ADVANCED TOPICS IN OR

Lecture Notes 7

Markov Decision Processes

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

Expected Average Cost Criterion

Costs are bounded

For any policy π, define

$$\phi_\pi(i) = \lim_{n\to\infty} E_\pi \frac{\left[\sum_{t=0}^{n} C(X_t, a_t) \middle| X_0 = i\right]}{n+1}$$

Policy π^* is average cost optimal if

$$\phi_{\pi^*}(i) = \min_{\pi} \phi_\pi(i) \qquad \text{for all } i$$

Question: whether an optimal policy exists?

Counterexample 1:

State space: {1, 1', 2, 2', 3, 3', …}, two actions

Expected Average Cost Criterion

Transition probability

$$P_{i\,i+1}(1) = P_{ii'}(2) = 1 \qquad P_{i'i'}(1) = P_{i'i'}(2) = 1$$

Costs $\quad C(i,\cdot) = 1 \qquad C(i',\cdot) = 1/i$

$X_0 = 1$, let π be any policy

case 1: always choose action 1 $\qquad \phi_\pi(1) = 1 > 0$

case 2: choose action 2 at some time with probability $P_{\bar{n}}$ $\qquad \phi_\pi(1) \geq \dfrac{P_{\bar{n}}}{\bar{n}} > 0$

However, by choosing action 1 long enough and then choosing action 2, we may make our average cost as close to zero as we desire. Thus, an optimal policy does not exist.

Expected Average Cost Criterion

Question: whether we may restrict to stationary policies?

Counterexample 2:

State space: {1, 2, 3, …}, two actions

Transition probability $P_{ii+1}(1) = 1 = P_{ii}(2)$

Costs $C(i,1) = 1$ $C(i,2) = 1/i$

$X_0 = 1$, let π be any policy

case 1: always choose action 1 $\phi_\pi(1) = 1 > 0$

case 2: choose action 2 for the first time at state n

$$\phi_\pi(1) = 1/n$$

Expected Average Cost Criterion

Hence, for any stationary policy, $\phi_{\pi}(1) > 0$

π^*: nonstationary, first enter *i*, choose action 2, *i* consecutive times, then choose action 1

The cost: 1, 1, ½, ½, 1, 1/3, 1/3, 1/3, 1, ¼, ¼, ¼, ¼, 1, 1/5, …

$$\Rightarrow \quad \phi_{\pi^*}(1) = 0$$

Hence, the nonstationary policy π^* is better than every stationary policy

However, randomized stationary policy may be zero cost

But, in general, nonstationary policy may be better than randomized stationary policy

Expected Average Cost Criterion

Conditions under which optimal stationary policies exist

Theorem 6.17 If there exists a bounded function $h(i)$, and a constant g such that

$$g + h(i) = \min_a \left\{ C(i,a) + \sum_{j=0}^{\infty} P_{ij}(a) h(j) \right\}, \quad \text{for all } i$$

then there exists a stationary policy π^* such that

$$g = \phi_{\pi^*}(i) = \min_\pi \phi_\pi(i), \quad \text{for all } i$$

Proof. Let $H_t = (X_0, a_0, \ldots, X_t, a_t)$ denote the history of the process up to time t. For any policy π

$$E_\pi \left\{ \sum_{t=1}^{n} \left[h(X_t) - E_\pi \left(h(X_t) \middle| H_{t-1} \right) \right] \right\} = 0$$

Expected Average Cost Criterion

But
$$E_\pi\left(h\left(X_t\right)\middle|H_{t-1}\right)=\sum_{j=0}^{\infty}h\left(j\right)P_{X_{t-1}j}\left(a_{t-1}\right)$$

$$=C\left(X_{t-1},a_{t-1}\right)+\sum_{j=0}^{\infty}h\left(j\right)P_{X_{t-1}j}\left(a_{t-1}\right)-C\left(X_{t-1},a_{t-1}\right)$$

$$\geq\min_a\left\{C\left(X_{t-1},a\right)+\sum_{j=0}^{\infty}h\left(j\right)P_{X_{t-1}j}\left(a\right)\right\}-C\left(X_{t-1},a_{t-1}\right)$$

$$=g+h\left(X_{t-1}\right)-C\left(X_{t-1},a_{t-1}\right)$$

with equality for π^*. Hence

$$0\leq E_\pi\left\{\sum_{t=1}^{n}\left[h\left(X_t\right)-g-h\left(X_{t-1}\right)+C\left(X_{t-1},a_{t-1}\right)\right]\right\}$$

Expected Average Cost Criterion

or
$$g \le E_\pi \frac{h(X_n)}{n} - E_\pi \frac{h(X_0)}{n} + E_\pi \frac{\sum_{t=1}^{n} C(X_{t-1}, a_{t-1})}{n}$$

with equality for π^*. Letting $n \to \infty$ and using the fact that h is bounded, we have
$$g \le \phi_\pi(X_0)$$

with equality for π^*, and for all possible values of X_0. Proven.

