Chapter. 05

어깨넘어 배워서 세상 조종하기

Off-policy MC control

FAST CAMPUS ONLINE

강화학습 A-Z

강사. 박준영

지난 이야기…

"강화학습 문제"

MDP

<Ch 01. 마르코프 결정과정>

"강화학습 문제의 풀이기법"

<Ch 02. 동적계획법>

DP

환경에 대해서 알 때

- 정책 반복 (Policy iteration)
  - 정책 평가
  - 정책 개선
- 가치 반복 (Value iteration)
- 비동기 DP

\+ (상대적으로) 문제를 해결하기 쉬움

\+ 매우 효율적임

\- 현실적이지 않음

MC TD

환경에 대해서 모를 때

- 정책 평가 (Policy evaluation)
  - Monte Carlo PE
  - Temporal difference PE

<Ch 03. 모델없이 세상 알아가기>

- 정책 개선 (Policy improvement)
  - MC Control
  - SARSA Control

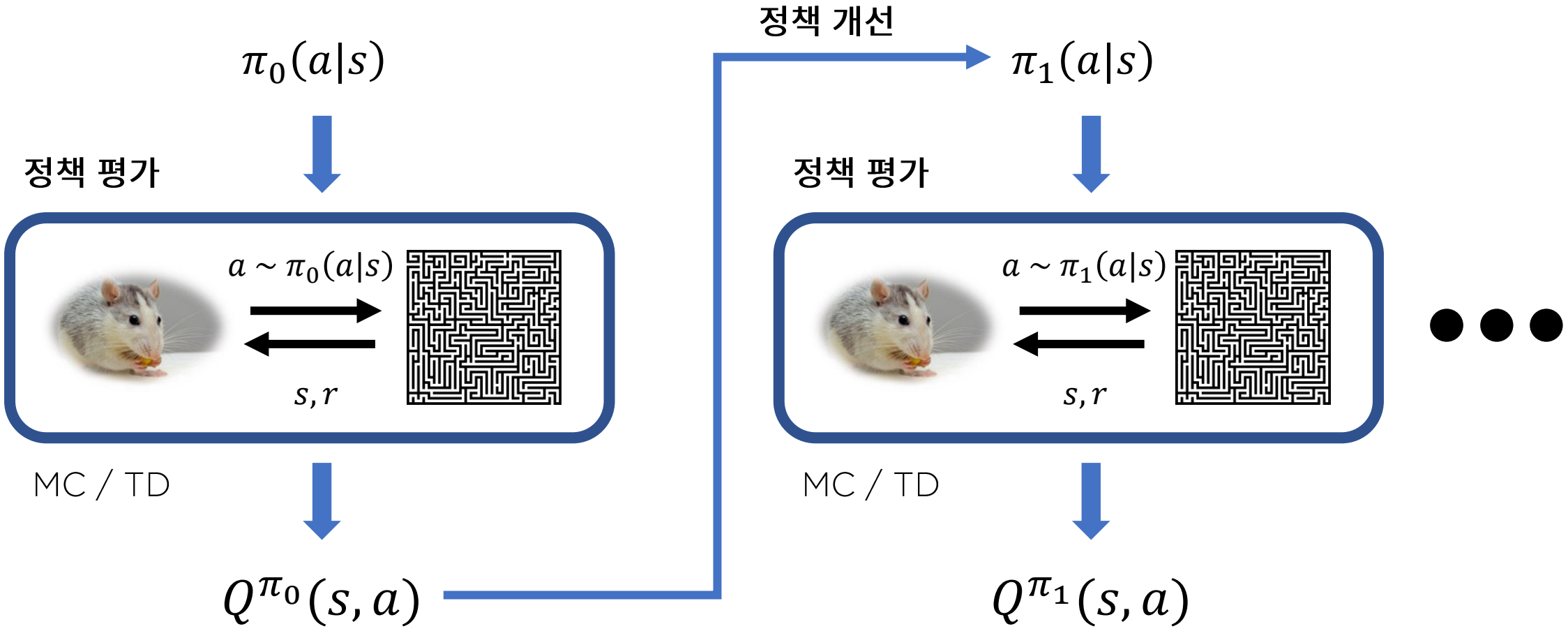
<Ch 04. 모델없이 세상 조종하기>

\- 문제를 해결하기 어려움

\- (DP에 비해) 효율성이 떨어짐

\+ 현실의 문제의 상황에 적용 가능

I 지금까지 배운 강화학습 기법의 템플릿



정책 개선마다 새롭게 Q^{π_k} 를 추산하기 위해서 **또 새롭게** 샘플들을 모음.

I 사람의 학습?

타산지석:

다른 사람의 행동이나 말도 자신의 지식과 인격을 수행하는데 도움이 됨.

