各レベルに複数定常環境をもつ可変階層構造学習オートマトン

最上 義夫*・馬場 則夫**

Variable Hierarchical Structure Learning Automata with Stationary Random Environments at Each Level

Yoshio MOGAMI* and Norio BABA**

In this paper, a learning algorithm is constructed for variable hierarchical structure learning automata with P-model stationary random environments at each level. The learning property of our algorithm is considered theoretically, and it is proved that the probability finding the optimal objective path can be approached 1 as much as possible by using our algorithm. In numerical simulation, the usefulness of our algorithm is shown.

Key Words: variable hierarchical structure learning automata, P-model stationary random environments, reward parameter, learning algorithm

1. ま え が き

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り、大規模かつ複雑なシステムへの適用を目的として、それを木構造に組み合わせた階層構造学習オートマトンが注目されている。階層構造学習オートマトンは、Thathachar and Ramakrishnan¹⁾によって最初に提案され、その後、種々の研究が行われている^{2)~6)}。ところで、ある問題に階層構造学習オートマトンを適用するときはその出力となり得るもの、すなわち、問題の解となり得るものを学習目標と呼び、また、それらの集合を目標集合と呼ぶことにすると、従来行われてきた研究はいずれも、目標集合の要素が学習開始時にすべて与えられて確定しており、それらに対する reward(評価)を与える確率のみが未知である場合、すなわち、所定の学習を遂行するためにはどのような階層構造学習オートマトンを構成すればよいかということが、事前に明確である場合について考察している。

しかし、実際の問題に適用することを考えるとき、rewardを与える確率のみならず目標集合の要素数および内容もが未知であるような環境中において動作する階層構造学習オートマトンについても考察することが必要である。このような問題に対して階層構造学習オートマトンを適用する場合、既知の情報に基づいて構成される階層構造学習オートマトンから学習を開始し、徐々に階層構造を変更しながら学習を行うこ

とが考えられる。これに対して、最上⁷⁾は可変階層構造学習オートマトンを提案し、その学習アルゴリズムを構築するとともに学習特性について考察したが、より広範囲の問題を考えるならば、複数の未知環境が存在する場合⁸⁾においても可変階層構造学習オートマトンを構築することが重要になる。

そこで本論文では、各レベルにP-モデル複数定常環境をもつ可変階層構造学習オートマトンについて考察し、Thathachar and Sastry^{3)~5)}が提案した Estimator Algorithm を修正することによって可変階層構造学習アルゴリズムを構築するとともに、その学習特性について検討する。

2. 複数環境中で動作する可変階層構造学習オートマトン

各レベルに複数定常環境を持つ可変階層構造学習オートマトンを Fig. 1に示す。まず構造であるが、各学習オートマトンはそれぞれ r 個の動作を持ち、各動作の下にはそれぞれ 1 個の学習オートマトンが結合可能である。よって、レベル s には r^{s-1} 個の学習オートマトンが存在可能であり、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 には r_s 個の未知環境 $\{R_s^1, R_s^2, \dots, R_s^{r_s}\}$ が対応している。ここで、レベル 1 から最下レベル N に至る動作の列を $\{\alpha_{j_1}, \alpha_{j_1 j_2}, \dots, \alpha_{j_1 j_2 \dots j_N}\}$ とするとき、これをパスと呼び $\phi_{j_1 j_2 \dots j_N}$ と書く。そして、パス上の動作の動作確率(その動作が選択される確率)の積をパス確率といい $\pi_{j_1 j_2 \dots j_N}(t)$ と表す。さらに、あるパスのパス確率が収束判定値(事前に与えられている)を越えたとき、可変階層構造学習オートマトンはそのパスに収束したという。

ところで、いま、何らかの方法で目標集合の要素が事前に

* 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

* The University of Tokushima, Tokushima

** Osaka Educational University, Kashiwara

(Received July 7, 1999)

(Revised December 20, 1999)

Fig. 1 Variable hierarchical structure learning automata

確定できたとすると、すべての学習目標にそれぞれパスを割り当てた完全な形の階層構造が構成できる。本論文では、このような階層構造を完全構造と呼ぶ。そして、完全構造から1個以上の学習オートマトンが欠けた階層構造を部分構造と呼ぶ。さらに、完全構造上のパスを完全パス、部分構造上のパスで完全パスではないものを部分パス、また、学習目標を出力とするパスを目標パス、最良の学習目標を出力とするパスを最良目標パスと呼ぶ（目標パスはすべて完全パスである）。そして、次の条件C.1～C.3が成立するものとする。

- [C.1] 目標集合は、学習開始時には確定していない。
- [C.2] 最良目標は唯一に存在し、最良目標パスは完全構造の最適パスである。
- [C.3] 可変階層構造学習オートマトンは、収束したパスが目標パスであるかどうかは認識できる（注1）。

すなわち本論文では、条件C.1～C.3が成り立つ場合において、各レベルにP-モデル複数定常環境をもつ可変階層構造学習オートマトンについて考察することになる。

つぎに、その動作について述べる。

(a) 既知の初期構造 Σ^0 として n_0 レベル階層構造学習オートマトンを構成し、学習を開始する (Fig. 1では、 $n_0 = 1$)。初期時刻 $t = 0$ における動作確率は、すべてのレベルの学習オートマトンについて等しいとする（つまり、 $1/r$ となる）。さて、

（注1）移動ロボットの構造未知迷路通過問題⁷⁾を考える。この問題では、迷路を通過できる移動ルートが学習目標であるが、条件C.3は、ある移動ルートに従って移動した結果、迷路を通過できたかどうか移動ロボットに分かるということに対応している。

今、レベル1の学習オートマトン Z が時刻 t において、動作確率分布 $(p_1(t), p_2(t), \dots, p_r(t))'$ に基づいて動作 α_{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、レベル1の r_1 個の未知環境 $\{R_1^1, R_2^1, \dots, R_{r_1}^1\}$ はそれぞれ reward 確率 $\{d_{j_1}^1, d_{j_2}^1, \dots, d_{j_{r_1}}^1\}$ に基づいて r_1 個の reward $\{\beta_1^1(t), \beta_2^1(t), \dots, \beta_{r_1}^1(t)\} = \{\beta_{j_1}^1, \beta_{j_2}^1, \dots, \beta_{j_{r_1}}^1\}$ を生成しレベル1の学習オートマトン Z に与える。すると、 Z はこの reward を用いて学習アルゴリズムに従って動作確率分布を更新する。そして、時刻を $t+1$ とし Z は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た結果、ある時刻において可変階層構造学習オートマトンがパス ϕ_1 に収束したとする。

