UNIVERSITÉ CATHOLIQUE DE LILLE

ÉCOLE DU NUMÉRIQUE BY FGES

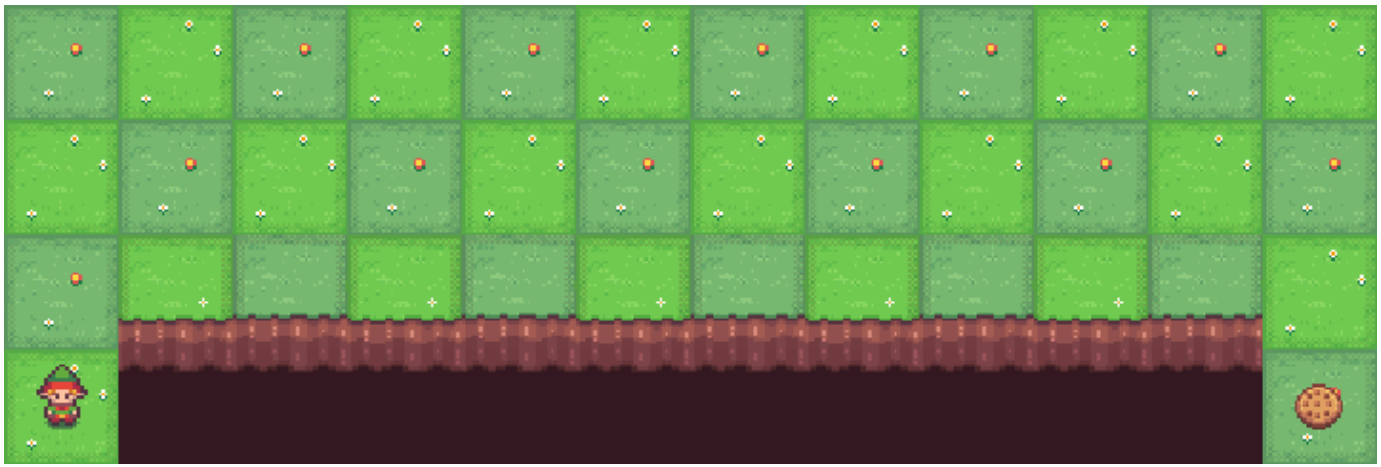
Apprentissage par renforcement

Sarsa VS Qlearning

Ali ABDALLAH
Mohamad Ali GHADDAR

Modèle

Cliff Walking



Le jeu commence avec le joueur à l'emplacement S36 du monde de la grille 4x12 avec le but situé à S47. Si le joueur atteint l'objectif, l'épisode se termine.

Si le joueur se déplace vers un emplacement de falaise (s37 --> s46), il revient à l'emplacement de départ.

Qualité de la démarche

10000 éxecutions pour chaque script

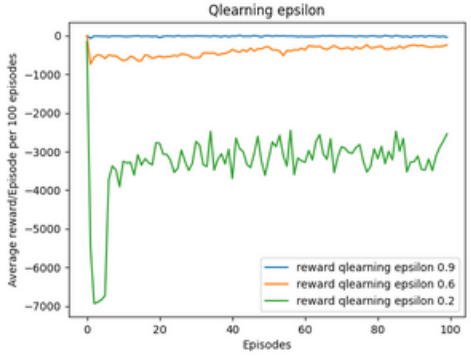
Comparaison des récomponses / episode pour les paramétres différentes pour chaque algorithmes et entre les 2 algorithmes sur les paramétres suivant :
Epsilon = 0.9 / Gamma = 0.9 / Alpha = 0.1

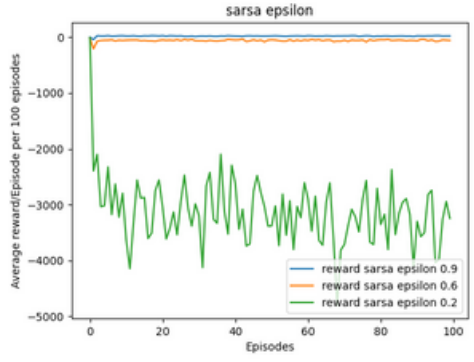
Paramètres:

