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

Yuki Goto and Katsunari Shibata

Dept. of Electrical and Electronic Engineering, Oita University,
700 Dannoharu, Oita 870-1192, Japan
iwishdayss@gmail.com,shibata@oita-u.ac.jp

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.

Keywords: reinforcement learning, chaotic neural network, higher exploration, emergence of intelligence, obstacle avoidance

1 Introduction

Our group has pointed out the difficulty of developing a program by hand for such massively parallel and highly flexible computation that our brain is doing, and proposed the approach that a Neural Network (NN) is responsible for the whole process from sensors to motors and various functions emerge in the NN through Reinforcement Learning (RL)[1][2]. Recent excellent performance of "Deep Learning" especially in the area of recognition[3] and the surprising result in the TV games by combining it with RL[4] are thanks to its emergence ability of useful internal representations, and support the significance of our approach.

Because higher functions such as "memory", "prediction", "logical thinking" and so on, need to cope with dynamics, a Recurrent Neural Network (RNN) is used on behalf of a layered NN. The emergence of "memory" or "prediction" has been confirmed in a simple task[5][6]. However, the learning of a task requiring multiple state transitions is not easy[7], and the emergence of what we can call "logical thinking" has not been shown yet.

Therefore, we have felt the need of another approach in which desired dynamics is not obtained from scratch in a non-chaotic "silent" NN, but is reformed

from chaotic dynamics through learning in a chaotic NN. We have also thought that "exploration" should be considered as a function based on internal dynamics as well as "memory" or "prediction", and random-like "exploration" is expected to grow up in "logical thinking" through learning. According to the hypothesis, we have proposed a completely novel RL where exploration is performed based on the internal chaotic dynamics without adding external random numbers[8].

On the other hand, recently, the ability of reservoir computing has been unveiled, and it was surprising that complicated dynamic patterns are easily learned using a chaotic NN by FORCE Learning[9]. In addition, it was shown that by adding exploration noises from the outside, a chaotic NN can learn complicated dynamic patterns based on a reward signal without giving any target signal directly[10]. From the above ability of chaotic NNs, RL using a chaotic NN is expected to develop greatly hereafter.

Authors thought that in the process of growing from "exploration" to "logical thinking", the level of exploration changes from "lower", which is motor-level, to "higher", which is more abstract level. For example in a forked road, we don't move our each muscle randomly, but choose whether to go the right way or left way in more abstract action space. That is because we have already learned that though we go on a non-road area, we cannot get a good result.

Therefore in this paper, aiming to show the possibility of emergence of the higher exploration, we replace the forked road situation with an obstacle avoidance task in which an agent learns to reach a goal while avoiding an obstacle, and whether the agent passes the right side or the left side of the obstacle is focused on. The task refers to [11], in which an agent learned appropriate actions based on regular RL using a layered NN, but there was a place where the agent could not move before the obstacle when no random number for exploration is added.

2 Reinforcement Learning using a Chaotic Neural Network

Reinforcement learning is autonomous and purposive learning of appropriate actions to get more reward and less punishment. Generally, an agent explores stochastically based on random numbers. However here, as mentioned in Introduction, an agent explores by chaotic dynamics that a chaotic NN produces without adding noises or random numbers. In this paper, for continuous input-output mappings, Actor-Critic is used as a RL architecture. A chaotic NN and a non-chaotic layered NN are used for actor and critic respectively as shown in Fig.1, and the sensor signals are the input for both NNs. Here, the neuron model used in both NNs is static that is different from [9] or [10] as

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

where $u_{j,t}^{(l)}$ and $o_{j,t}^{(l)}$ are the internal state and the output of the j-th neuron in the l-th layer at time t, $w_{j,i}^{(l)}$ is the synaptic weight from the i-th neuron in

