2021 인공지능 소수전공

96~100차시: 강화학습

2021.08.14 17:30~22:15

Seokhwan Yang

Artificial Intelligence &
Data Analysis Laboratory

강화학습 개요

- 강화학습(Reinforcement Learning)이란?

쉽게, 추상적으로 말하면

시행착오를 통해 발전해 나가는 과정

조금 더 정확하게 말하면

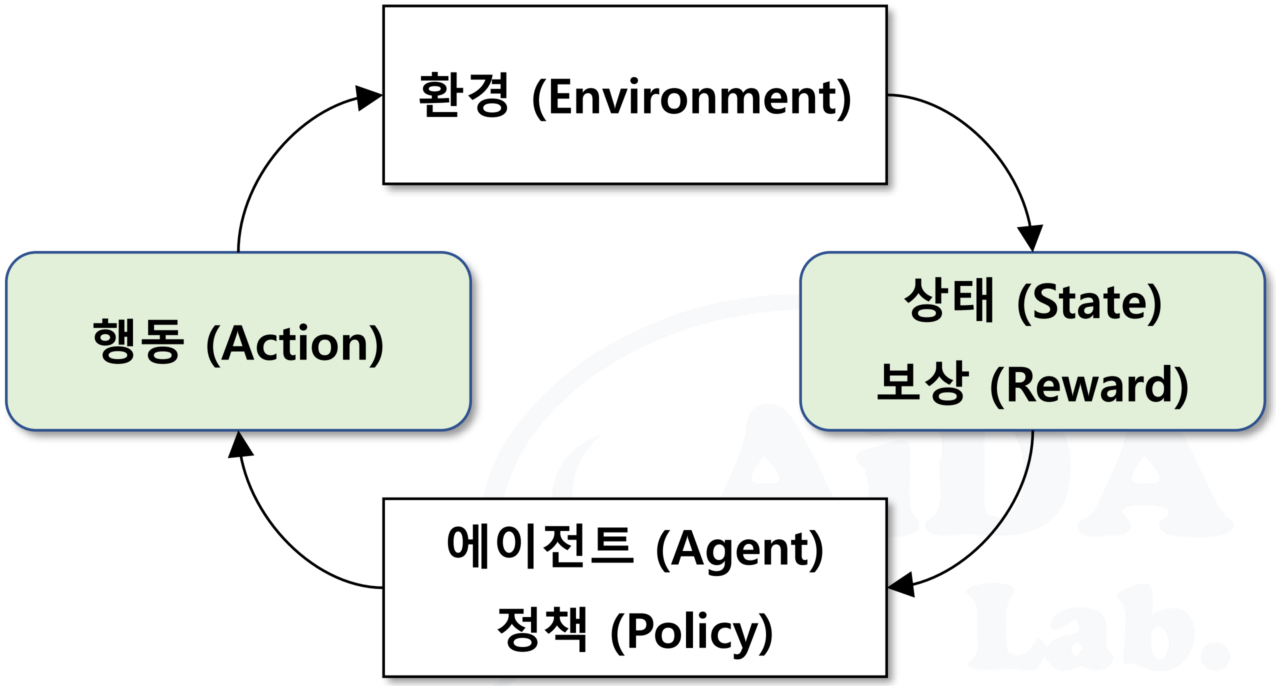
순차적 의사결정 문제에서
누적 보상을 최대화 하기 위하여
시행착오를 통해
행동을 교정하는
학습과정

강화학습 개요

- 강화학습(Reinforcement Learning)이란?
  - 적절히 설계된 보상 체계를 활용해
  - 에이전트가 긍정적인 행동을 할 수 있도록
  - 에이전트의 행동을 제어하는 정책을
  - 찾아내는 최적화 기법

강화학습 개요

- 강화학습에서
  - 에이전트(Agent)는
  - 정책(Policy)에 따라
  - 어떤 환경(Environment)에서
  - 특정 행동(Action)을 한다.
  - 그 행동에 따라 환경의
    상태(State)가 바뀌고
  - 상태가 긍정적으로 바뀌었는지
    부정적으로 바뀌었는지에 따라 보상(Reward)을 받는다.



강화 학습의 구성 요소

강화학습 개요

- 강화학습의 목적
  - 행동의 결과로 받는 모든 보상을 누적해서 합산하고, 그 값이 최대가 될 수 있는 정책을 찾는 것
    → 강화학습은 가장 좋은 정책을 찾는 것이 목적이고
    → 가장 좋은 정책은 누적 보상의 합을 최대로 만든다

강화학습 개요

발 밑을 볼 수 없고
앞만 볼 수 있는 아이가
꽃밭을 건너가는 경우

꽃을 잘 피해서 건너가면
밟힌 꽃이 없을 것이고

잘못 발을 디뎌 꽃을 밟는다면
건너간 후에 많은 꽃이
쓰러져 있을 것이다.

아이가 한 발자국
걸을 때 마다
꽃을 밟으면 꾸중을 하고
꽃을 밟지 않으면 칭찬을 한다면

무수히 많은 시도를 한 후에는
꽃을 전혀 밟지 않고
꽃밭을 건널 수 있을 것이다.

그림 1-2 강화학습 기본 개념(https://pixabay.com/)

(그림 출처: 프로그래머를 위한 강화학습(제이펍))

강화학습 개요

- **꽃밭을 건너는 아이의 예**
  - **무수한 시도 후에는 꽃을 밟지 않고 건널 수 있을 것이다 → 경험**
  - **강화학습에서의 요소라면**
    - 꽃밭 → 환경 (Environment)
    - 꽃밭의 꽃 → 상태 (State)
    - 아이 → 에이전트 (Agent)
    - 아이의 걸음 방식 → 정책 (Policy)
    - 발걸음 → 행동 (Action)
    - 걸을 때의 꾸중과 칭찬 → 보상 (Reward)

강화학습 개요

- 학습의 난이도
  - 행동과 상태의 종류가 적다면
    → 계산을 통해 쉽게 최적의 정책을 찾을 수 있다
  - 행동과 상태의 종류가 많아지면
    → 계산을 통해 최적의 정책을 찾기가 어렵다
    → 인공 신경망 도입

학습을 수행하는 강화학습 모델의 내부 요소는 신경망 모델을 사용한다

순차적 의사결정 (Sequential Decision Making)

- 강화학습이 풀고자 하는 문제는 **순차적 의사결정** 문제이다.
- 샤워 단계의 예

옷을 벗는다 →	샤워를 한다 →	물기를 닦는다 →	옷을 입는다	
옷을 입는다 →	샤워를 한다 →	물기를 닦는다 →	옷을 벗는다	
물기를 닦는다 →	샤워를 한다 →	옷을 벗는다 →	옷을 입는다	
옷을 벗는다 →	샤워를 한다 →	옷을 입는다 →	물기를 닦는다	

