Reinforcement Learning WS 2024 Assignment 8

Carla López Martínez (6637484)
Gaurav Niranjan (6599177)
Apoorv Agnihotri (6604679)

1 Exercise 1

The policy gradient with importance weighting, used for instance in PPO, is given by:

$$\nabla_\theta J(\theta) = E_{\tau \sim \pi_{\theta_{\text{old}}}} \left[\frac{\nabla_\theta \pi_\theta(s_t, a_t)}{\pi_{\theta_{\text{old}}}(s_t, a_t)} A(s_t, a_t) \right] \tag{1}$$

However, so far we have studied policy gradient formulations containing the score function:

$$\nabla_\theta J(\theta) = E_{\pi_\theta} \left[\nabla_\theta \log \pi_\theta(s, a) \cdot A(s, a) \right] \tag{2}$$

Why is there no logarithm in Eq. (1)? Show that Eq. (1) is correct under the assumption that the difference in the state visitation distribution $\mu(s)$ between the old and new policies can be ignored. Consider the expectation that needs to be computed, referencing slide 6 ("Recap: Policy Gradient") in lecture notes 8, but with respect to the old parameters.

Solution

In Equation (1), we directly compute the gradient of the policy $\nabla_\theta \pi_\theta(s, a)$, scaled by the importance weight $\pi_\theta / \pi_{\theta_{\text{old}}}$. There is no logarithm because the derivation directly uses the re-weighting to handle the expectation shift instead of using the score function.

We know that:

$$E_{\pi_\theta}[\cdot] = \sum_s \mu_\pi(s) \sum_a \pi_\theta(a|s)[\cdot]$$

If we assume that the state visitation distributions under π_θ and $\pi_{\theta_{\text{old}}}$ are similar ($\mu_\pi(s) \approx \mu_{\pi_{\text{old}}}(s)$), we can approximate the expectations:

$$E_{\pi_\theta}[\cdot] \approx E_{\pi_{\theta_{\text{old}}}}[\cdot]$$

This approximation justifies using samples from the old policy without needing to account for differences in $\mu(s)$, so instead of sampling from π_θ, we sample from $\pi_{\theta_{\text{old}}}$ and apply importance sampling:

$$E_{\pi_\theta}[\cdot] = E_{\pi_{\theta_{\text{old}}}}\left[\frac{\pi_\theta(s,a)}{\pi_{\theta_{\text{old}}}(s,a)}[\cdot]\right]$$

We know that:

$$\nabla_\theta \log \pi_\theta(s,a) = \frac{\nabla_\theta \pi_\theta(s,a)}{\pi_\theta(s,a)}$$

Substituting this into Eq.2 and applying the previous rewriting:

$$\nabla_\theta J(\theta) = E_{\pi_{\theta_{\text{old}}}}\left[\frac{\pi_\theta(s,a)}{\pi_{\theta_{\text{old}}}(s,a)}\frac{\nabla_\theta \pi_\theta(s,a)}{\pi_\theta(s,a)} \cdot A(s,a)\right]$$

Simplifying:

$$\nabla_\theta J(\theta) = E_{\pi_{\theta_{\text{old}}}}\left[\frac{\nabla_\theta \pi_\theta(s,a)}{\pi_{\theta_{\text{old}}}(s,a)} \cdot A(s,a)\right]$$

This matches Equation (1).