Difference between On-policy, Off-policy methods ; Value gradients vs Policy gradients

2021 Summer Internship Program

202011875 김현우

Flow

1. *Understanding RL Structure*

2. *Value Gradient advanced examples*

3. Policy Gradient advanced examples

Contents

1. DQN

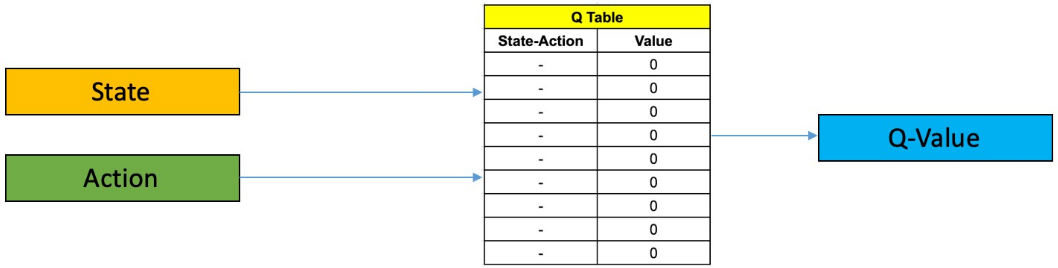
2. Policy Gradient ; REINFORCE, Actor-Critic

3. CartPole-v1 ; Value gradient vs Policy gradient

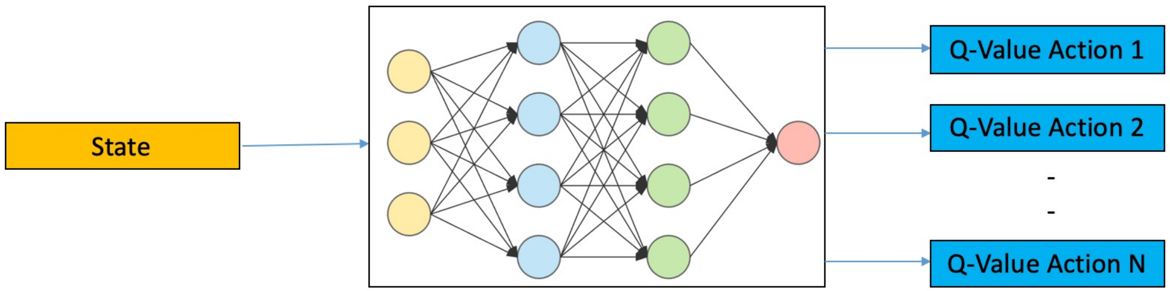
4. Pendulum-v0 ; On-policy vs Off-policy

