Mathematics of Reinforcement Learning

Exercise Class 8

Exercise 1. Let $\gamma < 1$, π an arbitrary policy and define an operator T^π acting on functions $w\colon S \to \mathbb{R}$ by

$$T^\pi[w](s) \coloneqq \sum_{s' \in S,\, a \in A} \Big[r(s,a) + \gamma w(s') \Big] \pi(a \mid s) p(s' \mid s, a), \quad \text{for } s \in S.$$

Show that T^π has a unique fixed point $W\colon S \to \mathbb{R}$, that is, $W(s) = T^\pi[W](s)$ for all $s \in S$.

Exercise 2. Let $M = (S, A, D, p, r, \gamma)$ be a Markov Decision Model, with $\gamma < 1$. Show that there exists a Markov Decision Model $\tilde{M} = (S, A, D, \tilde{p}, \tilde{r}, \gamma)$ such that

$$\sup_{\pi \in \Pi^M \; \varepsilon\text{-soft}} V^\pi(s) = \sup_{\pi \in \Pi^{\tilde{M}}} V^\pi(s), \quad \text{for all } s \in S,$$

Also works for ε-greedy?

where Π^M and $\Pi^{\tilde{M}}$ denote the set of all policies in the Markov Decision Models M and $\tilde{M}$.
Hints: Start by defining a new transition probability function $\tilde{p}$ that incorporates the ε-softness into the new MDM and adjust the reward function. Show that there exists a transformation of the policies such that the value function remains invariant with respect to these changes. For the last point, use the results of exercise 1. ◇

Definition 1 (ε-soft optimal)**.** An ε-soft policy π^* is called *ε-soft optimal* if

$$V^{\pi^*}(s) = \sup_{\pi \; \varepsilon\text{-soft}} V^\pi(s) =: \tilde{V}^*(s), \quad \text{for all } s \in S.$$

Exercise 3. Let $\gamma < 1$ and π_0 be an arbitrary ε-soft policy and $\{\pi_n\}_{n \in \mathbb{N}} \subseteq \Pi$ be a sequence of ε-soft policies, where π_n is chosen to be ε-greedy with respect to $Q^{\pi_{n-1}}$, for all $n > 0$. Show that for some $N \in \mathbb{N}$, for all $m \geq N$, the policy π_m is ε-soft optimal.
Hint: Use exercise 2. ◇