PERGAMON

Neural Networks 15 (2002) 665–687

Neural Networks

www.elsevier.com/locate/neunet

2002 Special issue

Control of exploitation–exploration meta-parameter in reinforcement learning

Shin Ishii[a,b,*], Wako Yoshida[a,b], Junichiro Yoshimoto[a,b]

[a]*Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0101, Japan*
[b]*CREST, Japan Science and Technology Corporation, Japan*

Received 10 October 2001; accepted 16 April 2002

Abstract

In reinforcement learning (RL), the duality between exploitation and exploration has long been an important issue. This paper presents a new method that controls the balance between exploitation and exploration. Our learning scheme is based on model-based RL, in which the Bayes inference with forgetting effect estimates the state-transition probability of the environment. The balance parameter, which corresponds to the randomness in action selection, is controlled based on variation of action results and perception of environmental change. When applied to maze tasks, our method successfully obtains good controls by adapting to environmental changes. Recently, Usher et al. [Science 283 (1999) 549] has suggested that noradrenergic neurons in the locus coeruleus may control the exploitation–exploration balance in a real brain and that the balance may correspond to the level of animal's selective attention. According to this scenario, we also discuss a possible implementation in the brain.

Keywords: Reinforcement learning; Exploitation–exploration problem; Neuromodulator; Attention; Partially observable Markov decision process

1. Introduction

Reinforcement learning (RL) (Sutton & Barto, 1998) is a learning framework in order to adapt to an environment based on trial and error. This paper discusses an RL scheme for dynamic environments, i.e. environments that change with time. Conventional RL schemes are formulated in terms of Markov decision process (MDP), that is, a decision-making problem or an optimal control problem in a stochastic but static environment. Since an optimal control problem in a dynamic environment can approximately be formulated as an MDP when RL is faster than the environmental change, this study also adopts that approximation. In addition, we also use a formulation of partially observable Markov decision process (POMDP). A POMDP assumes that the environment involves unobservable information, typically, unobservable state variables.

Although RL is a machine learning framework, recent studies (Schultz, Dayan, & Montague, 1997; Waelti, Dickinson, & Schultz, 2001) showed that in a real brain a dopaminergic system including the basal ganglia and the frontal cortex seems to realize a similar learning scheme. Doya (2000b) has suggested that parameters used in RL, which are called 'meta-parameters', may correspond to neuromodulators such as serotonin, noradrenaline and acetylcholine. Thus, the motivation of our study is not only on the machine learning but also on the brain learning.

In RL, an agent is provided by the environment with a scalar reward corresponding to a behavior (action) for each sensory state. The reward indicates instantaneous goodness of the action at the state. The objective of the agent is to maximize the rewards accumulated toward the future, and the maximization is done by improving its strategy to select an action for each state. Such a strategy is called a policy. The estimation and prediction of the accumulated rewards are important for improving the policy. Therefore, a standard RL scheme estimates the reward accumulation which is called the value function.

In order to make a good prediction, it is important to know the dynamics of the environment, i.e. how the current state changes by an action. Model-free RL methods like the actor–critic learning (Barto, Sutton, & Anderson, 1983) and the Q-learning (Watkins & Dayan, 1992) require no model of the environmental dynamics; instead, they try to directly estimate the value function. In contrast, model-based RL methods (Dayan & Sejnowski, 1996; Dearden, Friedman, & Andre, 1999; Doya, 2000a; Matsuno, Yamazaki, Matsuda,

* Corresponding author. Tel.: +81-743-72-5980; fax: +81-743-72-5989.
E-mail address: ishii@is.aist-nara.ac.jp (S. Ishii).

0893-6080/02/$ - see front matter
PII: S0893-6080(02)00056-4

& Ishii, 2001; Moore & Atkeson, 1993; Sutton, 1990) try to model the environmental dynamics and the value function is approximated using the model. Especially when the environment is complicated, e.g. partially observable, a model-based RL has an advantage, because the environmental model can explicitly deal with the complexity. A model-based RL learns faster than a model-free alternate. Our study presents a model-based RL method using the Bayes inference.

If the agent knows the correct optimal value function including the correct estimation of the environmental dynamics, the optimal policy is the one that just selects a 'greedy' action that maximizes the value function at each state. If the estimation and prediction are fairly good, therefore, a good policy is the one that selects a greedy action; this is called exploitation. During the process of trial and error, however, the agent does not know the correct optimal value function. Especially in a POMDP, an approximated value function may be apart from the correct optimal one, due to the uncertain estimation of unobservable state variables. In such a situation, the greedy action is not necessarily optimal. In addition, when the environment changes with time, the value function approximated using the past experiences will not be optimal. In order to know the optimal value function, the agent should execute trial actions, i.e. actions that are not optimal with respect to the current value function; this is called exploration. Since these two strategies, exploitation and exploration, cannot be operated at once, their control has long been an important issue in the control fields (Fe'ldbaum, 1965).

Methods for exploration can roughly be classified into two: undirected exploration methods and directed exploration methods (Thrun, 1992). Undirected exploration methods try to explore the whole state–action space by assigning positive probabilities to all possible actions. For example, semi-uniform (ϵ-greedy) exploration and the Boltzmann exploration (Sutton & Barto, 1998) are undirected methods.

Directed exploration methods use the statistics obtained through the past experiences in order to execute efficient exploration. Kearns and Singh (1998) proposed an exploration algorithm called E^3 algorithm, in which states were classified into known or unknown states based on the visit number. At a known state the agent executes directed exploration under a specific condition, while at an unknown state the agent mainly executes undirected exploration. R-max algorithm by Brafman and Tennenholtz (2001) is a modification of the E^3 algorithm so that the agent executes directed 'optimistic' exploration at an unknown state.

Exploration bonus is one popular technique for directed exploration. In Sutton's DYNA system (Sutton, 1990), exploration bonus is added to the immediate reward based on the time period that has passed since the state–action pair was previously experienced. Kaelbling (1993) proposed the interval estimation algorithm using exploration bonus based on the upper bound of the confidence interval for the value function. Moore and Atkeson (1993) also proposed exploration bonus in their learning algorithm called prioritized sweeping. In this method, an unfamiliar state is connected to a fictitious absorbing state with a high value and the agent is encouraged to visit such unfamiliar states. In the method by Dayan and Sejnowski (1996), due to the forgetting effect of the environmental dynamics, the agent comes to try an action that is not optimal with respect to the current estimation of the value function.

We discuss in this paper a new control method of the exploitation–exploration balance. The balance control was also studied by Thrun (1992). Our method is mainly an undirected method, in which the balancing parameter is controlled depending on the current state. Our method also uses exploration bonus. Usher, Cohen, Servan-Schreiber, Rajkowski, and Aston-Jones (1999) has suggested that the exploitation–exploration balance in a real brain may be controlled by noradrenergic neurons in the locus coeruleus (LC) and that the balance may correspond to the level of animal's selective attention. According to this scenario, we will discuss a possible implementation in the brain, which realizes our learning scheme.

Section 2 describes preliminaries to the RL. We propose in Section 3 a Bayes inference method for identifying the current environment. We next propose in Section 4 a control method of the exploitation–exploration balance. An exploration bonus is also introduced in the same section. Section 5 shows computer simulation results. Section 6 discusses a possible implementation in the brain, and Section 7 concludes the paper.

2. Reinforcement learning preliminaries

2.1. Markov decision process

We first consider Markov environments; $P(s'|s,a)$ gives the probability of reaching state s' by selecting action a at state s. If the state-transition probability $P(s'|s,a)$ is known, the value function for state s, $V(s)$, should satisfy the following (optimal) Bellman's equation:

$$V(s) = \max_a Q(s,a), \tag{1a}$$

$$Q(s,a) \equiv r(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s'). \tag{1b}$$

$Q(s,a)$ is often called the action-value function. The reward function $r(s,a)$ defines the immediate reward for a state–action pair (s,a). The reward function is assumed to be deterministic for simplicity, although the extension to stochastic reward functions is straightforward. $0 \leq \gamma \leq 1$ is a discount constant. The value function defines the summation of the discounted rewards accumulated toward the future. The action-value function $Q(s,a)$ represents the reward accumulation when the agent takes action a at state s and the optimal actions at the subsequent states.

A policy is a function that outputs an action for a given state s; a policy is called optimal when it outputs the action maximizing $Q(s,a)$. The objective of RL, which is often termed MDP, is to obtain the optimal policy. When the state-transition probability $P(s'|s,a)$ is known, this problem can be solved by a dynamic programming approach. In many RL problems, however, the state-transition probability is unknown. Temporal-difference (TD) learning (Sutton, 1988) tries to approximate the value function based on the agent's experiences without directly modeling the environment; it is a model-free approach. The actor–critic learning (Barto et al., 1983) and the Q-learning (Watkins & Dayan, 1992) are model-free TD learning. TD learning makes use of the so-called TD-error:

$$\delta = (r(s,a) + \gamma V(s')) - V(s). \tag{2}$$

The second term is the value of state s based on the present prediction, while the first term is the value of state–action pair (s,a) using one-ply actual state transition to s'; the TD-error is the difference between them. TD learning tries to approximate the value function by decreasing probabilistically the TD-error based on a stochastic approximation method (Sutton, 1988). Then, the state-transition probability is indirectly obtained.

On the other hand, model-based RL (Dayan & Sejnowski, 1996; Doya, 2000a; Moore & Atkeson, 1993; Sutton, 1990) tries to directly model the environment by approximating the state-transition probability $P(s'|s,a)$ based on experiences of past state transitions. The model-based RL is suitable for dealing with partially observable environments and/or dynamic environments. It is also suitable for multiagents environments (Matsuno et al., 2001). In the model-based RL, the learning of the value function and the learning of the environmental model are conducted concurrently but independently.

2.2. Partially observable Markov decision process

Since the problems considered in our study can be formulated as POMDPs (Kaelbling, Littman, & Cassandra, 1998), we explain the notion.

A typical POMDP deals with a Markov environment with unobservable (hidden) state variables. Let $s \equiv (y,z)$ be an environmental state, where y and z denote observable and unobservable state variables, respectively. Due to the unobservable variables, the environment with respect to the observable variables does not have a Markov property. Although standard RL algorithms have often been applied even to POMDPs by ignoring unobservable variables, such a 'naive' approach is sometimes very slow (Singh, Jaakkola, & Jordan, 1994). Another way to deal with a POMDP is called a belief state MDP, in which the (optimal) Bellman's equation is given by

$$V(b) = \max_a Q(b,a), \tag{3a}$$

$$Q(b,a) \equiv r(b,a) + \gamma \sum_{b'} P(b'|b,a)V(b'). \tag{3b}$$

The difference from the MDP Bellman's equations (1a) and (1b) is that state s is replaced by a belief state b. A belief state is represented by estimated probability distribution of states. Since there is no probabilistic factor for the observable variables, $b = [y, \hat{P}(z)]$ where $\hat{P}(z)$ is the estimated probability distribution of the unobservable state variables. We assume that a state estimator (SE) is able to estimate a new belief state b', after the agent experiences a state transition from the previous belief state b by an action a and at a new state it observes y'; i.e. $\mathrm{SE}(b,a,y') \equiv b' = [y', \hat{P}'(z)]$. It should be noted that the probability distribution of the unobservable variables, $\hat{P}(z)$, may change after the new observation. In addition, we assume for simplicity that the reward function does not depend on the unobservable variables. In this case, Eqs. (3a) and (3b) becomes

$$V([y,\hat{P}(z)]) = \max_a Q([y,\hat{P}(z)],a), \tag{4a}$$

$$Q([y,\hat{P}(z)],a) = r(y,a) + \gamma \sum_{y'} P(y'|[y,\hat{P}(z)],a)V([y',\hat{P}'(z)]). \tag{4b}$$

This study assumes a finite world; both the state and action spaces are discrete and finite. Even in such a finite world, the belief state MDPs (4a) and (4b) is hard to solve, because the belief state value function is defined for the probability distribution of the unobservable variables and is often intractable. Therefore, we need an approximation. If an RL agent is certain of the estimation of the unobservable variables, $[y,\hat{P}(z)]$ is equivalent to $[y,\hat{z}]$, where $\hat{z}$ denotes the most probable value of z. With this approximation,

$$V([y,\hat{z}]) = \max_a Q([y,\hat{z}],a), \tag{5a}$$

$$Q([y,\hat{z}],a) = r(y,a) + \gamma \sum_{y'} P(y'|[y,\hat{P}(z)],a)V([y',\hat{z}']). \tag{5b}$$

This approximation may not be appropriate when the RL agent is not certain of the estimation of the unobservable variables. Namely, when the uncertainty of the unobservable variables is high, the 'best' policy based on the approximated Bellman's equations (5a) and (5b) may not actually be optimal. Considering this problem, we will later propose an additional mechanism called exploration bonus.

3. Model of environment

For the time being, we assume that there exists one unobservable multinomial variable z in the environment. The distribution of the variable, $\hat{P}(z)$, is estimated by a Bayes inference.

Fig. 1 shows an example problem. This is a very simple maze task; the agent is required to get to a goal point (G) from a start point (S). There may be a barrier denoted by the

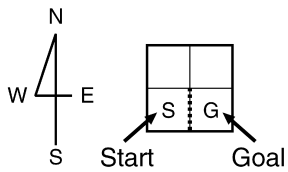


Fig. 1. An example maze with a 2×2 grid. 'S' and 'G' denote the start point and the goal point, respectively. The thick dotted line denotes an invisible bi-directional barrier that exists probabilistically.

vertical dotted bar but the barrier cannot be seen by the agent. The existence of the barrier can be 'perceived' when the agent fails to go beyond the barrier. If the existence of the barrier is regarded as a stochastic event, then it is a binomial event. The shortest path needs only one step without the barrier, while it needs three steps with the barrier. This problem can be formulated in two ways. One is a stochastic MDP, in which the existence of the barrier is a stochastic nature of the environment. The other is a deterministic POMDP, in which the barrier existence is described by an unobservable variable, and if the agent is able to observe all the variables including the barrier existence, the environment is deterministic. We first explain our method to identify the environment according to the formulation of a deterministic POMDP. After that, we discuss according to the formulation of a stochastic MDP.

