Simulation of rat behavior by a reinforcement learning algorithm in consideration of appearance probabilities of reinforcement signals

Kazushi Murakoshi *, Takuya Noguchi

Department of Knowledge-based Information Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka, Tenpaku-cho, Toyohashi 441-8580, Japan

Received 3 August 2004; revised 27 October 2004; accepted 29 October 2004

Abstract

Brown and Wanger [J. Exp. Psychol. 68 (1964) 503] investigated rat behaviors with the following features: (1) rats were exposed to reward and punishment at the same time, (2) environment changed and rats relearned, and (3) rats were stochastically exposed to reward and punishment. The results are that exposure to nonreinforcement produces resistance to the decremental effects of behavior after stochastic reward schedule and that exposure to both punishment and reinforcement produces resistance to the decremental effects of behavior after stochastic punishment schedule. This paper aims to simulate the rat behaviors by a reinforcement learning algorithm in consideration of appearance probabilities of reinforcement signals. The former algorithms of reinforcement learning were unable to simulate the behavior of the feature (3). We improve the former reinforcement learning algorithms by controlling learning parameters in consideration of the acquisition probabilities of reinforcement signals. The proposed algorithm qualitatively simulates the result of the animal experiment of Brown and Wanger.

Key words:
reinforcement learning; rat behavior; simulation; appearance probability of reinforcement signal

* Corresponding author. phone: +81-532-44-6899; fax: +81-532-44-6873.
Email address: mura@tutkie.tut.ac.jp (Kazushi Murakoshi).

Article published in Biosystems, 80(1): 83-90, 2005. doi:10.1016/j.biosystems.2004.10.005

1 Introduction

Reinforcement learning (RL) (Sutton and Barto, 1998) is a theory for learning how to map situations to actions by trial-and-error so as to maximize a numerical reward signal. The theory has been applied to a variety of dynamic optimization problems such as game problems, robotic control, and dynamic allocation problems (Sutton and Barto, 1998). Schultz and Dayan (1997) and Montague et al. (1996), on the other hand, showed that the RL theory is able to account for dopamine neurons responses to prediction of reward *in vivo*. In this paper, we examine whether the RL theory is able to completely explain animal behaviors related to reward, punishment and, especially, appearance probabilities of reinforcement signals.

Brown and Wagner (1964) investigated rat behavior in conflict of approach and avoidance with the following features: (1) rats were exposed to reward and punishment at the same time, (2) environment changed and rats relearned, and (3) rats were stochastically exposed to reward and punishment. The results are that exposure to nonreinforcement produces resistance to the decremental effects of behavior after stochastic reward schedule and that exposure to both punishment and reinforcement produces resistance to the decremental effects of behavior after stochastic punishment schedule.

We confirm that previous proposed RL algorithms are able to account for the results of Brown and Wagner (1964). The conventional algorithms, such as Actor-Critic (AC) (Sutton and Barto, 1998) and Q-learning (Watkins and Dayan, 1992; Sutton and Barto, 1998), are unable to account for all features mentioned above. Accordingly, we evaluate several other recent algorithms, though based on the conventional algorithm. At first, we adapt two dimensional evaluation RL algorithm (Okada et al., 2001) for simulating the feature (1) of Brown and Wagner (1964) experiment. Second, we apply the RL algorithm for rapidly following unexpected environmental changes (Murakoshi and Mizuno, 2004) for corresponding to the feature (2) of Brown and Wagner (1964) experiment. However, we cannot find an algorithm according to the feature (3) of Brown and Wagner (1964) experiment.

This paper aims to simulate the rat behaviors in conflict of approach and avoidance by RL in consideration of appearance probabilities of reinforcement signals. The previous algorithms of RL were unable to simulate the behavior of the feature (3) of Brown and Wagner (1964) experiment, rats are stochastically exposed to reward and punishment. We improve the previous RL algorithms by controlling learning parameters in consideration of the acquisition probabilities of reinforcement signals.

2 Method

In order to simulate the rat behaviors of Brown and Wagner (1964) experiment, firstly, we briefly introduce the experiment, and model the experiment for reinforcement learning algorithms in Sec. 2.1. Secondly, in Sec. 2.2, we briefly introduce two dimensional evaluation RL by Okada et al. (2001), and show that the algorithm is expected to simulate the feature (1) of Brown and Wagner (1964) experiment, namely, (1) rats were exposed to reward and punishment at the same time. Thirdly, in Sec. 2.3, we extend RL algorithm for rapidly following unexpected environmental changes by Murakoshi and Mizuno (2004) with the two dimensional evaluation algorithm (Okada et al., 2001), and show that the RL algorithm is able to simulate the feature (2) of Brown and Wagner (1964) experiment, namely, (2) environment changed and rats relearned. Finally, in Sec. 2.4, we propose a RL algorithm in consideration of the acquisition probabilities of reinforcement signals for simulating the feature (3) of Brown and Wagner (1964) experiment, namely, (3) rats were stochastically exposed to reward and punishment.

