各レベルに S-モデル定常環境をもつ階層構造学習オートマトン に対する L_{R-I} 型学習アルゴリズム

最上 義夫*・馬場 則夫**
塩尻 晃***・田上 隆徳****

L_{R-I} Type Learning Algorithm for Hierarchical Structure Learning Automata
with S-model Stationary Random Environment at Each Level

Yoshio MOGAMI*, Norio BABA**,
Akira SHIOJIRI*** and Takanori TAGAMI****

In this paper, for variable hierarchical structure learning automata with S-model stationary random environment at each level, a new definition of optimal path is proposed based on the arithmetic mean of average rewards, and an L_{R-I} type learning algorithm is constructed. The learning property of our algorithm is considered theoretically, and it is proved that the probability of finding the optimal path can approach 1 as much as possible by using our algorithm. In numerical simulations, the usefulness of our algorithm is shown.

Key Words: hierarchical structure learning automata, S-model stationary random environment, arithmetic mean of average rewards, reward parameter, L_{R-I} type learning algorithm

1. ま え が き

1981 年, 学習オートマトンを木構造に組み合わせた階層構造学習オートマトンが Thathachar and Ramakrishnan¹⁾によって提案された。そして, L_{R-I} 型学習アルゴリズムが与えられ, その学習アルゴリズムが良好な学習特性を持つことが示された。馬場, 最上, 三宅²⁾は, これを拡張することによって, 各レベルに非定常複数未知環境が存在するような階層構造学習オートマトンに対する学習アルゴリズムを提案した。また最上³⁾が, Thathachar and Ramakrishnan の学習アルゴリズムを可変階層構造学習オートマトンに適用し, 可変階層構造学習アルゴリズムを構築した。そして, その学習特性について理論的に考察した。

ところで, 学習オートマトンは未知環境の自律的学習システムであって, 自分が環境に与えた出力とそれに対する未知環境からの報酬(評価値)とに基づいて学習を進めて最適な

出力を見出すものである。このとき, 報酬の値が 1 (成功) または 0 (失敗) のいずれかであるような環境を P-モデル環境, $[0,1]$ 区間内の有限個の有理数で与えられる環境を Q-モデル環境, $[0,1]$ 区間内の実数で与えられる環境を S-モデル環境と呼ぶが, P-モデル環境が最もシンプルな環境であることから, 従来, P-モデル環境に関する研究が主に行なわれてきたし, 上にあげた研究も P-モデル環境を対象としている。しかし, 学習オートマトンのより広範囲な問題への適用を考えると, 最も一般的な未知環境である S-モデル環境を対象とすることが必要であろう。

そこで, 本論文では, 各レベルに S-モデル定常環境をもつ各レベル入出力階層構造学習オートマトンについて考察する。そして, S-モデル環境において自然であろうと思われる形で最適パスを定義し, そのもとで階層構造学習アルゴリズムを構築する。この学習アルゴリズムは, 単体の学習オートマトンに対する学習アルゴリズムとして最もよく用いられる L_{R-I} 型学習アルゴリズム⁴⁾⁻⁷⁾の中の SL_{R-I} 強化法⁷⁾に修正を加えることによって得られる。さらに, その学習特性についても検討する。

2. 各レベルに S-モデル定常環境をもつ階層構造学習オートマトン

2.1 各レベル入出力階層構造学習オートマトン

各レベルに定常環境を持つ各レベル入出力階層構造学習

* 徳島大学 徳島市南常三島町 2-1

** 大阪教育大学 柏原市旭ヶ丘 4

*** 徳島大学 徳島市南常三島町 2-1

**** 阿南工業高等専門学校 阿南市見能林町青木 265

* The University of Tokushima, Tokushima

** Osaka Educational University, Kashiwara

*** The University of Tokushima, Tokushima

**** Anan College of Technology, Anan

(Received May 7, 2001)

(Revised October 15, 2001)

Fig. 1 Hierarchical structure learning automata

オートマトンを Fig. 1 に示す。\$r\$ 個の動作を持つ学習オートマトンが木構造に結合された構造をしており、各レベルにそれぞれ未知環境が対応している。レベル \$s\$ には \$r^{s-1}\$ 個の学習オートマトンが存在し、レベル \$s\$ の学習オートマトンを \$Z_{i_1 i_2 \dots i_{s-1}}\$、その動作を \$\alpha_{i_1 i_2 \dots i_{s-1} i_s}\$ (\$i_s = 1, 2, \dots, r\$) と表記する。また、レベル \$s\$ に対応している未知環境は \$R_s\$ と書かれる。ここで、レベル 1 から最下レベル \$n\$ に至る動作の列を \$\{\alpha_{i_1}, \alpha_{i_1 i_2}, \dots, \alpha_{i_1 i_2 \dots i_n}\}\$ とするとき、これを \$\phi_{i_1 i_2 \dots i_n}\$ と書いてパスと呼ぶ。

つぎに、その動作について述べる。初期時刻 \$t = 0\$ における各動作の動作確率（その動作が選択される確率）は、すべての学習オートマトンにおいて等しいとする（つまり、\$1/r\$ となる）。そして、時刻 \$t\$ において、レベル 1 の学習オートマトン \$Z\$ が動作確率分布 \$(p_1(t), p_2(t), \dots, p_r(t))'\$ に基づいて動作 \$\alpha_{j_1}\$ を選択したとすると、\$\alpha_1(t) = \alpha_{j_1}\$ が環境に出力される。この出力に対して、未知環境 \$R_1\$ が報酬 \$\beta_1(t) = \beta_{j_1}\$ を生成し \$Z\$ に与える。すると、\$Z\$ はこの報酬を用いて学習アルゴリズムにしたがって動作確率分布を更新する。これとは別に、レベル 2 においてはレベル 1 で選択された動作 \$\alpha_{j_1}\$ の下に存在する学習オートマトン \$Z_{j_1}\$ が動作し、\$Z\$ と同様、時刻 \$t\$ における動作確率分布 \$(p_{j_1 1}(t), p_{j_1 2}(t), \dots, p_{j_1 r}(t))'\$ に基づいて動作 \$\alpha_2(t) = \alpha_{j_1 j_2}\$ を選択しレベル 2 の未知環境 \$R_2\$ へ出力する。この出力に対して \$R_2\$ は、報酬 \$\beta_2(t) = \beta_{j_1 j_2}\$ を生成し \$Z_{j_1}\$ に与える。そして、この報酬に基づいて \$Z_{j_1}\$ もその動作確率分布を更新する。さらに、レベル 3 ではレベル 2 で選択された動作 \$\alpha_{j_1 j_2}\$ の下に存在する学習オートマトン \$Z_{j_1 j_2}\$ が動作する。このような一連の動作が最下レベル \$n\$ まで続けられると、時刻を \$t+1\$ としてレベル 1 の学習オートマトン \$Z\$ が新たな動作を選択する。このような学習サイクルを繰り返しながら階層構造学習オートマトンの学習は進められる。このとき、パス上の動作の動作確率の積をパス確率といい、\$\pi_{i_1 i_2 \dots i_n}(t) = p_{i_1}(t)p_{i_1 i_2}(t) \dots p_{i_1 i_2 \dots i_n}(t)\$ と表わす。

