

시간차 학습 알고리즘을 개선한 알고리즘:

- 특징 2개: ,

가장 좋은것만을 고르는 정책:

ϵ 의 확률만큼 랜덤하게 움직이는 것:

SARSA의 정책:

Q-Learning의 정책:

행동하는 정책과 학습할 때 사용하는 정책이 다른 것:

행동하는 정책이 학습할 때 사용하는 정책과 같은 것:

SARSA 문제점: ,

- Q-Learning?:

행동가치 추정 시 분산이 높은 SARSA 단점 보완:

절벽이 있는 그리드 월드에서 안전한 경로 선택: ,

절벽이 있는 그리드 월드에서 최적 경로 선택:

학습 과정에서 입실론 감소:

주어진 목적을 달성하기 위해 수행해야 할 일에 대한 방법 또는 절차 마련:

- 예시: ,

주어진 대상과 직접 상호작용하며 목적을 달성하는 방법 또는 절차:

- ,

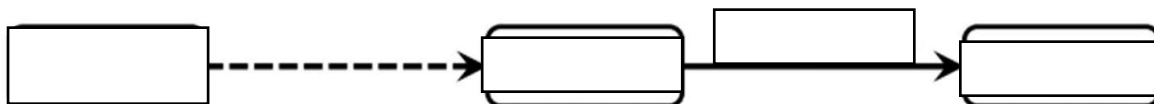
에이전트가 어떤 행동시 환경이 어떻게 반응할 지 예측 위하여 사용하는 것:

- 종류 2개: ,

모든 발생 가능한 상황 및 확률분포가 정확하게 기술:

발생 가능 상황 중 임의의 가능성만 샘플링 되는 모델:

둘 중 더 용이한 것:



임의의 정책을 통한 행동을 수행하여 상태 공간을 탐색:

가치함수 계산 위해, 시뮬레이션 위한 경험데이터 기반 () 과정 수행

모델에 의하여 생성된 경험 샘플 데이터 사용:

환경과의 직접 상호 작용을 통해 얻는 실제 경험 데이터 사용:

온라인 계획 및 강화 학습이 동시에 수행되는 알고리즘:

- 특징:

간접 모델 학습 및 계획:

- 장점:

- 단점:

직접 강화학습: Q-Learning

- 환경과 ()을 통하여 얻는 경험 바로 활용 하여()

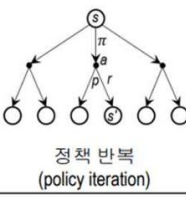
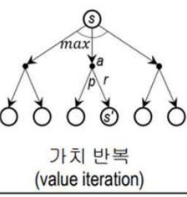
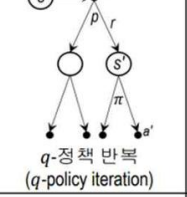
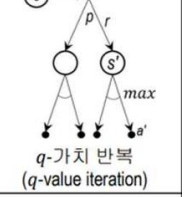
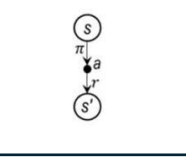
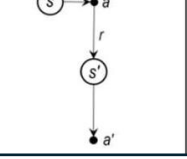
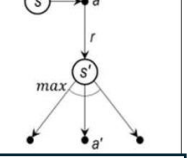
Dyna-Q 초기 시작:

- 행위 정책:

갱신 대상에 다른 갱신 방법 2개:

발생할 수 있는 모든 상황 고려 갱신:

발생할 수 있는 간단한 단일 샘플 활용 갱신:

추정 가치 (Value Estimated)	$v_{\pi}(s)$	$v_*(s)$	$q_{\pi}(s, a)$	$q_*(s, a)$
기대 갱신 (Expected Updates) - 동적 프로그래밍	 정책 반복 (policy iteration)	 가치 반복 (value iteration)	 q-정책 반복 (q-policy iteration)	 q-가치 반복 (q-value iteration)
샘플 갱신 (Sample Updates) - 단일 스텝 시간차 학습				

O: 해당 척도에 대하여 더 좋은 방법

	샘플 갱신 (Sample Update)	기대 갱신 (Expected Update)
계산량		
소요 시간		
정확도		
대표적인 기법		

환경 또는환경 모델과 지속적으로 상호작용 하며 상태와 보상 얻어오는 일련의 과정:

샘플링은 어떠한 ()에 입각하여 수행됨

샘플링 대상 후보들에 대해 어떠한 편중없이 임의로 선택:

현재 주어진 정책을 기반으로 각 상태의 행동 선택, 샘플링:

- 효과:

:

