論文の内容の一部訂正とお詫び

2015 年 6 月 1 日
大分大学工学部電気電子工学科
柴田克成

下記論文中のデータを取得したシミュレーションに誤りがありました。ここに訂正させて頂くとともに，間違ったデータを論文に掲載してしまいましたことをここに深くお詫び申し上げます。ただし，論文で主張していること自体が覆る訂正ではありませんので，本論文で主張していることは非常に重要であるとの認識の下，引き続き主張して参りたいと存じます。

[訂正対象論文]
電子情報通信学会 ニューロコンピューティング研究会 技術報告（信学技報）
柴田克成，坂下悠太
「カオスニューラルネットを用いた内部ダイナミクス由来の探索に基づく強化学習」
Vol. 114, No. 515, NC2014-117, pp. 277-282, 2015 年 3 月

[訂正の概要]
学習結果を検証するプログラム中に複数のバグがありました。これを修正したことに伴い，図３，図４，図５，図６に掲載した結果，および，本文中におけるその説明箇所に変更が入ることになりました。

[訂正内容]
・図３，図４，図５，図６は図の差し替えを行いました。
・４ページ右コラム最後の段落の５行目
(訂正前)
８試行中 2 試行でゴールに到達した。その他の 6 試行も，100 ステップで試行を終了しなければ，1,000 ステップ以内でゴールに到達した。
(訂正後)
８試行中 3 試行でゴールに到達した。その他の 5 試行も，100 ステップで試行を終了しなければ，2,000 ステップ以内でゴールに到達した。
・４ページ右コラム最後の段落の９行目
(訂正前)
最適ステップ数は 9 であるが，10-13 ステップでゴールした。
(訂正前)
少し回り道をすることもあるが，より少ないステップ数でゴールした。
・４ページ右コラム下から２行目
(訂正前)
10 または 11 ステップでゴール
(訂正後)
10 から 12 ステップでゴール
・５ページ左コラム下から２行目
(訂正前)
大きく値が変化している
(訂正後)
(c)の場合より値が変化している
・ ６ページ左コラム５行目
(訂正前)
むしろ増加傾向にあった。さらに面白いことに，環境が変わった直後では，どちらの環境の場合でも，指数はかなり大きな値になった。
(訂正後)
少し減っているもののあまり変化はなかった。また，環境が変わるとすぐに，どちらの環境の場合でも，指数は大きな値になっている。

以上

社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

カオスニューラルネットを用いた 内部ダイナミクス由来の探索に基づく強化学習

柴田　克成†　　坂下　悠太†

† 大分大学工学部電気電子工学科　大分市大字旦野原 700 番地
E-mail: †shibata@oita-u.ac.jp

あらまし　本稿ではまず，自律学習による知能創発における基本的なコンセプトとして，“強化学習で必須の「探索」は学習者内部のダイナミクスの一側面であり，学習とともに「思考」のような合目的でダイナミックな高次機能へと発展する” という新たな考え方を導入する。そして，動作生成をカオスニューラルネットで行って，外部からの乱数の付加なしで内部ダイナミクスに基づいて探索を行うことで,「カオス的遍歴」のようなダイナミクスに基づいた効果的な探索と，それが発展することで安定と遷移の両立が必要なダイナミックな高次機能の学習が期待される。さらに，探索成分を分離できない本手法において，状態価値の TD(Temporal Difference) 誤差と各ニューロンの時間変化との相関に注目した新たな強化学習法を提案し，ゴール到達行動の学習ができることを簡単なタスクで示した。
キーワード　カオスニューラルネット，探索，強化学習，ダイナミクス，思考

Reinforcement Learning Based on Internal-Dynamics-Derived Exploration Using a Chaotic Neural Network

Katsunari SHIBATA† and Yuta SAKASHITA†

† Dept. Electrical & Electronic Engineering, Oita University, 700 Dannoharu, Oita, JAPAN
E-mail: †shibata@oita-u.ac.jp

Abstract　As a basic concept for emergence of intelligence through autonomous learning, exploration that is essential in reinforcement learning is considered as one aspect of the learner's internal dynamics, which is expected to develop through learning towards purposive and dynamic higher functions such as 'thinking'. A chaotic neural network is used to generate motions that include exploratory factors based on the chaotic dynamics without adding external random noises. Effective exploration is expected based on the dynamics such as “chaotic itinerancy”, which is also the key to learn dynamic higher functions more easily that needs both stable and transitive dynamics. Furthermore, a reinforcement learning method without external random noises by focusing on the correlation between the TD (Temporal Difference) error of the state value and the output change of each neuron is newly proposed. It was confirmed that an agent learned goal-directed behaviors in a simple task.
Key words　chaotic neural network, exploration, reinforcement learning, dynamics, thinking

