マルコフ過程の漸近最適な学習制御†

佐藤 光男*・阿部 健一**・竹田 宏*

Asymptotically Optimal Learning Control of Markov Processes

Mitsuo SATO*, Ken'ichi ABE** and Hiroshi TAKEDA*

This paper is concerned with a learning control problem of finite Markov chains with unknown transition probabilities. The learning control consists of two different tasks of estimating the unknown probabilities and controlling the system. The control objective is here to maximize the total expected reward discounted in time, which is, as well known, accomplished by consecutive application of an optimal policy. But, of course, the optimal policy cannot be found unless the transition probabilities are previously known. Accordingly, the ultimate goal of the learning is to ensure the convergence to the optimal policy, in other words; asymptotic optimality.

Generally speaking, however, it is very difficult to achieve the asymptotic optimality, because there is an apparent conflict between estimation and control. It is pointed out that some of the learning control schemes which have been presented for this problem do not ensure the convergence to the optimal policy. The others dissolve that difficulty in their own ways and achieve the asymptotic optimality. But these are not so practical by reason that in their implementation computational procedures for determining the control policy are very complicated.

This paper presents a learning control scheme such that, in addition to its being asymptotically optimal, practical procedures involved by its policy determination, which is reduced to maximization of a performance criterion, are much simpler. This criterion is designed with consideration of the conflict

between estimation and control. For this scheme it is shown that the control policy converges to the optimal in the sense of a frequency of application.

Key Words: Markovian decision process, transition probability, learning control, asymptotic optimality

1. ま え が き

状態数が有限のマルコフ決定過程は、確率制御問題に対する重要なモデルの一つとして、工学から経済の分野にわたる幅広い応用範囲を持つことが知られている¹⁾。ただし、その現実的な応用においては、ほとんどの場合、状態遷移確率は未知であると考えられる。

したがって、未知確率の推定を行いながら制御の目的を追求するというマルコフ過程の学習制御問題は、理論面からだけでなく応用面から見ても重要な意味をもつ。制御の目的とは、ここでは過程においてもたらされる総期待利得を最大ならしめることであり、周知のようにこれは最適政策を継続的に適用することにより達成される。ところが遷移確率が既知でなければ最適政策を見出すことはできないのであるから、当然ながら、学習の最終的な目標は、最適政策への収束を保証すること、言い換えれば漸近最適性である。しかしながら、一般に、未知確率の推定とシステムの制御との間には明白な矛盾が存在し、そのために漸近最適性を達成するのは非常に困難である。

マルコフ過程の学習制御問題に対しては、これまでに種々の学習制御方式が提案されている^{2)~6)}。このうち、Martin²⁾による方式は、ベイズ決定理論に基づいて構成されたものである。ところが、この方式は、政策の決定に要する計算が膨大であるばかりでなく、推定と制御の矛盾が完全には解決されていないために、最適でない政策に収束する可能性のあることが指摘されている⁷⁾。また、Vorkar ら³⁾による方式は、システムが、あるきわめて限定的な可同定性の条件を満たしていなければ漸近最適とはならない。一方、

† 第 23 回計測自動制御学会学術講演会で発表(昭 59・7)

* 東北大学工学部 仙台市荒巻字青葉

** 豊橋技術科学大学 豊橋市天伯町字雲雀ヶ丘 1-1

* Faculty of Engineering, Tohoku University, Sendai

“ Toyohashi University of Technology, Toyohashi
(Received September 18, 1984)
(Revised January 31, 1985)

El-Fattah⁴⁾, Sato⁵⁾, および Kumar⁶⁾らの方式は、いずれも、独自の方法で推定と制御の矛盾を解決し、比較的弱い条件のもとで漸近最適性を達成している。しかし、これら漸近最適な方式も、政策を決定するための計算上の操作が非常に複雑であるという点で、実際的ではない。

本論文では、これに対して、可同定性を仮定しないなど、やはり比較的弱い条件のもとで漸近最適であるだけでなく、政策決定の操作およびそれに要する計算がはるかに簡単であるような方式を提案する。この方式では、推定の目的と制御の目的とを簡潔な形で反映した評価基準が政策決定のために導入される。この評価基準は両目的がともに達成されるように工夫される。本論文では、政策の適用頻度を表わす指数によってこの方式の漸近特性を調べ、この方式がこの頻度の収束の意味において漸近最適であることを示す。なお、この方式は、 ε -最適の意味での漸近最適性を確立した佐藤ら⁸⁾の方式を拡張したものである。

