Chapitre 4.

Apprentissage dans l'incertain

Rappel

$MDP = (S, A, p, r)$

- $S = \{s_1, s_2, \dots, s_n\}$
- $A = \{a_1, a_2, \dots, a_m\}$
- $p : S \times A \times S \longrightarrow [0,1]$
- $r : S \times A \times S \longrightarrow \mathbb{R}$

Ce qu'on a vu jusque là :

p et r sont connues.

$\Rightarrow$ Equation de Bellman

+ policy iteration / value iteration

Dans ce qui suit :

incertain = p et r ne sont pas connus

L'agent va devoir interagir avec l'envir-

- ouvement pour apprendre p et r.

Soit π une politique.

l'eq. de Bellman :

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,r) \left(r + \gamma v_\pi(s') \right)$$

Si π, p et r sont déterminés et si on instantie l'eq. à une trajectoire :

$$v_\pi(s_t) = r_t + \gamma v_\pi(s_{t+1})$$

$$\Rightarrow r_t + \gamma v_\pi(s_{t+1}) - v_\pi(s_t) = 0$$

<u>idée</u> : v_π est inconnue mais au fil de l'apprentissage, on essaie s'approcher de sa vraie valeur...

On commence donc avec une valeur arbitraire $\hat{v}$

- Quand l'agent effectue une interaction,

Algorithme Temporal Difference (TD)

Input: S, A, γ et π

$\hat{V}_\pi(s) \leftarrow 0$, $\forall s \in S$

$n(s) \leftarrow 0$, $\forall s \in S$

Répéter

initialiser l'état initial s_0

$t \leftarrow 0$

Répéter

émettre l'action $a_t = \pi(s_t)$

observer r_t et s_{t+1}

$$\hat{V}_\pi(s_t) \leftarrow V_\pi(s_t) + \alpha(s_t, n(s_t)) \times \left[r_t + \gamma \hat{V}_\pi(s_{t+1}) - \hat{V}_\pi(s_t) \right]$$

$n(s) \leftarrow n(s) + 1$

$t \leftarrow t + 1$

jusqu'à s_t état final

jusqu'à critère d'arrêt vérifié

Propriété : TD converge (asymptotiquement) si

- chaque état est visité une infinité de fois
- $\forall s \in S$

$$\sum_{t \text{ auquels } s \text{ est visité}} \alpha_t(s, a(s)) = +\infty$$

$$\sum \alpha_t^2(s, a(s)) < +\infty$$

Une solution :

$$\alpha_t(s, a(s)) = \frac{1}{a(s)+1}$$

α est nommé le pas d'apprentissage.