2.1 experiment by Brown and Wagner (1964) and modeling

Brown and Wagner (1964) experimented behaviors of rats as follows. 3 groups of 30 rats were trained in a simple runaway to a goal box where they obtained electrical punishment or food reward. During acquisition, Group N was exposed to nonreinforcement on a 50% reward schedule, Group P was exposed to gradually increasing punishment along with consistent reward, while Group C was never punished and received reward only on all trials. After the acquisition, half of each group was then tested for the decremental effects of either consistent nonreinforcement (Nonreinforcement condition) or consistent both strong punishment and reward (Punishment condition).

Brown and Wagner (1964) reported running speeds as results. During the acquisition, mean running speeds of rats of each group were increasing with blocks of trials; the only difference among the groups was a repression of Group P speeds below those of Groups C and N. After the acquisition, several features of the results are apparent. First, the two test conditions produced similar response decrements in the two C subgroups that had received no prior experience with nonreinforcement or punishment. Secondly, exposure to either nonreinforcement or punishment during acquisition produced resistance to the decremental effects: the running speeds of Group P Subjects in Punishment condition and Group N Subjects in Nonreinforcement condition decreased only negligibly.

We model the above situation to simulate the behaviors of rats by a RL algorithm. Studies in rats have shown that neurons in the hippocampus have spatial firing fields (O'Keefe and Dostrovsky, 1971; Wilson and McNaughton, 1994). These cells are called place cells. Each place cell fires when an animal find itself in a particular location. Then, we define a discrete position of a rat as a state in RL algorithm. The length of the position is set to be a body length of the rat. The number of states is set to 6 including the goal state. An action in RL algorithm is, alternatively, advance or stay for simplicity.

2.2 Two dimensional evaluation RL (Okada et al., 2001)

To solve the problem of trade-off between exploration and exploitation actions in reinforcement learning, Okada et al. (2001) proposed two-dimensional evaluation reinforcement learning, based on conventional AC architecture, which distinguishes between reward and punishment evaluation forecasts. Critic consists of a reward section and a punishment section. Each section receives a state (s), a reward evaluation r_R, and a punishment evaluation r_P according to the environment. Interest (δ^+) and Utility (δ^-) are defined according to the temporal difference (TD) errors (δ_R and δ_P) as follows:

$$\delta^- = \delta_R - \delta_P \tag{1}$$

$$\delta^+ = |\delta_R| + |\delta_P|. \tag{2}$$

Actor learns an action strategy using δ^- (Utility) as a de facto reinforcement signal and δ^+ (Interest) to determine the ration of exploitation action to environmental search action.

Advance probability of a rat *Go* can be defined as follows:

$$Go = \frac{\exp\left(\frac{p(s,0)}{\delta^+(t)}\right)}{\exp\left(\frac{p(s,0)}{\delta^+(t)}\right) + \exp\left(\frac{p(s,1)}{\delta^+(t)}\right)}, \tag{3}$$

where $p(s,a)$ indicates the desirability of executing action a in state s at time t, and $a = 0$ means advance of the rat while $a = 1$ means stay of the rat. $p(s,a)$ is calculated as expressed below:

$$p(s,a) \leftarrow p(s,a) + \alpha_p \delta^-(t) \tag{4}$$

where positive constant α_p represents the learning rate of actions.

Conventional RL methods such as Actor-Critic (AC) (Sutton and Barto, 1998) utilize only the difference between reward and punishment. In comparison, the two dimensional evaluation method (Okada et al., 2001) determines the sum of reward and punishment to determine an action. Thus, we expect that the two dimensional evaluation RL algorithm (Okada et al., 2001) can simulate the feature (1) of Brown and Wagner (1964) experiment, namely, (1) rats were exposed to reward and punishment at the same time.

2.3 RL algorithm for rapidly following unexpected environmental changes (Murakoshi and Mizuno, 2004) and its expansion

In order to rapidly follow unexpected environmental changes, Murakoshi and Mizuno (2004) proposed a parameter control method in RL that changes each of learning parameters in appropriate directions by considering an emergency as a key word. To recognize unexpected environmental changes, Murakoshi and Mizuno (2004) simply computed the decrease in the current sum of reward from the previous sum of reward as follows:

$$\text{if} \quad (down\_r_{t-1} < down\_r_t) \quad \text{then} \quad sum\_r_{t-1} = 0 \tag{5}$$

$$down\_r_{t+1} = down\_r_t + (sum\_r_t - sum\_r_{t-1}) \tag{6}$$

$$\text{if} \quad (down\_r_{t+1} > 0) \quad \text{then} \quad down\_r_{t+1} = 0, \tag{7}$$

where sum_r_t is the current sum total of reward for a step interval n, and sum_r_{t-1} is the one previous sum total of reward; $down\_r$ is the variable indicating how much sum_r_t decreases compared with sum_r_{t-1}. Using this $down\_r$, Weight $w_{R\chi}$ which is attached to the learning parameter χ ($\chi = \alpha$, α_{act}, β, or γ) in RL is calculated to change the learning parameters after environmental changes as follows:

$$w_{R\chi} = 1 + \frac{h_{R\chi}}{1 + \exp((6/s_{R\chi})(down\_r + s_{R\chi}))}. \tag{8}$$