1. Epsilon (ε)
   - Valeurs : 0.9 (haute exploration), 0.6 (équilibre), 0.2 (haute exploitation)
2. Gamma (γ)
   - Valeurs : 0.9 (importance aux récompenses futures), 0.6 (équilibre), 0.2 (importance aux récompenses immédiates)
3. Alpha (α)
   - Valeurs : 0.1 (apprentissage lent et stable), 0.5 (équilibre), 0.9 (apprentissage rapide)

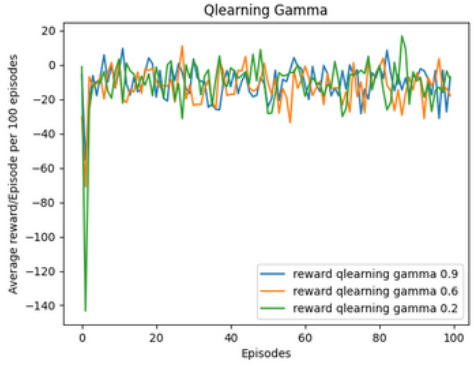
Résultats

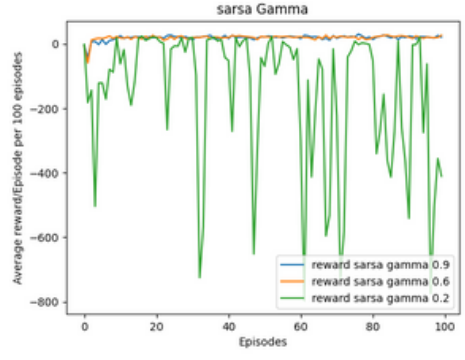
Epsilon :



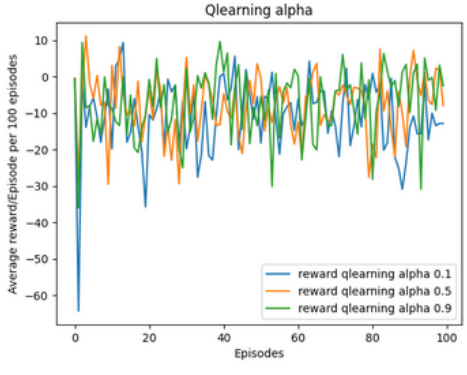


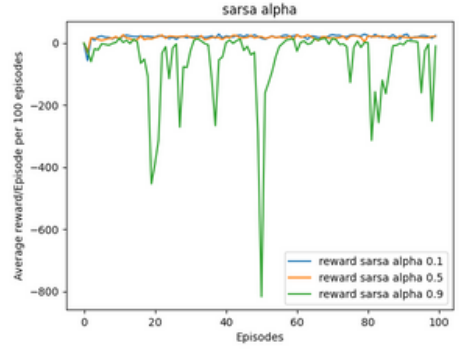
Gamma :



