Rapport

Choix d'implémentation

Les algorithmes utilisent une représentation tabulaire des Q-values (un dictionnaire indexé par état et action). Les hyperparamètres initiaux choisis sont :

- taux d'apprentissage (α) : 0.5
- facteur de réduction (γ) : 0.99
- exploration ε:
  - constant pour Q-Learning et SARSA (ε = 0.25)
  - décroissant linéairement pour Q-Learning-EpsScheduling (de 1.0 ⟶ 0.05 sur 10 000 pas)

La fonction `play_and_train` applique l'apprentissage en ligne à chaque étape: l'agent choisit une action via une stratégie ε-gloutonne, puis met à jour la valeur Q correspondante.

Pour la création des vidéos, un épisode final sans exploration est joué après l'entraînement afin de montrer le comportement appris par chaque agent.

Résultats

L'entraînement se fait sur 1000 épisodes. Nous suivons la moyenne des récompenses sur les 100 derniers épisodes.

Q-Learning: Convergence rapide vers des récompenses positives ⟶ politique efficace

Q-Learning avec epsilon scheduling: Performance légèrement meilleure à long terme grâce à une exploration contrôlée

SARSA: Apprentissage plus conservateur ⟶ progression plus lente

Ces différences viennent du fait que Q-Learning est hors-politique (il se dirige vers la meilleure action possible), tandis que SARSA est en-politique (il tient compte de la stratégie réellement suivie, donc de l'exploration).

Conclusion

Les trois agents parviennent à résoudre l'environnement Taxi-v3. Néanmoins :

Q-Learning-EpsScheduling offre le meilleur compromis exploration/exploitation SARSA apprend des politiques plus sûres mais moins optimales

Les vidéos enregistrées démontrent visuellement la capacité des agents à planifier leurs déplacements et atteindre l'objectif efficacement.

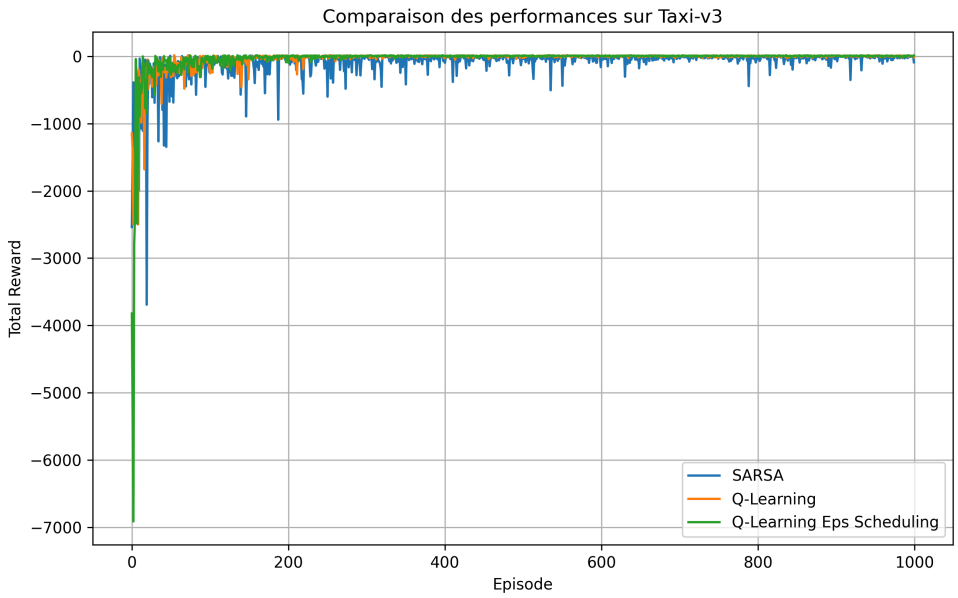

Figure 1: performance