Compte Rendu TP 02: MDP

Sorbonne Université 2018/2019

Master sciences pour l'ingenieur

SOUALHI Takieddine
3872967

Abstrait—Ce Compte rendu présente une étude du processus de decision markovien, dans la suite de ce document nous allons présenter et étudier les algorithmes iteration de valeur et iteration de politque*,

Keywords—Value iteration, Policy iteration, MDP

I. INTRODUCTION

Dans le domaine de l'intelligence Artificielle, il existe de nombreux types d'Agents, nous allons nous intéressé par un certain type appelé Agent Orienté utilité, cet agent utilise la notion de l'utilité pour déterminer une politique a exécuter pour chaque état.

Avant de d'aborder sur les deux algorithmes, nous devons définir ce qui est un processus de décision markovien, un processus de décision markovien s'agit de Spécification d'un problème de décision séquentiel dans un environnement entièrement observable qui satisfait l'hypothèse de Markov et dans lequel les récompenses sont additives

Formellement il est défini comme un tuple <S,A,T,R> dont la solution est une politique (sorte de plan universel) qui est une fonction de S dans A

Une Politique est un choix a priori d'une séquence d'actions, C'est une indication de l'action à exécuter dans chaque et contrairement a ce qu'on a vu dans les algorithmes de recherches la politique n'est pas un plan d'action qui doit êtreexécuté strictement et qui peut échouer; elle est poursuivie malgré des actions dont le résultat n'est pas le plus probable.

Donc l'objective d'un MDP est Trouver la politique optimale (notée π*) liant les états S aux actions A pour maximiser une valeur U(s) (récompense totale, ou utilité)..

II. CALCUL DE Π*

La politique optimale est donnée par la formule :

$$\pi^*(s) = argmax_a \sum_{s'} T(s, a, s')U(s')$$

$T(s, a, s')$ = Probabilité d'atteindre un état s' à partir de l'état s

$U(s')$ = Utilité de l'état s'.

Il existe deux algorithmes pour calculer la politique optimale, on les présentera dans a suite.

A. *Itération de valeur :*

Il s'agit de calcul préalable des utilités, puis calcul de la politique optimale.

Démarche :

- Calculer itérativement l'utilité de chaque état
- Utiliser les valeurs pour sélectionner une action optimale

Algorigramme:

ITERATION_Valeur (mdp)
Entrées: $mdp(S,A,T,R),\gamma,\varepsilon$
Initialiser $U_0(s)$ à $R(s)$ pour tout s
répéter
Pour chaque état $s \neq$ but **faire**
$U_{i+1}(s) \leftarrow R(s) + \gamma \max_a\{\Sigma_{s'} T(s,a,s')U_i(s')\}$
jusqu'à ($U_{i+1} \approx U_i$ "à ε près") [cf. convergence]
ou jusqu'à N itérations - horizon fini.
Sortie: Utilités $U(s)$

Figure.1 Algorigramme de la méthode Itération de valeur

B. *Itération de politique:*

Il s'agit d'un choix d'une politique puis calcul de l'utilité de chaque état pour cette politique avec une mise à jour de la politique à chaque état en utilisant les utilités des états successeurs. On Répéte cette operation jusqu'à obtenir une politique stable.

Démarche :

- On initialise aleaoirement une politique quelconque.
- Pour chaque état:

•Evaluation de la politique
•Calculer l'utilité Ui de chaque état si πi devait êtreexécutée
•Amélioration de la politique
•Calculer une nouvelle politique πi+1 à partir de πi

ITERATION_POLITIQUE (mdp)
entrée: *mdp(S,A,T,R)*
Initialiser *U* (utilités des états de *S*) à *R*
π (politique) à une valeur arbitraire
répéter jusqu'à *π* inchangé
Calculer l'utilité *U* de chaque état si *π* devait être exécutée (évaluation de *π*)
Pour chaque *s* **faire**
si $\max_a\{\Sigma_{s'}T(s,a,s')\ U[s']\} > \Sigma_{s'}T(s,\pi[s],s')\ U[s']$ **alors**
$\pi[s] \leftarrow \text{argmax}_a\{\Sigma_{s'}T(s,a,s')\ U[j]\}$
retourner *π*

Figure.2 Algorigramme de la méthode iteration de politique

III. PARITE PARTIQUE DE TP :

QUESTION 1: Iteration de valeur

L'implementation de l'algorithme se décompose en deux parties :

1. Calcul de l'utilité optimale : pour chaque itération :

- Avant tout, on initialise les état avec une utilité constante.
- Deuxiemement : on determine les voisinage de notre état selon ses particularités.
- Ensuite on calcul le gain de l'état
- On calcul l'utilité de l'état on met dans un vecteur
- On suit les mêmes étapes pour l'état prochaine.

2.Calcul de la politique optimale : Pour chaque état :

- Avant tout, on initialise les états avec une politique nulle.
- on détermine les voisinages de notre état selon ses particularités.
- Ensuite on calcul le gain de l'état
- On calcul la politique optimale comme définie dans la formule dans la Section II.

