カオスベース強化学習への誤差逆伝播法の適用

大分大学　○佐藤　克樹，後藤　祐樹，柴田　克成

Introduction of Error Backpropagation to Chaos-based Reinforcement Learning

Oita University　○ Katsuki SATO, Yuki GOTO, Katsunari SHIBATA

Abstract :　Aiming for the emergence of “thinking”, we have proposed new reinforcement learning (RL) using a chaotic neural network where exploration is generated inside the network together with the actions. In this paper, expecting to improve the performance, error backpropagation (BP) is employed in the proposed RL instead of the previous method using causality traces. The training signal for each actor output is generated using the product of the output itself and TD-error. It is shown that the network could learn a simple goal-reaching task by our new chaos-based RL using BP. For small recurrent weight size, recovery from failure episodes was faster than the case of using causality traces.

1.　序論

汎用人工知能 (Artificial General Intelligence: AGI) の実現を目指すために、我々の研究室ではセンサ入力からモータ出力までをニューラルネット (NN) で構成し、強化学習を行うことで様々な機能を創発する End-to-End 強化学習アプローチを提唱し [1]、この枠組みで DeepMind が TV ゲームの学習などで人間を超えるような結果を出している [2]。さらに我々の研究室では時系列データを扱えるリカレントニューラルネット (RNN) を用いて強化学習を行うことで、記憶や予測の機能が創発することを確認した [3][4]。しかしながら、典型的な高次機能である思考と呼べるような機能の創発は未だ確認できていない。

私たち人間は外部からの入力がなくとも頭の中で考えることができることから、論理的思考は自律的に状態遷移する一種の内部ダイナミクスであると考えられる。また強化学習に必要な探索はランダムに近い行動選択ではあるが、自律的な状態遷移が生み出すという点において思考に類似している。この論理的思考と探索の類似性から、我々の研究室では「ランダムに近い探索が学習を通して思考へと成長する」という仮説を立てた。そして、この仮説の実現に向け、カオスニューラルネット (ChNN) を用いることで、従来の強化学習の枠組みを大きく変えた新しい強化学習法を提案してきた。ここでは、エージェントは外部からの乱数付与による確率的選択なしに、内部のカオスダイナミクスによって探索を行う。そして、思考とは程遠いものの、この新しい強化学習でまず、簡単なゴール到達タスクや障害物回避タスクの学習ができることを確認した [5][6]。

この強化学習では探索は NN 内部で生成されるため、従来の強化学習のように探索と行動を分離することができない。したがって、従来と同様な方法では教師信号を生成できず、強力な学習方法である誤差逆伝播法 (BP 法) を適用することができなかった。そのため、我々の研究室では因果トレースという一種の短期記憶を用いて NN を学習させてきた。

しかし、因果トレースを用いる方法では内部ダイナミクスを司るリカレント部の重み値の学習が上手くできていない。さらに、近年の人工知能分野の発展には強力な学習方法である BP 法の存在が大きく、ChNN の学習にも適用できれば、探索から思考へ成長させる『学習』の部分のパフォーマンスを向上することができ、AGI 実現へ近づくと期待している。

本論文では、探索成分と出力を分離できない ChNN に BP 法を適用するために、教師信号の生成に探索成分ではなくモータ出力をそのまま利用し、我々が提案しているカオスベースの新しい強化学習を行う。リカレント部の学習は今後の課題として、まずは、簡単なゴール到達タスクの学習が可能かどうかを確認する。そして、我々が今まで用いてきた因果トレースによる学習と学習性能を比較する。

2.　カオスニューラルネットを用いた強化学習

強化学習を行うことで、エージェントは報酬や罰により動作を自律的に学習することができる。より良い動作を学習するためには探索 (試行錯誤) を行う必要があり、一般的には外部から乱数を与えて確率的な行動を行う。当研究室で提案してきた新しい強化学習では確率的な行動選択は行わず、ニューラルネット内部のカオスダイナミクスに基づいて探索を行う。連続動作の学習を可能にする Actor-Critic において、図 1 のように動作生成部である Actor 部をカオスニューラルネットで構成し、状態評価部である Critic 部はカオスの発生しない通常の階層型ニューラル

