博士論文

論文題目　**Effects of Information Trace to Cerebral Learning**
(情報トレースの脳の学習への効果**)**

氏名　齋藤　大

Effects of Information Trace to Cerebral Learning

Doctor's Thesis Presented

by

Hiroshi Saito

Supervised by

Professor, Masato Okada

Submitted to

the Graduate School of Frontier Sciences, the University of Tokyo

in partial fulfilment of the requirements for the degree of

DOCTOR OF SCIENCE

February, 2014

Department of Complexity Science and Engineering

Abstract

Humans and animals can adapt their behaviors or responses to a ever-changing environment. The brain's learning ability plays a principal role on the adaptation and underlies important brain functions such as motor control, memory retention and retrieval, future prediction, decision making, language acquisition and so on. The brain preserves various kind of information in different time scales in the forms of neural activities, strength of plastic synapses and distribution of neurotransmitters. It has been believed that the cerebral learning is mainly realized by the plastic synapses. However, the other types of information can have measurable effect to the learning.

In this dissertation, we theoretically investigate how various traces of information affect to the cerebral learning. Although many of theoretical studies assume the stability of a learning system except adaptive parameters between trials in learning, we consider the effect of past trials by introducing information trace. We focus on three major research topics in learning. First, we analyze a neural network model which can solve the structural and temporal credit assignment problems of reward. We clarify how the eligibility trace, in which is a kind of information trace, resolves these problems. Second, a heavy-tailed reward trace dependency of choice which is observed in matching behavior of monkey is accounted via Bayesian decision making models. A possible computational principle behind the reward trace dependency is proposed. Third, we analyze a neural network model which can learn the transition probabilities of states in environment and can predict future state. We discuss how the eligibility trace makes the learning robust to abrupt change of environment.

Acknowledgments

First, my thanks goes to my supervisor Prof. Masato Okada, as he gave me helpful comments and suggestions. He also taught me a range of things important to work on as a researcher. I also owe my deepest gratitude to Dr. Kentaro Katahira who has been supporting my study patiently for five years, and Dr. Ken Takiyama who has been passionately discussing with me. Without their advice, direction and encouragement, I would never have completed this study. The responsibility for the final formulation, and any errors that it may concern, are entirely mine.

I want to express my deepest gratitude to all my colleagues and friends for insightful discussion, invaluable suggestions and everyday affairs. Dr. Masafumi Oizumi effectively provided me the opportunity to belong Okada laboratory and taught me Go. Yasuhiko Igarashi initiated me to some interesting books and I've learned something about the working world from his attitudes. Yosuke Otsubo taught me the significance of language and offered a glimpse into the profound scientific history. Hiroki Terashima gave me a chance to learn about Matcha and showed me ideal attitude of intellectual figure.

I especially wish to thank Noriko Aramaki who has devotedly supported my study and gave me peace of mind. I would also like to express my gratitude to my family for their moral support over twenty-eight years.

This work was partially supported by a Grant-in-Aid from the Japan Society for the Promotion of Science (JSPS) Fellows of the Ministry of Education, Culture, Sports, Science and Technology (No. 11J06433).

Contents

Chapter 1

Introduction

The brain can be regarded as a biological computational system which has side-effects. The computation with side-effects produces varied output even if identical input is given [Hughes, 1989], and hence investigation of the underlying computation becomes difficult. One cause of the side-effects is existence of internal states in a computational system. The brain's internal states are represented by membrane potential of neuron, connection weight of plastic synapse, density distribution of neurotransmitter and other types of transient biophysical matters. The mass of such internal states has been supposed to ingenerate diverse brain functions such as perception, motor control, memory retention and retrieval, future prediction, decision making, language acquisition and so on. To elucidate the underlying computation of the brain, it is important to understand how those internal states evolve.

This dissertation focuses on cerebral learning which is a typical example of internal state change in the brain. There are mainly two forms of behavioral learning: classical conditioning and operant conditioning. The classical conditioning is that neutral stimulus (conditioned stimulus; CS) become predictive signal for stimulus eliciting innate response (unconditioned stimulus; US) after US are presented following CS repeatedly [Pavlov, 1927]. There are a lot of evidences that the association between US and CS is constructed inside the brain [Mazur, 2002]. The operant conditioning is phenomenon that the association between a stimulus and a response is strengthened when a subject receives a reinforcer after the response to the stimulus [Skinner, 1938]. It seems that the classical conditioning is reflexive learning and the operant conditioning is the learning of spontaneous behavior. Because a schedule of reinforcement significantly affects the operant conditioning, various reinforcement schedules have been proposed. The foraging schedules which

resemble animal's foraging environment have been actively studied from 1960s [Chung & Herrnstein, 1967, Glimcher, 2004]. The schedule extends the deterministic reward (reinforcement) delivery in conventional operant conditioning procedure; the amount of reward and an interval between rewards is determined stochastically. In task with foraging schedule (foraging task), the behavior of humans and animals is proposed to obey the matching law (matching behavior) [Herrnstein, 1961]. The matching behavior is important for understanding decision making behaviors and cerebral learning because the matching behavior deviates from economic behavior, i.e., maximization of future total reward. It is necessary to consider not only the macroscopic phenomenon such as behavior but also the microscopic components of the brain architecture to understand the cerebral learning. The neuron is a kind of electrical unit that emits a electrical pulse, called spike, when its membrane potential exceeds a certain threshold. Neurons release neurotransmitters from axon terminal after a spike and they are received by another neuron. This biochemical connection is known as synapse and the strength of the connection (synaptic weight) is increased when connected neurons fired simultaneously [Hebb, 1949]. In recent studies, it is proposed that precise temporal order of spikes are important for strengthening and weakening of synaptic weight [Levy & Steward, 1983, Dan & Poo, 2004]. The synaptic plasticity has been believed to play a principle role in cerebral learning. However, there are other kind of components which preserve information brain and preserved information has various lifetimes from milliseconds to months [Abraham, 2003]. In a learning task, the trace of information on a time scale of several trials should have significant effects to learning.

An intelligent computer which resembles or rather exceeds the brain capabilities has been steadily expected to be developed from midst of a significant evolution of computer science. In the earliest stage, McCulloch and Pitts proposed an artificial neuron model which resembles neural integration and firing [McCulloch & Pitts, 1943]. This model is extended to acquire unknown mapping between input and its binary category by adjusting its synaptic weights, known as *perceptron* [Rosenblatt, 1958]. Although single layer perceptrons only have a capability to approximate the linearly separable boolean-valued functions, it was proved that multilayer perceptrons can approximate arbitrary continuous function to any desired accuracy [Irie & Miyake, 1988, Hornik et al., 1989, Cybenko, 1989, Funahashi, 1989, Hornik, 1993]. Various artificial neural network models have been devised for engineering application such as pattern recognition and regression [Bishop, 1995, Bishop, 2006], and for elucidating underlying processes of the brain [Tsodyks et al., 1998, Brunel & Hakim, 1999, Wang, 2002,

Seung, 2003, Fusi et al., 2007]. Reinforcement learning, which can be one of a rigorous mathematical formulation of the operant conditioning, had been diligently studied in 1980s [Sutton, 1984, Tesauro, 1995, Kaelbling et al., 1996, Barto, 1998, Sutton & Barto, 1998]. In reinforcement learning, an agent in some uncertain environment is expected to acquire appropriate behavior which maximizes future total reward by trial and error. The environment is formulated by states of the environment, transition probabilities between states for an action, and a mapping from states/actions and reward. An agent is expected to learn these components either implicitly or explicitly. An agent falls into exploration-exploitation dilemma due to two exclusive objectives: exploration of the environment for learning and exploitation of learned knowledge for reward maximization [Gittins & Jones, 1974, Daw et al., 2006]. Besides, temporal and structural credit assignment problems occur if a reward depends on past states and actions, and an agent is consisted on several components respectively, i.e., the contribution of each state, each action and each agents for reward should be determined for learning [Minsky, 1961, Sutton, 1984]. Same issues should be considered also in the cerebral learning.

1.1 Main Contributions

The works presented in this dissertation have been already published in international journals or presented at conferences. In this dissertation, we reviewed them from the perspective of cerebral learning with trial-wise information trace in macroscopic and microscopic scales. Our main contributions are as follows:

- **Structural and temporal credit assignment problems on learning in neural networks**. We analyzed macroscopic learning dynamics of microscopic neural network model with eligibility trace by statistical mechanics approach. The network consists of several linear perceptrons and a feedback signal delays. Thus, there are structural and temporal credit assignment problems. We elucidated the joint effect of the structural-temporal uncertainties to learning and quantitative effects of the eligibility trace to the learning (Chapter 2).

  Original contents are found in [Saito et al., 2010, Saito et al., 2011].

- **Bayesian account for the matching law and the heavy-tailed reward trace dependency**. We investigated macroscopic Bayesian deterministic decision making models to elucidate the computational principle underlying the undermatching and the heavy-tailed reward trace dependency. Our models show

matching behavior although they have incorrect but conceivable postulate for reward delivery. The undermatching and the heavy-tailed reward trace dependency are observed when we introduced an belief about environment volatility. We proposed that the reward trace may be an effective implementation of the belief (Chapter 3).

Original contents are found in [Saito et al., 2014].

- **Future predicting neural network model**. We proposed a microscopic neural network model which can learn the state transition probabilities and can predict future states. Our learning algorithm is based on Hebb rule and activity-dependent weight decay. The neural activities and behaviors of our model resembles those of monkeys' in a randomdot motion stimulus discrimination task. We investigated the effect of the eligibility trace to the learning of transition probabilities and adaptation for abrupt change of environment (Chapter 4).

1.2 Summary of the remaining chapters

This dissertation is structured as follows. Effects of the eligibility trace to neural network learning under structural and temporal credit assignment problems are described in Chapter 2. In Chapter 3, we show that a belief of environment volatility is essential for matching behavior by analyzing Bayesian decision making models. A possible interpretation of reward trace dependency of monkey's choice is given. Chapter 4 presents the effect of input trace to learning of neural network for future state prediction. Finally, we conclude this dissertation in Chapter 5.

Chapter 2

Neural Network Model under Structural-Temporal Uncertainties of Reward

Reward may not be well-informative for an adaptive agent such as neural networks. Generally, there is a latency of reward delivery (temporal uncertainty) and a reward would not tell how each modifiable parameter of an agent should be updated (structural uncertainty). For learning, an agent is necessary to resolve the structural and temporal credit assignment problems caused by these uncertainties. We investigate effects of eligibility trace to structural-temporal uncertainties by analyzing a neural network model which can solve the credit assignment problems by the mechanism of eligibility trace and the node perturbation learning.

2.1 Introduction

Adaptive systems, including the brain, may be received a little information to accomplish a learning. In reinforcement learning, a system modulates its parameters according to a scalar instruction signal delivered from environment. Learning will be a relatively easy task if a system knows how an evaluation mechanism of the environment depends on its behavior in advance. Hence, a system has to estimate the evaluation mechanism, *objective function* in other words [Kaelbling et al., 1996, Sutton & Barto, 1998]. There are two types of credit assignment problems [Minsky, 1961, Sutton, 1988] in this estimation. A system

needs to determine which past behaviors or activities are significant for the evaluation when the objective function depends on the time sequence of system's behavior. In other words, the system should solve the *temporal credit assignment problem* [Sutton, 1984, Houk, 2005]. The *structural credit assignment problem* [Chapman & Kaelbling, 1991, Legenstein et al., 2009] arises if a system has many modifiable parameters; the system needs to know how each parameter depends on the evaluation. The brain also should solve these credit assignment problems.

The delay of an instruction signal, e.g., reward, is a simple instance of temporal credit assignment problems. In the field of neuroscience, this is classically known as "distal reward problem"; how a brain distinguishes the relation between a reward and the neuronal activities that triggered the reward [Hull, 1943]. In the field of reinforcement learning, temporal difference learning [Sutton, 1988] and the eligibility trace [Klopf, 1972, Singh & Sutton, 1996] are proposed to cope with such a temporal uncertainty of evaluation. The eligibility trace accumulates the recent states or actions (events) of a system and quantifies how eligible they are for a given reward, i.e., more credit is given to more recent events and events that have occurred many times. The distal reward problem is proposed to be solved by a network of spiking neurons with the eligibility trace [Izhikevich, 2007].

Structural credit assignment problems, also known as *generalization problem* [Lin, 1992], are related to gradient estimation of the objective function. Hence, several levels of structural credit assignment problems arise depending on the groups of related parameters such as across synaptic weights, multi-agent outputs, temporal parameters, and so on. There are two approaches to estimate the gradient: *model-based* and *model-free*. In the former approach, a model is trained and used for the estimation. The estimation is reliable if the model suits to the objective function. In contrast, the latter approach does not require specific model. Model-free methods are applicable if the explicit form of the objective function is not known in advance such as the case of reinforcement learning. Therefore, it is expected to be a plausible mechanism of flexible learning systems, such as the brain [Dembo & Kailath, 1990, Xie & Seung, 2004, Fiete & Seung, 2006, Legenstein et al., 2009]. A simple model-free method is to estimate the gradient by the deviation of the objective function value caused by introducing perturbation into parameters. There are two variants of these stochastic gradient following methods for neural networks. One is the node perturbation [Widrow & Lehr, 1990, Flower & Jabri, 1993, Werfel et al., 2005, Katahira et al., 2010, Hara et al., 2011] which induces a perturbation into an output of a network, and the other is weight

perturbation [Jabri & Flower, 1992] which perturbs synaptic weights independently. In the weight perturbation scheme, the performance degrades and the learning speed is slower than that of node perturbation when the network size increases [Werfel et al., 2005]. The node perturbation is proposed to be an underlying mechanism of song learning in songbird [Fiete et al., 2007]. Thus, the brain may employ the node perturbation as a solution for structural credit assignment problems in neural network level.

In this chapter, we investigate effects of the eligibility trace to learning under structural-temporal credit assignment problems by analyzing adaptive linear neural network models. Learning task is formulated as a student network imitates an output of a teacher network by a delayed scalar feedback signal (reward). Student neural network model employs the node perturbation learning and the eligibility trace to cope with the uncertain reward. First, we study effects of temporal uncertainty alone, i.e., the network consists of a linear perceptron (section 2.2). Quantitative effects of the eligibility trace are shown by studying derived macroscopic behaviors of our model. Then, we study joint effects of structural-temporal uncertainties by extending our model to multi-perceptrons (section 2.3). It is shown that both the structural and the temporal uncertainties influence the convergence of learning and change the optimal time constant of the eligibility trace non-linearly.

2.2 Model without Structural Uncertainty

In this section, we analyze a linear perceptron with the eligibility trace to elucidate the quantitative effects of temporal uncertainty of reward to learning.

2.2.1 Model

Both of student and teacher neural networks consist of N input units and one output unit (Fig. 2.1). The i-th input unit is connected to the output unit with a synaptic weight J_i and B_i for the student and teacher networks respectively. Each teacher synaptic weight B_i is drawn from a standard Gaussian distribution, then normalized as $\|\boldsymbol{B}\| = \sqrt{N}$. The activity of i-th input unit of both networks, $x_i(m)$, obeys a Gaussian distribution of mean 0 and variance $1/N$ where m ($m \in \{0, 1, \dots\}$) represents a time step. The activities of output units are determined by weighted

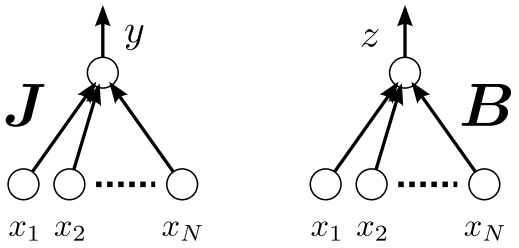


Figure 2.1: Schematic of our neural network models. The left is the student network and the right is the teacher.

sums of the inputs. The student output y and the teacher output z are

$$y(m) = \boldsymbol{J}(m) \cdot \boldsymbol{x}(m), \tag{2.1}$$
$$z(m) = \boldsymbol{B} \cdot \boldsymbol{x}(m), \tag{2.2}$$

where a bold symbol represents a N-dimensional vector and the binary operator $(\cdot)$ is the inner product. The objective function is defined by a squared error between the outputs:

$$E(m) = \frac{1}{2}(y(m) - z(m))^2. \tag{2.3}$$

The student is expected to minimize this error. Thus, the synaptic weights of the student should be updated to descend the gradient of Eq.(2.3):

$$\frac{\delta}{\delta \boldsymbol{J}} E(m) = (y(m) - z(m))\boldsymbol{x}(m). \tag{2.4}$$

However, we assume that the student network cannot calculate the gradient directly. Instead, it employs the node perturbation learning to update its synaptic weights [Widrow & Lehr, 1990]. Node perturbation introduces a perturbation into the network output and a perturbed objective function value is obtained:

$$E_{NP}(m) = \frac{1}{2}(y(m) + \xi(m) - z(m))^2. \tag{2.5}$$

The perturbation $\xi(m)$ is sampled from a Gaussian distribution of mean 0 and variance σ^2. Then the student can obtain an approximate gradient by the deviation from the value without the perturbation. Thus, an instruction signal (reward) is defined as:

$$\begin{aligned} d(m) &= -(E_{NP}(m) - E(m)) \\ &= -\frac{1}{2}\left[\xi^2(m) + 2\xi(m)(y(m) - z(m))\right]. \end{aligned} \tag{2.6}$$

Let us confirm that the student can obtain the gradient by using the reward in the standard form of the node perturbation learning rule. The standard form is

$$\begin{aligned} \boldsymbol{J}(m+1) &= \boldsymbol{J}(m) + \Delta\boldsymbol{J}^{\text{std}}(m), & (2.7)\\ \Delta\boldsymbol{J}^{\text{std}}(m) &= \eta d(m)\xi(m)\boldsymbol{x}(m) \\ &= -\frac{1}{2}\eta\xi^3(m)\boldsymbol{x}(m) - \eta\xi^2(m)(y(m)-z(m))\boldsymbol{x}(m), & (2.8) \end{aligned}$$

where η is a learning coefficient. The ensemble average of $\Delta\boldsymbol{J}^{\text{std}}(m)$ becomes:

$$\langle\Delta\boldsymbol{J}^{\text{std}}(m)\rangle \propto \langle(y(m)-z(m))\boldsymbol{x}(m)\rangle, \tag{2.9}$$

where $\langle\cdot\rangle$ describes an ensemble average. Thus, $\Delta\boldsymbol{J}^{\text{std}}(m)$ can averagely reconstruct the true differential (Eq. (2.4)).

Here, we extend the standard node perturbation learning rule of Eq.(2.7) to deal with unknown delay of reward. The student requires past ξ and $\boldsymbol{x}$ to extract a gradient as well as Eq. (2.8) from a delayed reward. The eligibility trace can be used to preserve each of past states with temporal credit assignment. The eligibility trace $\boldsymbol{e}$ is defined as:

$$\boldsymbol{e}(m) = \sum_{k=0}^{\infty}\varepsilon(k)\xi(m-k)\boldsymbol{x}(m-k), \tag{2.10}$$

where $\varepsilon(t) = \exp(-t/\tau)$ is a kernel and τ is a time constant. The kernel represents the credit assigned to the past states and also the retention period of the past states; a large τ gives more credit to far past states and preserves the past states longer. Thus, our learning rule is defined with the eligibility trace:

$$\begin{aligned} \boldsymbol{J}(m+1) &= \boldsymbol{J}(m) + \eta\tilde{d}(m)\boldsymbol{e}(m), & (2.11)\\ \tilde{d}(m) &= d(m-m_d), & (2.12) \end{aligned}$$

where $\eta > 0$ is a learning coefficient and m_d represents the delay of the reward. We assume that η is sufficiently small for later analysis.

2.2.2 Analysis

Statistical mechanics approach has been used to derive the macroscopic dynamics of learning from microscopic learning rule [Watkin et al., 1993, Saad, 1999, Kinzel et al., 2001, Biehl et al., 2009, Katahira et al., 2010, Hara et al., 2011]. By

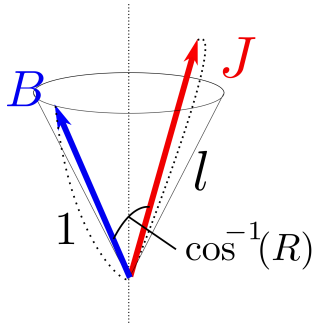


Figure 2.2: Schematic of the macroscopic variables using three dimensional vectors of student and teacher synaptic weights.

seeing in a macroscopic perspective, a convergence condition, learning speed and performance, and also effects of the (microscopic) eligibility trace can be discussed.

