

HW6

Note The goal of this homework is to gain exposure to reinforcement learning in the concrete context of language modeling.

Problem 1 Reinforcement learning from human feedback. Let x be the question or instruction (state) and y be the answer or completion (action). Assume $x \sim p(x)$, where $p(x)$ is the data distribution of questions. Let $\pi_{\text{teacher}}(y|x)$ be the teacher's policy. The teacher may be human expert or a more capable language model. Let $\pi_{\theta}(y|x)$ be the language model to be fine tuned. We may consider it student model.

(1) Suppose we want $\pi_{\theta}(y|x)$ to learn from $\pi_{\text{teacher}}(y|x)$. Our objective function is

$$J_{\text{MLE}}(\theta) = \mathbb{E}_{p(x)} \mathbb{E}_{\pi_{\text{teacher}}(y|x)} [\log \pi_{\theta}(y|x)].$$

Calculate the gradient $J'_{\text{MLE}}(\theta)$. Describe the stochastic gradient algorithm for maximizing J , where we replace expectations by sampling from $p(x)$ and $\pi(y|x)$, thus “stochastic”, where expectations are approximated by averaging over time. This is called imitation learning or behavior cloning.

(2) Suppose we want $\pi_{\theta}(y|x)$ to learn by itself based on a given reward model $r(x, y)$. Our objective function is

$$J_{\text{RL}}(\theta) = \mathbb{E}_{p(x)} \mathbb{E}_{\pi_{\theta}(y|x)} [r(x, y)].$$

Prove

$$J'_{\text{RL}}(\theta) = \mathbb{E}_{p(x)} \mathbb{E}_{\pi_{\theta}(y|x)} \left[r(x, y) \frac{\partial}{\partial \theta} \log \pi_{\theta}(y|x) \right].$$

Describe the stochastic gradient algorithm for maximizing J .

Explain that J' remains the same if we change $r(x, y)$ to $r(x, y) - b(x)$ for a baseline $b(x)$ that only depends on x . If $b(x) = V(x) = \mathbb{E}_{\pi(y|x)} [r(x, y)]$ for a policy π , then $V(x)$ is called the value function under π , and $A(x, y) = r(x, y) - V(x)$ is called advantage of action y at state x .

Also explain that

$$\mathbb{E}_{\pi_{\theta}(y|x)} \left[\frac{\partial}{\partial \theta} \log \pi_{\theta}(y|x) \right] = 0$$

by setting $r(x, y) = 1$ in J_{RL} and J'_{RL} above.

(3) Explain the similarity and difference between the two stochastic gradient algorithms in (1) and (2).

Note: reinforcement learning is to optimize the policy π assuming the reward model $r(x, y)$ is given. We can also learn $r(x, y)$ by observing the behavior or preference of a policy π , and this is called inverse reinforcement learning.

Problem 2 Play with the PyTorch code provided by the following webpage:

https://github.com/HumanSignal/RLHF/blob/master/tutorials/RLHF_with_Custom_Datasets.ipynb

Write a brief explanation of the code and show your results.