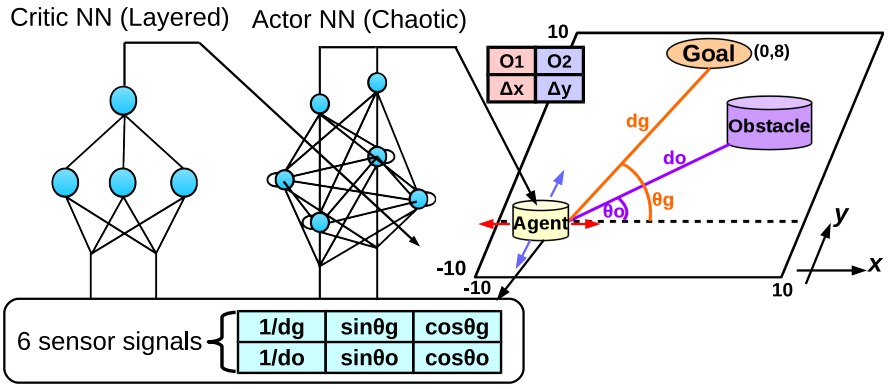


Fig. 1. Reinforcement learning system and the obstacle avoidance task in this paper

the (l-1)-th layer to the j-th neuron in the l-th layer. The second term in the right-hand side is only for the hidden layer in the chaotic NN, and $w_{j,i}^{\mathrm{FB}}$ is the weight for the recurrent connection from the i-th neuron in the hidden layer. The activation function is the sigmoid (tanh) function $f(\ )$ whose value ranges from -0.5 to 0.5, and the output is $o_{j,t}^{(l)} = f(u_{j,t}^{(l)})$. The chaotic NN has two actor outputs $\mathbf{A}(\mathbf{S}_t)$ that are used as motion signals, and the non-chaotic NN has a critic output $V(\mathbf{S}_t)$ where $\mathbf{S}_t$ is the sensor inputs at time t.

For learning, TD-error $\hat{r}_t$ is represented as

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

where r_{t+1} is the reward given at time t+1, γ is a discount factor. T_{V_t} is the target for the critic output at time t and is computed as

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

The critic NN is trained once according to Error Back Propagation using this target signal. To adjust the value range, 0.5 is added to the output of the critic NN and 0.5 is subtracted from the target T_{V_t} before using them actually.

In the chaotic NN in this paper, chaotic dynamics is produced by strong feedback connections between hidden neurons, and there is no feedback connections from the output. For the input-hidden and hidden-output connection (synaptic) weight $w_{j,i}^{(l)}$ in the chaotic NN, is modified using the trace $c_{j,i,t}^{(l)}$ as

$$\Delta w_{j,i,t}^{(l)} = \eta_A^{(l)} \hat{r}_t c_{j,i,t}^{(l)} \tag{4}$$

where $\Delta w_{j,i,t}^{(l)}$ is the modification of the weight $w_{j,i}^{(l)}$ at time t, $\eta_A^{(l)}$ is a learning rate for the l-th layer of the actor chaotic NN. The trace $c_{j,i,t}^{(l)}$ holds the past contribution of the pre-synaptic signal to the output increase in the post-synaptic neuron, and at each time step, it takes in the pre-synaptic signal $o_{i,t}^{(l-1)}$ and

Table 1. The parameters used in the simulation

Name		Actor Net	Critic Net
Step Limit in Each Episode		1,000	
Number of Layers		3	
Number of Inputs		6	
Number of Hidden Neurons		100	10
Number of Outputs		2	1
Value Range of Sigmoid Function		-0.5 - 0.5	
Gain of Sigmoid Function	Output	1	
	Hidden	2	1
Learning Rate η	Output <- Hidden	0.00001	1
	Hidden <- Input	0.001	1
	Hidden <- Hidden	0.0	—
Range of Initial Weights (uniformly random)	Hidden <- Hidden (feedback)	± 20	—
	others	± 1	
Discount Factor γ		—	0.95

forgets the past trace value according to the change in the post-synaptic neuron $\Delta o_{j,t}^{(l)} = o_{j,t}^{(l)} - o_{j,t-1}^{(l)}$ as