First, we define several macroscopic variables that describe the gross state of the system. The relative length of $\boldsymbol{J}$ against $\boldsymbol{B}$ is represented by a macroscopic variable l:

$$l(m) \equiv \frac{\|\boldsymbol{J}(m)\|}{\|\boldsymbol{B}\|} = \frac{\|\boldsymbol{J}(m)\|}{\sqrt{N}}. \tag{2.13}$$

The probability distribution of J_i is unknown but we assume $J_i \sim \mathcal{O}(1)$ since J_i approaches to $B_i \sim \mathcal{O}(1)$. From this assumption, we obtain $\|\boldsymbol{J}\| \sim \mathcal{O}(\sqrt{N})$ and hence $l \sim \mathcal{O}(1)$. A macroscopic variable R represents the overlap of $\boldsymbol{J}$ and $\boldsymbol{B}$,

$$R(m) \equiv \frac{\boldsymbol{J}(m) \cdot \boldsymbol{B}}{\|\boldsymbol{J}(m)\|\|\boldsymbol{B}\|} = \frac{1}{l(m)N}\boldsymbol{J}(m) \cdot \boldsymbol{B}. \tag{2.14}$$

Since R depends on l, we remove this dependency by defining the macroscopic variable r as follows.

$$r(m) \equiv l(m)R(m) = \frac{1}{N}\boldsymbol{J}(m) \cdot \boldsymbol{B}. \tag{2.15}$$

Obviously, $r \sim \mathcal{O}(1)$.

We derive time evolution equations of the macroscopic variables in the thermodynamic limit of $N \to \infty$. By squaring both sides of the Eq.(2.11), we obtain a microscopic update rule for the macroscopic variable l:

$$Nl^2(m+1) = Nl^2(m) + 2\eta\tilde{d}(m)\boldsymbol{J}(m) \cdot \boldsymbol{e}(m) + \eta^2\tilde{d}^2(m)\|\boldsymbol{e}(m)\|^2. \tag{2.16}$$

In Eq.(2.16), the change of l^2 at one update is only $\mathcal{O}(1/N)$. Therefore, we consider a continuous time $t = m/N$ to observe the change of $\mathcal{O}(1)$. The change from t to $t + dt$ for a small dt corresponds to the change from m to $m + Ndt$. We iteratively substitute the Eq. (2.16), and we obtain

$$l^2(m+Ndt) - l^2(m) = \frac{1}{N}\sum_{k=0}^{Ndt-1}\left\{\begin{array}{l} 2\,\eta\,\tilde{d}(m+k)\boldsymbol{J}(m+k)\cdot\boldsymbol{e}(m+k) \\ +\,\eta^2\tilde{d}^2(m+k)\|\boldsymbol{e}(m+k)\|^2 \end{array}\right\}. \tag{2.17}$$

Within the short period dt that spans the $\mathcal{O}(N)$ update, the weight change of $\mathcal{O}(1/N)$ can be neglected, and the self-averaging property holds. Hence, the summation of the above equation can be replaced by ensemble average. Thus, the evolution of l^2 obeys an ordinary differential equation:

$$\frac{dl^2}{dt} \simeq 2\eta\langle\tilde{d}\boldsymbol{J}\cdot\boldsymbol{e}\rangle + \eta^2\langle\tilde{d}^2\|\boldsymbol{e}\|^2\rangle. \tag{2.18}$$

Similarly, Eq.(2.11) multiplied by $\boldsymbol{B}$ is

$$Nr(m+1) = Nr(m) + \eta\tilde{d}(m)\boldsymbol{B}\cdot\boldsymbol{e}(m). \tag{2.19}$$

Therefore,

$$\frac{dr}{dt} = \eta\boldsymbol{B}\cdot\langle\tilde{d}\boldsymbol{e}\rangle. \tag{2.20}$$

The ensemble averages becomes (see appendix A for details of derivation),

$$\boldsymbol{B}\cdot\langle\tilde{d}\boldsymbol{e}\rangle = \sigma^2\epsilon(m_d)(1-r(t)), \tag{2.21}$$

$$\langle\tilde{d}\boldsymbol{J}\cdot\boldsymbol{e}\rangle = \frac{1}{4}\eta\sigma^6\left[D_1 S + 2F\right] - \sigma^2\varepsilon(m_d)(l^2(t)-r(t)), \tag{2.22}$$

$$\langle\tilde{d}^2\|\boldsymbol{e}\|^2\rangle = \frac{3}{4}\sigma^6 E_1 + \sigma^4 D_1(l^2(t) - 2r(t) + 1), \tag{2.23}$$

where the constants are

$$S \equiv \sum_{p=1}^{\infty}\varepsilon(p), \quad I \equiv \sum_{p=0}^{\infty}\varepsilon^2(p), \quad F \equiv \varepsilon(m_d)\sum_{p=0}^{m_d-1}\varepsilon(p),$$
$$D_k \equiv 2\varepsilon^2(m_d) + kI, \quad E_k \equiv 4\varepsilon^2(m_d) + kI.$$

By using Eqs. (2.21), (2.22) and (2.23), the differential equations of Eqs. (2.18) and (2.20) become

$$\frac{dl^2}{dt} = -H_1 l^2(t) + 2(H_1 - \eta\sigma^2\varepsilon(m_d))r(t) + (2\eta\sigma^2\varepsilon(m_d) - H_1) + \eta^2 G_1 \tag{2.24}$$

$$\frac{dr}{dt} = \eta\sigma^2\epsilon(m_d)(1-r(t)), \tag{2.25}$$

where the constants are

$$\begin{aligned} H_k &\equiv 2\eta\sigma^2\varepsilon(m_d) - \eta^2\sigma^4 D_k, \\ G_k &\equiv \frac{1}{2}\sigma^6 k\left[D_k S + 2F\right] + \frac{1}{4}\sigma^6(k+2)E_k. \end{aligned}$$

These differential equations are solvable:

$$l^2(t) = \left(l^2(0) - 2r(0) + 1 - \frac{\eta^2}{H_1}G_1\right)\exp(-H_1 t) + 2r(t) - 1 + \frac{\eta^2}{H_1}G_1, \tag{2.26}$$

$$r(t) = 1 - (1 - r(0))\exp(-\eta\,\sigma^2\varepsilon(m_d)t). \tag{2.27}$$

Equations (2.26) and (2.27) are closed-form macroscopic equations. These equations describe the macroscopic behavior of the system irrelevant to the microscopic value of J_i, x_i and so on.

By definition, r always converge to 1 for $t \to \infty$. However, convergence of l^2 depends on the sign of H_1. Therefore, the convergence condition of the system is

$$H_1 > 0. \tag{2.28}$$

The first and second terms of H_1 are considered as a signal that is extracted from the reward and a noise respectively.

An ensemble average of the squared error over $\boldsymbol{x}$ is the generalization error:

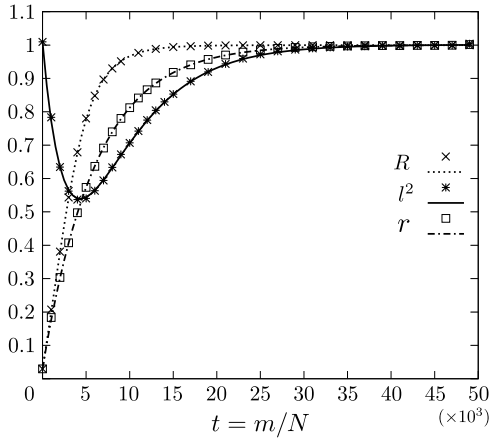
$$\begin{aligned} \epsilon_g(t) &= \frac{1}{2}\langle(y(t) - z(t))^2\rangle_{\boldsymbol{x}} = \frac{1}{2}(l^2(t) - 2r(t) + 1) \\ &= \frac{1}{2}(l^2(0) - 2r(0) + 1 - \frac{\eta^2}{H_1}G_1)\exp(-H_1 t) + \frac{\eta^2}{2H_1}G_1. \end{aligned} \tag{2.29}$$

If the convergence condition is satisfied, the generalization error decays exponentially. After the convergence, a residual error remains due to the perturbation and the credit assignment inefficiency. The residual error is

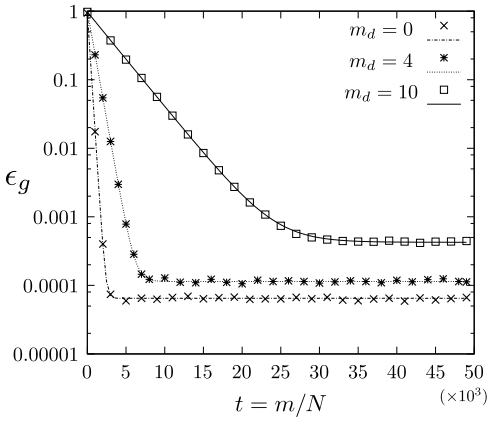
$$\epsilon_r \equiv \lim_{t\to\infty}\epsilon_g(t), \tag{2.30}$$

for $H_1 > 0$.

Thus, the student network can learn teacher's output by our learning rule even if the reward delays.

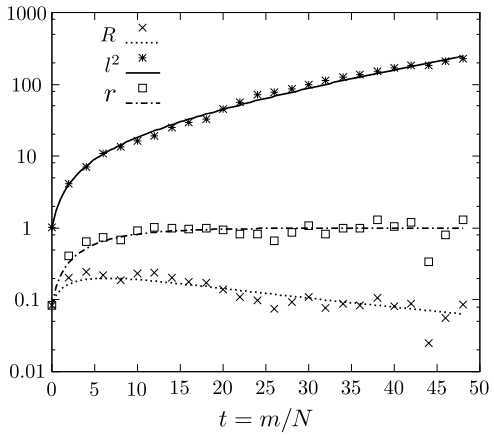


(a) Dynamics of macroscopic variables where $m_d = 10$.

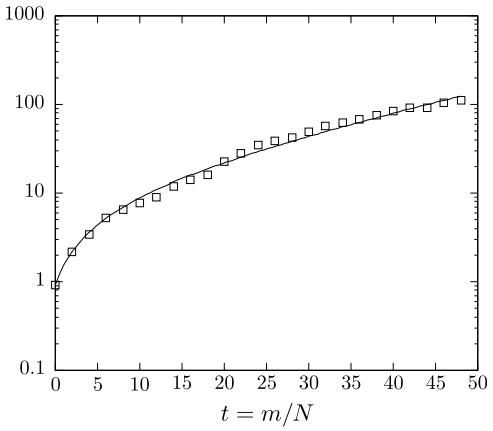


(b) Learning curves (time course of generalization error) in different reward delays.

Figure 2.3: Simulation results in convergent case. Symbols are simulated values and lines show the theoretical values. We set $N = 1000$, $\tau = 4$, $\sigma = 0.1$ and $\eta = 0.2$.



(a) Dynamics of macroscopic variables.



(b) Learning curve (time course of generalization error).

Figure 2.4: Simulation results in divergent case. Symbols are simulated values and lines show the theoretical values. We set $N = 1000$, $\tau = 9.7$, $\sigma = 0.7$, $\eta = 0.5$ and $m_d = 4$.

2.2.3 Simulation Results

Behavior of Macroscopic Variables

We confirm the correctness of our analysis by comparing the theoretical results with simulated ones. In the convergent case, the macroscopic variables and the generalization error are well-predicted by the theory (Fig. 2.3). The relative length l decays at the beginning and then increases as it depends on r. A delay of the reward slows down the learning and increases the residual error. The simulated values are well-matched to the theoretical values even as N is finite in the simulation. Similarly, the theoretical lines match to the simulated lines also in divergent case (Fig. 2.4).

Phase Diagram and Optimal Time Constant

Figure 2.5 shows a phase diagram for learnability in the essential parameters of temporal credit assignment, m_d and τ. The two phases, one convergent and the other divergent, are divided by the sign of H_1. Lower and upper bounds for τ seem to exist against m_d in terms of the convergence. The lower bound can increase because the network has to preserve the past states corresponding to a m_d-delayed reward for learning. On the other hand, if the time constant τ is too large, the signal of reward related states weakens due to the noise, i.e., reward irrelevant states. Because the signal also weakens when m_d increases, the upper bound can decrease. In some point, the bounds cross each other hence there is a region where the system is unstable for any τ. This result matches the intuition that an adaptive system cannot correctly learn if there is a large delay in a feedback signal.

In addition, we explore the parameters where "learning speed" is fastest and the residual error is minimum against m_d. We define the learning speed as the decay speed of the generalization error, H_1 (Eq. (2.29)). The symbols in Fig. 2.5 represent the numerically calculated optimal time constants that maximize the learning speed and minimize the residual error. Each trajectory of the optimal time constants in Fig. 2.5 seems to be fitted with a smooth monotonic function of m_d. Therefore, if the delay of a reward is fixed or does not largely vary, it would be easy to optimize the time constant of the eligibility trace. However, it seems to be impossible to accomplish both a learning speed maximization and a residual error minimization simultaneously because the trajectories do not intersect each other except the endpoints. Figure 2.6 shows the values of maximum H_1 and minimum ϵ_r with optimized time constant

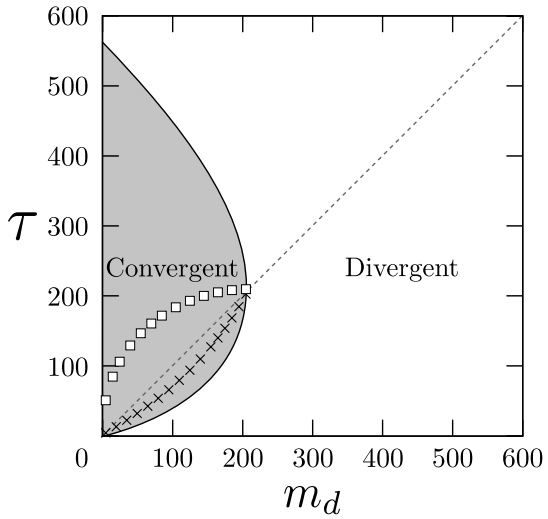
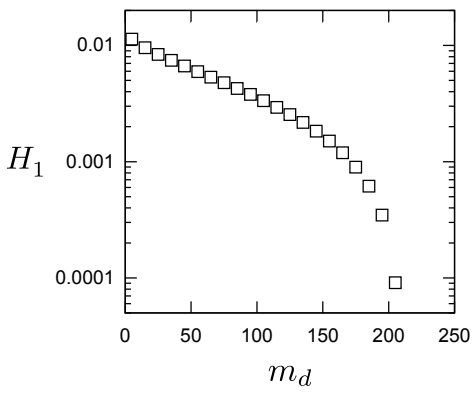
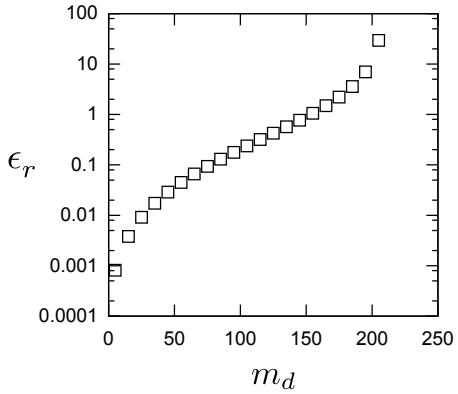


Figure 2.5: Phase diagram of learnability: the convergent regime (grayed region) and the divergent regime (empty region). The square and cross symbols describe optimal time constants τ of eligibility trace that maximizes learning speed and minimizes residual error respectively. We set $\eta = 0.7$ and $\sigma = 0.1$.



(a) Maximum value of H_1 against m_d.



(b) Minimum value of ϵ_r against m_d.

Figure 2.6: The value of H_1 and a residual error ϵ_r with the optimized time constant of the eligibility trace against m_d. We set $\eta = 0.7$ and $\sigma = 0.1$.

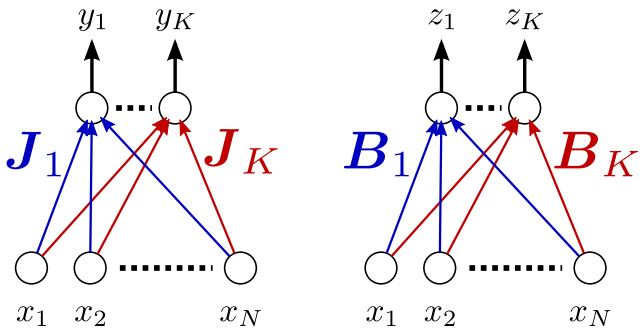


Figure 2.7: Schematic of neural network models. Left is the student and the right one is the teacher network.

against m_d. Maximum learning speed is a monotonically decreasing function of m_d because the amplitude of the signal, $\varepsilon(m_d)$, decreases (Fig. 2.6(a)). On the other hand, the minimum residual error is a monotonically increasing function of m_d (Fig. 2.6(b)).

2.3 Model with Structural Uncertainty

In this section, we show joint effects of structural and temporal uncertainties by increasing the number of linear perceptrons. In this model, the node perturbation works not only for calculating a gradient stochastically but also for solving the structural credit assignment problem.

2.3.1 Model

We suppose that student and teacher networks are composed of K linear perceptrons that share N input units (Fig. 2.7; one perceptron is referred to as an "agent" hereafter). The input $x_j(m)$ $(j = 1, 2, \ldots, N)$ is generated same as previous section: $x_j(m) \sim \mathcal{N}(0, 1/N)$. Student and teacher outputs of i-th $(i = 1, 2, \ldots, K)$ agent are denoted by y_i and z_i respectively:

$$y_i(m) = \boldsymbol{J}_i(m) \cdot \boldsymbol{x}(m), \tag{2.31}$$

$$z_i(m) = \boldsymbol{B}_i \cdot \boldsymbol{x}(m), \tag{2.32}$$

where $\boldsymbol{J}_i$ and $\boldsymbol{B}_i$ are the student and teacher synaptic weight vectors of the i-th agent respectively. As well as the previous section, $B_{ij} \sim \mathcal{N}(0, 1)$ and $\|\boldsymbol{B}_i\| = \sqrt{N}$.

Node perturbation introduces a perturbation $\xi_i \sim \mathcal{N}(0, \sigma^2)$ into the output of i-th agent respectively. The objective function values are:

$$E(m) = \frac{1}{2}\|\boldsymbol{y}(m) - \boldsymbol{z}(m)\|^2, \tag{2.33}$$

$$E_{NP}(m) = \frac{1}{2}\|\boldsymbol{y}(m) + \boldsymbol{\xi}(m) - \boldsymbol{z}(m)\|^2. \tag{2.34}$$

The reward d is defined by the deviation of Eqs. (2.33) and (2.34)

$$\begin{aligned} d(m) &= -\left[E_{NP}(m) - E(m)\right] \\ &= -\frac{1}{2}\left\{\|\boldsymbol{\xi}(m)\|^2 + 2\boldsymbol{\xi}(m)\cdot[\boldsymbol{y}(m) - \boldsymbol{z}(m)]\right\}. \end{aligned} \tag{2.35}$$

To follow the gradient of the objective function, each agent must know its output deviation from the objective value, $y_i(m) - z_i(m)$. However, the reward also contains errors from other agent as shown in Eq.(2.35). Each agent thus needs to estimate the own deviation from the reward. The node perturbation naturally solves this problem by assigning a random credit to respective agent.

As well as previous section, we confirm whether the node perturbation learning rule without delayed reward can extract the respective gradient even in this multi-agent model. By taking the ensemble average of $\Delta \boldsymbol{J}_i^{\text{std}}(m) \equiv \eta d(m)\xi_i(m)\boldsymbol{x}(m)$,

$$\begin{aligned} \langle \Delta \boldsymbol{J}_i^{\text{std}}(m)\rangle &= \eta\langle d(m)\xi_i(m)\boldsymbol{x}(m)\rangle \\ &\propto \langle [y_i(m) - z_i(m)]\boldsymbol{x}(m)\rangle, \end{aligned} \tag{2.36}$$

where η is a learning coefficient and $\langle\cdot\rangle$ represents the ensemble average. The quantity is proportional to the gradient of the objective function.

The agent requires past ξ_i and $\boldsymbol{x}$ to extract the gradient when the reward is delayed. Accordingly, the learning rule is defined with the eligibility trace $\boldsymbol{e}_i$ as follows.

$$\boldsymbol{J}_i(m+1) = \boldsymbol{J}_i(m) + \eta d(m - m_d)\boldsymbol{e}_i(m), \tag{2.37}$$

$$\boldsymbol{e}_i(m) = \sum_{k=0}^{\infty} \varepsilon(k)\xi_i(m-k)\boldsymbol{x}(m-k), \tag{2.38}$$

$$\varepsilon(t) = \exp\left(-\frac{t}{\tau}\right),$$

where m_d represents a delay and τ is a time constant of the eligibility trace. For $\tau \to 0$, our model becomes equivalent to those of previous study [Hara et al., 2011].

2.3.2 Analysis

To elucidate how the structural and temporal uncertainties influence learning, we analyzed the model under the thermodynamic limit $N \to \infty$. Same as previous section, we assume that the learning coefficient η is sufficiently small.

