Preuves value et policy iteration

Soit l'opérateur de Bellman T^π qui, appliqué à un ensemble de valeurs V est défini selon :

$$(T^\pi\ V)(s) = \sum_{s'} P(s'|s,\pi(s))(r(s,\pi(s),s') + \gamma V(s'))$$

On commence par 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.

Pour tout ensemble V, on veut alors montrer $\lim_{n\to\infty} ||(T^\pi)^{n+1}V - (T^\pi)^n V||_\infty = 0$.
Pour tout état s, on a pour $n \geq 1$:

$$\begin{aligned}
|(T^\pi)^{n+1}V(s) - (T^\pi)^n V(s)| &= |\sum_{s'} P(s'|s,\pi(s))(r(s,\pi(s),s') + \gamma(T^\pi)^n V(s')) \\
&\quad - \sum_{s'} P(s'|s,\pi(s))(r(s,\pi(s),s') + \gamma(T^\pi)^{n-1} V(s'))| \\
&= \gamma|\sum_{s'} P(s'|s,\pi(s))((T^\pi)^n V(s') - (T^\pi)^{n-1} V(s'))| \\
&\leq \gamma \sum_{s'} P(s'|s,\pi(s))|(T^\pi)^n V(s') - (T^\pi)^{n-1} V(s')| \\
&\leq \gamma||(T^\pi)^n V - (T^\pi)^{n-1} V||_\infty \sum_{s'} P(s'|s,\pi(s)) \\
&= \gamma||(T^\pi)^n V - (T^\pi)^{n-1} V||_\infty
\end{aligned}$$

On a alors $||(T^\pi)^{n+1}V - (T^\pi)^n V||_\infty \leq \gamma||(T^\pi)^n V - (T^\pi)^{n-1} V||_\infty$, ce qui permet d'affirmer que l'application répétée de l'opérateur de Bellman fait converger V vers un point fixe (car $\gamma < 1$ et donc dans ce cas $\lim_{n\to\infty} ||(T^\pi)^{n+1}V - (T^\pi)^n V||_\infty = 0$).

Et ce point fixe correspond à la vraie valeur V^π pour tous les états : si $V \neq V^\pi$, alors au moins pour 1 des états s on a $V(s) \neq \sum_{s'} P(s'|s,\pi(s))(r(s,\pi(s),s') + \gamma V(s'))$, et donc l'application de l'opérateur de Bellman modifie la valeur correspondante. Le seul point fixe est donc V^π.

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 : $\pi_{k+1}(s) = arg\max_a \sum_{s'} P(s'|s,a)(r(s,a,s') + \gamma V^{(k)}(s'))$

On souhaite démontrer la convergence de cet algorithme.

Puisqu'on a fait une évaluation complète de $V^{(k)}$ selon la politique π_k, on a : $V^{(k)} = T^{\pi_k} V^{(k)}$ (on est dans un etat stable après convergence)
Or : $T^{\pi_k} V^{(k)} \leq T^{\pi_{k+1}} V^{(k)}$ (du fait de la mise à jour greedy qui choisit l'action max)
On a alors, grâce à la propriété de monotonicité de l'opérateur de Bellman (si $V_1(s) \leq V_2(s)\ \forall s$, alors pour tout s on a : $T^\pi V_1(s) \leq T^\pi V_2(s)$, qui se démontre de manière triviale car on peut

retrouver l'expression de l'opérateur dans V_1 et V_2) :

$$V^{(k)} \leq T^{\pi_{k+1}} V^{(k)}, \tag{1}$$
$$T^{\pi_{k+1}} V^{(k)} \leq (T^{\pi_{k+1}})^2 V^{(k)}, \tag{2}$$
$$... \leq ..., \tag{3}$$
$$(T^{\pi_{k+1}})^{n-1} V^{(k)} \leq (T^{\pi_{k+1}})^n V^{(k)}. \tag{4}$$