2. マルコフ過程の学習制御問題

まず、本論文で扱うマルコフ決定モデルについて述べる。各時刻 $t=0, 1, 2, \dots$ において、システムは、 N 個の状態 $1, \dots, N$ のうちどれか一つの状態にあることが観測されるとする。システムが状態にあると観測されたとき、それに応じて K_i 個の決定 $1, \dots, K_i$ が与えられ、その中から一つの決定が選ばれて用いられる。ここに、上記 N および K_i はすべて有限とする。いま、状態 i において決定 k が用いられたとする。するとシステムはある一定の確率 p_{ij}^k で状態 j に遷移する ($j=1, \dots, N$)。これに伴い時刻が単位時間だけ更新され、そこで r_{ij}^k なる利得が与えられる ($|r_{ij}^k| < \infty$)。遷移確率 p_{ij}^k については、あきらかに、すべての i, j および k に対して次式が成立する。

$$0 \leq p_{ij}^k \leq 1, \quad \sum_j p_{ij}^k = 1$$

各状態 i において選ぶべき決定 u_i を指定した規則を政策と呼び、 N 次元ベクトル $U=(u_1, \dots, u_N)$ で表わす。この過程においては政策の選び方によって利得の増減を期待値の上で操作することができる。簡単に言えば、政策の選び方によってこの過程を制御することができる。このような制御過程をマルコフ決定過程と言う。制御の目的は今後得られるであろう総期待利得を最大ならしめることにある。ただし、将来の利得についてはその価値を割引いて考えるものとし、単位時間当りの割引き率 β を設定する ($0 < \beta < 1$)。周知のように、この目的は、最適政策と呼ばれる政策を継続的に適用することによって達成される。初期状態 i に

おいて今後政策 U を継続的に適用するときの総期待割引き利得を v_i^U で表わすとき、 i に関して一様に v_i^U を最大ならしめる $U=U^*=(u_1^*, \dots, u_N^*)$ が存在する。これが最適政策である(注1)。 $v_1^U, \dots, v_N^U$ は連立方程式

$$v_i^U = \sum_j p_{ij}^{u_i} r_{ij}^{u_i} + \beta \sum_j p_{ij}^{u_i} v_j^U \quad (i=1, \dots, N) \quad (1)$$

のただ1組の有界な解で与えられる。

さて、本論文ではすべての遷移確率 p_{ij}^k が未知であるとする。上述からあきらかなように、このような条件下では最適政策 U^* を事前に見い出すことができない。そこで、未知なる p_{ij}^k の推定を行いながら制御の目的を追求するという学習制御問題が設定される。当然ながら、学習の最終的な目標は、各時刻で選択する政策を確実に U^* に漸近させること、すなわち漸近最適性である。以下、漸近最適性を目標とした学習制御方式の構成を試みる。ここでつぎの仮定をおく。

[仮定 1]

$$p_{ij}^k > 0 \quad (k=1, \dots, K_i; i, j=1, \dots, N)$$

[仮定 2]

$$v_i^{U^*} > v_i^U \quad (U \neq U^*; i=1, \dots, N)$$

仮定 1 は、いかなる制御則のもとでもすべての状態が無限回生起することを保証するもので、これはすべての遷移確率を推定するのに必要である。一方、仮定 2 は最適政策が唯一であることを意味するが、これは議論を簡潔にするための便宜的なものである。

3. 推定と制御の矛盾

$\omega_t=(i_0, i_1, \dots, i_t)$ を t 回の状態遷移の標本とする。ここに i_t は時刻 t において生起した状態である。また、時刻 t までの間に決定 k のもとで状態 i から状態 j に遷移した回数を $n_{ij}^k(t)$ で表わし、 $n_{ij}^k(t)=\sum_j n_{ij}^k(t)$ および $n_i(t)=\sum_k n_{ij}^k(t)$ とする。本論文では、各時刻 t において用いられる決定 k_t が ω_t に依存して唯一的に定まる制御則のみを考える。すると、 ω_t に対する尤度は

$$L_t = \prod_{i,k,j} (p_{ij}^k)^{n_{ij}^k(t)}$$

と表わされる。これにより p_{ij}^k の最尤推定値 $\hat{p}_{ij}^k(t)$ が

$$\hat{p}_{ij}^k(t) = n_{ij}^k(t) / n_i(t)$$

と求められる。ただし、 $\hat{p}_{ij}^k(t)$ は $n_i(t) \neq 0$ なる i ,

(注 1) β はもともと総期待利得が確実に有限な値をとるために導入されたものであり、 β の値が1に近いほどそのときの最適政策は厳密な意味での総期待利得を最大ならしめる政策に近くなる。