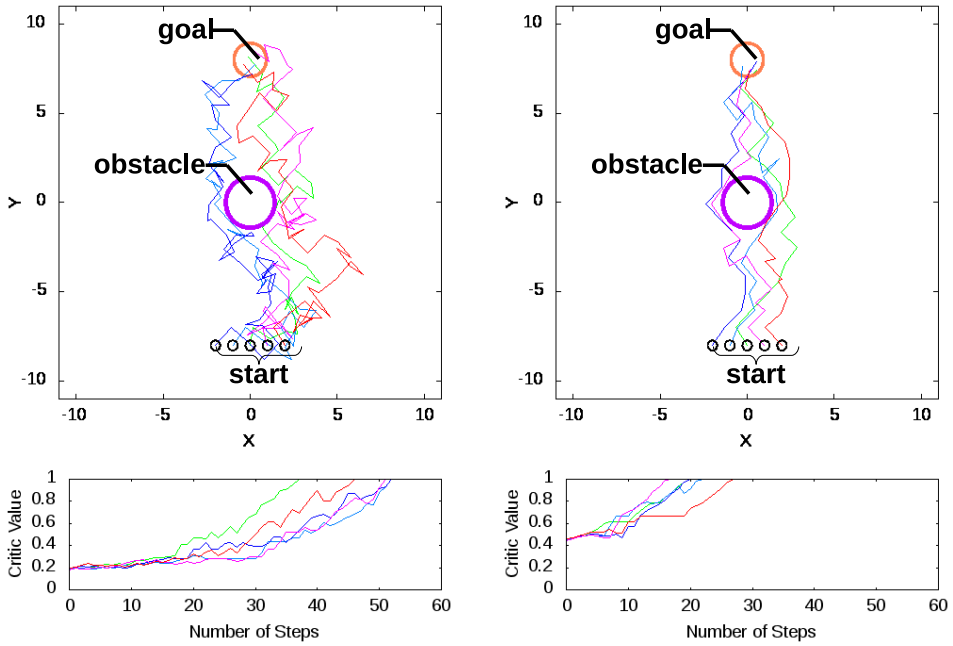
$$c_{j,i,t}^{(l)} = (1 - |\Delta o_{j,t}^{(l)}|) \cdot c_{j,i,t-1}^{(l)} + \Delta o_{j,t}^{(l)} \cdot o_{i,t}^{(l-1)}. \tag{5}$$

The feedback connection weights $w_{j,i}^{\mathrm{FB}}$ are not modified here.

3 Simulation

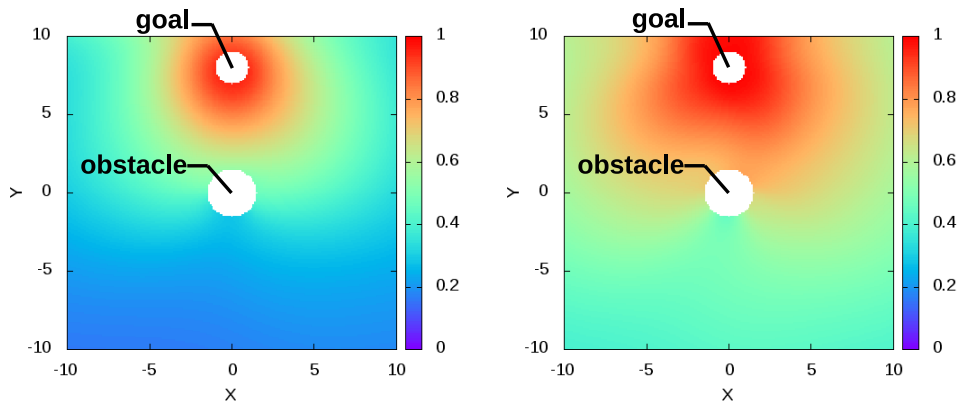
In this paper, to examine the acquisition of higher exploration, an obstacle avoidance task is simulated referring to the task in [11]. In this simulation, as shown in Fig.1, there is a 20×20 field, and a goal is fixed at the upper center area $(0, 8)$. An obstacle and an agent are located randomly at the beginning of every episode. The agent moves according to the outputs of the actor chaotic NN, and when it reaches the circle with a radius of 1.0 around the goal, 1.0 is given as a reward. When it reaches the circle with a radius of 1.5 around the obstacle or it collides with a wall at the boundary of the field, -0.01 is given as a punishment. The episode is terminated when the agent either reaches the goal or fails to do so in 1,000 steps. 6 sensor signals as shown in Fig.1 are sent to the both networks as input. Each of the two actor outputs decides the one-step move in x or y direction. The parameters used in the simulation are shown in Table 1.