3.1. Bayes inference of multinomial model

If there are M possible values for the unobservable variable z, it is represented by an M-dimensional vector; $z_i \in \{0,1\}$ $(i = 1,\ldots,M)$ and $\sum_{i=1}^{M} z_i = 1$. $z_i = 1$ indicates z takes the ith value. Let parameter $g \equiv (g_1,\ldots,g_M)$ define a probabilistic model of the multinominal model. From its definition, $\sum_{i=1}^{M} g_i = 1$. In the maze example $(M = 2)$, $z_1 = 1$ and $z_2 = 1$ indicate the existence and non-existence of the barrier, respectively, and g_1 and g_2 denote the probabilities of the existence and non-existence, respectively. After observing T events for the multinomial variable, $Z \equiv \{z(t)|t = 1,\ldots,T\}$, the likelihood of the events is given by

$$P(Z|g) = \prod_{t=1}^{T}\prod_{i=1}^{M} g_i^{z_i(t)} = \exp\left(T\sum_{i=1}^{M}\langle z_i\rangle_D \log g_i\right), \tag{6}$$

where $\langle z_i\rangle_D \equiv (1/T)\sum_{t=1}^{T} z_i(t)$. Eq. (6) indicates that the likelihood of the multinomial variable has an exponential form; the sufficient statistics is $T\langle z_i\rangle_D$ and the natural parameter is $\log g_i$.

A Bayes inference considers the posterior distribution of the parameter, $P(g|Z)$. A convenient method for the Bayes inference is that a trial posterior $Q(g)$ is prepared to approximate the true posterior $P(g|Z)$ and the following variational free energy is maximized with respect to the trial posterior:

$$F(Q) = \int Q(g)\log\frac{P(Z|g)P(g)}{Q(g)}\mathrm{d}g, \tag{7}$$

where $P(g)$ is a prior distribution for parameter g. If the trial posterior includes the true posterior, as a consequence of the maximization, the Kullback–Leibler (KL) divergence between the trial posterior and the true posterior becomes zero; namely, the two probabilistic distributions are equivalent to each other. The maximization is easily achieved by taking the variational condition: $\delta F/\delta Q = 0$. A detailed explanation is described in Appendix A.

If we assume a natural conjugate posterior distribution for parameter g, the posterior distribution becomes a Dirichlet distribution (Heckerman, 1999):

$$Q(g|\nu) = \exp\left(\sum_{i=1}^{M}\nu_i \log g_i - \Phi(\nu)\right), \tag{8}$$

where $\nu \equiv (\nu_1,\ldots,\nu_M)$ is a hyperparameter[1] that specifies the parameter distribution, and $\Phi(\nu)$ is the normalization term.

If no a priori knowledge on the prior distribution $P(g)$ is available, it is natural to choose a non-informative prior. In the maze example, a non-informative prior means that the agent has no a priori idea about the barrier existence and it considers $P(g_1) = P(g_2) = 1/2$. With a non-informative prior, the exact Bayes inference is given by

$$\nu_i = T\langle z_i\rangle_D, \tag{9}$$

with Eq. (8). The parameter expectation with respect to the Dirichlet posterior distribution is given by

$$\bar{g}_i \equiv \int g_i Q(g|\nu)\mathrm{d}g = \frac{\nu_i + 1}{\sum_{j=1}^{M}\nu_j + M}. \tag{10}$$

Using Eq. (9), expectation (10) becomes

$$\bar{g}_i = \frac{T\langle z_i\rangle_D + 1}{T + M}. \tag{11}$$

In our POMDP formulation (Eqs. (5a) and (5b)), the estimation of $\hat{P}(z)$ and $\hat{z}$ is necessary. $\hat{P}(z_i = 1)$ is estimated as $\bar{g}_i$ $(i = 1,\ldots,M)$, and $\hat{z}$ is estimated as $z_k = 1$ (signifying the kth value) such that $k = \arg\max_i \bar{g}_i$. This batch estimation is appropriate when the environment is static. In a dynamic environment, however, the estimation should be done in an on-line manner.

3.2. On-line learning and forgetting

In a dynamic environment, an inference based on observations in the past may not be correct due to the environmental change. Therefore, the inference should put an emphasis on recent observations. Such an inference can be done by defining a weighted variational free energy

[1] Note that a hyperparameter is different in its notion from a meta-parameter.

(Sato, 2001) as

$$F(Q|\tau) \equiv \tau\eta(T)\sum_{t=1}^{T}\left(\prod_{u=t+1}^{T}\lambda(u)\right)\int Q(g)\log P(z(t)|g)\mathrm{d}g + \int Q(g)\log\frac{P(g)}{Q(g)}\mathrm{d}g, \quad (12)$$

where $\eta(T) = [\sum_{t=1}^{T}(\prod_{u=t+1}^{T}\lambda(u))]^{-1}$ is the normalization term. The time-dependent discount factor $\lambda(t)$ $(0 \leq \lambda(t) \leq 1)$ is scheduled so that it approaches 1 as t increases; e.g. $1-\lambda(t) \sim 1/t$. Since the weight value for a data point becomes small as time passes, Eq. (12) puts an emphasis on recent data. In other words, it introduces 'forgetting' effect on old data. The weighted variational free energy thus introduces an on-line Bayes inference. In addition, parameter τ corresponds to the effective data number in the weighted free energy. If τ is smaller than T, the new free energy (12) respects the prior more than the original free energy. Namely, parameter τ balances the weight of the likelihood against the prior. Since our Bayes inference uses a non-informative prior, the decrease of parameter τ means a random inference of the unobservable variable; namely, it corresponds to further forgetting effect on past perceptions of the unobservable variable.

Based on the discussion above, we use the following Bayes inference method instead of the original one given by Eq. (9):

$$\nu_i = \tau\langle z_i\rangle(t), \quad (13a)$$

$$\tau^{\mathrm{new}} := \begin{cases} \tau^{\mathrm{old}} + 1 & \text{(after one perception of variable } z\text{)} \\ \kappa\cdot\tau^{\mathrm{old}} & \text{(after an episode)} \end{cases}, \quad (13b)$$

where the sufficient statistics after the tth perception, $\langle z_i\rangle(t)$, is calculated in an on-line manner (Sato, 2001):

$$\langle z_i\rangle(t) = (1-\eta(t))\langle z_i\rangle(t-1) + \eta(t)z_i(t), \quad (14a)$$

$$\eta(t) = (1+\lambda(t)/\eta(t-1))^{-1}. \quad (14b)$$

The effective data number τ is incremented after a single perception of the unobservable variable. After an episode, however, τ value for every unobservable variable is decreased by a discount factor $0 < \kappa \leq 1$ which is the forgetting coefficient. Here, an 'episode' denotes a series of state transitions, typically from a start state to an end state.

Accordingly, the SE estimates

$$\hat{P}(z_i = 1) = \bar{g}_i = \frac{\tau\langle z_i\rangle + 1}{\tau + M}, \quad (15)$$

where $\langle z_i\rangle$ is the current sufficient statistics. $\hat{z}$ is estimated as $z_k = 1$ such that $k = \arg\max_i \bar{g}_i$. If the unobservable variable has often been perceived, the corresponding effective data number becomes large. In this case, the inference by Eq. (15) almost becomes that by maximum likelihood; this is natural because the agent has much and recent knowledge on the unobservable variable. In contrast, if variable z has not been perceived lately, the corresponding effective data number becomes small. The inference becomes random in this case because the agent has little recent knowledge on the unobservable variable and it is natural for the agent to guess that its value may change during his absence. In the maze example, the agent becomes uncertain of the barrier existence if the agent has not tried to go beyond the barrier lately. The agent assumes that the unobservable variable, i.e. the environment, will change with time-constant $1/(1-\kappa)$.

It should be noted that our Bayes formulation can use an informative prior instead of the non-informative prior. Dayan and Sejnowski (1996) used an informative prior (the barrier will disappear with a high probability) in order to encourage the agent's exploration (attempting to go beyond the barrier).

3.3. Inference of state transition

In the example maze task, the probability that the start state reaches the goal state by an action 'go east' is identical to the probability of the barrier existence. Namely, the probability of the unobservable variable is equivalent to the state-transition probability. Therefore, the above-mentioned Bayes inference of the unobservable variables is naturally extended to the inference of the state transitions.

Let S and A denote the set of states and the set of actions, respectively. $P(s_i|s_j,a)$ denotes the probability that state $s_j \in S$ reaches state $s_i \in S$ by action $a \in A$. Here, we consider events with a fixed state–action pair (s_j,a). A multinomial variable z signifies an occurrence of a single state-transition event. If there are M possible states that are reachable from the state–action pair, z is represented by an M-dimensional vector; if a state transition event to state $s_i \in S$ occurs, $z_i = 1$, $z_k = 0$ $(k \neq i)$. Parameter g_i defines a probabilistic model of the multinomial model. In the example maze, when the agent tries to 'go east' (a = 'go east') at the start point ($s_j = S$), the possible new state (s_i) is either the goal point or the start point.

Like in the discussion in Section 3.2, the state-transition probability $P(s_i|s_j,a)$ is estimated as $\bar{g}_i$ given by Eq. (15), i.e. $\hat{P}(s_i|s_j,a) = \bar{g}_i$. The effective data number τ is updated by

$$\tau^{\mathrm{new}} := \begin{cases} \tau^{\mathrm{old}} + 1 & \text{(action } a \text{ is selected at state } s_j\text{)} \\ \kappa\cdot\tau^{\mathrm{old}} & \text{(after an episode)} \end{cases}, \quad (16)$$

instead of Eq. (13b). Eq. (15) means that $\hat{P}(s_i|s_j,a)$ approaches the maximum likelihood estimation $\langle z_i\rangle$, as the effective data number τ increases. When the effective data number is small, on the other hand, the estimation nearly becomes $1/M$, implying that the transition is regarded as random for every possible new state. Thus, the estimation reflects the information amount that the agent has. Note that τ denotes the effective data number of a state–action pair

(s_j, a), and it should be defined for every state–action pair individually.

3.4. Model-based reinforcement learning

The discussions above suggest that the inference of the state transition is equivalent to the inference of the unobservable variables, at least in the case of the example maze task.

In this study, we use a model-based RL method, in which Eq. (5b) is replaced by

$$Q([y, \hat{z}], a) = r(y, a) + \gamma \sum_{y'} \hat{P}(y'|y, a) V([y', \hat{z}']), \tag{17}$$

where $\hat{P}(y'|y, a)$ is determined by the method described in Section 3.3. In a deterministic POMDP, this is equivalent to that the SE conducts a Bayes inference of the unobservable variables. In a stochastic MDP, the model-based RL estimates the model of the stochastic environment based on a Bayes inference.

This model-based RL method can be applied to more general problems, like stochastic POMDPs; namely, the state transition is stochastic and there are unobservable variables. For example, provided in the maze example that each action is emitted or not with probability p or $1 - p$, respectively, and an emitted action is effective (i.e. changes the agent's state) if there is no barrier in the moving direction. This task can be formulated as a stochastic POMDP.

In the model-based RL method, even in such a case, the state transition for the observable state variable y, $\hat{P}(y'|y, a)$, is estimated by a Bayes inference, regarding the existence of the unobservable variables as stochastic nature of the environment. It should be noted that our model-based RL is not a naive MDP approximation, because the value function and the action-value function consider the estimation of the unobservable variables, $\hat{z}$.

4. Control of randomness in action selection

Although the objective of RL is to obtain the optimal policy that maximizes the value function (Eq. (1a) or (5a)), a simple maximization procedure often results in a semi-optimal policy and the lack of adaptability to the environmental change. This section discusses the way to overcome this problem. Although we assume a finite world, equations in this section often use integral notations for description convenience. In this section, $[y, \hat{z}]$ is represented as s.

4.1. Inverse-temperature

We define a stochastic policy π by a conditional probability $P^{\pi}(a|s)$ of action a for state s. From its definition, $\int P^{\pi}(a|s)\mathrm{d}a = 1$. Using the current action-value function, which may differ from the really optimal one, the greedy policy maximizes

$$\int Q(s, a) P^{\pi}(a|s)\mathrm{d}a. \tag{18}$$

Especially when the state and action spaces are finite, the greedy policy will assign probability zero to the possible actions except one or several. Then, it becomes difficult for the agent to adapt its policy to the environmental change, and/or to improve the present best (i.e. semi-optimal) policy. This is one aspect of the exploitation–exploration problem.

In order to preserve the exploration ability of the policy, we define the free energy[2]

$$J(P^{\pi}) = \int Q(s, a) P^{\pi}(a|s)\mathrm{d}a - \frac{1}{\beta} \int P^{\pi}(a|s) \log P^{\pi}(a|s)\mathrm{d}a. \tag{19}$$

The first and second terms in Eq. (19) are called the energy term and the entropy term, respectively. The coefficient of the entropy, $1/\beta$, is called the (thermo-dynamical) temperature. β is then called the inverse-temperature. If the temperature is large, the randomness of the probability $P^{\pi}(a|s)$ is large; namely, the policy becomes random and hence exploration is encouraged. If the temperature is small, the policy randomness becomes small so that exploitation is encouraged. Therefore, the inverse-temperature parameter β controls the balance between exploitation and exploration. Since the way many parameters of the agent are changed by learning is dependent on the parameter β, β is called a meta-parameter (Doya, 2000b).

