Mini-projet 5

GUERIN Cyril, POT Eline

1 DQN versus PPO on Lunar Lander

1.1 DQN

En recodant l'algorithme DQN et en utilisant les hyperparamètres ci-dessous, nous avons pu tester ses performances plusieurs fois.

```
Hyper-paramètres :
    buffer size: 100_000
    batch size: 400
    discount_factor: 0.99
    updates rate: 50
    learning rate: 0.0001
    nombre d'updates par episode: 10
```

Pour Lunar Lander, on fixe le score maximal à 200 en moyenne sur les 200 derniers épisodes de l'agent. L'algorithme s'arrête seulement lorsqu'il atteint ce seuil.

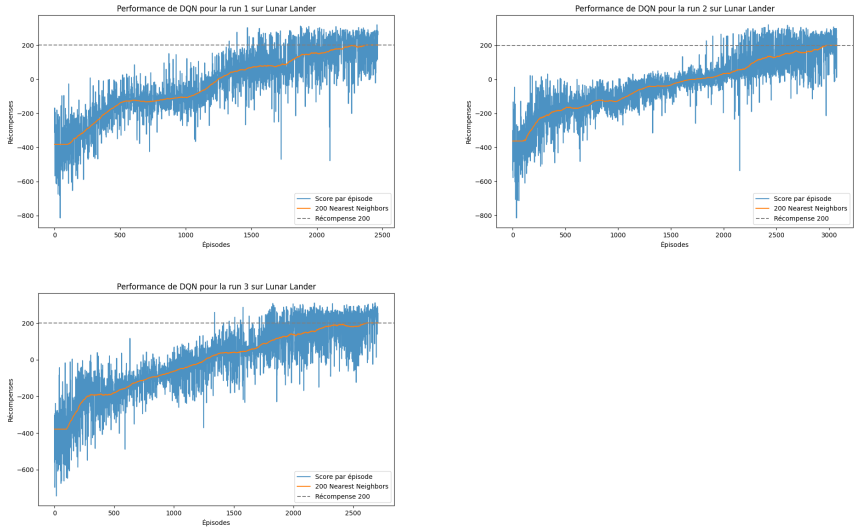

Figure 1: Performances de DQN sur Lunar Lander

1.2 PPO

En recodant l'algorithme PPO et en utilisant les hyperparamètres ci-dessous, nous avons pu tester ses performances plusieurs fois.

```
Hyper-paramètres :
    batch size (pour update value): 400
    batch epochs: 10
    policy epochs: 3
    discount_factor: 0.99
    learning rate: 0.0005
    delta: 0.001
    nombre de trajectoires par épisode: 50
```

Pour Lunar Lander, on fixe le score maximal à 200 en moyenne sur les 200 derniers épisodes de l'agent. L'algorithme s'arrête seulement lorsqu'il atteint ce seuil.

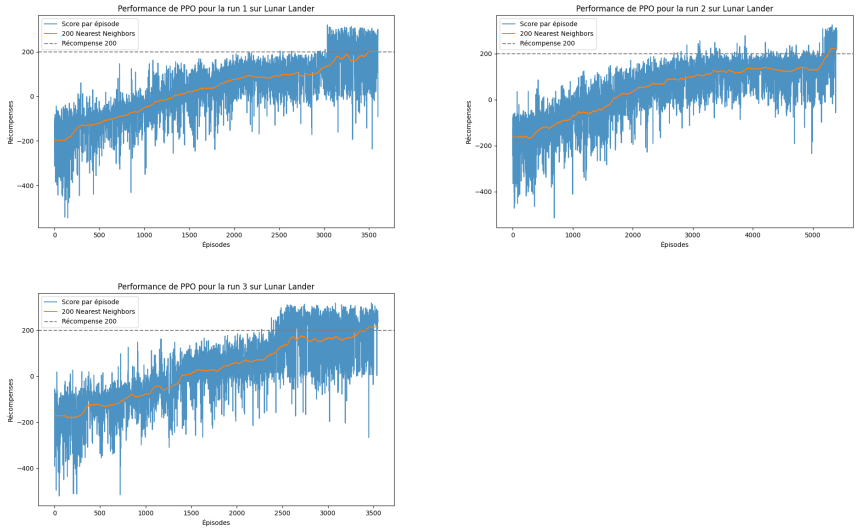

Figure 2: Performances de PPO sur Lunar Lander

1.3 Comparaison des deux algorithmes

De manière générale, on constate que l'algorithme DQN permet une convergence plus rapide en nombre d'épisodes vers le score maximal moyen.
Nous avions aussi remarqué que l'algorithme PPO pouvait parfois avoir du mal à atteindre ce seuil, et que sa moyenne sur les 200 derniers épisodes restait aux alentours de 150. Cela s'explique par le fait que l'algorithme PPO ne promet par une convergence optimale, et l'agent peut rester coincé en un maximum local.
Voici quelques éléments statistiques obtenus sur les performances de DQN et PPO sur l'environnement Lunar Lander:

```
Mean DQN: -29.382440683653112
Mean PPO: -26.951718788289035
Median DQN: -24.687595393197
Median PPO: -20.538851688148995
Std DQN: 196.835227
Std PPO: 117.553636
```