반면교사:

사람이나 사물 따위의 부정적인 면에서 얻는 깨달음이나 가르침을 주는 대상을 이르는 말

내가 한 행동이 아니어도 그것에서 무언가를 배울 수 있다!

- 강화학습 찍찍이도 가능할까요?

Off-Policy 학습

(일반적으로 Target policy 라고 불립니다)

우리의 목적: 주어진 정책 $\pi(a|s)$ 에 대한 $Q^{\pi}(s,a)$ 를 계산하는 것.

임의로 정한 행동 정책 $\mu(a|s)$ 으로 구한 episode에서도 $Q^{\pi}(s,a)$를 추산 할 수 있을까?

(Behavior policy)

$$\{S_1, A_1, R_2, \dots, S_T\} \sim \mu$$

만약 가능하다면!

- 다른 좋은 정책 (예를 들면 사람의 정책) 으로 부터 얻어진 샘플에서 학습 가능
- 정책 개선과정에서, 이전 정책의 가치평가를 위해 모은 샘플을 재활용 가능
- $\pi(a|s)$은 계속적으로 최적정책을 찾지만, $\mu(a|s)$ 조금 더 exploration을 하게 만들 수 있음

더 짧은 시간내에 더 좋은 RL 에이전트를 학습시킬 가능성이 있음.

Off-Policy 학습

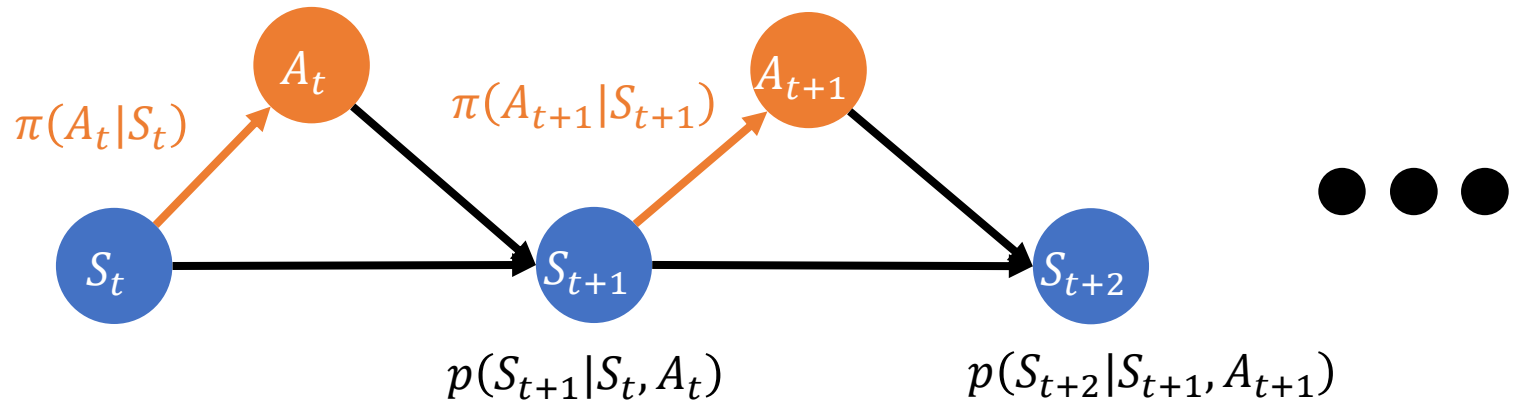
우리의 목적:

$Q^{\pi}(s,a)$를 효율적으로 잘 추산하는 것! ($Q^{\pi}(s,a)$ 을 알면 정책을 만들 수 있음)

$$Q^{\pi}(s,a) \overset{\text{def}}{=} \mathbb{E}_{\pi}[G_t \mid S_t = s, A_t = a]$$
$$= \mathbb{E}_{\pi}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ \ldots \mid S_t = s, A_t = a]$$

- 상태와 행동 그리고 보상의 Trajectory가 모두 π에 따라서 결정!
- $\mathbb{E}_{\pi}[\cdot]$ 가 무슨 의미? -> 정책 π 를 따르면 발생할 Trajectory의 분포에 대한 $[\cdot]$ 의 평균치.

