EE473 Deep Reinforcement Learning Homework 3

Excercise 3.26

For all $s \in \mathcal{S}$ and $a \in \mathcal{A}(\int)$, for a state-action pair (s, a), we can find the expected return for taking action a in state s as $q(s, a)$ that

$$\begin{aligned} q_*(s,a) &= \mathbb{E}\left[R_{t+1} + \gamma v_*(S_{t+1}) \,|S_t = s, A_t = a\right] \\ &= \sum_{s',r} p(s', r|\, s, a)\left[r + \gamma v_*(s')\right] \end{aligned}$$

Excercise 3.27

$$\begin{aligned} a_* &= \arg\max_a q_*(s,a) \\ \pi_*(s) &= \pi(a_*|s) \\ &= \pi\left(\arg\max_a q_*(s,a) \;\middle|\; s\right) \end{aligned}$$

Excercise 3.29

$$\begin{aligned} v_\pi(s) &= \mathbb{E}\left[G_t|S_t = s\right] \\ &= \sum_a \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a)\, v_\pi(s')\right] \pi(s,a) \end{aligned}$$

$$\begin{aligned} v_*(s) &= \max_\pi v_\pi(s) \\ &= \sum_a \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a)\, v_*(s')\right] \pi_*(s,a) \end{aligned}$$

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}_\pi\left[G_t \;\middle|\; S_t = s, A_t = a\right] \\ &= \mathbb{E}_\pi\left[R_{t+1} + \gamma G_{t+1} \;\middle|\; S_{t+1} = s', A_t = a\right] \\ &= r(s,a) + \gamma \sum_{s'} p(s'|s,a) \sum_{a'} q_\pi(a', s')\pi(s'|a') \end{aligned}$$

$$\begin{aligned} q_*(s,a) &= \max_\pi q_\pi(s,a) \\ &= \max_\pi \mathbb{E}_\pi\left[R_{t+1} + \gamma G_{t+1} \;\middle|\; S_{t+1} = s', A_t = a\right] \\ &= \mathbb{E}_{\pi_*}\left[R_{t+1} + \gamma G_{t+1} \;\middle|\; S_{t+1} = s', A_t = a\right] \\ &= r(s,a) + \gamma \sum_{s'} p(s'|s,a) \sum_{a'} q_*(a', s')\pi_*(s'|a') \end{aligned}$$

Excercise 4.1

Let state T denote the ternimal state

$$\begin{aligned} q_\pi(11, \texttt{down}) &= r(11, \texttt{down}) + v_\pi(T) \\ &= -1 + 0 \\ &= -1 \end{aligned}$$

$$\begin{aligned} q_\pi(7, \texttt{down}) &= r(7, \texttt{down}) + v_\pi(11) \\ &= -1 + (-14) \\ &= -15 \end{aligned}$$

Policy Iteration vs Value Iteration

Policy iteration requires a explicit representation of the policy π while the value representation does not so that it is simplier to solve using value iteration. But the policy iteration requires less steps to converge so that it is more effective. Even though the value iteration is simplier to implement, it is more effective using policy iteration. Hence I favor policy iteration rather than value iteration.