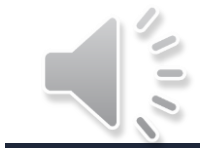
Deterministic Policy Gradient Algorithm

2024.11.01

Data Mining & Quality Analytics Labs

이준범



발표자 소개

❖ 이준범 (Junbeom LEE)

- 고려대학교 산업경영공학과 석사과정 (2024.03 ~)
- Data Mining & Quality Analytics Lab. (김성범 교수님)

❖ 관심 연구 분야

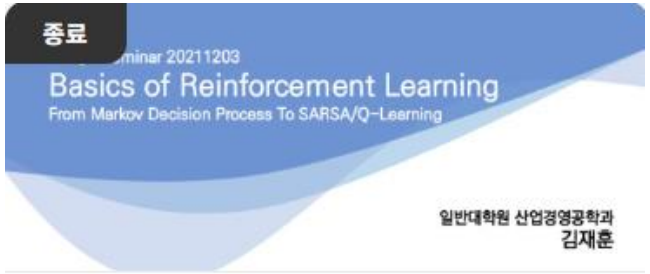
- Reinforcement learning

❖ E-mail

- junbeom99@korea.ac.kr

관련 세미나

- ❖ Basics of Reinforcement Learning : 강화학습의 이해 할 수 있는 기본 지식들을 소개
- ❖ Value-based Learning : 가치 기반 강화학습에 대한 전반적인 내용을 자세히 소개



Basics of Reinforcement Learning

발표자: 김재훈

2021년 12월 3일

오후 1시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶

Value-based Learning

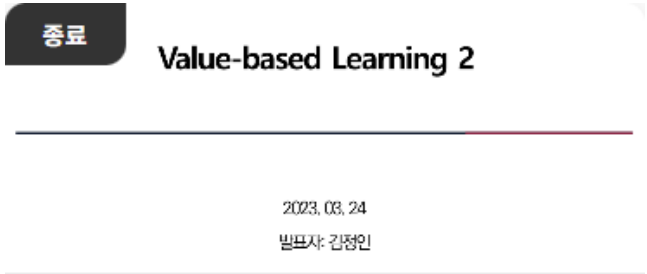
발표자: 허종국

2021년 7월 16일

오후 1시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶



Value-based Learning 2

발표자: 김정인

2023년 3월 24일

오전 12시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶

Contents

- ❖ **Background**
- ❖ **Deterministic Policy Gradient Algorithm**
  - DDPG
  - TD3
- ❖ **Conclusions**

Background
Reinforcement Learning

❖ 강화학습(Reinforcement Learning)

- 목표: 마리오가 장애물을 피해 목표 지점인 깃발에 도달

에이전트 (Agent)

환경 (environment)

Background
Reinforcement Learning

❖ **강화학습(Reinforcement Learning)**

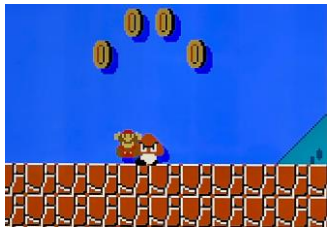
- 마리오는 게임 상황을 학습하며 목표를 달성하기 위해 적절한 행동을 선택함
- 행동에 따라 긍정적 또는 부정적 보상을 받으면서 점점 더 나은 전략을 학습함

에이전트 (Agent)

상태(state)

행동 (Action)

+1



-50

보상 (Reward)

Background
Reinforcement Learning

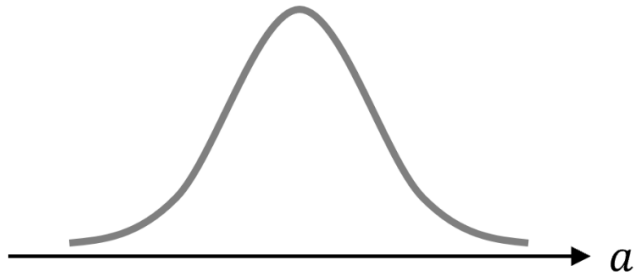
❖ 강화학습(Reinforcement Learning)

- 순차적인 문제 상황에서 에이전트(agent)가 환경(environment)과 상호작용하며 행동(Action)을 수행하고, 그 결과의 누적보상을 최대화하기 위한 정책을 학습하는 알고리즘

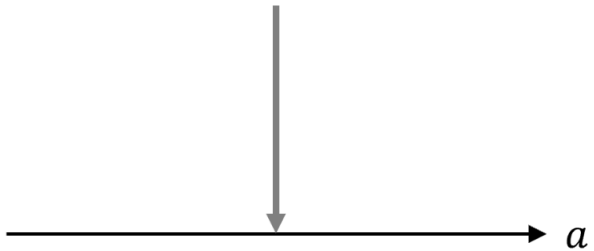
Background
Deterministic Policy Gradient

❖ Deterministic Policy

- 주어진 상태에서 행동을 확률적(stochastic)으로 선택하지 않고, **결정론적(Deterministic)**으로 선택
- Deterministic 정책은 불확실성을 줄여 학습을 효율적으로 실행가능



$$\boldsymbol{Stochastic}$$
$$\boldsymbol{\pi_\theta(a|s)}$$



$$\boldsymbol{Deterministic}$$
$$\boldsymbol{\mu_\theta(s)}$$

Background
Deterministic Policy Gradient

❖ **On–Policy & Off-Policy**

- On-policy : 행동 정책과 타겟 정책이 일치하는 경우
- Off-policy : 행동 정책과 타겟 정책이 일치하지 않는 경우

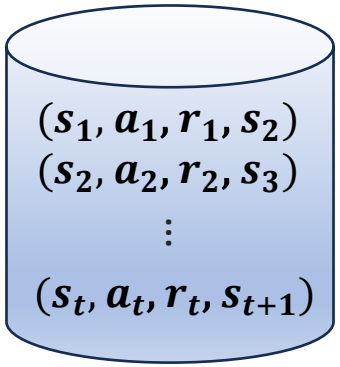
On-policy

Off-policy

Background
Deterministic Policy Gradient

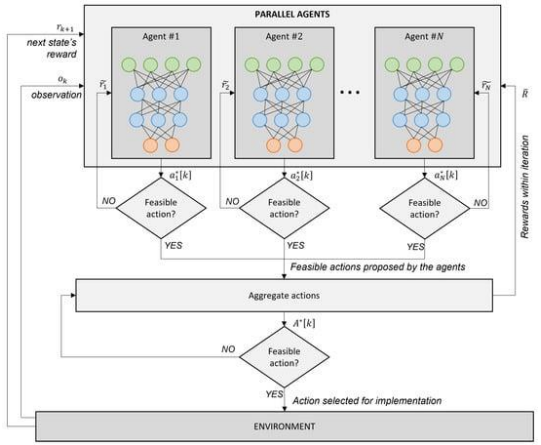
❖ **Off-Policy 의 장점**

1. 과거의 경험을 재사용하여 학습할 수 있음
2. 사람 데이터나 기존의 데이터를 가지고 학습할 수 있음
3. 동시에 여러 개의 모델을 한번에 학습 시킬 수 있음



