Advanced Topics in OR

Lecture Notes 6

Markov Decision Processes

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

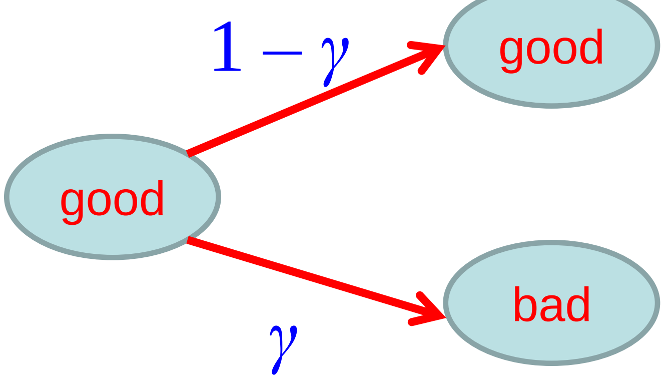
Some Examples

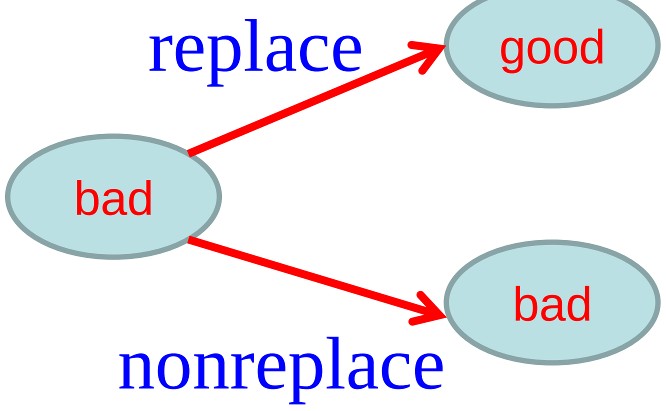
A quality control model

A machine has two states: good, bad

Produce an item each day:

good state → good item, bad state → bad item





After an item is produced, an option of inspecting or not

Some Examples

A quality control model

Produce a bad item with cost C

Inspecting costs I

If the item is inspected and found bad, the machine is replaced with cost R

The process is in state p, the posterior probability the machine in bad: the state space [0, 1]

If state is p and inspect action, then the expected cost

$$I + p(C + R)$$

The next state is γ

Some Examples

A quality control model

If state is p and not inspecting, then the expected cost pC

The next state is $p + (1 - p)\gamma$ The optimal function

$$V_\alpha(p) = \min\left\{I + p(C+R) + \alpha V_\alpha(\gamma);\, pC + \alpha V_\alpha\left(p + (1-p)\gamma\right)\right\}$$

$p \in [0, 1]$

Selling an asset

An individual sales his house

An offer at the beginning of each day

value i with probability P_i, $i = 0, 1, \ldots, N$

Some Examples

Selling an asset

The individual must immediately decide whether or not to accept the offer

accept: receive value i

reject: maintenance cost C

Discounting rate: α

State: the offer

The optimality equation

$$V_\alpha(i) = \min\left\{ -i; C + \alpha \sum_{j=0}^{N} P_j V_\alpha(j) \right\}$$

Some Examples

Selling an asset

Let
$$i^* = \min\left\{ i : -i < C + \alpha \sum_{j=0}^{N} P_j V_\alpha(j) \right\}$$

The α-optimal policy: accept any offer greater than or equal to i^*, and reject all offers less than i^*

Determine the optimal policy (or equivalently, i^*)

f_i: the policy which accepts any offer greater than or equal to i.

T: the number of rejected offers

Some Examples

Selling an asset

$$C+\alpha C+\cdots+\alpha^{T-1}C-\alpha^{T}\frac{\sum_{j=i}^{N} jP_j}{\sum_{j=i}^{N} P_j}=\frac{C\left(1-\alpha^{T}\right)}{1-\alpha}-\alpha^{T}\frac{\sum_{j=i}^{N} jP_j}{\sum_{j=i}^{N} P_j}$$

T is geometric with mean $\sum_{j=0}^{i-1} P_j \Big/ \sum_{j=i}^{N} P_j$

The expected discounted cost under f_i

$$\sum_{j=0}^{N} P_j V_{f_i}\left(j\right)=\frac{C\sum_{j=0}^{i-1} P_j-\sum_{j=i}^{N} jP_j}{1-\alpha\sum_{j=0}^{i-1} P_j}$$

i^*: chosen to minimize the right side

Positive Costs, No Discounting

Suppose all costs are nonnegative, $C(i, a) \geq 0$ for all i, a

No discount factor

Not required that the costs be bounded

For any policy π, let

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

Let $\quad V(i) = \inf_\pi V_\pi(i)$

It is possible that $V(i)$ might be infinite

The nature of the problem is such that $V(i) < \infty$ for at least some values of i

Positive Costs, No Discounting

A policy π^* is said to be optimal if

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

Theorem 6.10 $$V(i) = \min_a \left\{ C(i,a) + \sum_{j=0}^{\infty} P_{ij}(a) V(j) \right\}$$