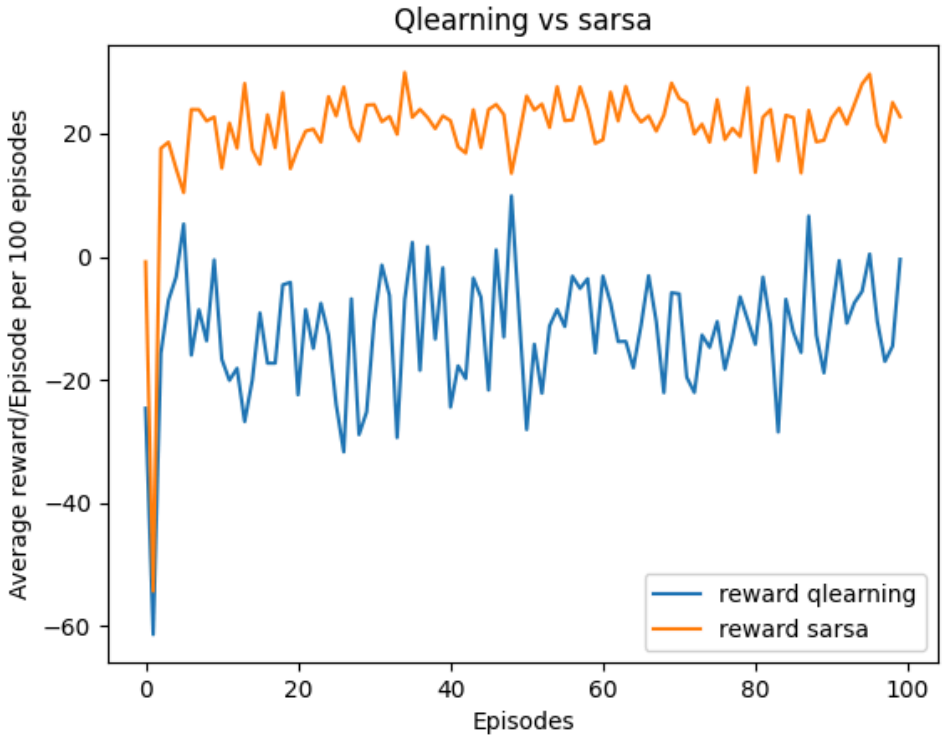


Alpha :



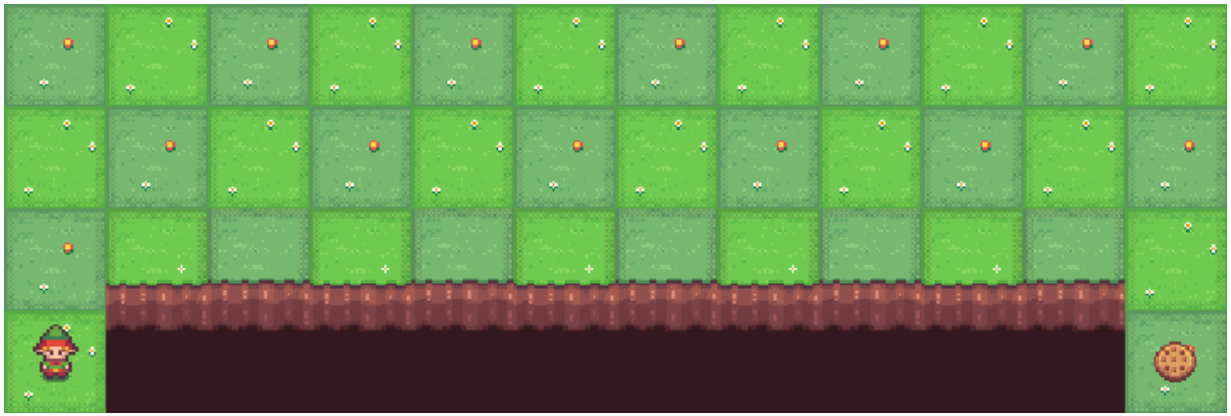


Résultats



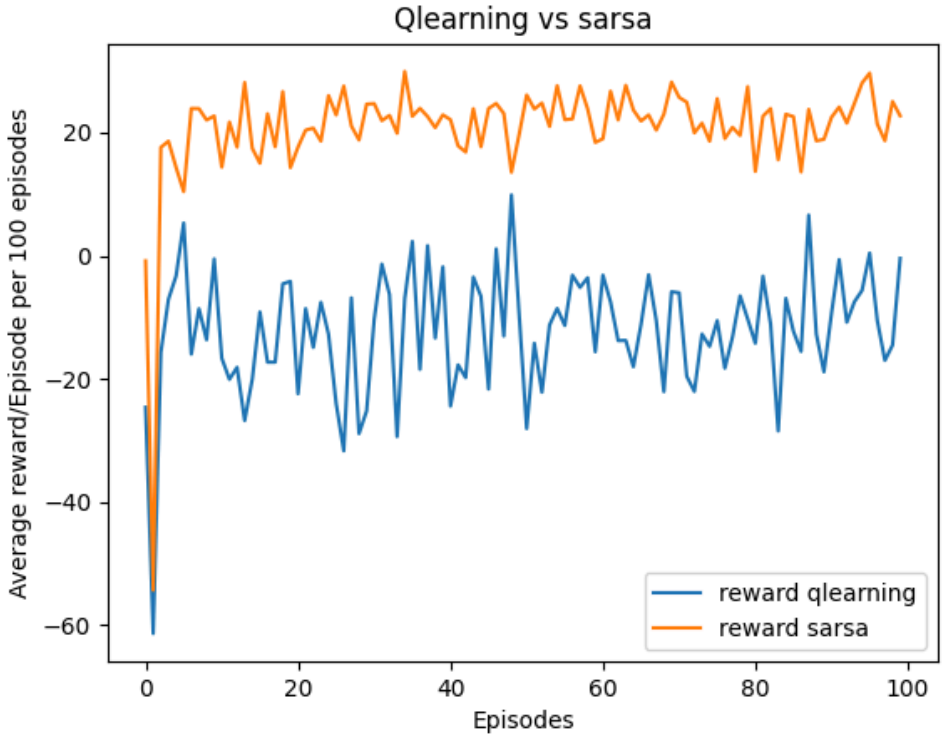
- SARSA a tendance à obtenir des récompenses moyennes plus élevées par rapport au Q-Learning.
- Q-Learning montre plus de fluctuations dans les récompenses au fil des épisodes.

