Chapter. 04

모델 없이 세상 조종하기

MC Control: MC 기법을 활용한 최적 정책 찾기

fast campus

FAST CAMPUS ONLINE
강화학습 A-Z I

강사. 박준영

Generalized Policy Iteration (복습)

정책 평가:

<u>어떠한</u> 알고리즘을 활용해서라도,

주어진 정책 π 에 대한 가치함수 $V^{\pi}(s)$ 를 계산

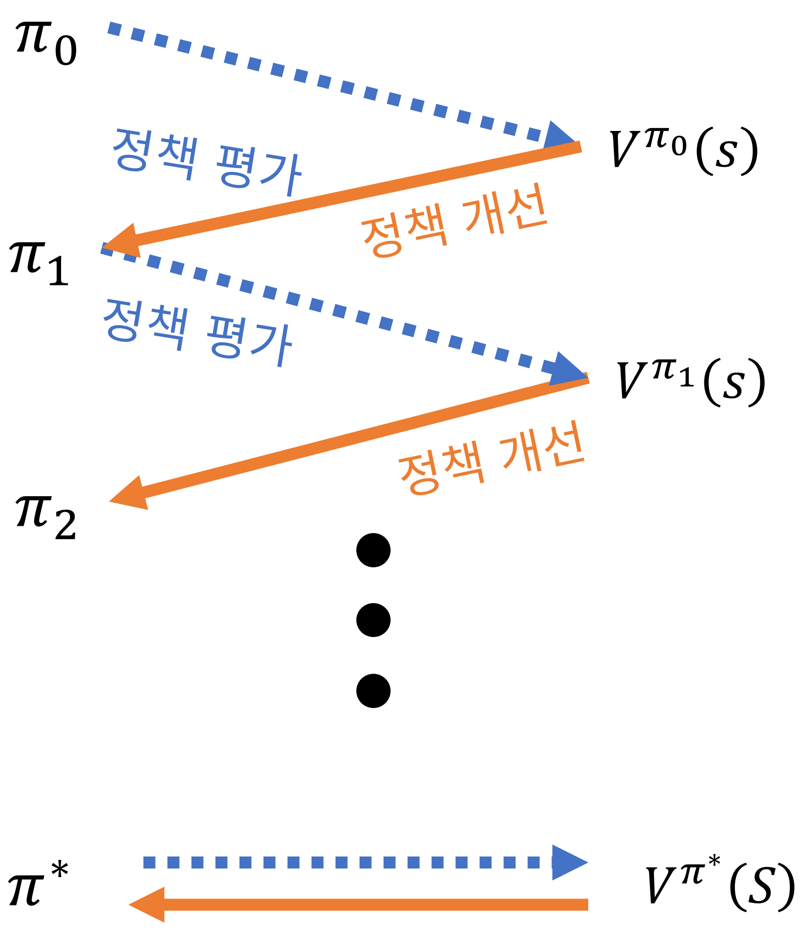
예시) DP 를 활용한 가치평가 / MC-PE/ TD-PE

정책 개선:

$\pi_{i+1} \geq \pi_i$ 를 만족시키는 π_{i+1} 을 생성

예시) 탐욕적 정책 개선

(정책함수 π 간의 대소관계를 다음과 같이 정의 한다.
$\pi' \geq \pi$ 만약 $V_{\pi'}(s) \geq V_{\pi}(s), \forall s \in \mathcal{S}$)



