나는 강화학습으로 축구한다

기초 이론

2025. 01. 19 - 20

01. 강화학습이란?

- 지도학습 vs 비지도학습 vs 강화학습
- 구성요소

02. 강화학습 이론

- 개요
- 마르코브 의사결정 과정
- 가치함수
- 벨만 기대 방정식
- 벨만 최적 방정식
- 최적 정책

03. 강화학습 알고리즘 분류

- Model-based vs. Model-free
- On-policy vs. Off-policy
- Value-based vs. Policy-based
- Actor-Critic
- Deterministic vs. Stochastic

04. 강화학습의 주요 알고리즘

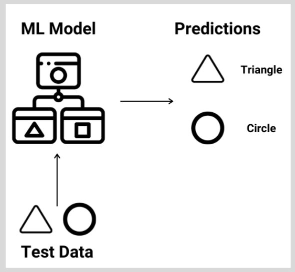
- Q-learning
- SARSA
- DQN
- Actor-Critic
- DDPG
- PPO

1. 강화학습이란?

1. 강화학습이란? : 지도학습 vs. 비지도학습 vs. 강화학습

지도학습

$$f^* = \text{argmin}_f \, \mathbb{E}\left[\mathcal{L}(f(\boldsymbol{x}), \boldsymbol{y})\right]$$



Note

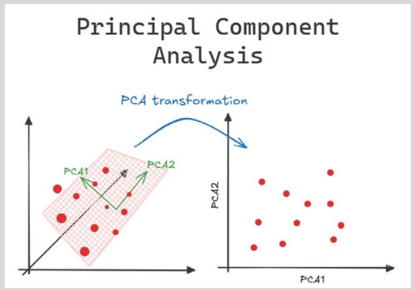
$$\mathbb{E}[X] = \sum_{i=1}^{K} P(X = x_i) \cdot x_i \approx \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\min_x(f(x)) \quad vs. \quad \text{argmin}_x(f(x))$$

https://medium.com/@dhara732002/supervised-machine-learning-a-beginners-guide-9ac0b07eccbb

비지도학습

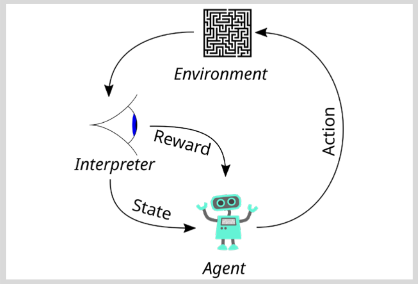
Latent structure in $\{x_1, x_2, \ldots, x_n\}$



https://ps.mjstudio.net/clustering-methods
https://mlpills.substack.com/p/issue-91-principal-component-analysis

