RL Report TP 3 - Octobre 2024

mael.reynaud
alexandre.devaux-riviere

Choix d'implémentation :

Voici les choix d'implémentation de notre TP :

— Nous avons décidé de modifier le paramètre epsilon, responsable de l'exploration, afin d'obtenir de meilleurs résultats finaux pour chacun des modèles, notamment en le baissant.

— L'implémentation du modèle SARSA a été légerement différente des 2 autres modèles, car nous n'avons pas implémenté la fonction **get_value()**, puisqu'elle ne nous aurait pas été utile, sachant qu'avec le modèle SARSA on prend en compte l'action à l'état+1 et pas l'action qui maximiserait la prochaine q-value (comme le fait **get_value()**).

— Pour observer les performances de nos modèles (partie Résultats), nous avons décidé de faire la moyenne des récompenses en fonction des époques sur plusieurs simulations afin de réduire le caractère stochastique de l'entraînement sur une simulation, avec l'axe **x** logarithmique pour bien se rendre compte de l'évolution des récompenses.

Résultats :

Vous pourrez trouver dans le dossier */report* de notre repository des vidéos de démonstration de chacun de nos algorithmes.

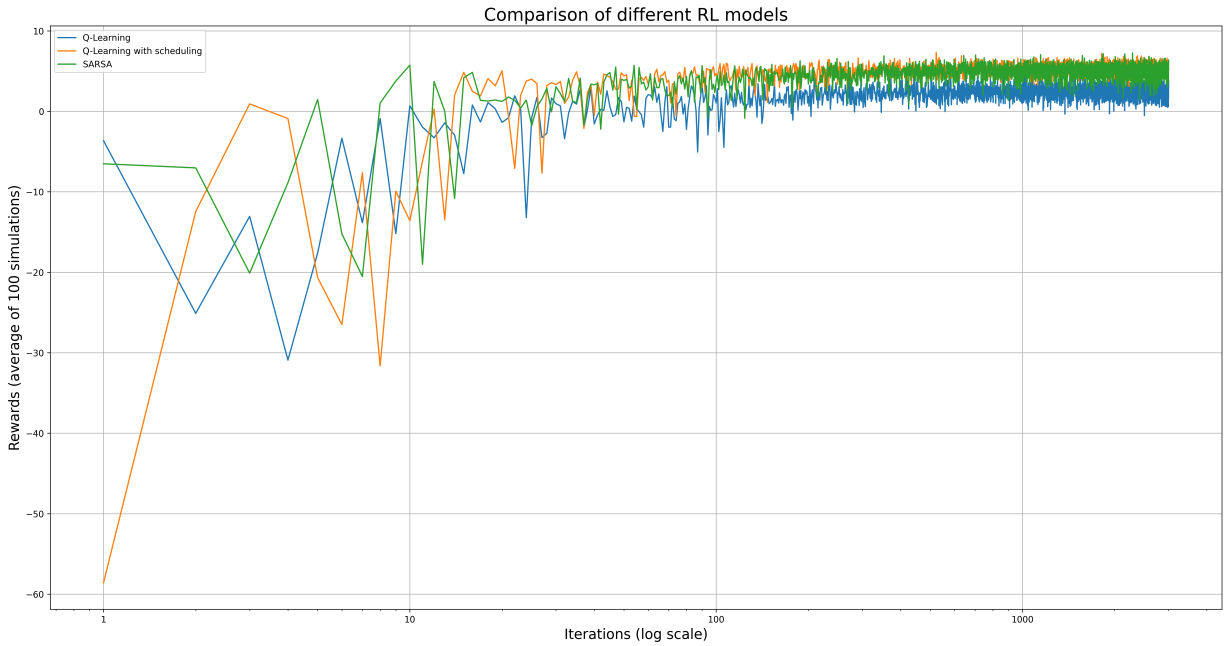

FIGURE 1 – Comparaison de la moyenne des récompenses sur 100 simulations entre différents modèles de RL au file des epochs pour le jeu TaxiV3

Environ 100 époques sont nécessaires pour se stabiliser autour d'une constante positive (représentant la solution). Le modèle Q-learning avec planification commence avec une récompense initiale très basse (indiquant une exploration très intense, qui diminue progressivement).

Les modèles Q-learning avec planification et SARSA offrent une meilleure approche du problème que le Q-learning de base, car ils convergent vers une constante plus élevée que celle obtenue par l'algorithme Q-learning standard.