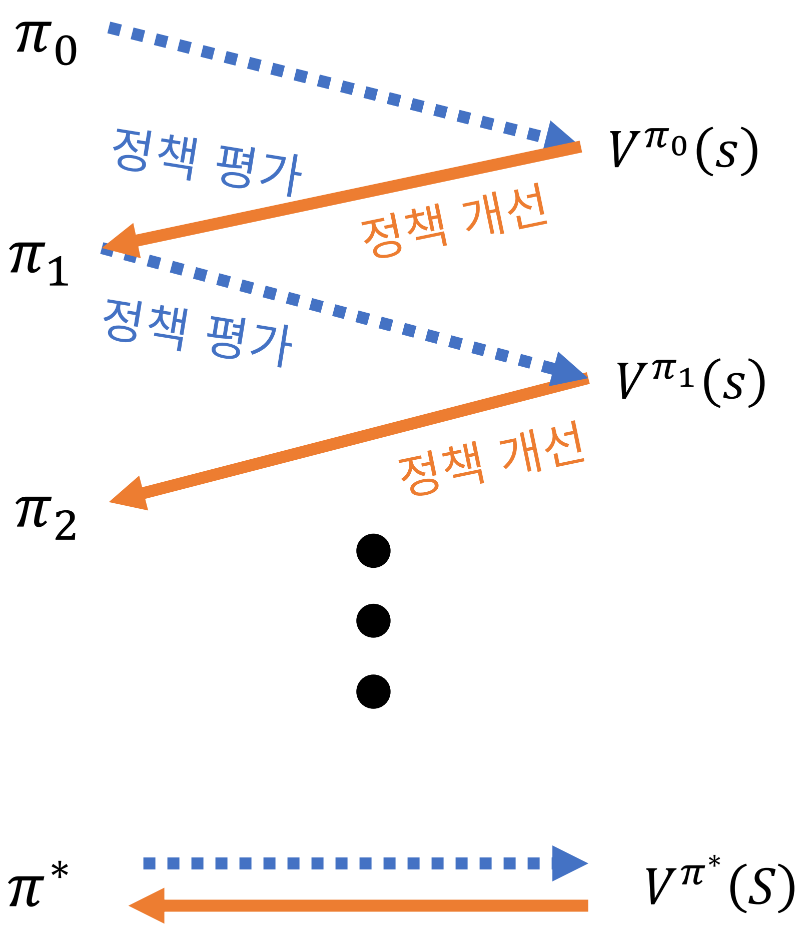
I MC policy evaluation + 탐욕적 개선 ??

정책 평가:

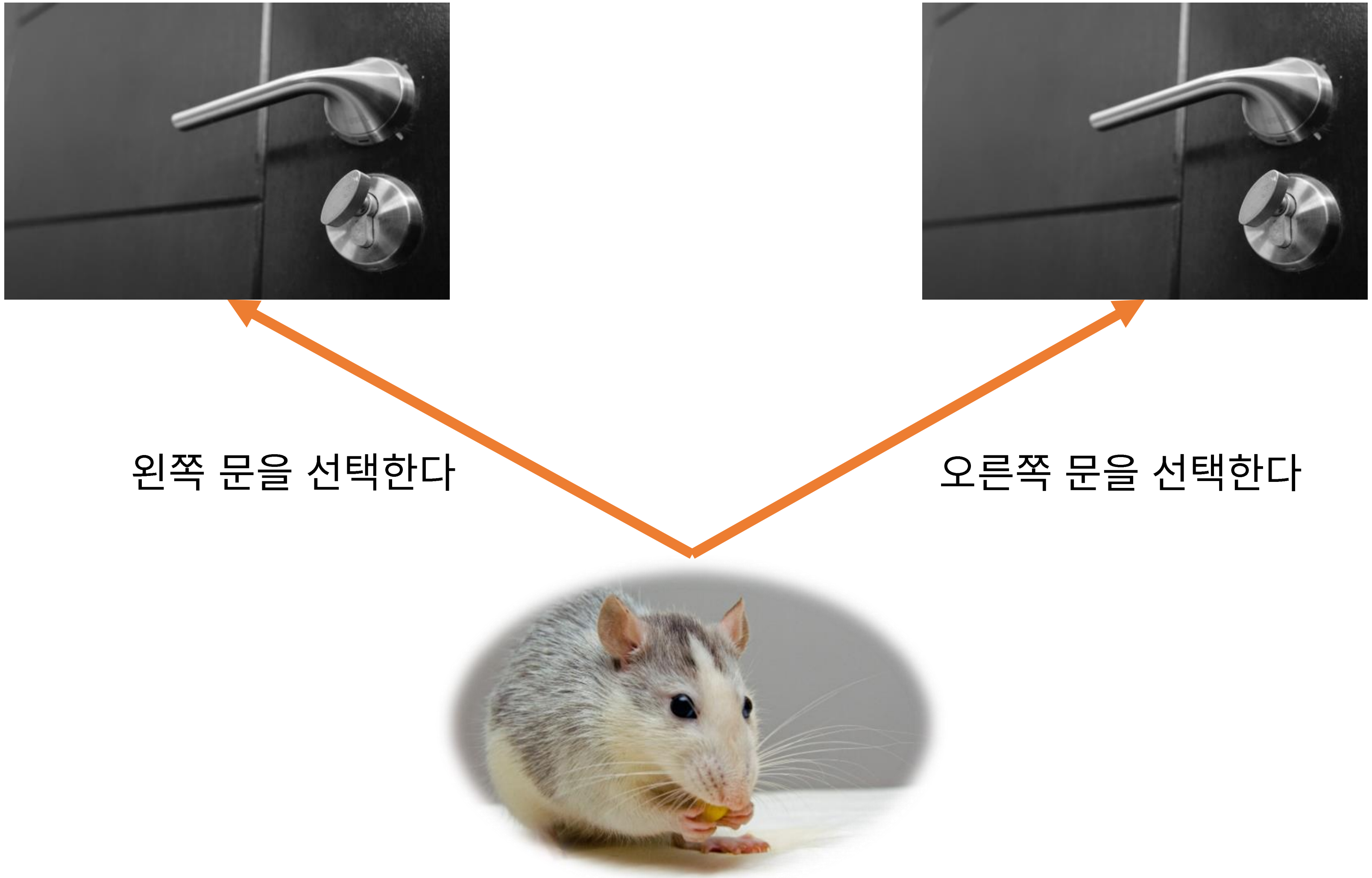
MC policy evaluation 을 활용해 $Q^{\pi}(s,a)$ 추산