(b) ここで、 ϕ_1 が目標パスではないとすると、可変階層構造学習オートマトンは α_1 の下に新たに1個の学習オートマトン Z_1 を追加して階層構造を Σ^1 に更新し学習を継続する。すなわち、時刻 t においてレベル1の学習オートマトン Z が動作し、出力を未知環境に出力し、未知環境から reward を得る。そして、この返答に基づいて Z はその動作確率分布を更新する。このとき、レベル1で選択された動作 $\alpha_1(t) = \alpha_{j_1}$ （ここでは α_1 ）の下に学習オートマトンが存在する場合、レベル2の学習オートマトン Z_{j_1} （ここでは Z_1 ）が動作し、 Z と同様、時刻 t における動作確率分布 $(p_{j_11}(t), p_{j_12}(t), \dots, p_{j_1r}(t))'$ に基づいて動作 $\alpha_2(t) = \alpha_{j_1j_2}$ を選択しレベル2の未知環境 $\{R_1^2, R_2^2, \dots, R_{r_2}^2\}$ へ出力する。この出力に対して、レベル2の未知環境はそれぞれ reward 確率 $\{d_{j_1j_2}^1, d_{j_1j_2}^2, \dots, d_{j_1j_{r_2}}^2\}$ に基づいて r_2 個の reward $\{\beta_1^2(t), \beta_2^2(t), \dots, \beta_{r_2}^2(t)\} = \{\beta_{j_1j_2}^2, \beta_{j_1j_3}^2, \dots, \beta_{j_1j_{r_2}}^2\}$ を生成しレベル2の学習オートマトン Z_{j_1} に与える。そして、この reward に基づいて Z_{j_1} もその動作確率分布を更新し、時刻を $t+1$ として 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た結果、ある時刻においてパス ϕ_{1r} に収束したとする。

(c) ϕ_{1r} も目標パスではないとすると、可変階層構造学習オートマトンは α_{1r} の下に新たに1個の学習オートマトン Z_{1r} を追加して構造 Σ^2 を構成し、学習を継続する。

以上のように学習が進められて行き、目標パスに収束すると可変階層構造学習オートマトンは学習を終了する。ここで、初期構造から m 回の変更を経た構造を Σ^m ($m = 0, 1, \dots$) と記述し、構造 Σ^m のパスの総数を Ψ^m 、Fig. 1のように階層構造を図示したときに左から数えて k 番目のパスを $\phi_{i_1 i_2 \dots i_{N_k^m}}$ ($k = 1, 2, \dots, \Psi^m$) と記述する（ただし、 N_k^m はこのパスのレベル数である）。そして、これに対応するパス確率は

$$\pi_{i_1 i_2 \dots i_{N_k^m}}(t) = p_{i_1}(t) p_{i_1 i_2}(t) \dots p_{i_1 i_2 \dots i_{N_k^m}}(t) \quad (1)$$

である。また、構造 Σ^m 上のパス確率の最大値を Σ^m 上の最大パス確率と呼んで $\tilde{\pi}_{i_1 i_2 \dots i_{N_k^m}}(t)$ と記述し、これに対応するパスを Σ^m 上の最大パス $\tilde{\phi}_{i_1 i_2 \dots i_{N_k^m}}$ と呼ぶ。

3. P-モデル複数定常環境

各レベルの未知環境として、本論文では、P-モデル複数定常環境を考える。すなわち、レベル s ($s = 1, 2, \dots, N_k^m$) には r_s 個の定常環境が存在し、時刻 t におけるレベル s の出力 $\alpha_s(t) = \alpha_{j_1 j_2 \dots j_s}$ に対して、reward 確率 $d_{j_1 j_2 \dots j_s}^i$ ($i = 1, 2, \dots, r_s$) で $\beta_s^i(t) = \beta_{j_1 j_2 \dots j_s}^i = 1$ (成功) の reward が、また、確率 $1 - d_{j_1 j_2 \dots j_s}^i$ で $\beta_s^i(t) = \beta_{j_1 j_2 \dots j_s}^i = 0$ (失敗) の reward が合計 r_s 個与えられることになる。ここで、

$$\bar{\beta}_{j_1 j_2 \dots j_s} = \frac{\beta_{j_1 j_2 \dots j_s}^1 + \beta_{j_1 j_2 \dots j_s}^2 + \dots + \beta_{j_1 j_2 \dots j_s}^{r_s}}{r_s} \quad (2)$$

を、出力 $\alpha_{j_1 j_2 \dots j_s}$ に対する平均 reward と呼び、また、

$$e_{j_1 j_2 \dots j_s} = \frac{d_{j_1 j_2 \dots j_s}^1 + d_{j_1 j_2 \dots j_s}^2 + \dots + d_{j_1 j_2 \dots j_s}^{r_s}}{r_s} \quad (3)$$

を、平均 reward 確率と呼ぶことにする。さらに、

$$X_{j_1 j_2 \dots j_{N_k^m}} = e_{j_1} e_{j_2} \dots e_{j_{N_k^m}} \quad (4)$$

を、パス $\phi_{j_1 j_2 \dots j_{N_k^m}}$ に対するパス平均 reward 確率と呼ぶことにする。

本論文では、定常環境が次の特性を持つ場合について考察する。

[N.1] 構造 Σ^m において、パス平均 reward 確率が最大であるパス $\phi_{j_1^* j_2^* \dots j_{N_k^m}^*}$ を構造 Σ^m の最適パスと呼ぶ。すなわち、つぎの不等式を満足する。

$$X_{j_1^* j_2^* \dots j_{N_k^m}^*} > X_{i_1 i_2 \dots i_{N_k^m}}, \quad (k = 1, 2, \dots, \Psi^m, k \neq a) \quad (5)$$

このとき、最適パスが唯一に存在する。

ここで、完全構造を Σ^P とするとき、 Σ^P の最適パスを最良目標パス $\phi_{h_1^* h_2^* \dots h_{N_a^P}^*}$ と言う。すなわち、次式が成立する。

$$X_{h_1^* h_2^* \dots h_{N_a^P}^*} > X_{i_1 i_2 \dots i_{N_k^P}}, \quad (k = 1, 2, \dots, \Psi^P, k \neq a) \quad (6)$$

そして、可変階層構造学習オートマトンの学習の目的は、この最良目標パスを見出すことである。さらに、定常環境はつぎの特性を持つものとする。

[N.2] 構造 Σ^m の最適パス $\phi_{j_1^* j_2^* \dots j_{N_k^m}^*}$ 上のある学習オートマトン $Z_{j_1^* j_2^* \dots j_{s-1}^*}$ において、最適パス上の動作に対する平均 reward 確率が、他のある動作に対する平均 reward 確率よりも小さいか等しい。すなわち、次式が成立する。

$$e_{j_1^* j_2^* \dots j_{s-1}^*} \leq e_{j_1^* j_2^* \dots j_{s-1}^* i_s}, \quad \exists i_s (i_s \neq j_s^*) \quad (7)$$

4. reward パラメータと可変階層構造学習アルゴリズム

Thathachar and Sastry によって提案された Estimator Algorithm は、定常未知環境から得る reward を用いて未知環境の reward 確率のサンプル平均値を計算し、この値を学習に用いるものであり、また、このサンプル平均値が最大の動作に収束するという学習特性を持っている。そして、reward

確率のサンプル平均値の大小関係が reward 確率の大小関係と同様に成り立つことに着目すると、収束した動作が最大の reward 確率を持つ動作、すなわち、最適動作ということになる。

