Chapter. 01

마르코프 결정과정

강화학습의 놀이터: MP, MRP

FAST CAMPUS ONLINE
강화학습 A-Z I

강사. 박준영

Chapter. 01

마르코프 결정과정

| 파트 2: 강화학습 문제와 가치기반 강화학습 문제의 풀이기법

"강화학습 문제"

MDP

Markov Decision Process
(마르코프 결정과정)

"강화학습 문제의 풀이기법"

Dynamic Programming
(동적최적화)

DP

MC

TD

환경에 대해서 알 때

환경에 대해서 모를 때

- + (상대적으로) 문제를 해결하기 쉬움
- + 매우 효율적임
- − 현실적이지 않음

- − (DP에 비해) 효율성이 떨어짐
- + 현실의 문제의 상황에 적용 가능

마르코프 결정과정 (Markov Decision Process: MDP)

마르코프 결정 과정은 "강화학습 문제"를 기술하는 수학적 표현방법!

MDP 를 최대한 쉽게 이해하기 몇 가지 전 단계:

- 마르코프 과정 (Markov Processes) 혹은 마르코프 연쇄 (Markov Chain) 으로도 불림
- 마르코프 보상 과정 (**M**arkov **R**eward **P**rocesses: MRP)
- 마르코프 결정 과정 (**M**arkov **D**ecision **P**rocess: MDP)

| 마르코프 특성 (Markov property)

안드레이 마르코프
(1856~1922)

"어떤 상태 s_t 는 Markov 하다" 의 정의:

$$P(s_{t+1}|s_t) = P(s_{t+1}|s_t, s_{t-1}, \dots, s_0)$$

현재 상태 s_t 를 알면, **역사**를 아는 것과 동일한 수준으로 미래 상태를 추론할 수 있다.
다른 말로, 미래의 상태는 과거와 무관하게 현재의 상태만으로 결정된다.

마르코프 과정 (Markov process)

마르코프 과정:

마르코프 과정 (마르코프 연쇄)는 $<\mathcal{S}, P>$ 인 튜플이다.

- $\mathcal{S}$ 은 (유한한) 상태의 집합
- P 는 상태 천이 행렬

한반도 문제에서,

$$\mathcal{S} = \{서울, 대전, 원주, 광주, 대구, 울산, 부산\}$$

$P =$

	서울	원주	대전	광주	대구	울산	부산
서울							
원주							
대전							
광주							
대구							
울산							
부산							

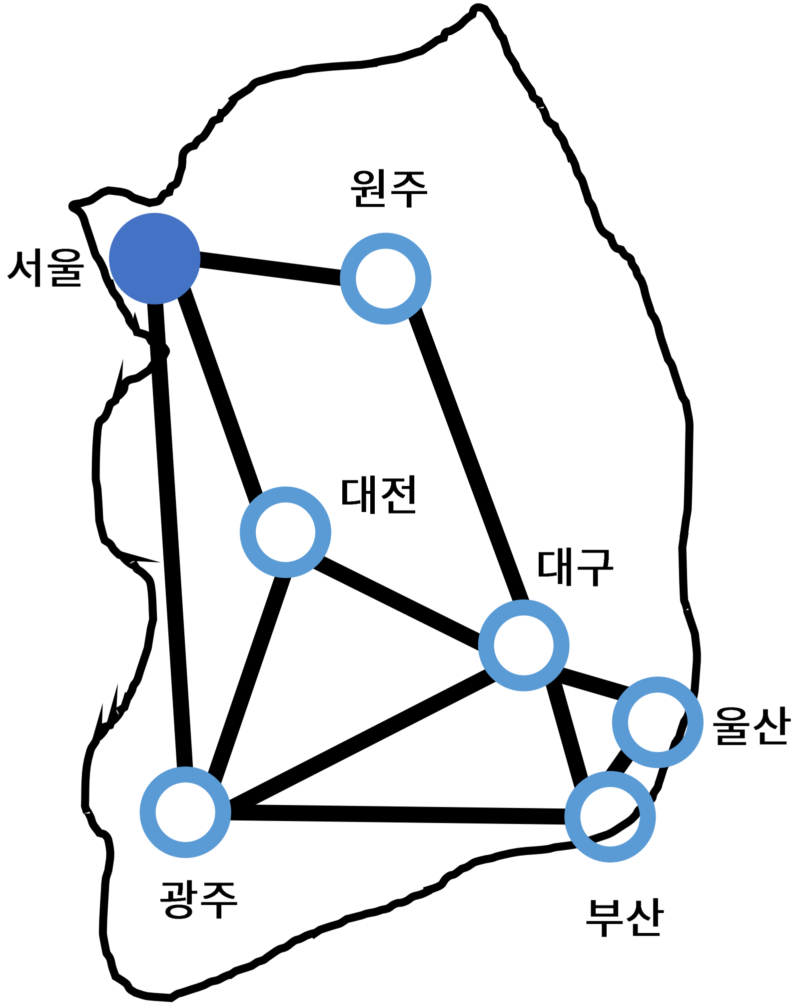
I 상태 천이 행렬 (State Transition Matrix)

