RL Eng Policy Gradient

Friday, December 9, 2022 11:48 PM

Q.1

1. **2 points.** (RL2e 13.2) *Generalize REINFORCE*

Written: Generalize the box on page 199, the policy gradient theorem (13.5), the proof of the policy gradient theorem (page 325), and the steps leading to the REINFORCE update equation (13.8), so that (13.8) ends up with a factor of γ^t and thus aligns with the general algorithm given in the pseudocode.

Eqn from the box of page 199

$$\eta(s) = h(s) + \sum_{\bar{s}} \eta(\bar{s}) \sum_a \pi(a|\bar{s}) p(s|\bar{s}, a), \quad \text{for all } s \in \mathcal{S}.$$

Generalize

$$\eta(s) = h(s) + \gamma \sum_{\bar{s}, a} \pi(a|\bar{s}) p(s|\bar{s}, a) + \gamma^2 \sum_{\bar{s}, a} \pi(a|\bar{s}) p(s|\bar{s}, a) \sum_{x, a'} \pi(a'|x) p(\bar{s}|x, a) \ldots$$

$$\mu(s) = \frac{\eta(s)}{\sum_{s'} \eta(s')}$$

Generalizing the proof of the policy gradient theorm (pg 325)

$$\begin{aligned}
\nabla v_\pi(s) &= \nabla \left[\sum_a \pi(a|s) q_\pi(s,a) \right], \quad \text{for all } s \in \mathcal{S} && \text{(Exercise 3.18)} \\
&= \sum_a \Big[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla q_\pi(s,a) \Big] && \text{(product rule of calculus)} \\
&= \sum_a \Big[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla \sum_{s',r} p(s',r|s,a) \big(r + v_\pi(s')\big) \Big] \\
& && \text{(Exercise 3.19 and Equation 3.2)} \\
&= \sum_a \Big[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s'} p(s'|s,a) \nabla v_\pi(s') \Big] && \text{(Eq. 3.4)} \\
&= \sum_a \Big[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s'} p(s'|s,a) && \text{(unrolling)} \\
&\qquad \sum_{a'} \big[\nabla \pi(a'|s') q_\pi(s',a') + \pi(a'|s') \sum_{s''} p(s''|s',a') \nabla v_\pi(s'') \big] \Big] \\
&= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \Pr(s \to x, k, \pi) \sum_a \nabla \pi(a|x) q_\pi(x,a),
\end{aligned}$$

$$\nabla_\theta v_\pi(s) = \sum_{x \in S} \sum_{k=0}^{\infty} P(s \to x, k, \pi) \gamma^k \sum_a \nabla_\theta \pi(a|x) q_\pi(x, a)$$

When we view in the form of termination

$$\nabla J(\boldsymbol{\theta}) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s, \boldsymbol{\theta}), \tag{13.5}$$

$$\nabla_\theta J(\theta) = \mathbb{E}_\pi \left[\gamma_t \sum_a q_\pi(S_t, a) \nabla_\theta \pi(a|S_t, \theta) \right]$$

Q.2

2. **2 points.** (RL2e 13.3) *Eligibility Vector for Softmax Policy*

Written: In Section 13.1 we considered policy parameterizations using the soft-max in action preferences (13.2) with linear action preferences (13.3). For this parameterization, prove that the eligibility vector is

$$\nabla \ln \pi(a \mid s, \boldsymbol{\theta}) = \mathbf{x}(s,a) - \sum_b \pi(b \mid s, \boldsymbol{\theta}) \mathbf{x}(s,b),$$

with softmax

$$\pi(a|s, \boldsymbol{\theta}) \doteq \frac{e^{h(s,a,\boldsymbol{\theta})}}{\sum_b e^{h(s,b,\boldsymbol{\theta})}}, \tag{13.2}$$

preference simply linear in features

$$h(s,a,\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}(s,a), \tag{13.3}$$

$\theta \leftarrow$ vector of the all the connection weights of the network

Using above 2 eqns

$$\nabla_\theta \log(\pi) = x(s,a) - \frac{\sum_b x(s,b)\, e^{(\theta^T x(s,b))}}{\sum_b e^{(\theta^T x(s,b))}}$$

$\downarrow$ Using eqn (13.2)

$$\nabla_\theta \log(\pi) = x(s,a) - \sum x(s,b)\, \pi(b|s,\theta)$$

Q.3

3. **3 points.** (RL2e 13.4) *Eligibility Vector for Gaussian Policy*

Written: Show that for the gaussian policy parameterization (13.19) the eligibility vector has the following two parts:

$$\nabla \ln \pi(a \mid s, \boldsymbol{\theta}_\mu) = \frac{\nabla \pi(a \mid s, \boldsymbol{\theta}_\mu)}{\pi(a \mid s, \boldsymbol{\theta})} = \frac{1}{\sigma(s,\boldsymbol{\theta})^2} (a - \mu(s,\boldsymbol{\theta})) \mathbf{x}_\mu(s), \text{ and}$$

$$\nabla \ln \pi(a \mid s, \boldsymbol{\theta}_\sigma) = \frac{\nabla \pi(a \mid s, \boldsymbol{\theta}_\sigma)}{\pi(a \mid s, \boldsymbol{\theta})} = \left(\frac{(a - \mu(s,\boldsymbol{\theta}))^2}{\sigma(s,\boldsymbol{\theta})^2} - 1 \right) \mathbf{x}_\sigma(s)$$

$$\nabla \ln \pi(a|s, \boldsymbol{\theta}) = \mathbf{x}(s,a) - \sum_b \pi(b|s, \boldsymbol{\theta}) \mathbf{x}(s,b), \quad \leftarrow \quad 13.9$$

$$\pi(a|s,\theta) = \frac{1}{\sigma(s,\theta)\sqrt{2\pi}} e^{\left(\frac{-(a - \mu(s,\theta))^2}{2\sigma(s,\theta)^2}\right)}$$

$$\mu(s,\theta) = \theta_\mu^T x_\mu(s)$$

$$\sigma(s,\theta_\sigma) = e^{(\theta_\sigma^T x_\sigma(s))}$$

$$\ln \pi(a|s,\theta) = -\ln\sqrt{2\pi} - \ln\sigma - \frac{(a-\mu)^2}{2\sigma(s,\theta)^2}$$

$\Downarrow$

$$\nabla \ln \pi(a|s,\theta) = \frac{a-\mu}{\sigma(s,\theta)^2} \nabla_{\theta_\mu} \mu(s,\theta_\mu)$$

$$= \frac{1}{\sigma(s,\theta)^2} (a - \mu(s,\theta)) x_\mu(s)$$

$$\nabla \ln \pi(a|s,\theta_\sigma) = -\frac{\nabla_{\theta_\sigma} \sigma(s,\theta)}{\sigma(s,\theta)} + \frac{1}{\sigma(s,\theta)^2} (a - \mu(s,\theta))^2 \nabla_{\theta_\sigma}\sigma$$

$$\nabla \ln \pi(a|s,\theta_\sigma) = \left(\frac{(a - \mu(s,\theta))^2}{\sigma(s,\theta)^2} - 1 \right) x_\sigma(s)$$