

Examen *Programmation pour l'IA*

Documents de cours autorisés, pas de calculatrice

MDP

On veut formaliser un jeu de dés avec un MDP, et utiliser l'algorithme de *policy iteration* pour trouver la meilleure stratégie.

Exercice 1

- (8 points) On considère un dé (non-biaisé) à 4 faces. On peut lancer le dé 3 fois au plus, et on cherche à maximiser son score. Le score est défini comme suit :
 - Si après un lancer à l'instant t , on obtient la même valeur v qu'au lancer précédent (de l'instant $t - 1$), la partie est finie et on gagne en tout la somme des lancers jusqu'à $t - 1$ plus $-2v$ (et on ne peut plus lancer). Par exemple si on a lancé 3, 1, 1 cela permet de gagner $2 = 3 + 1 - 2$ et termine la partie.
 - Sinon, après chaque lancer, le joueur peut s'arrêter et gagner la somme des valeurs des lancers (par exemple, la séquence 1, 4 permet de gagner $5 = 1 + 4$), ou bien continuer à lancer (pour espérer gagner plus), s'il n'a pas encore lancé 3 fois. Le joueur peut décider de ne jamais lancer, auquel cas son score est zéro.

Formulez ce problème comme un MDP (avec $\gamma = 1$) à deux actions, avec horizon infini. Indiquez l'état initial et les états finaux. Donnez toutes les récompenses $R(s, a, s')$, ainsi que les probabilités de transition $p_T(s'|s, a)$.

- (6 points) On suppose que la politique initiale consiste à toujours s'arrêter. Appliquer trois étapes de *policy iteration* pour améliorer cette politique. À chaque appel de *policy evaluation* vous ferez deux itérations de mise à jour des valeurs d'état seulement. Vous indiquerez bien les récurrences entre valeurs d'état à chaque itération, ainsi que les valeurs d'états à chaque tour de boucle.
En cas d'égalité de gain pendant l'amélioration de la politique, on choisira de s'arrêter plutôt que de continuer
- (2 points) Peut-on conclure que la politique obtenue est optimale ? Justifiez votre réponse

Approximation par fonction

On considère le problème du chariot de montagne (*mountain car*) vu en TP, que l'on veut résoudre par SARSA avec approximation par fonction linéaire. L'état s_t de l'environnement à chaque instant t , caractérisé par la position p et la vitesse v du chariot, est représenté par un vecteur de caractéristiques $x_{(p,v)} \in \mathbb{R}^{10}$ obtenu par un extracteur quelconque, par exemple l'encodage par tuile (*tile coding*). On supposera qu'on n'a que deux actions, aller à droite $\rightarrow$ et aller à gauche $\leftarrow$.

Exercice 2

- (6 points) On suppose que le chariot suit la trajectoire suivante :
 - $s_0 = (0; 0)$, $\rightarrow, -1$, $s_1 = (2, 5; 10)$, $\leftarrow, -1$, $s_2 = (1, 5; 7, 5)$, $\rightarrow, 0$, $s_3 = (6; 1, 2)$Donner les mises à jour successives des vecteurs de paramètres $w^{\rightarrow}, w^{\leftarrow}$ initialisés à zéro. Vous donnerez leurs valeurs en fonction des vecteurs x_s .