システム制御情報学会論文誌, 4 巻, 7 号, pp. 259~266 (1991)

複雑な構造をした非線形システムの同定を目的としたニューラルネットワーク

近藤 正

本研究では、構造が複雑でしかもその物理構造がよくわからない非線形システムの同定問題に対して、対象システムから得られる入出力データを用いて学習を繰り返すことにより、そのシステムの特徴をネットワークの内部に記憶保存させることができるニューラルネットワークを構成し、従来から用いられてきた統計的非線形システム同定手法 (GMDH 手法) との比較を行ない、その有効性を明らかにする。

システム制御情報学会論文誌, 4 巻, 7 号, pp. 267~276 (1991)

定性ネットワークモデルとそのシミュレーションアルゴリズム

堂園 浩・石田 好輝・得丸 英勝

プラントなどの大規模物理系に対する、定性シミュレーションアルゴリズムを開発した。プラントのような大規模物理系においては、状態変数が多く複雑に作用しあうため、数値的なシミュレーションは不可能であったり、数式に基づくモデルが得られない場合がある。そのため、本論文では、変数間の作用関係をアークとする定性ネットワークモデルを用い、ネットワーク上での定性シミュレーションについて考察した。そのアルゴリズムは基本的には deKleer のアルゴリズムを定性ネットワーク上で応用したもので次に述べる特徴を持つ。(1) 値伝播を静的伝播と動的伝播に分けることで定性シミュレーションに特徴的な分岐の数を減らした。(2) 静的伝播に伴う決定不能な変数と発生をバランス機能により抑えた。(3) ネットワーク上で一貫性を保ちながら値伝播を行なうことができる。

システム制御情報学会論文誌, 4 巻, 7 号 pp. 277~285 (1991)

制御を前提とした設計手法とその車両操舵機構への応用

河辺 徹・得丸 英勝

機械システムの設計において、従来行なわれている手順では、制御性からの要求により機械要素の値を決定できないという点に問題があると思われる。そこで、制御を前提として機械システムを設計する手法を提案する。

本手法では、機械設計に対する評価規範として、入力偏差による評価を行ない、この結果と制約条件に基づいて機械要素の値などを決定する。さらに、この手法を自動車の操舵システムの設計に応用した。

システム制御情報学会論文誌, 4 巻, 7 号 pp. 286~293 (1991)

拡散需要を伴う生産管理システムの最適制御

土肥 正・海生 直人・尾崎 俊治

従来の生産管理システムに関する多くの研究は、不確実性下での需要量を時間と独立な確率変数として取り扱ってきた。これに対し、本論文では需要量が Markov 拡散過程に従う場合の生産管理システムについて議論する。最適生産量を解析的に導出する手法として、確率制御の手法が用いられる。本論文の主な目的は、従来の決定論的モデルを確率場に拡張し、さらに需要分布の形状の違いによる生産管理システムの制御特性を明確に示すことである。

システム制御情報学会論文誌 4 巻, 7 号, pp. 294~301 (1991)

協力環境下での学習オートマトンの集団モデル

銭 飛・平田 廣則

複数の目標が存在する最適制御においては、制御システムは、これらの目標の実現を目指す複数の制御者からなる。複数の制約条件のもとで環境の変動が事前に予知できない場合に、競争的あるいは協力的に共存している各々の制御者が、ある種の集団行動を形成する現象は社会システム、経済システムなどの中でよく見られる。ゲーム理論の立場から見れば、このようなシステムは協力、非協力ゲームに帰着される。

本研究では、このような複数の目標のもとで動作しているシステムの最適化問題を取り上げ、複数の静的なランダム環境 (教師) 下で動作している学習オートマトンの集団モデルについて考察し、モデルの定義、評価基準などを与え、ランダム環境に適応するための各オートマトンの強化法の構築方法を示す。

集団モデルとしては、種々のタイプが考えられるが、本文では、とくに、協力環境下での非協力ゲームのモデルについて考察し、このモデルの静特性および動特性を議論したうえで P-model と S-model の二つのタイプの学習オートマトンに対応できる強化法を提案する。また、強化法の収束性の証明を与え、簡単なシミュレーション例を用いて、それを検証する。

協力環境下での学習オートマトンの集団モデル*

銭 飛**・平 田 廣 則***

A Collective Model of Learning Automata with N -Cooperative Random Environments*

Fei QIAN** and Hironori HIRATA***

In this paper, we consider a collective model of learning automata operating on N -static random environments. Each automaton has a behavioural tactic directed towards the realization of its own goal, estimating the magnitude of the expected value of utility function that depends explicitly on the automaton's strategy and the corresponding responses from environments. From game theoretic viewpoint, we can construct several types of environments for this model.

In here, a non-cooperative game on N -cooperative environments is proposed. For this model, the definitions, evaluation metrics and a distance diminishing reinforcement scheme are introduced. As an application, a simple numerical example is shown.

