マルコフ過程の有限時間学習制御[†]

佐藤 光 男*・竹 田 宏*

Finite-Time Learning Control of Markov Processes

Mitsuo SATO* and Hiroshi TAKEDA*

Many studies have been devoted to the Markovian decision problem in which the transition probabilities are unknown, and various learning schemes have been presented for it. Most of these works are concerned with seeking convergence to optimal control inputs under presupposition that the time length is infinite. Considering the practical phase, however, we can say that a finite-time problem is important as well as the above mentioned infinite-time problem. In a finite-time problem, the goal of learning is not assurance of the optimal convergence but minimization of the total loss in the whole process.

Previously, for the finite-time two-armed bandit problem, which is the simplest model of learning control problems, we presented a learning scheme that was designed from introducing an idea of deviding the whole process into the first part of estimating the unknown parameters and the subsequent part of controlling the system.

In this paper, we design a finite-time learning scheme with the same idea for a Markovian decision model, and present a stopping rule of the estimation. The Bayesian method is used in the estimation part. As a performance criterion for the stopping, the maximum posterior probability is considered, and then a loss function defined on the basis of the posterior probability distribution is applied. Using the result of examination of a statistic property of the maximum posterior probability, we attempt to find a time point that the loss function attains its minimum in the sense of an approximation. The stopping rule present-

ed here is such that the estimation should be stopped at the time point. Furthermore, we show some results of computer simulation to inspect accuracy of the approximation.

Key Words: Markovian decision process, transition probability, optimal policy, learning control, Bayesian estimation

1. ま え が き

未知パラメータを含む確率システムの学習制御問題として、2個のコインからなるベルヌーイ試行過程の最適化を考える two-armed bandit 問題、それをさらに拡張した multiarmed bandit 問題、学習オートマトンの問題、および未知遷移確率の推定を伴うマルコフ決定問題などがある。これらに対しては従来より数多くの研究がなされているが、特にマルコフ決定モデルは遷移確率が既知ならばきわめて広範な応用分野をもつことが知られている¹⁾。したがって、未知遷移確率の推定を伴うマルコフ決定問題は理論面からだけでなく実際面から見ても興味深いものである^{2), 3)}。

ところで、上記の各問題に対する従来の研究のほとんどは過程が無限に続くことを前提としており、そこでは最適な制御入力への収束を追求することが主題となっている。しかしながら、これら学習問題を現実的な立場から見ると、時間の制限が必ず存在するので、上記の収束が厳密に成り立つかどうかというような漸近的特性よりもむしろ、ある有限な時期までの過渡的特性が重要な意味をもつ。このような点から、有限時間の学習問題に対しても無限時間のそれと同様に積極的な研究がなされるべきである。

有限時間の学習問題においては、最適な制御入力へ収束させることなく、全期間の損失を最小化することが学習の目標となる。そのための学習方式として代表的なものは、ベイズ決定理論に基づき、制御入力を動的計画法により厳密に求めていく方式であ

[†] 第24回計測自動制御学会学術講演会で発表(昭60・7)

* 東北大学工学部 仙台市荒巻字青葉

* Faculty of Engineering, Tohoku University, Sendai

(Received May 19, 1986)

(Revised January 10, 1987)

る^{4),5)}. この方式はごく単純なモデルのもとでも全期間が長くなるにつれてその運用に伴う計算量が著しく増大するので、全期間がかなり短い場合にしか實際上適用できない。一方、著者らはさきに、これまでに提案されたいくつかの無限時間学習方式について、種々の条件のもとで収束速度を見るための計算機シミュレーションを実施した。その結果から、初期のうちに未知パラメータの推定を集中的に行うような学習様式が最適入力への速い収束を与え、その帰結として全体の損失を小さくするという結論を得た。これを参考に、著者らは、有限時間 two-armed bandit 問題に関して、未知パラメータの推定とそれに基づく制御とを時間軸上で 2 分割する考えを導入した。この考えのもと全期間の損失がなんらかの意味で最小になるような推定打ち切りについて考察した^{6),7)}。このごく単純な考えは全期間が比較的長い場合の有限時間問題に対して非常に有効な学習方式を与えるものと思われる。

そこで、本論文では、マルコフ決定モデルのもとで推定と制御とを時間軸上で 2 分割する学習方式を構成し、その推定を打ち切る方法について考察する。ただし、全期間は有限であるが十分長いものとする。推定の部分にはベイズ推定法を用いる。まず、推定打ち切りのための行動基準として、推定精度を表わす一つの指標である最大事後確率を用いるほか、事後分布を考慮した期待損失である損失関数を定義する。つぎに、最大事後確率の平均的な振舞を調べ、その結果のもとに損失関数が近似的な意味で最小となる時点で推定を打ち切るという方法を提案する。さらに、計算機シミュレーションによりその近似精度について検討する。

