数理解析研究所講究録
32
第747巻 1991年 32-42

Functional Characterization for Average Cost Markov Decision Processes with Doeblin's Conditions

千葉大・教育　蔵野正美 (Masami Kurano)

Kurano [5,6] は Doeblin's Condition の考え方 [4] を用いて，Compact Metric Space 上の平均コスト基準のマルコフ決定過程（MDPs）を取扱い，任意の randomized stationary policy によって誘導される（induced）マルコフ過程に，幾つかの ergodic classes と transient set が許される general (multichain) case に対して，最適定常政策の存在定理を与えている．

本報告では，同じ問題設定のもとでの平均コスト基準のマルコフ決定過程に対する関数的特徴づけ（最適方程式の導出，及び，その有効性）を行なう。 なお，この報告は，Kurano [7] の内容を一部手直ししたものである．

1. 定式化

ある complete separable な距離空間のボレル部分集合を挙

こボレル集合とよぶ。ボレル集合 X のボレル部分集合の全体を $\mathcal{B}_X$ で表す。

マルコフ決定過程 (Markov Decision Processes, MDPs) は次の4つの要素 S, A, Q, c から成る：

(i) S はボレル集合で状態空間を表す。

(ii) 各 $x \in S$ に対して，$A(x)$ はボレル集合 A の部分集合で，状態 x においてとりうる行動の全体を表す。

(iii) $c : S \times A \to (-\infty, \infty)$ は，有界なボレル可測関数で，直接費用関数 (immediate cost function) を表す。

(iv) Q は $\mathcal{B}_S \times S \times A$ 上の確率核で次の条件 (a), (b) を満す。

(a) 各 $(x,a) \in S \times A$ に対して，$Q(\cdot \mid x,a)$ は $\mathcal{B}_S$ 上の確率測度である。

(b) 各 $D \in \mathcal{B}_S$ に対して，$Q(D \mid \cdot)$ は $S \times A$ 上のボレル可測関数である。

この報告を通じて，次が仮定される。

<u>仮定</u>

(i) S と $R := \{(x,a) \mid x \in S,\ a \in A(x)\}$ は共にコンパクト集合である。

(ii) コスト関数 c は非負値有界かつ下半連続である。

(iii) $x_n \to x$, $a_n \to a$ のとき，$Q(\cdot \mid x_n, a_n)$ は $Q(\cdot \mid x, a)$ に弱収束する。

考察する決定過程の標本空間は $\Omega = (S \times A)^\infty$ で，t 期の状態と行動は，確率変数 X_t，Δ_t で表す．$(t \geq 0)$

各 $t \geq 0$ に対して，$\mathcal{B}_A \times S \times (A \times S)^t$ 上の確率核 π_t で

$$\pi_t(A(x_t) \mid x_0, a_0, \cdots, a_{t-1}, x_t) = 1$$

$$\text{for all } (x_0, a_0, \cdots, a_{t-1}, x_t) \in S \times (A \times S)^t$$

を満たすものの集合 $\pi = (\pi_0, \pi_1, \cdots)$ を政策という．

今，$D \in \mathcal{B}_S$ に対して，$T(A \mid D)$ を $\mathcal{B}_A \times D$ 上の確率核全体で，すべての $x \in D$ で，$\varphi(A(x) \mid x) = 1$ を満たす φ の全体とする．

政策 $\pi = (\pi_0, \pi_1, \cdots)$ がランダム定常政策であるとは，ある $\varphi \in T(A \mid S)$ が存在して，すべての $(x_0, a_0, \cdots, x_t) \in S \times (A \times S)^t$ とすべての $t \geq 0$ に対して，

$$\pi_t(\cdot \mid x_0, a_0, \cdots, x_t) = \varphi(\cdot \mid x_t)$$

が成り立つときをいう．この場，π を単に $\varphi^{(\infty)}$ で表す．

任意の $D \in \mathcal{B}_S$ に対して，$u(x) \in A(x)$，$x \in D$ を満たすボレル可測 (analytically measurable) 関数 $u : D \to A$ の全体を $\mathcal{B}(D \to A)$ $(\mathcal{B}_a(D \to A))$ で表す．

ランダム定常政策 $\varphi^{(\infty)}$ が定常であるとは，$f \in \mathcal{B}(S \to A)$ が存在して，$\varphi(\{f(x)\} \mid x) = 1$，$x \in S$ が成り立つときをいう．

そのような政策を $f^{(\infty)}$ で表す．

t 期までの履歴を $H_t = (X_0, \Delta_0, \cdots, \Delta_{t-1}, X_t)$ とする．

任意に与えられた政策 $\pi = (\pi_0, \pi_1, \cdots)$ に対して，次を仮定

する： すべての $D_1 \in \mathcal{B}_A$, $D_2 \in \mathcal{B}_S$ に対して

$$\mathrm{Prob}\,(\Delta_t \in D_1 \mid H_t) = \pi_t(D_1 \mid H_t)$$

$$\mathrm{Prob}\,(X_{t+1} \in D_2 \mid H_{t-1}, \Delta_{t-1}, X_t = x, \Delta_t = a) = Q(D_2 \mid x, a)$$

$(t \geq 0)$.

