Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor

키워드

Entropy , sample complexity , stable

Abstract

Model-free deep reinforcement learning (RL) 의 난제 2개 :
sample complexity와 brittle convergence properties 해결 어려움
->액션의 차원이 높으면 높을수록 다양한 형태의 샘플이 잔뜩 필요하고, 자그만한 하이퍼 파라미터 변화에도 수렴이 잘 안되는 불안정한 현상이 보인다.

여기서 저자가 제안한 Soft Actor Critic은 maximum entropy reinforcement learning framework을 기반으로 한 off-policy actor critic deep RL을 제안하였다.

여기서 Actor는 entropy를 최대로 하는 동시에 Expected reward를 최대로 목표한다.
또 다른 이 프레임워크는 안정적이다.

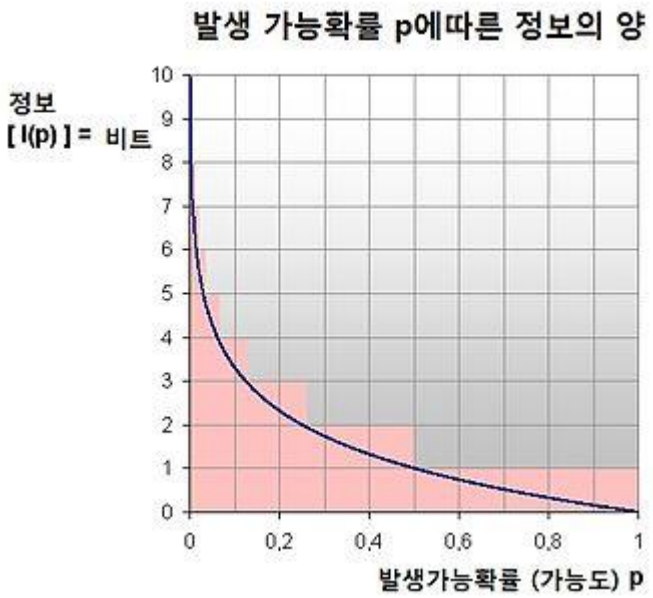
엔트로피 란?

사전적 정의 (열역학): 유용하지 않은 (일로 변환할 수 없는) 에너지의 흐름을 설명할 때 이용하는 상태 함수

통계적 정의 (정보 이론): 어떤 사건이 정보적 측면에서 얼마나 중요한가를 반영한 로그 지표에 대한 기댓값
-> 정보량

$$H(X) = \mathrm{E}\left[\log \frac{1}{P_i(x)}\right]$$

엔트로피가 커진다?

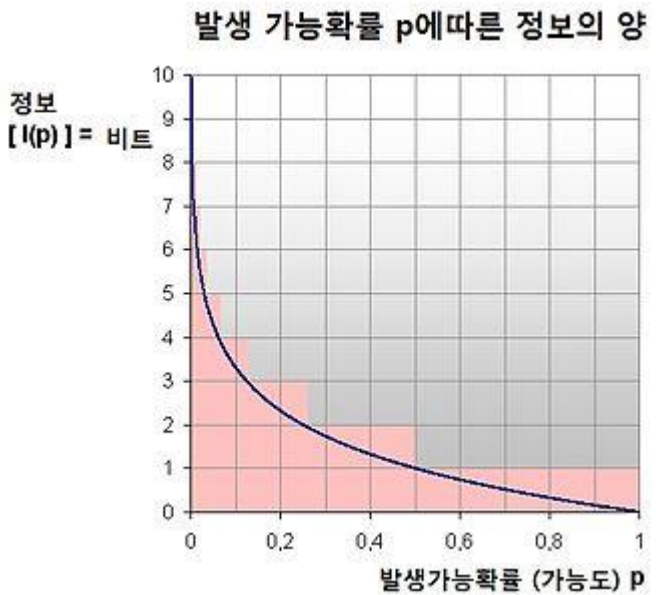


$$H(X) = \mathrm{E}\left[\log \frac{1}{P_i(x)}\right]$$

요약 : 확률이 낮을수록, 어떤 정보일지는 불확실하게 되고,
우리는 이때 '정보가 많다', '엔트로피가 높다'고 표현한다.

