CS 285 hw2

oliver song

October 2023

1

1.1

Using policy gradient, we get:

$$\begin{aligned}
\nabla_\theta E_{\pi_\theta} R(\tau) &= E_{\tau\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(\tau) R(\tau)] \\
&= E_{\tau\sim\pi_\theta}[(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t))(\sum_{t=1}^{T} r(s_t, a_t)] \\
&= \sum_{T=1}^{\infty} \theta^{T-1} T^2 \\
&= \frac{1+\theta}{(1-\theta)^3}
\end{aligned} \tag{1}$$

1.2

Directly calculate the expected return using expectation over the return, we get:

$$\begin{aligned}
\nabla_\theta E_{\tau\sim\pi_\theta} R(\tau) &= \nabla_\theta \sum_{n=1}^{\infty} n\theta^n \\
&= \sum_{n=1}^{\infty} n^2 \theta^{n-1} \\
&= \frac{1+\theta}{(1-\theta)^3}
\end{aligned} \tag{2}$$

This result matches that computed in policy gradient.

2

The variance of policy gradient is:

$$\begin{aligned}
Var &= E_{\tau\sim p_\theta(\tau)}[(\nabla_\theta \log p_{\pi_\theta}(\tau) r(\tau))^2] - E_{\tau\sim p_\theta(\tau)}[\nabla_\theta \log p_{\pi_\theta}(\tau) r(\tau)]^2 \\
&= E_{\tau\sim p_\theta(\tau)}[(T^2)^2] - [\frac{1+\theta}{(1-\theta)^3}]^2 \\
&= \sum_{T=1}^{\infty} \theta^{T-1} T^4 - [\frac{1+\theta}{(1-\theta)^3}]^2 \\
&= \frac{1+11\theta+11\theta^2+\theta^3}{(1-\theta)^5} - [\frac{1+\theta}{(1-\theta)^3}]^2 \\
&= \frac{2+12\theta+\theta^2-10\theta^3-\theta^4}{(1-\theta)^6}
\end{aligned} \tag{3}$$

Minimum $Var = 2$ is achieved when $\theta = 0$. Doesn't have maximum since $Var \to \infty$ as $\theta \to 1$

3

3.1

$$\nabla_\theta J(\theta) = E_{\tau\sim\pi_\theta}[\sum_{t=1}^{T} \log \pi_\theta(a_t|s_t)(\sum_{t'=t}^{T} r(s_{t'}, a_{t'}))] \tag{4}$$

Since we know that a baseline policy is unbiased, subtracting the sum of reward from previous steps would also be unbiased.

3.2

$$\begin{aligned}
Var &= E_{\tau\sim\pi_\theta}[(\sum_{t=1}^{T} \log \pi_\theta(a_t|s_t)(\sum_{t'=t}^{T} r(s_{t'}, a_{t'})))^2] - E_{\tau\sim\pi_\theta}[\sum_{t=1}^{T} \log \pi_\theta(a_t|s_t)(\sum_{t'=t}^{T} r(s_{t'}, a_{t'}))]^2 \\
&= E_{\tau\sim\pi_\theta}[\sum_{t=1}^{T}(T-t)^2] - E_{\tau\sim\pi_\theta}[\sum_{t=1}^{T}(T-t)]^2 \\
&= \frac{\theta(1+\theta)}{(1-\theta)^4} - [\frac{\theta}{(1-\theta)^3}]^2 \\
&= \frac{\theta(1-2\theta-\theta^2+\theta^3)}{(1-\theta)^6}
\end{aligned} \tag{5}$$

4

4.1

$$
\begin{aligned}
\nabla_{\theta'} J(\theta') &= E_{\tau \sim \pi_{\theta'}(\tau)}[\sum_{t=1}^{T} \frac{\pi_{\theta'}(s_t, a_t)}{\pi_{\theta}(s_t, a_t)} \nabla_{\theta'} \log \pi_{\theta'}(a_t|s_t) \hat{Q}_t] \\
&= P_{\tau = \tau_H}[\frac{\theta'}{\theta}] \\
&= \frac{\theta'^H}{\theta}
\end{aligned}
\tag{6}
$$

4.2

$$
\begin{aligned}
Var &= \frac{\theta'^{H+1}}{\theta^2} - \frac{\theta'^{2H}}{\theta^2} \\
&= \theta'^{H+1} \frac{1 - \theta'^{H-1}}{\theta^2}
\end{aligned}
\tag{7}
$$

We can see from above that $Var \to 0$ as $H \to \infty$