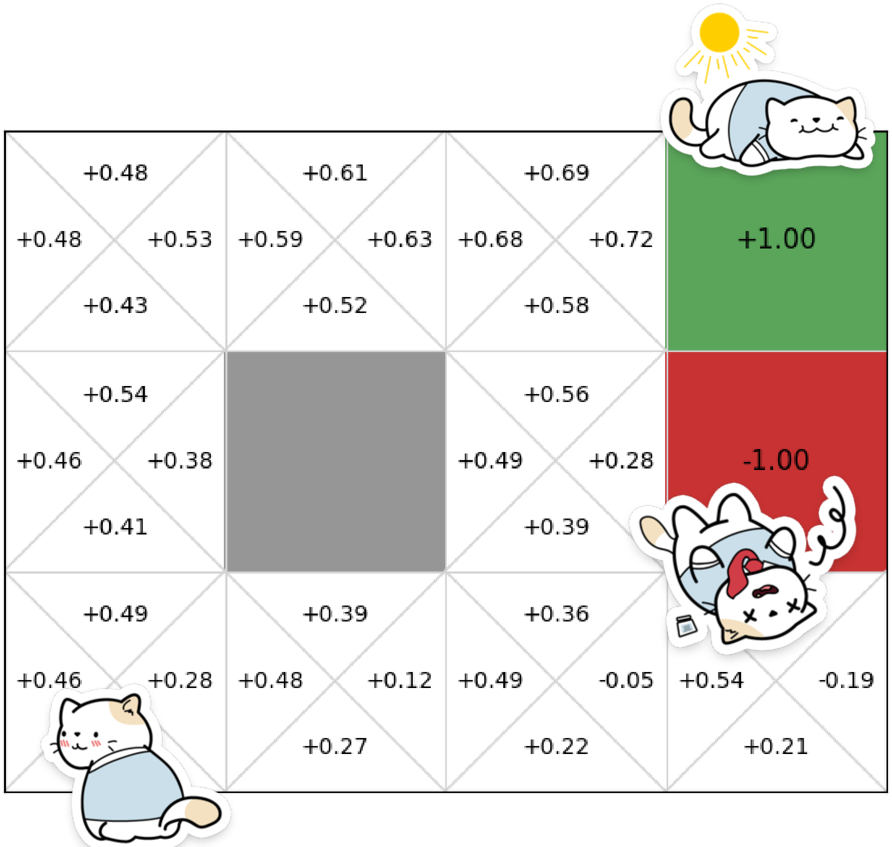
강화학습

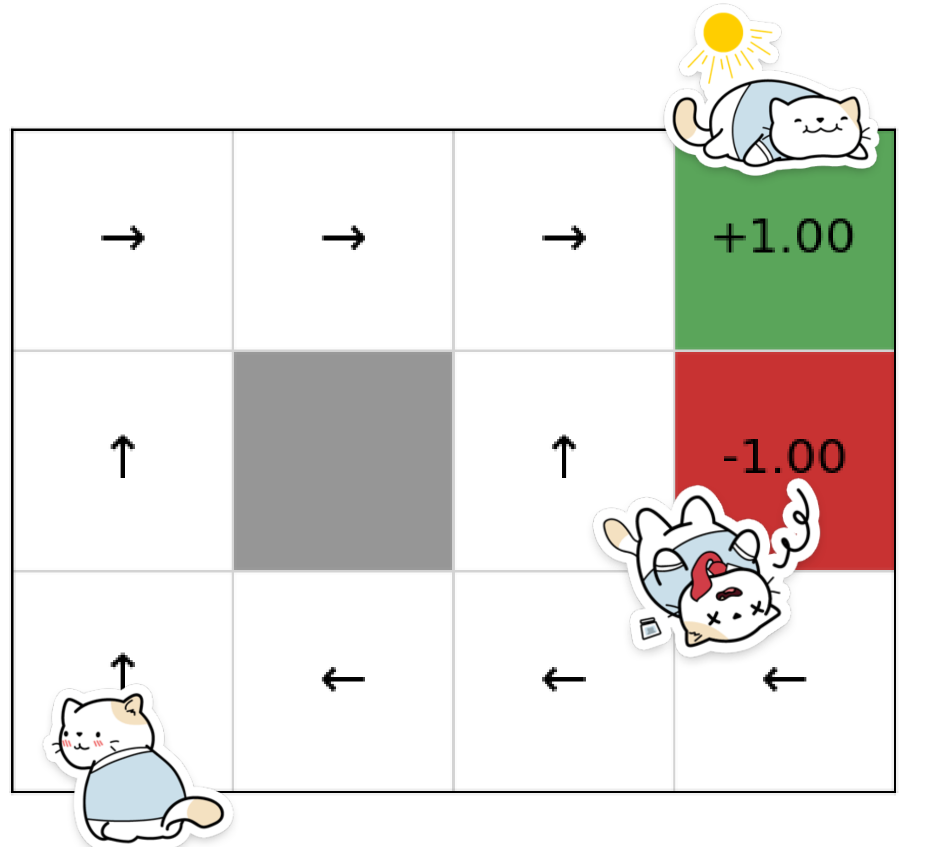
$$\pi^*(\boldsymbol{s}_t) \in \text{argmax}_{\boldsymbol{a}_t} \, Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t)$$



https://en.wikipedia.org/wiki/Reinforcement_learning

1. 강화학습이란? : 구성요소

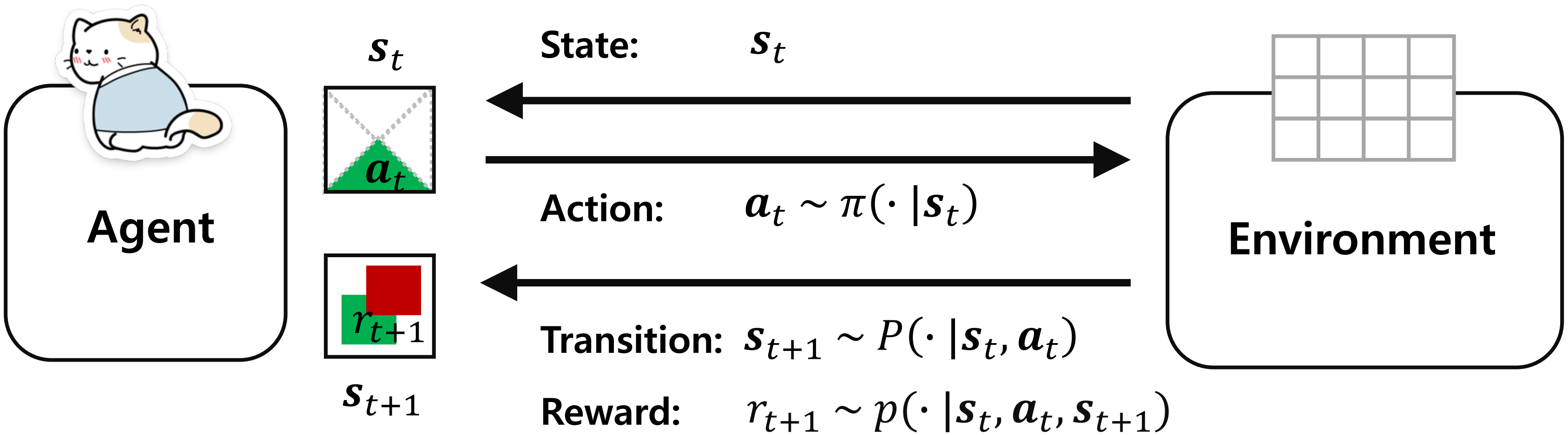




https://uq.pressbooks.pub/mastering-reinforcement-learning/chapter/q-function-approximation-mastering-reinforcement-learni

02. 강화학습 이론

2. 강화학습 이론 : 개요



Trajectory: $\mathcal{T} = (\boldsymbol{s}_0, \boldsymbol{a}_0, r_1, \boldsymbol{s}_1, \boldsymbol{a}_1, r_2, \boldsymbol{s}_2, \dots)$

Return: $G_t = \sum_{i=0}^{\infty} \gamma^i \cdot r_{t+i+1} = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots$ **Discount rate:** $\gamma \in [0,1)$

Q: $Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t) = \mathbb{E}_\pi[G_t | \boldsymbol{s}_t, \boldsymbol{a}_t]$ **Optimal Q:** $Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t) = \max_\pi Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t)$

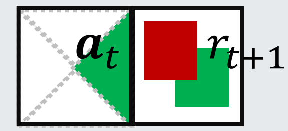
Optimal Policy: $\pi^*(\boldsymbol{s}_t) \in \mathrm{argmax}_{\boldsymbol{a}_t} Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t)$

02. 강화학습 이론 : 마르코브 의사결정 과정 (MDP)

Markov Decision Process:

$\boldsymbol{s}_t$

$\boldsymbol{s}_t \quad \boldsymbol{s}_{t+1}$



$$\boldsymbol{s}_t \xrightarrow[\boldsymbol{a}_t \sim \pi(\cdot\,|\boldsymbol{s}_t)]{} \boldsymbol{a}_t \xrightarrow[r_{t+1} \sim p(\cdot\,|\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})]{\boldsymbol{s}_{t+1} \sim P(\cdot\,|\boldsymbol{s}_t, \boldsymbol{a}_t)} r_{t+1}, \boldsymbol{s}_{t+1}$$

Trajectory: $\mathcal{T} = (\boldsymbol{s}_0, \boldsymbol{a}_0, r_1, \boldsymbol{s}_1, \boldsymbol{a}_1, r_2, \boldsymbol{s}_2, \dots)$

Markov Property:

$$\boxed{P(\boldsymbol{s}_{t+1}, r_{t+1} | \boldsymbol{s}_t, \boldsymbol{a}_t) = P(\boldsymbol{s}_{t+1}, r_{t+1} | \boldsymbol{s}_0, \boldsymbol{a}_0, r_1, \boldsymbol{s}_1, \boldsymbol{a}_1, r_2, \boldsymbol{s}_2, \dots, \boldsymbol{s}_t, \boldsymbol{a}_t)}$$

02. 강화학습 이론 : 가치함수 (Value-Function)

상태가치함수 (State-Value Function)

- 정책 π 를 따르는 경우, 상태 $\boldsymbol{s}_t$ 에서의 기대 리턴 G_t
- 이 상태는 얼마나 가치가 있는가?

$$V_\pi(\boldsymbol{s}_t) = \mathbb{E}_\pi[G_t | \boldsymbol{s}_t]$$

행동가치함수 (Action-Value Function)

- 상태 $\boldsymbol{s}_t$ 에서 행동 $\boldsymbol{a}_t$ 를 하고, 이후 정책 π 를 따르는 경우에 대한 기대 리턴 G_t
- 이 상태에서 이 행동을 얼마나 가치가 있는가?

$$Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t) = \mathbb{E}_\pi[G_t | \boldsymbol{s}_t, \boldsymbol{a}_t]$$

Note

$$V_\pi(\boldsymbol{s}_t) = \sum_{\boldsymbol{a}_t} \pi(\boldsymbol{a}_t | \boldsymbol{s}_t) Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t)$$

Note

$$G_t = \sum_{i=0}^{\infty} \gamma^i \cdot r_{t+i+1} = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ \dots$$

02. 강화학습 이론 : 벨만 기대방정식 (Bellman Expectation Equation)

상태가치함수 (State-Value Function)

Bellman Expectation Equation in Expectation Form

$$
\begin{aligned}
V_\pi(\boldsymbol{s}_t) &= \mathbb{E}_\pi[G_t|\boldsymbol{s}_t] \\
&= \mathbb{E}_\pi[r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \cdots)|\boldsymbol{s}_t] \\
&= \mathbb{E}_\pi[r_{t+1} + \gamma G_{t+1}|\boldsymbol{s}_t] \\
&= \mathbb{E}_\pi[r_{t+1}|\boldsymbol{s}_t] + \gamma\mathbb{E}_\pi[G_{t+1}|\boldsymbol{s}_t] \\
&= \mathbb{E}_\pi[r_{t+1}|\boldsymbol{s}_t] + \gamma\mathbb{E}_\pi[\,\mathbb{E}_\pi[G_{t+1}|\boldsymbol{s}_{t+1}]\mid \boldsymbol{s}_t] \\
&= \mathbb{E}_\pi[r_{t+1}|\boldsymbol{s}_t] + \gamma\mathbb{E}_\pi[V_\pi(\boldsymbol{s}_{t+\mathbf{1}})|\boldsymbol{s}_t]
\end{aligned}
$$

$$
\boxed{= \mathbb{E}_\pi[r_{t+1} + \gamma V_\pi(\boldsymbol{s}_{t+\mathbf{1}})|\boldsymbol{s}_t]}
$$

Bellman Expectation Equation in Summation Form

$$
\boxed{V_\pi(\boldsymbol{s}_t) = \sum_{\boldsymbol{a}_t} \pi(\boldsymbol{a}_t|\boldsymbol{s}_t) \sum_{\boldsymbol{s}_{t+1}} P(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t) \sum_{r_{t+1}} p(r_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})\{r_{t+1} + \gamma V_\pi(\boldsymbol{s}_{t+\mathbf{1}})\}}
$$

02. 강화학습 이론 : 벨만 기대방정식 (Bellman Expectation Equation)

행동가치함수 (*Action-Value Function*)

Note

$$V_\pi(\boldsymbol{s}_t) = \sum_{\boldsymbol{a}_t} \pi(\boldsymbol{a}_t|\boldsymbol{s}_t) Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t)$$

Bellman Expectation Equation in Expectation Form

$$\begin{aligned}
Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t) &= \mathbb{E}_\pi[G_t|\boldsymbol{s}_t, \boldsymbol{a}_t] \\
&= \mathbb{E}_\pi[r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \cdots)|\boldsymbol{s}_t, \boldsymbol{a}_t] \\
&= \mathbb{E}_\pi[r_{t+1} + \gamma G_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t] \\
&= \mathbb{E}[r_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t] + \gamma \mathbb{E}_\pi[G_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t] \\
&= \mathbb{E}[r_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t] + \gamma \mathbb{E}[\, \mathbb{E}_\pi[G_{t+1}|\boldsymbol{s}_{t+1}] \, |\boldsymbol{s}_t, \boldsymbol{a}_t] \\
&= \mathbb{E}[r_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t] + \gamma \mathbb{E}[V_\pi(\boldsymbol{s}_{t+1})|\boldsymbol{s}_t, \boldsymbol{a}_t] \\
&= \mathbb{E}[r_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t] + \gamma \mathbb{E}\left[\sum_{\boldsymbol{a}_{t+1}} \pi(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1}) Q_\pi(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}) \,|\boldsymbol{s}_t, \boldsymbol{a}_t\right] \\
&= \mathbb{E}\left[r_{t+1} + \gamma \sum_{\boldsymbol{a}_{t+1}} \pi(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1}) Q_\pi(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}) \,|\boldsymbol{s}_t, \boldsymbol{a}_t\right] \\
&= \boxed{\mathbb{E}[r_{t+1} + \gamma \mathbb{E}_\pi[Q_\pi(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})]|\boldsymbol{s}_t, \boldsymbol{a}_t]}
\end{aligned}$$

