A Meta-learning Method Based on Temporal Difference Error

Kunikazu Kobayashi, Hiroyuki Mizoue, Takashi Kuremoto, and Masanao Obayashi

Yamaguchi University, Tokiwadai 2-16-1, Ube, Yamaguchi 755-8611, Japan
{koba,wu,m.obayas}@yamaguchi-u.ac.jp
http://www.nn.csse.yamaguchi-u.ac.jp/k/

Abstract. In general, meta-parameters in a reinforcement learning system, such as a learning rate and a discount rate, are empirically determined and fixed during learning. When an external environment is therefore changed, the sytem cannot adapt itself to the variation. Meanwhile, it is suggested that the biological brain might conduct reinforcement learning and adapt itself to the external environment by controlling neuromodulators corresponding to the meta-parameters. In the present paper, based on the above suggestion, a method to adjust meta-parameters using a temporal difference (TD) error is proposed. Through various computer simulations using a maze search problem and an inverted pendulum control problem, it is verified that the proposed method could appropriately adjust meta-parameters according to the variation of the external environment.

Keywords: reinforcement learning, meta-parameter, meta-learning, TD-error, maze search problem, inverted pendulum control problem.

1 Introduction

Reinforcement learning is a famous model of animal learning [1]. Schultz et al. found that dopamine neurons in the basal ganglia have the formal characteristics of the teaching signal known as the temporal difference (TD) error through an animal experiment [2].

In this context, Doya proposed the hypotheses between meta-parameters in reinforcement learning and neuromodulators in the basal ganglia based on the review of experimental data and theoretical models. That is, he presented that dopamine signals the error in reward prediction, serotonin controls the time scale of reward prediction, noradrenaline controls the randomness in action selection, and acetylcholine controls the speed of memory update [3]. Successful reinforcement learning highly depends on the careful setting of meta-parameters in reinforcement learning. Schweighofer et al. proposed a meta-learning method based on rewards, which not only finds appropriate meta-parameters but also controls the time course of these meta-parameters in an adaptive manner [4]. However, their method has some parameters to be pre-determined.

In the present paper, we propose a meta-learning method based on a TD-error. The proposed method has only one parameter to be pre-determined and is easy to apply to reinforcement learning. Through various computer simulations using a maze search problem and an inverted pendulum control problem, it is verified that the proposed

C.S. Leung, M. Lee, and J.H. Chan (Eds.): ICONIP 2009, Part I, LNCS 5863, pp. 530–537, 2009.

method allows meta-parameters to be appropriately adjusted according to the variation of the external environment.

2 Reinforcement Learning

Reinforcement learning is a method that agents acquire the optimum behavior with a repeating process of exploration and exploitation by being given rewards in an environment as a compensation for its behavior [1]. In this section, we explain two kinds of temporal difference (TD) learning, i.e. a Q-learning method [5] and an actor-critic method [1] and also describe a policy in reinforcement learning.

2.1 Q-Learning Method

The Q-learning method guarantees that every state converges to the optimal solution by appropriately adjusting a learning rate in an MDP environment [5]. The state-action value function $Q(s(t), a(t))$ for a state $s(t)$ at time t and an action $a(t)$ at time t is updated so as to take the optimal action by exploring it in a learning space and defined as follows:

$$Q(s(t), a(t)) \leftarrow Q(s(t), a(t)) + \alpha\delta(t), \tag{1}$$

$$\delta(t) = r(t) + \gamma \max_{b \in A} Q(s(t+1), b) - Q(s(t), a(t)), \tag{2}$$

where $\delta(t)$ and $r(t)$ denote a TD-error at time t and a reward at time t, respectively, meta-parameters α and γ refer to a learning rate and a discount rate, respectively, and A is a set of actions to be taken.