1. はじめに

「知能創発」の実現のためには，学習者自らの試行錯誤を通した自律学習能力を持ち，リカレントニューラルネット (RNN) と組み合わせた際の機能創発能力が示されている強化学習が有望である [1] [2]。強化学習において試行錯誤を実現するために，従来確率的な探索が利用されて来た。離散行動生成の場合には，Q 値を用いた ϵ-greedy やボルツマン選択のように，学習者の出力を元にして一つの行動を選択する [3]。一方，連続値動作生成の場合には，Actor-Critic での Actor のように，学習者の出力によって決まる確率分布から動作が決定される [1] [2]。

著者の一人は，従来の探索に対して以下のような問題点を感じ，自律学習システムにおける知的な探索の実現を目指して来た。1) 人間の探索はもっと知的であるように見える。われわれが分かれ道に遭遇したとき，各アクチュエータ単位でばらばらに確率的動作を行う訳ではなく，分かれ道のどちらかを選ぶであろうが，その行動自体は，たくさんのアクチュエータの動きを統合した優れた動作である。 2) われわれ生物は，探

索のためだけの乱数発生器を有しているのだろうか？ 3) 探索 (exploration) と活用 (exploitation) の割合を決める適切な基準は何であるか？ 4) 探索が制御間隔に影響される。たとえば，付加する乱数の大きさが同じであっても，制御間隔が 10 倍になれば全く違う探索となる。 といった点である。

そこで筆者らのグループでは，リカレントニューラルネット (RNN) を用いた強化学習で，知的な探索を獲得できることを示した [4] [5]。しかし，その探索を学習するために，外部の乱数を利用した試行錯誤が必要であった。また，確率的探索なしでの学習を目指し,「楽観的初期値」[3] の拡張として,「状態または行動価値の逐次増加法」を考えたが，価値と動作の学習後に未知の状態へ行くことができるようになるため，スピィーディな探索ができなかった [6]。さらに，抽象表現空間での知的探索を目指し，RNN の中間層ニューロンに外部から摂動を与えることを試みたが，ダイナミクスの効果的な移行を引き起こすことは難しい上，摂動の与え方が問題として残った [7]。

一方，カオスダイナミクスが効果的な探索を実現する能力を持つことは良く知られている。昔からたくさんの研究がされて来たが，その能力は主に連想記憶の領域で検証されて来た。強化学習において，乱数の代わりにカオス系列を用いる研究も行われている [8] [9]。しかし，カオス系列生成器は行動生成の外部に置かれ，行動生成の内部ダイナミクスそのものによって生成されるものではなかった。

有江らは，通常の RNN とカオスネット (CNN) を結合した学習システムを提案した。ここでは， CNN の初期値鋭敏性を利用して，基本動作系列を組み合わせたゴールを目指す新たな動作系列を探索した [10]。内部ダイナミクスを利用した探索という点で先駆的な研究であるが，最初に，人間が提示した基本動作系列を再現するように，RNN が現在の状態から次の時刻の状態を予測する学習を行った後，CNN の初期状態を変化させることで，基本動作の組み合わせを探索した。この際，CNN は探索にその役割を特化されており，それ自身は学習せず，系列を生成する RNN との間で予め役割分担がされている。

本稿では，探索は，行動生成のニューラルネット (Actor) において形成される内部ダイナミクスの一側面であり，そのダイナミクスが学習を通して成長することで,「思考」のような高次の動的機能につながるという新しい考え方を提起する。別の言い方をすると,「探索」と「思考」はともに「内部ダイナミクス」という同じ線上で扱われる。学習は報酬を得るために有用なダイナミクス，たとえば，世の中の様々な因果関係を取り込んだりすることで,「探索」から「思考」へと成長して行く。そして，学習者が不確かな，または，未知の状況に遭遇すると，そのダイナミクスは「探索」に戻る。「探索」「思いめぐらす」「思考」の類似性からこの考えに至った。「思考」は状態の「安定」と状態間の「遷移」の両者を必要とする。この 2 つの要求は一件矛盾しているように見え，両者を共存させることは難しい [11]。

カオスダイナミクスはこれらの要求を満たしているように見える。特に,「カオス的遍歴」[12] は，前述の分かれ道での効果的な探索や思考のダイナミクスの起源であり，カオスダイナミクスが「探索」と「思考」の橋渡しをしていると考えることができる。「カオス的遍歴」は生物でも実験的に観察されており，既知と未知の刺激がダイナミクスの違いとして観察されている [13]。似たような結果が，CNN を用いたシミュレーションでも見られており，さらに，その CNN は未知のものを学習することができる [14]。前述の [10] と同様な枠組みで状態系列を学習させる際に，ある状態での系列を 2 つ用意し，毎回ランダムに選んで学習させると，カオスダイナミクスが発生したという報告もある [15]。これらは，学習によるカオスダイナミクスの合目的な制御の可能性を示唆していて大変面白い。

ここまでの議論にしたがって，著者らは，動作生成自体に CNN を用い，それを強化学習で学習させることを提案する。内部のカオスダイナミクスによって，学習者は外部からの摂動，雑音としての乱数の付加なく探索行動を生成でき，学習を通してそのダイナミクスが合目的的に変化することが期待される。

