Policy Gradient and REINFORCE

Hongshan Li(hongshal@amazon.com)

August 17, 2021

Last time we discussed policy gradient and how it is related to the relative importance of the state and state action value

Theorem 1 *Policy Gradient Theorem For a parametrized policy π_θ, let $J(\theta)$ denote the value of the policy $v_\pi(s_0)$, where s_0 is the initial state of the MDP. Then*

$$\nabla J(\theta) \propto \int_{\mathbb{S}} \mu(s) \int_{\mathbb{A}} \nabla \pi(a|s) q_\pi(s,a) ds da \tag{1}$$

where $\mathbb{S}$ denote the state space of the MDP and $\mathbb{A}$ denote the action space of the policy.

The theorem says the policy gradient, i.e. the direction that improves the policy should be propotional to the importance of the state and the state action value. This should make intuitive sense. The gist of policy gradient is that it maximizes the likelihood of good action, and "good action" means action with high state action value at important state.

1 Estimate Policy Gradient with Monte Carlo Method

In order to estimate the policy gradient with Monte Carlo sampling, the first thing we need to do is to write the integration in the policy gradient theorem into an expectation over policy runs. So that we estimate it by executing the policy a couple of times in the env.
Recall to compute the expectation of a function $f(x)$ over certain distribution $p(x)$ of $x \in \mathbb{R}$, we have

$$\mathbb{E}_{x\sim p(x)}[f(x)] = \int_{\mathbb{R}} p(x) f(x) dx$$

Therefore to make Eq 1 an expection over the policy, we do

$$\begin{aligned}\nabla J(\theta) &\propto \int_{\mathbb{S}} \mu(s) \int_{\mathbb{A}} \nabla \pi(a|s) q_\pi(s,a) ds da \\ &= \int_{\mathbb{S}} \mu(s) \int_{\mathbb{A}} \pi(a|s) \frac{\nabla \pi(a|s)}{\pi(a|s)} q_\pi(s,a) ds da \\ &= \mathbb{E}_{\tau\sim\pi}(\frac{\nabla \pi(a|s)}{\pi(a|s)} q_\pi(s,a)) \\ &= \mathbb{E}_{\tau\sim\pi}(\nabla \log(\pi(a|s)) q_\pi(s,a))\end{aligned}$$

Recall

$$\frac{d \log f(x)}{dx} = \frac{1}{f(x)} \frac{df(x)}{dx}$$

This trick of estimating gradient of through its log derivative is called *log derivative trick* in statistical learning.
This means by following the current policy π for a few epsidoes and compute the average of $\nabla \log(\pi(a|s)) q_\pi(s,a)$ we would get an unbiased estimate of the policy gradient.

2 REINFORCE

Now, we have enough ingredients for a vanilla policy gradient algorithm: REINFORCE (REward Increment = nonnegative factor × Offset Reinforcement × Characteristic Eligibility)

$$\overset{\text{reward increment}}{\epsilon \nabla(J(\theta))} = \overset{\text{non-negative factor}}{\alpha} \times \overset{\text{offset reinforcement}}{q_\pi(s,a)} \times \overset{\text{characteristic eligibility}}{\nabla \log \pi(a|s)}$$

Algorithm 1: REINFORCE

: INPUT: a differentiable policy π_θ; a learning rate α;
Initialze the parameters θ
repeat
  Generate an episode $s_0, a_0, r_1, \cdots, s_{T-1}, a_{T-1}, r_T$
  for $t = T-1, T-2, \cdots, 0$ **do**
    Compute

$$q_\pi(a_t, s_t) = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-t-1} R_T$$

    Estiamte policy gradient

$$\nabla \hat{J}(\theta) = \nabla \log \pi(a_t|s_t) q_\pi(a_t, s_t)$$

    Update parameter

$$\theta \leftarrow \theta + \alpha \frac{1}{T} \nabla \hat{J}(\theta)$$

9: **end for**
10: **until** Policy is good enough