强化学习 HW1

18300290007 加兴华

1. (5 pts) Show that the infinite horizon discounted state value $v_\pi(s)$ has the following alternative expression:

$$v_\pi(s) = \mathbb{E}_{N\sim Geo(1-\gamma)}\left[\mathbb{E}\left[\sum_{t=0}^{N-1} r(s_t, a_t, s_{t+1})|s_0 = s\right]\right],$$

where $Geo(1-\gamma)$ denotes the geometric distribution with parameter $1-\gamma$. In word, we can rewrite $v_\pi(s)$ into an undiscounted form where the length of trajectory obeys the geometric distribution. In addition, compute $\mathbb{E}[N]$ which is referred to as planning horizon.

记 $r_t = r(s_t, a_t, s_{t+1})$

$$V_\pi(s) = \mathop{E}_{\substack{a_t\sim\pi(\cdot|s_t)\\ s_{t+1}\sim P(\cdot|s_t,a_t)}}\left[\sum_{t=0}^{\infty}\gamma^t r_t \,|\, s_0=s\right]$$

$$= E\left[\sum_{t=0}^{\infty}\left(\frac{\gamma^t-\gamma^\infty}{1-\gamma}\right)(1-\gamma) r_t \,|\, s_0=s\right]$$

$$= E\left[\sum_{t=0}^{\infty}\left(\sum_{k=t}^{\infty}\gamma^k\right)(1-\gamma) r_t \,|\, s_0=s\right]$$

$$= \sum_{t=0}^{\infty}\sum_{k=t}^{\infty} E\left[\gamma^k(1-\gamma) r_t \,|\, s_0=s\right]$$

$$= \sum_{k=0}^{\infty}\sum_{t=0}^{k} E\left[\gamma^k(1-\gamma) r_t \,|\, s_0=s\right]$$

$$= \sum_{k=0}^{\infty}\gamma^k(1-\gamma) E\left[\sum_{t=0}^{k} r_t \,|\, s_0=s\right] =: E_{N\sim Geo(1-\gamma)} E\left[\sum_{t=0}^{N-1} r_t \,|\, s_0=s\right]$$

$$EN = \sum_{k=1}^{\infty} k\cdot\gamma^{k-1}(1-\gamma)$$

$$= \frac{1}{\gamma}\sum_{k=1}^{\infty}(k+1)\cdot\gamma^k(1-\gamma) - \sum_{k=1}^{\infty}\gamma^{k-1}(1-\gamma)$$

$$= \frac{1}{\gamma}(EN-(1-\gamma)) - 1$$

$$= \frac{1}{\gamma}EN - \frac{1}{\gamma}$$

$$= \frac{1}{1-\gamma}$$

2. (10 pts) Prove the Bellman equation for state value and action value functions (i.e., Theorem 1 in the latest version of Lecture1.pdf).

$$V_\pi(s) = \mathop{E}_{\substack{a_t\sim\pi(\cdot|s_t)\\ s_{t+1}\sim P(\cdot|s_t,a_t)}}\left[\sum_{t=0}^{\infty}\gamma^t r_t \,|\, s_0=s\right]$$

$$= E\left[r_0 + \gamma\cdot\sum_{t=1}^{\infty}\gamma^{t-1} r_t \,|\, s_0=s\right]$$

$$= E[r_0 + \gamma \cdot E[E(\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s')] | s_0 = s]$$

$$= E[r_0 + \gamma \cdot E V_\pi(s') | s_0 = s]$$

$$= E[r_0 + \gamma \cdot V_\pi(s') | s_0 = s]$$ 由于 V_π 为常数

$$= \sum_a \pi(a|s) \sum_{s'} P(s'|s,a) (r(s,a,s') + \gamma \cdot V_\pi(s'))$$

$$q_\pi(s,a) = E[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a]$$

$$= E[r_0 + \gamma E[E(\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s')] | s_0 = s, a_0 = a]$$

$$= E[r_0 + \gamma \cdot V_\pi(s') | s_0 = s, a_0 = a]$$

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

3. (5 pts) Show that the Bellman operator

$$\mathcal{T}_\pi v = r_\pi + \gamma P^\pi v$$

is a contraction with respect to infinity norm.

$\| T_\pi v_1 - T_\pi v_2 \|_\infty = \gamma \| P^\pi (v_1 - v_2) \|_\infty, \ \forall v_1, v_2.$

考虑 $\| P^\pi x \|_\infty$：

$\because$ P^π 为非负矩阵且行和 $=1$.

$$\therefore \| P^\pi x \|_\infty = \max_i | \sum_j P^\pi_{ij} x_j |$$

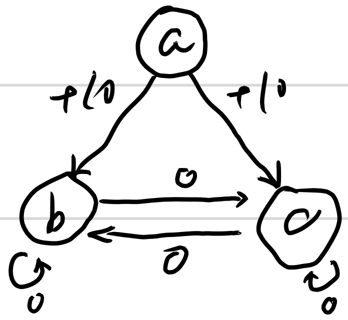
$$\leq \max_i | \sum_j P^\pi_{ij} |x_j| |$$

$$= \max_i \sum_j P^\pi_{ij} |x_j|$$

$$\leq \max_j |x_j| = \|x\|_\infty$$ 由于 $\max_j |x_j| = (\sum_j P^\pi_{ij}) \cdot \max_j |x_j| \geq \sum_j P^\pi_{ij} |x_j|$

$\therefore$ T_π 为收缩算子

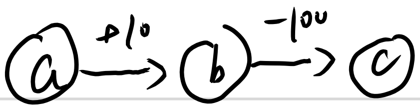
4. (5 pts) Whether the optimal policy is unique? Prove or disprove by a counter example.



如图. 对a,b,c而言，任何策略的一步回报都相同.

因此任何策略都最优，不唯一

5. (optional) Let π_k be the policy extracted from the k-th iteration of the value iteration. Is it always that $v_{\pi_{k+1}}(s) \geq v_{\pi_k}(s),\ \forall s$? Prove or disprove by a counter example.



$V_{\pi_0}=0$. 显然. $V_{\pi_1}(b) = -100 < V_{\pi_0}(b)$.

6. (5 pts) Prove the policy improvement result (i.e., Theorem 2 in the latest version of Lecture2.pdf).

$$q_\pi(s, \pi'(s)) = \sum_{s'} p(s'|s,\pi'(s))\left(r(s,\pi'(s),s') + \gamma v_\pi(s')\right)$$

$$= \max_a \sum_{s'} p(s'|s,a)\left(r(s,a,s') + \gamma v_\pi(s')\right)$$

$$=: \max_a f(a) =: f(a^*)$$

$$= \left(\sum_a \pi(a|s)\right) \cdot f(a^*) \qquad \text{由于 } \sum \pi(a|s) = 1$$

$$\geq \sum_a \pi(a|s) \cdot f(a) \qquad \text{由于 } f(a) \geq 0,\ \forall a$$

$$= v_\pi(s).$$

得证.