Chapter. 03

모델없이 세상 알아가기

Temporal Difference (TD) 정책추정

FAST CAMPUS ONLINE

강화학습 A-Z I

강사. 박준영

지난 이야기…

- 모델 free 가치 추산 알고리즘
  - **Monte Carlo (MC) 방식**
  - Temporal-difference (TD) 방식
- Monte Carlo 추정
- Monte Carlo 정책 추정
  - $V^{\pi}(s)$ 추정
  - $Q^{\pi}(s, a)$ 추정

I DP 와 MC 기법의 장단점

동적 계획법

장점: 현재 알고있는 값을 활용해 모르는 값을 추정.
(각 상태와 행동의 관계를 최대한 활용해 계산 량을 줄임)

단점: 환경에 대한 모델을 필요로 함

몬테 카를로 기법

장점: 환경에 대한 선험적 지식이 필요 없음,
시간이 충분히 주어지면 (이론적으로) 정확한 값을 계산할 수 있음.

단점: 각 상태와 행동의 관계에 대해서 전혀 활용하지 않음

Temporal-difference (TD) 기법

장점: 현재 알고있는 값을 활용해 모르는 값을 추정.
(각 상태와 행동의 관계를 최대한 활용해 계산량을 줄임)
환경에 대한 선험적 지식이 필요 없음

단점: (현실적으로) TD 기법은 불편 추정량이 아님.

어느 정도의 추산 오차가 발생할 수 있음

Incremental MC 돌아보기

$$V(s) \leftarrow V(s) + \alpha\big(G_t - V(s)\big)$$

MC기법이 G_t결정하는 방식: $G_t \overset{\text{def}}{=} R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ \ldots + \gamma^{T-1} R_T$

Temporal-difference (TD) 기법

$$V(s) \leftarrow V(s) + \alpha\big(G_t - V(s)\big)$$

MC기법이 G_t결정하는 방식: $G_t \overset{\text{def}}{=} R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ \ldots + \gamma^{T-1} R_T$

가장 간단한 TD 학습 알고리즘, **TD(0)** 의 경우: $G_t \overset{\text{def}}{=} R_{t+1} + \gamma V(S_{t+1})$

- $R_{t+1} + \gamma V(S_{t+1})$ 를 **TD target** 이라 부름. → $V(s)$ 를 **TD target** 에 가까워지게 조정
- $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ 를 **TD error** 라고 부름. → $V(s)$ 과 **TD target** 얼마나 차이 나는가?

(벨만 에러와 비슷하죠?)
DP 에서 prioritized sweeping 을 할때 Bellman error 가 기준이 됐었는데?

I MC vs. TD

	적용 가능 Episode종류	Bias/Variance	Markovian 특성 활용 여부
MC 기법	Terminating episode	편향 X / 분산 ↑	X
TD 기법	Terminating episode + Continuing episode	편향 O / 분산 ↓	O

- 편향/분산 관계

MC 기법은 편향이 없기 때문에, 수 많은 시행을 거치면 참 V^{π}, Q^{π} 을 찾을 수 있다.
하지만, (TD 에 비해) 추정치가 높은 분산을 가지기 때문에 좋은 추정을 얻기 위해서는 많은 시행이 필요.

일반적인 TD 기법은 알고리즘이 편향을 내재함. 따라서 시행횟수와 무관하게 V^{π}, Q^{π} 의 추정에 오류가 있다.
반면에, (MC에 비해) 추정치의 분산이 낮기 때문에, 빠른 속도로 충분히 괜찮은 추정치를 얻을 수 있다.

- Markovian 특성 활용 여부 (일반적으로 상태 및 보상함수는 사용자가 정의, 따라서 상태의 정의에 따라 강화학습 문제가 엄밀한 MDP가 아닐 수도 있음.)

MC 기법은 주어진 문제가 정확하게 Markovian이 **아니어도** 정확한 추산치를 계산할 수 있다. 하지만 효율성이 떨어짐

TD 기법은 주어진 문제가 정확하게 Markovian이 **아니면** 정확한 추산치를 계산 할 수 없다. 하지만 빠르게 추산할 수 있음

I Full-width back up vs. Sample-back up

DP 강의에서 남았던 두개의 문제

Q1) 과연 우리는 진짜로 **큰 문제**를 주어진 시간 내에 DP 로 풀 수 있을 것인가?

