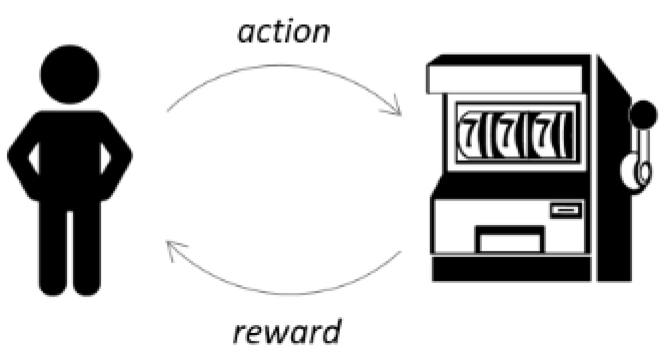
Multi-armed Bandit 문제

정 태 수
고려대학교 산업경영공학부
tcheong@korea.ac.kr

k-armed Bandit 문제

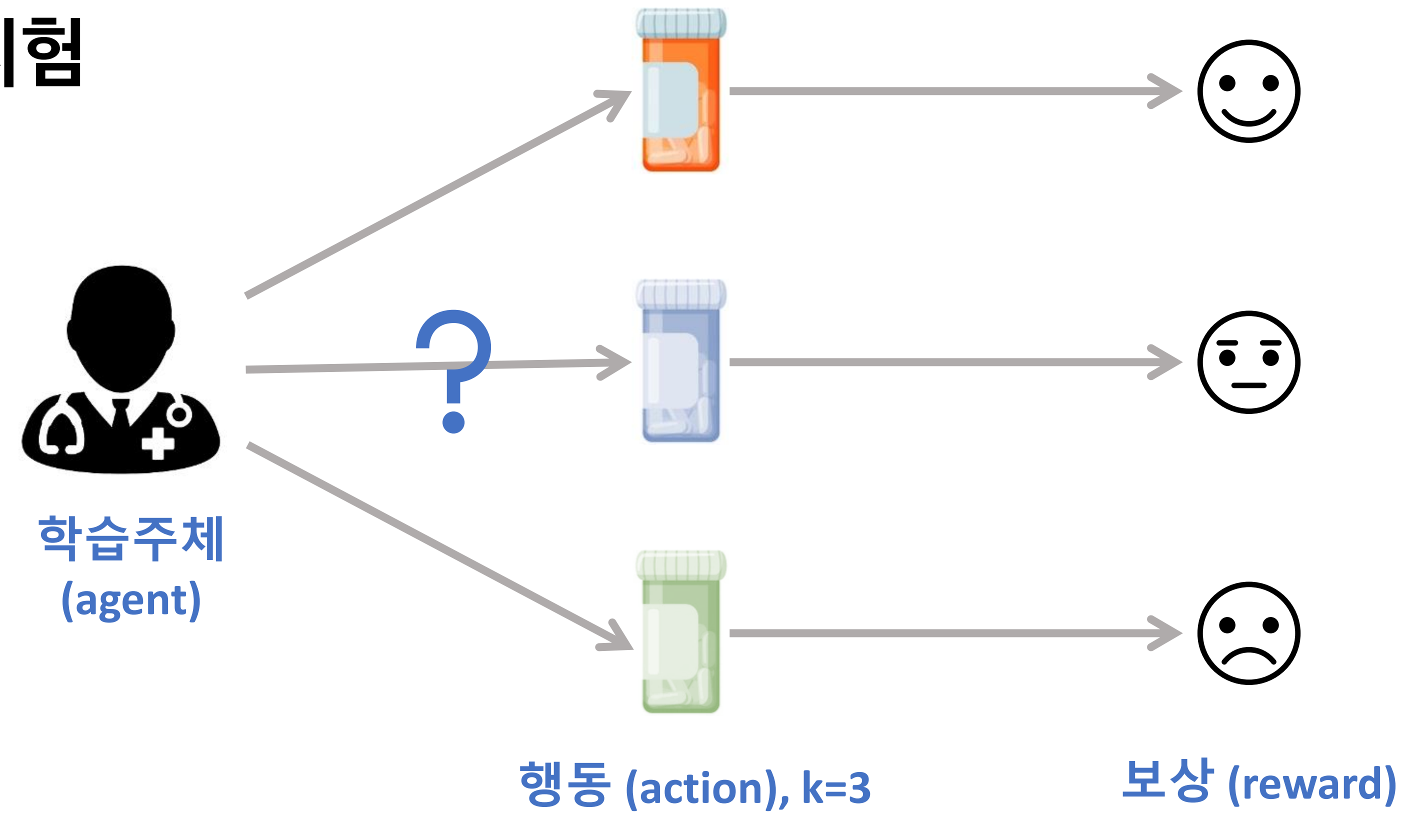
- 각 bandit machine에 대한 상태(state) 정보는 부재
- 행동(action)에 대한 즉각적인 보상 (reward)