Using the variational method, the maximization of the free energy $J(P^{\pi})$ with respect to the stochastic policy $P^{\pi}(a|s)$ is achieved by

$$P^{\pi}(a|s) = \frac{\exp(\beta Q(s, a))}{\int \exp(\beta Q(s, a))\mathrm{d}a}, \tag{20}$$

which is called the soft-max policy or the Boltzmann policy (Sutton & Barto, 1998). When the inverse-temperature meta-parameter is small, the soft-max policy randomly selects one of the possible actions. When the inverse-temperature parameter is large, in contrast, it selects the greedy action that maximizes the current action-value function.

4.2. Local control of randomness

A constant inverse-temperature means that the randomness induced by the entropy is constant against the energy term, while the energy term depends on the variation of the action-value function. For example, if the action-value function for a certain state s does not vary with respect to action a, on one hand, the soft-max policy becomes random

[2] Although we use the words 'free energy' for Eq. (7) or (12), and (19), their definitions are different from each other.

even with a large β. If the action-value function significantly varies, on the other hand, the soft-max policy likely selects the greedy action even with a small β. The policy randomness is thus dependent on the variation of the action-value function with respect to possible actions.

By considering the variation of the action-value function, we define a normalized soft-max policy:

$$P^{\pi}(a|s) = \frac{\exp(\beta_0 \tilde{Q}(s,a))}{\int \exp(\beta_0 \tilde{Q}(s,a))\mathrm{d}a}, \tag{21a}$$

$$\tilde{Q}(s,a) \equiv \frac{Q(s,a) - E[Q(s,a)]}{\sqrt{E[Q(s,a)^2] - (E[Q(s,a)])^2}}, \tag{21b}$$

where β_0 is a new inverse-temperature meta-parameter and is constant. $E[\cdot]$ denotes the expectation with respect to the current policy and it is approximated based on actual experiences using the current policy. Using the normalized soft-max policy, the action randomness is normalized so that exploratory actions do not significantly depend on the variation of their expected results.

The normalized soft-max policies (21a) and (21b) is equivalent to the original soft-max policy (20) with a new inverse-temperature:

$$\beta(s) = \beta_0 \cdot \beta_{\mathrm{l}}(s) = \frac{\beta_0}{\sqrt{E[Q(s,a)^2] - (E[Q(s,a)])^2}}. \tag{22}$$

Note that $\beta(s)$ does not depend on action a. Accordingly, in order to introduce the randomness normalized with respect to the variation of the action-value function, the inverse-temperature β becomes dependent on state s. $\beta_{\mathrm{l}}(s)$ is then called the local coefficient of the inverse-temperature.

4.3. Global control of randomness

Exploration is important especially when the agent perceives that the environment has probably changed. If the agent believes that the environment has not changed, in contrast, exploitation is more important than exploration. Therefore, the inverse-temperature should be controlled based on the perception of the environmental change.

One such control can be done by

$$\beta_{\mathrm{g}} := \begin{cases} \alpha + (1-\alpha)\beta_{\mathrm{g}} & (\text{if } \hat{z}' = \hat{z}) \\ \beta_r & (\text{otherwise}) \end{cases}, \tag{23}$$

where $\hat{z}$ ($\hat{z}'$) is the estimation of the unobservable variables before the action (after the action). $0 < \alpha < 1$ is a constant that determines how fast β_{g} approaches its maximum value ($= 1.0$) from its minimum value ($= \beta_r$). When the estimation of the unobservable variables does not change after an actual experience (action), the agent guesses that the environment represented by the unobservable variables has not changed. The upper condition in Eq. (23) says that β_{g} gradually increases in such a case (see Fig. 6). The agent then prefers exploitation. When the estimation of the unobservable variables changes after an actual experience, in contrast, the agent guesses that the environment has changed. The lower condition in Eq. (23) says that β_{g} is set to its minimum value in such a case. The agent then prefers exploration in order to quickly adapt to the new environment.

If the environment is deterministic like in the maze example, the following control will work well:

$$\beta_{\mathrm{g}} := \begin{cases} \alpha + (1-\alpha)\beta_{\mathrm{g}} & (\text{if } z = \hat{z}) \\ \beta_r & (\text{otherwise}) \end{cases}. \tag{24}$$

When an actual perception of the unobservable variables, z, is different from its expected one $\hat{z}$, the agent guesses that the environment has changed. This control is not appropriate for stochastic environments such that the perception of unobservable variables may differ from their expectation due to the stochastic nature.

With either of the controls above, inverse-temperature β in the original soft-max policy (20) is replaced by

$$\beta(s) = \beta_0 \cdot \beta_{\mathrm{g}} \cdot \beta_{\mathrm{l}}(s). \tag{25}$$

$\beta_{\mathrm{l}}(s)$ considers the variation of the action-value function and locally controls the randomness, while β_{g} attempts to perceive the environmental change and globally controls the randomness. Then, β_{g} is called the global coefficient of the inverse-temperature.

In the later experiments, we use the second control method given by Eq. (24).

4.4. Exploration bonus

In RL, the reward function is determined according to the task that is to be accomplished by the agent. It is usually independent of the amount of environmental information that the agent has. For actual animals, however, information of the environment is very important. Exploration is nothing but acquiring information from the environment. By assuming an additional reward term corresponding to the information that will be acquired from the environment, the agent is encouraged to take exploratory actions; this is the idea of our exploration bonus.

The bonus is given in proportional to the entropy of the posterior distribution of the state-transition, $H_D(s,a)$, where the definition of entropy H_D is described in Appendix A (see Eq. (A9)). Using the bonus, the action-value function used in the soft-max policy (20) is modified into

$$r^{+}(s,a) = r(s,a) + \epsilon H_D(s,a), \tag{26a}$$

$$Q^{+}(s,a) = r^{+}(s,a) + \gamma \sum_{s'} \hat{P}(s'|s,a) V(s'), \tag{26b}$$

where ϵ is a constant.

A small entropy means that the information acquired from the environment by taking action a at state s, is expected to be small with respect to the current estimation of the environment; in this case, the probability to take the action is decreased. When the acquired information is

expected to be large, the probability to take the action is increased. This results in an encouragement of exploration.

The POMDP formulation provides us with another interpretation of the exploration bonus. In a POMDP, an action determined by the approximated Bellman's equations (5a) and (5b) may be different from the optimal action in the belief state MDP. Such a situation occurs especially when the agent is uncertain of the estimation of the unobservable variables. As discussed in Section 3.4, the entropy of the state transitions is similar to the entropy of the unobservable variables. Therefore, the exploration bonus can be interpreted as follows. A small entropy of the state transition means that an action determined by our model-based RL is close to the optimal action in the belief state MDP. Therefore, the action selection should not be disturbed by the exploration bonus. On the other hand, a large entropy of the state transition means that an action determined by our model-based RL may be apart from the optimal action in the belief state MDP. In such a case, the agent prefers acquiring a large information so as to be certain of the environment.

Although the exploration bonus modifies the policy, it does not affect the Bellman's equations (1a) and (1b) or (5a) and (5b). Namely, the bonus does not introduce any bias to the estimation of the value function.

4.5. Reinforcement learning algorithm

Here we summarize the whole RL algorithm for a single learning episode.

1. Set the agent to a start state.
2. For a specific number of state transitions, the following steps are conducted.
   (a) Let y be the current observable state. Each unobservable variable $\hat{z}$, which is relevant to y, is estimated as $z_k = 1$ such that $k = \arg\max_i \bar{g}_i$, where $\bar{g}$ is given by Eq. (15).
   (b) For every action a possible at y, conduct the following steps.
      (i) For every possible observable state y' that is reachable from y by a, the state-transition probability $\hat{P}(y'|y,a)$ is calculated by Eq. (15).
      (ii) Using $s = [y, \hat{z}]$ and $\hat{P}(y'|y,a)$, obtain $Q(s,a)$ by Eq. (17).
      (iii) Using $s = [y, \hat{z}]$ and $\hat{P}(y'|y,a)$, obtain $Q^+(s,a)$ by Eqs. (26a) and (26b).
   (c) Update $V(s)$ based on Eq. (5a).[3]
   (d) Obtain $\beta(s)$ by Eq. (25).
   (e) Calculate $P^\pi(a|s)$ according to Eq. (20) with the replacement of $Q(s,a)$ by $Q^+(s,a)$ and the replacement of β by $\beta(s)$.
   (f) An action a is selected with probability $P^\pi(a|s)$. Observable state y changes to a new observable state y'' according to the real dynamics of the environment.
   (g) Update the sufficient statistics corresponding to y'' by Eqs. (14a) and (14b).
   (h) According to the upper rule in Eq. (16), increment the effective data number for the state–action pair (s,a).
   (i) If y'' is a goal state, exit the loop. Otherwise, $y := y''$ and go to step (a).
3. According to the lower rule in Eq. (16), decrease the effective data number for every state–action pair.

In our RL scheme, the Boltzmann policy with a modified inverse-temperature realizes undirected exploration, while the exploration bonus realizes directed but local exploration. It should be noted that our RL scheme does not use imaginary value-iteration steps based on the current environmental model,[4] although such imaginary steps were used in DYNA system of Sutton (1990) and Dayan and Sejnowski (1996).

5. Simulation results

Our RL scheme is applied to two-dimensional maze tasks.

5.1. Task setting

The first maze has a 16×16 grid (Fig. 2). This task is a modification of the one used by Dayan and Sejnowski (1996), which originated from Sutton (1990). At each grid point, the agent takes one of four actions: $a \in \{N, S, E, W\}$, though an action going beyond the maze boundary is not allowed. The maze boundary is visible. For every action, an immediate reward -1 is given, i.e. a cost 1 is given. The agent moves from the start point to the goal point. When the agent arrives at the goal point, the episode ends. If the agent cannot get to the goal point within 200 action steps, the episode also ends. Since the objective of the agent is to search for the optimal policy maximizing the reward accumulation, it is required to find out the shortest path from the start point to the goal point.

There are bi-directional barriers in the maze. If the agent tries to take an action going beyond a barrier, it stays at the current grid point and receives a reward -1. Barriers are invisible; namely, whether a barrier exists or not can be perceived only by executing an action. Since the existence of barriers is deterministic but temporally variant, it is assumed to be a stochastic event. The existence of a barrier portion is represented by an unobservable probabilistic variable. The number of possible values for each

[3] A gradual updating method is preferable in a stochastic environment.

[4] Although imaginary steps are useful for quickly propagating a local change of the environment to the whole state space, they need additional computational time.

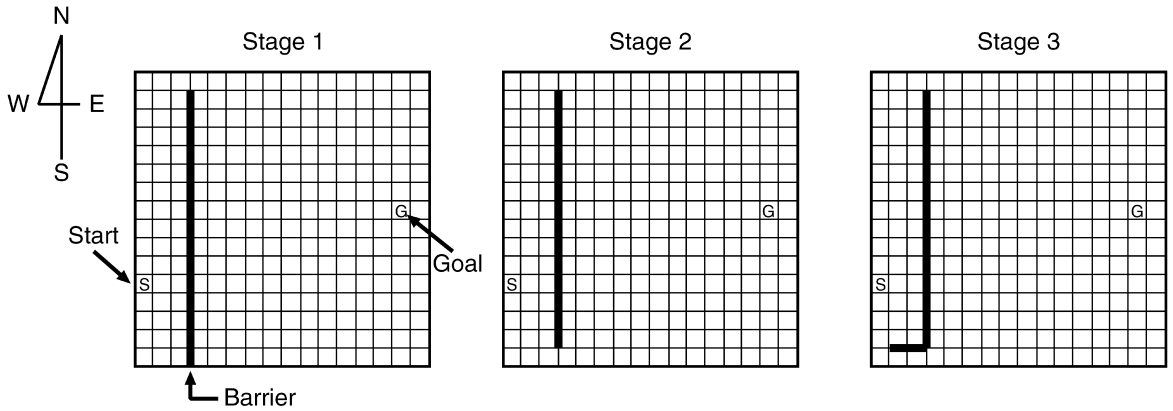


Fig. 2. A maze task with a 16 × 16 grid. 'S' and 'G' denote the start point and the goal point, respectively. The thick line denotes bi-directional barriers. The maze boundary is visible, while the barriers are invisible. The objective of the agent is to find out the shortest path from the start point to the goal point.

unobservable variable is two; existence or non-existence. With respect to the state transition, the number of possible transitions for each state–action pair is also two; moving to a new grid point by a successful action or staying at the current grid point due to an unsuccessful action. Therefore, the Bayes inference estimates a binomial probabilistic model for every unobservable variable (or state–action pair). This is the inference of the environmental model.

In order to see the agent's ability to deal with the exploitation–exploration problem, the environment changes with time. In this maze task, the existence of barriers changes. The series of learning episodes are divided into three stages, as shown in Fig. 2. At the first stage (1 ∼ 400 learning episodes), there are vertical barriers and the shortest paths to the goal go around the northernmost part of the barriers. The length of the shortest paths is 32. At the second stage (401 ∼ 800 learning episodes), the southernmost portion of the barriers is removed. The length of the shortest paths, which go through the removed barrier portion, turns to be 26. At the third stage (801 ∼ 1200 learning episodes), new barriers appear to hinder the agent from going along the barriers. At this stage, the length of the shortest paths does not change, while the variation of the shortest paths becomes small; the agent needs to go straightly south from the start point.

In our RL scheme, if a certain time period has passed after the last perception of the barrier existence, i.e. failure to go beyond the barrier, the agent comes to forget the existence. Due to this effect, the estimation of the value function is likely to involve the possibility of barrier disappearance. At the second stage, therefore, the value function on a grid point along the barriers is larger than that on a grid point apart from the barriers, even if the distance to the goal is the same. In order to find out the shortest path at the third stage, therefore, the agent first needs to go around the new barriers and to recognize that the actual shortest path is the one that goes straightly south from the start point. This is the difficulty of this maze task.

