Reinforcement Learning (RL)

Tendances IT

Abdelhakim Zetati – Lahcen Ezzara

zetati.abdehakim@ensam-casa.ma – ezzara.lahcen@ensam-casa.ma
Université Hassan II de Casablanca
ENSAM Casablanca

Mercredi 18 Octobre 2024

Comprendre l'apprentissage par renforcement

- **L'apprentissage par renforcement (RL)** se distingue en apprenant à travers des interactions avec un environnement pour maximiser une fonction de récompense, sans étiquettes correctes prédéfinies, ce qui le rend utile pour la prise de décision dans des environnements complexes.

L'interface agent-environnement d'un système d'apprentissage par renforcement

L'état de l'agent est composé de ses variables, et il interagit avec l'environnement à travers des actions, recevant des récompenses qui guident ses transitions d'état. Le processus d'apprentissage implique de trouver un équilibre entre l'exploration (essayer de nouvelles actions) et l'exploitation (choisir des actions avec des récompenses connues) pour maximiser les récompenses cumulées au fil du temps.

L'interaction entre l'agent et son environnement

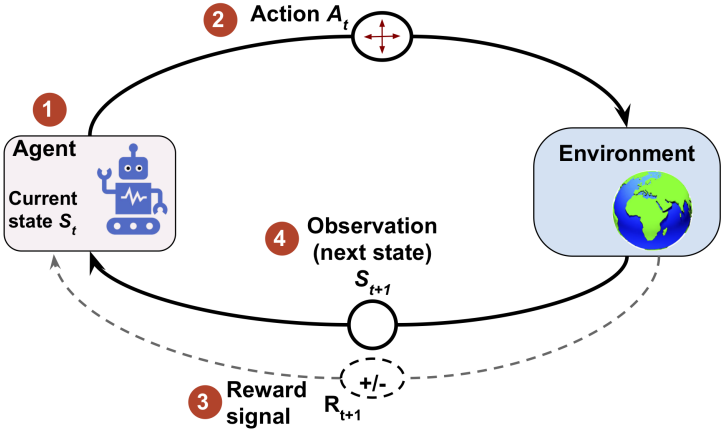

Figure 1: L'interaction entre l'agent et son environnement

Les processus de décision de Markov

En général, le type de problèmes que l'apprentissage par renforcement (RL) traite est généralement formulé comme des processus de décision de Markov (MDP). L'approche standard pour résoudre les problèmes MDP est d'utiliser la programmation dynamique, mais l'apprentissage par renforcement offre certains avantages clés par rapport à la programmation dynamique.

La formulation mathématique des processus de décision de Markov

La distribution de probabilité pour $S_{t+1} = s'$ et $R_{t+1} = r$ peut être écrite comme une probabilité conditionnelle sur l'état précédent S_t et l'action prise A_t comme suit :

$$p(s', r \mid s, a) \equiv P(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a)$$

Deux approches principales pour traiter ce problème sont les méthodes Monte Carlo (MC) sans modèle et les méthodes de différence temporelle (TD). Le tableau suivant présente les deux principales catégories et les branches de chaque méthode :

Les différents modèles à utiliser en fonction de la dynamique de l'environnement

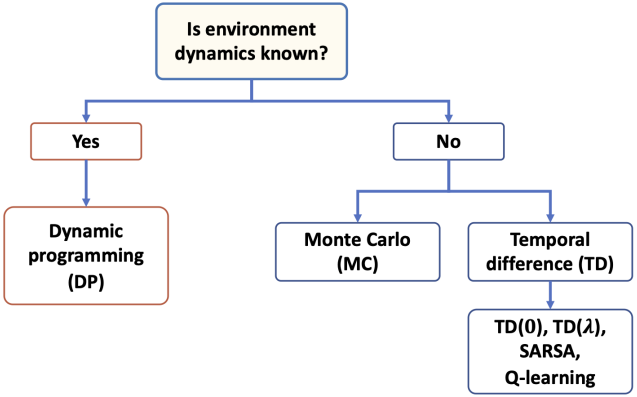

Figure 2: Les différents modèles à utiliser en fonction de la dynamique de l'environnement

