Value Evaluation and Policy Iteration

2021311360 안철균

Bellman Equation

- Bellman Expectation Equation

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

  - 선형 방정식 → Closed form solution 존재

- Bellman Optimality Equation

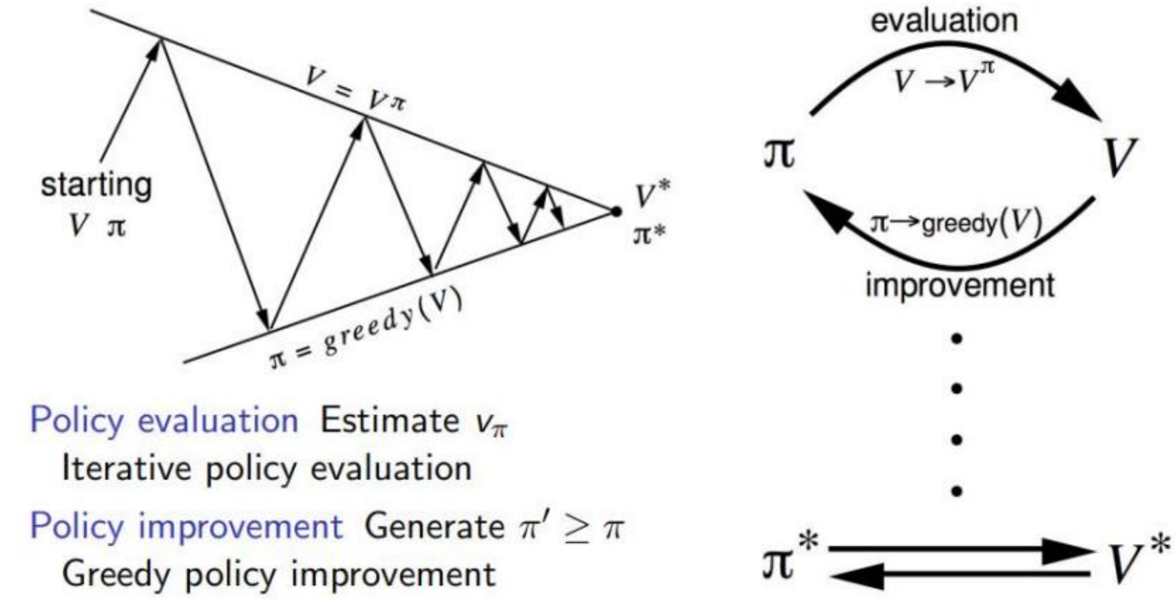
$$V^{*}(s) = \max_{a} \left\{ R(s,a) + \gamma \sum_{s'} P(s' | s, a) V^{*}(s') \right\}$$

  - 비선형 방정식 → Closed form solution 존재하지 않음

Solving the Bellman Optimality Equation (Many Iterative Solution Methods)

- **Dynamic programming :** MDP 모델이 정확히 주어졌을 경우에 사용함
  - Policy Iteration : Bellman Expectation Equation 을 통해 Policy evaluation 을 수행해주고, 구해진 value function을 통해 policy improvement 를 수행함
  - Value Iteration : Bellman Optimality Equation 을 통해 업데이트 됨
- **Reinforcement Learning :** MDP 모델이 정확히 주어지지 않았을 경우 혹은 MDP 모델이 정확히 주어졌으나 모델이 너무 클 경우에 사용함
  - SARSA
  - Q-Learning

Policy Iteration



- Policy Evaluation

$$v_{k+1}(s) = \sum_{a \in A} \pi(a|s) \left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_k(s') \right)$$
$$\rightarrow \boldsymbol{v}^{k+1} = \boldsymbol{R}^\pi + \gamma \boldsymbol{P}^\pi \boldsymbol{v}^k$$

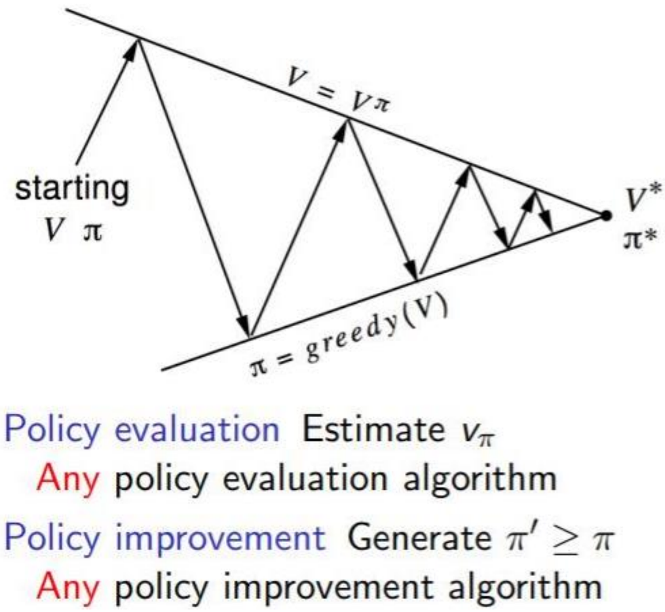
- Policy Improvement

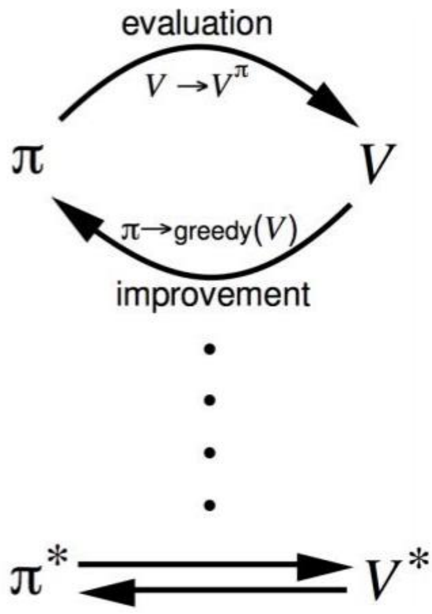
$$\pi'(s) = greedy(v_\pi)$$
$$= arg\max_{a \in A} q_\pi(s, a)$$

