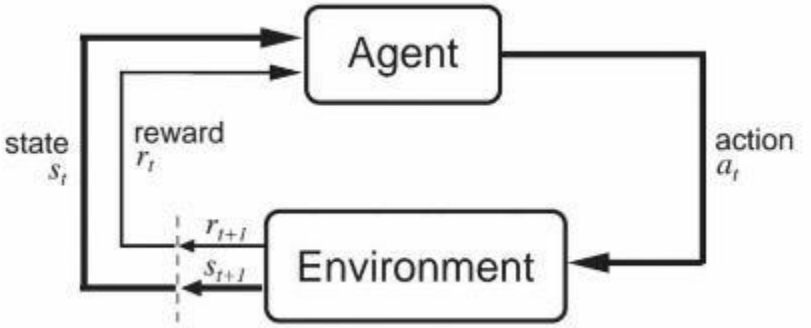
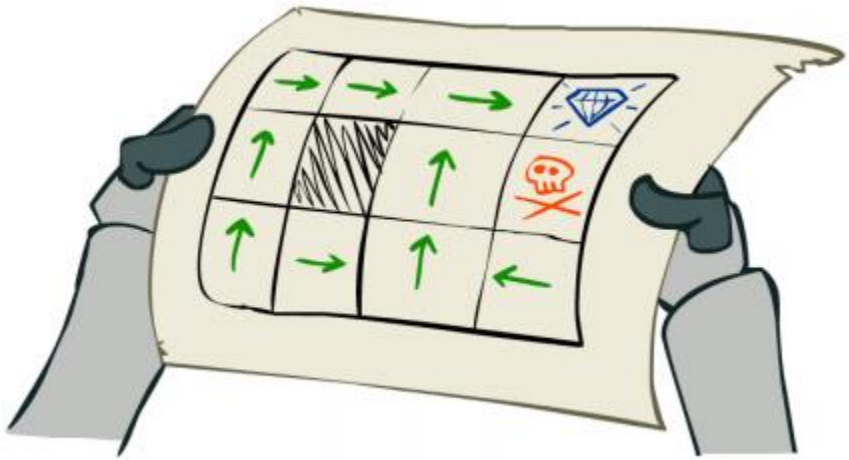
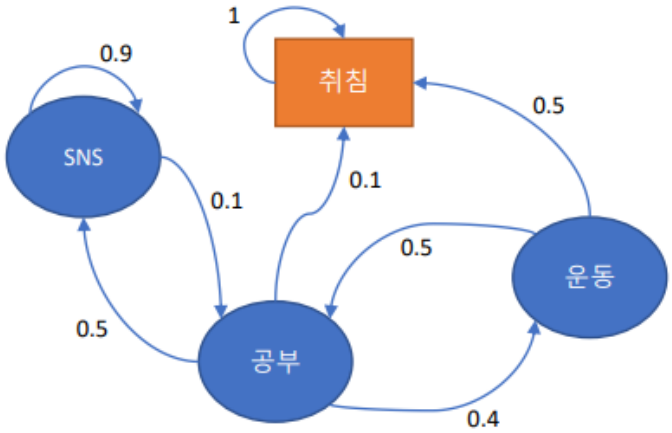
마르코브 보상 프로세스 예시



