The Sixteenth International Symposium on Artificial Life and Robotics 2011 (AROB 16th '11),
B-Con Plaza, Beppu,Oita, Japan, January 27-29, 2011

Reinforcement learning with the mechanism of short-term depression for learning rate

S. Kubota

Department of Biomedical Information Engineering, Yamagata University
4-3-16 Jonan, Yonezawa, Yamagata, 992-8510, Japan
(Tel : 81-238-26-3356; Fax : 81-238-26-3299)
(kubota@yz.yamagata-u.ac.jp)

Abstract: The temporal-difference (TD) reinforcement learning (RL), typically formulated in discrete state space, is frequently applied to the control problems represented by continuous state variables. The use of a coarse space discretization to describe the RL algorithm may degrade the control performance, whereas a fine discretization requires a large number of iteration steps to complete learning. In this study, I examine a novel RL algorithm by which the learning rate is dynamically and spatially modulated to improve the learning performance even when a relatively coarse space discretization is employed. This method is inspired by physiological phenomenon of short-term depression observed in biological synapses, and aims to produce a bias in the TD learning toward the states that have not been recently visited. The proposed algorithm is incorporated with Sarsa-lambda and tested in a nonlinear control task of swinging up a pendulum using limited torque. The simulation results show that the proposed dynamic learning rate can robustly reduce the number of trials required before accomplishing the task by facilitating efficient exploration in the RL process.

***Keywords*:** Temporal difference, Reinforcement learning, Dynamic learning rate, Synapse.

I. INTRODUCTION

The temporal-difference-(TD-)based reinforcement learning (RL) not only provides an efficient approach to control and decision problems, but also can be considered as an attractive model for studying the brain [1]. In fact, a broad range of experimental evidence suggests the involvement of neural activities occurring at a variety of brain areas in the mediation of reward processing as well as the TD error signal [1-3]. A close connection between the RL theory and the relevant physiological data implies that the theoretical scheme of RL could provide a quantitative framework for future studies in the related area of neuroscience [2]. Furthermore, it also appears possible to improve the RL algorithm by incorporating neurobiological mechanisms underlying adaptive behavior of animals, as this study aims to do.

An important application of RL technique is to control a strong nonlinear dynamical system, which would be difficult to be solved by a conventional engineering approach. In many interesting real-world control tasks, the state variables evolve with time in continuous space, although the progress of RL theory has been mainly restricted to the problems described by the Markov decision process (MDP) formulated in discrete state space [4]. Therefore, in the application studies of RL, the most common approach is first to discretize the state space and then to apply the learning algorithm described in a discrete stochastic system. In this approach, a fine discretization of state space necessarily requires a large number of memory storage as well as many iteration steps to complete learning the value functions. In contrast, a coarse space discretization leads to a situation where, when the state variables evolve slowly with time, an update of value functions takes place many times repeatedly at identical states; in such a case, it is likely that wrong actions repetitively taken at the same state are excessively reinforced, particularly during early learning phases.

Ideally, one would like to develop an algorithm that can improve the RL performance even when a relatively coarse space discretization is employed. In this study, I propose a novel method, which incorporates spatiotemporal modulation of learning rate to enhance the learning efficiency in such coarse discretized space. This method is inspired by physiological phenomenon of short-term depression (STD) observed in central synapses, i.e., a transient decrease in the strength of synaptic inputs following their repetitive activation. The proposed algorithm is introduced in Sarsa (λ) [5] and applied to a nonlinear control task of swinging up a pendulum with limited torque. I show in simulations that the task can be accomplished by employing the proposed dynamic learning-rate modulation in a number of trials less than by employing a conventional static learning rate. Further, I demonstrate that the proposed

method can decrease the sensitivity of the task performance to the scaling of learning rate, implying that this method can make the tuning of the learning parameter easier.

II. METHODS

1. Sarsa (λ)

As a basic RL algorithm to which the proposed method is to be incorporated, the author used Sarsa (λ) [5]. In this algorithm, the action-value function $Q(s,a)$, for each state s and action a, is updated as follows:

$$Q(s,a) \leftarrow Q(s,a) + \alpha e(s,a)\delta . \quad (1)$$

Here, α (>0) denotes the learning rate. δ represents the TD error described as

$$\delta = r + \gamma Q(s(t+1), a(t+1)) - Q(s(t), a(t)), \quad (2)$$

where r is the reward obtained by the state transition from $s(t)$ to $s(t+1)$ through action $a(t)$. γ is a discount rate satisfying $0 \le \gamma \le 1$. Each action is assumed to be decided from the ε-greedy policy with respect to $Q(s,a)$. The eligibility trace $e(s,a)$ for all the state-action pairs, *s*, *a*, is updated by the following equation:

$$e(s,a) \leftarrow \begin{cases} \gamma\lambda e(s,a) + 1, & (s = s(t) \text{ and } a = a(t)) \\ \gamma\lambda e(s,a). & (\text{otherwise}) \end{cases} \quad (3)$$

