به نام خدا

علی قاسم زاده ۴۰۱۱۶۴۳۳۹

تمرین ۸ RL

۱-

a)

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{\tau\sim\pi_\theta}\left\{\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right\} = \nabla_\theta \int P(\tau|\theta) \sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\, d\tau$$

$$= \int \left((\nabla_\theta P(\tau|\theta)) \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right) d\tau$$

$$P(\tau|\theta) = P(s_0) \prod_{t=0}^{\infty} \pi_\theta(a_t|s_t) P(s_{t+1}|s_t, a_t) \xrightarrow{\ln}$$

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

$$\xrightarrow{\partial/\partial\theta} \sum_{t=0}^{\infty} \nabla_\theta \ln \pi_\theta(a_t|s_t)$$

$$\Rightarrow \int \underbrace{\nabla_\theta \ln P(\tau|\theta)} P(\tau|\theta) \left(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right) d\tau =$$

$$\mathbb{E}_{\tau\sim\pi_\theta}\left\{\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \sum_{k=0}^{\infty} \nabla_\theta \ln \pi_\theta(a_k|s_k)\right\}$$

با causality و از اسلاید ها داریم که می شود:

$$\cancel{\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left\{\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \sum_{k=t}^{\infty} \nabla_\theta \ln \pi_\theta(a_k|s_k)\right\}}$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left\{ \sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t) \left(\sum_{k=0}^{\infty} \gamma^k r(s_{t+k}, a_{t+k}) \right) \right\}$$

$$Q^{\pi_\theta}(s_t, a_t) = \mathbb{E}_{\tau \sim \pi_\theta}\left\{ \sum_{k=0}^{\infty} \gamma^k r(s_{t+k}, a_{t+k}) \,\middle|\, s_t, a_t \right\}$$

$$\Rightarrow \nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left\{ \sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t)\, Q^{\pi_\theta}(s_t, a_t) \right\}$$

$$\mu_{s_0}^{\pi_\theta}(s) := \sum_{t=0}^{\infty} \gamma^t \Pr^{\pi_\theta}(s_t = s | s_0)$$

move براساس trajectory-based عبارت است از

$$\longrightarrow \nabla_\theta J(\theta) = \sum_s \mu_{s_0}^{\pi_\theta}(s) \sum_a \pi_\theta(a|s)\, \nabla_\theta \log \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a)$$

از آنجایی که $\mu_{s_0}^{\pi_\theta}(s)$ توزیع نیست و باید داشته باشیم که:

$$\sum_s \mu_{s_0}^{\pi_\theta}(s) = 1 \longrightarrow \sum_{t=0}^{\infty} \gamma^t \sum_s \Pr^{\pi_\theta}(s_t = s | s_0) =$$

$$\sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma} \longrightarrow d_{s_0}^{\pi_\theta}(s) = (1-\gamma)\, \mu_{s_0}^{\pi_\theta}(s) \checkmark$$

حالا داریم که:

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma} \sum_s d_{s_0}^{\pi_\theta}(s) \sum_a \pi_\theta(a|s)\, \nabla_\theta \log \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a)$$

$$= \frac{1}{1-\gamma}\, \mathbb{E}_{s \sim d_{s_0}^{\pi_\theta}(s)}\left\{ \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}\left\{ \nabla_\theta \log \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a) \right\} \right\}$$

:Compatible Function Approx Theorem-۳۰۱

b) داریم که،

$$\nabla_\phi(\varepsilon) = \nabla_\phi \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[\left(Q^{\pi_\theta}(s,a) - Q_\phi(s,a)\right)^2\right] =$$

$$-2\, \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left\{\left(Q^{\pi_\theta}(s,a) - Q_\phi(s,a)\right)\nabla_\phi Q_\phi(s,a)\right\} = 0$$

$$\Rightarrow \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[Q^{\pi_\theta}(s,a)\nabla_\phi Q_\phi(s,a)\right] =$$

$$\mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left[Q_\phi(s,a)\nabla_\phi Q_\phi(s,a)\right] =$$

$$\mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \left\{Q_\phi(s,a)\nabla_\theta \log\pi_\theta(a|s)\right\}$$

اینم با $\frac{1}{1-\gamma}$ اسکیل بشه حکم سوال ثابت میشه.

