# RL_Basic 알고리즘

## 알고리즘: 반복 정책 평가

1. **초기화**

   모든 $s \in S$에 대해 V(s)$\in R$과 $\pi(s) \in A(s)$를 임의로 설정

2. **정책 평가(Policy Evaluation)**

$\Delta < \theta$(작은 양수)가 될 때까지 반복:

$\Delta \leftarrow 0$

모든 $s \in S$에 대해:

$v \leftarrow V(s)$

$V(s) \leftarrow \sum_a \pi(a|s) \sum_{s'} P(s'|s,a)[r(s,a,s') \ + \ \gamma V(s')]$

$\Delta \leftarrow max(\Delta, |v - V(s)|)$

3. **정책 개선(Policy Improvement)**

$policyStable \leftarrow true$

모든 $s \in S$에 대해:

$old - action \leftarrow \pi(s)$

$\pi(s) \leftarrow \arg\max_a \sum_{s'} P(s'|s,a)[r(s,a,s') \ + \ \gamma V(s')]$

만약 $old - action \neq \pi(s)$라면 $policyStable \leftarrow False$

만약 $policyStable = true$라면 $V \approx v*, \pi \approx \pi*$반환; 아니면 2부터 반복

## 알고리즘: 가치 반복 알고리즘

초기화

모든 $s \in S^+$에 대해 $V(s) \in R$를 임의로 설정

반복: 최적 상태 가치 찾기(finding optimal state value)

$\Delta \leftarrow 0$

모든 $s \in S$에 대해:

$v \leftarrow V(s)$

$V(s) \leftarrow \max_a(P(s'|s,a)[r(s,a,s') + \gamma V(s'))$

$\Delta \leftarrow \theta$(작은 양수) 일 때까지 반복

정책 $\pi$를 다음과 같이 추출: 정책 추출(policy extraction)

$\pi(s) = \arg\max_a \sum_{s'} P(s'|s,a)[r(s,a,s') + \gamma V(s')]$

## 알고리즘: First-visti MC의 Prediction

입력:

초기화:

$\pi \leftarrow$평가할 정책

$V \leftarrow$임의의 상태 가치 함수

$Return(s) \leftarrow$빈 리스트(모든 $s \in S$에 대해)

반복:

정책 $\pi$를 이용해 에피소드 생성

에피소드에 출현한 각 상태 $s$에 대해:

$G \leftarrow$처음 $s$에 의해 발생한 수익

$G$를 $Return(s)$에 추가$(append)$

$V(s) \leftarrow average(Return(s))$

## 알고리즘: 몬테카를로 방법의 Control

모든 $s \in S, a \in A(S)$에 대해 초기화:

$Q(s,a) \leftarrow$ 임의의 값

$Return(s,a) \leftarrow$ 빈 리스트

$\pi(s,a) \leftarrow$ 임의의 $\epsilon-$탐욕정책

무한 반복:

(a) $\pi$를 사용해 에피소드 1개 생성

(b) 에피소드에 출현한 각 s, a에 대해:

$R \leftarrow s,\ a$의 처음 발생한 수익

$R$을 $Return(s,\ a)$에 추가

$Q(s,\ a) \leftarrow average(Return(s,\ a))$

(c) 에피소드 안의 각 $s$에 대해:

$a^* \leftarrow \arg\max_a Q(s,a)$

모든 a∈A(S)에 대해

$$\pi(s,a) = \begin{cases} 1-\epsilon+\dfrac{\epsilon}{|A(s)|} & (a=a^*) \\ \dfrac{\epsilon}{|A(s)|} & (a \neq a^*) \end{cases}$$

## TD(0)의 Prediction

초기화:

$\pi \leftarrow$평가할 정책

$V \leftarrow$임의의 상태가치 함수

각 에피소드에 대해 반복:

$s$를 초기화

에피소드의 각 스텝에 대해 반복:

$a \leftarrow$상태 $s$에서 정책 $\pi$에 대해 결정된 행동

행동 $a$를 취한 후 보상 $r$과 다음 상태 $s'$를 관측

$V(s) \leftarrow V(s) + a[r + \gamma V(s') - V(s)]$

$s \leftarrow s'$

$s$가 마지막 상태라면 종료

## 알고리즘: TD(0) SARSA

모든 $s \in S, a \in A(S)$에 대해 초기화:

$Q(s, a) \leftarrow$임의의 값

$Q(terminal - state, \cdot) = 0$

각 에피소드에 대해 반복:

$s$를 초기화

$s$에서 행동 정책(Behavior policy)으로 행동 $a$를 선택(예: $\epsilon-$탐욕정책)

에피소드의 각 스텝에 대해 반복:

행동 $a$를 취한 후 보상 $r$과 다음 상태 $s'$를 관측

$s'$에서 타깃 정책(Target policy)으로 행동 $a'$를 선택(예: $\epsilon-$탐욕정책)

$Q(s, a) \leftarrow Q(s, a) + a[r + rQ(s', a') - Q(s, a)]$

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

$s$가 마지막 상태라면 종료

## 알고리즘: TD(0) Q-learning

모든 $s \in S, \ a \in A(S)$에 대해 초기화:

$Q(s,a)$ ←임의의 값

$Q(terminal-state,a)=0$

각 에피소드에 대해 반복:

$s$를 초기화

에피소드의 각 스텝에 대해 반복:

$s$에서 행동 정책(Behavior policy)으로 행동 $a$를 선택(예: $\epsilon-$탐욕정책)

행동 $a$를 취한 후 보상 $r$과 다음 상태 $s'$를 관측

$s'$에서 타깃 정책(Target policy)으로 행동 $a'$를 선택(예: 탐욕정책)

$Q(s,a) \leftarrow Q(s,a)+a[r+\gamma \max_{a'} Q(s',a')-Q(s,a)]$

$s \leftarrow s'$

$s$가 마지막 상태라면 종료

## 알고리즘: Double Q-learning

모든 $s \in S,\ a \in A(s)$에 대해 초기화:

$Q_1(s,a), Q_2(s,a)$ ←임의의 값

$Q_1(terminal-state,a)=Q_2(terminal-state,a)=0$

각 에피소드에 대해 반복:

$S$를 초기화

에피소드의 각 스텝에 대해 반복:

$Q_1$과 $Q_2$로부터 $a$를 선택(예: $\epsilon-$탐욕정책 $Q_1+Q_2$)

행동 $a$를 취한 후 보상 $r$과 다음 상태 $s'$를 관측

확률이 0.5보다 작다면

$Q_1(s,a) \leftarrow Q_1(s,a)+\alpha(r+\gamma Q_2(s', \arg\max_{a'} Q_1(s',a'))-Q_1(s,a)]$

else

$Q_2(s,a) \leftarrow Q_2(s,a) + \alpha[r + \gamma Q_1(s', \arg\max_{a'} Q_2(s',a')) - Q_2(s,a)]$

$s \leftarrow s'$

$s$가 마지막 상태라면 종료

## 알고리즘: 액터-크리틱

모든 $s \in S, a \in A(S)$에 대해 초기화:

$p(s,a), V(s)$ ←임의의 값

각 에피소드에 대해 반복:

$s$를 초기화

에피소드의 각 스텝에 대해 반복:

Actor : $p(s,a)$로부터 $a$를 선택(예: Gibbs 소프트맥스 함수)

$$\pi(s,a) = Pr\{A_t = a | S_t = s\} = \frac{e^{p(s,a)}}{\sum_b e^{P(s,b)}}$$

행동 $a$를 취한 후 보상 $r$과 다음 상태 $s'$를 관측

Critic 학습:

$\delta_t = r_{t+1} + \gamma V(S_{t+1}) - V(S_t)$

$V(s) \leftarrow V(s) + a\delta_t$

Actor 학습:

$p(s_t, a_t) = p(s_t, a_t) + \beta\delta_t$

$s \leftarrow s'$

$s$가 마지막 상태라면 종료

## 함수근사 TD(0) Prediction

초기화:

$V(s|w)$ ← 미분 가능한 함수

$w$ ← 함수의 가중치를 임의의 값으로 초기화

각 에피소드에 대해 반복:

$s$를 초기화

에피소드의 각 스텝에 대해 반복:

$a$ ← $s$에서 $\pi$에 의해 결정된 행동(random selection)

행동 $a$를 취한 후 보상 $r$과 다음 상태 $s'$를 관측

$w \leftarrow w + \alpha[r + \gamma V(s'|w) - V(s|w)] \frac{\partial V(s|w)}{\partial w}$

$s \leftarrow s'$

$s$가 마지막 상태라면 종료

## 알고리즘: Q-learning

초기화:

$Q(s, a|w)$ ← 미분 가능한 함수

$w$ ← 함수의 가중치를 임의의 값으로 초기화

각 에피소드에 대해 반복:

$s$를 초기화

에피소드의 각 스텝에 대해 반복:

$s$에서 행동 정책을 이용해 행동 $a$를 선택(예: Gibbs 소프트맥스 함수)

행동 $a$를 취한 후 보상 $r$과 다음 상태 $s'$를 관측

$s'$에서 타깃 정책으로 행동 $a'$를 선택

$$w \leftarrow w + \alpha[r + \gamma \max_{a'} Q(s', a'|w) - Q(s, a|w)] \frac{\partial Q(s,a|w)}{\partial w}$$

$s$가 마지막 상태라면 종료