At first, critic (state value) and actions when the obstacle is put at $(0, 0)$ are observed in the two cases after 100,000 episodes (a) and after 1,000,000 episodes (b) of learning. The agent was located at $x = -2, -1, 0, 1, 2$, $y = -8$ and the trajectories and the change in the critic values along the trajectories are shown in Fig.2. It can be seen that after 1,000,000 episodes (b), the trajectories are smoother and the agent reaches the goal in smaller steps than in the case after 100,000 episodes (a). However after 1,000,000 episodes (b) when the agent starts



(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



(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

from $(2, -8)$ (red trajectory), the agent collided with the obstacle and could not move for 8 steps. Therefore, the number of steps to the goal when the agent moved along the red trajectory is larger than the others.

Fig.3 shows the distribution of the critic output as a function of the agent location when the obstacle is put as the above. In both cases, the critic value is larger as the agent location is closer to the goal and lower around $(0, -2)$ where the obstacle disturbs the agent to go to the goal. This result shows that the agent learned that when the agent is close to the goal, the state is good, and when the obstacle exists around the line segment from the agent to the goal, the state is not good. In (b) after 1,000,000 episodes, the critic value is higher in total than (a) after 100,000 episodes, and that shows the agent can reach the goal in smaller number of steps in the case of (b).

The learning curve is shown in Fig.4. The red trace shows the number of steps from the initial location of the agent to the goal for each episode, and the blue trace shows the average number of steps over every 100 episodes. Since the agent learns how to go to the goal and avoid the obstacle, the number of steps is decreased. However, after 200,000 episodes, although the average number of steps (blue trace) still continues to decrease, the number of steps looks to increase. This mean that, when the agent collided with the obstacle, it was sometimes trapped at the place for a while such as the red trajectory in Fig.2(b).

In this paper, as an index of chaotic property, Lyapunov exponent, which shows the sensitivity to small perturbations, is computed. When the Lyapunov exponent is positive, the dynamics is chaotic. Here, every 1,000 episodes, a random vector whose size is normalized to 0.001 is added to the internal state of the hidden neurons in the chaotic NN. After one-step action according to the actor outputs, the Euclidean distance d of the hidden states from the case when no perturbation is added was compared between before and after the action. The above is performed in 400 situations in which the agent's location varies as $x = -9, -7, \cdots, 9, y = -2, -8$ and the obstacle location varies as $x = -9, -7, \cdots, 9$, $y = 0, 5$, and the Lyapunov exponent λ is calculated by

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

The change in Lyapunov exponent according to the learning progress is shown in Fig.5. The Lyapunov exponent is decreased quickly before the 100,000th episode and slowly after the 100,000th episode keeping the value positive. As shown in Fig.2, in the case of after 100,000 episodes (a), the influence of the chaotic dynamics looks large, but around the end of the learning (after 1,000,000 episodes (b)), it looks smaller though still some irregularities can be seen.

In order to discuss whether the "higher exploration" emerges or not, Fig.6 shows how the side of the obstacle through which the agent passed to avoid it varies depending on the initial location in the area $y < -2$ where the agent is located farther than the obstacle from the goal. In the both cases, after 100,000 episodes (a) and after 1,000,000 episodes (b), the agent is likely to pass through the right side of the obstacle when the initial location is in the right part of the

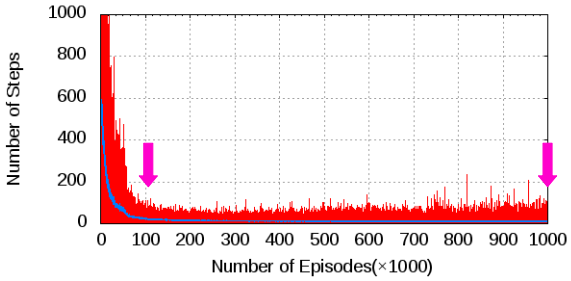


Fig. 4. Learning curve: change in the number of steps to the goal (red trace: steps at every episode, blue trace: average steps for every 100 episodes, pink arrows: the detail performances are shown in Fig.2,3,6)