순차적 의사결정 (Sequential Decision Making)

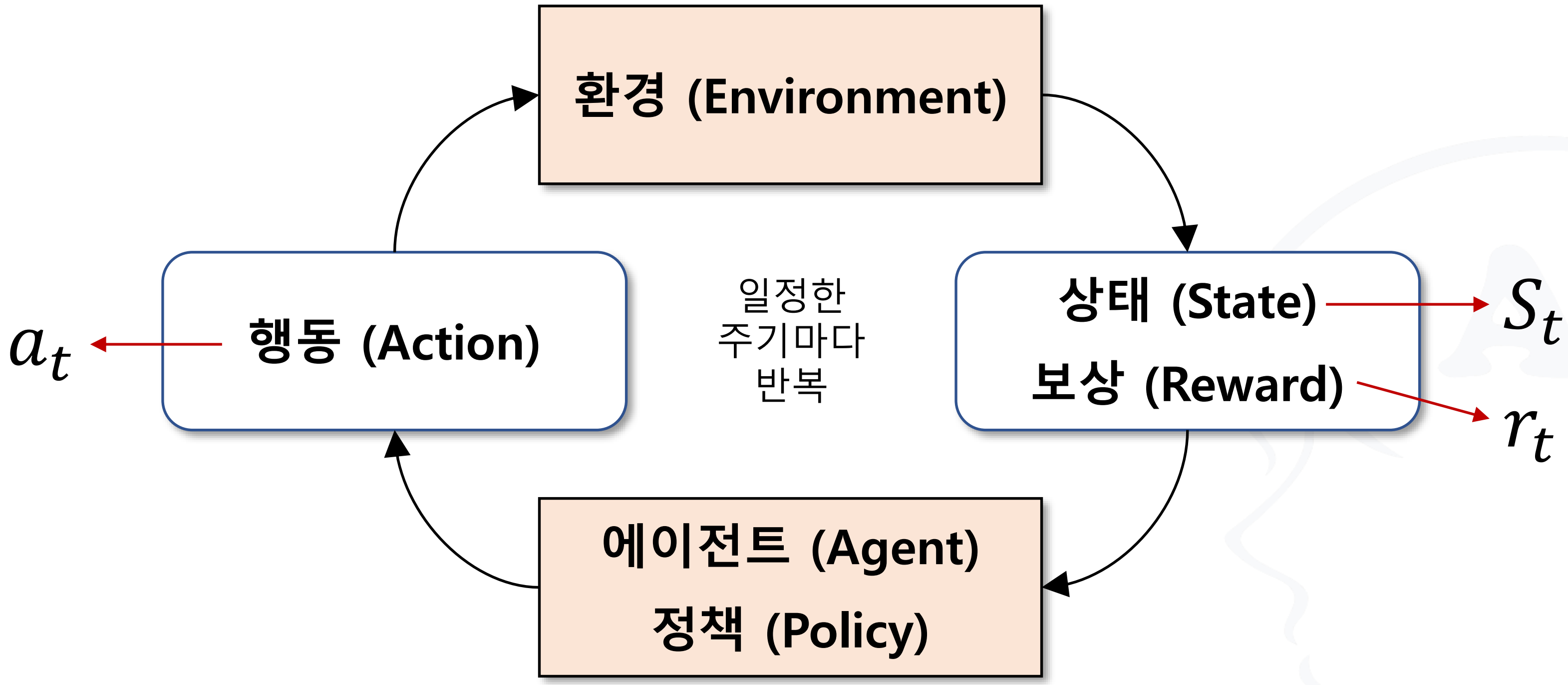
- 아무리 간단한 과정이라고 해도 이를 성공적으로 마치려면
  - 우리는 몇 가지의 의사결정을 순차적으로 해 주어야 한다.
  - 어떤 행동(의사결정)을 하고 → 그로 인해 상황이 바뀌고 → 다음 상황에서 또 다시 어떤 행동을 하고 → 또 상황이 바뀌고…
  - 각 상황에 따라 취하는 행동이 다음 상황에 영향을 줌에 따라 → 결국 연이은 행동을 잘 선택해야 상황이 잘 풀리는 문제가 → 순차적 의사결정 문제

순차적 의사결정 (Sequential Decision Making)

- **순차적 의사결정 문제의 예시**
  - **주식 투자에서의 포트폴리오 관리**
    - 정해진 예산으로 → 주식을 사는 순간부터 → 어떤 주식을 사고, 팔 것인지 매 순간 결정
    - 매 순간의 결정에 따라 수익률이 변화
  - **운전**
    - 운전을 하는 동안 변화되는 매 상황에 따라 → 도로 선택, 주행과 멈춤, 방향 전환 등 매 순간 결정
  - **게임**
    - 변화하는 게임 상황에 맞추어 → 게임의 운영, 조작, 제어 방향을 매 순간 결정

에이전트 (Agent)와 환경 (Environment)

- 순차적 의사결정 문제의 도식화



에이전트 (Agent)와 환경 (Environment)

- 순차적 의사결정 문제
  - 에이전트가 행동을 하고 그에 따라 상황이 바뀌는 것을 하나의 Loop라고 하면
  - 이 Loop가 끊임없이 반복되는 것이 순차적 의사결정 문제이다

에이전트 (Agent)와 환경 (Environment)

- 에이전트 (Agent)
  - 강화학습의 주체
  - **학습하는 대상이며, 동시에 환경속에서 행동하는 개체**를 가리킴

- 에이전트의 입장에서 보는 Loop의 동작단계
  1. 현재 상황 S_t에서 어떤 행동 a_t를 해야 할지 결정
  2. 결정된 행동 a_t를 환경으로 보냄
  3. 환경으로부터 그에 따른 보상과 다음 상태의 정보를 받음

에이전트 (Agent)와 환경 (Environment)

- **환경 (Environment)**
  - **에이전트를 제외한 모든 요소는 환경이다**
  - **상태 (State)**
    - 환경 상태에 대한 모든 정보를 숫자로 표현해서 기록해 놓은 것
    - 각각의 환경 상태에 대한 정보를 하나의 벡터로 볼 수 있다
  - **상태변화 (state Transition)**
    - 환경의 역할은 상태 변화를 일으키고
    - 행동의 결과를 알려주는 것

에이전트 (Agent)와 환경 (Environment)

- 환경이 하는 일의 단계
  1. 에이전트로부터 받은 행동 a_t를 통해서 상태 변화를 일으킴
  2. 그 결과, 상태는 $S_t \rightarrow S_{t+1}$로 바뀜
  3. 에이전트에게 줄 보상 r_{t+1}도 함께 계산
  4. S_{t+1}과 r_{t+1}을 에이전트에게 전달

보상 (Reward)