2.2 Actor-Critic Method

The actor-critic method has a separate memory structure to explicitly represent the policy independent of the value function [1]. The policy structure is known as the actor, because it is used to select actions, and the estimated value function is known as the critic, because it criticizes the actions made by the actor. The values in actor-critic method are updated as follows:

$$V(s(t)) \leftarrow V(s(t)) + \alpha\delta(t), \tag{3}$$

$$Q(s(t), a(t)) \leftarrow Q(s(t), a(t)) + \alpha\delta(t), \tag{4}$$

$$\delta(t) = r(t) + \gamma V(s(t+1)) - V(s(t)), \tag{5}$$

where $V(s(t))$ is a value function for a state $s(t)$ at time t.

2.3 Policy

The policy is a mapping from the states in an external environment to the actions to take in those states. Throughout the present paper, we suppose that an action is selected by the Boltzmann distribution. That is, the policy $\pi(s(t), a(t))$ is defined as follows [1]:

$$\pi(s(t), a(t)) = \frac{\exp\left(Q(s(t), a(t))/T\right)}{\sum_{b \in A} \exp\left(Q(s(t), b)/T\right)}, \tag{6}$$

where T refers to a temperature parameter. The policy realizes a random selection if $T \to \infty$ and is a greedy selection if $T \to 0$.

3 Meta-learning

Generally speaking, it is crucial that all the meta-parameters such as a learning rate and a discount rate are carefully tuned to elicit good performance in advance. It therefore is beneficial that the meta-parameters could be changed according to the situation. In this section, we describe the conventional meta-learning method based on rewards (Section 3.1) and propose a new meta-learning method based on the TD-error (Section 3.2).

3.1 Meta-learning Based on Reward

Schweighofer et al. proposed the meta-learning method based on mid-term and long-term rewards [4]. In their method, the mid-term reward $r_{MT}(t)$ at time t and the long-term reward $r_{LT}(t)$ at time t are defined as follows:

$$r_{MT}(t) = \left(1 - \frac{1}{\tau_{MT}}\right) r_{MT}(t-1) + r(t), \tag{7}$$

$$r_{LT}(t) = \left(1 - \frac{1}{\tau_{LT}}\right) r_{LT}(t-1) + r_{MT}(t), \tag{8}$$

where $r(t)$ refers to an instant reward at time t, τ_{MT} and τ_{LT} denote the time constants for $r_{MT}(t)$ and $r_{LT}(t)$, respectively. In the present paper, we assume that $\tau_{MT} < \tau_{LT}$ and $r_{MT}(0) = r_{LT}(0) = 0$. If an agent tends to take the desired actions than before, then the mid-term reward has a larger value than the long-term one. If not so, the mid-term reward has a smaller value than the long-term one. The conventional method updates meta-parameters such as a discount rate, a learning rate, and a temperature parameter using this characteristic.

A discount rate is defined as a function of time t.

$$\gamma(t) = 1 - e^{-\epsilon(t)}, \tag{9}$$

where $\epsilon(t)$ refers to a variable represented by the rewards and an exploration noise $\sigma(t)$, and is defined as follows:

$$\epsilon(t) = \epsilon'(t) + \sigma(t). \tag{10}$$

In (10), $\epsilon'(t)$ is a variable depending on the rewards and updated by the follwing equation ($\epsilon'(0) = 0$).

$$\epsilon'(t) = \epsilon'(t-1) + \mu \left\{ r_{MT}(t) - r_{LT}(t) \right\} \sigma(t), \tag{11}$$

where μ represents an updating rate for $\epsilon'(t)$.

Although the other meta-parameters, i.e. a learning rate $\alpha(t)$ and a temperature parameter $T(t)$, are also updated like the above, the literature [4] does not present their updating rules at all. In the present paper, we therefore propose the updating rules for $\alpha(t)$ and $T(t)$. At first, we propose that $\alpha(t)$ and $T(t)$ are defined as follows:

$$\alpha(t) = e^{-\epsilon(t)} \tag{12}$$

$$T(t) = \frac{1}{e^{\epsilon(t)} - 1} \tag{13}$$

From (10), the value of $\epsilon(t)$ increases if $r_{MT}(t)$ is larger than $r_{LT}(t)$, decreases if $r_{MT}(t) < r_{LT}(t)$, and has no change if $r_{MT}(t) \approx r_{LT}(t)$. The $\alpha(t)$ and $T(t)$ therefore increases if $r_{MT}(t) > r_{LT}(t)$ because of the growth of $\epsilon(t)$ and decreases if not so. This corresponds that $\alpha(t)$ and $T(t)$ should take large values if learning is required, e.g. in the beginning of learning or when the external environment has changed, and they should take small values if not so.

