

# 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.

## 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<sup>1</sup> 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<sup>2</sup> 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  $s_t$ ,  $R \leftarrow []$ ,  $S \leftarrow []$ ,  $A \leftarrow []$ 
  for each time step do
     $a_t \leftarrow \pi(s)$  ▷ 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
       $G \leftarrow \text{discounted } R$ 
       $advantages \leftarrow G - \hat{v}(S)$ 
      Fit  $\pi$  using  $S$  and  $A$  with  $advantages$  as sample weights
      Fit  $\hat{v}$  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  $\epsilon$ -greedy, cette technique consiste à choisir une action aléatoirement en fonction des probabilités déterminées par l'acteur. [1]

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 des gains réels pour évaluer les avantages obtenus par l'acteur. Les

---

<sup>1</sup>Expected Discounted Return

<sup>2</sup>Generative Adversarial Network

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

## Références

- [1] Richard S. SUTTON et Andrew G. BARTO. *Reinforcement Learning : An Introduction*. Second. The MIT Press, 2018.