k に対してのみ定義される。状態遷移の標本 $\omega = (i_0, i_1, \dots)$ から、状態 i において決定 k が用いられたという条件のもとで生じた状態列を取り出して考えてみると、これは制御則にかかわらず状態 $1, \dots, N$ がそれぞれ一定確率 $p_{i1}^k, \dots, p_{iN}^k$ で生起する独立過程を形成している。そこで、この過程に大数の強法則を適用することにより、推定値 $\hat{p}_{ij}^k(t)$ の一貫性に関して下記の補題が導かれる。なお、上記に定義した $i, k, n_{ij}^k(t)$ およびその他の諸量はいずれも ω に依存した確率変数であり、よってこれらに関する確率的表現もすべて ω に依存する。しかし、本論文では、表示の簡単のために、以下のすべての確率的表現は与えられた制御則のもとで出現するであろう1標本 ω に限定されているものとし、それらの表現の中にその ω を明示しないものとする。

【補題 1】 いかなる制御則のもとでも、 $\lim_{t \rightarrow \infty} n_{ij}^k(t) = \infty$ となるようなあらゆる i, k に対して

$$\lim_{t \rightarrow \infty} \hat{p}_{ij}^k(t) = p_{ij}^k \quad (j=1, \dots, N)$$

が成立する。

補題1からあきらかなように、すべての i, k に対して $\lim_{t \rightarrow \infty} n_{ij}^k(t) = \infty$ とならなければ、すべての p_{ij}^k の推定値が真値に収束するという保証はない。言い換えれば、推定の目的が完全に達成されるには、各状態においてあらゆる決定 $1, \dots, K_i$ を無限回用いるという、いわば探索的な行動をとらなければならない。これに対して、制御の目的は、各状態において最適な決定のみを用いるという指向的な行動により達成されるものである。このように推定と制御の間には明白な矛盾が存在する。ところが制御の目的の達成には推定の目的の達成が必要であるから、結局、制御の目的を厳密に達成することは不可能である。そして、実現可能な漸近最適性とは、上記のような探索的行動をとり続けるという制約のなかで、制御の目的を漸近的に達成するという形をとる。

前述のように、文献^{4)~6)}の学習制御方式は、それぞれ独自の方法で推定と制御の矛盾を解決し、漸近最適性を確立している。ただし、文献4)と文献5)の方式では政策の選択が確率的に行われる。このように選択の仕方を確率的なものにまで拡張すると、確率の属性を利用できることにより、漸近最適性の達成がはるかに容易になる。しかしながら、制御を実施するという立場から考えると、確率的な政策選択は、どの政策が実際に適用されるかが事前にはわからないという点で、現実的でない。また、確率的な選択を指示通りに実行することは、理想的な乱数を発生させるという付随的

な操作を伴う。そこで、もし確定的な政策選択によっても同様の学習特性が得られるならば、その方が実際上望ましいことは言うまでもない。

文献⁶⁾の方式は確定的な政策選択によって漸近最適性を確立するものである。この方式の特徴は、未知パラメータの推定値を求める際に、最尤推定値のみを選ぶのではなく、パラメータ空間のある範囲の中から探索的に推定値を選ぶという点にある。しかしながら、その推定値を求めることは実際問題としてきわめて困難である。なお、文献4)と文献5)の方式も、政策を決定するための計算上の操作が非常に複雑である。以上のような状況にかんがみ、本論文で学習制御方式を開発するに当たっては、確定的な政策選択によって漸近最適性の確立を図るだけでなく、運用上の容易さも合わせて追求する。

4. 学習制御方式

本論文で提案される学習制御方式は、時刻において適用すべき政策が、そのときのシステムの状態 i とするとき、評価基準

$$w_i^q(t) = \hat{v}_i^q(t) + \frac{c_i^{u_i}(t)}{n_i^{u_i}(t)} \quad (2)$$

を最大ならしめる U で与えられるというものである。ここに、 $\hat{v}_i^q(t)$ は $p_{ij}^{u_i} = \hat{p}_{ij}^{u_i}(t)$ とおいたときの(1)式の有界な解であり、 $c_i^{u_i}(t)$ は以下に示すように逐次的に定められる実数である(初期値 $c_i^{u_i}(0) = 0$ とする)。 $\{\theta(n)\}_{n=1}^{\infty}$ をつぎの性質を満たす任意の正数列とする。

$$\lim_{n \rightarrow \infty} \theta(n) = 0 \quad (3)$$

$$\sum_{n=1}^{\infty} \theta(n) = \infty \quad (4)$$

いま、時刻 t である政策が選ばれた結果、状態 i において決定 d が用いられたとする。このとき

$$\begin{cases} c_i^{u_i}(t+1) = c_i^{u_i}(t) + \theta(n_i(t+1) - n_i^{u_i}(t+1)) & (u_i \neq d) \\ c_i^d(t+1) = c_i^d(t) \end{cases}$$