Dynamique de l'environnement : déterministe ou stochastique

La dynamique de l'environnement peut être considérée comme déterministe si des actions particulières pour des états donnés sont toujours ou jamais prises, c'est-à-dire $p(s', r \mid s, a) \in \{0, 1\}$. Sinon, dans le cas plus général, l'environnement aurait un comportement stochastique.

Terminologie de l'apprentissage par renforcement

En apprentissage par renforcement, plusieurs concepts doivent s'envisager pour comprendre comment un agent interagit avec son environnement et maximise une fonction de récompense.

- **Retour (Return)**.
- **Politique (Policy)**.
- **Fonction de Valeur (Value Function)**.

Retour (Return)

Le retour au temps t en apprentissage par renforcement est la récompense cumulée totale d'un épisode, calculée comme suit :

$$G_t \equiv R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0} \gamma^k R_{t+k+1}$$

$$G_t = R_{t+1} + \gamma G_{t+1} = r + \gamma G_{t+1}$$

Politique (Policy)

Une politique, généralement notée $\pi(a \mid s)$, est une fonction qui détermine la prochaine action à prendre. Elle peut être soit déterministe, soit stochastique (c'est-à-dire la probabilité de prendre la prochaine action). Une politique stochastique a alors une distribution de probabilité sur les actions qu'un agent peut prendre dans un état donné :

$$\pi(a \mid s) \equiv P[A_t = a \mid S_t = s]$$

La politique optimale $\pi_*(a \mid s)$ est celle qui génère le retour le plus élevé.

Fonction de Valeur (Value Function)

La fonction de valeur, également appelée fonction de valeur d'état, mesure la qualité de chaque état, c'est-à-dire à quel point il est bon ou mauvais d'être dans un état particulier. Notez que le critère de qualité est basé sur le retour. Maintenant, sur la base du retour G_t, nous définissons la fonction de valeur de l'état s comme le retour attendu (le retour moyen sur tous les épisodes possibles) après avoir suivi la politique π :

$$v_\pi(s) \equiv E_\pi[G_t \mid S_t = s] = E_\pi\left[\sum_{k=0} \gamma^{k+1} R_{t+k+1} \mid S_t = s\right]$$

Fonction de Valeur (Value Function)

De plus, nous pouvons également définir une valeur pour chaque paire état-action, qui est appelée la fonction de valeur d'action et est notée $q_\pi(s,a)$. La fonction de valeur d'action fait référence au retour attendu G_t lorsque l'agent est à l'état $S_t = s$ et prend l'action $A_t = a$. En étendant la définition de la fonction de valeur d'état aux paires état-action, nous obtenons ce qui suit :

$$q_\pi(s,a) \equiv E_\pi[G_t \mid S_t = s, A_t = a] = E_\pi\left[\sum_{k=0} \gamma^{k+1} R_{t+k+1} \mid S_t = s, A_t = a\right]$$

Ceci est similaire à la référence à la politique optimale en tant que $\pi_*(a \mid s)$; $v_*(s)$ et $q_*(s,a)$ désignent également les fonctions de valeur d'état et de valeur d'action optimales.

Quelle est la différence entre la récompense, le retour et la fonction de valeur?

