A Parameter Control Method in Reinforcement Learning to Rapidly Follow Unexpected Environmental Changes

Kazushi Murakoshi *, Junya Mizuno

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

Received 9 May 2004; revised 3 May 2004; accepted 7 May 2004

Abstract

In order to rapidly follow unexpected environmental changes, we propose a parameter control method in reinforcement learning that changes each of learning parameters in appropriate directions. We determine each appropriate direction on the basis of relationships between behaviors and neuromodulators by considering an emergency as a key word. Computer experiments show that the agents using our proposed method could rapidly respond to unexpected environmental changes, not depending on either two reinforcement learning algorithms (Q-learning and actor-critic architecture) or two learning problems (discontinuous and continuous state-action problems).

Key words:
emergency; relearning; neuromodulator; serotonin; noradrenaline; acetylcholine

1 Introduction

Reinforcement learning (Sutton & 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 & Barto, 1998). In such dynamic environments, consider a

* 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, 77(1-3): 109-117, 2004. doi:10.1016/j.biosystems.2004.05.001

case in which an agent controlled by reinforcement learning encounters unexpected environmental changes after the agent has almost learned the environment. Can the agent rapidly follow the unexpected environmental changes?

Schweighofer and Doya (2003) proposed a learning algorithm based on a conceptual theory (Doya, 2002) in which parameters in reinforcement learning are adjusted by neuromodulators. In their algorithm, three learning parameters are dynamically adjusted. If all three learning parameters are changed appropriately, the agent will be able to follow unexpected environment changes rapidly and, simultaneously, flexibly. The algorithm of Schweighofer and Doya (2003), however, could not rapidly respond to unexpected environmental changes. The reason is that not all learning parameters are always improved in their method since they are changed by stochastic method.

In order to rapidly follow unexpected environmental changes, we propose a parameter control method that changes each of the parameters in appropriate directions. We determine each appropriate direction on the basis of relationships between behaviors and neuromodulators by considering an emergency as a key word. To evaluate the performance of our method, we compare it to the algorithm of Schweighofer and Doya (2003) and the algorithms with fixed parameters, by means of computer experiments.

Section 2 explains the appropriate directions of changing parameters in an emergency such as an unexpected environmental change. We propose in Section 3 a parameter control method of the rapidly following unexpected environmental changes. Section 4 shows the results of computer experiments. Section 5 concludes this paper.

2 Parameters in reinforcement learning and neuromodulators in an emergency

In order to rapidly follow unexpected environmental changes, we discuss how the appropriate directions of changing parameters are decided. The number of combination of the directions of changing three parameters is eight. When the time has been limited, it is difficult to find the optimum combination. Then, we decide the directions of the change by considering an emergency as a key word. An emergency is an unexpected and difficult situation, which happens suddenly and which requires rapid actions to deal with it. We describe below the correspondence of the parameters in the reinforcement learning algorithms and neuromodulators in an emergency such as an unexpected environmental change.

An agent in reinforcement learning to learn to obtain reward r corresponding

to its action. The state value function V_t is then defined as

$$V_t = r_t + \gamma \cdot V_{t+1}. \tag{1}$$

where γ is a parameter, $0 \leq \gamma \leq 1$, called discount factor. The discount factor γ, which is controlled by serotonin (5-HT), determines how far into the future the agent should consider in reward prediction and action selection (Doya, 2002). A low level of serotonin is often associated with impulsive behaviors, such as aggression (Wolff & Leander, 2002; Doya, 2002). Wolff and Leander (2002), for example, showed that selective serotonin reuptake inhibitors (SSRIs) decrease impulsive behavior as measured the length of delay to a large reinforcement in the pigeon. In the experiment, the pigeon mostly chose not a small immediate reward but a larger delayed reinforcement in a higher level of serotonin. In an emergency, on the other hand, rapid actions are required. Such rapid actions will lead to a choice of a smaller immediate reward. Thus, we consider that the decrease of serotonin (γ) will be appropriate in an emergency.

Any deviation from the consistency condition in Eq. (1), expressed as

$$\delta = r_t + \gamma V_{t+1} - V_t, \tag{2}$$

should be zero on average. This temporal difference (TD) error δ, which is signaled by dopamine (DA), is the essential learning signal for reward prediction and action selection. We think this TD error δ does not need not to be changed because its change is included in the conventional reinforcement learning algorithms.

The state value function V_t is updated as

$$V_t \leftarrow V_t + \alpha \cdot \delta, \tag{3}$$

where α is the learning rate. The learning rate α is controlled by acetylcholine (ACh) (Doya, 2002). What has already been learned could be rapidly overwritten if α is set to a very large value. Fadda, Cocco, and Stancampiano (2000) found acetylcholine release increased during learning tasks. The beginnings of learning tasks happen suddenly and require rapid actions to deal with it. In a word, the beginnings of learning tasks are a kind of beginnings of emergency. Thus, we consider that the increase of acetylcholine (α) is required in an emergency.