と定める。また、このとき、すべての状態 $j \neq i$ に対して $c_j^{u_j}(t+1) = c_j^{u_j}(t)$ と定める。

(2)式右辺第1項は、すべての遷移確率が既知のときの評価基準である(1)式から導かれるものである。この項だけが評価基準として与えられると、それによって決定される政策は、未知確率の推定値がかりにその真値であったときの最適政策となる。この意味で、第1項は単純に制御の目的を反映したものとと言える。しかしながら、この項には推定を促進させるための要素は含まれていない。この不備を補うために“推定の進み具合”あるいは“推定の遅れ具合”を数量化し

た項を第1項と同格の形で評価基準の中にとり入れ、推定を促進させようというのが、(2)式右辺第2項の導入の意図である。この項は、その分子 $c_i^{u_i}(t)$ の改変方法からあきらかなように、その増加によって未知確率 $p_{ij}^{u_i}$ の推定の遅れを、またその減少によって推定の進みを表わすものである。この効果により、結局、どの未知確率の推定も進むことが予想される。このように、第2項は推定の目的を反映したものである。ただし、推定が制御に対比して過不足なく行われるには、すなわち漸近最適性が達成されるには、第2項は、第1項がいかなる有限値をとろうともそれに対して増加しすぎてもまた減少しすぎてもいけない。この理由により、 $c_i^{u_i}(t)$ の増加性には(3)式と(4)式の制限が必要となる。

本方式のいま一つの特徴として、実際の政策決定の容易さが挙げられる。これは(2)式で与えられる評価基準の算出が容易であることによる。すなわち、各 U に対して $w_i^U(t)$ の値を実際に求めようとする場合、多少手間のかかる操作と言え、それは $\hat{v}_i^U(t)$ の値を求めること、言い換えれば(1)式の形の方程式を解くことだけである。従来の漸近最適な方式における実際的な政策決定の非常に困難さを考えれば、本方式はこの面ではるかに改善されていることがわかる。

5. 漸近特性

本章では政策の適用頻度によって本方式の漸近特性を調べる。ここでその頻度を意味する指数 $f_i^*(t) = n_i^*(t)/n_i(t)$ および $f^*(t) = (1/t) \sum_{i=1}^{K_i} n_i^{u_i^*}(t)$ を定義する。あきらかに $f_i^*(t)$ および $f^*(t)$ はすべて0と1の間の値をとり、またあらゆる t, i に対して $\sum_{k=1}^{K_i} f_i^k(t) = 1$ が成立する。これらの指数によって漸近最適性は、

$$\lim_{t \rightarrow \infty} f_i^{u_i^*}(t) = 1 \quad (i=1, \dots, N) \quad (5)$$

あるいはより端的に

$$\lim_{t \rightarrow \infty} f^*(t) = 1 \quad (6)$$

なる式で表現される。(5)式が成り立つならば(6)式も成り立つことが容易に示される。以下、本方式がこれらの成立の意味で漸近最適であることを示す。なお、文献⁸⁾の方式がそうであるように、前記の $c_i^{u_i}(t)$ の逐次の決定においてその増加項 $\theta(\cdot)$ を一定値 $\theta > 0$ に保つものとする、そのときには ε -最適の意味での漸近最適性を表わす $\lim_{\theta \rightarrow 0} \lim_{t \rightarrow \infty} f_i^{u_i^*}(t) = 1$ ($i=1, \dots, N$) あるいはまた $\lim_{\theta \rightarrow 0} \lim_{t \rightarrow \infty} f^*(t) = 1$ なる式が成立する。ただし、以下に示す漸近最適性の証明はこの ε -最適性の証明と大幅に異なるものである。

