Programmation Dynamique et Apprentissage par Renforcement

Annexe

Question

Expliquez comment le **Q-learning** utilise les principes de la **programmation dynamique** pour optimiser les actions d'un agent. Dans votre réponse, détaillez l'équation de mise à jour de $Q(s, a)$ en précisant le rôle de chaque terme (récompense immédiate, récompense future, et taux d'apprentissage), et comparez cette approche à celle des **équations de Bellman** dans un MDP (Processus de Décision Markovien).

Réponse

Oui, Q-learning, TD Learning (Temporal Difference Learning), et les équations de Bellman sont des techniques d'apprentissage par renforcement qui utilisent les principes de la programmation dynamique.

Programmation dynamique et apprentissage par renforcement

La programmation dynamique est un cadre général pour résoudre des problèmes de décision séquentiels, où l'on cherche une séquence d'actions qui maximise une récompense cumulative.

En apprentissage par renforcement, on utilise ces techniques pour optimiser les actions d'un agent dans un environnement, afin de maximiser la récompense obtenue sur le long terme.

Q-learning et TD Learning

1. **Q-learning** :

C'est une méthode de Temporal Difference (TD) Learning. Elle met à jour une fonction de valeur d'état-action $Q(s, a)$ en utilisant une approximation de la valeur future.

$$Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') \right]$$

2. **Equations de Bellman** :

Les équations de Bellman définissent une relation de récursivité pour la valeur d'un état ou d'une action en décomposant la récompense totale en immédiate et future.

$$V(s) = \max_a \left[r(s, a) + \gamma \sum_{s'} P(s'|s, a)V(s') \right]$$

$$Q(s, a) = r(s, a) + \gamma \sum_{s'} P(s'|s, a) \max_{a'} Q(s', a')$$

Pourquoi ces méthodes relèvent de la programmation dynamique

Ces méthodes relèvent de la programmation dynamique parce qu'elles reposent sur des **équations récursives** pour estimer les valeurs d'états ou d'actions en fonction des valeurs futures, tout en **mémorisant** (ou en réutilisant) ces valeurs pour éviter de recalculer les mêmes sous-problèmes.