2. STD mechanism for controlling learning rate

The author designed an algorithm for spatiotemporal control of learning rate, which models short-term activity-dependent modification of synapses. In the central nervous system, the synaptic inputs that are activated with higher frequency are rapidly and temporarily depressed so that the input-output gain for high-rate inputs decreases [6]. This phenomenon, called STD, occurs through the transmitter vesicle depletion at the presynaptic terminals, and can be modeled by the following equation [6,7]:

$$dD(t)/dt = -\rho_v D(t) \sum_j \delta(t - t_j) + [1 - D(t)]/\tau_r , \quad (4)$$

where $D(t)$ ($0 \le D(t) \le 1$) represents the synaptic strength normalized by its maximum value, and t_j is the *j* th activation time of the synapse. Equation 4 indicates that the synaptic strength is weakened just following activation such that $D(t_j^+) = (1-\rho_v)D(t_j^-)$ ($0 \le \rho_v \le 1$), whereas it recovers toward 1 with the time constant τ_r in the absence of activation [7]. If time is discretized by using Euler's method, Eq. 4 can be written, by defining parameters $\rho \equiv 1 - \rho_v$ and $\mu \equiv \Delta t / \tau_r$ ($0 \le \rho, \mu \le 1$), as follows:

$$D(t+1) = \begin{cases} \rho D(t) + \mu[1 - D(t)], \\ \qquad (\text{following synaptic activation}) \\ D(t) + \mu[1 - D(t)]. \qquad (\text{otherwise}) \end{cases} \quad (5)$$

Assume that, based on the classical cell assembly hypothesis [8], firing activity of a group of interconnected neurons encodes a specific information regarding environment. Then, the repetition of the same state in MDP could be represented in the brain by the repetitive spiking of the same cell group, which will weaken input activity for such group of neurons through STD [9]. Further, experimental evidence suggests that the resultant weakened activation of NMDA receptors (NMDARs), one of major receptor subtypes for glutamatergic synapses, may cause the suppression of long-term plasticity underlying learning in the brain [10]. This may correspond, in the terminology of RL, to the decreased value of learning rate.

Therefore, to incorporate the mechanism resembling STD to an RL algorithm, the author introduces a function $d(s)$ ($0 \le d(s) \le 1$) and considers that $d(s)$ for all the states *s* are updated, similar to Eq. 5, as follows:

$$d(s) \leftarrow \begin{cases} \rho d(s) + \mu[1 - d(s)], & (\text{for } s = s(t)) \\ d(s) + \mu[1 - d(s)]. & (\text{otherwise}) \end{cases} \quad (6)$$

The value of α (Eq. 1) is multiplied by $d(s)$ so that the function Q is updated, instead of Eq. 1, as

$$Q(s,a) \leftarrow Q(s,a) + \alpha d(s) e(s,a)\delta . \quad (7)$$

According to Eq. 6, $d(s)$ for a given state *s* will decrease each time the state *s* is visited, whereas it gradually recovers to 1 in the absence of visiting *s*. Therefore, it can be expected that $d(s)$ acts to decrease the change in the action-value function $Q(s,a)$

Table 1. The learning parameters

Parameter	Value
Parameter to decide the level of STD ρ	0.6
Rate of recovery from STD μ	0.04
Scale of learning rate α	0.8
Discount rate γ	0.98
Decay rate of eligibility trace λ	0.8
Randomness of policy ε	0.1