Figure 1 indicates $w_{R\chi}$. $w_{R\chi}$ is prevented from the divergence of learning by setting the maximum $h_{R\chi} + 1$ as shown in the sigmoid function of Fig. 1. The initial value of $w_{R\chi}$ is approximately one because $down\_r$ equals zero . When $down\_r$ efficiently decreases, the variable $w_{R\chi}$ increases $h_{R\chi} + 1$ times. To any $down\_r$ depending on learning problems, $w_{R\chi}$ is approximately maximum at the minimum of $down\_r$ owing to $6/s_{R\chi}$ in Eq. (8).

We expand the algorithm of Murakoshi and Mizuno (2004) with two dimensional evaluation algorithm (Okada et al., 2001). For two dimensional evaluation, we also compute the increase in the current sum of punishment from the previous sum of punishment. Because we think that the information of the increase of punishment is seriously important for animals. Therefore, the following calculation is proposed:

$$\text{if} \quad (up\_p_{t-1} > up\_p_t) \quad \text{then} \quad sum\_p_{t-1} = 0 \tag{9}$$

$$up\_p_{t+1} = up\_p_t + (sum\_p_t - sum\_p_{t-1}) \tag{10}$$

$$\text{if} \quad (up\_p_{t+1} < 0) \quad \text{then} \quad up\_p_{t+1} = 0 \tag{11}$$

$$w_P\chi = 1 + \frac{h_{P\chi}}{1 + \exp((6/s_{P\chi})(-up\_p + s_{P\chi}))}, \tag{12}$$

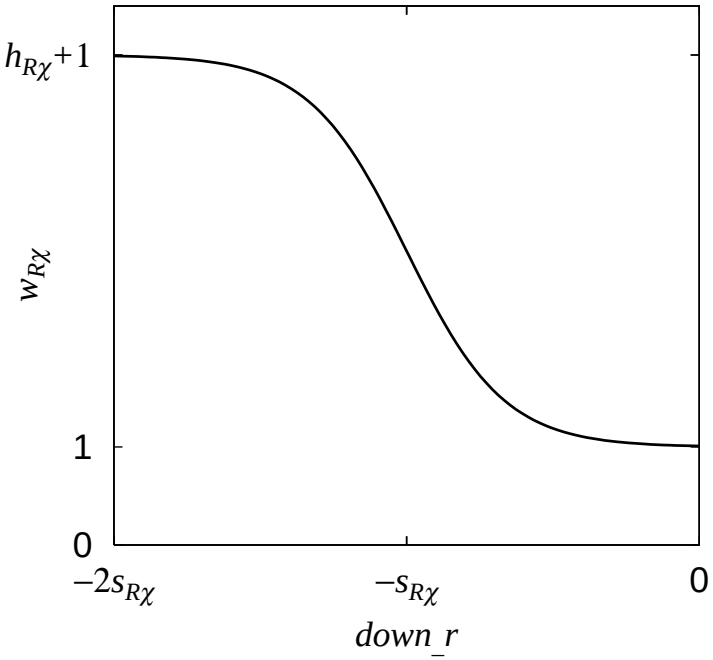

Fig. 1. Relationship of $down\_r$ to $w_{R\chi}$.

where sum_p_t is the current sum total of punishment for a step interval n, and sum_p_{t-1} is the one previous sum total of punishment; up_p is the variable indicating how much sum_p_t increases compared with sum_p_{t-1}. Figure 2 indicates $w_{P\chi}$. $w_{P\chi}$ is prevented from the divergence of learning by setting the maximum $h_{P\chi} + 1$ as shown in the sigmoid function of Fig. 2. When up_r efficiently increases, the variable $w_{P\chi}$ increases $h_{P\chi} + 1$ times. To any up_p depending on learning problems, $w_{P\chi}$ is approximately maximum at the minimum of up_p owing to $6/s_{P\chi}$ in Eq. (12).

We describe the equations involving the learning parameters varied by adopting the above proposed method as follows:

$$\delta_R = r_R(t) + (\gamma_R / w_{R\gamma}) \cdot V_R(t+1) - V_R(t) \tag{13}$$

$$\delta_P = r_P(t) + (\gamma_P / w_{P\gamma}) \cdot V_P(t+1) - V_P(t) \tag{14}$$

$$V_R(t) = V_R(t) + (\alpha_R \cdot w_{R\alpha}) \tag{15}$$

$$V_P(t) = V_P(t) + (\alpha_P \cdot w_{P\alpha}) \tag{16}$$

$$Go = \frac{\exp\left(\frac{p(s,0)\cdot w_\beta}{\delta^+(t)}\right)}{\exp\left(\frac{p(s,0)\cdot w_\beta}{\delta^+(t)}\right) + \exp\left(\frac{p(s,1)\cdot w_\beta}{\delta^+(t)}\right)} \tag{17}$$

$$p(s,a) \leftarrow p(s,a) + (\alpha_{act} \cdot w_{\alpha_{act}})\delta^-(t), \tag{18}$$

where w_β equals the average of $w_{R\beta}$ and $w_{P\beta}$; $w_{\alpha_{act}}$ equals the average of $w_{R\alpha_{act}}$ and $w_{P\alpha_{act}}$.