2.2 S-モデル定常環境と最適パス

本論文では、各レベルに S-モデル定常環境が存在する場合について考察する。すなわち、未知環境から与えられる報酬 \$\beta_{i_1 i_2 \dots i_s}\$ の平均値（平均報酬）\$E\{\beta_{i_1 i_2 \dots i_s}\}\$ を、

$$\bar{\beta}_{i_1 i_2 \dots i_s} = E\{\beta_{i_1 i_2 \dots i_s}\} \quad (1)$$

と書くことにすると、時刻 \$t\$ におけるレベル \$s\$ (\$s = 1, 2, \dots, n\$) の出力 \$\alpha_s(t) = \alpha_{i_1 i_2 \dots i_s}\$ に対する報酬 \$\beta_s(t) = \beta_{i_1 i_2 \dots i_s}\$ (\$0 \leq \beta_{i_1 i_2 \dots i_s} \leq 1\$) が、平均値 \$\bar{\beta}_{i_1 i_2 \dots i_s}\$ をもつ確率分布にしたがって与えられるような環境である。

ところで、階層構造学習オートマトンの学習の目的は最適パス（\$\phi_{j_1^* j_2^* \dots j_n^*}\$ と書く）を見出すことである。ここで、P-モデル環境における最適パスについて見てみよう。P-モデル環境では、動作 \$\alpha_{i_1 i_2 \dots i_s}\$ に対する成功の報酬 \$\beta_{i_1 i_2 \dots i_s} = 1\$ が報酬確率 \$d_{i_1 i_2 \dots i_s}\$ でもって与えられることから、最適パスは成功の報酬を得る確率が最も大きいパスとして定義される。すなわち、パス \$\phi_{i_1 i_2 \dots i_n}\$ 上の各動作 \$\{\alpha_{i_1}, \alpha_{i_1 i_2}, \dots, \alpha_{i_1 i_2 \dots i_n}\}\$ に対する報酬確率を \$\{d_{i_1}, d_{i_1 i_2}, \dots, d_{i_1 i_2 \dots i_n}\}\$ とすると、

$$d_{j_1^*} d_{j_1^* j_2^*} \times \dots \times d_{j_1^* j_2^* \dots j_n^*} > d_{i_1} d_{i_1 i_2} \times \dots \times d_{i_1 i_2 \dots i_n} \quad (2)$$

を満たすパス \$\phi_{j_1^* j_2^* \dots j_n^*}\$ が唯一に存在するとき、このパスが最適パスとなる。よって、この定義を準用して S-モデル環境における最適パスを定義しようとする、(1) 式の平均報酬を用いて、

$$\bar{\beta}_{j_1^*} \bar{\beta}_{j_1^* j_2^*} \times \dots \times \bar{\beta}_{j_1^* j_2^* \dots j_n^*} > \bar{\beta}_{i_1} \bar{\beta}_{i_1 i_2} \times \dots \times \bar{\beta}_{i_1 i_2 \dots i_n} \quad (3)$$

を満たすパス \$\phi_{j_1^* j_2^* \dots j_n^*}\$ が唯一に存在するとき、このパスが最適パスとなるということになる^(注1)。ここで、(2) 式および (3) 式の意味について考えてみよう。(2) 式は、上述したように成功の報酬を得る確率が最も大きいパスが最適パスであるということを表わしている。これに対して (3) 式はどうであろうか。これは P-モデル環境における定義の拡張として受け入れることの出来るものではあるが、平均報酬の積が最大のパスが最適パスであるということを述べている。しかし報酬の場合には、一般的な感覚としては、積が最大というよりも和が最大になるという方が自然であろうと思われる。そこで本論文では、パス上の各動作に対応している平均報酬の算術平均が最大のパスを最適パスと定義することにする。すなわち、パス \$\phi_{i_1 i_2 \dots i_n}\$ 上の各動作 \$\{\alpha_{i_1}, \alpha_{i_1 i_2}, \dots, \alpha_{i_1 i_2 \dots i_n}\}\$ に対応している平均報酬 \$\{\bar{\beta}_{i_1}, \bar{\beta}_{i_1 i_2}, \dots, \bar{\beta}_{i_1 i_2 \dots i_n}\}\$ の算術平均

(注1) P-モデル環境においては、(1) 式はつぎのように変換される。\$\bar{\beta}_{i_1 i_2 \dots i_s} = 1 \times d_{i_1 i_2 \dots i_s} + 0 \times (1 - d_{i_1 i_2 \dots i_s}) = d_{i_1 i_2 \dots i_s}\$。すなわち、報酬確率と平均報酬とは同値となる。このことから、S-モデル環境における最適パスの定義式は、(2) 式において報酬確率の代わりに平均報酬を用いて (3) 式とするということになる。

$$X_{i_1 i_2 \dots i_n} = \frac{\bar{\beta}_{i_1} + \bar{\beta}_{i_1 i_2} + \dots + \bar{\beta}_{i_1 i_2 \dots i_n}}{n} \quad (4)$$

をパス $\phi_{i_1 i_2 \dots i_n}$ に対するパス平均報酬と呼び、これを用いてつぎのように定義する。

【定義】 つぎの不等式を満たすパス $\phi_{j_1^* j_2^* \dots j_n^*}$ が唯一に存在するとき、このパスを最適パスと呼ぶ。

$$X_{j_1^* j_2^* \dots j_n^*} > X_{i_1 i_2 \dots i_n} \quad (5)$$

ただし、 $\max\{|j_1^* - i_1|, |j_2^* - i_2|, \dots, |j_n^* - i_n|\} > 0$ である。

さらに、定常環境はつぎの特性を持つものとする。

[N.1] 最適パス $\phi_{j_1^* j_2^* \dots j_n^*}$ 上の最下レベルにないある学習オートマトン $Z_{j_1^* j_2^* \dots j_{s-1}^*}$ において、次式が成立する場合が存在する。

$$\bar{\beta}_{j_1^* j_2^* \dots j_{s-1}^* j_s^*} \leq \bar{\beta}_{j_1^* j_2^* \dots j_{s-1}^* i_s}, \quad \exists i_s (i_s \neq j_s^*) \quad (6)$$

