Notes - Chapter 6

Policy Gradient Methods

Policy Gradient (PG) algorithms exhibit incredible potential in enviroments with a large number of actions or when the action space is continuous.

The objective of RL is to maximize the expected return of a trajectory. The objective function can then be expressed as:

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)] \tag{1}$$

Where θ is the parameter of the policy, such as the trainable variables of a deep neural network.

In PG methods, the maximization of the objective function is done through the gradient of the objective function $\nabla_\theta J(\theta)$. Using gradient ascent, we can improve $J(\theta)$ by moving the parameters toward the direction of the gradient, as the gradient points in the direction in which the function increases.

Using equation (1), the gradient of the objective function is defined as follows:

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)] \tag{2}$$