1. ま え が き

有限決定性オートマトンの学習特性を見つけ、学習オートマトンの基本概念を最初に与えたのは Tsetlin である¹⁾。その後、確率オートマトンの学習の性質に関する研究が行なわれた²⁾。それ以来、学習オートマトンは学習システムのモデル化に使われ、生体、学習制御システム^{3),4)}、情報通信^{5),6)}、パターン認識⁷⁾などの領域上での応用例が多数発表されている。それと同時に、単一の目標の元での学習オートマトンの集団特性、あるいは学習オートマトン間のゲームについての研究が行なわれ⁸⁾、いくつかの応用例も見られる⁹⁾。

本文では、複数の目標のもとでの学習オートマトンの集団モデルを考察する。

複数の目標が存在する最適制御においては、制御システムは、これらの目標の実現を目指す複数の制御者からなる。複数の制約条件のもとで環境の変動が事前に予知できない場合に、競争的あるいは協力的に共存している各々の制御者が、ある種の集団行動を形成する現象は社会システム、経済システムなどの中でよく見られる。ゲ

ーム理論の立場から見れば、このようなシステムは協力、非協力ゲームに帰着される。

本文では、このような複数の目標のもとで動作しているシステムの最適化問題を取り上げ、前述したように、複数の静的なランダム環境（教師）下で動作している学習オートマトンの集団モデルについて考察する。モデルの定義、評価基準などを与え、ランダム環境に適応するための各オートマトンの強化法の構築方法を示す。とくに、協力環境下での非協力ゲームのモデルについて考察し、このモデルの静特性および動特性を議論したうえで P -model と S -model の二つのタイプの学習オートマトンに対応できる強化法を提案する。また、強化法の収束性の証明を与え、簡単なシミュレーション例を用いて、それを検証する。

2. 学習オートマトンの M 人ゲーム

まず、本文の中で使っている各変数の意味を以下において説明する。具体的な定義は各々扱うところで述べる。

$A^i : \{x^i, a^i, \phi^i, P^i, T^i, o^i\}$ により定義された確率オートマトン。

a^i : 確率オートマトン A^i の出力集合。

$a^i = \{a_1^i, a_2^i, \dots, a_{m^i}^i\}$ 。

m^i : 確率オートマトン A^i の出力集合の要素数。

x^i : 確率オートマトン A^i の入力ベクトル。

$x^i = (x_1^i, \dots, x_N^i)^T \in \{0, 1\}^N$ 。

* 原稿受付 1990. 7. 26

** 広島電機大学 電子工学科 Department of Electronics Engineering, Hiroshima-Denki Institute of Technology

*** 千葉大学 工学部 Faculty of Engineering, Chiba University ; Chiba-shi 260, JAPAN

Key Words : learning automaton, reinforcement scheme, game theory, rash game, distance diminishing property.

$x_j^i = 1$: 未知環境 j から確率オートマトン A^i へのペナルティ応答を意味する。

$x_j^i = 0$: 未知環境 j から確率オートマトン A^i へのノンペナルティ応答を意味する。

ϕ^i : 確率オートマトン A^i の状態集合。

$\phi^i = \{\phi_1^i, \phi_2^i, \dots, \phi_{k_i}^i\}$ 。

k^i : 確率オートマトン A^i の状態集合の要素数。

P^i : 確率オートマトン A^i の状態遷移確率分布。

$P^i = (p_1^i, p_2^i, \dots, p_{m_i}^i)^T$ 。

T^i : 確率オートマトン A^i の状態遷移関数 (局所的状態遷移関数)。 $T^i: \phi^i \times x^i \rightarrow \phi^i$ 。

o^i : 確率オートマトン A^i の出力関数。 $o^i: \phi^i \rightarrow \alpha^i$ 。

A : 確率オートマトンの集合。 $A = \{A^1, \dots, A^M\}$ 。

$\alpha(t)$: 総合出力ベクトル。

$\alpha(t) = (\alpha^1(t), \alpha^2(t), \dots, \alpha^M(t))^T$ 。

$X(t)$: 総合入力行列。

$X(t) = (x^1(t), x^2(t), \dots, x^M(t))$ 。

$\Phi(t)$: 状態空間。

$\Phi(t) = \phi^1(t) \otimes \phi^2(t) \otimes \dots \otimes \phi^M(t)^\dagger$ 。

T : 大局的状态遷移関数。 $T = (T^1, T^2, \dots, T^M)$ 。

O : 出力関数空間。 $O = (o^1, o^2, \dots, o^M)$ 。

R : 未知環境の集合。 $R = \{R^1, \dots, R^N\}$ 。

c_{ik}^j : 未知環境 R^j が A^i から α_k^i を受けたとき、 A^i へペナルティ応答 ($x_j^i = 1$) を返す確率。

f^i : ゲームの中でのプレーヤである学習オートマトン A^i の効用関数 (Utility Function)。

【定義1】 N 個の目標のもとで行なう学習オートマトン $A^1, \dots, A^M$ の M 人ゲーム (以下, $MLAG/N$ と略す) は, 各オートマトンからの出力

$$\alpha(t) = (\alpha^1(t), \dots, \alpha^M(t))^T, \quad \alpha^i(t) = o^i(\phi^i(t)), \quad i = 1, 2, \dots, M \quad (1)$$

