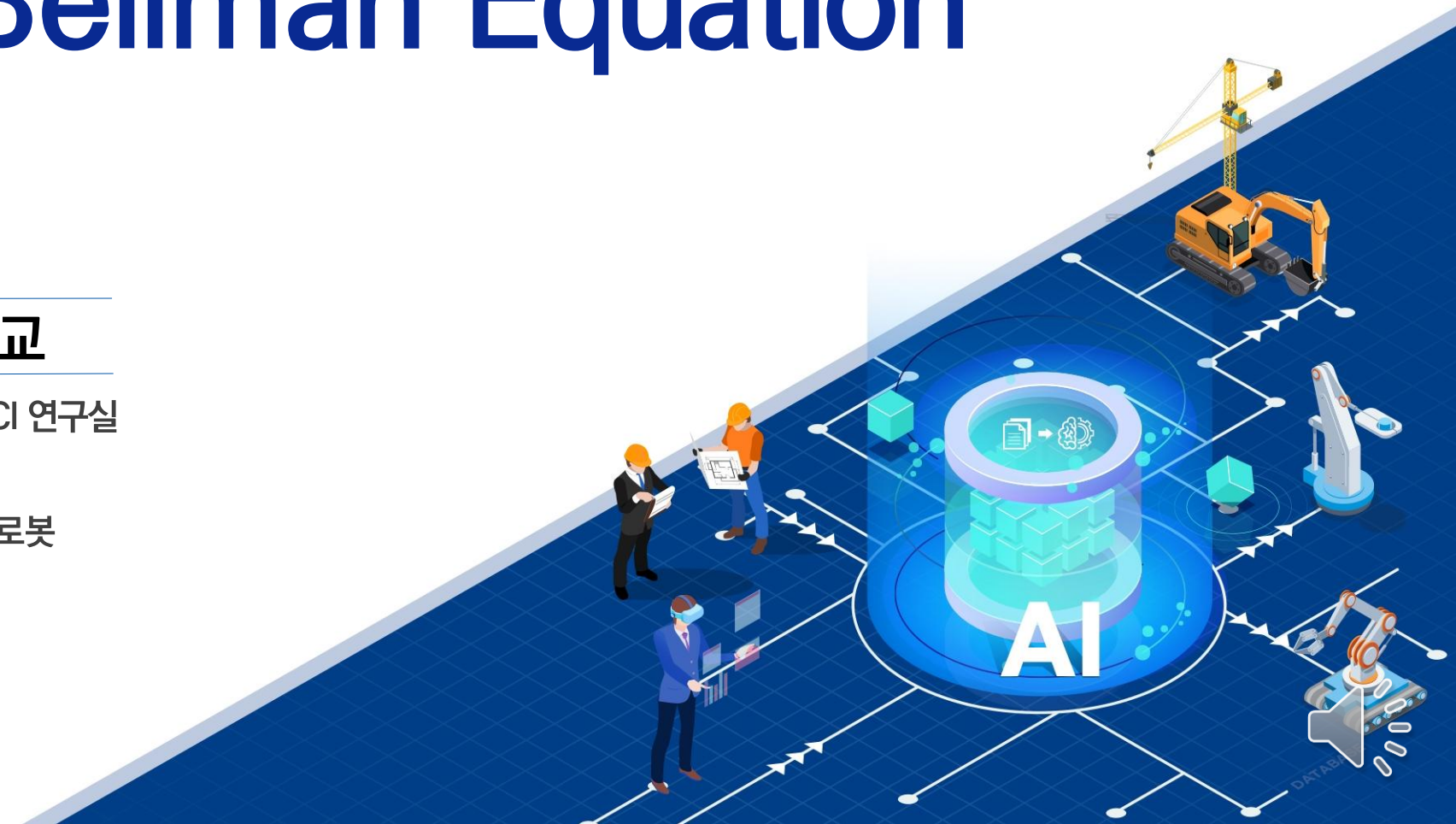
2. Bellman Equation

경희대학교

기계공학과 RCI 연구실
박보형

2025-2 이동로봇



목차

- Value Function
- Bellman Expectation Equation
- Bellman Optimality Equation

Value Function

Value Functions

- 하나의 Policy π가 주어졌을 때, 각 State S의 가치를 평가하는 지침
  - Value Function은 Policy π를 따를 때 각 State S의 가치를 G_t의 Expectation으로 평가
  - State-Value Function과 Action-Value Function이 있다.

State-Value Function

- State-Value Function $v_\pi(s)$ for the Policy π is expected return starting from State S when following Policy π

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s]$$

  - 장점 :

    Action-Value Function에 비해서 계산량이 적다.

  - 단점 :

    현재 State에 대한 Value Function 값만 알고 있기 때문에 다음에 어떤 Action 을 취하는 것이 가장 유리한지 알 수 없다.

