Reinforcement learning - TP3

Lilian SCHALL<lilian.schall@epita.fr>, Guillaume LALIRE <guillaume.lalire@epita.fr>

Octobre 2024

Table des matières

1 Introduction

Dans cette étude, nous examinons trois techniques d'apprentissage par renforcement appliquées à l'environnement Taxi, tel que défini dans la bibliothèque Gym. Les modèles utilisés sont Q-learning, Q-learning avec planification epsilon, et SARSA. L'objectif est d'analyser leurs performances en termes de rapidité d'apprentissage, du nombre d'époques nécessaires pour converger, ainsi que d'autres mesures pertinentes.

2 Choix d'implémentation

Pour trouver au mieux les différents hyperparamètres qui composent chaque modèle, nous avons décidé d'effectuer un grid search. Ce grid search permet également d'effectuer une analyse de sensibilité des modèles sur leurs différents paramètres. Pour éviter d'explorer un trop grand nombre de valeurs, nous avons sélectionné des valeurs pertinentes en se basant sur les plages de valeurs recommandées pour ce genre de modèles.

Chaque entraînement de modèle fait usage des mêmes conditions de sessions : nous utilisons 2000 époques pour entraîner chaque modèle, et enregistrons une vidéo sur 10 scénarios lors de la phase de testing. L'analyse de chaque modèle contiendra le score moyen sur les 10 scénarios de testing.

2.1 QLearning

La figure 1 correspond à un grid search sur les trois hyperparamètres du modèle Q-Learning : epsilon, learning rate et gamma. On observe qu'un learning rate trop faible ou trop élevé ne permet pas au modèle de converger vers une performance optimale. Quant à gamma, le modèle a en moyenne plus de facilité à converger lorsque gamma tend vers 1. Si l'on fixe un learning rate et un gamma optimaux (i.e. gamma = 0.999 et learning rate = 0.1), on s'aperçoit qu'epsilon doit être suffisamment petit pour converger à une récompense totale positive. Ici, lorsque epsilon = 0.1, on obtient 0.41 de récompense totale en moyenne.

2.2 QLearning Epsilon Scheduling

Pour Q-Learning Epsilon Scheduling, nous avons réutilisé les paramètres optimaux dans le grid search de Q-Learning pour gamma et learning rate, et avons entraîné le modèle avec les valeurs par défaut pour epsilon start / end et decay_steps. Nous avons obtenu une récompense totale de 3.

2.3 SARSA

Après grid search, représenté par la figure 2, nous observons que le choix du taux d'apprentissage est déterminant pour la performance du modèle sur 2000 époques : le modèle SARSA nécessite un taux d'apprentissage élevé (0.5) pour performer dans ces conditions. La valeur exacte de gamma ne semble à l'inverse pas énormément influencer les résultats tant que sa valeur reste dans l'intervalle [0.9, 1].

3 Conclusion

Nous avons pu ainsi entraîner plusieurs modèles de reinforcement learning pour une tache commune. Au travers de ce résultat, et grace a une recherche hyperparamètres, nous avons pu constater que le modele SARSA performait mieux que les modèles QLearning et QLearningEpsScheduling, avec un score maximal de 7.

4 Annexe

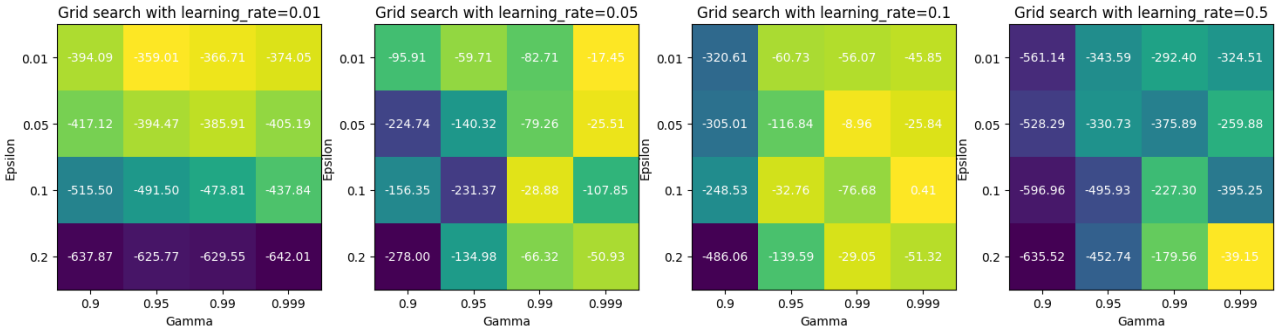

Figure 1 – Grid search pour QLearning

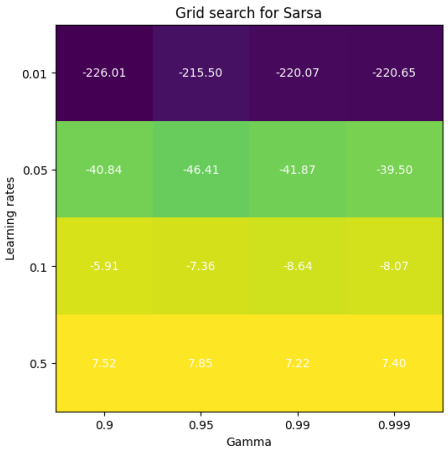

Figure 2 – Grid search pour SARSA