Exercise Set 5 - Reinforcement Learning

Advanced policy-based methods

Giulio Starace - 13010840

October 11, 2022

9.4 Homework: Limits of policy gradients

1. Given our policy:

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

We can compute $\nabla \log \pi(a|s,\theta)$ w.r.t. θ_μ and θ_σ by applying the chain rule. Let $\log \pi(a|s,\theta)$ be $L(\theta)$, then w.r.t. a given param θ_i we have:

$$\begin{aligned} \nabla_{\theta_i} \log \pi(a|s,\theta) &= \nabla_{\theta_i} L(\theta) \\ &= \frac{\partial L(\theta)}{\partial \pi} \cdot \frac{\partial \pi(a|s,\theta)}{\partial \theta_i}. \end{aligned}$$

When $\theta_i = \theta_\mu$, we have:

$$\begin{aligned} \nabla_{\theta_\mu} \log \pi(a|s,\theta) &= \frac{\partial L(\theta)}{\partial \pi,} \cdot \frac{\partial \pi(a|s,\theta)}{\partial \theta_\mu} \\ &= \frac{\partial L(\theta)}{\partial \pi(a|s,\theta)} \cdot \frac{\partial \pi(a|s,\theta)}{\partial \mu} \cdot \frac{\partial \mu(\theta_\mu)}{\partial \theta_\mu}. \end{aligned} \tag{2}$$

When $\theta_i = \theta_\sigma$, we have:

$$\begin{aligned} \nabla_{\theta_\sigma} \log \pi(a|s,\theta) &= \frac{\partial L(\theta)}{\partial \pi,} \cdot \frac{\partial \pi(a|s,\theta)}{\partial \theta_\sigma} \\ &= \frac{\partial L(\theta)}{\partial \pi(a|s,\theta)} \cdot \frac{\partial \pi(a|s,\theta)}{\partial \sigma} \cdot \frac{\partial \sigma(\theta_\sigma)}{\partial \theta_\sigma}. \end{aligned} \tag{3}$$

The first two terms of equations (2) and (3) will be the same regardless of parametrization. We get

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{1}{\pi}, \tag{4}$$

$$\frac{\partial \pi(a|s,\theta)}{\partial \mu} = \frac{a-\mu(\theta_\mu)}{\sqrt{2\pi}\sigma(\theta_\sigma)^3} \exp\left[-\frac{(a-\mu(\theta_\mu))^2}{2\sigma(\theta_\sigma)^2}\right], \tag{5}$$

$$\frac{\partial \pi(a|s,\theta)}{\partial \sigma} = \frac{(a-\mu(\theta_\mu))^2 - \sigma(\theta_\sigma)^2}{\sqrt{2\pi}\sigma(\theta_\sigma)^4} \exp\left[-\frac{(a-\mu(\theta_\mu))^2}{2\sigma(\theta_\sigma)^2}\right]. \tag{6}$$

We are then left with determining the final terms of equations (2) and (3) for different parametrizations.

(a) When $\mu(\theta_\mu) = \theta_\mu$ and $\sigma(\theta_\sigma) = \exp(\theta_\sigma)$, we get

$$\frac{\partial \mu(\theta_\mu)}{\partial \theta_\mu} = 1 \tag{7}$$

$$\frac{\partial \sigma(\theta_\sigma)}{\partial \theta_\sigma} = \exp(\theta_\sigma). \tag{8}$$

We can plug this into equations (2) and (3) along with the updated terms from previously and get:

(b) hello world

PS: I can't help but comment that this answer was extremely tedious to compute and write, and amounted mostly to busy work, particularly factoring in the LaTeX typesetting. Please consider this feedback for future assignments.

2. hello world

3. hello world

9.5 Homework: Coding Assignment - Policy Gradients

1. Two advantages of using policy based methods over value based methods are:

   (a) Policy based methods can more easily be applied to problems with large and/or continuous state spaces. Unlike value based methods, policy based methods do not need to compute the value of each state, and furthermore do not need to find a way to compute a maximum over all possible state values, which can be prohibitively expensive. Policy based methods circumvent this issue by directly adjusting the parameters of the policy function, directly estimating the action probability distribution for a given input state.

   (b) Policy based methods provide a natural way of learning stochastic policies. In value based methods, stochasticity is typically manually governed by the ϵ hyperparameter used in *epsilon*-greedy policies. In policy based methods, the optimal stochasticity with arbitrary action probabilities can be learned.

2. Coding answers have been submitted on codegra under the group "stalwart cocky sawly".

10.3 Homework: Update Directions

1. hello world

2. hello world

3. hello world