Advanced Topics in OR

Lecture Notes 5

Markov Decision Processes

Zhao Xiaobo
Department of IE
Tsinghua University
Beijing 100084, China
Tel. 010-62784898
Email. xbzhao@tsinghua.edu.cn

Introduction

A process:

Observed at time points: $t = 0, 1, 2, \ldots\ldots$

States: countable, 0, 1, 2, ……

Action: be chosen after observing state, the set A

If the process is in state i at time t and action a is chosen, then

(i): incur a cost $C(i, a)$

(ii): state transition probability $P_{ij}(a)$

$$P\left\{X_{t+1} = j \,\middle|\, X_0, a_0, X_1, a_1, \cdots, X_t = i, a_t = a\right\} = P_{ij}(a)$$

Suppose: bounded cost, $|C(i, a)| < M$ for all t and i

Introduction

Policy: a rule for choosing actions

may be randomized: P_a, $a \in A$

Stationary policy: the action only depends on the state

a function f mapping the state space into action space

Under policy f, the sequence of states $\{X_t, t = 0, 1, \ldots\}$ forms a Markov chain, $P_{ij} = P_{ij}[f(i)]$

Criterion:

(i): total expected discounted cost

(ii): average cost

Expected Discounted Cost

Functional equation

For a policy π, define

$$V_\pi(i) = E_\pi\left[\sum_{t=0}^{\infty} \alpha^t C(X_t, a_t) \middle| X_0 = i\right], \qquad t \geq 0$$

where $\alpha \in (0, 1)$

Let $$V_\alpha(i) = \inf_\pi V_\pi(i), \qquad i \geq 0$$

A policy π^* is α – optimal, if

$$V_{\pi^*}(i) = V_\alpha(i), \qquad \text{for all } i \geq 0$$

An α – optimal policy minimizes the cost for every initial state

Expected Discounted Cost

Functional equation

Theorem 6.1

$$V_\alpha(i) = \min_a \left\{ C(i,a) + \alpha \sum_{j=0}^{\infty} P_{ij}(a) V_\alpha(j) \right\} \quad , \qquad t \geq 0$$

Proof. Let π be an arbitrary policy; action a at time 0 with probability P_a. Then

$$V_\pi(i) = \sum_{a \in A} P_a \left[C(i,a) + \sum_{j=0}^{\infty} P_{ij}(a) W_\pi(j) \right]$$

where $W_\pi(j)$, cost from time 1 onward

Since $\quad W_\pi(j) \geq \alpha V_\alpha(j)$

Expected Discounted Cost

Functional equation

we have

$$V_\pi(i) \ge \sum_{a\in A} P_a \left[C(i,a) + \alpha \sum_{j=0}^{\infty} P_{ij}(a) V_\alpha(j) \right]$$

$$\ge \sum_{a\in A} P_a \min_{a'\in A} \left[C\left(i,a'\right) + \alpha \sum_{j=0}^{\infty} P_{ij}\left(a'\right) V_\alpha(j) \right]$$

$$= \min_{a\in A} \left[C(i,a) + \alpha \sum_{j=0}^{\infty} P_{ij}(a) V_\alpha(j) \right]$$

The arbitrary of π implies that

$$V_\alpha(i) \ge \min_{a\in A} \left[C(i,a) + \alpha \sum_{j=0}^{\infty} P_{ij}(a) V_\alpha(j) \right]$$

Expected Discounted Cost

Functional equation

To go the other way, let a_0 be such that

$$C(i,a_0)+\alpha\sum_{j=0}^{\infty}P_{ij}(a_0)V_\alpha(j)=\min_{a\in A}\left[C(i,a)+\alpha\sum_{j=0}^{\infty}P_{ij}(a)V_\alpha(j)\right]$$

Policy π: chooses a_0 at time 0; next state j, follows policy π_j such that $V_{\pi_j}(j)\le V_\alpha(j)+\varepsilon$

Hence $$V_\pi(i)=C(i,a_0)+\alpha\sum_{j=0}^{\infty}P_{ij}(a_0)V_{\pi_j}(j)$$

$$\le C(i,a_0)+\alpha\sum_{j=0}^{\infty}P_{ij}(a_0)V_\alpha(j)+\alpha\varepsilon$$

Because $V_\alpha(i)\le V_\pi(i)$

Expected Discounted Cost

Functional equation

we have $$V_\alpha(i) \le C(i, a_0) + \alpha \sum_{j=0}^{\infty} P_{ij}(a_0) V_\alpha(j) + \alpha\varepsilon$$

Hence $$V_\alpha(i) \le \min_a \left\{ C(i,a) + \alpha \sum_{j=0}^{\infty} P_{ij}(a) V_\alpha(j) \right\} + \alpha\varepsilon$$

$B(I)$: the set of all bounded functions on the state space.

