M0000.026500

학습기반 공정 동적최적화

Lecture 9: Infinite Horizon MDP

JONG MIN LEE

School of Chemical & Biological Engineering

Problem Setup: Discounted Total Cost

① States: $s \in \mathcal{S}$

② Decisions: $a \in \mathcal{A}(s)$

③ Transitions: $p[j|s,a]$ or $p(s_{t+1} = j | s_t = i) = p_{ij}$

④ Costs: $c(s,a)$

⑤ Objective

$$\min_{a_0, a_1, a_2, \cdots} \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t c(s_t, a_t) \middle| s_0 \right] \qquad \alpha \in (0,1)$$

Note:

1. MDP formulated in discrete time.

2. $\mathcal{A}(s)$, $p[j|s,a], c(s,a)$ do not depend on time (stationary)

3. s is countable

4. $c(s, a)$ are bounded $|c(s,a)| \leq M$ for all s, a

Policies and Value Functions

History-dependent policy vs. memoryless-randomized policy

For any $\pi \in \Pi^{HR}$ and any history h_t

$$U^\pi(h_t) := \mathbb{E}^\pi\left[\sum_{\tau=t}^{\infty} \alpha^{\tau-t} c(s_\tau, a_\tau) \middle| h_t\right]$$

$$U^*(h_t) := \sup_{\pi \in \Pi^{HR}} U^\pi(h_t)$$

Policy π^* is optimal if $U^{\pi^*}(h_t) = U^*(h_t) \qquad \forall h_t$

$$V^\pi(s) := \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \alpha^t c(s_t, a_t) \middle| s_0 = s\right] = U^\pi(h_0) \quad \text{where } h_0 = s$$

$$\boxed{V^*(s) := \sum_{\pi \in \Pi^{HR}} V^\pi(s) = U^*(h_0)}$$

initial time에서는 같다.

We want to show that for any history $h_t = (s_0, a_0, \cdots, s_{t-1}, a_{t-1}, s_t)$

$$\boxed{U^*(h_t) = V^*(s_t)}$$

History → Memoryless

Theorem

Consider any $\pi \in \Pi^{HR}$. Fix $s \in \mathcal{S}$ then there exists $\pi' \in \Pi^{MR}$ such that

$$p^{\pi'}\left[s_t = y, A_t = a \middle| s_0 = s\right] = p^{\pi}\left[s_t = y, A_t = a \middle| s_0 = s\right]$$

for all y, a, and t.

Corollary

For any h_t

$$U^*(h_t) := \sup_{\pi \in \Pi^{HR}} U^\pi(h_t) = \sup_{\pi \in \Pi^{MR}} U^\pi(s_t, t)$$

Value Fcn Under a Decision Rule

Stationary Policy

Decision Rule (Policy는 좀 더 general한 개념임. Sequence of d = policy)

① Deterministic decision rule

$$d : \mathcal{S} \rightarrow \mathcal{A} \qquad (d(s) \in \mathcal{A})$$

Note: deterministic decision rule is the same as a stationary deterministic policy.

② Randomized decision rule

$$d : \mathcal{S} \rightarrow \text{ Probability distiribution on } \mathcal{A}$$

Note: randomized decision rule is the same as a stationary randomized policy.

$$d(s,a) = \text{ prob. of choosing } a \in \mathcal{A} \text{ in } s \in \mathcal{S} \qquad \sum_{a \in \mathcal{A}} d(s,a) = 1$$

- A policy $\pi \in \Pi^{MD}$ is a sequence (d_0, d_1, d_2, …) of deterministic decision rules.

- A policy $\pi \in \Pi^{MR}$ is a sequence (d_0, d_1, d_2, …) of randomized decision rules.

Value Fcn Under a Decision Rule

For any deterministic decision rule *d*, let

$$c_d(s) = c(s, d(s)) \quad \text{; single-stage cost}$$

$$p_d[y|s] = p[y|s, d(s)]$$

For any randomized decision rule *d*,

$$c_d(s) = \sum_{a \in \mathcal{A}} d(s,a) c(s,a)$$

$$d(s,a): \text{ prob. of taking } a$$

$$p_d[y|s] = \sum_{a \in \mathcal{A}} d(s,a) p[y|s,a]$$

Value Fcn Under a Decision Rule

Memoryless Policy → Infinite horizon value fcn

$$\pi \in \Pi^{MR} \qquad \pi = (d_0,\ d_1,\ \cdots)$$

t: state-transition probability

$$p^{\pi,t} = p_{d_0} \times p_{d_1} \times p_{d_2} \times \cdots \times p_{d_{t-1}} = \prod_{\tau=0}^{t-1} p_{d_\tau}$$