A typical method of action-selection according to the action value p_{ij} (i and j indicate a state and an action, respectively) is Boltzmann selection. It chooses

action with probability

$$P_{ij} = \frac{\exp(\beta p_{ij})}{\sum_{k=1}^{n} \exp(\beta p_{ik})}, \tag{4}$$

where the parameter β, which is called the inverse temperature, is controlled by noradrenaline (NA) (Doya, 2002). High inverse temperatures cause a greater difference in selection probability for actions that differ in their value estimates. This means that an agent does exploitation by using the probability of actions effectively. The noradrenergic neurons in the locus coeruleus (LC) have been known to be activated when an animal encounters an unfamiliar environment (Vankov, Hervé-Minvielle, & Sara, 1995). This situation is exactly an emergency. Thus, noradrenaline (β) increases in an emergency.

In the other algorithms (e.g. Ishii, Yoshida, & Yoshimoto, 2002), the parameter β is set small in order to encourage exploratory behaviors when environment changes. This method is effective when there is enough time. If there is no enough time for exploration, however, the agent could not have much reward for a short time. To the contrary, β in our method is changed to the opposite direction in order to obtain reward rapidly.

From the above consideration, we summarize relationships between neuromodulators and behaviors in emergency in Table 1. In our method, we change the parameters in these directions as a hypothesis, respectively, in order to obtain reward rapidly. This hypothesis is verified by computer experiments in Section 4.

3 Parameter control method

In consideration of the discussion in Section 2, we design an algorithm with the capability of relearning rapidly to follow environmental changes by changing the learning parameters, α, γ, and β. In order to recognize unexpected environmental changes, we simply compute the decrease in the current sum of reward from the previous sum of reward.

Therefore, the following algorithm is proposed.

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

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

Table 1
Neuromodulators and behaviors in emergency.

parameter	neuromodulator	behavior
α up	acetylcholine up	renewal
β up	noradrenaline up	exploitation
γ down	serotonin down	impulsive behavior

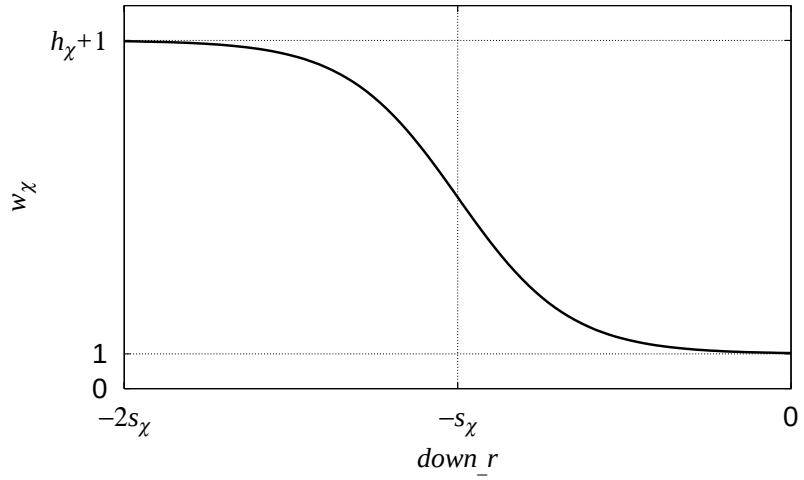


Fig. 1. Relationship of *down_r* to weight of parameters.

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

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}. When the agent does not obtain reward owing to unexpected environmental changes, $down\_r$ is decreased. The amount of reduction of obtained reward is expressed by adding the difference between sum_r_t and sum_r_{t-1} to $down\_r$, as shown in Eq. (6). Since there is no necessity for relearning when $down\_r$ is positive, it is set to zero in Eq. (7). Equation (5) is necessary for returning $down\_r$ to zero after relearning.

To change the learning parameters after environmental changes, the weight w_χ which is attached to the learning parameter χ (χ =α, β, or γ) is calculated as

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

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

In the algorithm of Schweighofer and Doya (2003), the difference between a short-term and a long-term running average of the reward to update learn-

ing parameters, which leads to detect environmental changes. The difference value, however, gradually approaches to zero even if obtained reward is not increased after an environmental change. This would be one of reasons why the algorithm cannot rapidly respond to environmental changes. In our method, the decrease of obtained reward, $down\_r$, as calculated in Eqs. (5), (6), and (7) does not return to zero until obtained reward is increased after an environmental change. At this point, thus, an agent in our method is expected to learn earlier than the other method.