A1) ??

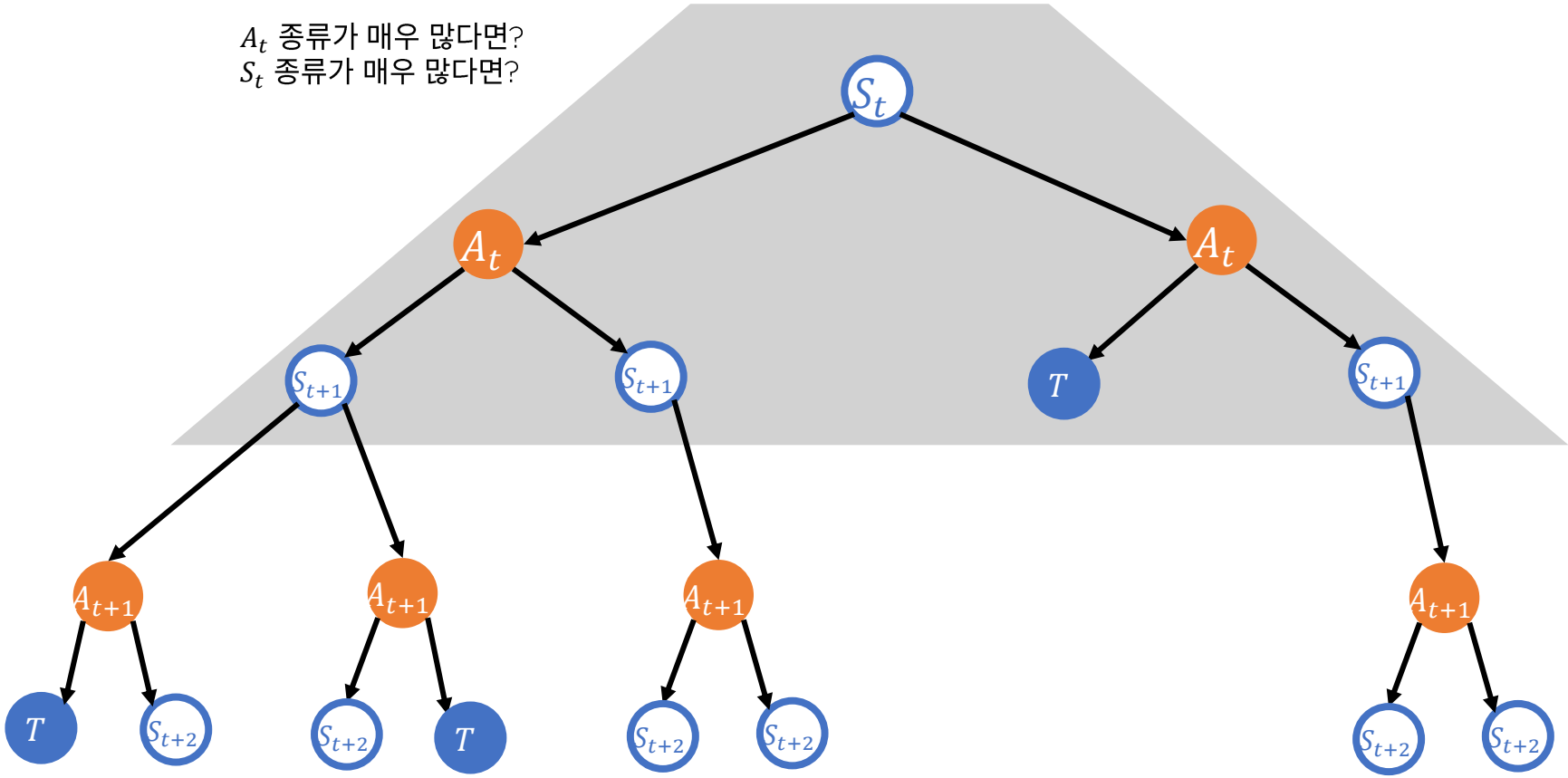
Q2) 현실에서 R, P 로 표현되는 문제의 정보를 언제나 알 수 있을까?