Actor의 엔트로피가 최대?

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[r(\mathbf{s}_t, \mathbf{a}_t) + \alpha \mathcal{H}(\pi(\cdot | \mathbf{s}_t)) \right].$$



$$H(X) \;=\; \mathrm{E}\left[\log \frac{1}{P_i(x)}\right]$$

요약 : 확률이 낮은 action을 최대한 많이 선택한다.

→ 각 state에서 잘 선택하지 않던 action(선택할 확률이 낮은 action)도 선택함으로서 exploration 향상

Derivation of Soft Policy Iteration

$$\mathcal{T}^{\pi}Q(\mathbf{s}_t, \mathbf{a}_t) \triangleq r(\mathbf{s}_t, \mathbf{a}_t) + \gamma\, \mathbb{E}_{\mathbf{s}_{t+1}\sim p}\left[V(\mathbf{s}_{t+1})\right]$$

$$\text{※ } V(\mathbf{s}_t) = \mathbb{E}_{\mathbf{a}_t\sim\pi}\left[Q(\mathbf{s}_t, \mathbf{a}_t) - \log \pi(\mathbf{a}_t|\mathbf{s}_t)\right]$$

Lemma 1 (Soft Policy Evaluation)**.** *Consider the soft Bellman backup operator* $\mathcal{T}^{\pi}$ *in Equation 2 and a mapping* $Q^0 : \mathcal{S}\times\mathcal{A} \rightarrow \mathbb{R}$ *with* $|\mathcal{A}| < \infty$*, and define* $Q^{k+1} = \mathcal{T}^{\pi}Q^k$*. Then the sequence* Q^k *will converge to the soft Q-value of* π *as* $k \rightarrow \infty$*.*

Derivation of Soft Policy Iteration

$$\pi_{\text{new}} = \arg\min_{\pi' \in \Pi} \mathrm{D}_{\mathrm{KL}} \left(\pi'(\cdot|\mathbf{s}_t) \,\middle\|\, \frac{\exp\left(Q^{\pi_{\text{old}}}(\mathbf{s}_t, \cdot)\right)}{Z^{\pi_{\text{old}}}(\mathbf{s}_t)} \right). \tag{4}$$

Lemma 2 (Soft Policy Improvement)**.** *Let* $\pi_{\text{old}} \in \Pi$ *and let* π_{new} *be the optimizer of the minimization problem defined in Equation 4. Then* $Q^{\pi_{\text{new}}}(\mathbf{s}_t, \mathbf{a}_t) \geq Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t)$ *for all* $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$ *with* $|\mathcal{A}| < \infty$.

Derivation of Soft Policy Iteration

Theorem 1 (Soft Policy Iteration). *Repeated application of soft policy evaluation and soft policy improvement from any* $\pi \in \Pi$ *converges to a policy* π^* *such that* $Q^{\pi^*}(\mathbf{s}_t, \mathbf{a}_t) \geq Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t)$ *for all* $\pi \in \Pi$ *and* $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$*, assuming* $|\mathcal{A}| < \infty$.

Soft Actor Critic

Algorithm 1 Soft Actor-Critic

Initialize parameter vectors ψ, $\bar{\psi}$, θ, ϕ.
for each iteration **do**
  for each environment step **do**
    $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)$
    $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$
  end for
  for each gradient step **do**
    $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi J_V(\psi)$
    $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$
    $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
    $\bar{\psi} \leftarrow \tau\psi + (1 - \tau)\bar{\psi}$
  end for
end for

Soft Actor Critic

V value loss function

$$J_V(\psi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}} \left[\frac{1}{2} \left(V_\psi(\mathbf{s}_t) - \mathbb{E}_{\mathbf{a}_t \sim \pi_\phi} \left[Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \log \pi_\phi(\mathbf{a}_t | \mathbf{s}_t) \right] \right)^2 \right] \tag{5}$$

Q value loss function

$$J_Q(\theta) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \mathcal{D}} \left[\frac{1}{2} \left(Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \hat{Q}(\mathbf{s}_t, \mathbf{a}_t) \right)^2 \right], \tag{7}$$

