Experimental Study of the Eligibility Traces in Complex Valued Reinforcement Learning

Takeshi Shibuya, Shingo Shimada and Tomoki Hamagami

Abstract— **Effectiveness of eligibility traces in complex valued reinforcement learning is studied. Complex valued reinforcement learning is a new method inspired by complex valued nerual networks. In this study, it is desired that various approaches in the ordinally real valued reinforcement learning are applied to the complex valued reinforcement learning. This paper focuses attention on an experimental study of the eligibility traces. Simulation results infer that there is a possibility of overcoming tight perceptual aliasing with long trace back up.**

I. INTRODUCTION

Reinforcement Learnings (RLs) [1][2], learning framework based on trial-and-error, have potential application to the problem of acquiring suitable behaviors for autonomous mobile robots. For example, in our previous studies about the intelligent wheelchair as an autonomous robot, the RL algorithms were applied to an agent learning how to avoid obstacles and to evolve cooperative behavior with other robots. [3][4]

An agent controlling the robots observes its state through sensors and effects its action with actuators. Available sensors and actuators depend on an application. In particular, applying to real world simple robots such as microrobots limits these resources tightly.

Equipping only poor sensors causes a serious problem called perceptual aliasing for the agent. Facing this problem, the agent can not distinguish multiple states and learn suitable behavior.

In order to resolve the problem, some representative approaches have been conducted [5][6][7][8]. The most direct approach is to use memory of the contexts called episodes to disambiguate current state and to keep trace of previous information [5]. Supposing the environment is stable and the agent has sufficient memory, this memory based approach brings out very strong performance. However, as many real world environment belongs to a dynamic class, the memory of experience has to be revised frequently. This revised algorithm often becomes complication and depends on the task.

We have developed Complex Valued Reinforcement Learnings (CVRLs)[9], new reinforcement learning algorithm using complex valued functions. A strong point using complex value to reinforcement learning is easily to extend state-acition function in time series. The algorithm has two practical advantages. First, the idea is easy to apply to conventional simple alrogithm such as Q-learning and profit sharing. Secondly, the idea does not memorize episodes of experience directly. We have confirmed the effectiveness of the algorithm in several simple tasks.

In previous work, eligibility trace is also introduced but veiled. This paper introduces two tasks to declare the effectiveness of the eligibility traces in CVRLs.

In the remainder of this paper, the basic idea of complex valued reinforcement is introduced and applied to the Q-learning. We also introduce eligibility traces for CVRLs. Then, simulation experiments are conducted for experimental study of effectiveness of eligibility traces.

Graduate School of Engineering, Yokohama National University
79-5 Tokiwadai, Hodogayaku, Yokohama JAPAN
t-shibuya@mail.dislab.ynu.ac.jp
s-shimada@mail.dislab.ynu.ac.jp
hamagami@ynu.ac.jp

II. COMPLEX VALUED REINFORCEMENT LEARNING

A. Theory

The idea of complex valued reinforcement learning has been inspired by the CVNNs. The complex valued neural networks (CVNNs)[10] has been defined as an extension of an usual, real valued neural network. The inputs, outputs, and parameters such as weights and thresholds are defined as all complex numbers, and the activation function is inevitably a complex valued function. The strong point of the network is it allows you to represent context dependencies of information by using not only the amplitude but the phase structure. The amplitude is corresponding to the energy, and the phase is represented as the lag and lead of the time. That is, as like CVNNs, the complex valued reinforcement learnings (CVRLs) are defined as expansion of complex values from real value in ordinary RLs. One of the motivation to introduce complex values is to compensate for perceptual aliasing with employing the phase representing the context of the behavior. The novel originality of the CVRLs and algorithm are as follows:

- The value functions: Q-value: $Q(s, a)$ are defined as complex values.
- These complex values are revised by proposed learning algorithm in which framework is almost same as conventional one.
- The difference of them is that the proposed algorithm shifts the phase during an exploring environment.
- Instead of the select of an action according to the value function, this algorithm adopts the maximum inner products between a complex value of the function and an internal vector.
- The internal vector represents a context of an agent. That is, the algorithm enables an agent to select an

1-4244-0991-8/07/$25.00/©2007 IEEE

action according to the coherency between function values and internal context.