- 보상이란
  - 의사결정을 얼마나 잘 하고 있는지 알려주는 신호
  - 강화학습의 목적은 → 과정에서 받는 보상의 총합(누적 보상, Cumulative Reward)을 최대화 하는 것
  - (예) 혼자 자전거 타기를 연습하는 아이에게 보상이란?
    - 넘어지지 않고 1m를 갈 때마다 +1 이라는 식으로 보상을 결정할 수 있다
    - 이런 경우, 넘어지지 않고 최대한 멀리 달려가는 것이 학습의 목적이 됨
    - 보상을 통해 아이는 행동을 교정할 방향에 대한 힌트를 얻게 됨
  - 보상은 강화학습에 있어서 가장 중요한 개념

보상 (Reward)

- **보상의 특징**
  - **"어떻게"에 대한 정보를 가지지 않는다**
    - 어떠한 행동을 하면 그것에 대해 "얼마나" 잘 하고 있는지 평가해 줄 뿐

      → 지도학습에서의 정답과 근본적으로 다르다
    - 학습의 방향
      - 아이는 자전거를 타면서 넘어지고 달리기를 반복하면서
      - 어떻게 하면 넘어지지 않고 멀리 갈 수 있는가?
      - 어떻게 하면 더 잘 넘어지는가?

      → 와 같이 양방향의 상황에 대하여 학습할 수 있다

보상 (Reward)

- **보상의 특징**
  - **보상의 값은 스칼라(Scalar) 값이다**
    - 보상은 벡터가 아니라 크기를 나타내는 값 하나로 이루어진 스칼라 값이다
    - 보상이 벡터라면 동시에 2개 이상의 값을 목표로 할 수 있겠지만
    - 스칼라 값이기때문에 한 번에 오직 하나의 목적만을 가져야 한다

    - 현실은 다양한 목적으로 행동할 수 있지만 강화학습에 있어서 다수의 목적은 학습을 방해하는 요인이 된다
    - 강화학습은 단 하나의 목표를 최대화하도록 모델을 최적화 한다

보상 (Reward)

- **보상의 특징**
  - **보상은 희소(Sparse)할 수 있으며 지연(Delay)될 수 있다**
    - 행동 하나하나마다 일대일 대응 되지 않고, 즉각적으로 반응하지 않을 수도 있다
      → 보상이 주어질 때, 어떤 행동에 따른 보상인지 책임소재가 불분명하다
      → 학습이 어려워진다

    - 이러한 특성에 따른 문제 해결을 위하여 밸류 네트워크(Value Network) 등의 다양한 아이디어가 연구되고 있다

확률

- **확률과 확률 과정**
  - **강화학습을 이해하려면 가장 먼저 확률을 이해하여야 함**
  - **확률**
    - 어떤 사건이 실제로 일어날 것인지, 혹은 일어났는지에 대한 지식, 믿음을 표현하는 방법
    - 같은 원인에서 특정한 결과가 나타나는 비율
    - 확률의 개념에는 **무작위**라는 개념이 섞여 있다
      - 예, 주사위 던지기

확률

- 조건부 확률
  - 어떤 특정한 조건 아래에서 발생하는 확률
  - 조건부 확률의 표현: A 사건이 발생했을 때 B 사건이 발생할 확률 = P(B|A)

그림 1-3 조건부 확률(https://pixabay.com/)

(그림 출처: 프로그래머를 위한 강화학습(제이펍))

확률

- **확률 과정(Stochastic Process)**
  - **확률(Stochastic) + 과정(Process)**
  - **확률은**
    - 짧은 시간 동안에는 무작위 적이지만 긴 시간을 두고 보면 일종의 규칙을 가지고 있다
  - **과정은**
    - 시간과 연관되어 있다. 모든 과정은 시간의 흐름에 따라 결정되는 것이다.
  - **확률 과정은 $\{X_t\}$로 나타낼 수 있다 → 시간의 흐름에 따라 발생하는 랜덤 변수의 집합**
      - X : 랜덤 변수
      - t : 시간
      - { } : 집합

확률

- 확률 과정이라는 개념을 만든 이유
  - 과학적으로 어떤 개념을 해결하기 위해 가장 먼저 해야 할 일은
    → 수학적으로 현상을 표현하는 것
  - 수학적으로 표현할 수 있다면 프로그래밍을 통해 문제를 쉽게 해결 가능
    → 확률 과정이란
    시간에 따라 무작위로 변화하는 상태 또는 환경을 수학적으로 표현한 것

확률

- **확률 과정이 활용된 대표적인 사례: 브라운 운동**
  - **1827년 스코틀랜드 식물학자 로버트 브라운이 발견한 현상**
  - **물 위에서 꽃가루 입자가 불규칙적으로 운동하는 현상을 이론적으로 설명**
    - 생물체의 자발적인 움직임이라고 생각했으나 돌가루 등의 무기물도 동일하게 움직임
  - **한 지점에서 출발한 꽃가루가 일정 시간 간격으로 멋대로 움직일 때**
    - n회 움직인 후, 출발점으로부터의 거리를 측정할 수 있다
    - n이 충분히 크면 꽃가루가 어디에 위치할지에 대한 확률을 구할 수 있다

- 브라운 운동

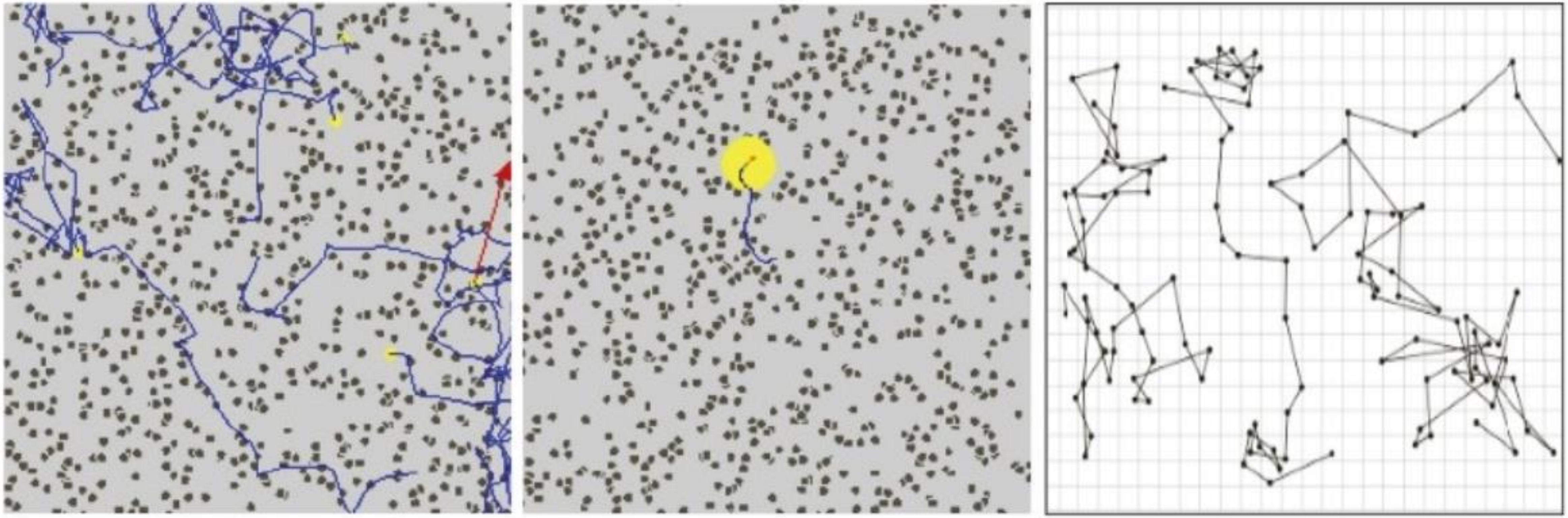


그림 1-5 브라운 운동 사례(https://en.wikipedia.org/wiki/Brownian_motion)