ここで、つぎの定理を得る。

《定理1》 最適パス $\phi_{j_1^* j_2^* \dots j_n^*}$ 上の任意のレベル s ($s = 1, 2, \dots, n$) において、つぎの関係式が成立する。

$$\begin{aligned} & \bar{\beta}_{j_1^* j_2^* \dots j_{s-1}^* j_s^*} + \bar{\beta}_{j_1^* j_2^* \dots j_{s-1}^* j_{s+1}^*} + \dots + \bar{\beta}_{j_1^* j_2^* \dots j_n^*} \\ & > \bar{\beta}_{j_1^* j_2^* \dots j_{s-1}^* i_s} + \bar{\beta}_{j_1^* j_2^* \dots j_{s-1}^* i_{s+1}} \\ & \quad + \dots + \bar{\beta}_{j_1^* j_2^* \dots j_{s-1}^* i_n} \quad (i_s \neq j_s^*) \quad (7) \end{aligned}$$

(証明) (4) 式および (5) 式より、 $s = 1$ に対して次式が成立する。

$$\begin{aligned} & \bar{\beta}_{j_1^*} + \bar{\beta}_{j_1^* j_2^*} + \dots + \bar{\beta}_{j_1^* j_2^* \dots j_n^*} \\ & > \bar{\beta}_{i_1} + \bar{\beta}_{i_1 i_2} + \dots + \bar{\beta}_{i_1 i_2 \dots i_n} \quad (8) \end{aligned}$$

ここで、(8) 式において $i_1 = j_1^*$ すると、

$$\begin{aligned} & \bar{\beta}_{j_1^*} + \bar{\beta}_{j_1^* j_2^*} + \dots + \bar{\beta}_{j_1^* j_2^* \dots j_n^*} \\ & > \bar{\beta}_{j_1^*} + \bar{\beta}_{j_1^* i_2} + \dots + \bar{\beta}_{j_1^* i_2 \dots i_n} \quad (9) \end{aligned}$$

となる。よって、(9) 式の両辺から $\bar{\beta}_{j_1^*}$ を引くと、 $s = 2$ に対して次式を得る。

$$\begin{aligned} & \bar{\beta}_{j_1^* j_2^*} + \bar{\beta}_{j_1^* j_2^* j_3^*} + \dots + \bar{\beta}_{j_1^* j_2^* \dots j_n^*} \\ & > \bar{\beta}_{j_1^* i_2} + \bar{\beta}_{j_1^* i_2 i_3} + \dots + \bar{\beta}_{j_1^* i_2 \dots i_n} \quad (i_2 \neq j_2^*) \quad (10) \end{aligned}$$

さらに、(10) 式において $i_2 = j_2^*$ とし、その両辺から $\bar{\beta}_{j_1^* j_2^*}$ を引くと、 $s = 3$ に対して次式を得る。

$$\begin{aligned} & \bar{\beta}_{j_1^* j_2^* j_3^*} + \bar{\beta}_{j_1^* j_2^* j_3^* j_4^*} + \dots + \bar{\beta}_{j_1^* j_2^* \dots j_n^*} \\ & > \bar{\beta}_{j_1^* j_2^* i_3} + \bar{\beta}_{j_1^* j_2^* i_3 i_4} + \dots + \bar{\beta}_{j_1^* j_2^* i_3 \dots i_n} \quad (i_3 \neq j_3^*) \quad (11) \end{aligned}$$

以下同様にすると、最適パス上の任意のレベル s ($s = 1, 2, \dots, n$) において (7) 式が成り立つ。 Q.E.D.

3. 報酬パラメータと階層構造学習アルゴリズム

本論文では、各レベルに S-モデル定常環境が存在する階層構造学習オートマトンについて考察し、最適パスの新しい定義の下で SL_{R-I} 強化法⁷⁾の概念に基づいて階層構造学習アルゴリズムを構築する。

ここで、特性 N.1 を満たす未知環境において SL_{R-I} 強化法を動作させることを考えよう。この強化法は平均報酬が最大の動作に収束する特性をもつ^{6), 8)}ので、学習オートマトン $Z_{j_1^* j_2^* \dots j_{s-1}^*}$ において、最適パス上にない動作 $\alpha_{j_1^* j_2^* \dots j_{s-1}^* i_s}$ に収束してしまつて、その結果、最適パスを見出せないことになる。この問題点を克服するためには、最適パス上のすべての学習オートマトン $Z_{j_1^* j_2^* \dots j_{s-1}^*}$ ($s = 1, 2, \dots, n$) において、最適パス上の動作 $\alpha_{j_1^* j_2^* \dots j_s^*}$ に対応する値が他の動作に対応する値よりも、平均値の意味で常に大きくなるようなパラメータを導入することが必要である。このパラメータを報酬パラメータと呼ぶこととし、一般にレベル s ($s = 1, 2, \dots, n$) の学習オートマトン $Z_{i_1 i_2 \dots i_{s-1}}$ における報酬パラメータを $\{\gamma_{i_1 i_2 \dots i_{s-1} 1}(t), \gamma_{i_1 i_2 \dots i_{s-1} 2}(t), \dots, \gamma_{i_1 i_2 \dots i_{s-1} r}(t)\}$ と書く。そして、各動作 $\alpha_{i_1 i_2 \dots i_s}$ が時刻 t までに得た報酬のうちで最新のものを最新報酬 $v_{i_1 i_2 \dots i_s}(t)$ と呼んで、

$$v_{i_1 i_2 \dots i_s}(t) = \beta_{i_1 i_2 \dots i_s} \quad (12)$$

と与える。このとき、つぎの関係が成り立つ。

$$E\{v_{i_1 i_2 \dots i_s}(t)\} = E\{\beta_{i_1 i_2 \dots i_s}\} = \bar{\beta}_{i_1 i_2 \dots i_s} \quad (13)$$

この最新報酬を用いて報酬パラメータ γ の各要素を、つぎのように与える^(注2)。

(i) 最下レベル n の動作に対する報酬パラメータ：

$$\gamma_{i_1 i_2 \dots i_n}(t) = v_{i_1 i_2 \dots i_n}(t) \quad (14)$$

(ii) レベル s の動作に対する報酬パラメータ ($s = 1, 2, \dots, n-1$)：

$$\begin{aligned} \gamma_{i_1 i_2 \dots i_s}(t) = & \frac{1}{n-s+1} [v_{i_1 i_2 \dots i_s}(t) \\ & + (n-s) \cdot \max_{i_{s+1}} \{\gamma_{i_1 i_2 \dots i_s i_{s+1}}(t)\}] \quad (15) \end{aligned}$$

この報酬パラメータ γ を用いて、以下のように学習アルゴリズムを構成する。

《階層構造学習アルゴリズム》

Step 1. 初期時刻を $t = 0$ とし、各学習オートマトンの動作確率を $p_{i_1 i_2 \dots i_s}(0) = 1/r$ ($i_s = 1, 2, \dots, r$; $s = 1, 2, \dots, n$) とおく。また、収束判定値 ρ を 1 に近い小数にとる。