Replay memory

AlphaGo



Parallel Reinforcement Learning

DDPG
Deep Deterministic Policy Gradient

❖ **DDPG : CONTINUOUS CONTROL WITH DEEP REINFORCEMENT (2016, ICLR)**

- DPG알고리즘을 Deep Neural Network을 사용해 높은 행동공간의 문제를 효과적으로 해결

Published as a conference paper at ICLR 2016

CONTINUOUS CONTROL WITH DEEP REINFORCEMENT LEARNING

Timothy P. Lillicrap*, Jonathan J. Hunt*, Alexander Pritzel, Nicolas Heess,
Tom Erez, Yuval Tassa, David Silver & Daan Wierstra
Google Deepmind
London, UK
{countzero, jjhunt, apritzel, heess,
etom, tassa, davidsilver, wierstra} @ google.com

ABSTRACT

We adapt the ideas underlying the success of Deep Q-Learning to the continuous action domain. We present an actor-critic, model-free algorithm based on the deterministic policy gradient that can operate over continuous action spaces. Using the same learning algorithm, network architecture and hyper-parameters, our algorithm robustly solves more than 20 simulated physics tasks, including classic

DDPG
Deep Deterministic Policy Gradient

❖ **연속행동공간(continuous action space)의 필요성**

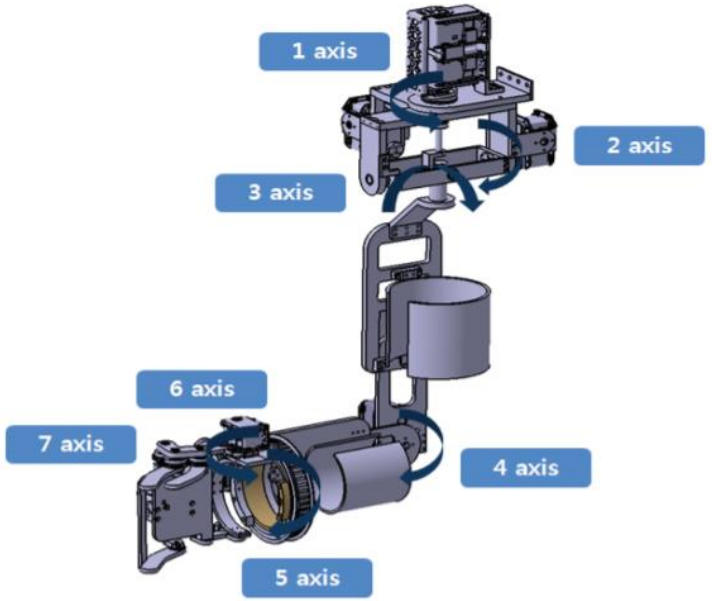
- DQN의 단점 : 이산적(Discrete)이고 차원이 낮은 행동 공간을 지닌 task만 해결가능

Action = {오른쪽, 정지, 왼쪽}

Action space = $3^7 = 2187$

높은 행동공간으로 인한 차원의 저주로 인해 학습이 어려움

7자유도 시스템



DDPG
Deep Deterministic Policy Gradient

❖ **연속행동공간(continuous action space)의 필요성**

- DQN의 단점 : 이산적(discrete)이고 차원이 낮은 행동 공간을 지닌 task만 해결가능

7자유도 시스템

Action = {오른쪽, 정지, 왼쪽}

Action space = $3^7 = 2187$

높은 행동공간으로 인한 차원의 저주로 인해 학습이 어려움

고차원의 행동공간을 처리하기 위해서는
연속적인 행동공간을 사용하는 알고리즘이 필요

DDPG
Deep Deterministic Policy Gradient

❖ **DDPG (Deep Deterministic Policy Gradient)**

- Deep Neural Network을 DPG알고리즘에 적용하여 **높은 차원의 연속적인 행동공간을** 해결
- 효과적인 학습을 위해 **DQN 알고리즘에 사용된 트릭**을 사용

DPG + Neural Network

DQN (experiment buffer+ target network)

Deep Deterministic Policy Gradient

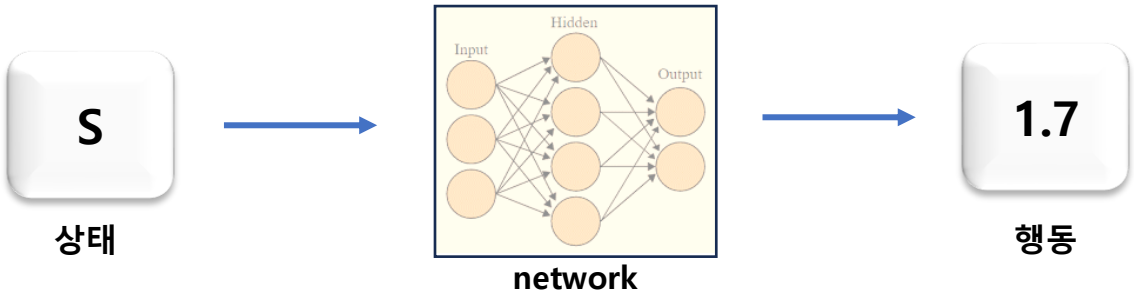
DDPG
Deep Deterministic Policy Gradient

❖ Deterministic Policy

- 주어진 상태에서 행동을 확률적(stochastic)으로 선택하지 않고, **결정론적(Deterministic)**으로 선택
- Deterministic 정책은 불확실성을 줄여 학습을 효율적으로 실행가능

Stochastic policy

Deterministic policy



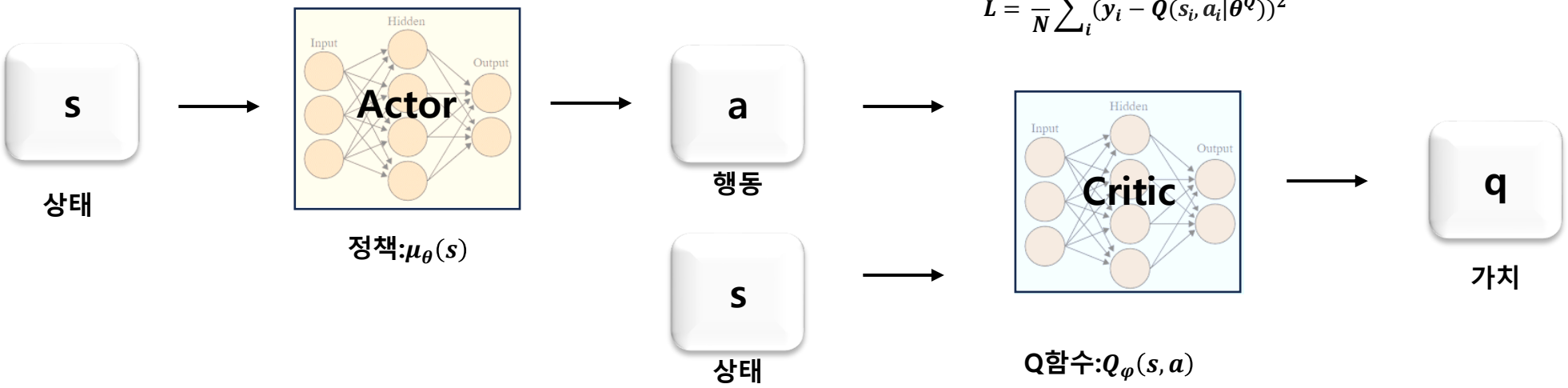
DDPG
Deep Deterministic Policy Gradient

❖ DDPG (Deep Deterministic Policy Gradient)