5. Further Study

DQN



Q Learning



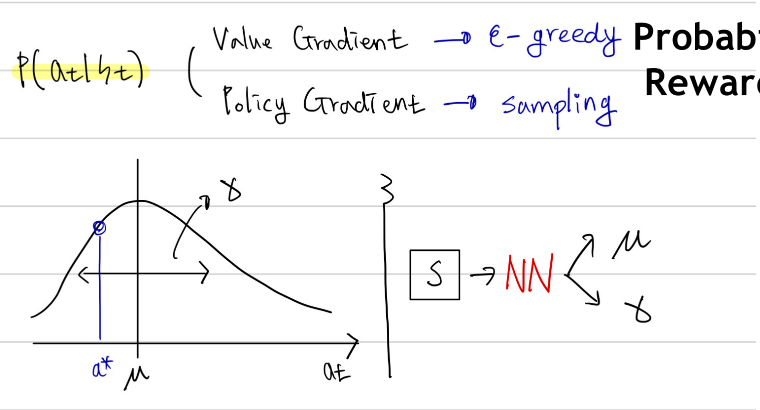
Deep Q Learning

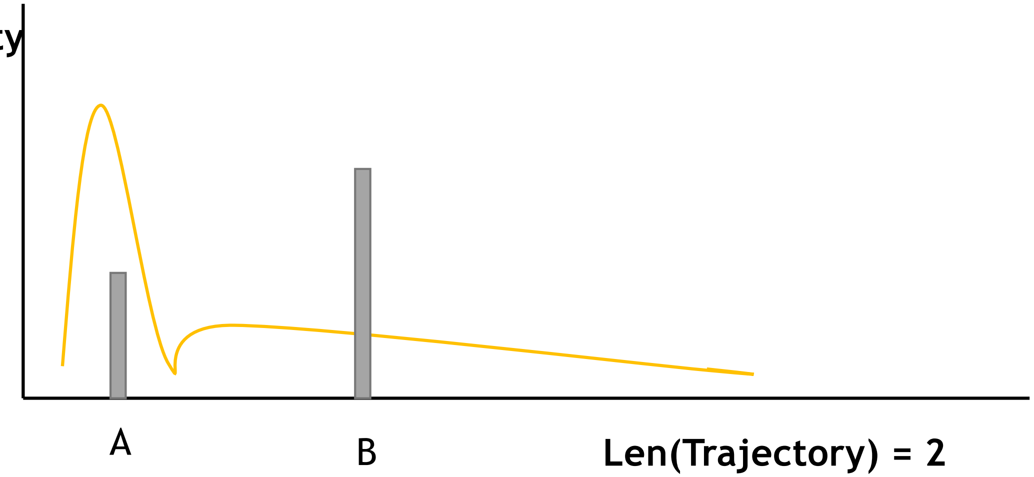
DQN

Algorithm 1 Deep Q-learning with Experience Replay

Initialize replay memory $\mathcal{D}$ to capacity N
Initialize action-value function Q with random weights
for episode $= 1, M$ **do**
  Initialise sequence $s_1 = \{x_1\}$ and preprocessed sequenced $\phi_1 = \phi(s_1)$
  for $t = 1, T$ **do**
    With probability $\boxed{\epsilon}$ select a random action a_t
    otherwise select $a_t = \boxed{\max_a Q^*(\phi(s_t), a; \theta)}$
    Execute action a_t in emulator and observe reward r_t and image x_{t+1}
    Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$
    Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in $\mathcal{D}$
    Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from $\mathcal{D}$
    Set $y_j = \begin{cases} r_j & \text{for terminal } \phi_{j+1} \\ r_j + \gamma \boxed{\max_{a'}} Q(\phi_{j+1}, \boxed{a'}; \theta) & \text{for non-terminal } \phi_{j+1} \end{cases}$
    Perform a gradient descent step on $(y_j - Q(\phi_j, \boxed{a_j}; \theta))^2$ according to equation $\boxed{3}$
  end for
end for

Policy Gradient





REINFORCE

① $J_\theta = E[G_0] = \int_\tau G_0 \cdot P_\theta(\tau)\, d\tau$

Find policy maximizes J_θ .... Gradient ascent

expectation form

② $\nabla_\theta J_\theta = \nabla_\theta \int_\tau G_0 \cdot P_\theta(\tau)\, d\tau = \int_\tau G_0 \nabla_\theta P_\theta(\tau)\, d\tau = \int_\tau G_0 \cdot P_\theta(\tau) \cdot \nabla_\theta \ln P_\theta(\tau)\, d\tau$

$\nabla_\theta P_\theta(\tau) = P_\theta(\tau) \cdot \nabla_\theta \cdot \ln P_\theta(\tau)$ .... $\nabla_\theta \ln P_\theta(\tau) = \frac{\nabla_\theta P_\theta(\tau)}{P_\theta(\tau)}$

③ $\nabla_\theta J_\theta = \int_\tau G_0 P_\theta(\tau) \cdot \nabla_\theta \sum \ln P_\theta(a_t|s_t)\, d\tau$

$P_\theta(\tau) = \cancel{P(s_0)} \times P_\theta(a_0|s_0) \times \cancel{P(s_1|s_0, a_0)} \times P_\theta(a_1|s_1) \times \cdots$ Bayesian, Log rule

④ $\nabla_\theta J_\theta = \int_\tau \sum_{t=0}^{\infty} \{ \nabla_\theta \ln P_\theta(a_t|s_t) \times \sum_{k=t}^{\infty} \gamma^k \cdot R_k \} P_\theta(\tau)\, d\tau$

$G_0 \times \nabla_\theta \sum \ln P_\theta(a_t|s_t) = \{ R_0 + \gamma R_1 + \gamma^2 R_2 + \cdots \} \times \nabla_\theta \{ \ln P_\theta(a_0|s_0) + \ln P_\theta(a_1|s_1) + \cdots \}$