만약 현재 **'서울'** 상태에 있을 때,
다음 시점에 '원주', '대전', '광주' 로 이동할 확률이 각각 $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ 이라면,

$P(S_{t+1} = 원주 \mid S_t = 서울) = {}^1/_4$
$P(S_{t+1} = 대전 \mid S_t = 서울) = {}^1/_2$
$P(S_{t+1} = 광주 \mid S_t = 서울) = {}^1/_4$ 로 표현 가능!

	서울	원주	대전	광주	대구	울산	부산
서울	0	1/4	1/2	1/4	0	0	0
원주							
대전							
광주							
대구							
울산							
부산							

가로합 =1.0



I 상태 천이 행렬 (State Transition Matrix)

현재 상태 s 에서 다음 상태 s' 로 이동할 확률 $P_{ss'}$ 이라 부르고

$$P_{ss'} \stackrel{\text{def}}{=} P(S_{t+1} = s' | S_t = s)$$

상태 천이 행렬 (State Transition Matrix) 는 모든 현재 상태에서 다음 상태로 이동할 확률을 정의한다.

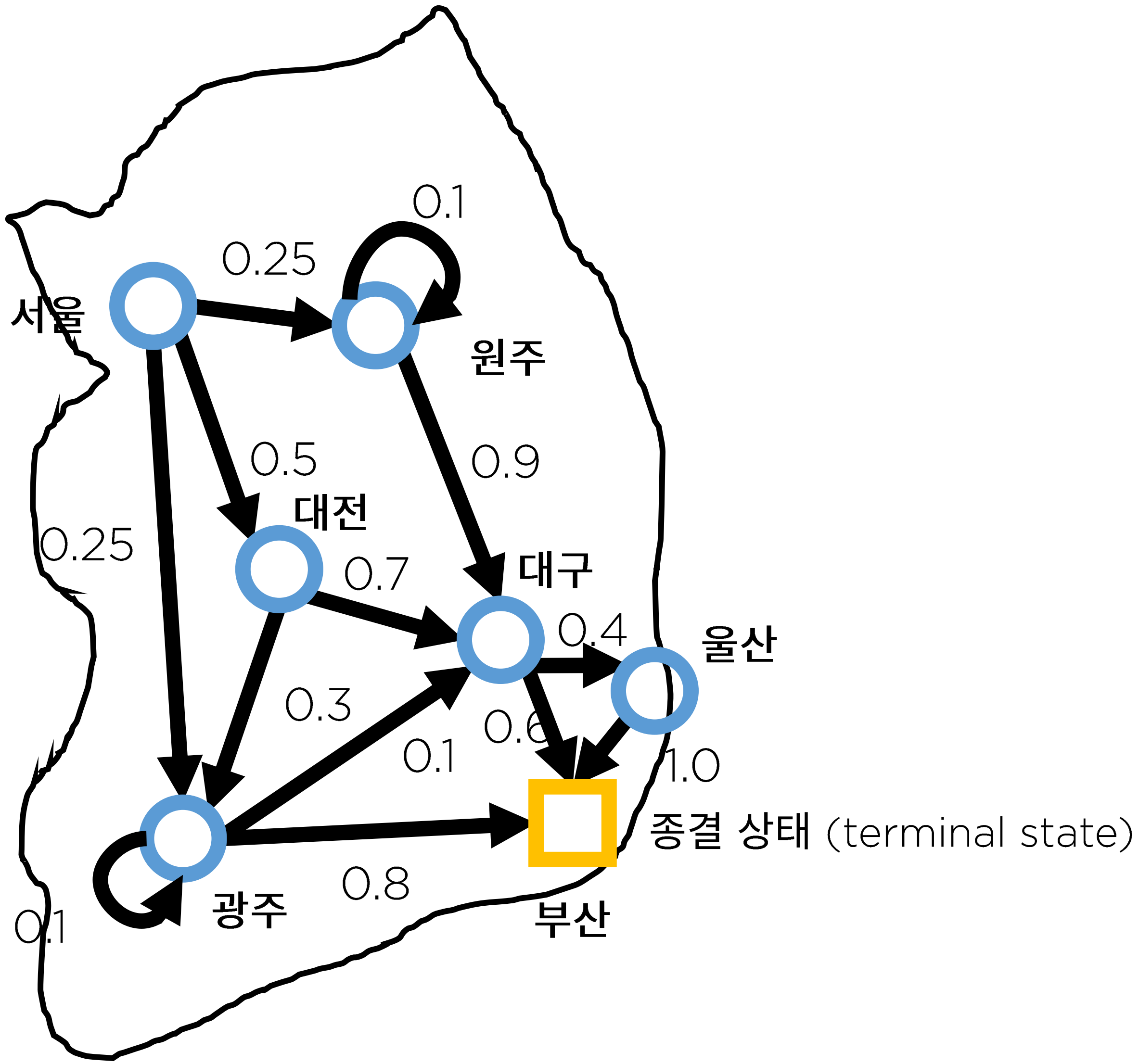
$$\begin{array}{cc} & \text{~로} \\ P = & \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix} \\ \text{~에서} & \end{array}$$