しかしながら，一つの大きな問題が存在する。従来の強化学習では，与えた確率的な摂動または雑音と，その結果としての状態価値または行動価値の変化の期待値との差である TD(Temporal Difference) 誤差との相関 (積) を用いて適切な動作を学習している。その際，行動生成とは別に摂動を与えているため，それを行動や動作と分離することは容易であった。しかし，提案システムでは，探索は内部ダイナミクスとして生成されるので，摂動と元の動作とを区別することは元々不可能である。そこで，状態価値における TD 誤差と各ニューロンの出力変化の相関に注目し，CNN の内部ダイナミクスによって生成される探索に基づく全く新しく，簡単な強化学習アルゴリズムを提案する。そして，簡単なタスクで，エージェントがゴールへ向かう動作を学習できることを示す。

2. カオスニューラルネットを用いた強化学習

ここでは，アクチュエータに直接連続値の動作信号が送られると考え，その学習に適している，Actor-Critic タイプの強化学習を考える。従来のニューラルネットを用いた連続値動作の強化学習では，図 1(a) のように，各 Actor に外部乱数を加えることで確率的に一つの動作が決定される。提案する学習システムでは，図 1(b) のように，CNN が Actor として学習者の動作を生成する。前述のように，学習者は，外部からの乱数の提供なくして CNN の内部ダイナミクスにしたがって探索する。

より良い探索のため，より良いダイナミクスの学習のためにどのようなタイプの CNN を用い，どのようなパラメータを使うかは重要な問題である。しかし，ここでは，この新たな考え方の導入の第一段階として，とにかく強化学習が CNN を用いた外部乱数なしで学習できることを示すことに焦点を当てる。ここで使う CNN は 3 層構造で，中間層に 50 個のニューロンがあり，互いに結合されている。CNN としては，不応期を持つカオスニューロンモデル [16] を使うことが多いが，ここでは簡単のため，通常の静的なニューロンモデルを用い，ニューロンの出力関数であるシグモイド関数のゲインを 7.0 とし，大きめのランダムな重み値で相互結合することでカオスダイナミクスを発生させる。中間層，出力層の各ニューロンの出力 o は，次の式のように計算する。

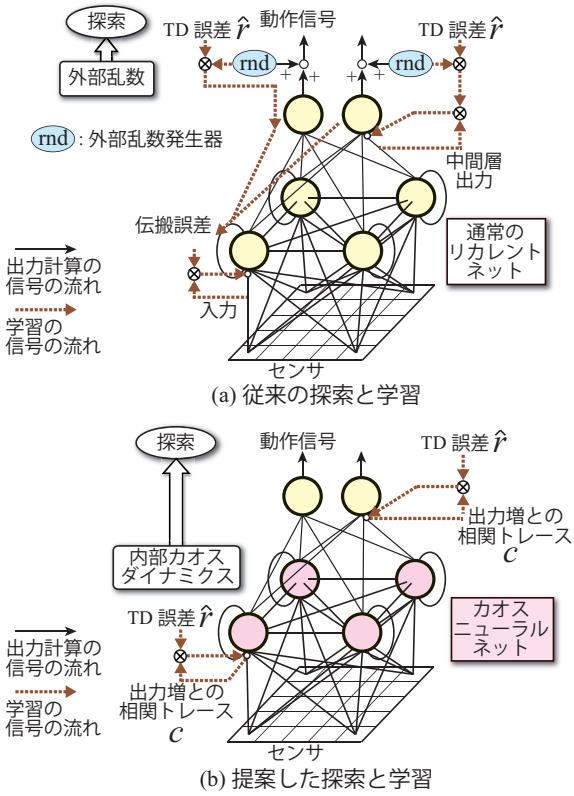

図 1　外部摂動による探索に基づく従来の強化学習と本稿で提案するカオスニューラルネット (CNN) を用いた内部ダイナミクスによる探索に基づく強化学習の比較

$$u_{j,t}^{(2)} = \sum_{i=1}^{N^{(1)}} w_{j,i}^{(2)} o_{i,t}^{(1)} + \sum_{i=1}^{N^{(2)}} w_{\mathrm{fb}j,i}^{(2)} o_{i,t-1}^{(2)} \tag{1}$$

$$u_{j,t}^{(3)} = \sum_{i=1}^{N^{(2)}} w_{j,i}^{(3)} o_{i,t}^{(2)} \tag{2}$$

$$o_{j,t}^{(l)} = \frac{1}{1 + exp(-g \cdot u_{j,t}^{(l)})} - 0.5 \tag{3}$$

ここで，上付きの添字 (1)〜(3) は，それぞれ入力層，中間層，出力層を表す。$u_{j,t}^{(l)}$，$o_{j,t}^{(l)}$ は時刻 t の第 l 層の j 番目のニューロンの内部状態と出力 (ただし，$o^{(1)}$ はセンサからの入力信号)，$w_{j,i}^{(l)}$ は第 l 層の j 番目のニューロンへの第 $l-1$ 層からの i 番目の入力に対する結合の重み値，$N^{(l)}$ は第 l 層のニューロン数である。$w_{\mathrm{fb},j,i}^{(2)}$ は，j 番目の中間層ニューロンの i 番目の中間層ニューロンからのフィードバック結合の重み値である。g はシグモイド関数のゲインであり，中間層，出力層ニューロンともに 7.0 とした。また，シグモイド関数は，値域が-0.5 から 0.5 のものを用いた。