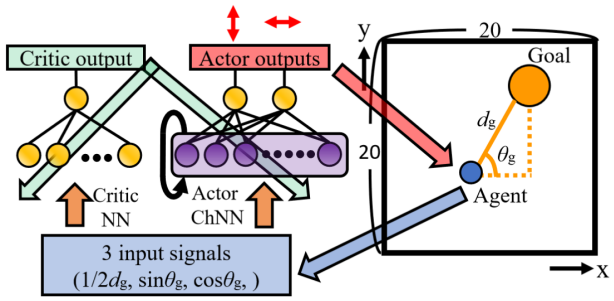


Fig. 1: カオスニューラルネット (ChNN) を用いた強化学習システムとゴール到達タスク

ネットを用いる。本論文で用いるニューラルネットの式を示す。各ニューロンの内部状態 u は、Actor 部の ChNN の中間層では (1) 式、それ以外は (2) 式で計算し、(3) 式によって出力 o を求める。

$$u_{j,t}^{A,h} = \sum_{i=1}^{N^{in}} w_{j,i}^{A,h} o_{i,t}^{A,in} + \sum_{i=1}^{N^{h}} w_{j,i}^{A,REC} o_{i,t-\Delta t}^{A,h} \quad (1)$$

$$u_{j,t}^{n,l} = \sum_{i=1}^{N^{(l-1)}} w_{j,i}^{n,l} \cdot o_{i,t}^{n,l-1} \quad (2)$$

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

ここで、$n = A$ or C であり、A は Actor 部、C は Critic 部を表している。N^l は l 層のニューロン数、$h(=1)$ と $in(=0)$ はそれぞれ中間層と入力層を表す。$u_{j,t}^{n,l}$、$o_{j,t}^{n,l}$ はそれぞれ時刻 t での l 層 j 番目ニューロンの内部状態と出力を表す。$w_{j,i}^{n,l}$ は l 層 j 番目ニューロンの $l-1$ 層の i 番目ニューロンからの結合重み値であり、REC はリカレント結合を意味する。全ての重み値は一様乱数によって決定される。g はシグモイド関数のゲインであり Actor 部の中間層のみ $g = 2$ その他は $g = 1$ とし、Δt はステップ幅であり $\Delta t = 1$ としている。TD 誤差 $\hat{r}_t$ は学習に使用し、(4) 式で計算する。

$$\hat{r}_t = r_{t+\Delta t} + \gamma \cdot V_{t+\Delta t} - V_t \quad (4)$$

ここで $r_{t+\Delta t}$ は時刻 $t+\Delta t$ で与えられる報酬であり、ここでは $r_{t+\Delta t} = 0.4$ とした。γ は割引率であり ここでは 0.96 とした。$V_{t+\Delta t} = O_{t+\Delta t}^{C,L}$ は Critic 部の出力であり, $L(=2)$ は出力層を表す。従来の Actor-Critic における Actor 部の教師信号 T_A には確率的行動選択として出力に付加する乱数と TD 誤差を掛けたものが用いられる。しかしこの手法では、Actor 部の出力層に外部から乱数を付与しないので、代わりに Actor 出力自身と TD 誤差を掛けたものを使用し、以下のように計算する。

$$T_A = \boldsymbol{A(S_t)} \cdot \hat{r}_t + \boldsymbol{A(S_t)}. \quad (5)$$

Critic 部の教師信号 T_{V_t} は、従来と同様な手法で求める。

$$T_{V_t} = r_{t+\Delta t} + \gamma \cdot V_{t+\Delta t}. \quad (6)$$

それぞれの教師信号を用いて Actor 部、Critic 部のネットワークを通常の誤差逆伝播法で 1 回だけ学習させる。

一方、比較対象である従来の因果トレース [5] を用いる方法では、まず、各ニューロンの各結合部に配置した因果トレース $c_{j,i,t}^{l}$ を、ニューロンの出力の増加に寄与した過去の入力 $o_{i,t}^{l-1}$ を保持するようにニューロンの出力の変化 $\Delta o_{j,t}^{l} = o_{j,t}^{l} - o_{j,t-\Delta t}^{l}$ を用いて以下の式で計算する。

$$c_{j,i,t}^{l} = (1 - |\Delta o_{j,t}^{l}|) c_{j,i,t-\Delta t}^{l} + \Delta o_{j,t}^{l} o_{i,t}^{l-1} \quad (7)$$