전체 상태의 개수가 n개

P_{ij} : i 에서 j 로 갈 확률

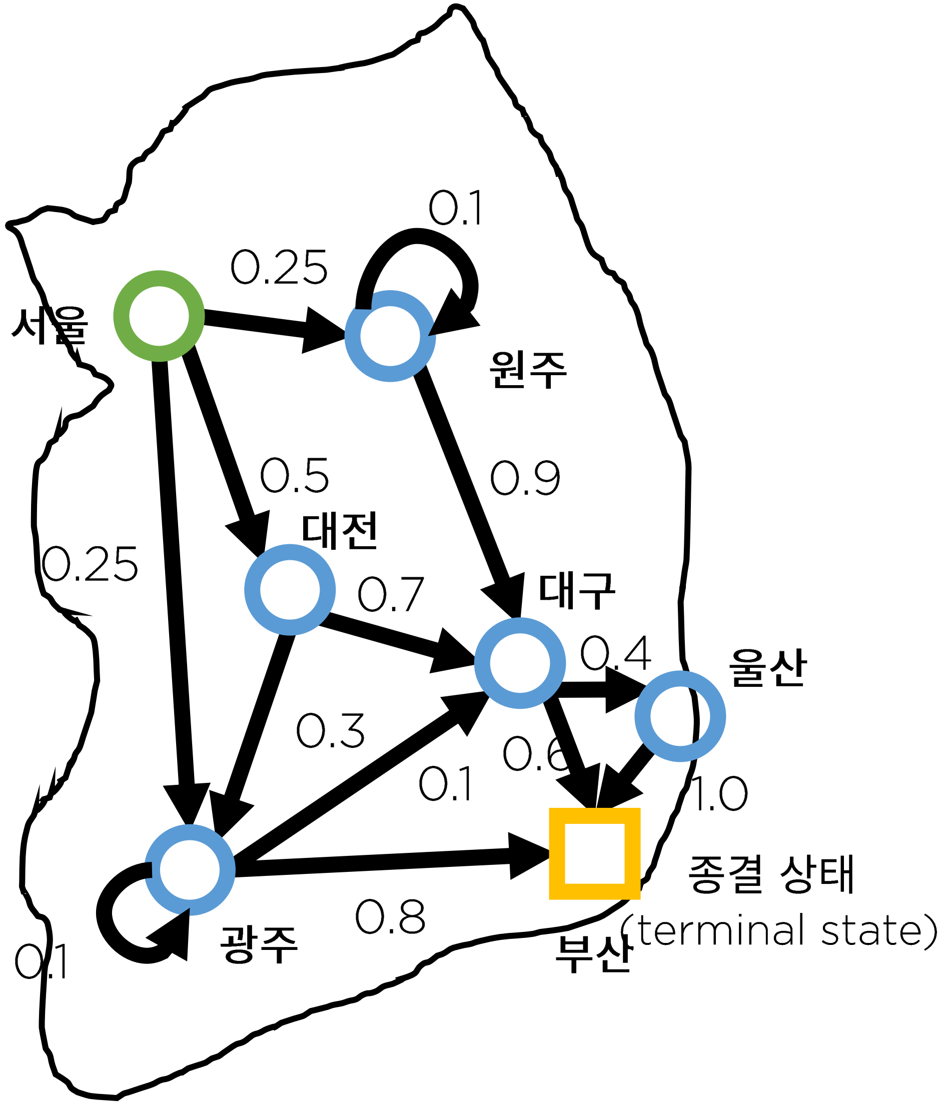
$A \stackrel{\text{def}}{=} B$ 의 의미: 'A는 B라고 정의한다'