$p^{\pi,0} = I$ (처음 initial state에서 t = 0에 그 initial state에 있을 확률)

$$\begin{aligned} \therefore\ V^\pi &= \sum_{t=0}^{\infty} \alpha^t p^{\pi,t} c_{d_t} \\ &= c_{d_0} + \alpha p_{d_0} c_{d_1} + \alpha^2 p_{d_0} p_{d_1} c_{d_2} + \cdots \\ &= c_{d_0} + \alpha p_{d_0} (c_{d_1} + \alpha p_{d_1} c_{d_2} + \cdots) \end{aligned} \tag{1}$$

Let $\pi^1 = (d_1,\ d_2,\ \cdots)$

Then, $V^{\pi^1} = c_{d_1} + \alpha p_{d_1} c_{d_2} + \cdots$

(1) becomes

$$V^{\pi} = c_{d_0} + \alpha p_{d_0} V^{\pi_1}$$

Suppose $\pi \in \Pi^{SR}$, $\pi = (d,\ d\ , d,\ \cdots)$

$$V^{\pi} = c_d + \alpha p_d V^{\pi}$$

Let $\mathcal{V}$ be the set of functions (bounded) $\mathcal{V}:\ \mathcal{S} \rightarrow \mathbb{R}$

For any decision rule d (randomized or deterministic), let $T_d:\ \mathcal{V} \rightarrow \mathcal{V}$ be defined by $V^{\pi} = T_d(V^{\pi})$ (system of equations)

In other words, V^π is a fixed-point of T_d

Theorem

For any $\pi \in \Pi^{SR}$, $\quad \pi = (d,\ d,\ \cdots)$

V^π is the unique solution (fixed point) of $V = T_d(V)$

and $V^\pi = (I - \alpha p_d)^{-1} c_d$

Stationary Deterministic Policy & Its Optimality

Decision rule is same at every time period.

$$\pi \in \Pi^{SR}, \quad \pi \in (d, d, d, \cdots)$$

$$V^{\pi} = c_d + \alpha p_d V^{\pi} = T_d(V^{\pi})$$

$$V^{*} = \sup_{\pi \in \Pi^{HR}} V^{\pi}(s) = \sup_{\pi \in \Pi^{MR}} V^{\pi}(s)$$ (This is what we know so far)

Optimality Equation

$$V^{*} = \sup_{d \in \Pi^{SD}} [c_d + \alpha p_d V^{*}]$$

Dynamic Programming Operator (T)

$$T : \mathcal{V} \to \mathcal{V}$$

$$T(V) := \sup_{d \in \Pi^{SD}} \{c_d + \alpha p_d V\}$$

We want to show that T has a unique fixed point; that is, there is a unique value function $V' : \mathcal{S} \rightarrow \mathbb{R}$ such that $T(V') = V'$

One can also show that $V' = V'^*$

(유인물: Banach Fixed Point Theorem 참고)

+ several other proofs

Value Iteration and Policy Iteration

Value Iteration: See the handout

Policy Iteration

Choose a policy

Find the infinite horizon, discounted value of the policy

This value is then used to choose a new policy

Algorithm: Policy Iteration

Step 0. Choose any (deterministic) decision rule $d_0, \ \varepsilon > 0, \ \text{set } i = 0$

Step 1. Policy $\pi_i \in \Pi^{SD}$ given by $\pi_i = (d_i, \ d_i, \ d_i, \ \cdots)$

Policy Evaluation

Compute V^{π_i} by solving $T_{d_i}(V^{\pi_i}) = V^{\pi_i}$

Note: T_{d_i} is contraction mapping, and thus V^{π_i} is unique.

# of eqns
= # of states

$$V^{\pi_i}(s) = c(s, d_i(s)) + \alpha \sum_{y \in \mathcal{S}} p[y|s,a] V^{\pi_i}(y) \qquad \forall s \in \mathcal{S}$$

or $$V^{\pi_i}(s) = c_{d_i} + \alpha p_{d_i} V^{\pi_i} \qquad (*)$$

Step 2. **Policy Improvement**

Choose decision rule (deterministic) d_{i+1} such that

$$d_{i+1}(s) \in \arg\max_{a \in \mathcal{A}(s)} \left\{ c(s,a) + \alpha \sum_{y \in \mathcal{S}} p[y|s,a] V^{\pi_i}(y) \right\}$$

Step 3. If $d_{i+1}(s) = d_i(s), \quad \forall s \in \mathcal{S}$ or if $\|V^{\pi_i} - V^{\pi_{i-1}}\|_\infty < \frac{1-\alpha}{2\alpha}\varepsilon$

