数理解析研究所講究録
726巻 1990年 57-66

Average Cost Markov Decision Processes under the Hypothesis of Doeblin

千葉大・教育　蔵野正美（Masami Kurano）

§1. はじめに

一般の状態および行動空間をもつ平均コスト基準のマルコフ決定過程の研究は，多くの研究者（たとえば，Ross[8], Tijms[9], Kurano[5]を参照）によってなされた。そこでは，任意の政策によって生成されるマルコフ過程は単一のエルゴディック集合からなり，かつ適当な再帰性を仮定しての議論であった。最近，kurano[6]は，任意のランダム定常政策（randomized stationary policy）によって生成されるマルコフ過程にDoeblin条件のもとでいくつかのエルゴディック集合と過渡的状態の集合が許されるより一般的な場合について考察した。そして，平均期待コストをすべての初期分布と政策のもとで最小にする初期状態と定常政策の対の存在定理を与えた。証明には，Borkar[2,3]の経験分布による方法とDoeblin条件を満たすマルコフ過程の諸定理が用いられている。

本報告は，Kurano[6]で議論された問題を再度取り扱い，いくつかの新しい結果を証明する。それは，ある適当な条件のもとでの最適定常政策の存在定理を与えるものである。

Wijngaard[10]はDoeblin条件と同値な作用素の準コンパクト性の仮定のもとで，位相的な接近法により同じ問題を考察しているが，本報告で用いられる方法は彼の方法とは異なるものである。また，本報告はKurano[7]の内容を手直しして，かつ加筆したものである。

§2. 定式化

ある　可分な距離空間のボレル部分集合を単にボレル集合とよぶ。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 Process, MDP）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)を満たす。

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

(b) 各 $D\in\mathcal{B}_S$ に対して，$Q(D|\cdot)$ 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 $Q(\cdot|x_n,a_n)$ は $Q(\cdot|x,a)$ 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

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

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，$T(A|S)$ を $\mathcal{B}_A\times S$ 上の確率核 Φ で，すべての $x\in S$ で，$\Phi(A(x)|x)=1$ 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 $\Phi\in T(A|B)$ 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

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

が成り立つときをいう。この場合，π を単に $\Phi^{(\infty)}$ 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関数 $u : D \to A$ の全体を $B(D \to A)$ で表す。

ランダム定常政策 $\Phi^{(\infty)}$ が定常であるとは，$f \in B(S \to A)$ が存在して，$\Phi(\{f(x)\} \mid x) = 1$, $x \in S$ 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

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

$$\mathrm{Prob}(X_{t+1} \in D_2 \mid H_{t-1}, \Delta_{t-1}, X_t = x, \Delta_t = a) = Q(D_2 \mid x, a) \quad (t \geq 0)$$

このとき，任意の政策 $\pi \in \Pi$ と初期分布 $\nu \in P(S)$ 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，$P(D)$ 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 $\nu \in P(S)$ に対して，

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

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．

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

$$\psi^* := \inf_{\nu\in P(S)} \psi(\nu)$$

任意の $D\in\mathcal{B}_S$ に対して，

$$\psi(x,\pi^*) \leq \psi(x,\pi),\quad x\in D,\ \pi\in\Pi$$

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。

S の任意の真部分集合 $D\in\mathcal{B}_S$ に対して，

$$Q(D|x,a) = 1,\quad x\in D,\ a\in A(x)$$

が成り立つとき，D は部分マルコフ決定過程（sub-MDP）を構成するという。

§3. では，状態空間 S の任意のボレル真部分集合は，sub-MDP をつくらない場合についての最適定常政策の存在定理を与える。§4 では，いくつかの sub-MDPs が存在する場合について考察する。

§3. 存在定理（1）

この節は，Doeblin [4] の仮説を用いて，状態空間 S のどんなボレル真部分集合も sub-MDP を構成しない場合についての最適政策の特徴づけと存在定理を与える。

任意の $\Phi\in T(A|S)$ に対して，t 期の推移確率 $Q^{(t)}$ を次で定

義する：

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

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

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$\Phi\in\Gamma(AIS)$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，$Q^{(\ell)}(D|x,\Phi)\leq 1-\varepsilon$がすべての$x\in S$に対して成り立つ。

次のLemmaはすでに証明されている。

Lemma 3.1. ([6]. 定理2.2)

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

(i) $\bar{f}^{(\infty)}$はCにおいて最適である，

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

(iii) すべての$x\in C$において，$\psi^* = \psi(x,\bar{f}^{(\infty)})$.