Question 1



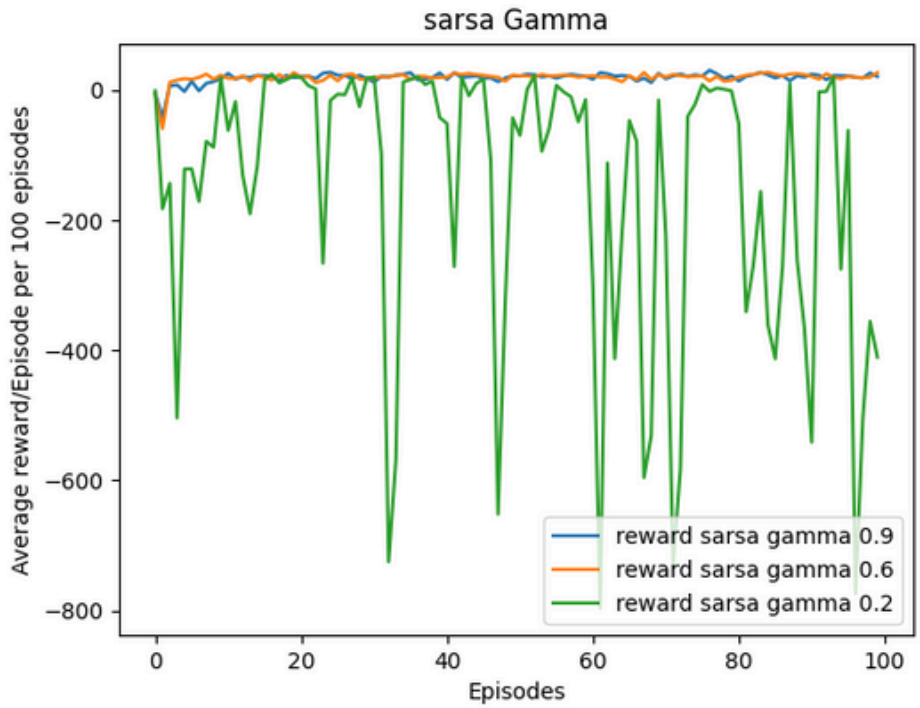
Si on change le point de départ, ça n'affecte pas l'apprentissage des 2 algorithmes parce que la liste des actions et des états est déjà définie.

Question 2



SARSA utilise une stratégie d'apprentissage on-policy, ce qui signifie qu'il apprend la valeur de la politique suivie, y compris les actions exploratoires. Q-learning, étant off-policy, apprend la valeur de la meilleure politique possible, ce qui peut parfois conduire à une exploration excessive et à des valeurs de récompense inférieures à court terme.

Question 3



Un faible gamma (0.2) signifie que l'algorithme SARSA privilégie fortement les récompenses immédiates au détriment des récompenses futures. Cela peut conduire à des politiques qui choisissent des actions sous-optimales sur le long terme, causant des chutes brutales dans les performances. l'algorithme peut se retrouver à constamment changer de politique en réponse aux fluctuations des récompenses immédiates, ce qui provoque une instabilité dans les récomponses.


THANK
YOU