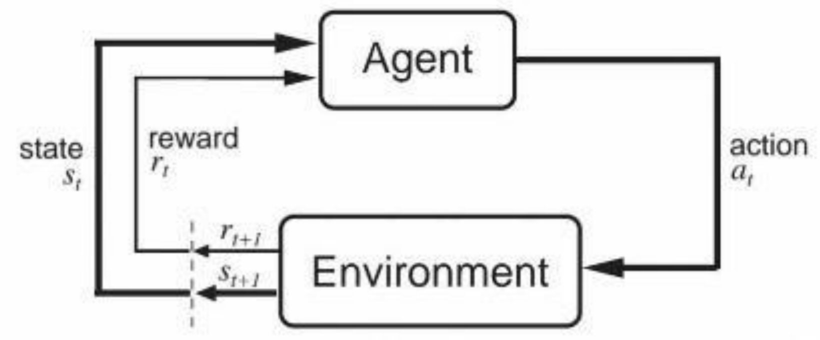
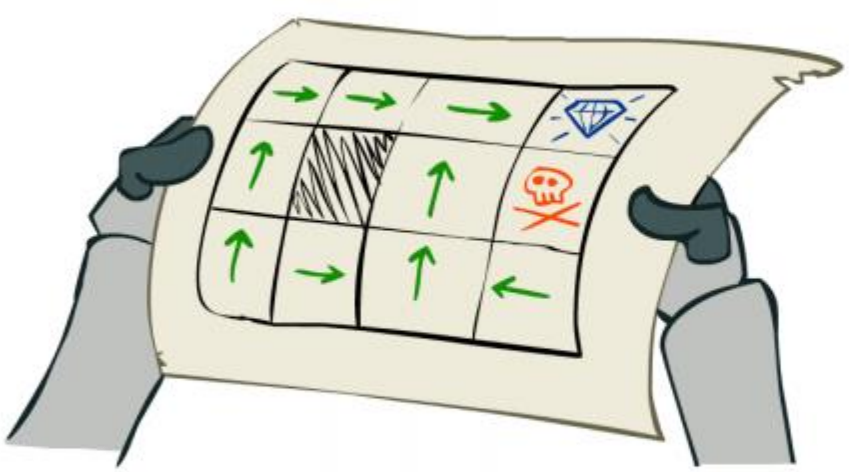
# Infinite-horizon MDP



# Infinite-horizon MDP

- **강화학습의 기본 수학적 모델 (모델기반 RL)**

- **Infinite-horizon**
  - 프로세스가 무한이 지속된다고 가정

- **정상성 (stationary) 가정**
  - 보상과 상태전이행렬이 의사결정시점 (단계) $t$에 의존적이지 않음
    - $r_t(s,a) \Longrightarrow r(s,a)$
    - $p_t(s'|s,a) \Longrightarrow p(s'|s,a)$
  - 정상 정책 (stationary policy) 고려
    - 이론적으로 최적의 정상 정책 (optimal stationary policy)이 존재
    - $\pi = (\delta, \delta, \dots) = \delta^{\infty}$: 모든 단계에서 동일한 의사결정 규칙 $\delta$ 적용
    - $\delta(s) \Leftrightarrow \pi(s), \delta(a|s) \Leftrightarrow \pi(a|s)$

# Infinite-horizon MDP

- **정책 $\pi$가 주어졌을 때 상태 $s$의 가치함수 $v^{\pi}(s)$**
  - 상태 $s$로 부터 정책 $\pi$를 따라갔을 때 받을 것으로 예상되는 감가율이 고려된 총 보상합의 기대치

$$
\begin{aligned}
v^{\pi}(s) &\equiv E_s^{\pi}\left[r\big(s_1, \pi(s_1)\big) + \gamma r\big(s_2, \pi(s_2)\big) + \gamma^2 r\big(s_3, \pi(s_3)\big) + \cdots \mid s_1 = s\right] \\
&\equiv E_s^{\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r\big(s_t, \pi(s_t)\big) \mid s_1 = s\right]
\end{aligned}
$$

$\star\ \gamma \in [0,1)$ 는 감가율

- 감가율 (discount factor) $\gamma$
  - $\gamma = 0$: 오직 현재 시점의 보상만을 중요시함
  - $\gamma \approx 1$: 먼 미래의 보상을 현재와 가까운 미래의 보상만큼 중요시함