Define the mapping

$$\left(T_f u\right)(i) = C\left[i, f(i)\right] + \alpha \sum_{j=0}^{\infty} P_{ij}\left[f(i)\right] u(j)$$

Expected Discounted Cost

Notation:

Let $T_f^1 = T_f$, and for $n > 1$, let $T_f^n = T_f\left(T_f^{n-1}\right)$

Definitions:

Any two functions u and v, $u \leq v$ if $u(i) \leq v(i)$ for all i

Similar for $u = v$

For u_n and u, $u_n \longrightarrow u$ if $u_n(i) \longrightarrow u(i)$ uniformly in i, for all i

Lemma 6.2: For $u, v \in B(I)$, and f a stationary policy

(i) $u \leq v \longrightarrow T_f u \leq T_f v$

(ii) $T_f V_f = V_f$

(iii) $T_f{}^n u \longrightarrow V_f$ for all $u \in B(I)$

Expected Discounted Cost

Proof:

Part (i): follows directly from the definition of T_f

Part (ii): is just the statement that

$$V_f(i) = C[i, f(i)] + \alpha \sum_{j=0}^{\infty} P_{ij}[f(i)] V_f(j)$$

Part (iii): $$\left(T_f^2 u\right)(i) = C[i, f(i)] + \alpha \sum_{j=0}^{\infty} P_{ij}[f(i)] \left(T_f u\right)(j)$$

$$= C[i, f(i)] + \alpha \sum_{j=0}^{\infty} P_{ij}[f(i)] \left[C[j, f(j)] + \alpha \sum_{k=0}^{\infty} P_{jk}[f(j)] u(k) \right]$$

$$= C[i, f(i)] + \alpha \sum_{j=0}^{\infty} P_{ij}[f(i)] C[j, f(j)] + \alpha^2 \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} P_{ij}[f(i)] P_{jk}[f(j)] u(k)$$

Expected Discounted Cost

Proof:

$T_f^2 u$: the cost using f for two periods with final cost $\alpha^2 u$

⇩

$T_f^n u$: the cost using f for n periods with final cost $\alpha^n u$

Since $\alpha < 1$ and u is bounded, the result follows

Lemma 6.3: f_α: in state i, select action such that

$$C\left[i, f_\alpha(i)\right] + \alpha \sum_{j=0}^{\infty} P_{ij}\left[f_\alpha(i)\right] V_\alpha(j) = \min_a \left\{ C(i,a) + \alpha \sum_{j=0}^{\infty} P_{ij}(a) V_\alpha(j) \right\}$$

Then $$V_{f_\alpha}(j) = V_\alpha(j)$$

Expected Discounted Cost

Proof: We can obtain

$$\left(T_{f_\alpha} V_\alpha\right)(i) = C\left[i, f_\alpha(i)\right] + \alpha \sum_{j=0}^{\infty} P_{ij}\left[f_\alpha(i)\right] V_\alpha(j)$$

$$= \min_a \left\{ C(i,a) + \alpha \sum_{j=0}^{\infty} P_{ij}(a) V_\alpha(j) \right\} = V_\alpha(i)$$

Hence $T_{f_\alpha} V_\alpha = V_\alpha$

$$\Rightarrow \quad T_{f_\alpha}^2 V_\alpha = T_{f_\alpha}\left(T_{f_\alpha} V_\alpha\right) = T_{f_\alpha} V_\alpha = V_\alpha$$

$$\Rightarrow \quad T_{f_\alpha}^n V_\alpha = V_\alpha$$ for all n

$$\Rightarrow \quad V_{f_\alpha} = V_\alpha$$

Expected Discounted Cost

The following situation

Suppose that $f \rightarrow V_f$

Let f^* be such that

$$C\left[i, f^*(i)\right] + \alpha \sum_{j=0}^{\infty} P_{ij}\left[f^*(i)\right] V_f(j) = \min_a \left\{ C(i,a) + \alpha \sum_{j=0}^{\infty} P_{ij}(a) V_f(j) \right\}$$

How good is f^* compared with f ?

