数理解析研究所講究録
第 99 巻 1970 年 29-45

Optimal Stochastic Control

九大 理　古川 長太

§1. 離散時間の場合

1.1 一般的な定義

X, Y ；ある完備可分距離空間の Borel 部分集合

$P(X)$ ； X 上のすべての確率測度の族

$q(y|x)$ ； x を与えたときの Y 上の条件付確率測度

$Q(Y|X)$ ；上記の $q(y|x)$ の族

$M(X)$ ； X 上の実数値有界 Baire 関数の族

以上の諸定義は，定義空間を拡張して

$P(X_1 X_2 \cdots X_n)$,　　$P(X_1 X_2 \cdots)$,

$Q(Y|X_1 \cdots X_n)$,　$Q(Y_1 Y_2 \cdots Y_m | X_1 X_2 \cdots X_n)$,

$Q(Y_1 Y_2 \cdots | X_1 X_2 \cdots X_n)$

の様に拡張される。

pu ；　$pu = \int_X u(x)\, dp(x)$　　for $p \in P(X)$,　$u \in M(X)$

qu ; $qu(x) = \int_Y u(x,y)\,dq(y|x)$ for $q \in Q(Y|X)$, $u \in M(XY)$

この定義も，直積空間を定義空間として拡張される。

$p \in P(X)$ なる p が degenerate ; $p\{x\} = 1$ for some $x \in X$

$q \in Q(Y|X)$ なる q が degenerate ; $q(\cdot|x)$ が 各xに対して degenerate

1.2 最適化問題の定義

S ; 状態空間（ある完備可分距離空間の Borel 部分集合）

A ; 行動空間（ある完備可分距離空間の Borel 部分集合）

$S \ni s$: 状態（state）

$A \ni a$: 行動（action）

$H_n \equiv SASA \cdots SAS$ （$2n-1$ 個）

$q^N \equiv \{q_1, q_2, \cdots, q_N\}$ ただし $q_n \in Q(S|H_nA)$ $(1 \leqq n \leqq N)$

$\pi^N \equiv \{\pi_1, \pi_2, \cdots, \pi_N\}$ ただし $\pi_n \in Q(A|H_n)$ $(1 \leqq n \leqq N)$

$(1 \leqq N \leqq +\infty)$

このような π^N を 政策（policy）という。

マルコフ政策（Markov policy）; $\pi^N = \{\pi_1, \pi_2, \cdots, \pi_N\}$ において，各 $\pi_n \in Q(A|S)$ でかつ，各 π_n が degenerate

マルコフ政策を $\pi^N = \{f_1, f_2, \cdots, f_N\}$ と書く。

定常政策 (stationary policy) ; マルコフ政策において

$$f_1 = f_2 = \cdots = f_N$$

$r_n \in M(H_{n+1})$ $(n=1,2,\cdots,N)$; 利益関数 (reward)

評価関数としてつぎのものを考える。

$$I_N(\pi^N, q^N) = \sum_{n=1}^{N} \pi_1 q_1 \pi_2 q_2 \cdots \pi_n q_n r_n$$

又は, $p \in P(S)$ として

$$p\, I_N(\pi^N, q^N).$$

政策 π^N を採用したとき system の運動は (π^N, q^N) で決定される H_{N+1} 上の確率過程になる。

Δ を政策のある与えられた族として, Δ の中で

$$I_N(\pi^N, q^N) \to \mathrm{Max}\ (\text{又は } \mathrm{Min})$$

$$p\, I_N(\pi^N, q^N) \to \mathrm{Max}\ (\text{又は } \mathrm{Min})$$

$$I_\infty(\pi^\infty, q^\infty) \to C\ (\text{given constant})$$

$$p\, I_\infty(\pi^\infty, q^\infty) \to C\ (\text{given constant})$$

ならしめる問題を離散時間確率的系における (one person) 最適化問題という。

1.3 マルコフ型の場合

$$q_1 = q_2 = \cdots = q_n = \cdots = q$$

$$q \in Q(S \mid SA)$$

$$r \in M(SAS) \qquad \text{(reward)}$$

$$\beta \quad (0 < \beta < 1) \qquad \text{(discount factor)}$$

$$I(\pi) \equiv \sum_{n=1}^{\infty} \beta^{n-1} \pi_1 q \pi_2 q \cdots \pi_n q r \in M(S)$$