۲-

a) داریم که

$$A_\pi(s_t, a_t) = Q_\pi(s_t,a_t) - V_\pi(s_t).$$

طبق معادله بلمن نیز داریم،

$$Q_\pi(s_t,a_t) = r(s_t,a_t) + \gamma\, \mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\left[V_\pi(s_{t+1})\right]$$

Scanned with CamScanner

$$\Rightarrow \underbrace{\cancel{Q_\pi(s_t, a_t)}}_{A_\pi(s_t, a_t)} = r(s_t, a_t) + \gamma \mathop{\mathbb{E}}_{s_{t+1}} V_\pi(s_{t+1}) - V_\pi(s_t)$$

$$\longrightarrow r(s_t, a_t) = A_\pi(s_t, a_t) + V_\pi(s_t) - \gamma \mathop{\mathbb{E}}_{s_{t+1}} V_\pi(s_{t+1})$$

expected return of π'

حالا داریم که

$$\mathbb{E}\{R(\pi')\} = \mathop{\mathbb{E}}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] =$$

با $\eta(\pi')$ نشونش میدیم

$$\mathop{\mathbb{E}}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)\right] + \mathop{\mathbb{E}}_{\pi'}\left[\sum \gamma^t V_\pi(s_t)\right] - \mathop{\mathbb{E}}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^{t+1} V_\pi(s_{t+1})\right]$$

$$\mathop{\mathbb{E}}_{s_0 \sim \rho_0}\left[V_\pi(s_0)\right]$$

$$= \mathop{\mathbb{E}}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)\right] + \mathop{\mathbb{E}}_{s_0 \sim \rho_0}\left[V_\pi(s_0)\right] \Rightarrow$$

$$\eta(\pi') = \eta(\pi) + \mathop{\mathbb{E}}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)\right]$$

$$\Rightarrow \eta(\pi') = \eta(\pi) + \mathop{\mathbb{E}}_{s_0, a_0, \ldots \sim \pi'}\left\{\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)\right\} \checkmark$$

b)

$$\mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)\right] = \sum_{t=0}^{\infty} \sum_{s_t} \gamma^t p(s_t) \sum_{a_t} \pi'(a_t|s) A_\pi(s_t, a_t)$$

$$= \sum_{s} P_{\pi'}(s_t) \sum_{a} \pi'(a|s) A_\pi(s,a)$$

c) داریم که، $V_\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q_\pi(s,a)\right]$ و از آنجایی که $\mathbb{E}_{a\sim\pi(\cdot|s)}\left[A_\pi(s,a)\right]$ صفر است، خواهیم داشت که:

$$\bar{A}(s) = \mathbb{E}_{a'\sim\pi'(\cdot|s)}\left[A_\pi(s,a')\right] =$$

$$\mathbb{E}_{(a,a')\sim(\pi,\pi')|s}\left[A_\pi(s,a') - A_\pi(s,a)\right] =$$

$$p(a\neq a'|s)\, \mathbb{E}_{(a,a')\sim(\pi,\pi')|s,\, a\neq a'}\left[A_\pi(s,a') - A_\pi(s,a)\right]$$

$$\leq \alpha\, \mathbb{E}_{(a,a')\sim(\pi,\pi')|s,\, a\neq a'}\left[A_\pi(s,a') - A_\pi(s,a)\right]$$

$$\Rightarrow |\bar{A}(s)| \leq \alpha\left(\max_{s,a'} |A_\pi(s,a')| + \max_{s,a} |A_\pi(s,a)|\right)$$

$$= 2\alpha \max_{s,a} |A_\pi(s,a)|$$

d) با احتمال $(1-\alpha)^t$ تمام action ها در دو policy یکسان خواهد بود

← با احتمال $1-(1-\alpha)^t$ متفاوت خواهد بود.