02. 강화학습 이론 : 벨만 기대방정식 (Bellman Expectation Equation)

행동가치함수 (*Action-Value Function*)

Bellman Expectation Equation in Summation Form

$$Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t) = \mathbb{E}[r_{t+1} + \gamma \mathbb{E}_\pi[Q_\pi(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})] | \boldsymbol{s}_t, \boldsymbol{a}_t]$$

$$= \mathbb{E}_{\boldsymbol{s}_{t+1} \sim P(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)}\left[\mathbb{E}_{r_{t+1} \sim p(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})}\left[\left[r_{t+1} + \gamma \mathbb{E}_{\boldsymbol{a}_{t+1} \sim \pi(\cdot|\boldsymbol{s}_{t+1})}[Q_\pi(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})]\right]\right] | \boldsymbol{s}_t, \boldsymbol{a}_t\right]$$

$$\boxed{= \sum_{\boldsymbol{s}_{t+1}} P(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t) \sum_{r_{t+1}} p(r_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}) \left\{r_{t+1} + \gamma \sum_{\boldsymbol{a}_{t+1}} \pi(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1}) Q_\pi(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})\right\}}$$

02. 강화학습 이론 : 벨만 최적 방정식 (Bellman Optimality Equation)

상태가치함수 (State-Value Function)

$$V^*(\boldsymbol{s}_t) = \max_{\pi} V_\pi(\boldsymbol{s}_t)$$

$$= \max_{a_t} \mathbb{E}[r_{t+1} + \gamma V^*(\boldsymbol{s}_{t+\mathbf{1}}) | \boldsymbol{s}_t, \boldsymbol{a}_t]$$

$$\boxed{= \max_{a_t} Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t)}$$

행동가치함수 (Action-Value Function)

$$Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t) = \max_{\pi} Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t)$$

$$= \mathbb{E}[r_{t+1} + \gamma \mathbb{E}_\pi[Q_\pi(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})] | \boldsymbol{s}_t, \boldsymbol{a}_t]$$

$$\boxed{= \mathbb{E}\left[r_{t+1} + \gamma \max_{\boldsymbol{a}_{t+1}} [Q^*(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})] | \boldsymbol{s}_t, \boldsymbol{a}_t\right]}$$

02. 강화학습 이론 : 최적정책 (Optimal Policy)

결정론적 최적정책 (Deterministic Optimal Policy)

$$A^*(\boldsymbol{s}_t) \;\coloneqq \operatorname{argmax}_{\boldsymbol{a}_t \in A}\; Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t)$$

$$\boxed{\pi^*(\boldsymbol{s}_t) \in A^*(\boldsymbol{s}_t)}$$

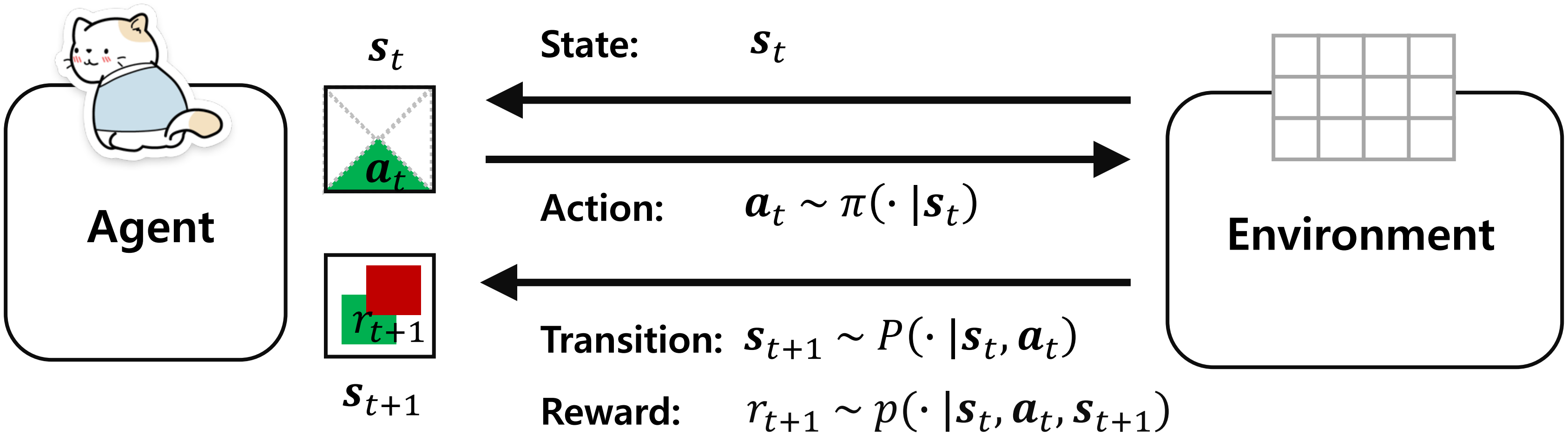
확률론적 최적정책 (Stochastic Optimal Policy)

$$\pi^*(\boldsymbol{a}_t | \boldsymbol{s}_t) = 0 \quad \forall a \notin A^*(\boldsymbol{s}_t)\;, \qquad \sum_{a \in A^*(\boldsymbol{s}_t)} \pi^*(\boldsymbol{a}_t | \boldsymbol{s}_t) = 1$$