Two questions: why such a theorem should indeed be true?

when are the conditions satisfied?

Approach 1: it seems reasonable that under certain conditions, the average cost case should be in some sense a limit of the discount factor approaches unity.

Expected Average Cost Criterion

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

One possible means, minimizing
$$\lim_{\alpha \to 1} \left\{ C(i,a) + \alpha \sum_{j=0}^{\infty} P_{ij}(a) V_\alpha(j) \right\}$$

However, this limit need not exist and indeed would often be infinite. So, this direct approach is not fruitful.

Indirect approach: fix state 0, and define
$$h_\alpha(i) = V_\alpha(i) - V_\alpha(0)$$

Expected Average Cost Criterion

Then, we have

$$(1-\alpha)V_\alpha(0)+h_\alpha(i)=\min_a\left\{C(i,a)+\alpha\sum_{j=0}^{\infty}P_{ij}(a)h_\alpha(j)\right\}$$

Minimizing the right side is an α – optimal policy

If for some sequence $\alpha_n \to 1$, $\Rightarrow$ $h_{\alpha_n}(j)\to h(j)$

$\Rightarrow$ $(1-\alpha_n)V_{\alpha_n}(0)\to g$

Then we have

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

The policy is the average cost optimal

Expected Average Cost Criterion

Theorem 6.18

If there exists an $N < \infty$, such that

$$\left|V_\alpha(i) - V_\alpha(0)\right| < N \quad \text{for all } \alpha, \quad \text{all } i$$

then: (i) There exists a bounded function $h(i)$ and a constant g satisfying the optimal function

(ii) For some sequence $\alpha_n \to 1$, we have

$$h(i) = \lim_{n \to \infty}\left[V_{\alpha_n}(i) - V_{\alpha_n}(0)\right]$$

(iii) $\lim_{\alpha \to 1}(1-\alpha)V_\alpha(0) = g$

Remark: $h(i)$ inherits the structural form of $V_\alpha(i)$

Expected Average Cost Criterion

An example: machine replacement

We have

$$V_\alpha(0) = C(0) + \alpha \sum_{j=0}^{\infty} P_{0j}(a) V_\alpha(j)$$

$$\Rightarrow \quad V_\alpha(i) \le R + C(0) + \alpha \sum_{j=0}^{\infty} P_{0j}(a) V_\alpha(j) = R + V_\alpha(0)$$

Since $V_\alpha(i)$ is increasing in i (Lemma 6.8), it holds that

$$\left| V_\alpha(i) - V_\alpha(0) \right| \le R$$

$$\Downarrow$$

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

Expected Average Cost Criterion

An example: machine replacement

The average cost optimal policy

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

Theorem: Jensen's inequality

If $g(x)$ is a convex function and X a random variable, then

$$Eg(X) \ge g(EX)$$

Let $M_{i0}(f_\alpha)$, the mean recurrence time from i to 0 when using the α – optimal policy f_α

Expected Average Cost Criterion

Theorem 6.19

If for some state (state 0) there is a constant N such that

$$M_{i0}(f_\alpha) < N, \quad \text{for all } i, \text{ all } \alpha$$

then $V_\alpha(i) - V_\alpha(0)$ is uniformly bounded

Proof. Without loss of generality, all costs are nonnegative.

Let $T = \min\{t : X_t = 0\}$

$$\Rightarrow \quad V_\alpha(i) = E_{f_\alpha} \sum_{n=0}^{T-1} C(X_n, a_n)\alpha^n + E_{f_\alpha} \sum_{n=T}^{\infty} C(X_n, a_n)\alpha^n$$

$$\Rightarrow \quad V_\alpha(i) \le M E_{f_\alpha} T + V_\alpha(0) E_{f_\alpha}\left[\alpha^T\right] \le MN + V_\alpha(0)$$

where M is the bound on costs

Expected Average Cost Criterion

Theorem 6.19

On the other hand, we have

$$V_\alpha(i) \geq V_\alpha(0) E_{f_\alpha}\left[\alpha^T\right]$$

$$\Rightarrow \quad V_\alpha(0) \leq V_\alpha(i) + \left(1 - E_{f_\alpha}\left[\alpha^T\right]\right) V_\alpha(0)$$

Since $V_\alpha(0) \leq M/(1-\alpha)$ and $E\alpha^T \geq \alpha^{ET} \geq \alpha^N$, hence

$$V_\alpha(0) \leq V_\alpha(i) + \left(1 - \alpha^N\right)\frac{M}{1-\alpha} < V_\alpha(i) + MN$$