First, we define macroscopic variables that quantify the macroscopic properties of the system: l_i is the relative length of $\boldsymbol{J}_i$ to $\boldsymbol{B}_i$, and R_i is the overlap of the $\boldsymbol{J}_i$ and $\boldsymbol{B}_i$.

$$l_i(m) = \frac{\|\boldsymbol{J}_i(m)\|}{\|\boldsymbol{B}_i\|}, \tag{2.39}$$

$$R_i(m) = \frac{\boldsymbol{J}_i(m) \cdot \boldsymbol{B}_i}{\|\boldsymbol{J}_i(m)\|\|\boldsymbol{B}_i\|}, \tag{2.40}$$

$$r_i(m) = l_i(m)R_i(m) = \frac{1}{N}\boldsymbol{J}_i(m) \cdot \boldsymbol{B}_i. \tag{2.41}$$

The differential equations of the macroscopic variables are derived by the statistical mechanics approach (see previous section for more detail).

$$\frac{dr_i}{dt} = \eta \boldsymbol{B}_i \cdot \langle d(t)\boldsymbol{e}_i(t)\rangle, \tag{2.42}$$

$$\frac{dl_i^2}{dt} = 2\eta\langle d(t)\boldsymbol{J}_i(t) \cdot \boldsymbol{e}_i(t)\rangle + \eta^2\langle d^2(t)\|\boldsymbol{e}_i(t)\|^2\rangle. \tag{2.43}$$

The solution of the ensemble averages of Eqs.(2.42) and (2.43) are a little complex to derive due to the effects of the eligibility trace but they can be solved (see appendix A for more detail). Thus, we obtain the closed-form macroscopic equations

$$\frac{dr_i}{dt} = \eta\sigma^2\varepsilon(m_d)(1 - r_i(t)), \tag{2.44}$$

$$\frac{dl_i^2}{dt} = -H_1 l_i^2(t) + 2(H_1 - \eta\sigma^2\varepsilon(m_d))r_i(t) + (2\eta\sigma^2\varepsilon(m_d) - H_1) + \eta^2 G_K + \eta^2\sigma^4 I\sum_{k\neq i}^{K}[l_k^2(t) - 2r_k(t) + 1], \tag{2.45}$$

where the constants are

$$\begin{aligned} H_k &\equiv 2\eta\sigma^2\varepsilon(m_d) - \eta^2\sigma^4 D_k, \;\; S \equiv \sum_{p=1}^{\infty}\varepsilon(p), \;\; I \equiv \sum_{p=0}^{\infty}\varepsilon^2(p), \\ D_k &\equiv 2\varepsilon^2(m_d) + kI, \;\; E_k \equiv 4\varepsilon^2(m_d) + kI, \;\; F \equiv \varepsilon(m_d)\sum_{p=0}^{m_d-1}\varepsilon(p), \\ G_k &\equiv \frac{1}{2}\sigma^6 k\left[D_k S + 2F\right] + \frac{1}{4}\sigma^6(k+2)E_k. \end{aligned}$$

The Eqs. (2.44) and (2.45) can be written as

$$\dot{X} = -AX + \boldsymbol{c}, \tag{2.46}$$

where

$$\begin{aligned} X &= \left(l_1^2, l_2^2, \ldots l_K^2, r_1, r_2, \ldots, r_K\right)^T, \;\; c_i = \begin{cases} \text{const.} & (1 \le i \le K) \\ 0 & (K+1 \le i \le 2K) \end{cases}, \\ A &= \begin{pmatrix} A_1 & A_2 \\ O & A_3 \end{pmatrix}, \\ (A_1)_{ij} &= \begin{cases} H_1 & (i=j) \\ -\eta^2\sigma^4 I & (i \ne j) \end{cases}, \;\; (A_3)_{ij} = \begin{cases} \eta\sigma^2\varepsilon(m_d) & (i=j) \\ 0 & (i \ne j) \end{cases}, \\ (A_2)_{ij} &= -2(A_1 - A_3), \\ &\text{where} \;\; 1 \le i, j \le K. \end{aligned}$$

The eigenvalues of A are H_K, H_0 ($(K-1)$-fold), $\eta\sigma^2\varepsilon(m_d)$ (K-fold). Therefore,

$$A = U\Lambda U^{-1}, \tag{2.47}$$

where

$$\begin{aligned} &\Lambda = \text{diag}(H_K, H_0, \ldots, H_0, \eta\sigma^2\varepsilon(m_d), \ldots, \eta\sigma^2\varepsilon(m_d)), \\ U &= \begin{pmatrix} P & 2P^{-1} \\ O & P \end{pmatrix}, \\ P &= \begin{pmatrix} 1 & -1 & \ldots & -1 \\ \vdots & \ddots & & \\ \vdots & & \ddots & \\ 1 & & & 1 \end{pmatrix}, \end{aligned}$$

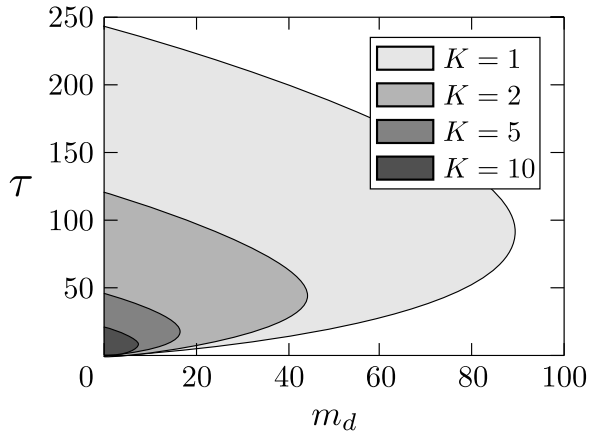


Figure 2.8: Convergent region (gray colors) shrinks as K increases. We set $\eta = 1.1, \sigma = 0.2$.

The solution of differential equation (2.46) is

$$X(t) = -U \exp(-t\Lambda)U^{-1}\left(A^{-1}\boldsymbol{c} - X(0)\right) + A^{-1}\boldsymbol{c}. \tag{2.48}$$

From (2.48), the dynamics of r_i becomes:

$$r_i(t) = 1 - (1 - r_i(0)) \exp(-\eta\sigma^2\varepsilon(m_d)t), \tag{2.49}$$

Thus, r_i always converges to 1 for $t \to \infty$. On the other hand, the convergence of l_i^2 is determined by the sign of $H_K = 2\eta\sigma^2\varepsilon(m_d) - \eta^2\sigma^4(2\varepsilon^2(m_d) + KI)$. The first term of H_K is considered as a signal term that is extracted by using the correlation between the perturbation and the reward, and the second term seems to represent a noise that contains the effects of both structural (K) and temporal (m_d, τ) uncertainties of the reward.

The exact time course of generalization error is expressed by the macroscopic variables. Let ϵ_g be the generalization error:

$$\begin{aligned}\frac{d}{dt}\epsilon_g &= \frac{1}{2}\sum_i^K \left(\frac{dl_i^2}{dt} - 2\frac{dr_i}{dt}\right) \\ &= \frac{1}{2}\sum_i^K \left\{ \begin{array}{l} -H_1 l_i^2(t) + 2H_1 r_i(t) - H_1 + \eta^2 G_K \\ +\eta^2\sigma^4 I \sum_{k\neq i}^K [l_k^2(t) - 2r_k(t) + 1] \end{array} \right\},\end{aligned} \tag{2.50}$$

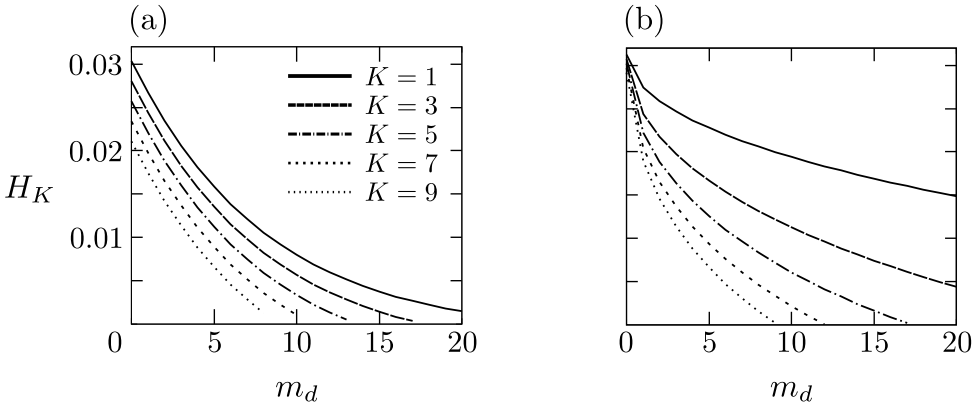


Figure 2.9: Learning speed H_K against delay m_d and number of agents K with (a) fixed time constant of eligibility trace, $\tau = 0.8$, and (b) τ that maximizes H_K. We set $\sigma = 0.2, \eta = 0.4$.

2.3.3 Results

Figure 2.8 shows that the convergent region shrinks according to the increase of K. We confirmed that this qualitative result is independent of the learning coefficient and the variance of the perturbation. Interestingly, τ, which is introduced to deal with temporal uncertainty of the reward, is affected by the change of K, which only varies the structural uncertainty. This implies that structural uncertainty interact with temporal uncertainty.

Figure 2.9 shows the change of the learning speed H_K for delay m_d where τ is fixed and where τ is optimized to maximize H_K. When τ is fixed, H_K simply decreases same amount for any m_d as K increases (Fig. 2.9(a)). Thus, in this case, the structural uncertainty is independent of the temporal uncertainty. However, when τ is optimized, the gradient of H_K are different (Fig. 2.9(b)). Thus, the structural and temporal uncertainties interact each other via the eligibility trace. Figure 2.10 shows the contour of the optimal τ for m_d and K. In the figure, optimal τ decreases with the increase of K. These results again suggest that the structural and the temporal uncertainties interact with each other, and thus they cannot be simply separated. It seems that the signal of gradient for an agent contained in the reward weakens as K increases because the signals of other agents are noise for one agent.

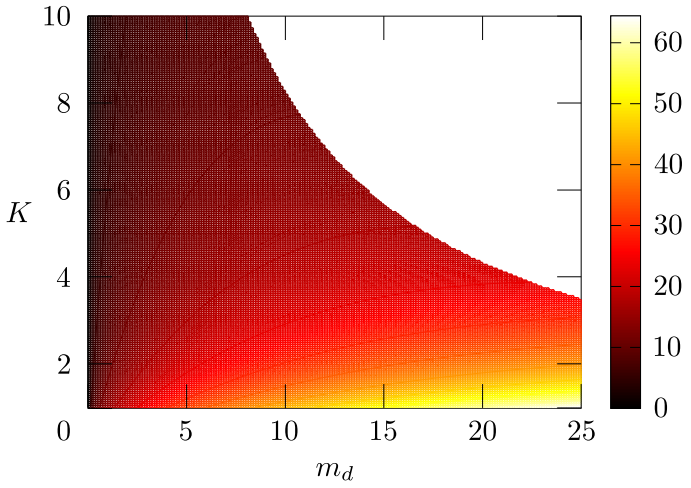


Figure 2.10: Optimal time constant of eligibility trace τ that maximizes learning speed H_K for delay m_d and number of agents K.

2.4 Discussion

We theoretically investigated the quantitative effects of structural-temporal uncertainties to neural network models which have the eligibility trace and adapted by node perturbation algorithm. We derived the learning dynamics in closed form macroscopic equations and confirmed the derived results by numerical simulations. We discovered that the learnability depends on three factors: an amplitude of signal and temporal and structural noises. Besides, we discovered interactions between the structural and temporal uncertainties. This result is consistent with that the temporal credit assignment problem can be converted to the structural one in the Markov decision process [Agogino & Tumer, 2004]. Our analysis relies on the linearity of output unit while conventional neural networks employ the non-linear one. However, in the case of the sigmoidal function, which is often adopted in neuroscience and engineering fields, it can be thought that the learning takes place in linear regime at least during the initial phase of training [Baldi & Hornik, 1995]. Thus, we expect that this qualitative property holds in any types of learning algorithms facing the structural and temporal uncertainties. Thus, we elucidated the quantitative effect of trace to structural-temporal uncertainties.

The node perturbation is proposed to be a neural mechanism of songbird's learning [Fiete et al., 2007]. A juvenile songbird learns how to sing by imitating the tutor's song. There are three regions involved in producing a song in a songbird's brain:

HVC, RA, and LMAN. The song varies with the change in synaptic weights from HVC neurons to RA neurons. Therefore, the synaptic weights are modulated to produce a song that is similar to the tutor's one through learning. LMAN neurons perturb the RA neural activities to induce song variability [Hessler & Doupe, 1999]. As a result of the perturbation, a feedback will be obtained. The reward seems to reach the HVC-RA synapses with a 50-75 ms delay [Fee et al., 2004, Troyer & Doupe, 2000]. Hence, a credit assignment mechanism, like the eligibility trace, is required. As our investigation of the optimal trajectories of time constants, the time constant of a credit assignment mechanism in a songbird brain may be optimized.

Retroaxonal signals [Hamburger, 1992] are reverse transmission on an axon, from a presynaptic terminal to the soma of the presynaptic cell, contrary to the classical form of action potentials. It is proposed that the retroaxonal signals are released according to changes in synaptic strength and then they stabilize the recent synaptic changes [Harris, 2008, Salihoglu et al., 2009]. The signals may have similar mechanisms to the node perturbation: transient changes in a presynaptic weight of a neuron are stabilized only if new spiking pattern of the neuron strengthen post synaptic weights; Another notable feature of the retroaxonal signaling is that it is very slow process compared with action potentials. Thus, the signals seem to realize a trace of past neural activities. Therefore, we expect that our learning model captures fundamental feature that would be observed in learning based on the retroaxonal signals.

Chapter 3

Computational Model for Matching Law in Volatile Environment

The decision making behaviors of humans and animals adapt and then satisfy an "matching law" in foraging tasks. In recent studies, it has been shown that a choice of monkey depends on reward trace double-exponentially in a foraging task. We analyze several deterministic Bayesian decision making models to elucidate underlying computational principle of matching behavior and the reward trace dependency.

3.1 Introduction

Does the brain play dice? This is a controversial question about the underlying processes of the brain in making a choice from several alternatives: Does the brain decide deterministically with some internal decision variables? Or does it calculate the probability of choosing individual alternatives and cast a 'biased die' [Sugrue et al., 2005]? The former strategy is suggested according to our everyday experience. However, it is possible to think that choices emerge probabilistically by observing a sequence of decisions in a repetitive task. Herrnstein conducted a foraging experiment where a pigeon was placed into a box that was equipped with two keys and when a key was pressed it was rewarded with concurrent variable-interval schedules. He found a relationship between rewards and choices known as the "matching law" [Herrnstein, 1961]. The law states that the fraction of the number of times one alternative is chosen against the total number of choices matches the fraction of the cumulative re-

ward obtained from the alternative against the total reward. Behaviors satisfying the law have been observed in a variety of task paradigms and across species [Anderson et al., 2002, Gallistel, 1994, de Villiers & Herrnstein, 1976]. Several learning models have been proposed to account for matching behavior [Loewenstein & Seung, 2006, Soltani & Wang, 2006, Simen & Cohen, 2009, Sakai & Fukai, 2008a, Corrado et al., 2005, Lau & Glimcher, 2005]. These models have a commonality in that a model learns the probabilities of choosing each alternative directly, and then a choice is made stochastically. However, it is yet unknown whether matching behaviors can be accounted for by a deterministic model.

Here, we propose deterministic Bayesian decision making models for a two-alternative choice task. Our models stand on the incorrect but conceivable postulate that animals have a belief that the choice made in one trial does not affect a reward in subsequent trials. The models estimate the unknown reward probabilities for each alternative and deterministically choose the alternative that has the highest reward probability according to the *winner-take-all* principle. We first study a model with belief that the environment does not change. Note that this is an extension of the fixed belief model (FBM) [Yu & Cohen, 2009] for the two-alternative choice task. We demonstrate that this model satisfies the matching law in a steady state in static foraging tasks, in which reward baiting probabilities are fixed, but not in dynamic foraging tasks, in which the reward baiting probabilities change abruptly. Then, we devise two models that forget past experience and exhibit matching behaviors even in dynamic tasks. Moreover, these models can explain *undermatching*, which is a phenomenon observed across different species [Baum, 1974, Baum, 1979, Anderson et al., 2002, Gallistel, 1994, de Villiers & Herrnstein, 1976, Sugrue et al., 2004, Lau & Glimcher, 2005]. We test these models by comparing their behaviors with those of monkeys'. Besides, we discuss the computational principle behind the matching behavior and reward trace dependency.

3.2 Foraging Task

The foraging task is a decision making task that simulates a foraging environment where an animal chooses one out of several foraging alternatives. There are two alternatives in this study although our results do not depend on this. We employ discrete trial-to-trial tasks that have often been used in recent experiments [Corrado et al., 2005, Lau & Glimcher, 2005, Sugrue et al., 2004]. Each alternative

has binary baiting state f_i ($i \in \{1, 2\}$ is the index of an alternative), where $f_i = 1$ if a reward is baited and $f_i = 0$ otherwise. If $f_i = 0$, a reward is baited ($f_i = 1$) at the beginning of each trial by baiting probability λ_i^t, where t represents the number of the trial. If the baiting probabilities are fixed across trials, the task is called a *static* foraging task, otherwise it is called a *dynamic* foraging task [Sugrue et al., 2004]. Suppose that r_i^t indicates whether a subject receives a reward ($r_i^t = 1$) or not ($r_i^t = 0$), and c_i^t indicates whether the subject chooses alternative i ($c_i^t = 1$) or not ($c_i^t = 0$) in trial t. When the subject chooses a baited alternative, i.e. $f_i = 1$ and $c_i^t = 1$, the baited reward is consumed ($f_i \leftarrow 0$). This reward schedule is known as a "concurrent variable-interval schedule"[Baum & Rachlin, 1969].

Whichever alternative the subject chooses in the foraging task, the choice can affect the reward probabilities of both alternatives in the future. Therefore, the optimal strategy is not to exclusively choose the foraging alternative that has the highest baiting probability. A behavioral strategy obeying the matching law is known to be nearly optimal for this task [Baum, 1981]. Formally, the law states that

$$\frac{\bar{R}_i^t}{\sum_j \bar{R}_j^t} = \frac{\bar{C}_i^t}{\sum_j \bar{C}_j^t}, \tag{3.1}$$

where $\bar{R}_i^t$ and $\bar{C}_i^t$ correspond to the total reward obtained from alternative i and the number of choices of alternative i until trial t. It is known that human and animal behaviors in these kinds of tasks are well described by the generalized matching law [Baum, 1974]

$$\log(\bar{R}_1^t/\bar{R}_2^t) = s \log(\bar{C}_1^t/\bar{C}_2^t) + \log k, \tag{3.2}$$

where s is sensitivity and k is bias. Eq. (3.2) is equivalent to Eq. (3.1) if both s and k are unities.

Here is the configurations of simulation. The reward schedule is analogous to the experiment by Corrado *et al.* (2005). We randomly set the baiting probabilities that satisfied $\lambda_1 + \lambda_2 = 0.3$ and their ratios were 1:8, 1:6, 1:3, 1:2, 1:1, 2:1, 3:1, 6:1, and 8:1 in a static setting. There are $10,000$ trials in the simulations. The baiting schedule in the dynamic setting is divided into blocks, in which the baiting probabilities were fixed, and their sum and ratios were the same as those in the static setting. The block length is uniformly sampled from $[50, 300]$ and there were 300 blocks in the simulations. Change-over-delay (COD), in which is the cost to switch from one alternative to another, is not included differently from Corrado *et al.* (2005).