ところで本論文では、この Estimator Algorithm の概念に基づいて可変階層構造学習アルゴリズムを構築する。このとき、reward 確率 $d_{j_1 j_2 \dots j_s}^i$ の時刻 t におけるサンプル平均値を $\hat{d}_{j_1 j_2 \dots j_s}^i(t)$ と書いて次式で計算する。

$$\hat{d}_{j_1 j_2 \dots j_s}^i(t) = \frac{w_{j_1 j_2 \dots j_s}^i(t)}{z_{j_1 j_2 \dots j_s}(t)} \quad (j_s = 1, 2, \dots, r) \quad (8)$$

ただし、 $z_{j_1 j_2 \dots j_s}(t)$ は、レベル s の出力 $\alpha_{j_1 j_2 \dots j_s}$ が時刻 t までに選択された回数であり、また、 $w_{j_1 j_2 \dots j_s}^i(t)$ はその出力に対して時刻 t までに i 番目の環境が与えた reward の和である。

そして、(3) 式で定義した平均 reward 確率の時刻 t におけるサンプル平均値を

$$\begin{aligned} \hat{e}_{j_1 j_2 \dots j_s}(t) &= \frac{\hat{d}_{j_1 j_2 \dots j_s}^1(t) + \hat{d}_{j_1 j_2 \dots j_s}^2(t) + \dots + \hat{d}_{j_1 j_2 \dots j_s}^{r_s}(t)}{r_s} \end{aligned} \quad (9)$$

と定義して、これを reward 確率のサンプル平均値の代わりに学習に用いることにすると、Estimator Algorithm は $\hat{e}_{j_1 j_2 \dots j_s}(t)$ が最大の動作に収束することになるであろう。

ここで、未知環境が特性 N.2 を満たすことに注意しよう。これは、構造 Σ^m の最適パス上のある学習オートマトン $Z_{j_1^* j_2^* \dots j_{s-1}^*}$ において、最適パス上の動作 $\alpha_{j_1^* j_2^* \dots j_{s-1}^* j_s^*}$ に対応する平均 reward 確率 $e_{j_1^* j_2^* \dots j_{s-1}^* j_s^*}$ が、これ以外のある動作 $\alpha_{j_1^* j_2^* \dots j_{s-1}^* i_s}$ に対応する平均 reward 確率 $e_{j_1^* j_2^* \dots j_{s-1}^* i_s}$ よりも小さいことを述べている。このような未知環境において Estimator Algorithm を動作させると、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、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、構造 Σ^m の最適パス上のすべての学習オートマトン $Z_{j_1^* j_2^* \dots j_{s-1}^*}$ ($s = 1, 2, \dots, N_a^m$) 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 reward パラメータと呼ぶこととし、一般にレベル s ($s = 1, 2, \dots, N_k^m$) の学習オートマトン $Z_{i_1 i_2 \dots i_{s-1}}$ における reward パラメータを $\gamma_{i_1 i_2 \dots i_{s-1} i_s}(t)$ ($i_s = 1, 2, \dots, r$) と書いて、次のように与える。

(i) 最下レベル N_k^m の動作に対する reward パラメータ：

$$\gamma_{i_1 i_2 \dots i_{N_k^m}}(t) = \hat{e}_{i_1 i_2 \dots i_{N_k^m}}(t) \quad (10)$$

(ii) レベル s の動作に対する reward パラメータ ($s = 1, 2, \dots, N_k^m - 1$)：

$$\gamma_{i_1 i_2 \dots i_s}(t) = \hat{e}_{i_1 i_2 \dots i_s}(t) \max_{l_{s+1}} \{\gamma_{i_1 i_2 \dots i_s l_{s+1}}(t)\} \quad (11)$$

(注意) (reward パラメータについて) 任意の動作 $\alpha_{i_1 i_2 \dots i_s}$ を通るすべてのパス ($\alpha_{i_1 i_2 \dots i_s}$ -パス群と呼ぶ) を考える。このとき、このパス群のパスそれぞれにパス平均 reward 確率のサンプル平均値 ((15) 式参照) が 1 つずつ対応しているが、この値はパス上の動作に対する平均 reward 確率のサンプル平均値 $\hat{e}$ の積によって与えられる。ところで、 $\alpha_{i_1 i_2 \dots i_s}$ -パス群において、積 $\hat{e}_{i_1}(t) \hat{e}_{i_1 i_2}(t) \times \dots \times \hat{e}_{i_1 i_2 \dots i_{s-1}}(t)$ は共通であるので、これを除いた積は、一般に、 $\hat{e}_{i_1 i_2 \dots i_s}(t) \hat{e}_{i_1 i_2 \dots i_{s+1}}(t) \hat{e}_{i_1 i_2 \dots i_{s+1} i_{s+2}}(t) \times \dots \times \hat{e}_{i_1 i_2 \dots i_{s+1} i_{s+2} \dots i_{n_k^m-1} i_{n_k^m}}(t)$ と書ける。そして、(10) 式および (11) 式より、この積のレベル $n_k^m - 1$ からレベル $s + 1$ までの各レベルで最大化を行うと動作 $\alpha_{i_1 i_2 \dots i_s}$ に対する reward パラメータ $\gamma_{i_1 i_2 \dots i_s}(t)$ が得られる。すなわち、reward パラメータ $\gamma_{i_1 i_2 \dots i_s}(t)$ は、 $\alpha_{i_1 i_2 \dots i_s}$ -パス群に対応するパス平均 reward 確率のサンプル平均値の中の最大値からこのパス群において共通な積である $\hat{e}_{i_1}(t) \hat{e}_{i_1 i_2}(t) \times \dots \times \hat{e}_{i_1 i_2 \dots i_{s-1}}(t)$ を除いたものである。

この reward 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。

《 可変階層構造学習アルゴリズム 》

Step 1. 初期時刻 $t = 0$ として、初期構造 Σ^0 (N^0 レベル階層構造学習オートマトン) を構成する ($m = 0, \Psi^0 = r^{N^0}$, $p_{j_1 j_2 \dots j_s}(0) = 1/r$ ($s = 1, 2, \dots, N^0$)). また、収束判定値 ρ_0 (1 に近い小数) を与える。

Step 2. 最大パス $\phi_{i_1 i_2 \dots i_{N_k^m}}$ 上の各動作の動作確率を $\bar{p}_{i_1 i_2 \dots i_s}(t)$ ($s = 1, 2, \dots, N_k^m$) とする。

(i) $\bar{p}_{i_1 i_2 \dots i_s}(t) > \rho_m, \forall s$ のとき：最大パスが目標パスであるならば、本学習アルゴリズムを終了する。それ以外のときは Step 4 へ。