と, それと対応するランダム環境からの反応

$$X(t+1) = (x^1(t+1), \dots, x^M(t+1)) = (x_j^i)_{i=1, \dots, M, j=1, \dots, N}, \quad x_j^i \in \{0, 1\} \quad (2)$$

に関する確率分布 $p(\alpha(t), X(t+1))$ が存在し, かつ

$$\sum_X p(\alpha(t), X(t+1)) = 1 \quad (3)$$

が満たされるとき成立する。

このとき, $\alpha(t)$, $X(t+1)$ をそれぞれ時刻 t でのプレー, プレーの結果と呼び, $\alpha^i(t)$ をオートマトン A^i からの戦略と呼ぶ。確率分布 $p(\alpha(t), X(t+1))$ は $\{c_{ik}^j\}$ により定められる。

【定義2】 $MLAG/N$ のプレーの環境を以下のように定義する。

$$R = \{R^j | R^j = \otimes_{i=1}^M \{c_{ik}^j, \dots, c_{im}^j\}, j = 1, \dots, N\} \quad (4)$$

ただし,

$$c_{ik}^j = \Pr(x_j^i(t+1) = 1 | \alpha^i(t) = \alpha_k^i) \quad (5)$$

$\alpha(t) = (\alpha^1(t), \dots, \alpha^M(t))^T$ が出されたとき, オートマトンへ反応 $X(t+1)$ を返す環境の確率分布は $p(\alpha(t), X(t+1))$ で表われ, 各プレーヤのプレーの効用関数を以下のように定義する。

$$f^i(\alpha(t)) = \frac{1}{2} \sum_{j=1}^N \sum_{x_j^1, \dots, x_j^{i-1}, x_j^{i+1}, \dots, x_j^M} (p(\alpha(t), x_j^{i+}) - p(\alpha(t), x_j^{i-})) \quad (6)$$

x_j^{i-} : 未知環境 R^j から A^i への入力ベクトル。ただし, $x_j^i = 0$ である。

$$x_j^{i-} = (x_j^1, \dots, x_j^{i-1}, 0, x_j^{i+1}, \dots, x_j^M)^T \quad (7)$$

x_j^{i+} : 未知環境 R^j から A^i への入力ベクトル。ただし, $x_j^i = 1$ である。

$$x_j^{i+} = (x_j^1, \dots, x_j^{i-1}, 1, x_j^{i+1}, \dots, x_j^M)^T \quad (8)$$

$f^i(\alpha(t))$ はオートマトン A^i の時刻 t での戦略について, 未知環境から平均ペナルティと平均ノンペナルティの差を意味している。このような効用関数により定められる学習オートマトンのゲームはプレーの系列 $\{\alpha(t), t = 1, 2, \dots\}$ からなり, 各プレーの結果 $X(t+1)$ は確率 $p(\alpha(t), X(t+1))$ により決められる。(2) 式と対応する未知環境は以下のようなペナルティ反応確率行列で表現できる。

$$C = (c_{ik}^j)_{ijk}, \quad c_{ik}^j = \Pr(x_j^i(t+1) = 1 | \alpha^i(t) = \alpha_k^i), \quad i = 1, \dots, M, j = 1, \dots, N, k = 1, \dots, m^i \quad (9)$$

c_{ik}^j は, 入力 α_k^i を受けたとき, 未知環境 R^j がオートマトン A^i にペナルティ応答を返す確率を表す。

理論分析を容易にするために, 行列 $X(t)$ を写像 $\xi: \{0, 1\}^N \rightarrow \{0, \dots, N\}$ によってベクトル化する。ベクトル化した $X(t)$ を $\bar{X}(t)$ と表わすと,

$$\bar{X}(t) = (\bar{x}^1(t), \dots, \bar{x}^M(t)), \quad \bar{x}^i(t) = E[\xi(x_j^i(t), \dots, x_N^i(t))] \quad (10)$$

となる。ここで, $E(\cdot)$ は期待値である。写像 ξ をペナルティ応答写像と呼ぶことにする。ここで, $\xi(x_1^i, \dots, x_N^i)$ は, A^i が $\alpha^i(t) = \alpha_j^i$ を出したときのペナルティ応答を生じさせる未知環境の数を表わし, 次のように表現できる。