(그림 출처: 프로그래머를 위한 강화학습(제이펍))

마르코프 속성

- **마르코프 속성 (Markov Property)**
  - **미래는 오로지 현재에 의해 결정된다**
  - **과거에 일어났던 모든 일을 무시하고 현재의 상황만으로 미래를 예측하는 것**
  - **왜 과거의 일을 무시하는가?**
    - 사건을 단순화하기 위해서
    - 과거와 현재의 모든 상황을 고려해서 미래를 예측한다면
      → 고려해야할 문제가 감당하기 어려울 만큼 증가할 것

마르코프 속성

- 마르코프 속성을 조건부 확률로 나타내면

$$P[S_{t+1}|S_t] = P[S_{t+1}|S_1, \dots, S_t]$$

  - 시간 t 에서 상태가 S_t일 때 시간 $t+1$에서 상태가 S_{t+1}일 확률을 의미
  - 즉 S_{t+1}은 S_t에 의해서만 결정되므로 S_t만으로 S_{t+1}을 알 수 있다.

마르코프 속성

• 자루에 담긴 공의 예시

- 오늘 하나의 공을 꺼내서 다른 곳에 보관하고
- 내일 또 다른 공을 꺼내서 다른 곳에 보관하면
- 모레 나올 수 있는 공은 오늘과 내일 꺼낸 공 모두에게 영향을 받는다

→ **마르코프 속성을 만족하지 않음**

- 오늘 하나의 공을 꺼내서 다른 곳에 보관하고
- 내일 또 다른 공을 꺼낸 후 오늘 꺼낸 공을 다시 자루에 집어넣는다면
- 모레 나올 수 있는 공은 내일 꺼낸 공에게만 영향을 받는다

→ **마르코프 속성을 만족함**

자루에는 빨간색 2개, 파란색 1개, 노란색 1개, 이렇게 총 4개의 공이 들어있다

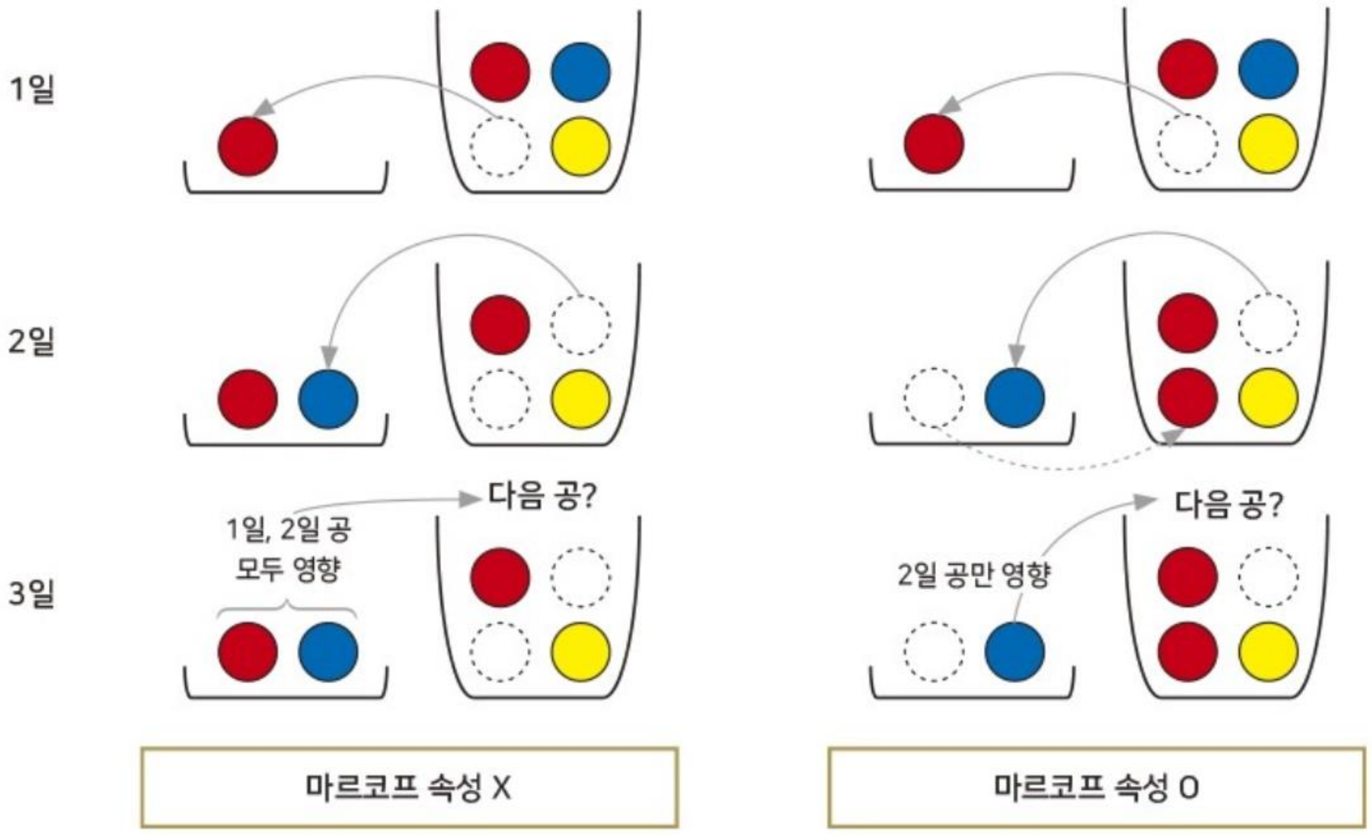


그림 1-7 마르코프 속성

(그림 출처: 프로그래머를 위한 강화학습(제이펍))