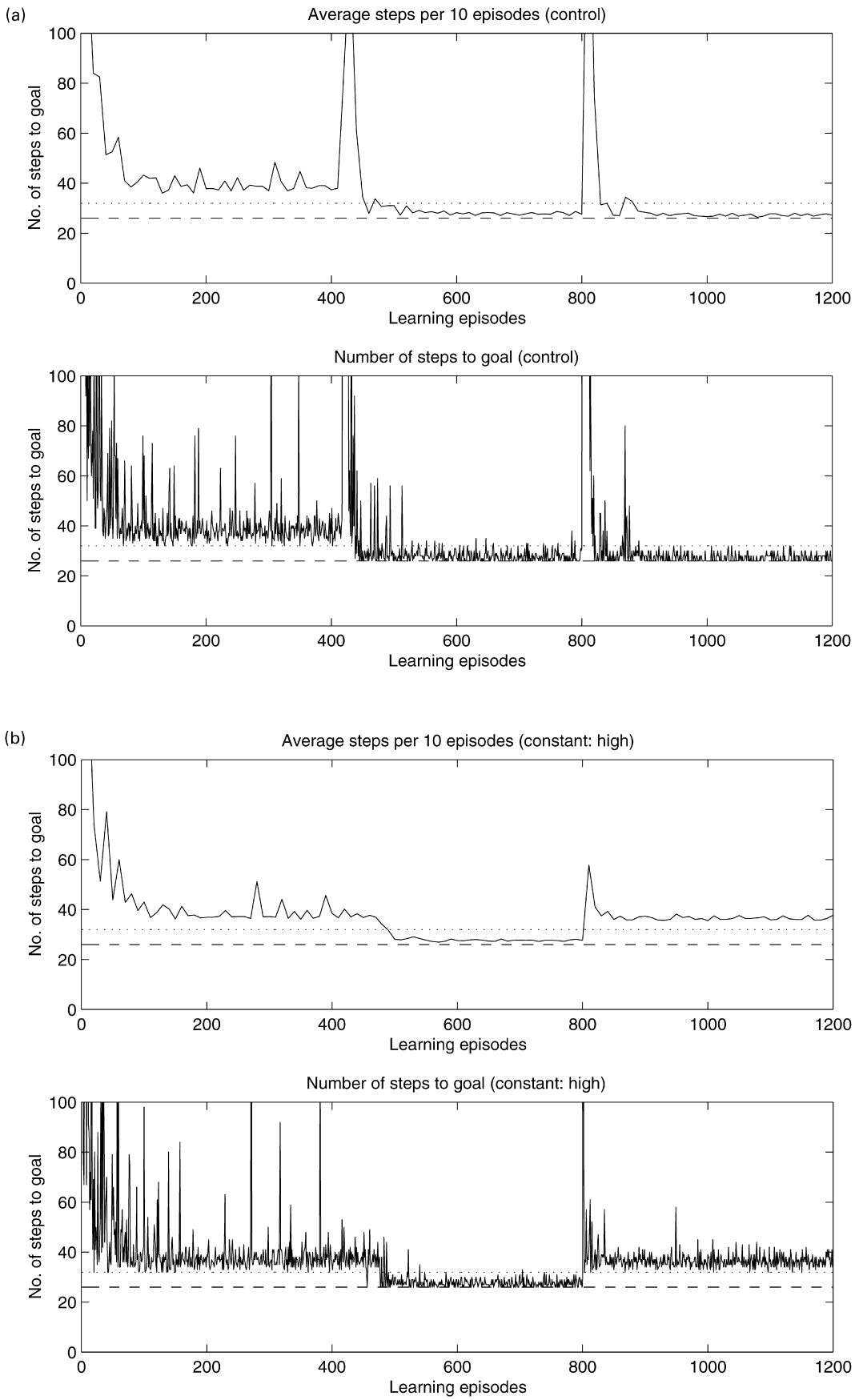
The value function $V(s)$ is initialized to be 0.0 for every state. Since the initial value of the value function is smaller than its real value, possible actions are tried several times at every state at the early learning stage. Although this 'optimistic' initialization is a very simple heuristic method to encourage the exploration, it is not effective in adapting to the environmental change.

5.2. Simulation result

Our RL scheme is applied to the maze task above. Fig. 3(a) shows the number of actions during 1200 learning episodes. At the first stage, the action number significantly varies, because the estimated value function is distant from the real one and hence the action randomness is large. Another reason is the effect of the optimistic initialization. At the second and third stages, the variation of action number becomes small because the improvement of the value function suppresses the action randomness.

Fig. 3(a, lower) shows that the agent successfully finds out the shortest path, whose length is 32 at the first stage and 26 at the second and third stages. The randomness in its actions is small so that the average number of actions is slightly larger than the shortest path length. When the new barriers appear at the beginning of the third stage, the number of actions grows considerably. After a short trial-and-error period, however, the agent successfully finds out the new shortest path.

The role of the control of the inverse-temperature is examined by comparing the result with that by a similar method with a fixed inverse-temperature. Fig. 3(b) and (c) show the results with the inverse-temperature values fixed at a large value ($\beta = 100$) and a small value ($\beta = 1.0$), respectively. Even with a fixed inverse-temperature, the exploration bonus is used. If it is not used, the agent with a large inverse-temperature cannot adapt to the environmental change. With a large constant value for the inverse-temperature, the agent prefers exploitation to exploration. Fig. 3(b, upper) shows that the averaged action number at the third stage is larger than that at the second stage. The agent does not find out the shortest path at the third stage and it selects the semi-optimal path going around the


(a)
Average steps per 10 episodes (control)
No. of steps to goal
Learning episodes
Number of steps to goal (control)
No. of steps to goal
Learning episodes
(b)
Average steps per 10 episodes (constant: high)
No. of steps to goal
Learning episodes
Number of steps to goal (constant: high)
No. of steps to goal
Learning episodes

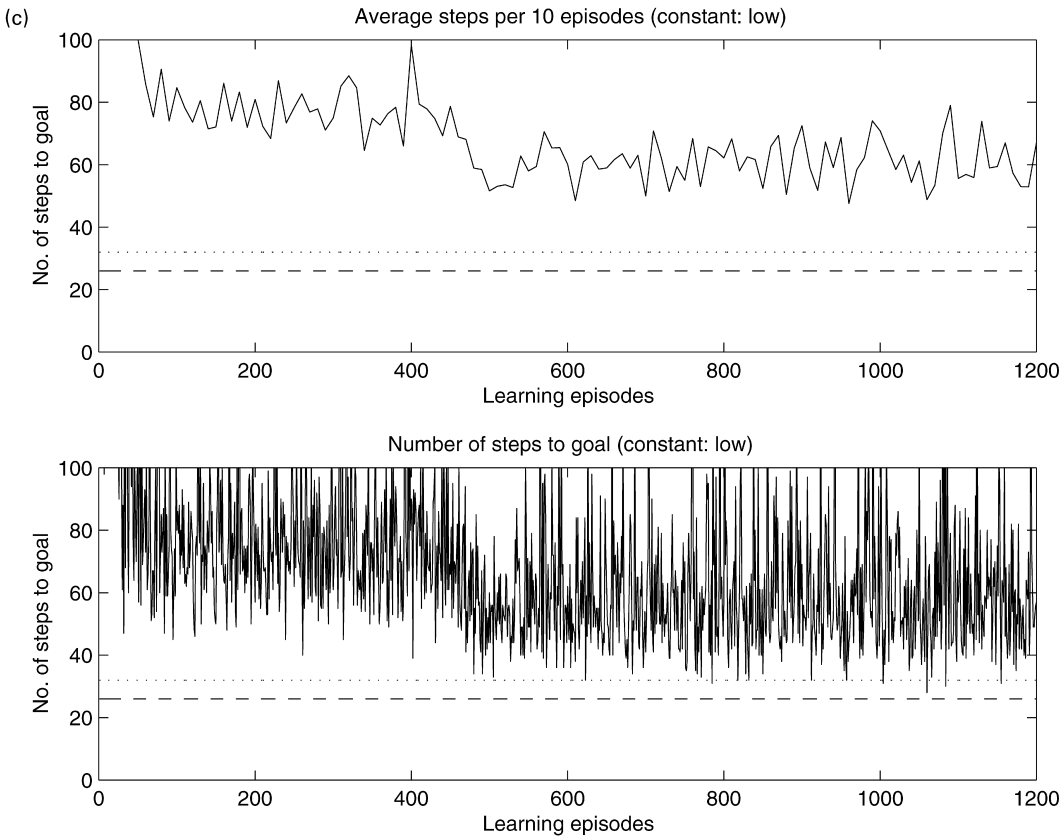


Fig. 3 (*continued*)

northernmost part of the barriers. The ability to adapt to the environmental change is thus small in this agent. With a small constant value for the inverse-temperature, in contrast, the averaged action number stays large throughout the three stages (Fig. 3(c)).

Fig. 4 shows the average performance for the three agents. We executed 100 training runs by varying initial conditions and the ordinate in Fig. 4 denotes the average number of actions over the 100 runs.

Fig. 5 shows the position distribution of the three agents. Each sub-figure shows the logarithm of the number of visits to each grid point. The upper three sub-figures show that the agent with the inverse-temperature control follows the shortest path at the three stages. Especially at the third stage, the path variation in the western part of the barriers is small. This part is important for following the shortest path. In contrast, the variation in the eastern part of the barriers is large. This part is not very important for following the shortest path. The middle three sub-figures show that the path variation of the agent with a large constant for the inverse-temperature is small. This agent cannot find out the shortest path at the third stage. The lower three sub-figures show that the path variation of the agent with a small inverse-temperature is so large that at the third stage it almost randomly selects a path going around the northernmost part or the southernmost part of the barriers.

Fig. 6 shows the global coefficient of the inverse-temperature, β_g, during a single training run. When the agent perceives the environmental change, exploration is encouraged in order to adapt to the new environment by making β_g a small value. When the RL agent does not perceive the environmental change, it prefers exploitation by increasing β_g. Fig. 7 shows the reciprocal of the local coefficient of the inverse-temperature at each grid point, i.e. $1/\beta_l(s)$. On the grid points except for those adjacent to the barriers, the inverse-temperature is large so that the agent prefers exploitation. On the grid points adjacent to the barriers, in contrast, the inverse-temperature is small so that

Fig. 3. Number of actions taken by the three agents. (a)–(c) The abscissa denotes the number of learning episodes. The ordinate denotes the number of actions averaged over 10 episodes in the upper figure, and the number of actions in each episode in the lower figure. The dotted (dash) line denotes the shortest path length at stage 1 (stages 2 and 3). We conducted experiments many times by varying the random seeds, and these figures are the most typical ones among them. (a) Learning process of an agent with the control of the inverse-temperature. Parameters are set at $\kappa = 0.98$, $\beta_0 = 10$, $\beta_r = 0.001$, $\alpha = 0.0005$, and $\epsilon = 3$. (b) Learning process of an agent with a large constant ($\beta = 100$) for the inverse-temperature. (c) Learning process of an agent with a small constant ($\beta = 1.0$) for the inverse-temperature.

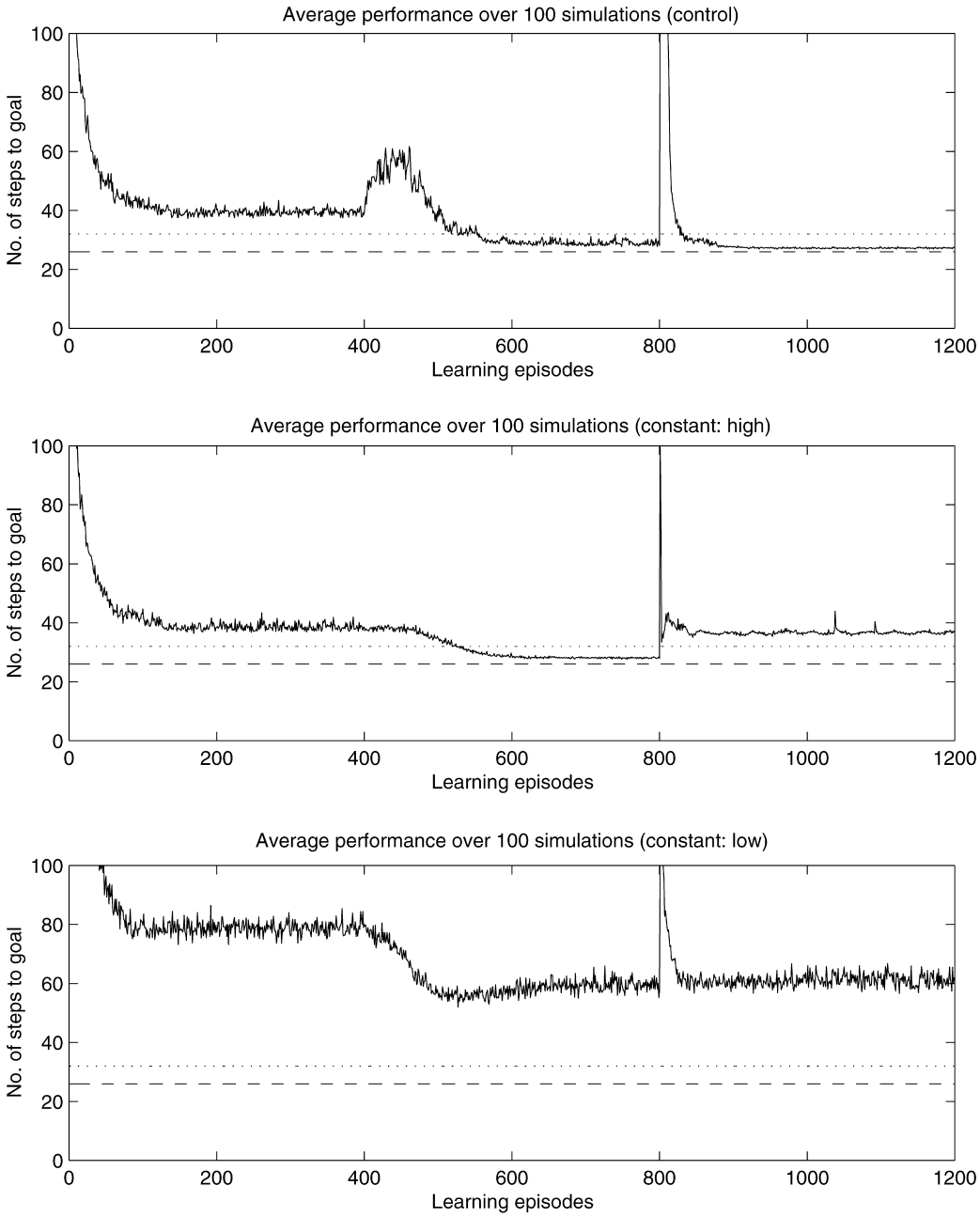


Fig. 4. Number of actions taken by the three agents. We executed 100 training runs by varying initial conditions and the ordinate denotes the average number of actions over the 100 runs. (upper) Learning process of an agent with the control of the inverse-temperature. (middle) Learning process of an agent with a large inverse-temperature. (lower) Learning process of an agent with a small inverse-temperature.

the agent prefers exploration; namely, the agent expects the barriers to disappear.