In the present paper, it is assumed that the conventional method includes the two proposed equations (12) and (13) besides the conventional equation (9) in the literature [4]. Note that the conventional method does not completely correspond with Schweighofer et al.'s method.

3.2 Meta-learning Based on TD-Error

The optimal values of meta-parameters in reinforcement learning might change according to the progress of learning. We therefore focus on the TD-error which could be changed by the progress of learning and propose a meta-learning method based on the TD-error. We define a variable $\delta'(t)$ which depends on the absolute value of the TD-error. Then, meta-parameters are updated based on it. The $\delta'(t)$ is defined as follows ($\delta'(0) = 0$):

$$\delta'(t) = \left(1 - \frac{1}{\tau}\right)\delta'(t-1) + \frac{1}{\tau}|\delta(t)|, \tag{14}$$

where $\delta(t)$ refers to a TD-error at time t and τ is a time constant.

A learning rate $\alpha(t)$ and a temperature parameter $T(t)$ are expected to be a large value for exploration in the beginning of learning. On the other hand, their values are desired to be small for exploitation at the matured stage of learning. In addition, if relearning is required according to an environmental change, a discount rate $\gamma(t)$ are expected to be a small value. On the other hand, the value of $\delta'(t)$ becomes large if relearning is required because of the environmental change and converges to 0 at the matured stage. The meta-parameters $\alpha(t)$, $\gamma(t)$, and $T(t)$ at time t are therefore defined based on $\delta'(t)$ as follows:

$$\alpha(t) = \frac{2}{1 + e^{-\delta'(t)}} - 1, \tag{15}$$

$$\gamma(t) = \frac{2}{1 + e^{\delta'(t)}}, \tag{16}$$

$$T(t) = e^{\delta'(t)} - 1. \tag{17}$$

Based on the above equations, $\alpha(t)$ and $T(t)$ becomes small according to the decrease of $\delta'(t)$ and becomes large according to the increase of $\delta'(t)$. This allows that meta-parameters are appropriately updated in accordance with the change of $\delta'(t)$.

4 Computer Simulation

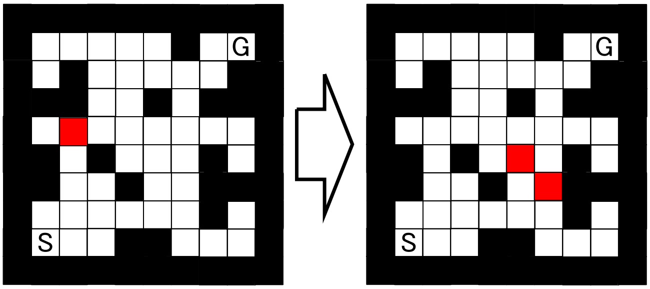
The performance of the proposed method is verified through computer simulation. In the simulation, we prepare two reinforcement learning tasks, i.e. a maze search problem and an inverted pendulum control problem. Accordingly, we apply the proposed method to the Q-learning method (Section 2.1) and the actor-critic method (Section 2.2). The proposed method applied to the Q-learning system is compared with the conventional method (Section 3.1) and the Q-learning without meta-learning (Section 2.1). Then, our method applied to the actor-critic system is compared with the conventional method and the actor-critic method without meta-learning (Section 2.2).

4.1 Maze Search Problem

To evaluate the performance for the discrete task, we use the maze search problem. In this simulation, we apply the proposed method to the Q-learning system.

Simulation Setting. Figure 1 shows a maze used in the simulation. In this figure, the black and gray squares correspond to walls and the white squares correspond to paths. The structure of the maze in Fig.1(a) will change to that in Fig.1(b) at 301 episodes. Namely, one gray square and two white squares are changed to a white one and two gray ones, respectively. The shortest path for both mazes is 14 steps. An agent is able to perceive only the adjacent eight squares. In this task, the Markov property is guaranteed because there is no aliasing problem. But, since the structure of the maze is changed, meta-parameters should be adjusted again. In this simulation, one episode is assumed that an agent starts from a start point and arrives at a goal point. The failure of maze search is defined when an agent cannot arrive at a goal point within 10,000 steps.