Step 2. パス確率が最大のパス上の各動作の動作確率を $\bar{p}_{i_1 i_2 \dots i_s}(t)$ ($s = 1, 2, \dots, n$) とする。

(注2) (14) 式および (15) 式から明らかなように、ある動作の報酬パラメータは、その動作から始まり最下レベルに至る動作列 (たとえば $\alpha_{i_1 i_2 \dots i_s}, \alpha_{i_1 i_2 \dots i_s i_{s+1}}, \dots, \alpha_{i_1 i_2 \dots i_n}$) の各動作に対応する最新報酬の和 $v_{i_1 i_2 \dots i_s}(t) + v_{i_1 i_2 \dots i_s i_{s+1}}(t) + \dots + v_{i_1 i_2 \dots i_n}(t)$ に基づいて構成される。ところで、(13) 式に注意して最新報酬の和に期待値演算を施すと、 $\bar{\beta}_{i_1 i_2 \dots i_s} + \bar{\beta}_{i_1 i_2 \dots i_s i_{s+1}} + \dots + \bar{\beta}_{i_1 i_2 \dots i_n}$ のようにその動作列上の動作に対応する平均報酬の和となる。これは定理1の(7)式の両辺に現れるものと同様のものであることから、報酬パラメータを(14)式および(15)式によって与えると、定理1を以後の議論に適用することが可能となる。

そして、 $\bar{p}_{i_1 i_2 \dots i_s}(t) > \rho$, $\forall s$ のとき、本学習アルゴリズムを終了する。これ以外のときは、Step 3へ。

Step 3. 時刻 t において、パス $\phi_{j_1 j_2 \dots j_n}$ 上の各動作 $\{\alpha_{j_1 j_2 \dots j_s} : s = 1, 2, \dots, n\}$ が選択され、これに対して各レベルの環境から $\{\beta_{j_1 j_2 \dots j_s} : s = 1, 2, \dots, n\}$ なる返答が与えられたとする。

- (i) 選択された動作の最新報酬を $\{v_{j_1 j_2 \dots j_s}(t) = \beta_{j_1 j_2 \dots j_s} : s = 1, 2, \dots, n\}$ と更新する。
- (ii) (14) 式および (15) 式にしたがって報酬パラメータ $\{\gamma_{j_1 j_2 \dots j_s}(t) : s = 1, 2, \dots, n\}$ を計算する。
- (iii) パス $\phi_{j_1 j_2 \dots j_n}$ 上の各学習オートマトン $Z_{j_1 j_2 \dots j_{s-1}}$ ($s = 1, 2, \dots, n$) の動作確率分布を次式によって更新する^(注3)。

$$\left. \begin{aligned} p_{j_1 j_2 \dots j_{s-1} j_s}(t+1) &= p_{j_1 j_2 \dots j_{s-1} j_s}(t) \\ &+ \lambda \gamma_{j_1 j_2 \dots j_{s-1} j_s}(t) [1 - p_{j_1 j_2 \dots j_{s-1} j_s}(t)] \\ p_{j_1 j_2 \dots j_{s-1} i_s}(t+1) &= p_{j_1 j_2 \dots j_{s-1} i_s}(t) \\ &\times [1 - \lambda \gamma_{j_1 j_2 \dots j_{s-1} j_s}(t)] \quad (i_s \neq j_s) \end{aligned} \right\} \quad (16)$$

ただし、 λ は $0 < \lambda < 1$ なる実数である。

- (iv) $t = t + 1$ とし、Step 2へ。

なお、パス $\phi_{j_1 j_2 \dots j_n}$ 上にない学習オートマトンの最新報酬、報酬パラメータおよび動作確率の値はまったく変更されない。

4. 学習特性

定理 1 から、報酬パラメータに関してつぎの定理を得る。

《定理 2》 報酬パラメータ γ が (14) 式および (15) 式によって与えられるものとする。このとき、最適パス上の各学習オートマトン $Z_{j_1^* j_2^* \dots j_{s-1}^*}$ ($s = 1, 2, \dots, n$) において、つぎの関係式が成立する。

$$E\{\gamma_{j_1^* j_2^* \dots j_{s-1}^* j_s^*}(t)\} > E\{\gamma_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t)\}, \quad \forall t \\ (i_s = 1, 2, \dots, r; i_s \neq j_s^*) \quad (17)$$

(証明) 最適パス $\phi_{j_1^* j_2^* \dots j_n^*}$ 上の各学習オートマトンに着目する。まず、最下レベル n の学習オートマトン $Z_{j_1^* j_2^* \dots j_{n-1}^*}$ において、最適パス上の動作 $\alpha_{j_1^* j_2^* \dots j_{n-1}^* j_n^*}$ に対する報酬パラメータについて、(14) 式および (13) 式より次式が得られる。

$$E\{\gamma_{j_1^* j_2^* \dots j_{n-1}^* j_n^*}(t)\} = \bar{\beta}_{j_1^* j_2^* \dots j_{n-1}^* j_n^*} \quad (18)$$

また、 $\alpha_{j_1^* j_2^* \dots j_{n-1}^* j_n^*}$ 以外の動作 $\alpha_{j_1^* j_2^* \dots j_{n-1}^* i_n}$ ($i_n = 1, 2, \dots, r; i_n \neq j_n^*$) に対しては、つぎの関係式が成立する。すなわち、(14) 式より、

$$E\{\gamma_{j_1^* j_2^* \dots j_{n-1}^* i_n}(t)\} = \bar{\beta}_{j_1^* j_2^* \dots j_{n-1}^* i_n} \quad (19)$$

そして、定理 1 に注意すると、(18) 式および (19) 式より、次

式が成り立つ。

$$\begin{aligned} E\{\gamma_{j_1^* j_2^* \dots j_{n-1}^* j_n^*}(t)\} &= \bar{\beta}_{j_1^* j_2^* \dots j_{n-1}^* j_n^*} \\ &> \bar{\beta}_{j_1^* j_2^* \dots j_{n-1}^* i_n} = E\{\gamma_{j_1^* j_2^* \dots j_{n-1}^* i_n}(t)\} \\ &\quad (i_n = 1, 2, \dots, r; i_n \neq j_n^*) \quad (20) \end{aligned}$$

つぎに、学習オートマトン $Z_{j_1^* j_2^* \dots j_{n-1}^*}$ が結合されている動作 $\alpha_{j_1^* j_2^* \dots j_{n-1}^*}$ を持つレベル $(n-1)$ の学習オートマトン $Z_{j_1^* j_2^* \dots j_{n-2}^*}$ について考える。まず、動作 $\alpha_{j_1^* j_2^* \dots j_{n-2}^* j_{n-1}^*}$ に対して、(15) 式より、つぎの関係が成り立つ。

$$\begin{aligned} E\{\gamma_{j_1^* j_2^* \dots j_{n-1}^*}(t)\} &= \frac{1}{2} \left[\bar{\beta}_{j_1^* j_2^* \dots j_{n-1}^*} \right. \\ &\quad \left. + \max_{i_n} \left\{ E\{\gamma_{j_1^* j_2^* \dots j_{n-1}^* i_n}(t)\} \right\} \right] \quad (21) \end{aligned}$$