In our method, the inverse-temperature control and the exploration bonus cooperatively control the exploitation–exploration balance. Fig. 8 shows a result for an RL agent with the inverse-temperature control but without the exploration bonus. Since the action randomness should be large in order for this agent to adapt to the environmental changes, the average steps to the goal becomes larger than those of an agent with the exploration bonus (see Fig. 4(upper)).

5.3. Zig-zag maze

Our RL scheme is next applied to a more complicated 'zig-zag' maze (Fig. 9). This task is also a modification of the one used by Dayan and Sejnowski (1996). At the first stage (1 ∼ 500 learning episodes), the shortest path to the goal is a zig-zag one and its length is 41. At the second stage (501 ∼ 1000 learning episodes), a barrier portion is removed so that the shortest path becomes the straight one whose length is 21. At the third stage (1000 ∼ 1500 learning episodes), the barrier portion appears again but

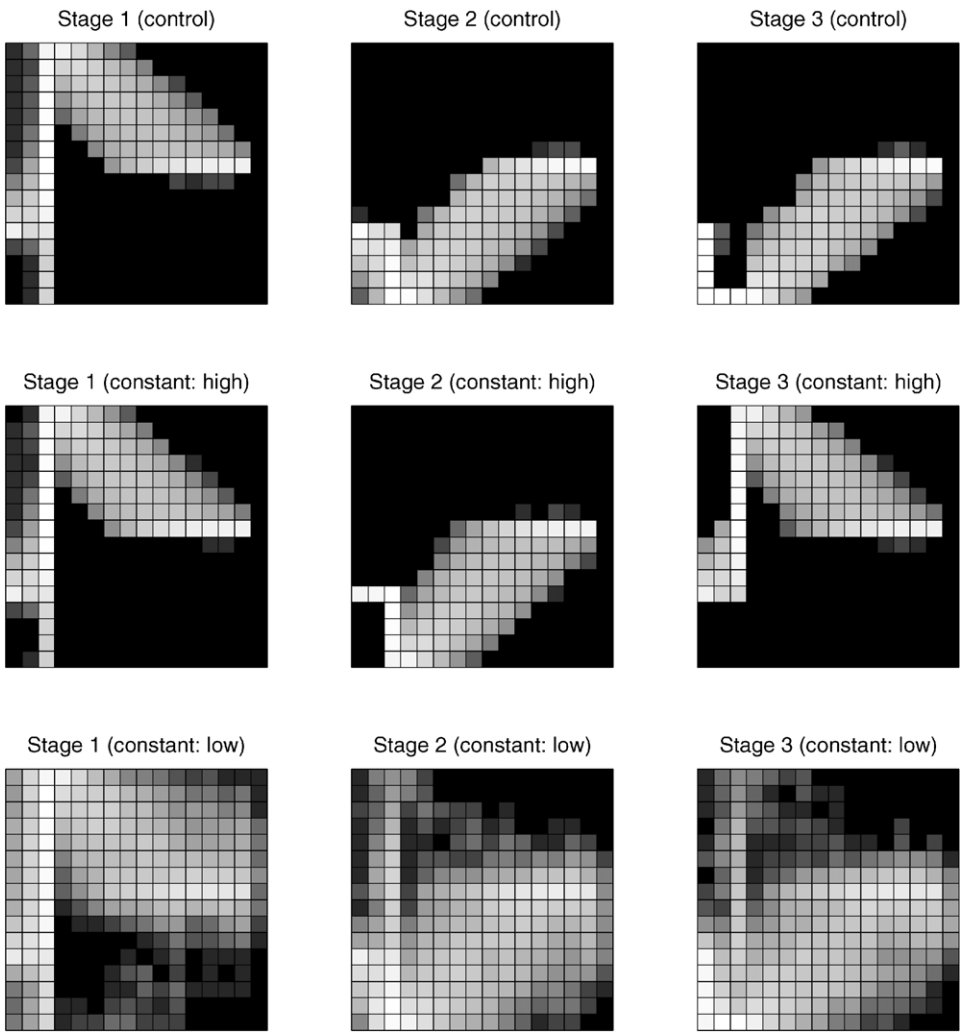


Fig. 5. Logarithm of the number of visits to each grid point for the three agents. The lighter a grid point is, the more frequently the agent visits the grid point. In order to see the behavior of the agents after adapting to the current environment, each sub-figure shows the average in the last 100 learning episodes at each of the three stages. (upper) An agent with the control of the inverse-temperature. (middle) An agent with a large inverse-temperature. (lower) An agent with a small inverse-temperature.

three other portions are removed, so that the shortest path passes the easternmost part of the maze and its length is 31.

Fig. 10 shows results by the three learning agents: (a) with the control of the inverse-temperature; (b) with a large constant ($\beta = 100$) for the inverse-temperature; and (c) with a small constant ($\beta = 0.85$) for the inverse-temperature. The agent with the control successfully adapts to the environmental changes, while the agent with a large constant cannot adapt to the second change of the environment.

Accordingly, setting the inverse-temperature at a large value corresponds to respecting exploitation, while setting it at a small value corresponds to respecting exploration. With a fixed value, the agent cannot change the balance between them, although the balance control is important especially when the environment changes with time.

6. Exploitation–exploration problem in the brain

6.1. Selective attention

Attention is a cognitive function, whose aim is to focus the consciousness on one of the targets of sensation, perception or thought. Attention can be divided into two operations: one is selective attention and the other is sustained attention. They can be validated by different psychological tasks. In a selective attention task, on one hand, a subject is required to process one of the two or more stimuli provided simultaneously. In a sustained attention task, on the other hand, a subject is required to focus on a specific stimulus for a certain period. Selective attention is important for selecting information in order to achieve an objective, whereas sustained attention is important for maintaining the objective itself.

This section discusses selective attention. Awake human

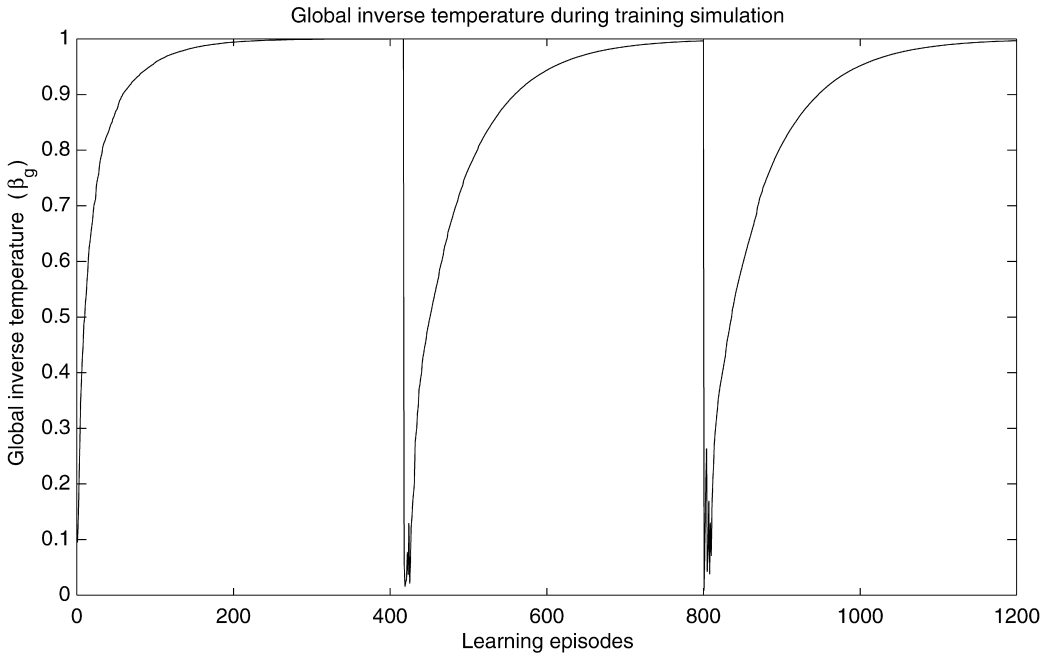


Fig. 6. Global coefficient of the inverse-temperature, β_g, during a single training run. After 400 and 800 episodes, the environment changes.

brain is confronted with a flood of information, i.e. thoughts, memories, emotions and innumerable sensory inputs via various modality channels. Selective attention processes only appropriate portion among vast amount of information, and is a necessary ability in order to rapidly execute appropriate behaviors in a real environment.

Attention is believed to have three major functions: orientation to stimuli, executive function and maintenance of an alert state (Posner & Raichle, 1996). The orientation to stimuli is to orient a part of a body to the direction of a novel stimulus. The executive function is related to control of goal-directed behaviors, detection of targets, resolution of conflicts, suppression of unconscious reaction, and so on. The executive function is necessary in a novel or highly competitive situation, and is important especially in a selective attention task. The maintenance of an alert state involves the establishment of a vigilant state and the readiness for a rapid reaction. This function is necessary not only in a sustained attention task, but also for sustaining the objective in a selective attention task.

6.2. Locus coeruleus and inverse-temperature

Tonic activities of noradrenergic LC neurons depend on sleep-awake stages; namely, they are active in an awake state, less active in a slow-wave sleep state and nearly silent in a rapid-eye-movement sleep state (Aston-Jones, Chiang, & Alexinsky, 1991). Therefore, LC neurons have long been thought to regulate arousality of the brain. However, a recent study based on multicellular recordings of LC neurons in monkeys performing a visual discrimination task has suggested that the LC neurons also have relevance to selective attention (Aston-Jones, Rajkowski, Kubiak, & Alexinsky, 1994). Since the response latency of the LC neurons correlates with the behavioral response time, it is suggested that the LC activity induced by the target facilitates the behavioral response to the target.

The LC, which is located by the forth ventricle in the mid-pontine region of the brain stem, is the major noradrenergic nucleus in the brain. LC neurons have widespread projection on the telencephalic cortical structures

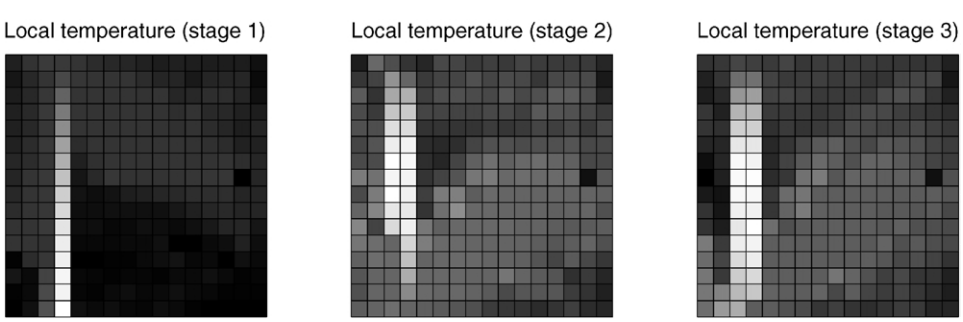


Fig. 7. Reciprocal of local coefficient of the inverse-temperature at each grid point, $1/\beta_l(s)$, for an agent with the control of the inverse-temperature; the lighter a grid point is, the higher the temperature is.

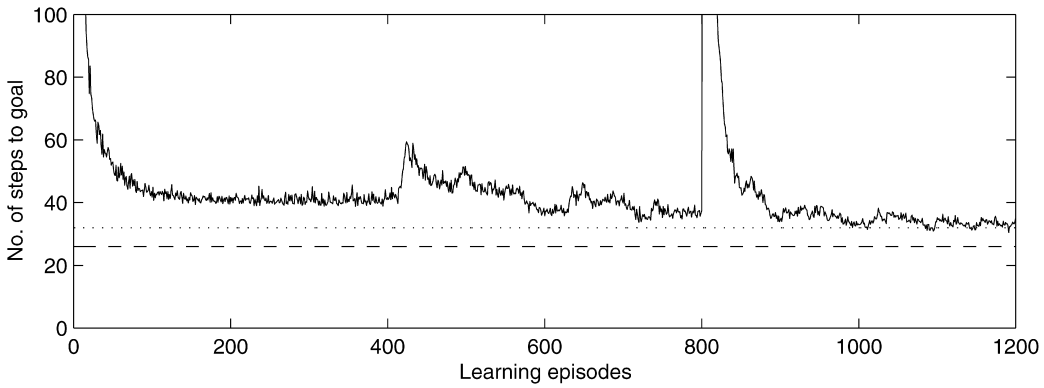


Fig. 8. Number of actions taken by an agent with the control of the inverse-temperature but without the exploration bonus. We executed 100 training runs by varying initial conditions and the ordinate denotes the average number of actions over the 100 runs. Parameters are set at $\kappa = 0.98$, $\beta_0 = 1.6$, $\beta_r = 0.001$, and $\alpha = 0.0005$. In order to make the agent adapt to the environmental changes, β_0 is set at a smaller value than in the simulation in Fig. 4.

and the cerebellar cortex (Foote, Bloom, & Aston-Jones, 1983).