  - $v_{\pi'}(s) \geq v_\pi(s)$: 새롭게 policy improvement 가 됐을 경우 해당 policy 에 따른 가치 함수는 더 높은 값을 가짐

$$\begin{aligned} v_\pi &\leq q_\pi(s, \pi'(s)) \\ &= E_{\pi'}[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s] \\ &\leq E_{\pi'}[R_{t+1} + \gamma q_\pi(S_{t+1}, \pi'(S_{t+1})) | S_t = s] \\ &\leq E_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 q_\pi(S_{t+2}, \pi'(S_{t+2})) | S_t = s] \\ &\leq E_{\pi'}[R_{t+1} + \gamma R_{t+2} + \cdots | S_t = s] \\ &= v_{\pi'}(s) \end{aligned}$$

- Generalized Policy Iteration

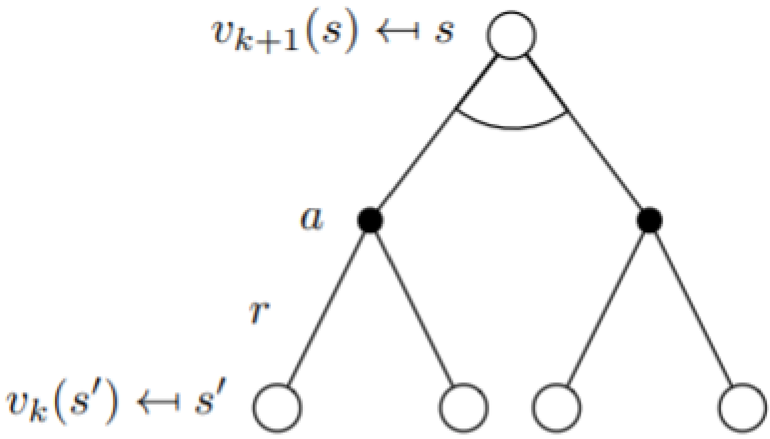




Value Iteration

$$v_{k+1}(s) = \max_{a \in A}\left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_k(s')\right)$$
$$\rightarrow \boldsymbol{v}_{k+1} = \max_{a \in A} \boldsymbol{R}^a + \gamma \boldsymbol{P}^a \boldsymbol{v}_k$$

- Problem : Find optimal policy π
- Solution : Iterative application of Bellman optimality backup



- Synchronous backups
- No explicit policy

Synchronous DP Algorithms

Problem	Bellman Equation	Algorithm
Prediction	Bellman Expectation Equation	Iterative Policy Evaluation
Control	Bellman Expectation Equation + Greedy Policy Improvement	Policy Iteration
Control	Bellman Optimality Equation	Value Iteration

Asynchronous DP Algorithms

- In-place dynamic programming : 이전과 이후 가치함수를 구분없이 업데이트함

$$v(s) \leftarrow \max_{a \in \mathcal{A}} \left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v(s') \right)$$

- Prioritized sweeping : Bellman error 를 업데이트함

$$\left| \max_{a \in \mathcal{A}} \left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v(s') \right) - v(s) \right|$$

- Real-time DP : 오직 Agent에게 관련된 state 에 대해서만 업데이트함

$$v(S_t) \leftarrow \max_{a \in \mathcal{A}} \left(\mathcal{R}_{S_t}^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{S_t s'}^a v(s') \right)$$

Convergence of VI, PI

- The Bellman expectation backup operator

$$T^{\pi}(v) = R^{\pi} + \gamma P^{\pi} v$$

  - This operator is a γ-contraction, i.e.

$$\begin{aligned} \|T^{\pi}(u) - T^{\pi}(v)\|_{\infty} &= \|(R^{\pi} + \gamma P^{\pi} u) - (R^{\pi} + \gamma P^{\pi} v)\|_{\infty} \\ &= \|\gamma P^{\pi}(u - v)\|_{\infty} \\ &\le \|\gamma P^{\pi}\|u - v\|_{\infty}\|_{\infty} \\ &\le \gamma\|u - v\|_{\infty} \end{aligned}$$

  - Contraction Mapping Theorem

> **Theorem (Contraction Mapping Theorem)**
>
> *For any metric space $\mathcal{V}$ that is complete (i.e. closed) under an operator $T(v)$, where T is a γ-contraction,*
>
> - *T converges to a unique fixed point*
> - *At a linear convergence rate of γ*

    - Convergence of PI

      T^{π} has a unique fixed point

      → v_{π} is a fixed point of T^{π}

      → Iterative policy evaluation converges on v_{π} (∵ Contraction Mapping Theorem)

      → Policy iteration converges on v_{*}

- The Bellman optimality backup operator

$$T^{*}(v) = \max_{a \in A} R^{a} + \gamma P^{a} v$$

  - This operator is a γ-contraction, i.e.

$$\|T^{*}(u) - T^{*}(v)\|_{\infty} \le \gamma\|u - v\|_{\infty}$$

    - Convergence of VI

      T^{*} has a unique fixed point

      → v_{*} is a fixed point of T^{*}

      → Value iteration converges on v_{*} (∵ Contraction Mapping Theorem)