$\int_\tau \{ P_\theta(\tau) \times (R_0 \times \ln P_\theta(a_1|s_1)) + \cdots \} d\tau = \int_{\tau - a_1} R_0 \int_{a_1} \nabla_\theta \ln P_\theta(a_1|s_1)\, da_1\, P(\tau - a_1)\, d\tau\text{-}a_1$

$\exists\ \nabla_\theta \cdot \int_{a_1} P_\theta(a_1|s_1)\, da_1 = \nabla_\theta \cdot 1 = 0$

⑤ $\nabla_\theta J_\theta = \cancel{\int_\tau} \sum_{t=0}^{\infty} (\nabla_\theta \ln P_\theta(a_t|s_t) \cdot G_t) P_\theta(\tau)\, d\tau$

$\int x \cdot p(x)\, dx \approx \frac{1}{N} \cdot \sum_{i=1}^{N} x_i$ (Sample mean)

$(N=1) \rightarrow$ REINFORCE

Actor-Critic

① $\nabla_\theta J_\theta = \int_\tau \sum_{t=0}^{\infty} \nabla_\theta \ln P_\theta(a_t|s_t) G_t P_\theta(\tau) d\tau$

$P_\theta(\tau) = P_\theta(s_{t+1}, a_{t+1}, \cdots | s_t, a_t) \times P(s_0, a_0, \cdots, s_t, a_t)$

② $\nabla_\theta J_\theta = \int \sum_{t=0}^{\infty} \nabla_\theta \ln P_\theta(a_t|s_t) \int G_t P_\theta(s_{t+1}, a_{t+1} \cdots | s_t, a_t) d s_{t+1}, \cdots \times P_\theta(s_0, \cdots, s_t, a_t) d s_0, \cdots, a_t$

$Q(s_t, a_t) = \int_{s_{t+1}:a_\infty} G_t \cdot P_\theta(s_{t+1}, a_{t+1} \cdots | s_t, a_t) d s_{t+1}:a_\infty$

③ $\nabla_\theta J_\theta = \sum_{t=0}^{\infty} \int_{s_t, a_t} \nabla_\theta \ln P_\theta(a_t|s_t) Q(s_t, a_t) P_\theta(a_t|s_t) d s_t, a_t$

$\Rightarrow$ Return $\rightarrow$ "Q" : Episodic Free $\Rightarrow$ TD 계열

$\theta \leftarrow \theta + \nabla_\theta \ln P_\theta(a_t|s_t) Q(s_t, a_t)$

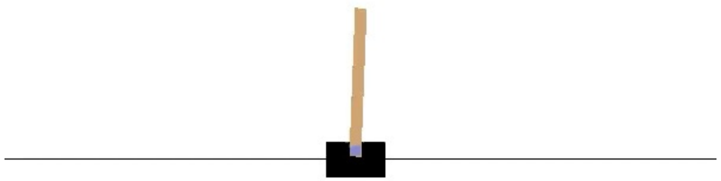
$w \leftarrow w - \beta \nabla_w \left(R_0 + \gamma Q_w(s_{t+1}, a_{t+1}) - Q_w(s_t, a_t)\right)^2$