Spontaneous activities of LC neurons play a role in the maintenance of the arousal state. Since the brain areas associated with attentional processing exhibit particularly dense LC innervations (Morrison & Foote, 1986), the LC is probably related to selective attention by controlling the signal-to-noise ratio of the brain processing. In a computational model by Servan-Schreiber, Printz, and Cohen (1990) and Usher et al. (1999), noradrenaline sharpens the response tuning of neurons by increasing the gain of the sigmoidal transfer function.

A recent study on LC neuron recordings showed that spontaneous and stimulus-induced discharge patterns are correlated with behavioral performance (Usher et al., 1999). Phasic LC discharges, which selectively respond to target stimuli, are associated with good behavioral performance. Good performance will be achieved by focusing the consciousness on the target stimuli. On the other hand, a higher level of tonic LC discharges is associated with a higher false alarm error rate, implying a low attentional level.

Usher et al. (1999) suggested that the phasic and the tonic discharges seem to correspond to the exploitation operation and the exploration operation, respectively. Their model assumed that the two modes in LC neuron discharges are controlled by the strength of electrotonic couplings among the LC neurons; namely, the strong and weak couplings induce phasic and tonic firings, respectively. Therefore, the coupling strength controls the balance between exploitation and exploration. The existence of electrotonic couplings in the LC has been suggested in adult rats (Ishimaru & Williams, 1996) and the couplings are considered to regulate activities of LC neurons (Christie, Williams, & North, 1989).

The idea by Usher et al. is similar to that of our model in which the exploitation–exploration balance is controlled by a single parameter, i.e. the inverse-temperature meta-parameter β. Actually, the sigmoidal transfer function (Servan-Schreiber et al., 1990) of behavioral neurons in the Usher's model is similar to the soft-max policy in our model, when the number of possible actions is two as in the visual discrimination task used in Usher et al. (lever release and hold).

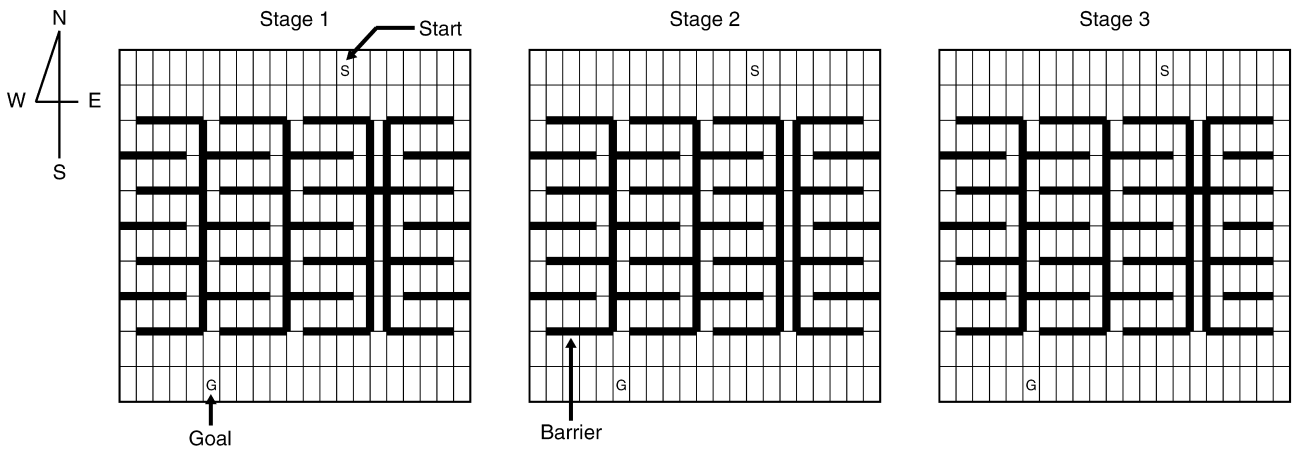


Fig. 9. A 'zig-zag' maze task. 'S' and 'G' denote the start point and the goal point, respectively. The thick line denotes bi-directional barriers. The objective of the agent is to find out the shortest path from the start point to the goal point.

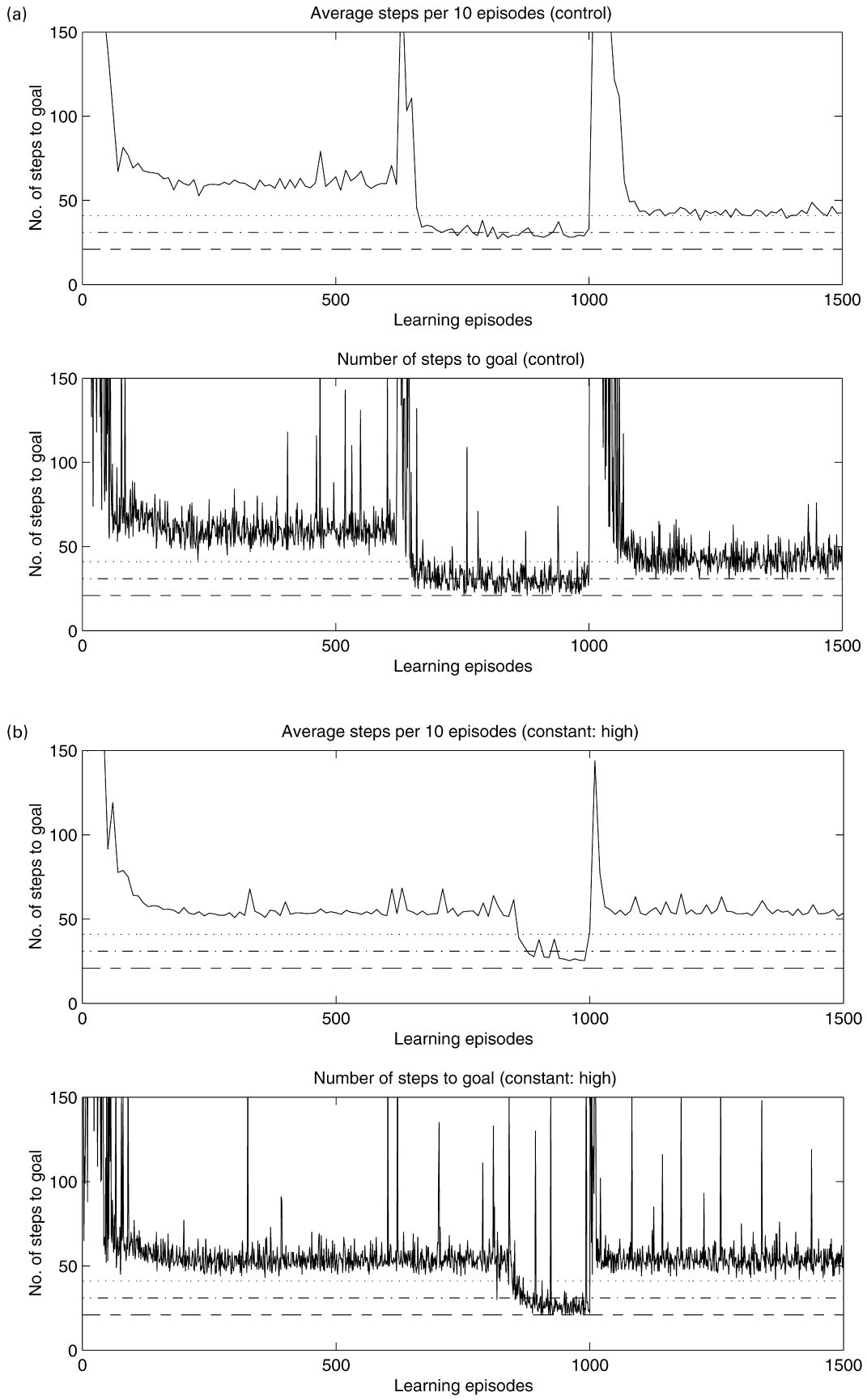

(a)
Average steps per 10 episodes (control)
No. of steps to goal
Learning episodes
Number of steps to goal (control)
No. of steps to goal
Learning episodes
(b)
Average steps per 10 episodes (constant: high)
No. of steps to goal
Learning episodes
Number of steps to goal (constant: high)
No. of steps to goal
Learning episodes

Fig. 10

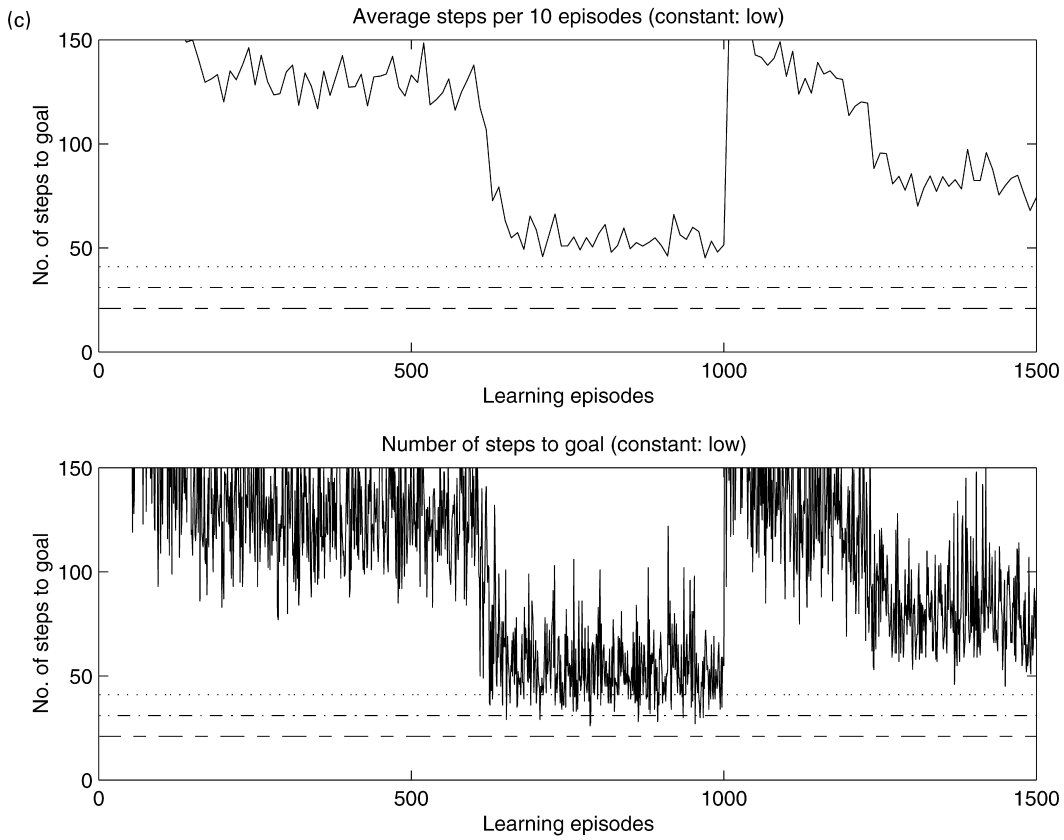


Fig. 10. Number of actions taken by the three agents. (a)–(c) The abscissa denotes the number of learning episodes. The ordinate denotes the number of actions averaged over 10 episodes in the upper figure, and the number of actions in each episode in the lower figure. The dotted, dash and dash-dotted lines denote the shortest path lengths at stages 1, 2 and 3, respectively. (a) Learning process of an agent with the control of the inverse-temperature. Parameters are set at $\kappa = 0.995$, $\beta_0 = 10$, $\beta_r = 0.0003$, $\alpha = 0.0003$, and $\epsilon = 8$. (b) Learning process of an agent with a large constant ($\beta = 100$) for the inverse-temperature. (c) Learning process of an agent with a small constant ($\beta = 0.85$) for the inverse-temperature.

Therefore, we assume that the control of the inverse-temperature is implemented as an activity pattern of LC neurons.

6.3. Calculation of inverse-temperature

Since projection from anterior cingulate cortex (ACC), in particular, Brodmann's area 24 and the rostral part of area 32, to the LC has recently been found (Rajkowski, Lu, Zhu, Cohen, & Aston-Jones, 2000), it is considered that the ACC regulates the activities of LC neurons, possibly by controlling the couplings among them.

The ACC (areas 24 and 32) is located on the medial surface of the frontal lobe and superior to the corpus callosum. Since the ACC is activated especially when the action selection requires a 'top-down' supervisory system, the ACC has been linked to the executive function of attention (Vogt, Finch, & Olson, 1992). Since the ACC is not activated in a vigilance task, which is used to validate sustained attention, the ACC is mainly involved in selective attention. A neuroanatomical study suggested that the ACC can be divided into different functional subregions (Picard & Strick, 1996), implying that the ACC has multiple functions (Bush et al., 2002). We assume here that both the local and the global coefficients of the inverse-temperature are calculated and represented in the ACC.