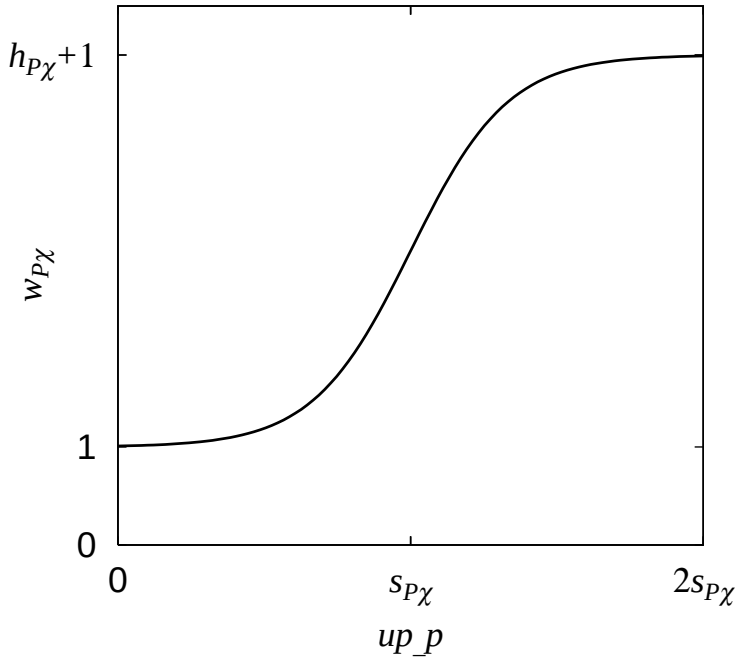

Fig. 2. Relationship of up_p to $w_{P\chi}$.

The learning parameters are not flexibly altered on conventional RL methods such as Actor-Critic (AC) (Sutton and Barto, 1998). In comparison, the RL algorithm for rapidly following unexpected environmental changes (Murakoshi and Mizuno, 2004) flexibly alters the learning parameters after environmental changes. Thus, we expect that the RL algorithm for rapidly following unexpected environmental changes (Murakoshi and Mizuno, 2004) is able to simulate the feature (2) of Brown and Wagner (1964) experiment, namely, (2) environment changed and rats relearned.

2.4 RL algorithm in consideration of appearance probabilities of reinforcement signals

In the experiment by Brown and Wagner (1964), rats were stochastically exposed to reward and punishment. Desiring to correspond to this situation, we hypothesize on appearance probabilities of reinforcement signals as follows. Rats are able to recognize appearance probabilities of nonreinforcement and punishment. Moreover, the rats can suppose the maximum reward and punishment when they are invariably exposed to reward and punishment. From the hypothesis, we formulate the rule updating $S_{R\chi}$ in Eq. (8) and $S_{P\chi}$ in Eq. (12) as follows:

$$\text{if} \quad (s_{R\chi} < 0.6 \cdot sum\_r/prob\_R) \quad \text{then} \quad s_{R\chi} = 0.6 \cdot sum\_r/prob\_R \qquad (19)$$

$$\text{if } (s_{P\chi} < 0.6 \cdot sum\_p/prob\_P) \text{ then } s_{P\chi} = 0.6 \cdot sum\_p/prob\_P, \quad (20)$$

where $prob\_R$ and $prob\_P$ ($0 < prob\_R,\ prob\_P \leq 1$) indicate appearance probabilities of reward and punishment, respectively, for step n. $prob\_R$ and $prob\_P$ are calculated as following equations:

$$\text{if } (count\_r! = 0) \text{ then } prob\_R = count\_r/count\_g \quad (21)$$

$$\text{if } (count\_p! = 0) \text{ then } prob\_P = count\_p/count\_g, \quad (22)$$

where $count\_g$, $count\_r$, and $count\_p$ are frequency of reaching the goal, frequency of obtaining reward, and frequency of obtaining punishment, respectively.

We discuss a possibility to assess the occurrence of appearance probabilities of reinforcement signals. In addition to the place cells as mentioned in Sec. 2.1, stimulation of certain regions of the brain of a rat acts as obtaining reward or punishment (Olds, 1976). If these reinforcement-related neurons are bound to the place cells by a certain method, it is possible to assess the occurrence of appearance probabilities of reinforcement signals. Although the biding method have not yet been perfectly clarified, there is a practicable method such as neuronal spike binding (Gray et al., 1989; Vaadia et al., 1995).

