カオスニューラルネットを用いた強化学習における高次探索の創発

大分大学　○後藤祐樹　柴田克成

Emergence of Higher Exploration in Reinforcement Learning using a Chaotic Neural Network

Yuki Goto and Katsunari Shibata, Oita University

Abstract: Aiming for the emergence of higher functions such as "logical thinking", our group has proposed completely novel reinforcement learning where exploration is performed based on the internal dynamics of a chaotic neural network. In this paper, in the learning of an obstacle avoidance task, it was examined that in the process of growing the dynamics through learning, the level of exploration changes from "lower" to "higher", in other words, from "motor level" to "more abstract level". It was shown that the agent learned to reach the goal while avoiding the obstacle and there is an area where the agent looks to pass through the right side or left side of the obstacle randomly. The result shows the possibility of the "higher exploration" though the agent sometimes collided with the obstacle and was trapped for a while as learning progressed.

1　序論

ロボットは人間が与えたプログラムに基づいた行動を実行できるが、実世界のあらゆる状況に柔軟に対応できる行動を与えることは困難である。これに対し、我々の研究室は脳のように超並列で柔軟な処理ができるニューラルネットを使って強化学習することで、ニューラルネット内に様々な機能を自律的に獲得することを提唱してきた[1)]。また、これによって「思考」などの高次機能がニューラルネット内部に創発されることも期待される。近年では多層ニューラルネットである Deep Learning が特徴空間を自律形成することで認識分野で力を発揮しており[2)]、さらに Deep Learning に強化学習を組み合わせて TV ゲームを学習させ、人間以上のスコアを出した[3)]ことなどは、我々のアプローチの有効性を支持している。

高次機能は記憶、予測、思考など、時間軸やダイナミクスを扱っていくことが必須であり、そのためにリカレントニューラルネットの導入を試みてきた。そして簡単なタスクで記憶や予測の学習ができることを確認した[4)][5)]が、複数の状態遷移が必要なタスクの学習[6)]は簡単ではなく、最も典型的な高次機能である「思考」と呼べる機能の創発は確認できていない。

強化学習では「探索」を必要とする。そして、「探索」と「思考」はどちらも自発的な状態遷移を伴う内部ダイナミクスであるという類似点がある。そこで、カオスニューラルネットを用い、その内部のカオスダイナミクスで「探索」をし、学習することでそれが「思考」へと成長すると考えた。そこで、従来の強化学習のように探索のために外部から乱数を付加するのではなく、カオスニューラルネットのカオスダイナミクスに基づいた探索による全く新しい強化学習を提案し、簡単な物体到達タスクを学習できることを示した[7)]。

我々は分かれ道の前で「右へ行く」か「左へ行く」かの選択をする。この時我々は、「道に沿って行った方が良い」という過去に学習したことを利用しており、モータレベルのランダムな探索とは異なる。我々はこれを「高次探索」と呼び,「探索」から「思考」へと成長する途中段階と位置付けている。本論文では、この分かれ道を障害物回避タスクでの障害物の右側を通るか左側を通るかという探索に置き換え、高次探索の創発の可能性について検討した結果を示す[8)]。

2　カオスニューラルネットを用いた強化学習

強化学習を用いることで、学習者が探索（試行錯誤）を通して、より多くの報酬を獲得し、罰を避けるような行動を自律的に学習することができる。一般的にエージェントは学習部の外から供給される乱数に基づいた探索を行うが、前節で述べたように本論文では乱数は付加せず、ニューラルネット内部のカオスダイナミクスに基づいた探索を行う。本論文では連続値を扱うために Actor-Critic と呼ばれる強化学習を用い、動作出力を行う Actor 部にカオスニューラルネットを用い、状態評価値を出力する Critic 部はカオスでない通常の階層型ニューラルネットを用いる。時刻 t での l 層 j 番目ニューロンの内部状態 $u_{j,t}^{(l)}$ は

$$u_{j,t}^{(l)} = \sum_{i=1}^{N^{(l-1)}} w_{j,i}^{(l)} o_{i,t}^{(l-1)} \left(+ \sum_{i=1}^{N^{(l)}} w_{j,i}^{\mathrm{FB}} o_{i,t-1}^{(l)} \right) \qquad (1)$$