筆者らは，Critic の計算も最終的には同一のニューラルネットで行うべきと考えているが，ここでは，フィードバック結合のない階層型ニューラルネットで Actor ネットと同じ入力の Critic ネットが別にあり，Critic のみを出力するとした。使用するシグモイド関数のゲインは 1.0，値域は-0.5 から 0.5 とした。

Critic ネットは従来と同様な方法で学習する。教師信号 tr_{critic} を TD (Temporal Difference) 学習に基づいて

$$tr_{critic,t} = \gamma o_{critic,t+1} + r_{t+1} = \hat{r}_t + o_{critic,t} \tag{4}$$

と計算する。ここで，γ は割引率，$o_{critic,t+1}$ は時刻 $t+1$ での Critic ネットの出力，r_{t+1} は時刻 $t+1$ で与えられた報酬，$\hat{r}_t$ は TD 誤差で

$$\hat{r}_t = \gamma o_{critic,t+1} + r_{t+1} - o_{critic,t} \tag{5}$$

と表される。ここでは，学習者がゴールして報酬を得ると当該試行 (エピーソード) を終了するタスクを用いるが，その際には，

$$tr_{critic,t} = r_t \tag{6}$$

を教師信号とする。そして，これらの教師信号を用いて，CNN 内のすべての重み値を誤差逆伝搬法で毎ステップ更新する。

Actor CNN では，前述のように，内部ダイナミクスで探索行動を生成するため，探索成分を Actor の出力と分離できない。そのため，Actor CNN の学習として，従来のものと全く異なる方法をここで新しく提案する。中間層の相互結合を除いたすべての結合は，TD 誤差を用いて

$$\Delta w_{j,i,t}^{(l)} = \eta \hat{r}_t c_{j,i,t}^{(l)} \tag{7}$$

と更新する。ここで，$\Delta w_{j,i,t}^{(l)}$ は第 l 層の j 番目のニューロンの第 $l-1$ 層の i 番目のニューロンからの重み値の時刻 t での更新量である。$c_{j,i,t}^{(l)}$ は第 l 層の j 番目のニューロンの出力の増加と第 $l-1$ 層からの入力との相関のトレースである。連続時間で書くと，トレース $c_{j,i,t}^{(l)}$ の変化は

$$dc_{j,i,t}^{(l)} = -c_{j,i,t}^{(l)} |do_{j,t}^{(l)}| + o_{i,t}^{(l-1)} do_{j,t}^{(l)}, \tag{8}$$

となり，離散時間で書くと，各ステップで

$$c_{j,i,t}^{(l)} = (1 - |\Delta o_{j,t}^{(l)}|) c_{j,i,t-1}^{(l)} + \Delta o_{j,t}^{(l)} o_{i,t}^{(l-1)}. \tag{9}$$

と更新される。そのトレースは当該入力とそのニューロンの増加との相関を表現する。したがって，結合重み値は，式 (7) によって，TD 誤差が正のときはそのニューロンの変化を促進し，TD 誤差が負のときは，その変化を抑制するように変化する。

従来法との大きな違いとして，ニューラルネット内の信号の逆伝搬がないことも挙げられる。個々の中間層ニューロンにとって，入力信号は他のニューロンと同じであるが，トレース $c_{j,i}$ は中間層ニューロン同士で異なり，それが信号の逆伝搬のない学習を可能にしている。しかし，その妥当性についてはより検証を積み重ねる必要がある。また，学習によって，安定と遷移を両立する有用なダイナミクスを獲得することが CNN 導入の大きな目的であるが，ここでは，第一段階として，中間層ニューロン同士の相互結合の重み値は学習させなかった。

3. シミュレーション

この節では，前述の学習が機能するかどうかを簡単なタスクで確認する。20 × 20 のフィールドがあり，エージェントは各試行ごとにランダムな位置に置かれる。エージェントが動いてフィールド中央の半径 2.0 の円の中に入ったら 0.4 の報酬をもらい，試行は終了する。また，2 × 2 の受容野を持つ視覚センサ

表 1　シミュレーションで用いたパラメータ

		Actor CNN	Critic NN
層数		3	3
入力の信号数		121	121
中間層ニューロン数		50	30
出力数		2	1
シグモイド関数の傾き係数		7.0	1.0
シグモイド関数の値域		-0.5 - 0.5	-0.5 - 0.5
学習係数	出力　←中間層	4.0	4.0
	中間層←中間層	0.0	-
	中間層←入力	0.4	4.0
初期重み値	出力　←中間層	±0.1	0.0
	中間層←中間層	±1.5	-
	中間層←入力	±0.1	±0.1
割引率 γ		0.98	
ゴール時の報酬 r		0.4	

