Sorbonne Université
Master ANDROIDE

Rapport

UE de COCOMA

Félix Savarit
Kim Saïdi

2024

Chapitre 1

1.1 Contexte

L'objectif de ce projet a été d'explorer l'apprentissage par renforcement dans un environnement multi-agent en utilisant des approches personnalisées pour entraîner des agents distincts et observer leur comportement collectif. Nous avons choisi de travailler avec l'environnement Knights-Archers-Zombies (KAZ), une simulation multijoueurs de type "PettingZoo" qui permet d'expérimenter des situations de coopération et de compétition entre agents.

Pour l'entraînement des agents, nous utilisons Stable-Baselines3, une bibliothèque qui fournit des implémentations stables et optimisées de divers algorithmes de RL. Parmi ces algorithmes, nous avons opté pour Proximal Policy Optimization (PPO), qui est bien adapté aux environnements stochastiques et complexes comme KAZ.

1.2 Approche technique

1.2.1 Entraînement séparé

Nous avons choisi d'entraîner les agents séparément afin de favoriser un apprentissage plus efficace. Cette approche consiste à isoler un groupe d'agents en fonction de leur type et à appliquer des algorithmes d'apprentissage par renforcement adaptés à leurs tâches spécifiques. Ainsi, une stratégie d'entraînement distincte a été mise en place pour chaque type d'agent : les **archers** et les **chevaliers**.

Pour les **archers**, un environnement filtré a été utilisé afin de conserver uniquement les agents correspondant à ce rôle. Cela a été possible grâce au wrapper `FilterAgentWrapper`, qui sélectionne exclusivement les agents de type "archer".

Le même procédé a été appliqué pour entraîner les **chevaliers** dans un environnement filtré ne contenant que ce type d'agent. Une fois les environnements prêts, les agents ont été entraînés à l'aide de l'algorithme PPO (Proximal Policy Optimization).

1.2.2 Wrappers Personnalisés

Pour enrichir les interactions des agents avec l'environnement et ajuster leur apprentissage, nous avons conçus et intégrés plusieurs wrappers personnalisés :

Wrapper de récompense basée sur la distance : Ce wrapper attribue aux agents un bonus proportionnel à leur proximité avec une cible précise (ex : un zombie). Par exemple, plus un agent s'approche de sa cible, plus il reçoit de points de récompense. C'est une méthode pour orienter les agents vers des comportements spécifiques, comme la poursuite active ou la défense rapprochée, en fonction de leurs rôles.

Wrapper de récompense partagée : Ici, une partie des récompenses gagnées par un agent est redistribuée entre les autres agents. Ce mécanisme favorise la coopération entre les agents, même si leurs objectifs individuels diffèrent.

Wrapper de filtrage des agents : Ce wrapper a été conçu pour isoler un type particulier d'agent dans l'environnement, comme les archers ou les chevaliers. Grâce à lui, il a été possible de limiter les actions et interactions uniquement aux agents sélectionnés, facilitant ainsi leur entraînement individuel.

En combinant ces différents wrappers, nous avons pu améliorer le comportement des agents et tester différentes dynamiques, comme l'individualité, la coopération, et l'adaptation à des situations variées.

1.2.3 Suivi des performances

Nous avons implémenté un callback `RewardLoggerCallback`, pour enregistrer et tracer les récompenses moyennes au fil du temps. Cela a permis de générer des courbes d'apprentissage.

1.3 Résultats

1.3.1 Environnement simple

Nous allons voir dans ce paragraphe les résultats pour l'environnement standard. Pour évaluer les performances des agents, deux types d'évaluations ont été réalisés : une **évaluation aléatoire**, où les agents n'ont subi aucun entraînement, et une **évaluation après entraînement**, où les agents ont été formés à l'aide de l'algorithme PPO. Les résultats sont présentés ci-dessous.

TABLE 1.1 – Comparaison des récompenses moyennes entre évaluation aléatoire et après entraînement

Agent	Évaluation aléatoire	Après entraînement
Archer 0	0.681	1.126
Archer 1	0.736	0.804
Chevalier 0	0.005	0.088
Chevalier 1	0.002	0.056

Lors de l'évaluation aléatoire, les agents ont joué sans avoir été entraînés. Les résultats montrent que les archers ont obtenu des récompenses modérées, ce qui indique qu'ils réussissent parfois à accomplir certaines tâches sans stratégie précise. En revanche, les chevaliers ont obtenu des récompenses très faibles, ce qui suggère qu'ils ont beaucoup de mal à interagir efficacement avec l'environnement sans apprentissage préalable.