A2) MC, TD를 활용해서 해결

I “큰 문제”에서 Full-width backup

DP 를 활용한 Policy evaluation : $V^{\pi}(S_t) \leftarrow \mathbb{E}_{\pi}[R_{t+1} + \gamma V^{\pi}(S_{t+1})]$

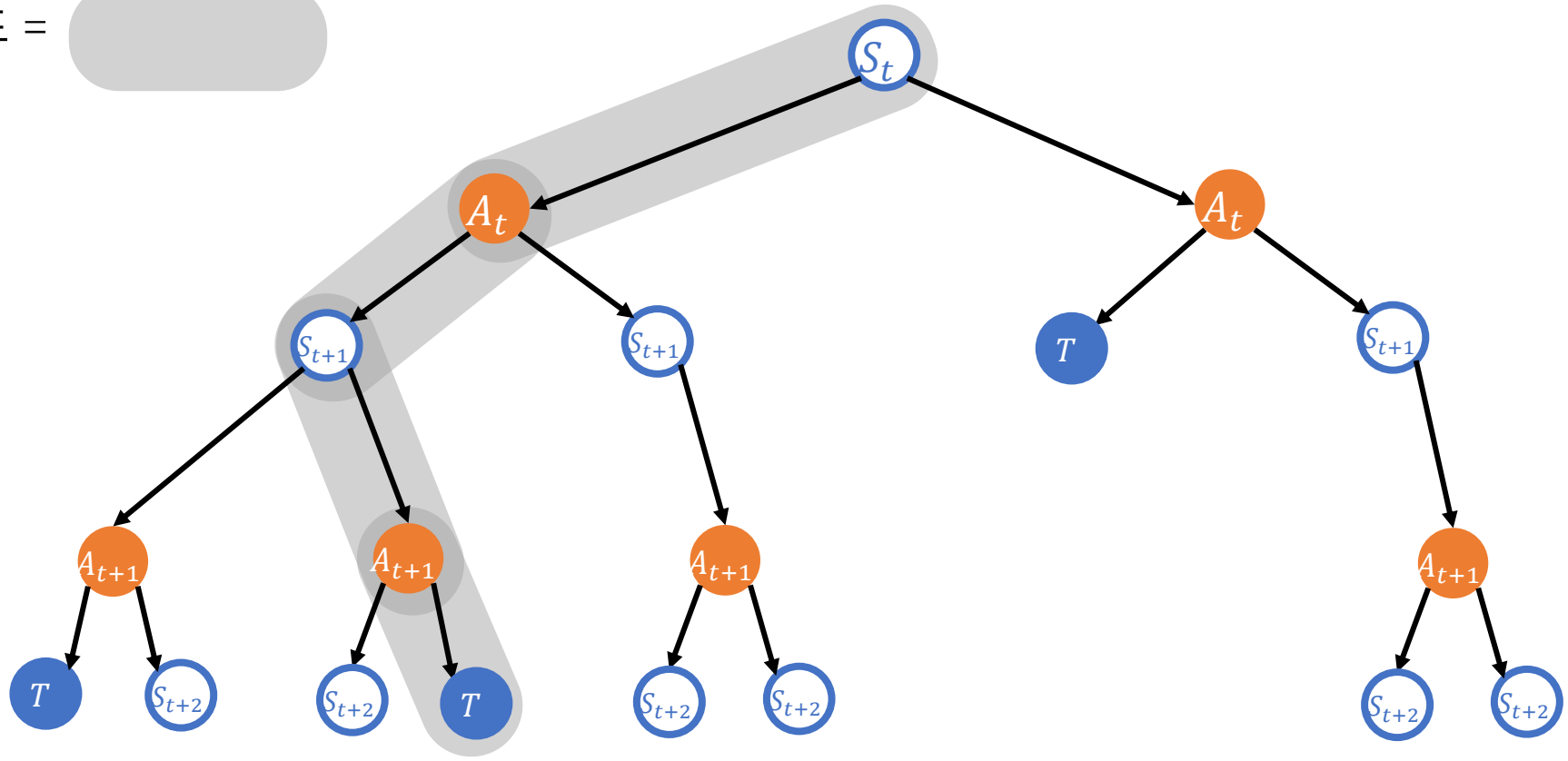
A_t 종류가 매우 많다면?
S_t 종류가 매우 많다면?