※ $\hat{Q}(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \gamma \, \mathbb{E}_{\mathbf{s}_{t+1} \sim p} \left[V_{\bar{\psi}}(\mathbf{s}_{t+1}) \right],$

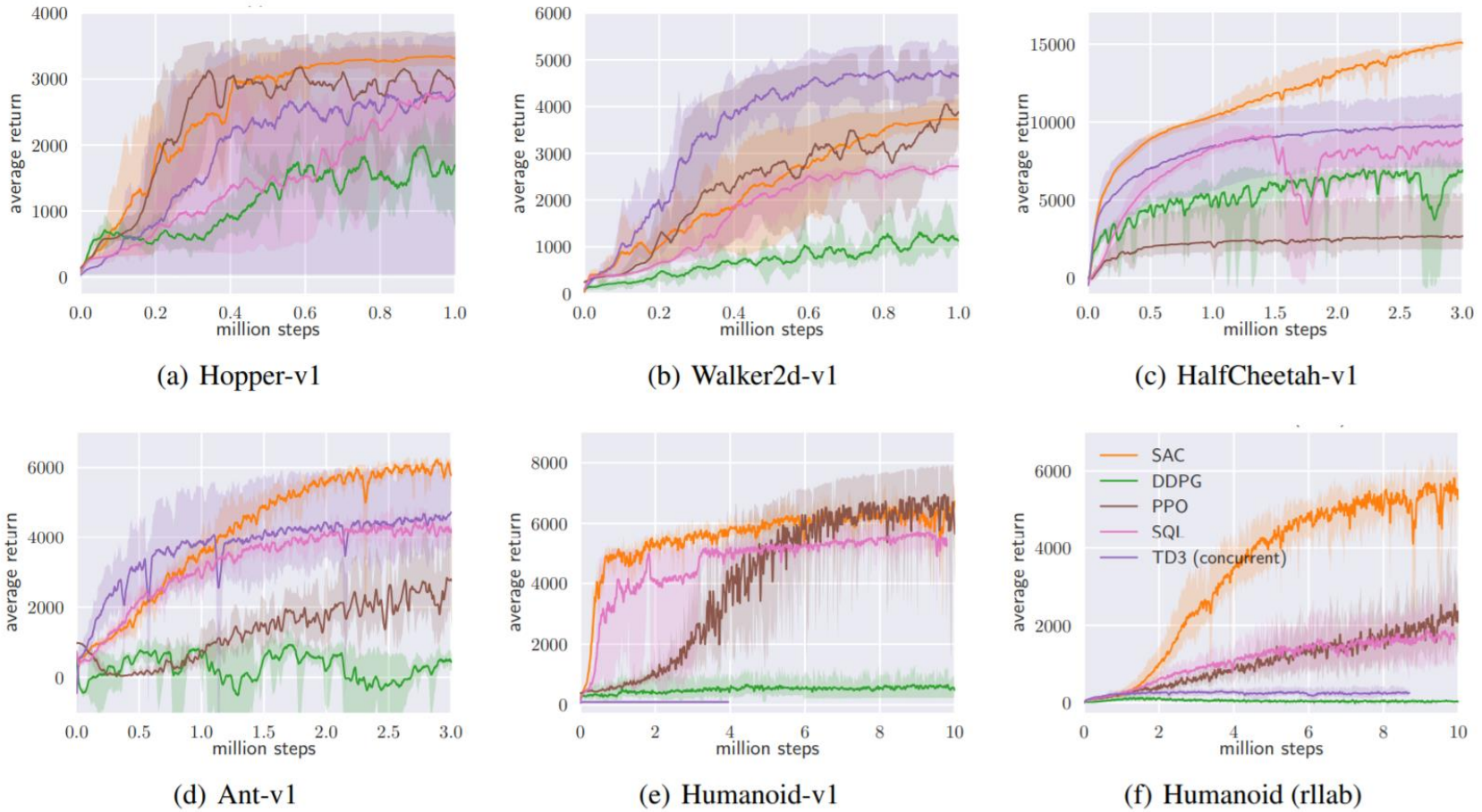
Soft Actor Critic

Policy loss function

$$J_\pi(\phi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}} \left[\mathrm{D}_{\mathrm{KL}} \left(\pi_\phi(\cdot|\mathbf{s}_t) \,\middle\|\, \frac{\exp\left(Q_\theta(\mathbf{s}_t, \cdot)\right)}{Z_\theta(\mathbf{s}_t)} \right) \right]. \tag{10}$$