3.3 Simple Bernoulli Estimators

First, we study a simple normative Bayesian decision making model to clarify the underlying feasible computation for matching behaviors. Suppose that a subject makes a decision simply depending on its estimates of the reward probabilities for the alternatives. The estimate can be formally described as

$$P_i^{t+1} = p(r_i^{t+1} = 1 | R^t, C^t), \tag{3.3}$$

where R^t is a list of reward vectors $\boldsymbol{r}^t = (r_1^t, r_2^t)$ from trials 1 to t and C^t is a list of choice vectors $\boldsymbol{c}^t = (c_1^t, c_2^t)$ from trials 1 to t. The model employs a *winner-take-all* (WTA) strategy, i.e., it chooses the alternative that has the highest P_i^t. The model requires an assumption about a reward assignment mechanism to estimate P_i^{t+1}. One simple and conceivable assumption is that a choice is rewarded according to hidden reward probability μ_i^t that is irrelevant to the past reward and choice trace,

$$p(r_i^t = 1) = \mu_i^t. \tag{3.4}$$

This assumption is incorrect for our tasks but we have assumed that the model employs it and predicts μ_i^t by Bayesian inference. Hence, P_i^{t+1} is given by the predictive distribution over μ_i^t:

$$P_i^{t+1} = \int_0^1 d\mu \, \mu \, p(\mu_i^{t+1} = \mu | R^t, C^t). \tag{3.5}$$

Note that $p(\mu_i^{t+1} = \mu | R^t, C^t)$ can include a model's belief about the change of μ_i^t in between trials. Our first model assumes that μ_i^t is time invariant,

$$p(\mu_i^{t+1} = \mu | R^t, C^t) = p(\mu_i^t = \mu | R^t, C^t). \tag{3.6}$$

The posterior distribution for an alternative is not updated if the alternative is not chosen. If it is chosen, the posterior distribution is updated

$$\begin{aligned} p(\mu_i^t = \mu | R^t, C^t) &\propto p(r_i^t | \mu_i^t = \mu) p(\mu_i^{t-1} = \mu | R^{t-1}, C^{t-1}) \\ &= \mu^{r_i^t} (1-\mu)^{1-r_i^t} p(\mu_i^{t-1} = \mu | R^{t-1}, C^{t-1}). \end{aligned} \tag{3.7}$$

We employ the Beta prior, $p(\mu_i^0 = \mu) = \text{Beta}(\mu | \text{a}, \text{b})$, which is a conjugate for the likelihood. Note that we set the hyper-parameters, $a = b = 1$, to make the prior non-informative in all simulations. Therefore, the posterior becomes a Beta distribution:

$$p(\mu_i = \mu | \bar{R}_i^t, \bar{C}_i^t) = \text{Beta}(\mu | \bar{R}_i^t + a, \bar{C}_i^t - \bar{R}_i^t + b). \tag{3.8}$$

From Eqs. (3.5) and (3.8), we obtain

$$P_i^{t+1} = \frac{\bar{R}_i^t + a}{\bar{C}_i^t + a + b}. \tag{3.9}$$

This model is a natural extension of FBM [Yu & Cohen, 2009] to the two-alternative choice task (for this reason, we will refer to our model as FBM). An alternative is repeatedly chosen while its predictive distribution is higher than those of the other due to the WTA strategy. Because the empirical probability of reward for an alternative converges to its baiting probability in repeated choices, P_i^t gradually approaches to λ_i and the variance of P_i^t decreases. As a result, FBM tends to choose exclusively the high payoff alternative after a large number of observations. Hence, the matching law (Eq. (3.1)) is satisfied in $t \to \infty$ because such a exclusive choice unboundedly increases both $\bar{R}_i^t$ and $\bar{C}_i^t$ of the high payoff alternative.

3.3.1 Results

We simulate FBM in static and dynamic foraging tasks. The time course for the predictive distributions is shown in Figure 3.1**A**. As can be expected, both predictive distributions approach the respective baiting probabilities and FBM behavior converges to exclusive choice of the high payoff alternative in static foraging tasks. However, the steady-state choice behavior of animals in static concurrent VI schedules has not been thought to be exclusive [Baum, 1982, Davison & McCarthy, 1988, Baum et al., 1999]. It might be that there are not enough trials for choice behavior to actually reach a steady state. Figures 3.1**B** and **C** plot the log ratios of rewards and choices in both tasks. The marginal histograms indicate the FBM's strong preference for the alternative that has the highest baiting probability, because most pairs of log ratios lie near the endpoints of the matching line. It is shown that bias is nearly zero and sensitivity is nearly one in the static foraging tasks (Fig. 3.1**B**) by least-square fitting the generalized matching law (Eq. (3.2)) to the data. Therefore, the model exhibits matching behavior in the static foraging tasks. However, the model no longer exhibits matching behavior in dynamic foraging tasks, a result that is inconsistent with the behavior of monkeys [Corrado et al., 2005] (Fig. 3.1**C**). This can be because the model adheres to past experience and cannot adapt rapidly to changes in the environment.

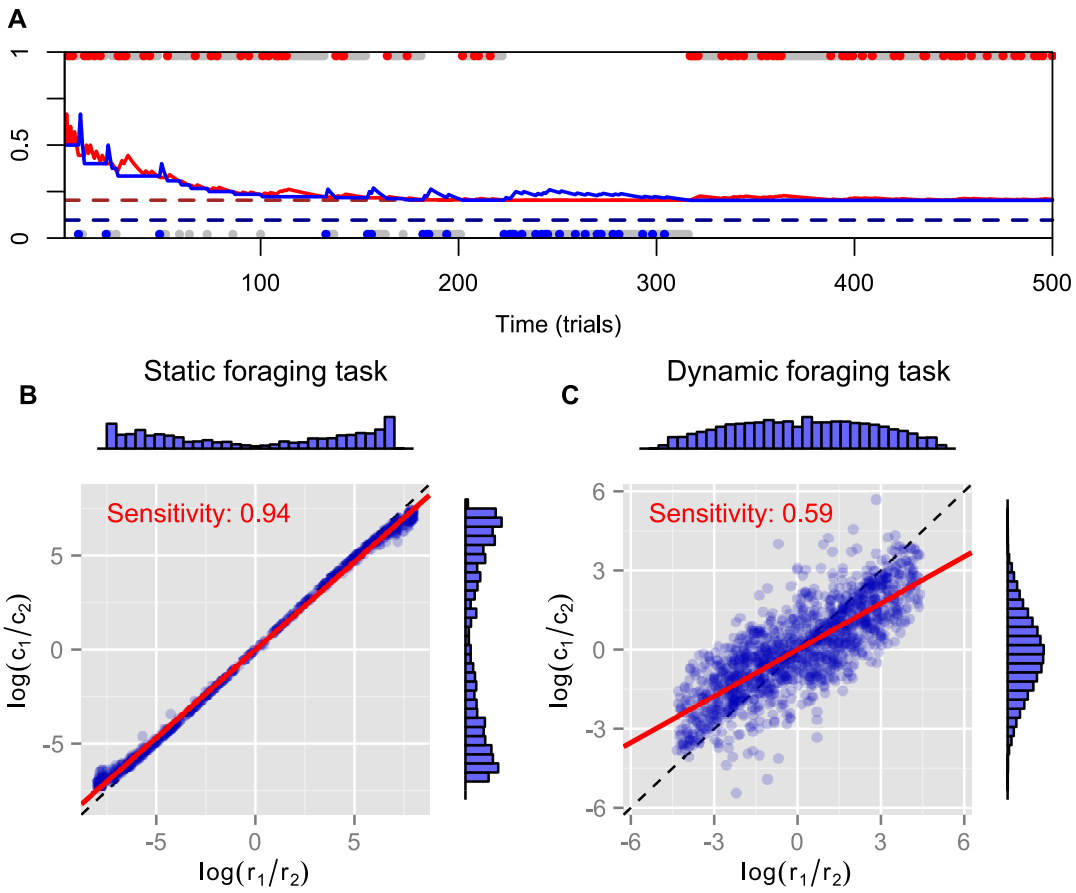


Figure 3.1: **Simulation results for FBM**. (**A**) Time course of predictive distributions for alternatives #1 (red solid line) and #2 (blue solid line) in static foraging task. Dashed lines indicate baiting probabilities of alternatives #1 (dark red) and #2 (dark blue). Upper and lower dots respectively represent choices for alternatives #1 and #2 in that trial and colored dots (red or blue) represent that the model received a reward at that trials. (**B**) Reward log ratios as a function of count log ratios in static and (**C**) dynamic foraging tasks. Blue symbols represent pairs of log ratios calculated in block where baiting probabilities are fixed and distributions of dots are represented by marginal histograms. Red line indicates best-fitted line to points and inner text shows its slope, i.e., sensitivity parameter of generalized matching law. Dashed line is identity line.

3.4 Extended Bernoulli Estimators

One possible way of improving the model to enable it to rapidly adapt to changes in the environment is to introduce a forgetting mechanism for past rewards and choice trace. We therefore assume a simple extended model, which utilizes only the L most recent rewards and choices for the estimates. Hence, the predictive distribution becomes

$$P_i^{t+1} = \frac{(\sum_{l=0}^{L-1} r_i^{t-l}) + a}{(\sum_{l=0}^{L-1} c_i^{t-l}) + a + b}. \tag{3.10}$$

We refer to this model as windowed FBM (WFBM).

Another possibility may be derived from the idea that humans and animals may innately believe their environment is volatile. Here, we propose a model that estimates time-varying reward probabilities. Although there are several ways to model a belief of a volatile environment, we assume our model believes that μ_i^t remains unchanged with probability α, or else (with probability $1-\alpha$) changes completely. This idea is derived from the dynamic belief model (DBM), proposed by Yu & Cohen as a model of sequential effect [Yu & Cohen, 2009]. Our model is a natural extension of DBM to a two-alternative choice task. Thus, we refer to our model as DBM. The transition of μ_i^t is modeled as a mixture of the posterior and prior distributions

$$p(\mu_i^{t+1} = \mu | R^t, C^t) = \alpha p(\mu_i^t = \mu | R^t, C^t) + (1-\alpha)\text{Beta}(\mu | a, b), \tag{3.11}$$

where $0 \leq \alpha \leq 1$ represents the model's expectations of the stability of the environment. However, the posterior distribution is no longer a Beta distribution:

$$\begin{aligned}
&p(\mu_i^t = \mu | R^t, C^t) \\
&= p(\mu_i^t = \mu | r_i^t, c_i^t = 1, R^{t-1}, C^{t-1})^{c_i^t} p(\mu_i^t = \mu | R^{t-1}, C^{t-1})^{1-c_i^t} \\
&= \left[\left(\frac{p(r_i^t = 1 | \mu_i^t = \mu)}{p(r_i^t = 1 | R^{t-1}, C^{t-1})} \right)^{r_i^t} \left(\frac{p(r_i^t = 0 | \mu_i^t = \mu)}{p(r_i^t = 0 | R^{t-1}, C^{t-1})} \right)^{1-r_i^t} \right]^{c_i^t} p(\mu_i^t = \mu | R^{t-1}, C^{t-1}) \\
&= \left[\left(\frac{\mu}{P_i^t} \right)^{r_i^t} \left(\frac{1-\mu}{1-P_i^t} \right)^{1-r_i^t} \right]^{c_i^t} p(\mu_i^t = \mu | R^{t-1}, C^{t-1}),
\end{aligned} \tag{3.12}$$

where Eq. (3.3) is used. Then, predictive distribution P_i^t is calculated with Eqs. (3.5), (3.11), and (3.12). Note that these models are equivalent to FBM when $L \to \infty$ and $\alpha = 1$.

3.4.1 Results

Matching Behavior

Figure 3.2 has the time courses for the predictive distributions of WFBM and DBM, and the posterior distributions of DBM in the dynamic foraging task. Neither model is stuck on one alternative and can follow the changes in schedules as expected. There is a clear difference in the predictive distribution trajectories. Because WFBM exploits recent samples, its predictive distribution for the unchosen alternative can approach the true baiting probability. DBM's predictive distribution for the unchosen alternative, on the other hand, is only retracted to the mean of the prior, i.e., 0.5. Both models demonstrate matching behaviors even in the dynamic foraging task (Fig. 3.3). More precisely, the behaviors slightly deviate from the matching law toward an unbiased choice. This phenomenon is known as *undermatching* [Baum, 1979]. Because the models' parameters L and α control the effect of past experience, the degree of undermatching is controlled by the parameters. The sensitivities that were fitted in the experiments were in a range of about 0.44 to 0.91 [Lau & Glimcher, 2005, Corrado et al., 2005, Hinson & Staddon, 1983]. Hence, we basically focus on parameter regions $10 \leq L$ and $0.9 \leq \alpha$.

Run-length Distribution

It is known that the probability of switching alternatives is nearly constant against the number of consecutive choices for one alternative (run length) in the concurrent VI schedule [Heyman & Luce, 1979]. Hence, run lengths are distributed exponentially but, in a dynamic foraging task, the distribution seems to be a mixture of exponentials [Corrado et al., 2005]. The distribution of WFBM does not monotonically decrease and there is a peak where the run length is nearly equal to L. Therefore, the distribution is neither an exponential nor a mixture of exponentials. This nature is consistent on different values of L. However, DBM demonstrates an exponential like distribution. We fit single and double exponential functions,

$$\phi_1(l) = \nu_0 \exp(-\nu_0(l-1)), \tag{3.13}$$

$$\phi_2(l) = \gamma\nu_1 \exp(-\nu_1(l-1)) + (1-\gamma)\nu_2 \exp(-\nu_2(l-1)), \tag{3.14}$$

to the distribution, where $l \geq 1$ is the run length, ν_0 and $\nu_1 < \nu_2$ are the rate parameters and γ is the combining rate. The distribution is well-fitted by the double

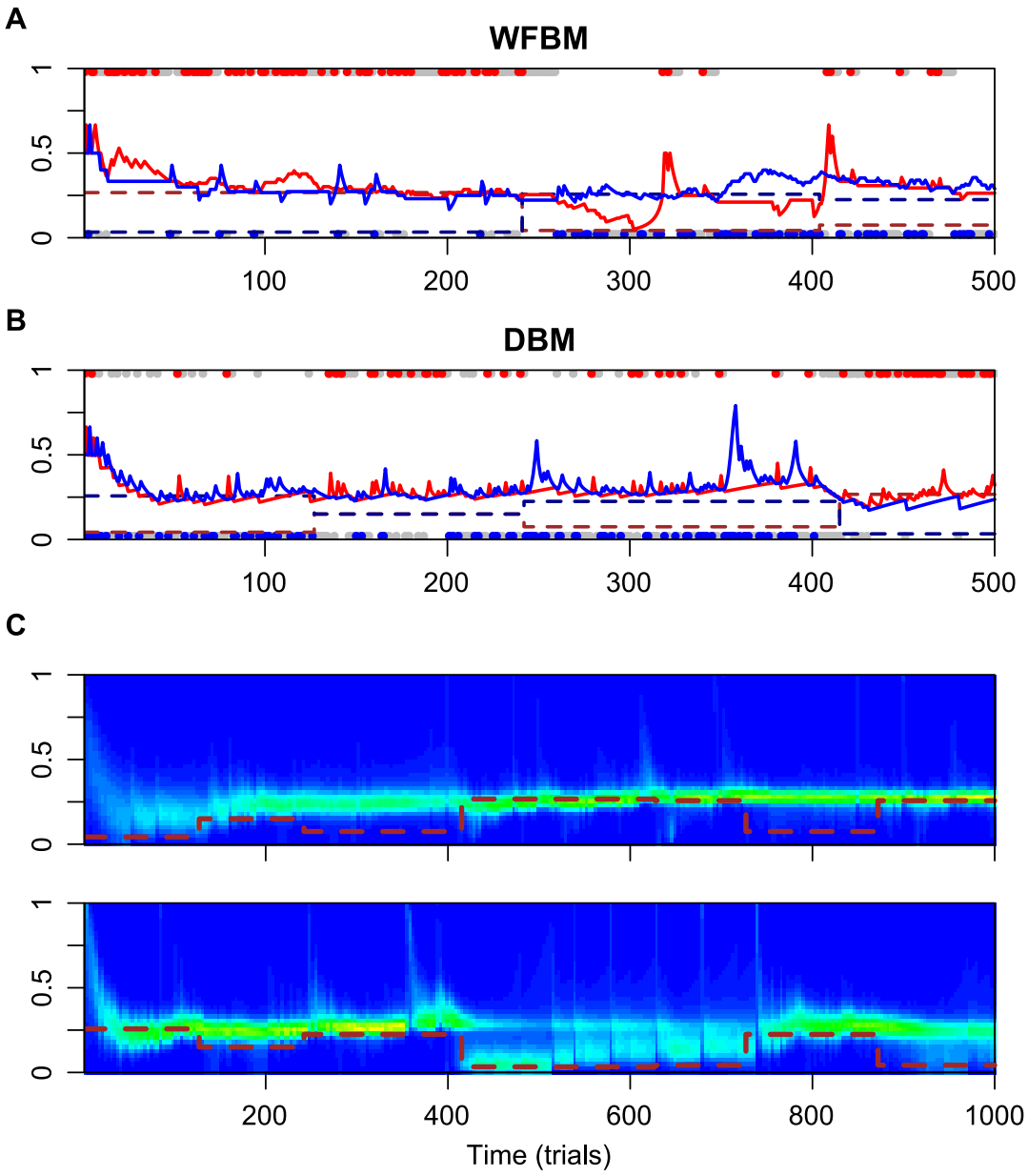


Figure 3.2: **Simulation results for WFBM and DBM in dynamic foraging task**. Simulation parameters were set to $L = 60$ and $\alpha = 0.99$. (**A**) Time course of predictive distributions of WFBM and (**B**) DBM. Details in figure are described in caption of Fig. 3.1**A**. (**C**) Time course of posterior distributions of DBM for reward probabilities of alternative #1 (top) and #2 (bottom). Brown dashed lines are baiting probabilities for respective alternatives.

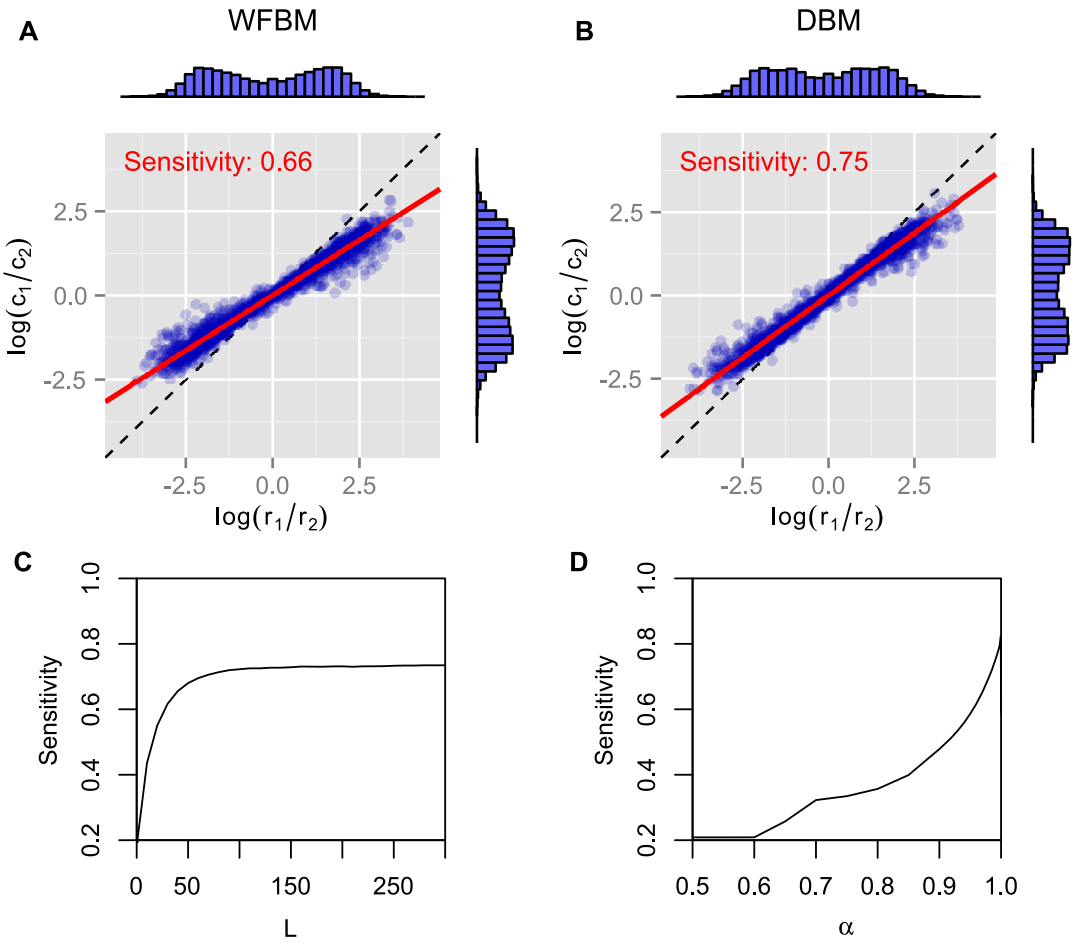


Figure 3.3: **Analytical results for matching behavior of WFBM and DBM in dynamic foraging tasks**. (**A** and **B**) Reward log ratios as a function of count log ratios. Details in figure are described in caption of Figs. 3.1**B** and **C**. Simulation parameters were set to $L = 40$ and $\alpha = 0.99$. (**C** and **D**) Sensitivity as a function of parameters of WFBM and DBM.