4 Computer experiments

4.1 Fundamental learning algorithms for experiments

With respect to the method proposed in Section 3, it is desirable that it is broadly applicable. Our method, therefore, is applied to two primary reinforcement learning algorithms, Q-learning (Watkins & Dayan, 1992; Sutton & Barto, 1998) and Actor-Critic (AC) (Sutton & Barto, 1998), to verify its independence from reinforcement learning algorithms. We describe the equations involving the learning parameters made variable by adopting the proposed method in Section 3 to each reinforcement learning algorithm as follows.

4.1.1 Q-learning

Q-learning is a method for learning all state action values. w_α, w_γ, and w_β are attached to the parameters α, γ, and β in the Q-learning algorithm, respectively, as

$$\delta = r_t + (\gamma / w_\gamma) \cdot \max_j Q_{i'j} - Q_{ij} \tag{9}$$

$$Q_{ij} \leftarrow Q_{ij} + (\alpha \cdot w_\alpha) \cdot \delta \tag{10}$$

$$P_{ij} = \frac{\exp((\beta \cdot w_\beta) Q_{ij})}{\sum_{k=1}^{n} \exp((\beta \cdot w_\beta) Q_{ik})}. \tag{11}$$

4.1.2 Actor-Critic (AC)

AC architecture consists of the critic evaluating state and the actor selecting actions. The actor selects actions by using the error δ from the critic. w_α, w_γ, and w_β are attached to the parameters α, γ, and β in the AC algorithm,

respectively, as

$$\delta = r_t + (\gamma / w_\gamma) \cdot V_{t+1} - V_t \tag{12}$$

$$V_t \leftarrow V_t + (\alpha \cdot w_\alpha) \cdot \delta \tag{13}$$

$$P_{ij} = \frac{\exp((\beta \cdot w_\beta) p_{ij})}{\sum_{k=1}^{n} \exp((\beta \cdot w_\beta) p_{ik})}, \tag{14}$$

In the AC algorithm, the critic and the actor have learning speed parameters, α and α_p, respectively. Here, the parameter α_p in the AC algorithm is multiplied by w_{α_p} as

$$p_{ij} \leftarrow p_{ij} + (\alpha_p \cdot w_{\alpha_p}) \cdot \delta. \tag{15}$$

4.2 Learning problems

In order to verify the independence of the proposed method from learning problems, we perform two experiments: a maze problem and a two-linked arm robot moving forward problem. The former is a typical problem in which states and actions are discontinuous, while the latter is a problem in which the states and actions are continuous. We examine whether the proposed method is effective for solving these two problems which include unexpected environmental changes.

The setting parameters of our proposed method are described in Table 2. These parameters are the best in various parameters which we tried. The parameters in Q-learning were $\alpha = 0.1$, $\gamma = 0.9$, and $\beta = 3.3$; those in AC were $\alpha = 0.1$, $\alpha_p = 0.02$, $\gamma = 0.9$, and the $\beta = 3.3$ after preliminary experiments to learn adequately. In the algorithm, we used $\tau_1 = 100$, $\tau_2 = 100$, $\mu = 0.2$, $\nu^2 = 0.3$, and $n = 100$. Please refer to Schweighofer and Doya (2003) for the details of these parameters and the algorithm. The learning parameters at unexpected environmental changes were reset as follows for comparison with other methods: $\alpha = 0.1$, ($\alpha_p = 0.02$ in AC), $\gamma = 0.9$, and $\beta = 3.3$.

4.2.1 Maze problem

This robot learns to find the path from the start to the goal (Kimura, 2002) as shown in Fig. 2(a). A reward 100 is received at reaching the goal. The robot restarts from the start point when it reaches the goal point. The states (25) are positions of the robot; the actions (4) involve moving in four directions: up, down, right, and left.

Table 2
Setting parameters.

	h_χ	updating steps n	s_χ
w_γ	1	100	60% of
w_β	1	100	maximum
w_α in Q-learning	9	100	sum_r_t
w_α in AC	2	100	(initial value = 1)
w_{α_p} in AC	4	100	

