

CS285: Deep Reinforcement Learning

Assignment 2

Written Report

Alan Sorani

April 22, 2025

1 Analysis

1. (a) Using policy gradients, we have

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\left(\sum_{t=1}^{\infty} \nabla_\theta \log \pi_\theta(a_{i,t} \mid s_{i,t}) \right) R(\tau) \right]. \quad (1)$$

Due to the simplicity of our MDP, we can easily enumerate all the possible trajectories τ based on the first time-step t for which $s_t = S_F$. Let τ_i be the trajectory for which $s_{i,j} = s_1$ for all $j < i$ and for which $s_{i,i} = s_F$ (and therefore $s_{i,j} = S_F$ for all $j > i$). The probability of the trajectory τ_i under the policy π_θ is then $\theta^{i-1}(1-\theta)$. We see that $R(\tau_i) = \sum_{j=1}^i r(s_{i,j} \mid a_{i,j}) = \sum_{j=1}^{i-1} 1 = i-1$.

Writing the expectation of (1) explicitly, we get

$$\begin{aligned} \nabla J(\theta) &= \sum_{i=1}^{\infty} p_\theta(\tau_i) \left[\left(\left(\sum_{t=1}^{i-1} \nabla_\theta \log(\theta) \right) + \nabla_\theta \log(1-\theta) \right) (i-1) \right] \\ &= \sum_{i=1}^{\infty} \theta^{i-1} (1-\theta) \left[\left(\frac{i-1}{\theta} - \frac{1}{1-\theta} \right) (i-1) \right] \\ &= \sum_{i=1}^{\infty} \theta^{i-1} \left[\left(\frac{(i-1)(1-\theta)}{\theta} - 1 \right) (i-1) \right] \\ &= \sum_{i=1}^{\infty} \theta^{i-1} \left[\left(\frac{(i-1)(1-\theta) - \theta}{\theta} \right) (i-1) \right] \\ &= \sum_{i=1}^{\infty} \theta^{i-2} [(i-\theta i - 1 + \theta - \theta)(i-1)] \\ &= \sum_{i=1}^{\infty} \theta^{i-2} (i - \theta i - 1)(i-1) \\ &= (1-\theta) \sum_{i=1}^{\infty} i(i-1)\theta^{i-2} - \sum_{i=1}^{\infty} (i-1)\theta^{i-2}. \end{aligned}$$

Finally, we have

$$\begin{aligned}
(1-\theta) \sum_{i=1}^{\infty} i(i-1) \theta^{i-2} &= (1-\theta) \sum_{i=1}^{\infty} \frac{d^2}{d^2\theta} \theta^i \\
&= (1-\theta) \frac{d^2}{d^2\theta} \sum_{i=1}^{\infty} \theta^i \\
&= (1-\theta) \frac{d^2}{d^2\theta} \frac{\theta}{1-\theta} \\
&= (1-\theta) \frac{d}{d\theta} \frac{1}{(1-\theta)^2} \\
&= (1-\theta) \cdot \left(-\frac{2}{(1-\theta)^3} \right) \\
&= \frac{2}{(1-\theta)^2}
\end{aligned}$$

and

$$\begin{aligned}
\sum_{i=1}^{\infty} (i-1) \theta^{i-2} &= \sum_{i=1}^{\infty} \frac{d}{d\theta} \theta^{i-1} \\
&= \frac{d}{d\theta} \sum_{i=1}^{\infty} \theta^{i-1} \\
&= \frac{d}{d\theta} \frac{1}{1-\theta} \\
&= \frac{1}{(1-\theta)^2},
\end{aligned}$$

so we get

$$\nabla J(\theta) = \frac{2}{(1-\theta)^2} - \frac{1}{(1-\theta)^2} = \frac{1}{(1-\theta)^2}.$$

(b) We shall now compute $\mathbb{E}_{\tau \sim \pi_\theta} [R(\tau)]$ directly and verify that

$$\nabla \mathbb{E}_{\tau \sim \pi_\theta} [R(\tau)] = \frac{1}{1-\theta}.$$

We have

$$\begin{aligned}
\mathbb{E}_{\tau \sim \pi_\theta} [R(\tau)] &= \sum_{i=1}^{\infty} p_\theta(\tau_i) R(\tau_i) \\
&= \sum_{i=1}^{\infty} \theta^{i-1} (1-\theta) (i-1) \\
&= (1-\theta) \sum_{i=1}^{\infty} \theta^{i-1} (i-1) \\
&= (1-\theta) \sum_{i=1}^{\infty} \frac{d}{d\theta} \theta^i \\
&= (1-\theta) \frac{d}{d\theta} \sum_{i=1}^{\infty} \theta^i \\
&= (1-\theta) \frac{d}{d\theta} \frac{\theta}{1-\theta} \\
&= (1-\theta) \cdot \frac{1}{(1-\theta)^2} \\
&= \frac{1}{1-\theta}
\end{aligned}$$