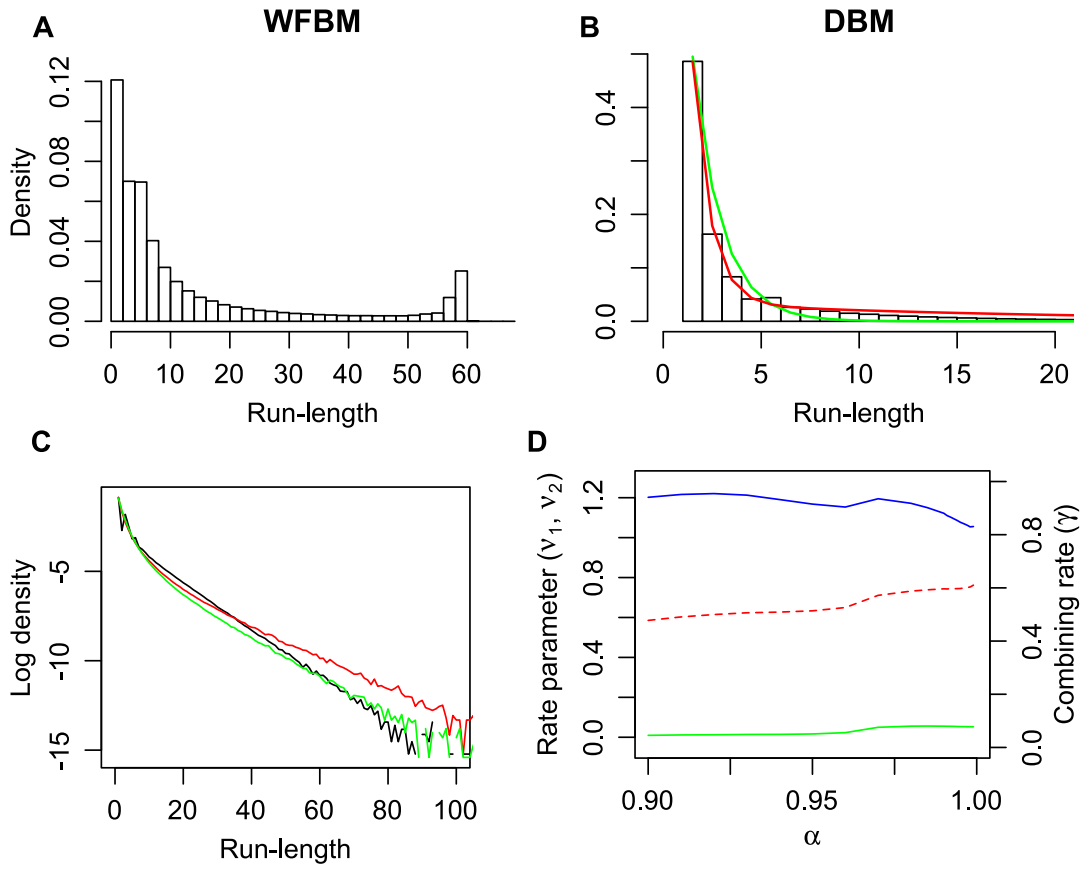


Figure 3.4: **Run-length distributions of windowed FBM and DBM in dynamic foraging task**. Simulation parameters were set to $L = 60$ and $\alpha = 0.99$. (**A** and **B**) Bars represent densities of run length for alternative #1. Single (green line) and double-exponential (red line) functions fitted to run-length distributions of DBM. Double-exponential function is fitted better than single one (likelihood ratio test, $p \ll 0.001$). (**C**) Log probability density of run-length distribution of DBM (black line) and linear-nonlinear Poisson models (red and green lines) that are fitted to monkeys' experimental data in Corrado et al. [Corrado et al., 2005]. (**D**) Fitted parameters of double-exponential function with different values of α. Left ordinate indicates value of rate parameters ν_1 (green line) and ν_2 (blue line), and right indicates value of combining rate γ (red line).

exponential function (Fig. 3.4**B**; likelihood ratio test, $p \ll 0.001$; r^2 for the double and single exponential functions are 0.99 for the former and 0.96 for the latter). The run-length distribution in monkey experiments has few frequencies of a very short run length; however our models have the largest frequency at the run length of 1 (Fig. 3.4**A** and **B**). This difference can be due to the absence of change-over-delay (COD) in our schedule. If our model had and exploited prior knowledge about COD as well as the proposed model for the previous experiment [Corrado et al., 2005], the frequency at a run length of 1 could disappear. We simulate linear-nonlinear-Poisson (LNP) models that are fitted to the monkeys' experimental data in Corrado et al. [Corrado et al., 2005] and compare run-length distributions (Fig. 3.4**C**). Note that COD is not considered for the LNP models that is different from Corrado et al.'s approach [Corrado et al., 2005]. Because the absence of COD could affect the occurrence of short run lengths, log probability densities are compared to count differences at long run lengths. The calculated mean squared differences of DBM against LNP models for two monkeys correspond to ~ 0.67 and 0.16. The double-exponential function is better than the single one in different α and the fitted parameters are slightly affected by α (Fig. 3.4**D**).

Reward Trace Dependency

The dependence of choices on reward trace has been studied in several monkey experiments. An exponential shaped dependency was first reported [Sugrue et al., 2004] and then heavier-tailed dependencies were reported [Lau & Glimcher, 2005, Corrado et al., 2005]. We test our models by calculating the dependence of choices on reward trace. Suppose that dependency is expressed with a linear filter kernel $\kappa(i)$ as in previous studies. The kernel is calculated by minimizing the following Wiener-Hopf equation,

$$\frac{1}{2}\sum_t \left[(c_1^t - c_2^t) - \sum_{i=1}^{K} \kappa(i)(r_1^{t-i} - r_2^{t-i}) \right]^2 . \tag{3.15}$$

Then, we fit the exponential filter and double-exponential filter that were introduced by Corrado et al. [Corrado et al., 2005] to the normalized kernel:

$$\begin{aligned} \epsilon_1(i) &= \frac{\exp(-i/\tau_0)}{\sum_{k=1}^{K}\exp(-k/\tau_0)}, \\ \epsilon_2(i) &= \rho\,\frac{\exp(-i/\tau_1)}{\sum_{k=1}^{K}\exp(-k/\tau_1)} + (1-\rho)\,\frac{\exp(-i/\tau_2)}{\sum_{k=1}^{K}\exp(-k/\tau_2)}, \end{aligned} \tag{3.16}$$

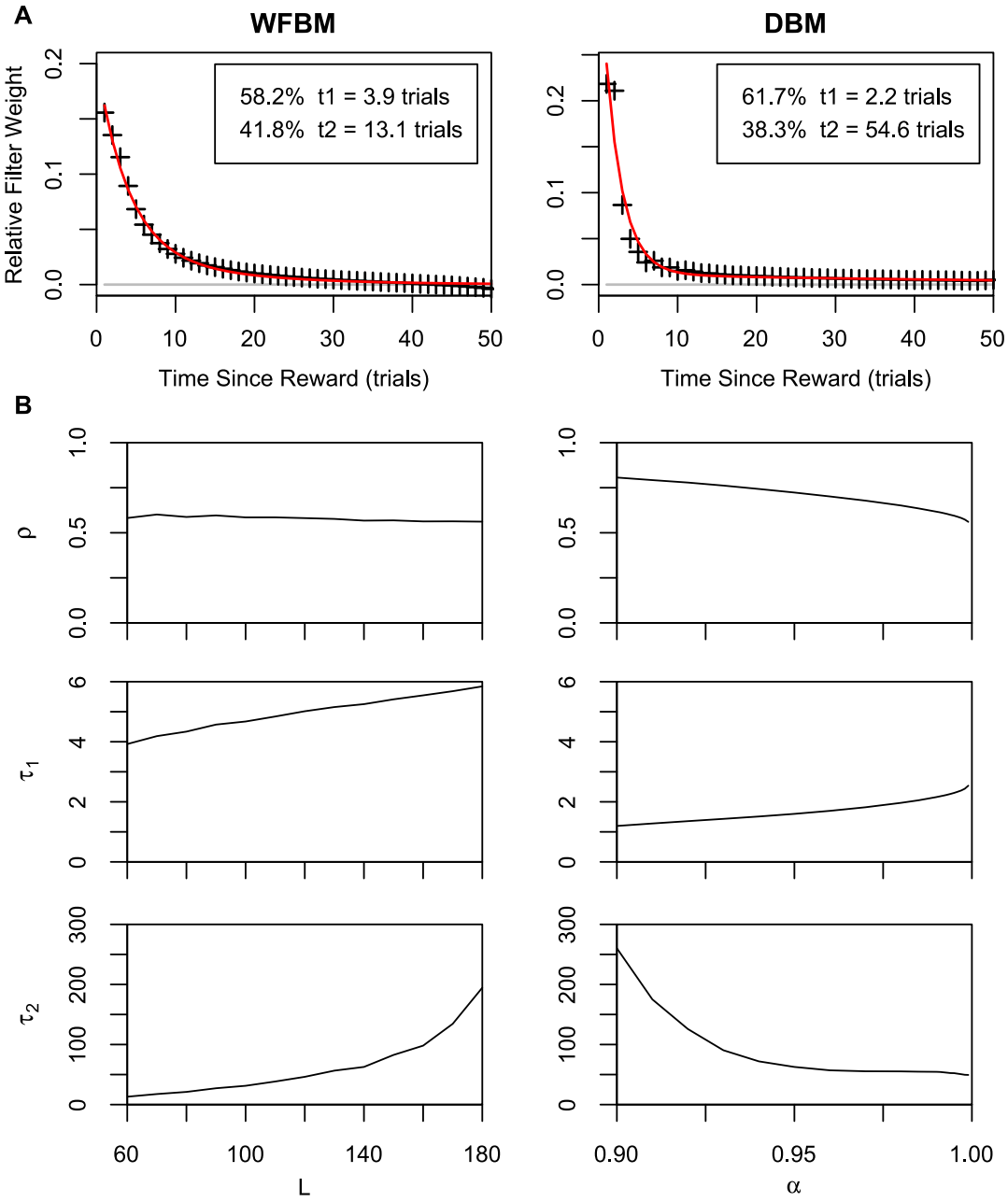


Figure 3.5: **Results of Wiener-Hopf analysis for WFBM and DBM in dynamic foraging task**. (**A**) Symbols represent normalized Wiener-Hopf kernel and red line represents best fitted double-exponential filter. Double-exponential filters are better fitted to data than single-exponential filter (likelihood ratio test, $p \ll 0.001$). Insets show time constants for each exponential component and their combining rate. Simulation parameters were set to $L = 60$ and $\alpha = 0.99$. (**B**) Fitted parameters of double-exponential filter ρ, τ_1, and τ_2 to simulation data of WFBM (left column) and DBM (right column). Abscissas represent parameters of WFBM or DBM.

where τ_0 and $\tau_1 \leq \tau_2$ are time constants and $0 < \rho < 1$ is the combining rate. Note that ϵ_2 is identical to ϵ_1 when $\tau_1 = \tau_2$. The double-exponential filter is rather more well-fitted than the single one for WFBM and DBM (likelihood ratio test, $p \ll 0.001$; adjusted r^2 for double and single exponential filters are 0.99 and 0.98 for WFBM, and 0.94 and 0.85 for DBM). The kernel for WFBM has a negative value around L but it disappears if L is much longer than K. The kernel for DBM drops sharply and decays slowly. The sharp drop probably arose from the exponential decay of reward trace, which is embedded in the posterior distributions (Eq. (3.11)). Because a decision is made due to the difference in two predictive distributions and both distributions decay at the same rate, the effect of one predictive distribution would have persisted slightly longer and hence the kernel included a longer exponential component. This characteristic is qualitatively consistent with the experimental results [Corrado et al., 2005]. The fitting parameters for the two monkeys in Corrado et al. [Corrado et al., 2005] were $\rho = 0.4$, $\tau_1 = 2.2$, and $\tau_2 = 17.0$ (monkey F), and $\rho = 0.25$, $\tau_1 = 0.9$, and $\tau_2 = 12.6$ (monkey G). Although there were no suitable WFBM and DBM parameters that exactly matched their fitting parameters to those of the monkeys, similar values were obtained for smaller L and larger α (Fig. 3.5**B**).

Harvesting performance

Figure 3.6**A** compares the harvesting performance of the models, which is normalized by the performance of a near-optimal probabilistic decision making model. The near-optimal model knows the details of the schedules, i.e., both the baiting probabilities and the change points. It distributes its choices according to the choice probabilities that on average maximize the total reward [Sakai & Fukai, 2008a]. Due to such given knowledge, none of the other models can exceed the performance of the near-optimal model. We carry out paired t-tests between the models, in which the means of total reward for an identical schedule are paired. The FBM and WFBM ($L = 60$) are more inferior than the random choice model that chooses by tossing an unbiased coin. The DBM ($\alpha = 0.99$) outperforms FBM, WFBM, and LNP models ($p \ll 0.001$) but the differences from the LNP models are very small. Harvesting performance is less when a model memorizes a more distant past (Fig. 3.6**B**).

3.5 Discussion

We demonstrated that deterministic Bayesian decision making models can account for the matching law. We confirmed that a simple Bernoulli estimator with a deter-

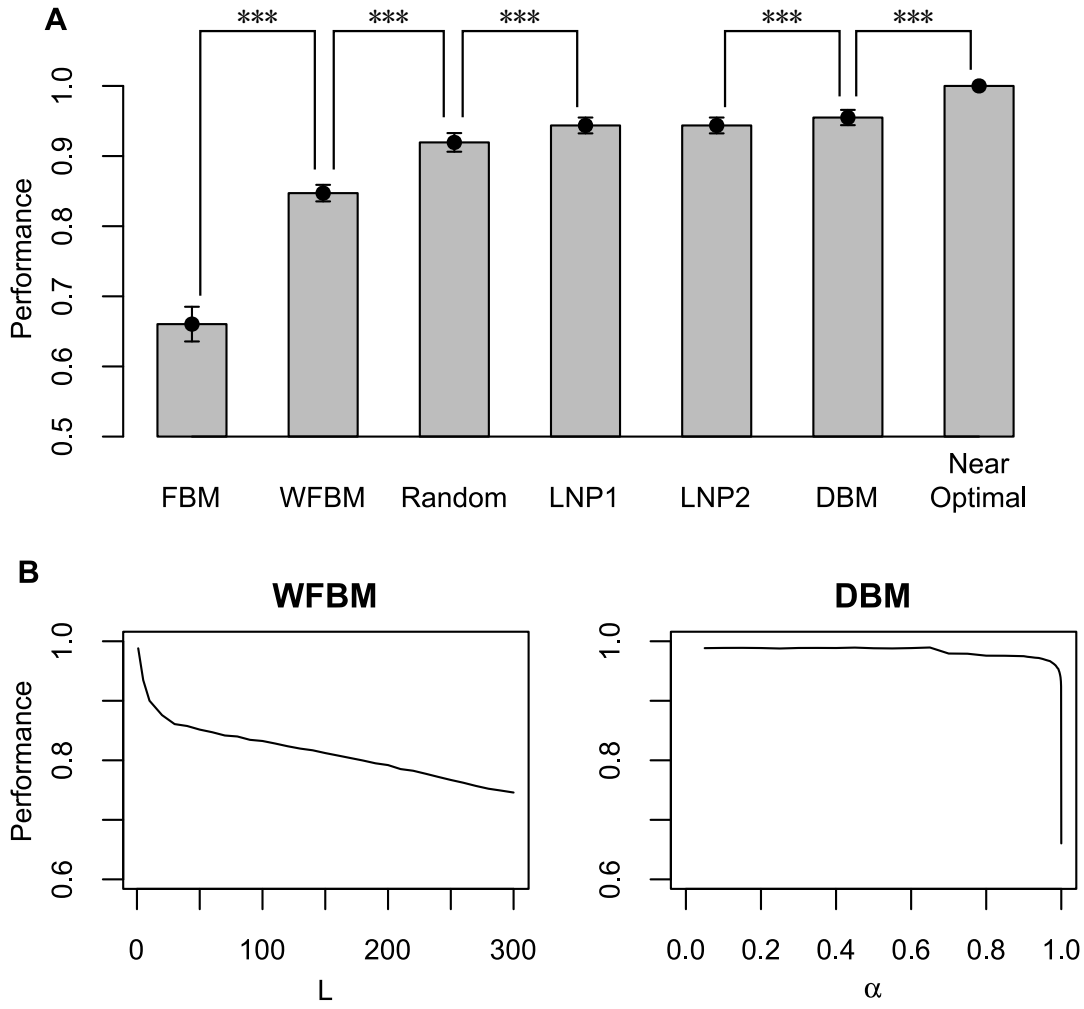


Figure 3.6: **Normalized harvesting performance of each model**. (**A**) Average normalized total rewards earned by each model divided by average total rewards of near-optimal model. Near-optimal model uses strategy that maximizes average total rewards proposed by Sakai and Fukai [Sakai & Fukai, 2008a] with previous knowledge on details of schedule. Error bars indicate standard deviations around mean. Simulation parameters were set to $L = 60$ and $\alpha = 0.99$. (**B**) Harvesting performance of WFBM and DBM as a function of their parameters.

ministic decision policy demonstrated matching behavior in a static foraging task. We also studied an extended model that includes a belief about a changing environment. The belief effectively works to wipe out the past experience of the model and hence the model can capture three characteristics of behaviors observed in the experiments. First, our model accounts for undermatching, which is a well-known phenomenon in which choices deviate slightly from the matching law [Baum, 1974, Baum, 1979, Sugrue et al., 2004]. Several studies have addressed possible causes of undermatching, i.e., limitations in the learning rule [Soltani & Wang, 2006], mistuning of parameters [Loewenstein, 2008], and diffusion of synaptic weights [Katahira et al., 2012]. We suggest a cause from a computational perspective, i.e., undermatching is the consequence of a belief in environmental volatility. Second, the run-length distribution of our model is better fitted by a double-exponential function than a single exponential function. This is also consistent with the previous study [Corrado et al., 2005] although our task did not include changeover delay, which can strongly affect the frequency of shorter run lengths. Third, our model exhibits double-exponential shaped reward trace dependency. This is consistent with recent monkey experiments [Corrado et al., 2005, Lau & Glimcher, 2005]. We propose that a reward trace may be an implementation of computation for a volatile environment.

The previous models implicitly or explicitly use the strategy of probabilistic choice selection and they learn the choice probability of respective alternatives that satisfy the matching law [Corrado et al., 2005, Lau & Glimcher, 2005, Loewenstein & Seung, 2006, Soltani & Wang, 2006, Simen & Cohen, 2009, Sakai & Fukai, 2008a, Katahira et al., 2012]. Such probabilistic models use a scaling parameter that maps internal decision variables to appropriate choice probabilities and the parameter generally requires fine-tuning [Soltani & Wang, 2006, Fusi et al., 2007]. In contrast, as our models act deterministically according to decision variables, no tuning is required for a parameter at the decision stage.

We argued that matching behavior can be explained by a deterministic choice strategy at the computational level. Loewenstein and Seung (2006) proposed biologically inspired synaptic learning rules for neural networks at the neural implementation level. They proved that neural networks developed by covariance-based learning with the assumption of a low learning rate demonstrated matching behaviors. However, this assumption causes the choice to be affected by relatively distant past rewards and the kernel for reward trace dependency consequently flattens. A more mi-

croscopic spiking neural network model, in which demonstrates double-exponential dependency in foraging tasks, has been proposed [Soltani & Wang, 2006]. However, there is a huge gap between the computational principles of our deterministic macroscopic models and their stochastic microscopic model. This gap can be filled by using a method of reducing spiking neuron models to the diffusion equation [Roxin & Ledberg, 2008]. There have been some other neural network models that can show heavy-tailed dependency of choices on past experience. A reservoir network [Jaeger et al., 2007], which can reproduce neural activity in the monkey prefrontal cortex, preserves the memory trace of a reward with one or two time constants [Bernacchia et al., 2011]. The composite learning system of faster and slower components is flexible to abrupt changes in the environment [Fusi et al., 2007]. These models could be a possible neural implementation for our model. Furthermore, our models are an extension of that by Yu & Cohen who argued that decision variables of their model can be approximated by a linear exponential filter, and that there are neural implementations for that operation [Yu & Cohen, 2009].

Because matching behavior often deviates from optimal behavior in the sense of total reward maximization [Vaughan Jr, 1981], it is not likely to be a consequence of optimization. However, our model acts optimally in terms of Bayesian decision making with an incorrect assumption about the environment, indicating that matching behavior is a bounded optimal behavior. This idea is consistent with the theory of Sakai and Fukai (2008) who found any learning method neglecting the effect of a choice on future rewards displays matching behavior if choice probabilities are differentiable with respect to parameters [Sakai & Fukai, 2008b]. Note that the choice probabilities of our model are not differentiable. Hence, we confirmed that their theory could be correct in such extreme cases.

Chapter 4

Neural Network Model for Future State Prediction

Recent behavioral, imaging, and computational studies suggest that humans and animals determine their actions based on the prediction about what will occur, or how states in an environment change, after the actions. However, it is still remain unclear how our brains learn to predict the state changes from one state to another. We propose and analyze an algorithm to estimate the probabilities of the state transitions based on a Hebbian learning algorithm. In addition, we confirm how eligibility trace affects our algorithm.

4.1 Introduction

Humans and animals can predict future states of environment based on acquired knowledge through experiences. Assuming that state is weather for instance, we predict that rainy days continue in rainy season and sunny days continue in dry season. Various predictive signals have been found in experiments [Duhamel et al., 1992, Schultz et al., 1997, Eskandar & Assad, 1999], and a recent study suggested that transition probabilities, which are essential to estimate future state, are encoded in the brain [Gläscher et al., 2010].