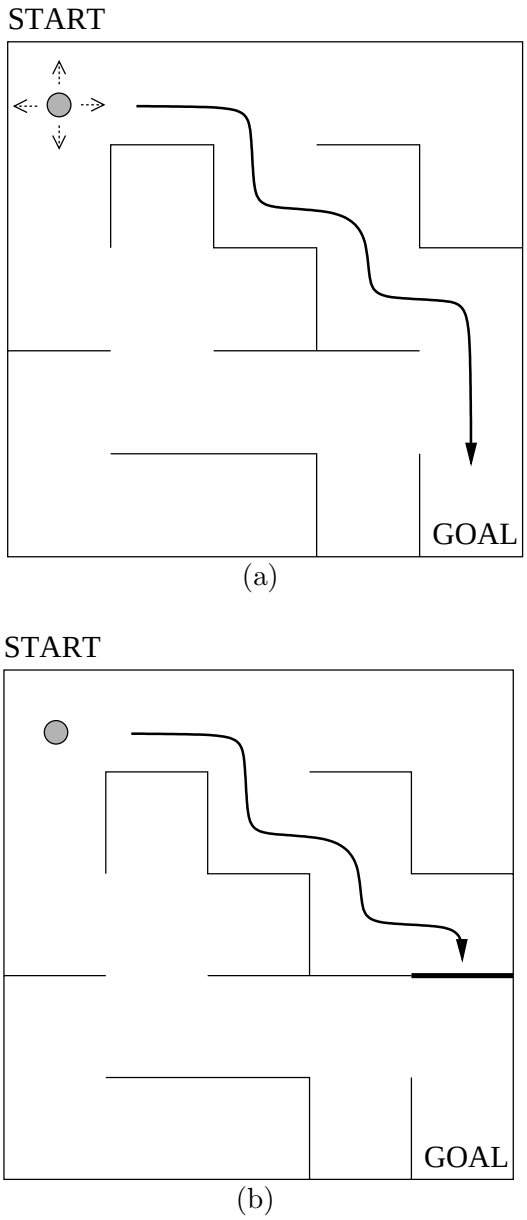


Fig. 2. Maze problem. (a) Normal situation. (b) New wall as an unexpected environmental change.

First, we make the robot learn to perform almost convergent actions for 5000 steps in the environment as shown in Fig. 2(a). Next, a new wall which blocks the robot appears in the learned path as shown in Fig. 2(b) as an unexpected environmental change. The robot must relearn an other path to the goal. After this change, we make the robot learn for 2000 steps with the parameters maintained.

We have repeated the experiments twenty times in each case: the fixed learning

parameters, Schweighofer and Doya (2003), and our proposed method. The results in Q-learning and AC are shown in Table 3.

Table 3 shows the mean ± SEM (standard error of the mean) obtained reward after the environmental change for 2000 steps in Q-learning and AC. The mean reward of our method was obviously the best from among all the methods. The data of the first step of obtaining the reward cannot be calculated in the fixed parameters and in the Schweighofer and Doya (2003) algorithm because the agents in those methods almost could not obtain the reward in 2000 steps after the environmental change.

Figure 3 indicate an example of changes in $down\_r$, $\alpha \cdot w_\alpha$, γ / w_γ, and $\beta \cdot w_\beta$, and subtotal of reward for 100 steps around the environmental change in Q-learning by our method. $\alpha \cdot w_\alpha$, γ / w_γ, and $\beta \cdot w_\beta$ are the learning rate, the discount factor, and the inverse temperature, respectively. After the environmental change, obtained reward returned efficiently by three parameters controlled by $down_r$.

4.2.2 Two-linked arm robot moving forward problem

A two-linked arm robot learns to move forward (Kimura & Kobayashi, 1997; Kimura, 2002) as shown in Fig. 4(a). The reward is defined as the length that the body moved forward in the current step (Moving backward is a negative reward). We apply Q-learning with CMAC (cerebellar model articulation controller) (Albus, 1975; Sutton, 1996) to this problem for continuous states and actions. The CMACs consisted of six tilings with eight divisions of each dimension and a constant offset. The learning parameter in this Q-learning with CMAC was $\alpha = 0.05$ because the robot with $\alpha = 0.1$ cannot learn well even in the first constant environment. The remainder of parameters is described in the previous experiment.

First, we make the robot learn to perform almost convergent actions for 20000 steps in the environment as shown in Fig. 4(a). Next, a tunnel appears in front of the robot as shown in Fig. 4(b) as an unexpected environmental change. The entrance of the tunnel becomes the obstacle when the arm of the robot catches on it. After this environmental change, we make the robot to learn for 2000 steps with the parameters maintained.

We have repeated the experiments twenty times in each case: the fixed learning parameters, Schweighofer and Doya (2003) algorithm, and our proposed method. The results are shown in Table 4.

Table 4 shows the mean ± SEM obtained reward after the environmental change for 2000 steps. The mean reward of our method is significantly higher than those of any other methods. The mean first step of our method was the

Table 3
Results of the maze problem. (a) Obtained reward after the environmental change for 2000 steps. (b) First step of obtaining reward after the environmental change. Data are mean ± SEM <number of times of reward acquisition> of twenty repetitions. $^{\dagger}p < 0.05$ is significantly better than the Schweighofer and Doya (2003) algorithm.

