Chapter 13

Akshay Gupta

June 2, 2020

Exercise 13.1
We know that.

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

Consider i^{th} state as s_i where $i \,\epsilon\, \{1, 2, 3, 4\}$. s_1 is starting state and s_4 is terminating state.
And consider.

$$\pi_\theta(a|s) = \begin{cases} \theta = p_r & a = right \\ (1 - theta) = p_l & a = left \end{cases}$$

Now we can write.

$$V(s_1) = p_l * (-1 + V(s_1)) + p_r * (-1 + V(s_2)) \tag{1}$$
$$V(s_2) = p_l * (-1 + V(s_3)) + p_r * (-1 + V(s_1)) \tag{2}$$
$$V(s_3) = p_l * (-1 + V(s_2)) + p_r * (-1 + V(s_4)) \tag{3}$$

Where $V(s_4) = 0$. So by solving the equation 1,2 and 3 we get.

$$V(s_1) = \frac{-2(1 + p_l)}{p_l(1 - p_l)}$$

So for optimal policy the $V(s_1)$ should be maximum. Therefore

$$\frac{dV(s_1)}{dp_l} = \frac{-2(p_l^2 + 2p_l - 1)}{(p_l(1 - p_l))^2} = 0$$

$$(p_l^2 + 2p_l - 1) = 0 \tag{4}$$

Considering the positive value only.

$$p_l = \sqrt{2} - 1$$

As $p_l + p_r = 1$. We get $p_r = 2 - \sqrt{2} \approx 0.59$.Also the corresponding $V(s_1) \approx -11.6$

Hence the equation 4 is the symbolic representation of the optimal probability and this is verified by calculating $V(s_1)$.

Exercise 13.2

The equation on the page 199 in the book is.

$$\eta(s) = h(s) + \sum_{\bar{s}} \eta(\bar{s}) \sum_{a} \pi(a|\bar{s}) p(s|\bar{s}, a)$$

The generalization for $\gamma \neq 1$ is.

$$\eta(s) = h(s) + \gamma \sum_{\bar{s}} \eta(\bar{s}) \sum_{a} \pi(a|\bar{s}) p(s|\bar{s}, a)$$

Also

$$\mu(s) = \frac{\eta(s)}{\sum_s \eta(s)}$$

Generalization of policy gradient theorem is as follows.

$$
\begin{aligned}
\nabla V_\pi(s) &= \nabla \sum_a \pi(a|s) q_\pi(s,a) \\
&= \sum_a \nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla q_\pi(s,a) \\
&= \sum_a \nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla \sum_{s'} p(s', r|s,a)(r + \gamma V(s')) \\
&= \sum_a \nabla \pi(a|s) q_\pi(s,a) + \gamma \pi(a|s) \sum_{s'} p(s'|s,a) \nabla V(s') \\
&= \sum_a [\nabla \pi(a|s) q_\pi(s,a) + \gamma \pi(a|s) \sum_{s'} p(s'|s,a) \sum_{a'} [\nabla \pi(a'|s') q_\pi(s',a') \\
&+ \gamma \pi(a|s) \sum_{s''} p(s''|s',a') \nabla V_\pi(s'')]] \\
&= \sum_{x \epsilon \mathcal{S}} \sum_{k=0}^{\infty} Pr(s \to x, k, \pi) \gamma^k \sum_a \nabla \pi(a,x) q_\pi(x,a)
\end{aligned}
$$

Now

$$
\begin{aligned}
\nabla J(\theta) &= \nabla V_\pi(s_0) \\
&= \sum_s \sum_{k=0}^{\infty} Pr(s_0 \to s, k, \pi) \gamma^k \sum_a \nabla \pi(a,s) q_\pi(s,a) \\
&= \sum_s \eta(s) \sum_a \nabla \pi(a,s) q_\pi(s,a) \qquad (\because \mathit{ergodicity\ assumption}) \\
&= \sum_{s'} \eta(s') \sum_s \frac{\eta(s)}{\sum_{s'} \eta(s')} \sum_a \nabla \pi(a,s) q_\pi(s,a) \\
&= \sum_{s'} \eta(s') \sum_s \mu(s) \sum_a \nabla \pi(a,s) q_\pi(s,a)
\end{aligned}
$$

$$\begin{aligned}
\nabla J(\theta) &\propto \sum_s \mu(s) \sum_a \nabla \pi(a,s) q_\pi(s,a) \\
&= \mathbb{E}_\pi\{\gamma^t \sum_a \nabla \pi(a,S_t) q_\pi(S_t,a)\} \\
&= \mathbb{E}_\pi\{\gamma^t \sum_a \pi(a,S_t) q_\pi(S_t,a) \frac{\nabla \pi(a,S_t)}{\pi(a,S_t)}\} \\
&= \mathbb{E}_\pi\{\gamma^t q_\pi(S_t,A_t) \nabla ln(\pi(A_t,S_t))\} \\
&= \mathbb{E}_\pi\{\gamma^t G_t \nabla ln(\pi(A_t,S_t))\}
\end{aligned}$$

So now we can write the generalized update equation for REINFORCE policy Gradient.

$$\theta_{t+1} = \theta_t + \alpha \gamma^t G_t \nabla ln(\pi(A_t,S_t))$$

Exercise 13.3

Let $\pi(a|s,\theta) = \dfrac{\exp(\theta^T x(s,a))}{\sum_{a'} exp(\theta^T x(s,a'))}$

