TP6: Apprentissage par démonstration.

Fait par: Imène TARAKLI

Master ISI - 2020/2021

Objectif

L'objectif de ce TP est de développer des stratégies d'apprentissage par démonstration et de comparer leurs performances sur différentes configurations de l'environnement Gridword.

Développement de l'environnement

L'environnement considéré dans ce TP est une grille (5x6) comportant une case "récompense" jaune, valant 10 points, des cases "danger" rouges, valant une pénalité de 2 points, et des cases blanches "neutre" ne valant aucun point. Ces cases peuvent être disposées selon différentes configurations comme le montre la figure 1 tirée du TP.

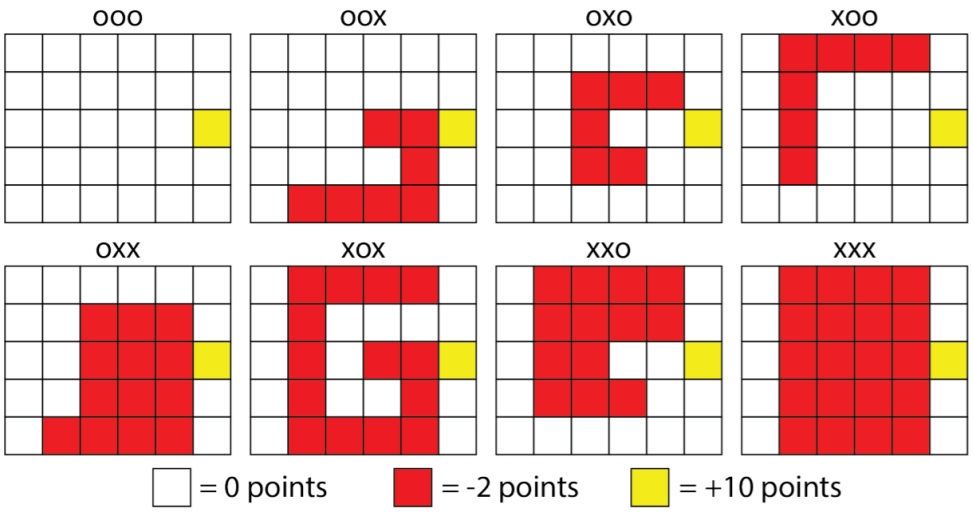

Figure 1: Configurations de l'environnement.

Cet environnement est implémenté sous forme de classe composée d'un attribut grille qui n'est autre que la version matricielle de l'environnement, comme montré dans la figure 2. Plusieurs méthodes de cette classe permettent de se déplacer dans la grille (up, down, left et right), d'obtenir la récompense d'une action et d'afficher le résultat d'une politique suivie.

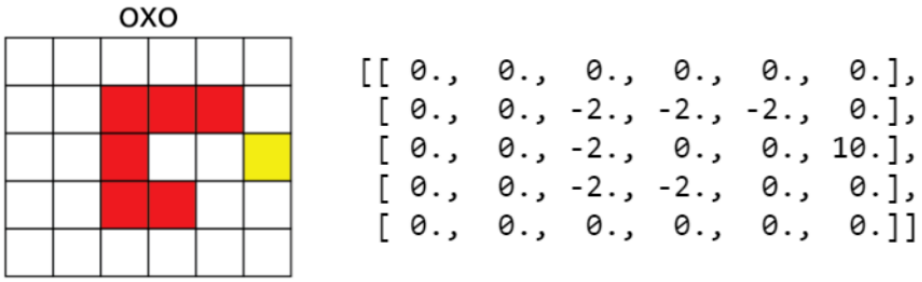

Figure 2: Format matriciel de l'environnement.

Afin d'avoir la politique optimale qui servira de démonstration à l'agent, une méthode reprenant l'algorithme Q-learning a été ajoutée. Cette dernière renvoie les Q-values ainsi que les états parcourus dans la politique optimale. La figure 3 reprend le résultat de cette dernière pour la configuration "oxo" où -1 représente la position de départ et 1 les cases parcourues avant d'atteindre la case "récompense".

```
[[-1.,  1.,  1.,  1.,  1.,  1.],
 [ 0.,  0., -2., -2., -2.,  1.],
 [ 0.,  0., -2.,  0.,  0., 10.],
 [ 0.,  0., -2., -2.,  0.,  0.],
 [ 0.,  0.,  0.,  0.,  0.,  0.]]
```

Figure 3: Politique optimale pour la configuration "oxo".

