强化学习 HW2

1830029007 加兴华

1. (5 pts) In a finite state MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, suppose every reward function $r(s, a, s')$ is changed by an affine transformation to $a \cdot r(s, a, s') + b$, where $a > 0$. Show that the optimal policies remain unchanged.

$\forall \pi.$

$$V(s) = E_{a \sim \pi(a|s)} E_{s'} \left\{ \sum \gamma^t \cdot r_t \mid s_0 = s \right\}$$

$$V'(s) = E_{a \sim \pi(a|s)} E_{s'} \left\{ \sum \gamma^t \cdot (a\, r_t + b) \mid s_0 = s \right\}$$

$$= a \cdot E_{a \sim \pi(a|s)} E_{s'} \left\{ \sum \gamma^t r_t \mid s_0 = s \right\} + b \sum \gamma^t$$

$$= a \cdot V(s) + b \cdot \sum \gamma^t$$

$$\pi(s) = \arg\max_{\pi} E_{a \sim \pi(\cdot|s)} E_{s'} \left\{ r(s,a,s') + \gamma \cdot V(s') \right\}$$

$$\pi'(s) = \arg\max_{\pi} E_{a \sim \pi(\cdot|s)} E_{s'} \left\{ a\, r(s,a,s') + b + \gamma \cdot V'(s') \right\}$$

$$= \arg\max_{\pi} E_{a \sim \pi(\cdot|s)} E_{s'} \left\{ a\, r(s,a,s') + b + \gamma \cdot [a \cdot V(s) + b \sum \gamma^t] \right\}$$

$$= \arg\max_{\pi} E_{a \sim \pi(\cdot|s)} E_{s'} \left\{ a \cdot [r(s,a,s') + \gamma \cdot V(s)] + c \right\}$$

$$= \pi(s).$$

∴在给定相同初始策略的情况下，两种回报函数下求得的最优策略相同.

2. (10 pts) Recall the definition of the advantage function in Lecture 2:

$$g(\pi', \pi) = \mathcal{T}_{\pi'} v_\pi - v_\pi,$$

where $\mathcal{T}_{\pi'}$ is the Bellman operator associated with the policy π'. Show that π^* is the optimal policy if and only if for any π there holds $g(\pi, \pi^*) \leq 0$ (elementwise).

① π^* optimal $\Rightarrow$ $\forall \pi$ s.t. $g(\pi, \pi^*) \leq 0.$

$\forall \pi, s$

$$g(\pi,\pi^*)(s) = T_\pi V_{\pi^*}(s) - V_{\pi^*}(s)$$

$$= T_\pi V_{\pi^*}(s) - T_{\pi^*} V_{\pi^*}(s)$$

$$= E_{a\sim\pi(\cdot|s)} E_{s'}\{r(s,a,s') + \gamma \cdot V_{\pi^*}(s')\} - \max_a E_{s'}\{r(s,a,s') + \gamma \cdot V_{\pi^*}(s')\}$$

$$\leq 0. \quad (\text{elementwise})$$

$\therefore \forall \pi,\ g(\pi,\pi^*) \leq 0.$

② $\forall \pi,\ g(\pi,\pi^a) \leq 0 \Rightarrow \pi^a$ optimal

考虑其逆否：π^a not opimal $\Rightarrow \exists s,\pi$ s.t. $g(\pi,\pi^a)(s) > 0$.

取 $\pi = \pi^*$ 为最优策略

$$g(\pi^*,\pi^a)(s) = [T V_{\pi^a} - V_{\pi^a}](s)$$

$$= [V_{\pi^*} - V_{\pi^a} - (T V_{\pi^*} - T V_{\pi^a})](s) > 0$$

取 $s_0 = \arg\max_s \{V_{\pi^*}(s) - V_{\pi^a}(s)\}$

$\Rightarrow g(\pi^*,\pi^a)(s_0) = |V_{\pi^*}(s_0) - V_{\pi^a}(s_0)| - (T V_{\pi^*}(s_0) - T V_{\pi^a}(s_0))$ Note $\forall s,\pi$, s.t. $V_{\pi^*}(s) \geq V_\pi(s)$

$$= \|V_{\pi^*} - V_{\pi^a}\|_\infty - (T V_{\pi^*}(s_0) - T V_{\pi^a}(s_0))$$

$$\geq \|V_{\pi^*} - V_{\pi^a}\|_\infty - |T V_{\pi^*}(s_0) - T V_{\pi^a}(s_0)|$$

$$\geq \|V_{\pi^*} - V_{\pi^a}\|_\infty - \|T V_{\pi^*} - T V_{\pi^a}\|_\infty$$ Note $\|T V_{\pi^*} - T V_{\pi^a}\| = \arg\max |T V_{\pi^*}(s) - T V_{\pi^a}(s)|$

$$\geq \|V_{\pi^*} - V_{\pi^a}\|_\infty - \gamma \|V_{\pi^*} - V_{\pi^a}\|_\infty$$ Note T 为 γ 一压缩算子

$$= (1-\gamma)\|V_{\pi^*} - V_{\pi^a}\|_\infty > 0.$$

$\therefore \exists s = s_0,\ \pi = \pi^*$ s.t. $g(\pi,\pi^a)(s) > 0$. 逆否命题得证

$\therefore \forall \pi,\ g(\pi,\pi^a) \leq 0 \Rightarrow \pi^a$ optimal

4. (5 pts) Present and prove a general policy improvement lemma that has been used in the proof of Theorem 1 (about the improvement of ϵ-greedy policy) in Lecture 4.

Present: $q_\pi(s, \pi'(s)) \geq V_\pi(s).\ \forall s \Rightarrow V_{\pi'}(s) \geq V_\pi(s)\ \forall s$

Proof: $q_\pi(s, \pi'(s)) = \underset{a_0 \sim \pi'}{E}\{q_\pi(s, a_0)\}$

$$= \underset{a_0 \sim \pi'}{E}\ \underset{s_1}{E}\{r(s, a_0, s_1) + \gamma V_\pi(s_1)\}$$

$$\leq \underset{a_0 \sim \pi'}{E}\ \underset{s_1}{E}\{r(s, a, s_1) + \gamma q_\pi(s_1, \pi'(s_1))\}$$

$$= \underset{a_0 \sim \pi'}{E}\ \underset{s_1}{E}\{r(s, a_0, s_1)\} + \gamma \underset{\substack{a_0 \sim \pi' \\ s_1}}{E}\{q_\pi(s_1, \pi'(s_1))\}$$

$$= \cdots + \gamma \underset{\substack{a_0, a_1 \sim \pi' \\ s_1}}{E}\{q_\pi(s_1, a_1)\}$$

$\cdots$

$$\leq \underbrace{\underset{\substack{a_i \sim \pi' \\ s_i \sim p}}{E}\{\sum^{k} r_t \cdot \gamma^t \mid s_0 = s\} + \gamma^k \underset{\substack{a_i \sim \pi' \\ s_i \sim p}}{E}\{q_\pi(s_\infty, \pi'(s_\infty))\}}$$

$$\leq V_{\pi'}(s) + 0 \qquad \text{Note: } V_{\pi'}(s) = \lim_{k \to \infty}[\downarrow]$$

$\therefore\ V_{\pi'}(s) \geq q_\pi(s, \pi'(s)) \geq V_\pi(s).$