$$J_\pi(\phi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}, \epsilon_t \sim \mathcal{N}} \left[\log \pi_\phi(f_\phi(\epsilon_t; \mathbf{s}_t)|\mathbf{s}_t) - Q_\theta(\mathbf{s}_t, f_\phi(\epsilon_t; \mathbf{s}_t)) \right], \tag{12}$$

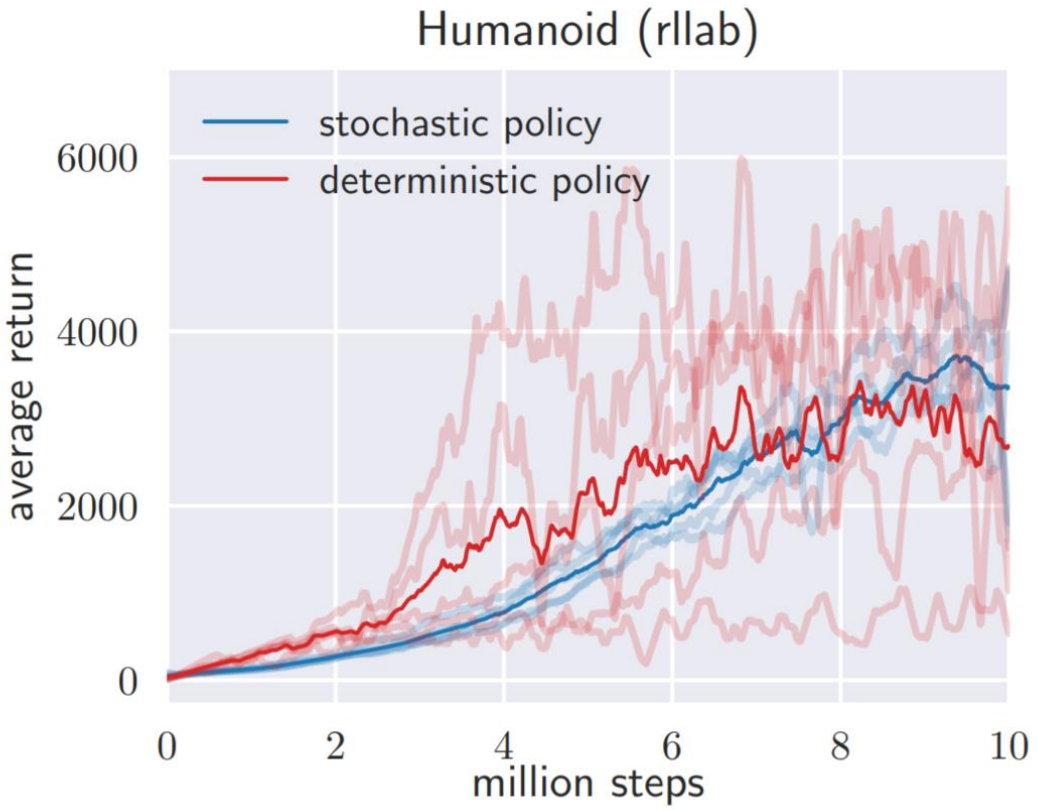
Experiment



(a) Hopper-v1 (b) Walker2d-v1 (c) HalfCheetah-v1

(d) Ant-v1 (e) Humanoid-v1 (f) Humanoid (rllab)