마르코프 연쇄

- **마르코프 연쇄 (Markov Chain)**
  - **마르코프 속성을 지닌 시스템의 시간에 따른 상태 변화를 나타냄**
  - **과거와 현재의 상태가 주어졌을 때,**
    - 미래 상태의 조건부 확률 분포가
    - 과거 상태와는 독립적으로 현재 상태에 의해서만 결정되는 환경
  - **이러한 상태 공간이**
    - **이산적(Discrete)**일 때: **마르코프 연쇄 (Markov Chain)**
    - **연속적(Continuous)**일 때: **마르코프 과정 (Markov Process)**

마르코프 연쇄

- **마르코프 연쇄의 두 가지 구성 요소**
  - **상태 집합 (S: Set of States)**
  - **상태 전이 매트릭스 (P: State Transition Matrix)**
    - 각 상태 별 확률을 매트릭스(행렬) 형태로 모아 놓은 것

$$P_{ss'} = P[S_{t+1} = s' | S_t = s]$$

마르코프 연쇄

- **마르코프 연쇄 – 상태 전이 매트릭스**
  - **날씨 예측 시스템의 예**

날씨 상태는 맑음과 강우 2가지
→ 조건부 확률은 모두 4가지

- 맑음 → 맑음
- 맑음 → 강우
- 강우 → 맑음
- 강우 → 강우

날씨 예측		내일날씨	
		맑음	강우
오늘날씨	맑음	0.6	0.4
	강우	0.7	0.3

조건부 확률

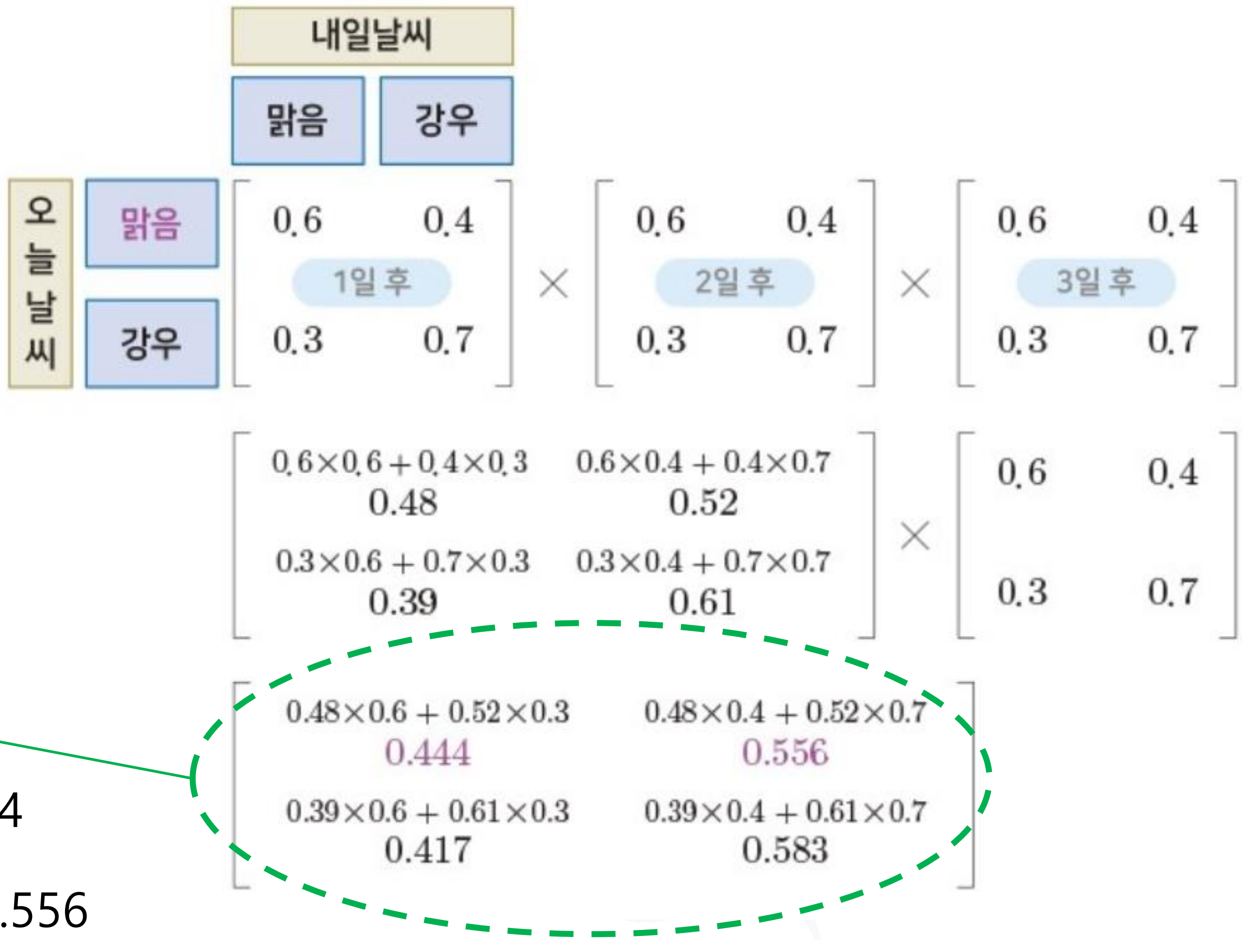
$$P = \begin{bmatrix} 0.6 & 0.4 \\ 0.7 & 0.3 \end{bmatrix}$$

상태 전이 매트릭스

(그림 출처: 프로그래머를 위한 강화학습(제이펍))

마르코프 연쇄

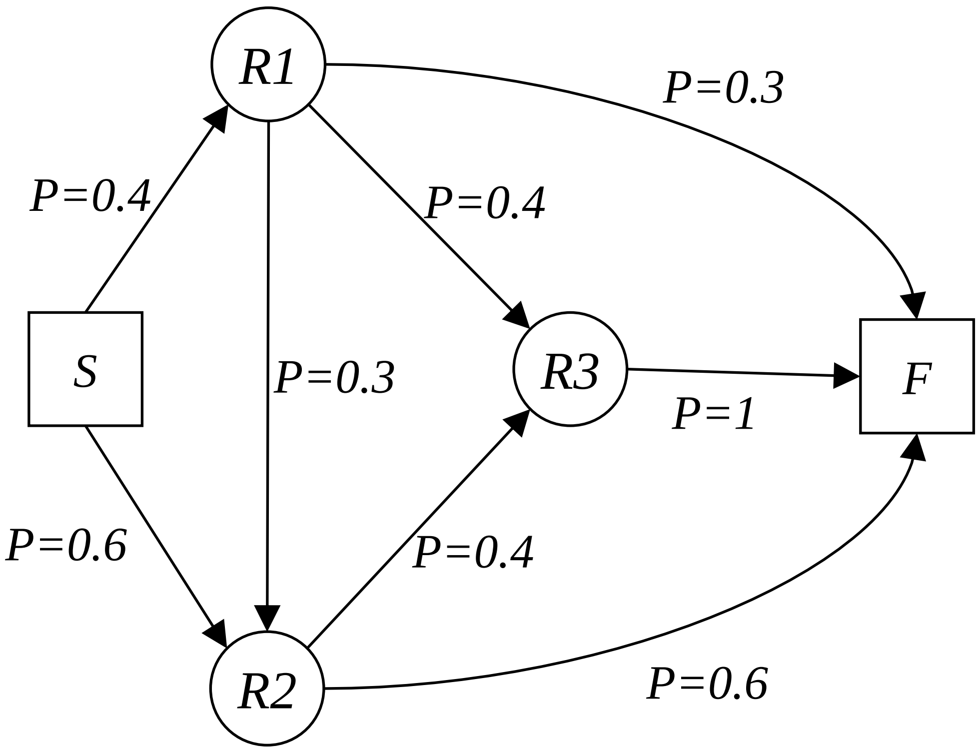
- **3일 후 날씨 예측**
  - 과거의 데이터는 고려하지 않음
  - 앞으로 일어날 일에 대한 조건부 확률만 고려하면 됨
  - 3일 후 날씨를 예측하기 위해서는 상태 전이 매트릭스를 모두 3번 곱해주면 됨
  - 3일 후의 상태 전이 매트릭스
    - 오늘 맑다면: 3일 후 맑을 확률 0.444
    - 오늘 맑다면: 3일 후 비가 올 확률 0.556