A recent study on single-cell recordings from monkeys performing a reward-based decision-making task reported that cingulate motor area (CMA)[5] has relevance to reward-based behaviors (Shima & Tanji, 1998). A monkey continued a particular behavior during constant-reward trials, while the reward decrement led to an active switching of the behavioral rule. It was found that neurons in the rostral CMA were saliently activated when a monkey switched the behavioral rule. Furthermore, the blocking of this area by muscimol injection induced a failure of smooth switching or a needless switching. These results suggest that the CMA plays an important role in behavior selection by detecting the distinction between an expected reward and an actual reward. A recent human neuroimaging study using a similar task also observed ACC activation (Bush et al., 2002). According to electrophysiological studies measuring error-related negativity (ERN; Gehring, Goss, Coles,

[5] The CMA of primates resides in the banks of the cingulate sulcus in the medial surface of the cerebral hemisphere and overlaps the ACC in humans.

Meyer, & Donchin, 1993), the ACC is involved in monitoring of errors. The ERN is a large negative polarity peak in an event-related potential waveform that occurs when a subject makes an error in a reaction time task. A recent imaging study also found that a rostral inferior ACC region is mainly related to the error detection (Braver, Barch, Gray, Molfese, & Snyder, 2001). These studies suggest that the ACC detects the environmental change using the result of own response, i.e. error or error prediction, so that changes in behavioral rules are induced. This control of the behavior selection is consistent with our control method of the global coefficient of the inverse-temperature.

On the other hand, several neuroimaging studies suggested that error-related ACC activities are likely due to detection of a conflict among incompatible responses (Botvinick, Nystrom, Fissell, Carter, & Cohen, 1999; Carter et al., 1998). This is called the conflict monitoring theory. While performing a visual discrimination task, i.e. a variation of continuous performance tests or a flanker task, the ACC exhibited transient activity increase during incorrect responses. However, greater ACC activity was also observed during correct responses in a situation with a high level of conflict. The ACC was also activated significantly in a novel environment. According to a positron emission tomography (PET) study on motor sequence learning, the ACC was activated during a learning of new sequences but not during an automatic execution after the learning (Jenkins, Brooks, Nixon, Frackowiak, & Passingham, 1994). A later study indicated that the ACC was activated more when a subject learned a new sequence than when the subject simply paid attention to a prelearned sequence (Jueptner et al., 1997). Thus, the ACC activities depend on the state of the subject. They are linked to the variation of possible results, i.e. the response conflict, and are related to reward-based learning processes especially in an unfamiliar environment.

In our RL scheme, the randomness due to the control of the local coefficient $\beta_1(s)$ is dependent on the agent's state, and it is large when the variation of the action-value function with respect to the current policy is large. The large variation of the action-value function is mainly due to the large variation of the policy, implying that the current state is unfamiliar or conflicting. Accordingly, the above-mentioned activities of the ACC seem to be consistent with our control method for the local coefficient of the inverse-temperature.

Based on the conflict monitoring theory, Cohen, Botvinick, and Carter (2000) presented a mechanism on how the ACC controls cognitive functions. If competing responses are simultaneously represented in the prefrontal cortex (PFC), the ACC detects a conflict. Subsequently, the LC system responds to the conflict detected by the ACC, and competitively suppresses the irrelevant representation activated by a distractor. The response conflict is thus reduced. Namely, this mechanism increases the level of selective attention when the ACC detects a conflict. However, the model by Cohen et al. focused on an exploitation operation and did not consider an active exploration operation, because they assumed a static environment, i.e. the task does not change.

Our study assumes dynamic environments, where a conflict occurs due not only to the forgetting of the environment but also to the environmental change. Similarly to the model by Cohen et al., in our method, the level of selective attention is controlled based on the conflict detection. However, the control depends on expectation of resultant value, i.e. prediction of consequence, of the conflict. If the resultant value will vary so much, the level of attention is rather decreased so that active exploratory behaviors are encouraged. As discussed earlier, the ACC is related to an active change of policy (Shima & Tanji, 1998), and we suggest that the system incorporating the ACC and the LC is related to inducing active exploratory behaviors.

6.4. Evaluation of environment

If the ACC evaluates the variation of the current action-value function, it should be provided with the evaluation of the environment. The ACC mainly receives innervations from the frontal association cortex (or PFC). It is known that connections of the cingulate cortices with other fronto-cortical areas are not limited to immediate neighbors, but also more distant prefrontal regions, particularly those in dorsolateral PFC (Barbas & Pandya, 1989).

The PFC receives sensory inputs processed by other association cortices, whereas the other association cortices directly receive sensory inputs. The major areas to which the PFC outputs are motor systems such as the striatum and the motor association cortex. The PFC is considered to direct various higher-order functions, e.g. decision-making, behavioral inhibition, planning of behavior, action evaluation, and maintenance of working memories. Since the PFC function cannot be explained by a unitary theory, the PFC should be divided into several functionally different subregions. Here, we introduce three important subregions: dorsolateral prefrontal cortex (DLPF, areas 9 and 46), orbitofrontal cortex (OFC, area 47/12) and anterior prefrontal cortex (APF, area 10). We speculate that the functions used in RL are expressed, maintained and learned within these brain regions.

6.4.1. DLPF/OFC and value function

Studies on the DLPF have been mainly focused on a working memory function, i.e. the active maintenance of necessary information for a certain period of time. Rao, Rainer, and Miller (1997) recorded activities of DLPF neurons from monkeys performing a visually

guided saccade task. First, a sample object was presented at the center of gaze. After a delay (the what delay), the sample object and a distractor were simultaneously presented at two different locations among four possible locations. After another delay (the where delay), the monkey was required to make a saccade to the remembered location where the sample had appeared. During this task, DLPF neurons showed sustained activities like the working memory during the what and/or where delay. In the what delay, on one hand, the monkey should keep a partial information to achieve its behavior. In the where delay, on the other hand, the monkey should keep the action to do. Thus, DLPF neurons exhibit state-dependent and action-dependent sustained activities (Hoshi, Shima, & Tanji, 2000).

Recent recording studies have revealed that DLPF neurons predict the quality and the quantity of the future reward (Leon & Shadlen, 1999; Watanabe, 1996). DLPF neurons of monkeys performing a delayed response task exhibited large activity when a preferred reward was expected, while the activity was small when a non-preferred reward was expected (Watanabe, 1996). Among such reward-dependent neurons, some were independent of the action to be selected, but the others were dependent on the action, i.e. which button to press (Watanabe, 1996). A later study using monkeys performing a memory-guided eye movement task showed that DLPF neurons exhibited larger activities when the monkey expected a larger reward (Leon & Shadlen, 1999). In this experiment, the monkey was informed in advance the amount of reward received by a successful completion of the task. The expected quantity of the reward also affected the success rate and the reaction time.

These experimental results imply that DLPF neurons are activated depending on state and/or action, and the activities represent the estimation of accumulated reward (total future reward), i.e. the value function or the action-value function in RL.

According to a recent view, the DLPF constructs automata, i.e. cascade networks representing transitions of states, in order to successively achieve a behavioral goal (Tanji & Hoshi, 2001). A physiological recording study using monkeys performing a delayed motor task investigated movement-related neuronal activities in the DLPF (Hoshi et al., 2000). The findings of neurons that were selectively active in different task phases showed that integration of movement information and behavioral planning are executed within an automaton in the DLPF. Since behavioral planning requires an environmental model, we assume that environmental models are, at least partly, expressed in the DLPF.

The OFC has dense connections with basolateral amygdala (ABL) and ventral tagmental area (VTA) which are involved in emotion and motivation functions. It is considered that the OFC is crucially involved in the motivational control of goal-directed behaviors (Rolls, 1996). For example, monkeys with an OFC damage showed performance impairment in an object-reversal task; the monkeys continued to respond to an object which was no longer rewarded (Meunier, Bachevalier, & Mishkin, 1997). A lesion study with humans also showed similar results (Rolls, Hornak, Wade, & McGrath, 1994). In addition, the OFC seems to have a role in monitoring rewards in order to select appropriate actions (Elliott, Dolan, & Frith, 2000). According to a study on neural activities of rats in an olfactory discrimination task, OFC neurons were activated selectively during the anticipation of rewarding or aversive outcomes (Schoenbaum, Chiba, & Gallagher, 1998). Furthermore, a functional magnetic resonance imaging (f-MRI) study using an emotion-related visual reversal-learning task found that the activation magnitude of the OFC was correlated with the magnitude of received rewards (O'Doherty, Kringelbach, Rolls, Hornak, & Andrews, 2001). These evidences suggest that the OFC is related to rapid stimulus–reward association learning, and we assume that the OFC maintains the evaluation of immediate or short-term accumulated rewards in order to execute a long-term planning.

6.4.2. APF and state estimation

Many PFC studies have concentrated on the posterior regions including the DLPF and the OFC, and there has been far less consideration to the APF. Using a branching task, in which the maintenance of a primary task was necessary while performing a subtask, Koechlin, Corrado, Pietrini, and Grafman (2000) showed that the APF was activated when a subject could not predict whether the forthcoming task would be the primary task or the subtask. Another imaging study using an explicit categorization task suggested that a rule change evoked an activation in the APF (Strange, Henson, Friston, & Dolan, 2001). These results enable us to make a speculation that the APF is involved in the prediction of a (significant) change of the environment. We assume that the estimation of unobservable states (environment) is related to the function of the APF.

Accordingly, we assume that the reward-based environmental model, i.e. the value function, the action-value function and the environmental model with the estimation of unobservable states, used in RL, are maintained in the PFC.

6.5. Dopaminergic system and novelty bonus

An animal placed in a novel environment is likely to display exploratory behaviors in order to analyze the new situation. Exploration of novel stimuli can be rewarding, and we have introduced in this study an exploration bonus added to the immediate reward.

Dopaminergic (DA) neurons of the VTA and substantia nigra have long been engaged on the processing of reward

stimuli. Recording studies using alert monkeys showed that DA neurons in the VTA were activated by stimuli associated with reward prediction and it was suggested that the neurons represent error in the prediction of future reward (Schultz et al., 1997). Since DA neurons were also activated when provided novel and salient stimuli (Schultz, 1998), signals transferred by DA neurons may be modified by the novelty bonus information.

DA neurons receive massive inputs from amygdala (Gonzales & Chesselet, 1990) that responds to primary rewards and reward-predicting stimuli. In addition, the amygdala responds to relatively novel stimuli (Wilson & Rolls, 1993) like the VTA. The amygdala is directly interconnected with the hippocampus that is involved in memory functions. Recent neuroimaging studies have shown that the hippocampal region is also critically involved in novelty detection (Stern et al., 1996; Tulving, Markowitsch, Craik, Habib, & Houle, 1996), and the memory system may provide the rewarding system with the information whether the current stimulus is novel or not. Therefore, it can be considered that novelty is added to the primary reward information represented in the amygdala.

Neurons in the ABL and the VTA directly project to nucleus accumbens (NAc) in the ventral striatum. The NAc may control motor systems via the ventral pallidum, and it is considered that this is a route through which limbic information is transferred to output systems (Pennartz, Groenewegen, & Lopez de Silva, 1994). By stimulating the ABL, animals tend to explore novel objects or situations. In addition, a local infusion into the NAc of drugs that release DA increased the magnitude of conditioned reinforcement in an operant task (Taylor & Robbins, 1986). DA innervations of the NAc is considered necessary for exploratory behaviors (Yim & Mogenson, 1989).

Thus, we currently assume that the amygdala associates the stimuli and its biological value including the novelty, i.e. the state and its value function (action-value function) modified by the exploration bonus, and that the system incorporating the ABL, the VTA and the NAc is related to producing exploratory behaviors. However, this assumption would need further discussion in the future.

7. Conclusion

This paper presented a new RL method in which the balance between exploitation and exploration is controlled. Our RL method is a model-based one in which the environment is estimated based on a Bayes inference. In the estimation, the forgetting of the environment encourages exploration. The exploitation–exploration balance is controlled by the inverse-temperature meta-parameter. The control is dependent on the agent's state; the dependence is due to the variation of the action-value function. The control is also dependent on the perception of the environmental changes. This method is one of the undirected exploration methods. The exploration bonus is also used as a directed exploration method. Our RL method is suitable especially when the environment is partially observable and dynamic. When applied to maze tasks, our method exhibited good adaptability to the environmental changes.

We also discussed a possible implementation in the brain. According to our assumption, the inverse-temperature is represented as the activity of the LC neurons, and the activity is controlled by the ACC. In order to achieve the control, the PFC maintains and provides the ACC with the value function, the action-value function and the environmental model. Accordingly, we consider that the control of randomness in RL is realized in the PFC–ACC–LC system and that it is related to selective attention.

Acknowledgments

The authors thank the reviewers for their valuable comments in improving this paper. The authors also thank Dr Kenji Doya and Dr Masa-aki Sato for their insightful suggestions concerning this study.

Appendix A

This Appendix section describes a detail of the Bayes inference method we use.

A Bayes inference considers the posterior distribution of the parameter. The Bayes theorem states that the posterior distribution is given by

$$P(g|Z) = \frac{P(Z|g)P(g)}{P(Z)}, \tag{A1}$$

where $P(g)$ is a prior distribution. $P(Z) \equiv \int P(Z|g)P(g)\mathrm{d}g$ is the normalization factor, which is called the marginal likelihood.

We prepare a trial posterior $Q(g)$ in order to approximate the posterior $P(g|Z)$. $Q(g)$ is determined based on the minimization of the following KL divergence between $Q(g)$ and the true posterior $P(g|Z)$.

$$\begin{aligned}\mathrm{KL}(Q|P) &\equiv \int Q(g)\log\frac{Q(g)}{P(g|Z)}\mathrm{d}g \\ &= \log P(Z) - \int Q(g)\log\frac{P(Z|g)P(g)}{Q(g)}\mathrm{d}g \\ &\equiv \log P(Z) - F(Q). \end{aligned} \tag{A2}$$

$F(Q)$ is called the variational free energy. Since $P(Z)$ does not depend on $Q(g)$, the minimization of the KL divergence is equivalent to the maximization of the

variational free energy $F(Q)$. The maximization is easily achieved by taking the variational condition: $\delta F/\delta Q = 0$. From the condition, the posterior distribution is analytically obtained as

$$\log Q(g) = \log P(Z|g) + \log P(g) + \text{const.}, \tag{A3}$$

where the constant term is determined by the distribution condition $\int Q(g)\mathrm{d}g = 1$.