In the following sections, more specific algorithm using the complex valued function based on Q-learning is described.

B. Complex valued Q-learning: $\dot{Q}$-learning

Q-learning is a typical reinforcement learnibng algorithm classified into the learning of environment state identification. Q-value: $Q(s,a)$ indicates an expected value of reward , when an agent selects action a at state s. In the learning phase, $Q(s,a)$ is revised as follows.

$$\begin{aligned} Q(s_t,a_t) \quad \leftarrow \quad & (1-\alpha)Q(s_t,a_t) \\ & +\alpha(r_{t+1}+\gamma \max_{a'} Q(s_{t+1},a')) \end{aligned} \tag{1}$$

Where, α denotes a learning rate, γ denotes a discount rate, and r denotes a reward. In this paper, we assume that the agent obtains the reward iff it acheive a terminal state.

The proposed algorithm causes the Q-value function to expand to complex value as follows:

$$\dot{Q}(s_t,a_t) \leftarrow (1-\alpha)\dot{Q}(s_t,a_t)+\alpha(r_{t+1}+\gamma\dot{Q}^{(t)}_{\max})\dot{\beta}_t \tag{2}$$

The *dot* mark denotes the value is a complex value. $\dot{\beta}_t$ means an amount of rotation at the time t. The maximum of inner product: $\dot{Q}^{(t)}_{\max}$ is defined as follows:

$$\dot{Q}^{(t)}_{\max} = \dot{Q}(s_{t+1},a) \tag{3}$$

$$a = \underset{a'\in\mathcal{A}(s_{t+1})}{\mathrm{argmax}} \left(Re[\dot{Q}(s_{t+1},a')\overline{\dot{I}_t}]\right) \tag{4}$$

$$\dot{I}_t = \dot{Q}(s_t,a_t)/\dot{\beta}_t \tag{5}$$

In this equation, $*$ means the inner product of complex values. The reference vector: $\dot{I}_t(t)$ is an internal complex value of an agent, the vector which is estimated from the adjacent $\dot{Q}$-value shifting according to $\dot{\beta}_t$. In this paper, the $\dot{\beta}_t$ is assumed to be a time constant value.

Equations (2)-(5) imply that the complex value: $\dot{Q}(s_{t+1},a_{t+1})$ is diffused to adjacent value $\dot{Q}(s_t,a_t)$ along with $\dot{\beta}_t$. This phase function enables an agent to represent time series context into value functions. Fig. 1 shows the example of $\dot{Q}(s_t,a)$s on the complex value plain, and illustrates the phase of internal reference vector $\dot{I}$s shifting according to $\dot{\beta}_t$.

Furthermore, in order to improve the learning efficiencies, the idea of the eligibility trace[11] is applied. $\dot{Q}(s,a)$ is revised a trace length N_e times with k ($0 \leq k < N_e$) as follows:

$$\begin{aligned} \dot{Q}(s_{t-k},a_{t-k}) \quad \leftarrow \quad & (1-\alpha)\dot{Q}(s_{t-k},a_{t-k}) \\ & +\alpha(r_{t+1}+\gamma\dot{Q}^{(t)}_{\max})\dot{u}_t(k) \end{aligned} \tag{6}$$

.$\dot{u}_t(k)$ denotes an eligibility parameter. In this paper, we determine $\dot{u}(k)$ as

$$\dot{u}_t(k) = \dot{\beta}_t(k)\dot{u}_t(k-1) \quad k \geq 1 \tag{7}$$

$\dot{u}_t(0)$ is determined as an experimental parameter. In ordinary RLs, it is known that the eligibility trace makes learning speed fast[2]. It is expected that the eligibility trace in CVRLs produces an similar effect.