Value Function

Action-Value Function

- Action-Value Function $q_\pi(s,a)$ for Policy π is the expected return starting from State S, taking Action A and following Policy π

$$q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$$

  - 차이점은 현재 State에서 특정 Action을 취한 후 부터 시작한다.
    - 이에 의해 조건부 $A_t = a$가 추가되었다.
  - 장점 :

    현재 State에서 취할 수 있는 각 Action들에 대해서 return의 Expectation을 계산하기 때문에 어떤 Action을 취하는 것이 가장 좋은 Policy인지 알 수 있다.
  - 단점 :

    State-Action Pair (s, a)에 대해서 계산해야 하기 때문에 계산량이 State-Value Function 보다 훨씬 많다.

Bellman Expectation Equation

Bellman Expectation Equation

- Agent가 Policy π를 따라 Stochastic Action을 취한다면 현재 State의 가치는 보상 + 미래의 기대값이다.

State-Value Function

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi[G_t \mid S_t = s] \\
&\dots \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s].
\end{aligned}
$$

- 원래 Return G_t를 계산해야 하기 때문에 하나의 Episode가 종료되어야 Value Function 사용 가능
  - 위와 같이 식을 변형해서 Return G_t 대신에 Immediate Reward R_{t+1}와 Discounted next State-Value Function $\gamma v_\pi(S_{t+1})$의 합으로 나타낼 수 있다.
  - 따라서 Episode가 끝나지 않더라도 바로바로 State-Value Function 값을 계산할 수 있다.

Bellman Expectation Equation

Action-Value Function

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] \\
&= \sum_{s',r} p(s',r \mid s,a) \left[r + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s',a') \right] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a].
\end{aligned}
$$

- State-Value Function과 같은 방법으로 전개, 변형할 수 있는데, Action만 조건부로 추가된 형태이다.

Bellman Optimality Equation

Bellman Optimality Equation

- Optimal Policy π^*를 가정 : π가 고정되지 않고 최적의 Action을 선택한다고 가정
- Agent가 현재 State에서 Optimal Action만 선택한다면, 현재 State-Value는 그 Action에서 얻을 수 있는 최대 기대값이다.

Bellman Optimality Equation

Optimal State-Value Function

$$
\begin{aligned}
v_*(s) &= \max_{a \in \mathcal{A}(s)} q_*(s,a) = \max_a \mathbb{E}_{\pi_*}[G_t \mid S_t = s, A_t = a] \\
&= \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a] \\
&= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a] \\
&= \max_a \sum_{s',r} p(s', r \mid s, a)\,[r + \gamma v_*(s')] \\
&\qquad \left(\Leftarrow \; v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_\pi(s')] \right) \\
&= \max_a \left[R_s^a + \gamma \sum_{s'} P_{ss'}^a v_*(s') \right].
\end{aligned}
$$

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi[G_t \mid S_t = s] \\
&\cdots \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s].
\end{aligned}
$$

- 위와 같이 식을 변형해서 앞선 State-Value Function과 마찬가지로 Return G_t 대신에 Immediate Reward R_{t+1}와 Discounted next State-Value Function $\gamma v_\pi(S_{t+1})$의 합으로 나타낼 수 있다.
- Bellman Expectation과 비교하면
  - Action : Value를 maximize 시키는 Action
  - Next State-Value Function : Optimal State-Value Function

Bellman Optimality Equation

Optimal Action-Value Function

$$q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a\right]$$

$$= \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma \max_{a'} q_*(s', a')\right]$$

$$= R_s^a + \gamma \sum_{s'} P_{ss'}^a \max_{a'} q_*(s', a').$$

$$q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$$

$$= \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s', a')\right]$$

$$= \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a].$$

- Optimal Action-Value Function도 동일한 과정으로 식 변형
- Optimal State-Value Function은
  - 상태 S 이후에 항상 Optimal Action을 취했을 때 받을 수 있는 최대 기대 Return
- Optimal Action-Value Function은
  - 상태 S에서 특정 Action a를 하고 나서, 그 이후로 Optimal Action을 취했을 때 받을 수 있는 최대 기대 Return
- Model-Based Method인 known MDP에서는 p와 r(s, a)를 모두 알고 있기 때문에 두 번째 수식을 그대로 사용해서 Bellman Optimality Equation을 계산 : Dynamic Programming

Next

Monte Carlo

- Bellman Equation을 통해서 Value Function을 재귀적으로 정의
- 하지만 Bellman Equation을 계산하기 위해서는 환경의 Transition Probability를 알아야 함
- Transition Probability 없이 Sample Data 만으로 Value Function을 추정하는 방법이 Monte Carlo

감사합니다