and then indeed

$$\nabla \mathbb{E}_{\tau \sim \pi_\theta} [R(\tau)] = \nabla \frac{1}{1-\theta} = \frac{1}{(1-\theta)^2}.$$

2. We have

$$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2,$$

so

$$\begin{aligned} \text{Var}_{\tau \sim p_\theta(\tau)} [\nabla_\theta \log p_\theta(\tau) r(\tau)] &= \mathbb{E}_{\tau \sim p_\theta(\tau)} [(\nabla_\theta \log p_\theta(\tau) r(\tau))^2] - \mathbb{E}_{\tau \sim p_\theta(\tau)} [\nabla_\theta \log p_\theta(\tau) r(\tau)]^2 \\ &= \mathbb{E}_{\tau \sim p_\theta(\tau)} [(\nabla_\theta \log p_\theta(\tau) r(\tau))^2] - \nabla J(\theta)^2 \\ &= \mathbb{E}_{\tau \sim p_\theta(\tau)} [(\nabla_\theta \log p_\theta(\tau) r(\tau))^2] - \frac{1}{(1-\theta)^4}. \end{aligned}$$

Now, using our computations from the previous part, we get

$$\begin{aligned} \mathbb{E}_{\tau \sim p_\theta(\tau)} [(\nabla_\theta \log p_\theta(\tau) r(\tau))^2] &= \sum_{i=1}^{\infty} p_\theta(\tau_i) \left[\left(\left(\sum_{t=1}^{i-1} \nabla_\theta \log(\theta) \right) + \nabla_\theta \log(1-\theta) \right) (i-1) \right]^2 \\ &= \sum_{i=1}^{\infty} \theta^{i-1} (1-\theta) \left[\left(\frac{i-1}{\theta} - \frac{1}{1-\theta} \right) (i-1) \right]^2 \\ &= \sum_{i=1}^{\infty} \frac{\theta^{i-3}}{1-\theta} (i(1-\theta) - 1)^2 (i-1)^2 \\ &= \frac{1}{1-\theta} \sum_{i=1}^{\infty} \theta^{i-3} (i(1-\theta) - 1)^2 (i-1)^2 \\ &= \frac{1}{1-\theta} \sum_{i=1}^{\infty} \theta^{i-3} (i^2(1-\theta)^2 - 2i(1-\theta) + 1) (i^2 - 2i + 1) \\ &= \frac{1}{\theta^3(1-\theta)} \sum_{i=1}^{\infty} \theta^i (i^4\theta^2 - 2i^4\theta + i^4 - 2i^3\theta^2 + 6i^3\theta - 4i^3 + i^2\theta^2 - 6i^2\theta + 6i^2 + 2i\theta - 4i + 1). \end{aligned}$$

We get that

$$\mathbb{E}_{\tau \sim p_\theta(\tau)} [(\nabla_\theta \log p_\theta(\tau) r(\tau))^2] = \frac{1}{\theta^3(1-\theta)} (S_1 - S_2 + S_3 - S_4 + S_5 - S_6 + S_7 - S_8 + S_9 + S_{10} - S_{11} + S_{12})$$

where

$$\begin{aligned} S_1 &= \sum_{i=1}^{\infty} i^4 \theta^{i+2}, & S_7 &= \sum_{i=1}^{\infty} i^2 \theta^{i+2} \\ S_2 &= 2 \sum_{i=1}^{\infty} i^4 \theta^{i+1}, & S_8 &= 6 \sum_{i=1}^{\infty} i^2 \theta^{i+1} \\ S_3 &= \sum_{i=1}^{\infty} i^4 \theta^i, & S_9 &= 6 \sum_{i=1}^{\infty} i^2 \theta^i \\ S_4 &= 2 \sum_{i=1}^{\infty} i^3 \theta^{i+2}, & S_{10} &= 2 \sum_{i=1}^{\infty} i \theta^{i+1} \\ S_5 &= 6 \sum_{i=1}^{\infty} i^3 \theta^{i+1}, & S_{11} &= 4 \sum_{i=1}^{\infty} i \theta^i \\ S_6 &= 4 \sum_{i=1}^{\infty} i^3 \theta^i, & S_{12} &= \sum_{i=1}^{\infty} \theta^i. \end{aligned}$$