(ii) (i) 以外のとき：Step 3 へ。

Step 3. 構造 Σ^m のもとで、下記の強化法を用いて学習を行う。

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

(i) パス $\phi_{j_1 j_2 \dots j_s \dots j_{N_k^m}}$ 上の各学習オートマトン $Z_{j_1 j_2 \dots j_{s-1}}$ ($s = 1, 2, \dots, N_k^m$) の動作確率分布を次式によって更新する。

$$\begin{aligned} p_{j_1 j_2 \dots j_{s-1} i_s}(t+1) &= p_{j_1 j_2 \dots j_{s-1} i_s}(t) \\ &\quad - \lambda_s [f(\gamma_{j_1 \dots j_{s-1} j_s}(t)) - f(\gamma_{j_1 \dots j_{s-1} i_s}(t))] \\ &\quad \times [u_{j_1 \dots j_{s-1} j_s}^{j_1 \dots j_{s-1} j_s}(t) p_{j_1 \dots j_{s-1} i_s}(t) \\ &\quad + u_{j_1 \dots j_{s-1} j_s}^{j_1 \dots j_{s-1} i_s}(t) \frac{(1 - p_{j_1 \dots j_{s-1} i_s}(t)) p_{j_1 \dots j_{s-1} j_s}(t)}{r - 1}] \\ &\quad (i_s \neq j_s) \end{aligned} \quad (12)$$

$$p_{j_1 j_2 \dots j_s}(t+1) = 1 - \sum_{\substack{i_s=1 \\ i_s \neq j_s}}^r p_{j_1 j_2 \dots j_{s-1} i_s}(t+1) \quad (13)$$

ただし、 λ_s は $0 < \lambda_s < 1$ を満たす実数であり、 f は $[0, 1] \rightarrow [0, 1]$ なる単調増加関数、 u は

$$u_{j_1 \dots j_{s-1} i_s}^{j_1 \dots j_{s-1} j_s}(t) = \begin{cases} 1, & \gamma_{j_1 \dots j_{s-1} j_s}(t) > \gamma_{j_1 \dots j_{s-1} i_s}(t) \\ 0, & \gamma_{j_1 \dots j_{s-1} j_s}(t) \leq \gamma_{j_1 \dots j_{s-1} i_s}(t) \end{cases} \quad (14)$$

なる関数である。

(ii) $t = t + 1$ とし、(8) 式によって reward 確率のサンプル平均値 $\{\hat{d}_{j_1 j_2 \dots j_s}^1(t), \hat{d}_{j_1 j_2 \dots j_s}^2(t), \dots, \hat{d}_{j_1 j_2 \dots j_s}^{r_s}(t) : s = 1, 2, \dots, N_k^m\}$ を求める。

(iii) (9) 式、(10) 式および (11) 式に従って reward パラメータ $\{\gamma_{j_1 j_2 \dots j_s}(t) : s = 1, 2, \dots, N_k^m\}$ を計算して、Step 2 へ。

なお、パス $\phi_{j_1 j_2 \dots j_{N_k^m}}$ 上にない学習オートマトンの動作確率分布および reward パラメータの値は変更されない。

Step 4. Σ^m 上の最大パスの下に学習オートマトンを 1 個追加することによって、構造を Σ^m から Σ^{m+1} に更新する。

Step 5. 新構造 Σ^{m+1} 上の全てのパスのパス確率を $1/\Psi^{m+1}$ と置く。

Step 6. パス確率をもとにパス上の各学習オートマトンの動作確率を計算し、この値を新構造での初期動作確率とする。 $m = m + 1$ とし、収束判定値 ρ_m (1 に近い小数) を与えて Step 2 へ。

(注意) (文献 7) 参照) 本論文では条件 C.1 が成り立つ問題を考察しており、そこでは階層構造を適宜更新しつつ学習を進めることが必要となる。本可変階層構造学習アルゴリズムはこのことに対応しており、構造 Σ^m の下で学習を進める部分 (Step 3) と階層構造を更新する部分 (Step 4~Step 6) とから成っている。ところで従来の階層構造学習オートマトンにおける構造更新は、すべてのパスに同時一律に学習オートマトンを追加することによって行われる。そのため、レベル数が大きくなるにつれてパス数が急速に増加し、その結果、学習効率の大幅な低下を招く。これに対して可変階層構造学習アルゴリズムにおいては、必要と思われる所に 1 個だけ学習オートマトンを追加する (Step 4)。このために、パス数の爆発的増加を招くことが無く、効率の良い学習を行うことが可能となる。

5. 学習特性

構造 Σ^m において、パス平均 reward 確率 $\chi_{j_1 j_2 \dots j_{N_k^m}}$ のサンプル平均値 $\hat{\chi}_{j_1 j_2 \dots j_{N_k^m}}(t)$ を次のように定義する。

$$\hat{\chi}_{j_1 j_2 \dots j_{N_k^m}}(t) = \hat{e}_{j_1}(t) \hat{e}_{j_1 j_2}(t) \dots \hat{e}_{j_1 j_2 \dots j_{N_k^m}}(t) \quad (15)$$

このとき、つぎの補題を得る。

[補題 1] 構造 Σ^m の最適パス上の任意のレベル s ($s = 1, 2, \dots, N_a^m$) において、つぎの関係式が成立する。すなわち、任意の実数 $\delta^m > 0$ に対してある時刻 $t^m < \infty$ が存在して、次式が成り立つ。

$$\begin{aligned} & \text{Prob}\{\hat{e}_{j_1^* j_2^* \dots j_s^*}(t) \hat{e}_{j_1^* j_2^* \dots j_{s+1}^*}(t) \dots \hat{e}_{j_1^* j_2^* \dots j_{N_a^m}^*}(t) \\ & > \hat{e}_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t) \hat{e}_{j_1^* j_2^* \dots j_{s-1}^* i_{s+1}}(t) \\ & \dots \hat{e}_{j_1^* j_2^* \dots j_{s-1}^* i_s \dots i_{N_k^m}}(t)\} > 1 - \delta^m, \\ & \forall t \geq t^m \quad (k \neq a) \quad (16) \end{aligned}$$

証明は付録 A 参照^(注 2)。

また、補題 1 から次の補題 2 を得る。補題 2 は、構造 Σ^m の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 $\alpha_{j_1^* j_2^* \dots j_{s-1}^* j_s^*}$ に対する reward パラメータ γ の値が、それ以外の動作 $\alpha_{j_1^* j_2^* \dots j_{s-1}^* i_s}$ ($i_s = 1, 2, \dots, r$; $i_s \neq j_s^*$) に対するその値よりも確率 $1 - \delta^m$ 以上の確率で大きいことを述べている。

[補題 2] reward パラメータ γ が (10) 式および (11) 式によって与えられるものとする。このとき、構造 Σ^m の最適パス上の各学習オートマトン $Z_{j_1^* j_2^* \dots j_{s-1}^*}$ ($s = 1, 2, \dots, N_a^m$) において、次の関係が成立する。すなわち、任意の実数 $\delta^m > 0$ に対してある時刻 $t^m < \infty$ が存在して、次式が成り立つ。

$$\begin{aligned} & \text{Prob}\{\gamma_{j_1^* j_2^* \dots j_{s-1}^* j_s^*}(t) > \gamma_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t)\} > 1 - \delta^m, \\ & \forall t \geq t^m \quad (i_s = 1, 2, \dots, r; i_s \neq j_s^*) \quad (17) \end{aligned}$$