π^* が (p, ε)-最適 ; $p\{ I(\pi^*) \geqq \sup_{\pi} I(\pi) - \varepsilon \} = 1$

π^* が ε-最適 ; $I(\pi^*) \geqq \sup_{\pi} I(\pi) - \varepsilon$

π^* が 最適 ; $I(\pi^*) \geqq \sup_{\pi} I(\pi)$

定理 1.1 (Strauch [7], Furukawa [2])

各 $p \in P(S)$, 各 $\varepsilon > 0$ に対して, (p, ε)-最適政策が存在する。

(この定理は, 最初 Strauch が証明し, 後に Furukawa が証明を簡単化した)

定理 1.2 (Blackwell [1])

各 $p \in P(S)$, 各 $\varepsilon > 0$ に対して (p, ε)-最適マルコフ政策が存在する。

定理 1.3 (Blackwell [1])

各 $p \in P(S)$, 各 $\varepsilon > 0$ に対して (p, ε)-最適定常政策が

存在する。

定理 1.4 （Blackwell [1]）

(i) Aが可算集合なら，各$p \in P(S)$，各$\varepsilon > 0$に対してε-最適定常政策が存在する。

(ii) Aが有限集合なら，各$p \in P(S)$，各$\varepsilon > 0$に対して最適定常政策が存在する。

定理 1.5 （Howard [3]）

A，Sが有限集合なら，政策反復法によって最適政策の構成が出来る。

1.4 Sが吸収点をもつ場合

$(\Omega, \mathcal{F}, P)$ 確率空間

$\{\mathcal{F}_n\}_{n=1,2,\cdots}$ ； $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_n \cdots$ なる$\mathcal{F}$の sub-σ-field の系列

$\{x_n\}_{n=1,2,\cdots}$ ； 各nにつきx_nが$\mathcal{F}_n$-可測であるような確率過程

$\tau(\omega)$ ； つぎの(i)～(iii)をみたす確率変数

(i) $\tau(\omega)$は正整数値をとる

(ii) $P\{\tau(\omega) < +\infty\} = 1$

(iii) $\{\omega ;\ \tau(\omega) = n\} \in \mathcal{F}_n$ for each n

このような $\tau(\omega)$ を stopping time という。

$E(x_\tau) \to$ Max (又は Min) ならしめる (stopping time τ に関して) 問題を Optimal Stopping Problem という。

$X = \{x_n, \mathcal{F}_n\}$, $\quad Y = \{y_n, \mathcal{F}_n\}$ に対して

$$Y < X \quad \text{w.p.1} \overset{\text{def}}{\Longleftrightarrow} y_n < x_n \quad \text{w.p.1} \quad \text{for all } n$$

maximal semi-Martingale ;

$Z = \{z_n, \mathcal{F}_n\}$ に対して, $X = \{x_n, \mathcal{F}_n\}$ があって

(i) X は semi-Martingale

(ii) $X < Z$ w.p.1

(iii) $Y < Z$ w.p.1 なるいかなる semi-Martingale Y に対しても, $Y < X$ w.p.1

なるとき, X は Z に関して maximal semi-Martingale であるという。

<u>定理 1.6</u> (Snell [6])

$Z = \{z_n, \mathcal{F}_n\}$ をある与えられた確率過程, $X = \{x_n, \mathcal{F}_n\}$ を Z に関して maximal semi-Martingle で

