Reinforcement Learning – TP 3

Barel, Ewan, ewan.barel@epita.fr, 25610

21 octobre 2024

1 Q-Learning

Configuration 1 Dans un premier temps, on étudie les résultats obtenus avec les hyperparamètres à leur valeur par défaut (figure 1). Le modèle ne converge pas, et on observe sur les vidéos produites que le taxi réalise régulièrement des écarts impromptus empêchant l'obtention d'une récompense moyenne[1] strictement positive. On suppose que ceux-ci sont dus à une valeur trop élevée de ε.

Configuration 2 La diminution de ε à 0.05 permet de converger après plus de 600 épisodes (figure 2). On en conclut qu'une valeur trop élevée de ε nuit aux performances en raison d'une proportion trop élevée de choix aléatoires tard dans l'apprentissage.

Configuration 3 Puisque le modèle ne semble pas apprendre suffisamment vite et que cet environnement est complètement déterministe, on étudie la convergence quand $\varepsilon = 0$ et $\alpha = 1$. Dans cette configuration, le modèle bat ses performances initiales et converge au bout de 400 épisodes en moyenne, atteignant une récompense moyenne maximale autour de 8 (figure 3).

Configuration 4 On cherche maintenant à évaluer l'impact du facteur d'actualisation γ sur les performances. Un facteur d'actualisation trop faible $\gamma = 0.01$ empêche complètement le modèle de converger (figure 4a) tandis qu'un facteur d'actualisation moyen $\gamma = 0.5$ diminue le temps d'apprentissage (environ 300 épisodes au lieu de 400) tout en conservant des performances semblables à la configuration 3 avec une récompense moyenne maximale toujours autour de 8 (figure 4b).

2 Q-Learning ε-scheduling

Configuration 1 Tout d'abord, on étudie les résultats obtenus avec les hyperparamètres à leur valeur par défaut. Grâce à l'affaiblissement de ε au fil du temps, le modèle parvient à converger au bout d'environ 600 épisodes mais présente de faibles performances avec une récompense moyenne autour de 1 (figure 5).

Configuration 2 Dans l'optique de faire apprendre au modèle plus vite, on étudie la convergence lorsque $\alpha = 1$, à l'instar de la précédente expérimentation. Le modèle converge au bout d'en moyenne 400 épisodes et ses performances augmentent jusqu'à 5 de récompense moyenne (figure 6). Cependant celles-ci ne sont pas comparables au pic de 8 obtenu auparavant lorsque $\varepsilon = 0$.

Configuration 3 On décide de désactiver complètement ε après ε-decay-steps, en donnant à ε-end la valeur de 0. Le temps d'apprentissage reste inchangé, mais les performances augmentent, approchant le pic de 8 de récompense moyenne (figure 7). On en conclut qu'avoir $\varepsilon > 0$ nuit à la récompense moyenne après un certain nombre d'étapes.

Configuration 4 On cherche maintenant à évaluer l'impact du facteur d'actualisation γ sur les performances. Avec la valeur précédemment identifiée de $\gamma = 0.5$, le modèle préserve ses performances et son temps d'apprentissage est diminué à environ 300 épisodes comme observé durant la première version de l'algorithme.

3 SARSA

Configuration 1 Comme dans les expérimentations précédentes, on commence par étudier les performances avec les hyperparamètres à leur valeur par défaut. Dans cette configuration, le modèle obtient des performances moyennes avec une convergence au bout d'environ 700 épisodes et une récompense moyenne autour de 4 (figure 9).

Configuration 2 Dans un environnement complètement déterministe comme le nôtre, on décide d'étudier l'impact d'un taux d'apprentissage maximal sur le modèle SARSA. Dans ce cas, les performances sont significativement améliorées, avec un temps d'apprentissage autour des 400 épisodes et une récompense moyenne approchant du pic de 8 (figure 10).

Configuration 3 Enfin, de même que pour les expérimentations précédentes, avec la valeur précédemment identifiée $\gamma = 0.5$, le modèle préserve ses performances et son temps d'apprentissage est diminué à environ 300 épisodes comme observé durant la première version de l'algorithme (figure 11).

[1] On considère la moyenne des récompenses par intervalle de 100 sur 1000 épisodes d'au plus 200 actions.

4 Annexes

Toutes les figures sont également disponibles sous le répertoire `export/figures` de ce dépôt.