証明は付録 B 参照。

さらに、構造 Σ^m の最適パスに関してつぎの定理を得る。

《定理 1》 特性 N.1 が成り立つとする。このとき任意の部分構造の最適パスは、部分パスか、あるいは、最良目標パスかのいずれかである。

(証明) 文献 7) の補題と同様の議論で示せるので、その記述を省略する。

ここで補題 2 と定理 1 に注意すると、本可変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立する。

《定理 2》 構造 Σ^m において、特性 N.1 および N.2 が成り立つとする。このとき、初期構造として部分構造を採用するならば、可変階層構造学習アルゴリズムによって最良目標パスが選択される確率は、学習アルゴリズムの強化法の更新則の

(注 2) 構造 Σ^m の最適パス上の任意レベルの学習オートマトン $Z_{j_1^* j_2^* \dots j_{s-1}^*}$ において、その各動作 $\alpha_{j_1^* j_2^* \dots j_{s-1}^* i_s}$ ($i_s = 1, 2, \dots, r$) に対する r 個の $\alpha_{j_1^* j_2^* \dots j_{s-1}^* i_s}$ パス群を考える。そして、これらのパス群に含まれる各パスに対応するパス平均 reward 確率のサンプル平均値から、それらすべてに共通である積 $\hat{e}_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t) \times \dots \times \hat{e}_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t)$ を除いた残りの積 $\hat{e}_{j_1^* j_2^* \dots j_{s-1}^* j_s^*}(t) \hat{e}_{j_1^* j_2^* \dots j_{s-1}^* j_{s+1}^*}(t) \times \dots \times \hat{e}_{j_1^* j_2^* \dots j_{s-1}^* j_{N_k^m}^*}(t)$ の値に着目する。このとき本補題は、最適パスに対するその値がそれ以外のパスに対するその値よりも大きいと言うことが $1 - \delta^m$ 以上の確率で成り立つことを保証している。

係数 $\{\lambda_s\}$ を適切に選ぶことによっていくらかでも 1 に近くすることができる。

(証明) 補題 2 より、構造 Σ^m の最適パス上の各学習オートマトン $Z_{j_1^* j_2^* \dots j_{s-1}^*}$ ($s = 1, 2, \dots, N_a^m$) において、最適パス上の動作 $\alpha_{j_1^* j_2^* \dots j_{s-1}^* j_s^*}$ に対する $\gamma_{j_1^* j_2^* \dots j_{s-1}^* j_s^*}(t)$ の値がその他の動作 $\alpha_{j_1^* j_2^* \dots j_{s-1}^* i_s}$ ($i_s \neq j_s^*$) に対する $\gamma_{j_1^* j_2^* \dots j_{s-1}^* i_s}(t)$ の値よりも大きいことがいえる。このことから、文献 4) の Theorem 2 と同様にして、次式が成立する。すなわち、任意の $\varepsilon^m > 0$ および $\delta^m > 0$ に対して、

$$\text{Prob}\{\pi_{j_1^* j_2^* \dots j_{N_a^m}^*}(t) > 1 - \varepsilon^m\} > 1 - \delta^m,$$

$$\forall t \geq t^m, \forall \lambda_s \in (0, \lambda_s^*) \quad (s = 1, 2, \dots, N_a^m) \quad (18)$$

なる正実数 $\lambda_1^*, \lambda_2^*, \dots, \lambda_{N_a^m}^*$ および t^m が存在する。よって、構造 Σ^m の最適パスが ε^m の精度で選択される確率を Q^m とすると、十分小さな ε^m と δ^m に対して、 $\{\lambda_s\}$ を適切に選ぶことによって、次式を得る。

$$Q^m > 1 - \delta^m \quad (19)$$

ところで、初期構造 Σ^0 として部分構造を採用しつぎのように学習が進むとき、最良目標パスが見出されることになる。「学習の結果、 Σ^0 の最適パスが選択される。このパスは、定理 1 より、最良目標パスか部分パスであるので、これが (最良) 目標パスなら学習は終了し、部分パスならば構造が Σ^1 に更新される。そして学習が継続され、 Σ^1 の最適パスが選択される。このパスも同様に最良目標パスか部分パスであるので、これが (最良) 目標パスなら学習は終了し、部分パスならば構造が Σ^2 に更新される。このように最適パスとしての部分パスの選択と構造更新とを繰り返しながら、ついには、ある構造 Σ^M において (最良) 目標パスを最適パスとして選択して学習を終了する。」よって、可変階層構造学習アルゴリズムによって最良目標パスが選択される確率を Q と書くと、(19) 式より次式を得る。

$$\begin{aligned} Q & \triangleq \text{Prob}\{\text{構造 } \Sigma^0 \text{ の最適パスが選択される}\} \\ & \times \text{Prob}\{\text{構造 } \Sigma^1 \text{ の最適パスが選択される}\} \\ & \vdots \\ & \times \text{Prob}\{\text{構造 } \Sigma^{M-1} \text{ の最適パスが選択される}\} \\ & \times \text{Prob}\{\text{構造 } \Sigma^M \text{ の最適パスが選択される}\} \\ & = Q^0 Q^1 \dots Q^M \\ & > (1 - \delta^0)(1 - \delta^1) \dots (1 - \delta^M) \\ & \geq 1 - (M + 1)\delta \quad (20) \end{aligned}$$

となる。ただし、 $\delta = \max\{\delta^0, \delta^1, \dots, \delta^M\}$ である。ところで、 δ の値はいくらでも小さくとれることに注意しよう。すると、(20) 式の右辺はいくらでも 1 に近くすることが出来る。よって、確率 Q の値はいくらでも 1 に近くすることが出来る。 Q.E.D.

6. 数値シミュレーション

3動作($r=3$)の学習オートマトンからなり、完全構造のレベル数が $N=5$ の可変階層構造学習オートマトンを考える。すると、パスの総数は $r^N=3^5=243$ 本となり、この中の最良目標パスに到達することが学習の目的となる。

いま、最良目標パスを ϕ_{11331} とし、第2位のパスを ϕ_{11111} 、第3位のパスを ϕ_{33333} とする。そして、各動作に対するreward確率がTable 1のようにになっているものとする。ただし、この表以外のreward確率の値は $[0.1, 0.6]$ の一樣乱数で与えた。このとき、表から明らかのように、Level 1およびLevel 3において特性N.2が成り立っている。すなわち、 $e_1 = 0.75 < 0.85 = e_3$ および $e_{113} = 0.75 < 0.85 = e_{111}$ となっている。

この問題に本可変階層構造学習アルゴリズムを適用した結果をTable 2に示す。実験は、 $\lambda_s \equiv \lambda$ の値を0.01刻みで変化させて、各 λ ごとに50回づつ行った。ここで、Success Rate(%)は最良目標パスに到達した割合を、また、Average Number of Iterationsはそのときに要した学習回数の平均値を示しており、 $\lambda = 0.26$ のとき、本学習アルゴリズムによって最良目標パスに必ず到達していることがわかる。ここで参考のために、オリジナルのEstimator Algorithmを動作させた場合(学習アルゴリズムにおいて、 γ の代わりに $\hat{e}$ を用いた場合)の結果もあわせて示す。このときは、50回とも最適パス以外のパスに収束してしまい、特性N.2が成り立つ場合には、オリジナルのEstimator Algorithmでは最良目標パスには全く到達できないことが確かめられた。

