Projet 4 : Apprentissage par renforcement

INFO-F-311 : Intelligence Artificielle

Bahja Ali
Matricule: 000534976
BA3-INFO

Année académique 2024-2025

Université Libre de Bruxelles
Faculté des Sciences
Département d'Informatique

Introduction

L'objectif de ce projet est d'explorer l'**apprentissage par renforcement (RL)** à travers deux méthodes : **Value Iteration**, un algorithme basé sur un modèle complet de l'environnement, et **Q-Learning**, un algorithme sans modèle apprenant directement à partir d'interactions.

L'environnement utilisé est un **labyrinthe** où l'agent doit atteindre une sortie tout en maximisant ses récompenses. Ce projet permet de comparer les performances des deux approches, d'analyser l'effet de l'exploration, ainsi que la sensibilité aux paramètres clés (facteur de discount γ, taux d'apprentissage α, et probabilité d'aléatoire p de l'environnement).

1 Mode opératoire

1.1 Environnement

Le labyrinthe est représenté comme une grille de taille 7×7, avec :

- Des états bloqués (murs) où l'agent ne peut pas se déplacer.
- Deux cases terminales (les sorties) avec une récompense positive (10 pour la sortie supérieur, 100 pour la sortie inférieure gauche).
- Un coût de déplacement négatif -1 pour encourager l'agent à trouver un chemin court.

1.2 Algorithmes implémentés

Value Iteration : Itère sur les valeurs d'états jusqu'à convergence, en utilisant l'équation de Bellman.

Q-Learning : Met à jour une table $Q(s, a)$ à partir des transitions observées, avec deux stratégies d'exploration :

- ε**-greedy** : choix aléatoire avec probabilité ε.
- **Softmax (Max-Boltzmann)** : choix probabiliste basé sur un paramètre de température τ.

1.3 Paramètres d'entraînement

- Nombre d'épisodes : 1 000 (20 répétitions en moyenne).
- $\gamma = 0.9$ (sauf analyse de sensibilité).
- $\alpha = 0.1$.
- Politiques d'exploration : ε fixe, décroissant, τ fixe, décroissant.

2 Résultats et discussion

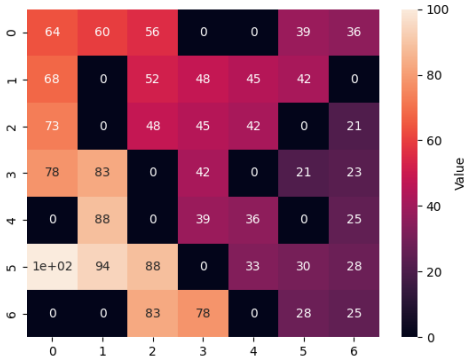

FIGURE 1 – Heatmap des valeurs d'états obtenues par Value Iteration.

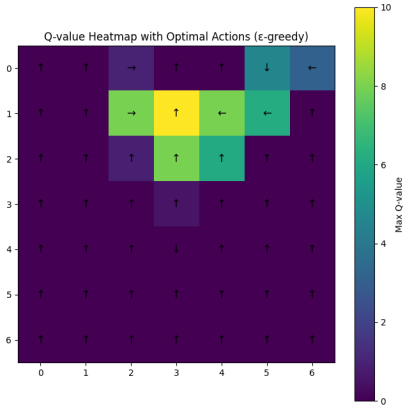

FIGURE 2 – ε-greedy

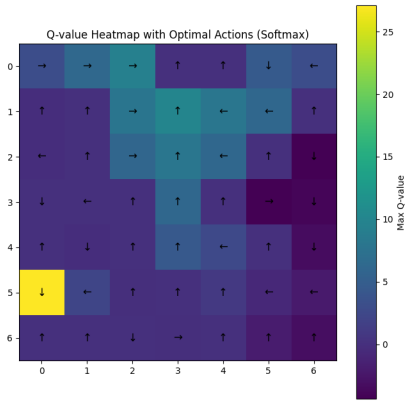

FIGURE 3 – Softmax

Lorsque la probabilité de non-déterminisme $p > 0$, une même action peut mener à des états différents. **Value Iteration** encode explicitement ce hasard dans l'équation de Bellman. **Q-learning** l'encode implicitement par échantillonnage : les valeurs $Q(s, a)$ deviennent des espérances à force de mises à jour.

On observe que plus p est grand, plus les valeurs diminuent (le coût -1 s'accumule et les récompenses terminales sont atteintes moins souvent).