Corollary 6.20

If the state space is finite and every stationary policy gives irreducible, then $V_\alpha(i) - V_\alpha(0)$ is uniformly bounded

Expected Average Cost Criterion

A special case the average cost criterion may be reduced to a discount cost criterion

Assumption There is a state (state 0), and $\beta > 0$, such that

$$P_{i0}(a) \geq \beta \quad \text{for all } i, \quad \text{all } a$$

Consider a new process, but with transition probabilities

$$P'_{ij}(a) = \begin{cases} \dfrac{P_{ij}(a)}{1-\beta} & j \neq 0 \\[2ex] \dfrac{P_{i0}(a)-\beta}{1-\beta} & j = 0 \end{cases}$$

Let $V_{1-\beta}'(i)$ be the $(1 - \beta)$ – optimal for the new process

Expected Average Cost Criterion

Let $h'(i) = V'_{1-\beta}(i) - V'_{1-\beta}(0)$

$$\beta V'_{1-\beta}(0) + h'(i) = \min_a \left\{ C(i,a) + (1-\beta) \sum_{j=0}^{\infty} P'_{ij}(a) h'(j) \right\}$$

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

It follows that $g = \beta V'_{1-\beta}(0)$ and the average cost optimal policy is to select the action minimizing the right side

The policy is the $(1-\beta)$ – optimal for the new process

⇨ Reduce the average cost problem to a discounted cost problem, the methods of policy improvement or successive approximations may be employed

Computational Approaches

Finite state space, finite action space

Discount case

Policy improvement technique

For any stationary policy *f*, we have

$$V_f(i) = C\left[i, f(i)\right] + \alpha \sum_{j=0}^{m} P_{ij}\left[f(i)\right] V_f(j), \qquad i = 0, 1, \cdots, m$$

m + 1 equations, m + 1 unknowns

Improve f by choosing actions to minimize

$$C(i,a) + \alpha \sum_{j=0}^{m} P_{ij}(a) V_f(j)$$

Computational Approaches

Finite state space, finite action space

Discount case

Change the present $f(i)$ if new action leads to strict improvement

(i) If the improved policy is the original policy f, then f is α – optimal

(ii) If the improved policy is not the original policy f, then the improved policy is strictly better than f

Since there are only a finite number of stationary policies, this policy improvement technique will eventually lead to an α – optimal

Computational Approaches

Finite state space, finite action space

Discount case

Another approach

Lemma 6.21

According to the definition of T_α, for any function u, we have

$$T_\alpha u \geq u \quad \Rightarrow \quad V_\alpha \geq u$$

Proof.

If $T_\alpha u \geq u$, then by the monotonicity of T_α, it follows that

$T_\alpha^n u \geq u$ and the result follows by letting $n \to \infty$

Computational Approaches

Finite state space, finite action space

Discount case

Another approach

Since $T_\alpha V_\alpha = V_\alpha$, it follows that V_α may be obtained by

$$\begin{aligned} &\text{Maximizing} \quad u \\ &\text{Subject to} \quad T_\alpha u \geq u \end{aligned}$$

Maximizing $u(i)$ for each $i \longrightarrow$ maximizing $\sum_{i=0}^{m} u(i)$

The problem reduces to

$$\text{maximizing } \sum_{i=0}^{m} u(i)$$

$$\text{subject to } \min\left\{C(i,a) + \alpha \sum_{j=0}^{m} P_{ij}(a) u(j)\right\} \geq u(i)$$

Computational Approaches

Finite state space, finite action space

Discount case

Another approach

Or equivalently

maximizing $\sum_{i=0}^{m} u(i)$

subject to $C(i,a) + \alpha \sum_{j=0}^{m} P_{ij}(a) u(j) \geq u(i)$ for all a, all i

A linear program

Average cost case

Assumption: all stationary policies give rise to an irreducible Markov chain

Computational Approaches

Finite state space, finite action space

Average cost case

Consider randomized stationary policy

P_i^a : probability of taking action a when in state i

z_i: $i = 0, 1, \ldots, m$, vector of stationary probability

Letting $z_i^a = z_i P_i^a$

It follows that the average cost is $\sum_i \sum_a z_i^a C(i,a)$

Subject to the restrictions $\sum_a z_i^a = \sum_j \sum_a z_i^a P_{ji}(a)$

Computational Approaches

Finite state space, finite action space

Average cost case

$$z_i^a = z_i P_i^a$$

$$\sum_i \sum_a z_i^a = 1$$

$$z_i^a \geq 0$$

$$\sum_a z_i^a = z_i$$

The problem reduces to the above linear program

It turns out that the minimal average cost can be achieved by a nonrandomized policy