If we assume a natural conjugate posterior distribution for parameter g, the posterior distribution becomes a Dirichlet distribution:

$$\begin{aligned} Q(g|\nu) &= \frac{\Gamma(\nu_1 + \cdots + \nu_M + M)}{\Gamma(\nu_1 + 1)\cdots\Gamma(\nu_M + 1)} g_1^{\nu_1}\cdots g_M^{\nu_M} \\ &\equiv \exp\left(\sum_{j=1}^{M} \nu_j \log g_j - \Phi(\nu)\right), \end{aligned} \tag{A4}$$

where ν is a hyperparameter.

$$\Phi(\nu) \equiv \sum_{j=1}^{M} \log \Gamma(\nu_j + 1) - \log \Gamma\left(\sum_{k=1}^{M} \nu_k + M\right), \tag{A5}$$

is the normalization term. $\Gamma(\cdot)$ is a Gamma function.

If no a priori knowledge on the prior distribution $P(g)$ is available, it is natural to choose a non-informative prior. In the multinomial model, a non-informative prior corresponds to regarding the $\log P(g)$ term in Eq. (A3) as constant. Therefore, Eq. (A3) becomes

$$\sum_{j=1}^{M} \nu_j \log g_j - \Phi(\nu) = T \sum_{j=1}^{M} \langle z_j \rangle_D \log g_j + \text{const.}, \tag{A6}$$

implying that

$$\nu_j = T\langle z_j \rangle_D, \tag{A7}$$

which is the exact Bayes solution for the multinomial model.

The entropy of the posterior distribution is obtained as follows. First we calculate

$$\begin{aligned} E_D[\log g_j] &\equiv \int Q(g|\nu) \log g_j \, \mathrm{d}g = \frac{\partial \Phi(\nu)}{\partial \nu_j} \\ &= \psi(\nu_j + 1) - \psi\left(\sum_{k=1}^{M} \nu_k + M\right), \end{aligned} \tag{A8}$$

where $\psi(x) \equiv \mathrm{d} \log \Gamma(x)/\mathrm{d}x$ is called the digamma function. Using Eq. (A8), the entropy of the Dirichlet posterior distribution $Q(g|\nu)$ is given by

$$\begin{aligned} H_D &\equiv -\int Q(g|\nu) \log Q(g|\nu) \mathrm{d}g = -\sum_{j=1}^{M} \nu_j E_D[\log g_j] + \Phi(\nu) \\ &= -\sum_{j=1}^{M} \nu_j \psi(\nu_j + 1) + \psi\left(\sum_{k=1}^{M} \nu_k + M\right) \sum_{j=1}^{M} \nu_j + \Phi(\nu). \end{aligned} \tag{A9}$$

References

Aston-Jones, G., Chiang, C., & Alexinsky, T. (1991). Discharge of noradrenergic locus coeruleus neurons in behaving rats and monkeys suggests a role in vigilance. *Progress in Brain Research*, *88*, 501–520.

Aston-Jones, G., Rajkowski, J., Kubiak, P., & Alexinsky, T. (1994). Locus coeruleus neurons in monkey are selectively activated by attended cues in a vigilance task. *The Journal of Neuroscience*, *14*, 4467–4480.

Barbas, H., & Pandya, D. N. (1989). Architecture and intrinsic connections of the pre-frontal cortex in the rhesus monkey. *The Journal of Comparative Neurology*, *286*, 353–375.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, *13*, 835–846.

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature*, *402*, 179–181.

Brafman, R.I., & Tennenholtz, M (2001). *R*-max: A general polynomial time algorithm for near-optimal reinforcement learning. *Proceedings of the 17th International Joint Conference on Artificial Intelligence* (pp. 953–958).

Braver, T. S., Barch, D. M., Gray, J. R., Molfese, D. J., & Snyder, A. (2001). Anterior cingulate cortex and response conflict: Effects of frequency, inhibition and errors. *Cerebral Cortex*, *11*, 825–836.

Bush, G., Vogt, B. A., Holmes, J., Dale, A. M., Greve, D., Jenike, M. A., & Rosen, B. R. (2002). Dorsal anterior cingulate cortex: A role in reward-based decision making. *Proceedings of the National Academy of Sciences, USA*, *99*, 507–512.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, *280*, 747–749.

Christie, M. J., Williams, J. T., & North, R. A. (1989). Electrical coupling synchronizes subthreshold activity in locus coeruleus neurons in vitro from neonatal rat. *The Journal of Neuroscience*, *9*, 3584–3589.

Cohen, J. D., Botvinick, M., & Carter, C. S. (2000). Anterior cingulate and prefrontal cortex: Who's in control? *Nature Neuroscience*, *3*, 421–423.

Dayan, P., & Sejnowski, T. J. (1996). Exploration bonuses and dual control. *Machine Learning*, *25*, 5–22.

Dearden, R., Friedman, N., & Andre, D. (1999). *Model based Bayesian exploration. Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kaufman, pp. 150–159.

Doya, K. (2000a). Reinforcement learning in continuous time and space. *Neural Computation*, *12*, 219–245.

Doya, K. (2000b). Metalearning, neuromodulation, and emotion. In G. Hatano, N. Okada, & H. Takabe (Eds.), *Affective minds* (pp. 101–104). Amsterdam: Elsevier.

Elliott, R., Dolan, R. J., & Frith, C. D. (2000). Dissociable functions in the medial and lateral orbitofrontal cortex: Evidence from human neuroimaging studies. *Cerebral Cortex*, *10*, 308–317.

Fe'ldbaum, A. A. (1965). *Optimal control systems*. New York, NY: Academic Press.

Foote, S. L., Bloom, F. E., & Aston-Jones, G. (1983). Nucleus locus coeruleus: New evidence of anatomical and physiological specificity. *Physiological Reviews*, *63*, 844–914.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*, 385–390.

Gonzales, C., & Chesselet, M.-F. (1990). Amygdalonigral pathway: An anterograde study in the rat with *Phaseolus vulgaris* leucoagglutinin (PHA-L). *The Journal of Comparative Neurology*, *297*, 182–200.

Heckerman, D. (1999). A tutorial on learning with Bayesian networks. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 301–354). Cambridge, MA: MIT Press.

Hoshi, E., Shima, K., & Tanji, J. (2000). Neuronal activity in the primate prefrontal cortex in the process of motor selection based on two behavioral rules. *Journal of Neurophysiology*, *83*, 2355–2373.

Ishimaru, M., & Williams, J. T. (1996). Synchronous activity in locus coeruleus results from dendritic interactions in pericoerulear regions. *The Journal of Neuroscience*, *16*, 5196–5204.

Jenkins, I. H., Brooks, D. J., Nixon, P. D., Frackowiak, R. S. J., & Passingham, R. E. (1994). Motor sequence learning: A study with positron emission tomography. *The Journal of Neuroscience*, *14*, 3775–3790.

Jueptner, M., Stephan, K. M., Frith, C. D., Brooks, D. J., Frackowiak, R. S., & Passingham, R. E. (1997). Anatomy of motor learning. I. Frontal cortex and attention to action. *The Journal of Neurophysiology*, *77*, 1313–1324.

Kaelbling, L. (1993). *Learning in embedded systems*. Cambridge, MA: MIT Press.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*, 99–134.

Kearns, M., & Singh, S. (1998). *Near-optimal performance for reinforcement learning in polynomial time*. *Proceedings of the 15th International Conference on Machine Learning*, San Mateo, CA: Morgan Kaufmann, pp. 260–268.

Koechlin, E., Corrado, G., Pietrini, P., & Grafman, J. (2000). Dissociating the role of the medial and lateral anterior prefrontal cortex in human planning. *Proceedings of the National Academy of Sciences, USA*, *97*, 7651–7656.

Leon, M. I., & Shadlen, M. N. (1999). Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. *Neuron*, *24*, 415–425.

Matsuno, Y., Yamazaki, T., Matsuda, J., & Ishii, S. (2001). *A multi-agent reinforcement learning method for a partially-observable competitive game*. *Proceedings of the Fifth International Conference on Autonomous Agents*, New York, NY: ACM, pp. 39–40.

Meunier, M., Bachevalier, J., & Mishkin, M. (1997). Effects of orbital frontal and anterior cingulate lesions on object and spatial memory in rhesus monkeys. *Neuropsychologia*, *35*, 999–1015.

Moore, A. W., & Atkeson, C. G. (1993). Prioritized sweeping: Reinforcement learning with less data and less real time. *Machine Learning*, *13*, 103–130.

Morrison, J., & Foote, S. (1986). Noradrenergic and serotonergic innervation of cortical, thalamic and tectal visual structures in old and new world monkeys. *The Journal of Comparative Neurology*, *243*, 117–128.

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience*, *4*, 95–102.

Pennartz, C. M., Groenewegen, H. J., & Lopez de Silva, F. H. (1994). The nucleus accumbens as a complex of functionally distinct neuronal ensembles: An integration of behavioural, electrophysiological and anatomical data. *Progress in Neurobiology*, *42*, 719–761.

Picard, N., & Strick, P. L. (1996). Motor areas of the medial wall: A review of their location and functional activation. *Cerebral Cortex*, *6*, 342–353.

Posner, M. I., & Raichle, M. (1996). *Images of mind*. Washington, DC: Scientific American Books, revised.

Rajkowski, J., Lu, W., Zhu, Y., Cohen, J., & Aston-Jones, G. (2000). Prominent projections from the anterior cingulate cortex to the locus coeruleus in rhesus monkey. *Society of Neuroscience Abstract*, *26*, 2230.

Rao, S. C., Rainer, G., & Miller, E. K. (1997). Integration of what and where in the primate prefrontal cortex. *Science*, *276*, 821–824.

Rolls, E. T. (1996). The orbitofrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological sciences*, *351*, 1433–1443.

Rolls, E. T., Hornak, J., Wade, D., & McGrath, J. (1994). Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *Journal of Neurology, Neurosurgery, and Psychiatry*, *57*, 1518–1524.

Sato, M. (2001). On-line model selection based on the variational Bayes. *Neural Computation*, *13*, 1649–1681.

Schoenbaum, G., Chiba, A. A., & Gallagher, M. (1998). Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nature Neuroscience*, *1*, 155–159.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *The Journal of Neurophysiology*, *80*, 1–27.

Schultz, W., Dayan, P., & Montague, R. P. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Servan-Schreiber, D., Printz, H., & Cohen, J. D. (1990). A network model if catecholamine effects: Gain, signal-to-noise ratio, and behavior. *Science*, *249*, 892–895.

Shima, K., & Tanji, J. (1998). Role for cingulate motor area cells in voluntary movement selection based on reward. *Science*, *282*, 1335–1338.

Singh, S. P., Jaakkola, T., & Jordan, M. I. (1994). *Learning without state-estimation in partially observable Markovian decision processes*. *Proceedings of the 11th International Conference on Machine Learning*, San Francisco, CA: Morgan Kaufmann, pp. 284–292.

Stern, C. E., Corkin, S., Gonzalez, R. G., Guimaraes, A. R., Baker, J. R., Jennings, P. J., Carr, C. A., Sugiura, R. M., Vedantham, V., & Rosen, B. R. (1996). The hippocampal formation participates in novel picture encoding: Evidence from functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences, USA*, *93*, 8660–8665.

Strange, B. A., Henson, R. N. A., Friston, K. J., & Dolan, R. J. (2001). Anterior prefrontal cortex mediates rule learning in humans. *Cerebral Cortex*, *11*, 1040–1046.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*, 9–44.

Sutton, R.S (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Machine Learning: Proceeding of the Seventh International Conference* (pp. 216–224).

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tanji, J., & Hoshi, E. (2001). Behavioral planning in the prefrontal cortex. *Current Opinion in Neurobiology*, *11*, 164–170.

Taylor, J. R., & Robbins, T. W. (1986). 6-Hydroxydopamine lesions of the nucleus accumbens, but not of the caudate nucleus, attenuate enhanced responding with reward-related stimuli produced by intra-accumbens d-amphetamine. *Psychopharmacology*, *90*, 390–397.

Thrun, S. B. (1992). *The role of exploration in learning control*. *Handbook of intelligent control: Neural, fuzzy and adaptive approaches*, Frorence, KY: Van Nostrand Reinhold.

Tulving, E., Markowitsch, H. J., Craik, F. E., Habib, R., & Houle, S. (1996). Novelty and familiarity activations in PET studies of memory encoding and retrieval. *Cerebral Cortex*, *6*, 71–79.

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., & Aston-Jones, G. (1999). The role of locus coeruleus in the regulation of cognitive performance. *Science*, *283*, 549–554.

Vogt, B. A., Finch, D. M., & Olson, C. R. (1992). Functional heterogeneity in cingulate cortex: The anterior executive and posterior evaluative regions. *Cerebral Cortex*, *2*, 435–443.

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, *412*, 43–48.

Watanabe, M. (1996). Reward expectancy in primate prefrontal neurons. *Nature*, *382*, 629–632.

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

Wilson, F. A. W., & Rolls, E. T. (1993). The effects of stimulus novelty and familiarity on neuronal activity in the amygdala of monkeys performing recognition memory tasks. *Experimental Brain Research*, *93*, 367–382.

Yim, C. Y., & Mogenson, G. J. (1989). Low doses of accumbens dopamine modulate amygdala suppression of spontaneous exploratory activity in rats. *Brain Research*, *477*, 202–210.