(그림 출처: 프로그래머를 위한 강화학습(제이펍))

마르코프 연쇄

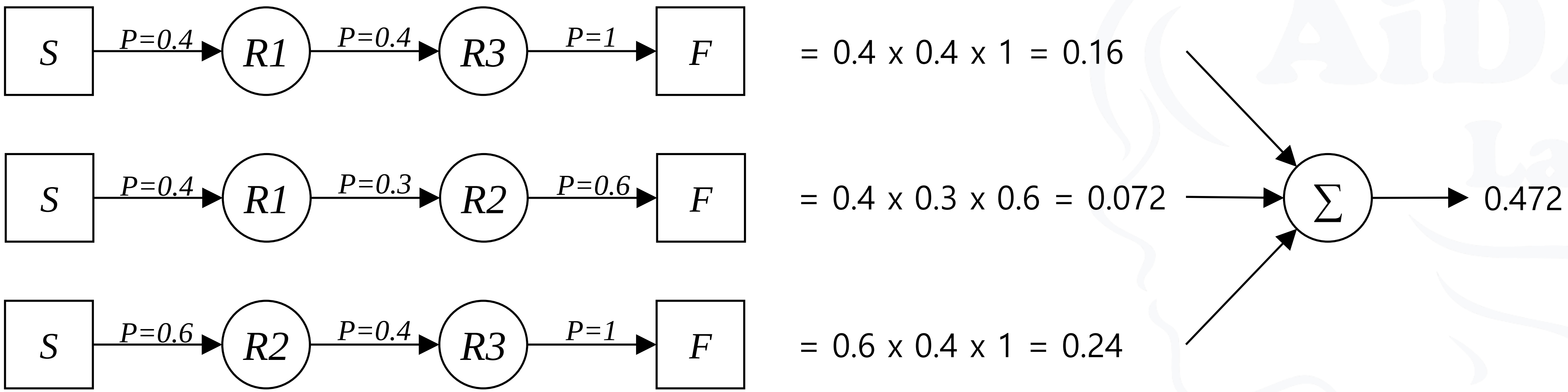
- **마르코프 연쇄의 다양한 표현**



	S	R1	R2	R3	F
S	0	0.4	0.6	0	0
R1	0	0	0.3	0.4	0.3
R2	0	0	0	0.4	0.6
R3	0	0	0	0	1
F	0	0	0	0	0

마르코프 연쇄

- **마르코프 연쇄 분석**
  - **한 타임(t)에 화살표 하나씩 이동한다고 하면** 한 타임에 하나씩 이동하는 것: 타임 스텝(Time Step)
  - **정확히 3타임만에 출발점(S)에서 목적지(F)까지 도달할 수 있는 확률은?**



마르코프 연쇄

- 마르코프 연쇄를 사용하는 목적
  - 해결하고자 하는 문제에 대한 발생 확률을 구하는 것

- 마르코프 연쇄의 활용
  - 다양한 분야에서 활용 중이며 특히 야구 통계 분야에서 널리 사용됨
  - 과거의 야구 통계 데이터를 분석해서 선수 별 평균 득점 확률을 얻고, 모델을 만들어 다음 경기에서의 예상 득점을 계산해서 어떤 선수를 등판 시킬지 결정하는 것 등

마르코프 보상 과정 (MRP)

- **마르코프 보상 과정 (MRP, Markov Reword Process)**
  - **마르코프 연쇄 + 보상(Reword) + 감마(γ, 시간에 따른 보상의 감가율)**
  - **마르코프 연쇄의 구성**
    - 상태 집합(S), 상태 전이 매트릭스(P)
    - 상태 변화에 대한 가치가 반영되지 않음
  - **마르포크 보상 과정의 구성**
    - 상태 집합(S), 상태 전이 매트릭스(P), 보상함수(R), 감가율(γ)
    - 상태 변화에 대한 가치가 반영됨

마르코프 보상 과정 (MRP)

- 마르코프 보상 과정의 구성 요소
  - S : 상태(State)의 집합
  - P : 상태 전이 매트릭스

    $$P_{ss'} = P[S_{t+1} = s' | S_t = s]$$

  마르코프 연쇄

  - R : 보상 함수

    $$R_s = E[R_{t+1} | S_t = s]$$

  - γ : 감가율

    $$\gamma \in [0, 1]$$

마르코프 보상 과정 (MRP)

- **마르코프 보상 과정의 구성 요소**
  - S : **상태 집합**
    - 다루고 있는 환경이 가질 수 있는 다양한 상태. MRP에서 상태는 유한
  - P : **상태 전이 매트릭스.**
    - 각각의 상태가 다른 상태로 변할 수 있는 조건부 확률을 매트릭스 형태로 표현한 것
    - 시간 t에서 상태가 s일 때, 시간 t+1에서 상태가 s′이 될 조건부 확률을 의미

마르코프 보상 과정 (MRP)

- R : **보상 함수**
  - 확률의 기댓값 형태로 표현
  - 시간 t에서 s일 때 시간 t+1에서 받을 수 있는 보상의 기댓값

- γ : **감가율(할인율)**
  - 시간의 흐름에 따라 가치를 얼마의 비율로 할인할 것인지 결정하는 비율
  - 지난 시간의 가치뿐만 아니라 아직 다가오지 않은 미래의 가치를 계산할 때도 사용
  - 감가율은 현재의 보상과 미래의 보상을 바라보는 관점과 관계가 있음
  - 감가율이 0 이면 미래의 보상을 전혀 고려하지 않는 것, 1이면 현재와 미래의 보상을 동일하게 평가하는 것