そして、それと TD 誤差から以下のように重み値を更新する。

$$\Delta w_{j,i,t}^{l} = \eta \cdot c_{j,i,t}^{l} \cdot \hat{r}_t \quad (8)$$

また、BP 法、因果トレース法いずれの場合も今回はリカレント結合の重み値 $w_{j,i}^{A,REC}$ は学習させない。

3. シミュレーション

Fig.1 のように中心座標を (0,0) とした 20×20 のフィールド内にゴールを設置したゴール到達タスクを行った。シミュレーションに用いたパラメータを Table 1 に示す。

Table 1: シミュレーションに用いたパラメータ

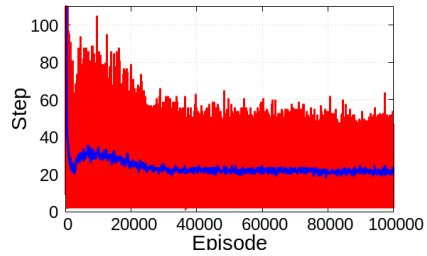
名前		Actor	Critic
中間層ニューロンの数		100	10
学習係数 η	出力	1.0	
	出力	1.0	
	中間 (REC)	0	-
重み値の値域	出力	[-1,1]	
	中間 (FW)	[-1,1]	
	中間 (REC)	[-2,2]	-

Fig. 1 に示した 3 個の入力情報 (エージェントとゴールとの相対距離、相対角度) をそれぞれ最大値が 1 となるように正規化してネットワークの入力とする。2 つの Actor 出力はそれぞれ x 方向、y 方向の移動量を表し、移動可能範囲が半径 0.5 の円になるように移動方向は変えずに大きさを調整した値にしたがってエージェントが移動する。

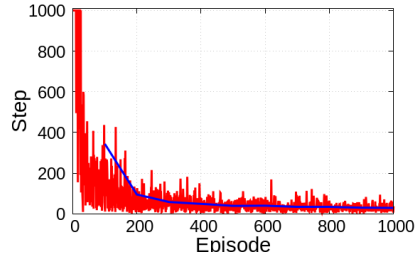
半径 1.0 のゴールと半径 0.5 のエージェントの初期位置は毎試行ランダムに決定する。エージェントはカオスニューラルネットの Actor 出力で動き、エージェントの中心がゴール内に入ると 0.4 の報酬を獲得する。エージェントがゴールに到達する、またはステップ上限である 1,000 ステップ経過するまでを 1 試行とし 100,000 試行の学習を行った。

Fig. 2 に学習曲線として、(a) 縦軸を拡大した学習全体の様子と (b) 学習初期の様子を見るために横軸を拡大したものを示す。学習曲線の赤い線は各試行のエージェントがスタートからゴールするまでのステップ数を表し、青い線は 100 試行毎の平均値である。また、(c) に学習終了 (100,000 試行) 後にゴールの位置を (0, 0) の位置に固定し、エージェントのスタート位置をずらして設置した 8 パターン分の軌道と (d) 学習後の Critic(状態評価) 値の分布を Fig.2 に示す。

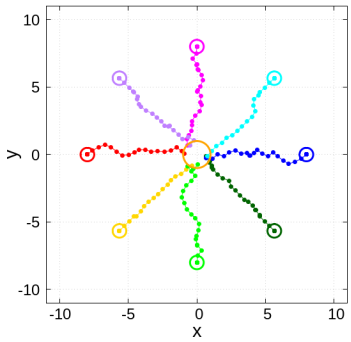
試行回数が増えるとゴールまでのステップ数 2000 試行あたりで一旦上昇傾向にあるものの全体的に下がっている。学習後のエージェントの軌道はゴールに向かっている様子がわかり、critic 値は中心のゴールに近いほど高い値であることがわかる。上記のことから、カオスベースの新しい強化学習に対し、Actor 出力と TD 誤差に基づいて生成した教師信号を用いた BP 法を適用することで学習ができることを確認した。



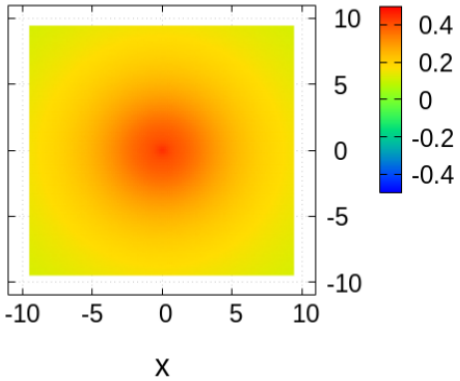
(a) 学習曲線 (縦軸を 100 ステップまで拡大)



