Reinforcement learning - HW1

Notations: A MDP (S, A, p, γ)

- $\forall V : S \to \mathbb{R}$, $T^* V(s) = \max_a \left[r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s, a) V(s') \right]$ (Bellman operator optimal)

- $\forall V : S \to \mathbb{R}$, $T^\pi V(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s' \mid s, \pi(s)) V(s')$ (Bellman operator for a policy π)

Exercice 1: 1) V^* is the optimal value function, hence $\forall s \in S$, $T^* V^*(s) = V^*(s)$.

So V^* belongs to the feasible set of the optimization problem and a solution exists.

Let us take $\tilde{V}$ a solution for the optimization problem. We have $\operatorname{argmin}_{V \geq T^* V} \sum_{s \in S} V(s) = \tilde{V}$

By hypothesis, we have that $\sum_{s \in S} T^* \tilde{V}(s) \leq \sum_{s \in S} \tilde{V}(s)$ but since $\tilde{V}$ is the solution of the problem, we have $T^* \tilde{V} = \tilde{V}$. Hence $\tilde{V}$ is fixed point of the Bellman operator and $\tilde{V} = V^*$ since the fixed point is unique.

$$\boxed{V^* \text{ is the unique solution of the problem } \min_{V \geq T^* V} \sum_{s \in S} V(s)}$$

2) Let $T_S = (p(s' \mid s, a))_{(s,a) \in S \times A}$ the matrix representation of the propability kernel transition.

et $V = (V(s))_{s \in S}$

Let us pose as well, $\mathbb{1} = (1)_{|S|}$ et $\bar{r} = (r(s,a))_{a \in A}$, then it comes the following problem equivalence:

$$\min_{V \geq T^* V} \sum_{s \in S} V(s) \iff \boxed{\begin{cases} \min_{\mathbb{R}^{|S|}} \mathbb{1}^T V \\ (I_d - \gamma T_S) V - \bar{r} \geq 0 \end{cases}}$$

Exercice 2 - $V : S \mapsto \mathbb{R}$ a fonction. $\pi : S \to \mathcal{A}$ the greedy policy w.r.t the value fonction V:

$$\pi(s) \in \underset{a}{\operatorname{argmax}} \left[r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V(s') \right]$$

1.a) We obviously have $T^* V \geq T^\pi V$.

Reciprocally, $\forall s, a \in S \times \mathcal{A}$, $T^\pi V(s) \geq r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V(s')$

hence $\forall s$, $T^\pi V(s) \geq \max_a \left[r(s,a) + \sum_{s' \in S} p(s'|s,a) V(s') \right] = T^* V(s)$

As a conclusion, $\boxed{T^* V = T^\pi V}$.

1.b) $\| V^* - V^{\hat{\pi}} \|_\infty = \| V^* - T^{\hat{\pi}} \hat{V} + T^{\hat{\pi}} \hat{V} - V^{\hat{\pi}} \|_\infty$ but since V^* is a fixed point of T^* and by Minkowski inequality

$\| V^* - V^{\hat{\pi}} \|_\infty \leq \| T^* V^* - T^* \hat{V} \|_\infty + \| T^{\hat{\pi}} \hat{V} - T^{\hat{\pi}} V^{\hat{\pi}} \|_\infty$ and $T^*, T^{\hat{\pi}}$ are γ-contraction

$\| V^* - V^{\hat{\pi}} \|_\infty \leq \gamma \left[\| V^* - \hat{V} \|_\infty + \| \hat{V} - V^{\hat{\pi}} \|_\infty \right]$

$\| V^* - V^{\hat{\pi}} \|_\infty \leq \gamma \left[\| V^* - \hat{V} \|_\infty + \| \hat{V} - V^* \|_\infty + \| V^* - V^{\hat{\pi}} \|_\infty \right]$

hence we have the result; $\boxed{\| V^* - V^{\hat{\pi}} \|_\infty \leq \frac{2\gamma}{1-\gamma} \| V^* - \hat{V} \|_\infty}$

1-c) i) If $V^* = \hat{V}$ then $\| V^* - \hat{V} \|_\infty = 0$ and since 1-b we have $\| V^* - V^{\hat{\pi}} \|_\infty = 0$ so $\forall s \in S$, $V^*(s) = V^{\hat{\pi}}(s) \Rightarrow V^* = V^{\hat{\pi}} \Rightarrow \boxed{\hat{V} = V^{\hat{\pi}}}$

ii) Reciprocally, $\hat{V} = V^{\hat{\pi}}$ hence $V^{\hat{\pi}} = T^* V^{\hat{\pi}}$, i.e $V^{\hat{\pi}}$ is a fixed point of T^*. by definition, the fixed point is unique so $\boxed{V^{\hat{\pi}} = V^*}$.