Even RL algorithm for rapidly following unexpected environmental changes (Murakoshi and Mizuno, 2004) could not perfectly simulate the behavior of rats in conflict of approach and avoidance because of lack of corresponding to appearance probabilities of reinforcement signals. In comparison, our current RL algorithm is taking appearance probabilities of reinforcement signals into consideration. Thus, we expect that the RL algorithm in consideration of appearance probabilities of reinforcement signals can simulate the feature (3) of Brown and Wagner (1964) experiment, namely, (3) rats were stochastically exposed to reward and punishment.

3 Simulation

We conduct simulations in the four reinforcement learning algorithms: conventional AC algorithm (Sutton and Barto, 1998), two dimensional evaluation algorithm (Okada et al., 2001), extended RL algorithm for rapidly following unexpected environmental changes (Murakoshi and Mizuno, 2004), and RL algorithm in consideration of appearance probabilities of reinforcement signals. AC architecture (Sutton and Barto, 1998) involves discount factor (γ), learning rate of the critic (α), learning rate of the actor (α_{act}), and inverse temperature (β). We obtained the parameter values, $\gamma_R = 0.9$, $\gamma_P = 0.9$, $\alpha_R = 0.1, \alpha_P = 0.1$, $\alpha_{act} = 0.05$, and $\beta = 3.3$, from the preliminary simulations in which rats learn to reach the goal similar to the behavior in the experiment for the acquisition by Brown and Wagner (1964).

Here we have necessity to decide the values of reward and punishment. The values, however, cannot be directly found in the experiment by Brown and Wagner (1964). Then, we obtain the ratio of the maximum value of reward to that of punishment from the preliminary simulations: we found the ratio which rats learn to reach the goal for the acquisition term. The ratio was 1.0 to 1.4; thus, we assign the value 1.0 and 1.4 to the maximum value of reward and punishment, respectively, in all simulations. In all acquisition simulations with punishment, value of the punishment is set to increase by 4% by learning middle stage at every block, and the value is set to increase by 8% after the middle stage at every block by adjusting to Brown and Wagner (1964) experiment. The remaining parameters are described in each subsection of simulation.

3.1 Conventional AC

Only the evaluation of reinforcement signal of this algorithm is one dimensional. Thus, the parameters are set: $\gamma = \gamma_R = \gamma_P$, $\alpha = \alpha_R = \alpha_P$, and the values of punishment are redefined as negative.

The simulation results are shown in Fig. 3 and 4. The left hand side from the dotted vertical line at 10 blocks is the term of the acquisition while the right hand side is the term of each condition. Although learning for the acquisition almost succeeded, the conventional AC could not simulate the experiment by Brown and Wagner (1964) in both conditions: especially, there was almost no difference among the groups.

3.2 Two dimensional evaluation algorithm (Okada et al., 2001)

The simulation results are shown in Fig. 5 and 6. Although the difference between the result of Nonreinforcement and Punishment expanded, two dimensional evaluation algorithm (Okada et al., 2001) could not simulate the experiment by Brown and Wagner (1964) in both conditions: there was no large difference among the groups.

3.3 extended RL algorithm for rapidly following unexpected environmental changes

The remaining parameters are set as follows: $n = 100$, $h_\gamma = 1.0$, $h_{R\beta} = h_{P\beta} = 1.0$, $h_{R\alpha} = h_{P\alpha} = 2.0$, $h_{R\alpha_{act}} = h_{P\alpha_{act}} = 4.0$ obtained from Murakoshi and Mizuno (2004).

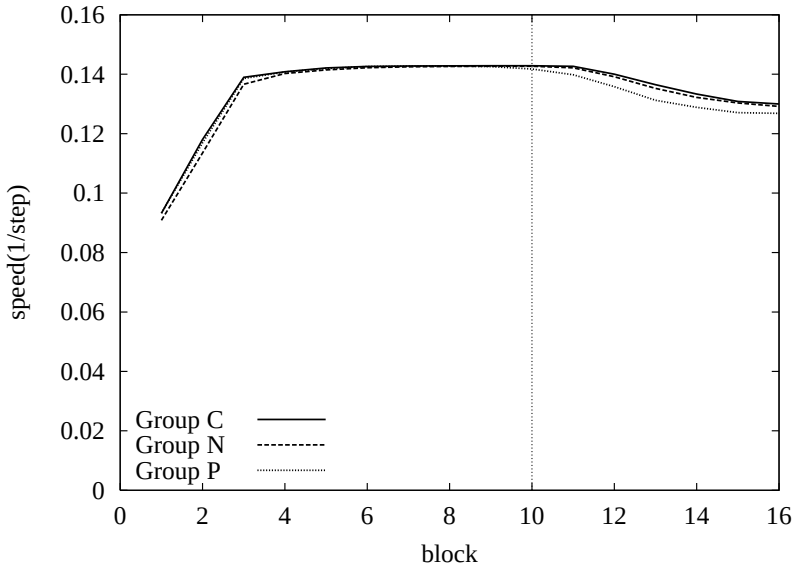

Fig. 3. Simulation result in Nonreinforcement condition by conventional AC.

