强化学习作业3

18300290007 加兴华

1. (10 pts) Prove the convergence of Q-Learning (i.e., Theorem 4 of lecture5.pdf).

Theorem 4. Q-Learning for finite-state and finite-action MDPs converges to the optimal action-value, i.e., $q_t \to q^*$ with probability one if the stepsizes α_t satisfy

$$\sum_{t=0}^{\infty} \alpha_t(s,a) = \infty, \quad \sum_{t=0}^{\infty} \alpha_t^2(s,a) < \infty$$

for all (s,a).

Q-learning 的更新式：

$$\Delta_{t+1}(s,a) = (1-\alpha_t(s,a))\,\Delta_t(s,a) + \alpha_t(s,a)\,F_t(s,a)$$

其中 $\Delta_t(s,a) = q_t(s,a) - q_*(s,a)$，$F_t(s,a) = r(s,a,s') + \gamma \cdot \max_{a'} q_t(s',a') - q_*(s,a)$

对于 $F_t(s,a)$，满足：

① $\| E[F_t(s,a) \mid F_t] \|_\infty \le \gamma \| \Delta_t \|_\infty$

$$E[F_t(s,a) \mid F_t] = \sum_{s'} P(s' \mid s,a)\left(r(s,a,s') + \gamma \cdot \max_{a'} q_t(s',a') - q_*(s,a) \right)$$

$$= (T q_t)(s,a) - q_*(s,a)$$

$$= (T q_t)(s,a) - (T q_*)(s,a)$$

由于 [1]：

$$\| E[F_t(s,a) \mid F_t] \|_\infty = \| T q_t - T q_* \|_\infty$$

$$\le \gamma \| q_t - q_* \|_\infty = \gamma \| \Delta_t \|_\infty$$

② $Var[F_t(s,a) \mid F_t] \le C \cdot (1 + \|\Delta_t\|_\infty^2)$

$$Var[F_t(s,a) \mid F_t] = E\{ (F_t(s,a) \mid F_t - E[F_t(s,a) \mid F_t])^2 \}$$

$$= E\left\{ \left(r(s,a,s') + \gamma \max_{a'} q_t(s',a') - (Tq_t)(s,a) \right)^2 \right\}$$

$$\leq E\left\{ \left(r(s,a,s') + \gamma \|\Delta_t\|_\infty + \max_{a'} q_*(s',a') - (Tq_t)(s,a) \right)^2 \right\}$$

$$\leq E\left\{ \left(m_1 + \gamma \|\Delta_t\|_\infty + m_2 - \underbrace{\min_{s'} \left(r(s',a') + \max_{a'} q_t(s',a') \right)}_{\overset{!!}{m_3}} \right)^2 \right\}$$

$$\leq M_1 + \gamma^2 \cdot M_2 \cdot \|\Delta_t\|_\infty^2$$

$$=: C(1 + \|\Delta_t\|_\infty^2)$$

根据随机逼近理论，①&②& $\sum \alpha_t(s,a) = \infty$ & $\sum \alpha_t^2(s,a) < \infty$ $\Rightarrow$ $\Delta_t \xrightarrow{P} 0$. 也即 $q_t \xrightarrow{P} q_*$

Proof [1]: $\|Tq_t - Tq_*\|_\infty \leq \gamma \|q_t - q_*\|_\infty$

$$\|Tq_t - Tq_*\|_\infty = \max_{s,a} \left| \sum_{s'} p(s'|s,a) \left(r(s,a,s') + \gamma \max_{a'} q_t(s',a') - r(s,a,s') - \gamma \max_{a'} q_*(s',a') \right) \right|$$

$$= \max_{s,a} \gamma \cdot \left| \sum_{s'} p(s'|s,a) \left[\max_{a'} (q_t(s',a') - q_*(s',a')) \right] \right|$$

$$\leq \max_{s,a} \gamma \cdot \sum_{s'} p(s'|s,a) \cdot \max_{a'} |q_t(s',a') - q_*(s',a')|$$

$$\leq \gamma \|q_t - q_*\|_\infty$$

3. (5 pts) Given a policy π, let v_π be the state value that corresponds to π and $\mathcal{T}_\pi$ be the corresponding Bellman operator. Recall the definition of Bellman error (under infinity norm) for a vector v:

$$\mathrm{BE}(v) = \|v - \mathcal{T}_\pi v\|_\infty.$$

Show that

$$\|v - v_\pi\|_\infty \leq \frac{\mathrm{BE}(v)}{1-\gamma}.$$

$$\|v - v_\pi\|_\infty = \|v - T_\pi v + T_\pi v - v_\pi\|_\infty$$

$$\leq \|v - T_\pi v\|_\infty + \|T_\pi v - v_\pi\|_\infty$$

$$= BE(v) + \|T_\pi v - T_\pi v_\pi\|_\infty$$

$$\leq BE(v) + \gamma \|v - v_\pi\|_\infty$$

移项得证.