마르코프 보상 과정 (MRP)

- MRP의 목적: 가치를 계산하는 것
  - 보상 함수를 계산하여 한 순간의 가치만을 계산하는 것이 아니라
  - 하나의 에피소드 혹은 전체 환경의 가치를 한꺼번에 모두 계산
  - 계산된 가치는 현재 가치로 환산되어야 함
  - 하나의 에피소드 전체 가치를 계산하기 위해서는 에피소드가 끝날 때까지 몇 개의 타임 스텝을 진행해야 함 → 그래서 감가율이 필요함
  - 감가율을 사용하여 몇 타임 스텝 후에 얻을 수 있는 가치를 현재 가치로 환산 → 현 시점에서 바라보는 에피소드의 가치를 구함

마르코프 보상 과정 (MRP)

- 반환 값(G, Return) 개념 도입
  - 타입 스텝 t에서 계산한 누적 보상의 합계
  - 누적 보상은 감가율로 할인되어 계산됨
  - 반환 값은 주로 전체 환경이 아닌 에피소드 단위로 계산됨
  - 에피소드의 효율성이나 가치를 반환 값을 통해 평가
  - 반환 값을 극대화 할 수 있도록 환경을 설계하는 것이 MRP의 목적 중 하나

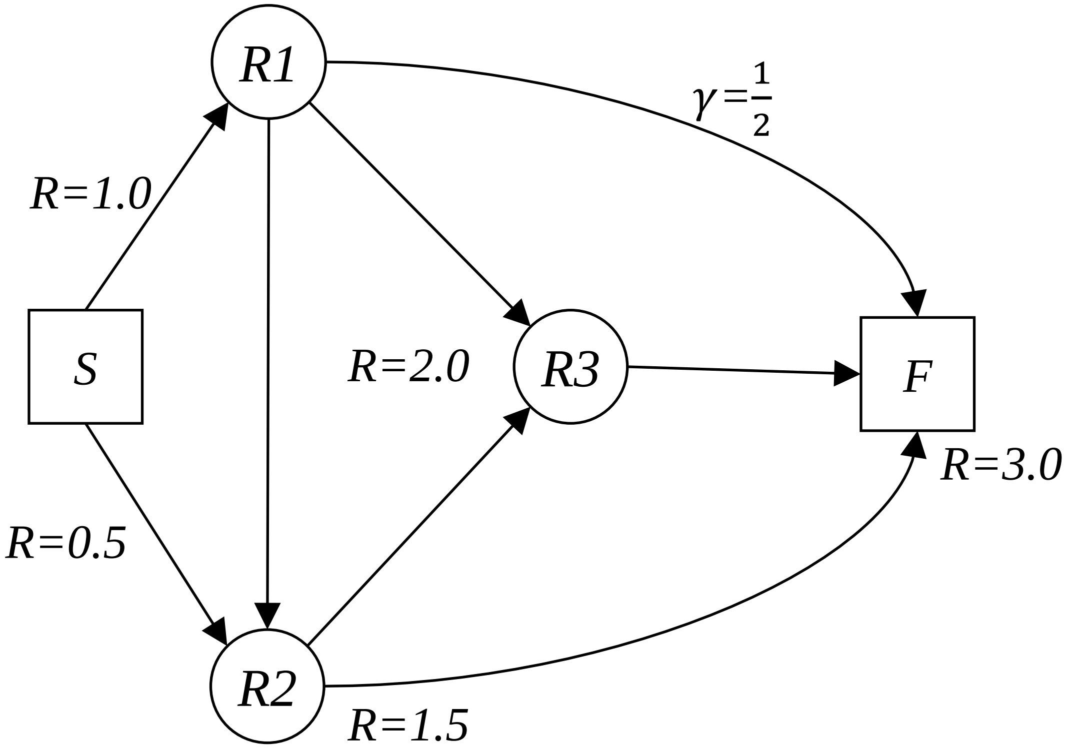
$$G_t = R_{t+1} + R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

마르코프 보상 과정 (MRP)

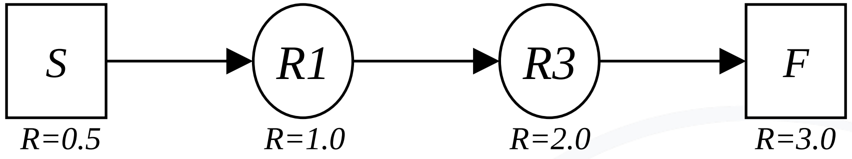
- 반환 값의 계산
  - 반환 값 계산식에서는 상태 전이 확률이 고려되지 않음
  - 반환 값은 하나의 선택된 경로(에피소드)에 대한 전체적인 보상을 계산하는 방식 → 이미 경로가 선택됨 → 상태 전이 확률을 사용할 필요가 없음

마르코프 보상 과정 (MRP)

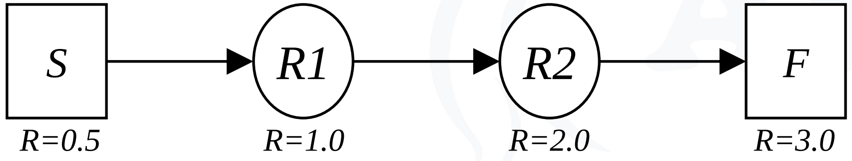
- 3 타임 스텝에 목적지에 도달하는 에피소드의 반환 값 계산



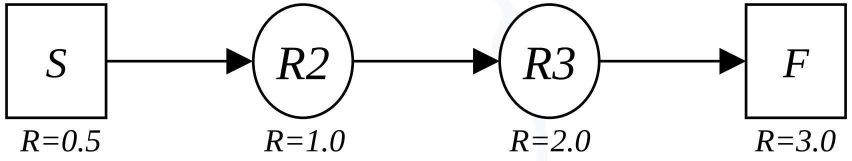
$\gamma = \frac{1}{2}$



= 0.5 + ½ x 1.0 + ½ x ½ x 2.0 + ½ x ½ x ½ x 3.0 = 1.875



= 0.5 + ½ x 1.0 + ½ x ½ x 1.5 + ½ x ½ x ½ x 3.0 = 1.75



= 0.5 + ½ x 1.5 + ½ x ½ x 2.0 + ½ x ½ x ½ x 3.0 = 2.215

마르코프 보상 과정 (MRP)

- 상태 가치 함수 (State Value Function)
  - 반환 값(G)으로 에피소드 하나에 대한 가치를 측정했다면
  - 상태 가치 함수로는 환경 전체에 대한 가치를 측정할 수 있다
  - 상태 가치 함수에서는 상태 전이 확률을 같이 고려한다

	측정 대상	특징	감가율 γ	상태 전이 확률 P
반환 값	에피소드	합계	사용	미사용
상태 가치 함수	전체 환경	기댓값	사용	사용

마르코프 보상 과정 (MRP)

