Student ID : 20194293

Name : Go, Kyeong Ryeol

[AI 602] Deterministic Policy Gradient Algorithms

1. Paper Summary

According to the policy gradient theorem, the computation of the performance gradient reduces to a simple expectation by the score function trick. As a previous work, the policy gradient algorithms like REINFORCE estimates the value function by the sample return. Moreover, the actor-critic algorithms try to adjust the parameter of the policy by the actor network and estimate the value function with the critic network. These function approximators may introduce bias if without compatibility.

Often, off-policy actor critic is considered where the behavior policy is distinct from the target policy, which resolves the sample efficiency issue. However, it then suffers from the variance as the importance sampling ratio should be computed. When used along with the stochastic policy, the training gets harder.

Usually, a parametric probability distribution is considered to express the stochastic policy as it is beneficial for exploring the full state and action spaces and moreover, it was believed that the deterministic policy gradient did not exist. Here, the author proposed the deterministic policy gradient algorithms which suggests to use a stochastic behavior policy and a deterministic target policy. This not only to encourage the exploration, but also to exploit the efficiency of the deterministic policy gradient.

In the continuous action spaces, the greedy policy improvement may be problematic and local gradient updates are considered rather than global maximization. Here, the author proposed the deterministic policy gradient theorem, which states the explicit form of this update as follow.

$$\mathrm{\theta}^{\mathrm{k+1}} = \theta^k + \alpha E_{s\sim\rho^\mu}\left[\nabla_\theta \mu_\theta(s) \nabla_\alpha Q^\mu(s,a)|_{a=\mu_\theta(s)}\right]$$

Furthermore, the author also proved that the stochastic policy gradient converges to the deterministic gradient as the variance gets 0. SARSA-based updates and Q-learning-based updates can be found at equation (11~13) and (16~18) in the paper, respectively. Moreover, $\mathrm{Q}^{\mathrm{w}}(s,a)$ is shown to be compatible with $\mathrm{\mu}_{\mathrm{\theta}}(s)$ if the following conditions are satisfied. (Second condition is usually relaxed in favour of efficient policy evaluation algorithms like temporal-difference learning.)

I. $\nabla_a \mathrm{Q}^{\mathrm{w}}(s,a)|_{a=\mu_\theta(s)} = \nabla_{\mathrm{\theta}}\mathrm{\mu}_{\mathrm{\theta}}(\mathrm{s})^{\mathrm{T}}\mathrm{w}$

II. $\mathrm{loss(w)} = \mathrm{E}[\mathrm{\epsilon(s;\theta,w)^T\epsilon(s;\theta,w)}],\ \epsilon(s;\theta,w) = \nabla_a \mathrm{Q}^{\mathrm{w}}(s,a)|_{a=\mu_\theta(s)} - \nabla_a \mathrm{Q}^{\mathrm{\mu}}(s,a)|_{a=\mu_\theta(s)}$

As a result, the author proposed a Compatible Off-Policy Deterministic Actor-Critic (COPDAC) algorithms, which uses the gradient TD learning that minimize MSPBE by stochastic gradient descent. The natural policy gradient can be also applied where it turns out to be simple as follow.

$$\mathrm{\theta}_{\mathrm{t+1}} = \theta_t + \alpha_\theta w_t \quad where \quad \mathrm{M}_{\mathrm{\mu}}(\theta) = E_{s\sim\rho^\mu}[\nabla_\theta \mu_\theta(s) \nabla_\theta \mu_\theta(s)^T]$$

2. In depth discussion

I. Rather than explicitly using the deterministic policy, are there any way to schedule the variance of the stochastic policy?

II. What may be some additional strategy for applying the deterministic policy gradient when handling really large state and action spaces that require more exploration?