RL, CS 2019: Homework 1

O. Darwiche, E. Oyallon

Notations:

We consider MDPs given by $(\mathcal{S}, \mathcal{A}, p, \gamma)$. Similarly, we define the optimal Bellman operator for any $V : \mathcal{S} \to \mathbb{R}$ by:

$$\forall s \in \mathcal{S}, \mathcal{T}^* V(s) = \max_a \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V(s') \right] \tag{1}$$

We write the Bellman operator for a deterministic policy π:

$$\forall s \in \mathcal{S}, \mathcal{T}^\pi V(s) = r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, \pi(s)) V(s') \tag{2}$$

Exercise 1

1. Let V^* being the optimal value function. Show that V^* is the solution of the following optimization problem:

$$\begin{aligned} &\min_V \sum_{s \in \mathcal{S}} V(s) \\ &\text{subject to } V \geq \mathcal{T}^* V \end{aligned} \tag{3}$$

2. (**Bonus**) Show it is a Linear Program. Propose an implementation. Is it practical? (see, for instance, `https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.linprog.html` ; the answer can be a *python* code or a pseudo-code.)

Exercise 2

1. a. Let $V : \mathcal{S} \to \mathbb{R}$ be any function. The greedy policy $\pi : \mathcal{S} \to \mathcal{A}$ with respect to V is defined by:

$$\pi(s) \in \arg\max_a \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V(s') \right] \tag{4}$$

Show that: $\mathcal{T}^* V = \mathcal{T}^\pi V$.

1. b. Let $\hat{V}$ be an approximation of V^* and $\hat{\pi}$ the greedy policy w.r.t. $\hat{V}$. We write $V^{\hat{\pi}}$ the value function corresponding to $\hat{\pi}$. Using the previous question, deduce that:

$$\|V^* - V^{\hat{\pi}}\|_\infty \leq \frac{2\gamma}{1-\gamma} \|V^* - \hat{V}\|_\infty \tag{5}$$

1. c. Show that $\hat{V} = V^*$ if and only if $\hat{V} = V^{\hat{\pi}}$.

2. a. Let $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be any function. Define the following greedy policy w.r.t. Q as:

$$\tilde{\pi}(s) \in \arg\max_a Q(s,a) \tag{6}$$

Let $Q_{\tilde{\pi}}$ be the action-value function of $\tilde{\pi}$, that is:

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

Let Q^* be the optimal value-function. Let $\epsilon = \sup_s Q^*(s, \pi^*(s)) - Q_{\tilde{\pi}}(s, \tilde{\pi}(s))$. Show that $\epsilon \geq 0$.

2. b. Show that:

$$\epsilon \leq \frac{2\|Q^* - Q\|_\infty}{1-\gamma} \tag{7}$$