(b) 学習曲線 (横軸を 1000 試行まで拡大)



(c) 学習後の軌道 (8 パターン)



(d) 学習後の Critic 分布

Fig. 2: 誤差逆伝播を適用した学習の結果
(リカレント結合重み値の範囲 [-2,2])

BP 法と因果トレースのそれぞれを用いた場合に、リカレント結合の重み値に用いる乱数の範囲 $[-W_{MAX}^{REC}, W_{MAX}^{REC}]$ を変化させながら、乱数系列 20 系列に対する成功回数と平均ステップ数をプロットしたものを Fig. 3 と 4 に示す。リカレント結合重み値の大きさ W_{MAX}^{REC} を横軸とし、右縦軸は 20 系列中の成功回数 (ゴールできれば成功とした)、左縦軸は平均ステップ数 (ただし、学習成功時のみ) を示した。ただし、因果トレース法は学習を安定させるため、中間層の学習係数を 0.01 とした。

Fig. 3(BP 法) では、どのリカレント結合重み値 W_{MAX}^{REC} の場合でも学習成功回数が高いことがわかる。また、W_{MAX}^{REC} が小さくなればなるほど平均ステップ数も下がっていることがわかる。一方、Fig. 4(因果トレース) では、W_{MAX}^{REC} が小さくなっていくと学習成功回数が急激に低くなっていることがわかる。また、学習ができているリカレント結合重み値の場合では平均ステップ数に大きな違いは見られない。Fig. 3 と Fig. 4 を比較すると、BP 法で重み値が小さい時 ($W_{MAX}^{REC} \leq 0.5$) は平均ステップ数が一番少ないことがわかる。一方その時、因果トレースでは学習できない場合がかなり多くなっていることがわかる。

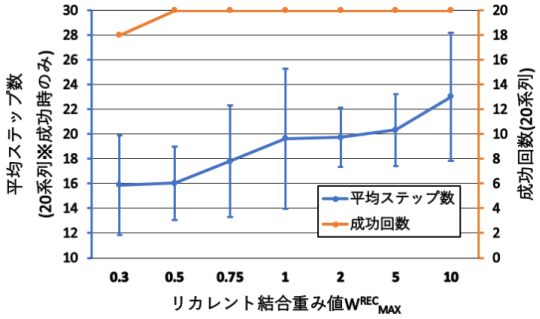


Fig. 3: 平均ステップ数と学習成功率 (BP 法)

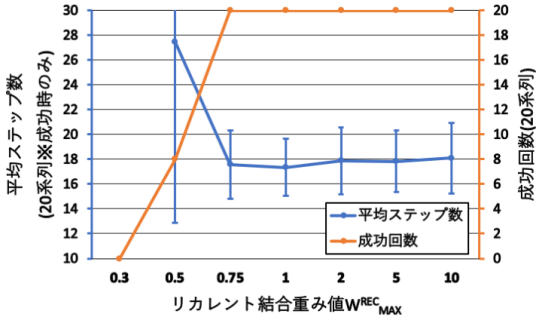


Fig. 4: 平均ステップ数と学習成功率 (因果トレース法)

BP 法と因果トレース法を比べて、重み値が小さい時に学習性能に差が出る原因を探るために、探索の様子とリアプノフ指数を見てみる。カオスベース強化学習は探索を内部のカオスダイナミクスに大きく依存しており、リアプノフ指数を見ることでカオス性の強弱を知ることができる。リアプノフ指数が正であればカオス性があることを示す。ここでは入力を 0、初期内部状態を 0.001 の一様乱数にし、中間層同士のリカレント結合のみで 50 ステップネットワークを回して内部状態を決定する。その後に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更に $T = 50$ ステップ回し、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を用いてリアプノフ指数 λ を求めた。リカレント結合重み値の大きさ W_{MAX}^{REC} を変えながら、乱数系列 $P = 20$ 個を用いて

以下のように計算し、平均と標準偏差を求めたものを Fig. 5 に示す。

$$\lambda = \frac{1}{50 \cdot 20} \sum_{p=1}^{20} \sum_{t=1}^{50} \ln \frac{d_{p,t+\Delta t}}{d_{p,t}} \tag{9}$$