Corollary 6.3: $V_{f^*}(i) \le V_f(i)$ for all i

Proof: $$\left(T_{f^*} V_f\right)(i) = C\left[i, f^*(i)\right] + \alpha \sum_{j=0}^{\infty} P_{ij}\left[f^*(i)\right] V_f(j)$$

Expected Discounted Cost

$$\Rightarrow \quad \le C\left[i, f(i)\right] + \alpha \sum_{j=0}^{\infty} P_{ij}\left[f(i)\right] V_f(j)$$

$$= V_f(i)$$

Hence $\quad T_{f^*} V_f \le V_f$

$$\Rightarrow \quad T_{f^*}^2 V_f \le T_{f^*} V_f \le V_f$$

$$\Rightarrow \quad T_{f^*}^n V_f \le V_f$$

$\Rightarrow$ Policy improvement algorithm

Expected Discounted Cost

Contraction Mappings

Definition: A mapping T: $B(I) \rightarrow B(I)$ is contraction if

$$\|Tu - Tv\| \leq \beta \|u - v\|$$

where $\beta < 1$, and $\|u\| = \sup_{i \geq 0} |u(i)|$

Theorem (Contraction Mapping Fixed Point Theorem)

If T: $B(I) \rightarrow B(I)$ is a contraction mapping, then there exists a unique function $g \in B(I)$ such that

$$Tg = g$$

Furthermore, for all $u \in B(I)$ such that $T^n u \rightarrow g$ as $n \rightarrow \infty$

Expected Discounted Cost

Defining the mapping T_α: $B(I) \to B(I)$ by

$$(T_\alpha u)(i) = \min_a \left\{ C(i,a) + \alpha \sum_{j=0}^{\infty} P_{ij}(a) u(j) \right\}$$

It follows that $\quad T_\alpha V_\alpha = V_\alpha$

$\Rightarrow$ Successive approximations

Theorem 6.5: The mapping T_α is a contraction mapping

Proof. $(T_\alpha u)(i) - (T_\alpha v)(i) =$

$$\min_a \left\{ C(i,a) + \alpha \sum_{j=0}^{\infty} P_{ij}(a) u(j) \right\} - \min_a \left\{ C(i,a) + \alpha \sum_{j=0}^{\infty} P_{ij}(a) v(j) \right\}$$

Expected Discounted Cost

$$= \min_a \left\{ C(i,a) + \alpha \sum_{j=0}^{\infty} P_{ij}(a) u(j) \right\} - C(i,\bar{a}) - \alpha \sum_{j=0}^{\infty} P_{ij}(\bar{a}) v(j)$$

where

$$C(i,\bar{a}) + \alpha \sum_{j=0}^{\infty} P_{ij}(\bar{a}) v(j) = \min_a \left\{ C(i,a) + \alpha \sum_{j=0}^{\infty} P_{ij}(a) v(j) \right\}$$

Hence

$$(T_\alpha u)(i) - (T_\alpha v)(i) \le \alpha \sum_{j=0}^{\infty} P_{ij}(\bar{a}) u(j) - \alpha \sum_{j=0}^{\infty} P_{ij}(\bar{a}) v(j)$$

$$= \alpha \sum_{j=0}^{\infty} P_{ij}(\bar{a}) \left[u(j) - v(j) \right] \le \alpha \sum_{j=0}^{\infty} P_{ij}(\bar{a}) \sup_j \left[u(j) - v(j) \right]$$

Expected Discounted Cost

$$\leq \alpha \|u - v\|$$

Thus, we obtain

$$\sup_{i \geq 0} \{(T_\alpha u)(i) - (T_\alpha v)(i)\} \leq \alpha \|u - v\|$$

By reversing the roles of *u* and *v*, we obtain

$$\sup_{i \geq 0} \{(T_\alpha v)(i) - (T_\alpha u)(i)\} \leq \alpha \|u - v\|$$

$$\Rightarrow \quad \sup_{i \geq 0} |(T_\alpha v)(i) - (T_\alpha u)(i)| \leq \alpha \|u - v\|$$

$$\Rightarrow \quad \|T_\alpha u - T_\alpha v\| \leq \alpha \|u - v\|$$

Expected Discounted Cost

Corollary 6.6: V_α is the unique solution to

$$V_\alpha(i) = \min_a \left\{ C(i,a) + \alpha \sum_{j=0}^{\infty} P_{ij}(a) V_\alpha(j) \right\}$$