各 i, k に対して、 $u_i = k$ なる政策 U に関する v_i^U , $\hat{v}_i^U(t)$ の最大値をそれぞれ v_i^k , $\hat{v}_i^k(t)$ で表わす。すると、本方式は、(2)式の最大化による政策の選択から

$$w_i^k(t) = \hat{v}_i^k(t) + \frac{c_i^k(t)}{n_i^k(t)} \quad (7)$$

の最大化による決定の選択に帰着される。言うまでもなく、(7)式の評価基準は $n_i^k(t) = 0$ の場合には適用できないので、初期の間はこれ以外の基準によらなければならない。そこで、たとえば、その間は各状態 i において決定 $1, \dots, K_i$ を循環的に用いるものとする。ただし $c_i^k(t)$ だけは本方式にしたがってはじめて逐次的に定めるものとする、すべての t および i, k に対して

$$c_i^k(t) = \theta(1) + \theta(2) + \dots + \theta(n_i(t) - n_i^k(t)) \quad (8)$$

となる。よって、ここで $\bar{\theta}(n) = (1/n) \sum_{\tau=1}^n \theta(\tau)$ とすると、すべての t および i, k に対して

$$c_i^k(t) = \bar{\theta}(n_i(t) - n_i^k(t)) [n_i(t) - n_i^k(t)]$$

と表わすことができる。これにより(7)式は

$$w_i^k(t) = \hat{v}_i^k(t) + \bar{\theta}(n_i(t) - n_i^k(t)) \frac{1 - f_i^k(t)}{f_i^k(t)} \quad (9)$$

と変形される。まず、本方式の推定面における効果を下記に示す。

[補題 2] 仮定1のもとで、本方式はすべての i, k に対して $\lim_{t \rightarrow \infty} n_i^k(t) = \infty$ を保証する。

(証明) 仮定1によりすべての状態が無限回生起することが保証されるので、補題を証明するに当っては、任意の状態 i を選んでそれを固定して考えればよい。そこで、証明のために、ある決定 a が状態 i においてたかだか有限回しに用いられないと仮定する。この仮定により、決定 a がある有限の時刻 T 以後用いられることはない、あきらかに $\lim_{t \rightarrow \infty} f_i^a(t) = 0$ であり、また(4)式の効果により $\lim_{t \rightarrow \infty} w_i^a(t) = \infty$ である。状態 i において少くとも一つの決定は無限回用いられるので、その一つを b とすると、決定 b が時刻 T 以後用いられる時刻列 $\{t_m\}_{m=1}^{\infty} \subset \{t\}$ を定義することができる。この定義によりすべての m に対して $w_i^b(t_m) = \max_{1 \leq k \leq K_i} w_i^k(t_m) \geq w_i^a(t_m)$ であるので、 $\lim_{t \rightarrow \infty} w_i^a(t) = \infty$ であることから、次式が成立する。

$$\lim_{m \rightarrow \infty} w_i^b(t_m) = \infty \quad (10)$$

(9)式において $\hat{v}_i^b(\cdot)$ および $\bar{\theta}(\cdot)$ がともに一様有界であることから、(10)式は次式を意味する。

$$\lim_{m \rightarrow \infty} f_i^b(t_m) = 0 \quad (11)$$

ふたたび $\{t_m\}$ の定義によれば、時刻 T 以後 $f_i^a(t)$ は時刻 $t = t_m + 1$ においてのみ ($m = 1, 2, \dots$) 増加するは

ずである。ところが、 $\lim_{m \rightarrow \infty} n_i(t_m) = \infty$ および (11) 式から、

$$\begin{aligned} f_i^b(t_m+1) &= \frac{n_i^b(t_m)+1}{n_i(t_m)+1} \\ &= \frac{n_i(t_m)}{n_i(t_m)+1} f_i^b(t_m) + \frac{1}{n_i(t_m)+1} \end{aligned}$$

もまた $m \rightarrow \infty$ のとき 0 に収束するので、よって $\lim_{t \rightarrow \infty} f_i^b(t) = 0$ である。このようにして、 $k=a, b$ に対して $\lim_{t \rightarrow \infty} f_i^k(t) = 0$ であることが示された。ところがすべての t に対して $\sum_{k=1}^{K_i} f_i^k(t) = 1$ であるので、これら以外に状態 i において無限回用いられる決定があるはずである。その一つを e とすると、決定 b の場合とまったく同様にして $\lim_{t \rightarrow \infty} f_i^e(t) = 0$ を得る。したがって、同様の操作をくり返すことにより、結局次式を得る。

$$\lim_{t \rightarrow \infty} f_i^k(t) = 0 \quad (k=1, \dots, K_i)$$

これはあきらかに $\sum_{k=1}^{K_i} f_i^k(t) = 1$ に矛盾する。よって i の任意性により補題は証明された。(証明終)

