TD-Learning, Q-Learning et Équations de Bellman

Rhouma Haythem

Avril 2025

1 TD(0)

TD(0) est la forme la plus simple de TD-Learning. Il met à jour la valeur d'un état immédiatement après chaque action.

$$V(S_t) \leftarrow (1-\alpha)V(S_t) + \alpha\left[R_{t+1} + \gamma V(S_{t+1})\right] \tag{1}$$

2 TD(2)

$$V(S_t) \leftarrow (1-\alpha)V(S_t) + \alpha\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2})\right] \tag{2}$$

3 TD(3)

$$V(S_t) \leftarrow (1-\alpha)V(S_t) + \alpha\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 V(S_{t+3})\right] \tag{3}$$

4 TD(4)

$$V(S_t) \leftarrow (1-\alpha)V(S_t) + \alpha\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \gamma^4 V(S_{t+4})\right] \tag{4}$$

5 TD(n) - Généralisation

$$V(S_t) \leftarrow (1-\alpha)V(S_t) + \alpha\left[\sum_{k=1}^{n} \gamma^{k-1} R_{t+k} + \gamma^n V(S_{t+n})\right] \tag{5}$$

Dans ces équations :

- $V(S_t)$ est la valeur estimée de l'état actuel
- α est le taux d'apprentissage

- R_{t+k} est la récompense reçue k pas dans le futur
- γ est le facteur d'actualisation
- $V(S_{t+n})$ est la valeur estimée de l'état n pas dans le futur

La méthode TD(n) utilise les n prochaines récompenses et la valeur estimée de l'état n pas plus loin pour mettre à jour la valeur de l'état actuel. Plus n est grand, plus la méthode prend en compte d'informations futures, ce qui peut améliorer la précision mais augmente aussi la variance des estimations.

6 Q-Learning

Q-Learning est une méthode *off-policy* qui apprend les valeurs des paires état-action.

$$Q(S_t, A_t) \leftarrow (1-\alpha)Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a)\right] \tag{6}$$

Où :

- $Q(S_t, A_t)$ est la valeur de la paire état-action actuelle
- $\max_a Q(S_{t+1}, a)$ est la valeur maximale de l'action dans l'état suivant

7 Équation de Bellman

L'équation de Bellman exprime la valeur d'un état comme la récompense immédiate attendue plus la valeur actualisée des futurs états, selon une politique π.

7.1 Valeur d'un état sous politique π

$$V^\pi(s) = \mathbb{E}_\pi \left[R_{t+1} + \gamma V^\pi(S_{t+1}) \mid S_t = s\right] \tag{7}$$

7.2 Valeur d'une paire état-action sous politique π

$$Q^\pi(s, a) = \mathbb{E}_\pi \left[R_{t+1} + \gamma Q^\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a\right] \tag{8}$$

7.3 Forme optimale (état)

$$V^*(s) = \max_a \mathbb{E} \left[R_{t+1} + \gamma V^*(S_{t+1}) \mid S_t = s, A_t = a\right] \tag{9}$$

7.4 Forme optimale (état-action)

$$Q^*(s, a) = \mathbb{E} \left[R_{t+1} + \gamma \max_{a'} Q^*(S_{t+1}, a') \mid S_t = s, A_t = a\right] \tag{10}$$

8 Comparaison des méthodes

Méthode	**Type**	**Mise à jour**	**Avantages**	**Inconvénients**
TD(0)	On-policy	État uniquement	Simple, rapide	Ne considère qu'un seul pas
TD(n)	N-step	n récompenses et 1 valeur future	Meilleure précision possible	Variance et complexité accrues
Q-Learning	Off-policy	Paire état-action	Trouve une politique optimale	Nécessite plus de mémoire