- Deep Neural Network로 근사화한 action-value function을 사용하는 **Off-policy actor-critic 알고리즘**
- Actor는 상태 s에서 행동을 결정 ,Critic은 상태 s에서 Actor가 취한 Action에 대한 행동가치를 평가

$$\nabla_{\theta^\mu} J = \frac{1}{N}\sum_i \nabla_a Q(s_i, a_i|\theta^Q)\nabla_{\theta^\mu}\mu(s|\theta^\mu)$$

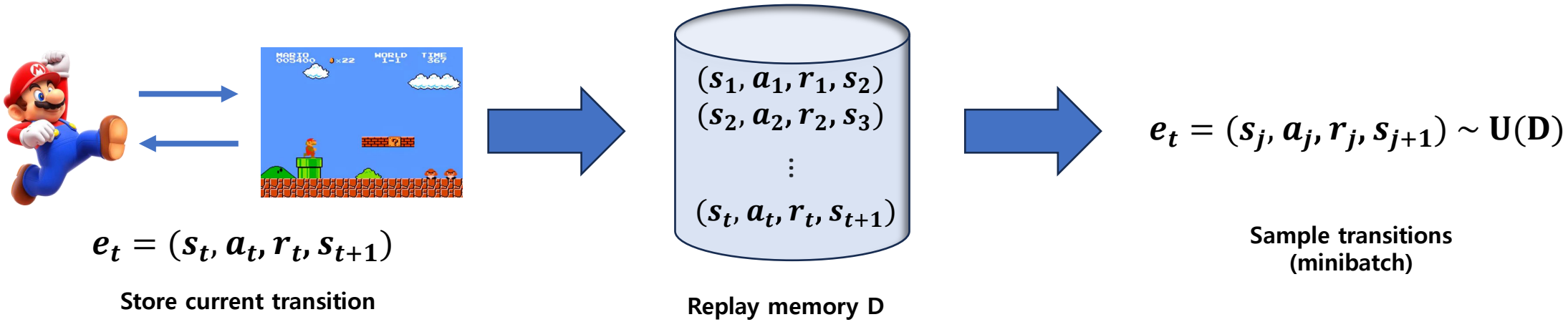
$$L = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$$



DDPG
Properties of DDPG

❖ Replay Buffer

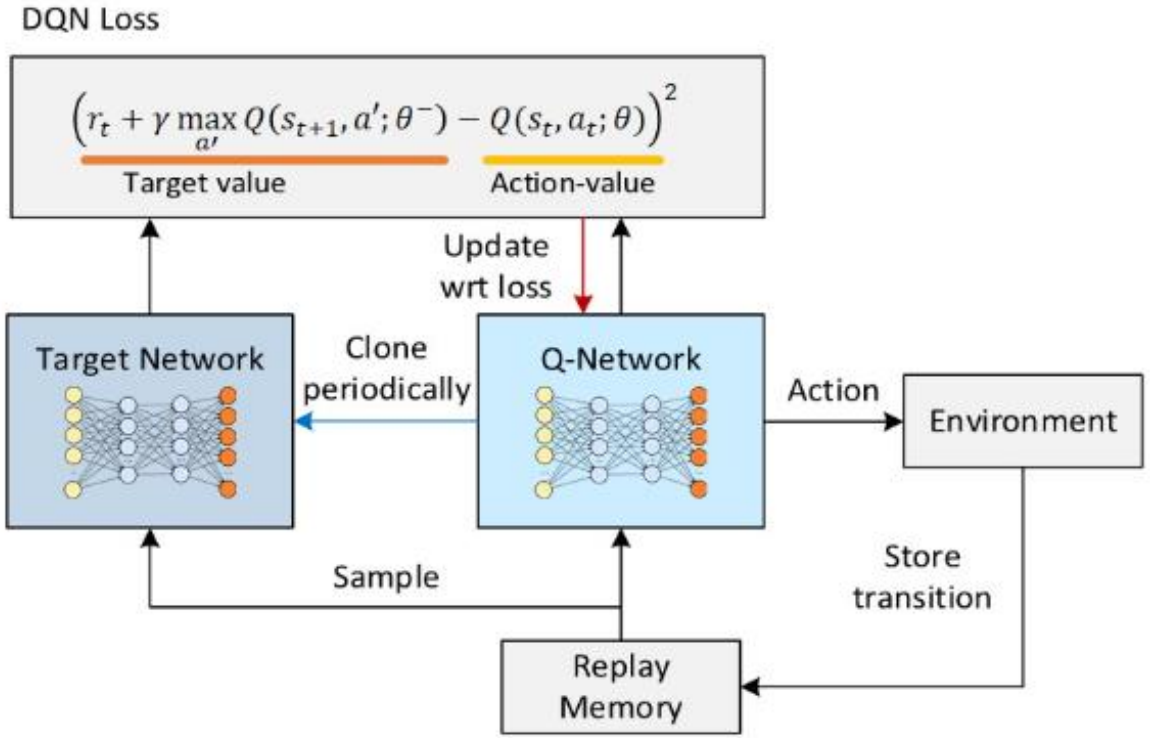
- 매 스탭마다 샘플을 Replay Buffer에 저장하고 랜덤하게 minibatch만큼 추출하여 학습에 이용
- DDPG는 **Off-policy 알고리즘**이기 때문에 이전 학습자료를 사용해 학습 할 수 있음
- 연속된 샘플 데이터 간의 상관성이 줄여 **Dependency 문제**를 해결할 수 있음
- 학습 분포의 급격한 변화를 줄여 **안정적 학습을** 가능하게 함



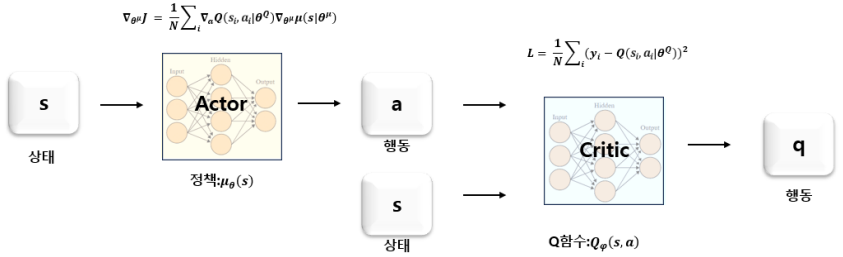
DDPG
Properties of DDPG

❖ Target Network

- 안정적인 학습을 진행하기 위해서 Target network를 도입
- N-step 마다 Q-Network를 복사하여 Target Network 업데이트 하는 방식으로 진행

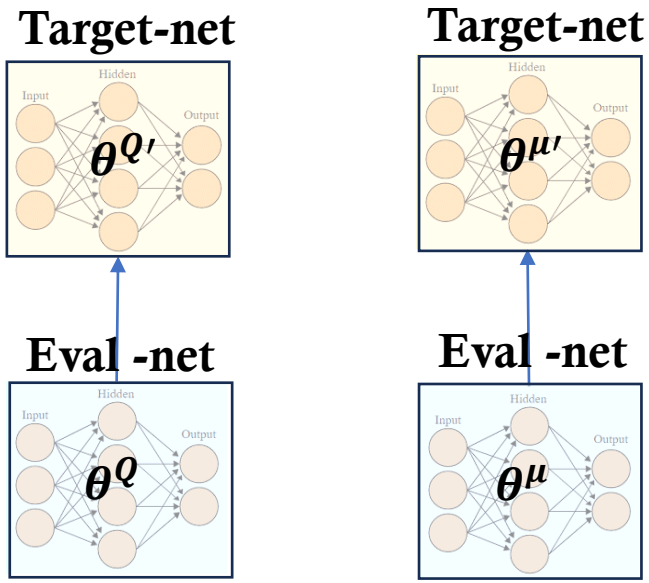


DDPG
Properties of DDPG



❖ Tagret network & Soft update

- Actor와 Critic 모두 target network를 구성하여 **target net과 eval net** 네트워크를 구성
- DQN에서 사용된 T Step 마다 한번에 업데이트 하는 것이 아닌 모든 스탭 마다 조금씩 업데이트는 **soft-update**를 사용
- Soft update로 인해 학습이 느리지만 안정적(robust)으로 학습



