02

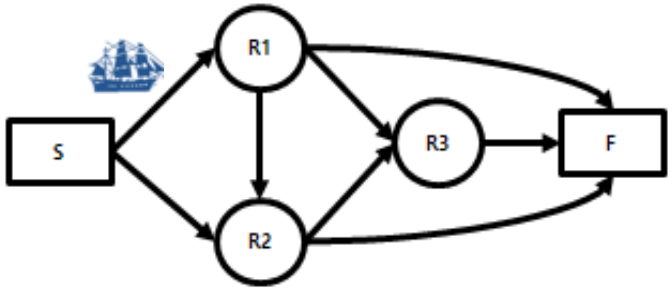
강화학습 기본개념

3. 마르코프 보상 과정

마르코프 보상과정 기본개념

마르코프 보상과정(Markov Reward Process)

- Markov Chain + Reward + 감가율(γ)
- 마르코프 체인이 시간에 따른 상태변화만을 다뤘다면 마르코프 보상과정은 변화에 따른 가치(Reward, 감가율)를 함께 다룬다.



마르코프 보상과정 구성요소

확률의 기댓값

이산확률분포의 기댓값 $E(X) = \sum xf(x)$

*f(x) : 사건이 일어날 확률
*x : 사건의 값(이득)

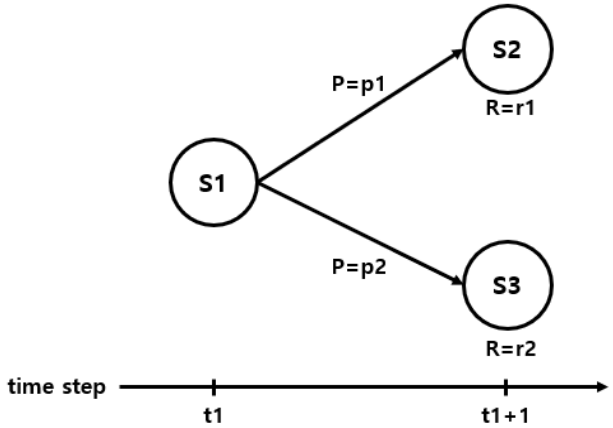
주사위의 기댓값 $1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$

연속확률분포의 기댓값 $E(X) = \int_{-\infty}^{\infty} xf(x)dx$

마르코프 보상과정 구성요소

보상함수(Reward Function)

- 상태가 시간 t에서 s일 때 시간 t+1에서 받을 수 있는 보상의 기대 값

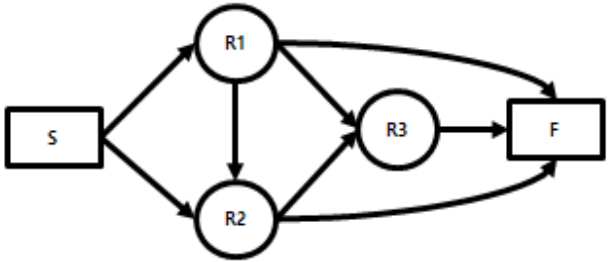


보상함수 R_{s1} = p1*r1 + p2*r2

마르코프 보상과정 구성요소

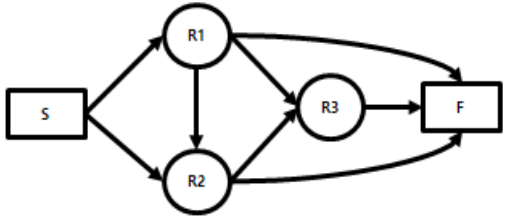
감가율 γ

- 시간의 흐름에 따라 가치를 얼마의 비율로 할인할 지를 결정하는 비율
- 감가율은 0과 1사이의 값을 가진다
- ex) 중고 자동차, 채권, 어음



마르코프 보상과정 구성요소

마르코프 보상과정(Markov Reward Process)