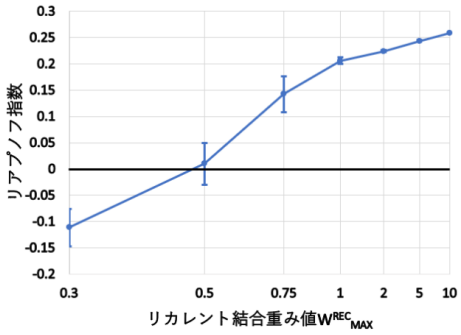


Fig. 5: 中間層 REC 部の重み値によるリアプノフ指数

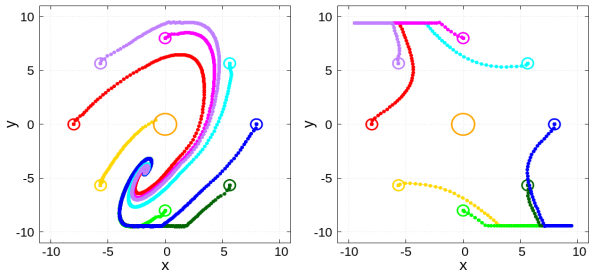


Fig. 6: (REC : 0.3) 学習前の軌道 (2 系列 16 パターン)

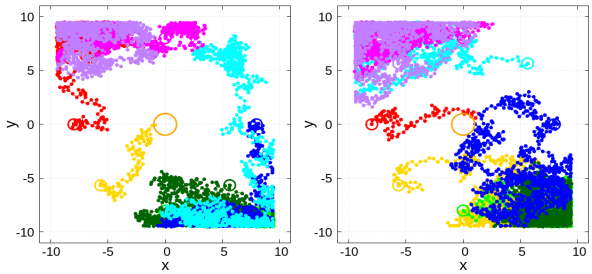


Fig. 7: (REC : 2.0) 学習前の軌道 (2 系列 16 パターン)

リカレント結合重み値が小さくなればなるほど、リアプノフ指数が下がっていることがわかる。また、リアプノフ指数が小さい時 (w^{REC}_{MAX} : 0.3) と大きい時 (w^{REC}_{MAX} : 2.0) の学習初期の軌道を見てみると、リアプノフ指数が小さい時の軌道は Fig .6 のように探索をほとんど行えておらず、リアプノフ指数が大きい時は Fig. 7 のようにある程度探索していることがわかる。BP 法と因果トレースいずれの場合もリアプノフ指数が小さい時は、ほとんど探索しないながらも偶然ゴールすることをきっかけに一旦ゴールに向かうようになるが、しばしばステップ上限の 1,000 ステップでゴールできないことがある。BP 法ではその後すぐに再びゴールするようになるが、因果トレースではなかなか再びゴールに向かうようにならない。一方、リアプノフ指数が大きい時は、学習が進みゴールに向かう行動を取り始めた時に安定的に学習することができた。

4. 結論

カオスベース強化学習に出力と TD 誤差を掛けたものを教師信号として誤差逆伝播法を適用して簡単なゴール到達タスクを学習させることができた。リカレント結合重み値を下げるとステップ数は少なくなり、因果トレースを用いた時のどの場合よりも少なかった。しかし、両手法ともにリアプノフ指数が小さい時はゴールに向かう行動を取り始めたにも関わらず、ステップ上限である 1,000 ステップ以内にゴールできないことが見られた。その時、BP 法では再びすぐにゴールできるようになったが、因果トレースではなかなか再びゴールに向かうようにならないという差があった。また、リアプノフ指数が高い時はそれらが見られず安定的な学習ができた。

今後の課題として、BP 法と因果トレースでの違いを解析するとともに、今回用いたゴール到達タスクよりも難しい問題でカオスベース強化学習における BP 法の有効性を見極める必要がある。また、記憶が必要なタスク等でリカレント結合の重み値を学習させる。その際、学習によるカオス性の低下や、BPTT(Backpropagation through time) での誤差発散の影響を見ていく。

謝辞

本論文は JSPS 科研費 (15K00360) の補助を受けた。

参考文献

[1] 柴田克成 : 深層学習が示唆する end-to-end 強化学習に基づく機能創発アプローチの重要性と思考の創発に向けたカオスニューラルネットを用いた新しい強化学習, 認知科学, Vol.24, No1, pp.96-117 (2017)

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

[3] 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)

[4] 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.ID 15 (2013)

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

[6] Y. Goto and K. Shibata : Influence of the Chaotic Property on Reinforcement Learning Using a Chaotic Neural Network, Proc. of ICONIP 2017, pp. 759-767, (2017)