Cours DQN – Section 4.2

4.2 – Étapes d'apprentissage complètes d'un DQN

Dans cette section, nous détaillons pas à pas le processus d'apprentissage d'un agent DQN. Chaque composant interagit avec les autres pour créer une boucle d'entraînement efficace.

Étape 1 – Initialisation

Avant de commencer l'apprentissage :
— Initialiser les poids du **réseau principal** $Q(s, a; \theta)$.
— Créer une copie $Q(s, a; \theta^-)$ pour le **réseau cible**.
— Créer une **mémoire de répétition** (Replay Memory) vide.
— Définir les hyperparamètres : taux d'apprentissage α, facteur de discount γ, stratégie ε-greedy, taille du mini-batch, fréquence de mise à jour du réseau cible, etc.

Étape 2 – Interaction avec l'environnement

Pour chaque épisode de jeu :

1. Observer l'état initial s_0.
2. Pour chaque pas de temps t :
   (a) Choisir une action a_t à partir de s_t selon une politique ε-greedy :

$$a_t = \begin{cases} \text{action aléatoire} & \text{avec probabilité } \varepsilon \\ \arg\max_a Q(s_t, a; \theta) & \text{avec probabilité } 1 - \varepsilon \end{cases}$$

   (b) Exécuter a_t, observer la récompense r_t et le nouvel état s_{t+1}.
   (c) Stocker la transition $(s_t, a_t, r_t, s_{t+1}, \text{done})$ dans la mémoire.

Étape 3 – Apprentissage (à chaque K étapes)

Si la mémoire contient assez d'expériences, alors :

1. Tirer un mini-batch d'expériences aléatoires :

$$\mathcal{B} = \{(s_i, a_i, r_i, s'_i, \text{done}_i)\}_{i=1}^{N}$$

2. Calculer les cibles pour chaque transition :

$$y_i = \begin{cases} r_i & \text{si done}_i = \text{True} \\ r_i + \gamma \cdot \max_{a'} Q(s'_i, a'; \theta^-) & \text{sinon} \end{cases}$$

3. Calculer la prédiction :

$$Q(s_i, a_i; \theta)$$

4. Calculer la perte :

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (y_i - Q(s_i, a_i; \theta))^2$$

5. Appliquer la descente de gradient pour mettre à jour θ.

Étape 4 – Mise à jour du réseau cible

Toutes les C itérations :

$$\theta^- \leftarrow \theta$$

Cela permet au réseau cible de suivre progressivement le réseau principal, sans changements brusques.

Étape 5 – Répéter

On répète ce processus pendant plusieurs épisodes (souvent des milliers), tout en diminuant progressivement ε (stratégie ε-décroissante) pour favoriser l'exploitation.

Conclusion

Le cycle d'apprentissage du DQN est donc :

Initialisation → Jeu et collecte d'expériences → Apprentissage → Mise à jour du réseau cible

Cette boucle itérative est au cœur de l'agent DQN et permet une amélioration progressive de sa politique d'action.