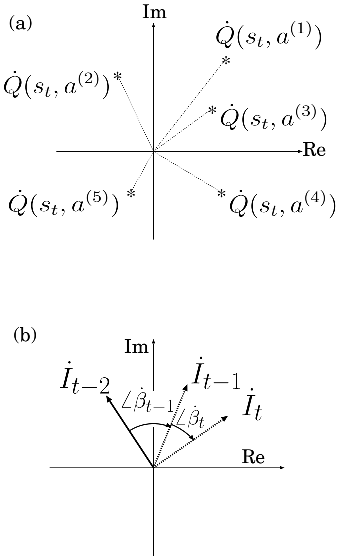

Fig. 1. An example of complex Q-values ($\dot{Q}$-values) and internal reference vectors on a complex plain. The internal reference vecrtor varies its phase with $\dot{\beta}_t$.

Fig. 2 shows a block diagram of complex valued Q-learning.

When an agent selects an action, the agent evaluates each action according to the stochastic policy as follows:

$$P(s_t,a) = \frac{\exp\left(Re[\dot{Q}(s_t,a)*\overline{\dot{I}_{t-1}}]/T\right)}{\sum_{a'\in\mathcal{A}(s_t)}\exp\left(Re[\dot{Q}(s_t,a')*\overline{\dot{I}_{t-1}}]/T\right)} \tag{8}$$

T denotes a scale parameter which controls the probability is varied from randomness to the roulette selection. This Boltzmann selection provides more high probability to the action corresponding to $\dot{Q}$-value which has not only greater norm: $|\dot{Q}|$ but more closer phase with $\dot{I}$.

III. Experiments and Results

We conduct two tasks; maze task and acrobot task.

In the maze task, the objective of an agent is to acquire behavior that leads the agent for the goal. The CVRL approach is applied to the maze task first.

Next, we introduce an acrobot task.[2] The acrobot is a double pendulum robot. To control the acrobot is known a one of hard task due to its nonlinearity and continious state space. The CVRL approach is applied to the acrobot task secound.

A. Maze task

Figure 3 shows the maze environment involving some perceptual aliasings. An agent starts from the cell indicated “start” and aims for the cell indicated “goal”. However, since agent’s sensors detect only the existence of walls around it, the agent can not distinguish some cells in the mazes.

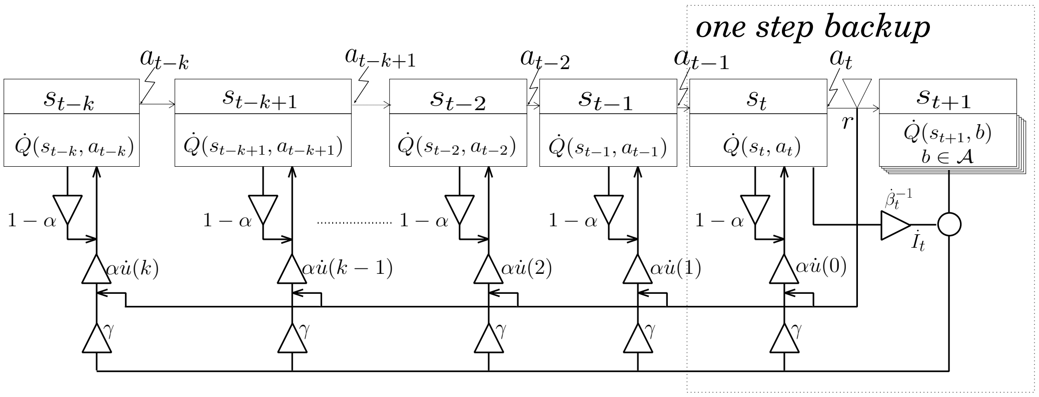

Fig. 2. Block diagram of the flows of complex values and the changes in the $\dot{Q}$-value with $\dot{I}$ in the learning phase.

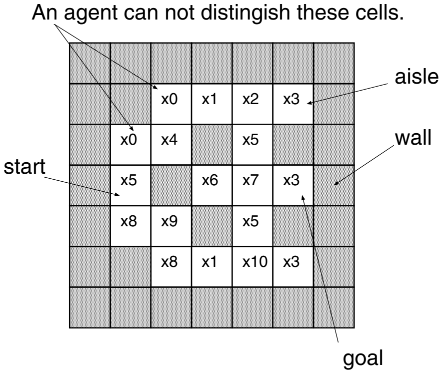

Fig. 3. Maze Tasks involving perceptual aliasings. Each label on the cell is an observation for the agent.