$$\theta^{Q\prime} \leftarrow \theta^{Q}$$

DQN update

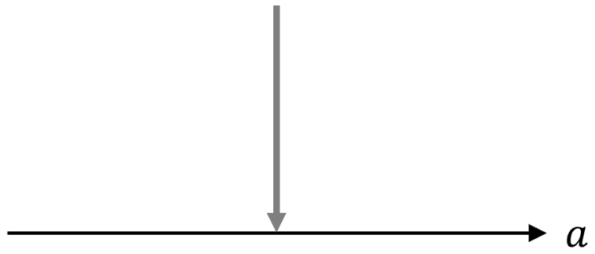
$$\theta^{Q\prime} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q\prime}$$
$$\theta^{\mu\prime} \leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu\prime}$$

Soft update

DDPG
Properties of DDPG

❖ OUNoise

- 행동에 노이즈를 추가하는 방법을 사용하여 **탐험 문제(exploration problem)**를 해결
- 안정적인 무작위성을 증가시키기 위해 **시간적 연관성을 가진 노이즈를** 추가

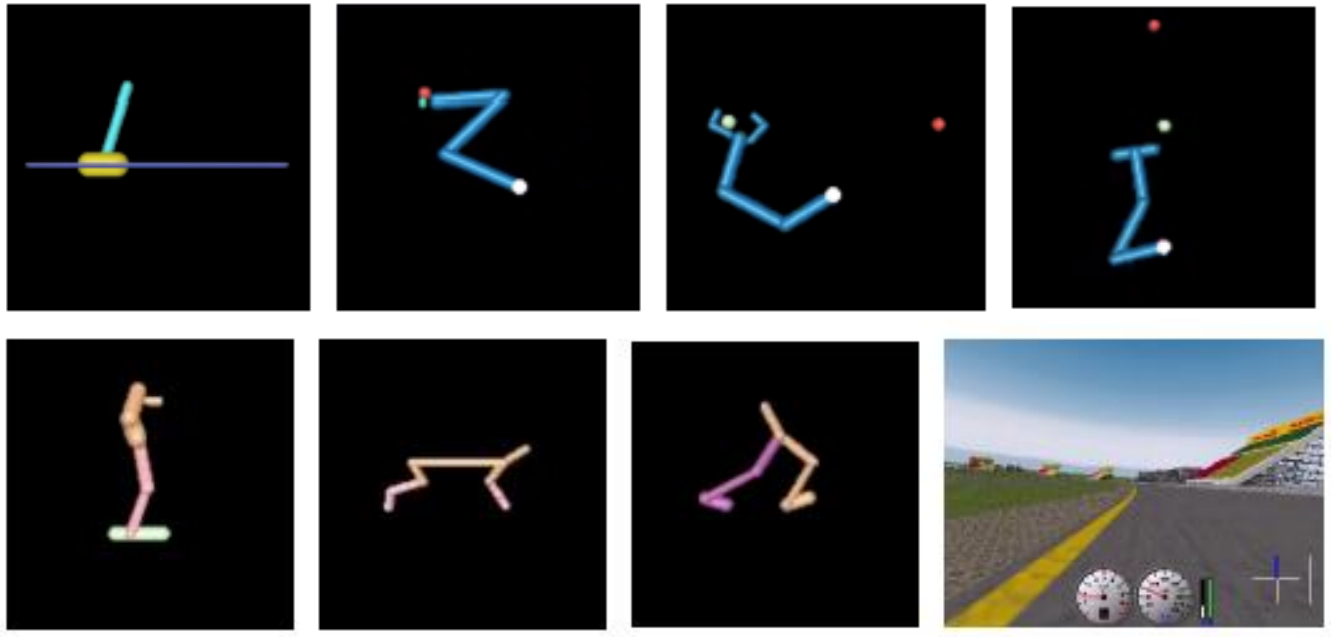


$$a_t = \mu(s_t|\theta^\mu) + N_t(noise),$$
$$N_t = N_{t-1} + \theta(\mu - N_{t-1}) + \sigma r, \qquad r \sim Normal(0, 1)$$

DDPG
Experiment

❖ **Experiment**

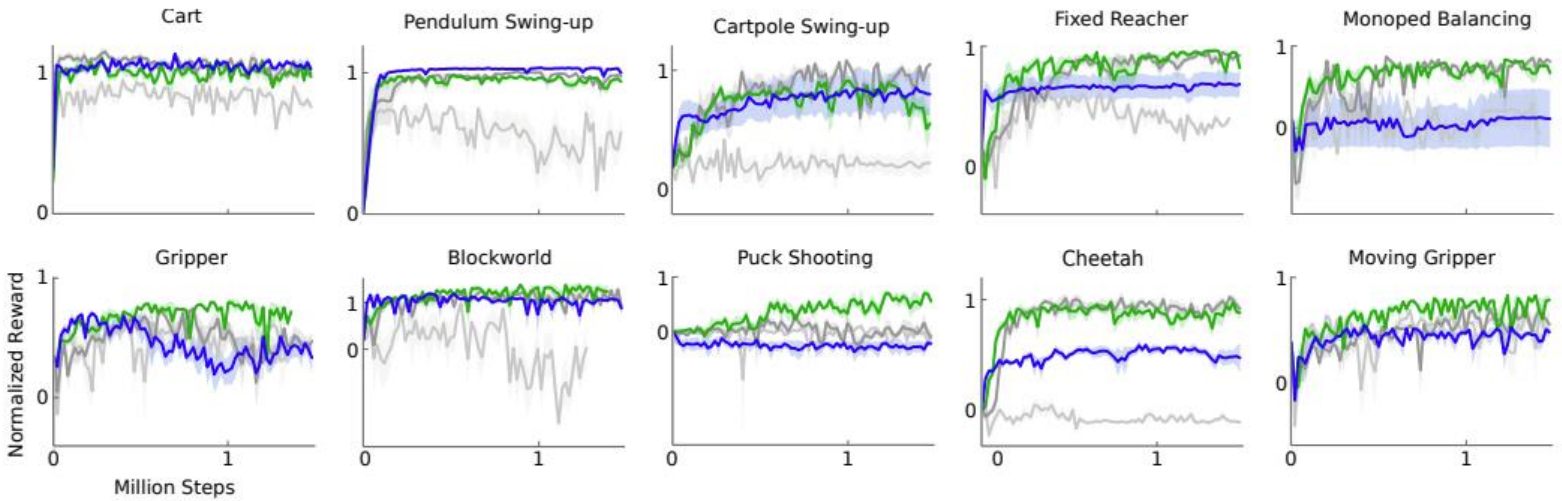
- 높은 차원의 상태공간을 가진 8개의 환경에서 실험 진행
- 관절 정보 뿐만 아니라 이미지 픽셀단위 정보만 가지고 학습을 진행