補題 1 と補題 2 から、本方式が推定の完遂を保証することが示される。そこで、つぎに、(7) 式で与えられる評価基準の漸近的ふるまいを調べることによって、本方式の制御面に対する学習効果を考察する。(7) 式右辺第 1 項については、補題 1 と補題 2 からすべての i, j および k に対して $\lim_{t \rightarrow \infty} \hat{p}_{ij}^k(t) = p_{ij}^k$ が成立するので、これにより

$$\lim_{t \rightarrow \infty} \hat{v}_i^k(t) = v_i^k \quad (k=1, \dots, K_i; i=1, \dots, N)$$

の成立すること、すなわち、任意の $\varepsilon > 0$ のもとで十分大きなすべての t に対して次式の成立することがわかる。

$$v_i^k - \varepsilon < \hat{v}_i^k(t) < v_i^k + \varepsilon \quad (k=1, \dots, K_i; i=1, \dots, N) \quad (12)$$

ところで、 $\theta(\cdot)$ が正であることと (8) 式より (7) 式右辺第 2 項はつねに正であるから、すべての t に対して $w_i^k(t) > \hat{v}_i^k(t)$ である ($k=1, \dots, K_i; i=1, \dots, N$)。よって、(12) 式より、十分大きなすべての t に対して次式が成立する。

$$w_i^k(t) > v_i^k - \varepsilon \quad (k=1, \dots, K_i; i=1, \dots, N) \quad (13)$$

下記の補題は (7) 式右辺第 2 項の漸近的ふるまいを示すものである。

【補題 3】 仮定 1 および仮定 2 のもとで、十分大きなすべての t に対して

$$w_i^k(t) - \hat{v}_i^k(t) = c_i^k(t)/n_i^k(t) > \delta \quad (k \neq u_i^*; i=1, \dots, N)$$

となるような $\delta > 0$ が存在する。

(証明) 定義により、仮定 2 のもとでは、すべての i および $k \neq u_i^*$ に対して $v_i^{u_i^*} > v_i^k$ であるから、次式が成り立つような $\delta_0 > 0$ が存在する。

$$v_i^{u_i^*} - v_i^k > \delta_0 \quad (k \neq u_i^*; i=1, \dots, N) \quad (14)$$

ここでまた任意の状態 i を選んでそれを固定して考える。 $a \neq u_i^*$ を状態 i において適用可能な一つの決定とし、 $\{t_l\}_{l=1}^{\infty} \subset \{t\}$ を決定 a が用いられる時刻列とする。すると、(13) 式から、任意の $\varepsilon > 0$ のもとで十分大きなすべての l に対して次式が成立する。

$$w_i^a(t_l) \geq w_i^{u_i^*}(t_l) > v_i^{u_i^*} - \varepsilon \quad (15)$$

(12) 式、(14) 式および (15) 式から、十分大きなすべての l に対して次式が成立する。

$$\begin{aligned} c_i^a(t_l)/n_i^a(t_l) &= w_i^a(t_l) - \hat{v}_i^a(t_l) > w_i^a(t_l) - v_i^a - \varepsilon \\ &> w_i^a(t_l) + v_i^{u_i^*} + \delta_0 - \varepsilon \\ &> \delta_0 - 2\varepsilon \end{aligned} \quad (16)$$

$\{t_l\}$ の定義により各 l に対して

$$\frac{c_i^a(t_l+1)}{n_i^a(t_l+1)} = \frac{c_i^a(t_l)}{n_i^a(t_l)+1} = \frac{n_i^a(t_l)}{n_i^a(t_l)+1} \frac{c_i^a(t_l)}{n_i^a(t_l)}$$

であるから、(16) 式より次式が成立する。

$$\frac{c_i^a(t_l+1)}{n_i^a(t_l+1)} > \frac{n_i^a(t_l)}{n_i^a(t_l)+1} (\delta_0 - 2\varepsilon) \quad (17)$$

(16) 式および (17) 式から、十分大きなすべての l に対して $c_i^a(t_l)/n_i^a(t_l) > \delta_1$ および $c_i^a(t_l+1)/n_i^a(t_l+1) > \delta_1$ となるような $\delta_1 > 0$ が存在する。ところが $c_i^a(t)/n_i^a(t)$ は区間 $t_l+1 \leq t < t_{l+1}$ において単調に増加するのであるから ($l=1, 2, \dots$)、上記の結果より、十分大きなすべての t に対して $c_i^a(t)/n_i^a(t) > \delta_1$ となることがわかる。したがって、状態数と決定数がともに有限であることから、すべての $i, k \neq u_i^*$ に対して $c_i^k(t)/n_i^k(t) > \delta$ となるような $\delta > 0$ が存在する。(証明終)