TABLE I

PARAMETER CONDITIONS IN EXPERIMENT OF MAZE TASK.

α	$\dot{\beta}_t(k)$	$\dot{u}_t(0)$	γ	T	N_e
0.25	$\exp(j\pi/6)$	$\exp(j\pi/6)$	0.9	20	2 - 5

Accordingly, an agent faces a serious problems in which the agent has to change its action according to the context.

The parameter condition is shown in Table I. The agent obtains a reward when agent arrives at the goal. In order to confirm the effectiveness of N_e, we conduct four experiments with various N_e.

Figure 4 reports the change of average steps through for accomplishing the task with various N_e. The average steps of random walk was 85.0. On the other hand, the agent with $\dot{Q}$-learning takes 12.0, 12.0, 13.2 and 15.2 steps averagely in each condition, whereas, the ideal steps are 8.0.

Figure 5 illustrates typical examples of acquired behaviors. We can confirm the agent needs redundant steps to adjust the internal phase for the contexts. The behavior in the case of $N_e = 5$ required more steps than $N_e = 2$.

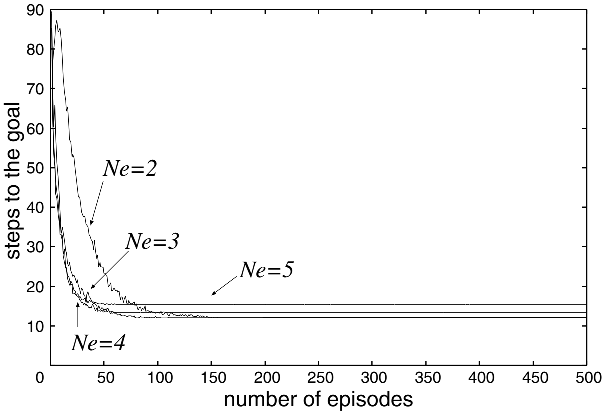

Fig. 4. Comparison of trace length toward goals in the maze. Each curve is an average of 1000 trials.

B. Acrobot task

Acrobot task is conducted in order to evaluate the effectiveness of eligibility traces in a harder task more than the previous maze task. The acrobot is a robot with two links and two joints shown in Fig.6 The objective of an agent is swinging up the toe. The joint 1 is a freejoint. The agent can torque only the joint 2. The equations of motion are as follows:

$$d_{11}\ddot{\theta}_1 + d_{12}\ddot{\theta}_2 + h_1 + \phi_1 = 0 \quad (9)$$

$$d_{21}\ddot{\theta}_1 + d_{22}\ddot{\theta}_2 + h_2 + \phi_2 = \tau \quad (10)$$

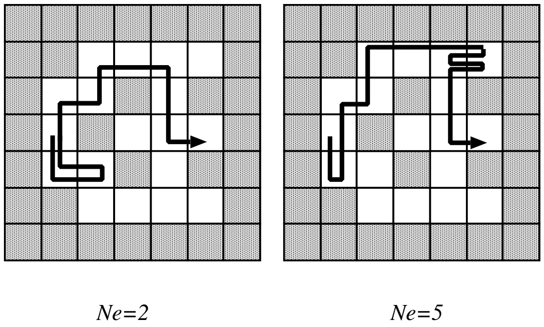

Fig. 5. Typical acquired behaviors in the maze task.

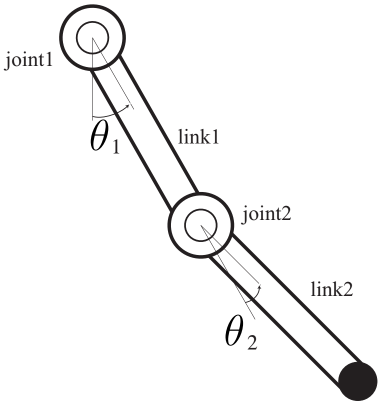

Fig. 6. Acrobot, a two-link and two-joint robot.