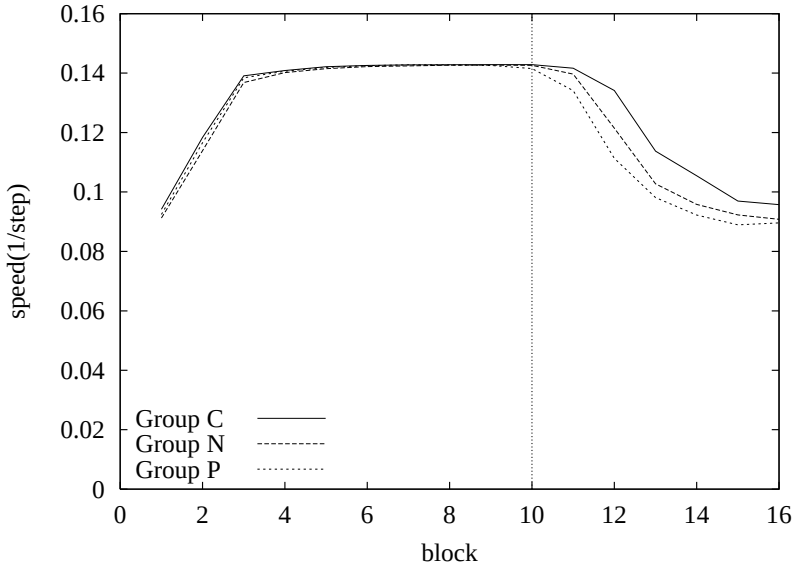

Fig. 4. Simulation result in Punishment condition by conventional AC.

The simulation results are shown in Fig. 7 and 8. Although the difference between the result of Nonreinforcement and Punishment spreaded, the extended RL algorithm for rapidly following unexpected environmental changes could not perfectly simulate the experiment by Brown and Wagner (1964) : the running speeds of Group C Subjects in Punishment condition did not decrease, and the running speeds of Group N Subjects in Nonreinforcement condition fell from the initial speeds.

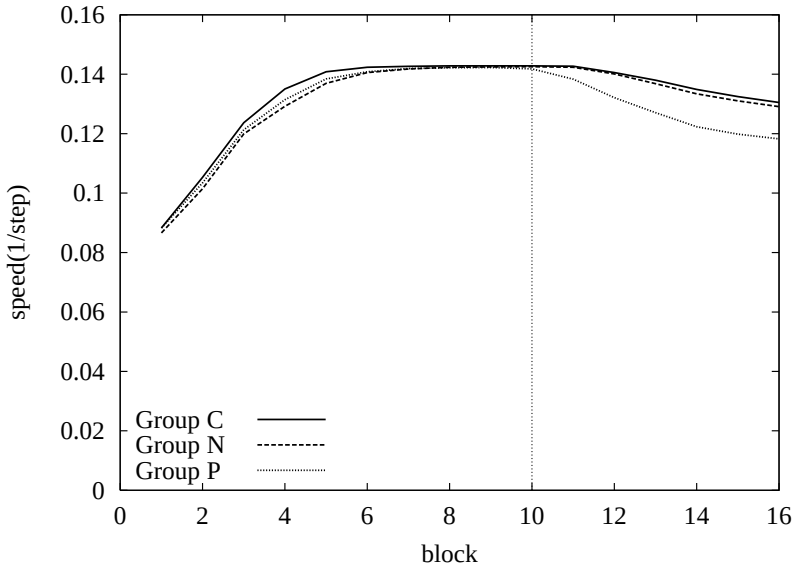

Fig. 5. Simulation result in Nonreinforcement condition by two dimensional evaluation algorithm (Okada et al., 2001)

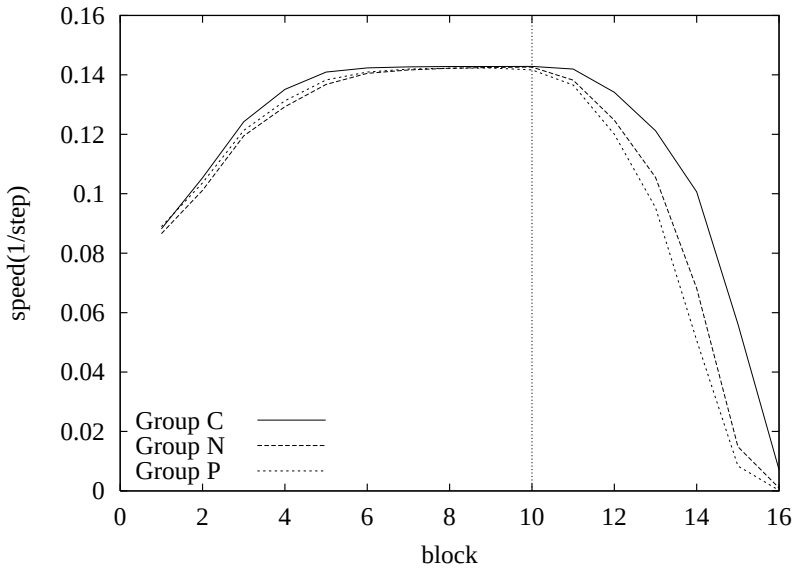

Fig. 6. Simulation result in Punishment condition by two dimensional evaluation algorithm (Okada et al., 2001).