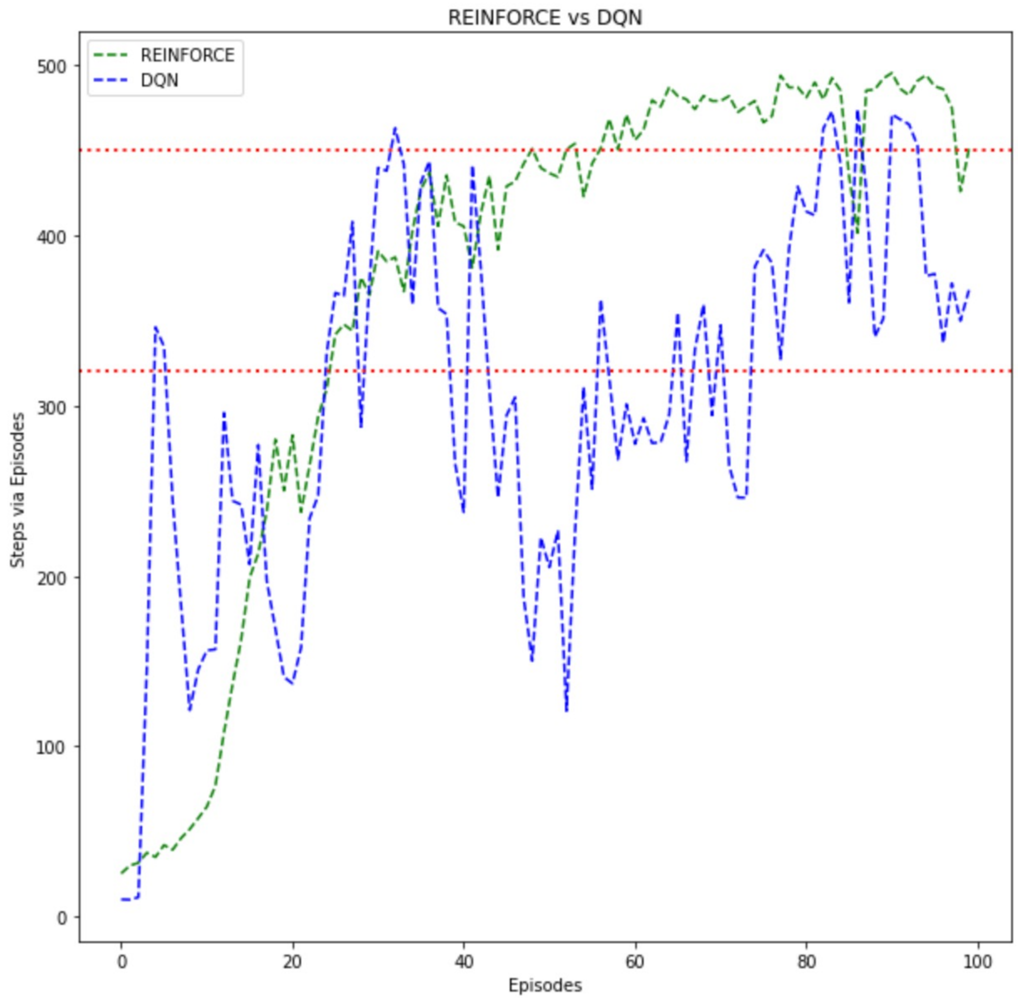
CartPole-v1

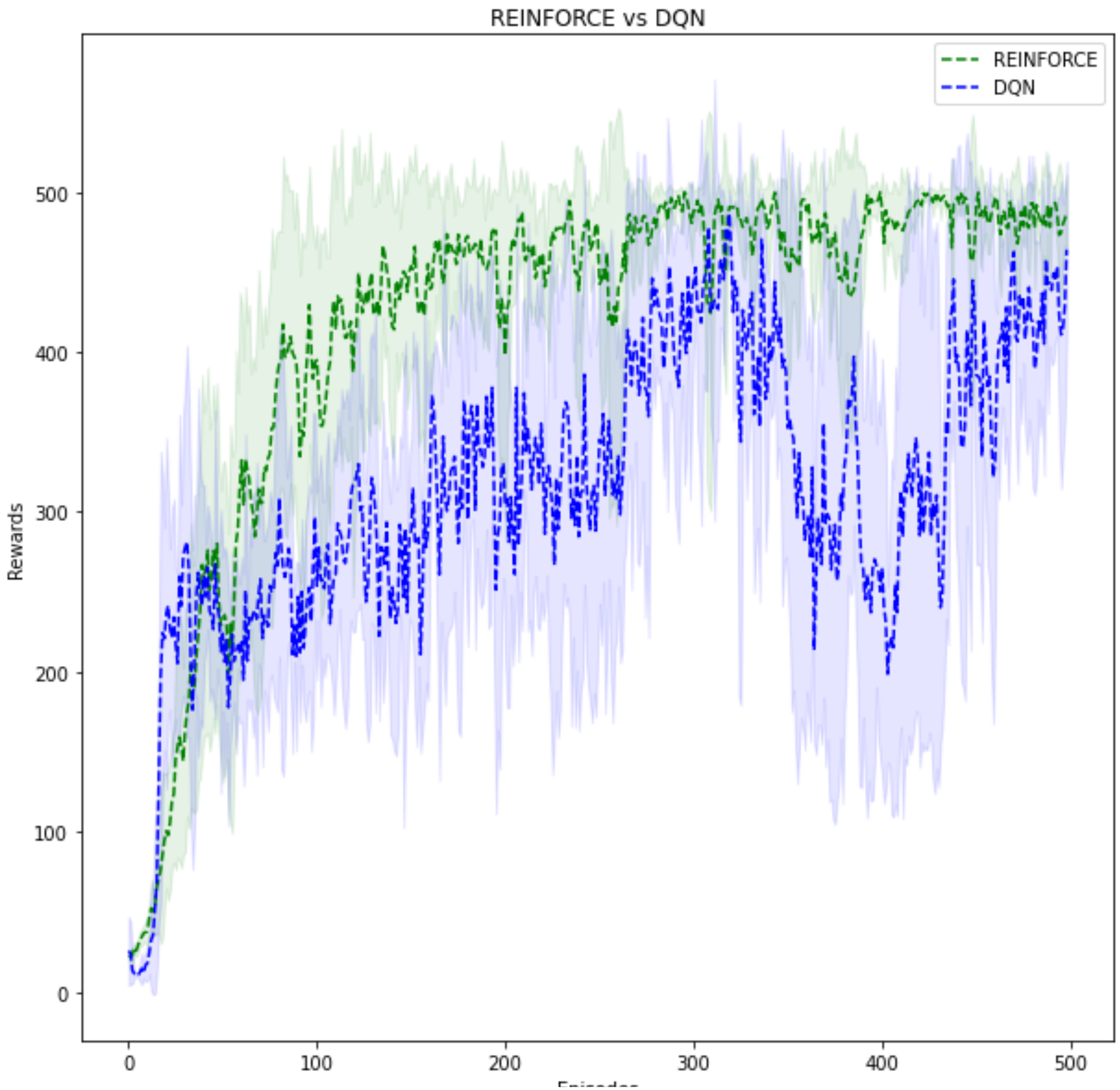
$[Cart\ Position, Cart\ Velocity, Pole\ Angle, Pole\ Angular\ Velocity]$

$[Push\ cart\ to\ the\ left, Push\ cart\ to\ the\ right]$

$= to\ make\ pole\ remain\ upright$




REINFORCE vs DQN
REINFORCE
DQN
Steps via Episodes
Episodes
0
100
200
300
400
500
0
20
40
60
80
100


REINFORCE vs DQN
REINFORCE
DQN
Rewards
0
100
200
300
400
500
0
100
200
300
400
500

A2C

IDEA > Q-function 자리에 임의의 state function을 대입한다면?

PROOF > $\nabla_\theta J_\theta \simeq \int_{s_t,a_t} \nabla_\theta \ln p_\theta(a_t|s_t)\, Q(s_t,a_t)\, p_\theta(s_t,a_t)\, ds_t,a_t$

$\ni \int_{s_t,a_t} \nabla_\theta \ln p_\theta(a_t|s_t) \cdot A \cdot p_\theta(a_t|s_t)\, P(s_t)\, ds_t,a_t$

v $\nabla_\theta \ln p_\theta(a_t|s_t) = \nabla_\theta p_\theta(a_t|s_t) / p_\theta(a_t|s_t)$

$= \int_{s_t} \nabla_\theta \int_{a_t} p_\theta(a_t|s_t)\, da_t \cdot P(s_t)\, ds_t$

v $\nabla_\theta \cdot 1 =$ ZERO

CONCLUSION > Q 대신 Q-V 함수 사용 : to lower variance

A2C

$$\nabla_\theta J_\theta \simeq \sum_{t=0}^{\infty} \int_{s_t, a_t} \nabla_\theta \ln P_\theta(a_t|s_t) \left[Q(s_t, a_t) - V(s_t) \right] P_\theta(s_t, a_t)\, ds_t, a_t$$