で求められ、ここで $N^{(l)}$ は l 層のニューロン数、$w_{j,i}^{(l)}$ は l 層 j 番目ニューロンの i 番目の結合重み値、$o_{j,t}^{(l)}$ は l 層 j 番目ニューロンの出力を表す。右辺第 2 項はカオスニュー

ラルネットの中間層のみに用いられ、$w^{FB}_{j,i}$ は中間層 j 番目ニューロンの i 番目のフィードバック結合の重み値を表す。ニューロンの出力関数には値域 (-0.5,0.5)、ゲインが ϵ のシグモイド関数 $f(x) = 1/\{1+\exp(-\epsilon \cdot x)\} - 0.5$ を通しており、$o^{(l)}_{j,t} = f(u^{(l)}_{j,t})$ である。カオスニューラルネットは動作信号として二つの Actor 出力 $\mathbf{A}(\mathbf{S}_t)$、階層型ニューラルネットは状態評価値として一つの Critic 出力 $V(\mathbf{S}_t)$ を持ち、$\mathbf{S}_t$ は時刻 t でのセンサー入力を表す。

学習のための TD 誤差 $\hat{r}_t$ は

$$\hat{r}_t = r_{t+1} + \gamma V(\mathbf{S}_{t+1}) - V(\mathbf{S}_t) \tag{2}$$

で求められ、r_{t+1} は時刻 $t+1$ で与えられる報酬、γ は割引率を表しており、0.95 としている。時刻 t での Critic 部の教師信号 T_{V_t} は次式で求められる。

$$T_{V_t} = V(\mathbf{S}_t) + \hat{r}_t = r_{t+1} + \gamma V(\mathbf{S}_{t+1}) \tag{3}$$

この教師信号 T_{V_t} を用いて誤差逆伝搬法によって階層型ニューラルネットを学習させる。式 (3) の T_{V_t}、$V(\mathbf{S}_t)$ は [0,1] の値域を前提としているため、Critic 部の出力との間で 0.5 の加減算をして値域を調整する。

本論文でのカオスニューラルネットは中間層ニューロン同士の強いフィードバック結合によりカオスダイナミクスを生成している。結合重み値の更新量 $\Delta w^{(l)}_{j,i,t}$ は因果トレース $c^{(l)}_{j,i,t}$ を用いて

$$\Delta w^{(l)}_{j,i,t} = \eta^{(l)}_A \hat{r}_t c^{(l)}_{j,i,t} \tag{4}$$

で求められ、$\eta^{(l)}_A$ はカオスニューラルネット l 層の学習係数を表す。因果トレース $c^{(l)}_{j,i,t}$ はニューロンの出力の変化 $\Delta o^{(l)}_{j,t} = o^{(l)}_{j,t} - o^{(l)}_{j,t-1}$ を用いて式 (5) で計算され、入力 $o^{(l-1)}_{i,t}$ が出力の増加にどれだけ影響しているのかを表す。

$$c^{(l)}_{j,i,t} = (1 - |\Delta o^{(l)}_{j,t}|) c^{(l)}_{j,i,t-1} + \Delta o^{(l)}_{j,t} \cdot o^{(l-1)}_{i,t} \tag{5}$$

また、本論文ではフィードバック結合の重み値 $w^{FB}_{j,i}$ は学習がうまくいかないので、更新を行っていない。

3　シミュレーション

本論文では高次探索の創発を確認するために、20×20 のフィールドにゴールと障害物を設置した Fig.1 のような障害物回避タスクのシミュレーションを行う。半径 1 のゴールを (0,8) の位置に固定し、半径 1.5 の障害物は毎試行の始めにランダムな位置に設置する。エージェントはカオスニューラルネットの出力により動き、ゴールに到達すると 1 の報酬を獲得する。また、障害物に衝突すると-0.01 の罰が与えられる。エージェントがゴールに到達する、またはステップ上限である 1,000 ステップとなるとその試行を終了する。Fig.1 に示した 6 個の入力情報

