시간차 학습

태그

6.1 TD 예측

MDP를 모를 때 탐색하는 방법
에피소드가 끝나기 전에 업데이트 하기
추측으로 추측을 업데이트 하기

몬테 카를로 방법은 에피소드가 끝나고 G_t를 알고난 후에 $V(S_t)$의 증가량을 업데이트 할 수 있다.

$$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$$

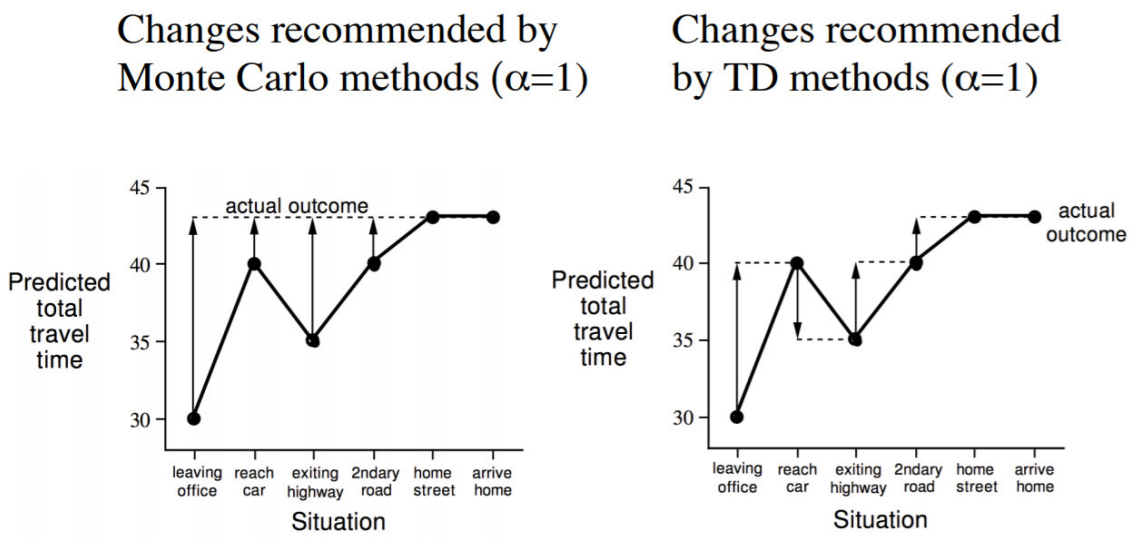
TD 방식은 그 중 TD(0) 방식은 다음 시간 단계 t+1에 가서 관측된 보상 R_{t+1}과 추정 값 $V(S_{t+1})$을 이용해 $V(S_t)$의 증가량을 업데이트 한다.

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

TD 오차 : t+1 시점이 되어서 알게된 $V(S_t)$ 의 오차

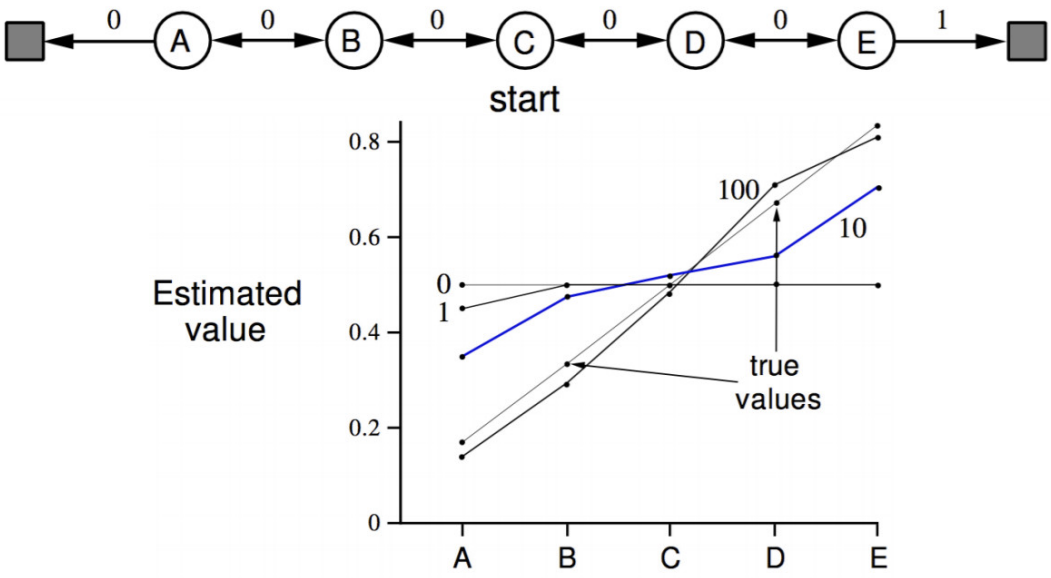
$$\delta_t = R_{t+1} + \gamma V(s_{t+1}) - V(S_t)$$