(a)

	fixed parameters	Schweighofer and Doya Algorithm	proposed method
Q-learning	0±0 < 0 >	20±16 < 2 >	†18595±495 < 20 >
AC	0±0 < 0 >	5±5 < 1 >	†13490±1326 < 20 >

(b)

	fixed parameters	Schweighofer and Doya Algorithm	proposed method
Q-learning	— < 0 >	— < 2 >	382±27 < 20 >
AC	— < 0 >	— < 1 >	366±38 < 20 >

Table 4
Results of the arm robot problem. (a) Obtained reward after the environmental change for 2000 steps. (b) First step of obtaining reward after the environmental change. Data are mean ± SEM <number of times of reward acquisition> of twenty repetitions. $^{*}p < 0.05$ is significantly better than the fixed parameters; $^{\dagger}p < 0.05$ is significantly better than the Schweighofer and Doya (2003) algorithm.

(a)

	fixed parameters	Schweighofer and Doya Algorithm	proposed method
Q-learning with CMAC	2485±370 < 18 >	*5462±378 < 20 >	†9494±532 < 20 >

(b)

	fixed parameters	Schweighofer and Doya Algorithm	proposed method
Q-learning with CMAC	1444±79 < 18 >	*545±53 < 20 >	†136±53 < 20 >

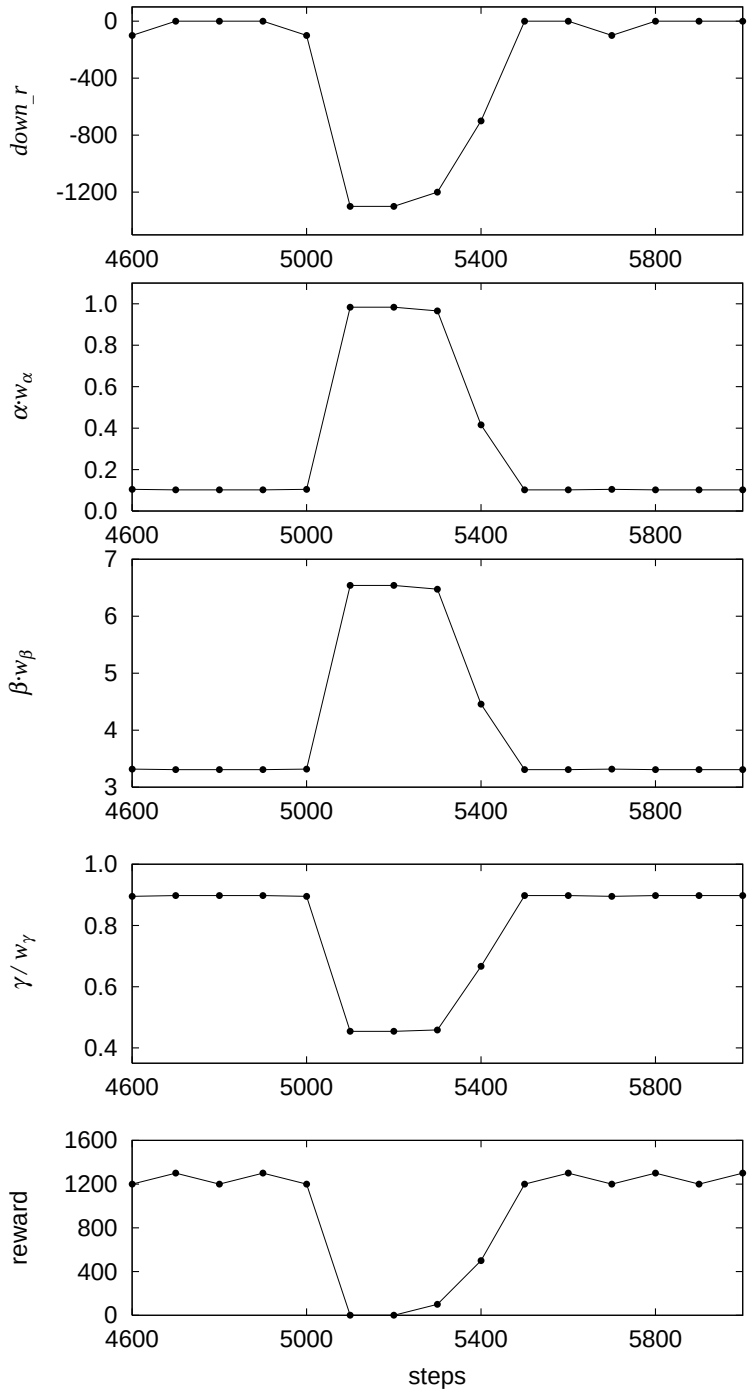
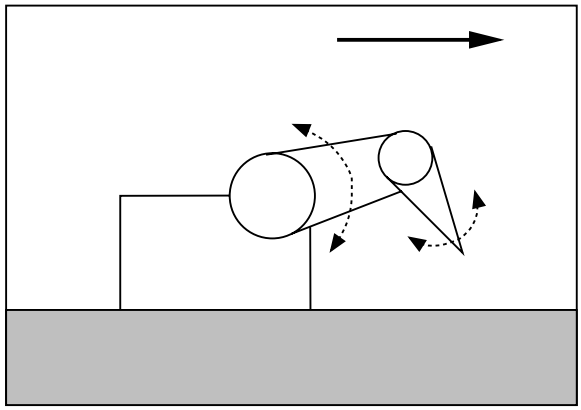
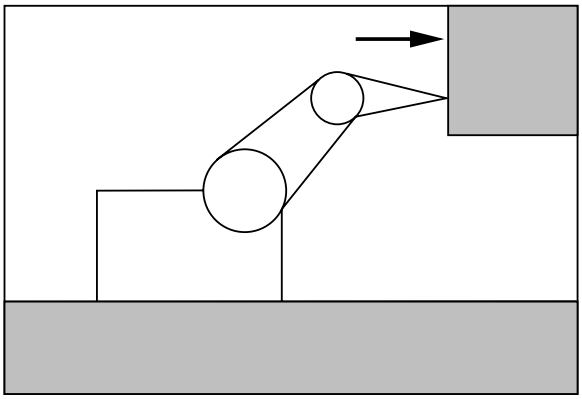