| “한반도” 마르코프 연쇄



0.1
0.25
서울
원주
0.5
0.9
대전
0.25
0.7
대구
0.4
울산
0.3
0.6
0.1
1.0
종결 상태 (terminal state)
0.8
광주
부산
0.1

| "한반도" 마르코프 연쇄의 상태천이 행렬



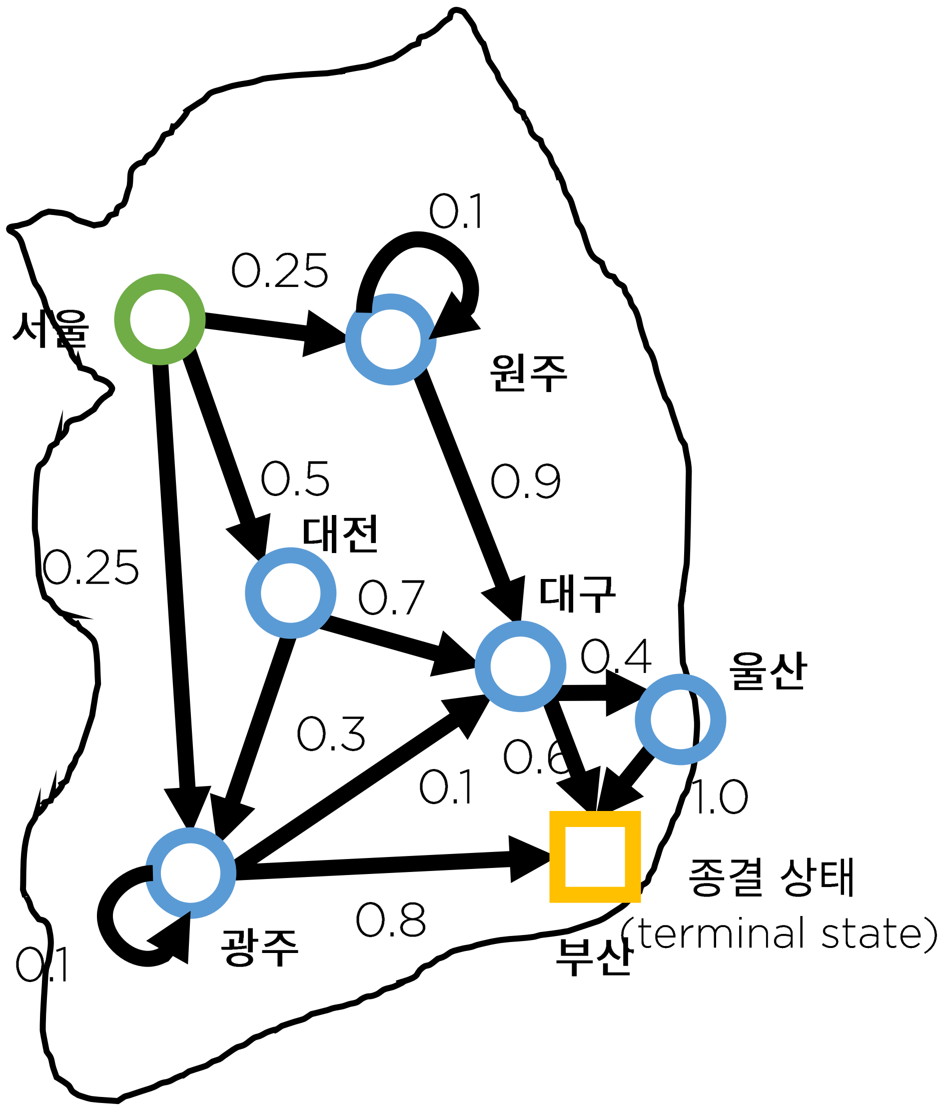
	서울	원주	대전	광주	대구	울산	부산
서울	0	0.25	0.5	0.25	0	0	0
원주	0	0.1	0	0	0.9	0	0
대전	0	0	0	0.3	0.7	0	0
광주	0	0	0	0.1	0.1	0	0.8
대구	0	0	0	0	0	0.4	0.6
울산	0	0	0	0	0	0	1.0
부산	0	0	0	0	0	0	1.0