$$\left|\mathbb{E}_{s_t\sim\pi'}[\bar{A}(s_t)]-\mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)]\right|=(1-(1-\alpha)^t)\left|\mathbb{E}_{\substack{s_t\sim\pi'|\\ a\neq a'}}[\bar{A}(s_t)]-\mathbb{E}_{s_t\sim\pi|a\neq a'}[\bar{A}(s_t)]\right|$$

$$\leq (1-(1-\alpha)^t)\left(\left|\mathbb{E}_{s_t\sim\pi'|a\neq a'}[\bar{A}(s_t)]\right|+\left|\mathbb{E}_{s_t\sim\pi|a\neq a'}[\bar{A}(s_t)]\right|\right)$$

$$\leq (1-(1-\alpha)^t)\left(2\alpha\max_{s,a}|A_\pi(s,a)|+2\alpha\max_{s,a}|A_\pi(s,a)|\right)$$

$$=4\alpha(1-(1-\alpha)^t)\max|A_\pi(s,a)|$$

e)

$$|\eta(\pi')-L_\pi(\pi')|=\left|\sum_{t=0}^{\infty}\gamma^t\left(\mathbb{E}_{s_t\sim\pi'}[\bar{A}(s_t)]-\mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)]\right)\right|$$

$$\leq\sum_{t=0}^{\infty}\gamma^t\left|\mathbb{E}_{s_t\sim\pi'}[\bar{A}(s_t)]-\mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)]\right|\leq$$

$$\sum_{t=0}^{\infty}\gamma^t\cdot 4\alpha(1-\cancel{\alpha}(1-\alpha)^t)\varepsilon=4\alpha\varepsilon\left(\sum_{t=0}\gamma^t-\sum_{t=0}\gamma^t(1-\alpha)^t\right)$$

$$=4\alpha\varepsilon\left(\frac{1}{1-\gamma}-\frac{1}{1-\gamma(1-\alpha)}\right)$$

$$\leq 4\alpha\varepsilon\left(\frac{\alpha\gamma}{(1-\gamma)^2}\right)=\frac{4\alpha^2\varepsilon\gamma}{(1-\gamma)^2}$$

۴) اگر α را طوری انتخاب کنیم که $D_{TV}^{max}(\pi,\pi') \le \alpha$، آنگاه داریم که

$$\left|\eta(\pi') - L_\pi(\pi')\right| \le \frac{4\alpha^2\epsilon\gamma}{(1-\gamma)^2} \Rightarrow$$

$$\eta(\pi') - L_\pi(\pi') \ge -\frac{4\alpha^2\epsilon\gamma}{(1-\gamma)^2} \Rightarrow$$

$$\eta(\pi') \ge L_\pi(\pi') - \frac{4\alpha^2\epsilon\gamma}{(1-\gamma)^2}$$

$$= L_\pi(\pi') - \frac{4\epsilon\gamma}{(1-\gamma)^2} D_{TV}^{max}(\pi,\pi')^2 \ge$$

$$L_\pi(\pi') - \frac{4\epsilon\gamma}{(1-\gamma)^2} D_{KL}^{max}(\pi,\pi')$$

$$\boxed{\begin{aligned} & D_{TV}(\pi(\cdot|s) \,\|\, \pi'(\cdot|s))^2 \le D_{KL}(\pi(\cdot|s) \,\|\, \pi'(\cdot|s)) \\ & \Rightarrow D_{TV}^{max}(\pi,\pi') \le D_{KL}^{max}(\pi,\pi') \end{aligned}}$$

به خاطر این نامساوی بالا برقرار است.