https://www.slideshare.net/stjunya/large-scale-bandit-poblems-and-kwik-learning

k-armed Bandit 문제

- 임상시험

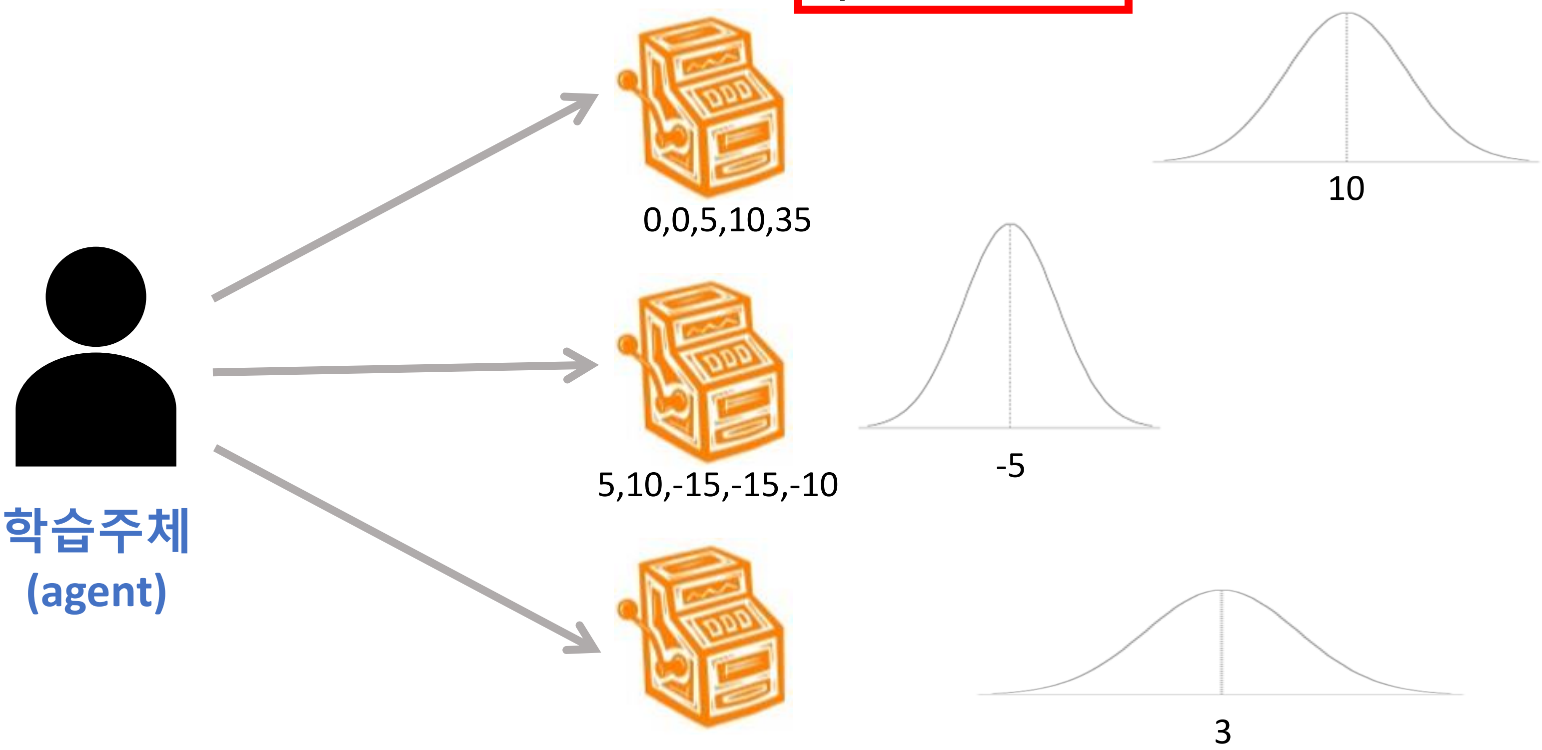


- k-armed bandit 문제는 학습주체(agent)가 k개의 행동(action) 중 하나를 선택하고 선택한 