Apprentissage par renforcement: Méthode Acteur-Critique

Sami EL KATEB Thomas PAUL

1 Introduction

Acteur-critique est une méthode d'apprentissage par renforcement de type policy-gradient. Contrairement à d'autres méthodes de ce type, acteur-critique utilise une approximation de la fonction de valeur d'état $\hat{v}(s_t)$ et la différence temporelle pour apprendre la politique. Dans ce document, nous détaillerons le fonctionnement de la méthode acteur-critique, ses avantages et ses inconvénients. Puis, nous examinerons une implémentation spécifique pour le jeu de Pong et explorerons des pistes d'amélioration [1].

2 Contexte

La méthode acteur-critique se base sur deux composants principaux : l'acteur et le critique. Ces deux composants interagissent pour apprendre la meilleure politique possible pour un environnement donné.

L'acteur est responsable du choix de l'action, il représente la politique $\pi(a|s)$. L'objectif de l'acteur est d'ajuster la politique pour maximiser les récompenses futures.

Le critique a pour objectif d'approximer la fonction de valeur d'un état $\hat{v}(s_t)$. Celle-ci estime la récompense réduite attendue[1] à partir d'un état. Le critique utilise la différence temporelle, qui est l'écart entre l'estimation actuelle de la récompense et celle obtenue, pour mettre à jour ses estimations.

Ce type d'architecture rappelle les réseaux de neurones de la famille des GANs[2] où un générateur et un discriminateur interagissent pour s'améliorer mutuellement.

Un exemple de pseudocode pour l'algorithme acteur-critique est donné dans la figure ci-dessous.

Algorithm 1 Algorithme Acteur-Critique

```
for each episode do
    Initialize s_t, R ← [ ], S ← [ ], A ← [ ]
    for each time step do
        a_t ← π(s)                ▷ Action aléatoire a_t sélectionnée selon la distribution définie par l'acteur
        Take action a_t, observe s_t and r_t
        append a_t to A, r_t to R, s_t to S
        if episode is done then
            G ← discounted R
            advantages ← G − v̂(S)
            Fit π using S and A with advantages as sample weights
            Fit v̂ using S and G
        end if
    end for
end for
```

À chaque épisode étape l'acteur retourne une distribution de probabilités en fonction de l'état. L'algorithme acteur-critique utilise alors la préférence d'actions soft-max pour trouver un équilibre entre l'exploration et l'exploitation. Contrairement à l'algorithme ϵ-greedy, cette technique consiste à choisir une action aléatoirement en fonction des probabilités déterminées par l'acteur. [2]

Une fois l'épisode terminé, le critique évalue les différents états de l'épisode et retourne les gains réduits attendus. Ces gains sont ensuite soustraits aux gains réels pour évaluer les avantages obtenus par l'acteur. Les

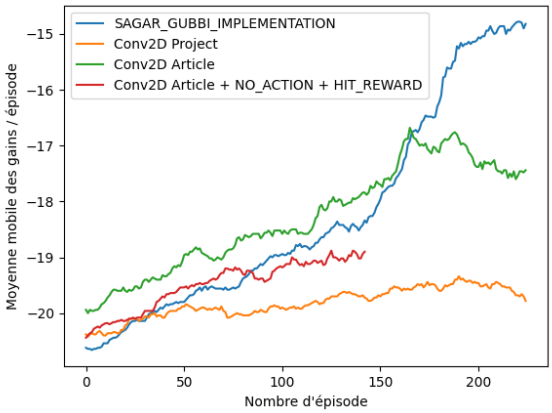
[1] Expected Discounted Return
[2] Generative Adversarial Network

avantages peuvent ensuite être utilisés pour améliorer la politique de l'acteur et les gains réels sont utilisés pour entrainer le critique.

3 Application au jeu de Pong

Pour appliquer l'algorithme Acteur-Critique au jeu de Pong nous avons pris pour référence le code source de *Learning to Play with Pong*[1] que nous avons adapté à la nouvelle api de gymnasium.

Nous avons par la suite tenté plusieurs modifications dans le but d'améliorer les performances de l'algorithme. Pour les évaluer, nous avons analysé l'évolution de la moyenne mobile des gains par épisode. Chaque entrainement a été effectué sur 300 épisodes et a duré en moyenne 3 heures. Bien que le nombre d'épisode nécessaire pour vaincre l'IA soit plus élevé, nous avons choisit un nombre d'épisode nous permettant d'entrainer plusieurs modèles pour pouvoir ainsi les comparer.



Notre première modification a consisté à reprendre le même réseau de convolution que celui utilisé pour le projet Pong. Cependant, cette approche n'a pas été concluante. En effet, pour le projet Pong, nous avons utilisé un réseau convolutionnel avec un nombre restreint de paramètres. Nous avons pris cette décision car nous disposions d'un nombre limité de données labellisées et voulions diminuer le risque de surapprentissage.

Ensuite, nous avons choisit d'utiliser le réseau de convolution décrit dans l'article *Learning to Play Pong using Gradient Learning*[3]. Celui-ci a donné de meilleurs résultats que le réseau de convolution de notre projet, cependant, il n'a pas surpassé les performances de l'implémentation originale.

Finalement, nous avons choisit d'entrainer l'acteur et le critique à la fin de chaque incrément de temps et non à la fin de chaque épisode. Cette approche est sensée être plus efficace en terme de d'échantillons.

4 Avantages & Inconvénients

4.1 Avantages

La méthode Acteur-Critique combine les avantages des méthodes de type value-based (Q-Learning, Sarsa) et policy-based :

- Elle offre une meilleure convergence que les méthodes basées uniquement sur la politique, grâce à l'utilisation de la différence temporelle.
- Elle ne nécessite pas de données labellisés contrairement à l'approximation de la fonction de valeur par apprentissage supervisé.

4.2 Inconvénients

Elle présente cependant quelques inconvénients :

- Elle nécessite d'apprendre à la fois une fonction de valeur et une politique, ce qui augmente la complexité et le coût de calcul de l'algorithme.
- Des erreurs dans l'estimation de la fonction de valeur par le critique peuvent entraîner de mauvaises mises à jour de la politique par l'acteur.

5 Conclusion

Références

[1] Sagar Gubbi. *Learning to play Pong*. 2019. URL : https://github.com/s-gv/pong-keras/blob/master/pong-actor-critic.py (visité le 15/10/2023).

[2] Richard S. Sutton et Andrew G. Barto. *Reinforcement Learning : An Introduction*. Second. The MIT Press, 2018.

[3] Somnuk Phon-Amnuaisuk. *Learning to Play Pong using Policy Gradient Learning*. 2018. arXiv : 1807.08452 [cs.LG].