Après l'entraînement avec l'algorithme PPO, les performances des agents se sont améliorées. Les archers ont obtenu des récompenses plus élevées, montrant qu'ils sont capables d'accomplir leurs tâches avec un peu plus de précision et d'efficacité. Les chevaliers ont également montré une nette amélioration, bien qu'ils restent moins performants que les archers. Cependant, leur progression est significative par rapport à l'évaluation aléatoire.

Les résultats obtenus après cet entraînement restent relativement faibles en moyenne, en particulier pour les chevaliers, dont les récompenses sont bien inférieures à celles des archers. Nous allons maintenant comparer ces résultats avec ceux obtenus dans un environnement plus coopératif, grâce à l'ajout des wrappers expliqués précédemment, et avec un entraînement plus spécifique.

1.3.2 Environnement modifié

Dans cette section, nous allons observer les performances des agents dans un environnement modifié, où des mécanismes de coopération et de partage des récompenses ont été introduits grâce aux wrappers décrits précédemment. Deux configurations principales sont étudiées :

— Un **apprentissage séparé**, où chaque agent suit une politique distincte.
— Un **partage des récompenses**, où les récompenses sont mutualisées entre tous les agents.

Apprentissage séparé

Évaluation avec un apprentissage séparé

TABLE 1.2 – Récompenses moyennes pour un apprentissage séparé

Agent	Récompenses moyennes
Archer 0	1.112
Archer 1	0.866
Chevalier 0	0.000
Chevalier 1	0.000

En suivant un apprentissage séparé, les archers obtiennent des récompenses plus élevées que celles observées dans l'environnement simple, indiquant une amélioration de leurs performances. Cependant, les chevaliers n'obtiennent aucune récompense, ce qui montre leur incapacité à interagir efficacement avec l'environnement, même dans cette configuration.

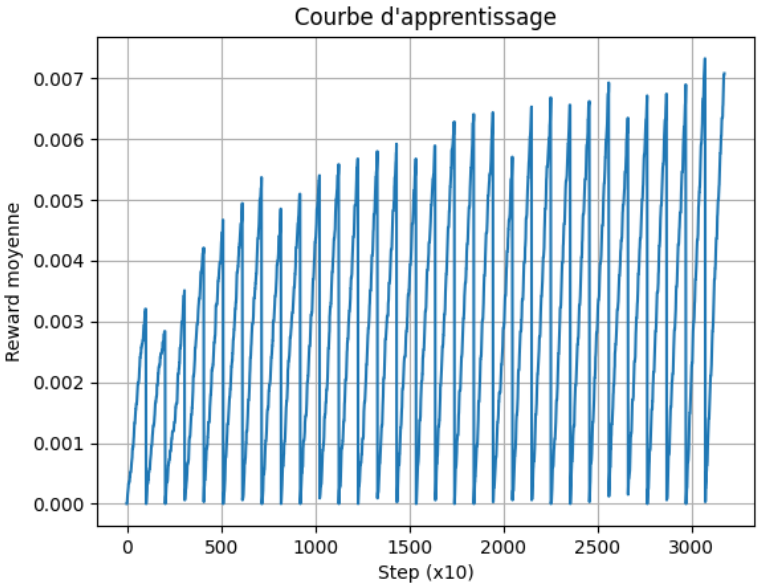

FIGURE 1.1 – Courbe d'apprentissage moyen des archers avec un apprentissage séparé.

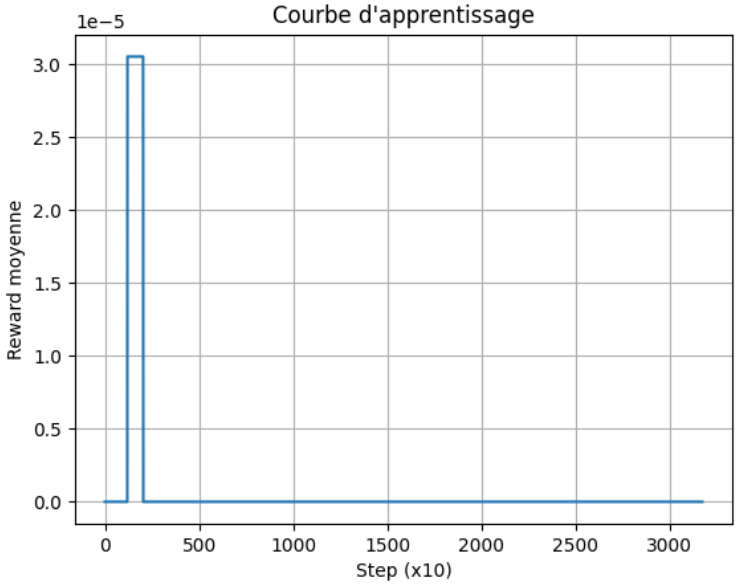

FIGURE 1.2 – Courbe d'apprentissage moyen des chevaliers avec un apprentissage séparé.