さて、以上の準備のもとに本方式の漸近最適性を示す。(9) 式よりすべての t に対して次式が成立する。

$$f_i^k(t) = \frac{\bar{\theta}(n_i(t) - n_i^k(t))}{w_i^k(t) - \hat{v}_i^k(t) + \bar{\theta}(n_i(t) - n_i^k(t))} \quad (k=1, \dots, K_i; i=1, \dots, N) \quad (18)$$

補題 2 からわかるようにすべての i, k に対して $\lim_{t \rightarrow \infty} [n_i(t) - n_i^k(t)] = \infty$ であるから、(3) 式の性質より次式が成立する。

$$\begin{aligned} \lim_{t \rightarrow \infty} \bar{\theta}(n_i(t) - n_i^k(t)) &= 0 \\ (k=1, \dots, K_i; i=1, \dots, N) \end{aligned} \quad (19)$$

そこで、補題 3 および (19) 式を (18) 式に適用すれば、次式を得る。

$$\lim_{t \rightarrow \infty} f_i^k(t) = 0 \quad (k \neq u_i^*; i=1, \dots, N)$$

これはとりも直さず (5) 式を意味し、よって (6) 式が示される。以上の結果をまとめてつぎの定理とする。

《定理》 仮定 1 および仮定 2 のもとで、本方式は $\lim_{t \rightarrow \infty} f^*(t) = 1$ の意味での漸近最適性を達成する。

6. 計算機シミュレーション

本方式の実験的な学習特性をあきらかにするために、ここでは、本方式を遷移確率および利得が **Table 1** に与えられるような 3 状態 2 決定のモデルに適用したときのシミュレーション結果を示す。このモデルに関しては政策は全部で 8 つある、割引き率を $\beta = 0.8$ として、Table 1 から、各政策に対する総期待割引き利得が Table 2 のように算出される。**Table 2** からわかるように、最適政策は $U^* = (1, 2, 1)$ である。したがって、本シミュレーションにおける本方式の学習特性は頻度指数 $f^*(t) = [n_1^*(t) + n_2^*(t) + n_3^*(t)]/t$ によって表わされる。また、数列 $\{\theta(n)\}$ の具体例として、つぎの 3 通りの場合を考えた。

(a) $\theta(n) = 1/n$ (b) $\theta(n) = 1/\sqrt{n}$

(c) $\theta(n) = 0.1$

前にも述べたように、上記(c)は、 ε -最適な方式に相当するものである。(a), (b), (c)の各場合に対応するシミュレーション結果を、 $f^*(t)$ の t に対する変化として表わし、それらをすべて **Fig. 1** に示す。

以上のほかにも、3つの各場合について、乱数の初

Table 1 Transition probabilities and rewards

Transition Probabilities			Rewards		
$p_{11}^1 = 0.2$	$p_{12}^1 = 0.3$	$p_{13}^1 = 0.5$	$r_{11}^1 = 1$	$r_{12}^1 = 2$	$r_{13}^1 = 3$
$p_{21}^1 = 0.4$	$p_{22}^1 = 0.2$	$p_{23}^1 = 0.4$	$r_{21}^1 = 1$	$r_{22}^1 = 3$	$r_{23}^1 = 2$
$p_{31}^1 = 0.6$	$p_{32}^1 = 0.3$	$p_{33}^1 = 0.1$	$r_{31}^1 = 2$	$r_{32}^1 = 1$	$r_{33}^1 = 3$
$p_{11}^2 = 0.1$	$p_{12}^2 = 0.7$	$p_{13}^2 = 0.2$	$r_{21}^2 = 2$	$r_{22}^2 = 3$	$r_{23}^2 = 1$
$p_{21}^2 = 0.5$	$p_{22}^2 = 0.1$	$p_{23}^2 = 0.4$	$r_{31}^2 = 3$	$r_{32}^2 = 1$	$r_{33}^2 = 2$
$p_{31}^2 = 0.2$	$p_{32}^2 = 0.2$	$p_{33}^2 = 0.6$	$r_{31}^3 = 3$	$r_{32}^3 = 2$	$r_{33}^3 = 1$