2. 問題の記述

2.1 マルコフ決定過程

各時刻 $t=0, 1, 2, \dots$ においてシステムは N 個の状態 $1, \dots, N$ のうちのどれか一つの状態にある ($N < \infty$)。各時刻ごとにシステムはある状態からある状態へ確率的に遷移し、それに伴いある利得が得られる。この遷移確率と利得は、つぎに述べるように、決定なるものを用いることによってそれらの値を変えることができる。たとえばシステムが状態 i にあるとき、それに応じて K_i 個の決定 $1, \dots, K_i$ が与えられ ($K_i < \infty$)、その中から一つの決定が選ばれて用いられる。そこでたとえば決定 k が用いられたとしよう。すると、つぎの時刻においてシステムは一定の確率 p_{ij}^k で状態 j に遷移し、それに応じて r_{ij}^k なる利得が得られる ($|r_{ij}^k| < \infty$)。

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、 u_i から添字を取って u で表わす。簡単のために $p_{ij}^{u_i}, r_{ij}^{u_i}$ をそれぞれ p_{ij}^i, r_{ij}^i の意味として用いる。上述の過程は政策を制御入力とする一つの確率制御過程であり、一般にマルコフ決定過程と呼ばれる。制御の目的は得られる利得をなんらかの意味で最大ならしめることであるが、長期間に得られる利得の 1 期間当たりの平均値を最大ならしめる政策は特に平均最適であるといわれる。全期間が有限でしかも十分長いときにはこの政策を適用するのが實際上最も効果がある。

2.2 平均最適政策

ある政策 u だけを続けて適用する過程を考える。過程が正則であると仮定すると、 $t \rightarrow \infty$ のときシステムは各状態 i に一定の確率 π_i^u で存在することが示される⁸⁾。確率 $\pi_1^u, \dots, \pi_N^u$ はつぎの連立方程式のただ 1 組の解で与えられる ($0 \leq \pi_i^u \leq 1$)。

$$\pi_j^u = \sum_i \pi_i^u p_{ij}^{u_i} \quad (j=1, \dots, N), \quad \sum_i \pi_i^u = 1$$

なお、過程の正則性はたとえば遷移確率 p_{ij}^i がすべて正であるということによりつねに保証される。さて、 π_i^u を用いて定義される

$$g^u = \sum_{i,j} \pi_i^u p_{ij}^{u_i} r_{ij}^{u_i}$$

は過程の 1 期間当たりの平均利得を表わす。この g^u を最大ならしめる $u=u^*$ が平均最適政策である。最大平均利得 g^* を単に g^* と書く。

つぎに、各状態 i ごとに決定 $1, \dots, K_i$ をすべて等頻度で用いる過程を考える。このときにも、上と同様に、1 期間当たりの平均利得 $\bar{g}$ が、連立方程式

$$\bar{\pi}_j = \sum_i \bar{\pi}_i (1/K_i) \sum_k p_{ij}^k \quad (j=1, \dots, N), \quad \sum_i \bar{\pi}_i = 1$$

の解 ($0 \leq \bar{\pi}_i \leq 1$) を用いてつぎのように表わされる。

$$\bar{g} = \sum_{i,j} \bar{\pi}_i (1/K_i) \sum_{k,j} p_{ij}^k r_{ij}^k$$

また、周知のように、 $\bar{g} \leq \max_u g^u = g^*$ である。

2.3 有限時間学習制御

遷移確率 p_{ij}^k の値がすべて既知ならば、平均最適政策など最適な制御入力容易に求められるので特に問題はない。しかし、多くの現実的諸問題におけるようにそれらが未知であるとする、最適入力もまた未知となるので問題はさらに難しくなる。このような状況に対しては、未知確率の推定を行い、その結果に基づいて制御を行うことが重要な方策となる。このように推定を伴う制御をここでは学習制御と呼ぶ。

本論文では、遷移確率 p_{ij}^k の値がすべて未知であるとして、マルコフ過程の学習制御問題を考察する。ただし、全期間 T は有限とし、平均最適政策をその

期間中適用するのが十分有効である程度の大きさであるとする。この問題に対して、つぎのようないわば推定と制御とを時間軸上で2分割した学習方式を提案する。まず、試行開始から時刻 t までは制御面を考慮せずに、適当な政策を適用しながら遷移確率の推定を行う。時刻 t になったら推定を打ち切り、そこでの推定結果に基づく平均最適政策を求める。その後、時刻 T までの残りの期間はその政策だけを用いる。