π를 정책으로 할 때 Trajectory 의 확률



시점 t 에서 상태 S_t 가 주어지고 정책 π를 사용할 때, t시점 이후의 Trajectory의 확률

$$p(A_t, S_{t+1}, A_{t+1}, \dots, S_T | S_t, A_{t:T-1} \sim \pi) = \pi(A_t|S_t)p(S_t|S_t, A_t)\pi(A_{t+1}|S_{t+1}) \dots p(S_T|S_{T-1}, A_{T-1})$$
$$= \prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k)$$

시점 t 에서 상태 S_t 가 주어지고 또 다른 정책 μ 를 사용할 때, t시점 이후의 Trajectory의 확률

$$p(A_t, S_{t+1}, A_{t+1}, \dots, S_T | S_t, A_{t:T-1} \sim \mu) = \prod_{k=t}^{T-1} \mu(A_k|S_k)p(S_{k+1}|S_k, A_k)$$

I 서로 다른 정책 π 와 μ 의 Trajectory 의 비율

$$\rho_{t:T-1} \overset{\text{def}}{=} \frac{p(A_t, S_{t+1}, A_{t+1}, \dots, S_T | S_t, A_{t:T-1} \sim \pi)}{p(A_t, S_{t+1}, A_{t+1}, \dots, S_T | S_t, A_{t:T-1} \sim \mu)} = \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k) p(S_{k+1}|S_k, A_k)}{\prod_{k=t}^{T-1} \mu(A_k|S_k) p(S_{k+1}|S_k, A_k)} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}$$

- 각 MDP Trajectory 의 확률은 MDP 구조 $p(S_{k+1}|S_k, A_k)$ 에 영향을 받지만,

두 개의 Trajectory 의 비율 $\rho_{t:T-1}$ 는 MDP 구조에 영향을 받지 않는다!

- 결과적으로 우리는 MDP 모델을 몰라도 이 값을 계산할 수 있다!

Importance sampling

$$\begin{aligned}\mathbb{E}_{x\sim P}[f(x)] &= \sum_{x\in X} p(x)f(x) \\ &= \sum_{x\in X} q(x)\frac{p(x)}{q(x)}f(x) \\ &= \mathbb{E}_{x\sim Q}\left[\frac{p(x)}{q(x)}f(x)\right]\end{aligned}$$

Importance sampling 원래 목적:
$p(x)$ 의 함수 형태는 알지만 샘플을 만들기 어려울 때,
$q(x)$ 라는 샘플링하기 쉬운 분포를 활용해 $\mathbb{E}_{x\sim P}[f(x)]$ 를 추산하는 기법.

P: 이산확률분포 Q: 이산확률분포 라고 가정하여 표현. 연속확률분포의 경우 적분으로 표현됨.

Off-policy MC

$$\mathbb{E}_{x\sim P}[f(x)] = \sum_{x\in X} p(x)f(x)$$
$$= \sum_{x\in X} q(x)\frac{p(x)}{q(x)}f(x)$$
$$= \mathbb{E}_{x\sim Q}\left[\frac{p(x)}{q(x)}f(x)\right]$$

$$Q^{\pi}(s,a) \overset{\text{def}}{=} \mathbb{E}_{\pi}[G_t|\, S_t = s, A_t = a]$$
$$= \mathbb{E}_{\mu}[\rho_{t:T-1}G_t|\, S_t = s, A_t = a]$$

$\mathbb{E}_{\pi}[G_t|\, S_t = s, A_t = a]$: 현재상태 S_t와 행동 A_t로 행동하고 그 뒤로 π로 인해서 발생할 Trajectory 에서 얻을 G_t의 기댓값

Off-policy MC

- μ 임의의 행동 정책함수로 에피소드 생성
- $G_t^{\pi/\mu} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} G_t$ 계산
- MC update 수행

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left(G_t^{\pi/\mu} - Q(s_t, a_t) \right)$$

- $\prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}$ 때문에, $G_t^{\pi/\mu}$ 의 분산이 커질 수 있음.
- $\pi(A_k|S_k) \neq 0$ 이면, $\mu(A_k|S_k) \neq 0$ 이어야만 함.

분산을 최소화 하는 optimal 한 μ 를 만들 수 있지만,
(1) Optimal μ 를 찾는데 노력이 필요.
(2) Optimal μ 는 exploration에 도움이 되지 않을 수도 있음

(현실적으로 많이 사용되지 않는 이유 ☺)

마무리 …

환경에 대해서 모를 때

On-policy learning

1. 현재 정책 π 로 세상에서, 데이터를 얻어 $Q^{\pi}(s,a)$를 추산.
2. 정책 개선기법 (e.g. $\epsilon-$ 탐욕적 정책개선) 으로 개선된 정책 π' 을 구함
3. (1), (2) 를 π' 가 수렴할 때까지 반복.

e.g.) Monte-Carlo Control, SARSA Control

- 이전 정책을 평가하기 위해 사용한 데이터를 재활용할 수 있을까?
- 다른 정책으로 얻어진 데이터를 학습에 사용할 수 있을까?

Off-policy learning

1. 임의의 행동 정책 μ 로 데이터를 얻음
2. Importance sampling 을 활용해서 타겟 정책 $Q^{\pi}(s,a)$을 추산
3. 정책 개선 수행

e.g) off-policy MC control