4. (5 pts) Consider the following softmax parameterization

$$\pi_\theta(a|s) = \frac{\exp(\theta_a)}{\sum_{a' \in \mathcal{A}} \exp(\theta_{a'})},$$

where $\theta = (\theta_a)_{a \in \mathcal{A}}$. Calculate $\nabla_\theta \log \pi_\theta(a|s)$.

$$\log \pi_\theta(a|s) = \theta_a - \log\left(\sum_{a'} \exp(\theta_{a'})\right)$$

$$\frac{\partial}{\partial \theta_i} \log \pi_\theta(a|s) = 1(i=a) - \frac{\exp(\theta_i)}{\sum_{a'} \exp(\theta_{a'})}$$

$$\nabla_\theta \log \pi_\theta(a|s) = \left(1(i=a) - \frac{\exp(\theta_i)}{\sum_{a'} \exp(\theta_{a'})}\right)_{1\times|A|}$$

5. (10 pts) Recall the definition of state visitation measure

$$d_\mu^\pi(s) = \mathbb{E}_{s_0 \sim \mu}\left[d_{s_0}^\pi(s)\right] = \mathbb{E}_{s_0 \sim \mu}\left[(1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}\left[s_t = s | s_0, \pi\right]\right],$$

where $(s_0, a_0, s_1, a_1, \cdots)$ is trajectory starting from initial distribution μ and then following policy π. Let T obey the geometric distribution, i.e., $\mathbb{P}[T = t] = \gamma^t(1-\gamma), \quad t = 0, 1 \cdots$. Show that

$$\mathbb{P}[s_T = s] = d_\mu^\pi(s).$$

Then suggest a way to sample from d_μ^π.

$$P(s_T = s) = \sum_{t=0}^{\infty} P(T=t) \cdot P(s_t = s)$$

$$= \sum_{t=0}^{\infty} \gamma^t (1-\gamma) \cdot P(s_t = s)$$

$$= \sum_{t=0}^{\infty} \gamma^t (1-\gamma) \cdot \sum_{x \sim \mu} P(s_0 = x) \cdot P(s_t = s | x, \pi)$$

$$= \sum_{x\sim\mu} P(s_0 = x)(1-\gamma) \sum_{t=0}^{\infty} \gamma^t P(s_t = s | x, \pi)$$

$$= d_\mu^\pi(s)$$

6. (5 pts) Show the following expression for the Fisher information matrix in NPG for policy optimization (see pg. 8 of lecture8.pdf):

$$F(\theta) = \mathbb{E}_{\tau \sim \mathbb{P}_\mu^{\pi_\theta}}\left[\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t)(\nabla_\theta \log \pi_\theta(a_t|s_t))^T\right].$$

$$F(\theta) = E_{p_\theta}\left[\nabla_\theta \log p_\theta(x) (\nabla_\theta \log p_\theta(x))^T\right]$$

$$= E_{\tau \sim p_\mu^{\pi_\theta}}\left[\nabla_\theta \log p_\mu^{\pi_\theta}(\tau) (\nabla_\theta \log p_\mu^{\pi_\theta}(\tau))^T\right]$$

$$\log p_\mu^{\pi_\theta}(\tau) = \log \mu(s_0) + \sum_{t=0}^{\infty} \log \pi_\theta(a_t|s_t) + \sum_{t=0}^{\infty} \log P(s_{t+1}|s_t, a_t)$$

$$\Rightarrow \nabla_\theta \log p_\mu^{\pi_\theta}(\tau) = \sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t)$$

$$\therefore F(\theta) = E_{\tau\sim p_\mu^{\pi_\theta}}\left[\left(\sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(a_t|s_t)\right)(\cdots)^T\right]$$

$$= E_{\tau\sim p_\mu^{\pi_\theta}}\Big[\underbrace{\sum_{t=0}^{\infty}(\nabla_\theta \log \pi_\theta(a_t|s_t))(\cdots)^T}_{\text{Part I}} + \underbrace{\sum_{\substack{u,v=0\\ u\neq v}}^{\infty}(\nabla_\theta \log \pi_\theta(a_u|s_u))(\nabla_\theta \log \pi_\theta(a_v|s_v))^T}_{\text{Part II}}\Big]$$