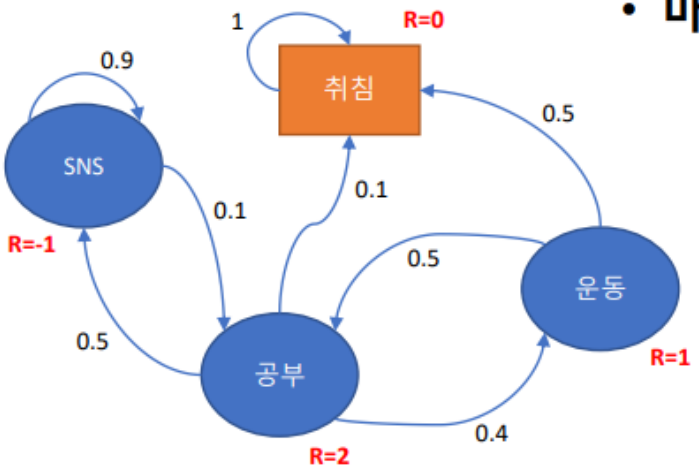


마르코브 보상 프로세스



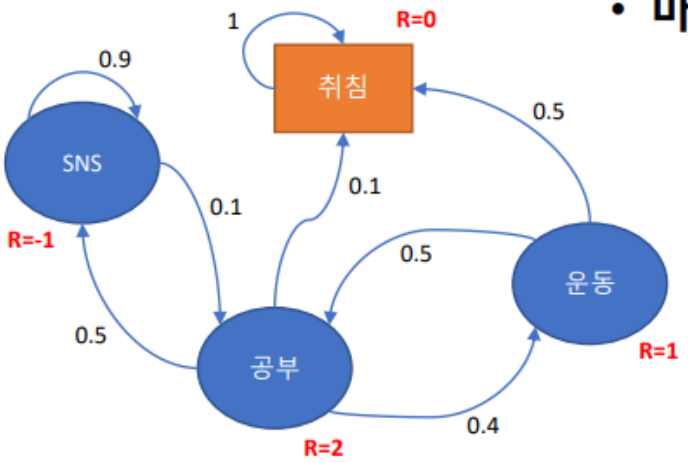
	공부	SNS	운동	취침
공부				
SNS				
운동				
취침				

마르코브 보상 프로세스



- **마르코브 보상 프로세스 (Markov Reward Process)**
  - 구성요소
    - S=상태집합
    - P=상태전이확률. $P[S_{t+1} = s' | S_t = s]$
    - R=보상(reward)함수. $R_s = E[R_t | S_t = s]$
    - γ=감가율 (discounting factor)
  - 에피소드 (episode)
    - 특정 상태로부터 시작하여 종료 상태까지의 (상태,보상) sequence를 의미
    - 현재 상태 S_t = 공부 일 경우,
      - 공부 – SNS – SNS – 공부 – 운동 – 취침
      - 공부 – SNS – 공부 – 운동 – 취침
      - 공부 – 운동 – 공부 – 취침
      - 공부 – 운동 – 공부 – SNS – 공부 – 취침
      - 공부 – 취침
      - …

마르코브 보상 프로세스

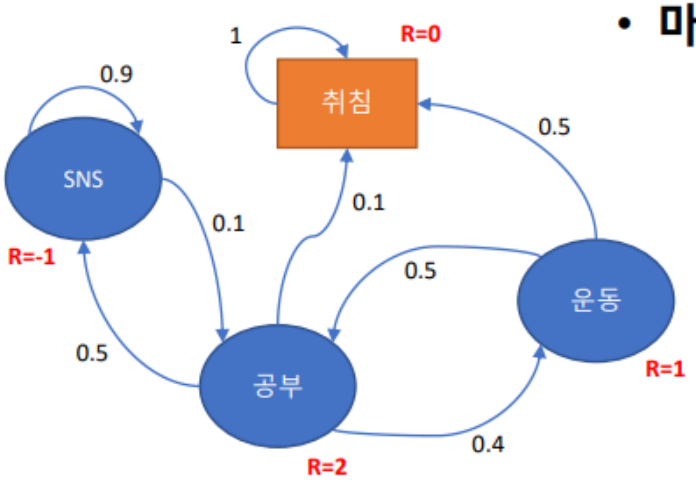


- **마르코브 보상 프로세스 (Markov Reward Process)**
  - 리턴 (Return) G_t
    - t 번째 시각 이후의 (감가율이 반영된) 누적 보상

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

에피소드 (S_t =공부, $\gamma = 1/2$)	리턴 G_t
공부 – SNS – SNS – 공부 – 운동 – 취침	$2 + 0.5 \times (-1) + 0.5^2(-1) + 0.5^3 \times 2 + 0.5^4 \times 1 + 0.5^5 \times 0 = 1.5625$
공부 – SNS – 공부 – 운동 – 취침	$2 + 0.5 \times (-1) + 0.5^2 \times 2 + 0.5^3 \times 1 + 0.5^4 \times 0 = 2.15$
공부 – 운동 – 취침	$2 + 0.5 \times 1 + 0.5^2 \times 0 = 2.5$
공부 – 취침	$2 + 0.5 \times 0 = 2$