$N(I)$: the set of all nonnegative (possibly infinite-valued) functions

For any stationary policy f, define the mapping

$$T_f : N(I) \rightarrow N(I)$$

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

Positive Costs, No Discounting

Lemma 6.11

For $u, v \in N(I)$, and f a stationary policy

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

(ii) $T_f V_f = V_f$

(iii) $(T_f^n 0)(i) \rightarrow V_f(i)$ as $n \rightarrow \infty$ for each i, where 0 represents the function which is identically zero

Note that (iii) is only true for the zero function and not for any $u \in B(I)$

For discount function α, the final cost is $\alpha^n u$, which uniformly goes to zero if $u \in B(I)$

Without discounting, the only way is to let it be zero

Positive Costs, No Discounting

Theorem 6.12

Let f_1 be the stationary policy which, when the process is in state i, selects the action minimizing

$$C(i,a)+\sum_{j=0}^{\infty}P_{ij}(a)V(j)$$

Then $V_{f_1}(i) = V(i)$, for all i, and hence f_1 is optimal.

Proof. We have $$\left(T_{f_1}V\right)(i)=C\left[i,f_1(i)\right]+\sum_{j=0}^{\infty}P_{ij}\left[f_1(i)\right]V(j)$$

$$=\min_{a}\left\{C(i,a)+\sum_{j=0}^{\infty}P_{ij}(a)V(j)\right\}\quad=V(i)$$

Positive Costs, No Discounting

Hence $$T_{f_1}V = V$$

$C(i, a) \geq 0 \rightarrow V \geq 0$. By the monotonicity, we obtain

$$T_{f_1}0 \leq T_{f_1}V = V$$

$$\Longrightarrow \quad T_{f_1}^n 0 \leq V$$

Letting $n \rightarrow \infty$, we arrive at $V_{f_1} \leq V$

Since $V_{f_1} \geq V$ by the definition, yields the desired result

Thus, an optimal policy exists and is determined by

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

Applications

Optimal stopping problems

States: 0, 1, 2, …

Actions:

1 → stop, a terminal reward $R(i)$

2 → continue, pay a cost $C(i)$, transition probability

MDP

$$C(i,1) = -R(i) \qquad C(i,2) = C(i) \qquad C(\infty,\cdot) = 0$$

$$P_{i\infty}(1) = 1 \qquad P_{ij}(2) = P_{ij} \qquad P_{\infty\infty}(\cdot) = 1$$

Suppose $\inf_{i\geq 0} C(i) > 0 \qquad \sup_{i\geq 0} R(i) < \infty$

Applications

Optimal stopping problems

It is not the case that all costs are nonnegative

Let $R = \sup_{i \geq 0} R(i)$

A related process:

stop and pay a terminal cost $R - R(i)$

pay a cost $C(i)$ and go to the next state with P_{ij}

For any policy π, we have $V'_{\pi}(i) = V_{\pi}(i) + R$

Any policy π does not stop in finite expected time

$\Longrightarrow$ $V'_{\pi}(i) = V_{\pi}(i) = \infty$

So only consider policies stop in finite expected time

Applications

Optimal stopping problems

The related process, nonnegative costs

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

⇒ The original process

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

Let $\quad V_0(i) = -R(i)$

and for $n > 0$ $\quad V_n(i) = \min\left\{-R(i); C(i) + \sum_{j=0}^{\infty} P_{ij} V_{n-1}(j)\right\}$

Applications

Optimal stopping problems

It follows that

$$V_n(i) \geq V_{n+1}(i) \geq V(i) \quad \Longrightarrow \quad \lim_{n\to\infty} V_n(i) \geq V(i)$$

The process is stable if $\lim_{n\to\infty} V_n(i) = V(i)$

Let $R = \sup_i R(i) \qquad C = \inf_i C(i)$

Theorem 6.13

$$V_n(i) - V(i) \leq \frac{(R-C)\left[R - R(i)\right]}{(n+1)C}$$

Proof. f: optimal policy, T stop time.

f_n: same as f but stop at time n (if not stopped so far).

Applications

Optimal stopping problems

$$V(i)=V_f(i)=E_f\left[X\,|T\le n\right]P\{T\le n\}+E_f\left[X\,|T>n\right]P\{T>n\}$$

$$V_n(i)\le V_{f_n}(i)=E_f\left[X\,|T\le n\right]P\{T\le n\}+E_{f_n}\left[X\,|T>n\right]P\{T>n\}$$

$$V_n(i)-V(i)\le\left[E_{f_n}\left(X\,|T>n\right)-E_f\left(X\,|T>n\right)\right]P\{T>n\}$$

$$\le(R-C)P\{T>n\}$$

From the first line above, we have

$$-R(i)\ge V(i)\ge -RP\{T\le n\}+\left(-R+(n+1)C\right)P\{T>n\}$$

$$=-R+(n+1)CP\{T>n\}$$

Applications

Optimal stopping problems

or $$P\{T > n\} \le \frac{R - R(i)}{(n+1)C}$$

Let

$$B = \left\{ i : -R(i) \le C(i) - \sum_{j=0}^{\infty} P_{ij} R(j) \right\} = \left\{ i : R(i) \ge \sum_{j=0}^{\infty} P_{ij} R(j) - C(i) \right\}$$

B: the set of states for which stopping is at least as good as continuing for exactly one more period and then stopping

Theorem 6.14

If the process is stable, and if $P_{ij} = 0$ for $i \in B$: $j \notin B$, then the optimal policy stops at i if and only if $i \in B$.

