강화학습 실습

목차

❖ Introduction

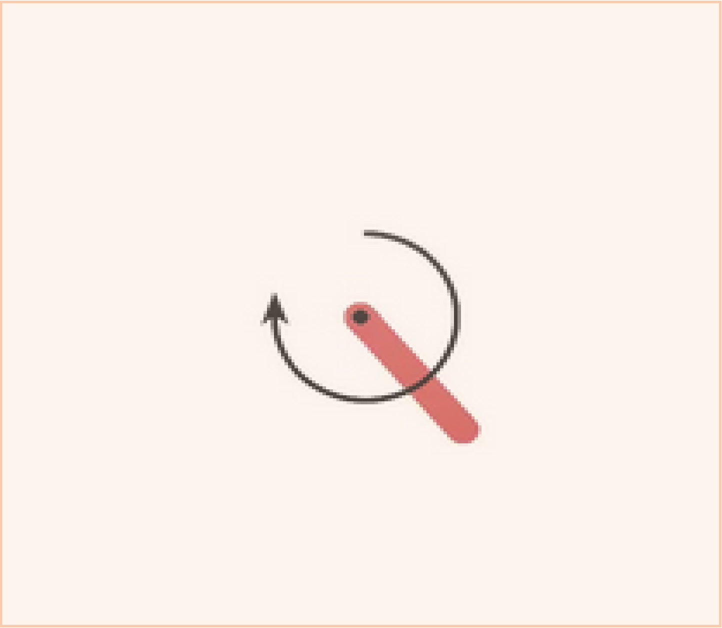
❖ Deep Deterministic Policy Gradient(DDPG)

강화 학습 실습

Jupyter Notebook 실습

❖ Introduction

- Pendulum
  a. 입력 상태: 회전 각에 따른 막대기의 상태(cos(theta), sin(theta), theta dot)
  b. 행동: cos(theta), sin(theta)를 고려한 행동



- Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... & Wierstra, D. (2015). Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.

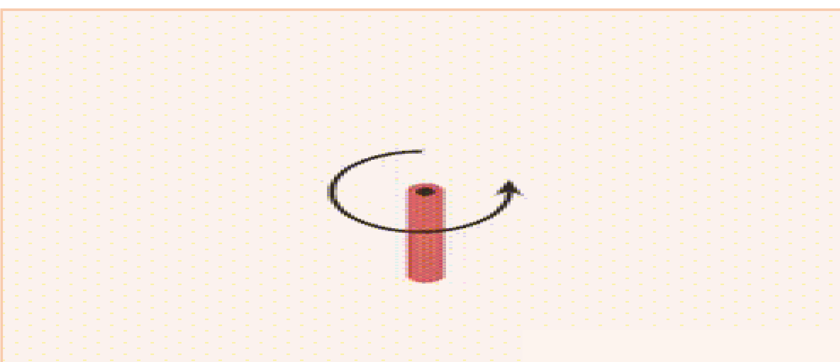
강화 학습 실습

Jupyter Notebook 실습

❖ Introduction

- Pendulum
  a. Pendulum 게임에 대한 상세 설명
  b. 관측 상태, 행동, 보상 등 상세 설명

게임의 특성



게임의 목표

- Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... & Wierstra, D. (2015). Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.

강화 학습 실습

Jupyter Notebook 실습

❖ Introduction

- Pendulum

  a. Pendulum 게임에 대한 상세 설명

  b. 관측 상태, 행동, 보상 등 상세 설명

- **Observation:** [cos(theta), sin(theta), theta dot]
  - Theta : $(-\pi, \pi)$
- **Ending condition(of episode)**
  - No specified termination
  - Adding a maximum number of steps
- **Action:** 회전력(-2,2)
- **Reward:** $-(\pi^2 + 0.1 \times 8^2 + 0.001 \times 2^2)$
- **Objective:** 제한된 timestep안에 보상을 0으로(pole의 균형을 수직)