DDPG
Experiment

❖ Experiment

- 높은 차원의 공간과 연속행동공간에서의 학습을 처음으로 성공
- 운동량, 각도등의 추가적인 정보 없이 pixel로만 이루어진 이미지에서 성공적인 학습이 가능

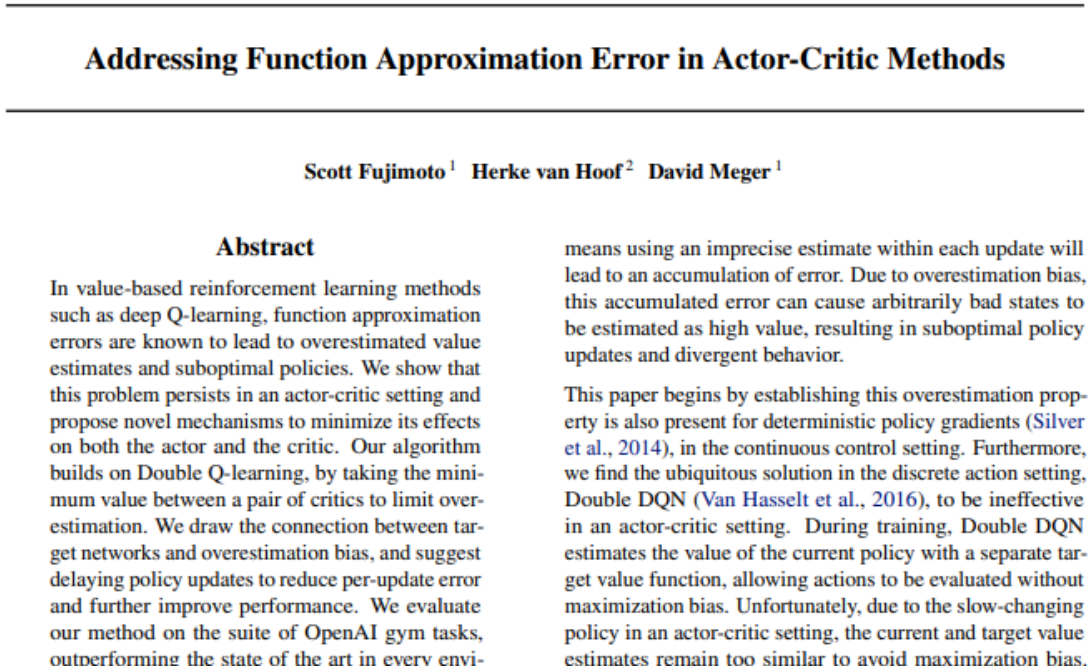


- original DPG with batch normalization
- original DPG with targetnetwork
- original DPG with batch normalization, targetnetwork
- original DPG with batch normalization, targetnetwork+ pixel-only inputs

TD3
Twin Delayed Deep Deterministic Policy Gradient

❖ **TD3 : Addressing Function Approximation Error in Actor-Critic Methods (2018, ICML)**

- DDPG 알고리즘의 Q-value의 **overestimate** 문제를 해결하기 위해 여러 trick을 도입
- Q-value의 Overestimation을 문제를 효과적으로 개선하여 높은 성능을 나타냄



Addressing Function Approximation Error in Actor-Critic Methods

Scott Fujimoto [1] Herke van Hoof [2] David Meger [1]

Abstract

In value-based reinforcement learning methods such as deep Q-learning, function approximation errors are known to lead to overestimated value estimates and suboptimal policies. We show that this problem persists in an actor-critic setting and propose novel mechanisms to minimize its effects on both the actor and the critic. Our algorithm builds on Double Q-learning, by taking the minimum value between a pair of critics to limit overestimation. We draw the connection between target networks and overestimation bias, and suggest delaying policy updates to reduce per-update error and further improve performance. We evaluate our method on the suite of OpenAI gym tasks, outperforming the state of the art in every envi-

means using an imprecise estimate within each update will lead to an accumulation of error. Due to overestimation bias, this accumulated error can cause arbitrarily bad states to be estimated as high value, resulting in suboptimal policy updates and divergent behavior.

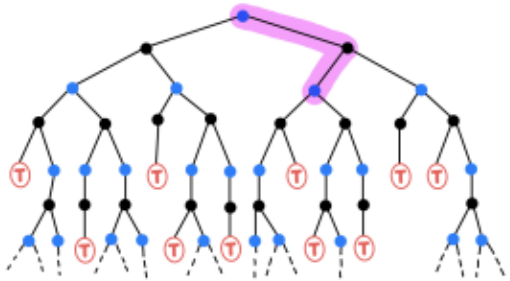
This paper begins by establishing this overestimation property is also present for deterministic policy gradients (Silver et al., 2014), in the continuous control setting. Furthermore, we find the ubiquitous solution in the discrete action setting, Double DQN (Van Hasselt et al., 2016), to be ineffective in an actor-critic setting. During training, Double DQN estimates the value of the current policy with a separate target value function, allowing actions to be evaluated without maximization bias. Unfortunately, due to the slow-changing policy in an actor-critic setting, the current and target value estimates remain too similar to avoid maximization bias.

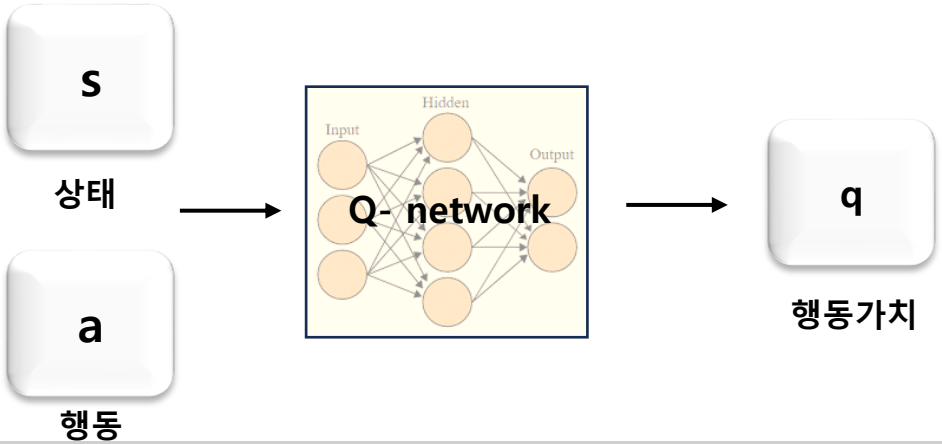
TD3
Twin Delayed Deep Deterministic Policy Gradient

❖ Q-Value의 과대평가 문제

- Q-Value를 추정할 때 뉴럴 네트워크를 통해 함수로 근사하기 때문에 발생하는 TD에러로 필연적인 노이즈가 발생
- 발생한 노이즈는 MAX 연산자로 인해 일관되게 과대 추정이 발생