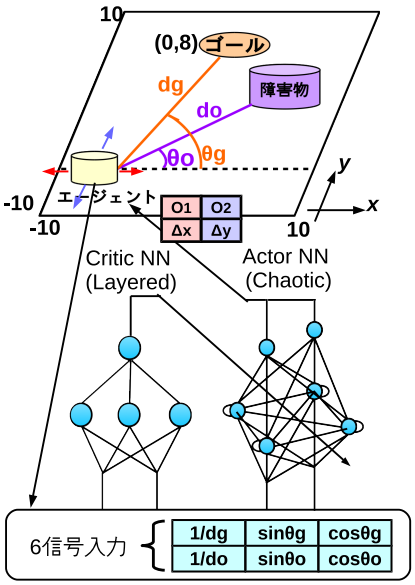


Fig. 1　Reinforcement learning system and the obstacle avoidance task used in this paper

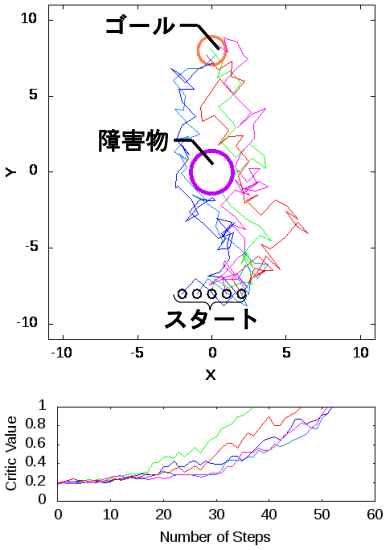
Table 1　The parameters used in the simulation

		Actor 部	Critic 部
中間層ニューロン数		100	10
出力関数の値域		[-0.5,0.5]	
出力関数のゲイン ϵ	出力層	1	
	中間層	2	1
学習係数 η	出力層	0.00001	1
	中間層 FW 部	0.001	1
初期重み値の値域	中間層 FB 部	[-20,20]	—
	その他	[-1,1]	

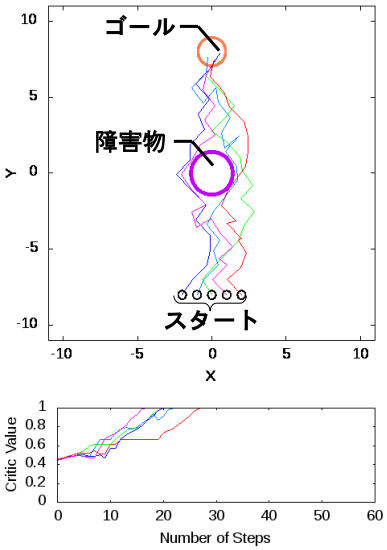
をそれぞれのネットワークに送り、Actor の二つの出力の分だけ x 方向、y 方向にエージェントが移動する。シミュレーションに用いたパラメータを Table 1 に示す。

100,000 試行後と 1,000,000 試行後のそれぞれの場合について、障害物を (0,0) の位置に固定し、エージェントのスタート位置を x=-2,-1,0,1,2、y=-8 とした 5 つの場合のエージェントがゴールするまでの軌道と、その際の critic（状態評価）値の変化を Fig.2 に示す。(b) の軌道の方が (a) と比べて滑らかに動き、短いステップでゴールできていることが確認できる。しかしながら、スタート位置が (2,-8) の軌道（赤線）は障害物に衝突し、8 ステップ動くことができなかった。そのため、(b) の他の軌道と比べてゴールするまでのステップ数が多くなっている。

次に状態評価値がどう学習されているか確認した結果を Fig.3 に示す。先程と同じように障害物を固定して、フィールド内にエージェントを設置したときの critic の出力を表している。いずれの場合もエージェントがゴー



