Policy Gradient

- Basic 분류 Policy를 구하는법

Value Method — 1) value function을 구하고 2) policy 구하기
ex. 지금까지의 알고리즘

Policy Method — 1) Policy 구하기

Policy Gradient
- Finite Difference Policy Gradient
- Monte-Carlo Policy Gradient
- Actor-Critic Policy Gradient

* Policy Based RL Find θ that maximises $J(\theta)$

장) better convergence properties
effective in high-dimensional action spaces
can learn stochastic policies

단) converge to local
high variance

$$\theta_{t+1} = \theta_t + \Delta\theta$$
$$= \theta_t + \alpha \nabla_\theta J(\theta)$$
$$= \theta_t + \alpha (E_{\pi_\theta} [\nabla_\theta \log \pi_\theta (s,a) \, Q^{\pi_\theta}(s,a)])$$

↳ Policy Gradient Theorem

- Monte-Carlo Policy Gradient (REINFORCE) (θ).

$$\Delta\theta_t = \alpha \nabla_\theta \log \pi_\theta (s_t, a_t) v_t$$

using stochastic gradient ascent, policy gradient theorem

- Actor-Critic Policy Gradient (θ, w)

update
- Actor (update action-value function) — parameter θ 목적
- Critic (to estimate action-value function) — parameter w 목적의 목적

$$\Delta\theta = \alpha \nabla_\theta \log \pi_\theta (s,a) \, Q_w (s,a)$$ (approximate policy gradient)

- Policy Gradient algorithms 강화학습 find θ ← use $J(\theta)$!

$\nabla_\theta J(\theta) = E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a)\, V_t]$ REINFORCE

$E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a)\, Q^w(s,a)]$ Q Actor-Critic

$E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a)\, A^w(s,a)]$ Advantage Actor-Critic

$E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a)\, \delta]$ TD Actor-Critic

$E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a)\, \delta_e]$ TD(λ) Actor-Critic

$G_\theta^{-1} \nabla_\theta J(\theta) = w$ Natural Actor-Critic