03. 강화학습 알고리즘 분류

03. 강화학습 알고리즘 분류 : Model-based vs. Model-free

Model-based vs. Model-free



$\pi(a_t|s_t)$ $\quad$ $P(s_{t+1}|s_t, a_t)$ $\quad$ $p(r_{t+1}|s_t, a_t, s_{t+1})$

Known

Unknown

Model-based

Learned by planning

Model-free

Learned by transition:
(s_t, a_t, r_t, s_{t+1})

03. 강화학습 알고리즘 분류 : Model-based vs. Model-free

Model-based *vs.* *Model-free*

Planning

$$V_\pi(\boldsymbol{s}_t) = \sum_{\boldsymbol{a}_t} \pi(\boldsymbol{a}_t|\boldsymbol{s}_t) \sum_{\boldsymbol{s}_{t+1}} P(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t) \sum_{r_{t+1}} p(r_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})\{r_{t+1} + \gamma V_\pi(\boldsymbol{s}_{t+1})\}$$

$$V^*(\boldsymbol{s}_t) = \max_{a_t} Q_\pi^*(\boldsymbol{s}_t, \boldsymbol{a}_t)$$

$$Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t) == \sum_{\boldsymbol{s}_{t+1}} P(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t) \sum_{r_{t+1}} p(r_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}) \left\{ r_{t+1} + \gamma \sum_{\boldsymbol{a}_{t+1}} \pi(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1})Q_\pi(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}) \right\}$$

$$Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t) = \mathbb{E}\left[r_{t+1} + \gamma \max_{\boldsymbol{a}_{t+1}} [Q^*(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})] | \boldsymbol{s}_t, \boldsymbol{a}_t \right]$$

$$A^*(\boldsymbol{s}_t) \coloneqq \mathrm{argmax}_{\boldsymbol{a}_t \in A}\ Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t)$$

$$\pi^*(\boldsymbol{s}_t) \in A^*(\boldsymbol{s}_t)$$

$$\pi^*(\boldsymbol{a}_t|\boldsymbol{s}_t) = 0 \quad \forall a \notin A^*(\boldsymbol{s}_t)$$

$$\sum_{a \in A^*(\boldsymbol{s}_t)} \pi^*(\boldsymbol{a}_t|\boldsymbol{s}_t) = 1$$

03. 강화학습 알고리즘 분류 : Model-based vs. Model-free

Model-based vs. ***Model-free***

$$V^*(\boldsymbol{s}_t) = \max_{a_t} \mathbb{E}[r_{t+1} + \gamma V^*(\boldsymbol{s}_{t+1}) | \boldsymbol{s}_t, \boldsymbol{a}_t]$$

$$(\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}_{t+1})$$

$$V^*(\boldsymbol{s}_t) \approx r_{t+1} + \gamma V^*(\boldsymbol{s}_{t+1})$$

$$TD = r_{t+1} + \gamma \hat{V}_{\boldsymbol{\pi}}(\boldsymbol{s}_{t+1}) - \hat{V}_{\boldsymbol{\pi}}(\boldsymbol{s}_t)$$

$$\hat{V}_{\boldsymbol{\pi}}(\boldsymbol{s}_t) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\boldsymbol{s}_t) + \alpha\left(r_{t+1} + \gamma \widehat{\hat{V}}_{\boldsymbol{\pi}}(\boldsymbol{s}_{t+1}) - \hat{V}_{\boldsymbol{\pi}}(\boldsymbol{s}_t)\right)$$

** Q-learning, DQN, DDPG, PPO*

03. 강화학습 알고리즘 분류 : Model-based vs. Model-free

Model-based vs. Model-free

$$Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t) = \mathbb{E}\left[r_{t+1} + \gamma \max_{\boldsymbol{a}_{t+1}}[Q^*(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})] | \boldsymbol{s}_t, \boldsymbol{a}_t\right]$$

$$Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t) = r_{t+1} + \gamma \max_{\boldsymbol{a}_{t+1}} Q^*(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})$$

$$TD = r_{t+1} + \gamma \max_{\boldsymbol{a}_{t+1}} \hat{Q}_{\boldsymbol{\pi}}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}) - \hat{Q}_{\boldsymbol{\pi}}(\boldsymbol{s}_t, \boldsymbol{a}_t)$$

$$\hat{Q}_{\boldsymbol{\pi}}(\boldsymbol{s}_t, \boldsymbol{a}_t) \leftarrow \hat{Q}_{\boldsymbol{\pi}}(\boldsymbol{s}_t, \boldsymbol{a}_t) + \alpha\left(r_{t+1} + \gamma \max_{\boldsymbol{a}_{t+1}} \hat{Q}_{\boldsymbol{\pi}}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}) - \hat{Q}_{\boldsymbol{\pi}}(\boldsymbol{s}_t, \boldsymbol{a}_t)\right)$$

** Q-learning, DQN, DDPG, PPO*

03. 강화학습 알고리즘 분류 : On-policy vs. Off-policy

On-policy vs. Off-policy

$$\mathcal{T} = (\boldsymbol{s}_0, \boldsymbol{a}_0, r_1, \boldsymbol{s}_1, \boldsymbol{a}_1, r_2, \boldsymbol{s}_2, \dots)$$

On-policy

$$\hat{Q}_{\boldsymbol{\pi}}(\boldsymbol{s}_t, \boldsymbol{a}_t) \longleftarrow \hat{Q}_{\boldsymbol{\pi}}(\boldsymbol{s}_t, \boldsymbol{a}_t) + \alpha\left(r_{t+1} + \gamma \hat{Q}_{\boldsymbol{\pi}}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}) - \hat{Q}_{\boldsymbol{\pi}}(\boldsymbol{s}_t, \boldsymbol{a}_t)\right)$$