몬테카를로 방식이 G_t에 가깝게 value function을 예측하는 것을 목표로 한다면 TD 방식은 다음기에 관측 가능한 $R_{t+1} + \gamma V(S_{t+1})$ 에 가깝게 value function을 예측하는 것을 목표로 한다.

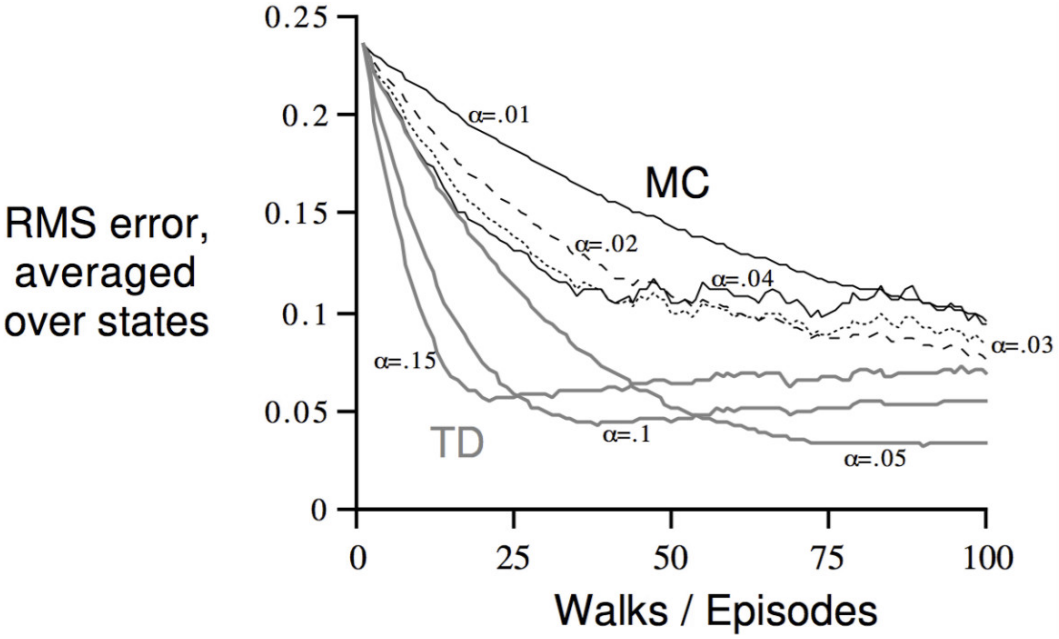


6.2 TD 예측 방법의 좋은점

- 최종 return이 나오기 전에 업데이트가 가능하다. 에피소드가 길 때도 전이마다 업데이트 가능.
- TD 방법으로도 수렴한다. 고정 정책에 대해, 고정 α를 사용하고, 그 값이 충분히 작다면 TD(0)은 평균적으로 v_π로 수렴한다.



- 확률론적 문제에서 TD 방식이 고정 αMC 방식보다 보통 빨리 수렴하는 현상이 있다.