かつ regular な確率過程とする。$\tau^*(\omega)$ を

$$\tau^*(\omega) = \begin{cases} 1 & \text{if} \quad x_1(\omega) = \infty \\ j & \text{if} \quad x_k(\omega) < z_k(\omega) - \varepsilon \ \text{for} \ k < j \ \text{and} \ x_j(\omega) \geqq z_j(\omega) - \varepsilon \\ \infty & \text{その他} \end{cases}$$

で定義すると，$\tau^*(\omega)$ は stopping time でかつ

$$E(x_\tau) \geqq E(x_{\tau^*}) - \varepsilon \quad \text{for all } \tau \in T$$

(ただし T は，すべての stopping time の族)

上のような考えを control model に適用する。

$Z^{(i)} = \{z_n^{(i)}, \mathcal{F}_n\}_{n=1,2,\cdots} \qquad (i=1,2,\cdots)$

ただし 各 $Z^{(i)}$ は Ω 上の実数値確率過程で，$z_n^{(i)}$ は $\mathcal{F}_n$-可測とする。

$A = \{1, 2, 3, \cdots\}$; action space

$a_n(\omega)$; $\Omega \to A$ への $\mathcal{F}_{n-1}$-可測な mapping $(n=1,2,\cdots)$

$\tau(\omega)$; stopping time

$\pi \equiv \{a_1(\omega), a_2(\omega), \cdots, a_{\tau(\omega)}(\omega)\}$; policy

Φ_n ; R^n 上に定義された $\mathcal{B}_n$-可測 実数値関数 $(n=1,2,\cdots)$

δ_{ij} ; Kronecker のデルタ

<u>問題</u> $\mathcal{Z} \equiv \{Z^{(1)}, Z^{(2)}, \cdots, Z^{(i)}, \cdots\}$ (ただし $Z^{(i)} = \{z_n^{(i)}\}_{n=1,2,\cdots}$) を与えられた確率過程の系列

とし,

$$\Psi_\tau(Z,\pi) \equiv \Phi_\tau\left(\sum_{i_1=1}^{\infty}\delta_{i_1}a_1 z_1^{(i_1)}, \sum_{i_2=1}^{\infty}\delta_{i_2}a_2 z_2^{(i_2)}, \cdots, \sum_{i_\tau=1}^{\infty}\delta_{i_\tau}a_\tau z_\tau^{(i_\tau)}\right)$$

とおいて,

$$E[\Psi_\tau(Z,\pi)] \longrightarrow \mathrm{Min}$$

ならしめる((π, τ) に関して) 問題を考える。

<u>Def</u>　$X^{(i)} = \{x_n^{(i)}\}_{n=1,2,\cdots}$ を Ω 上の確率過程とし

$$X \equiv \{X^{(1)}, X^{(2)}, \cdots, X^{(i)}, \cdots\}$$ とおく。

$\{\Phi_n(x_1^{(i_1)}, x_2^{(i_2)}, \cdots, x_n^{(i_n)})\}_{n=1,2,\cdots}$ が各 $(i_1, i_2, \cdots)$ に対して semi-Martingale であるとき, X は $\{\Phi_n\}$ に関して uniform semi-Martingale であるという。

<u>Def</u>　$\tilde{\pi} \equiv \{a_1(\omega), a_2(\omega), \cdots\}$

$$\Psi_n(X,\tilde{\pi}) \equiv \Phi_n\left(\sum_{i_1=1}^{\infty}\delta_{i_1}a_1 x_1^{(i_1)}, \sum_{i_2=1}^{\infty}\delta_{i_2}a_2 x_2^{(i_2)}, \cdots, \sum_{i_n=1}^{\infty}\delta_{i_n}a_n x_n^{(i_n)}\right)$$

$\{\Psi_n(X,\tilde{\pi})\}$ を $\tilde{\pi}$ によって変換された X 上の process という。

<u>定理 1.7</u> (Furukawa)