Applications

Optimal stopping problems

Proof. For $n = 0$, it follows $V_n(i) = - R(i)$.

Suppose it for $n - 1$. Then, for $i \in B$,

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

$$= \min\left\{ -R(i); C(i) + \sum_{j \in B} P_{ij} V_{n-1}(j) \right\}$$

$$= \min\left\{ -R(i); C(i) - \sum_{j \in B} P_{ij} R(j) \right\}$$

$$= -R(i)$$

Applications

Optimal stopping problems

Proof. Hence, $V_n(i) = -R(i)$ for all $i \in B$, all n.

By letting $n \to \infty$ and using the stability hypothesis, we obtain

$$V(i) = -R(i) \qquad \text{for } i \in B$$

For $i \notin B$, the policy which continues for exactly one stage and then stops has $C(i) - \sum_{j=0}^{\infty} P_{ij} R(j)$

which is strictly less than $-R(i)$ (since $i \notin B$)

Hence $$V(i) \begin{cases} = -R(i) & \text{for } i \in B \\ < -R(i) & \text{for } i \notin B \end{cases}$$

One-stage lookahead policy

Applications

Optimal stopping problems

Example 4: A house selling example

P_j: the successive offers, $j = 0, 1, \ldots, N$

Any offer not immediately accepted is not lost but may be accepted at any later date.

C: maintenance cost each day

Hence
$$P_{ij} = \begin{cases} 0 & j < i \\ \sum_{k=0}^{i} P_k & j = i \\ P_j & j > i \end{cases}$$

Applications

Optimal stopping problems

Example 4: A house selling example

$$\Rightarrow \quad B=\left\{i:-i\le C-i\sum_{k=0}^{i}P_k-\sum_{j=i+1}^{N}jP_j\right\}$$

$$=\left\{i:C\ge\sum_{j=i+1}^{N}jP_j-i\sum_{k=i+1}^{N}P_k\right\}=\left\{i:C\ge\sum_{j=i+1}^{N}(j-i)P_j\right\}$$

Since the right side is decreasing in *i*, it follows that

$$B=\left\{i^*,i^*+1,\cdots,N\right\}\quad \text{where}\quad i^*=\min\left\{i:C\ge\sum_{j=i+1}^{N}(j-i)P_j\right\}$$

New problem: once an offer is rejected, it is no longer available. The above policy is also optimal

Applications

Sequential analysis

$Y_1, Y_2, \ldots$: sequence of iid random variables

Probability density function of Y_i's is either f_0 or f_1

At time t, after observing $Y_1, Y_2, \ldots, Y_t$

⇨ stop observing, choose either f_0 or f_1

incur cost 0 if choice is correct

incur cost L if choice is incorrect

⇨ or pay a cost C and observe Y_{t+1}

Initial probability p_0: the true density is f_0

State at time t: p, the posterior probability, the true density is f_0

Applications

Sequential analysis

MDP: 3 action, nonnegative cost, uncountable state space [0, 1]

If state p, we stop and choose f_0

⇨ Expected cost $(1 - p)L$

If state p, we stop and choose f_1

⇨ Expected cost pL

If state p, we take another observation

⇨ value x with probability (density) $pf_0(x) + (1 - p)f_1(x)$

⇨ state $X_{t+1} = \dfrac{pf_0(x)}{pf_0(x)+(1-p)f_1(x)}$

Applications

Sequential analysis

Optimal function

$$V(p)=\min\left\{(1-p)L, pL, C+\int_{-\infty}^{\infty} V\left(\frac{pf_0(x)}{pf_0(x)+(1-p)f_1(x)}\right)\left[pf_0(x)+(1-p)f_1(x)\right]dx\right\}$$

Lemma 6.15 $V(p)$ is a concave function of p

Proof. For $\lambda \in (0, 1)$

$$V\left[\lambda p_1+(1-\lambda)p_2\right]=\min_{\pi\in\Delta} V_\pi\left[\lambda p_1+(1-\lambda)p_2\right]$$

$$\Rightarrow \quad V_\pi\left[\lambda p_1+(1-\lambda)p_2\right]=\lambda V_\pi(p_1)+(1-\lambda)V_\pi(p_2)$$

$$\Rightarrow \quad V\left[\lambda p_1+(1-\lambda)p_2\right]\geq \lambda V(p_1)+(1-\lambda)V(p_2)$$

Applications

Sequential analysis

Theorem 6.16 There exist numbers p^*, p^{**}

If $p > p^{**}$, stop and choose f_0

If $p < p^*$, stop and choose f_1

If $p^{**} < p < p^*$, continue