† $A \otimes B$ は集合 A と B の直積とする。

$$\xi(x_1^i, \dots, x_N^i) = x_1^i + x_2^i + \dots + x_N^i \quad (11)$$

$\bar{x}^i$ は確率

$$\begin{aligned} p(\bar{x}^i | \alpha^i(t) = \alpha_j^i) \\ = c_{ij}^1 \dots c_{ij}^{\bar{x}^i} (1 - c_{ij}^{\bar{x}^i+1}) \dots (1 - c_{ij}^N) + \dots \\ + (1 - c_{ij}^1) \dots (1 - c_{ij}^{N-\bar{x}^i}) c_{ij}^{N-\bar{x}^i+1} \dots c_{ij}^N \end{aligned} \quad (12)$$

により決まる。

【定義3】時刻 t で、総合力 $a(t) = (\alpha_{k1}^1, \dots, \alpha_{kM}^M)^T$ を出した MLAG/N のシステム状態 $\phi_a \in \Phi$ を次のように定義する。

$$\phi_a = (\phi_{k1}^1, \dots, \phi_{kM}^M) \quad (13)$$

ここで、 ϕ_{ki}^i は $\phi^i(t)$ である。 $i = 1, \dots, M$ 。

【定義4】学習オートマトン A^i の状態遷移確率行列 a^i を以下のように定義する。

$$a^i = (a_{qq'}^i(x^i))_{qq'} \quad (14)$$

ただし、 $a_{qq'}^i$ は入力 x^i が与えられたとき、学習オートマトン A^i が状態 $\phi^i(t) = q$ から状態 $\phi^i(t+1) = q'$ に遷移する確率を表わす。

$p(a(t), X(t+1))$ の形式により、 N 個の環境下で競争している M 個の学習オートマトン集団モデルを以下

(a) MODEL 1: A collective model of learning automata with N -cooperative random environments

$$p(a(t), X(t)) = \prod_{i=1}^N p(a(t), x_1^i(t+1), \dots, x_N^i(t+1))$$

(b) MODEL 2: A collective model of learning automata with N -noncooperative random environments

$$p(a(t), X(t)) = \prod_{j=1}^M p(a(t), x_1^j(t+1), \dots, x_N^j(t+1))$$

Fig. 1 A collective model of learning automata

のように分類することが出来る。

モデル1:

$$\begin{aligned} p(a(t), X(t+1)) \\ = \prod_{i=1}^M p(a(t), x_1^i(t+1), \dots, x_N^i(t+1)) \end{aligned} \quad (15)$$

これは、依存関係のある N 個の環境の全体に対して、 M 個のオートマトンへの反応は各々独立していることを意味している (Fig.1 (a)). これを N 教師付き学習オートマトンの間のゲームと見直すことが出来る。

モデル2:

$$\begin{aligned} p(a(t), X(t+1)) \\ = \prod_{j=1}^N p(a(t), x_j^1(t+1), \dots, x_j^M(t+1)) \end{aligned} \quad (16)$$

この場合、オートマトンの集団に対する環境 j からの反応は環境 $s (s \neq j)$ からの反応と独立している (Fig.1 (b)).

本文では、モデル1のような協力環境上での非協力ゲームについて考察する。

3. 協力環境上での非協力ゲーム

このモデルでは、 $p(a(t), X(t+1))$ は以下になる。

$$\begin{aligned} p(a(t), X(t+1)) \\ = p(a(t), x_1^1, \dots, x_N^1, \dots, x_1^M, \dots, x_N^M) \\ = \prod_{i=1}^M p(a(t), x_1^i(t+1), \dots, x_N^i(t+1)) \end{aligned} \quad (17)$$

つまり、 M 個のオートマトンへの N 個の環境の全体からの反応は各々独立している (Fig.1 (a)). これを N 教師付き学習オートマトン⁹⁾ のゲームと見直すことが出来る。

3.1 環境の性質

このモデルに対して、システムの状態空間を具体的に $\Phi \triangleq \{a, X\}$ と表わし、それは M 個の同値類 (equivalence class) に分けられる。各同値類 $\Omega^i (i = 1, \dots, M)$ は $(\alpha^i, \bar{x}^i)$ からなる。 $\Phi \otimes \Omega^i$ 上で定義されている同値関係 (equivalence relation) η は $\Phi \otimes \Omega^i$ の部分集合であり、以下のように表わされる。

$$\begin{aligned} \eta = \{(\phi, \omega^i) | \phi = (a, X), \\ \omega^i = (\alpha^i, \bar{x}^i) = (\alpha^i, E[\xi(x_1^i(t), \dots, x_N^i(t))])\} \\ \omega^i \in \Omega^i, \phi \in \Phi \end{aligned} \quad (18)$$

η の各要素は、確率分布 $p(\bar{x}^i | \alpha^i)$ により一意に決められる。

また、外部戦略¹⁰⁾

$$\bar{\alpha}^i = (\alpha_{j1}^i, \dots, \alpha_{j,i-1}^i, \alpha_{j,i+1}^i, \dots, \alpha_{jM}^i)$$

が与えられたとき、プレイヤー i の最適反応戦略集合 Q^i

は以下のように定義される。

$$Q^i(\bar{a}^i) = \{a_{ji}^* | a_{ji}^* \in a^i, f^i(a_{ji}^*, \bar{a}^i) = \min_{j^i} f^i(a_{ji}^*, \bar{a}^i)\}$$

