TD-Learning : TD(0), TD(1), TD(2), TD(3), TD(4), TD(n) et Q-Learning

Rhouma Haythem

Avril 2025

1 Introduction

Ce document présente les méthodes TD(0), TD(1), TD(2), TD(3), TD(4), la généralisation TD(n) ainsi que Q-Learning. Chaque mise à jour est exprimée sous les deux formes :

— **Forme Erreur TD :** $V \leftarrow V + \alpha(\text{cible} - V)$

— **Forme pondérée** $(1-\alpha)$ **:** $V \leftarrow (1-\alpha)V + \alpha(\text{cible})$

2 TD(0)

Forme Erreur TD

$$V(S_t) \leftarrow V(S_t) + \alpha\big(\gamma V(S_{t+1}) - V(S_t)\big) \tag{1}$$

Forme $(1-\alpha)$

$$V(S_t) \leftarrow (1-\alpha)V(S_t) + \alpha\gamma V(S_{t+1}) \tag{2}$$

3 TD(1)

Forme Erreur TD

$$V(S_t) \leftarrow V(S_t) + \alpha\big(R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\big) \tag{3}$$

Forme $(1-\alpha)$

$$V(S_t) \leftarrow (1-\alpha)V(S_t) + \alpha\big(R_{t+1} + \gamma V(S_{t+1})\big) \tag{4}$$

4 TD(2)

Forme Erreur TD

$$V(S_t) \leftarrow V(S_t) + \alpha\big(R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2}) - V(S_t)\big) \tag{5}$$

Forme $(1-\alpha)$

$$V(S_t) \leftarrow (1-\alpha)V(S_t) + \alpha\big(R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2})\big) \tag{6}$$

5 TD(3)

Forme Erreur TD

$$V(S_t) \leftarrow V(S_t) + \alpha\big(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 V(S_{t+3}) - V(S_t)\big) \tag{7}$$

Forme $(1-\alpha)$

$$V(S_t) \leftarrow (1-\alpha)V(S_t) + \alpha\big(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 V(S_{t+3})\big) \tag{8}$$

6 TD(4)

Forme Erreur TD

$$V(S_t) \leftarrow V(S_t) + \alpha\big(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \gamma^4 V(S_{t+4}) - V(S_t)\big) \tag{9}$$

Forme $(1-\alpha)$

$$V(S_t) \leftarrow (1-\alpha)V(S_t) + \alpha\big(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \gamma^4 V(S_{t+4})\big) \tag{10}$$

7 TD(n) Général

La cible n-step est :

$$G_t^{(n)} = \sum_{k=1}^{n} \gamma^{k-1} R_{t+k} + \gamma^n V(S_{t+n})$$

Forme Erreur TD

$$V(S_t) \leftarrow V(S_t) + \alpha\big(G_t^{(n)} - V(S_t)\big) \tag{11}$$

Forme $(1-\alpha)$

$$V(S_t) \leftarrow (1-\alpha)V(S_t) + \alpha\left[\sum_{k=1}^{n} \gamma^{k-1} R_{t+k} + \gamma^n V(S_{t+n})\right] \tag{12}$$

8 Q-Learning

Forme Erreur TD

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\big(R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\big) \tag{13}$$

Forme $(1 - \alpha)$

$$Q(S_t, A_t) \leftarrow (1 - \alpha)Q(S_t, A_t) + \alpha\big(R_{t+1} + \gamma \max_a Q(S_{t+1}, a)\big) \tag{14}$$

9 Résumé global

Méthode	Type	Utilise	Objectif
TD(0)	On-policy	Bootstrap immédiat (aucune récompense utilisée)	Stable, rapide
TD(1)	On-policy	1 récompense + $V(S_{t+1})$	Estimation simple et efficace
TD(2)	On-policy	2 récompenses + $V(S_{t+2})$	Intermédiaire entre court et moyen terme
TD(3)	On-policy	3 récompenses + $V(S_{t+3})$	Meilleure utilisation des informations futures
TD(4)	On-policy	4 récompenses + $V(S_{t+4})$	Encore plus d'information, variance plus haute
TD(n)	N-step	n récompenses + $V(S_{t+n})$	Estimation générale n-step
Q-Learning	Off-policy	Valeurs état-action + $\max_a Q$	Trouver la politique optimale