4.1 Q-Learning

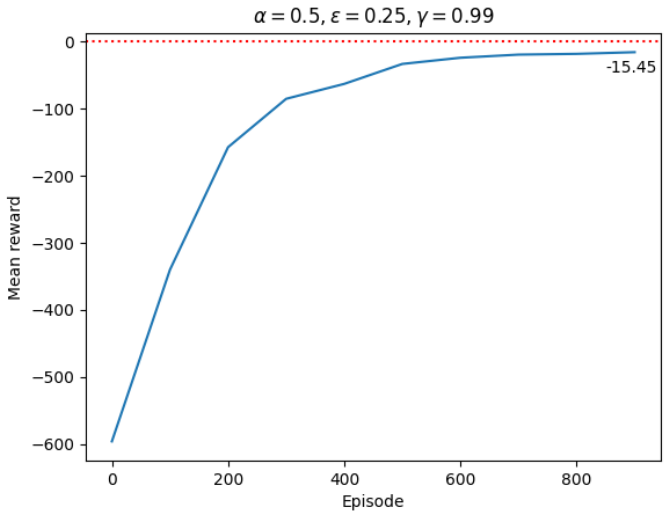


Figure 1: Résultats pour "Configuration 1" de Q-Learning

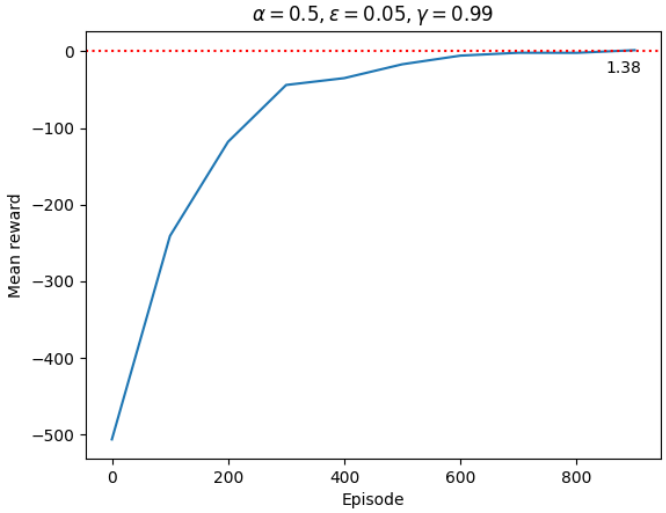


Figure 2: Résultats pour "Configuration 2" de Q-Learning

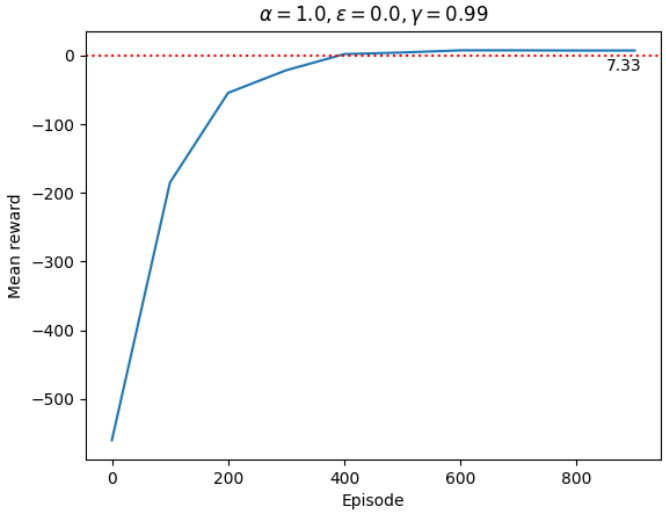
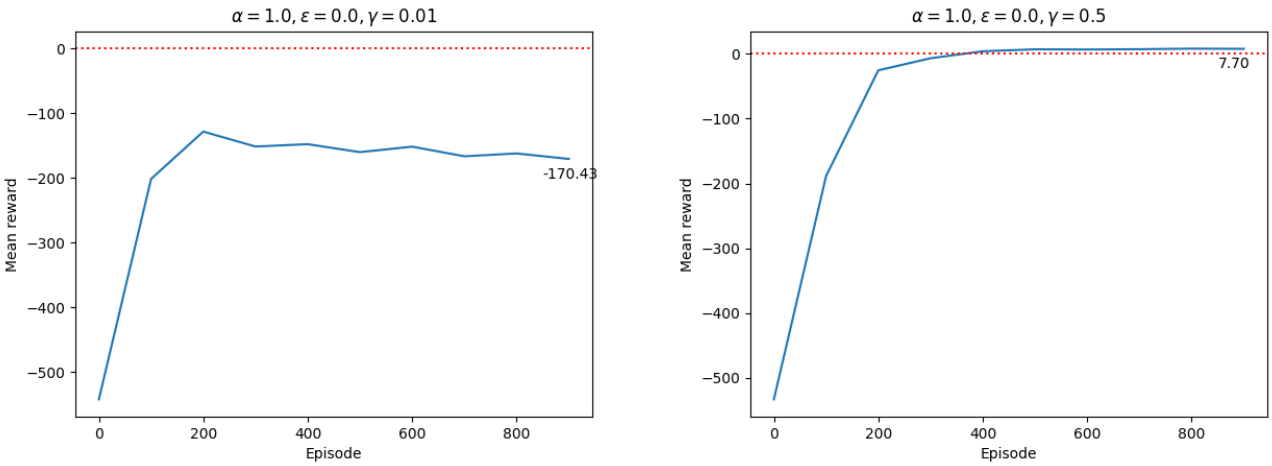