セルを 11×11 個配置した視覚センサがあり，重なりなくフィールド全体をカバーしている。エージェントのサイズも 2×2 であり，各セルはその受容野中にエージェントが占める面積を出力する。エージェントの画像がちょうどセルと重なると，そのセルの出力は 4.0，その他のセルの出力はすべて 0 となる。全部で 121 個の信号が両ネットワークに入力として送られる。

Actor CNN には 2 つの出力があり，それぞれエージェントの x または y のどちらかの方向の移動を担当する。5,000 試行ごとに環境が変わり，各出力が担当する方向が入れ替わる。各出力は，加速度として，

$$v_{x,t} = 0.9v_{x,t-1} + o^{(3)}_{1or2,t} \tag{10}$$

$$v_{y,t} = 0.9v_{y,t-1} + o^{(3)}_{2or1,t}, \tag{11}$$

と働く。ただし，$v_{x,t}$ と $v_{y,t}$ はそれぞれ x，y 方向の速さを表すが, -1.0 から 1.0 の範囲から出た場合は-1.0 または 1.0 とした。各ステップでエージェントは動くが，移動方向によって最大速度が変わらないようにするため，その速度は，その方向の最大可能速さ max_v_t で

$$loc_{x,t} = loc_{x,t-1} + v_{x,t}/max\_v_t \tag{12}$$

$$loc_{y,t} = loc_{y,t-1} + v_{y,t}/max\_v_t \tag{13}$$

と割って正規化する。たとえば，v_x と v_y のどちらかが 0.0 のときは，$max\_v = \sqrt{1.0^2 + 0.0^2} = 1.0$ となり，v_x と v_y が等しいときは，max_v は $\sqrt{1.0^2 + 1.0^2} = \sqrt{2}$ となる。フィールド境界の壁に衝突した場合は，そこで止まり，v_x，v_y ともに 0.0 とした。罰は与えなかった。エージェントの動きは動的であるが，エージェントが速度を知る入力がないため，知覚混同 (perceptual aliasing) の状態になる。表 1 は使用したパラメータを示す。各試行でのステップ数が多くなって，$0.125^{\frac{1}{step}}$ が割引率 γ を越えた場合は，critic の値が小さくなり過ぎないように，割引率 γ を一時的に $0.125^{\frac{1}{step}}$ とした。

1 回の学習で 20,000 試行行った。さらに，初期重み値やエージェントのスタート位置を決める乱数系列を変えて 20 回シミュレーションを行った。以下の結果は，その中で一番良い結果である。学習中，5,000 試行に一回環境が変わったにも関わらず，

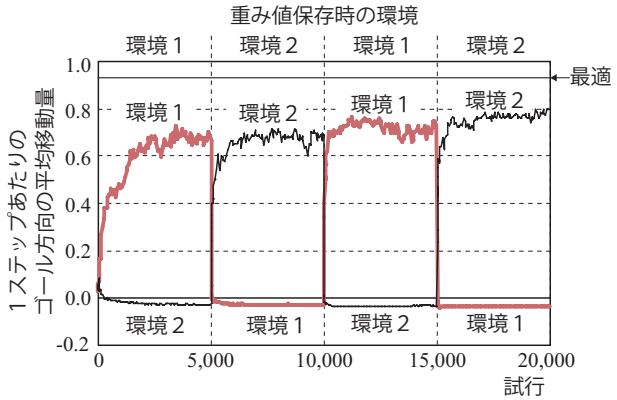

図 2　学習曲線として，1 ステップあたりのゴール方向への平均移動距離の変化を示す。横軸が表す試行回数の学習終了後に保存したネットワーク用いてテスト試行を行い，そのときの値をプロットした。学習時は 5,000 試行ごとに環境が変化した。

ゴール到達までに 2,000 ステップ以上かかった試行はなかった。図 2 に学習曲線を示す。このグラフは，図の横軸に書かれた試行終了後に保存した Actor CNN を用い，200 回のテスト試行における 1 ステップあたりのゴール方向への平均移動距離をプロットした。テスト試行は，フィールド中央から半径 9 の円周を 200 分割した各点をスタートとし，100 ステップでゴールできない場合は，試行を終了した。グラフ中の 2 つの線は，環境 1 の場合と環境 2 の場合の結果を示している。移動距離は最適値に届かないものの，学習していた環境においてはゴールに到達することを学習していることがわかる。環境が変わった後も，外部からの乱数の供給がないにもかかわらず，エージェントは動作を学習できていることがわかる。ネットワーク保存時と異なる環境では，平均移動量は負の値となっている。これは，100 ステップ終了後に，元の位置よりゴールから遠いところにいたことを表している。