$X = \{X^{(1)}, X^{(2)}, \cdots\}$ は $\{\Phi_n\}$ に関して uniform semi-Martingale とし,

$$\sum_{i_1=1}^{\infty}\sum_{i_2=1}^{\infty}\cdots\sum_{i_n=1}^{\infty} E\left[\,\left|\Phi_n(x_1^{(i_1)}, x_2^{(i_2)}, \cdots; x_n^{(i_n)})\right|\,\right] < \infty$$

for each n,

$\tilde{\pi}$によって変換されたX上のprocess $\{\Psi_n(X, \tilde{\pi})\}$が

Pに関して一様可積分とする。

このとき，

$\{\Psi_n(X, \tilde{\pi})\}$は収束して，かつ，$\{\Psi_n(X, \tilde{\pi}), n \geq 1, \Psi_\infty(X, \tilde{\pi})\}$は semi-Martingale である。

（上に設定した問題，あるいは，もっと拡張された問題に対する解答は次回に発表する予定）

1.5 2-person の最適化問題

S ; state space

A_1 ; player I の action space

A_2 ; player II の action space

$A \equiv A_1 \times A_2$

$q \in Q(S \mid SA)$; transition probability measure

$r \in M(SAS)$; reward function

$0 < \beta < 1$; discount factor

$H_n^{(1)} \equiv S A_1 S A_1 \cdots S A_1 S$ $(2n-1$ 个$)$

$H_n^{(2)} \equiv S A_2 S A_2 \cdots S A_2 S$ $(2n-1$ 个$)$

$\pi = \{\pi_1, \pi_2, \cdots\}$; $\pi_n \in Q(A_1 \mid H_n^{(1)})$ $(n=1,2,\cdots)$

.... player I の policy

$\sigma = \{\sigma_1, \sigma_2, \cdots\}$; $\sigma_n \in Q(A_2 \mid H_n^{(2)})$ $(n=1,2,\cdots)$

.... player II の policy

$\pi^\vee\sigma \equiv \{\pi_1, \sigma_1, \pi_2, \sigma_2, \cdots\}$ paired policy

$$I(\pi^\vee\sigma) = \sum_{n=1}^{\infty} \beta^{n-1} \pi_1\sigma_1 q \pi_2\sigma_2 q \cdots \pi_n\sigma_n q r$$

player I は $I(\pi^\vee\sigma) \to$ Max ならしめる（πについて）

player II は $I(\pi^\vee\sigma) \to$ Min ならしめる（σについて）

<u>Def</u>

$\pi^{*\vee}\sigma^*$ が $(p, \varepsilon_1, \varepsilon_2)$－最適 ;

$$p\{ I(\pi^{*\vee}\sigma) + \varepsilon_1 \geqq I(\pi^{*\vee}\sigma^*) \geqq I(\pi^\vee\sigma^*) - \varepsilon_2 \} = 1 \quad \text{for all } \pi^\vee\sigma$$

<u>Def</u> $\{\pi^m\}$を player I の policy の系列とする。このとき

$\hat{\pi}$ が $\{\pi^m\}$ で生成される $\overset{def}{\Longleftrightarrow}$ $\exists\{S_m\}$ partition of S ;

$\hat{\pi} = \pi^m$ for $s_1 \in S_m$

$\{\pi^m\}$ で生成される policy の全体を $\mathfrak{G}(\{\pi^m\})$ と書く。

同様にして, player II の policy の系列 $\{\sigma^n\}$ で生成される policyの全体 $\mathfrak{G}(\{\sigma^n\})$ も定義される。

定理 1.8 (Furukawa)

$$\sup_{\pi} pI(\pi \vee \sigma) = \sup_{m} pI(\pi^m \vee \sigma) \quad \text{for every } \sigma \in \mathcal{O}\mathcal{J}(\{\sigma^n\})$$

$$\inf_{\sigma} pI(\pi \vee \sigma) = \inf_{n} pI(\pi \vee \sigma^n) \quad \text{for every } \pi \in \mathcal{O}\mathcal{J}(\{\pi^m\})$$

ならば，各 $\varepsilon_1 > 0$，各 $\varepsilon_2 > 0$ に対して $(p, \varepsilon_1, \varepsilon_2)$ − 最適な paired policy が存在する。