The parameters to be pre-determined are as follows: In the conventional method, time constant in (7) and (8): $\tau_{MT} = \tau_{LT} = 300$, updating rate in (11): $\mu = 0.1$. In the proposed method, time constant in (14): $\tau = 300$. The exploration noise $\sigma(t)$ in (10) is set as the Gaussian distribution with mean 0 and variance 1. In the standard method with fixed meta-parameters, we set as $\alpha = 0.2$, $\gamma = 0.95$, and $T = 0.3$.



(a) Maze from the 1st to 300th step

(b) Maze over the 301th step

Fig. 1. Dynamical structure change in the maze search problem

Table 1. The total number of steps in the maze search problem

Method	From 1 to 300 episodes	From 1 to 600 episodes
Proposed method	9,855	17,622
Conventional method	9,922	23,226
Q-learning	10,664	—

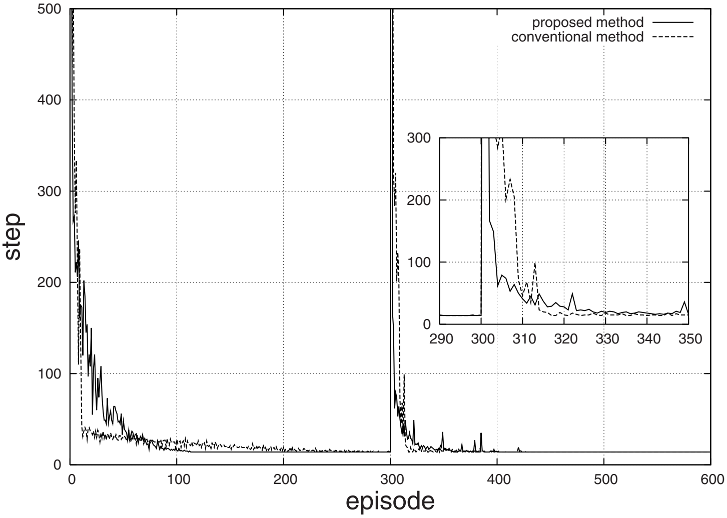


Fig. 2. Development of the number of steps for the conventional and proposed methods in the maze search problem

Simulation Result. Table 1 shows the total number of steps and Fig.2 illustrates the development of the number of steps. From these results, the proposed method could adjust to the structure change of the maze. That is, although the number of steps increases significantly when the structure is changed at 301 steps, both the proposed and conventional methods can find a new shortest path but the Q-learning method without meta-learning cannot find it. Furthermore, as seen from Table 1, the proposed method takes smaller steps than the conventional method.

4.2 Inverted Pendulum Control Problem

To evaluate the performance for the continuous task, we use the inverted pendulum control problem. In this simulation, we apply the proposed method to the actor-critic system.

Simulation Setting. Let θ, $\dot{\theta}(= d\theta/dt)$, and τ_c be angle, angular velocity, and torque, respectively. The dynamics of the inverted pendulum is represented by

$$ml^2\ddot{\theta} = -mgl\sin(\theta) - \mu\dot{\theta} + \tau_c, \tag{18}$$

where m denotes the mass of the pendulum, l is the length of the pendulum, g is the acceleration of gravity, μ is the friction coefficient of axis. In the simulation, we set $m = 0.5[kg]$, $l = 0.5[m]$, $g = 9.8[m/s^2]$, and $\mu = 0.1$. Then, reward $r(t)$ is defined as follows:

$$r(t) = \cos(\theta) - 0.005\,\tau_c^2(t). \tag{19}$$

At the initial state, we set $\theta(0) = 0$ and $\dot{\theta}(t) = 0$. We can only observe θ and set the torque to the pendulum as -1 or $+1$. The success of control is assumed that the pendulum can be controled within $\pm 5[deg]$ for $15[s]$. Then, the trial that the angle of the pendulum exceeds $\pm 45[deg]$ is assumed to be failure and go to the next trial. A time step is set as $\Delta t = 0.01[s]$. The success rate of controlling the pendulum is calculated as the average of 100 trials after $1,000$ training episodes.