各プレーヤ A^i に対するゲームの中の純戦略 (pure strategy)¹⁰⁾

$$a_{ji}^* \in Q^i(\bar{a}^i) \quad (19)$$

は部分空間 Ω^i 上で各々独立に決めることが出来る。明らかに、 Ω^i は有界閉空間である。

総合出力 α, β とそれに対応する状態をそれぞれ $\phi_\alpha \in \Phi(t), \phi_\beta \in \Phi(t+1)$ とする

$$\alpha = O(\phi_\alpha), \beta = O(\phi_\beta) \quad (20)$$

ここで、

$$\begin{aligned} \phi_\alpha &= (\phi_{\alpha 1}, \dots, \phi_{\alpha M})^T, \phi_\beta = (\phi_{\beta 1}, \dots, \phi_{\beta M})^T \\ \alpha_i &= o^i(\phi_{\alpha i}), \beta_i = o^i(\phi_{\beta i}), i = 1, \dots, M \end{aligned} \quad (21)$$

である。

システムが状態 ϕ_α から状態 ϕ_β へ遷移する確率 $p(\phi_\beta | \phi_\alpha)$ は以下のように求められる。

$$\begin{aligned} p(\phi_\beta | \phi_\alpha) &= p(\alpha, \bar{X}(t+1) | \alpha_{\phi_\alpha}, \bar{X}(t+1)) \\ &= \sum_{\bar{x}^1, \dots, \bar{x}^M} \prod_{i=1}^M p(\alpha_{k^1}^1, \dots, \alpha_{k^M}^M, \bar{x}^i | \alpha_{\phi_\alpha}^i, \bar{x}^i) \end{aligned} \quad (22)$$

このモデルの内部構造とシステム状態遷移フローをそれぞれ Fig. 2, 3 に示す。

Fig. 2 The inner structure of MODEL 1

時刻 $t+1$ での遷移確率 $p(\phi_\beta | \phi_\alpha)$ は時刻 t での遷移確率だけに依存しているので、システムを有限マルコフ連鎖で表現できる。

$$\sum_{\phi_\beta} p(\phi(t+1) = \phi_\beta | \phi(t) = \phi_\alpha) = 1 \quad (23)$$

(23) 式から、行列 $(p(\phi_\beta | \phi_\alpha))_{\phi_\alpha, \phi_\beta}$ は確率マトリクスで、その要素は各遷移状態が発生する確率測度であることが分かる。ゆえに、マルコフ過程の定義¹¹⁾ により、このような非協力的 MLAG/ N 上での状態遷移確率の遷移過程は直積空間 $S^{m^1} \oplus [0, 1]^{m^1}$ 上でのエルゴード的マルコフ過程であることが明らかとなる。

【定義 5】 状態遷移過程がエルゴード的マルコフ連鎖をなす MLAG/ N をエルゴード的 MLAG/ N と呼ぶ。

エルゴード的 MLAG/ N (以下 E-MLAG/ N と略す) においては、平均利得、あるいは平均損失は初期状態とは無関係である。

【定理 1】 E-MLAG/ N には少なくとも一つのナッシュ均衡点¹⁰⁾ が存在する。

定理 1 の証明に関しては、参考文献 6), 12) と同じ方法で簡単に証明できる。

3.2 オートマトンモデル

E-MLAG/ N 中の各オートマトン $A^i (i = 1, \dots, M)$ の出力を選択する確率分布が単体

$$S^{m^i} = \{P^i | P^i = (p_1^i, \dots, p_{m^i}^i), \sum_{j=1}^{m^i} p_j^i = 1, 0 \leq p_j^i \leq 1\} \quad (24)$$

上で定義されているものとする。

時刻 t で A^i は、 p_k^i の確率で a_k^i を未知環境 R に作用した結果、 N 個の未知環境から $\bar{x}^i$ 個のペナルティ応答を受け取ったとする。時刻 $t+1$ での出力選択確率分布を以下のような強化法 T_{N-C} を用いて修正する。

強化法 T_{N-C} :

$$p_k^i(t+1) = p_k^i(t) - \theta_1 \left(1 - \frac{\bar{x}^i}{N}\right) \sum_{j \neq k} p_j^i(t) + \theta_2 \frac{\bar{x}^i}{N} p_k^i(t) \quad (25)$$

$$\begin{aligned} p_j^i(t+1) &= p_j^i(t) + \theta_1 \left(1 - \frac{\bar{x}^i}{N}\right) p_j^i(t) \\ &\quad - \theta_2 \frac{\bar{x}^i}{N} \left(\frac{1}{m^i - 1} - p_j^i(t)\right), j \neq k \end{aligned} \quad (26)$$