Modèle de Boltzman

Le modèle de Boltzman a pour but de dériver des politiques à partir d'une, optimale, donnée. Ces dernières constituerons les différentes démonstrations pour l'expérience considérée.

Le modèle de Boltzman calcule la probabilité d'une action pour un état donné :

$$\pi\left(a_t \mid s_t\right) = \frac{e^{Q^*(s_t,a_t)/\tau}}{\sum_{a' \in A(s_t)} e^{Q^*(s_t,a')/\tau}} \tag{1}$$

où: $Q^*(s_t, a_t)$ est la politique optimale et τ, le facteur de température.

La facteur de température influence la distribution de probabilités de l'espace action. Quand ce facteur est inférieur à 1, les probabilités sont accentués de façon à favoriser une action plus que d'autres. Quand il est supérieur à 1, la distribution devient uniforme; les actions sont donc choisies aléatoirement.

Pour un état donné, la figure 4 montre les résultats obtenus pour différents facteurs de température.

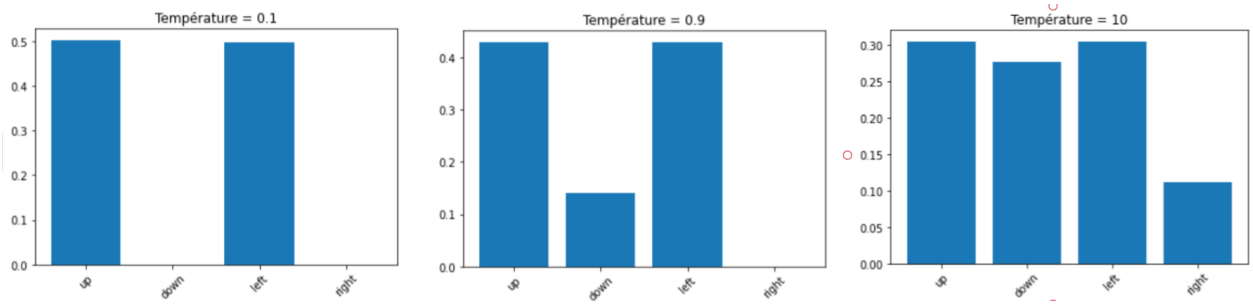

Figure 4: Influence du facteur température sur la distribution de probabilités.

La politique dérivé est le résultat d'un échantillonage de ces distributions. Un facteur de temprératue inférieur à 1 fournit une politique différente de celle optimale mais toutefois correcte tandis qu'un facteur supérieur à 1 résulte en une politique aléatoire n'évitant pas toujours les obstacles. Cela est illustré par les figures 5 et 6 pour les configurations "OXO" et "OOX", respectivement.

