Homework #12

Junwu Zhang
CME 241: Reinforcement Learning for Finance

Februray 14, 2020

Problem 1.
Prove the Epsilon-Greedy Policy Improvement Theorem

Solution.

The ϵ-Greedy Policy Improvement Theorem states: *For any ϵ-greedy policy π, the ϵ-greedy policy π' with respect to q_π is an improvement, $v_{\pi'}(s) \geq v_\pi(s)$*

To prove this, we can expand q_π and incorporate the ϵ-greedy idea as:

$$\begin{aligned} q_\pi\left(s, \pi'(s)\right) &= \sum_{a\in\mathcal{A}} \pi'(a|s) q_\pi(s,a) \\ &= \epsilon/m \sum_{a\in\mathcal{A}} q_\pi(s,a) + (1-\epsilon) \max_{a\in\mathcal{A}} q_\pi(s,a) \end{aligned} \tag{1}$$

Expanding the "greedy" part of the equation, we have:

$$(1-\epsilon) \max_{a\in\mathcal{A}} q_\pi(s,a) \geq (1-\epsilon) \sum_{a\in\mathcal{A}} \frac{\pi(a|s) - \epsilon/m}{1-\epsilon} q_\pi(s,a) \tag{2}$$

Relating it with previous expressions, we have:

$$\begin{aligned} q_\pi\left(s, \pi'(s)\right) &= \epsilon/m \sum_{a\in\mathcal{A}} q_\pi(s,a) + (1-\epsilon) \max_{a\in\mathcal{A}} q_\pi(s,a) \\ &\geq \epsilon/m \sum_{a\in\mathcal{A}} q_\pi(s,a) + (1-\epsilon) \sum_{a\in\mathcal{A}} \frac{\pi(a|s) - \epsilon/m}{1-\epsilon} q_\pi(s,a) \\ &= \sum_{a\in\mathcal{A}} \pi(a|s) q_\pi(s,a) = v_\pi(s) \end{aligned} \tag{3}$$

Using policy improvement theorem, we can see that $v_{\pi'}(s) \geq v_\pi(s)$. ∎

Problem 2.

Provide (with clear mathematical notation) the definition of Greedy in the Limit with Infinite Exploration (GLIE)

Solution. GLIE states that:

- All state-action pairs are explored infinitely many times,

$$\lim_{k\to\infty} N_k(s,a) = \infty \tag{4}$$

- The policy converges on a greedy policy,

$$\lim_{k\to\infty} \pi_k(a|s) = 1\left(a = \underset{a'\in\mathcal{A}}{\operatorname{argmax}} Q_k\left(s,a'\right)\right) \tag{5}$$

■