Processus de décision markoviens : contrôle de stock

Travaux Pratiques

- Objectif de la séance -

Le but de ce TP est de contrôler le stock d'un vendeur de machines à laver de manière à optimiser son profit. Le cadre est le suivant :

- on note X_t le nombre de machines à laver qu'il détient dans son magasin à la fin de la semaine numéro numéro t ;
- au maximum, il peut stocker M machines ; mais chaque semaine, le coût d'entretien et de lavage de chaque machine est de h ;
- il peut acheter A_t machines au prix unitaire c : il les reçoit alors au début de la semaine suivante ; les frais de livraison sont de K quel que soit le nombre de machines livrées, sauf bien sûr si $A_t = 0$;
- il vend ses machines au prix p ; le nombre de clients au cours de la semaine t est une variable aléatoire D_t, et on suppose que la suite $(D_t)_{t\geq 1}$ est i.i.d.
- on note R_t son chiffre d'affaire sur la semaine t ;
- on suppose qu'au temps initial $t = 1$, son stock est plein : $X_0 = M$;
- soit $\gamma \in]0, 1[$ tel que le taux d'inflation par semaine soit égal à $\gamma^{-1} - 1$: on suppose donc que le vendeur cherche à maximiser en espérance son profit actualisé

$$\mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t R_{t+1}\Big] .$$

On pourra prendre, par exemple : $M = 15$, $K = 0.8$, $c = 0.5$, $h = 0.3$, $p = 1$, $\gamma = 0.99$ et pour la distribution de D_t une loi géométrique de paramètre 0.1.

- Simulation -

1. Montrer que

$$X_{t+1} = \Big((X_t + A_t) \wedge M - D_{t+1}\Big)_+ ,$$

$$R_{t+1} = -K\mathbb{1}\{A_t > 0\} - c\Big((X_t + A_t) \wedge M - X_t\Big)_+ - hX_t + p\Big(D_{t+1} \wedge (X_t + A_t) \wedge M\Big) .$$

2. Comment évolue son chiffre d'affaire s'il commande chaque semaine exactement deux machines ? Faire des simulations.

- PARAMÈTRES DU MDP -

3. Pour chaque $(x, y, a) \in \{0, \ldots, M\}^3$, calculer $P(X_{t+1} = y | X_t = x, A_t = a)$. Construire en R une liste `trans` telle que

$$\texttt{trans[[a]][x,y]} = P(X_{t+1} = y | X_t = x, A_t = a)$$

4. Pour chaque $(x, a) \in \{0, \ldots, M\}^2$, calculer $\mathbb{E}[R_{t+1} | X_t = x, A_t = a]$. Construire en R une liste `rew` telle que

$$\texttt{rew[[a]][x]} = \mathbb{E}[R_{t+1} | X_t = x, A_t = a]$$

- EVALUATION D'UNE POLITIQUE -

On code une politique `pol` par un tableau tel que `pol[1+x]` désigne le nombre de machines à acheter quand $X_t = x$.

5. Ecrire une fonction `policyValue <- function(pol)` qui calcule la fonction valeur d'une politique `pol`.
6. Que peut-il espérer gagner en achetant exactement deux machines chaque semaine ?

- ITÉRATION SUR LES VALEURS -

7. Programmer l'opérateur de Bellman de sorte que la fonction `BellmanOperator <- function(V)` renvoie `res$V`$=T^*(V)$ ainsi que la politique gloutonne associée `res$pol`.
8. Programmer l'algorithme d'itération sur les valeurs pour trouver la stratégie optimale pour le vendeur.

- ITÉRATION SUR LES POLITIQUES -

9. Programmer l'algorithme d'itération sur les politiques, et commenter le résultat. En particulier, on comparera le temps de calcul avec l'algorithme d'itération sur les valeurs.

- Q-LEARNING -

On suppose maintenant que le vendeur ne connaît pas la loi de D_t.

10. Implémenter la méthode de Q-learning, et vérifier sa convergence quand t tend vers l'infini.

Références

[1] M.L. Puterman. *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc. New York, NY, USA, 1994.

[2] Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.