2- REINFORCE Algorithm

Behzad Khamidehi

According to the policy gradient theorem, if we define the performance measure for our policy gradient algorithm as $J(\theta) = v_{\pi_\theta}(s_0)$, the value of the start state, we have

$$\nabla_\theta J(\theta) = \mathbb{E}_\pi\{Q_{\pi_\theta}(s,a)\nabla_\theta \ln \pi_\theta(a|s)\} \tag{1}$$

As can be seen in 1, to evaluate the gradient of function $J(\theta)$, we need to know the value of $Q_{\pi_\theta}(s,a)$. However, it is difficult and impossible to have the accurate value of the Q-function. Instead, we can use the Monte-Carlo method to estimate the Q-function by the return function. In what follows, we discuss how to use the return function.

It is worth mentioning that in the policy gradient theorem, we use a performance measure, i.e., $J(\theta) = v_{\pi_\theta}(s_0)$. According to the definition of the state value function,

$$v_\pi(s) = \mathbb{E}_\pi\{G_t|s_t = s\} \tag{2}$$

where $G_t = \sum_{k=0}^{T-1} \gamma^k r_{t+k+1}$ is the return function defined for the episodic task. We can rewrite $J(\theta)$ as

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left\{\sum_{t=0}^{T-1} \gamma^t r_t\right\} \tag{3}$$

If we define a trajectory τ of length T as $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_{T-1}, a_{T-1}, r_{T-1}, s_T)$, where $s_0 \sim p(s_0)$, $a_i \sim \pi_\theta(a_i|s_i)$, and $s_i \sim \mathbf{P}(s_i|s_{i-1}, a_{i-1})$, and $\mathbf{P}$ is the dynamics of the model, our goal in the policy gradient approaches is to directly find a policy to maximize the expected return over all possible trajectories, i.e., $\max_\theta \mathbb{E}_{\pi_\theta}\left\{\sum_{t=0}^{T-1} \gamma^t r_t\right\}$. To do this maximiztion, we need to compute $\nabla_\theta J(\theta)$. In what follows, we we show how to evaluate $\nabla_\theta J(\theta)$.

We remember

$$\begin{aligned}
\nabla_\theta \mathbb{E}\{f(x)\} &= \nabla_\theta \int p_\theta(x) f(x) dx \\
&= \int \frac{p_\theta(x)}{p_\theta(x)} \nabla_\theta p_\theta(x) f(x) dx \\
&= \int p_\theta(x) \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} f(x) dx \\
&= \mathbb{E}\left\{f(x) \nabla_\theta \ln p_\theta(x)\right\}.
\end{aligned} \tag{4}$$

As a result, $\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left\{\sum_{t=0}^{T-1} \gamma^t r_t \nabla_\theta \ln p_\theta(\tau)\right\}$, where $\tau \sim \pi_\theta$ indicates that the trajectory τ is derived by following policy π_θ.

Since

$$p_\theta(\tau) = \mu(s_0) \prod_{t=0}^{T-1} \pi_\theta(a_t|s_t)\mathbf{P}(s_{t+1}|s_t, a_t),$$

we have

$$\nabla_\theta \ln p_\theta(\tau) = \nabla_\theta \sum_{t=0}^{T-1} \ln \pi_\theta(a_t|s_t).$$

As a result, the gradient of $J(\theta)$ is expressed as

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left\{\left(\sum_{t=0}^{T-1}\gamma^t r_t\right)\sum_{t=0}^{T-1}\nabla_\theta \ln \pi_\theta(a_t|s_t).\right\} \tag{5}$$

Now, we can use the Monte-Carlo method to approximate the expectation. According to this approach,

$$\mathbb{E}_\pi\{f(x)\} = \lim_{N\longrightarrow\infty}\frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

where $x_i \sim \pi(x)$. Hence,

$$\nabla_\theta J(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=0}^{T-1}\nabla_\theta \ln \pi_\theta(a_{t,i}|s_{t,i})\right)\left(\sum_{t=0}^{T-1}\gamma^t r_{t,i}\right). \tag{6}$$