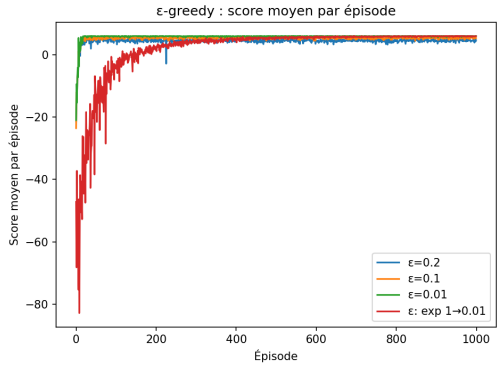

FIGURE 4 – Score moyen par épisode pour différentes valeurs de ε

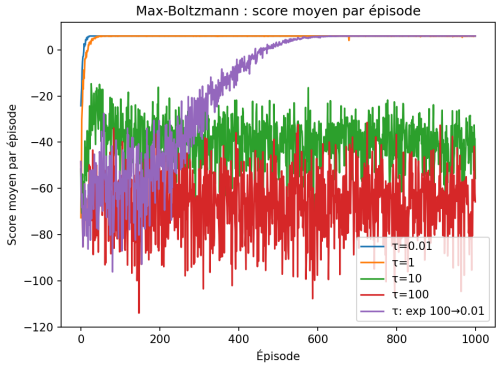

FIGURE 5 – Score moyen par épisode pour différentes valeurs de τ

Quand p augmente, les retours moyens chutent et la convergence ralentit. Value Iteration reste plus robuste car il optimise directement l'espérance des transitions. Q-learning est plus bruité : il lui faut davantage d'épisodes pour lisser l'aléa, et ses performances se dégradent plus rapidement dans les environnements très stochastiques.

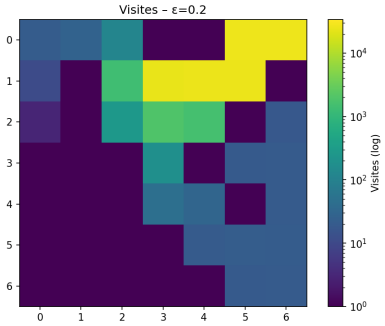

FIGURE 6 – Heatmap des visites pour $\varepsilon = 0.2$

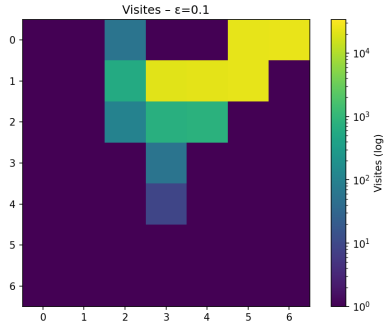

FIGURE 7 – Heatmap des visites pour $\varepsilon = 0.1$

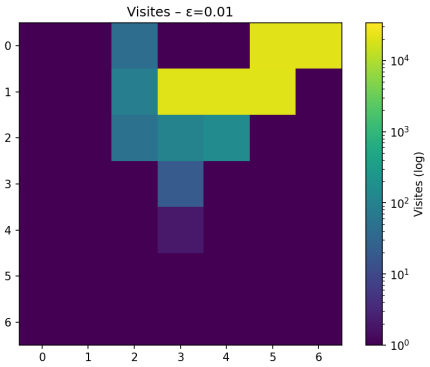

FIGURE 8 – Heatmap des visites pour $\varepsilon = 0.01$

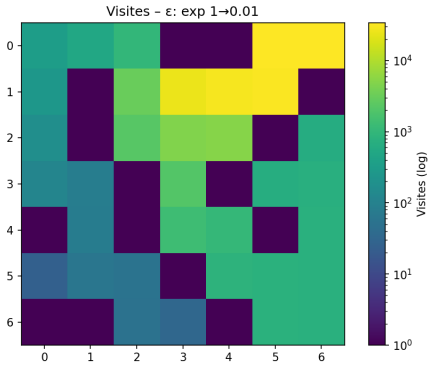

FIGURE 9 – Heatmap des visites pour ε décroissant de 1 à 0.01

La stratégie ε-greedy explore uniformément : un ε trop grand ralentit l'apprentissage, tandis qu'un ε décroissant donne un bon compromis.

