Différences entre TD(0), TD(n) et Q-Learning

Rhouma Haythem

Avril 2025

1 Introduction

Ce document présente les différences fondamentales entre TD(0), TD(n) et Q-Learning. Pour chaque méthode, les deux formes d'équations sont fournies :

— **forme "erreur TD"** : ancienne valeur + α (cible – ancienne valeur),

— **forme "mélange pondéré"** : $(1-\alpha) \times$ ancienne valeur + $\alpha \times$ cible.

2 TD(0) — Temporal Difference à 0-step

Type : On-policy

Idée principale : bootstrap immédiat

TD(0) utilise uniquement :

— la récompense immédiate R_{t+1},

— la valeur estimée du prochain état $V(S_{t+1})$.

Forme 1 : Erreur TD

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

Forme 2 : Mélange pondéré $(1-\alpha)$

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

Avantages : simple, rapide, faible variance.

Limite : sous-utilise les récompenses futures.

3 TD(n) — Méthode n-step

Type : N-step (on-policy)

Idée principale : utiliser plusieurs récompenses futures

On prend les n récompenses futures :

$$R_{t+1}, R_{t+2}, \ldots, R_{t+n}$$

puis on fait un bootstrap sur la valeur $V(S_{t+n})$.

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 1 : Erreur TD (générale)

$$V(S_t) \leftarrow V(S_t) + \alpha\big(G_t^{(n)} - V(S_t)\big) \tag{3}$$

Forme 2 : Mélange pondéré $(1-\alpha)$

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

Avantages : plus précis si n est grand, combine effets immédiats et futurs.

Inconvénients : variance plus élevée, mise à jour plus lente (attendre n pas).

4 Q-Learning

Type : Off-policy

Idée principale : apprendre la politique optimale (même si on explore autrement)

La cible optimale est :

$$R_{t+1} + \gamma \max_a Q(S_{t+1}, a)$$

Forme 1 : Erreur TD

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right] \tag{5}$$

Forme 2 : Mélange pondéré $(1-\alpha)$

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

Avantages : converge vers la politique optimale, robuste.

Inconvénients : nécessite plus de mémoire et une exploration suffisante.

5 Résumé global

Méthode	Type	Utilise	Objectif
TD(0)	On-policy	1 récompense + bootstrap immédiat	Mise à jour rapide, faible variance
TD(n)	N-step	n récompenses futures + $V(S_{t+n})$	Estimation plus précise grâce à plus d'information
Q-Learning	Off-policy	Valeurs état-action + $\max_a Q(S_{t+1}, a)$	Trouver la politique optimale