Si on assemble toutes ces inégalités (qui correspondent en quelque sorte aux différentes passes de l'algo d'évaluation de la politique), on a : $V^{(k)} \leq \lim_{n\to\infty} (T^{\pi_{k+1}})^n V^{(k)} = V^{(k+1)}$
On a donc une séquence de valeurs qui s'accroissent à chaque itération jusqu'à stabilité. Avec un nombre de politiques fini, cette stabilité est sûre d'être atteinte au bout d'un certain temps, c'est l'avantage de policy iteration. Le problème de cet algo c'est qu'il demande à chaque itération une évaluation complète des valeurs, ce qui peut être très coûteux.

On s'intéresse maintenant à l'équation d'optimalité de Bellman : $V^*(s) = max_\pi V^\pi(s)$.

On considère une politique $\pi = (a, \pi')$ (a est la première action, π' définit les suivantes en fonction des etats)

$$\begin{aligned}
V^*(s) &= \max_\pi V^\pi(s) \\
&= \max_{(a,\pi')} \sum_{s'} P(s'|s,a)(r(s,a,s') + \gamma V^{\pi'}(s')) \\
&= \max_a \sum_{s'} P(s'|s,a)(r(s,a,s') + \gamma \max_{\pi'} V^{\pi'}(s')) \\
&= \max_a \sum_{s'} P(s'|s,a)(r(s,a,s') + \gamma V^*(s'))
\end{aligned}$$

Où le déplacement du $\max_{\pi'}$ est autorisé ici car la politique est définie de manière indépendante sur les différents états (ce ne serait pas vrai par exemple avec de l'approché, si on défini le meilleur π en fonction de paramètres communs θ utilisés pour tous les états).

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

$$(T^*\ V)(s) = \max_a \sum_{s'} P(s'|s,a)(r(s,a,s') + \gamma V(s'))$$

Soit deux ensembles de valeurs V_1 et V_2 définis pour tout etat s. On souhaite montrer la propriété de contraction : $\|T^*V_1 - T^*V_2\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$.

Pour tout etat s on a :

$$\begin{aligned}
|T^*V_1(s) - T^*V_2(s)| &= |\max_a \sum_{s'} P(s'|s,a)(r(s,a,s') + \gamma V_1(s')) - \max_{a'} \sum_{s'} P(s'|s,a')(r(s,a',s') + \gamma V_2(s'))| \\
&\leq \max_a |\sum_{s'} P(s'|s,a)(r(s,a,s') + \gamma V_1(s')) - \sum_{s'} P(s'|s,a)(r(s,a,s') + \gamma V_2(s'))| \\
&= \gamma \max_a \sum_{s'} P(s'|s,a)|V_1(s') - V_2(s')| \\
&\leq \gamma \|V_1 - V_2\|_\infty \max_a \sum_{s'} P(s'|s,a) \\
&\leq \gamma \|V_1 - V_2\|_\infty
\end{aligned}$$

Où la première inégalité est due au fait que $\max_a f(a) - \max_b g(b) \leq \max_a (f(a) - g(a))$

Et si c'est vrai pour tout état s, alors c'est vrai pour l'état s de différence maximale et donc $\|T^*V_1 - T^*V_2\|_\infty \leq \gamma\|V_1 - V_2\|_\infty$.

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 la convergence de cet algorithme (en supposant la propriété de point fixe $TV^* = V^*$).

$\|V^{(k+1)} - V^*\|_\infty = \|T^*V^{(k)} - T^*V^*\|_\infty \leq \gamma\|V^{(k)} - V^*\|_\infty \leq ... \leq \gamma^{(k+1)}\|V^{(0)} - V^*\|_\infty \to 0$
On a donc une convergence asymptotique de l'algo (contrairement à policy iteration dont on est sûr qu'il converge vers la politique optimale si on considère un ensemble de politiques fini). L'avantage de cet algo c'est qu'on évite une évaluation complète des valeurs à chaque itération, ce qui peut être très coûteux.