Furthermore, for any $u \in B(I)$

$$T_\alpha^n u \to V_\alpha \quad \text{as} \quad n \to \infty$$

Remark 1: Let $u(i) = 0$ for all i. Let

$$V_\alpha(i,n) = \left(T_\alpha^n 0\right)(i)$$

$V_\alpha(i, n)$: cost of an n-period problem

To obtain some property, first prove $V_\alpha(i, n)$, then prove $V_\alpha(i)$

Expected Discounted Cost

Remark 2:

Corollary 6.6 shows that for the policy improvement technique, either the new policy is strictly better than the old one, or else they are both optimal.

This follows since if $V_{f^*} = V_f$, then we have V_f satisfies the optimality equation, and hence by uniqueness $V_f = V_\alpha$

Remark 3:

It can be shown that T_f is a contraction mapping. Hence, V_f is the unique solution to

$$V_f(i) = C\left[i, f(i)\right] + \alpha \sum_{j=0}^{\infty} P_{ij}\left[f(i)\right] V_f(j)$$

Some Examples

A machine replacement model

State: 0, 1, 2, ……

At the beginning of each day: observe the state

Action: 1 – replace, 2 – nonreplace (if replaced, the new machine with state 0)

Cost:

replacing a machine, R

maintenance in state i, $C(i)$

Transition probability: P_{ij}

Some Examples

A machine replacement model

$$C(i,1) = R + C(0) \qquad C(i,2) = C(i)$$

$$P_{ij}(1) = P_{0j} \qquad P_{ij}(2) = P_{ij}$$

Assumptions:

(I) $\{C(i), i \geq 0\}$ is bounded, increasing sequence

(II) $\sum^{\infty}_{j=k} P_{ij}$ is an increasing function of i, for each $k \geq 0$

Lemma 6.7 Assumptions (ii) implies that for any increasing function $h(i)$, the function:

$$\sum_{j=0}^{\infty} P_{ij} h(j)$$ is also increasing in i.

Some Examples

A machine replacement model

Lemma 6.8 Under (i) and (ii), $V_\alpha(i)$ is increasing in i.

Proof. Let

$$V_\alpha(i,1) = \min\{R + C(0); C(i)\}\ ,\quad i \ge 1$$

and for $n > 1$

$$V_\alpha(i,n) = \min\left\{R + C(0) + \alpha \sum_{j=0}^{\infty} P_{0j} V_\alpha(j,n-1); C(i) + \alpha \sum_{j=0}^{\infty} P_{ij} V_\alpha(j,n-1)\right\}$$

Assumption (i) $\rightarrow$ $V_\alpha(i, 1)$ is increasing in i

Assume $V_\alpha(i, n-1)$ is increasing in i, then Lemma 6.7 $\rightarrow$ $V_\alpha(i, n)$ is also increasing in i

Some Examples

A machine replacement model

By induction, $V_\alpha(i, n)$ is increasing in i for all n. Hence $V_\alpha(i) = \lim_n V_\alpha(i, n)$ is increasing in i

Theorem 6.9 Under (i) and (ii), there exists an integer i^*, such that an α – optimal policy replaces for $i > i^*$, and does not replace for $i \le i^*$

Proof. By Theorem 6.1, we have

$$V_\alpha(i) = \min\left\{ R + C(0) + \alpha \sum_{j=0}^{\infty} P_{0j} V_\alpha(j); C(i) + \alpha \sum_{j=0}^{\infty} P_{ij} V_\alpha(j) \right\}$$

Some Examples

A machine replacement model

Let

$$i^* = \max\left\{ i : C(i) + \alpha \sum_{j=0}^{\infty} P_{ij} V_\alpha(j) \le R + C(0) + \alpha \sum_{j=0}^{\infty} P_{0j} V_\alpha(j) \right\}$$

It follows that $C(i) + \alpha \sum_{j=0}^{\infty} P_{ij} V_\alpha(j)$ is increasing in i

$$\Rightarrow \quad V_\alpha(i) = \begin{cases} C(i) + \alpha \sum_{j=0}^{\infty} P_{ij} V_\alpha(j) & \text{for } i \le i^* \\ R + C(0) + \alpha \sum_{j=0}^{\infty} P_{0j} V_\alpha(j) & \text{for } i > i^* \end{cases}$$