次の連続性の条件Dが必要で，以後はこれを仮定する。

条件D. 次のD1－D2が成り立つ：

D1. 任意の$G\in\mathcal{B}_S$に対して，$Q(G|x,a)$は(x,a)の

連続関数である。

D2. 集合値関数 $A(\cdot)$ は 下半連続である。すなわち，

$x_n \to x$ なる点列 $\{x_n\}$ と $a \in A(x)$ に対して

$a_n \to a$, $a_n \in A(x_n)$ なる $\{a_n\}$ が存在する。

最初の存在定理は次の仮定のもとで証明される。

仮定A

任意の $D \in \mathcal{B}_S$ $(D \neq S, \gamma(D) > 0)$ に対して

$Q(S-D \mid x, a) > 0$ なる $x \in D$, $a \in A(x)$ が存在する。

定理3.1

仮定Aが成立するとき，次を満たす最適定常政策 $f^{(\omega)}$ が存在する：

すべての $x \in S$ に対して

$$\psi(x, f^{(\omega)}) = \psi^*$$

略証

仮定Aを使えば，Lemma 3.1 の集合 C は，γ - 測度が ε より大きい任意の集合から到達可能で，あたかも吸収状態のようにみなすことができる。この考えを用いて定理は証明される。詳しくは，[7] を参照のこと。 ■

§4. 存在定理(2)

この節では，§3の仮定Aが成立しない場合について，

定理3.1を拡張する。

次の仮定を必要とする。

仮定B

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

$Q(\partial D|x,a)=0$ $(\forall x\in S,\ \forall a\in A(x))$ が成立する。

但し，∂D は D の境界を表す。γ,ε は Doeblin の条件の中に表われている。

任意の $D\in\mathcal{B}_S$ に対して，次を定義する。

$$\Gamma(D):=\{(\nu,\pi)\in P(D)\times\Pi\mid P_\pi^\nu(X_t\in D \text{ for all } t\geq 0)=1\}$$

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

但し，$\Gamma(D)=\phi$ ならば $\psi^*(D)=\infty$.

次の Lemma が得られる。

Lemma 4.1　仮定Bが成り立つとする。そのとき，

$\gamma(G)>\varepsilon$, $\Gamma(G)\neq\phi$ なる任意の $G\in\mathcal{B}_S$ に対して，

次の(i)(ii)を満たす $C\in\mathcal{B}_S$ $(\gamma(C)>0)$ と定常政策 $f^{(\infty)}$ が存在する：

(i)　すべての $x\in C$ に対して，$\psi(x,f^{(\infty)})=\psi^*(G)$

(ii)　すべての $x\in C$ に対して，$Q(C|x,f(x))=1$.

略証

G の閉包を $\bar{G}$ とすると，$\bar{G}$ はコンパクトであるから，[6]の定理2.2の証明で用いた方法が使える。

Doeblin条件を満たすマルコフ過程の性質を用いるが，詳細な証明については [7] を参照のこと ■

Lemma 4.1 を用いて，次の定理を得る。証明は[7]に表われている。

定理4.1 仮定Bが成立するとする。 このとき，

Sの有限分割： $S = S_1 \cup S_2 \cup \cdots \cup S_r \cup F$,

$S_i \in \mathcal{B}_S$, $F \in \mathcal{B}_S$ $S_i \cap S_j = \phi$ ($i \neq j$) $S_i \cap F = \phi$

と定常政策 $f^{(\infty)}$ が存在して，次が成り立つ。

(i) $\gamma(S_i) > \varepsilon$ かつ $f^{(\infty)}$ は S_1 で最適である。

(ii) 各 i ($1 \leq i \leq r$) と任意の $x \in S_i$ に対して，$Q(S_i | x, f(x)) = 1$

$\gamma(x, f^{(\infty)}) = \gamma^*(S_i^*)$ が成り立つ。但し $S_i^* = \bigcup_{j \neq i}^{r} S_j \cup F$

(iii) すべての $x \in F$ と $\pi \in \Pi$ に対して，$P_\pi^x(T_{S-F} < \infty) = 1$,

$\gamma(x, f^{(\infty)}) \leq \sum_{j=1}^{r} \gamma_j^*(S_j^*) P_\pi^x(X_{T_{S-F}} \in S_j)$ が成り立つ。

注意

定理3.1および定理4.1により，最適定常政策が存在するための十分条件が与えられた。この定理の証明においては，最適方程式が使われなかった。Doeblin条件を満たす平均コスト基準のMDPに対する最適方程式については，次の機会に論究するつもりである。

[参考文献]

[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 Optimization*. 22 965-978.

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

[5] Kurano, M. (1986). Markov decision processes with a Borel measurable cost function-the average case. *Math. Oper. Res.* 11 309-320.

[6] ————————.(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.

[7] ————————.(1989) Average Cost Markov Decision Processes under the Hypothesis of Doeblin. Technical Reports of Mathematical Sciences, Chiba University. Vol.4(1988) No.9. Submitted to Annals of Operations Research.

[8] Ross, S.M. (1968). Arbitrary state Markovian decision processes. *Ann. Math. Statist.* 39 2118-2122.

[9] Tijms, H.C. (1975). On dynamic programming with arbitrary state space, compact action space and the average return as criterion *Report* BW 55/75, Math. Centre, Amsterdam.

[10]Wijngaard, J. (1977). Stationary Markovian decision problems and perturbation theory of quasi-compact linear operators. Math. Oper. Res. 2 91-102.