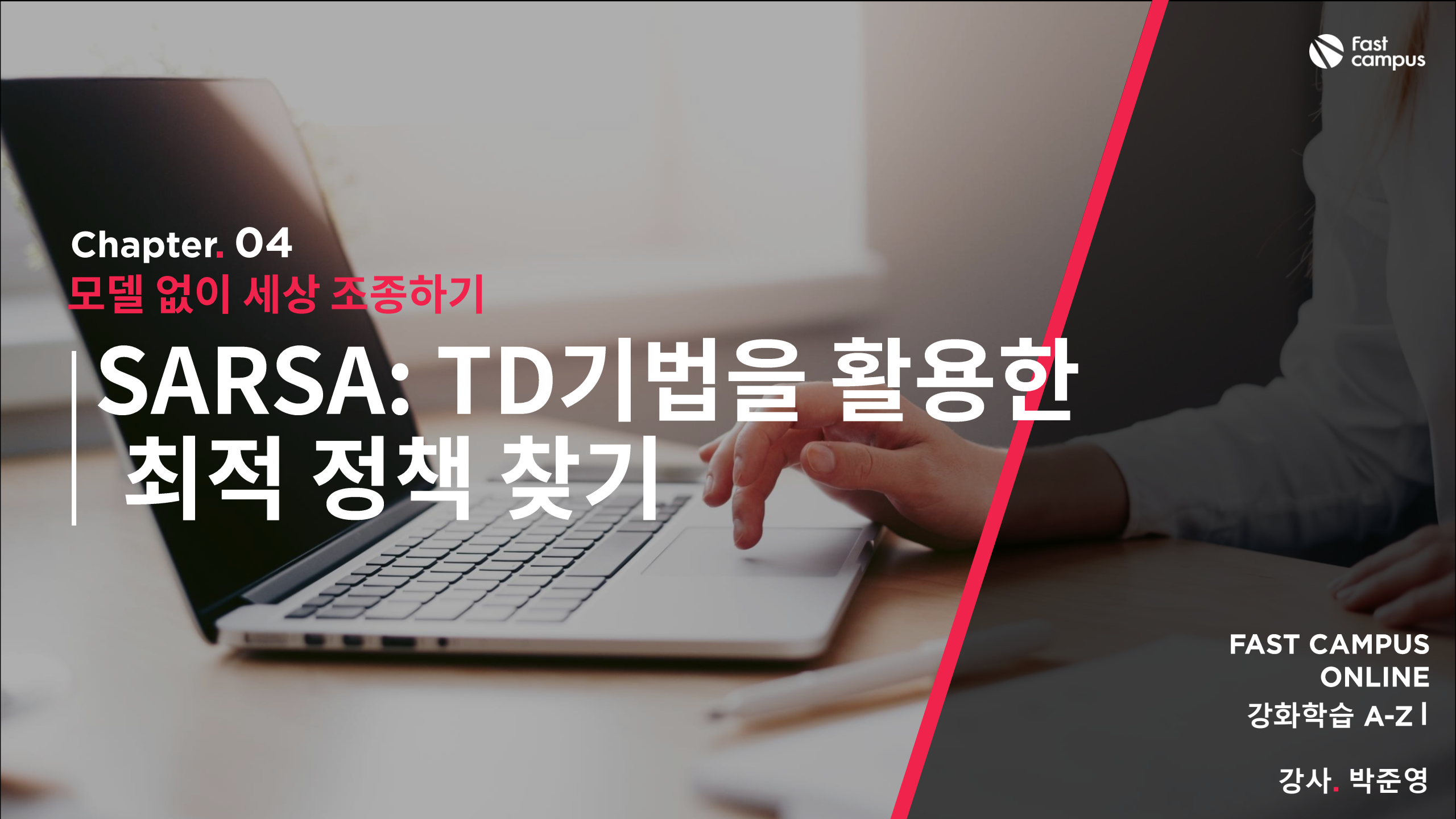


Chapter. 04

모델 없이 세상 조종하기

SARSA: TD기법을 활용한 최적 정책 찾기

FAST CAMPUS ONLINE

강화학습 A-Z I

강사. 박준영

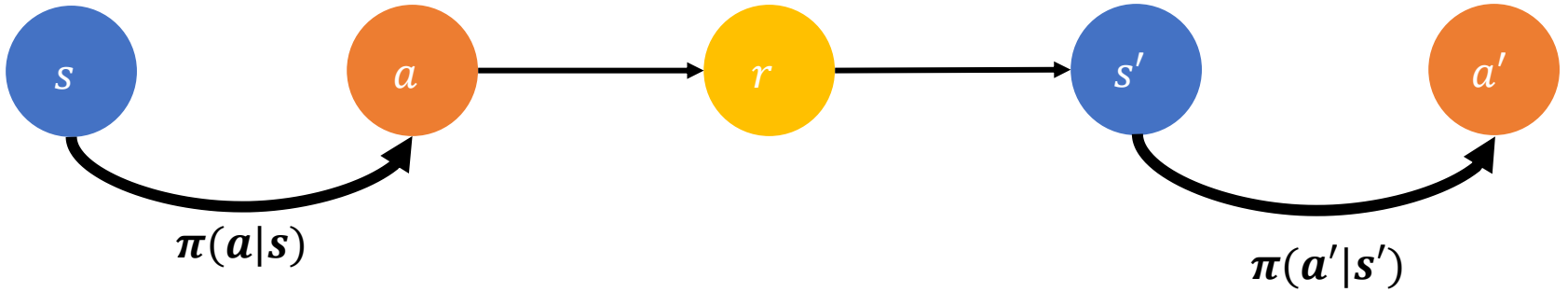
I (복습) Temporal-difference (TD) 기법

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

TD(0) 의 경우: $G_t \overset{\text{def}}{=} R_{t+1} + \gamma V(S_{t+1})$

비슷한 방식을 활용해서 $Q(s, a)$ 를 추산할 수 있지 않을까?