Figure 3: Résultats pour "Configuration 3" de Q-Learning



(a) Configuration 4a

(b) Configuration 4b

Figure 4: Résultats pour la "Configuration 4" de Q-Learning

4.2 Q-Learning ε-scheduling

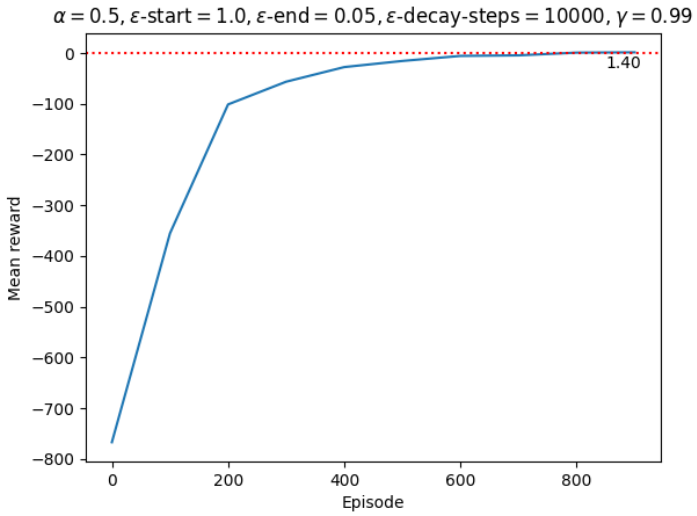


Figure 5: Résultats pour "Configuration 1" de Q-Learning ε-scheduling

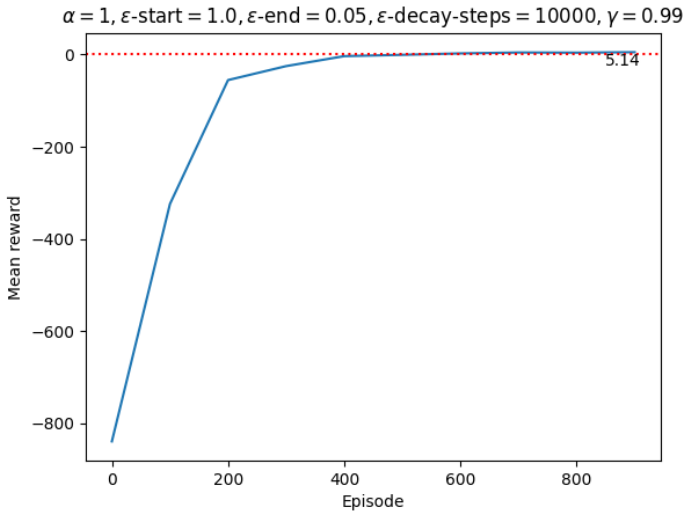


Figure 6: Résultats pour "Configuration 2" de Q-Learning ε-scheduling

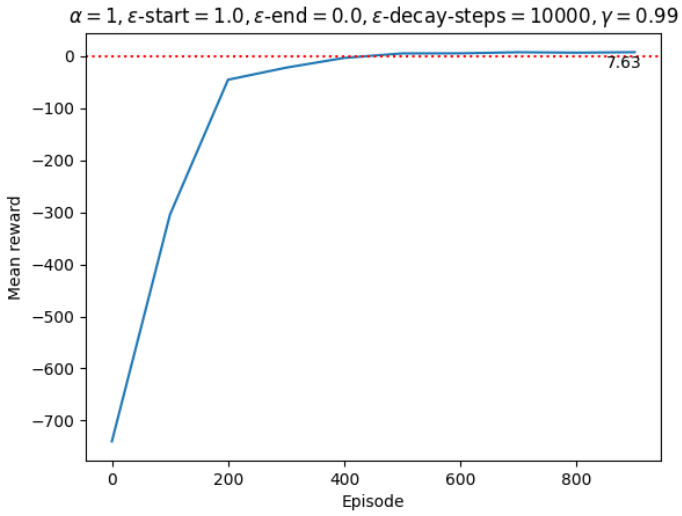