Les courbes ci-dessus montrent l'évolution des performances des archers et des chevaliers au cours d'un entraînement séparé. Comme dit précédemment, on voit bien que les chevaliers n'apprennent pas du tout comparé aux archers qui ont une courbe d'apprentissage qui augmente légèrement au cours des jeux.

Évaluation avec partage des récompenses

Dans le tableau ci-dessous se trouvent les résultats sans les partages des récompenses. Ceux-ci se sont fait durant l'entraînement mais pas pendant l'évaluation.

TABLE 1.3 – Récompenses moyennes

Agent	**Récompenses moyennes**
Archer 0	1.297
Archer 1	1.013
Chevalier 0	0.000
Chevalier 1	0.000

On observe que les performances des archers sont légèrement meilleures que celles obtenue avec un entraînement distinct. Les chevaliers, en revanche, n'obtiennent toujours aucune récompense.

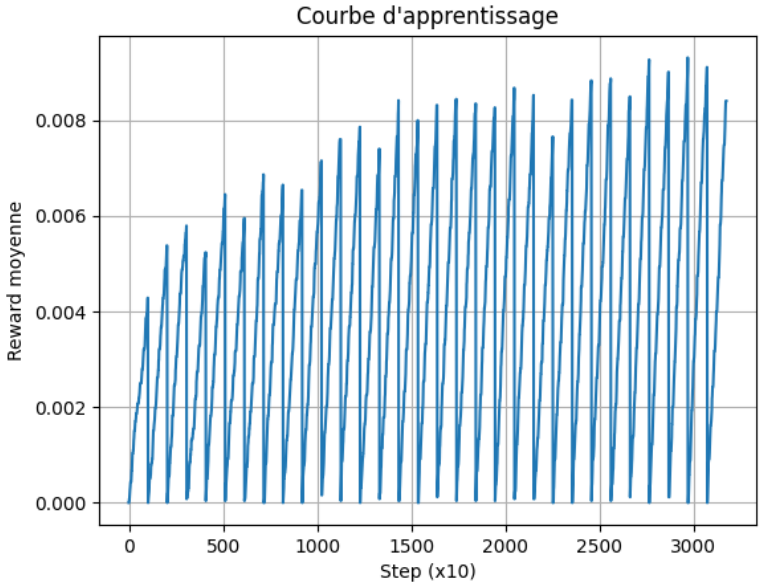

FIGURE 1.3 – Courbe d'apprentissage moyen pour les récompenses partagées.

La courbe ci-dessus illustre les performances des agents dans le scénario des partages de récompense. On y observe la légère amélioration de l'apprentissage décrite précédemment.

En comparaison, la courbe ci-dessous illustre les performances des agents avec un entraînement basique :

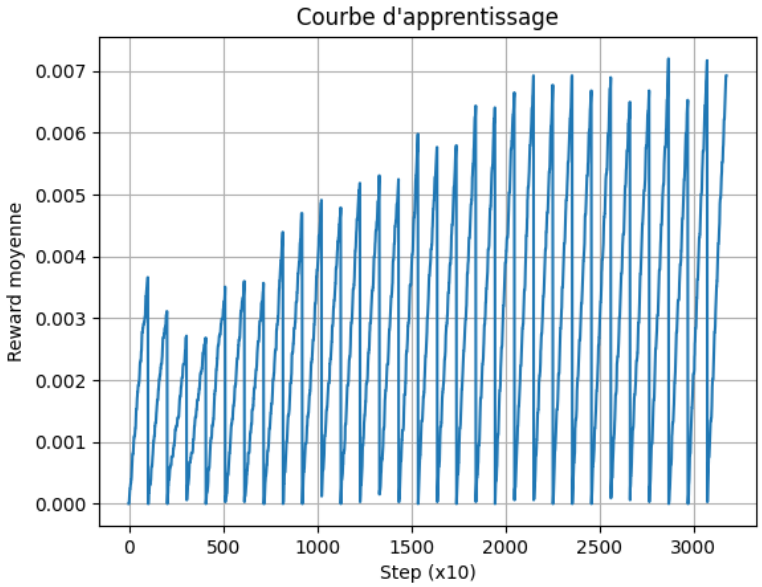

FIGURE 1.4 – Courbe d'apprentissage moyen pour un entraînement classique.

On observe que dans le scénario des récompenses partagé ou des entraînement séparés, il y a une amélioration de l'apprentissage par rapport à l'environnement de base.

1.4 Conclusion

Pour conclure, on note une légère amélioration de l'apprentissage dans l'environnement modifié par rapport à l'environnement de base. En continuant à ajuster l'environnement et à encourager une meilleure coopération entre les agents, il serait possible d'obtenir de meilleurs résultats, notamment avec les chevaliers qui n'ont pas réussi à obtenir de récompense lors de l'évaluation.