$$= \sum_{t=0}^{\infty} E\left[\nabla_\theta \ln P_\theta(a_t|s_t) \left[Q(s_t, a_t) - V(s_t) \right] \right] \qquad Q(s_t, a_t) = \mathbb{E}_{s_{t+1}}\left[R_t + \gamma V(s_{t+1}) \mid \overset{\text{given}}{s_t} \right]$$

$$= \sum_{t=0}^{\infty} E_{s_t, a_t}\left[\nabla_\theta \ln P_\theta(a_t|s_t)\, \mathbb{E}_{s_{t+1}}\left[R_t + \gamma V(s_{t+1}) - V(s_t) \mid s_t \right] \right] \qquad \text{"} V \rightarrow V_w \text{"}$$

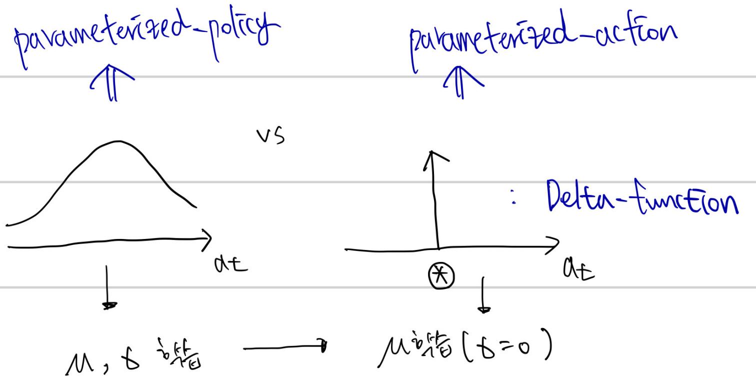
$$\downarrow$$

$$= \sum_{t=0}^{\infty} \int_{s_t, a_t} \nabla_\theta \ln P_\theta(a_t|s_t) \int_{s_{t+1}} \left(R_t + \gamma V(s_{t+1}) - V(s_t) \right) P(s_{t+1}|s_t, a_t)\, ds_{t+1}\, P_\theta(s_t, a_t)\, ds_t, a_t$$

$$= \sum_{t=0}^{\infty} \int_{s_t, a_t, s_{t+1}} \underbrace{\nabla_\theta \ln P_\theta(a_t|s_t) \left(R_t + \gamma V(s_{t+1}) - V(s_t) \right)}_{\text{sample}} P(s_t) P_\theta(a_t|s_t) P(s_{t+1}|s_t, a_t)$$

(on-policy : update에 사용된 N개의 sample은 폐기해야함. → 우리 θ에서 뽑은 sample 이여서)

DDPG



$$J_\theta = E[G_0] = \int_{s_0:a_\infty} G_0 \cdot P(s_0:a_\infty)\, ds_0:a_\infty = \int_{s_0:a_\infty} \cdot G_0 \cdot P(a_0:a_\infty \mid s_0)\, da_0:a_\infty\, P(s_0)\, ds_0 = \int_{s_0} V(s_0) P(s_0)\, ds_0$$

$$V(s_0) = \int_{a_0:a_\infty} G_0 \cdot P(a_0, \cdots, a_\infty \mid s_0)\, da_0:a_\infty = \int_{a_0} \int_{s_1:a_\infty} G_0 P(s_1:a_\infty \mid s_0, a_0)\, ds_1:a_\infty\, P_\theta(a_0 \mid s_0)\, da_0 = \int_{a_0} Q(s_0, a_0) P_\theta(a_0 \mid s_0)\, da_0$$

$$= Q(s_0, a_{\theta,0})$$

cf) $Q(s_0, a_0) = \int_{s_1:a_\infty} (R_0 + \gamma G_1) P(s_1:a_\infty \mid s_0, a_0)\, ds_1:a_\infty = \int_{s_1,a_1} \int_{s_2:a_\infty} (R_0 + \gamma G_1) P(s_2:a_\infty \mid s_1, a_1) P_\theta(a_1 \mid s_1) P(s_1 \mid s_0, a_0)\, ds_1, a_1\, ds_2:a_\infty$

$$= \int_{s_1,a_1} (R_0 + \gamma Q(s_1, a_1)) P_\theta(a_1 \mid s_1) P(s_1 \mid s_0, a_0)\, ds_1, a_1$$