図 3 中の各グラフでは，8 個の場所からスタートした 8 回のテスト試行におけるエージェントの軌道を表している。100 ステップを越えた場合は，試行を打ち切った。学習前の (a) を見ると，エージェントはカオスダイナミクスによって動き回り，8 試行中 3 試行でゴールに到達した。その他の 5 試行も，100 ステップで試行を終了させなければ，2,000 ステップ以内でゴールに到達した。100 試行の学習後の (b) では，ゴール方向に進む傾向が見え，8 試行すべてで 100 ステップ以内にゴールした。5,000 試行後 (c) は，エージェントは，少し回り道をすることもあるが，より少ないステップ数でゴールした。同じ CNN を用いて，エージェントを環境 2 に置くと，つまり，出力が担当する x と y を入れ替えると (d)，エージェントは 30,000 ステップかかってもゴールすることができなかった。しかし，両ネットを学習しながらテスト試行を行うと，いずれも 3,000 ステップ以内にゴールすることができた。さらに，5,040 試行後，つまり，学習環境が変わって 40 試行後の (e) では，新しい環境で 8 試行いずれもゴールすることができるようになっており，10,000 試行後の (f) では，10 から 12 ステップでゴールすることができた。

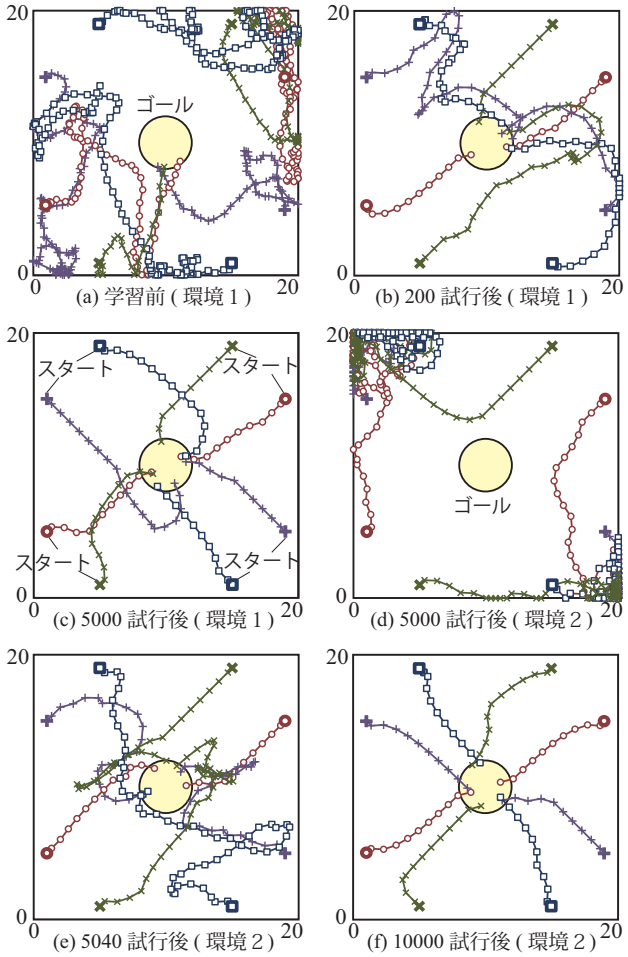

図 3　学習による，8 カ所からのエージェントの軌道の変化。(c) と (d) では，5,000 試行後の環境が 1 から 2 に変わる直前に保存されたネットワークを用いているが，テスト環境が異なっている。少し大きな太いマークはエージェントのスタート地点を表す。

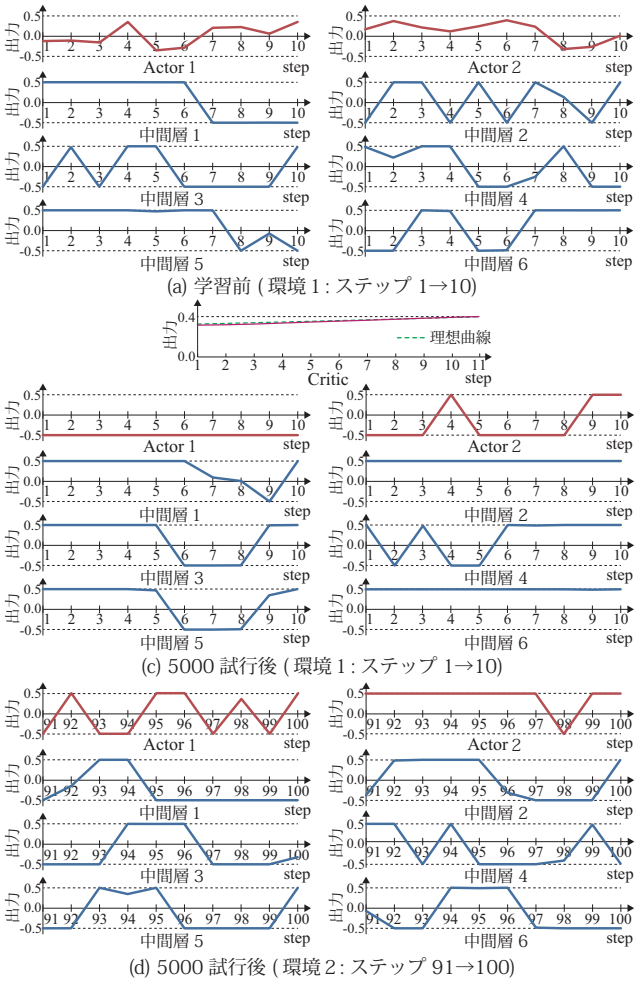

図 4　エージェントが (19.0, 15.0) からスタートしたときの 2 つの Actor 出力と最初の 6 個の中間層ニューロンの出力の試行中の変化。(c) では，Critic の時間変化のグラフも掲載した。また，(a) と (c) ではスタート後の 10 ステップ，(d) ではスタートから時間が経過した後の 10 ステップ分の変化である。