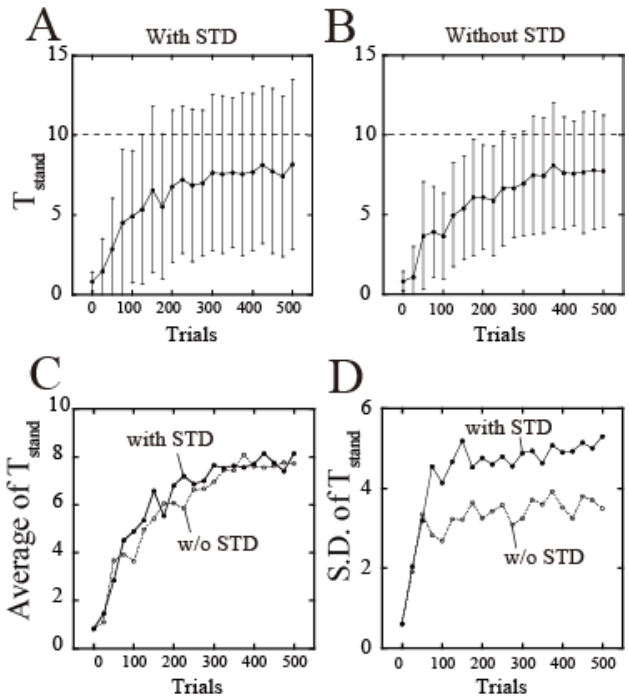

Fig.1. (A and B) The change in the average and standard deviation (error bar) of T_{stand} as a function of the number of trials in the presence (A; ρ = 0.6) and absence (B; ρ = 1) of the STD mechanism. The dashed lines show the level of T_{stand} above which a trial is considered successful ($T_{\text{stand}} > 10$). (C and D) The average (C) and standard deviation (D) of T_{stand} are plotted for both the cases with (solid) and without (dashed) STD in the same graph. The parameters for the system dynamics are as follows: m = l = 1, g = 9.8, c = 0.01, and $u_{\max}$ = 5 [4].

for the states that have been visited in the near past. This also implies that by the inclusion of the STD mechanism, the TD learning tends to proceed with a bias that assigns a higher weight to the states that have not been recently visited. The learning parameters in Table 1 are used unless otherwise stated.

III. RESULTS

To test the proposed algorithm, the author performed simulations for the control task of swinging up a pendulum with limited torque [4]. The dynamics of the pendulum are described as $d\theta/dt = \omega$ and $d\omega/dt = (-c\omega + mgl\sin\theta + u)/ml^2$ with the external torque $u = \pm u_{\max}$, implying that only its direction can be controlled. The physical parameters used are summarized in the Fig. 1 caption. For numerical integration, the Euler's method was used with the time step size Δt = 0.02. The reward given was set to be r = 1 for $|\theta| < \pi/4$, $r = -75/\Delta t$ for $|\theta| > 4\pi$, and r = 0 otherwise, where the negative reward for large θ is to prevent the over rotation. The state space $\{(\theta, \omega) \mid -4\pi < \theta, \omega < 4\pi\}$ was digitized into subspaces with the length of $\Delta\theta = \Delta\omega = \pi/3$. A trial ended at t = 20 or when the pendulum became over-rotated ($|\theta| > 4\pi$). To quantify the task performance, the total length of time at which the pendulum stands up ($|\theta| < \pi/4$) was defined as T_{stand}. The trial was considered to be successful when $T_{\text{stand}} > 10$, and the learning speed was measured by using the number of trials, $N_{success}$, required before achieving 10 successful trials.

Figure 1 shows the comparison of the time course of T_{stand} obtained by using and not using the STD mechanism of learning rate (Figs. 1A and 1B, respectively). Note that although the mean value of T_{stand} is largely the same regardless of the inclusion of STD (Fig. 1C), the temporal variation of T_{stand} becomes significantly increased by the STD function (Fig. 1D). This would be attributed to the fact that STD will contribute to assigning higher weight to the learning of 'novel' states that have not been recently visited, as mentioned above. The result here shows that the STD function will be effective to facilitate the exploration in RL almost without changing the average task performance.

When similar simulations were performed by using various values of ρ and α, the value of $N_{success}$ was found to take a minimum at an intermediate value of α for all ρ (Fig. 2A). Importantly, the minimum $N_{success}$ value for each ρ with ρ < 1 is considerably smaller than that for ρ = 1 (i.e., the case of no STD) (Fig. 2B), suggesting that the inclusion of STD can robustly improve the learning performance. To further explore the robustness of the outcome, the upper and lower limits of the α range, where the value of $N_{success}$ is less than a threshold (= 250), were defined as α_U and α_L, respectively, and the α_U/α_L ratio was plotted as a function of ρ (Fig. 2C). The figure shows that this ratio becomes greater than that obtained without STD when ρ is relatively large ($0.7 \le \rho \le 0.9$). This implies that in this range of ρ, STD can decrease the sensitivity of the task performance to the scale of

learning rate, which will make the tuning of the learning parameter easier. Note that around similar values of ρ ($0.5 \leq \rho \leq 0.9$), the minimum $N_{success}$ becomes quite small (Fig. 2B), indicating that this range of ρ will be near-optimal in that the task performance is both robustly and significantly enhanced.