3.4 RL algorithm in consideration of appearance probabilities of reinforcement signals

All parameters are the same as shown in Sec. 3.3. The simulation results are shown in Fig. 9 and 10. The RL algorithm in consideration of appearance probabilities of reinforcement signals could quantitatively simulate the experiment by Brown and Wagner (1964): the running speeds of Group C Subjects

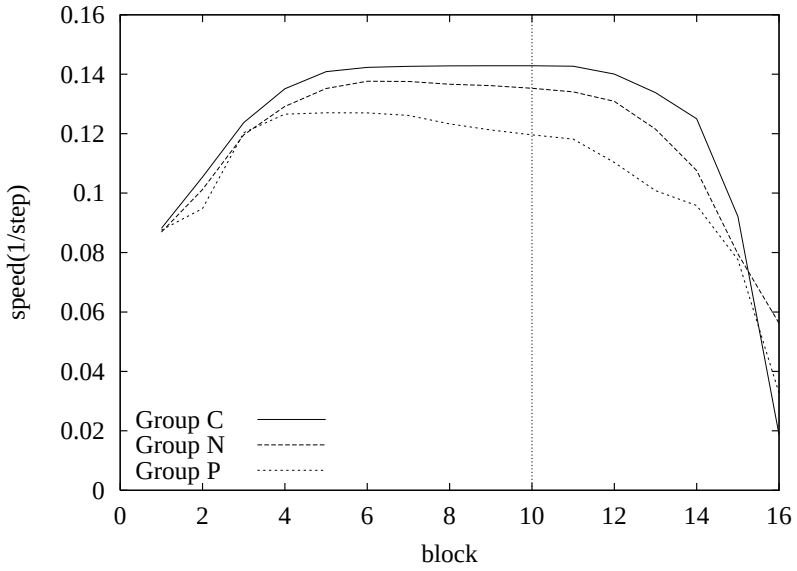

Fig. 7. Simulation result in Nonreinforcement condition by extended RL algorithm for rapidly following unexpected environmental changes.

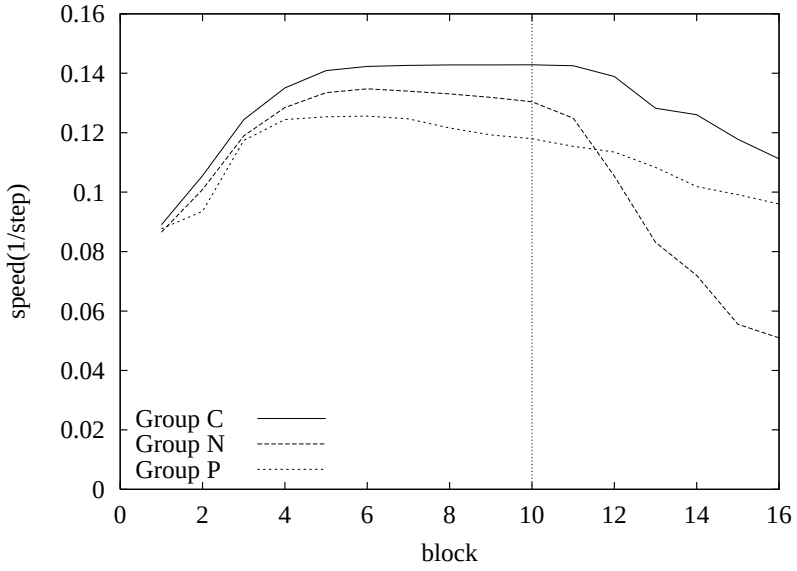

Fig. 8. Simulation result in Punishment condition by extended RL algorithm for rapidly following unexpected environmental changes.

in both conditions decreased, and the running speeds of both Group N Subjects in Nonreinforcement condition and Group P in Punishment condition did not decrease from those of the other conditions.