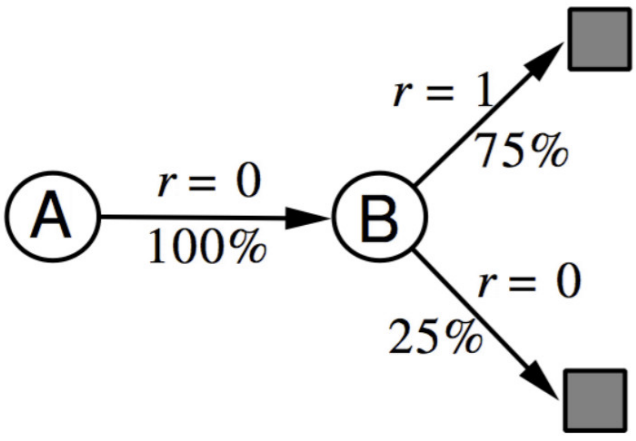
6.3 TD(0)의 최적성

k개의 유한한 에피소드를 일괄로 학습하여 한번에 갱신하는 경우 TD(0)는 몬테카를로 방식보다 성능이 좋다.

- 일괄 몬테카를로 방법은 마르코프 과정에 대한 고려 없이 평균 제곱 오차를 최소화 하는 추정량을 찾는다.
- 일괄 TD(0)은 항상 마르코프 과정의 최대우도모델에 대해 올바른 추정 값을 찾는다.

Two states A, B; no discounting; 8 episodes of experience

$A, 0, B, 0$
$B, 1$
$B, 1$
$B, 1$
$B, 1$
$B, 1$
$B, 1$
$B, 0$



What is $V(A), V(B)$?

6.4 SARSA : 활성 정책 TD 제어

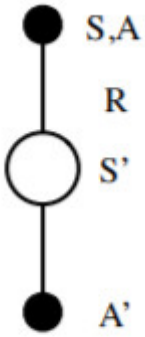
제어 문제 : 최적 policy를 찾는 문제
활성 정책 : 현재 정책과 타겟 정책이 동일한 경우
SARSA : 상태 + 액션 + 보상 + 다음 상태 + 다음 액션

- **TD 예측**

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

- **SARSA : 활성 정책 TD 제어**

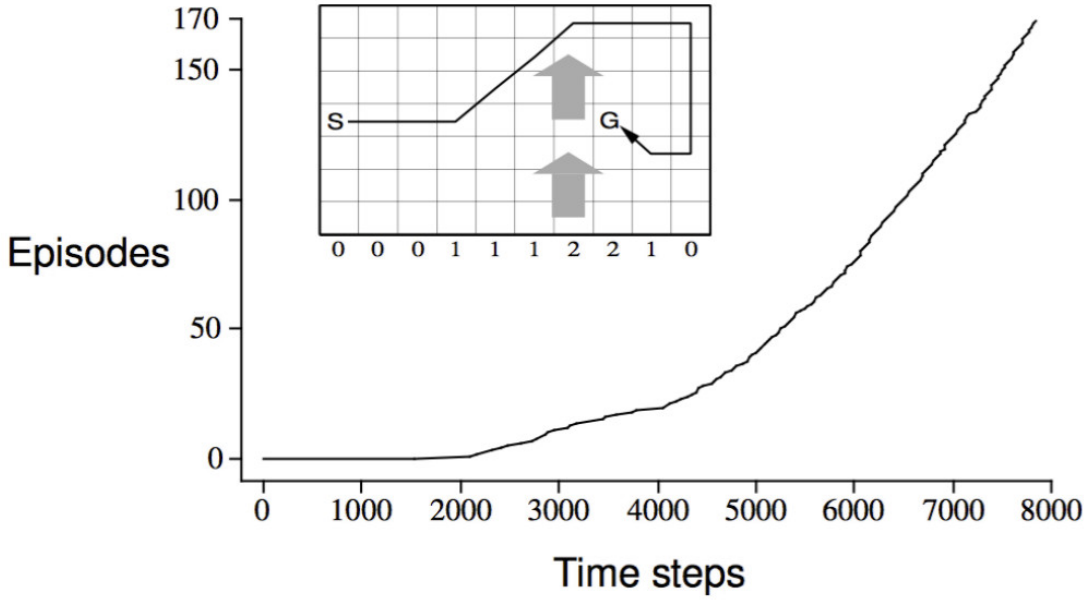
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$$



Sarsa

상태-행동 쌍의 가치를 학습한다.