Table 1 Reward probabilities

Level	Reward Probability		
1	$d_1^1 = 0.8$ $d_2^2 = 0.7$	$d_3^1 = 0.8$ $d_2^2 = 0.9$	
2	$d_{11}^1 = 0.9$ $d_{11}^2 = 0.85$ $d_{11}^3 = 0.95$	$d_{33}^1 = 0.8$ $d_{33}^2 = 0.8$ $d_{33}^3 = 0.8$	
3	$d_{111}^1 = 0.85$ $d_{111}^2 = 0.9$ $d_{111}^3 = 0.8$	$d_{113}^1 = 0.75$ $d_{113}^2 = 0.75$ $d_{113}^3 = 0.75$	$d_{333}^1 = 0.7$ $d_{333}^2 = 0.7$ $d_{333}^3 = 0.7$
4	$d_{1111}^1 = 0.9$ $d_{1111}^2 = 0.8$	$d_{1133}^1 = 0.8$ $d_{1133}^2 = 0.9$	$d_{3333}^1 = 0.85$ $d_{3333}^2 = 0.85$
5	$d_{11111}^1 = 0.7$ $d_{11111}^2 = 0.7$	$d_{11331}^1 = 0.8$ $d_{11331}^2 = 0.9$	$d_{33333}^1 = 0.8$ $d_{33333}^2 = 0.8$

Table 2 Simulation results

λ	Our Algorithm		Estimator Algorithm
	Average Number of Iterations	Success Rate(%)	Success Rate(%)
0.25	15989	100	0
0.26	15658	100	0
0.27	15157	98	0
0.28	13382	99	0

7. む す び

本論文においては、各レベルにP-モデル複数定常環境をもつ可変階層構造学習オートマトンについて考察し、特性N.2が成り立つようなより一般的な定常環境中において生じる問題点について指摘した。そして、この問題点を解決するためにrewardパラメータ γ を導入し、このパラメータを組み込んだ可変階層構造学習アルゴリズムを構築するとともに、本可変階層構造学習アルゴリズムによって最良目標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、本可変階層構造学習オートマトンが確かに最良目標パスに到達することを確認した。

今後は、本可変階層構造学習アルゴリズムの種々の実問題への適用を計るとともに、強化法の更新則の係数 $\{\lambda_s\}$ の適切な値を解析的に決定する方法を確立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) N. Baba: New Topics in Learning Automata Theory and Applications, Lecture Note in Control and Information Sciences, Springer-Verlag (1985)
- 3) M.A.L. Thathachar and P.S. Sastry: A Class of Rapidly Converging Algorithms for Learning Automata, IEEE International Conference on Cybernetics and Society, 602/606 (1984)
- 4) M.A.L. Thathachar and P.S. Sastry: A New Approach to the Design of Reinforcement Schemes for Learning Automata, IEEE Trans. Systems, Man, and Cybernetics, SMC-15-1, 168/175 (1985)
- 5) M.A.L. Thathachar and P.S. Sastry: A Hierarchical System of Learning Automata That Can Learn the Globally Optimal Path, Information Sciences, 42, 143/166 (1987)
- 6) N. Baba and Y. Mogami: Learning Behaviours of Hierarchical Structure Stochastic Automata Operating In a Non-stationary Multi-teacher Environment, Int. J. Systems Sci., 19-7, 1345/1350 (1988)
- 7) 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)
- 8) 馬場則夫, 最上義夫, 三宅隆浩: 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)

《付 録》

A. 補題1の証明

(9)式より、 $\hat{e}_{i_1 i_2 \dots i_s}(t)$ は $\{d_{i_1 i_2 \dots i_s}^l(t) : l=1, 2, \dots, r_s\}$ の算術平均によって与えられるわけであるが、この $d_{i_1 i_2 \dots i_s}^l(t)$ について、つぎの関係が成立することに注意しよう³⁾。すなわち、任意の実数 $\Delta^m > 0$ および $\delta^m > 0$ に対してある時刻 $t^m < \infty$ が存在して、

$$\text{Prob}\{|d_{i_1 i_2 \dots i_s}^l(t) - d_{i_1 i_2 \dots i_s}^l| < \Delta^m\} > 1 - \delta^m, \quad \forall t \geq t^m \quad (\text{A.1})$$

となる. すると, (3) 式と (9) 式より, 以下の関係が $1 - \delta^m$ 以上の確率で成立することになる.

$$e_{i_1 i_2 \dots i_s} - \Delta^m < \hat{e}_{i_1 i_2 \dots i_s}(t) < e_{i_1 i_2 \dots i_s} + \Delta^m, \quad \forall t \geq t^m \quad (\text{A.2})$$

よって, (15) 式および (A.2) 式より, 次式が $1 - \delta^m$ 以上の確率で成立する.

$$\begin{aligned} & \hat{\chi}_{j_1 j_2 \dots j_{N_a^m}}(t) \\ & < (e_{j_1} + \Delta^m)(e_{j_1 j_2} + \Delta^m) \times \dots \times (e_{j_1 j_2 \dots j_{N_a^m}} + \Delta^m) \\ & = \chi_{j_1 j_2 \dots j_{N_a^m}} + \Delta^m, \quad \forall t \geq t^m \end{aligned} \quad (\text{A.3})$$

$$\begin{aligned} & \hat{\chi}_{j_1 j_2 \dots j_{N_a^m}}(t) \\ & > (e_{j_1} - \Delta^m)(e_{j_1 j_2} - \Delta^m) \times \dots \times (e_{j_1 j_2 \dots j_{N_a^m}} - \Delta^m) \\ & = \chi_{j_1 j_2 \dots j_{N_a^m}} - \Delta^m, \quad \forall t \geq t^m \end{aligned} \quad (\text{A.4})$$

ただし, $e_{j_1} e_{j_1 j_2} \times \dots \times e_{j_1 j_2 \dots j_{N_a^m}}$ 以外の微小項を改めて Δ^m と記述した. ここで, Δ^m を最適パスのパス平均 reward 確率と第 2 位のパスのパス平均 reward 確率との差の $1/2$ としよう. すると, (A.3) 式および (A.4) 式より,

$$\begin{aligned} & \hat{\chi}_{j_1^* j_2^* \dots j_{N_a^m}^*}(t) > \chi_{j_1^* j_2^* \dots j_{N_a^m}^*} - \Delta^m \\ & \geq \chi_{i_1 i_2 \dots i_{N_k^m}} + \Delta^m \\ & > \hat{\chi}_{i_1 i_2 \dots i_{N_k^m}}(t), \quad \forall t \geq t^m \end{aligned} \quad (\text{A.5})$$

となる^{3),4)}ことから, 次式を得る.