- **S** : 상태(State)의 집합
- ***P***: 상태 전이 매트릭스

  $P_{ss'} = P[S_{t+1} = s' \mid S_t = s]$
- ***R***: 보상 함수

  $R_s = E[R_{t+1} \mid S_t = s]$
- **γ** : 감가율

  $\gamma \in [0, 1]$

} 마르코프 체인

*보상 함수
상태가 시간 t에서 s일 때 시간 t+1에서 받을 수 있는 보상의 기대 값

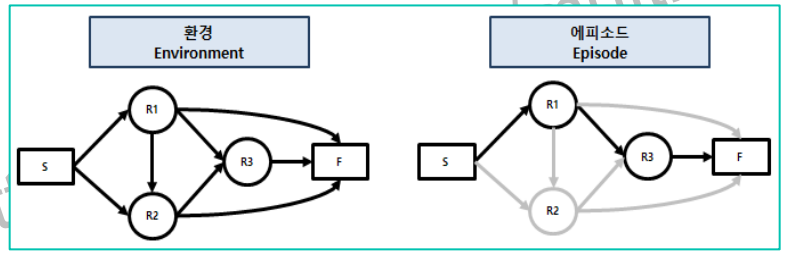
*감가율
시간의 흐름에 따라 가치를 얼마의 비율로 할인할 지를 결정하는 비율

마르코프 보상과정 응용(1)

반환값 G

- 반환값은 타임스텝 t에서 계산한 누적 보상의 합계이다.
- 이 누적 보상은 감가율로 할인되어 계산된다.
- 반환값은 주로 전체 환경이 아닌 에피소드 단위로 계산되는데 에피소드의 효율성이나 가치를 반환값을 가지고 평가한다.
- 반환값은 하나의 선택된 경로(에피소드)에 대한 전체적인 보상을 계산하는 방식이다. 이미 경로가 선택되었기 때문에 상태전이확률을 사용할 필요가 없다.

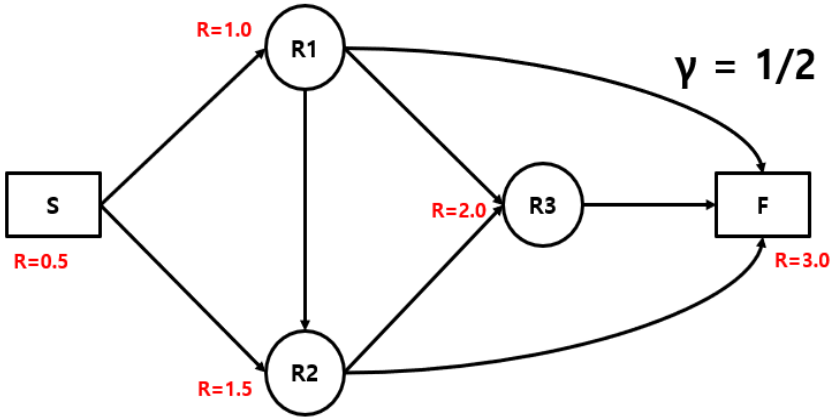
$$G_t = R_{t+1} + R_{t+2} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$



마르코프 보상과정 응용(1)

반환값 G 사례

*감가율은 ½로 가정



3 타임스텝에 목적지에 도달하는 에피소드의 반환값 계산

① S → R1 → R3 → F = 0.5 + ½x1.0 + ½x½x2.0 + ½x½x½x3.0 = 1.5
(R=0.5, R=1.0, R=2.0, R=3.0)

② S → R1 → R2 → F = 0.5 + ½x1.0 + ½x½x1.5 + ½x½x½x3.0 = 1.21875
(R=0.5, R=1.0, R=1.5, R=3.0)

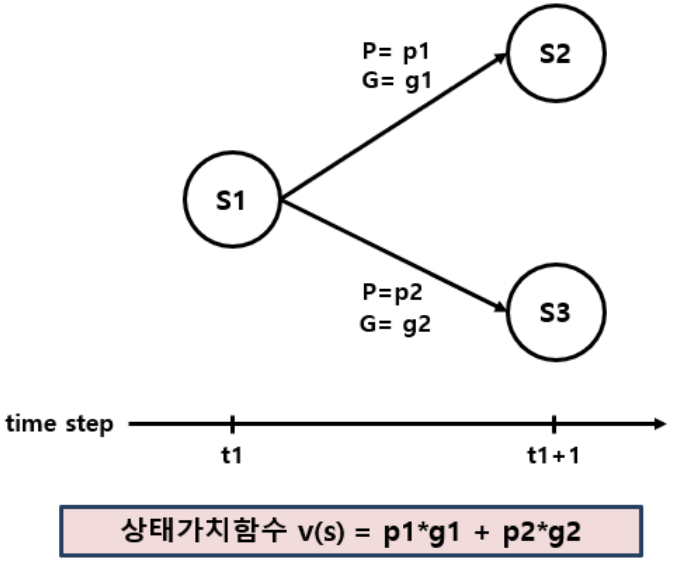
③ S → R2 → R3 → F = 0.5 + ½x1.5 + ½x½x2.0 + ½x½x½x3.0 = 1.875
(R=0.5, R=1.5, R=2.0, R=3.0)