정책 개선:

탐욕적 정책 개선 ??



I 강화학습의 오랜 숙적: Exploration



I 강화학습의 오랜 숙적: Exploration

왼쪽 문을 선택한다

보상 3을 받았다!

오른쪽 문을 선택한다

보상 2을 받았다!

"왼쪽 문" 은 좋은 선택이었구나!

I 찍찍이는 알 수 없지만 사실은 …

$R^{\text{왼쪽}} \sim \mathcal{N}(1.0, 2.0)$

$R^{\text{오른쪽}} \sim \mathcal{N}(5.0, 1.0)$

왼쪽 문을 선택한다

오른쪽 문을 선택한다

“탐욕적 정책 개선”은 에이전트가
새로운 선택을 할 수 없게 만듦.

Exploitation ↑ Exploration ↓

I ϵ-Greedy policy

Greedy policy (탐욕적 정책) $\pi(a|s) = \begin{cases} 1, & if\ a = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q^{\pi}(s,a) \\ 0, & otherwise \end{cases}$

ϵ-Greedy policy $\pi(a|s) = \begin{cases} \dfrac{\epsilon}{|\mathcal{A}|} + 1 - \epsilon, & if\ a = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q^{\pi}(s,a) \\ \epsilon/|\mathcal{A}|, & otherwise \end{cases}$ $|\mathcal{A}|$: 가능한 action 갯수

ϵ-Greedy policy 는 매우 간단하지만, 매우 잘 작동하는 알고리즘!

- $1 - \epsilon$ 의 확률로 “가장 좋은” 행동을 선택.
- ϵ 의 확률로 모든 가능한 행동 중 하나를 임의로 선택.

I ϵ-Greedy policy 톺아보기

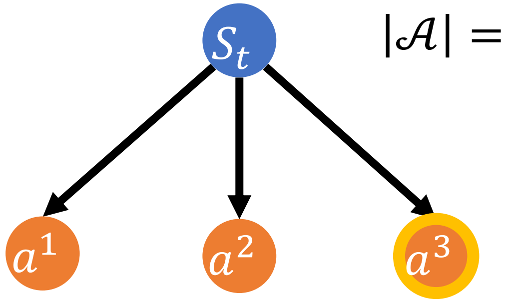
ϵ-Greedy policy

$$\pi(a|s) = \begin{cases} \frac{\epsilon}{|\mathcal{A}|} + 1 - \epsilon, & if\ a = \operatorname{argmax}_{a \in \mathcal{A}} Q^{\pi}(s,a) \\ \epsilon/|\mathcal{A}|, & otherwise \end{cases}$$

