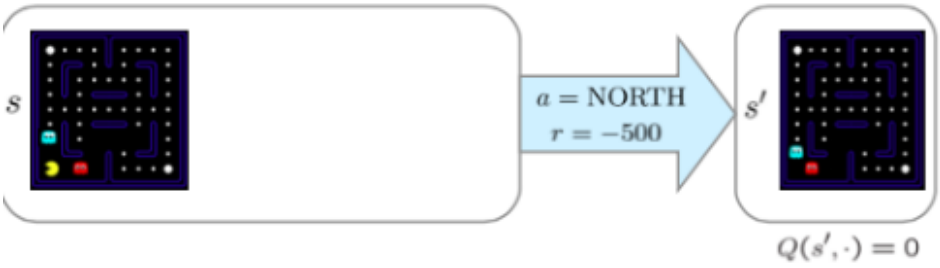
On utilise 2 fonctions caractéristiques

- $f_{DOT}(s, a)$ renvoie $\frac{1}{\texttt{distance au pac gomme le plus proche}}$ lorsque l'on fait a dans s
- $f_{GST}(s, a)$ renvoie la distance au fantôme le plus proche lorsque l'on fait a dans s

L'approximation linéaire pour estimer la Q-fonction est actuellement

$$Q(s, a) = 4.0 f_{DOT}(s, a) - 1.0 f_{GST}(s, a)$$

Soit l'étape ci-dessous :



Dans l'état s, le Pacman fait l'action NORTH, arrive dans l'état s' qui est absorbant (fin de la partie) et reçoit la récompense -500 (partie perdue).

Question 1 *Donner l'expression de la Q-fonction après cette étape. On prendra* $\alpha = 0.004$.