書　評

HOWARD, R. A.: DYNAMIC PROGRAMMING AND MARKOV PROCESSES. *Technology Press and Wiley, New York, 1960.*

本書は，著書が 1958 年 MIT 工学部に提出した学位論文を中心にその後の発展をまとめたものである．逐次決定の問題を dynamic programming で解くことはよく知られている方法であるが，MIT 電気工学科の大学院 OR コースでもこの主題がとり上げられた．本書は故に，大学院学生程度の数学知識で理解できるように書かれてある．

第1章: Markov 過程，第2章: rewards のある Markov 過程，(約25頁)は Markov 過程，特に漸近的性質の説明である．第3章: value iteration による逐次決定過程の解法，第4章: 逐次決定過程の解法に対する policy-iteration method，第5章: 前章の方法の応用例3つ(約35頁)が本書の中心をなす部分である．残りの第6章: multiple chain process に対する policy-iteration method，第7章: discounting のある逐次決定過程，第8章: 時間につき連続な決定過程(約60頁)は前記中心部分の種々の拡張である．

本書で扱っている最も代表的な問題を一つ説明しよう．N個の状態をもつ Markov chain があってその遷移確率行列が各 choice q によって $(p_{ij}(q)|i,j=1,\cdots,N)$, $q=1,2,\cdots$, choice q により遷移 $i\to j$ が起ったときの return を $r_{ij}(q)$ とする．

$f_n(i)$…状態 i より出発して，n 期間に亘り最適政策をとったときの全期待利益

とおくと容易に

$$f_n(i)=\max_q \sum_{j=1}^{N} p_{ij}(q)\,(r_{ij}(q)+f_{n-1}(j))$$
$$(i=1,\cdots,N;\ n=1,2,\cdots)$$
$$f_0(i)\equiv 0 \qquad (*)$$

が成立する．右辺の max. をとり去った式を考えれば，Markov chain の一般理論により，もし全状態が同一 chain に属すれば，漸近的に $f_n(i)\sim v_i+ng$ $(n\to\infty)$ とかける．そこでこれを代入して得る

$$v_i+g=\max_q \sum_{j=1}^{N} p_{ij}(q)\,(r_{ij}(q)+v_j)\quad (i=1,\cdots,N)$$

を考える．著者の policy iteration technique はこれの一つの解法である．

先ず適当にとったある policy に対して連立方程式

$$v_i+g=\sum_{j=1}^{N} p_{ij}(r_{ij}+v_j)\quad (i=1,\cdots,N;\ v_N=0)$$

を解く(value determination operation) この解 $(v_1^{(1)},\cdots,v_N^{(1)};\ g^{(1)})$ にもとづいて $\max_q \sum p_{ij}(q)(r_{ij}(q)+v_j^{(1)})$ を求め(policy improvement routine)新しい連立方程式

$$v_i^{(2)}+g^{(2)}=\max_q \sum_{j=1}^{N} p_{ij}(q)\,(r_{ij}(q)+v_j^{(1)}),$$
$$(i=1,\cdots,N;\ v_N^{(2)}=0)$$

を解く．必らず $g^{(2)}\geqq g^{(1)}$ になっている．これを何回も繰返して始めて $g^{(s)}=g^{(s-1)}$ になったらやめる．そのときの maximizing vector q が最適政策である．

例題としてタクシーの問題: 3つの市を仕事場にしているタクシーあり，choices は各市で

q_1: 流して客に拾われるのを待つ

q_2: もよりの cab stand へゆき待つ

q_3: そのままそこに park して radio-call を待つ．

transition および return matrices は

$$(p_{ij}(q_1))=\begin{array}{c|ccc} \nearrow & 1 & 2 & 3 \\ \hline 1 & 1/2 & 1/4 & 1/4 \\ 2 & 1/2 & 0 & 1/2 \\ 3 & 1/4 & 1/4 & 1/2 \end{array},$$

$$(p_{ij}(q_2))=\begin{bmatrix} \frac{1}{16} & \frac{3}{4} & \frac{3}{16} \\ \frac{1}{16} & \frac{7}{8} & \frac{1}{16} \\ \frac{1}{8} & \frac{3}{4} & \frac{1}{8} \end{bmatrix},\qquad \text{etc.}$$

$$(r_{ij}(q_1))=\begin{pmatrix} 10 & 4 & 8 \\ 14 & 0 & 18 \\ 10 & 2 & 8 \end{pmatrix},\quad (r_{ij}(q_2))=\begin{pmatrix} 8 & 2 & 4 \\ 8 & 16 & 8 \\ 6 & 4 & 2 \end{pmatrix}$$

etc.

として total return を最大にする政策を求めよ．immediate return を最大にするのは政策(1, 1, 1)である．これから出発して3 step 目で最適政策(2, 2, 2)すなわち，「常に q_2 をとれ」(これは immediate return を最小にする!)に達する．

なお蛇足を加えると，著者 Dr. Howard は M. I T. 電気工学科の新鋭助教授で当年27才，4人の子持で，Prof. Morse の愛弟子の由．1961 年8月来日したが，氏の談によれば，問題(*)は線型計画法でも解ける(q が有限個ならば)のであって，だから氏の方法は L. P. の simplex 法と本質は同じである．

(坂口　実)