Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Summary of Lecture 16: Policy-Based Theory

Summarized By: Amirhossein Asadi

- In the continuation of our discussion, we aim to optimize the policy improvement process. Specifically, we define an objective function subject to a constraint that ensures the new policy, denoted as $\pi_{\theta'}$, remains close to the π_θ. This constraint prevents significant deviations between the two policies, promoting stable and reliable learning.

$$\pi_\theta \approx \pi_{\theta'}$$

- If $P(\theta)$ is deterministic and $P(\theta')$ is stochastic, then their outcomes cannot be directly compared. One solution is to partition the outcome space. Alternatively, we can define a metric over trajectories τ, such that the distance between the distributions over τ for θ and θ' is minimized. we have :

$$p_{\theta'}(s_t) = \underbrace{(1-\epsilon)^t}_{\text{probability we made no mistakes}} p_\theta(s_t) + \underbrace{\left(1-(1-\epsilon)^t\right)}_{\text{some other distribution}} p_{\text{mistake}}(s_t)$$

We then use the total variation distance and write:

$$|p_{\theta'}(s_t) - p_\theta(s_t)| = \left(1-(1-\epsilon)^t\right) |p_{\text{mistake}}(s_t) - p_\theta(s_t)| \le 2\left(1-(1-\epsilon)^t\right)$$

Using the identity $(1-\epsilon)^t \ge 1-\epsilon t$ for $\epsilon \in [0,1]$, we get:

$$|p_{\theta'}(s_t) - p_\theta(s_t)| \le 2\epsilon t$$

In this way, we obtain an upper bound on the difference between the two distributions.
Intuitively, as time progresses, the divergence between the two distributions increases. However, by selecting a smaller value for ϵ, we can mitigate this effect, leading to a tighter bound on their difference.

- In the more complex scenario, we consider both distributions to be stochastic. We assume that the total variation distance between them is bounded by a small ϵ:

$$D_{\text{TV}}(p_{\theta'}, p_\theta) \le \epsilon$$

We use thr lemma and define the joint distribution such that:

$$p(x) = p_X(x), \quad p(y) = p_Y(y), \quad \text{and} \quad p(x=y) = 1-\epsilon$$

It can then be shown that, similar to the previous case, we have:

$$|p_{\theta'}(s_t) - p_\theta(s_t)| = \left(1-(1-\epsilon)^t\right) |p_{\text{mistake}}(s_t) - p_\theta(s_t)| \le 2\left(1-(1-\epsilon)^t\right) \le 2\epsilon t$$

- In the subsequent step, we utilize this bound to estimate the difference between $J(\theta)$ and $J(\theta')$. In this case, the expectation of any arbitrary function f under the distribution $p_{\theta'}(s_t)$ can be written as:

$$\sum_s f(s) p_\theta(s) = \sum_s f(s) \left[p_{\theta'}(s) + p_\theta(s) - p_{\theta'}(s)\right]$$

$$= \sum_s f(s) p_{\theta'}(s) + \sum_s f(s) \left[p_\theta(s) - p_{\theta'}(s) \right]$$

$$\leq \sum_s f(s) p_{\theta'}(s) + \sum_s |f(s)| \cdot |p_\theta(s) - p_{\theta'}(s)|$$

$$\leq \sum_s f(s) p_{\theta'}(s) + \max_s |f(s)| \sum_s |p_\theta(s) - p_{\theta'}(s)|$$

$$= \mathbb{E}_{p_{\theta'}}[f(s)] + \max_s |f(s)| \cdot \|p_\theta - p_{\theta'}\|_1$$

Then, by rearranging terms, we obtain the following inequality, which provides a lower bound:

$$\mathbb{E}_{p_{\theta'}(s_t)}[f(s_t)] \geq \mathbb{E}_{p_\theta(s_t)}[f(s_t)] - 2\epsilon t \cdot \max_{s_t} f(s_t)$$

This corresponds to the difference between $J(\theta)$ and $J(\theta')$.
Therefore, we can derive a lower bound for policy improvement,
and use this bound to ensure that a policy improvement actually occurs.

- We can adjust the value of ϵ over time to control the exploration-exploitation trade-off. By scheduling ϵ appropriately, we can determine its order and ensure effective policy improvement.
- All of these considerations lead us to use the following objective function:

$$\sum_t \mathbb{E}_{s_t \sim \rho_{\theta'}(s_t)} \left[\mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t) \right] \right]$$

$$\geq \sum_t \mathbb{E}_{s_t \sim \rho_\theta(s_t)} \left[\mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t) \right] \right] - \sum_t 2\epsilon t C$$

This is precisely the objective function utilized in PPO.

- SAC can be viewed as a form of policy iteration within the maximum entropy framework. It alternates between soft policy evaluation and soft policy improvement steps, and under certain conditions, this iterative process is guaranteed to converge to the optimal policy within the considered policy class.