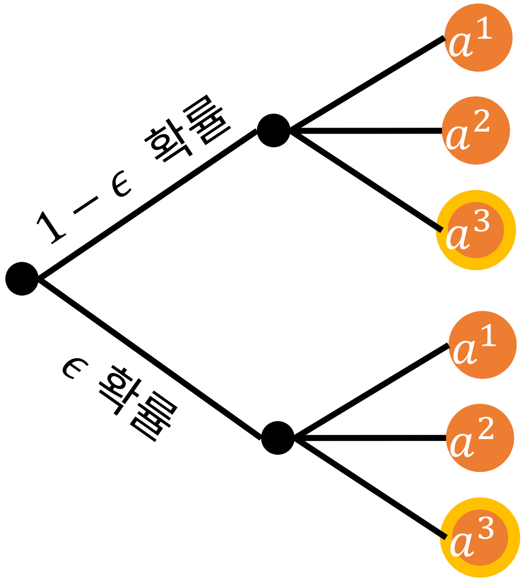
$|\mathcal{A}|$: 가능한 action 갯수

- $1-\epsilon$ 의 확률로 "가장 좋은" 행동을 선택.
- ϵ 의 확률로 모든 가능한 행동 중 하나를 임의로 선택.

$a^3 = \underset{a_t \in \mathcal{A}}{argmax}\, Q^{\pi}(s_t, a_t)$ 라고 가정



$|\mathcal{A}| = 3$



0.0

0.0

$(1-\epsilon) \times 1.0$

$\epsilon \times 1/3$

$\epsilon \times 1/3$

$\epsilon \times 1/3$

$$\pi(a|s_t) = \begin{cases} \epsilon/3, & a = a^1 \\ \epsilon/3, & a = a^2 \\ (1-\epsilon) + \epsilon/3, & a = a^3 \end{cases}$$

I ϵ-Greedy 정책개선이 정말 "정책개선"인가요?

정책 개선: $\pi' \geq \pi$ 를 만족시키는 π 을 생성

(정책함수 π 간의 대소관계를 다음과 같이 정의 한다.

$\pi' \geq \pi$ 만약 $V_{\pi'}(s) \geq V_{\pi}(s), \forall s \in \mathcal{S}$)

정리) ϵ-Greedy 정책 π 와 개선된 ϵ-Greedy 정책 π' 일때, 모든 s 에 대하여 $V^{\pi'}(s) \geq V^{\pi}(s)$ 이다.

$Q^{\pi}(s, \pi'(s))$ 의 의미?

"현재 가치함수 $Q^{\pi}(s,a)$ 에서 만약에 a 를 $\pi'(s)$ 로 고르면" 그 가치는 얼마인가?

$$
\begin{aligned}
Q^{\pi}(s, \pi'(s)) &= \sum_{a \in \mathcal{A}} \pi'(a|s) Q^{\pi}(s,a) \\
&= \sum_{a \in \mathcal{A}} \frac{\epsilon}{|\mathcal{A}|} Q^{\pi}(s,a) + (1-\epsilon) \max_{a \in \mathcal{A}} Q^{\pi}(s,a) \\
&= \frac{\epsilon}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} Q^{\pi}(s,a) + (1-\epsilon) \max_{a \in \mathcal{A}} Q^{\pi}(s,a) \\
&\geq \frac{\epsilon}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} Q^{\pi}(s,a) + (1-\epsilon) \sum_{a \in \mathcal{A}} \frac{\pi(a|s) - \frac{\epsilon}{|\mathcal{A}|}}{1-\epsilon} Q^{\pi}(s,a) \\
&= \sum_{a \in \mathcal{A}} \pi(a|s) Q^{\pi}(s,a) \\
&= V^{\pi}(s)
\end{aligned}
$$

I ϵ-Greedy 정책개선이 정말 정책개선인가요?