3. 遷移確率のベイズ推定

前章に提案した学習方式の推定の部分には、ベイズ推定法を用いる。各 i および k について、遷移確率ベクトル $\mathbf{p}_i^k = (p_{i1}^k, \dots, p_{iN}^k)$ は有限な M 個の値 $\mathbf{p}_i^k(1), \dots, \mathbf{p}_i^k(M)$ のうちのいずれか一つをとるものとし、その事前分布 $\delta(\mathbf{p}_i^k)$ を仮定する ($\sum_m \delta(\mathbf{p}_i^k(m)) = 1$)。いま、状態 i において決定 k を用いたところ、状態 j に遷移したとする。すると、事前分布 δ はベイズの法則により、下記のように事後分布 δ' に修正される。

$$\begin{cases} \delta'(\mathbf{p}_i^k(m)) = c p_{ij}^k(m) \delta(\mathbf{p}_i^k(m)) & (m=1, \dots, M) \\ \delta'(\mathbf{p}_i^k(m)) = \delta(\mathbf{p}_i^k(m)) & (m=1, \dots, M; \ell \neq i) \end{cases}$$

または $k \neq k$

ここに、 $p_{ij}^k(m)$ はベクトル $\mathbf{p}_i^k(m)$ の第 j 成分であり、また c は $\sum_m \delta'(\mathbf{p}_i^k(m)) = 1$ より定まる正規化定数である。このように、ある状態においてある決定を用いると、実質的にはその状態と決定に対応した遷移確率ベクトルの事前分布だけが変化する。

N 次元行ベクトルである遷移確率ベクトルは全部で $K = \sum_i K_i$ 個存在する。それらを適当な順序で縦に並べて得られる $K \times N$ 行列 $\mathbf{P}$ を遷移確率行列という。 K 個の遷移確率ベクトルはおのおの独立に M 通りの値をとりうるので、行列 $\mathbf{P}$ は $L = M^K$ 通りの値をとりうる。これらの値を $\mathbf{P}(1), \dots, \mathbf{P}(L)$ で表わす。このとき、上記 δ と δ' を用いて $\mathbf{P}$ の事前分布と事後分布をそれぞれ

$$\gamma(\mathbf{P}) = \prod_{i,k} \delta(\mathbf{p}_i^k), \quad \gamma'(\mathbf{P}) = \prod_{i,k} \delta'(\mathbf{p}_i^k)$$

と定めることができる。あきらかに $\sum_L \gamma(\mathbf{P}(L)) = \sum_L \gamma'(\mathbf{P}(L)) = 1$ が成り立つ。時刻 t における δ' 、 γ' をそれぞれ δ_t 、 γ_t と書き、 $\gamma_t(\mathbf{P}(L))$ を簡単に $\gamma_t(L)$ と書く。

さて、事前分布の修正の仕方からわかるように、遷移確率全部の実質的な推定を行うには、状態 i において決定 k を用いる回数がある i と k の組に対して互いに等しくなるようにしなければならない。しかし、たとえ $\mathbf{P}$ が既知であっても、そうならしめるような政策選択を行うのはかなり困難である。そこ

で、本学習方式の推定の部分では、そのような政策選択にできるだけ近づけるために各状態ごとに決定をすべて等頻度で用いて $\mathbf{P}$ のベイズ推定を行うこととし、時刻 t において $\gamma_t(L)$ の最大化により推定値 $\mathbf{P}(L_t)$ を求める ($\gamma_t(L_t) = \max_L \gamma_t(L)$)。

4. 推定打ち切り点の決定

さきに定義した g^* 、 u^* 、 g^* 、 $\bar{g}$ の値はいずれも遷移確率行列 $\mathbf{P}$ の値に依存するから、 $\mathbf{P} = \mathbf{P}(L)$ に対するそれらの値をそれぞれの記号の後に (L) を付けて表わす。本学習方式では、時刻 t において推定値 $\mathbf{P}(L_t)$ が得られたら、時刻 T までの残りの期間はその推定結果に基づく平均最適政策 $u^*(L_t)$ だけを適用する。ここで当然問題となるのは推定打ち切り点 t をどのように定めたらよいかということである。本方式では、試行以前に打ち切り点を定めるのではなく、試行過程で打ち切りの判断を行うという方法をとる。これは後者の方が実際の面で有利だからである⁷⁾。以下、この問題を考察していくが、 t には変数としての推定打ち切り点の意味をもたせ、実際には t を推定が行われている任意の時刻とする。これに応じて t^* を実際の打ち切り点とする。