가로 합 = 1.0

종결 상태는
항상 자기 자신으로 되돌아온다고 정의.

FAST CAMPUS ONLINE
박준영 강사.

| "한반도" 마르코프 연쇄에서의 "Episode"



서울에서 시작해서 마르코브 체인을 시작해
매 도시에 도착할 때마다 확률적으로 다음 도시를 선택하면...

Ep1. 서울 → 대전 → 대구 → 부산

Ep2. 서울 → 대전 → 대구 → 울산 → 부산

Ep3. 서울 → 대전 → 대구 → 울산 → 부산

Ep4. 서울 → 원주 → 원주 → 대구 → 울산 → 부산

Ep5. 서울 → 광주 → 부산

| 마르코브 보상 과정

마르코프 보상 과정은 마르코프 과정에 보상을 추가한 확률 과정:

마르코브 보상 과정 (MRP)는 $<\mathcal{S}, P, R, \gamma>$ 인 튜플이다.

- $\mathcal{S}$ 은 (유한한) 상태의 집합
- P 는 상태 천이 행렬
- R 는 보상 함수, $R: \mathcal{S} \rightarrow \mathbb{R}$ 확률적/결정적일 수 있음
- γ 는 감소율, $\gamma \in [0,1]$ 0이상 1이하의 실수 중 하나!

ı 리턴 (Return)

차후에, 강화학습 agent가 현재 상태로부터 미래에 일을 고려할 수 있게 해주는 장치!

리턴 G_t 는 현재 시점 t부터 전체 미래에 대한 "감가된 보상의 합"

$$G_t \overset{\text{def}}{=} R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ \dots = \sum_{\tau=1}^{\infty} \gamma^{\tau-1} R_t$$

$\gamma = 0$: 미래에 대한 고려 X

$\gamma \rightarrow 1$: 미래에 대한 고려가 커짐

R_t 는 확률 **변수**, 즉 아직 하나의 결정된 값이 아님.

연금???

I 리턴 (Return) 은 왜 감가 하나요? ($\gamma < 1.0$)

$$G_t \overset{\text{def}}{=} R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ \dots = \sum_{\tau=1}^{\infty} \gamma^{\tau-1} R_t$$

1. 감가하면 값이 무한히 커지는 것을 방지해, 계산하기에 쉬워진다.
2. 수학적인 분석이 쉬워진다.
3. '먼 미래는 불확실하다'는 철학을 반영
4. 사람은 먼 미래에 실현되는 이익을 선호하지 않는다.

하.지.만. 한반도 마르코프 연쇄와 같이, 항상 <u>에피소드가 끝나는 경우</u>에는 $\boldsymbol{\gamma = 1.0}$ 를 사용하기도 한다!

I 가치 함수 (Value function)

가치 함수 $V(s)$ 는 현재 상태 s에서 미래의 $\underbrace{\text{"감가된 보상의 합"}}_{\text{리턴 } (G_t)}$ 의 기댓값

$$V(s) = \mathbb{E}[G_t \mid S_t = s]$$

현재 t의 상태가 s 일 때, 미래의 기대 리턴은 얼마인가?