$Q(s_0, a_{\theta,0}) = R(s_0, a_{\theta,0}) + \int_{s_1} \gamma Q(s_1, a_{\theta,1}) P(s_1 \mid s_0, a_{\theta,0})\, ds_1$ (DDPG Ver)

DDPG

① $\nabla_\theta J_\theta = \int_{s_0} \nabla_\theta V(s) P(s_0)\, ds_0 = \int_{s_0} \nabla_\theta Q(s_0, a_{\theta,0}) P(s_0)\, ds_0$

$$\nabla_\theta Q(s_0, a_{\theta,0}) = \nabla_{a_{\theta,0}} R(s_0, a_{\theta,0}) \cdot \nabla_\theta a_{\theta,0} + \int_{s_1} \gamma Q(s_1, a_{\theta,1}) \nabla_{a_{\theta,0}} P(s_1 | s_0, a_{\theta,0}) \cdot \nabla_\theta a_{\theta,0}\, ds_1 + \int_{s_1} \gamma \nabla_\theta Q(s_1, a_{\theta,1}) P(s_1 | s_0, a_{\theta,0})\, ds_1$$

$$\nabla_\theta a_{\theta,0} \cdot \nabla_{a_{\theta,0}} \left(R(s_0, a_{\theta,0}) + \int_{s_1} \gamma Q(s_1, a_{\theta,1}) P(s_1 | s_0, a_{\theta,0})\, ds_1 \right) + \int_{s_1} \gamma \nabla_\theta Q(s_1, a_{\theta,1}) P(s_1 | s_0, a_{\theta,0})\, ds_1$$

$$\nabla_\theta a_{\theta,0} \cdot \nabla_{a_{\theta,0}} \cdot Q(s_0, a_{\theta,0}) + \int_{s_1} \gamma \nabla_\theta Q(s_1, a_{\theta,1}) P(s_1 | s_0, a_{\theta,0})\, ds_1$$

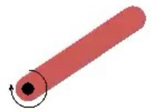
② $\nabla_\theta J_\theta = \int_{s_0} \nabla_\theta a_{\theta,0} \nabla_{a_{\theta,0}} Q(s_0, a_{\theta,0}) P(s_0)\, ds_0 + \int_{s_0}\int_{s_1} \nabla_\theta a_{\theta,1} \nabla_{a_{\theta,1}} Q(s_1, a_{\theta,1}) P(s_1 | s_0, a_{\theta,0})\, ds_1\, P(s_0)\, ds_0$

$$+ \int_{s_0}\int_{s_1}\int_{s_2} \nabla_\theta a_{\theta,2} \nabla_{a_{\theta,2}} Q(s_2, a_{\theta,2}) P(s_2 | s_1, a_{\theta,1})\, ds_2\, P(s_1 | s_0, a_{\theta,0})\, ds_1\, P(s_0)\, ds_0$$

$$= \sum_{t=0}^{\infty} \int_{s_0:s_t} \nabla_\theta a_{\theta,t} \nabla_{a_{\theta,t}} Q(s_t, a_{\theta,t}) P(s_0) P(s_1 | s_0, a_{\theta,0}) \cdots P(s_t | s_{t-1}, a_{\theta,t-1})\, ds_{0:t}$$