Fig. 3. An example of changes in *down_r*, three actual learning parameters ($\alpha \cdot w_\alpha$, γ/w_γ, and $\beta \cdot w_\beta$), and subtotal of reward for 100 steps in the maze problem. The environmental change occurred at 5000 steps.



(a)



(b)

Fig. 4. Two-linked arm robot moving forward problem. (a) Normal situation. (b) Entrance of tunnel as an unexpected environmental change.

fastest by a significant amount in all methods although that of the Schweighofer and Doya (2003) algorithm was significantly faster than that of the fixed parameters.

Figure 5 indicate an example of changes in $down\_r$, $\alpha \cdot w_\alpha$, γ / w_γ, and $\beta \cdot w_\beta$, and subtotal of reward for 100 steps around the environmental change in Q-learning by our method. Temporal traces of these values similar to that of the arm robot problem are seen.

4.3 Contribution of each parameter

In order to verify contribution of each parameter, we perform more experiments with various combination of fixed and variable parameters. We examine whether the proposed directions of changing parameters as in Section 2 is effective.

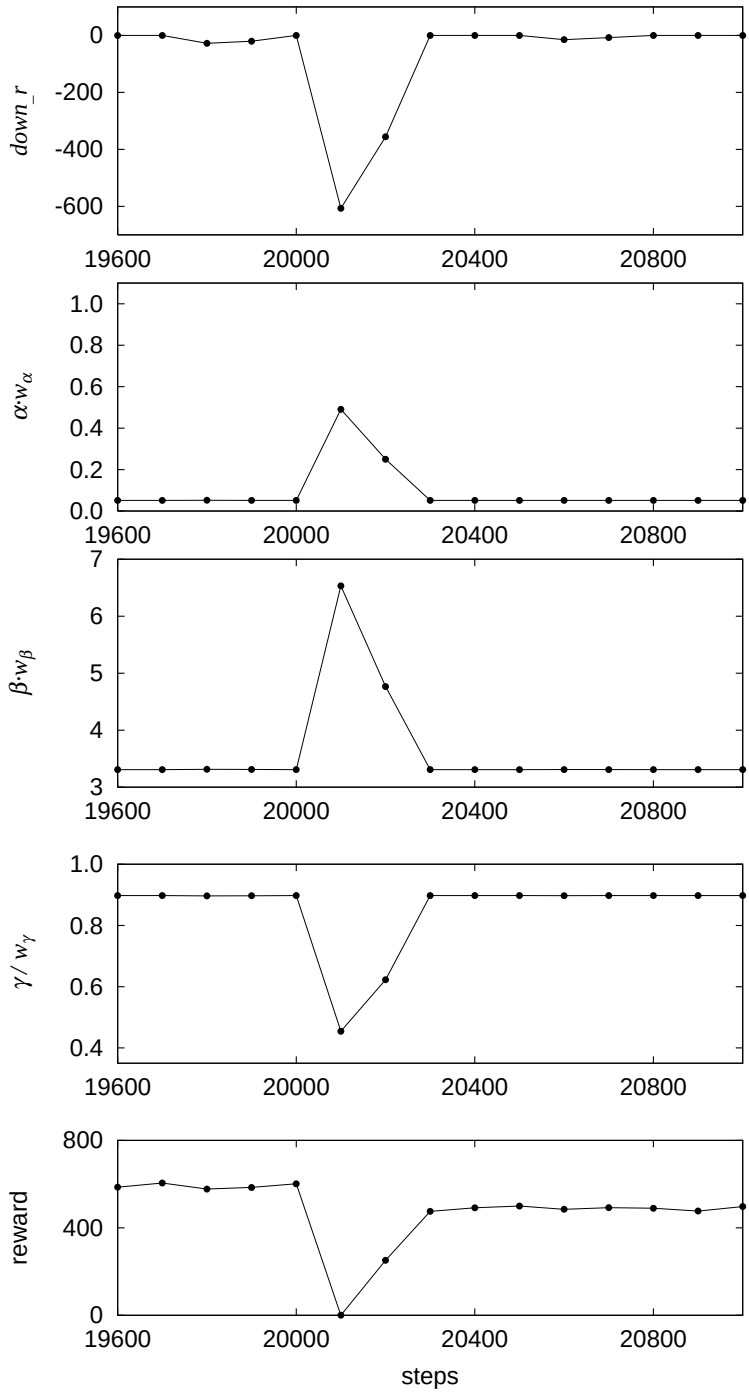


Fig. 5. An example of changes in *down_r*, three actual learning parameters ($\alpha \cdot w_\alpha$, γ/w_γ, and $\beta \cdot w_\beta$), and subtotal of reward for 100 steps in the arm robot problem. The environmental change occurred at 20000 steps.

Table 5 shows the mean ± SEM obtained reward after the environmental change for 2000 steps in the maze problem with various combination of fixed and variable parameters. $\bar{\alpha}$, $\bar{\beta}$, and $\bar{\gamma}$ mean that each parameter reversely changes against our proposed method, where only h_γ is restricted so that γ does not exceeded 1. In AC, it is shown that each parameter contributes to rapidly follow the unexpected environmental change. In Q-learning, the change of β in the opposite direction against our proposed method significantly disturb the relearning although the change of β in the direction with our proposed method does not significantly contribute to the relearning. These results support that our proposed directions of changing parameters as in Section 2 is effective.

Table 6 shows the mean ± SEM obtained reward after the environmental change for 2000 steps in the arm robot problem with various combination of fixed and variable parameters. α and γ obviously contribute. β does not disturb the relearning at least although β is not significantly effect. These results support that our proposed directions of changing parameters as in Section 2 is almost effective.

5 Conclusion