Many cerebral learning models which utilize the transition probabilities have been proposed [Rao, 2004, Beck & Pouget, 2007, Wacongne et al., 2012]. A previous model [Rao, 2004] can predict a future state and its neural activities are similar to

those in lateral intraparietal (LIP) area, however, this model assumes that the transition probabilities are already learned. There is another model which can predict a future state [Wacongne et al., 2012], but, in this model, the state transitions are learned in an unsupervised manner, contrasting to a supervised manner assumed in recent studies of reinforcement learning [Gläscher et al., 2010, Daw et al., 2005]. Besides, a delay line architecture is necessary for estimating a future state. The architecture has not yet being reported in predicting a future state. Hence, it is still unclear how the transition probabilities are learned in our brains.

We propose a synaptic learning rule for a feed-forward neural network so that the network learns the transition probabilities and predicts future states. Our learning rule is based on Hebbian learning algorithm [Hebb, 1949] and an presynaptic activity dependent weight decay [Kempter et al., 1999, Abbott & Nelson, 2000]. Numerical simulations and analytical calculations show that our model can learn the transition probabilities and predict future states. To compare our model and experimentally reported results, we simulate a random dots motion discrimination task. The predicted dot motion and neural activities of our model resemble those experimentally reported [Rao, 2004, Britten et al., 1992, Shadlen & Newsome, 2001, Roitman & Shadlen, 2002, Gold & Shadlen, 2003, Law & Gold, 2008]. In addition, we introduce the eligibility trace, in which is a widely assumed neural mechanism to keep memory [Klopf, 1972, Izhikevich, 2007], to our model. We show effects of the eligibility trace to the learning performance and an abrupt change of environment.

4.2 Markovian Environment

In our model, a state of environment, $\boldsymbol{z}$, stochastically changes from trial to trial following transition probabilities, $\mathcal{T}$. This state transition can be written as,

$$p(z_i^{t+1} = 1 | z_j^t = 1) = \mathcal{T}_{ij}, \tag{4.1}$$

where z_i^t represents whether the state is $i \in \{1, \ldots, N\}$ at trial t ($z_i^t = 1$) or not ($z_i^t = 0$), and N is number of states. Note that $\mathcal{T}$ is defined as a left stochastic matrix in which every sum of column vector is 1. Without a loss of generality, we assume that any states are reachable in finite transition from any other states, i.e., the Markov chain is irreducible.

There are several important statistics regarding the states and the transition probabilities for later analysis. The stationary distribution $\boldsymbol{\pi}$ of the Markov chain is defined

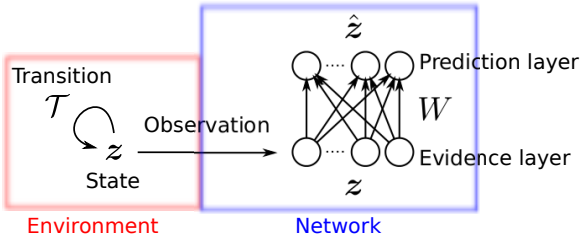


Figure 4.1: Schematic of environment and our neural network model.

as,

$$\pi_i \equiv \langle z_i \rangle = p(z_i = 1) = \sum_{j=1}^{N} p(z_i^\tau = 1 | z_j^{\tau-1} = 1) p(z_j = 1) = \sum_{j=1}^{N} \mathcal{T}_{ij} \pi_j, \quad (4.2)$$

where $\langle f(\boldsymbol{x}) \rangle$ represents an average, $\int d\boldsymbol{x}\, p(\boldsymbol{x}) f(\boldsymbol{x})$, and τ is a pseudo variable for representing temporal difference. Hence, the stationary distribution is the eigenvector of $\mathcal{T}$ with eigenvalue 1. Second moments of the state variable are,

$$\langle z_i z_j \rangle = \delta_{ij} \pi_i, \quad (4.3)$$

$$\langle z_i^\tau z_j^{\tau-l} \rangle = \sum_{k=1}^{N} \mathcal{T}_{ik} \langle z_k^\tau z_j^{\tau-(l-1)} \rangle = (\mathcal{T}^l)_{ij} \pi_j, \quad (4.4)$$

where δ is the Kronecker's delta.

4.3 Fully observable model

There are "evidence" and "prediction" layers in our network (Fig.4.1). Our network observes the state and the observed state determines neural activities in the evidence layer, $\boldsymbol{x} = \boldsymbol{z}$. Neural activities in the prediction layer, $\hat{\boldsymbol{x}}$, are determined by the weighted sum of the activities in the evidence layer, $\hat{\boldsymbol{x}}^t = W^t \boldsymbol{x}^t$, where $W \in R^{N \times N}$ is a matrix of synaptic weights. The constrained task to our model is to predict future state from current state by learning the transition probabilities. In other words, our model learns a model of the environment and, based on internal simulations of the model, generates a prediction of what will happen in the near (or far) future.

Following a previous MRI study [Gläscher et al., 2010], a learning rule is defined to minimize the average prediction error,

$$\langle E^t \rangle = \frac{1}{2} \langle \| \boldsymbol{x}^t - W^{t-1} \boldsymbol{x}^{t-1} \|^2 \rangle. \quad (4.5)$$

This average prediction error can be minimized by the following learning rule,

$$W_{ij}^t = (1 - \eta x_j^{t-1}) W_{ij}^{t-1} + \eta x_i^t x_j^{t-1}, \tag{4.6}$$

where $0 < \eta < 1$ is a learning rate. This update rule consists of activity dependent weight decay [Kempter et al., 1999, Abbott & Nelson, 2000] and Hebbian learning [Hebb, 1949].

4.3.1 Analysis

The average dynamics of our neural network model can be derived as well as [Werfel et al., 2005]. For simplicity, we assume that the correlation between W and $\boldsymbol{x}$ is negligible. The sufficient condition of satisfying this assumption is that the learning rate is infinitesimally small. We also suppose that $\boldsymbol{x}^t$ is defined also in $t < 0$. The average dynamics of W^t is as follows.

$$\begin{aligned} \langle W_{ij}^t \rangle &= \beta_j \langle W_{ij}^{t-1} \rangle + \eta \langle z_i^\tau z_j^{\tau-1} \rangle = \beta_j^t W_{ij}^0 + \eta \mathcal{T}_{ij} \pi_j \frac{1 - \beta_j^t}{1 - \beta_j} \\ &= \beta_j^t \left(W_{ij}^0 - \mathcal{T}_{ij} \right) + \mathcal{T}_{ij}, \end{aligned} \tag{4.7}$$

where $\beta_j = 1 - \eta \pi_j$. Because $0 < \beta_j < 1$ by definition, $\lim_{t \to \infty} \langle W_{ij}^t \rangle = \mathcal{T}_{ij}$. Thus, the transition probabilities are encoded in the connection weights by our learning rule. The average dynamics of second moment of the weight variable is as follows.

$$\begin{aligned} &\langle W_{ki}^t W_{kj}^t \rangle \\ &= \alpha_{ij} \langle W_{ki}^{t-1} W_{kj}^{t-1} \rangle + \eta C_{kij} \langle W_{ki}^{t-1} \rangle + \eta C_{kji} \langle W_{kj}^{t-1} \rangle + \delta_{ij} \eta^2 \langle z_k^\tau z_i^{\tau-1} \rangle \\ &= \alpha_{ij} \langle W_{ki}^{t-1} W_{kj}^{t-1} \rangle + \beta_i^{t-1} \eta C_{kij} (W_{ki}^0 - \mathcal{T}_{ki}) + \beta_j^{t-1} \eta C_{kji} (W_{kj}^0 - \mathcal{T}_{kj}) \\ &\quad + \eta C_{kij} \mathcal{T}_{ki} + \eta C_{kji} \mathcal{T}_{kj} + \delta_{ij} \eta^2 \langle z_k^\tau z_i^{\tau-1} \rangle \\ &= \alpha_{ij}^t W_{ki}^0 W_{kj}^0 + \frac{\beta_i^t - \alpha_{ij}^t}{\beta_i - \alpha_{ij}} \eta C_{kij} (W_{ki}^0 - \mathcal{T}_{ki}) + \frac{\beta_j^t - \alpha_{ij}^t}{\beta_j - \alpha_{ij}} \eta C_{kji} (W_{kj}^0 - \mathcal{T}_{kj}) \\ &\quad + \frac{1 - \alpha_{ij}^t}{1 - \alpha_{ij}} [\eta C_{kij} \mathcal{T}_{ki} + \eta C_{kji} \mathcal{T}_{kj} + \delta_{ij} \eta^2 \langle z_k^\tau z_i^{\tau-1} \rangle] \\ &= \alpha_{ij}^t [W_{ki}^0 W_{kj}^0 - C'_{kij} - C'_{kji} - C''_{kij}] + \beta_i^t C'_{kij} + \beta_j^t C'_{kji} + C''_{kij}, \end{aligned} \tag{4.8}$$

where the constants are

$$\begin{aligned}
\alpha_{ij} &= 1 - \eta(\pi_i + \pi_j) + \delta_{ij}\eta^2\pi_i, \\
C_{kij} &= (1 - \delta_{ij}\eta)\mathcal{T}_{kj}\pi_j, \\
C'_{kij} &= \frac{\eta C_{kij}(W^0_{ki} - \mathcal{T}_{ki})}{\beta_i - \alpha_{ij}}, \\
C''_{kij} &= \frac{\eta C_{kij}\mathcal{T}_{ki} + \eta C_{kji}\mathcal{T}_{kj} + \delta_{ij}\eta^2\mathcal{T}_{ki}\langle z_i\rangle}{1 - \alpha_{ij}}.
\end{aligned}$$

The second moment also converges because of $0 < \alpha_{ij} < 1$. The average prediction error is

$$\begin{aligned}
\langle E^t\rangle &= \frac{1}{2}\sum_j \left[\pi_j - 2\sum_i \langle W^{t-1}_{ji}\rangle \mathcal{T}_{ji}\pi_i + \sum_i \langle (W^{t-1}_{ji})^2\rangle \pi_i\right] \\
&= \frac{1}{2}\sum_i \alpha^{t-1}_{ii}\pi_i \sum_j [W^0_{ji}W^0_{ji} - 2C'_{jii} - C''_{jii}] \\
&\quad + \sum_i \beta^{t-1}_i \pi_i \sum_j \left[C'_{jii} - (W^0_{ji} - \mathcal{T}_{ji})\mathcal{T}_{ji}\right] + E_r, \qquad (4.9)
\end{aligned}$$

where $E_r \equiv \lim_{t\to\infty}\langle E^t\rangle$ is the residual error,

$$E_r = \frac{1}{2} + \frac{1}{2}\sum_i \pi_i \sum_k C''_{kii} - \sum_{i,j}\mathcal{T}^2_{ij}\pi_j. \qquad (4.10)$$

Thus, the learning curve (Eq.(4.9)) has two kinds of decaying components which correspond to the convergence of first and the second moments of connection weights.

4.3.2 Simulation Results

Learning of Transition Probabilities

At first, we confirm the analytical results based on numerical simulations. In a simulation, an initial state of environment has been sampled from the stationary distribution and initial weights are $W^0_{ij} = 0$ (same in other simulations unless otherwise specified). When the mean squared error between $\boldsymbol{W}$ and $\mathcal{T}$ is defined as $\mathrm{tr}[(\boldsymbol{W} - \mathcal{T})(\boldsymbol{W} - \mathcal{T})^T]/N^2$, this error decreases exponentially (Fig. 4.2**A**). The transition probabilities are thus successfully encoded into the weights (Fig. 4.2**B**). Although the transition probabilities is completely learned, the prediction error fluctuates around an asymptotic value (Figure 4.2**C**). This fluctuation and residual error

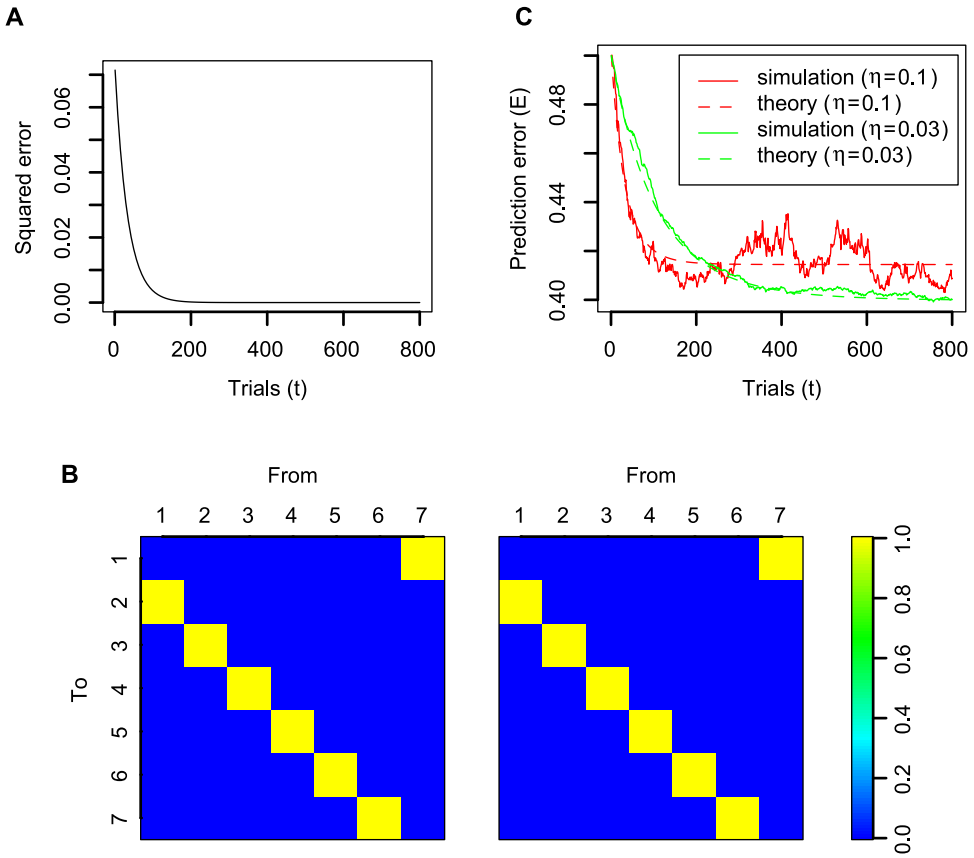


Figure 4.2: Simulation results when the state of environment is fully observable. We set the number of states $N = 7$ and the learning rate η is fixed to 0.1 in **A** and **B**, and varied in **C**. **A**: Mean squared error between the transition probability matrix and the weights matrix. **B**: Values of the transition probability matrix which all transitions are deterministic (left) and the weights matrix after 800 trials of learning (right). Labels indicate the state index. **C**: Solid and dashed lines show prediction errors and theoretically derived learning curves respectively. Color difference represents difference of learning rates (see the inset). A randomly generated transition probability matrix is used.

is unavoidable due to the stochasticity of the state transitions, however we could confirm that our algorithm can learn the environment.

Random Dot Discrimination Task

Our algorithm can estimate the probabilities of observing states, however it is still uncertain whether our algorithm is similar to an actual state prediction algorithm used by monkeys, humans, or other kinds of animals. To discuss the similarity, we conduct a numerical simulation which mimics a random dot discrimination task [Shadlen & Newsome, 2001]. If the neural activities are similar between our model and actual data and if estimated state can explain behavioral aspects of the animals, we can expect that our algorithm is a similar one to an actual state prediction algorithm used by the animals.

In the random dot discrimination task, subjects need to answer which direction the dots move after watching continually displayed dots. Some fraction of dots move toward the same direction, but the other dots move randomly. The more dots move coherently, the easier to decide the direction. After deciding the direction of the movement of the dots, subjects make a saccade towards the decided direction. Following previous studies, we assume two states, i.e., subjects need to make a decision whether dots move right or left.

Random dot stimulus has been simulated as follows. We suppose that a trial corresponds to observation of a motion direction of a dot. Let coherence be $0 < c < 1$, the probability of observing the coherent direction is $c + (1 - c)/2 = (1 + c)/2$. Without loss of generality, the coherent direction corresponds to state #1. Because there is no dependence between successive motion direction observations, the transition probability matrix is

$$\mathcal{T} = \begin{pmatrix} (1+c)/2 & (1+c)/2 \\ (1-c)/2 & (1-c)/2 \end{pmatrix}. \tag{4.11}$$

Suppose that subject's expectation to motion directions is neutral at the beginning of stimulus presentation hence we set $W_{ij}^0 = 0.5$.

At first, we perform 30 number of simulations and investigate neural activities of our model. When the random dots move rightward, an activity of the prediction unit increases if the unit is responsible for predicting the rightward direction and decreases otherwise (figure 4.3**A**). In addition, the larger the coherence, the larger (smaller) the activity. Activities of neurons in monkey's LIP (lateral intraparietal) show evidence

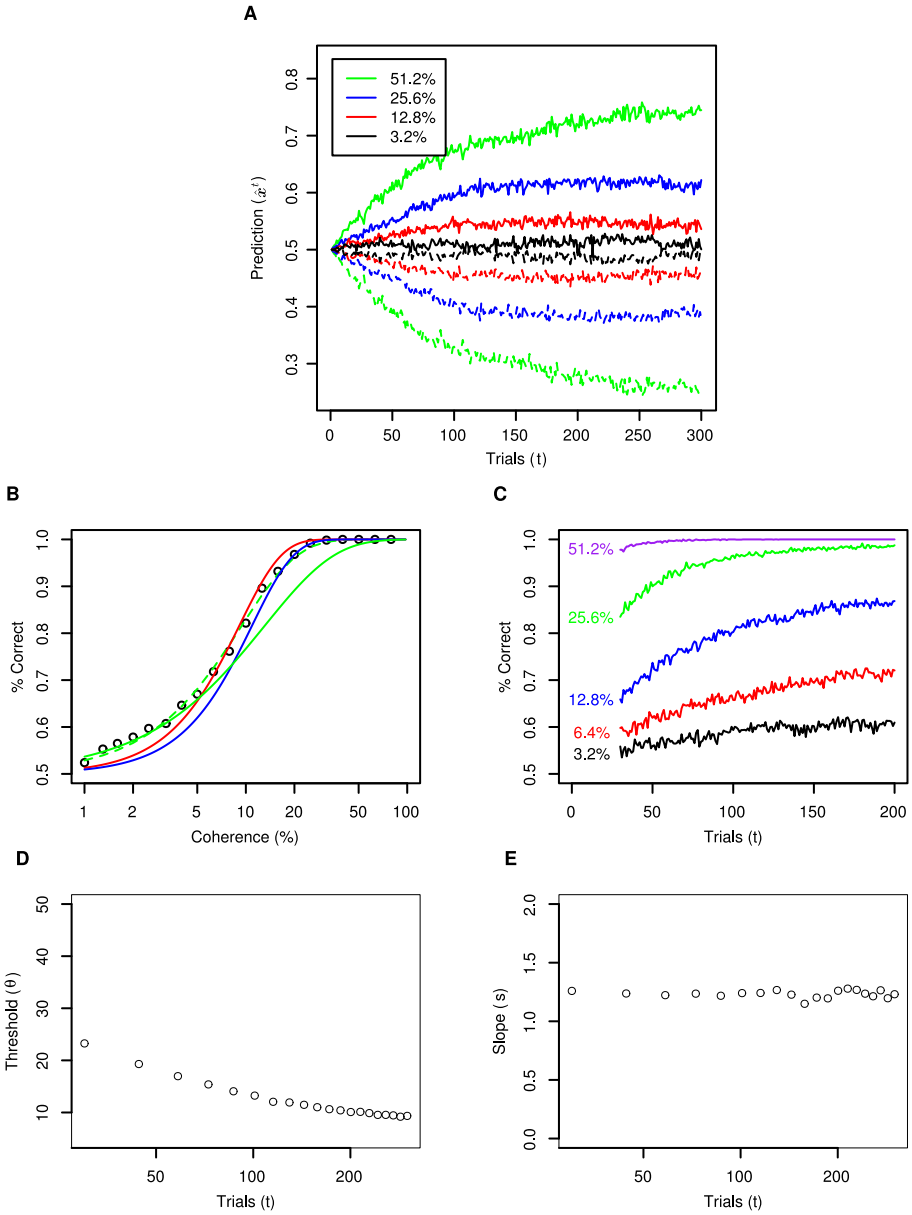


Figure 4.3: Simulation results of random dot stimulus motion discrimination task. We set the learning rate $\eta = 0.02$ and the number of motion directions is two ($N = 2$). **A**, The average estimated probabilities for coherent (solid) and incoherent (dashed) motion directions are plotted as a function of trial. Motion coherence is varied (see the inset). **B**, Probability of correct choice at 300-th trial is plotted as a function of stimulus coherence (symbols). Solid lines are psychometric functions reported in past monkey experiments: (green) [Shadlen & Newsome, 2001] and (red and blue) [Roitman & Shadlen, 2002]. **C**, Probability of correct choice increased over time and the slope is graded on coherence. **D** and **E** show the threshold θ and the shape s, respectively, of a cumulative Weibull function fitted to simulation data.

accumulation, e.g., when monkeys continue to look at the random dots that tend to move rightward, the activities of the neurons continue to increase (decrease) if the receptive field of the neurons are located to the rightward (leftward) to the fixation point [Roitman & Shadlen, 2002]. In addition, the speed of the increase (decrease) depends on a coherence of the random dots. Neural units in our model therefore show similar activities to neurons in monkey's LIP area.

Second, we compared the predicted states in our model with results of psychophysical experiments by 2000 number of simulations. We employed a *winner-take-all* policy as a decision making strategy for our network, i.e., our model decides that random dots move rightward (leftward) when the neural activity is higher in the neural unit responsible for predicting the rightward (leftward) movement than in the other unit. Our model makes a decision following a sigmoid function of coherence (Fig. 4.3**B**), this function corresponding to widely-reported psychometric functions in this task [Shadlen & Newsome, 2001, Roitman & Shadlen, 2002]. In addition, by fitting a Weibull function [Quick Jr., 1974] to simulation data (Fig. 4.3**C**), we investigate how the duration of the observation affects the decision. The Weibull function is defined as

$$p = 1 - \frac{1}{2}\exp(-(c/\theta)^s), \tag{4.12}$$

where p is a probability of correct choice, c is a coherence (%), θ is the threshold that a subject makes 82% correct choice and s is the slope of the psychometric function. When the duration of the observation increases, the threshold decreases log-linearly and the shape is invariant (Fig. 4.3**D**,**E**). Not only psychometric function but also these parameter dependencies coincide to those of experimentally reported [Gold & Shadlen, 2003, Law & Gold, 2008].

4.4 Partially observable models

In the previous section, we discuss under the assumption that our neural network model can observe and represent complete information of environment. However, in general, only partial information is available due to sensory limitations of organs, inherent neuronal variabilities [Stein et al., 2005], and so on. We investigate several partially observable models to show how our learning rule works in more realistic situations in this section.

Here, we consider two separate processes for the partial observation. First is the observation process, $\mathcal{A}$, which modifies input to the network,

$$\boldsymbol{x} = \mathcal{A}[\boldsymbol{z}]. \tag{4.13}$$

Second is the representation process, $\mathcal{R}$, which determines how the input is represented by neural activities, $\bar{\boldsymbol{x}}$. Because the neural activities can be affected by past neural activities,

$$\bar{\boldsymbol{x}}^\tau = \mathcal{R}[\boldsymbol{x}^\tau; \bar{\boldsymbol{x}}^{\tau-1}]. \tag{4.14}$$

Thus, the learning rule is extended by using $\bar{\boldsymbol{x}}$,

$$W_{ij}^{t+1} = (1 - \eta\bar{x}_j^t)W_{ij}^t + \eta x_i^{t+1}\bar{x}_j^t. \tag{4.15}$$

This learning rule is identical to Eq.(4.6) if the two processes are both identity transformation. Note that this learning rule is not guaranteed to minimize the average prediction error (Eq.(4.5)).

4.4.1 Analysis

As well as the previous section, we can derive the network dynamics by the extended learning rule (Eq.(4.15)). The convergence conditions are

$$\forall i, j|\bar{\alpha}_{ij}| < 1, \quad \forall i|\bar{\beta}_i| < 1, \tag{4.16}$$

where $\bar{\alpha}_{ij} \equiv \langle(1-\eta\bar{x}_i)(1-\eta\bar{x}_j)\rangle$ and $\bar{\beta}_i \equiv 1-\eta\langle\bar{x}_i\rangle$. The first and second moments of connection weights are

$$\langle W_{ij}^t\rangle = \bar{\beta}_j^t\left(W_{ij}^0 - W_{ij}^*\right) + W_{ij}^*, \tag{4.17}$$

$$\langle W_{ki}^t W_{kj}^t\rangle = \bar{\alpha}_{ij}^t[W_{ki}^0 W_{kj}^0 - C'_{kij} - C'_{kji} - C''_{kij}] + \bar{\beta}_i^t C'_{kij} + \bar{\beta}_j^t C'_{kji} + C''_{kij}, \tag{4.18}$$

where the constants are

$$\begin{aligned}
W_{ij}^* &= \frac{\langle x_i^\tau \bar{x}_j^{\tau-1}\rangle}{\langle\bar{x}_j\rangle}, \\
\bar{\alpha}_{ij} &= 1 - \eta(\langle\bar{x}_i\rangle + \langle\bar{x}_j\rangle) + \eta^2\langle\bar{x}_i\bar{x}_j\rangle, \\
C_{kij} &= \langle x_k^\tau \bar{x}_j^{\tau-1}\rangle - \eta\langle x_k^\tau \bar{x}_i^{\tau-1}\bar{x}_j^{\tau-1}\rangle, \\
C'_{kij} &= \frac{\eta C_{kij}(W_{ki}^0 - W_{ki}^*)}{\bar{\beta}_i - \bar{\alpha}_{ij}}, \\
C''_{kij} &= \frac{\eta C_{kij} W_{ki}^* + \eta C_{kji} W_{kj}^* + \eta^2\langle(x_k^\tau)^2 \bar{x}_i^{\tau-1}\bar{x}_j^{\tau-1}\rangle}{1 - \bar{\alpha}_{ij}}.
\end{aligned}$$

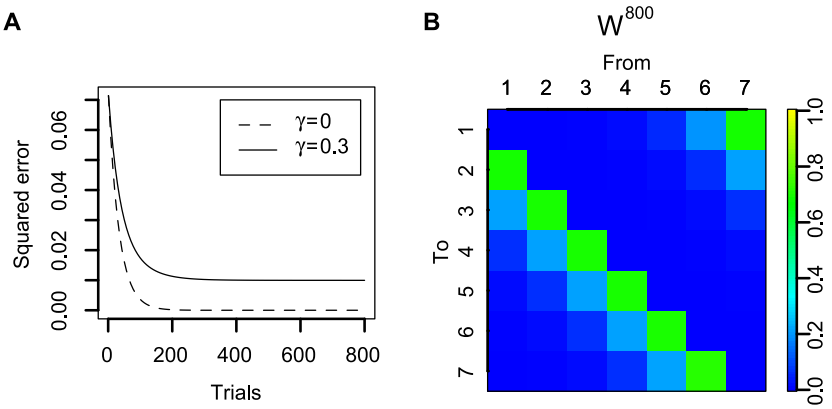


Figure 4.4: Learning results of model with the eligibility trace. We set the number of states $N = 7$ and the learning rate $\eta = 0.1$. The transition probability matrix is identical to Fig. 4.2**B**. **A**: Mean squared error between the transition probability matrix and the weights matrix where the time constant of eligibility trace $\gamma = 0.3$ (solid line) and $\gamma = 0$ (dashed line; same as Fig. 4.2**A**). **B**: Value of the weights matrix W after 800 trials of learning. We set $\gamma = 0.3$.

Hence, the average prediction error is

$$\begin{aligned}\langle E^t \rangle &= \frac{1}{2}\sum_{i,j}\bar{\alpha}_{ij}^{t-1}\langle x_i x_j\rangle \sum_k [W_{ki}^0 W_{kj}^0 - 2C'_{kij} - C''_{kij}] \\ &\quad + \sum_i \bar{\beta}_i^{t-1} \sum_j \left[\langle x_i x_j\rangle \sum_k C'_{kij} - (W_{ji}^0 - W_{ji}^*)\langle x_j^\tau x_i^{\tau-1}\rangle\right] + E_r\end{aligned} \tag{4.19}$$

where $E_r \equiv \lim_{t\to\infty}\langle E^t\rangle$ is the residual error,

$$E_r = \frac{1}{2}\sum_i \langle x_i^2\rangle + \frac{1}{2}\sum_{i,j}\langle x_i x_j\rangle \sum_k C''_{kij} - \sum_{i,j} W_{ij}^* \langle x_i^\tau x_j^{\tau-1}\rangle. \tag{4.20}$$

4.4.2 The effect of the eligibility trace

Recent studies suggested that "eligibility trace" [Klopf, 1972], a memory trace of inputs, is assumed to be available in a variety of brain regions [Pan et al., 2005, Izhikevich, 2007]. Here, we investigate the effect of the eligibility trace on our algorithm.

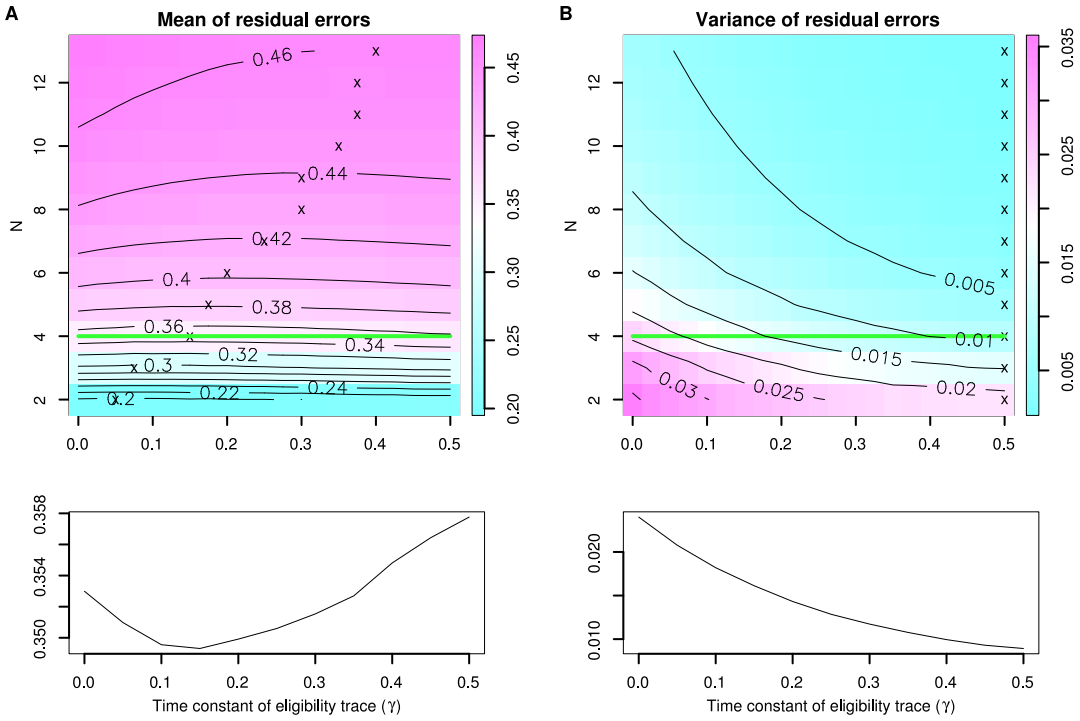


Figure 4.5: Effect of eligibility trace for residual error. Mean (**A**) and variance (**B**) of residual errors with respect to the time constant of eligibility trace γ and the number of states N. Symbol represent the minimum value against respective N. Lower panels are cutaway views of **A** and **B** where $N = 4$ (green lines), respectively. We set the learning rate $\eta = 0.1$.

The eligibility trace is defined as low-pass filtered input, thus the activities of evidence units are

$$\bar{\boldsymbol{x}}^t = \gamma \bar{\boldsymbol{x}}^{t-1} + (1-\gamma)\boldsymbol{x}^{t-l}, \tag{4.21}$$

where $0 < \gamma \leq 1/2$ is a time constant of eligibility trace. In this model, the important statistics become

$$\bar{\beta}_i = 1 - \eta\pi_i, \tag{4.22}$$

$$\bar{\alpha}_{ii} = 1 - 2\eta\pi_i + \eta^2\langle \bar{x}_i^2 \rangle, \tag{4.23}$$

$$W^*_{ij} = (1-\gamma)\sum_{k=0}^{\infty} \gamma^k (\mathcal{T}^{k+1})_{ij}. \tag{4.24}$$

It is not necessary to consider $\bar{\alpha}_{ij}$ $(i \neq j)$ and the convergence condition is satisfied by definition. The connection weights matrix, W, at steady state is a convolution of transition probabilities in different steps.

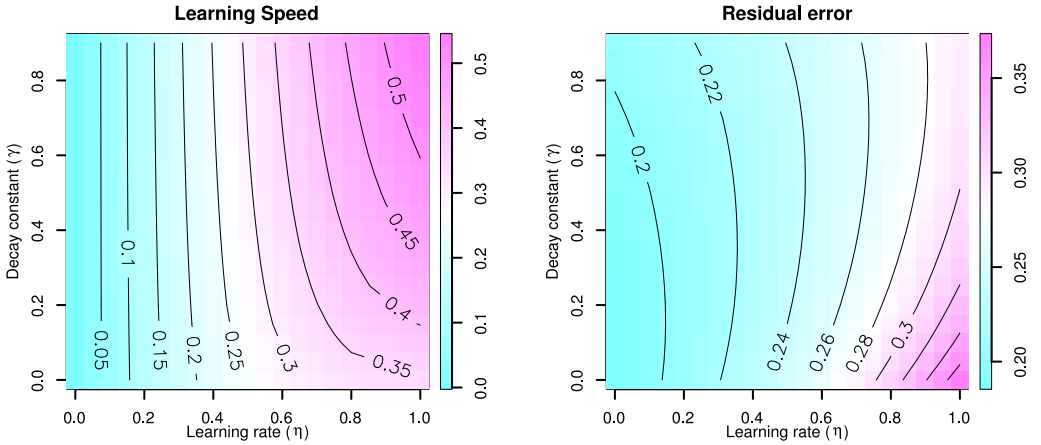


Figure 4.6: Learning speed and residual error as functions of the decay constant and the learning rate. We set the number of states $N = 2$.

Since several steps of transitions are memorized in the eligibility trace, this network model can also predict future states in several steps further. However, due to the change of the auto-covariance matrix, the weights no longer encode the transition probabilities completely (Fig. 4.4**A**, **B**). Although the learning slows down (Fig. 4.4**A**), the mean and variance of residual error can be decreased by eligibility trace (Fig. 4.5). The optimal time constant which minimizes the mean of residual error lies linearly against the number of states N (Fig. 4.5**A**). The eligibility trace plays as a low-pass filter, and a nature of a low-pass filter may exploit the cyclicity of state transitions and hence the error may decrease. Also due to a nature of a low-pass filter, the variance of residual error consistently decreases against increase of the time constant (Fig. 4.5**B**). However, the eligibility trace can degrade the learning speed, in which is defined by $\max_i 1-\bar{\alpha}_{ii}$. We simulated our network to show how the learning speed and the residual error change by the decay constant of eligibility trace and the learning rate (Fig. 4.6). We found that the learning speed are not greatly reduced by introduction of the eligibility trace when the learning rate is small. This is consistent that the effect of eligibility trace to the learning speed is $\mathcal{O}(\eta^2)$ (Eq.(4.23)). Hence, the eligibility trace can improve the learning performance without loss of learning speed in slow learning.

Additionally, we investigate effects of the eligibility trace when the environment abruptly changes. Based on the memory of state transitions in the eligibility trace, we can suppose that the eligibility trace enables to easily detect the change of the environment. We generate 8000 pairs of random transition probabilities. For each of pairs, 100 number of simulations are performed to calculate mean and variance of

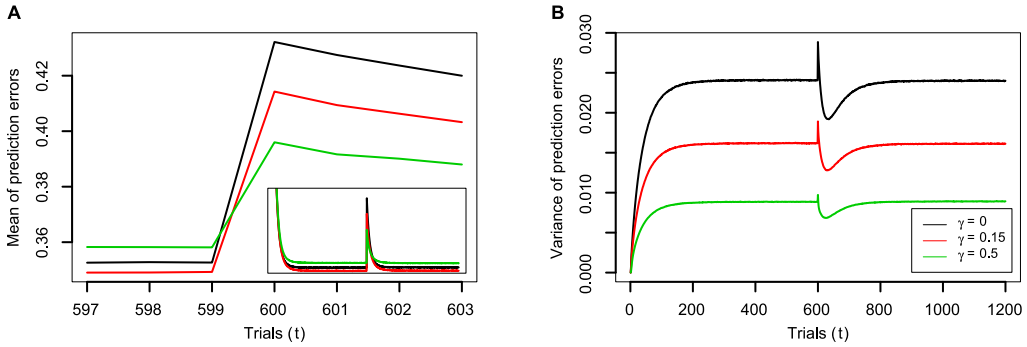


Figure 4.7: Effect of eligibility trace for prediction error where a transition probabilities changes at trial 600. Average time course of mean (**A**) and variance (**B**) of prediction errors against the time constant γ. **A** shows a magnified view around change point and the inset shows the overall time course. Color difference represents the difference of time constant (see the inset of **B**). We set the number of states $N = 4$ and the learning rate $\eta = 0.1$.

prediction errors. These statistics are again averaged over all pairs. As a result, it is shown that the eligibility trace can detect the rapid change of the environment, and the prediction error caused by the change of environment decreases (figure 4.7**A,B**). Thus, the eligibility trace can decrease the residual error and its variance, and can make our learning algorithm robust to the change of the environment.

4.4.3 Noisy linear transform

Here, we discuss the behavior of our neural network model when the observation process $\mathcal{A}$ is a noisy linear transform

$$\boldsymbol{x}^t = A\boldsymbol{z}^t + \boldsymbol{\epsilon}^t, \tag{4.25}$$

where $A \in R^{N\times N}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})$ and $\boldsymbol{I} \in R^{N\times N}$ is the identity matrix. Thus, we obtain

$$\bar{\beta}_i = 1 - \eta \sum_j A_{ij}\pi_j, \tag{4.26}$$

$$\bar{\alpha}_{ij} = 2 + N\sigma^2 - \sum_k \langle z_k\rangle(1+\eta A_{ik})(1+\eta A_{jk}), \tag{4.27}$$

$$W^* = A\mathcal{T}R, \tag{4.28}$$

$$R = \langle Z\rangle A^T \langle X\rangle^{-1}, \tag{4.29}$$

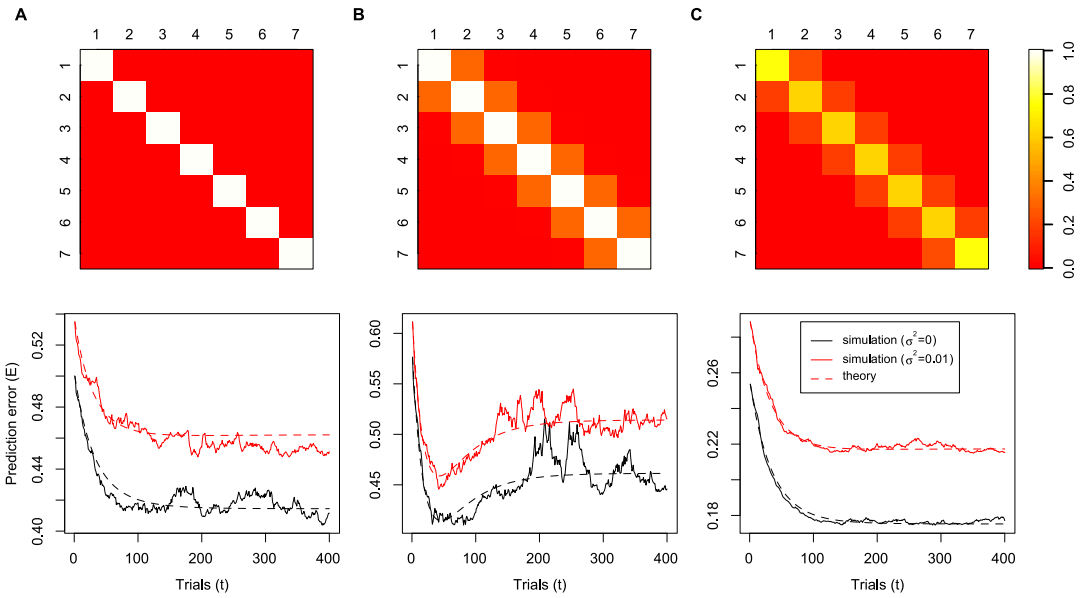


Figure 4.8: Each column shows simulation results with different linear transformation matrix A. The upper figures show the A_{ij} and lower figures show prediction errors (solid line) and theoretically derived learning curves (dashed line) in different noise levels: $\sigma^2 = 0$ (black lines) and $\sigma^2 = 0.01$ (red lines). We set the number of states $N = 7$, the learning rate $\eta = 0.1$ and an identical transition probabilities is used in all simulations.

where $\langle Z \rangle = \mathrm{diag}(\langle z_1 \rangle, \ldots, \langle z_N \rangle)$ and $\langle X \rangle = \mathrm{diag}(\langle x_1 \rangle, \ldots, \langle x_N \rangle)$. Thus, the convergence depends on the linear transformation and the noise level. Note that $W^* \boldsymbol{x}$ is expected to be an estimate of next input given present input $\boldsymbol{x}$. Because $A\mathcal{T}\boldsymbol{z}$ is an expected next input given present state $\boldsymbol{z}$, R seems to be an recovery transform of $\boldsymbol{z}$ given $\boldsymbol{x}$. Hence, the condition $\langle \boldsymbol{z} \rangle \equiv R\langle \boldsymbol{x} \rangle$ is expected to be satisfied for the network to correctly estimate the hidden state without bias. The recovery condition reduces to $\forall j \in \mathcal{S} \left(\sum_{i=1}^{N} A_{ij} = 1 \right)$.