마르코브 보상 프로세스



- **마르코브 보상 프로세스 (Markov Reward Process)**
  - 가치함수 (value function) $v(s)$
    - 특정 상태에서의 리턴의 기대값
      - 상태 s로 부터 시작하는 프로세스로 부터 기대할 수 있는 누적보상의 평균

$$v(s) = E[G_t | S_t = s]$$

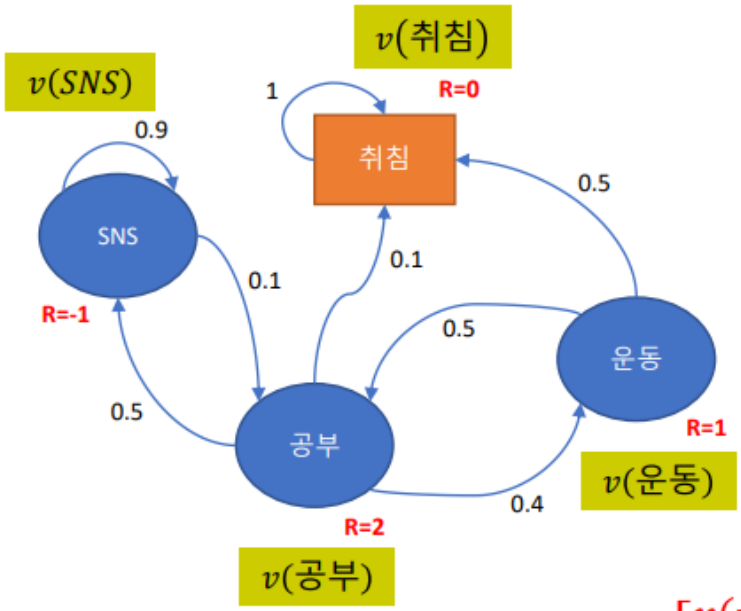
    - 프로세스가 진행됨에 있어, 상태 s가 보상측면에서 얼마나 좋은 상태인지를 평가하기 위한 지표로 활용

마르코브 보상 프로세스

- **마르코브 보상 프로세스 (Markov Reward Process)**
  - 가치함수 $v(s) = E[G_t|S_t = s]$를 어떻게 계산할 것인가?
    - 각 상태 s에 대해 모든 에피소드와 그의 확률을 계산하여 이들을 활용해 평균을 계산할 것인가?
    - 대신 다음과 같은 재귀식을 통해 계산

$$
\begin{aligned}
v(s) &= E[G_t|S_t = s] \\
&= E[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots |S_t = s] \\
&= E[R_t + \gamma(R_{t+1} + \gamma R_{t+2} + \cdots)|S_t = s] \\
&= E[R_t + \gamma G_{t+1}|S_t = s] \\
&= E[R_t + \gamma v(S_{t+1})|S_t = s] \\
&= E[R_t|S_t = s] + \gamma E[v(S_{t+1})|S_t = s] \\
&= R_s + \gamma \sum_{s' \in S} P[S_{t+1} = s'|S_t = s] v(s')
\end{aligned}
$$

마르코브 보상 프로세스



$$v(s) = R_s + \gamma \sum_{s' \in S} P[S_{t+1} = s' | S_t = s] v(s')$$

$$v(\text{공부}) = 2 + \gamma[0.5 \times v(SNS) + 0.4 \times v(\text{운동}) + 0.1 \times v(\text{취침})]$$

$$v(SNS) = (-1) + \gamma[0.1 \times v(\text{공부}) + 0.9 \times v(SNS)\,]$$

$$v(\text{운동}) = 1 + \gamma[0.5 \times v(\text{공부}) + 0.5 \times v(\text{취침})\,]$$

$$v(\text{취침}) = 0 + \gamma[1 \times v(\text{취침})\,]$$

$$\begin{bmatrix} v(\text{공부}) \\ v(SNS) \\ v(\text{운동}) \\ v(\text{취침}) \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \\ 1 \\ 0 \end{bmatrix} + \gamma \begin{bmatrix} 0 & 0.5 & 0.4 & 0.1 \\ 0.1 & 0.9 & 0 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v(\text{공부}) \\ v(SNS) \\ v(\text{운동}) \\ v(\text{취침}) \end{bmatrix}$$

$$V = R + \gamma P V$$

마르코브 보상 프로세스

$$V = R + \gamma PV$$

$$\Rightarrow \quad IV = R + \gamma PV$$

$$\Rightarrow \quad (I - \gamma P)V = R$$

$$\Rightarrow \quad V = (I - \gamma P)^{-1} R$$

$$\begin{bmatrix} v(\text{공부}) \\ v(SNS) \\ v(\text{운동}) \\ v(\text{취침}) \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - \gamma \begin{bmatrix} 0 & 0.5 & 0.4 & 0.1 \\ 0.1 & 0.9 & 0 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 2 \\ -1 \\ 1 \\ 0 \end{bmatrix}$$