4.1 行動基準の設定

推定打ち切りはなんらかの行動基準をもとに行われなければならない。行動基準の設定のために、まず、種々の動作に伴う損失を考える。 T が十分大きいことから、全期間にわたって真の平均最適政策を適用することをもって理想の動作とし、それ以外の政策を適用するときにはその動作をつぎに定める損失によって評価する。すなわち、かりに遷移確率行列の真値を $\mathbf{P}(L)$ とするとき、理想の動作による1期間当たりの平均利得が $g^*(L)$ であることから、最初の t 期間決定を等頻度で用いることを損失 $t[g^*(L) - \bar{g}(L)]$ の動作、また残りの $T-t$ 期間政策 $u^*(L_t)$ を適用することを損失 $(T-t)[g^*(L) - g^{u^*(L_t)}(L)]$ の動作と定める。ここで、事後確率 $\gamma_t(L)$ を重みとする上記2種類の損失の期待値

$$\begin{cases} U(t) = t \sum_L \gamma_t(L) [g^*(L) - \bar{g}(L)] \\ V(t) = (T-t) \sum_L \gamma_t(L) [g^*(L) - g^{u^*(L_t)}(L)] \end{cases} \quad (1)$$

を定義する。すると、これらの和で定義される

$$W(t) = U(t) + V(t) \quad (2)$$

は時刻 t での推定情報から算出できる全期間の損失の期待値を表わす。そこで、 $W(t)$ を t の損失関数と名づける。 $U(t)$ は推定に伴う損失を意味し、本質的には t の増加につれて直線的に増加すべき性質のものであ

る。なお、推定に要する計算量を t に比例する損失として $U(t)$ に含めることができるが、ここではそれを考慮しないことにする。一方、 $V(t)$ は推定の誤りによる損失を意味し、 $l=l_i$ に対して $g^*(l)=g^{*(l)}(l)$ であることから、 t の増加につれて $T-t$ の減少と下に述べるような $\gamma_i(l_i)$ の増加の効果により $V(t)$ は直線よりも急な形で減少する。したがって、 $W(t)$ においては、 t が小さければ $V(t)$ が支配的となり、逆に $T-t$ が小さければ $U(t)$ が支配的となる。

本方式では、推定値の事後確率すなわち最大事後確率 $\gamma_i(l_i) \triangleq \gamma^*(t)$ と上に定義した損失関数 $W(t)$ を推定打ち切りのための行動基準として用いる。もちろん、 $W(t)$ は $\gamma^*(t)$ にも依存しており、両者とも時刻 t ではじめてその値が定まる確率変数である。いくつかのモデルについて推定精度の変化を見るための計算機シミュレーションを実施してみると、一般的な傾向として、 t が小さいうちは推定値や $\gamma^*(t)$ の値にかなりのばらつきがあるものの、次第にそれらのばらつきがなくなり、 t の増加につれて $\gamma^*(t)$ のトレンドはほぼなめらかな曲線を描きながら 1 に漸近することがわかる。このような変化に伴い、 $W(t)$ もはじめのうちはかなり変動するが、漸近的には $U(t)$ の支配的な効果によりほぼ直線的に増加する。

基本的には、損失を最小化したいという目的により、 $W(t)$ を行動基準とすべきである。ただし、 $\gamma^*(t)$ の値がまだ小さく推定値がばらついたりする段階では、 $W(t)$ がどんな小さい値をとったとしてもそこで推定を打ち切るのは時期尚早である。それでまず、推定精度が不十分な段階で推定を打ち切らないようにするために、推定精度に関する下記の条件をおく。

条件 1: $\gamma^*(t) > 1/2$

これが満たされたときに限り $W(t)$ を行動基準とするが、確率変数である $W(t)$ をそのままの形で用いるのには無理がある。そこで、 $\gamma^*(t)$ のトレンドがほぼなめらかな曲線を描くという事実から、その曲線を何らかの関数で近似し、それに基づいて $W(t)$ もその将来のトレンドを予測するような関数に書き換える。そのための準備として、次節において、理論的な観点から $\gamma^*(t)$ の t に関する平均的な変化を調べる。

4.2 $\gamma^*(t)$ の平均的な変化

ここでは、議論の簡単のために、状態と決定の数とともに 2 で、各遷移確率ベクトルが 2 通りの値をとる場合 ($M=2$) について調べる。各 i および k に対して

$$q_i^k(t) = \max [\delta_i(\mathbf{p}^k(1)), \delta_i(\mathbf{p}^k(2))]$$

とすると、定義からあきらかに

$$\gamma^*(t) = q_1^1(t)q_2^1(t)q_3^1(t)q_4^1(t)$$

となる。時刻 t までの間に状態 i において決定 k が用いられた回数を $n_i^k(t)$ で表わすと、事前分布の修正の仕方から、 $\delta_i(\mathbf{p}^k(1))$ や $\delta_i(\mathbf{p}^k(2))$ は $n_i^k(t)$ に依存し、その結果 $q_i^k(t)$ も $n_i^k(t)$ に依存することがわかる。よって、 $\gamma^*(t)$ の期待値がつぎのように表わされる。