$$\bar{x}^i = \xi(x_1^i, \dots, x_N^i)$$

$$0 < N\theta_1 < 1, \theta_2 < \theta_1$$

$$\begin{aligned} \phi(t) = \begin{bmatrix} \phi^1(t) \\ \vdots \\ \phi^i(t) \\ \vdots \\ \phi^M(t) \end{bmatrix} &\Rightarrow \begin{bmatrix} \alpha^1(t) = o(\phi^1) \\ \vdots \\ \alpha^i(t) = o(\phi^i) \\ \vdots \\ \alpha^M(t) = o(\phi^M) \end{bmatrix} \Rightarrow \begin{bmatrix} p(\alpha^1(t), \bar{x}^1), E[\xi(x_1^1(t), \dots, x_N^1(t))] \\ \vdots \\ p(\alpha^i(t), \bar{x}^i), E[\xi(x_1^i(t), \dots, x_N^i(t))] \\ \vdots \\ p(\alpha^M(t), \bar{x}^M), E[\xi(x_1^M(t), \dots, x_N^M(t))] \end{bmatrix} \\ &\Rightarrow \begin{bmatrix} \phi^1(t+1) = T^1(\alpha^1(t), \bar{X}^1(t+1)) \\ \vdots \\ \phi^i(t+1) = T^i(\alpha^i(t), \bar{X}^i(t+1)) \\ \vdots \\ \phi^M(t+1) = T^M(\alpha^M(t), \bar{X}^M(t+1)) \end{bmatrix} = \phi(t+1) \end{aligned}$$

Fig. 3 The transition flow of MODEL 1's system states

ここで, θ_1, θ_2 はそれぞれペナルティとナンペナルティパラメータである。

この強化法は, 従来の GNA 強化法⁹⁾ と MGAE⁹⁾ 強化法の混合形で, 定義から分かるように,

$$\bar{x}^i(t) = E[\xi(x_1^i(t), \dots, x_N^i(t))]$$

の取り方により, P-model の学習オートマトンと S-model の学習オートマトンのどちらかに対しても対応できる。

強化法に関する評価基準がいろいろあるが, ここでは, 学習システムの距離短縮性 (distance diminishing property, 以下 d.d.p. と略す)^{13), 14)} について考える。

一般には, 静的な未知環境上で動作する強化法 T が d.d.p. を持つということは, 任意の初期状態 ϕ と ϕ' に関して, $d(T(\phi), T(\phi')) < d(\phi, \phi')$ を満たすような縮小写像 $d(\cdot, \cdot)$ が存在していることである。ここで, $d(\cdot, \cdot)$ はある評価基準に相当する。本文では $d(P, P') = \sum_j |P_j - P'_j|$ について考える。

【定理 2】 $E\text{-MLAG}/N$ において, 強化法 T_{N-C} は d.d.p. を持つ。

証明: 以下, A^i について考えるが, A^i を簡単に A と略す。

1°. 各オートマトンの出力集合と各未知環境の出力集合は有限集合であるので, 可能な事象空間 $E = (a, X)$ は有限空間となる。

2°. 状態 $\phi = (p(t), p(a(t), X(t+1)))$, $\phi \in \Phi$ における事象 $e_t = (a, \bar{x})$, $e_t \in E$ が生起する確率 $Pr(e_t)$ に対して,

$$Pr(e_{t+1} | e_j, 1 \leq j \leq t) = Pr(T_{N-C_{e_1 \dots e_t}}(\phi)), t \geq 1 \quad (27)$$

ここで,

$$T_{N-C_{e_1 \dots e_t}}(\phi) = T_{N-C_{e_t}}(\dots(T_{N-C_{e_1}}(\phi))) \quad (28)$$

な状態遷移関数となる。

3°. 測度 d を持つ距離空間 (metric space) (Φ, d) においては, Φ の中の任意の一点 o を中心に距離 d を半径とする $r = \prod_{i=1}^M m^i$ 次元の球体 B^r が Φ に含まれている。また, Φ は d の意味での閉空間であるので, コンパクト空間をなしている。

4°. 本文では, 静的な未知環境について考察しているので, 異なる確率 p と p' により選択したオートマトンの同じ出力に対する環境からの反応 $\bar{x}$ は変わらない。このことから, 次の関係式が求まる。

$$\begin{aligned} \Delta p_k(t+1) &= p_k(t+1) - p'_k(t+1) \\ &= (1 - \theta_2 \bar{x}^i / N) \Delta p_k(t) \\ &\quad + \theta_1 (1 - \bar{x}^i / N) \sum_{j \neq k} \Delta p_j(t), \quad a^i(t) = a_k^i \end{aligned} \quad (29)$$

$$\begin{aligned} \Delta p_j(t+1) &= p_j(t+1) - p'_j(t+1) \\ &= (1 - \theta_1 + \bar{x}^i / N (\theta_1 - \theta_2)) \Delta p_j(t), \\ a^i(t) &= a_j^i, \quad j \neq k \end{aligned} \quad (30)$$