Temporal Difference(TD)





$$max_{a'}(Q(s',a')+\varepsilon)]$$

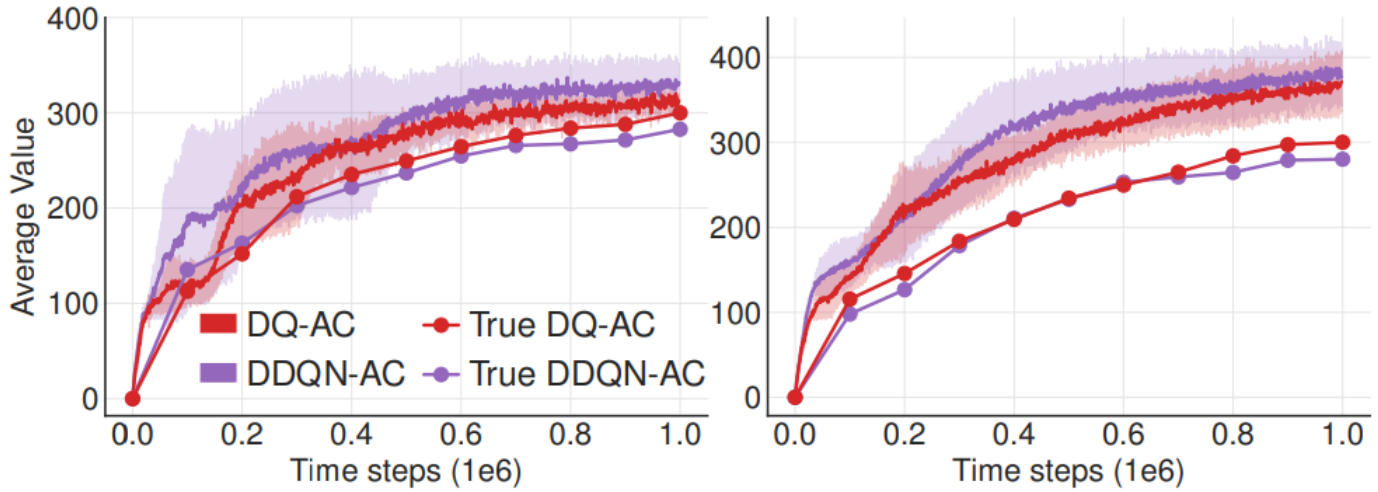
$$E[max_{a'}(Q(s',a')+\varepsilon)] \geq max_{a'}Q(s',a')$$

$$E[max_{a'}(Q(s',a')+\varepsilon)] > max_{a'}Q(s',a')$$

TD3
Twin Delayed Deep Deterministic Policy Gradient

❖ Q-Value의 overestimation 문제

- Q-Value 과대평가문제 발생함을 실험적으로 보여줌
- Actor-critic 환경에서 적용한 DQ-AC와 DDQN-AC가 큰 효과가 없음

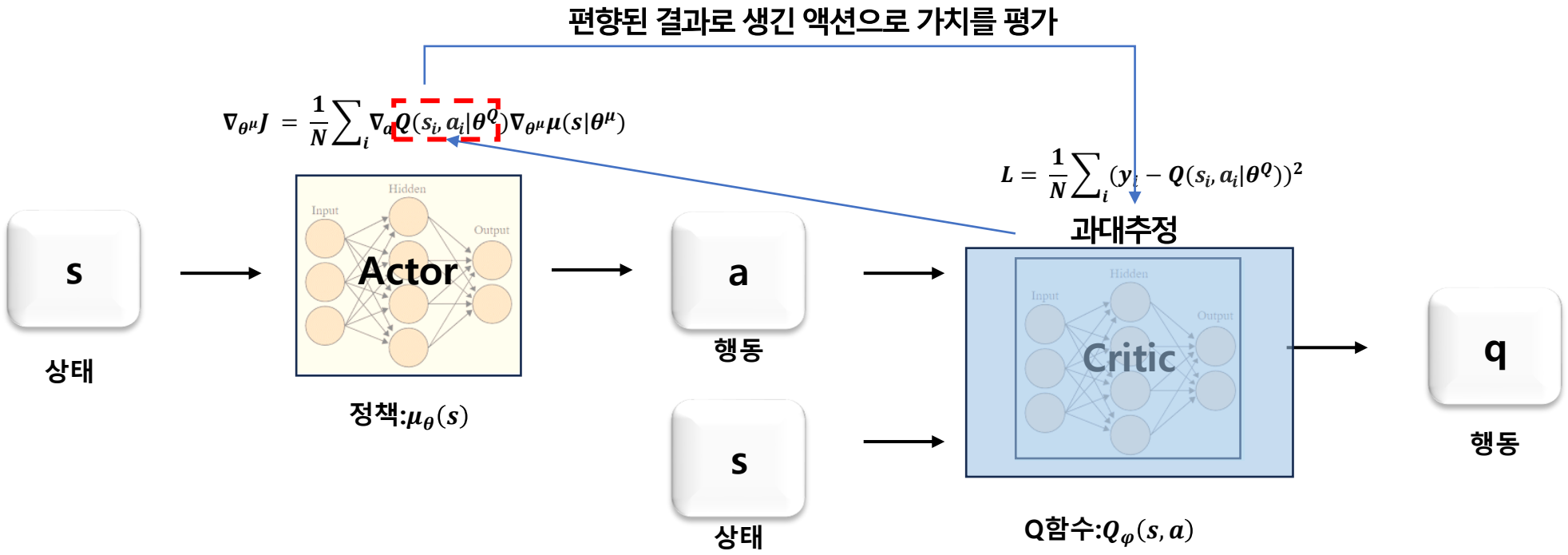


(a) Hopper-v1 (b) Walker2d-v1

TD3
Twin Delayed Deep Deterministic Policy Gradient

❖ **Q-Value의 overestimation 문제**

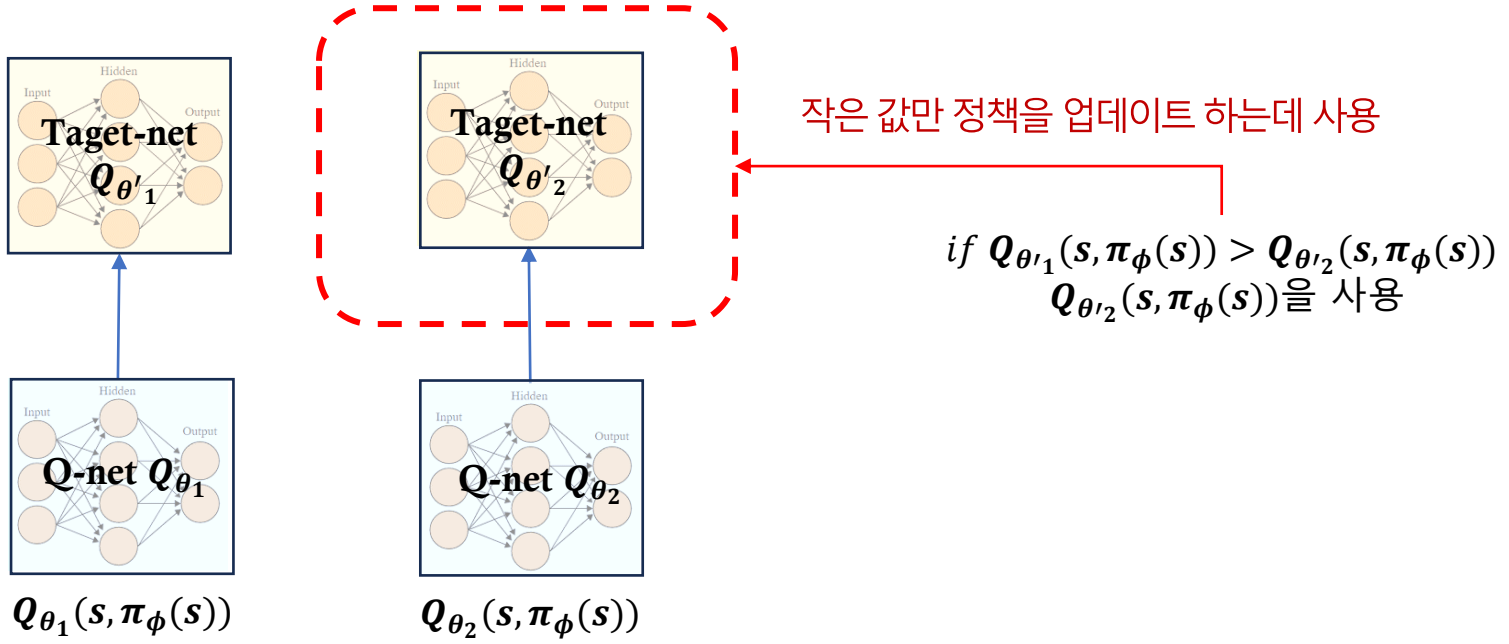
1. 부정확한 추정치로 엑터 네트워크의 학습을 어렵게 만듦
2. 과대평가가 일어나면 더 큰 편향을 일어날 가능성이 높아짐