ここで、(20) 式に注意すると、

$$\max_{i_n} \left\{ E\{\gamma_{j_1^* j_2^* \dots j_{n-1}^* i_n}(t)\} \right\} = E\{\gamma_{j_1^* j_2^* \dots j_{n-1}^* j_n^*}(t)\} \quad (22)$$

であるので、(21) 式より次式を得る。

$$\begin{aligned} E\{\gamma_{j_1^* j_2^* \dots j_{n-1}^*}(t)\} &= \frac{1}{2} \left[\bar{\beta}_{j_1^* j_2^* \dots j_{n-1}^*} + E\{\gamma_{j_1^* j_2^* \dots j_{n-1}^* j_n^*}(t)\} \right] \\ &= \frac{1}{2} \left[\bar{\beta}_{j_1^* j_2^* \dots j_{n-1}^*} + E\{v_{j_1^* j_2^* \dots j_{n-1}^* j_n^*}(t)\} \right] \\ &= \frac{1}{2} \left[\bar{\beta}_{j_1^* j_2^* \dots j_{n-1}^*} + \bar{\beta}_{j_1^* j_2^* \dots j_{n-1}^* j_n^*} \right] \quad (23) \end{aligned}$$

また、 $\alpha_{j_1^* j_2^* \dots j_{n-2}^* j_{n-1}^*}$ 以外の動作 $\alpha_{j_1^* j_2^* \dots j_{n-2}^* i_{n-1}}$ ($i_{n-1} = 1, 2, \dots, r; i_{n-1} \neq j_{n-1}^*$) に対しては、つぎの関係式が成立する。(15) 式より、

$$\begin{aligned} E\{\gamma_{j_1^* j_2^* \dots j_{n-2}^* i_{n-1}}(t)\} &= \frac{1}{2} \left[\bar{\beta}_{j_1^* j_2^* \dots j_{n-2}^* i_{n-1}} \right. \\ &\quad \left. + \max_{i_n} \left\{ E\{\gamma_{j_1^* j_2^* \dots j_{n-2}^* i_{n-1} i_n}(t)\} \right\} \right] \\ &= \frac{1}{2} \left[\bar{\beta}_{j_1^* j_2^* \dots j_{n-2}^* i_{n-1}} \right. \\ &\quad \left. + \max_{i_n} \left\{ E\{v_{j_1^* j_2^* \dots j_{n-2}^* i_{n-1} i_n}(t)\} \right\} \right] \\ &= \frac{1}{2} \left[\bar{\beta}_{j_1^* j_2^* \dots j_{n-2}^* i_{n-1}} + \bar{\beta}_{j_1^* j_2^* \dots j_{n-2}^* i_{n-1} k_n} \right] \quad (24) \end{aligned}$$

ゆえに、定理 1 と (23) 式および (24) 式とより、

$$\begin{aligned} E\{\gamma_{j_1^* j_2^* \dots j_{n-2}^* j_{n-1}^*}(t)\} &> E\{\gamma_{j_1^* j_2^* \dots j_{n-2}^* i_{n-1}}(t)\}, \\ &\quad (i_{n-1} = 1, 2, \dots, r; i_{n-1} \neq j_{n-1}^*) \quad (25) \end{aligned}$$

よって、レベル $(n-1)$ の学習オートマトン $Z_{j_1^* j_2^* \dots j_{n-2}^*}$ において、(17) 式が成り立つ。

さらに、学習オートマトン $Z_{j_1^* j_2^* \dots j_{n-2}^*}$ が結合されている動作 $\alpha_{j_1^* j_2^* \dots j_{n-2}^*}$ を持つレベル $(n-2)$ の学習オートマトン $Z_{j_1^* j_2^* \dots j_{n-3}^*}$ について考える。まず、動作 $\alpha_{j_1^* j_2^* \dots j_{n-2}^*}$ に対して、(15) 式より、つぎの関係が成り立つ。

$$E\{\gamma_{j_1^* j_2^* \dots j_{n-2}^*}(t)\} = \frac{1}{3} \left[\bar{\beta}_{j_1^* j_2^* \dots j_{n-2}^*} \right]$$

(注3) 選択された動作以外の $r-1$ 個の動作の動作確率が第 2 式にしたがって減少し、その減少分の合計だけ選択された動作の動作確率が増加する (第 1 式)。

$$+2 \max_{i_{n-1}} \left\{ E\{\gamma_{j_1^* j_2^* \dots j_{n-2}^* i_{n-1}}(t)\} \right\} \quad (26)$$

ここで, (25) 式に注意すると,

$$\begin{aligned} \max_{i_{n-1}} \left\{ E\{\gamma_{j_1^* j_2^* \dots j_{n-2}^* i_{n-1}}(t)\} \right\} \\ = E\{\gamma_{j_1^* j_2^* \dots j_{n-2}^* j_{n-1}^*}(t)\} \end{aligned} \quad (27)$$

であるので, (23) 式と (26) 式とより次式を得る.

$$\begin{aligned} E\{\gamma_{j_1^* j_2^* \dots j_{n-2}^*}(t)\} \\ = \frac{1}{3} \left[\bar{\beta}_{j_1^* j_2^* \dots j_{n-2}^*} + 2E\{\gamma_{j_1^* j_2^* \dots j_{n-2}^* j_{n-1}^*}(t)\} \right] \\ = \frac{1}{3} \left[\bar{\beta}_{j_1^* j_2^* \dots j_{n-2}^*} + \bar{\beta}_{j_1^* j_2^* \dots j_{n-1}^*} + \bar{\beta}_{j_1^* j_2^* \dots j_n^*} \right] \end{aligned} \quad (28)$$