このとき，任意の政策 $\pi \in \Pi$ と初期分布 $\nu \in \mathcal{P}(S)$ に対して，Ω 上の確率測度 P_π^ν が通常の方法で定義される．

但し，$D \in \mathcal{B}_S$ に対して，$\mathcal{P}(D)$ は D 上の確率分布の全体を表す．

次の平均コスト基準を考察する．

任意の $\pi \in \Pi$ と初期分布 $\nu \in \mathcal{P}(S)$ に対して，

$$\gamma(\nu, \pi) := \limsup_{T \to \infty} E_\pi^\nu \left[\sum_{t=0}^{T-1} c(X_t, \Delta_t)\right] / T.$$

但し，E_π^ν は P_π^ν に関する期待値を表す．

さらに，次を定義する．

$$\gamma(\nu) := \inf_{\pi \in \Pi} \gamma(\nu, \pi),$$

$$\gamma^* := \inf_{\nu \in \mathcal{P}(S)} \gamma(\nu).$$

任意の $D \in \mathcal{B}_S$ に対して，

$$\gamma(x, \pi^*) \leqq \gamma(x, \pi) \quad , x \in D,\ \pi \in \Pi$$

が成り立つとき，π^* を D において最適 (optimal in D) という．

S において最適な政策を単に最適であるという．

2 最適方程式 (1)

この節では，MDPs に対する positive recurrence の条件のもとで，最適方程式を導出し，その有効性を議論する。

任意の $\pi \in \Gamma(A|S)$ に対して，t 期の推移確率 $Q^{(t)}$ を次で定義する：

$$Q^{(1)}(\cdot|x,\pi) = \int Q(\cdot|x,a)\pi(da|x)$$

$$Q^{(t+1)}(\cdot|x,\pi) = \int Q^{(t)}(\cdot|x_1,\pi)Q^{(1)}(dx_1|x,\pi) \quad (t \geq 1).$$

この報告を通じて，次の Doeblin 条件が成り立つことを仮定する。

仮定 (Doeblin [4])

次を満足する $\mathcal{B}_S$ 上の有限測度 γ と $\varepsilon > 0$ が存在する：

任意の $\pi \in \Gamma(A|S)$ に対して，自然数 l が存在して，

$\gamma(D) \leq \varepsilon$ なる $D \in \mathcal{B}_S$ に対して，$Q^{(l)}(D|x,\pi) \leq 1-\varepsilon$

がすべての $x \in S$ に対して成り立つ。

次の定理はすでに証明されている。

定理 2.1 ([5])

次の (i), (ii) を満たす $\gamma(C) > \varepsilon$ なる $C \in \mathcal{B}_S$ と定常政策 $f^{(\infty)}$ が存在する。

(i) $f^{(\infty)}$ は C において最適で，かつ，すべての $x \in C$ において

$$\gamma^* = \gamma(x, f^{(\infty)}).$$

(ii) すべての $x \in C$ において，$Q(C|x,f(x)) = 1$ で，

$Q(\cdot|x,f(x))$ によって誘導される (induced) C 上のマルコ

$\mathcal{T}$過程は transient state を許さない。

定理2.1の $C \in \mathcal{B}_S$ に対して，次の Lemma 2.1 が成り立つ。

Lemma 2.1

定理2.1の $C \in \mathcal{B}_S$ と定常政策 $f^{(\infty)}$ に対して，C 上の一様有界なボレル可測関数 $u \in B(C)$ が存在して，次を満たす：

$$u(x) + \gamma^* = c(x, f(\infty)) + \int_C u(x') Q(dx' | x, f(\infty))$$

for all $x \in C$.