Pour un facteur d'escompte égale a 1 , un horizon infini on a obtenu la politique optimale ci-dessous

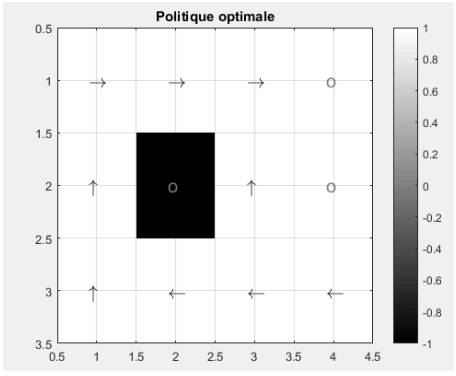

Figure.3 Politique optimale obtenue par la méthode itération de valeurs pour N=150, escompte=1,epsilon=0.00001

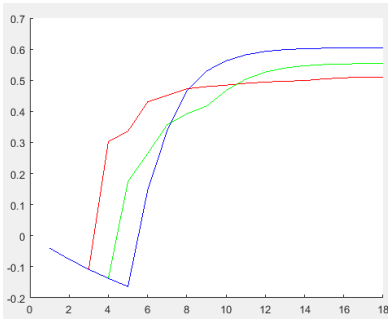

Figure.4 Evolution de U pour les case (3,6,9) en function de nombre d'iterations

On remarque que :

- l'algorithme converge pour N>13
- La récompense de chaque état est plus élevée que

la case "-1". Préférence de contourner la case "-1"pour éviter d'y tomber.

QUESTION 02: Evolution de la politique optimale

1- Cas de facteur d'escompte =0.8 :

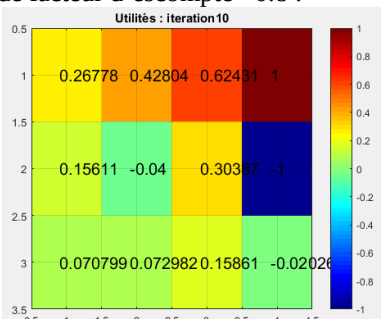

Figure.5 Utilité max obtenue pour escompte =0.8

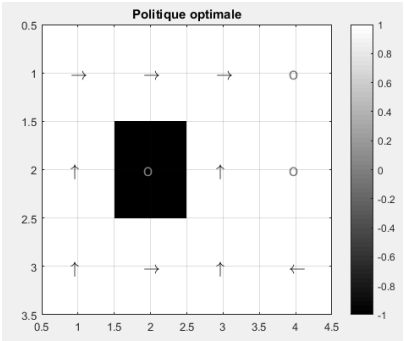

Figure.6 politique optimale obtenue pour escompte =0.8

On remarque que l'algorithme converge rapidement, et il prends le risque de tomber dans -1 pour arriver a l'état but.

Le facteur d'escompte décrit la préférence d'un agent pour les récompenses actuelles par rapport aux récompenses futures. Lorsque γ est proche de 0, les récompenses dans un avenir lointain sont considérées comme non significatives. Lorsque γ est égal à 1, les récompenses remises sont exactement

équivalentes aux récompenses additives. Les récompenses additives constituent donc un cas particulier des récompenses remises. Un facteur d'escompte de γ est équivalent à un taux d'intérêt
de (1 / γ) - 1.

2- Cas ou N=4

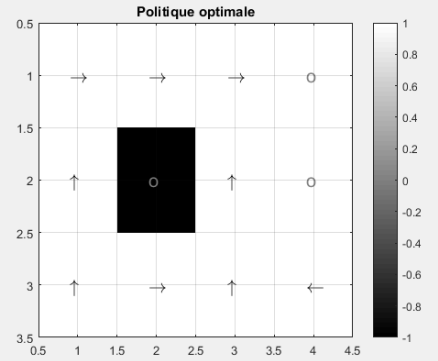

Figure.7 Politique optimale obtenue pour N=4

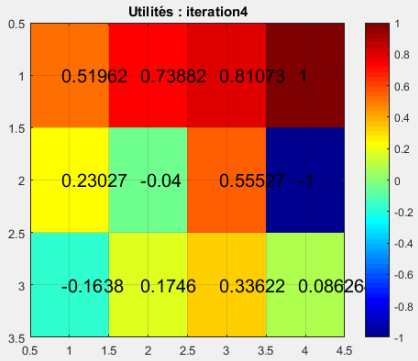

Figure.8 Utilité max obtenue Pour N=4

Dans ce cas, on a supposé que l'horizon est finie, donc l'agent choisis de aller le plus vite possible au but, c-à-d prendre le risque de tomber dans -1 quand même. et c'est ce qu'on a obtenu dans la figure 7.

