Actor Critic Method

Kihwan Lee

Contents

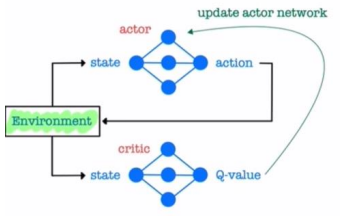
1. Actor Critic Method

Actor-Critic Method

 # Review

- value ft Q(s,a) > Deep Q-Network
- policy $\pi(a|s)$ > Policy Gradient (REINFORCE)

=> **value ft + policy > Actor-Critic (A3C)**



Actor-Critic Method

 # TD Actor-Critic Method

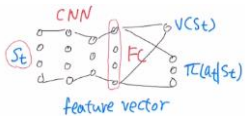
Gt는 끝까지 고려해야 하기에 baseline을 줌에도 불구하고 high variance

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{\pi_\theta}[r(\tau)] = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} (G_t - b(s_t)) \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right]$$

$$\Rightarrow \quad E_{\pi_\theta}\left[\left(r + \gamma V_\emptyset(s') - V_\emptyset(s)\right)\nabla_\theta \log \pi_\theta[a|s]\right]$$

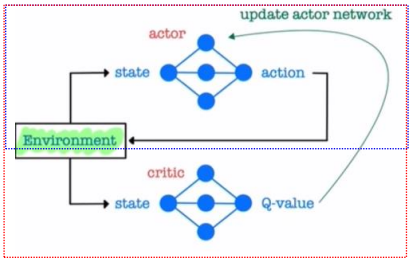
TD Error !

① Reduces variance
② Bootstrapping > Online Learning
③ Accelerates learning



Policy Gradient

Actor Critic



2. Variants of Critic

Actor-Critic Method

 # Variants of Critic

① High variance

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} G_t \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right] \quad \text{REINFORCE (Monte Carlo PG)}$$

$$= \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} (G_t - V_\phi(s_t)) \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right] \quad \text{REINFORCE with baseline}$$

$$= \mathbb{E}_{\pi_\theta}[Q_\phi(s,a) \nabla_\theta \log \pi_\theta(a \mid s)] \quad \text{Q-value Actor-Critic}$$

④ $Q_\emptyset$ - $V_\emptyset$

$$= \mathbb{E}_{\pi_\theta}[A_{\phi_1,\phi_2}(s,a) \nabla_\theta \log \pi_\theta(a \mid s)] \quad \text{Advantage Actor-Critic}$$

② TD Error $= \mathbb{E}_{\pi_\theta}[(r + \gamma V_\phi(s') - V_\phi(s)) \nabla_\theta \log \pi_\theta(a \mid s)]$ TD Actor-Critic

Critic (value function) Actor (policy)

③ $Q_\emptyset(s,a)$

Actor-Critic Method

 # 수도 코드

Initialize critic network $V(s;\phi)$ and actor network $\pi(a\,|\,s;\theta)$ randomly
Hyperparameters: stepsizes $\alpha > 0$, $\beta > 0$
for episode = $1, M$ **do**
  Initialize s, the first state of the episode
  $I \leftarrow 1$
  for s *is not terminal* **do**
    Select action a according to policy $\pi(\cdot\,|\,s;\theta)$
    Execute a and observe r, s'
    $\delta \leftarrow r + \gamma V(s';\phi) - V(s;\phi)$: TD Error!
    $\phi \leftarrow \phi + \beta\,\delta\,\nabla_\phi V(s;\phi)$
    $\theta \leftarrow \theta + \alpha I \delta\,\nabla_\theta \log \pi(a\,|\,s;\theta)$
    $I \leftarrow \gamma I$
    $s \leftarrow s'$
  end
end

$$\Delta\phi = \beta\left(r_{t+1} + \gamma\, V_\phi(s_{t+1}) - V_\phi(s_t)\right)\nabla_\phi V_\phi(s_t)$$

$$E_{\pi_\theta}\left[\left(r + \gamma V_\phi(s') - V_\phi(s)\right)\nabla_\theta \log \pi_\theta[a|s]\right]$$

TD method
Critic
Actor