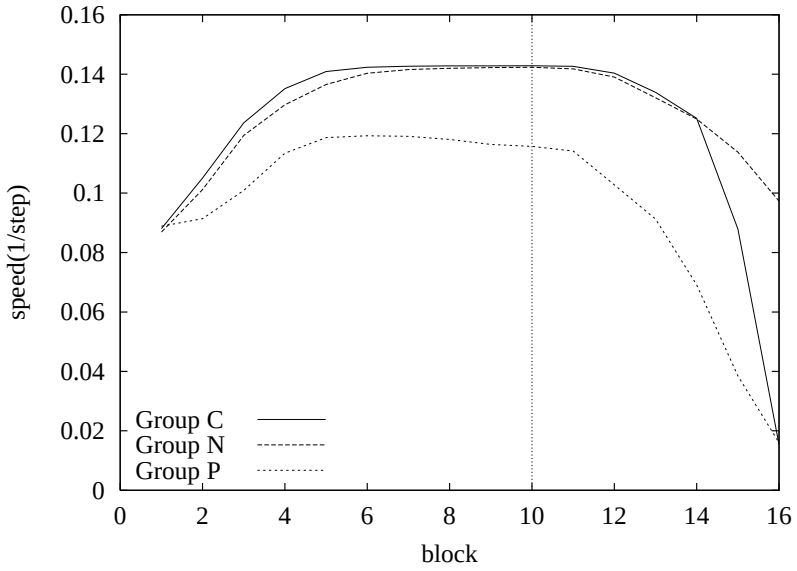

Fig. 9. Simulation result in Nonreinforcement condition by RL algorithm in consideration of appearance probabilities of reinforcement signals.

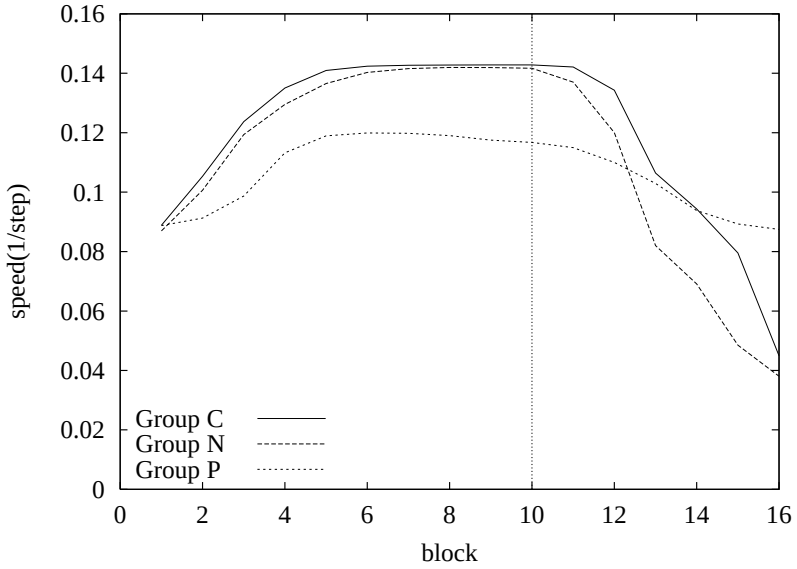

Fig. 10. Simulation result in Punishment condition by RL algorithm in consideration of appearance probabilities of reinforcement signals.

4 Conclusion

We propose a RL algorithm in consideration of the acquisition probabilities of reinforcement signals. The algorithm qualitatively simulates the results of the animal experiment of Brown and Wagner (1964) in which rats were stochastically exposed to reward and punishment.

In the experiment by Brown and Wagner (1964), the appearance probabilities of reinforcement signals are only 50%, 100% or 0%. Our proposed algorithm is able to forecast the results with the other appearance probabilities. The verification of the results will have to wait for the results of another behavioral experiments. Additionally, considering how the brain hardware recognizes appearance probabilities of reinforcement signals is a future work.

References

Brown, R. T., and Wagner, A. R., 1964. Resistance to punishment and extinction following training with shock or nonreinforcement. J. Exp. Psychol., 68, 503–507.

Gray, C. M., Kőnig, P., Engel, A. K., and Singer, W., 1989. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature, 338, 334–337.

Montague, P. R., Dayan, P., and Sejnowski, T. J., 1996. A framework for mesencephalic dopamine systems based on predictive hebbian learning. J. Neurosci., 16, 1936–1947.

Murakoshi, K., and Mizuno, J., 2004. A parameter control method in reinforcement learning to rapidly follow unexpected environmental changes. Biosystems, 77(1-3), 109-117.

Okada, H., Yamakawa, H., and Omori, T. h., 2001. Two dimensional evaluation reinforcement learning. Lect. Notes Comput. Sci., 2084, 370–377.

O'Keefe, J., and Dostrovsky, J., 1971. The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat. Brain Res., 34, 171–5.

Olds, J. 1976. Reward and drive neurons. In A. Wauquier and E. Rolls (Eds.), Brain-stimulation reward (pp. 1–27). North-Holland Publishers.

Schultz, W., and Dayan, P. R., P. Motague., 1997. A neural substrate of prediction and reward. Science, 275, 1593–1599.

Sutton, R. S., and Barto, A. G. 1998. Reinforcement learning: An introduction. The MIT Press.

Vaadia, E., Haalman, I., Abeles, M., Bergman, H., Prut, Y., Slovin, H., and Aertsen, A., 1995. Dynamics of neuronal interactions in monkey cortex in relation to behavioural events. Nature, 373, 515–518.

Watkins, C. J. C. H., and Dayan, P., 1992. Technical note: Q-learning. Machine Learning, 8, 279–292.

Wilson, M., and McNaughton, B., 1994. Reactivation of hippocampal ensemble memories during sleep. Science, 265, 676–679.