3- Cas ou R est égale a -0.5 :

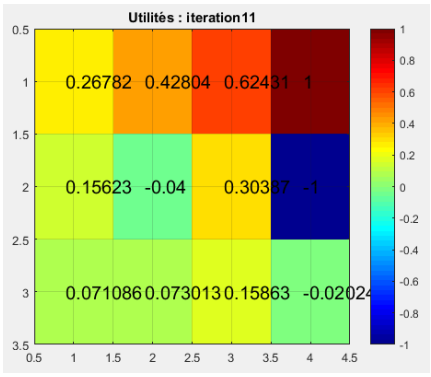

Figure.9 Utilité max pour R initiale =-0.5

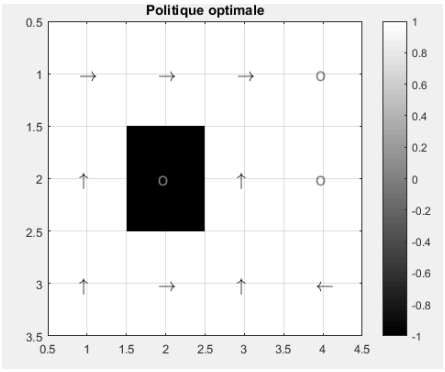

Figure.10 Politique optimale pour R initiale =-0.5

Dans ce cas La récompense est comparable à celle de l'état à éviter. L'agent prend le risque de tomber dans -1 en voulant atteindre +1.

4- Cas ou R(1,1)=1

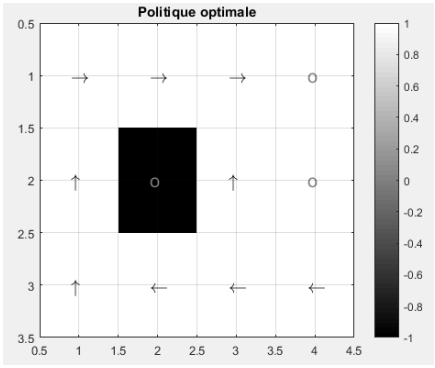

Figure.11 Politique optimale obtenue pour R(1,1)=1

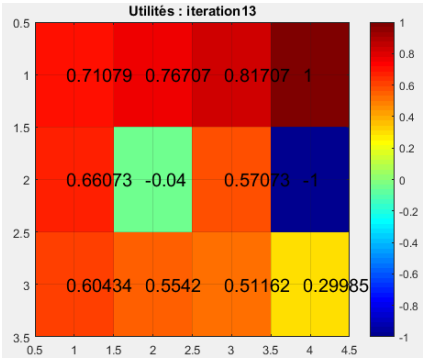

Figure.12 Utilité Max obtenue pour R(1,1)=1

On remarque que l'agent évite la case -1. Car l'agent essaye toujour de maximiser l'utilité donc il prefere le chemin qui passe par +1

QUESTION 03: Itération de politique

L'implémentation de l'algorithme Itération de politique se fait directement :

- On initialise la matrice M est les états a politique nulle
- On construit M et on résout le système d'équations linéaires
- On calcul le gain de l'utilité des états
- On calcul la politique optimale

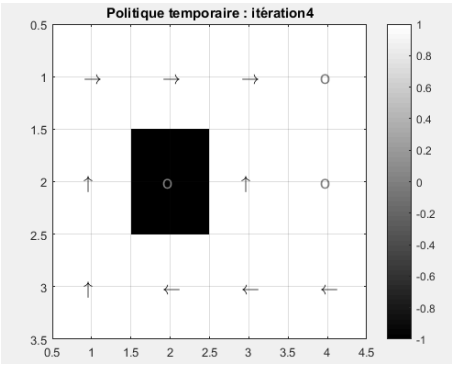

Figure.13 Politique optimale obtenue pour l'algorithme itération de politique

On remarque que cet algorithme est plus rapide que le premier, il a convergé à N=4.

QUESTION 04: Complexité en temps

Temps estimer par Iteration de politique :

```
Elapsed time is 3.608414 seconds.
```

Temps estimer par Iteration de valeur:

```
Elapsed time is 12.208383 seconds.
```

On conclut que l'algorithme d'itération de politique est beaucoup plus rapide que l'algorithme d'itération de valeur (Test effectué pour politique initiale égale à 1 puis à 4 pour toutes les états

IV. CONCLUSION

Lors de ce TP, on a pu mettre en pratique les connaissances acquises en cours sur la résolution des problèmes et le Processus de décision markovien. Et on était capable de noté plusieurs points importantes :

- L'algorithme d'itération de valeur pour la résolution de MDPs fonctionne en résolvant de manière itérative les équations relatives à l'utilité de chaque état à ceux de ses voisins.

- L'itération de politique alterne entre le calcul des utilités des États pour la politique actuelle et l'amélioration de la politique actuelle tout en prenant en compte les utilités actuelles .

- en pratique l'algorithme d'itération de politique est plus rapide et plus sophistiqué que l'algorithme d'itération de valeur

- le nombre d'itération, le facteur d'escompte et les valeurs des récompenses sont les paramètres qui :

* Caractérise l'optimalité d'un l'algorithme (itération de valeur ou de politique

* Détermine la convergence et la sortie de l'algorithme.

REFERENCES

[1] Andrew Schaefer Seminar , Markov decision processes .
[2] P. Norvig and S. J. Russel, "Artificial Intelligence : A modern Aproach".
[3] Cours de M. Raja Chatila