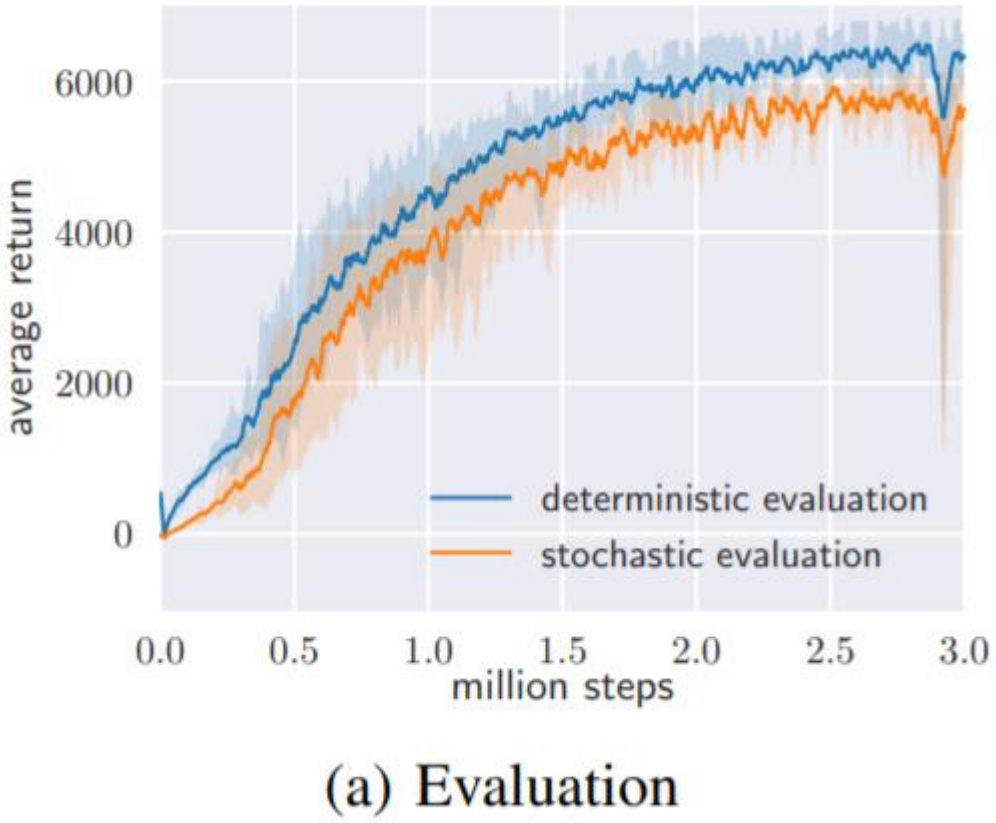
Stochastic vs deterministic policy



Stochastic은 안정적으로 expected return이 증가

Deterministic은 변동성이 너무 큰상태로 expected return 이 증가

Policy evaluation



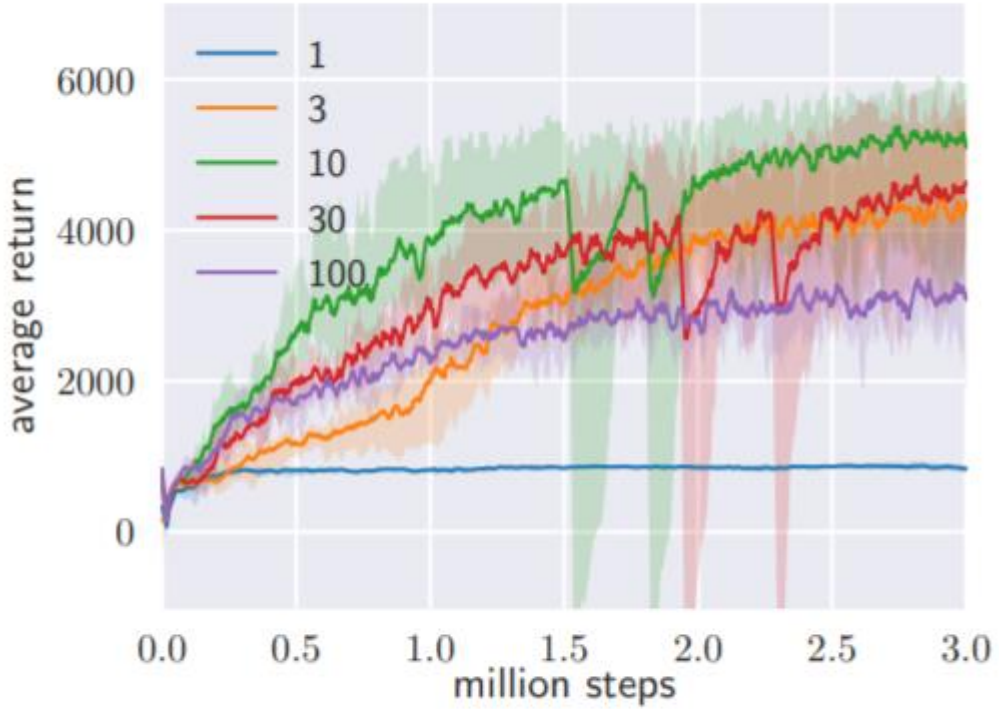
(a) Evaluation

최종적으로 stochastic policy를 deterministic하게 해야 더 나은 성능을 보여준다.

여기서 stochastic evalution은 원래 stochastic polic의 expected return을 평가하는거고

Deterministic evalution은 그 stochastic policy distribution의 mean 값에 해당되는 action을 선택해서 계산된 Expected return을 평가하는걸 의미한다.

