Background

- Notation
  - $\mathcal{S}$: state space
  - $\mathcal{A}$: action space
  - $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$: reward function
  - $G_t = r_t^\gamma = \sum_{k=t}^{\infty} \gamma^{k-t} r(s_k, a_k)$: total discounted reward from time-step t
  - $\pi_\theta : \mathcal{S} \to \mathcal{P}(\mathcal{A})$: policy ,$\pi_\theta(a_t \mid s_t)$ is the conditional probability at a_t associated with policy
  - $p_1(s_1)$: initial state distribution
  - $V^\pi(s_t) = \mathbb{E}[G_t \mid s_t; \pi]$: State Value Function
  - $Q^\pi(s_t, a_t) = \mathbb{E}[G_t \mid s_t, a_t; \pi]$: Action Value Function
  - $A^\pi(s_t) = Q^\pi(s_t) - V^\pi(s_t)$
  - $\rho_\pi(s) = P(s_0 = s) + \gamma P(s_1 = s) + \cdots$: unnormalized discounted visitation frequencies, improper discounted state distribution

- Basic Objective Function

$$J(\theta) = \mathbb{E}_{p_\theta(\tau)}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right] = \mathbb{E}_{p_\theta(\tau)}[G_0] \tag{1}$$

  - $\tau : (s_0, a_0, \cdots, s_T, a_T)$

$$\nabla_\theta J(\theta) = \int_\tau \nabla_\theta \log p_\theta(\tau) \sum_{t=0}^{T} \gamma^t r(s_t, a_t) p_\theta(\tau) d\tau \tag{2}$$

$$\nabla_\theta J(\theta) = \int_\tau \left(\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right) \left(\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right) p_\theta(\tau) d\tau \tag{3}$$

  - Transition probability disappears → model free

$$\nabla_\theta J(\theta) = \int_\tau \left(\sum_{t=0}^{T} \gamma^t \nabla_\theta \log \pi_\theta(a_t \mid s_t) \left(\sum_{k=t}^{T} \gamma^{k-t} r(s_k, a_k) \right) \right) p_\theta(\tau) d\tau \tag{4}$$

$$\nabla_\theta J(\theta) \approx \int_\tau \left(\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \left(\sum_{k=t}^{T} \gamma^{k-t} r(s_k, a_k) \right) \right) p_\theta(\tau) d\tau \tag{5}$$

- Because discount factor makes behind episode useless
- This is biased gradient of objective function

$$\begin{aligned}\nabla_\theta J(\theta) = \sum_{t=0}^{T} \int_{\tau_{s_0:a_t}} \int_{\tau_{s_{t+1}:a_T}} & p_\theta(\tau_{s_{t+1}:a_T} \mid \tau_{s_0:a_t}) p_\theta(\tau_{s_0:a_t}) \\ & \left(\gamma^t \nabla_\theta \log \pi_\theta(a_t \mid s_t) \left(\sum_{k=t}^{T} \gamma^{k-t} r(s_k, a_k) \right) \right) d\tau \end{aligned} \tag{6}$$

$$\begin{aligned} = \sum_{t=0}^{T} \int_{\tau_{s_0:a_t}} & \gamma^t \nabla_\theta \log \pi_\theta(a_t \mid s_t) p_\theta(\tau_{s_0:a_t}) \\ & \left[\int_{\tau_{s_{t+1}:a_T}} p_\theta(\tau_{s_{t+1}:a_T} \mid \tau_{s_0:a_t}) \left(\sum_{k=t}^{T} \gamma^{k-t} r(s_k, a_k) \right) \right] d\tau \end{aligned} \tag{7}$$

$$= \sum_{t=0}^{T} \int_{\tau_{s_0:a_t}} \gamma^t \nabla_\theta \log \pi_\theta(a_t \mid s_t) Q^\pi(s_t, a_t) p_\theta(\tau_{s_0:a_t}) d\tau_{s_0:a_t} \tag{8}$$

$$= \sum_{t=0}^{T} \int_{(s_t, a_t)} \gamma^t \nabla_\theta \log \pi_\theta(a_t \mid s_t) Q^\pi(s_t, a_t) p_\theta(s_t, a_t) ds_t da_t \tag{9}$$

$$\approx \sum_{t=0}^{T} \int_{(s_t, a_t)} \nabla_\theta \log \pi_\theta(a_t \mid s_t) Q^\pi(s_t, a_t) p_\theta(s_t, a_t) ds_t da_t \tag{10}$$

- Another approach for gradient of objective function

$$J(\theta) = \sum_s \rho_{\pi_\theta}(s) \sum_a \pi_\theta(a \mid s) Q^\pi(s, a) \tag{11}$$

$$\nabla_\theta J(\theta) \propto \sum_s \rho_{\pi_\theta}(s) \sum_a \nabla_\theta \pi_\theta(a \mid s) Q^\pi(s, a) \tag{12}$$