また, $\alpha_{j_1^* j_2^* \dots j_{n-3}^* j_{n-2}^*}$ 以外の動作 $\alpha_{j_1^* j_2^* \dots j_{n-3}^* i_{n-2}} (i_{n-2} = 1, 2, \dots, r; i_{n-2} \neq j_{n-2}^*)$ に対しては, つぎの関係式が成立する. すなわち, (15) 式を繰り返し用いることによって,

$$\begin{aligned} E\{\gamma_{j_1^* j_2^* \dots j_{n-3}^* i_{n-2}}(t)\} &= \frac{1}{3} \left[\bar{\beta}_{j_1^* j_2^* \dots j_{n-3}^* i_{n-2}} \right. \\ &\quad \left. + 2 \max_{i_{n-1}} \left\{ E\{\gamma_{j_1^* j_2^* \dots j_{n-3}^* i_{n-2} i_{n-1}}(t)\} \right\} \right] \\ &= \frac{1}{3} \left[\bar{\beta}_{j_1^* j_2^* \dots j_{n-3}^* i_{n-2}} \right. \\ &\quad \left. + 2 \max_{i_{n-1}} \left\{ \frac{1}{2} [E\{\gamma_{j_1^* j_2^* \dots j_{n-3}^* i_{n-2} i_{n-1}}(t)\} \right. \right. \\ &\quad \left. \left. + \max_{i_n} \{\gamma_{j_1^* j_2^* \dots j_{n-3}^* i_{n-2} i_{n-1} i_n}(t)\}] \right\} \right] \\ &= \frac{1}{3} \left[\bar{\beta}_{j_1^* j_2^* \dots j_{n-3}^* i_{n-2}} + \max_{i_{n-1}, i_n} \{\bar{\beta}_{j_1^* j_2^* \dots j_{n-3}^* i_{n-2} i_{n-1} i_n}\} \right. \\ &\quad \left. + \bar{\beta}_{j_1^* j_2^* \dots j_{n-3}^* i_{n-2} i_{n-1} i_n} \right] \\ &= \frac{1}{3} \left[\bar{\beta}_{j_1^* j_2^* \dots j_{n-3}^* i_{n-2}} + \bar{\beta}_{j_1^* j_2^* \dots j_{n-3}^* i_{n-2} k_{n-1}} \right. \\ &\quad \left. + \bar{\beta}_{j_1^* j_2^* \dots j_{n-3}^* i_{n-2} k_{n-1} k_n} \right] \end{aligned} \quad (29)$$

ゆえに, 定理1と (28) 式および (29) 式とより,

$$\begin{aligned} E\{\gamma_{j_1^* j_2^* \dots j_{n-3}^* j_{n-2}^*}(t)\} &> E\{\gamma_{j_1^* j_2^* \dots j_{n-3}^* i_{n-2}}(t)\}, \\ (i_{n-2} &= 1, 2, \dots, r; i_{n-2} \neq j_{n-2}^*) \end{aligned} \quad (30)$$

よって, レベル $(n-2)$ の学習オートマトン $Z_{j_1^* j_2^* \dots j_{n-3}^*}$ において, (17) 式が成り立つ.

以下同様にして, 一般にレベル s の学習オートマトン $Z_{j_1^* j_2^* \dots j_{s-1}^*}$ において, (17) 式が成り立つ. Q.E.D.

定理2の下で, 本階層構造学習アルゴリズムの学習特性についてつぎの定理が成り立つ.

《定理3》 特性N.1が成り立つものとする. このとき, 本階層構造学習アルゴリズムによって最適パスに到達する確率は, λ を十分小さくとることによっていくらでも1に近くすることができる.

(証明) 本階層構造学習アルゴリズムにおいては, 強化法(16)式が SL_{R-I} 強化法⁷⁾における報酬の値を報酬パラメータの値で置き換えたものであることに注意しよう. そして, 定理2から, 最適パス上の各学習オートマトンにおいて, 最適パス上の動作に対応する報酬パラメータの値がそれ以外の動作

に対応する報酬パラメータの値よりも, 平均値の意味で常に大きいことが言える. よって, 最適パスを $\phi_{j_1^* j_2^* \dots j_n^*}$ とすると, 文献6), 8)と同様の議論を行なうことによって, 最適パス上の各動作において次式が成立することを示すことができる(付録A参照).

$$\lim_{\lambda \rightarrow 0} \lim_{t \rightarrow \infty} E\{p_{j_1^* j_2^* \dots j_n^*}(t)\} = 1 \quad (31)$$

ここで, (31) 式を書き直すとつぎのようになる. すなわち, 任意の正数 ε に対して適当な正数 λ_0 を定めると, $|\lambda| < \lambda_0$ なる任意の λ に対して,

$$\lim_{t \rightarrow \infty} E\{p_{j_1^* j_2^* \dots j_n^*}(t)\} \geq 1 - \varepsilon \quad (32)$$

となる. また,

$$\lim_{t \rightarrow \infty} p_{j_1^* j_2^* \dots j_n^*}(t) = 0 \text{ or } 1 \text{ w.p.1} \quad (33)$$

を得る^{6), 8)}. よって, (32) 式と (33) 式より, $|\lambda| < \lambda_0$ なる λ に対して次式が成り立つ.

$$\begin{aligned} 1 - \varepsilon &\leq \lim_{t \rightarrow \infty} E\{p_{j_1^* j_2^* \dots j_n^*}(t)\} \\ &= 0 \times \text{Prob}\{\lim_{t \rightarrow \infty} p_{j_1^* j_2^* \dots j_n^*}(t) = 0\} \\ &\quad + 1 \times \text{Prob}\{\lim_{t \rightarrow \infty} p_{j_1^* j_2^* \dots j_n^*}(t) = 1\} \\ &= \text{Prob}\{\lim_{t \rightarrow \infty} p_{j_1^* j_2^* \dots j_n^*}(t) = 1\} \end{aligned} \quad (34)$$

すなわち, つぎを得る.

$$\begin{aligned} \text{Prob}\{\text{学習オートマトン } Z_{j_1^* j_2^* \dots j_{s-1}^*} \text{ において} \\ \text{最適パス上の動作が選択される}\} \\ = \text{Prob}\{\lim_{t \rightarrow \infty} p_{j_1^* j_2^* \dots j_s^*}(t) = 1\} \\ \geq 1 - \varepsilon \end{aligned} \quad (35)$$

このことから, 本階層構造学習アルゴリズムによって最適パスに到達する確率を Q と書くと, (35) 式より, $|\lambda| < \lambda_0$ なる λ に対して次式が成り立つ.

$$\begin{aligned} Q &\triangleq \text{Prob}\{\text{学習オートマトン } Z \text{ において} \\ &\quad \text{最適パス上の動作が選択される}\} \\ &\quad \times \text{Prob}\{\text{学習オートマトン } Z_{j_1^*} \text{ において} \\ &\quad \text{最適パス上の動作が選択される}\} \\ &\quad \times \text{Prob}\{\text{学習オートマトン } Z_{j_1^* j_2^*} \text{ において} \\ &\quad \text{最適パス上の動作が選択される}\} \\ &\quad \vdots \\ &\quad \times \text{Prob}\{\text{学習オートマトン } Z_{j_1^* j_2^* \dots j_{n-1}^*} \text{ において} \\ &\quad \text{最適パス上の動作が選択される}\} \\ &\geq (1 - \varepsilon)(1 - \varepsilon)(1 - \varepsilon) \cdots (1 - \varepsilon) \\ &\geq (1 - \varepsilon)^n \\ &\geq 1 - n\varepsilon \end{aligned} \quad (36)$$

ところで, ε の値はいくらでも小さくとれることに注意しよう.(このとき, λ_0 が小さくなるので λ も小さくとることが必要である.) すると, (36) 式の右辺はいくらでも1に近く

することができる。よって、本階層構造学習アルゴリズムによって最適パスに到達する確率は、 λ を十分小さくすることによっていくらでも 1 に近くすることができる。 Q.E.D.