TD3
Properties of TD3

❖ Clipped double Q-learning

- 2개의 critic 네트워크를 이용해 Q-value update시 **작은 값을 학습에 이용**
- 두 개의 Critic 네트워크를 사용하고, 각각의 **독립적인 타겟 네트워크를** 생성하여 업데이트
- Critic 네트워크는 Clipping된 타겟 Q-value에 대한 추정치를 사용하여 훈련되어 불안정성을 감소

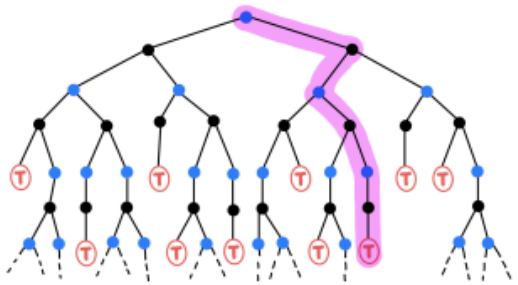


TD3
Properties of TD3

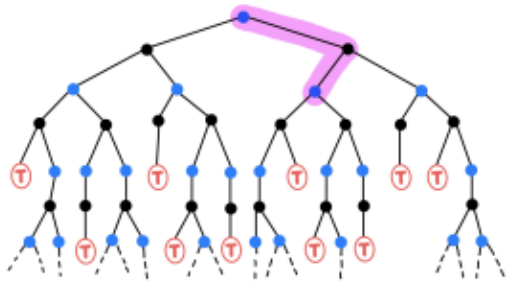
❖ Accumulate error

- 1 step씩 업데이트 하는 TD-update를 진행하기 때문에 가치 함수 추정하여 사용하기 때문에 오류가 누적
- 추정치의 분산은 미래 보상과 추정 오류의 분산에 비례하게 발생

Monte Carlo(MC)



Temporal Difference(TD)



1번 추정

TD 에러

$$Q_\theta(s,a) = r + \gamma E[Q_\theta(s',a')] + \delta_{(s,a)}.$$

n번 추정

TD 에러가 누적됨

$$Q_\theta(s_t a_t) = r_t + \gamma E[Q_\theta(s_{t+1}, a_{t+1}] - \delta_t,$$
$$= r_t + \gamma E[r_{t+1} + \gamma E[Q_\theta(s_{t+2}, a_{t+2}) - \delta_{t+1}]] - \delta_t,$$
$$= E_{s_i \sim p\,\pi,\, a_i \sim \pi} \sum_{i=t}^{T} [\gamma^i (r_i - \delta_i)]$$

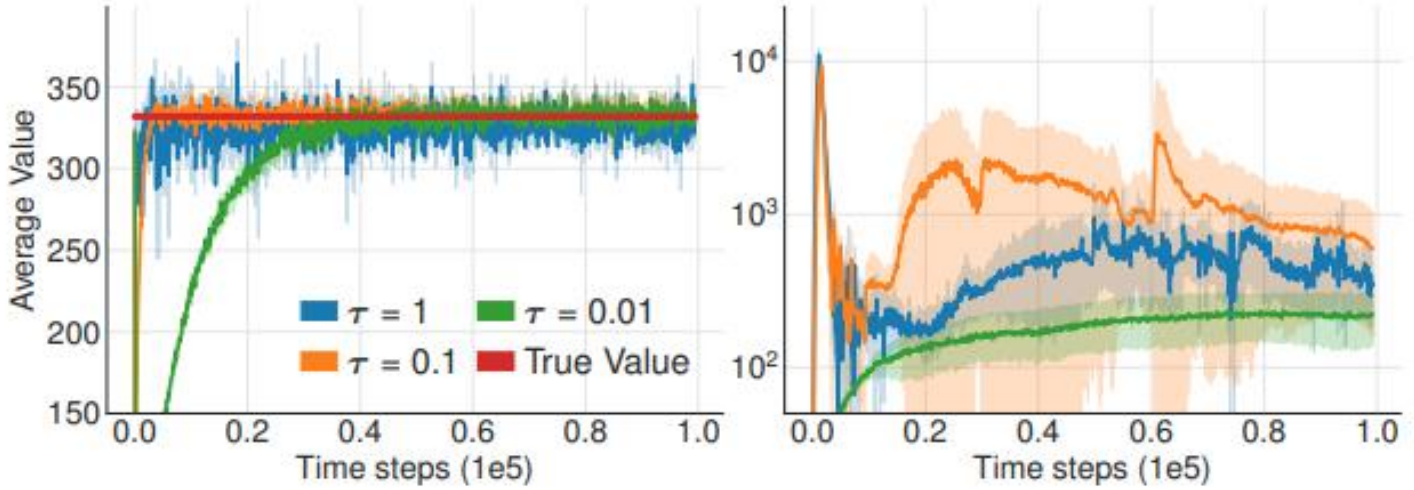
추정 오류의 분산이 추정치에 오류 비례하여 영향을 줌

각 업데이트에서 오류를 최소화 하는 것이 매우 중요

TD3
Properties of TD3

❖ **Delayed Policy update**

- Target 네트워크의 사용이 학습의 안정성을 부여해 variance를 낮추는 역할
- 타겟 네트워크가 급격히 업데이트 된다면 분산이 높아지는 문제 발생