$$\begin{aligned}
d_{11} &= m_1r_1^2 + m_2l_1^2 + m_2r_2^2 \\
&\quad +2m_2l_1r_2\cos\theta_2 + I_1 + I_2 && (11)\\
d_{12} &= d_{21} = m_2r_2^2 + m_2l_2r_2\cos\theta_2 + I_2 && (12)\\
d_{22} &= m_2r_2^2 + I_2 && (13)\\
h_1 &= -m_2l_1r_2(2\dot{\theta}_1 + \dot{\theta}_2)\dot{\theta}_2\sin\theta_2 && (14)\\
h_2 &= m_2l_1r_2\dot{\theta}_1^2\sin\theta_2 && (15)\\
\phi_1 &= (m_1r_1 + m_2l_1)g\sin\theta_1 \\
&\quad +m_2r_2g\sin(\theta_1+\theta_2) && (16)\\
\phi_2 &= m_2r_rg\sin(\theta_1+\theta_2) && (17)
\end{aligned}$$

where θ_1 denotes the angle of joint 1, θ_2 denots the angle of joint 2 and $\tau \in \{-1, 0, 1\}$ denotes the toque. The dot means the time derivative above equations (11)-(17) . The angular velocities $\dot{\theta}_1$ and $\dot{\theta}_2$ are limited to $\dot{\theta}_1 \in [-4\pi, 4\pi]$ and $\dot{\theta}_2 \in [-9\pi, 9\pi]$, respectively.

We assume that the agent can observe only with θ_1 and θ_2. This partially observation may cause decreasing energy of the system because the agent can not distinguish which direction the acrobot moves to and which direction the agent should torque for.

The initial state of the acrobot in each episode is hanging down at rest. When swinging the toe above the given level, the agent obtains a reward r. Each sensor quantizes real-valued angle as shown in Table II.

TABLE II

SENSOR QUANTIZATION OF THE ACROBOT.

sensor type	quantization threshold[%]	number of levels
angular	1,5,10,30,50,70,90,95,99	10

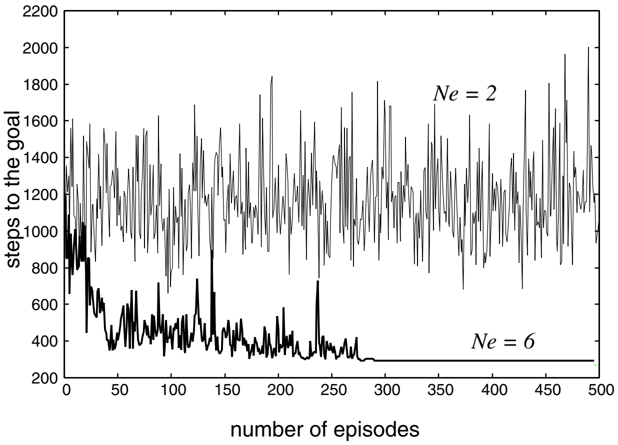

Fig. 7. Comparison of trace length toward goals in the acrobot task. Each curve is an average of 10 trials.

Figure 7 shows the changes in the steps with learnings to accomplish the task in the cases of $N_e = 2$ and $N_e = 6$. The average steps with random policy to accomplish the task was 1198.

When $N_e = 2$, agent can not improve steps to accomplish the task. In constant, steps decreased monotonically and converged to 295 when $N_e = 6$.

Typical example of acquired behavior with $N_e = 6$ is appeared in Figure 8. Each angle corresponding to each joint is plotted in time series. Amplitudes of the time series were on increase. Namely, the agent acquired suitable behavior in the task of acrobot with perceptual aliasing.

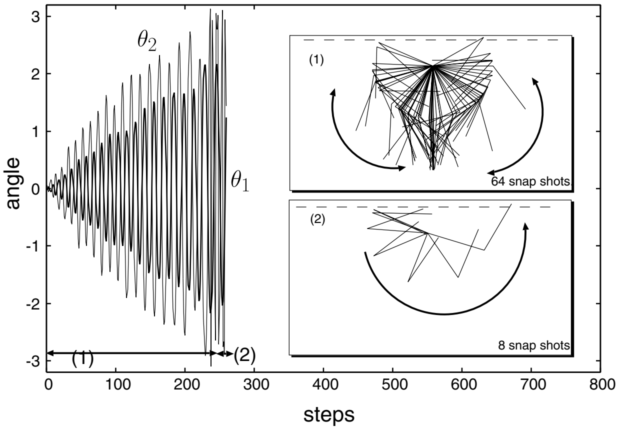

Fig. 8. Typical behavior of the acrobot.