5. 数値シミュレーション

本階層構造学習アルゴリズムによって、特性 N.1 が成り立つ未知環境においても確かに最適パスを見出せることを、数値シミュレーションによって確認する。階層構造学習オートマトンとしてはレベル数が $N = 10$ であるものを考え、また、それを構成する各学習オートマトンは 2 個の動作 ($r = 2$) を持つものとする。このとき、パスの総数は $r^N = 2^{10} = 1024$ 本であり、この中の最適パスに到達することが学習の目的となる。

時刻 t において、動作 $\alpha_{j_1 j_2 \dots j_{s-1} i_s}$ ($i_s = 1, 2, \dots, r$) に対する環境からの報酬 $\beta_{j_1 j_2 \dots j_{s-1} i_s}$ は、つぎの式にしたがって与えられるものとする：

$$\beta_{j_1 j_2 \dots j_{s-1} i_s} = \bar{\beta}_{j_1 j_2 \dots j_{s-1} i_s} + e_{j_1 j_2 \dots j_{s-1} i_s} \eta \quad (37)$$

ただし、 η は $[-0.5, 0.5]$ に値をとる一様乱数である。そして、係数 $e_{j_1 j_2 \dots j_{s-1} i_s}$ はすべての動作に対して 0.1 とした。ここでは、最適パスを $\phi_{1111111111}$ とし、そして、このパス上の各動作に対する平均報酬 (average reward) は Table 1 のようになっており、レベル 1~9 の各レベルにおいて特性 N.1 が成立している。すなわち、最適パス上の該当レベルの学習オートマトンにおいて、最適パス上の動作に対する平均報酬がもう一方の動作に対する平均報酬よりも小さくなっている。ただし、この Table 以外の平均報酬は $[0, 0.7]$ の一様乱数で与えた。

この問題に本階層構造学習アルゴリズムを適用する。100 回の実験を行なった結果、平均学習回数=8582.75 で 100 回とも最適パスを見出すことができた (ただし、 $\lambda = 0.011$ とした)。ここで比較のために、オリジナルの SL_{R-I} 強化法⁷⁾ を動作させた ((16) 式において γ の代わりに報酬 β をそのまま用いた) が、この場合には 100 回とも最適パス以外のパスに収束した。すなわち、特性 N.1 が成り立つ場合には、オリジナルの SL_{R-I} 強化法では最適パスをまったく発見できないことが確かめられた。

ところで、強化法に含まれる実数 λ の決定法についてであ

Table 1 Average rewards

Level	Average rewards	
1	$\bar{\beta}_1 = 0.92$	$\bar{\beta}_2 = 0.98$
2	$\bar{\beta}_{11} = 0.89$	$\bar{\beta}_{12} = 0.95$
3	$\bar{\beta}_{111} = 0.93$	$\bar{\beta}_{112} = 0.94$
4	$\bar{\beta}_{1111} = 0.88$	$\bar{\beta}_{1112} = 0.92$
5	$\bar{\beta}_{11111} = 0.91$	$\bar{\beta}_{11112} = 0.93$
6	$\bar{\beta}_{111111} = 0.87$	$\bar{\beta}_{111112} = 0.89$
7	$\bar{\beta}_{1111111} = 0.90$	$\bar{\beta}_{1111112} = 0.96$
8	$\bar{\beta}_{11111111} = 0.91$	$\bar{\beta}_{11111112} = 0.93$
9	$\bar{\beta}_{111111111} = 0.93$	$\bar{\beta}_{111111112} = 0.94$
10	$\bar{\beta}_{1111111111} = 0.90$	$\bar{\beta}_{1111111112} = 0.81$

Table 2 Simulation results

λ	Success rate (%)	Average number of iterations
0.008	100	11759.20
0.009	100	10623.48
0.010	100	8996.69
0.011	100	8582.75
0.012	99	7644.86
0.013	98	6693.23

るが、現在の所、解析的手法は得られていない。そこで、 λ の値と学習性能との関係を数値的に見ることにする。すなわち、 λ の値を種々変化させて本学習アルゴリズムによる学習を行ない、その結果を Table 2 に示す。これより λ は小さな値に取る必要があることは明らかであるが、さらに、 λ を大きくすると学習回数は少なくなるが正解率が下がり、逆に小さくすると学習回数は多くなるが正解率は 100% となることもわかる。このことから、適切な λ の値が存在すると思われるので、その決定法の確立が望まれる。なお、 $\lambda = 0.011$ は最も少ない平均学習回数で 100 回とも最適パスを見出したときの値である。

6. む す び

本論文では各レベルに S-モデル定常環境をもつ階層構造学習オートマトンについて考察し、各動作に対応する平均報酬の算術平均を用いて最適パスの定義を与えることを提案した。従来は平均報酬の積を用いて定義していたのであるが、本論文の定義の方がその意味するところがわかりやすいと考えている。さらに、対象としている定常環境は特性 N.1 が成り立つような一般的なものであり、その様な環境中においても良好な学習特性を持つ L_{R-I} 型階層構造学習アルゴリズムを、報酬パラメータ γ を導入することによって構築した。そして、本階層構造学習アルゴリズムによって最適パスが見出される確率はいくらでも 1 に近くすることができることを理論的に示した。さらに、数値シミュレーションによって、本階層構造学習アルゴリズムが確かに最適パスを見出すことを確認した。

今後は、本論文の最適パスの定義の下で、可変階層構造学習アルゴリズムを構築することが望まれる。

参 考 文 献

- 1) M.A.L. Thathachar and K.R. Ramakrishnan: A Hierarchical System of Learning Automata, IEEE Trans. Systems, Man, and Cybernetics, SMC-11-3, 236/241 (1981)
- 2) 馬場, 最上, 三宅: 非定常複数教師環境における階層構造確率オートマトンの学習特性, 電子通信学会論文誌, J69-D-11, 1724/1731 (1986)
- 3) 最上義夫: 一人教師定常環境中で動作する可変階層構造学習オートマトン, 計測自動制御学会論文集, 25-9, 1108/1025 (1989)
- 4) M.F. Norman: On the Linear Model with Two Absorbing Barriers, J. Mathematical Psychology, 5, 225/241 (1968)
- 5) R. Viswanathan and K.S. Narendra: A Note on the Linear Reinforcement Scheme for variable-Structure Stochas-