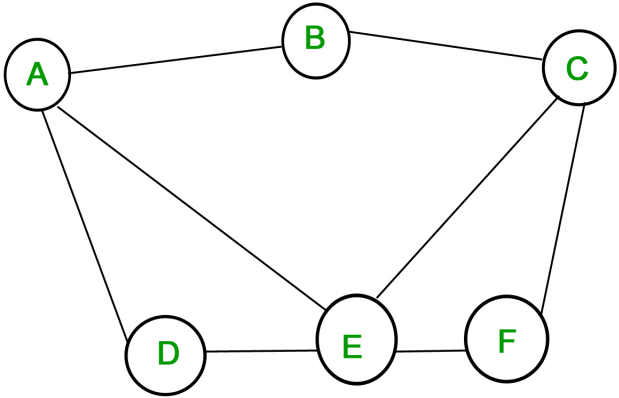

Figure 3: Graphe

Deep Q-Learning

Lorsque la fonction d'approximation, $q_w(x_s, a)$, est un réseau de neurones profond (DNN), le modèle résultant est appelé un réseau Q profond (DQN). Pour entraîner un modèle DQN, les poids sont mis à jour conformément à l'algorithme Q-learning.

Exemple de DQN

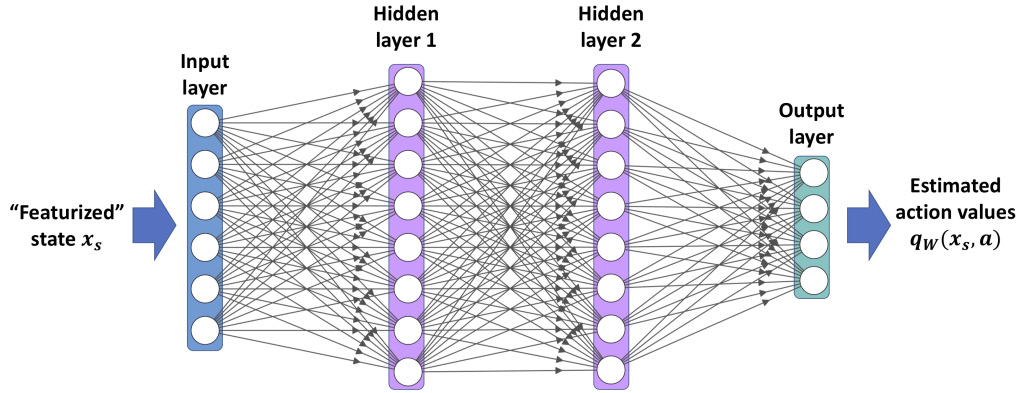

Figure 4: Exemple de DQN

Q-values

Pour ce faire, nous pouvons modifier la règle de mise à jour pour prendre en compte la valeur Q maximale en variant les différentes actions dans l'état immédiat suivant. L'équation modifiée pour la mise à jour des valeurs Q est la suivante :

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$$

Détermination de la valeur cible à l'aide du DQN

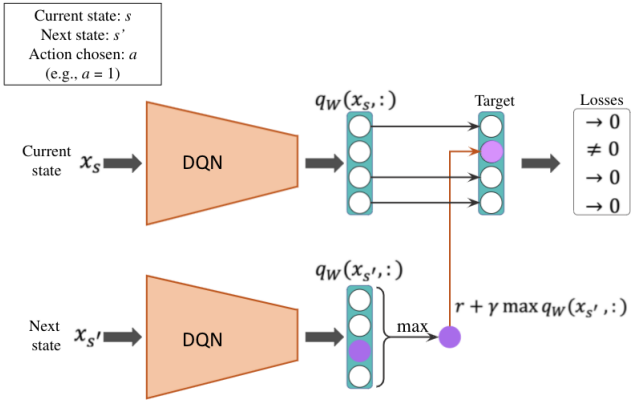

Figure 5: Détermination de la valeur cible à l'aide du DQN

Détermination de la valeur cible à l'aide du DQN

Nous traitons cela comme un problème de régression, en utilisant les trois quantités suivantes :

- Les valeurs prédites actuellement, $q_w(x_s, :)$
- Le vecteur de valeur cible tel que décrit
- La fonction de perte standard d'erreur quadratique moyenne (MSE)

Merci pour Votre Attention !

Abdelhakim Zetati – Lahcen Ezzara