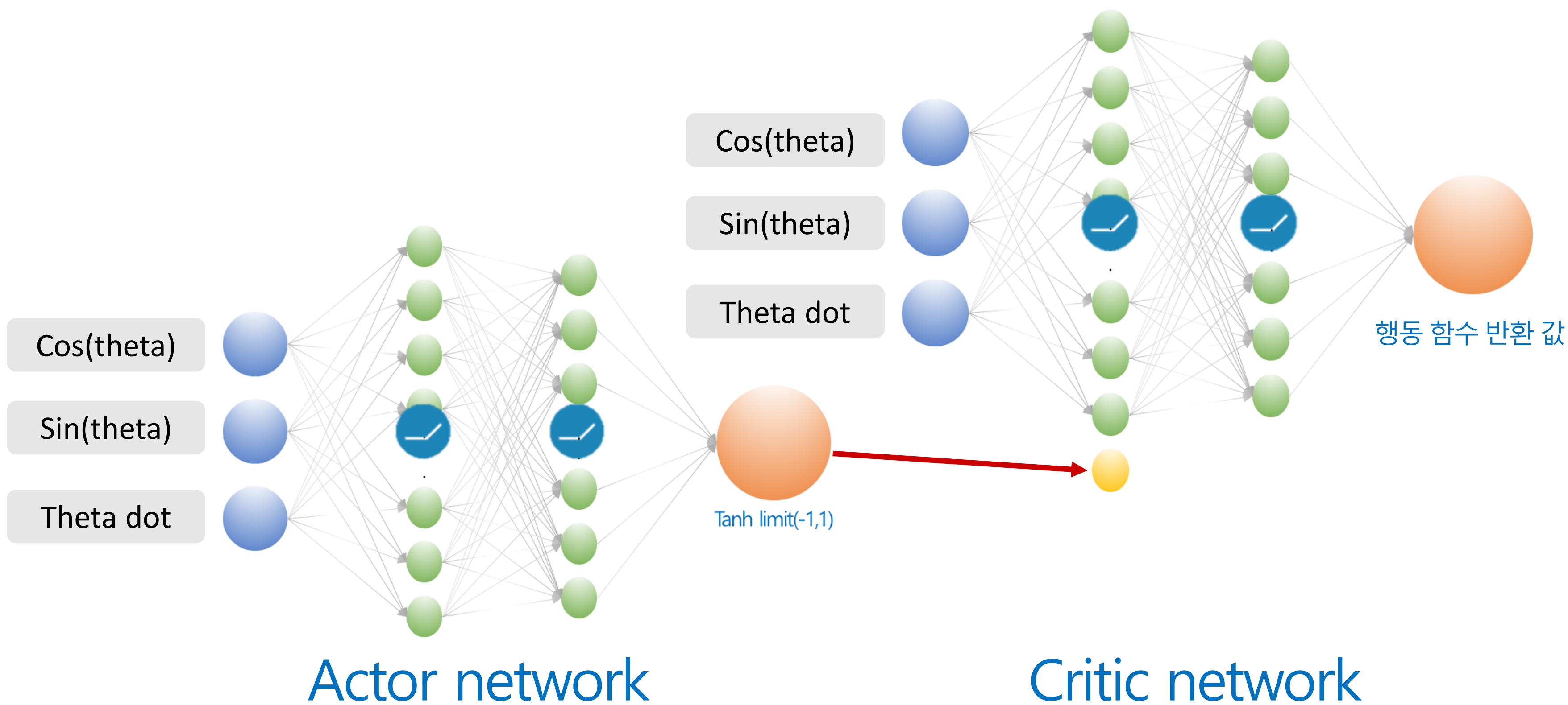
- Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... & Wierstra, D. (2015). Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.

강화 학습 실습

Jupyter Notebook 실습

❖ Introduction

- Deep Deterministic Policy Gradient(DDPG) 네트워크 구조



- Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... & Wierstra, D. (2015). Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.

DMQA

강화 학습 실습

Jupyter Notebook 실습

❖ Deep Deterministic Policy Gradient

- Weight update

  a. Critic network(train)의 손실 함수를 최소화하는 방향으로 가중치를 업데이트

  b. Actor network(train)를 학습시킬 때 Critic network(train)에서 업데이트한 Q값의 가중치를 포함하여 정책 함수의 가중치를 업데이트

$$\textit{Set target}:\ y_i = r_i + \gamma\, Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$$

$$\text{Update critic by minimizing the loss: } \mathcal{L} = \frac{1}{N}\sum_i \left(y_i - Q(s_i, a_i|\theta^Q)\right)^2$$

$$\text{Update the actor policy using the sample policy gradient: } \nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_i \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu}\mu(s|\theta^\mu)|_{s_i}$$

$$\text{Update the target networks: } \begin{aligned} \theta^{Q'} &\leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'} \\ \theta^{\mu'} &\leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'} \end{aligned}$$

- Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... & Wierstra, D. (2015). Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.

감사합니다