MC 기법을 활용한 Sample backup

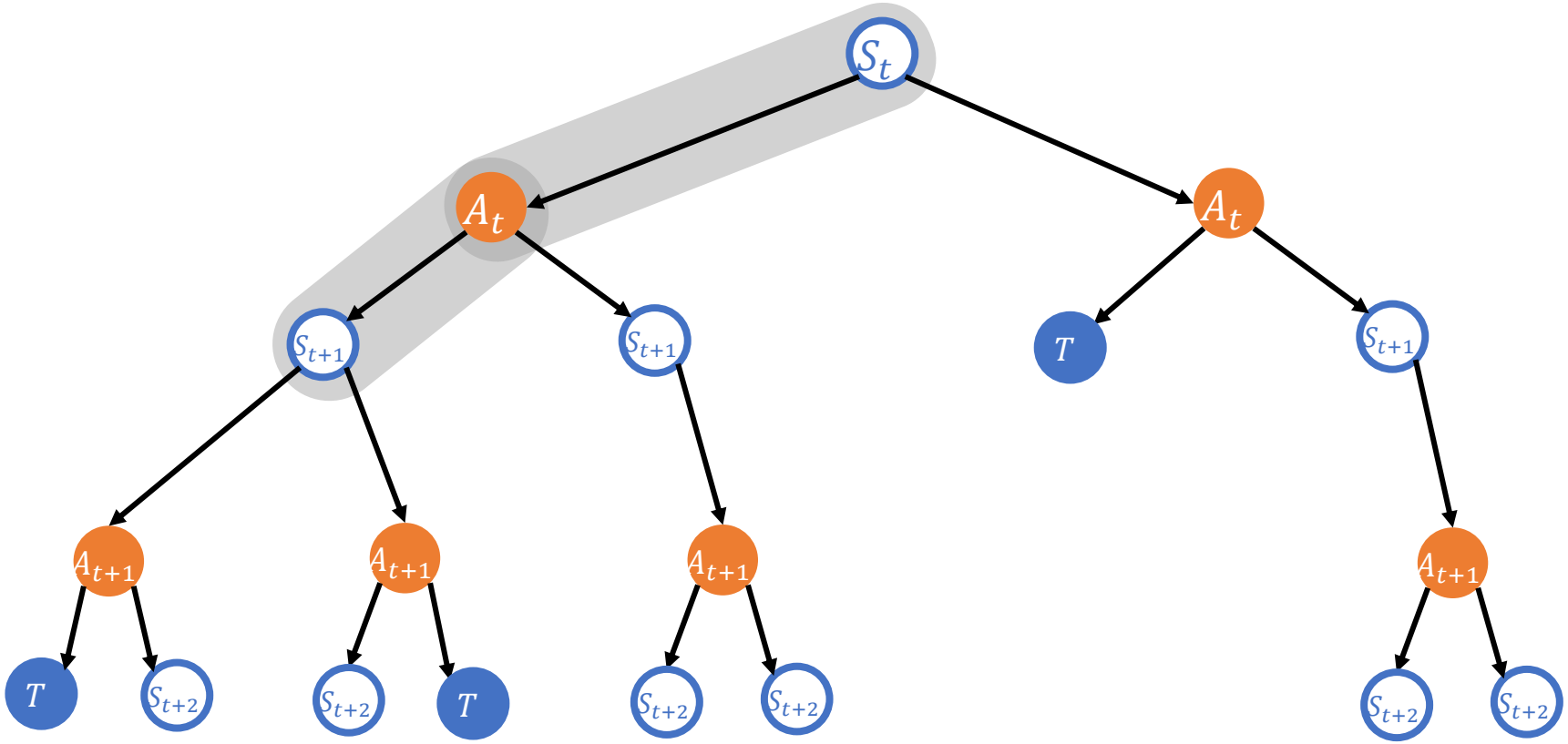
MC Policy evaluation : $V(s) \leftarrow V(s) + \alpha\big(G_t - V(s)\big)$