Now,

$$\sum_{i=1}^{\infty} \theta^i = \frac{\theta}{1-\theta}$$

so by differentiating

$$\sum_{i=1}^{\infty} i\theta^{i-1} = \frac{1}{(1-\theta)^2}$$

and by multiplying by θ

$$\sum_{i=1}^{\infty} i\theta^i = \frac{\theta}{(1-\theta)^2}.$$

By differentiating this and then multiplying by θ we get

$$\sum_{i=1}^{\infty} i^2\theta^i = \frac{\theta^2 + \theta}{(1-\theta)^3}.$$

By repeating this we get

$$\sum_{i=1}^{\infty} i^3\theta^i = \frac{\theta^3 + 4\theta^2 + \theta}{(1-\theta)^4}$$

and finally

$$\sum_{i=1}^{\infty} i^4\theta^i = \frac{\theta^4 + 11\theta^3 + 11\theta^2 + \theta}{(1-\theta)^5}.$$

Using these equations we get

$$\begin{aligned} S_1 &= \theta^2 \cdot \frac{\theta^4 + 11\theta^3 + 11\theta^2 + \theta}{(1-\theta)^5}, & S_7 &= \theta^2 \cdot \frac{\theta^2 + \theta}{(1-\theta)^3} \\ S_2 &= 2\theta \cdot \frac{\theta^4 + 11\theta^3 + 11\theta^2 + \theta}{(1-\theta)^5}, & S_8 &= 6\theta \cdot \frac{\theta^2 + \theta}{(1-\theta)^3} \\ S_3 &= \frac{\theta^4 + 11\theta^3 + 11\theta^2 + \theta}{(1-\theta)^5}, & S_9 &= 6 \cdot \frac{\theta^2 + \theta}{(1-\theta)^3} \\ S_4 &= 2\theta^2 \cdot \frac{\theta^3 + 4\theta^2 + \theta}{(1-\theta)^4}, & S_{10} &= 2\theta \cdot \frac{\theta}{(1-\theta)^2} \\ S_5 &= 6\theta \cdot \frac{\theta^3 + 4\theta^2 + \theta}{(1-\theta)^4}, & S_{11} &= 4 \cdot \frac{\theta}{(1-\theta)^2} \\ S_6 &= 4 \cdot \frac{\theta^3 + 4\theta^2 + \theta}{(1-\theta)^4}, & S_{12} &= \frac{\theta}{1-\theta}. \end{aligned}$$

Going back to the calculation of $\mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[(\nabla_{\theta} \log p_{\theta}(\tau) r(\tau))^2 \right]$ we get

$$\begin{aligned} \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[(\nabla_{\theta} \log p_{\theta}(\tau) r(\tau))^2 \right] &= \frac{1}{\theta^3 (1-\theta)} (S_1 - S_2 + S_3 - S_4 + S_5 - S_6 + S_7 - S_8 + S_9 + S_{10} - S_{11} + S_{12}) \\ &= \frac{1}{\theta^3 (1-\theta)} \cdot \frac{\theta^2 (4\theta^2 + 9\theta + 1)}{(1-\theta)^3} \\ &= \frac{4\theta^2 + 9\theta + 1}{\theta (1-\theta)^4} \end{aligned}$$

and so

$$\begin{aligned}\text{Var}_{\tau \sim p_{\theta}(\tau)} [\nabla_{\theta} \log p_{\theta}(\tau) r(\tau)] &= \frac{4\theta^2 + 9\theta + 1}{\theta(1-\theta)^4} - \frac{1}{(1-\theta)^4} \\ &= \frac{4\theta^2 + 8\theta + 1}{\theta(1-\theta)^4}.\end{aligned}$$

To find the values of θ which minimize or maximize the variance, we compare

$$\frac{d}{d\theta} \text{Var}_{\tau \sim p_{\theta}(\tau)} [\nabla_{\theta} \log p_{\theta}(\tau) r(\tau)] = \frac{12\theta^3 + 36\theta^2 + 5\theta - 1}{\theta^2(1-\theta)^5}$$

to 0. We get that the variance is minimal for $\theta \approx 0.10988$ and goes to infinity as θ approaches 0 or 1.

3. (a)
- (b)