$$E[\gamma^*(t)] = \sum_{t_1} \sum_{t_2} \sum_{t_3} \sum_{t_4} E[q_1^1(t)q_2^1(t)q_3^1(t)q_4^1(t) |$$

$$n_1^1(t)=t_1, n_2^1(t)=t_2, n_3^1(t)=t_3, n_4^1(t)=t_4]$$

$$\times P_r \{n_1^1(t)=t_1, n_2^1(t)=t_2, n_3^1(t)=t_3,$$

$$n_4^1(t)=t_4\} \quad (3)$$

ただし、上式の $t_1, \dots, t_4$ には、すべて非負整数であることおよび $t_1+t_2+t_3+t_4=t$ なる制限が加えられるほか、決定を等頻度で用いることによる

$$t_1=t_2, t_3=t_4 \quad (4)$$

なる制限が加えられる。これにより (3) 式はつぎのように変形される。

$$E[\gamma^*(t)] = \sum_{t_1=0}^{t/2} E[q_1^1(t)q_2^1(t)q_3^1(t)q_4^1(t) | n_1^1(t)$$

$$=n_1^1(t)=t_1, n_2^1(t)=n_3^1(t)$$

$$=t/2-t_1] P_r \{n_1^1(t)=t_1\}$$

$$= \sum_{t_1=0}^{t/2} E[q_1^1(t) | n_1^1(t)=t_1] E[q_2^1(t) | n_2^1(t)$$

$$=t_1] E[q_3^1(t) | n_3^1(t)=t/2-t_1] E[q_4^1(t) | n_4^1(t)$$

$$=t/2-t_1] P_r \{n_1^1(t)=t_1\} \quad (5)$$

なお、(4) 式の制限は時刻 t までの状態 1, 2 の生起回数がともに偶数であることを前提としているが、一方あるいは両方が奇数でもそれらの間に実質的な差が生じないので、このまま議論を進める。

$E[\gamma^*(t)]$ を t の関数として見ると、それは t の増加につれて 1 に漸近するはずであるが、簡単な関数としては記述されない。そこで、1 への漸近の仕方を t の連続関数の形で近似的に表現する、 $E[\gamma^*(t)]$ の下限を求める。Hellman ら⁹⁾は、各時刻で同一の母集団分布をもつ独立過程について、 m 個の母集団 $C_1, \dots, C_m$ の中からベイズ推定法に従って一つを決定する問題を考察し ($m \geq 2$)、推定の誤り確率の tight な上限を示した。すなわち、ベイズ推定によって時刻 t で得られる $C_1, \dots, C_m$ の事後確率のうち最大のものを $\alpha^*(t)$ とすると、推定の誤り確率 $P_e(t)$ の上限が

$$P_e(t) = E[1 - \alpha^*(t)] \leq a \rho^t$$

で与えられる。ここに、 a と ρ はあらかじめ求めることのできる定数である ($a > 0, 0 < \rho < 1$)。まず、この結果を本方式の推定過程に適用してみると、各 i および k に対して次式の成り立つことがわかる。

$$E[1 - q_i^k(t) | n_i^k(t)=t] \leq a_i^k(\rho_i^k)^t$$

ここに、 $a_i^k=1/2$ および $\rho_i^k = \inf_{0 \leq x \leq 1} \{[p_{i1}^k(1)]^x [p_{i1}^k(2)]^{1-x}\}$

$+ [p_{12}^*(1)]^* [p_{12}^*(2)]^{1-*}$ である。上式を(5)式に適用して次式を得る。

$$E[\gamma^*(t)] \geq \sum_{t_1=0}^{t/2} [1-a_1^2(\rho_1^2)^{t_1}] [1-a_2^2(\rho_2^2)^{t_1}] \cdot [1-a_3^2(\rho_3^2)^{t/2-t_1}] \times [1-a_4^2(\rho_4^2)^{t/2-t_1}] P_r\{n_1(t)=t_1\} \quad (6)$$

