TD 9 : RL

Exercice 1 – MDP et programmation dynamique

Rappel :

- un MDP est défini par un ensemble d'états $\mathcal{S}$, un ensemble d'actions $\mathcal{A}$, une fonction de transition $t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$: $t(s'|a,s)$ est la probabilité d'arriver en s' à partir de s en prenant l'action a, une fonction de récompense $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$.
- une politique $\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ (cas non déterministe) attribue une probabilité $\pi(a|s)$ qu'une action soit prise quand on est dans l'état s.
- $V^\pi(s) = \mathbb{E}_\pi[G_t|S_t = s]$ avec $G_t = \sum\limits_{k\geq 0} \gamma^k R_{t+k}$, R_{t+k} étant les récompenses obtenues à l'état $t+k$.

Q 1.1 Equation de Bellman

Q 1.1.1 Montrer $V^\pi(s) = \sum\limits_a \pi(a|s) \sum\limits_{s'} t(s'|s,a) \left[r(s,a,s') + \gamma v_\pi(s')\right] = \mathbb{E}_{a,s'\sim\pi}\left[r(s,a,s') + \gamma v(s')\right]$

Q 1.1.2 On appelle opérateur de Bellman T^π l'opérateur

$$V_{k+1}(s) = (T^\pi\ V_k)(s) = \sum_a \pi(a|s) \sum_{s'} t(s'|s,a) \left[r(s,a,s') + \gamma V_k(s')\right]$$

Montrer la propriété de monotonicité de l'opérateur : si $\forall s,\ v^1(s) \leq v^2(s)$, alors pour tout s on a : $T^\pi v^1(s) \leq T^\pi v^2(s)$.

Q 1.1.3 Montrer que l'application répétée de l'opérateur de Bellman sur un ensemble de valeurs V, fait converger cet ensemble de valeurs vers un point fixe.

Q 1.1.4 L'algorithme policy iteration travaille, à chaque itération k, en deux temps :

1. Évaluation de $V^{(k)}$ jusqu'à convergence selon la politique courante stationnaire π_k.
2. Mise à jour de la politique π_{k+1} selon une stratégie greedy. Pour tout etat s : $\pi_{k+1}(s) = argmax_a \sum\limits_{s'} t(s'|s,a) \left[r(s,a,s') + \gamma V^{(k)}(s')\right]$

Déduire de la propriété de monotonicité démontrée à la question précédente la convergence de cet algorithme.

Q 1.1.5 Montrer l'equation d'optimalité de Bellman :

$$V^*(s) = max_\pi V^\pi(s) = max_a \ \sum_{s'} t(s'|s,a) \left[r(s,a,s') + \gamma V^*(s')\right]$$

.

Q 1.1.6 On appelle opérateur de Bellman optimal T^* l'opérateur :

$$(T^*\ V)(s) = \max_a \sum_{s'} t(s'|s,a) \left[r(s,a,s') + \gamma V(s')\right]$$

Soit deux ensembles de valeurs V_1 et V_2 définis pour tout etat s. Montrer que

$$\|T^* V_1 - T^* V_2\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$$

.

Q 1.1.7 L'algorithme value iteration applique, pendant un nombre donné d'itérations, l'opérateur optimal de Bellman pour mettre à jour les valeurs des etats : $V^{(k+1)} = T^* V^{(k)}$, avec $V^{(k)}$ l'ensemble des valeurs à l'itération k. Déduire de la propriété de contraction démontrée à la question précédente la convergence de cet algorithme (en supposant la propriété de point fixe $TV^* = V^*$).

Q 1.2 On considère le MDP déterministe suivant :

- 6 états $s_1, \ldots, s_6$
- s_1 et s_6 comme états finaux.
- deux actions possibles g et d
- les transitions : $t(s_i, g) = s_{i-1}$ et $t(s_i, d) = s_{i+1}$ pour $1 < i < 6$
- La récompense est définie par $r(s_i, a, s_{i'}) = -10$ pour $1 < i, i' < 6$ pour $a \in \{g, d\}$, et $r(s_2, g, s_1) = 50$ et $r(s_5, d, s_6) = 100$.

Q 1.2.1 Représenter le mdp graphiquement.

Q 1.2.2 Soit la politique non déterministe uniforme $\pi(s_i, d) = \pi(s_i, g) = 0.5$ pour $1 < i < 6$. Déterminer $V^\pi(s)$ en utilisant un algorithme d'évaluation de cette politique.

Q 1.2.3 Déterminer la politique optimale π^* en utilisant l'algorithme Policy iteration.

Q 1.2.4 Déterminer la politique optimale π^* en utilisant l'algorithme Value iteration.

Q 1.3 On considère maintenant que l'on ne connaît pas la fonction de transition t

Q 1.3.1 Dire quels problèmes se posent alors avec les algorithmes précédents

Q 1.3.2 Quels sont les algorithmes vus en cours pour evaluer une politique π ?

Q 1.3.3 Si on souhaite maintenant trouver la politique optimale π^*, comment procéder ?

Q 1.3.4 Et si on a beaucoup trop d'états pour tout stocker en mémoire / tout considérer ?