Variantes des équations Bellman, TD-Learning et Q-Learning

Rhouma Haythem

April 2025

1 Introduction

Le TD-Learning (Temporal Difference Learning) et le Q-Learning sont des méthodes d'apprentissage par renforcement utilisées pour apprendre à partir d'expériences sans avoir besoin d'un modèle complet de l'environnement.

2 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. Elle constitue la base théorique du TD-Learning et du Q-Learning.

2.1 Valeur d'un état sous politique π

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

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

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[R_{t+1} + \gamma \cdot Q^{\pi}(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a\right] \tag{2}$$

2.3 Forme optimale (état)

$$V^{*}(s) = \max_{a} \mathbb{E}\left[R_{t+1} + \gamma \cdot V^{*}(S_{t+1}) \mid S_t = s, A_t = a\right] \tag{3}$$

2.4 Forme optimale (état-action)

$$Q^{*}(s,a) = \mathbb{E}\left[R_{t+1} + \gamma \cdot \max_{a'} Q^{*}(S_{t+1}, a') \mid S_t = s, A_t = a\right] \tag{4}$$

3 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.

3.1 Équation de mise à jour TD(0)

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

Où :

- $V(S_t)$ est la valeur de l'état actuel
- α est le taux d'apprentissage
- R_{t+1} est la récompense immédiate
- γ est le facteur d'actualisation
- $V(S_{t+1})$ est la valeur de l'état suivant

4 TD(n)

TD(n) est une généralisation de TD(0) qui prend en compte n étapes futures pour la mise à jour.

4.1 Équation de mise à jour TD(n)

$$V(S_t) \leftarrow (1-\alpha)V(S_t) + \alpha\left[\sum_{k=1}^{n} \gamma^{k-1} R_{t+k} + \gamma^n V(S_{t+n})\right] \tag{6}$$

5 Q-Learning

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

5.1 Équation de mise à jour Q-Learning

$$Q(S_t, A_t) \leftarrow (1-\alpha)Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a)\right] \tag{7}$$

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

6 Comparaison des méthodes

Méthode	**Type**	**Mise à jour**	**Avantages**	**Inconvé**
TD(0)	On-policy	État uniquement	Simple, rapide	Ne consi
TD(n)	N-step	État et n récompenses futures	Bon compromis exploration/efficacité	Complex
Q-Learning	Off-policy	Paire état-action	Trouve une politique optimale	Nécessite

7 Conclusion

Le TD-Learning et le Q-Learning sont des méthodes puissantes pour l'apprentissage par renforcement, permettant aux agents d'apprendre de manière efficace dans des environnements complexes sans modèle complet.