** SARSA, PPO*

Off-policy

$$\hat{Q}_{\boldsymbol{\pi}}(\boldsymbol{s}_t, \boldsymbol{a}_t) \longleftarrow \hat{Q}_{\boldsymbol{\pi}}(\boldsymbol{s}_t, \boldsymbol{a}_t) + \alpha\left(r_{t+1} + \gamma \max_{\boldsymbol{a}_{t+1}} \hat{Q}_{\boldsymbol{\pi}}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}) - \hat{Q}_{\boldsymbol{\pi}}(\boldsymbol{s}_t, \boldsymbol{a}_t)\right)$$

** Q-learning, DQN, DDPG*

03. 강화학습 알고리즘 분류 : Value-based vs. Policy-based

Value-based vs. Policy-based

Value-based

$$\hat{\boldsymbol{\pi}}(\boldsymbol{s}_t) = \mathrm{argmax}_{\boldsymbol{a}_t} \hat{Q}_{\boldsymbol{\pi}}(\boldsymbol{s}_t, \boldsymbol{a}_t)$$

** Q-learning, DQN*

Policy-based

$$\text{Maximize} \quad J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}}}[G_0]$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}}}\left[\sum_{t=0}^{T-1} \nabla_{\boldsymbol{\theta}} \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t) G_t\right] = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1} \nabla_{\boldsymbol{\theta}} \log \pi_\theta\left(\boldsymbol{a}_t^i|\boldsymbol{s}_t^i\right) G_t^i$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1} \nabla_{\boldsymbol{\theta}} \log \pi_\theta\left(\boldsymbol{a}_t^i|\boldsymbol{s}_t^i\right) G_t^i$$

$$\boldsymbol{a} \sim \pi_\theta(\cdot\,|\boldsymbol{s}_t)$$

** PPO*

03. 강화학습 알고리즘 분류 : Actor-Critic

Actor-Critic

Actor

$$\boldsymbol{a}_t \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}(\cdot \mid s_t)$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \times \delta \times \nabla_{\boldsymbol{\theta}} \log(\boldsymbol{\pi}_{\boldsymbol{\theta}}(\boldsymbol{a}_t \mid \boldsymbol{s}_t))$$

Critic

$$V_w(\boldsymbol{s}_t) \qquad \delta_t = r_{t+1} + \gamma V_w(\boldsymbol{s}_{t+1}) - V_w(\boldsymbol{s}_t)$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \beta \times \delta \times \nabla_{\boldsymbol{w}} V_w(s_t)$$

** DDPG, PPO*

03. 강화학습 알고리즘 분류 : Deterministic vs. Stochastic

Deterministic vs. Stochastic

Deterministic

$$\boldsymbol{s}_{t+1} = F(\boldsymbol{s}_t, \boldsymbol{a}_t)$$

$$\boldsymbol{a}_t = f(\boldsymbol{s}_t)$$

$$r_{t+1} = R(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})$$

** DDPG*

Stochastic

$$\boldsymbol{s}_{t+1} = P(\cdot\,|\boldsymbol{s}_t, \boldsymbol{a}_t)$$

$$\boldsymbol{a}_t \sim \pi(\cdot\,|\boldsymbol{s}_t)$$

$$r_{t+1} \sim p(\cdot\,|\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})$$

** PPO*

04. 강화학습 주요 알고리즘

04. 강화학습의 주요 알고리즘: Q-learning

Off-policy *Value-based* *Discrete Action* *Model-free*

Pseudo-code

$\hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t) = 0$

for each episode:

$\boldsymbol{s}_t$ = reset()

while not terminal:

$\boldsymbol{a}_t$ = epsilon_greedy(Q, $\boldsymbol{s}_t$)

$\boldsymbol{s}_{t+1}$, r_{t+1}, done = step($\boldsymbol{a}_t$)

target = $r_{t+1} + \gamma \times \max_{\boldsymbol{a}_{t+1}}\left[\hat{Q}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})\right]$ ×(not done)

$\hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t) = \hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t) + \alpha$(target $- \hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t)$)

$\boldsymbol{s}_t = \boldsymbol{s}_{t+1}$

$$\boldsymbol{a}_t = \begin{cases} \mathrm{argmax}_{\boldsymbol{a}_t} \hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t) & \text{with prob } 1-\varepsilon \\ \text{random action} & \text{with prob } \varepsilon \end{cases}$$

$$\boldsymbol{s}_t, \boldsymbol{a}_t, r_{t+1}, \boldsymbol{s}_{t+1}$$

$$\begin{aligned} Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t) &= \max_{\pi} Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t) \\ &= \mathbb{E}\left[r_{t+1} + \gamma \max_{\boldsymbol{a}_{t+1}}[Q^*(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})] | \boldsymbol{s}_t, \boldsymbol{a}_t\right] \\ &\approx r_{t+1} + \gamma \max_{\boldsymbol{a}_{t+1}}[Q^*(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})] \end{aligned}$$

(single-sample estimate of the expectation)

$$\hat{y} = r_{t+1} + \gamma \max_{\boldsymbol{a}_{t+1}}\left[\hat{Q}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})\right] \cdot 1[\text{not done}]$$

$$\hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t) \leftarrow \hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t) + \alpha\left(\hat{y} - \hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t)\right)$$

$$\hat{Q} \to Q^* \quad (\textit{as learning convergens})$$

$$\pi^*(\boldsymbol{s}_t) = \mathrm{argmax}_{\mathrm{a}} Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t)$$

04. 강화학습의 주요 알고리즘: SARSA

On-policy *Value-based* *Discrete Action* *Model-free*

Pseudo-code

$\hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ = 0

for each episode:

  $\boldsymbol{s}_t$ = reset()
  $\boldsymbol{a}_t$ = epsilon_greedy(Q, $\boldsymbol{s}_t$)

  while not terminal:
    $\boldsymbol{s}_{t+1}$, r_{t+1}, done = step($\boldsymbol{a}_t$)
    $\boldsymbol{a}_{t+1}$ = epsilon_greedy(Q, $\boldsymbol{s}_t + 1$)

    target = r_{t+1} + $\gamma \times \hat{Q}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})$ ×(not done)