Figure 7: Résultats pour "Configuration 3" de Q-Learning ε-scheduling

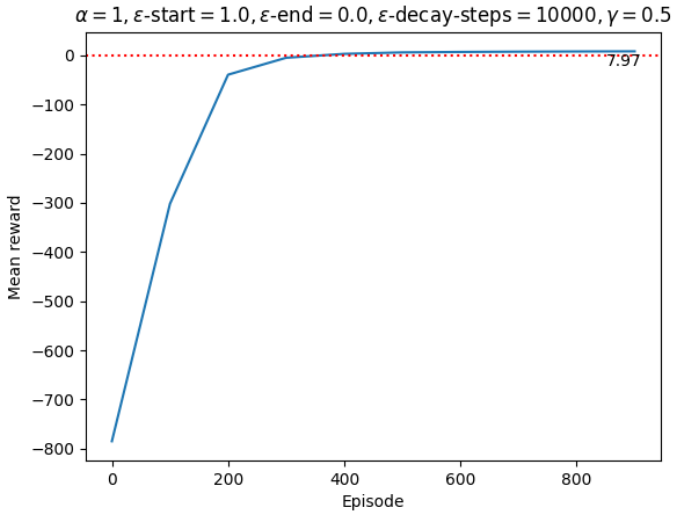


Figure 8: Résultats pour "Configuration 4" de Q-Learning ε-scheduling

4.3 SARSA

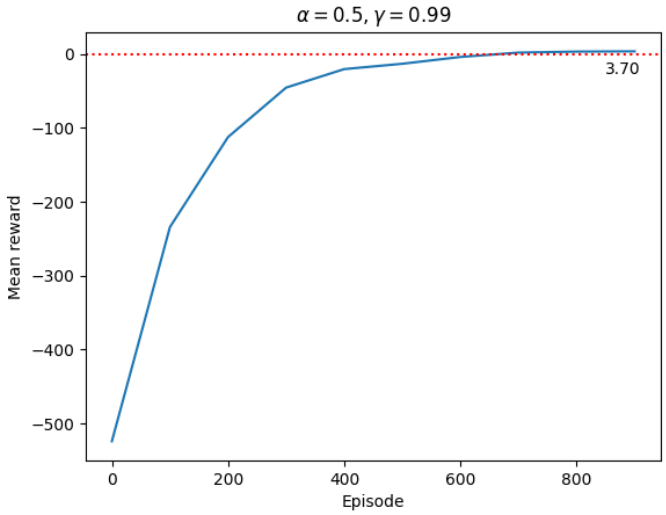


Figure 9: Résultats pour "Configuration 1" de SARSA

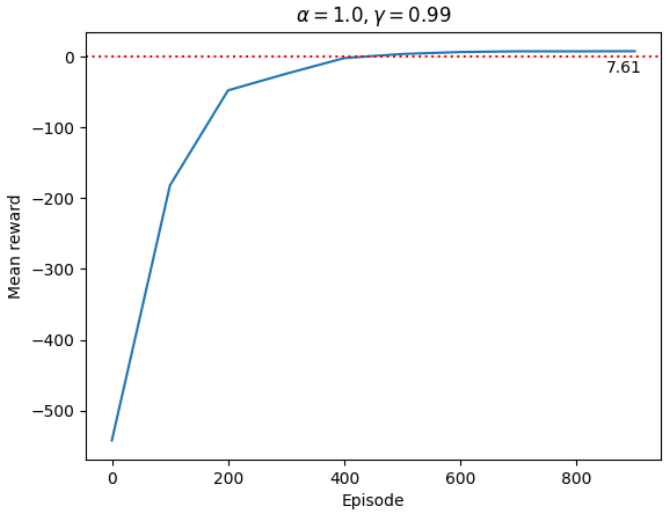


Figure 10: Résultats pour "Configuration 2" de SARSA

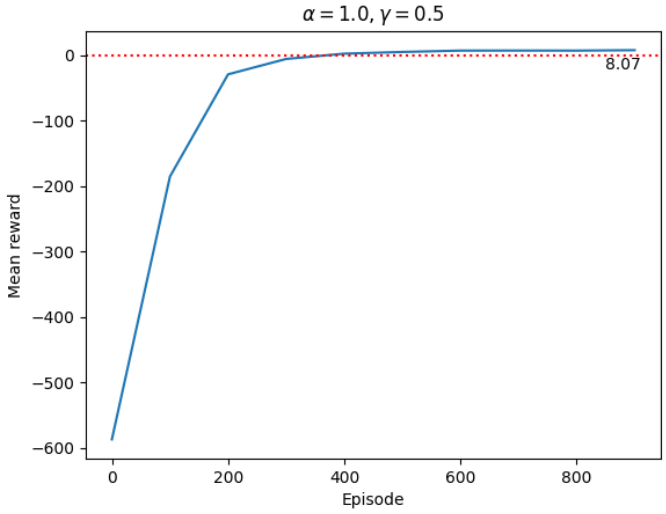


Figure 11: Résultats pour "Configuration 3" de SARSA