Reinforcement Learning

HW 1-2

309535018
許俊廷

Problem 1

(a) Q-value and discounted state visitation

$$\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} E_{s\sim d^{\pi_\theta}_\mu} E_{a\sim\pi_\theta(\cdot|s)} [Q^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a|s)]$$

subtract a baseline function $B(s)$ from the policy gradient

$$E_{s\sim d^{\pi}_{\theta}} E_{a\sim\pi_\theta(\cdot|s)} [(Q^{\pi_\theta}(s,a) - B(s)) \nabla_\theta \log \pi_\theta(a|s)]$$

$$E_{s\sim d^{\pi}_{\theta}} E_{a\sim\pi_\theta(\cdot|s)} [B(s) \nabla_\theta \log \pi_\theta(a|s)]$$

$$= \sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) B(s)$$

$$= \sum_s d^{\pi}(s) B(s) \nabla_\theta \sum_a \pi_\theta(a|s) = 0$$

(b) REINFORCE

$$\nabla_\theta V^{\pi_\theta}(\mu) = E_{\tau\sim p^{\pi_\theta}_\mu} [\sum_{t=0}^{\infty} \gamma^t Q^{\pi_\theta}(s_t,a_t) \nabla_\theta \log \pi_\theta(a_t|s_t)]$$

subtract a baseline function $B(s)$ from the policy gradient

$$E_{\tau\sim p^{\pi_\theta}_\mu} [\sum_{t=0}^{\infty} \gamma^t [Q^{\pi_\theta}(s_t,a_t) - B(s_t)] \nabla_\theta \log \pi_\theta(a_t|s_t)]$$

$$E_{\tau\sim p^{\pi_\theta}_\mu} [B(s_t) \nabla_\theta \log \pi_\theta(a_t|s_t)]$$

$$= \sum_s P(s_t=s) \sum_a \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) B(s)$$

$$= \sum_s P(s_t=s) B(s) \nabla_\theta \sum_a \pi_\theta(a|s) = 0$$

(c)

$$V\{(G(t) - B(s)) \frac{\delta}{\delta\theta_i} \log \pi_\theta(a_t|s_t)\}$$

$$= \sum_s P(s_t = s)\left(E[(G(t) - B(s))^2 (\tfrac{\delta}{\delta\theta_i} \log \pi_\theta(a_t|s))^2 | s]\right)$$

$$- \left(\sum P(s_t = s) E\{(G(t) - B(s)) \tfrac{\delta}{\delta\theta_i} \log \pi_\theta(a_t|s_t) | s]\right)^2$$

$$= \sum_s P(s_t = s)\left(\sum_a \pi_\theta(a|s)(E[(G(t) - B(s))^2 (\tfrac{\delta}{\delta\theta_i} \log \pi_\theta(a|s))^2 | s, a]\right)$$

$$- \left(\sum P(s_t = s) \sum_a \pi_\theta(a|s) E[(G(t) - B(s)) \tfrac{\delta}{\delta\theta_i} \log \pi_\theta(a|s) | s, a]\right)^2$$

$$= \cancel{\sum_s}$$

$$= \sum_s P(s_t = s)\left(\sum \pi_\theta(a|s)(\tfrac{\delta}{\delta\theta_i} \log \pi_\theta(a|s))^2 E[(G_t - B(s))^2 | s, a]\right)$$

$$- \left(\sum P(s_t = s) \sum_a \pi_\theta(a|s) \tfrac{\delta}{\delta\theta_i} \log \pi_\theta(a|s) E[G_t | s, a]\right)^2$$

$$= \sum_s P(s_t = s) \sum_a C_a (E[2B(s) G_t - B(s)^2 | s, a])$$

309505018
[illegible]

Problem 2

(a) Show that $E_{\tau \sim p_{\mu}^{\pi_\theta}}\left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t)\right] = \frac{1}{1-\gamma} E_{s \sim d^{\pi_\theta}} E_{a \sim \pi_\theta(\cdot|s)}\left[f(s,a)\right]$

pf:

$$RHS = \frac{1}{1-\gamma} \sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) f(s,a)$$

$$= \frac{1}{1-\gamma} \sum_s \left(\sum \mu(s)(1-\gamma) \sum_{t=0}^{\infty} \gamma^t p(s_t=s|s_0, \pi_\theta)\right) \cdot \sum_a \pi_\theta(a|s) f(s,a)$$

$$= \sum_{s_0} \mu(s_0) \sum_s \sum_a \pi_\theta(a|s) \sum_{t=0}^{\infty} \gamma^t p(s_t=s|s_0, \pi_\theta) f(s,a)$$

$$= \sum_{s_0} \mu(s_0) \sum_s \sum_a \sum_{t=0}^{\infty} \gamma^t p(s_t=s, a_t=a|s_0, \pi_\theta) f(s,a)$$

$$= \sum_{s_0} \mu(s_0) \sum_\tau p(\tau|s_0) \sum_{t=0}^{\infty} \gamma^t f(s_t, a_t)$$

$$= \sum_\tau p(\tau) \sum_{t=0}^{\infty} \gamma^t f(s_t, a_t)$$

$$= E_{\tau \sim p_{\mu}^{\pi_\theta}}\left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t)\right]$$

$$= LHS$$

(b)

For episodic environments, show that

$$\nabla_\theta V^{\pi_\theta}(\mu) = E_{\tau \sim p_\mu^{\pi_\theta}} \left[\sum_{t}^{T(t)-1} \gamma^t A^{\pi_\theta}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \right]$$

pf:

The major different between episodic and continuing environments is the existence of "terminate state"

Let s_* be the terminate state of an episodic environments

Once the agent reaches state s_*, it will stay at s_* forever (and hence the episode ends)

Moreover, $Q^\pi(s_*, a) = 0$, $V^\pi(s) = 0$, $A^\pi(s, a) = 0$, for all a and all π

Let $T(\tau)$ be the episode length of a trajectory τ,

then, we have

$$\nabla_\theta V^{\pi_\theta}(\mu) = E_{\tau \sim p_\mu^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t A^{\pi_\theta}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \right]$$

$$= E_{\tau \sim p_\mu^{\pi_\theta}} \left[\sum_{t=0}^{T(t)-1} \gamma^t A^{\pi_\theta}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \right]$$