행동에 따라 보상(reward)을 받는 일련의 과정을 통해 일정 기간동안 취득한 보상의 총합에 대한 기대값을 최대화하도록 어떠한 행동들을 취할지 결정하는 문제

k-armed Bandit 문제

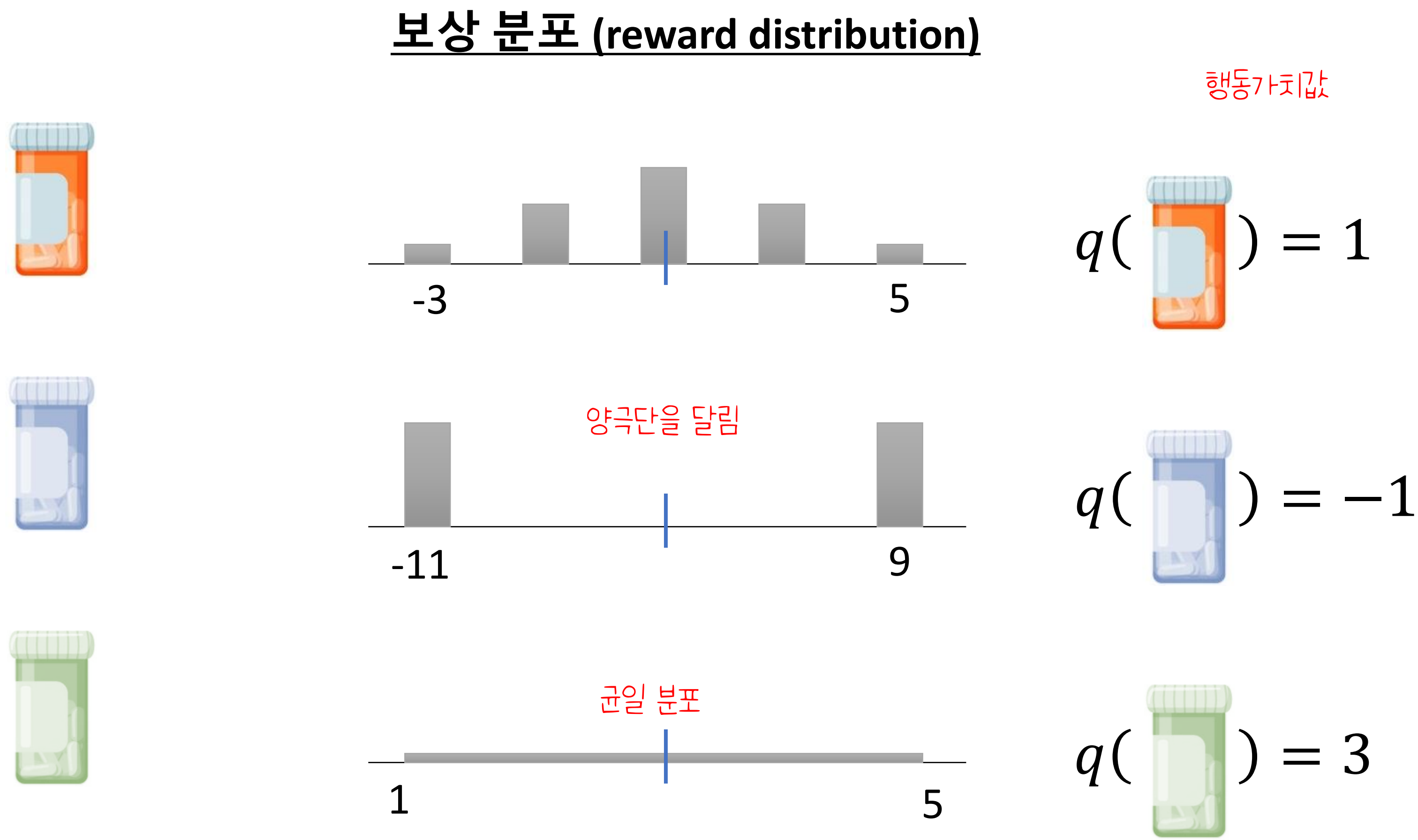
- **행동 가치 (action values)**
  - 특정 시점에서 어떠한 행동을 취했을 때의 보상에 대한 기댓값