(평균)

ㅣ흥미로운 항등식: Bellman (expectation) equation

$$V(s) = \mathbb{E}[G_t \mid S_t = s] \quad G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ \ldots \text{ 을 대입}$$

$$= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ \ldots \mid S_t = s]$$

$$= \mathbb{E}[R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \ \ldots) \mid S_t = s]$$

$$(4) = \mathbb{E}[R_{t+1} + \gamma \boldsymbol{G_{t+1}} \mid S_t = s] \quad \boldsymbol{G_{t+1}} = R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+2} + \ \ldots$$

$$(5) = \mathbb{E}[R_{t+1} + \gamma \boldsymbol{V(S_{t+1})} \mid S_t = s]$$

(4) → (5): $\mathbb{E}[A] = \mathbb{E}[\mathbb{E}[A]]$ 를 활용.

$$\begin{aligned} \mathbb{E}[R_{t+1} + \gamma \boldsymbol{G_{t+1}} \mid S_t = s] &= \mathbb{E}[R_{t+1} \mid S_t = s] + \mathbb{E}[\gamma \boldsymbol{G_{t+1}} \mid S_t = s] \\ &= \mathbb{E}[R_{t+1} \mid S_t = s] + \gamma \mathbb{E}[\boldsymbol{G_{t+1}} \mid S_t = s] \\ &= \mathbb{E}[R_{t+1} \mid S_t = s] + \gamma V(S_{t+1}) \\ &= \mathbb{E}[R_{t+1} \mid S_t = s] + \gamma \mathbb{E}[\boldsymbol{V(S_{t+1})} \mid S_t = s] \\ &= \mathbb{E}[R_{t+1} + \gamma \boldsymbol{V(S_{t+1})} \mid S_t = s] \end{aligned}$$

리처드 어니스트 벨먼 (1920-1984)

- 1953년에 동적 계획법을 고안

l 흥미로운 항등식 : Bellman (expectation) equation

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

$$\Downarrow$$

$$V(s) = R(s) + \gamma \sum_{s' \in S_{t+1}} P_{ss'} V(s')$$

'S가 유한하다 / $R(s)$ 가 결정적이다'를 가정

I 흥미로운 항등식 : Bellman (expectation) equation

Bellman (기대) 방정식은 선형 연립방정식을 활용해 표현 가능!

$$v = R + \gamma P v$$

(보통, 소문자 영어는 벡터를,
대문자 영어는 매트릭스를 표현하는데 사용됩니다.)

v 는 모든 상태의 value function 값을 담은 벡터. $v \in \mathbb{R}^n$ ($\mathbb{R}$은 실수공간을 의미함)

R 은 모든 상태의 reward function 값을 담은 벡터. $R \in \mathbb{R}^n$

γ 은 감가율. $\gamma \in \mathbb{R}$ (혹은 $\gamma \in \mathbb{R}^1$ 로 표현)

P 은 상태 천이 매트릭스. $P \in \mathbb{R}^{n \times n}$

$$\begin{bmatrix} V(1) \\ \vdots \\ V(n) \end{bmatrix} = \begin{bmatrix} R(1) \\ \vdots \\ R(n) \end{bmatrix} + \gamma \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix} \begin{bmatrix} V(1) \\ \vdots \\ V(n) \end{bmatrix}$$

Q1) 왜 복잡하게 행렬로 쓰나요 ☹
A1) 계산하기 편합니다 ☺

I Bellman 기대 방정식의 풀이법

I : 단위행렬
대각 성분을 제외하고 모두 0.
대각 성분은 1.

$$v = R + \gamma P v$$
$$(I - \gamma P)v = R$$
$$v = (I - \gamma P)^{-1} R$$

- Bellman 기대 방정식은 선형 방정식이기 때문에 직접적으로 해를 구하는 것 (수식을 만족하는 v 를 찾는 것) 이 가능.
- 하지만, n 커질수록 직접적으로 문제를 푸는 것이 어려워 짐
  - **DP, MC, TD** 등을 활용해 문제를 풀 수 있음

| 마무리!

- 마르코프 특성 (Markov property)
- 마르코프 과정 (Markov processes)
  - 상태 집합 $\mathcal{S}$
  - 상태 천이 행렬 P
- 마르코프 보상 과정 (Markov reward process: MRP)
  - 보상함수 R
  - 감가율 γ
  - 리턴 G_t / 가치 함수 $V(s)$
  - Bellman 방정식
  - Bellman 방정식 풀이기법