§2 連続時間の場合

2.1 マルコフ型の場合

はじめに Miller ([5]) の結果を紹介する。

$S = \{1, 2, \cdots, n\}$; state space

A ; action space である有限集合とする。

$F \equiv A \times A \times \cdots \times A$ (n 個)

$\pi(t)$ (n ベクトル) ; $[0, T] \to F$ なる L−可測な mapping

$f = (f_1, f_2, \cdots, f_n) \in F$ に対して

$$r(f) = \begin{pmatrix} r(1, f_1) \\ r(2, f_2) \\ \vdots \\ r(n, f_n) \end{pmatrix}$$

$$Q(f) = \Big(q(i, f_i, j) \Big) \begin{matrix} \downarrow \\ i \end{matrix} \; ; \text{ Markov infinitesimal generator matrix}$$

$$\longrightarrow j$$

ただし

$$q(i,a,j) \geqq 0 \quad \text{for } i \neq j, \text{ for all } a \in A,$$
$$\sum_{j=1}^{n} q(i,a,j) = 0.$$

$Q(\pi(\cdot))$ の L-可測性を仮定すれば殆んどすべての t について (1) をみたす絶対連続一意解 $P(s,t)$ が存在し，かつそれは Markov transition matrix になる。

$$(1) \quad \begin{cases} \frac{d}{dt}P(s,t) = P(s,t)Q(\pi(t)) \\ P(s,s) = I \end{cases}$$

評価関数： $v = \int_0^T P(0,t)r(\pi(t))\,dt \longrightarrow$ Max ならしめる。

<u>最大原理</u>

policy π が，補助方程式：

$$(2) \quad \begin{cases} -\frac{d\psi}{dt} = r(\pi(t)) + Q(\pi(t))\psi & 0 \leq t \leq T \\ \psi(T) = 0 \end{cases}$$

の一意解 $\psi(t)$ に対して

$$r(\pi(t)) + Q(\pi(t))\psi(t) = \max_{f \in F}\left[r(f) + Q(f)\psi(t)\right], \quad \text{a.e. } t \in [0,T]$$

をみたすとき，π は最大原理をみたすという。

上のことと equivalent に

$$(3) \qquad -\frac{d\Psi}{dt} = \max_{f\in F}\left[r(f) + Q(f)\Psi(t) \right]$$

が成立する。(3) はいわゆる Bellman の最適性の原理であるが，Bellman は (3) を導いたわけではなく，(3) を以て最適条件と定義した。

更に，Miller は，homogeneous Markov（Q が t を explicit に含まない）の場合について，π が最適であることと π が最大原理をみたすこととが同等であることを証明し，また，piecewise constant な最適政策の存在定理を与えた。

Martin-Löf（[4]）は，finite state space, compact action space で non-homogeneous Markov の場合（$Q(t,\omega)$），を扱い，$Q(t,\omega)$, $r(t,\omega)$ は共に有界かつリプシッツ条件をみたし，$Q(t,u(t))$ は t に関して L－可測を仮定して，最適政策の存在を示した。更に詳細な結果として，periodic case では最適な periodic policy の存在を，また homogenous case では最適な homogeneous policy の存在を示した。

Rykov（[8]）は，finite state, finite action で homogeneous Markov の場合，時間区間 $[0,+\infty)$ にわたっての平均（単位時間当りの）期待利得を Max ならしめる

問題を考え，(3) の analogy を導き，また stationary な最適政策の存在を示した。

2.2 ある非線型確率微分方程式系

(4) $$\frac{dx(\omega,t)}{dt} = f(t, x(\omega,t), u(\omega,t)) + \sigma(x(\omega,t), u(\omega,t))\frac{dz(\omega,t)}{dt}$$

$$(t \in R^1,\ x \in R^n,\ u \in R^m)$$

Ω ; sample space

T ; 任意に fix

$z(\omega,t)$; Ω 上の確率過程

$\Sigma(t)$; $\{z(\cdot,\tau),\ 0 \le \tau \le t\}$ が可測となる Ω 上の最小の σ-field

$\mathfrak{J}(T)$; $[0,T]$ 上の Borel field

$\tilde{\Sigma}(T) \equiv \mathcal{B}\left(\bigcup_{0\le t\le T}(\Sigma(t)\times\mathfrak{J}(T))\right)$