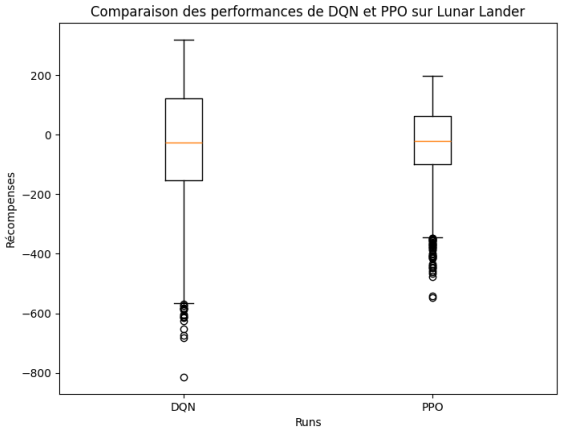

Figure 3: Box plot des performances DQN et PPO

Sur la figure 3, on constate que la médiane des scores obtenus par DQN et PPO sont semblables, même si PPO est un peu meilleur en terme de moyenne et de médiane. On remarque auss une variance plus élevée pour les performances de DQN par rapport à celles de PPO. Cela indique que les performances de DQN sont plus dispersées celles de PPO. En effet, DQN utilise une stratégie epsilon-greedy pour l'exploration, qui dépend du paramètre epsilon. La variation de la stratégie d'exploration peut entraîner des fluctuations dans les performances.
Pour obtenir le profil de performance de la figure 4, on a tronqué les relevés de score pour avoir le même nombre de données et on a moyenné sur les différentes

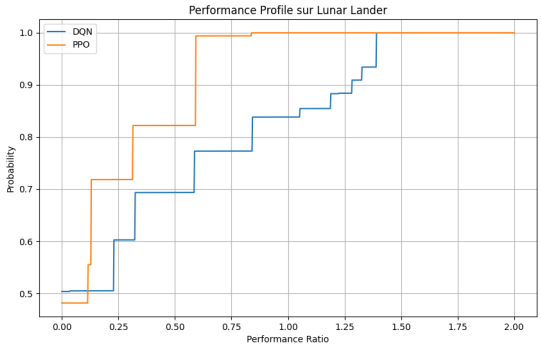

Figure 4: Profils de performance de DQN et PPO

runs pour DQN et PPO. Ici, on voit que PPO a un meilleur profil de performance que DQN.

2 DQN versus PPO on Cartpole

2.1 DQN

En recodant l'algorithme DQN et en utilisant les hyperparamètres ci-dessous, nous avons pu tester ses performances plusieurs fois.

```
Hyper-paramètres :
    buffer size: 100_000
    batch size: 400
    discount_factor: 0.99
    updates rate: 50
    learning rate: 0.0001
    nombre d'updates par episode: 10
```

Pour Cartpole, on fixe le score maximal à 450 en moyenne sur les 200 derniers épisodes de l'agent. L'algorithme s'arrête seulement lorsqu'il atteint ce seuil.

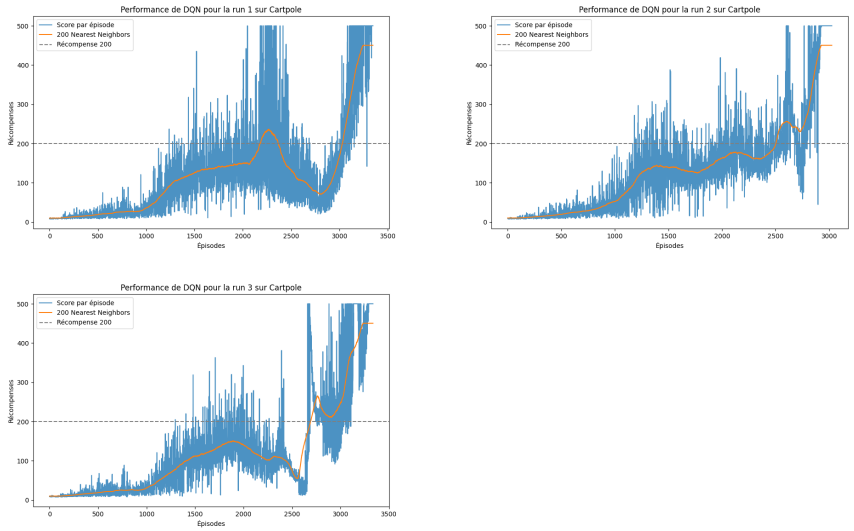

Figure 5: Performances de DQN sur Cartpole

2.2 PPO

En recodant l'algorithme PPO et en utilisant les hyperparamètres ci-dessous, nous avons pu tester ses performances plusieurs fois.

```
Hyper-paramètres :
    batch size (pour update value): 400
    batch epochs: 10
    policy epochs: 3
    discount_factor: 0.99
    learning rate: 0.0005
    delta: 0.001
    nombre de trajectoires par épisode: 50
```

