Policy Evaluation agent

Eden BELOUADAH

27 novembre 2017

1 Introduction

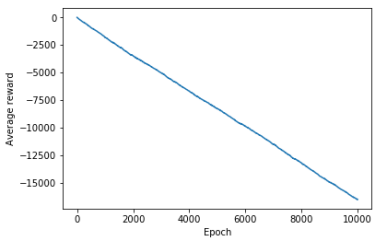
Le but de ce TP est d'implémenter un agent qui a pour tâche l'évaluation d'une politique. Nous avons pris comme exemple ici la politique aléatoire et nous avons appliqué deux algorithmes : *Temporal Differencing* et *Monte Carlo*.

Le problème étant d'avoir une grille de (5x5), le but est de passer du point le plus haut à gauche, au point le plus bas à droite qui a une récompense de 20.

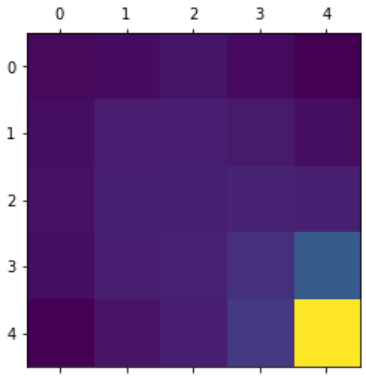
L'évaluation d'un politique se fait par le biais des fonctions de valeurs qu'on calcul pour chaque état du monde. La fonction de valeur d'un état indique à quel point on veut rester dans cet état.

2 Monte Carlo

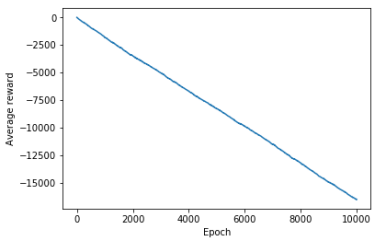
Etant donné une suite d'épisodes, l'algorithme de Monte Carlo consiste à parcourir les états de la grille et pour chaque état, calculer une moyenne sur les récompenses que nous avons obtenu depuis que nous sommes dans cet état. Le résultat est le suivant :



(a) Rewards with time



(b) Value function



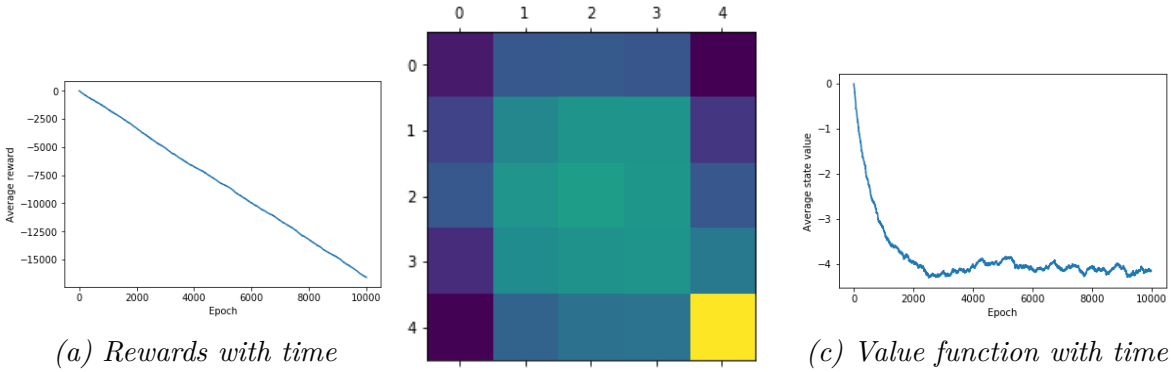
(c) Value function with time

Figure 1. Résultats obtenu par la méthode Monte Carlo

Malheureusement, cet algorithme nous oblige d'attendre la fin d'un épisode pour pouvoir calculer les fonctions de valeurs des états, hors que cela ne donne aucun sens pour les problème du monde réel.

3 Temporal Differencing

Cette méthode applique une équaltion très simple qui n'exige pas d'atteindre la fin d'un épisode. Ceci dit que la mise à jours des fonctions de valeurs se fait au fur et à mesure que l'agent bouge dans la grille. L'application de cette méthode a donné les résultats suivants :



(a) Rewards with time

(b) Value function

(c) Value function with time

Figure 2. Résultats obtenu par la méthode Temporal Differencing

On remarque que la politique aléatoire n'est pas efficace. En effet le mouvement de l'agent n'est pas intelligent, ce qui le mène à ne recevoir que des pénalités.

On remarque que la fonction de valeur est petite lorsqu'on est loin du but, et elle devient de plus en plus grande en nous approchons de celui ci.

4 Conclusion

Il est inutile de garder la même politique pour toute la partie du jeu, un bon système est un système qui observe les récompenses et les pénalités qu'il reçoit et qui change de politique afin de maximiser ses récompenses tout en faisant un bon compromis entre l'exploitation (aller toujours vers les cases avec une grande récompenses) et l'exploration (changer de région de recherche).

Ce TP avait pour but de mettre en pratique l'évaluation d'une politique sans essayer de l'améliorer au court du temps.