$$E_{\tau\sim p_\mu^{\pi_\theta}}[\text{Part II}] = 2\sum_{u=0}^{\infty}E_\tau\left[\sum_{v=0}^{u-1}(\nabla_\theta \log \pi_\theta(a_u|s_u))(\nabla_\theta \log \pi_\theta(a_v|s_v))^T\right]$$

$$= 2\sum_{u=0}^{\infty}E_\tau\left[(\nabla_\theta \log \pi_\theta(a_u|s_u))(\nabla_\theta \sum_{v=0}^{u-1}\log \pi_\theta(a_v|s_v))^T\right]$$

$$= 2\sum_{u=0}^{\infty}E_{a_u}E_{s_0\to s_u}\left[\nabla_\theta \log \pi_\theta(a_u|s_u)\cdot(\nabla_\theta \sum_{v=0}^{u-1}\log \pi_\theta(a_v|s_v))^T\right]$$

$$= 2\sum_{u=0}^{\infty}E_{a_u}\left[\nabla_\theta \log \pi_\theta(a_u|s_u)\cdot E_{s_0\to s_u}[\nabla_\theta \sum_{v=0}^{u-1}\log \pi_\theta(a_v|s_v)^T]\right]$$

$$= 2\sum_{u=0}^{\infty}E_{a_u}\left[\nabla_\theta \log \pi_\theta(a_u|s_u)\cdot k(u)\right]$$

$$= 0.$$

$$\therefore F(\theta) = E_{\tau\sim p_\mu^{\pi_\theta}}[\text{Part I}].$$

7. (5 pts) Show the statement "$J(\theta)$ and $L_{\theta_{\text{old}}}(\theta)$ matches at θ_{old} up to first derivative" in pg. 11 of lecture8.pdf.

$$J(\theta) = J(\theta_{old}) + \frac{1}{1-\gamma}E_{s\sim d_\mu^{\pi_\theta}}E_{a\sim\pi_\theta(\cdot|s)}[A^{\pi_{\theta_{old}}}(s,a)]$$

$$L_{\theta_{old}}(\theta) = J(\theta_{old}) + \frac{1}{1-\gamma}E_{s\sim d_\mu^{\pi_{\theta_{old}}}}E_{a\sim\pi_\theta(\cdot|s)}[A^{\pi_{\theta_{old}}}(s,a)]$$

易见 $J(\theta_{old}) = L_{old}(\theta_{old})$;

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\nabla_\theta\int d_\mu^{\pi_\theta}(s)\cdot\pi_\theta(a|s)\cdot(q_{\pi_{\theta_{old}}}(s,a) - V_{\pi_{\theta_{old}}}(s))\,ds\,da$$

$$= \frac{1}{1-\gamma}\int[\nabla_\theta d_\mu^{\pi_\theta}(s)\,\pi_\theta(a|s) + d_\mu^{\pi_\theta}(s)\nabla_\theta\pi_\theta(a|s)](q_{\pi_{\theta_{old}}}(s,a) - V_{\pi_{\theta_{old}}}(s))\,ds\,da$$

$$\nabla_\theta L_{\theta_{old}}(\theta) = \frac{1}{1-\gamma}\int d_\mu^{\pi_{\theta_{old}}}(s)\cdot\nabla_\theta\pi_\theta(a|s)\cdot(q_{\pi_{\theta_{old}}}(s,a) - V_{\pi_{\theta_{old}}}(s))\,ds\,da$$

$$\therefore \nabla_\theta J(\theta_{old}) - \nabla_\theta L_{\theta_{old}}(\theta_{old})$$

$$= \frac{1}{1-\gamma} \int \nabla_\theta d_\mu^{\pi_{\theta_{old}}}(s) \cdot \pi_{\theta_{old}}(a|s) \left(q_{\pi_{\theta_{old}}}(s,a) - V_{\pi_{\theta_{old}}}(s) \right) ds\, da$$

其中 $\int \pi_{\theta_{old}}(a|s) \left(q_{\pi_{\theta_{old}}}(s,a) - V_{\pi_{\theta_{old}}}(s) \right) da$

$$= \left[\sum_a \pi_{\theta_{old}}(a|s) \cdot q_{\pi_{\theta_{old}}}(s,a) \right] - V_{\pi_{\theta_{old}}}(s)$$

$$= V_{\pi_{\theta_{old}}}(s) - V_{\pi_{\theta_{old}}}(s) = 0.$$

$\therefore \nabla_\theta J(\theta_{old}) - \nabla_\theta L_{\theta_{old}}(\theta_{old}) = 0.$

$\therefore J(\theta)$ 与 $L_{old}(\theta)$ 在 θ_{old} 处匹配至一阶.