Pour Cartpole, on fixe le score maximal à 450 en moyenne sur les 200 derniers épisodes de l'agent. L'algorithme s'arrête seulement lorsqu'il atteint ce seuil.

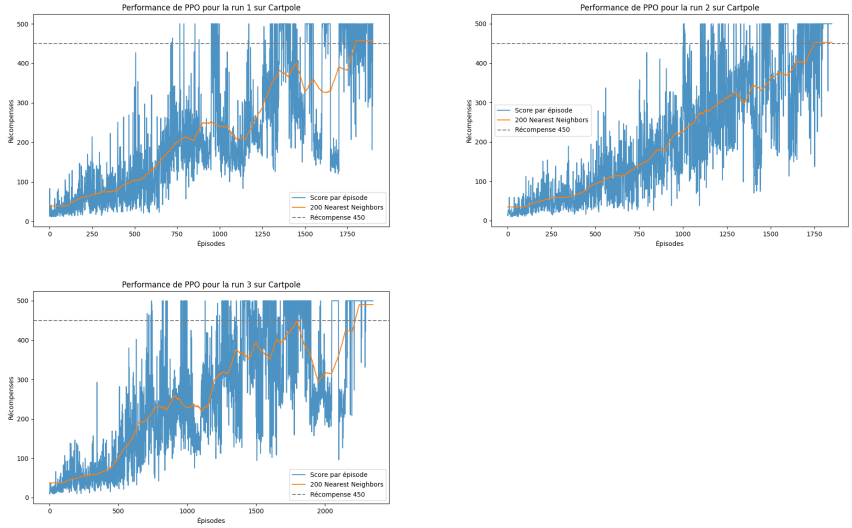

Figure 6: Performances de PPO sur Cartpole

2.3 Comparaison des deux algorithmes

Ici, on constate que l'algorithme PPO permet une convergence plus rapide en nombre d'épisodes vers le score maximal moyen.
Voici quelques éléments statistiques obtenus sur les performances de DQN et PPO sur l'environnement Cartpole:

```
Mean DQN: 61.82526315789474
Mean PPO: 225.28578947368422
Median DQN: 33.0
Median PPO: 186.0
Std DQN: 60.351183
Std PPO: 159.082309
```

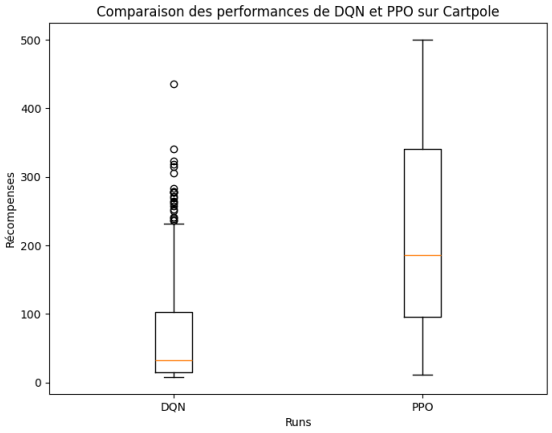

Figure 7: Box plot des performances DQN et PPO

Sur la figure 7, on constate que DQN fait moins bien que PPO en terme de médiane et de moyenne sur les scores, et que PPO a une variance plus importante.
Pour obtenir le profil de performance de la figure 8, on a tronqué les relevés de score pour avoir le même nombre de données et on a moyenné sur les différentes runs pour DQN et PPO. Ici, on voit que DQN a un meilleur profil de performance que PPO.

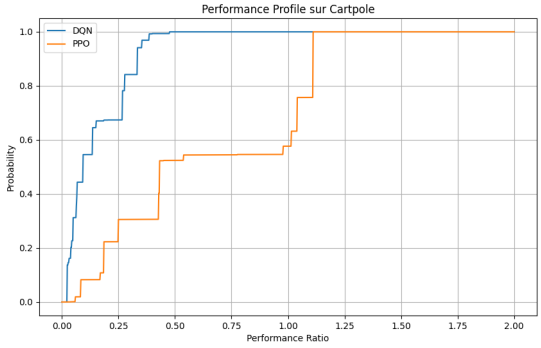

Figure 8: Profils de performance de DQN et PPO