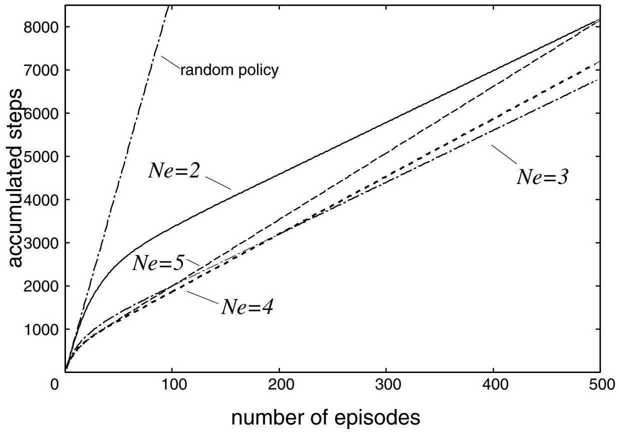

Fig. 9. Accumulated steps in the maze task.

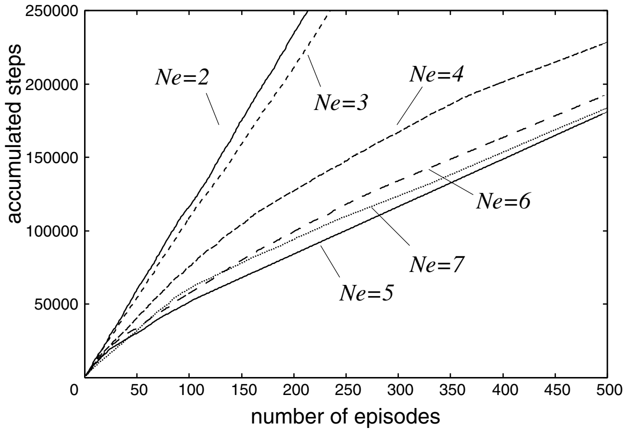

Fig. 10. Accumulated steps in the acrobot task.

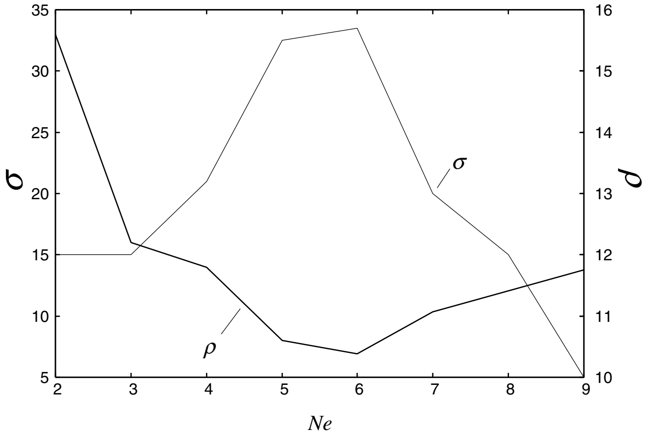

Fig. 11. Evaluation of convergence speed and performance in the maze task.

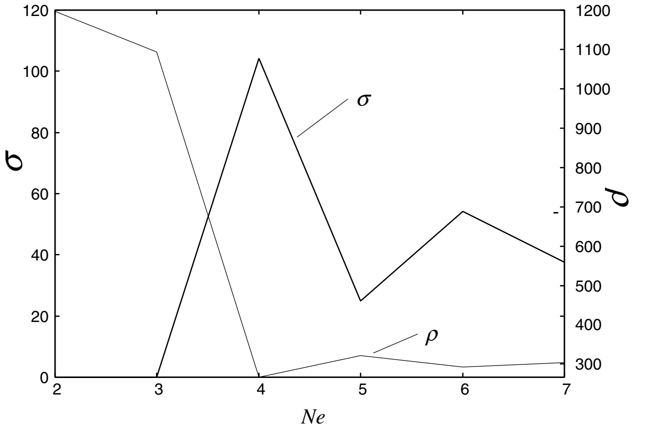

Fig. 12. Evaluation of convergence speed and performance in the acrobot task.

