Cahier des Charges - CEM vs RWR

Réalisé par
Damien LEGROS
Hector KOHLER

Dans le cadre du cours
PANDROIDE

Travail encadré par
Oliver SIGAUD

Master ANDROIDE
Université Sorbonne Université

1 Présentation du sujet

1.1 Contexte

Les méthodes évolutionnaires et les méthodes d'apprentissage par renforcement constituent deux alternatives pour résoudre des problèmes de recherche d'une politique optimale sur des problèmes où les actions sont continues.

1.2 Objectifs

L'objet de ce projet est de se livrer à une comparaison systématique des propriétés de ces deux approches dans un environnement et de comprendre les différences de performances entre les deux types de méthodes.

Dans un premier temps, nous prendrons en main l'interface d'OpenAI Gym et nous implémenterons les méthodes que nous utiliserons.

Dans un second temps, nous effectuerons des analyses via différents outils pour comparer les performances et visualiser les moments critiques montrant la divergence entre celles-ci.

1.3 Méthodes

Pour effectuer la comparaison nous nous intéresserons à deux méthodes simples. Pour les méthodes évolutionnaires nous nous focaliserons sur CEM (Cross Entropy Method) et sur PG (Policy Gradient) pour les méthodes d'apprentissage.

1.4 Environnement

Pour l'environnement, nous utiliserons Pendulum-v0 distribué par OpenAI :

Ce benchmark est disponible sur Gym : https ://gym.openai.com/envs/Pendulum-v0/

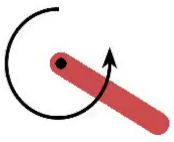

FIGURE 1 – Environnement Pendulum

Tout d'abord définissons les variables du problème du pendule inversé.

Soit les variables suivantes :

— L'accélération de la gravité : g
— La taille de la barre du pendule : l
— La masse du pendule : m
— La vitesse maximum du pendule : $max_{vitesse}$
— La vitesse maximum de l'effort de l'action : max_{effort}
— L'angle du pendule : θ
— La vélocité du pendule : θ_{dot}
— La vélocité de l'effort de l'action : u
— Le coefficient de l'accélération : dot

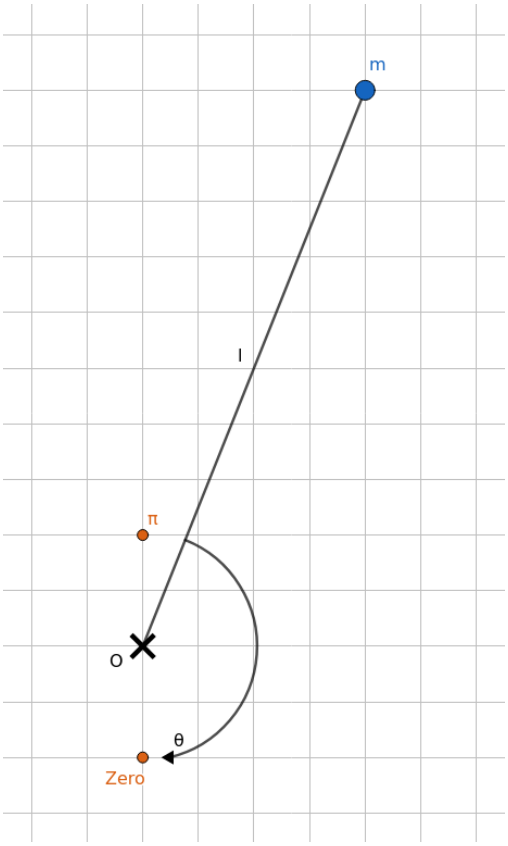

FIGURE 2 – Pendule

L'environnement pendulum commence à une position aléatoire et a pour but de garder le pendule en haut et en équilibre le plus longtemps possible.

Dans l'environnement pendulum certaines variables sont constantes :

— $g = 1.0$
— $l = 1.0$
— $m = 1.0$
— $max_{vitesse} = 8.0$
— $max_{effort} = 2.0$
— $dot = 0.5$

La position aléatoire initialise θ dans l'intervalle $[-\pi,\pi]$ et θ_{dot} dans l'intervalle $[-1.0,1.0]$

Il n'y a qu'une action :

— Aller à gauche ou à droite avec une vélocité. Cette action est comprise dans l'intervalle $[-2.0, 2.0]$. Nombre négatif pour un effort du pendule vers la droite avec une vélocité allant de 0.0 à 2.0 et positif pour un effort du pendule vers la gauche avec une vélocité allant aussi de 0.0 à 2.0.

Il y a 3 observations :

— Le cosinus de θ compris dans l'intervalle $[-1.0, 1.0]$
— Le sinus de θ compris dans l'intervalle $[-1.0, 1.0]$
— θ_{dot} compris dans l'intervalle $[-8.0, 8.0]$

Pour les calculs suivants θ est normalisé dans l'intervalle $[-\pi,\pi]$.

L'équation du reward est la suivante :

$$-(\theta^2 + 0.1 * \theta_{dot}^2 + 0.001 * action^2)$$

A chaque action θ et θ_{dot} sont mis à jour :

$$\theta_{new} = \theta_{old} + \theta_{dot_{old}} * dot$$

$$\theta_{dot_{new}} = \theta_{dot_{old}} + (-\frac{3 * g}{2 * l} * sin(\theta_{old} + \pi) + \frac{3}{m * 1^2 * u}) * dot$$

Il n'y a pas de statut de fin de l'environnement, le but est de rester à 0 le plus longtemps possible.

Pour faciliter les comparaisons, nous utiliserons toujours 200 épisodes.

2 Analyses à effectuer

2.1 Comparaison des graphes de performance

Une fois les méthodes implémentées, nous visualiserons les performances des méthodes pour trouver à quel moment les performances divergent.

2.2 Visualisation de l'évolution des politiques avec t-SNE

Une fois cela effectuée, nous utiliserons l'outil t-SNE pour effectuer une réduction de dimension nous permettant d'avoir une visualisation plus ludique des divergences de performances entre les méthodes.

2.3 Visualisation de l'évolution des politiques avec Vignettes

Nous utiliserons pour finir l'outil Vignettes qui sera developpé dans un autre projet PANDROIDE pour visualiser directement sur les points critiques de divergences entre les méthodes.

3 Tâches complémentaires

Si le temps nous le permet nous pourrons être amenés à implémenter de nouveaux outils pour effectuer des visualisations complémentaires. Il nous sera aussi possible de modifier les méthodes pour voir la réaction sur les performances ainsi que de chercher une politique experte sur Pendulum.

4 Planning du projet

Semaine	Tâche
25 Janvier	Découverte du Reinforcement Learning
8 Février	Prise en main de OpenAI Gym
22 Février	Implémentation des méthodes
8 Mars	Comparaison des graphes de performances
22 Mars	Visualisation avec t-SNE
2 Avril	Rendu du carnet de bord
5 Avril	Prise en main de Vignettes
12 Avril	Visualisation avec Vignettes
22 Mai	Remise du Rapport
28-29 Mai	Soutenance du Projet

Table 1 – Planning prévisionnel du projet