つぎに、(6)式の $P_r\{n_1(t)=t_1\}$ を評価するために、時刻 t までの状態1の生起回数 $n_1(t)=2n_1^*(t)$ について調べる。状態1の生起確率はつねに $\min_{i,k} p_{11}^*$ より大きく $\max_{i,k} p_{11}^*$ より小さいので、全期間にわたるその生起確率の上限 p_U と下限 p_L が区間 $(\min_{i,k} p_{11}^*, \max_{i,k} p_{11}^*)$ 内に存在する。いま、各時刻における状態1の生起確率がそれぞれ p_U と p_L で与えられる二つの独立過程を考え、 $n_U(t)$ と $n_L(t)$ をそれぞれの過程における時刻 t までの状態1の生起回数とする。すると、任意の整数 τ に対して ($0 < \tau < t$)、事象 $\{n_1(t) > \tau\}$ よりも事象 $\{n_U(t) > \tau\}$ が起こりやすく、また事象 $\{n_1(t) < \tau\}$ よりも事象 $\{n_L(t) < \tau\}$ が起こりやすいことから、次式が成り立つ。

$$P_r\{n_1(t) > \tau\} \leq P_r\{n_U(t) > \tau\} = \sum_{n < \tau} \binom{t}{n} p_U^n (1-p_U)^{t-n} \quad (7)$$

$$P_r\{n_1(t) < \tau\} \leq P_r\{n_L(t) < \tau\} = \sum_{n < \tau} \binom{t}{n} p_L^n (1-p_L)^{t-n} \quad (8)$$

ところで、Ash¹⁰⁾によれば、任意の実数 $\lambda \in (p_U, 1)$ に対して

$$\sum_{n > \lambda t} \binom{t}{n} p_U^n (1-p_U)^{t-n} < 2^{-D'}$$

また任意の実数 $\lambda' \in (0, p_L)$ に対して

$$\sum_{n < \lambda' t} \binom{t}{n} p_L^n (1-p_L)^{t-n} < 2^{-D'}$$

が成り立つ。ここに、 D は λ と p_U の間の、また D' は λ' と p_L の間のいずれも距離を表わす divergence と呼ばれる正数である。 $D=D'$ となるように λ と λ' を選び、上記2式をそれぞれ(7)式と(8)式に適用すれば

$$P_r\{n_1(t) > \lambda t\}, P_r\{n_1(t) < \lambda' t\} < 2^{-D'}$$

が得られる。よって

$$P_r\{\lambda' t \leq n_1(t) \leq \lambda t\} \geq 1 - 2 \cdot 2^{-D'}$$

あるいは

$$P_r\{\lambda' t/2 \leq n_1(t) \leq \lambda t/2\} \geq 1 - 2 \cdot 2^{-D'} \quad (9)$$

が成り立つ。さて、(6)式より

$$E[\gamma^*(t)] \geq \sum_{t_1=\lambda' t/2}^{\lambda t/2} [1-a_1^2(\rho_1^2)^{t_1}] [1-a_2^2(\rho_2^2)^{t_1}] \cdot [1-a_3^2(\rho_3^2)^{t/2-t_1}]$$

$$\begin{aligned} & \times [1-a_4^2(\rho_4^2)^{t/2-t_1}] P_r\{n_1(t)=t_1\} \\ & \geq \min_{\lambda' t/2 \leq t_1 \leq \lambda t/2} [1-a_1^2(\rho_1^2)^{t_1}] [1-a_2^2(\rho_2^2)^{t_1}] \\ & \quad \times [1-a_3^2(\rho_3^2)^{t/2-t_1}] [1-a_4^2(\rho_4^2)^{t/2-t_1}] \\ & \quad \times P_r\left\{\frac{\lambda' t}{2} \leq n_1(t) \leq \frac{\lambda t}{2}\right\} \\ & \geq [1-a_1^2(\rho_1^2)^{(\lambda/2)t}] [1-a_2^2(\rho_2^2)^{(\lambda/2)t}] \\ & \quad \cdot [1-a_3^2(\rho_3^2)^{(1-\lambda/2)t}] \\ & \quad \times [1-a_4^2(\rho_4^2)^{(1-\lambda/2)t}] \\ & \quad \cdot P_r\left\{\frac{\lambda' t}{2} \leq n_1(t) \leq \frac{\lambda t}{2}\right\} \end{aligned}$$

が成り立つので、これに(9)式を適用して $\rho_1=(\rho_1^*)^{\lambda/2}$, $\rho_2=(\rho_2^*)^{\lambda/2}$, $\rho_3=(\rho_3^*)^{1-\lambda/2}$, $\rho_4=(\rho_4^*)^{1-\lambda/2}$, $\rho_5=2^{-D}$ とおくことにより次式が得られる。

$$E[\gamma^*(t)] \geq (1-a_1^2 \rho_1^2)(1-a_2^2 \rho_2^2)(1-a_3^2 \rho_3^2)(1-a_4^2 \rho_4^2) \times (1-2\rho_5^2) \quad (10)$$