$$\begin{aligned} & \text{Prob}\{\hat{\chi}_{j_1^* j_2^* \dots j_{N_a^m}^*}(t) > \hat{\chi}_{i_1 i_2 \dots i_{N_k^m}}(t)\} > 1 - \delta^m, \\ & \quad \forall t \geq t^m \quad (k = 1, 2, \dots, \Psi^m, k \neq a) \end{aligned} \quad (\text{A.6})$$

よって, (15) 式および (A.6) 式より, $1 - \delta^m$ 以上の確率で $s = 1$ に対して次式が成立する.

$$\begin{aligned} & \hat{e}_{j_1^*}(t) \hat{e}_{j_1^* j_2^*}(t) \times \dots \times \hat{e}_{j_1^* j_2^* \dots j_{N_a^m}^*}(t) \\ & > \hat{e}_{i_1}(t) \hat{e}_{i_1 i_2}(t) \times \dots \times \hat{e}_{i_1 i_2 \dots i_{N_k^m}}(t), \\ & \quad \forall t \geq t^m \quad (k \neq a) \end{aligned} \quad (\text{A.7})$$

ここで, (A.7) 式において $i_1 = j_1^*$ とすると,

$$\begin{aligned} & \hat{e}_{j_1^*}(t) \hat{e}_{j_1^* j_2^*}(t) \times \dots \times \hat{e}_{j_1^* j_2^* \dots j_{N_a^m}^*}(t) \\ & > \hat{e}_{j_1^*}(t) \hat{e}_{j_1^* i_2}(t) \times \dots \times \hat{e}_{j_1^* i_2 \dots i_{N_k^m}}(t), \\ & \quad \forall t \geq t^m \quad (k \neq a) \end{aligned} \quad (\text{A.8})$$

となる. よって, (A.8) 式の両辺を $\hat{e}_{j_1^*}(t)$ で割ると, $s = 2$ に対して次式を得る.

$$\begin{aligned} & \hat{e}_{j_1^* j_2^*}(t) \hat{e}_{j_1^* j_2^* j_3^*}(t) \times \dots \times \hat{e}_{j_1^* j_2^* \dots j_{N_a^m}^*}(t) \\ & > \hat{e}_{j_1^* j_2^*}(t) \hat{e}_{j_1^* j_2^* i_3}(t) \times \dots \times \hat{e}_{j_1^* j_2^* i_3 \dots i_{N_k^m}}(t), \\ & \quad \forall t \geq t^m \quad (k \neq a) \end{aligned} \quad (\text{A.9})$$

さらに, (A.9) 式において $i_2 = j_2^*$ とし, その両辺を $\hat{e}_{j_1^* j_2^*}(t)$ で割ると, $s = 3$ に対して次式を得る.

$$\begin{aligned} & \hat{e}_{j_1^* j_2^* j_3^*}(t) \hat{e}_{j_1^* j_2^* j_3^* j_4^*}(t) \times \dots \times \hat{e}_{j_1^* j_2^* \dots j_{N_a^m}^*}(t) \\ & > \hat{e}_{j_1^* j_2^* j_3^*}(t) \hat{e}_{j_1^* j_2^* j_3^* i_4}(t) \times \dots \times \hat{e}_{j_1^* j_2^* j_3^* i_4 \dots i_{N_k^m}}(t), \end{aligned}$$

$$\forall t \geq t^m \quad (k \neq a) \quad (\text{A.10})$$

以下同様にすると, 構造 Σ^m の最適パス上の任意のレベル s ($s = 1, 2, \dots, N_a^m$) において (16) 式が成り立つ. Q.E.D.

B. 補題 2 の証明

構造 Σ^m の最適パス $\phi_{j_1^* j_2^* \dots j_{N_a^m}^*}$ 上の各学習オートマトンに着目する. まず, 最下レベル N_a^m の学習オートマトン $Z_{j_1^* j_2^* \dots j_{N_a^m}^*}$ において, 最適パス上の動作 $\alpha_{j_1^* j_2^* \dots j_{N_a^m}^*}$ に対する reward パラメータについて, (10) 式より次式が得られる.

$$\gamma_{j_1^* j_2^* \dots j_{N_a^m}^*}(t) = \hat{e}_{j_1^* j_2^* \dots j_{N_a^m}^*}(t) \quad (\text{B.1})$$

また, $\alpha_{j_1^* j_2^* \dots j_{N_a^m}^*}$ 以外の動作 $\alpha_{j_1^* j_2^* \dots j_{N_a^m}^* l_1}$ ($l_1 = 1, 2, \dots, r; l_1 \neq j_{N_a^m}^*$) に対しては, この動作が最下レベルの動作であるか否かによって次の関係式が成立する.

(i) $\alpha_{j_1^* j_2^* \dots j_{N_a^m}^* l_1}$ が最下レベルの動作であるとき:

$$\gamma_{j_1^* j_2^* \dots j_{N_a^m}^* l_1}(t) = \hat{e}_{j_1^* j_2^* \dots j_{N_a^m}^* l_1}(t) \quad (\text{B.2})$$

(ii) $\alpha_{j_1^* j_2^* \dots j_{N_a^m}^* l_1}$ が最下レベルの動作でないとき:

$$\begin{aligned} & \gamma_{j_1^* j_2^* \dots j_{N_a^m}^* l_1}(t) = \hat{e}_{j_1^* j_2^* \dots j_{N_a^m}^* l_1}(t) \\ & \quad \times \max_{l_2} \{\gamma_{j_1^* j_2^* \dots j_{N_a^m}^* l_1 l_2}(t)\} \end{aligned} \quad (\text{B.3})$$

さらに, (B.3) 式の右辺においても, $\gamma_{j_1^* j_2^* \dots j_{N_a^m}^* l_1 l_2}(t)$ が最下レベルの動作に対するものであるか否かによって, 上と同様の議論が必要である. すなわち,

(ii-a) $\gamma_{j_1^* j_2^* \dots j_{N_a^m}^* l_1 l_2}(t)$ が最下レベルの動作に対するものであるとき:

$$\gamma_{j_1^* j_2^* \dots j_{N_a^m}^* l_1 l_2}(t) = \hat{e}_{j_1^* j_2^* \dots j_{N_a^m}^* l_1 l_2}(t) \quad (\text{B.4})$$

を得る. よって, (B.3) 式および (B.4) 式より,

$$\begin{aligned} & \gamma_{j_1^* j_2^* \dots j_{N_a^m}^* l_1}(t) = \hat{e}_{j_1^* j_2^* \dots j_{N_a^m}^* l_1}(t) \\ & \quad \times \max_{l_2} \{\hat{e}_{j_1^* j_2^* \dots j_{N_a^m}^* l_1 l_2}(t)\} \end{aligned} \quad (\text{B.5})$$