정책 개선: $\pi' \geq \pi$ 를 만족시키는 π 을 생성

정리) ϵ-Greedy 정책 π 와 개선된 ϵ-Greedy 정책 π' 일때, 모든 s 에 대하여 $V^{\pi'}(s) \geq V^{\pi}(s)$ 이다.

$$
\begin{aligned}
Q^{\pi}(s, \pi'(s)) &= \sum_{a \in \mathcal{A}} \pi'(a|s) Q^{\pi}(s,a) \\
&= \sum_{a \in \mathcal{A}} \frac{\epsilon}{|\mathcal{A}|} Q^{\pi}(s,a) + (1-\epsilon) \max_{a \in \mathcal{A}} Q^{\pi}(s,a) \\
&= \frac{\epsilon}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} Q^{\pi}(s,a) + (1-\epsilon) \max_{a \in \mathcal{A}} Q^{\pi}(s,a) \\
&\geq \frac{\epsilon}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} Q^{\pi}(s,a) + (1-\epsilon) \sum_{a \in \mathcal{A}} \frac{\pi(a|s) - \frac{\epsilon}{|\mathcal{A}|}}{1-\epsilon} Q^{\pi}(s,a) \\
&= \sum_{a \in \mathcal{A}} \pi(a|s) Q^{\pi}(s,a) \\
&= V^{\pi}(s)
\end{aligned}
$$

$$
\max_{a \in \mathcal{A}} Q^{\pi}(s,a) \geq \sum_{a \in \mathcal{A}} \frac{\pi(a|s) - \frac{\epsilon}{|\mathcal{A}|}}{1-\epsilon} Q^{\pi}(s,a) \ ???
$$

FAST CAMPUS
ONLINE
박준영 강사.

| ϵ-Greedy 정책개선이 정말 정책개선인가요?

$w_i \geq 0$ 이고, $\sum w_i = 1$ 인 **임의의** w_i 를 가지고 있다고 했을 때

$$\max_i x_i \geq \sum w_i x_i$$

등호는 $w_{i^*} = 1.0, i^* = \text{argmax}_i\, x_i$ 일 때 성립.

I ϵ-Greedy 정책개선이 정말 정책개선인가요?

$$\pi(a|s) = \begin{cases} \frac{\epsilon}{|\mathcal{A}|} + 1 - \epsilon, & if\ a = \operatorname{argmax}_{a \in \mathcal{A}} Q^{\pi}(s,a) \\ \epsilon / |\mathcal{A}|, & otherwise \end{cases}$$

$$\sum_{a \in \mathcal{A}} \frac{\pi(a|s) - \frac{\epsilon}{|\mathcal{A}|}}{1 - \epsilon} Q^{\pi}(s,a) = \sum_{a \in \mathcal{A}} w(a) Q^{\pi}(s,a) \le \max_{a \in \mathcal{A}} Q^{\pi}(s,a)$$

"임의의 가중치"

$$w(a) \stackrel{\text{def}}{=} \frac{\pi(a|s) - \frac{\epsilon}{|\mathcal{A}|}}{1 - \epsilon}$$

(1) 모두 0 이상인가?

$$w(a) = \begin{cases} 0, & a \neq a^* \\ 1, & a = a^* \end{cases}$$

(2) 합이 1.0 인가?

$$\sum_{a \in \mathcal{A}} w(a) = \sum_{a \in \mathcal{A}} \frac{\pi(a|s) - \frac{\epsilon}{|\mathcal{A}|}}{1 - \epsilon} = \frac{1 - \sum_{a \in \mathcal{A}} \frac{\epsilon}{|\mathcal{A}|}}{1 - \epsilon} = \frac{1 - \epsilon}{1 - \epsilon} = 1$$