X^n ; R^n のすべての compact subset からなる空間

つぎの仮定をおく。

1° 各 $\omega \in \Omega$ に対し，$u(\omega,\cdot)$ が $[0,T]$ 上で L-可測，
各 $t \in [0,T]$ に対し，$u(\cdot,t)$ が Ω 上で $\Sigma(t)$ 可測
$u(\cdot,\cdot)$ は $\Omega\times[0,T]$ 上で $\tilde{\Sigma}(T)$ 可測

2° $A \equiv [0,T]\times C$ ただし C は R^n の closed subset

3° control space $U(t, x)$ は $A \to X^m$ なる mapping で、各 t に対し、x に関して連続（Hausdorff distance で）、各 x に対し、t に関して L-可測

4° $f(t, x, u) = (f_1, f_2, \cdots, f_n)$ は $A \times \bigcup_{(t,x)\in A} U(t, x) \to R^n$ の mapping で、各 t に対し、(x, u) に関して連続、各 (x, u) に対し、t に関して L-可測

5° cost function $f_0(t, x, u)$; $A \times \bigcup_{(t,x)\in A} U(t, x) \to R^1$ は、各 t に対し、(x, u) に関して連続、各 (x, u) に対し t に関して L-可測

6° $\tilde{f} = (f_0, f)$ とおくとき、

$$\|\tilde{f}(t, x, u)\| \leqq m(t) \quad \text{for } (t, x, u) \in A \times \bigcup_{(t,x)\in A} U(t, x)$$

なる $[0, T]$ 上 L-可積関数 $m(t)$ が存在する。

7° $Q(t, x) \equiv \{ (v^0, v) ; v^0 \geqq f_0(t, x, u),$

$$v = f(t, x, u) + \sigma(x, u)\frac{dz}{dt}, \quad u \in U(t, x) \}$$

が殆んどすべての $\omega \in \Omega$ に対して凸集合である。

$t_1(\omega)$, $t_2(\omega)$ を $t_1(\omega) \leqq t_2(\omega) \leqq T$ w.p.1 なる random time とし、

$$E\left[\int_{t_1(\omega)}^{t_2(\omega)} f_0(t, x(\omega, t), u(\omega, t))\,dt\right] \to \mathrm{Min}$$

ならしめる。

定理 2.1 (Furukawa)

1°〜7° の仮定のもとで，いかなる non-empty な complete class の中でも最適な control が存在する．

参 考 文 献

[1] D. Blackwell ; Discounted dynamic programming, Ann. Math. Statist., 36 (1965), 226-235.

[2] N. Furukawa ; A Markov decision process with non-stationary transition laws, Bulletin of Math. Statist., 13 (1968), 41-52.

[3] R.A. Howard ; Dynamic programming and Markov processes, Technology Press of M.I.T and John Wiley, (1960)

[4] A. Martin-Löf ; Optimal control of a continuous-time Markov chain with periodic transition probabilities, Operations Research, 15 (1967), 872-881

[5] B.L. Miller ; Finite state continuous time Markov decision processes with a finite planning horizon, SIAM Jour. on Control, 6 (1968), 266-280

[6] J. L. Snell ; Application of martingale system theorems, Trans. Amer. Math. Soc., 73 (1952), 293 – 312

[7] R. E. Strauch ; Negative dynamic programming, Ann. Math. Statist., 37 (1966), 871 – 890

[8] V. V. Rykov. ; Markov decision processes with finite state and decision spaces, Theory of Prob. and its Appl., 11 (1966), 302 – 311