(a) after 100,000 episodes

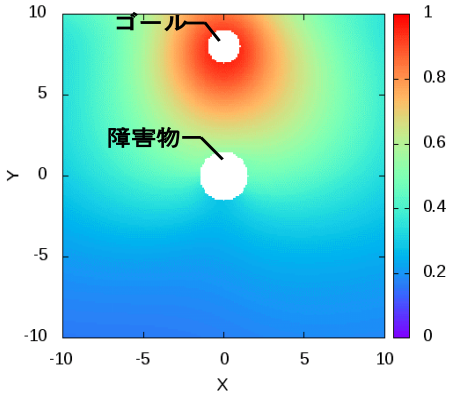


(b) after 1,000,000 episodes

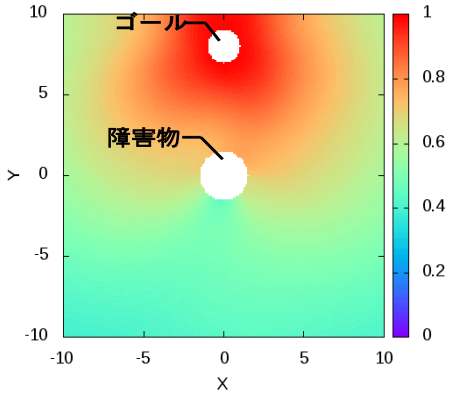
Fig. 2　Sample trajectories of the agent and change in the critic (state value) along the trajectories

ルに近いほど状態評価値は高くなっており、障害物によって直線的にゴールに向かうことができない (0,-2) 付近では低くなっている。この結果はエージェントがゴールに近いほど評価値はよく、障害物に邪魔されてしまう範囲は評価値はよくないことを学習できていることを表している。さらにより多く学習している (b) の方が (a) と比べて短いステップでゴールできているので、広い範囲で評価値が高くなっている。

学習曲線を Fig.4 に示す。赤い線はそれぞれの試行でのエージェントがスタートの位置からゴールするまでのステップ数を表し、青い線は 100 試行毎のその平均値を表している。エージェントが障害物を避けてゴールすることを学習するため、ゴールするまでのステップ数が減っ



(a) after 100,000 episodes



(b) after 1,000,000 episodes

Fig. 3　Distribution of critic (state value) output as a function of the agent location

ている。しかし 200,000 試行以降は平均値（青線）は減っているものの、各試行毎のステップ数（赤線）は増えているように見える。これは Fig.2(b) の赤色の軌道のようにエージェントが障害物に衝突し、しばらくの間動くことができなくなり、ステップ数が多くなることがしばしば現れてしまうためである。

カオス性を調べる指標として、微小な摂動の時間的広がりを表すリアプノフ指数を 1000 試行ごとに計算した。リアプノフ指数が正であれば、このダイナミクスはカオス性を持つことを表す。エージェントの位置を x=-9,-7,$\cdots$,9、y=-2,-8、障害物の位置を x=-9,-7,$\cdots$,9、y=0,5 の合計 400 パターンのそれぞれについて、カオスニューラルネットの中間層ニューロンの内部状態に 0.001 の大きさに正規化された乱数を加えて 1 ステップだけ動かす。そして、乱数を加えなかった場合と加えた場合の中間層ニューロンの内部状態の差 d_{after} を用いて次式よりリアプノフ指数 λ を計算する。

$$\lambda = \frac{1}{400}\sum_{p=1}^{400} ln\frac{d_{after}^{(p)}}{d_{before}^{(p)}} = \frac{1}{400}\sum_{p=1}^{400} ln\frac{d_{after}^{(p)}}{10^{-3}}. \qquad (6)$$

学習過程でのリアプノフ指数の変化を Fig.5 に示す。

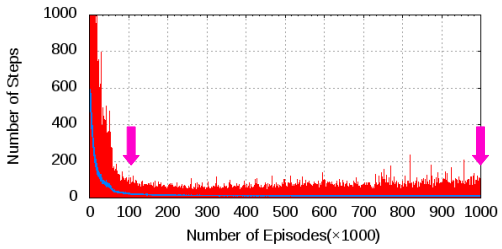


Fig. 4　Learning curve: change in the number of steps to the goal (red trace: steps at every episode, blue trace: average steps for every 100 episodes, pink arrows: the detail performances are shown in Fig.2,3,6)

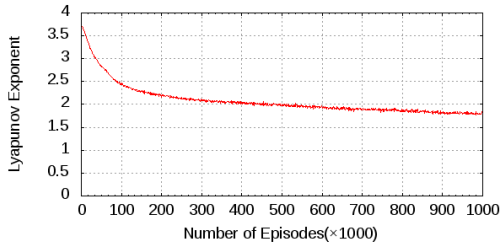


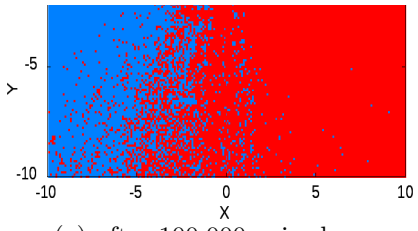
Fig. 5　the Lyapunov exponent during learning