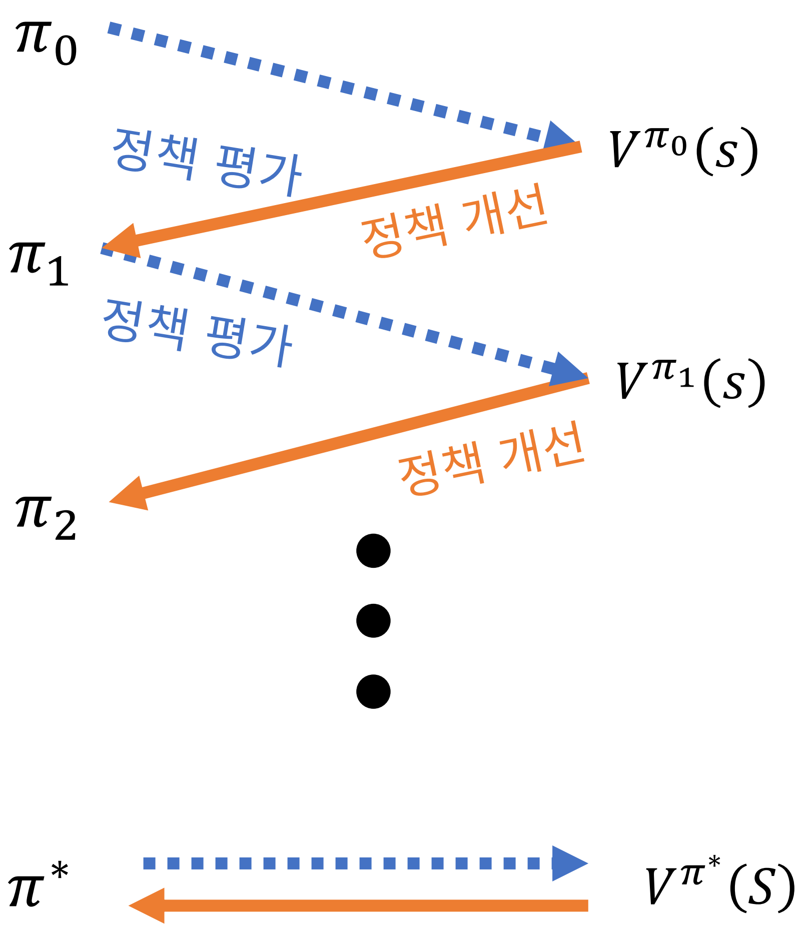
I MC policy evaluation + ϵ-탐욕적 개선 !

정책 평가:

MC policy evaluation 을 활용해 $Q^{\pi}(s,a)$ 추산

정책 개선:

ϵ-탐욕적 정책 개선



I GLIE 조건

"ϵ-탐욕적 정책" 의 단점

- ϵ 의 확률로 임의의 행동을 해야 한다!

그러면 혹시 학습 초기에는 ϵ 상대적으로 높게 하고 학습이 진행됨에 따라서 $\epsilon \rightarrow 0$ 으로 하면 어떨까??

- 학습이 진행되면 Greedy 정책으로 수렴한다.
- 어떤 조건이 필요할까?
  - **GLIE** (Greedy in the Limit of Infinite Exploration)
    - 모든 $N(s,a) \rightarrow \infty$ 이 되도록 ϵ를 학습진행에 따라 스케쥴링한다.
    - Ex) $\epsilon_k = 1/k$, k 는 에피소드 인덱스 (실제로 자주 쓰이는 트릭!)

I GLIE Monte-Carlo 제어

매 에피소드마다

Monte-Carlo
정책 평가

Incremental
MC policy evaluation

(s,a) (처음) 때마다,

$N(s,a) \leftarrow N(s,a) + 1$

$$Q(s,a) \leftarrow Q(s,a) + \frac{1}{N(s,a)}\left(G_t - Q(s,a)\right)$$

(현실에서는, $N(s)$ 을 세는 것조차 어려움)

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left(G_t - Q(s,a)\right)$$

수렴까지
반복

GLIE ϵ-greedy
정책개선

ex)

- $\epsilon \leftarrow \frac{1}{k}$
- $\pi \leftarrow \epsilon - greedy(Q)$

강화학습은 하이퍼 파라미터와의 싸움.

Monte-Carlo
정책 평가

$$Q(s,a) \leftarrow Q(s,a) + \alpha\big(G_t - Q(s,a)\big)$$

GLIE ϵ-greedy
정책개선

- $\epsilon \leftarrow \frac{1}{k}$
- $\pi \leftarrow \epsilon - greedy(Q)$

α 혹은 ϵ 같이 학습에 결과에 영향을 주지만, 학습 알고리즘에 의해 결정되지 않는 것들을 기계학습에서는 Hyperparameter 라고 부릅니다

성공적인 강화학습, 기계학습 모델을 만들기 위해서는 Hyperparameter 튜닝을 통해서 최적의 Hyperparameter 를 찾아내는 것이 관건!

(반대로 parameter는 학습알고리즘에 의해 최적값이 찾아지는 값들을 표현. 다음 파트에서 좀 더 자세히 다룸)