에피소드 =



I TD 기법을 활용한 Sample backup

TD Policy evaluation : $V(s) \leftarrow V(s) + \alpha\left(R_{t+1} + \gamma V(S_{t+1}) - V(s)\right)$



여러 스텝의 보상을 활용한 TD: n-step TD

$$V(s) \leftarrow V(s) + \alpha\left(G_t^{(n)} - V(s)\right)$$

Aka TD(0)

1-step TD : $G_t^{(1)} \overset{\text{def}}{=} R_{t+1} + \gamma V(S_{t+1})$: **1 스텝**까지의 보상 (새로운 정보) + 가치함수 (알고 있는 정보)

2-step TD :$G_t^{(2)} \overset{\text{def}}{=} R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2})$: **2 스텝**까지의 보상 (새로운 정보) + 가치함수 (알고 있는 정보)

3-step TD :$G_t^{(3)} \overset{\text{def}}{=} R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 V(S_{t+3})$

…

∞-step TD :$G_t^{(\infty)} \overset{\text{def}}{=} R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots + \gamma^{T-1} R_T$: 몬테카를로와 동일

$$G_t^{(n)} \overset{\text{def}}{=} R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n V(S_{t+n})$$

여러 스텝의 리턴을 어떻게 평균을 낼까?

$$V(s) \leftarrow V(s) + \alpha\left(G_t^{(n)} - V(s)\right)$$

G_t 의 추정치로서 다양한 n에 해당하는 $G_t^{(n)}$ 를 추산할 수 있다.

즉, 하나의 값을 추정하기 위해 여러가지 방식의 추정기법을 활용할 수 있다.

자연스러운 의문

서로 다른 추정치를 모두 사용하여 하나의 G_t 추정치를 만들 수는 없을까?

I 여러 개의 추정치를 하나의 추정치로 표현하기

산술 평균?

$$G_t \overset{\text{def}}{=} \frac{1}{T}\left(G_t^{(1)} + G_t^{(2)} + G_t^{(3)} + G_t^{(4)} + G_t^{(5)} + \cdots + G_t^{(T)}\right)$$

- 현재 정보만을 $G_t^{(1)}$는 상대적으로 적은 분산을 가지나, 편향이 있음.
- 미래의 정보를 모두 사용하는 $G_t^{(T)}$ 큰 분산을 가지나, 편향이 없음.

추가적으로 필요한 장치:

1. 편향-분산 트레이드오프를 조정할 수 있는 방법이 필요

2. 각의 추정치를 "**적합하게**" scaling 하는 요소가 필요: 산술평균의 경우 $\frac{1}{T}$

$$G_t^{dumb} \overset{\text{def}}{=} \left(G_t^{(1)} + G_t^{(2)} + G_t^{(3)} + G_t^{(4)} + G_t^{(5)} + \cdots + G_t^{(T)}\right)$$

I 여러 개의 추정치를 하나의 추정치로 표현하기 – TD(λ)

기하적으로 감소하는 scaling + 평균!

$$G_t^\lambda \overset{\text{def}}{=} (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{(n)} \qquad (0 \le \lambda \le 1)$$

기하급수 $1, r, r^2, \ldots$

$$\sum_{n=0}^{N}\lambda^n = \frac{(1-\lambda^N)}{1-\lambda} \quad (\lambda \neq 1)$$

N 이 충분히 클 때, $\lambda^N \approx 0$

즉, $\sum_{n=0}^{N}\lambda^n \approx \frac{1}{1-\lambda}$

λ 를 바꿔 줌으로써, 분산-편향 관계를 조정가능.

- TD(0) ($\lambda = 0$ 일 때), $G_t^0 = G_t^{(1)}$ 이 됨. **1-step TD**와 동일
- TD(1) 매-방문 MC와 유사.

I 마무리!

- 모델 free 가치 추산 알고리즘
  - Temporal-difference (TD) 방식
- n-step TD
- TD(λ)

(우리가 다루지 않은 것: Eligibility trace algorithm)