以上のことから, 学習モデルの最大差分係数 (maximum difference quotients) $l(T_{N-C})$ は以下のように求まる。

$$\begin{aligned} l(T_{N-C}) &= \frac{d(T_{N-C}(P(t)), T_{N-C}(P'(t)))}{d(P(t), P'(t))} \\ &= \frac{|\Delta p_s(t+1) + \sum_{j \neq s} \Delta p_j(t+1)|}{\sum_j |p_j(t) - p'_j(t)|} \\ &= \frac{|(1 - \theta_2 \bar{x}^i / N) \Delta p_s(t) + \theta_1 (1 - \bar{x}^i / N) \sum_{j \neq s} \Delta p_j(t)|}{\sum_j |p_j(t) - p'_j(t)|} \\ &\quad + \frac{\sum_{j \neq s} |(1 - \theta_1 + \bar{x}^i / N (\theta_1 - \theta_2)) \Delta p_j(t)|}{\sum_j |p_j(t) - p'_j(t)|} \\ \frac{N-1}{N} < |l(T_{N-C})| &= \left| 1 - \theta_2 \frac{\bar{x}^i}{N} \right| < 1 \end{aligned} \quad (31)$$

$1^\circ \sim 4^\circ$ から, 距離短縮モデルの定義¹⁴⁾, および Norman の定理¹³⁾ により T_{N-C} は d.d.p. を持つことが分かる。

□

4. シミュレーション例: 商品価格問題

以上で求められた強化法の正当性を検証するために, 次のような簡単な商品価格問題について考察する。

ある町では, 5 人の商人と 2 種類 (合理型, 贅沢型) の消費者が存在しているとす。ある必需商品に対して, 各商人はそれぞれ異なる価格を提示することにより, 消費者を獲得しようとするゲームを展開する。商人をオートマトンとし, 消費者の商品選択行動を未知環境とする。合理型の消費者は, 一番安い商品を探めに行くのに対して, 贅沢型の消費者では, 自分が前もって想定した価格ライン以下のものがあれば場所を問わずに買う。各商人はそれぞれ何種類かの価格を決めておき, ある確率分布に従って, 価格を確率的に消費者に提示することとする。全員が同等の利益を得るために, この確率分布をいかに決めるべきかを商人が目標とする問題を扱う。

これを 5 LAG/2 モデルを用いて考察しよう。シミュレーションの中の各パラメータは以下のとおりとする。

$A = \{A^1, \dots, A^5\}$: 5 人のプレーヤ。

各プレーヤの提示価格をプレーヤの戦略集合とし, Table 1 に示す。たとえば, 1 番目のプレーヤは a^1 の戦略を取る。戦略 a^1 は各商品に対して, $a_1^1 = 10, a_2^1 = 13, a_3^1 = 25, a_4^1 = 30, a_5^1 = 70$ のような価格を提示する。

Table 1 から分かるように、商人に対しても、 α^2 のように価格を抑えて、ばらつきの割合を小さくしようとするものがあれば、 α^4 のようにばらつきの大きい価格を提示するものもある。ここで、最大リスクを最小にするゲームを遂行していくので、前者を慎重型の商人、後者を冒険型の商人と各々呼ぶことにする。

各商人の戦略選択確率分布の初期分布を一様分布とする。

$\varepsilon = 25$: 贅沢型消費者の価格ラインとする。

環境のペナルティ確率分布を Table 2, 3 に示す。

合理型の消費者では、低い価格と対応するペナルティ確率 (Table 2) も全体的に低いのにに対して、贅沢型の

消費者の場合では、商人の提示価格は価格ライン ε を越えると、ペナルティ確率が急に大きくなる。

$\theta_1 = 0.025$, $\theta_2 = 0.001$, $T = 500$, 各オートマトンの確率分布の瞬時変動と 10 回試行の平均変動特性を Fig. 4~8 に示している。収束した時点での確率分布の平均値を Table 4 にまとめている。

Table 4 から分かるように、同じ利益を獲得するために、慎重型の商人 2 においては、消費者に好まれる価格の多数の商品を選択することが可能であるが、冒険型の商人 4 では、結局価格を選択する範囲が狭くなり、販売可能な商品数が少なくなる。

Table 1 The set of outputs

i	α^1	α^2	α^3	α^4	α^5
1	10	9	15	5	11
2	13	15	18	20	20
3	25	21	30	35	25
4	30	25	35	39	35
5	70	40	60	80	45

Table 2 Random environment 1

R_1 : Rational type					
i	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
c_{11}^i	0.15	0.14	0.20	0.10	0.17
c_{12}^i	0.18	0.20	0.22	0.24	0.24
c_{13}^i	0.28	0.25	0.30	0.36	0.28
c_{14}^i	0.30	0.28	0.36	0.40	0.36
c_{15}^i	0.89	0.42	0.82	0.95	0.46