(ii-b) $\gamma_{j_1^* j_2^* \dots j_{N_a^m}^* l_1 l_2}(t)$ が最下レベルの動作に対するものでないとき:

$$\begin{aligned} & \gamma_{j_1^* j_2^* \dots j_{N_a^m}^* l_1 l_2}(t) = \hat{e}_{j_1^* j_2^* \dots j_{N_a^m}^* l_1 l_2}(t) \\ & \quad \times \max_{l_3} \{\gamma_{j_1^* j_2^* \dots j_{N_a^m}^* l_1 l_2 l_3}(t)\} \end{aligned} \quad (\text{B.6})$$

となる. このような考察を, 動作 $\alpha_{j_1^* j_2^* \dots j_{N_a^m}^*}$ を通るすべてのパスの最下レベルに至るまで繰り返す. いま, $\gamma_{j_1^* j_2^* \dots j_{N_a^m}^* l_1 l_2 l_3}(t)$ が最下レベルの動作に対するものであるとすると, (10) 式, (B.6) 式および (B.3) 式より,

$$\begin{aligned} & \gamma_{j_1^* j_2^* \dots j_{N_a^m}^* l_1}(t) = \hat{e}_{j_1^* j_2^* \dots j_{N_a^m}^* l_1}(t) \\ & \quad \times \max_{l_2, l_3} \{\hat{e}_{j_1^* j_2^* \dots j_{N_a^m}^* l_1 l_2}(t) \hat{e}_{j_1^* j_2^* \dots j_{N_a^m}^* l_1 l_2 l_3}(t)\} \end{aligned} \quad (\text{B.7})$$

以上から, $\gamma_{j_1^* j_2^* \dots j_{N_a^m}^* l_1}(t)$ は, (B.2) 式, (B.5) 式あるいは (B.7) 式のように与えられることになる. そして, 補題 1 に注意すると, このそれぞれに対して次式が $1 - \delta^m$ 以上の

確率で成り立つ.

(B.1) 式および (B.2) 式より,

$$\begin{aligned} \gamma_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}(t) &= \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}(t) \\ &> \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}(t) = \gamma_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}(t), \\ &\quad \forall t \geq t^m \quad (l_1 \neq j_{N_a^m}^*) \quad (\text{B.8}) \end{aligned}$$

(B.1) 式および (B.5) 式より,

$$\begin{aligned} \gamma_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}(t) &= \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}(t) \\ &> \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1} \max\{\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1 l_2}\} \\ &= \gamma_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}(t), \quad \forall t \geq t^m \quad (l_1 \neq j_{N_a^m}^*) \quad (\text{B.9}) \end{aligned}$$

(B.1) 式および (B.7) 式より,

$$\begin{aligned} \gamma_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}(t) &= \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}(t) \\ &> \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}(t) \\ &\quad \times \max_{l_2, l_3}\{\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1 l_2}(t) \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1 l_2 l_3}(t)\} \\ &= \gamma_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}(t), \quad \forall t \geq t^m \quad (l_1 \neq j_{N_a^m}^*) \quad (\text{B.10}) \end{aligned}$$

ゆえに, (B.8) 式, (B.9) 式および (B.10) 式より,

$$\begin{aligned} \gamma_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}(t) &> \gamma_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}(t), \quad \forall t \geq t^m \\ &\quad (l_1 = 1, 2, \dots, r; l_1 \neq j_{N_a^m}^*) \quad (\text{B.11}) \end{aligned}$$

よって, 最下レベルの学習オートマトン $Z_{j_1^* j_2^* \dots j_{N_a^m-1}^*}$ において, (17) 式が成り立つ.

つぎに, 学習オートマトン $Z_{j_1^* j_2^* \dots j_{N_a^m-1}^*}$ が結合されている動作 $\alpha_{j_1^* j_2^* \dots j_{N_a^m-1}^*}$ を持つレベル $(N_a^m - 1)$ の学習オートマトン $Z_{j_1^* j_2^* \dots j_{N_a^m-2}^*}$ について考える. まず, 動作 $\alpha_{j_1^* j_2^* \dots j_{N_a^m-1}^*}$ に対して, (11) 式より, つぎの関係が成り立つ.

$$\begin{aligned} \gamma_{j_1^* j_2^* \dots j_{N_a^m-1}^*}(t) &= \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^*}(t) \max_{l_1}\{\gamma_{j_1^* j_2^* \dots j_{N_a^m-1}^* l_1}(t)\} \quad (\text{B.12}) \end{aligned}$$

ここで, (B.11) 式に注意すると, 次式を得る.

$$\begin{aligned} \gamma_{j_1^* j_2^* \dots j_{N_a^m-1}^*}(t) &= \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^*}(t) \gamma_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}(t) \\ &= \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^*}(t) \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-1}^* j_{N_a^m}^*}(t) \quad (\text{B.13}) \end{aligned}$$

また, $\alpha_{j_1^* j_2^* \dots j_{N_a^m-2}^* j_{N_a^m-1}^*}$ 以外の動作 $\alpha_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}$ ($l_1 = 1, 2, \dots, r; l_1 \neq j_{N_a^m-1}^*$) に対しては, この動作が最下レベルの動作であるか否かによって次の関係式が成立する.

(iii) $\alpha_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}$ が最下レベルの動作であるとき:

$$\gamma_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t) = \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t) \quad (\text{B.14})$$

(iv) $\alpha_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}$ が最下レベルの動作でないとき:

$$\begin{aligned} \gamma_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t) &= \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t) \\ &\quad \times \max_{l_2}\{\gamma_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1 l_2}(t)\} \quad (\text{B.15}) \end{aligned}$$

そして, (B.15) 式について (B.4) 式~(B.7) 式と同様の議論を行うことと (B.14) 式より, $\gamma_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t)$ はつぎの

いずれかの関係を満たすことになる.

$$\gamma_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t) = \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t) \quad (\text{B.16})$$

$$\begin{aligned} \gamma_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t) &= \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t) \\ &\quad \times \max_{l_2}\{\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1 l_2}(t)\} \quad (\text{B.17}) \end{aligned}$$

$$\begin{aligned} \gamma_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t) &= \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t) \\ &\quad \times \max_{l_2, l_3}\{\hat{e}_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1 l_2}(t) \hat{e}_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1 l_2 l_3}(t)\} \quad (\text{B.18}) \end{aligned}$$

ゆえに, 補題1と (B.13) 式, (B.16) 式, (B.17) 式および (B.18) 式とより, $1 - \delta^m$ 以上の確率で次式が成立する.

$$\begin{aligned} \gamma_{j_1^* j_2^* \dots j_{N_a^m-2}^* j_{N_a^m-1}^*}(t) &> \gamma_{j_1^* j_2^* \dots j_{N_a^m-2}^* l_1}(t), \\ &\quad \forall t \geq t^m \quad (l_1 = 1, 2, \dots, r; l_1 \neq j_{N_a^m-1}^*) \quad (\text{B.19}) \end{aligned}$$

よって, レベル $(N_a^m - 1)$ の学習オートマトン $Z_{j_1^* j_2^* \dots j_{N_a^m-2}^*}$ 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.

[著者紹介]

最上 義夫 (正会員)

1975年, 京都工芸繊維大学工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). 計測自動制御学会, 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, 計測自動制御学会, 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.