あきらかに $0 < \rho_1, \dots, \rho_5 < 1$ である。

(10)式は五つの項の積からなる t の連続関数を $E[\gamma^*(t)]$ の下限として与えるものである。これらの項はいずれも $1-a\rho^2$ なる形 ($a>0$, $0<\rho<1$) をとっている。(10)式を導いたのと同様にすれば、状態や決定の数あるいはまた M が大きい場合には、同じ形の項をより多く掛け合せたような下限が得られるであろう。そこで、 $1-a\rho^2$ なる形の項の積で表わされる曲線を用い、各項の a と ρ の値を調整することにより、 $\gamma^*(t)$ が実際に描くトレンドを近似することを考える。そのことに関する何回かの計算機シミュレーションの結果によれば、状態や決定の数および M が小さい場合には、ただ一つの項 $1-a\rho^2$ からなる曲線だけで十分良く近似できることが確かめられる。

4.3 推定打ち切りの方法

推定精度に関する条件1が満たされたら、そのときには、損失関数を推定打ち切りのための行動基準として考慮する。いまの時刻 t は条件1が満たされているものとする。まず、 $\tau=1$ から $\tau=t$ まで最大事後確率 $\gamma^*(\tau)$ が描いてきたトレンドを、前節の結果をもとに

$$f(\tau) = \prod_{i=1}^I (1-a_i \rho_i^2)$$

なる曲線で近似する。ただし、以下、モデルを $I=1$ ととれる程度の規模のものに限定し、 a_1 と ρ_1 は2乗誤差の和 $\sum_{\tau=1}^t |f(\tau) - \gamma^*(\tau)|^2$ が最小となるように定める ($a_1>0$, $0<\rho_1<1$)。つぎに、この曲線を将来の時刻 $\tau \geq t$ に外挿することにより、将来の事後確率 $\gamma_\tau(l)$ は $l=l_t$ に対しては $f(\tau) \triangleq \hat{\gamma}_\tau(l_t)$ なる曲線の形で増加し、各 $l \neq l_t$ に対しては $[1-f(\tau)]\gamma_\tau(l)/\sum_{l \neq l_t} \gamma_\tau(l) \triangleq \hat{\gamma}_\tau(l)$ なる曲線の形で減少するものと予測する。この予測に

従えば, (1)式と(2)式で定義される損失関数は将来の時刻 $\tau \geq t$ に対して

$$\hat{W}(\tau) = \tau \sum_l \hat{r}_\tau(l) [g^*(l) - \hat{g}(l)] + (T - \tau) \sum_l \hat{r}_\tau(l) [g^*(l) - g^{u^*(t, \tau)}(l)] \quad (11)$$

となるはずである。 $\hat{W}(\tau)$ は τ の連続関数と見なすことができ, その区間 $[t, T]$ における最小値は容易に求めることができる。もし $\hat{W}(t)$ がその最小値に等しければ, そのことは, これから先推定を続けても損失関数がいまより小さな値をとることはないを予測されたことを意味する。よってこのとき推定を打ち切る。

以上提案した推定打ち切り方法を要約するとつぎのようになる。各時刻 t において, 遷移確率のベイズ推定の後, まず条件 1 が満たされるかどうかを調べる。もしそれが満たされたならば, ついで(11)式の関数 $\hat{W}$ を構成し,

$$\text{条件 2: } \hat{W}(t) = \min_{t \leq \tau \leq T} \hat{W}(\tau)$$

が満たされるかどうかを調べる。このようにして二つの条件が同時に満たされたならば, そこで推定を打ち切る。これ以上の場合には, つぎの時刻に進み, 同じ操作を繰返す。

5. 計算機シミュレーション

本章では, 2 状態 2 決定および $M=2$ のモデルに対して本学習方式の推定打ち切り方法を適用した計算機シミュレーションの結果を示す。モデルの遷移確率ベクトルと利得の値はそれぞれ Table 1 と Table 2 の通りである。このとき, 遷移確率行列は下記の 16 通りの値をとりうるので, $\gamma_0(1) = \dots = \gamma_0(16) = 1/16$ とおく。

$$P(1) = \begin{bmatrix} p_1^1(1) \\ p_2^1(1) \\ p_1^2(1) \\ p_2^2(1) \end{bmatrix}, \quad P(2) = \begin{bmatrix} p_1^1(2) \\ p_2^1(2) \\ p_1^2(2) \\ p_2^2(2) \end{bmatrix}, \dots,$$

Table 1 Transition probability vectors

i	k	m	$p_i^k(m)$
1	1	1	$p_1^1(1) = (0.7, 0.3)$
		2	$p_1^1(2) = (0.1, 0.9)$
	2	1	$p_1^2(1) = (0.4, 0.6)$
		2	$p_1^2(2) = (0.9, 0.1)$
2	1	1	$p_2^1(1) = (0.6, 0.4)$
		2	$p_2^1(2) = (0.2, 0.8)$
	2	1	$p_2^2(1) = (0.2, 0.8)$
		2	$p_2^2(2) = (0.5, 0.5)$

$$P(16) = \begin{bmatrix} p_1^1(2) \\ p_2^1(2) \\ p_1^2(2) \\ p_2^2(2) \end{bmatrix}$$