(略証)

Doeblin条件のもとでのマルコフ過程の性質を利用する。

$Q(\cdot|x) := Q(\cdot|x, f(\infty))$

$\{Q(\cdot|x), x \in C\}$ による induced Markov Process on C は one ergodic set のみであると仮定してよい。

$C_1, C_2, \cdots, C_d$: the cyclically moving classes in C

$x \in C_1$ に対して，

$$\lim_{t\to\infty} Q^{(td+j-1)}(\cdot|x) = \nu_j(\cdot)$$

uniformly and exponentially fast

$$\nu(\cdot) := \frac{1}{d}\sum_{j=1}^{d} \nu_j(\cdot)$$

このとき，$\gamma^* = \int_C c(x, f(\infty))\, \nu(dx)$ が成り立つ。

$$u^T(x) := \sum_{t=0}^{[T/d]d-1} E^x_{f^{(\infty)}}[c(X_t, \Delta_t) - \gamma^*]$$

$$u(x) := \lim_{T\to\infty} u^T(x) \quad (\text{一様収束})$$

このとき，

$$u(x) = \sum_{t=0}^{\infty} E^x_{f^{(\infty)}}[\, c(X_t, \Delta_t) - \gamma^* \,]$$

上式を再帰式に書きかえれば，$u(x)$ は Lemma 2.1 の関係式を満たす．　(証終)

任意の $D \in \mathcal{B}_S$ に対する hitting time を次で定義する．

$$T_D := \inf\{\, t \geq 0 \mid X_t \in D \,\} \quad \text{但し } \inf\phi = \infty$$

Lemma 2.1 の結果を C から全空間に拡大するためには，MDP に対する次の positive recurrence の条件を必要とする．

仮定 A

$\gamma(D) > \varepsilon$ なる任意の $D \in \mathcal{B}_S$ と $x \in S$ に対して．

$E^x_\pi(T_D) < \infty$ なる $\pi \in \Pi$ が存在する．

Lemma 2.2 ([9])

定常政策 $f^{(\infty)}$ と $D \in \mathcal{B}_S$ に対して，次の不等式を満たす $\varphi \in B_e(S)$, $\varphi \geq 0$ と正の定数 α が存在するとする：

$$\varphi(x) \geq \alpha + \int_{S-D} \varphi(x')\, Q(dx' \mid x, f(x)) \quad , \forall x \notin D$$

このとき $E^x_{f^{(\infty)}}(T_D) \leq \varphi(x)/\alpha$ が成り立つ．

Lemma 2.3

仮定 A のもとで，$\gamma(D) > \varepsilon$ なる任意の $D \in \mathcal{B}_S$ に対して，$E^x_{f^{(\infty)}}(T_D) < \infty$, $x \notin D$ を満たす $f^{(\infty)} \in B_a(S \to A)$ が存在する．

ボレル集合 X に対して，X 上の universally measurable functions の全体を $B_u(X)$ で表す。記述を簡単にするために，$B_u(S)$ 上の operator U を次のように定義する：

各 $x \in S$, $a \in A(x)$ に対して

$$U(x,a,u) := c(x,a) + \int u(x') Q(dx' \mid x,a), \quad u \in B_u(S)$$

次の結果はよく知られている。

Lemma 2.4 (たとえば [8])

定常政策 $f^{(\infty)}$ に対して，次の2つの関係式を満たす $u \in B_u(S)$ が存在するとする：

$$u(x) + \gamma^* \geq U(x, f(x), u), \quad \forall x \in S$$

$$\lim_{T \to \infty} E^x_{f^{(\infty)}}(u(X_T))/T = 0$$

このとき，$f^{(\infty)}$ は最適政策である。

以上の Lemmas を用いて次の定理を得る。証明は [7] を参照のこと。

定理 2.2 仮定 A のもとで，次の (i)～(iii) を満たす $v \in B_u(S)$ が存在する：

(i) v は次の最適不等式を満たす。

$$v(x) + \gamma^* \geq \inf_{a \in A(x)} U(x,a,v), \quad \forall x \in S$$

(ii) 定常政策 $f^{(\infty)}$ ($f \in B_a(S \to A)$) が次の (*), (**) を満たすならば，$f^{(\infty)}$ は最適である。

(*) $\quad v(x) + \psi^* \geq \bigcup(x, f(x), v) \qquad \forall x \in S$

(**) $\quad \liminf_{T\to\infty} E^x_{f^{(\infty)}}(v(X_T))/T = 0$

(iii) (*) (**) を満たす定常政策 $f^{(\infty)}$ ($f \in B_a(S \to A)$) が存在する。

3. 最適方程式 (2)

この節は positive recurrence の仮定（前節の仮定 A）が成り立たない場合について考察する。

<u>定義</u>

任意の $D \in \mathcal{B}_S$ に対して,

$$\gamma(D) := \{ x \in S - D \mid E^x_\pi(T_D) < \infty \text{ for some } \pi \in \Pi \}$$

<u>定義</u>

任意の $D \in \mathcal{B}_S$ に対して.

$$\Gamma(D) := \{ (\nu, \pi) \in P(D) \times \Pi \mid P^\nu_\pi(X_T \in D \text{ for all } t \geq 0) \}$$

$$\gamma^*(D) := \inf_{(\nu,\pi)\in\Gamma(D)} \gamma(\nu, \pi)$$

但し $\Gamma(D) = \phi$ のとき. $\gamma^*(D) = \infty$ とする。

次の仮定が必要である：

<u>仮定 B</u> 次の B1, B2 が成り立つ.