Pendulum-v0

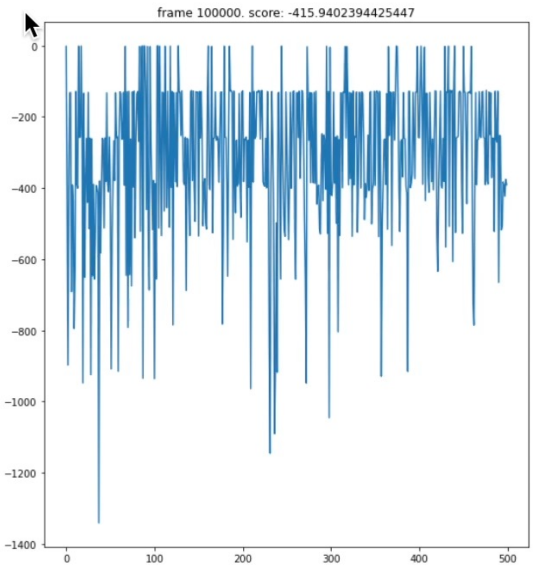
$[\cos(theta), \sin(theta), thetadot]$



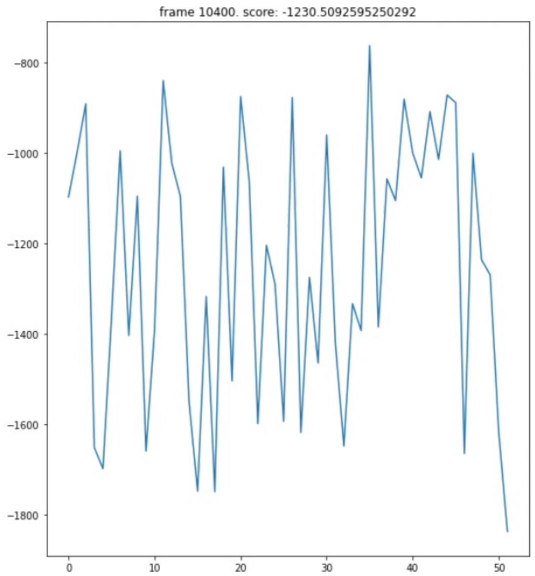
$[torque]$

$= swing\ the\ pendulum\ up\ so\ it\ stays\ upright$

Results

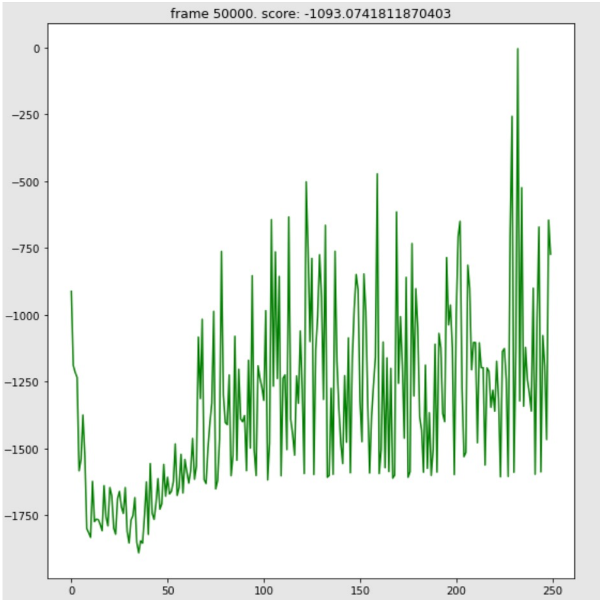


A2C

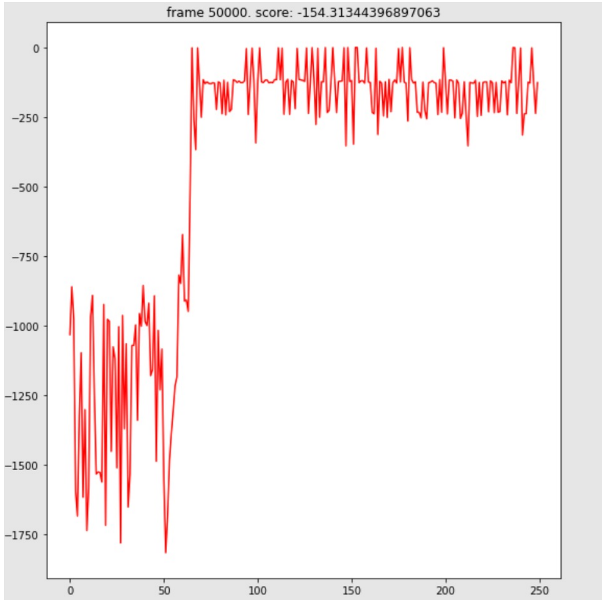


DDPG

Results



A2C



DDPG

Further Study

1) Off Policy Algorithm -> Different Policy (but almost similar)

2) Several Agents learning with totally different Algorithms / Policies

3) Sharing Results / Parameters of Learning -> Find out Optimal Policy / Agent

4) Apply to Sports Environment or Economic Model Simulator

감사합니다