Cours DQN – Section 3.2

3.2 – Target Network : stabiliser l'apprentissage

Lorsqu'un réseau de neurones est utilisé pour estimer les valeurs Q, l'apprentissage peut devenir instable, voire diverger. Le **Target Network** (ou *réseau cible*) est une solution simple et élégante pour stabiliser cet apprentissage.

Problème : mise à jour instable des cibles

Dans l'algorithme de Q-Learning, la cible d'apprentissage est :

$$\text{cible} = r + \gamma \cdot \max_{a'} Q(s', a')$$

Avec un réseau de neurones, cela devient :

$$\text{cible} = r + \gamma \cdot \max_{a'} Q(s', a'; \theta)$$

Mais comme les poids θ changent constamment pendant l'apprentissage :
— La cible change tout le temps.
— Le réseau apprend sur des bases mouvantes.
— Cela entraîne des oscillations, voire une divergence.

Solution : réseau cible figé temporairement

DQN introduit un second réseau, appelé **Target Network**, noté $Q(s, a; \theta^-)$, dont les poids sont mis à jour plus lentement.

Principe :
— Le réseau principal (Q_{online}) est mis à jour à chaque étape.
— Le réseau cible (Q_{target}) est une copie figée du réseau principal.
— Toutes les N étapes (par exemple, tous les 10 000 pas), les poids du réseau cible sont mis à jour :

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

Avantages du Target Network

— La cible devient plus stable pendant un certain nombre d'itérations.
— Le réseau principal peut s'ajuster à une référence fixe.
— Cela réduit l'instabilité et améliore la convergence.

Illustration pédagogique

C'est comme si un étudiant révisait à partir d'un corrigé temporairement figé (le réseau cible), et ne mettait à jour ce corrigé qu'après plusieurs sessions de travail. Cela évite de changer constamment de méthode sans évaluer les progrès.

Formule de la cible avec le réseau cible

$$y = r + \gamma \cdot \max_{a'} Q(s', a'; \theta^-)$$

La fonction de perte devient :

$$\mathcal{L} = (y - Q(s, a; \theta))^2$$

Conclusion

Le Target Network est un des piliers du DQN :
— Il évite la dérive des valeurs cibles.
— Il stabilise l'apprentissage en créant une référence fixe temporaire.
— Il est facile à implémenter et très efficace.

Dans la section suivante, nous verrons une dernière innovation importante : l'**échantillonnage aléatoire des expériences** pour éviter les biais de séquence.