- 상태 가치 함수 식 유도

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

마르코프 보상 과정 (MRP)

- 상태 가치 함수 식의 일반화 → 벨만 방정식

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

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

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

마르코프 결정 과정 (MDP)

- 마르코프 결정 과정
  - 마르코프 보상 과정(MRP) + 행동(A: Action) + 정책(π: Policy)
  - MRP가 에피소드나 환경 전체의 가치를 계산하는 것이 목적이라면
  - MDP는 환경의 가치를 극대화하는 정책을 결정하는 것이 목적이다

MRP	에이전트는 시간의 흐름(타임 스텝)에 따라 상태 전이 확률에 영향을 받으며 자연스럽게 이동
MDP	에이전트는 타임 스텝 별로 정책에 따라 행동을 선택하고 상태 전이 확률에 영향을 받아 이동

마르코프 결정 과정 (MDP)

- **MDP에서의**
  - **에이전트**
    - 행위자, 어떤 행동을 하는 주체
    - 정책($\boldsymbol{\pi}$)에 따라 행동(Action)을 하며
    - 상태(State)는 에이전트가 취한 행동과 상태 전이 확률(P)에 따라 바뀐다
  - **상태 전이 매트릭스(P):**
    - 시간 t에서 상태가 s였을 때 a라는 행동을 할 경우, 시간 t+1에서 상태가 s′일 조건부 확률
  - **보상함수(R):**
    - 시간 t에서 상태가 s 였을 때 a라는 행동을 할 경우, 시간 t+1에서 받는 보상의 기댓값

마르코프 결정 과정 (MDP)

- 마르코프 결정 과정의 구성 요소
  - S : 상태(State)의 집합
  - P : 상태 전이 매트릭스

    $$P_{ss'}^{a} = P[S_{t+1} = s' | S_t = s, A_t = a]$$

  - R : 보상 함수

    $$R_s^a = E[R_{t+1} | S_t = s, A_t = a]$$

  - γ : 감가율

    $$\gamma \in [0, 1]$$

  - A : 행동(Action)의 집합
  - π : 정책함수

마르코프 결정 과정 (MDP)

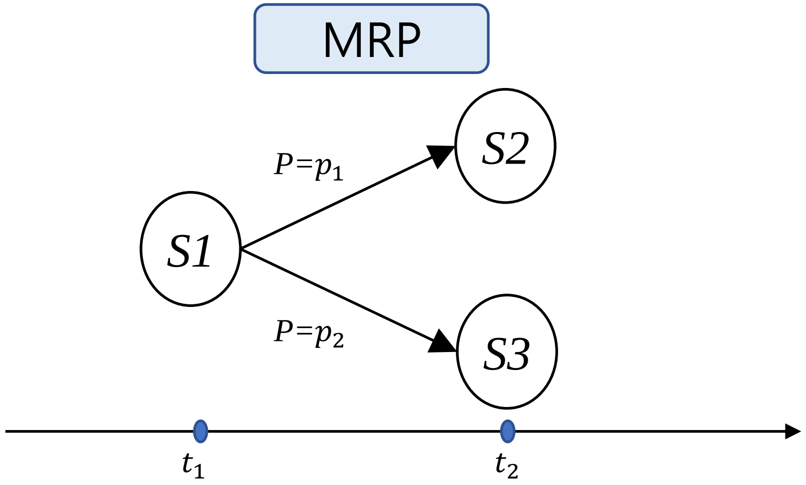
- MDP + 행동(A: Set of Actions)
  - MDP에는 행동이 추가되었기 때문에
  - 상태 전이 매트릭스와 보상 함수도 행동을 고려해야 함
  - 행동은 다음 상태에 영향을 미치는 행위이기 때문
  - MDP에서 취할 수 있는 행동의 개수는 상태와 마찬가지로 종류가 정해져 있다(유한 상태)
  - MDP 정책 공식

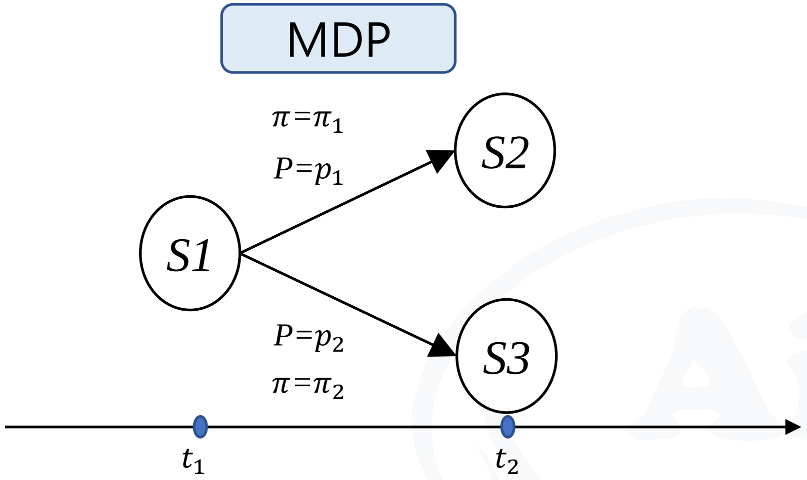
$$\pi = P[A_t = a | S_t = s]$$

마르코프 결정 과정 (MDP)

- MDP + 정책(π: Policy)
  - MDP에서의 정책: 행동을 선택하는 확률(상태 전이 매트릭스와 같은 형태)
  - 만약 4가지 종류의 행동이 있다면,
    에이전트가 한 상태에서 각각의 행동을 할 확률의 합은 1
  - 정책은 확률로 표현되기 때문에 에이전트가 정책에 따라 행동한다는 것은 항상 확률이 높은 행동을 하는 것이 아니라 확률이 높은 행동을 할 가능성이 크다는 의미이다

- **MRP와 MDP 비교**

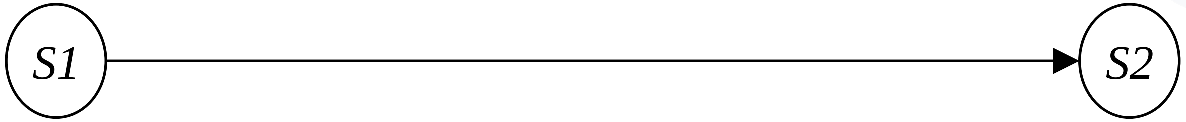




t_1에서 S_1일때 t_2에서 S_2일 확률

$\pi_1(0.4) + \pi_2(0.6) = 1$

$p_1(0.7) + p_2(0.3) = 1$



MRP $p_1 = 0.7$

MDP $\pi_1 * p_1 + \pi_2 * p_1 = 0.4 * 0.7 + 0.6 * 0.7 = 0.28 + 0.42 = 0.7$

마르코프 결정 과정 (MDP)