The parameters are set as $\tau_{MT} = \tau_{LT} = 300$ and $\mu = 0.1$ in the conventional method, and $\tau = 300$ in the proposed method. The exploration noise is set as the Gaussian distribution with mean 0 and variance 1. In the standard method with fixed meta-parameters, we set as $\alpha = 0.2$, $\gamma = 0.95$, and $T = 0.35$.

Simulation Result. As seen in Table 2, the success rate is 87.9% in the proposed method, 85.2% in the conventional method, and 76.8% in Q-learning method. As a result, it is shown that the proposed method could improve the performance for controlling the pendulum. Figure 3 shows the temporal development of the accumulated

Table 2. The success rate of controlling an inverted pendulum

Method	Success rate (%)
Proposed method	87.9
Conventional method	85.2
Actor-critic method	76.8

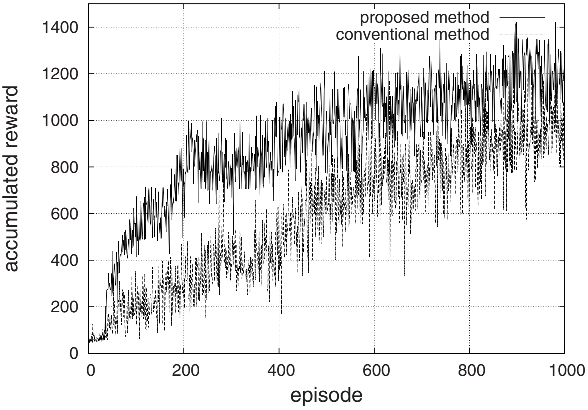


Fig. 3. Development of the accumulated reward for the proposed and conventional methods in the inverted pendulum control problem

reward. From this figure, it is clear that the accumulated reward in the proposed method is much larger than that in the conventional method. In addition, the number of parameters to be pre-determined is only one, i.e. τ in (14) in the proposed method. On the other hand, there are three such parameters, i.e. τ_{MT} in (7), τ_{LT} in (8), and μ in (11) in the conventional method. It therefore is much easier for the proposed method to apply the reinforcement learning system than the conventional method.

5 Summary

The present paper have proposed the meta-learning method based on the TD-error. Through various computer simulations, we investigated the performance of the proposed method using the discrete and continuous tasks. As a result, it is shown that the proposed method applied to the Q-learning system could improve the learning performance for the discrete task, the maze search problem because it allows meta-parameters to adjust according to the variation of the external environment. In addition, it is clarified that the proposed method applied to the actor-critic system could improve the control performance for the continuous task, the inverted pendulum control problem. Furthermore, it is shown that the proposed method could easily apply to reinforcement learning compared with the conventional method, the standard Q-learning and actor-critic methods because the proposed method can reduce the number of parameters to be pre-determined. In future work, the proposed method under a noisy environment needs to be evaluated.

Acknowledgments. This work was partly supported by three Grant-in-Aid projects for Scientific Research (No.20500207 and 20500277) from MEXT, Japan.

References

1. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998)
2. Schultz, W., Dayan, P., Montague, P.R.: A Neural Substrate of Prediction and Reward. Science 275, 1593–1599 (1997)
3. Doya, K.: Metalearning and Neuromodulation. Neural Networks 15, 495–506 (2002)
4. Schweighofer, N., Doya, K.: Meta-learning in Reinforcement Learning. Neural Networks 16(1), 5–9 (2003)
5. Watkins, C.J.C.H., Dayan, P.: Q-learning. Machine Learning 8(3-4), 279–292 (1992)
6. Ishii, S., Yoshida, W., Yoshimoto, J.: Control of Exploitation-Exploration Meta-parameter in Reinforcement Learning. Neural Networks 15(4-6), 665–687 (2002)