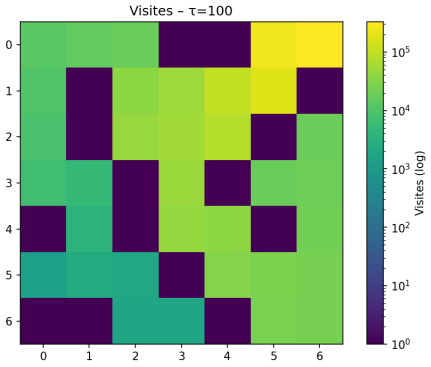

FIGURE 10 – Heatmap des visites pour $\tau = 100$

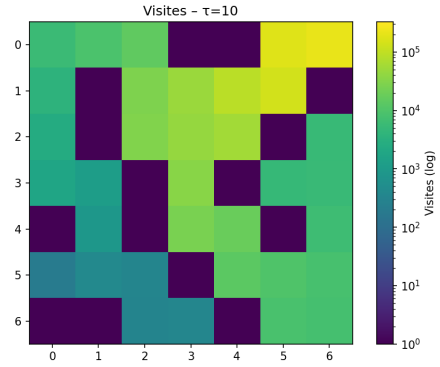

FIGURE 11 – Heatmap des visites pour $\tau = 10$

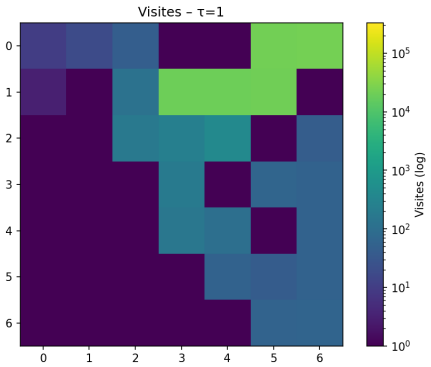

FIGURE 12 – Heatmap des visites pour $\tau = 1$

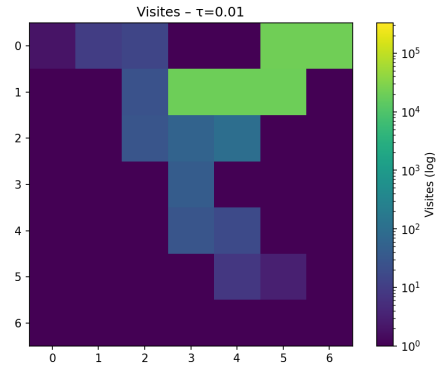

FIGURE 13 – Heatmap des visites pour $\tau = 0.01$

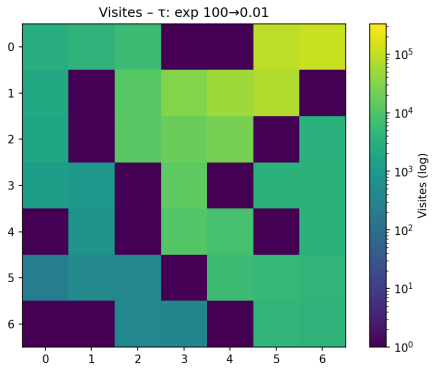

FIGURE 14 – Heatmap des visites pour τ décroissant de 100 à 0.01

La stratégie **Softmax (Max-Boltzmann)** explore de manière proportionnelle aux valeurs Q, ce qui permet un meilleur équilibre exploration/exploitation si τ est bien choisi. Un τ trop grand (≈ 100) conduit à une exploration quasi-aléatoire et de faibles scores, tandis qu'un τ trop petit revient à une politique gourmande.

Lorsque l'agent converge, il suit le chemin le plus court car les récompenses négatives s'accumulent sur les trajets plus longs, tandis que la **propagation des valeurs** ou la mise à jour des **Q-valeurs** favorise naturellement les actions menant le plus rapidement à la sortie.

3 Conclusion

Dans ce projet, nous avons implémenté et comparé deux approches d'apprentissage par renforcement : **Value Iteration**, méthode déterministe basée sur un modèle, et **Q-Learning**, méthode d'apprentissage sans modèle. Nos résultats montrent que **Value Iteration** converge rapidement vers une politique optimale lorsque l'environnement est déterministe, tandis que **Q-Learning** reste plus flexible et parvient à s'adapter même en présence d'aléatoire dans les transitions.

L'étude des stratégies d'exploration (ε-*greedy* et *Max-Boltzmann*) a mis en évidence leur rôle essentiel dans l'équilibre entre exploration et exploitation : un taux ε décroissant et une température τ adaptée permettent une exploration suffisante en début d'entraînement puis une exploitation efficace par la suite. Nous avons également montré que l'augmentation de la probabilité de non-déterminisme p ralentit l'apprentissage et dégrade la performance, mais que les valeurs apprises intègrent cet aspect aléatoire en ajustant les espérances de récompenses.

En conclusion, ce travail illustre l'importance des mécanismes de mise à jour des valeurs et des stratégies d'exploration pour atteindre un comportement optimal, même dans un environnement incertain.