Rapport d'Implémentation : Q-Learning vs SARSA dans Taxi-v3

simon.thuaud

October 9, 2024

1 Introduction

Dans ce travail pratique, nous avons implémenté trois agents utilisant les algorithmes d'apprentissage par renforcement Q-Learning, Q-Learning avec scheduling de l'exploration (ϵ-greedy), et SARSA pour jouer au jeu `Taxi-v3` d'OpenAI Gym. Le but du jeu est de déplacer un taxi sur une grille 5x5, ramasser un passager et le déposer à destination en un minimum d'actions.

2 Algorithmes

2.1 Q-Learning

L'algorithme de Q-Learning est un algorithme hors-politique (off-policy) qui met à jour la valeur d'une paire état-action en maximisant la récompense future attendue. Dans notre implémentation, l'agent utilise une stratégie ϵ-greedy où, avec une probabilité ϵ, il choisit une action aléatoire pour explorer de nouveaux états. Sinon, il exploite les connaissances actuelles en choisissant l'action qui maximise la récompense attendue.

2.2 QLearning avec Scheduling de l'Exploration

Cet agent utilise la même règle de mise à jour que le Q-Learning classique, mais avec un paramètre ϵ qui décroît progressivement au fil des épisodes. Cela permet à l'agent d'explorer davantage au début de l'entraînement, puis de se concentrer sur l'exploitation des actions optimales à mesure qu'il apprend une politique plus efficace.

2.3 SARSA

L'algorithme SARSA est un algorithme on-policy, ce qui signifie qu'il met à jour la valeur d'une paire état-action en fonction de l'action réellement choisie dans l'état suivant. Cet algorithme prend en compte les actions exploratoires lors de l'apprentissage, contrairement à Q-Learning qui mise uniquement sur les actions optimales dans l'état futur.

3 Comparaison des Performances

Pour comparer les performances des trois algorithmes, nous avons mesuré la récompense moyenne obtenue par chaque agent au cours de 1000 épisodes d'entraînement. Les résultats montrent que :

- L'agent **Q-Learning** converge rapidement, obtenant des récompenses positives dès les 200 premiers épisodes.
- L'agent **QLearning avec Scheduling** présente un comportement similaire, avec une légère amélioration de la stabilité à long terme.

- L'agent **SARSA** met plus de temps à converger et nécessite plus d'épisodes pour atteindre des performances comparables à celles de Q-Learning.

4 Conclusion

En conclusion, les algorithmes Q-Learning et Q-Learning avec scheduling de l'exploration convergent plus rapidement que SARSA dans l'environnement Taxi-v3. La différence de performances est due à la nature hors-politique de Q-Learning, qui permet à l'agent de toujours choisir les actions optimales, tandis que SARSA prend en compte les actions exploratoires. Bien que SARSA puisse être plus approprié dans certains environnements, Q-Learning reste plus efficace dans ce cas particulier.