ECE 567 – Reinforcement Learning Theory

Homework 7

Due: 11:59 PM on Mar. 20

1. Clipped PPO [30pt]

In this problem, we will compute the contribution of a single sample to the clipped PPO loss explained in lecture.

Consider a Markov decision process with states $\mathcal{X} = \{A, B\}$, actions $\mathcal{U} = \{\mathsf{stay}, \mathsf{go}\}$. The transition dynamics and reward are deterministic:

$$(A, \mathsf{go}) \to B \text{ with reward } r = 1, \qquad \text{all other } (s, a) \text{ self-loop with reward } 0.$$

We are interested in solving the following infinite-horizon discounted reinforcement learning problem

$$\max_{\mu} E\left[\sum_{k=0}^{\infty} 0.9^k r(x_k, \mu(x_k)) \middle| x_0\right],$$

where x_k is the state at time k, u_k is the action at time k, μ denotes a policy, and the discount factor is 0.9.

Consider the following softmax policy

$$\pi_w(u|x) = \frac{e^{w\phi(x,u)}}{e^{w\phi(x,\mathsf{stay})} + e^{w\phi(x,\mathsf{go})}},$$

where

$$\phi(x, u) = \mathbf{1}_{\{u=\mathsf{go}\}}.$$

Suppose the current Actor and previous Actor have $\tilde{w} = 1$ and $w = 0$ respectively. For the advantage calculation, assume the critic

$$V(x) = 0$$

along with one-step advantage. Given the following trace $(x_k, u_k, r(x_k, u_k))$:

$$(A, \mathsf{go}, 1) \to (B, ?, ?)$$

Please use the approximation $e = 3$ in this problem.

1.1 Compute the importance sampling ratio $\frac{\pi_{\tilde{w}}(u|x)}{\pi_w(u|x)}$.

1.2 Compute the advantage $A_w(x, u)$.

1.3 Compute the clipped PPO objective for the given trace.

2. Bipedal Walker Problem - Twin Delayed DDPG (TD3) [70pts]

Please download the files `hw_7.ipynb` from Canvas. Train your model using Google Colab following the instructions in the file. **After completing the training, save your actor network as** `actor.pth` (save the model's "state_dict"). **Then upload** `actor.pth` **to Gradescope**. We will test your policy (actor) for 50 episodes on Gradescope.