$$q(a) = E[R_t | A_t = a] = \boxed{\sum_r p(r|a)r}$$



k-armed Bandit 문제

- **행동 가치**



k-armed Bandit 문제

- 불행히도 대부분의 경우 최적의 행동 가치에 대한 정보가 부재함

- 따라서 일련의 반복된 실험을 통해 얻은 정보를 기반으로 행동가치를 추정할 수 있음
  - 그럼 어떻게 $q(a)$ 를 추정할 수 있을까? 표본평균 방법 (sample-average method)

$$Q_t(a) = \frac{t\text{시점까지 행동 } a\text{를 선택 시 취득한 보상 합}}{t\text{시점까지 행동 } a\text{를 선택한 횟수}}$$

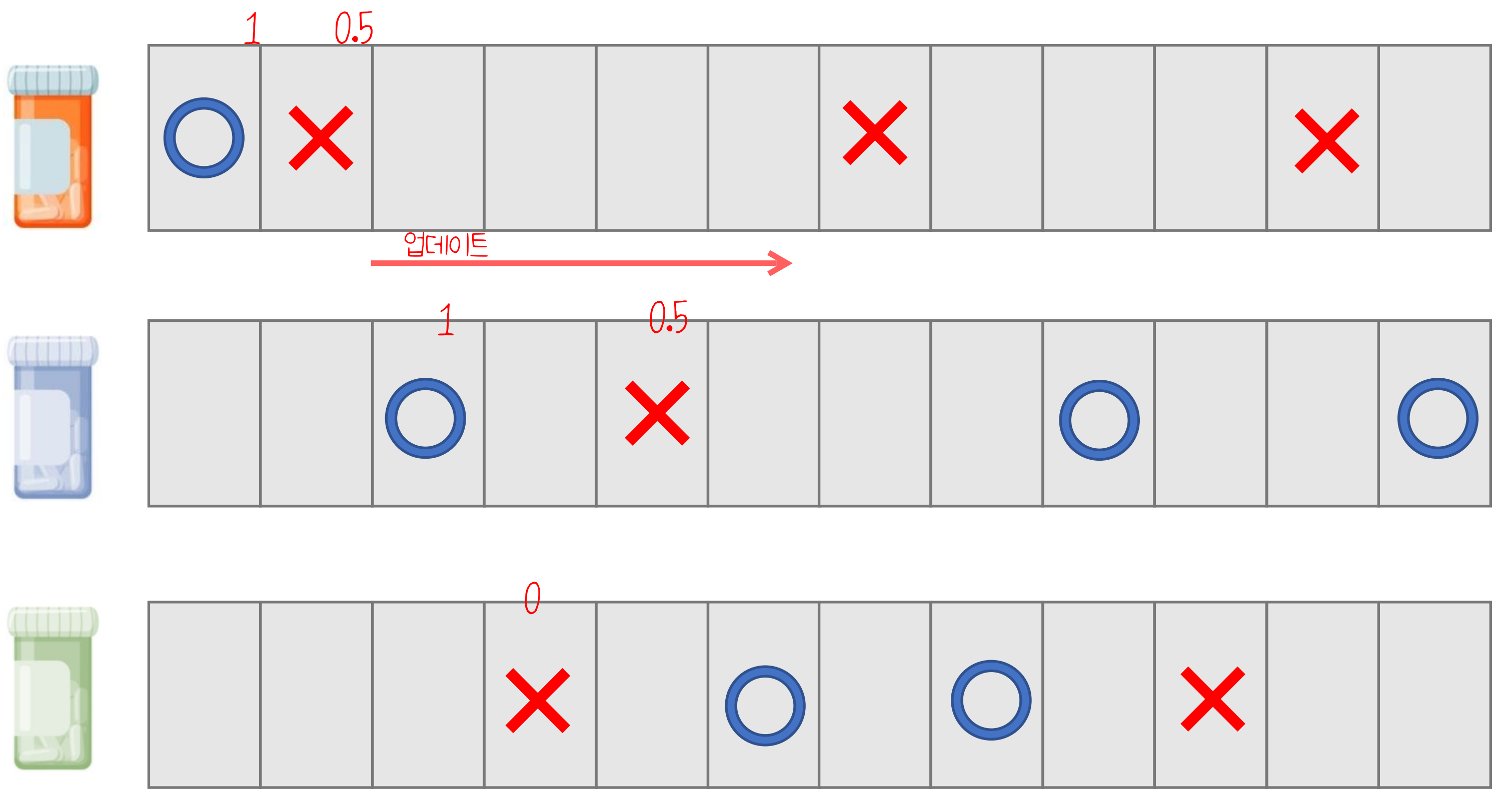
행동 가치를 추정하는 가장 단순한 방법

표본평균 방법
(Sample-mean method)

$$Q_t(a) = \frac{t\text{시점까지 행동 } a\text{를 선택 시 취득한 보상 합}}{t\text{시점까지 행동 } a\text{를 선택한 횟수}}$$

k-armed Bandit 문제

- **임상시험 예시**
  - 치료가 성공하면 보상이 1, 아니면 0



$Q_{12}(R) = 0.25$

$Q_{12}(B) = 0.75$

$Q_{12}(G) = 0.5$

파란색약을 많이 투여하려고 시도하게 될것이다.

k-armed Bandit 문제

- **행동가치 추정치 $Q_t(a)$를 보다 효율적을 계산할 수 있는 방법이 없을까?**
  - 증분 업데이트 규칙(Incremental update rule)

총 보상의 합

$$\boxed{Q_{n+1} = \frac{1}{n}\sum_{i=1}^{n} R_i} = \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) = \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right)$$

$$= \frac{1}{n}(R_n + (n-1)Q_n) = \frac{1}{n}(R_n + nQ_n - Q_n) = Q_n + \frac{1}{n}(R_n - Q_n)$$

$$Q_{n+1} = Q_n + \frac{1}{n}(R_n - Q_n)$$

전체 시도 분에 누적되는 보상들의 합, 이것들을 매 단계마다 매번 계산하는 것이 아니라

그냥 곱해줌 어차피 1

n-1번째 시도까지의 행동가치의 추정치

총 보상의 합

$$Q_{n+1} = \boxed{\frac{1}{n}\sum_{i=1}^{n} R_i} = \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) = \frac{1}{n}\left(R_n + (n-1)\boxed{\frac{1}{n-1}\sum_{i=1}^{n-1} R_i}\right)$$