- Use Advantage Function
  - In equation 10 replace Q as b which is not function of a_t

$$\begin{aligned} &\sum_{t=0}^{T} \int_{(s_t,a_t)} \nabla_\theta \log \pi_\theta(a_t \mid s_t) B \pi_\theta(a_t \mid s_t) p_\theta(s_t) ds_t da_t \\ &= \sum_{t=0}^{T} \int_{s_t,} B p_\theta(s_t) \nabla_\theta \int_{a_t} \pi_\theta(a_t \mid s_t) ds_t da_t \\ &= 0 \end{aligned} \tag{13}$$

$$\nabla_\theta J(\theta) \approx \sum_{t=0}^{T} \int_{(s_t,a_t)} \nabla_\theta \log \pi_\theta(a_t \mid s_t) A^\pi(s_t,a_t) p_\theta(s_t,a_t) ds_t da_t \tag{14}$$

Algorithm

A2C(Advantage Actor Critic)

- Q function 대신 Advantage function을 policy gradient에서 사용
- Policy를 추정하는 Actor와 Value를 추정하는 Critic 구조로 설계되어 있음
- Objective function
  - Actor

$$\nabla_\theta J(\theta) \approx \sum_{t=0}^{T} \int_{(s_t,a_t)} \nabla_\theta \log \pi_\theta(a_t \mid s_t) A^\pi(s_t,a_t) p_\theta(s_t,a_t) ds_t da_t \tag{15}$$

    - Gradient ascent
  - Critic

$$\nabla_\phi J(\phi) = \left[r(s_t,a_t) + \gamma V_\phi(s_{t+1}) - V_\phi(s_t) \right] \nabla_\phi V(s_t) \tag{16}$$

PPO(Proximal Policy Optimization)

- Sample efficiency를 위해 sample들을 재사용
- 이때 $\pi_\theta \approx \pi_{\theta_{old}}$ 를 위해 clipping 도입
- Objective function
    - Actor

$$L^{CLPI}(\theta) = \hat{\mathbb{E}}_t[\min(r_t(\theta)\hat{A}_t, \text{clip }(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \quad (17)$$

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1} \quad (18)$$

$$\text{where } \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \quad (19)$$

        - Advantage 함수 대신 GAE(Global Average Pooling) 사용
        - Gradient ascent
    - Critic

$$\nabla_\phi J(\phi) = [r(s_t, a_t) + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)] \nabla_\phi V(s_t) \quad (20)$$

DDPG(Deep Deterministic Policy Gradient)

- Policy를 deterministic하게 가져감
- DQN을 continuous action space에 적용한 개념으로도 볼 수 있다.
- Action에 대한 적분이 모두 사라지므로 계산이 쉬워진다.
- Off-Policy
- Objective function
    - Actor

$$\nabla_{\theta^\mu} J = \mathbb{E}_{s_t \sim \rho^\beta}[\nabla_a Q(s, a \mid \theta^Q)|_{s=s_t, a=\mu(s_t)} \nabla_{\theta^\mu}\mu(s \mid \theta^\mu)|_{s=s_t}] \quad (21)$$

        - Gradient ascent
    - Critic

$$L(\theta^Q) = \mathbb{E}_{s_t \sim \rho^\beta, a_t \sim \beta, r_t \sim E}\left[\left(Q(s_t, a_t \mid \theta^Q) - y_t\right)^2\right] \quad (22)$$

$$\text{where } y_t = r(s_t, a_t) + \gamma Q(s_{t+1}, \mu(s_{t+1}) \mid \theta^Q) \quad (23)$$

SAC(Soft Actor Critic)

- Policy가 가능하면 entropy가 커지도록 유도
- Objective function에 entropy항을 추가
- Off-Policy
- Objective function
  - Actor

$$\hat{\nabla}_\phi J_\pi(\phi) = \nabla_\phi \log \pi_\phi(a_t \mid s_t) + \\ (\nabla_{a_t} \log \pi_\phi(a_t \mid s_t) - \nabla_{a_t} Q(s_t, a_t)) \nabla_\phi f_\phi(\epsilon_t; s_t) \tag{24}$$

$$\text{where } a_t = f_\phi(\epsilon_t; s_t) \tag{25}$$

  - Critic

$$\hat{\nabla}_\psi J_V(\psi) = \nabla_\psi V_\psi(s_t)(V_\psi(s_t) - Q_\theta(s_t, a_t) + \log \pi_\phi(a_t \mid s_t)) \tag{26}$$

$$J_Q(\theta) = \mathbb{E}_{(s_t, a_t) \sim \mathcal{D}} \left[\frac{1}{2} \left(Q_\theta(s_t, a_t) - \hat{Q}(s_t, a_t) \right)^2 \right] \tag{27}$$