Table 3 Random environment 2

R_2 : Luxurious type					
i	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
c_{11}^i	0.24	0.20	0.31	0.15	0.25
c_{12}^i	0.27	0.31	0.34	0.37	0.37
c_{13}^i	0.41	0.39	0.75	0.78	0.41
c_{14}^i	0.75	0.41	0.78	0.81	0.78
c_{15}^i	0.92	0.83	0.91	0.95	0.86

 Table 4 The equilibrium probabilities $E(p_j^i(10))$

p^1	p^2	p^3	p^4	p^5
0.501411	0.524333	0.498656	0.662697	0.594212
0.426431	0.276062	0.478497	0.320152	0.387525
0.062315	0.167617	0.012915	0.009109	0.006476
0.007212	0.026118	0.006902	0.005571	0.006492
0.002631	0.005870	0.003030	0.002471	0.005294

(a) Transient properties

(b) Average properties

Fig. 4 Evolution of seller 1's price probabilities

Fig. 5 Evolution of seller 2's price probabilities

Fig. 6 Evolution of seller 3's price probabilities

Fig. 7 Evolution of seller 4's price probabilities

Fig. 8 Evolution of seller 5's price probabilities

5. まとめ

本文では、複数の目標のもとでの学習オートマトンの集団特性について考察している。この場合、目標間、オートマトン間、および目標とオートマトン間などの依存性などが考えられるが、ここでは、このようなシステムの状態空間が

$$p(a(t), X(t)) = \prod_{i=1}^N p(a(t), x_i^1(t+1), \dots, x_i^M(t+1)) \quad (32)$$

によって表現できるクラスの問題について分析した。このクラスの問題を N 教師付き学習オートマトン間のゲームに帰着することができ、システムの状態遷移過程がエルゴード的マルコフ過程になることを示した。また、エルゴード的ゲームのナッシュ均衡点が存在し、各プレイヤーのゲームの中での純戦略が (18) 式の意味での同値関係により分割した各部分空間上で求められることを明らかにした。各オートマトンの学習の性質は、d.d.p. を持つ強化法 T_{N-C} を使うことにより保証出来る。

求められた強化法の正当性を検証するために、本文の中では、簡単な商品価格問題について実験を行なった。同じような応用がマーケティングシミュレーション、並列処理マシン上の確率的な順序関係を持つジョブの振り分けなど、多目的最適化の分散処理問題においても考えられる。

今後の課題として、このようなゲームの性質をさらに理論的に解析するとともに、動的に変動する未知環境に対応するモデルについても考察する必要がある。

参 考 文 献

- 1) M. L. Tsetlin : On Behavior of Finite Automata in Random Media ; Automation and Remote Control, Vol. 22, No. 11, pp. 1345~1354 (1961)
- 2) V. I. Varshavskii and I. P. Vorontsova : On the Behavior of Stochastic Automata with Variable Structure ; Automation and Remote Control, Vol. 24, No. 3, pp. 327~333 (1963)
- 3) A. G. Barto and R. S. Sutton : Landmark Learning : An Illustration of Associative Search ; Biol. Cybern, Vol. 42, No. 1-8, pp. 1~8 (1981)
- 4) E. A. Wright, K. S. Narendra and L. G. Mason : Application of Learning Automata to Telephone Traffic Routing and Control ; IEEE Trans., SMC, Vol. 7, No. 11, pp. 785~792 (1977)
- 5) 栗林, 中村, 佐藤 : 大規模パケット通信網における学習的ルーティング方式 ; 電子情報通信学会論文誌, J 63-B (3), pp. 240~245 (1980)

- 6) 銭 平田, 深尾 : 学習オートマトンを用いた分散型ルーティングアルゴリズム ; 電子情報通信学会論文誌, D-I, J 72-D-I (9), pp. 669~677 (1989)
- 7) 銭, 平田 : パターン認識識別関数のパラメータ推定への学習オートマトンの適用 ; SICE '89 学術講演会予稿集, pp. 507~508 (1989)
- 8) S. Lakshmivarahan and K. S. Narendra : Learning Algorithms for Two-Person Zero-Sum Stochastic Games with Incomplete Information ; Mathematics of Operations Research, Vol. 6, No. 3, pp. 379~386 (1981)
- 9) N. Baba : New Topics in Learning Automata Theory and Applications, Springer-Verlag (1984)
- 10) M. Bacharach : Economics and the Theory of Games, Curtis Brown (1976)
- 11) M. Sharpe : General Theory of Markov Processes, Academic Press (1988)
- 12) 銭, 平田 : 2 人教師付き学習オートマトンを用いたコンピュータネットワークのフロー制御アルゴリズム ; 電子情報通信学会論文誌 D-I, J 73-D-I (5), pp. 521~528 (1990)
- 13) M. F. Norman : Markov Processes and Learning Models, Academic Press (1972)
- 14) M. F. Norman : Some Convergence Theorems for Stochastic Learning Models with Distance Diminishing Operators ; Journal of Math. Psychology, Vol. 5, pp. 61~101 (1968)