      $\hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ =$\hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ + α(target -$\hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t)$)

    $\boldsymbol{s}_t$, $\boldsymbol{a}_t$ = $\boldsymbol{s}_{t+1}$, $\boldsymbol{a}_{t+1}$

$$\boldsymbol{a}_t = \begin{cases} \mathrm{argmax}_{\boldsymbol{a}_t} \hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t) & with\ prob\ 1-\varepsilon \\ random\ action & with\ prob\ \varepsilon \end{cases}$$

$$\boldsymbol{s}_t, \boldsymbol{a}_t, r_{t+1}, \boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}$$

$$Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t) = \mathbb{E}[r_{t+1} + \gamma \mathbb{E}_\pi[Q_\pi(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})] | \boldsymbol{s}_t, \boldsymbol{a}_t]$$

$$\approx r_{t+1} + \gamma Q_\pi(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})$$

(single-sample estimate of the expectation)

$$\hat{y} = r_{t+1} + \gamma \hat{Q}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}) \cdot 1[not\ done]$$

$$\hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t) \leftarrow \hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t) + \alpha\big(\hat{y} - \hat{Q}(\boldsymbol{s}_t, \boldsymbol{a}_t)\big)$$

$$\hat{Q} \rightarrow Q^*$$

$$(if\ GLIE:\ \varepsilon \rightarrow 0\ with\ infinite\ exploration)$$

$$\pi^*(\boldsymbol{s}_t) = \mathrm{argmax}_a Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t)$$

04. 강화학습의 주요 알고리즘: DQN (Deep Q-Network)

Off-policy *Value-based* *Discrete Action* *Model-free* *Deep Neural Network* *Replay Buffer* *Target Network*

Pseudo-code

```
Replay buffer D = empty_buffer()
Online network parameters Q_θ = init_network()
Target network parameters Q_θ- = copy(Q_θ )

for each episode:

  s_t = reset()
  done = False

  while not done:

    a_t = epsilon_greedy(Q_θ, s_t)
    s_{t+1}, r_{t+1}, done = step(a_t)
    D.add(s_t, a_t, r_{t+1}, s_{t+1}, done)
    s_t = s_{t+1}

    if len(D) == B:
       batch = D.randomSample(B)
              # {s_t^i, a_t^i, r_{t+1}^i, s_{t+1}^i, done^i} for i =1..B
    target _i
         = r_{i+1} + γ × max_{a_{t+1}}[Q_θ-(s_{t+1}^i, a_{t+1})] ×(not done)
    L(θ) = (1/B) Σ_i(Q_θ(s_t^i, a_t^i) − target_i)^2
    θ = θ − l_r × ∇_θ L(θ)

    every C steps: θ⁻ ← θ
```

$$\boldsymbol{a}_t = \begin{cases} \mathrm{argmax}_{\boldsymbol{a}_t} Q_\theta(\boldsymbol{s}_t, \boldsymbol{a}_t) & \textit{with prob } 1-\varepsilon \\ \textit{random action} & \textit{with prob } \varepsilon \end{cases}$$

$$\boldsymbol{s}_t, \boldsymbol{a}_t, r_{t+1}, \boldsymbol{s}_{t+1}$$

$$Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t) = \max_\pi Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t)$$

$$= \mathbb{E}\left[r_{t+1} + \gamma \max_{\boldsymbol{a}_{t+1}}[Q^*(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})] | \boldsymbol{s}_t, \boldsymbol{a}_t\right]$$

(minibatch-sample estimate of the expectation)

$$\hat{y}_i = r_{i+1} + \gamma \max_{\boldsymbol{a}_{t+1}} Q_{\theta^-}(\boldsymbol{s}_{t+1}^i, \boldsymbol{a}_{t+1}) \cdot 1[\textit{not done}]$$

$$\hat{\mathcal{L}}(\theta) = \frac{1}{B}\sum_{i=1}^{B}\left(\hat{y}_i - \hat{Q}(\boldsymbol{s}_t^i, \boldsymbol{a}_t^i)\right)^2$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - L_r \nabla_\theta \hat{\mathcal{L}}(\theta) \qquad (\textit{for every steps})$$

$$\boldsymbol{\theta}^- \leftarrow \boldsymbol{\theta} \qquad (\textit{every } C \textit{ steps})$$

$$\hat{Q} \rightarrow Q^* \qquad (\textit{as learning convergens})$$

$$\pi^*(\boldsymbol{s}_t) = \mathrm{argmax}_{\mathrm{a}} Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t)$$

04. 강화학습의 주요 알고리즘: Actor–Critic

On/Off-policy *Value/Policy-based* *Discrete/Continuous Action* *Model-free* *Deep Neural Network*

Pseudo-code

actor: $\boldsymbol{\theta}$, $\boldsymbol{\pi_\theta}(\boldsymbol{a}_t|s_t)$
Critic: $\boldsymbol{w}$, $V_w(s_t)$

for each episode:
  $\boldsymbol{a}_t \sim \boldsymbol{\pi_\theta}(\cdot\,|s_t)$
  $(r_{t+1}, \boldsymbol{s}_{t+1}, done)$ = step($\boldsymbol{a}_t$)

  $\delta_t = r_{t+1} + \gamma\, V_{\boldsymbol{w}}(\boldsymbol{s}_{t+1})$ ×(not done) - $V_{\boldsymbol{w}}(\boldsymbol{s}_t)$

  # critic update
  $\boldsymbol{w} \leftarrow \boldsymbol{w} + \beta \times \delta \times \nabla_{\boldsymbol{w}} V_w(s_t)$

  # actor update (policy gradient)
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \times \delta \times \nabla_{\boldsymbol{\theta}} \log(\boldsymbol{\pi_\theta}(\boldsymbol{a}_t|\boldsymbol{s}_t))$