Reward Scale



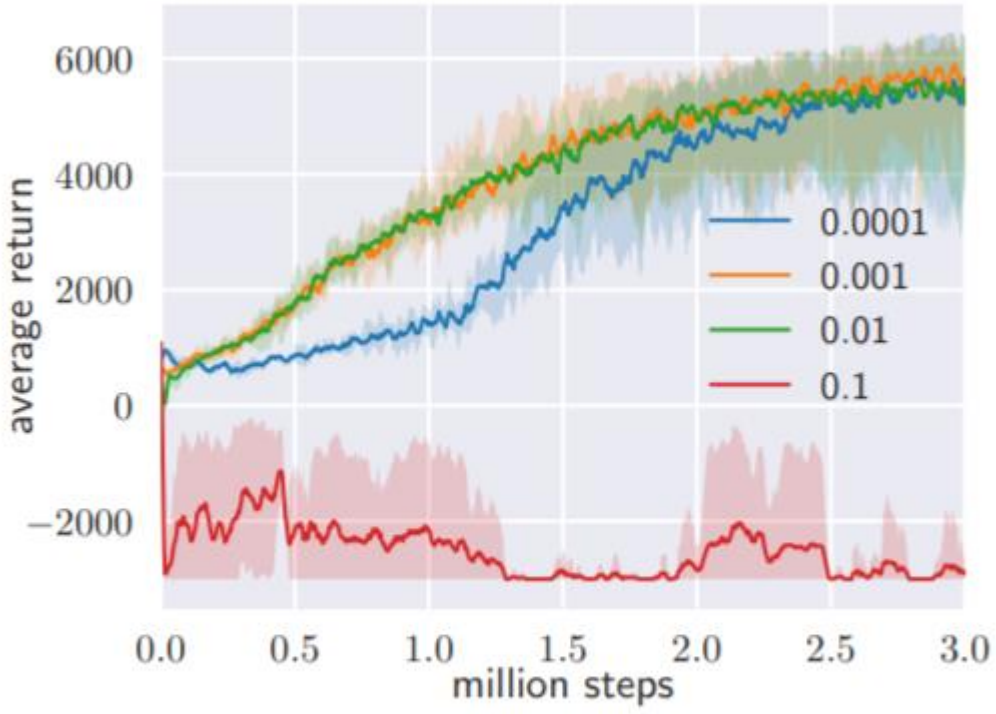
(b) Reward Scale

Soft Actor Critic은 reward singnal에 특히나 더 예민하다.

Reward magnitud가 크면 클수록 학습 속도가 빠를수 있지만, 이로 인해 exploration이 떨어지면서 local minima 에 빠지게 된다.

적절한 reward scale 로 조종해야 exploratio과 exploitation이 적절하게 균형이 맞추어진다.

Target network update



(c) Target Smoothing Coefficient (τ)

τ가 크면 training이 빠르게 진행되고, 작으면 느리게 진행된다.

감사합니다