Table 2 Total expected discounted rewards corresponding to the policies

Policies	Total Expected Discounted Rewards		
$U = (1, 1, 1)$	$v_1^U = 11.18$	$v_2^U = 10.63$	$v_3^U = 11.36$
$U = (1, 1, 2)$	$v_1^U = 9.71$	$v_2^U = 9.44$	$v_3^U = 8.97$
$U = (1, 2, 1)$	$v_1^U = 11.93$	$v_2^U = 12.21$	$v_3^U = 11.99$
$U = (1, 2, 2)$	$v_1^U = 10.57$	$v_2^U = 11.15$	$v_3^U = 9.76$
$U = (2, 1, 1)$	$v_1^U = 9.85$	$v_2^U = 9.69$	$v_3^U = 10.47$
$U = (2, 1, 2)$	$v_1^U = 8.67$	$v_2^U = 8.73$	$v_3^U = 8.43$
$U = (2, 2, 1)$	$v_1^U = 10.65$	$v_2^U = 11.68$	$v_3^U = 11.78$
$U = (2, 2, 2)$	$v_1^U = 9.60$	$v_2^U = 10.83$	$v_3^U = 9.37$

Fig. 1 Variations of $f^*(t)$ with respect to t in the cases of $\theta(n) = 1/n$, $\theta(n) = 1/\sqrt{n}$, and $\theta(n) = 0.1$

期値を変えて同じ実験を何回かくり返した。その結果、(b)と(c)についてはすべて、ここに示したそれぞれの結果とほとんど同じ特性が得られた。また、(c)の場合は $f^*(t)$ が飽和して ε -最適性の特徴を示しているのに対して、(b)の場合は $f^*(t)$ が次第に増加していきついには(c)の特性を上まわることが確認された。ところで、(a)についていくつかの実験例において、初期に誤動作が生じたために結果的にきわめて遅い収束に帰することが観測された。これはあきらかに $\theta(n) = 1/n$ の減少が急速すぎることにによる。 $\theta(n) = 1/n$ は、(3)式と(4)式を満たす正数列のうち最も代表的なものであるが、以上の例からみると、初期の減少が急速すぎることを除けばきわめて満足のいく結果を与えている。そこで、誤動作が生じやすい初期の間は、 $\theta(n)$ を、 $1/\sqrt{n}$ 以下の減少速度をもつ列もしくは正定数の列にとり、その後は $1/n$ 程度にとるのが望ましいであろう。

7. むすび

本論文では、マルコフ過程の学習制御問題に対して、1方式を提案した。本方式は、推定の目的と制御の目的とを簡潔な形で反映した評価基準を最大化することにより政策の決定を行うものである。本論文では、本方式が放策の適用頻度の収束という意味において漸近最適であることを示した。

実際面から見た場合、従来のほとんどの方程式が、政策の決定に伴う計算上の操作が複雑あるいは困難なためにシミュレーションの実施も困難であるのに対して、本方式はそれが十分に容易である。このように、本方式は、漸近最適性の確立および運用上の簡便さという点で、従来の方式よりもすぐれている。

参考文献

- 1) H. J. Kushner: Introduction to Stochastic Control,

- Holt, Rinehart and Winston (1971)
- 2) J. Martin: Bayesian Decision Problems and Markov Chains, Wiley (1967)
 - 3) V. Borkar and P. Varaiya: Adaptive Control of Finite Markov Chains, I: Finite Parameter Set, IEEE Trans. on Automatic Control, **AC-24-6**, 953/957 (1979)
 - 4) Y. M. El-Fattah: Recursive Algorithms for Adaptive Control of Finite Markov Chains, IEEE Trans. on Systems, Man, and Cybernetics, **SMC-11-2**, 135/144 (1981)
 - 5) M. Sato, K. Abe, and H. Takeda: Learning Control of Finite Markov Chains with Unknown Transition Probabilities, IEEE Trans. on Automatic Control, **AC-27-2**, 502/505 (1982)
 - 6) P. R. Kumar and W. Lin: Optimal Adaptive Controllers for Unknown Markov Chains, IEEE Trans. on Automatic Control, **AC-27-4**, 765/774 (1982)
 - 7) 佐藤, 阿部, 竹田: ベイズ的手法を用いた未知パラメータを含むマルコフ決定過程の漸近的性質, 電子通信学会論文誌D, **J 61-D-1**, 1/8 (1978)
 - 8) 佐藤, 阿部, 竹田: 確定的な手法によるマルコフ過程の学習制御, 電気学会論文誌C, **102-C-10**, 245/250 (1982)
-