We propose a new simple parameter control method in reinforcement learning. This method has the feature that a robot using the method is able to rapidly follow unexpected environmental changes by controlling the parameters according to the changes of neuromodulators in an emergency.

In order to verify whether the proposed method is independent of learning algorithms and learning problems, we have experimented with the Q-learning and the AC for learning algorithms, and with the maze, in which they are discontinuous, and the two-linked arm robot, in which states and actions are continuous, for learning problems. Consequently, in each situation, the robots with our proposed method could respond more rapidly to the unexpected environmental changes than the robots with the other methods.

In order to investigate the validity of changing each learning parameter in our method, comparative experiments were additionally conducted. These results show that our decision on directions of changing parameters in an emergency is appropriate. The change only of inverse temperature parameter β, however, is not effective although the change of β with the changes of α and β is effective. This shows that β is closely related to other parameters α or γ. If our proposed directions of changing neuromodulators are effective also in physiology, the physiological observation of effect like this research on NA would be difficult.

In our method, as long as a certain behavior has succeeded in comparison with the past, a better behavior will not be able to be found although the behavior can be improved rapidly in a deteriorating situation. In other words, the exploration capability of our method in an emergency is accelerated instead of losing wide exploration capability when having succeeded.

The computer experiments show that the agents using our proposed method could rapidly respond to unexpected environmental changes, as long as we have tested it in some environments in this issue where we especially take notice of the robustness in reinforcement learning algorithms and types of learning problem. We will still need further verification on the various types of robustness.

Considering how the brain hardware realizes our proposed method is a future work. Our method, however, is effective for rapid action improvement in machine learning.

References

Albus, J. S. (1975). A new approach to manipulator control: the cerebellar model articulation controller (CMAC). *Transactions of the ASME, Journal of Dynamic Systems, Measurement, and Control*, *97*, 220–227.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*, 495–506.