모든 상태-행동 쌍을 무한번 마주치고 정책의 극한이 탐욕적 정책으로 수렴한다는 조건이 있으면, SARSA는 최적 정책과 최적 행동 가치 함수로 수렴한다.

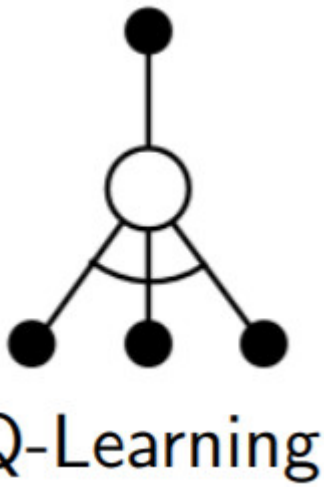


6.5 Q 학습 : 비활성 정책 TD 제어

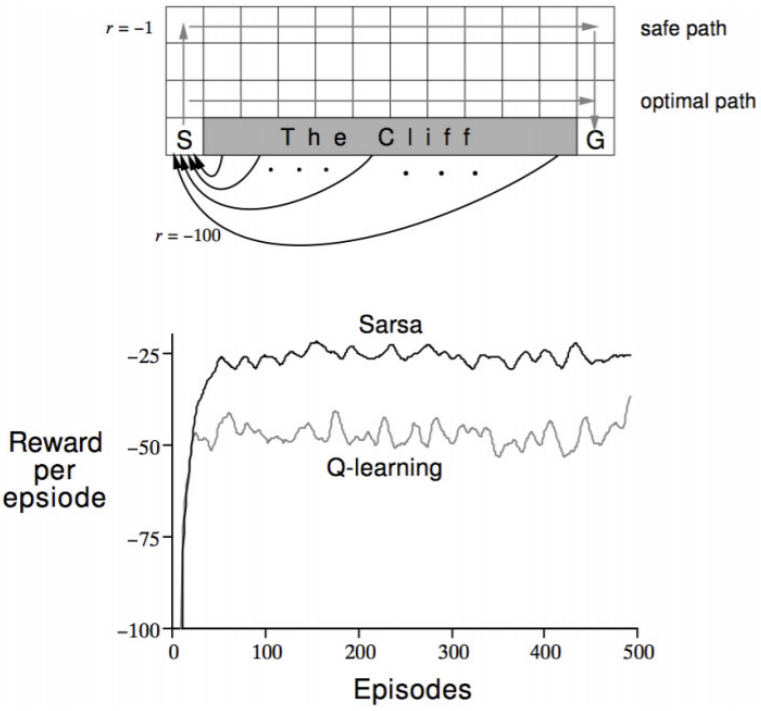
비활성 정책 : 현재 정책과 타겟 정책이 다른 경우
Q 학습 : max SARSA

Q 학습 : 현재 정책에 상관없이 최적 행동 가치함수 q_*를 근사한다.

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$$



올바른 수렴을 위해서 모든 쌍이 계속해서 갱신되어야 한다.



▼ 많이 쓰이는 정책

target : greedy

behavior : epsilon-greedy

6.6 기댓값 SARSA

Q 학습과 동일한 알고리즘이지만 max 대신 기댓값을 기준으로 한다.
Q 학습의 바탕을 이루며 Q 학습을 일반화한 것으로 SARSA 보다 향상된 성능이 면서 Q 학습이 될 수도 있다.

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \mathbb{E}[Q(S_{t+1}, A_{t+1}) | S_{t+1}] - Q(S_t, A_t) \right]$$

6.7 최대화 편차 및 이중 학습

최대화 편차의 문제 : 최대값의 추정치와 추정치의 최대값의 차이로 인한 실제 가치와 추정치의 최대값 간 편차
이중학습 : 독립적인 이중 추정값 학습

최대화 편차의 사례 $N(-0.1, 1)$

6.8 게임, 이후상태, 그 밖의 특별한 경우들