100,000 試行より前ではリアプノフ指数は急速に減少し、以降ではゆっくりと減少しているが、常に正の値をとっている。100,000 試行後ではカオス性が強いので、Fig.2(a) のような軌道をとり、1,000,000 試行後ではカオス性を持ちつつも弱いため、(b) のように滑らかな軌道をとる。

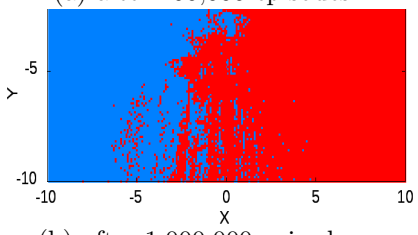
高次探索の創発について検討するために、エージェントのスタートの位置を $y < -2$ の範囲全ての場合でエージェントがゴールに到達するまでに障害物の右を通るか左を通るかの分布を Fig.6 に示す。(a)、(b) 両方の場合において、障害物より右側では右に多く避け、たまに左に避けており、逆も同様であることが確認できる。しかし、(a) では右に避ける場所と左に避ける場所が混在している領域が広いが、(b) では障害物の正面付近に狭まっている。ここでは、障害物に衝突しているため「高次探索」が創発したとは言えないが、学習を反映して不規則な行動が減り、障害物正面では確率的に近い行動をとっていることから、高次探索創発の可能性を示唆していると考えている。

4　結論

カオスニューラルネットを用いた強化学習により、障害物回避動作を学習した。エージェントが障害物の手前にいる場合、学習を反映して、障害物より右では障害物の右を、障害物より左では左を通ることが多くなった。また、障害物正面付近では完全にトラップされることなく、障害物の右を通るか左を通るかも確率的に選ばれているように見え、高次機能の創発の可能性を示唆する結果となった。しかし、学習が進むと障害物にしばらくトラップされた。今後、障害物に衝突することなく確率的に見える行動を維持できる学習方法を探っていきたい。



(a) after 100,000 episodes



(b) after 1,000,000 episodes

Fig. 6　Distribution of the agent initial location from which the agent passed the right or left side of the obstacle to reach the goal (blue: left side, red: right side)

謝辞

本論文は JSPS 科研費 (15K00360) の補助を受けた。

参考文献

1) K. Shibata: Emergence of Intelligence through Reinforcement Learning with a Neural Network, A. Mellouk (Ed.), “Advances in Reinforcement Learning” InTech, pp.99-120 (2011)

2) A. Krizhevsky et al.: ImageNet Classification with Deep Convolutional Neural Networks, in Adv. in NIPS 25, pp. 1097-1105 (2012)

3) V.Mnih, et al.: Playing Atari with Deep Reinforcement Learning, NIPS Deep Learning Workshop 2013 (2013)

4) K. Shibata and H. Utsunomiya: Discovery of Pattern Meaning from Delayed Rewards by Reinforcement Learning with a Recurrent Neural Network, Proc. of IJCNN 2011, pp. 1445-1452 (2011)

5) K. Shibata and K. Goto: Emergence of Flexible Prediction-Based Discrete Decision Making and Continuous Motion Generation through Actor-Q-Learning, Proc. of ICDL-Epirob 2013, ID 15 (2013)

6) Y. Sawatsubashi, et al.: Emergence of Discrete and Abstract State Representation in Continuous Input Task, Robot Intelligence Technology and Applications 2012, pp. 13-22 (2012)

7) K. Shibata and Y. Sakashita: Reinforcement Learning with Internal-Dynamics-based Exploration Using a Chaotic Neural Network, Proc. of IJCNN 2015, #15231 (2015)

8) Y. Goto and K. Shibata: Emergence of Higher Exploration in Reinforcement Learning using a Chaotic Neural Network, Proc. of ICONIP, (to appear)