図 4 には，図 3 で示した 6 個の場合のうちの 3 個の場合に，(19.0, 15.0) からスタートしたテスト試行での 2 つの出力ニューロンと最初の 6 個の中間層ニューロンの試行中の出力の変化を示す。ただし，(a) と (c) は，スタートから 10 ステップ分の変化を，(d) はスタートから時間が経過した 91 から 100 ステップまでの変化を示している。学習前 (a) は，2 つの Actor 出力は-0.5 や 0.5 までは行かないが，頻繁にその値を変化させている。また，中間層ニューロンの出力は，-0.5 と 0.5 を中心に頻繁に値を変化させていた。これは，中間層ニューロン間のフィードバック結合の初期値とシグモイド関数のゲインが大きい値だからと考えられる。5,000 試行後の (c) では，一番上のグラフのように，Critic の出力が理想値より少し小さいところがあるものの状態価値を表すようになっていた。また，Actor の出力は必要に応じて-0.5 か 0.5 の値を取っており，(a) のときのように頻繁に変化していない。中間層出力も，出力よりは変化が見られるものの，(a) のときのようには変化していない。一方，(c) と同じ CNN で出力の担当方向が入れ替わった環境 2 での (d) では，図 3(d) のように，エージェントは右下の方に行ってしまうが，この時の出力および中間層の出力ともに (c) の場合より値が変化していることがわかる。このことが探索行動の生成を助けていると考えられる。

探索行動の指標として，リアプノフ指数，つまり，微小摂動に対する感度を観察した。ここでは，前回同様，200 回のテスト試行を行った。200 試行中の各ステップにおいて，10^{-3} の大きさの乱数ベクトルを Actor CNN の 50 個の中間層ニューロンの出力に付加し，1 ステップ後に摂動を加えなかった場合との距離の広がり

$$\lambda = \frac{1}{\sum_{e=1}^{Episode} Step(e)} \sum_{e=1}^{Episode} \sum_{t=1}^{Step(e)} log_2 \frac{d(e,t)}{10^{-3}} \qquad (14)$$

を両環境の場合について観察した。ここで，$d(e,t)$ は，e 番目の試行で，時刻 $t-1$ で 10^{-3} の摂動を与えたときの時刻 t での中間層ニューロンの，摂動を与えないときに対するユークリッド距離，$Episode$ はテストの試行数であり，ここでは 200 である。また，$Step(e)$ は，e 番目の試行でのゴールまでに掛かったステップ数である。本稿では，この値をリアプノフ指数として用いた。このダイナミクスは，Actor CNN の中間層のループだけでなく，エージェントの知覚ー動作のループも一緒になって作り出されるものである。

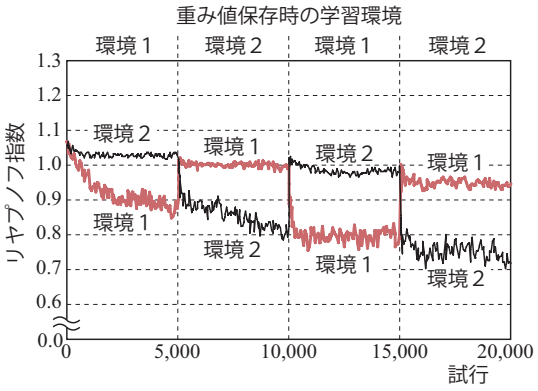

図 5　学習の進行による両環境でのリヤプノフ指数の変化。

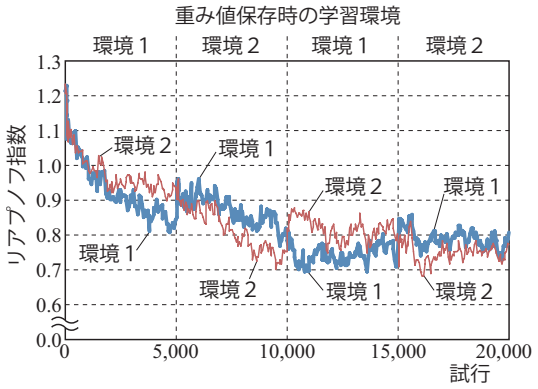

図 6　スタートから最初の 4 ステップの間のリヤプノフ指数の変化

図 5 は，学習の進展によって，CNN を保存した際にエージェントが学習していた環境とそうでない環境で，リアプノフ指数がどう変化したかを示す。エージェントが学習していた環境では，学習が進むとともに指数は小さくなり，学習していなかった環境では大きな値のままで，少し減っているもののあまり変化はなかった。また，環境が変わるとすぐに，どちらの環境の場合でも，指数は大きな値になっている。

