03

강화학습 기본 알고리즘

4. TD와 SARSA

TD와 SARSA

TD

TD 개념

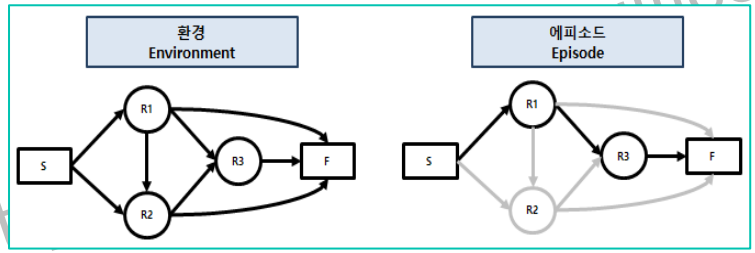
- 시간차학습(TD : Temporal Difference Learning)
- 반환값(G: Return)을 하나의 타임스텝이 완료되면 얻을 수 있는 값으로 대체
- 끝이 정해지지 않은 환경(Nonterminating Environment)에서도 사용 가능

MC $V(s_t) \leftarrow V(s_t) + \propto(G_t - V(s_t))$ ①

정책평가

TD $V(s_t) \leftarrow V(s_t) + \propto(R_{t+1} + V(s_{t+1}) - V(s_t))$ ②

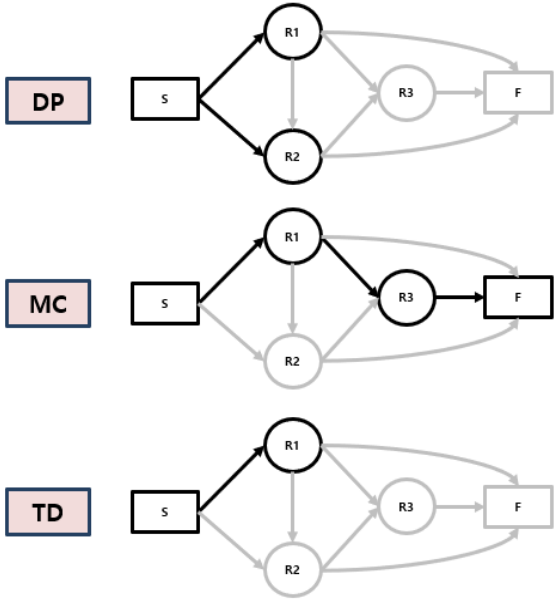
from G_t to $R_{t+1} + V(s_{t+1})$



TD와 SARSA

TD

TD 개념



기댓값

$$V(s_t) \leftarrow V(s_t) + \propto(R_{t+1} + V(s_{t+1}) - V(s_t))$$

정책평가

$$V(s_t) \leftarrow V(s_t) + \propto(G_t - V(s_t))$$

$$V(s_t) \leftarrow V(s_t) + \propto(R_{t+1} + V(s_{t+1}) - V(s_t))$$

TD와 SARSA

Q함수와 정책제어

MDP에서 Q함수와 정책제어

- 모델프리 환경에서 다음 상태를 알 수 없다. 하지만 Q함수는 계산할 수 있다.
- TD에서는 Q함수를 사용해서 정책제어 가능하다.

특정 행동(a)을 했을 때 가치

$$q_\pi(s, a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')$$

정책제어

$$\pi'(s) = \underset{a \in A}{\operatorname{argmax}}\ Q(s, a)$$

TD와 SARSA

SARSA

- Q함수 안에 행동과 상태에 대한 가치가 들어있기 때문에 Q함수를 가지고 정책을 평가할 수 있다.

TD $V(s_t) \leftarrow V(s_t) + \propto(R_{t+1} + \gamma V(s_{t+1}) - V(s_t))$

SARSA $Q(S,A) \leftarrow Q(S,A) + \propto(R_{t+1} + \gamma Q(S',A') - Q(S,A))$

$(S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}) \rightarrow S\ A\ R\ S\ A$

정책평가

정책제어

$$\pi'(s) = \underset{a \in A}{\operatorname{argmax}}\ Q(s,a)$$