Fadda, F., Cocco, S., & Stancampiano, R. (2000). Hippocampal acetylcholine release correlates with spatial learning performance in freely moving rats. *Neuroreport*, *11*, 2265–2269.

Ishii, S., Yoshida, W., & Yoshimoto, J. (2002). Control of exploitation-exploration meta-parameter in reinfocement learning. *Neural Networks*, *15*, 665–687.

Kimura, H. (2002). Hajime Kimura's official home page. Retrieved in 2002 from http://www.fe.dis.titech.ac.jp/˜gen/.

Kimura, H., & Kobayashi, S. (1997). Reinforcement learning for locomotion of a two-linked robot arm. *The 6th European Workshop on Learning Robots*, 144–153.

Schweighofer, N., & Doya, K. (2003). Meta-learning in reinforcement learning. *Neural Networks*, *16*, 5–9.

Sutton, R. S. (1996). Generalization in reinforcement learning: successful examples using sparse coarse coding. *Advances in Neural Information Processing System*, *8*, 1038–1044.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. The MIT Press.

Vankov, A., Hervé-Minvielle, A., & Sara, S. J. (1995). Response to novelty

and its rapid habituation in locus coeruleus neurons of the freely exploring rat. *European Journal of Neuroscience*, *7*, 1180–1187.

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

Wolff, M. C., & Leander, J. D. (2002). Selective serotonin reuptake inhibitors decrease impulsive behavior as measured by an adjusting delay procedure in the pigeon. *Neuropsychopharmacology*, *27*, 421–9.

Table 5
Results of the maze problem with various combination of fixed and variable parameters: obtained reward after the environmental change for 2000 steps. $\bar{\alpha}$, $\bar{\beta}$, and $\bar{\gamma}$ mean that each parameter changes perversely against our proposed method. Data are mean ± SEM <number of times of reward acquisition> of twenty repetitions. $^{*}p < 0.05$ is significantly better than fixed parameters; $^{\dagger}p < 0.05$ is significantly worse than proposed method.

	Q-learning	AC
no change (fixed parameters)	$0\pm0 < 0 >$	$0\pm0 < 0 >$
α	$^{*\dagger}10415\pm660 < 20 >$	$^{*\dagger}3795\pm830 < 20 >$
β	$^{\dagger}0\pm0 < 0 >$	$^{\dagger}0\pm0 < 0 >$
γ	$^{*\dagger}770\pm119 < 19 >$	$^{*\dagger}765\pm129 < 20 >$
α, γ	$^{*}17630\pm675 < 20 >$	$^{*\dagger}8295\pm1560 < 20 >$
α, β	$^{*\dagger}11405\pm486 < 20 >$	$^{*\dagger}2005\pm642 < 14 >$
β, γ	$^{*\dagger}940\pm184 < 20 >$	$^{*\dagger}115\pm25 < 14 >$
α, $\bar{\beta}$, γ	$^{*\dagger}15870\pm1069 < 20 >$	$^{*\dagger}295\pm58 < 18 >$
α, β, $\bar{\gamma}$	$^{\dagger}0\pm0 < 0 >$	$^{\dagger}0\pm0 < 0 >$
$\bar{\alpha}$, β, γ	$^{\dagger}0\pm0 < 0 >$	$^{\dagger}0\pm0 < 0 >$
α, β, γ (proposed method)	$18595\pm495 < 20 >$	$13490\pm1326 < 20 >$

Table 6
Results of the arm robot problem with various combination of fixed and variable parameters: obtained reward after the environmental change for 2000 steps. $\bar{\alpha}$, $\bar{\beta}$, $\bar{\gamma}$ mean that each parameter reversely changes with our proposed method. Data are mean ± SEM <number of times of reward acquisition> of twenty repetitions. $^{*}p < 0.05$ is significantly better than fixed parameters; $^{\dagger}p < 0.05$ is significantly worse than proposed method.

	Q-learning with CMAC
no change (fixed parameters)	$2485 \pm 370 < 18 >$
α	$^{*\dagger}7491 \pm 439 < 20 >$
β	$^{\dagger}505 \pm 225 < 5 >$
γ	$^{*\dagger}6593 \pm 262 < 20 >$
α, γ	$^{*}9473 \pm 430 < 20 >$
α, β	$^{*\dagger}5353 \pm 762 < 16 >$
β, γ	$^{*\dagger}6105 \pm 195 < 20 >$
α, $\bar{\beta}$, γ	$^{*}9033 \pm 510 < 20 >$
α, β, $\bar{\gamma}$	$^{\dagger}0 \pm 0 < 0 >$
$\bar{\alpha}$, β, γ	$^{\dagger}147 \pm 70 < 7 >$
α, β, γ (proposed method)	$9494 \pm 532 < 20 >$