마르코프 보상과정 응용(2)

상태가치함수

- 반환값(G: Return)이 에피소드 하나에 대한 가치를 측정할 수 있었다면, 상태 가치함수는 환경(Environment) 전체에 대한 가치를 측정할 수 있다

$$
\begin{aligned}
v(s) &= E[G_t \mid S_t = s] \\
&= E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + .. \mid S_t = s] \\
&= E[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + ..) \mid S_t = s] \\
&= E[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\
&= E[R_{t+1} + \gamma v(s_{t+1}) \mid S_t = s]
\end{aligned}
$$



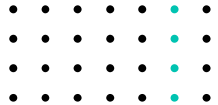
상태가치함수 v(s) = p1*g1 + p2*g2

마르코프 보상과정 응용(2)

상태가치함수

	측정대상	특징	감가율 γ	상태전이확률 P
반환값 G : Return	에피소드 Episode	합계	사용	미사용
상태 가치함수 v : State Value Function	전체 환경 Environment	기대값	사용	사용

마르코프 보상과정 응용(2)



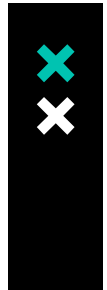
상태가치함수

- 벨만방정식은 일반적으로 기대 값을 시그마 기호를 사용한 수열의 합으로 표현하며 현재 상태와 다음 상태의 관계로 나타낸 수식이다.
- 상태가치 함수를 벨만 방정식으로 나타내면 다음과 같다.

$$v(s) = \mathbb{E}[R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s]$$

$$= R_{t+1} + \gamma \mathbb{E}[v(S_{t+1}) \mid S_t = s]$$

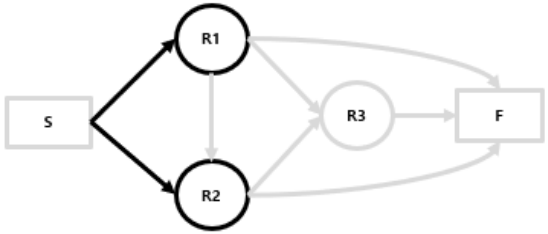
$$= R_{t+1} + \gamma \sum_{s' \in S} P_{ss'} v(s')$$



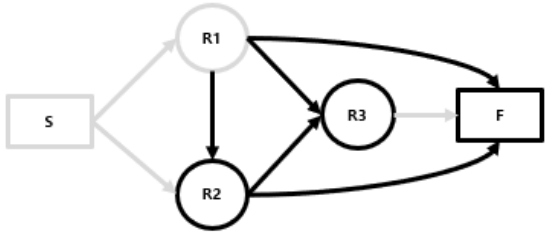
마르코프 보상과정 응용(2)

상태가치함수

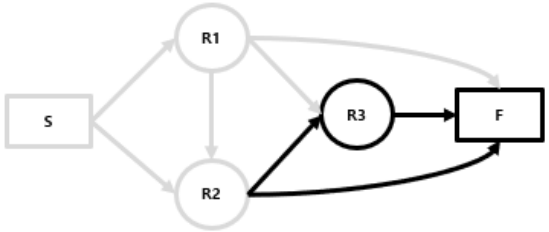
time step 1



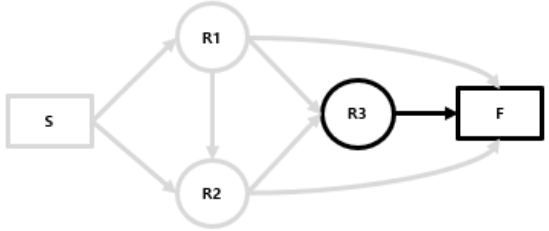
time step 2



time step 3



time step 4



*타임스텝별로 고려해야 하는 상태

*실제로 상태가치함수를 구하는 것은 어렵다

정 리

마르코프 보상과정

확률의 기댓값

보상함수와 감가율

반환값

상태가치함수