- tic Automata, IEEE Trans. Systems, Man, and Cybernetics, **SMC-2-2**, 292/294 (1972)
- 6) Y. Sawaragi and N. Baba: A Note on the Learning Behavior of Variable-Structure Stochastic Automata, IEEE Trans. Systems, Man, and Cybernetics, **SMC-3-6**, 644/647 (1973)
- 7) R. Viswanathan and Kumpati S. Narendra: Stochastic Automata Models with Applications to Learning Systems, IEEE Trans. Systems, Man, and Cybernetics, **SMC-3-1**, 107/111 (1973)
- 8) Norio Baba: On the Learning Behavior of the $SLR-I$ Reinforcement Scheme for Stochastic Automata, IEEE Trans. Systems, Man, and Cybernetics, **SMC-6-8**, 580/582 (1976)

《付 録》

A. (31) 式導出の概略⁸⁾

まず, (16) 式より, 次式を得る. ただし, $P(t) = (p_{j_1^* j_2^* \dots j_{s-1}^* 1}(t), p_{j_1^* j_2^* \dots j_{s-1}^* 2}(t), \dots, p_{j_1^* j_2^* \dots j_{s-1}^* r}(t))'$.

$$\begin{aligned} E\{p_{j_1^* j_2^* \dots j_s^*}(t+1)|P(t)\} &= \left[p_{j_1^* j_2^* \dots j_s^*}(t) \right. \\ &\quad + \lambda E\{\gamma_{j_1^* j_2^* \dots j_s^*}(t)\{1 - p_{j_1^* j_2^* \dots j_s^*}(t)\}\} p_{j_1^* j_2^* \dots j_s^*}(t) \\ &\quad + \sum_{i_s \neq j_s^*} p_{j_1^* j_2^* \dots j_s^*}(t) \left[1 - \lambda E\{\gamma_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t)\} \right. \\ &\quad \quad \times p_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t) \\ &\quad \left. \left. = p_{j_1^* j_2^* \dots j_s^*}(t) + \lambda p_{j_1^* j_2^* \dots j_s^*}(t) \sum_{i_s \neq j_s^*} [E\{\gamma_{j_1^* j_2^* \dots j_s^*}(t)\} \right. \right. \\ &\quad \quad \left. \left. - E\{\gamma_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t)\}] p_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t) \right] \right. \\ &\quad \left. \geq p_{j_1^* j_2^* \dots j_s^*}(t) \right] \quad (A.1) \end{aligned}$$

つぎに,

$$h_{x,\lambda}(p_{i_1 i_2 \dots i_s}(t)) = \frac{e^{\frac{x}{\lambda} p_{i_1 i_2 \dots i_s}(t)} - 1}{e^{\frac{x}{\lambda}} - 1} \quad (A.2)$$

において (ただし, $x > 0$),

$$p'_s(t) = \sum_{i_s \neq j_s^*} p_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t) = 1 - p_{j_1^* j_2^* \dots j_s^*}(t) \quad (A.3)$$

とすると, 次式を得る.

$$E\{h_{x,\lambda}(p'_s(t+1))|P(t)\} = \frac{1}{e^{\frac{x}{\lambda}} - 1} \{e^{\frac{x}{\lambda} p'_s(t)} g(t) - 1\} \quad (A.4)$$

ただし,

$$\begin{aligned} g(t) &= E\left\{ e^{-x \gamma_{j_1^* j_2^* \dots j_s^*}(t) [1 - p_{j_1^* j_2^* \dots j_s^*}(t)]} p_{j_1^* j_2^* \dots j_s^*}(t) \right. \\ &\quad + \sum_{i_s \neq j_s^*} E\left\{ e^{x \gamma_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t) p_{j_1^* j_2^* \dots j_s^*}(t)} \right. \\ &\quad \quad \times p_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t) \left. \right\} \quad (A.5) \end{aligned}$$

ここで, (A.5) 式の指数関数に Taylor の定理を適用 (2 次の項まで) し整理すると, 十分小さな $x > 0$ に対して次式を得る.

$$g(t) \leq 1 - p_{j_1^* j_2^* \dots j_s^*}(t)x \left[\sum_{i_s \neq j_s^*} (E\{\gamma_{j_1^* j_2^* \dots j_s^*}(t)\} \right.$$

$$\begin{aligned} &\quad - E\{\gamma_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t)\} p_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t) - \frac{x}{2!} \\ &\quad - \frac{x}{2!} \sum_{i_s \neq j_s^*} E\{\gamma_{j_1^* j_2^* \dots j_s^*}^2(t) e^{x \gamma_{j_1^* j_2^* \dots j_s^*}(t) p_{j_1^* j_2^* \dots j_s^*}(t)} \} \\ &\quad \quad \times p_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t) \left. \right] \quad (A.6) \end{aligned}$$

よって, 十分小さな $x > 0$ に対して, (A.6) 式より,

$$g(t) \leq 1 \quad (A.7)$$

となるので, (A.4) 式より次式を得る.

$$E\{h_{x,\lambda}(p'_s(t+1))|P(t)\} \leq h_{x,\lambda}(p'_s(t)) \quad (A.8)$$

ゆえに, (A.1) 式と (A.8) 式とから, (31) 式が成立する⁶⁾.

[著 者 紹 介]

最 上 義 夫 (正会員)

1975 年, 京都工芸繊維大学大学院工芸学研究科修士課程修了. 同年徳島大学工学部情報工学科助手, 97 年知能情報工学科講師, 現在に至る. 学習オートマトン, 強化学習, 確率分布定数システムの観測器・制御器の最適配置問題の研究に従事. 京都大学博士 (工学). IEEE, システム制御情報学会, 電子情報通信学会会員.

馬 場 則 夫 (正会員)

1974 年, 京都大学大学院博士課程修了. 同年徳島大学工学部情報工学科助手, 75 年同講師, 81 年同助教授, 88 年知能情報工学科教授, 90 年大阪教育大学教育学部教授, 現在に至る. 学習オートマトン, ゲーミング, 数理計画法, 環境工学, ロボット工学, ニューラルネット等の研究に従事. 工学博士. IEEE, システム制御情報学会, 日本ロボット学会, 日本オペレーションズ・リサーチ学会会員.

塩 尻 晃

2001 年, 徳島大学大学院工学研究科博士前期課程知能情報工学専攻入学. 現在に至る. 可変階層構造学習オートマトンの研究に従事. システム制御情報学会会員.

田 上 隆 徳 (正会員)

1996 年, 徳島大学大学院工学研究科システム工学専攻博士後期課程修了. 同年香川短期大学経営情報科講師, 99 年阿南工業高等専門学校一般教科講師, 現在に至る. 遺伝的アルゴリズムなどの最適化の手法に関する研究に従事. 徳島大学博士 (工学). システム制御情報学会, IEEE の各会員.