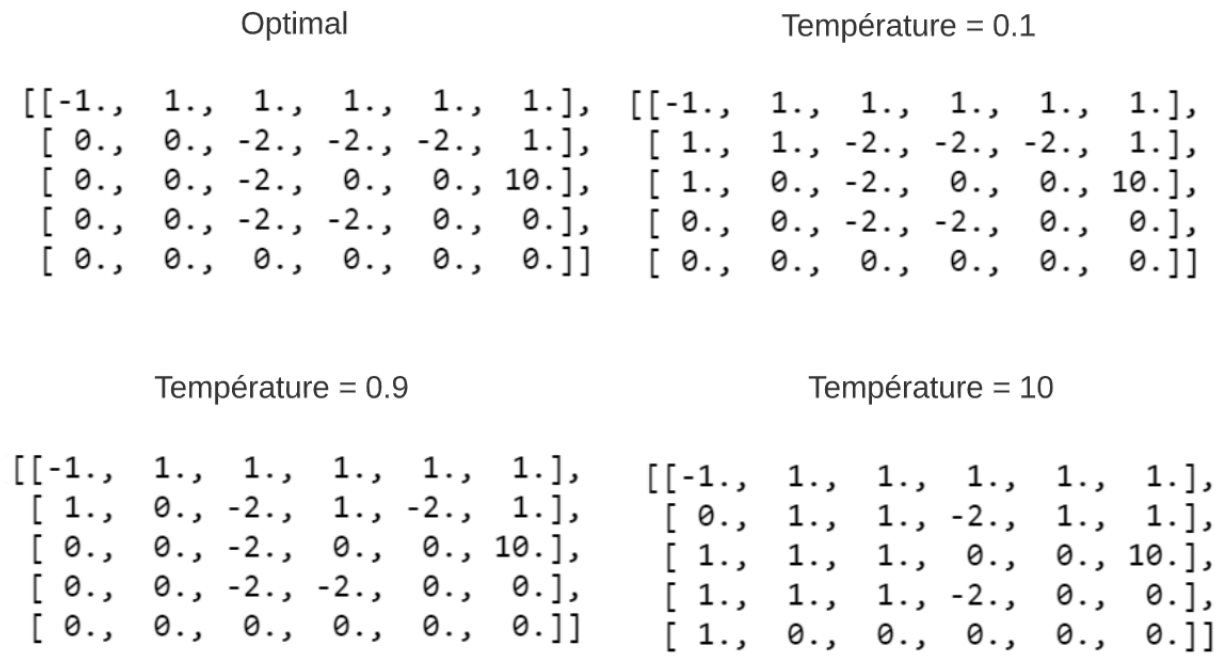

Figure 5: Influence du facteur température sur la politique pour la configuration 'OXO'.

LESS is more : Rethinking probabilistic models of human behavior

Dans l'étude menée par Bodu et al., un nouveau modèle générateur de démonstrations est présenté. Ce dernier est plus proche du comportement humain car l'espace des trajectoire n'est plus discret, comme c'est le cas pour le modèle de Boltzman, mais continu, s'approchant donc plus des trajectoires humaines. En effet, la récompense n'est plus l'unique paramètre entrant en jeu dans la prise de décision mais la distance et la similarité des trajectoires sont aussi considéré. Les résultats obtenus précédemment par le modèle de Boltzman montrent des politiques correctes, dans le sens où les obstacles sont en effet évitées, mais elles ne sont pas toujours optimales. En effet, l'homme aura tendance à choisir la plus courte politique évitant les obstacles.

Optimal

```
[[-1.,  1.,  1.,  1.,  0.,  0.],
 [ 0.,  0.,  0.,  1.,  1.,  1.],
 [ 0.,  0.,  0., -2., -2., 10.],
 [ 0.,  0.,  0.,  0., -2.,  0.],
 [ 0., -2., -2., -2., -2.,  0.]]
```

Température = 0.1

```
[[-1.,  1.,  1.,  1.,  1.,  1.],
 [ 1.,  1.,  1.,  1.,  1.,  1.],
 [ 0.,  0.,  0., -2., -2., 10.],
 [ 0.,  0.,  0.,  0., -2.,  0.],
 [ 0., -2., -2., -2., -2.,  0.]]
```

Température = 0.9

```
[[-1.,  1.,  1.,  1.,  1.,  1.],
 [ 1.,  1.,  1.,  1.,  1.,  1.],
 [ 0.,  0.,  0.,  1., -2., 10.],
 [ 0.,  0.,  0.,  0., -2.,  0.],
 [ 0., -2., -2., -2., -2.,  0.]]
```

Température = 10

```
[[-1.,  1.,  1.,  1.,  1.,  0.],
 [ 1.,  1.,  1.,  1.,  1.,  1.],
 [ 0.,  0.,  1.,  1., -2., 10.],
 [ 0.,  0.,  1.,  1., -2.,  0.],
 [ 0., -2.,  1., -2., -2.,  0.]]
```

Figure 6: Influence du facteur température sur la politique pour la configuration 'OOX'.