(a) Fixed Policy (b) Learned Policy

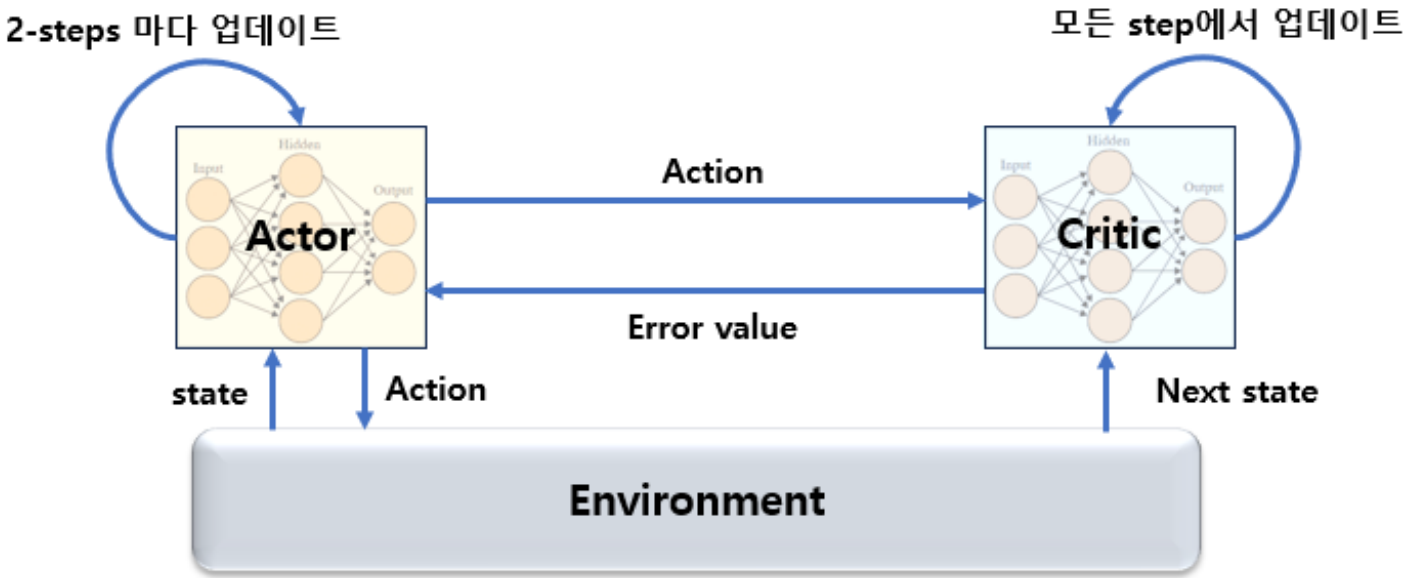
$$\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'}$$
$$\theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu'}$$

Soft update

TD3
Properties of TD3

❖ **Delayed Policy update**

- 정책 네트워크의 업데이트를 지연시키는 것을 통해 특정 시간 간격 동안 정책 네트워크가 업데이트되는 것을 방지
- 정책 네트워크의 지연으로 학습의 분산을 줄여 안정적인 훈련을 촉진하고 높은 성능을 달성하는 데 기여

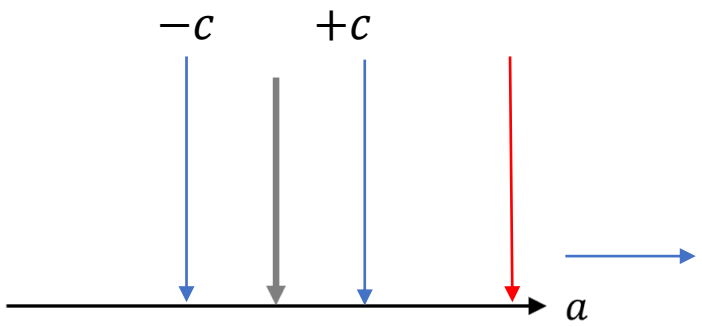


TD3
Properties of TD3

❖ **Target policy smoothing**

- 가까이 액션은 비슷한 Q-값을 가질거라는 가정하에 진행
- 추가된 노이즈는 클리핑하여 Q-값이 급격하게 변하지 않도록 방지

$$\varepsilon \sim clip(N(0,\sigma), -c, c)$$



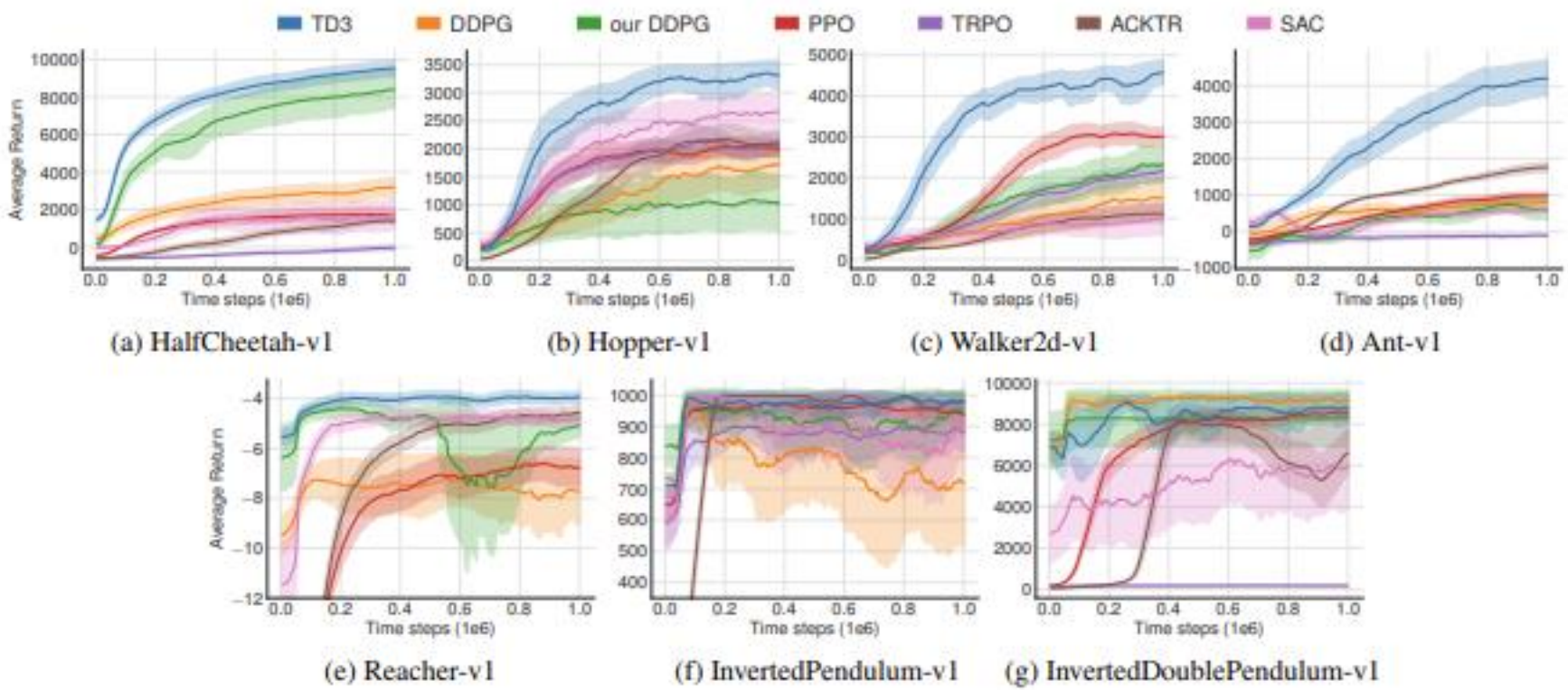
가까운 행동이 아닌 거리가 먼 행동은 Q-value의 많은 변화가 일어남

$$\boldsymbol{Deterministic}$$
$$\boldsymbol{\mu_\theta(s)}$$

TD3
Experiments

❖ Experiments

- TD3가 다른 알고리즘들에 비해 빠른 수렴을 보임
- 수렴한 average return의 값이 다른 Actor-Critic방법론들에 비해 높은 성능을 보임



Conclusion
Deterministic Policy Gradient

❖ **DDPG(Deep Deterministic Policy Gradient)**

- DPG 알고리즘을 Deep Neural Network에 적용한 연구
- continuous action space에 강화학습 알고리즘을 적용

❖ **TD3(Twin Delayed Deep Deterministic Policy Gradient)**

- TD3는 DDPG의 Q-value의 과대추정 문제를 개선하였음
- 다양한 Actor-Critic 알고리즘 중에도 높은 성능을 나타냄

고맙습니다