I SARSA: TD(0)를 활용한 행동 가치함수 Q^{π} 추산



SARSA update:

$$G_t \stackrel{\text{def}}{=} r + \gamma Q(s', a')$$

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma Q(s',a') - Q(s,a))$$

현재 Evaluation 하는 정책 π를 따라서 a' 이 결정.

$$a' = \pi(s')$$

I SARSA 의사 코드

SARSA

초기화 $Q(s,a) \leftarrow 0$ 모든 $(s,a) \in \mathcal{S} \times \mathcal{A}$

반복 (에피소드 $1, \ldots,$):

초기 상태 s 관찰

$Q(s,a)$를 활용해서 a 결정 (ex. ϵ greedy 정책)

반복:

a 를 환경에 가한 후, r과 s' 관측.

s' 에서 $Q(s',a)$ 를 활용해 a' 결정 (ex. ϵ greedy 정책)

$$Q(s,a) \leftarrow Q(s,a) + \alpha\big(r + \gamma Q(s',a') - Q(s,a)\big)$$

$s \leftarrow s'; a \leftarrow a'$

까지 s 는 종결상태

까지 $Q(s,a)$ 수렴.

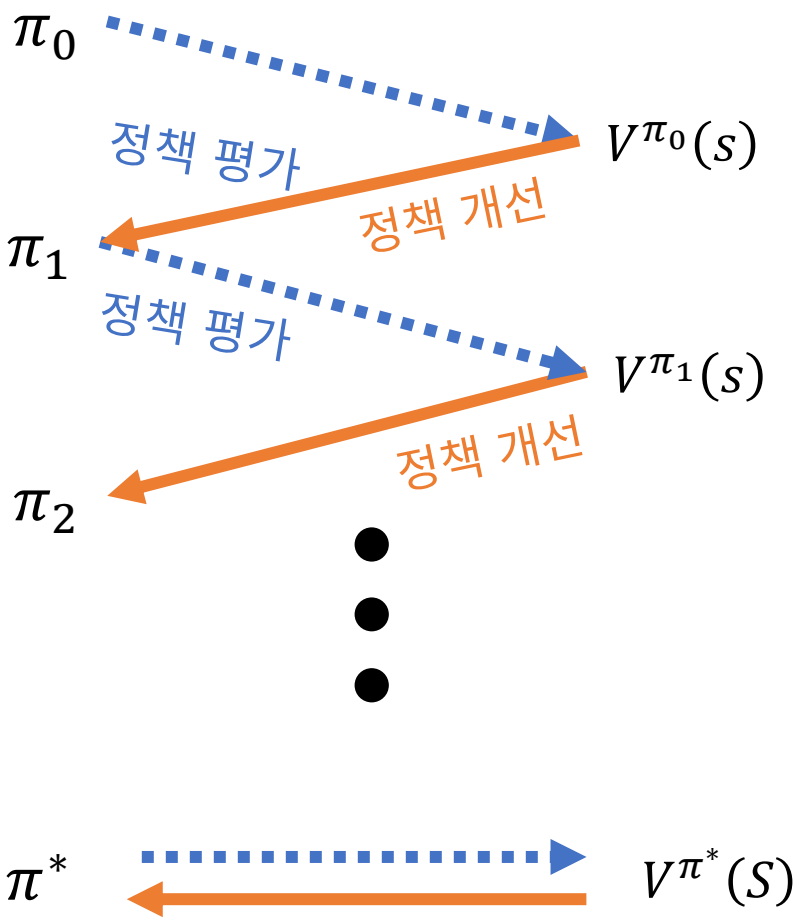
I SARSA control: SARSA policy evaluation + ϵ-탐욕적 개선 !

정책 평가:

SARSA 를 활용해 $Q^{\pi}(s,a)$ 추산

정책 개선:

ϵ-탐욕적 정책 개선



I 기억나시나요? (1) n-step TD

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

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

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

n-step SARSA

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left(q_t^{(n)} - Q(s_t, a_t)\right)$$

1-step TD : $q_t^{(1)} \overset{\text{def}}{=} R_{t+1} + \gamma \boldsymbol{Q(S_{t+1})}$: **1** 스텝까지의 보상 (새로운 정보) + 가치함수 (알고 있는 정보)

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

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

…

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

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

*$\boldsymbol{Q(S_{t+1})}$: 시점 $t+1$ 에서, S_{t+1} 이란 상태에서 주어진 정책을 따라서 고른 행동에 대한 행동 가치함수. Ex) ϵ-탐욕적 정책

SARSA(λ)

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

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left(q_t^{\lambda} - Q(s_t, a_t)\right)$$