Conclusion: $\boxed{V^{\hat{\pi}} = V^* \iff V^* = \hat{V}}$

2-a) - $Q : S \times A \to \mathbb{R}$ a function and the greedy policy $\hat{\pi} : S \to A$

$$\tilde{\pi}(s) = \underset{a}{\operatorname{argmax}}\, Q(s,a)$$

- $Q_{\tilde{\pi}} : S \times A \to \mathbb{R}$

$$Q_{\tilde{\pi}}(s,a) = r(s,a) + \gamma \sum_{s' \in S} p(s' | s,a) V_{\tilde{\pi}}(s')$$

- Q^* the optimal value function and $\varepsilon = \sup_{s} \left[Q^*(s,\pi^*(s)) - Q_{\tilde{\pi}}(s,\tilde{\pi}(s)) \right]$

Since π^* is the optimal policy, $V^* \geq V_{\tilde{\pi}} \Rightarrow \forall s \in S,\ Q^*(s,\pi^*(s)) \geq Q_{\tilde{\pi}}(s,\hat{\pi}(s))$

hence $\boxed{\varepsilon \geq 0}$

2-b) Let us pose $\forall (s,a) \in S \times A$, $\underline{Q^*(s,a) - \|Q^* - Q\|_\infty \leq Q(s,a) \leq Q^*(s,a) + \|Q^* - Q\|_\infty}$

Since $\tilde{\pi}$ is the greedy policy w.r.t Q, we have $\forall s \in S,\ Q(s,\pi^*(s)) \leq Q(s,\hat{\pi}(s))$

combining the inequations: $r(s,\pi^*(s)) + \gamma \sum_{s' \in S} p(s'|s,\pi^*(s))(Q(s',\pi^*(s')) - \varepsilon) \leq r(s,\tilde{\pi}(s)) + \gamma \sum_{s' \in S} p(s'|s,a)[Q(s',\tilde{\pi}(s')) - \varepsilon]$

i.e ; $Q^*(s,\pi^*(s)) - \|Q^* - Q\|_\infty \leq Q^*(s,\tilde{\pi}(s)) + \|Q^* - Q\|_\infty$

Then, $Q^*(s,\pi^*(s)) - Q_{\tilde{\pi}}(s,\hat{\pi}(s)) \leq 2\|Q^* - Q\|_\infty + Q^*(s,\hat{\pi}(s)) - Q_{\tilde{\pi}}(s,\hat{\pi}(s))$

by definition of $s \in \underset{s}{\operatorname{argmax}}\, Q^*(s,\pi^*(s)) - Q_{\tilde{\pi}}(s,\tilde{\pi}(s))$ $\quad \varepsilon \leq 2\|Q^* - Q\|_\infty + \gamma \sum_{s' \in S} p(s'|s,\hat{\pi}(s))[V^*(s') - V^{\tilde{\pi}}(s')]$

$$\varepsilon \leq 2\|Q^* - Q\|_\infty + \gamma \sum_{s' \in S} p(s'|s,\hat{\pi}(s))[Q^*(s',\pi^*(s')) - Q^{\tilde{\pi}}(s',\hat{\pi}(s'))]$$

$$\varepsilon \leq 2\|Q^* - Q\|_\infty + \gamma \varepsilon \sum_{s' \in S} p(s'|s,\hat{\pi}(s))$$

i.e $\quad (1-\gamma)\varepsilon \leq 2\|Q^* - Q\|_\infty$

and $\boxed{\varepsilon \leq \dfrac{2\|Q^* - Q\|_\infty}{1-\gamma}}$