B1. $\gamma(D) > 0$ なる任意の $D \in \mathcal{B}_S$ に対して.

$Q(\partial D \mid x, a) = 0$, $\forall x \in S$, $a \in A(x)$ 但し ∂D は D の boundary を表す.

B2. 任意の $D \in \mathcal{B}_S$ に対して. $Q(D \mid x, a)$ は $(x, a) \in S \times A$ の連続関数。

次の Lemma の証明は [6] の Lemma 3.4 および，前節の定理 2.2 と Lemma 2.3 に含まれている。

Lemma 3.1

$\Gamma(G)\neq\phi$ なる $G\in\mathcal{B}_S$ に対して，次の (i)～(iii) を満たす定常政策 $\bar{f}^{(\infty)}$，$\gamma(C)>0$ なる $C\in\mathcal{B}_G$ および，$v\in B_u(\gamma(C)\cup C)$ が存在する：

(i) $\gamma(x,\bar{f}^{(\infty)})=\gamma^*(G)\quad {}^\forall x\in\gamma(C)\cup C$.

(ii) $Q(\gamma(C)\cup C\mid x,\bar{f}(x))=1\quad {}^\forall x\in\gamma(C)\cup C$.

(iii) v は C 上で一様有界で次の 2 つの関係式を満す：

$$v(x)+\gamma^*(G)=c(x,\bar{f}(x))+\int_{\gamma(C)\cup C}v(x')Q(dx'\mid x,\bar{f}(x))$$
$$\text{for all } x\in\gamma(C)\cup C$$
$$\lim_{T\to\infty}E^x_{\bar{f}^{(\infty)}}(v(X_T))/T=0.$$

定義

任意の $D\in\mathcal{B}_S$ に対して，$A(x,D):=\{a\in A(x)\mid Q(D\mid x,a)=1\}$

以上の準備のもとで次の定理を得る。証明は [7] を参照のこと。

定理 3.1 仮定 B のもとで，次の (i) (ii) を満たす S の可測分割と $v\in B_u(S)$ が存在する：

$S=S_1\cup S_2\cup\cdots\cup S_r\cup F$，$F\in\mathcal{B}_S$ $S_i\in\mathcal{B}_S$ $S_i\cap S_j=\phi$ ($i\neq j$)

(i) 各 i ($1\leq i\leq r$) に対して，次の最適不等式が成立する。

$v(x) + \gamma^*(S_i^*) \geq \inf_{a \in A(x,S_i)} U(x,a,v)$ $x \in S_i$, 但し $S_i^* = \bigcup_{j \geq i}^{r} S_j$.

(ii) 下上では，次の最適方程式が成り立つ．

$$v(x) = \inf_{a \in A(x)} \{ \sum_{j=1}^{r} \gamma^*(S_j^*) Q(S_i | x,a) + \int v(x') Q(dx' | x,a) \}$$

Reference.

[1] Bertsekas, D.P. and Shreve, S.D. (1978). Stochastic Optimal Control - The Discrete Time Case, Academic Press.

[2] Borkar, V.S. (1983). Controlled Markov chains and stochastic networks. Siam J. Control and Optimization. 21 652-666.

[3] ————————. (1984). On minimum cost per unit time control of Markov chains. Siam J. Control and Optimzation. 22 965-978.

[4] Doob, J.L. (1953). Stochastic Processes, Wiley, New York.

[5] Kurano, M. (1989). The existence of a minimum pair of state and policy for Markov decision processes under the hypothesis of Doeblin. Siam J. Control and Optimization. 27 296-307.

[6] ————————. (1989) Average cost Markov decision processes under the hypothesis of Doeblin. Technical Reports of Mathematical Sciences, Chiba University. Vol.4 (1988) No.9.
To appear in Annals of Operations Research.

[7] ————————. (1990). Functional characterization for average cost Markov decision processes with Doeblin's conditions. Technical Reports of Mathematical Sciences, Chiba University, Vol.6 (1990) No.1. To appear in Comp.& Math. Appl..

[8] Ross, S.M. (1970). Applied Probability Models with Optimization Applications. Holden-Say, San Francisco.

[9] Tweedie, R.L. (1976). Criteria for classifying general Markov chains. Adv. Appl. Prob., 8, 737-771.