IV. DISCUSSIONS

The results show that higher N_e causes more redandant steps in the maze task experiment and that higher N_e supports more difficult tasks in the acrobot experiment. In order to discuss more details, we define convergence episode σ and convergence steps ρ. We determine them as follow procedure:

1) calculate accumlated steps from the first episode
2) draw tangent lines at the first episode and the last episode
3) define the incline of the tangent line at the last episode as ρ.
4) define an episode of an intersection between lines as σ.

Figure 9 and 10 show accumulated steps in maze task and acrobot task.

Figure 11 shows σ and ρ in the maze task. In this task, the agent with lower N_e can acquire much better behavior than random walk. Higher N_e causes better σ and worse ρ. Higher than $N_e = 7$ improves the ρ drastically because a minimum steps which the agent needs to move from start to an x5 near the goal.

Figure 12 also shows σ and ρ in the acrobot task. In this task, the agent with lower N_e can not acquire behavior. However, we confirmed that N_e greater than 4 enables the agent to acquire behavior with late σ.

V. CONCLUSIONS AND FUTURE WORKS

This paper focuses attention on the effectiveness of the eligibility ttaces in complex valued reinforcement learning. This results of simulation experiments infer that there is a possibility of overcoming tight perceptual aliasing with high N_e.

Our continuing work includes the expanded evaluation of the method against other reinforcement learning algorithms as well as several improvements to the method as follows:

- In order to apply to the dynamic environment, the phase of the internal vector has to be implemented as a time varying function.

- The method should be extended to more complex and high dimensional or continuous space.
- The method should be applied to real autonomous robots in an uncertain environment.

REFERENCES

[1] L.P.Kaelbling, M.L.Littman, A.W.Moore, “Reinforcement Learning: A Survey,” Journal of Artificial Intelligence Research, Vol.4, pp. 237–285, 1996.

[2] R.S.Sutton, A.G.Barto, Reinforcement Learning: An Introduction, MIT Press, 1998.

[3] T.Hamagami, H.Hirata, “State Space Partitioning and Clustering with Sensor Alignment for Autonomous Robots,” Proc of IEEE International Conference on Systems, Man and Cybernetics, pp.2655-2660, 2005.

[4] T.Hamagami, H.Hirata, T.Hamagami, H.Hirata, “Development of Intelligent Wheelchair Acquiring Autonomous, Cooperative, and Collaborative Behavior,” Proc of IEEE International Conference on Systems, Man and Cybernetics, pp.3235-3530, 2004.

[5] R. A. McCallum, “Instance-based utile distinctions for reinforcement learning with hidden state,” In Proceedings of the Twelfth International Conference on Machine Learning, pp. 387–395, 1995.

[6] M. Wiering and J. Schmidhuber, “HQ-learning,” Adaptive Behavior, vol. 6.2, pp.219-246, 1998.

[7] Satinder Singh, Michael L. Littman, Nicholas K. Jong, David Pardoe, and Peter Stone, “Learning predictive state representations,” In Proceedings of the Twentieth International Conference on Machine Learning, pp. 712–719, 2003.

[8] T.Hamagami, S.Koakutsu,;H. Hirata, ”Reinforcement learning to compensate for perceptual aliasing using dynamic additional parameter: motivational value,” Systems, Man and Cybernetics, 2002 IEEE International Conference on Volume 2, pp.1–6, 2002.

[9] T.Hamagami, T.Shibuya and S.Shimada, “Complex Valued Reinforcement Learning,” Systems, Man and Cybernetics, 2006 IEEE International Conference on Volume pp.3235-3530, 2006.

[10] ”Complex-Valued Neural Networks : Theories and Applications,” A. Hirose, ed., Series on Innovative Intelligence, World Scientific Publishing Co. Pte. Ltd., Singapore, Nov. 2003.

[11] A.G.Barto, R.S.Sutton, C.W.Anderson, “Neuronlike elements that can solve difficult learning control problems,” IEEE Trans. on Systems, Man, and Cybernetics, 13, pp.835–846, 1983.