아래 Qn을 의미한다

최근 보상

처음부터 이전 단계까지 얻었던 보상의 합

$$= \frac{1}{n}(R_n + (n-1)\boxed{Q_n}) = \frac{1}{n}(R_n + nQ_n - Q_n) = Q_n + \frac{1}{n}(R_n - Q_n)$$

$$\boxed{Q_{n+1}} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) = \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right)$$

$$= \frac{1}{n}(R_n + (n-1)\, Q_n) = \frac{1}{n}(R_n + nQ_n - Q_n) = \boxed{Q_n + \frac{1}{n}(R_n - Q_n)}$$

보상의 합을 매번 계산하는 것이 아닌

Q값을 트래킹

기존 정보와 새로운 정보의 차이 반영 → Q 값 업데이트

n번 시도의 보상값

$$\boxed{Q_{n+1}} = Q_n + \frac{1}{n}(R_n - Q_n)$$

정보 업데이트

모든 단계에서 얻은 보상의 가중치가 동일하다는 가정하에 평균치 계산

우리가 접하는 상황들은 시간에 따라 보상의 가치가 달라 질 수 있음 – 비정상적 상황

k-armed Bandit 문제

- **간략한 Bandit 알고리즘**

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Repeat forever:

$$A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

$R \leftarrow bandit(A)$

$N(A) \leftarrow N(A) + 1$

$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\left[R - Q(A)\right]$

누적되어있는 보상값을 계속 저장하고 있는 것이 아니라
이런 Q값을 가지고 계속 업데이트를 해 나가면서 실질적으로 업데이트를 하는 것 확인 할 수 있다.

k-armed Bandit 문제

- **Nonstationary (비정상성 문제)**
  - Step-size 인자 조절 방식

$$Q_{n+1} = Q_n + \frac{1}{n}(R_n - Q_n)$$

매 단계의 보상의 중요도가 모두 같다는 가정

일반화

$$\begin{aligned} Q_{n+1} &= Q_n + a_n(R_n - Q_n) \\ &= a_n R_n + (1 - \alpha_n)Q_n \end{aligned}$$

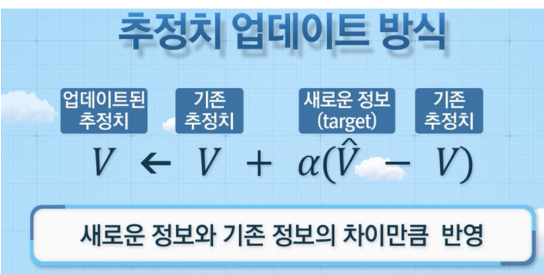
매시점 마다 보상의 가치가 달라져도 유연한 업데이트가 가능하다

(참고) 추정치 업데이트 방식

- **정보의 업데이트 방식**

업데이트된 추정치 / 기존 추정치 / 새로운 정보 (target) / 기존 추정치

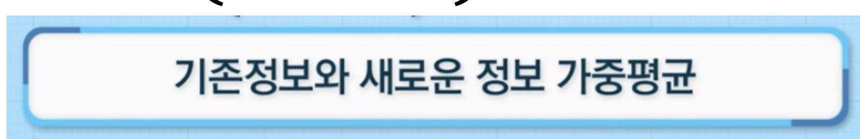
$$V \leftarrow V + \alpha(\hat{V} - V)$$



혹은

기존정보와 새로운 정보 가중평균

$$(1-\alpha)V + \alpha\hat{V}$$



Convex combination

THANK YOU

tcheong@korea.ac.kr