シミュレーションは, 全期間 $T=100$ のもとで, 学習制御者にとっては未知なる P の真値を最初は $P(1)$ ついで $P(16)$ に設定して行われた。もちろん学習方式としては本方法による推定打ち切り点 t^* で推定を打ち切るわけであるが, シミュレーションでは損失関数 $W(t)$ の変化を見るために全期間にわたって推定を行った。その結果を Fig. 1 と Fig. 2 に示す。Fig. 1 は $P(1)$ が真値のときの $W(t)$ と t^* を表わし, Fig. 2 は $P(16)$ が真値のときのそれらを表わす。なお, どちらが真値のときも時刻 t^* における推定値 $P(t^*)$ は

Table 2 Rewards

i	k	j	r_{ij}^k
1	1	1	$r_{11}^1 = 0$
		2	$r_{12}^1 = 1$
	2	1	$r_{11}^2 = 0$
		2	$r_{12}^2 = 1$
2	1	1	$r_{21}^1 = 0$
		2	$r_{22}^1 = 1$
	2	1	$r_{21}^2 = 0$
		2	$r_{22}^2 = 1$

Fig. 1 Variation of loss function $W(t)$ with respect to t and estimation stopping time t^* when $P(1)$ is the true value

Fig. 2 Variation of loss function $W(t)$ with respect to t and estimation stopping time t^* when $P(16)$ is the true value

真値であった。これらの図によれば、本方法による推定打ち切り点はほぼ $W(t)$ が直線的に増加し始めた時点となっており、そのあきらかな初期の変動を考慮に入ればほぼ $W(t)$ が最小になった時点となっている。また、利得の値をいくつか変えて行った同じ実験によってもやはり同様の結果が得られた。この意味で、本方法に用いた近似の精度は十分満足すべきものであった。

6. む す び

本論文では、全期間が有限であるマルコフ過程の学習制御問題を考察し、推定と制御とを時間軸上で2分割する一つの学習方式を示した。推定の打ち切りに関しては、そのための行動基準として、ベイズ推定による最大事後確率と、事後分布を考慮して定義される損失関数を用いた。ここに提案された打ち切りの方法は、最大事後確率の統計的性質を調べた結果をもとに、損失関数が近似的な意味で最小となるところで推定を打ち切るというものである。シミュレーション結果によればその近似精度は十分満足すべきものであった。

このほかにも行動基準のとり方などはいろいろ考えられるが、本方法の意義は、あまり複雑な計算手順を要しない、しかも妥当性のある推定打ち切りを行うための一つの目安を与えたことにある。本論文の学習方式は実際上は小規模なモデルのもとで構成されたが、

その基本となる考えは非常に単純であるので、より複雑なモデルあるいはマルコフ過程以外のモデルに対しても適当な簡単化手法を用いれば拡張が可能であろう。今後は、その拡張も含めて、本方式の実用化に向けての具体的な検討を行う必要がある。

参 考 文 献

- 1) H. J. Kushner: Introduction to Stochastic Control, Holt, Rinehart and Winston (1971)
- 2) P. R. Kumar and W. Lin: Optimal Adaptive Controllers for Unknown Markov Chains, IEEE Trans. Automatic Control, **AC-27-4**, 765/774 (1982)
- 3) 佐藤, 阿部, 竹田: マルコフ過程の漸近最適な学習制御, 計測自動制御学会論文集, **21-7**, 706/712 (1985)
- 4) R. N. Bradt, S. M. Johnson and S. Karlin: On Sequential Designs for Maximizing the Sum of n Observations, Ann. Math. Stat., **27**, 1060/1074 (1956)
- 5) J. Martin: Bayesian Decision Problems and Markov Chains, Wiley (1967)
- 6) 佐藤, 阿部, 竹田: 2個のコインからなる確率モデルの有限時間学習制御, 電気学会論文誌C, **104-C-2**, 35/42 (1984)
- 7) 佐藤, 竹田: 有限時間 Two-armed bandit 問題における推定打ち切り点の決定, 電気学会論文誌C, **105-C-12**, 235/241 (1985)
- 8) J. G. Kemeny and J. L. Snell: Finite Markov Chains, Van Nostrand (1963)
- 9) M. E. Hellman and J. Raviv: Probability of Error, Equivocation and the Chernoff Bound, IEEE Trans. Information Theory, **IT-16-4**, 368/372 (1970)
- 10) R. B. Ash: Information Theory, Interscience (1965)