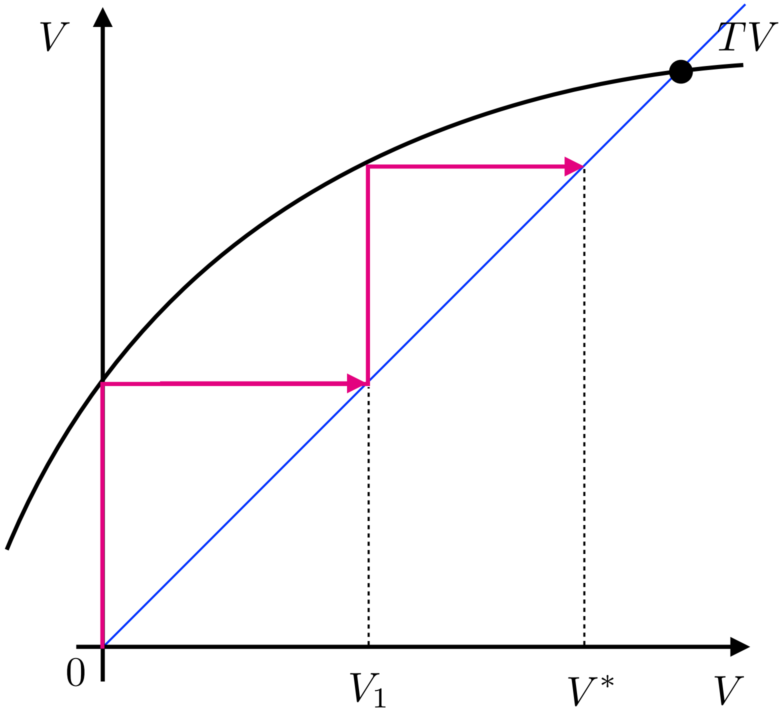
, then stop with optimal or ε-optimal policy π_{i+1}

otherwise, set $i \leftarrow i+1$ and go to step 1.

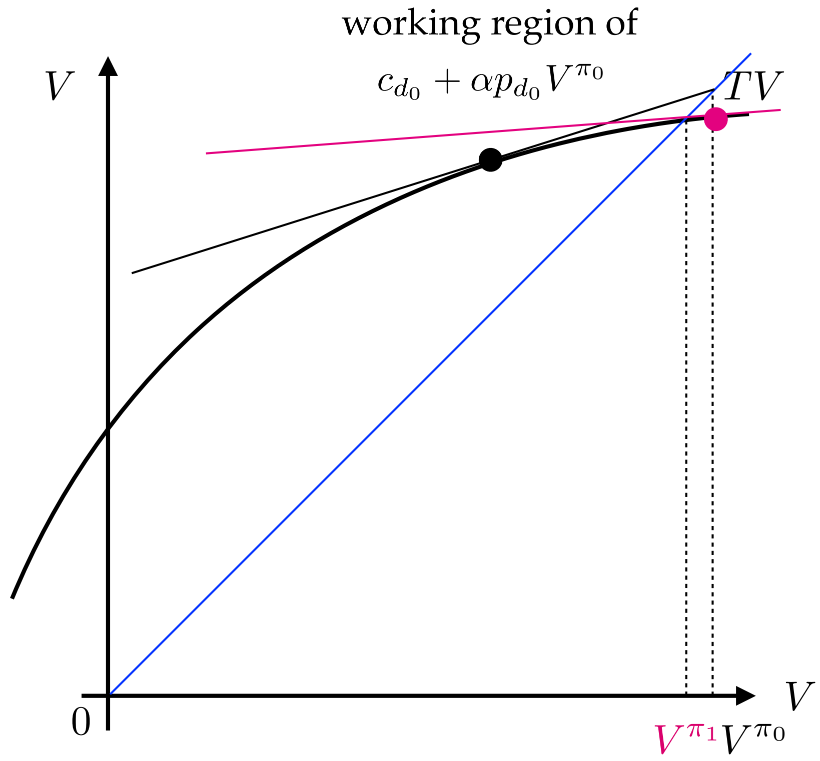
Comments on PI

- Faster convergence in terms of # of iterations
- Solving (*) is quite hard if the # of states is large.
- Matrix inversion can be computationally expensive.
- Value iteration updates the value at each iteration and then determines a new policy given the new estimate of the value function. At any iteration, the value fan is not the true, steady-state value of the policy. PI converges faster because it is doing a lot more work in each iteration.

Illustration of VI and PI



Value iteration



Policy iteration