# Infinite-horizon MDP

- **벨만 기대 방정식 (Bellman Expectation Equation)**
  - 정책 $\pi$가 주어졌을 때 상태 $s$의 가치함수 $v^{\pi}(s)$
  - $\pi \equiv \delta^{\infty}$

$$
\begin{aligned}
v^{\pi}(s) &\equiv E_s^{\pi}[r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots | s_1 = s] \\
&= E_s^{\pi}[r_1 | s_1 = s] + \gamma E_s^{\pi}[r_2 + \gamma r_3 + \cdots | s_1 = s] \\
&= r(s, \pi(s)) + \gamma \textstyle\sum_{s'} P\big(s' | s, \pi(s)\big)\, E_s^{\pi}[r_2 + \gamma r_3 + \cdots | s_2 = s'] \\
&= r(s, \pi(s)) + \gamma \textstyle\sum_{s'} P\big(s' | s, \pi(s)\big)\, v^{\pi}(s')
\end{aligned}
$$

  - 현재상태 가치 함수와 다음상태 가치함수 간의 관계 방정식

**<u>$\pi$ 가 확률적 정책일 경우,</u>**

$$
v^{\pi}(s) = \sum_{a \in A} \pi(a|s) \overbrace{\left[ r(s,a) + \gamma \sum_{s'} P(s'|s,a)\, v^{\pi}(s') \right]}^{\sum_{s'} P(s'|s,a)\,[r(s,a,s') + \gamma v^{\pi}(s')]}
$$

# Infinite-horizon MDP

- 정책 평가 (Policy evaluation)

$$v^{\pi}(s) = r(s,\pi(s)) + \gamma \sum_{s'} P\big(s' | s, \pi(s)\big)\, v^{\pi}(s')$$

$$\begin{bmatrix} v^{\pi}(s_1) \\ v^{\pi}(s_2) \\ v^{\pi}(s_3) \end{bmatrix} = \begin{bmatrix} r(s_1,\delta(s_1)) \\ r(s_2,\delta(s_2)) \\ r(s_3,\delta(s_3)) \end{bmatrix} + \gamma \begin{bmatrix} P(s_1|s_1,\delta) & P(s_2|s_1,\delta) & P(s_3|s_1,\delta) \\ P(s_1|s_2,\delta) & P(s_2|s_2,\delta) & P(s_3|s_2,\delta) \\ P(s_1|s_3,\delta) & P(s_2|s_3,\delta) & P(s_3|s_3,\delta) \end{bmatrix} \begin{bmatrix} v^{\pi}(s_1) \\ v^{\pi}(s_2) \\ v^{\pi}(s_3) \end{bmatrix}$$

$$V^{\pi} = R_{\pi} + \gamma P_{\pi} V^{\pi}$$

$$\Rightarrow \quad V^{\pi} = (I - \gamma P_{\pi})^{-1} R_{\pi}$$

# Infinite-horizon MDP

- **최적 가치함수 (optimal value function)** $v^*(s)$

$$(\text{모든 } s\text{에 대해})\ v^*(s) = \max_{\pi} v^{\pi}(s)$$

- **최적 정책 (optimal policy)** $\pi^*$ 모든 $s, \pi$에 대해 $v^{\pi^*}(s) \geq v^{\pi}(s)$

$$(\text{모든 } s\text{에 대해})\ \ v^{\pi^*}(s) = v^*(s)$$

  - 결국 $v^*(s)$는 상태 s로 부터 최적정책을 따르를 때 얻을 수 있는 기대값 (감가율이 고려된 누적보상합의 최대값)

# Infinite-horizon MDP

- 벨만 최적 방정식(Bellman optimality equation)

$$v_t(s_t) = \max_{a_t \in A_{s_t}} \left\{ r_t(s_t, a_t) + \gamma \sum_{j \in S} p(s_{t+1} | s_t, a_t) v_{t+1}(s_{t+1}) \right\}$$

$v^*(s) = \max_{\pi} v^{\pi}(s)$ for all $s$

$$v^*(s) = \max_{a} \left\{ r(s,a) + \gamma \sum_{s'} P(s'|s,a) \, v^*(s') \right\}$$

$$\delta^*(s) = \operatorname*{argmax}_{a} \left\{ r(s,a) + \gamma \sum_{s'} P(s'|s,a) \, v^*(s') \right\}$$

$\pi^* = (\delta^*)^{\infty}$: 최적 (stationary) 정책

# Infinite-horizon MDP

- 최적 행동-가치함수 $Q^*(s,a)$
  - 상태 $s$에서 행동 $a$를 결정한 이후, 최적정책을 따를 때 얻을 수 있는 기대값 (감가율이 고려된 누적보상합의 최대값)

$$Q^*(s,a) = r(s,a)+\gamma \sum_{s'} P(s'|s,a)\, v^*(s')$$

⬇

$$v^*(s) = \max_a \left\{ \underbrace{r(s,a)+\gamma \sum_{s'} P(s'|s,a)\, v^*(s')}_{Q^*(s,a)} \right\} = \max_a Q^*(s,a)$$

$$\delta^*(s) = \operatorname*{argmax}_a \left\{ r(s,a)+\gamma \sum_{s'} P(s'|s,a)\, v^*(s') \right\} = \operatorname*{argmax}_a Q^*(s,a)$$