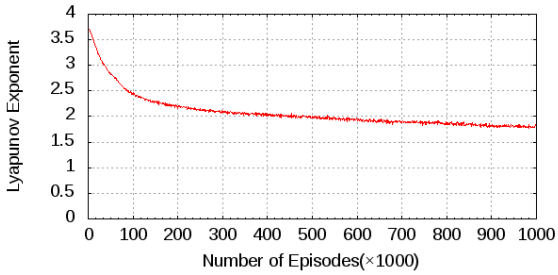
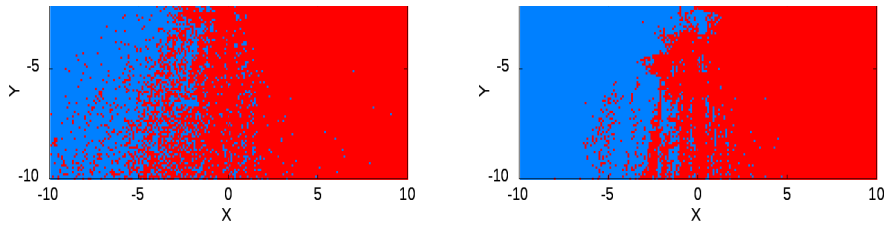


Fig. 5. Change in the Lyapunov exponent during learning

field, and vice versa. Around the boundary of the two areas, especially in (a), the side the agent passed varies frequently depending on the initial location and so the agent looks to choose the side randomly. In (b), the distribution of these two areas is more symmetrical and reasonable than in (a). The result also shows that the agent is not trapped completely in front of the obstacle even without adding any external random numbers to the actor output, and that is different from the result in[11]. It is thought that the possibility of the emergence of higher exploration in which learning is reflected could be shown although it is ideal not to collide with the obstacle.

4 Conclusion

It was shown that by RL using a chaotic NN, the agent learned to go to the goal while avoiding a randomly-located obstacle. The distribution of the agent initial location where the agent passed the right side or left side of the obstacle did not have a clear boundary and the agent looks to choose the side to pass randomly. There was no place where the agent could not move to the right or left to avoid the obstacle. These results suggest the emergence of higher exploration, which would appear on the way to the emergence of "thinking", we expect. In the latter half of learning, Lyapunov exponent was decreased, and the agent sometimes collided with the obstacle and was trapped at the place for a while.



(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 side or left side of the obstacle to reach the goal (blue: left side, red: right side)

Since the sensor inputs in this task are different from in [11], it is necessary to think about the solution of the problem from both sides of task setting and control of the chaotic property.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 15K00360.

References

1. K. Shibata and Y. Okabe: Reinforcement Learning When Visual Signals are Directly Given as Inputs, Proc. of ICNN '97, Vol. 3, pp.1716-1720 (1997)
2. K. Shibata: Emergence of Intelligence through Reinforcement Learning with a Neural Network, A. Mellouk (Ed.), "Advances in Reinforcement Learning" InTech, pp.99-120 (2011)
3. A. Krizhevsky et al.: ImageNet Classification with Deep Convolutional Neural Networks, in Adv. in NIPS 25, pp. 1097-1105 (2012)
4. V.Mnih, et al.: Playing Atari with Deep Reinforcement Learning, NIPS Deep Learning Workshop 2013 (2013)
5. K. Shibata and H. Utsunomiya: Discovery of Pattern Meaning from Delayed Rewards ... , Proc. of IJCNN 2011, pp. 1445-1452 (2011)
6. K. Shibata and K. Goto: Emergence of Flexible Prediction-Based Discrete Decision ... , Proc. of ICDL-Epirob 2013, ID 15 (2013)
7. Y. Sawatsubashi, et al.: Emergence of Discrete and Abstract State Representation ..., Robot Intelligence Technology and Applications 2012, pp. 13-22 (2012)
8. K. Shibata and Y. Sakashita: Reinforcement Learning with Internal-Dynamics-based Exploration Using a Chaotic Neural Network, Proc. of IJCNN 2015, #15231 (2015)
9. David C. Sussillo: Learning in Chaotic Recurrent Neural Networks, Ph.D.Thesis, Columbia University (2009)
10. Hoerzer GM et al.: Emergence of complex computational structures from chaotic neural networks through reward-modulated Hebbian learning, Cerebral Cortex, Vol.24, No.3, pp. 677-690 (2014)
11. K. Shibata, et al.: Direct-Vision-Based Reinforcement Learning in "Going a Target" Task ... , Proc. of NEURAP '98, pp.95-102 (1998)