スタートから 4 ステップだけ観察すると，図 6 のように，学習していた環境とそうでない環境で多少の差はあるものの，学習していない環境でも指数が学習とともに減少した。このことは，試行中のステップ数が増えた際に，何らかの理由により指数が大きくなったことを表している。この結果は，エージェントがゴールできなかった図 4(d) で，時間が経過すると中間層ニューロンの出力が大きく変動していたことと一致する。

4. 結　　論

カオスニューラルネット (CNN) を用いることで，エージェントが内部のカオスダイナミクスに基づいて，外部からの出力への乱数の付加なしに探索し，ゴールに向かう動作を学習することができた。提案学習法では，入力信号の出力増加との相関を計算し，それと TD 誤差から各重み値を更新した。また，微小な摂動に対する感度が，学習している環境では，学習の進行に応じて小さくなることを確認した。これは，学習を通した合理的なダイナミクスの制御の可能性を示していると考えている。本稿は，強化学習における全く新しい考え方の導入であり，まだまだ多くの課題が残されているものの，カオスダイナミクスによる「探索」の延長として「思考」のような合目的で複雑なダイナミクスを学習によって獲得することが期待される。

謝　　辞

本研究は，静岡理工科大学の金久保教授のカオスニューラルネットに関する Web ページに助けられた。ここに謝意を表す。

文　　献

[1] K. Shibata: Emergence of Intelligence through Reinforcement Learning with a Neural Network, *Advances in Reinforcement Learning*, Abdelhamid Mellouk (Ed.), InTech, 99–120, 2011

[2] 柴田克成: 強化学習とニューラルネットによる知能創発, 計測と制御, **48**(1):106–111, 2009

[3] R. S. Sutton & A. G. Barto: *Reinforcement Learning: An Introduction*, A Bradford Book, The MIT Press,1998

[4] K. Goto & K. Shibata: Acquisition of Deterministic Exploration and Purposive Memory through Reinforcement Learning with a Recurrent Neural Network, *Proc. of SICE Annual Conf. 2010*, FB03-1.pdf, 2010

[5] K. Shibata: Learning of Deterministic Exploration and Temporal Abstraction in Reinforcement Learning, *Proc. of SICE-ICCAS (SICE-ICASE Int'l Joint Conf.)*, 4569–4574, 2006

[6] 藤田剛: 強化学習における状態評価値の逐次調整に基づく適応的な一様探索行動の研究, 大分大学工学研究科電気電子工学専攻修士論文, 2006

[7] 品矢裕介: 強化学習を行うリカレントネットにおける合目的な非固定点収束ダイナミクスの形成, 大分大学工学研究科電気電子工学専攻修士論文, 2014

[8] A. B. Potapov & M. K. Ali: Nonlinear Dynamics and Chaos in Information Processing Neural Networks, *Differential Equations and Dynamical System*, **9**(3&4):259–319, 2001

[9] K. Morihiro, T. Isokawa, N. Matsui & H. Nishimura: Effects of Chaotic Exploration on Reinforcement Learning in Target Capturing Task, *Int. J. Knowledge-based and Intelligent Engineering Systems*, **12**:369–377, 2008

[10] H. Arie, T. Endo, T. Arakaki, S. Sugano & J. Tani: Creating Novel Goal-Directed Actions at Criticality: A Neuro-Robotic Experiment, *New Mathmatics and Computation*, **5**(1): 307–334, 2009

[11] 田口優馬，柴田克成: リカレントネットによる内部状態遷移を要する問題学習時の初期重み値の影響, SICE 九州支部学術講演会論文集, 87–90, 2011

[12] K. Kaneko & I. Tsuda: Chaotic Itinerancy, *Chaos*, **13**: 926–936, 2003

[13] C. A. Skarda & W. J. Freeman: How brains make chaos in order to make sense of the world, *Behavioral and Brain Science*, **10**,161–195, 1987

[14] Y. Osana & M. Hagiwara: Successive Learning in hetero-associative memory using chaotic neural networks. *Int. J. Neural Systems*, **9**(4):285–299, 1999

[15] J. Namikawa, R. Nishimoto & J. Tani: A Neurodynamic Account of Spontaneous Behaviour, *PLoS Computational Biology*, 7(10): e1002221, 2011

[16] K. Aihahara: Chaotic Neural Networks, *Bifurcation Phenomena in Nonlinear Systems and Theory of Dynamical Systems*, H. Kawakami ed., World Scientific, 143–161, 1990

[17] D.E. Rumelhart, G.E. Hinton, R.J. Williams: Learning Internal Representation by Error Propagation. In: Parallel Distributed Processing, MIT Press, Cambridge, **1**:318–364, 1986