Rapport sur l'implémentation d'agents d'apprentissage par renforcement pour le jeu Taxi-v3

Vincent SAUNIER Martin NATALE

1 Introduction

Ce rapport présente l'implémentation et les résultats de trois agents d'apprentissage par renforcement pour le jeu Taxi-v3 de OpenAI Gym : Q-Learning, Q-Learning avec planification d'epsilon, et SARSA.

2 Choix d'implémentation

2.1 Q-Learning et SARSA

Les deux algorithmes ont été implémentés avec les caractéristiques suivantes :

- Utilisation d'un dictionnaire pour stocker les Q-valeurs, permettant une mise à jour et un accès efficaces.
- Normalisation des récompenses (division par 20) pour stabiliser l'apprentissage et éviter des mises à jour trop importantes.
- Clipping de l'erreur TD entre -1 et 1 pour prévenir des changements brusques dans les Q-valeurs.
- Stratégie d'exploration epsilon-greedy avec décroissance exponentielle pour Q-Learning standard, favorisant l'exploration au début et l'exploitation à la fin.
- Pour SARSA, utilisation d'une décroissance d'epsilon plus lente pour maintenir un certain niveau d'exploration plus longtemps.

2.2 Q-Learning avec planification d'epsilon

Cet agent étend le Q-Learning standard avec :

- Une décroissance linéaire d'epsilon sur un nombre fixe d'étapes, offrant un contrôle plus précis de la transition exploration-exploitation.
- Un epsilon minimal pour maintenir une exploration résiduelle même après la période de décroissance.

2.3 Optimisations communes

Pour tous les agents :

- Réinitialisation de l'agent en cas de stagnation des performances, permettant de sortir des optima locaux.
- Utilisation de moyennes mobiles pour le suivi des performances, offrant une vue plus stable de l'évolution de l'apprentissage.

3 Résultats

Les graphiques montrent l'évolution des récompenses totales pour chaque agent sur 10 000 épisodes :

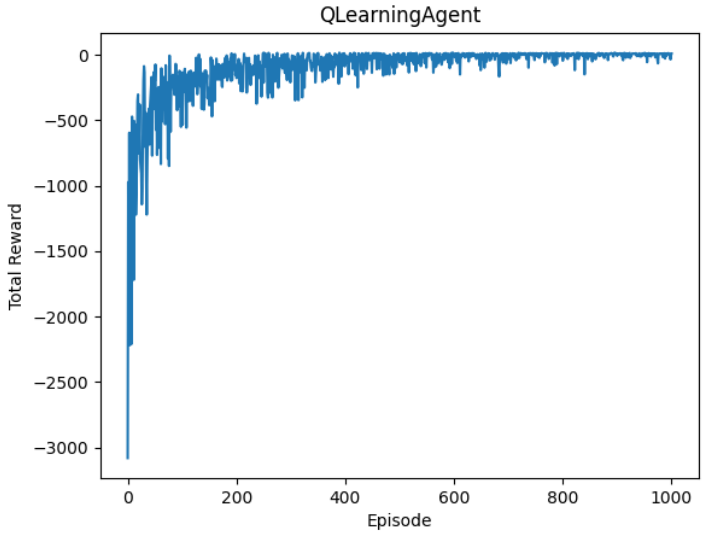

Figure 1: Performance de l'agent Q-Learning

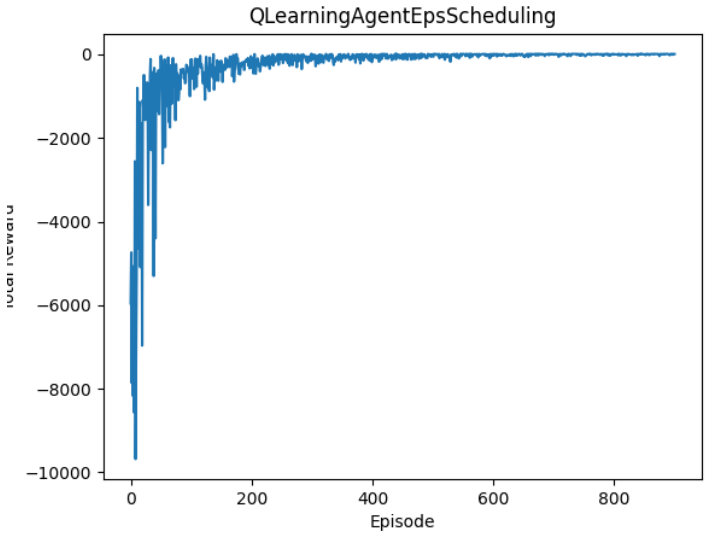

Figure 2: Performance de l'agent Q-Learning avec planification d'epsilon

4 Analyse

- Tous les agents montrent une amélioration significative des performances au fil du temps, indiquant un apprentissage effectif.
- L'agent Q-Learning standard converge le plus rapidement vers une politique stable, probablement dû à sa nature off-policy qui lui permet d'apprendre directement la politique optimale.
- L'agent Q-Learning avec planification d'epsilon montre une exploration plus prolongée et une convergence plus lisse, ce qui peut être bénéfique pour des environnements plus complexes ou dynamiques.
- L'agent SARSA présente une convergence plus lente mais potentiellement plus stable, caractéristique des méthodes on-policy qui peuvent être plus robustes dans certains environnements.

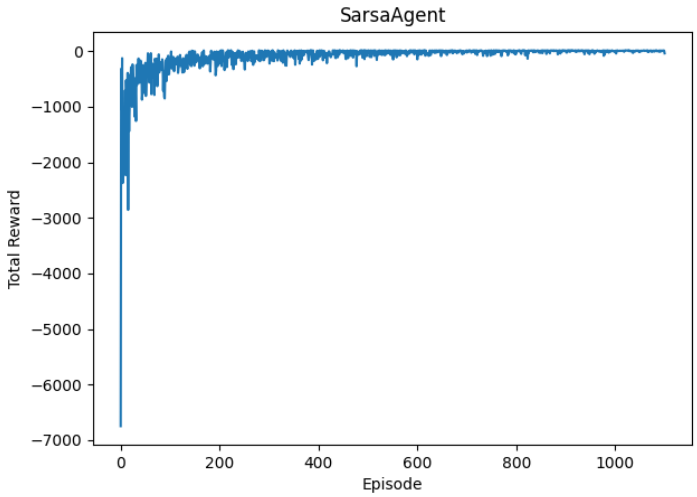

Figure 3: Performance de l'agent SARSA

5 Conclusion

Les trois agents ont réussi à apprendre une politique efficace pour le jeu Taxi-v3. Le choix entre ces méthodes dépendrait du compromis souhaité entre vitesse d'apprentissage, stabilité et capacité d'exploration. Pour ce problème spécifique, le Q-Learning standard semble offrir le meilleur équilibre, mais les autres méthodes pourraient être préférables dans des environnements plus complexes ou avec des dynamiques différentes.

6 Perspectives

Pour améliorer davantage les performances et l'applicabilité de ces agents :

- Implémenter des méthodes d'apprentissage par renforcement plus avancées comme DQN (Deep Q-Network) pour gérer des espaces d'états plus grands.
- Expérimenter avec des techniques d'exploration plus sophistiquées comme l'exploration basée sur l'incertitude.
- Tester les agents sur des variantes plus complexes du problème Taxi ou sur d'autres environnements pour évaluer leur généralisation.
- Intégrer des techniques d'apprentissage par transfert pour accélérer l'apprentissage sur des tâches similaires.