Now

$$\begin{aligned}
\nabla ln(\pi(a|s,\theta)) &= \nabla\{\theta^T x(s,a) - ln(\sum_{a'} exp(\theta^T x(s,a')))\} \\
&= x(s,a) - \nabla ln(\sum_{a'} exp(\theta^T x(s,a'))) \\
&= x(s,a) - \frac{\nabla \sum_{a'} exp(\theta^T x(s,a'))}{\sum_{a'} exp(\theta^T x(s,a'))} \\
&= x(s,a) - \frac{\sum_{a'} \nabla exp(\theta^T x(s,a'))}{\sum_{a'} exp(\theta^T x(s,a'))}
\end{aligned}$$

$$
\begin{aligned}
&= x(s,a) - \frac{\sum_{a'} exp(\theta^T x(s,a'))x(s,a')}{\sum_{a'} exp(\theta^T x(s,a'))} \\
&= x(s,a) - \sum_{a'} \frac{exp(\theta^T x(s,a'))}{\sum_{a'} exp(\theta^T x(s,a'))} x(s,a') \\
&= x(s,a) - \sum_{a'} \pi(a'|s,\theta) x(s,a')
\end{aligned}
$$

$Q.E.D$

Exercise 13.4

$$
\begin{aligned}
\theta &= (\theta_\mu, \theta_\sigma) \\
\pi(a|s,\theta) &= \frac{1}{\sigma(s,\theta)\sqrt{2\pi}} \exp\Big(-\frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2}\Big) \\
\mu(s,\theta) &= \theta_\mu^T x_\mu(s) \;\; \sigma(s,\theta) = \exp\big(\theta_\sigma^T x_\sigma(s)\big) \\
ln(\pi(a|s,\theta)) &= -\frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2} - ln(\sigma(s,\theta)) - ln(\sqrt{2\pi}) \\
\nabla ln(\pi(a|s,\theta_\mu)) &= \frac{2(a-\mu(s,\theta))}{2\sigma(s,\theta)^2} \nabla\mu(s,\theta) \\
&= \frac{(a-\mu(s,\theta))}{\sigma(s,\theta)^2} x_\mu(s)
\end{aligned}
$$

$$
\begin{aligned}
\nabla ln(\pi(a|s,\theta_\sigma)) &= \frac{2(a-\mu(s,\theta))}{2\sigma(s,\theta)^3} \nabla\sigma(s,\theta) - \frac{\nabla\sigma(s,\theta)}{\sigma(s,\theta)} \\
&= \frac{(a-\mu(s,\theta))}{\sigma(s,\theta)^3} \sigma(s,\theta) x_\sigma(s) - \frac{\sigma(s,\theta) x_\sigma(s)}{\sigma(s,\theta)} \qquad (\because \nabla\sigma(s,\theta) = \sigma(s,\theta) x_\sigma(s)) \\
&= \left(\frac{(a-\mu(s,\theta))}{\sigma(s,\theta)^2} - 1\right) x_\sigma(s)
\end{aligned}
$$

$Q.E.D$

- **Exercise 13.5**

$$
\begin{aligned}
Pr\{A_t = 1\} &= p_t \\
Pr\{A_t = 0\} &= 1 - p_t \\
h(s, 1, \theta) - h(s, 0, \theta) &= \theta^T x(s)
\end{aligned}
$$

(a)

Now

$$
\begin{aligned}
p_t = \pi(1|st, \theta) &= \frac{e^{h(s,1,\theta)}}{e^{h(s,1,\theta)} + e^{h(s,0,\theta)}} \\
&= \frac{1}{1 + e^{h(s,0,\theta) - h(s,1,\theta)}} \\
&= \frac{1}{1 + e^{-(h(s,1,\theta) - h(s,0,\theta))}} \\
&= \frac{1}{1 + e^{-\theta^T x(s)}}
\end{aligned}
$$

(b)

$$\theta_{t+1} = \theta_t + \alpha G_t \nabla ln(\pi(a|s, \theta))$$

(c)

We know that $\pi(a|s, \theta)$ is a sigmoid function and derivative of sigmoid function is $\pi(a|s, \theta)(1 - \pi(a|s, \theta))$

Now, when a=1

$$\pi(1|s,\theta) = p_t = \frac{1}{1+e^{-\theta^T x(s)}}$$
$$\nabla\pi(1|s,\theta) = p_t(1-p_t)x(s)$$

Now, when a=0

$$\pi(0|s,\theta) = 1 - p_t = 1 - \frac{1}{1+e^{-\theta^T x(s)}}$$
$$\nabla\pi(0|s,\theta) = -p_t(1-p_t)x(s)$$

$$\begin{aligned}
\nabla ln(\pi(a|s,\theta)) &= \frac{\nabla\pi(a|s,\theta)}{\pi(a|s,\theta)} \\
\nabla ln(\pi(1|s,\theta)) &= \frac{p_t(1-p_t)x(s)}{p_t} \\
&= (1-p_t)x(s) \\
\nabla ln(\pi(0|s,\theta)) &= \frac{-p_t(1-p_t)x(s)}{1-p_t} \\
&= -p_t x(s) \\
\nabla ln(\pi(a|s,\theta) &= a(1-\pi(1|s,\theta))x(s) + (1-a) - \pi(1|s,\theta)x(s) \\
&= (a-\pi(1|s,\theta))x(s)
\end{aligned}$$