IV. CONCLUSION

In this study, I have proposed an STD mechanism for TD-based RL, where the learning rate is spatiotemporally modulated, and have shown that this method can reduce the number of trials required before accomplishing the control task. This method is motivated by the physiological experimental findings on synapses that their activation are rapidly followed by depression [6]. The proposed learning rate modulation (represented by the change in $d(s)$) tends to assign a greater weight to the learning of value functions for the states that have not been recently visited. This appears to be similar to the observed response of dopamine neurons, which can reflect the novelty of the presented stimuli [11]. Therefore, given that the TD error δ closely resembles the dopamine cell response representing the unpredictability of reward [11], the term $d(s)\delta$ (Eq. 7) may correspond to the dopamine signal that encodes both the novelty and unpredictability of the reward. The present study may suggest a new framework of RL research in which the functional significance of a biological mechanism can be examined by constructing an RL algorithm that incorporates the mechanism and testing it through the application to control problems.

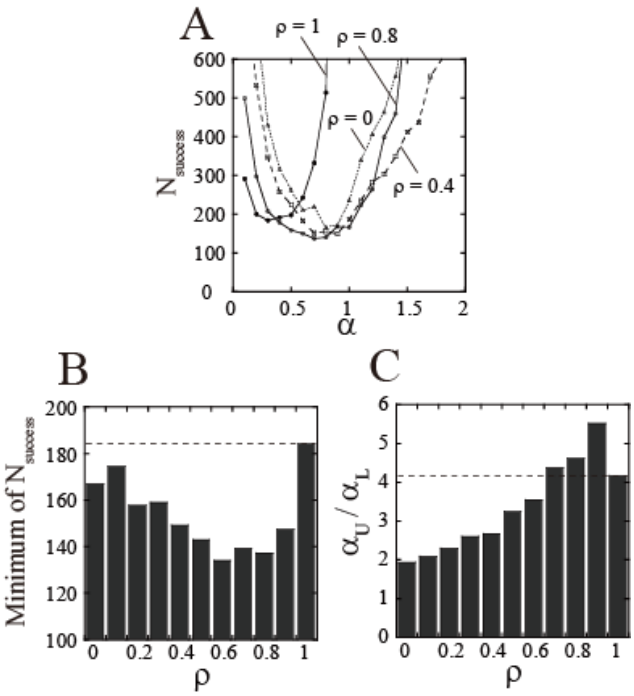

Fig.2. The comparison of task performance for various values of α and ρ. (A) The change in the value of N_{success}, which is averaged over 32 cases, as a function of α and ρ. (B and C) The minimum N_{success} value (B) and the α_U / α_L ratio (C), which are obtained by using various values of α, are plotted for each ρ. The cases of ρ = 1 (denoted by the horizontal dashed lines) correspond to those without the STD mechanism.

REFERENCES

[1] Doya K (2008), Modulators of decision making. Nature Neuroscience 11:410-416.
[2] Schultz W, Dayan P, Montague R (1997), A neural substrate of prediction and reward. Science 275:1593-1599.
[3] Schultz W, Apicella P, Ljungberg T (1993), Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. Journal of Neuroscience 13: 900-913.
[4] Doya K (2000), Reinforcement learning in continuous time and space. Neural Computation 12:219-245.
[5] Sutton R. S., Barto A. G. (1998) Reinforcement learning: An introduction. The MIT Press, Cambridge.
[6] Abbott L. F., Varela J. A., Sen K., et al. (1997) Synaptic depression and cortical gain control. Science 275: 220-224.
[7] Wang X.J. (1999) Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. Journal of Neuroscience 19: 9587-9603.
[8] Lansner A. (2009) Associative memory models: from the cell-assembly theory to biophysically detailed cortex simulations. Trends in Neuroscience 32:178-186.
[9] Zucker R. S. (1999) Calcium- and activity-dependent synaptic plasticity. Current Opinion in Neurobiology 9:305-313.
[10] Bear M. F. (1996) A synaptic basis for memory storage in the cerebral cortex. Proceedings of the National Academy of Sciences of the United States of America, 93:13453-13459.
[11] Shultz W. (1998) Predictive reward signal of dopamine neurons. Journal of Neurophysiology, 80: 1-27.