  $\boldsymbol{s}_t \leftarrow \boldsymbol{s}_{t+1}$ (or reset if done)

$$V^*(\boldsymbol{s}_t) = \max_{\pi} V_\pi(\boldsymbol{s}_t)$$

$$= \max_{a_t} \mathbb{E}[r_{t+1} + \gamma V_\pi(\boldsymbol{s}_{t+1})|\boldsymbol{s}_t, \boldsymbol{a}_t]$$

04. 강화학습의 주요 알고리즘: DDPG (Deep Deterministic Policy Gradient)

Off-policy *Actor-Critic* *Continuous Action* *Deep Neural Network*

Pseudo-code

actor $\mu_{\boldsymbol{\theta}}(\boldsymbol{s}_t)$ actor_target $\mu_{\boldsymbol{t}} = copy(\mu_{\boldsymbol{\theta}})$
critic $Q_{\boldsymbol{\phi}}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ critic_target $\boldsymbol{Q_t} = copy(Q_{\boldsymbol{\phi}})$
Replay Buffer: D

for each step:
  $\boldsymbol{a}_t = \mu_{\boldsymbol{\theta}}(\boldsymbol{s}_t)$ + noise
  $\boldsymbol{s}_{t+1}, r_{t+1}, done$ = step($\boldsymbol{a}_t$)
  D.add($\boldsymbol{s}_t, \boldsymbol{a}_t, r_{t+1}, \boldsymbol{s}_{t+1}, done$)

  batch = D.randomSample(B)

  $\hat{y}_i = r_{t+1,i} + \gamma\, \boldsymbol{Q_t}\left(\boldsymbol{s}_{t+1,i}, \mu_{\boldsymbol{t}}(\boldsymbol{s}_{t+1,i})\right) \times (\text{not } done)$

  $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \beta \nabla_{\boldsymbol{\phi}} \frac{\boldsymbol{1}}{\boldsymbol{B}} \sum_i^B \left(Q_{\boldsymbol{\phi}}(\boldsymbol{s}_{t,i}, \boldsymbol{a}_{t,i}) - \hat{y}_i\right)^2$

  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} Q_{\boldsymbol{\phi}}(\boldsymbol{s}_t, \mu_{\boldsymbol{\theta}}(\boldsymbol{s}_t))$
     $\boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} Q_{\boldsymbol{\phi}}(\boldsymbol{s}_t, \boldsymbol{a})|_{a=\mu_{\boldsymbol{\theta}}(\boldsymbol{s}_t)} \nabla_{\boldsymbol{\theta}} \mu_{\boldsymbol{\theta}}(\boldsymbol{s}_t)$

  # soft target update
  $Q_{\boldsymbol{t}} \leftarrow \tau\, Q_{\boldsymbol{\phi}} + (1-\tau) Q_{\boldsymbol{t}}$
  $\mu_{\boldsymbol{t}} \leftarrow \tau\, \mu_{\boldsymbol{\theta}} + (1-\tau) \mu_{\boldsymbol{t}}$

  $\boldsymbol{s}_t \leftarrow \boldsymbol{s}_{t+1}$ (or reset if done)

$$Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t) = \max_{\pi} Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t)$$
$$= \mathbb{E}\left[r_{t+1} + \gamma \max_{\boldsymbol{a}_{t+1}} [Q^*(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})] | \boldsymbol{s}_t, \boldsymbol{a}_t\right]$$

04. 강화학습의 주요 알고리즘: PPO (Proximal Policy Optimization)

On-policy *Actor-Critic* *Discrete/Continuous Action* *Deep Neural Network*

Pseudo-code

Initialize policy $\boldsymbol{\pi_\theta}$ and value function $\boldsymbol{V_\phi}$
for iter = 1..K:
$\boldsymbol{\theta_{old}} \leftarrow \boldsymbol{\theta}$
$\boldsymbol{\phi_{old}} \leftarrow \boldsymbol{\phi}$

\# Collect rollouts with old policy
$\boldsymbol{D} = \{(\boldsymbol{s}_{t,i}, \boldsymbol{a}_{t,i}, \boldsymbol{r}_{t+1,i}, \boldsymbol{s}_{t+1,i}, done_i)\}$ collected by $\boldsymbol{\pi_{\theta,old}}$

\# Compute return and advantages
$R_{t,i} = r_{t+1,i} + \gamma\, V_{\boldsymbol{old}}(\boldsymbol{s}_{t+1,i}) \times(\text{not } done_i)$
$\widehat{\boldsymbol{A}}_{t,i} = R_{t,i} - V_{\boldsymbol{old}}(\boldsymbol{s}_{t,i})$

for epoch = 1..E:
for minibatch $\boldsymbol{B} \subset \boldsymbol{D}$:

\# Actor loss (clip objective)
$r_{t,i} = \exp\big(log\boldsymbol{\pi_\theta}(\boldsymbol{a}_{t,i}|\boldsymbol{s}_{t,i}) - log\boldsymbol{\pi_{\theta.old}}(\boldsymbol{a}_{t,i}|\boldsymbol{s}_{t,i})\big)$
$\boldsymbol{L_{clip}} = \text{mean}_{i\in B}(\min(r_{t,i} \times \widehat{\boldsymbol{A}}_{t,i}, \text{clip}(r_{t,i}, 1\text{-}\varepsilon, 1+\varepsilon) \times \widehat{\boldsymbol{A}}_{t,i}))$

\# Critic loss (MSE)
$\boldsymbol{L_v} = \text{mean}_{i\in B}\big(\boldsymbol{V_\phi}(\boldsymbol{s}_{t,i}) - R_{t,i}\big)^{\boldsymbol{2}}$

\# Update
$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\nabla_{\boldsymbol{\theta}}\boldsymbol{L_{clip}}(\boldsymbol{\theta})$
$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \beta\nabla_{\boldsymbol{\phi}}\boldsymbol{L_v}(\boldsymbol{\phi})$

수고하셨습니다.

junsu@handong.edu