Figure 4.8 shows simulation results in different linear transformations and noise levels with an identical transition probabilities. Figure 4.8**A** clarify that the noise solely degrades the predictive performance of network and slows down the learning. These are consistent in other cases (Fig. 4.8**B**,**C**). A sensory neuron often respond to similar stimulus to its preferred stimulus [Maunsell & Van Essen, 1983, Hubel & Wiesel, 1962]. Two linear transforms imitating this characteristic has been used where one violates the recovery condition (Fig. 4.8**B**) and the other is normalized to satisfy the condition (Fig. 4.8**C**). The violation of the recovery condition can

cause a dip in learning curve and the residual error can increase in comparison with the case of identity transform (Fig. 4.8**A**). On the other hand, the dip vanishes and the residual error can drastically decrease if the recovery condition is satisfied (Fig. 4.8**C**).

4.5 Discussion

We proposed an algorithm for estimating the environment, or state transition probabilities, based on a Hebbian synaptic rule [Hebb, 1949] and activity dependent weight decay [Kempter et al., 1999, Abbott & Nelson, 2000]. By analytical and numerical calculations, our algorithm was confirmed to be successful to estimate the environment and predict future states. It is proposed that humans estimate the environment to minimize state prediction error [Gläscher et al., 2010], however, neural implementation of the estimation remained unclear. Our model is a candidate of the neural implementation to estimate the environment and predict future states.

In a simulated random dot discrimination task, neural units in our model show similar activities to those in monkey's LIP area. A previous study proposed a neural network model whose neural activities are similar to those in monkey's LIP area [Rao, 2004], however the previous model assumed that the transition probabilities has been already learned. In contrast, our model can learn the probabilities. Although previous studies assumed that the neural activities in monkey's LIP area can be fit well by log-probability of states, our model does not have such an assumption. Furthermore, predicted states by our model shows a similar psychometric function to those reported in many psychophysical experiments [Britten et al., 1992, Shadlen & Newsome, 2001, Roitman & Shadlen, 2002, Gold & Shadlen, 2003, Law & Gold, 2008].

As a memory trace, we used a eligibility trace which is assumed to be biologically plausible especially in context of reinforcement learning [Pan et al., 2005]. Although a previous study assumed that a memory trace is indispensable for estimating the environment [Wacongne et al., 2012], our model suggests that the trace is not necessary for the estimation. A memory trace plays, however, other roles in the estimation; decreasing the residual error and generating a robustness to a rapid change of the environment.

The transition probabilities learned with the eligibility trace implicated that sequence of future states is predicted. This characteristic can degrade a pinpoint prediction but it is more computationally efficient for sequence prediction than traversing possible

state paths. Therefore, our model can underlie a process which requires rapid sequence prediction such as spoken language comprehension and motor control.

Chapter 5

Conclusion

The brain is a mysterious information processing machine. A discriminating feature of the brain is its flexibility which enables us to acquire new knowledge, abilities and skills by modifying its internal states. It is important to elucidate the cerebral learning for improving our life and advancing some scientific fields such as medical care, engineering and so on. The brain is a complex system because it consists of considerable biological matters. Therefore, it is quite difficult to understand our macroscopic behaviors by only investigating microscopic dynamics of the system. It is necessary to investigate the system in different scales and connect them.

The work in this dissertation has presented effects of information traces such as history of rewards, past neural activities, input and noise in the cerebral learning. Our approach is to build a computational or algorithmic model for known cerebral learning problems and investigate it by analytical calculations and numerical simulations to elucidate general effects of information trace.

In reward-modulated learning, reward delivery is distant from responses or behaviors which cause the reward. Besides, scalar reward signal is not informative to determine which part of brain or component of adaptive agent has been dedicated to the reward. Therefore, the brain or an artificial agent should solve the temporal and structural credit assignment problems of reward, i.e., the relationships between the reward and past responses in specific components should be determined for successful learning. We analyzed a multi-agent neural network model with trace of input and noise. We found that there is interactions between the structural and the temporal uncertainties of reward. The trace is necessary for solving the temporal credit assignment problem and also useful to reduce a degradation of learning by the structural

uncertainty (Chapter 2).

Matching law states the tendency of decision making behavior of humans and animals in tasks imitating foraging environment; fractions of choice and obtained reward to an alternative are equal. Observed behavior is mostly consistent with the law but undermatching, in which the choice behavior is slightly biased toward even choice, is sometimes observed. Recent studies unveiled the strong choice dependency to the recent reward trace and the dependency is well regressed by double exponential or hyperbolic functions, in which have longer tail than exponential function. The computational principle underlying these phenomena is still controversial. Therefore, we investigated the principle by comparing several Bayesian decision making models and found that the belief about the environment volatility is an essential factor of undermatching and the long tail reward trace dependency (Chapter 3).

The state transition probabilities are essential to predict future state of environment. In recent imaging study, it has been shown that the transition probabilities are encoded in the brain. However, it has been still unclear how the brain learns the transition probabilities. We proposed a learning algorithm which is based on Hebb rule and activity dependent weight decay. The neural network model with our learning algorithm can learn the transition probabilities and predict future state. We found that, by the trace of input, the residual error of learning is reduced, multiple sequences of future states are simultaneously predicted and the neural network model can rapidly adapt to change of environment (Chapter 4).

Thus, we found that, by the trial-wise traces, a learning system can resolve the structural and temporal credit assignment problems and can rapidly adapt to the abrupt change of environment. It seems that an information trace has effect to partially resolve some uncertainties.

Bibliography

[Abbott & Nelson, 2000] Abbott, L. F. & Nelson, S. B. (2000). Synaptic plasticity: taming the beast. *Nature neuroscience*, 3, 1178–1183.

[Abraham, 2003] Abraham, W. C. (2003). How long will long-term potentiation last? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 358(1432), 735–744.

[Agogino & Tumer, 2004] Agogino, A. K. & Tumer, K. (2004). Unifying temporal and structural credit assignment problems. In *Third International Joint Conference on Autonomous Agents and Multiagent Systems*, volume 2 (pp. 980–987). Washington, DC: IEEE Computer Society.

[Anderson et al., 2002] Anderson, K. G., Velkey, A. J., & Woolverton, W. L. (2002). The generalized matching law as a predictor of choice between cocaine and food in rhesus monkeys. *Psychopharmacology*, 163(3-4), 319–326.

[Baldi & Hornik, 1995] Baldi, P. F. & Hornik, K. (1995). Learning in linear neural networks: A survey. *IEEE Transactions on Neural Networks*, 6(4), 837–858.

[Barto, 1998] Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

[Baum, 1974] Baum, W. M. (1974). On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior*, 22(1), 231–242.

[Baum, 1979] Baum, W. M. (1979). Matching, undermatching, and overmatching in studies of choice. *Journal of the Experimental Analysis of Behavior*, 32(2), 269–281.

[Baum, 1981] Baum, W. M. (1981). Optimization and the matching law as accounts of instrumental behavior. *Journal of the Experimental Analysis of Behavior*, 36(3), 387–403.

[Baum, 1982] Baum, W. M. (1982). Choice, changeover, and travel. *Journal of the Experimental Analysis of Behavior*, 38(1), 35–49.

[Baum & Rachlin, 1969] Baum, W. M. & Rachlin, H. C. (1969). Choice as time allocation1. *Journal of the Experimental Analysis of Behavior*, 12(6), 861–874.

[Baum et al., 1999] Baum, W. M., Schwendiman, J. W., & Bell, K. E. (1999). Choice, contingency discrimination, and foraging theory. *Journal of the Experimental Analysis of Behavior*, 71(3), 355–373.

[Beck & Pouget, 2007] Beck, J. M. & Pouget, A. (2007). Exact inferences in a neural implementation of a hidden markov model. *Neural computation*, 19(5), 1344–1361.

[Bernacchia et al., 2011] Bernacchia, A., Seo, H., Lee, D., & Wang, X. J. (2011). A reservoir of time constants for memory traces in cortical neurons. *Nature Neuroscience*, 14, 366–372.

[Biehl et al., 2009] Biehl, M., Caticha, N., & Riegler, P. (2009). Statistical mechanics of on-line learning. In M. Biehl, B. Hammer, M. Verleysen, & T. Villmann (Eds.), *Similarity-Based Clustering: Recent Developments and Biomedical Applications*, volume 5400 of *Lecture Notes in Computer Science* chapter 1, (pp. 1–22). Heidelberg: Springer.

[Bishop, 1995] Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer New York.

[Britten et al., 1992] Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *The Journal of Neuroscience*, 12(12), 4745–4765.

[Brunel & Hakim, 1999] Brunel, N. & Hakim, V. (1999). Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. *Neural computation*, 11(7), 1621–1671.

[Chapman & Kaelbling, 1991] Chapman, D. & Kaelbling, L. P. (1991). Input generalization in delayed reinforcement learning: An algorithm and performance comparisons. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence* (pp. 726–731). San Francisco, CA: Morgan Kaufmann.

[Chung & Herrnstein, 1967] Chung, S.-H. & Herrnstein, R. J. (1967). Choice and delay of reinforcement1. *Journal of the Experimental Analysis of Behavior*, 10(1), 67–74.

[Corrado et al., 2005] Corrado, G. S., Sugrue, L. P., Seung, H. S., & Newsome, W. T. (2005). Linear-nonlinear-poisson models of primate choice dynamics. *Journal of the experimental analysis of behavior*, 84(3), 581–617.

[Cybenko, 1989] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4), 303–314.

[Dan & Poo, 2004] Dan, Y. & Poo, M. (2004). Spike timing-dependent plasticity of neural circuits. *Neuron*, 44(1), 23–30.

[Davison & McCarthy, 1988] Davison, M. & McCarthy, D. (1988). *The matching law: A research review*. Lawrence Erlbaum Associates, Inc.

[Daw et al., 2005] Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8, 1704–1711.

[Daw et al., 2006] Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095), 876–879.

[de Villiers & Herrnstein, 1976] de Villiers, P. A. & Herrnstein, R. J. (1976). Toward a law of response strength. *Psychological Bulletin*, 83(6), 1131–1153.

[Dembo & Kailath, 1990] Dembo, A. & Kailath, T. (1990). Model-free distributed learning. *IEEE Transactions on Neural Networks*, 1(1), 58–70.

[Duhamel et al., 1992] Duhamel, J.-R., Colby, C. L., & Goldberg, M. E. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science*, 255(5040), 90–92.

[Eskandar & Assad, 1999] Eskandar, E. N. & Assad, J. A. (1999). Dissociation of visual, motor and predictive signals in parietal cortex during visual guidance. *Nature neuroscience*, 2, 88–93.

[Fee et al., 2004] Fee, M. S., Kozhevnikov, A. A., & Hahnloser, R. H. R. (2004). Neural mechanisms of vocal sequence generation in the songbird. *Annals of the New York Academy of Sciences*, 1016(1), 153–170.

[Fiete et al., 2007] Fiete, I. R., Fee, M. S., & Seung, H. S. (2007). Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances. *Journal of Neurophysiology*, 98(4), 2038.

[Fiete & Seung, 2006] Fiete, I. R. & Seung, H. S. (2006). Gradient learning in spiking neural networks by dynamic perturbation of conductances. *Physical Review Letters*, 97(4), 48104.

[Flower & Jabri, 1993] Flower, B. & Jabri, M. (1993). Summed weight neuron perturbation: An o(n) improvement over weight perturbation. In C. L. Giles, S. J. Hanson, & J. D. Cowan (Eds.), *Advances in Neural Information Processing Systems*, volume 5 (pp. 212–219). San Mateo, CA: Morgan Kaufmann.

[Funahashi, 1989] Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3), 183–192.

[Fusi et al., 2007] Fusi, S., Asaad, W. F., Miller, E. K., & Wang, X. J. (2007). A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple timescales. *Neuron*, 54(2), 319–333.

[Gallistel, 1994] Gallistel, C. R. (1994). Foraging for brain stimulation: toward a neurobiology of computation. *Cognition*, 50(1-3), 151–170.

[Gittins & Jones, 1974] Gittins, J. C. & Jones, D. M. (1974). *Progress in Statistics*. Amsterdam, NL: North-Holland.

[Gläscher et al., 2010] Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4), 585–595.

[Glimcher, 2004] Glimcher, P. W. (2004). *Decisions, uncertainty, and the brain: The science of neuroeconomics*. The MIT press.

[Gold & Shadlen, 2003] Gold, J. I. & Shadlen, M. N. (2003). The influence of behavioral context on the representation of a perceptual decision in developing oculomotor commands. *The Journal of neuroscience*, 23(2), 632–651.

[Hamburger, 1992] Hamburger, V. (1992). History of the discovery of neuronal death in embryos. *Journal of neurobiology*, 23(9), 1116–1123.

[Hara et al., 2011] Hara, K., Katahira, K., Okanoya, K., & Okada, M. (2011). Statistical mechanics of on-line node-perturbation learning. *IPSJ Transactions on Mathematical Modeling and Its Applications*, 4(1), 72–81.

[Harris, 2008] Harris, K. D. (2008). Stability of the fittest: organizing learning through retroaxonal signals. *Trends in neurosciences*, 31(3), 130–136.

[Hebb, 1949] Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.

[Herrnstein, 1961] Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, 4(3), 267–272.

[Hessler & Doupe, 1999] Hessler, N. A. & Doupe, A. J. (1999). Social context modulates singing-related neural activity in the songbird forebrain. *Nature Neuroscience*, 2(3), 209–211.

[Heyman & Luce, 1979] Heyman, G. M. & Luce, R. D. (1979). Operant matching is not a logical consequence of maximizing reinforcement rate. *Learning and Behavior*, 7(2), 133–140.

[Hinson & Staddon, 1983] Hinson, J. M. & Staddon, J. E. R. (1983). Matching, maximizing, and hill-climbing. *Journal of the Experimental Analysis of Behavior*, 40(3), 321–331.

[Hornik, 1993] Hornik, K. (1993). Some new results on neural network approximation. *Neural Networks*, 6(8), 1069–1072.

[Hornik et al., 1989] Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5), 359–366.

[Houk, 2005] Houk, J. C. (2005). Agents of the mind. *Biological Cybernetics*, 92(6), 427–437.

[Hubel & Wiesel, 1962] Hubel, D. H. & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1), 106–154.

[Hughes, 1989] Hughes, J. (1989). Why functional programming matters. *The computer journal*, 32(2), 98–107.

[Hull, 1943] Hull, C. L. (1943). *Principles of behavior: an introduction to behavior theory*. New York: D. Appleton-Century Co. Inc.

[Irie & Miyake, 1988] Irie, B. & Miyake, S. (1988). Capabilities of three-layered perceptrons. In *Neural Networks, 1988., IEEE International Conference on* (pp. 641–648).

[Izhikevich, 2007] Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of stdp and dopamine signaling. *Cerebral Cortex*, 17(10), 2443–2452.

[Jabri & Flower, 1992] Jabri, M. & Flower, B. (1992). Weight perturbation: An optimal architecture and learning technique for analog vlsi feedforward and recurrent multi-layer networks. *IEEE Transactions on Neural Networks*, 3(1), 154–157.

[Jaeger et al., 2007] Jaeger, H., Lukoševičius, M., Popovici, D., & Siewert, U. (2007). Optimization and applications of echo state networks with leaky-integrator neurons. *Neural Networks*, 20(3), 335–352.

[Kaelbling et al., 1996] Kaelbling, L. P., Littman, M., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4, 237–285.

[Katahira et al., 2010] Katahira, K., Cho, T., Okanoya, K., & Okada, M. (2010). Optimal node perturbation in linear perceptrons with uncertain eligibility trace. *Neural Networks*, 23(2), 219–225.

[Katahira et al., 2012] Katahira, K., Okanoya, K., & Okada, M. (2012). Statistical mechanics of reward-modulated learning in decision-making networks. *Neural computation*, 24(5), 1230–1270.

[Kempter et al., 1999] Kempter, R., Gerstner, W., & Van Hemmen, J. L. (1999). Hebbian learning and spiking neurons. *Physical Review E*, 59(4), 4498–4514.

[Kinzel et al., 2001] Kinzel, W., Metzler, R., & Kanter, I. (2001). Statistical physics of interacting neural networks. *Physica A: Statistical Mechanics and its Applications*, 302(1-4), 44–55.

[Klopf, 1972] Klopf, A. H. (1972). *Brain function and adaptive systems: A heterostatic theory*. Technical Report AFCRL-72-0164, Air Force Research Laboratories, Bedford, MA.

[Lau & Glimcher, 2005] Lau, B. & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, 84(3), 555–579.

[Law & Gold, 2008] Law, C.-T. & Gold, J. I. (2008). Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. *Nature neuroscience*, 11, 505–513.

[Legenstein et al., 2009] Legenstein, R., Chase, S. A., Schwartz, A. B., & Maass, W. (2009). Functional network reorganization in motor cortex can be explained by reward-modulated hebbian learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems*, volume 22 (pp. 1105–1113). Cambridge, MA: MIT Press.

[Levy & Steward, 1983] Levy, W. B. & Steward, O. (1983). Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience*, 8(4), 791–797.

[Lin, 1992] Lin, L. J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8(3-4), 293–321.

[Loewenstein, 2008] Loewenstein, Y. (2008). Robustness of learning that is based on covariance-driven synaptic plasticity. *PLoS computational biology*, 4(3), e1000007.

[Loewenstein & Seung, 2006] Loewenstein, Y. & Seung, H. S. (2006). Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity. In *Proceedings of the National Academy of Sciences*, volume 103 (pp. 15224–15229). U.S.A.: National Academy of Sciences.

[Maunsell & Van Essen, 1983] Maunsell, J. H. & Van Essen, D. C. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. i. selectivity for stimulus direction, speed, and orientation. *Journal of neurophysiology*, 49(5), 1127–1147.

[Mazur, 2002] Mazur, J. E. (2002). *Learning and behavior*. Upper Saddle River, NJ, US: Prentice Hall/Pearson Education.

[McCulloch & Pitts, 1943] McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of mathematical biology*, 5(4), 115–133.

[Minsky, 1961] Minsky, M. L. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1), 8–30.

[Pan et al., 2005] Pan, W. X., Schmidt, R., Wickens, J. R., & Hyland, B. I. (2005). Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. *The Journal of neuroscience*, 25(26), 6235–6242.

[Pavlov, 1927] Pavlov, I. P. (1927). *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*. Oxford: Oxford University Press.

[Quick Jr., 1974] Quick Jr., R. F. (1974). A vector-magnitude model of contrast detection. *Kybernetik*, 16(2), 65–67.

[Rao, 2004] Rao, R. P. N. (2004). Bayesian computation in recurrent neural circuits. *Neural computation*, 16(1), 1–38.

[Roitman & Shadlen, 2002] Roitman, J. D. & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *The Journal of neuroscience*, 22(21), 9475–9489.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6), 386–408.

[Roxin & Ledberg, 2008] Roxin, A. & Ledberg, A. (2008). Neurobiological models of two-choice decision making can be reduced to a one-dimensional nonlinear diffusion equation. *PLoS Computational Biology*, 4(3), e1000046.

[Saad, 1999] Saad, D. (1999). *On-Line Learning in Neural Networks*. Cambridge, MA: Cambridge University Press.

[Saito et al., 2010] Saito, H., Katahira, K., Okanoya, K., & Okada, M. (2010). Statistical mechanics of the delayed reward-based learning with node perturbation. *Journal of the Physical Society of Japan*, 79(6), 064003.

[Saito et al., 2011] Saito, H., Katahira, K., Okanoya, K., & Okada, M. (2011). Statistical mechanics of structural and temporal credit assignment effects on learning in neural networks. *Physical Review E*, 83(5), 051125.

[Saito et al., 2014] Saito, H., Katahira, K., Okanoya, K., & Okada, M. (2014). Bayesian deterministic decision making: A normative account of the operant

matching law and heavy-tailed reward history dependency of choices. *Frontiers in Computational Neuroscience*. Accepted for publication.

[Sakai & Fukai, 2008a] Sakai, Y. & Fukai, T. (2008a). The actor-critic learning is behind the matching law: Matching versus optimal behaviors. *Neural computation*, 20(1), 227–251.

[Sakai & Fukai, 2008b] Sakai, Y. & Fukai, T. (2008b). When does reward maximization lead to matching law? *PLoS One*, 3(11), e3795.

[Salihoglu et al., 2009] Salihoglu, U., Bersini, H., Yamaguchi, Y., & Molter, C. (2009). Online organization of chaotic cell assemblies. a model for the cognitive map formation? In *Neural Networks, 2009. IJCNN 2009. International Joint Conference on* (pp. 2771–2777).

[Schultz et al., 1997] Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593–1599.

[Seung, 2003] Seung, H. S. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron*, 40(6), 1063–1073.

[Shadlen & Newsome, 2001] Shadlen, M. N. & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area lip) of the rhesus monkey. *Journal of Neurophysiology*, 86(4), 1916–1936.

[Simen & Cohen, 2009] Simen, P. & Cohen, J. D. (2009). Explicit melioration by a neural diffusion model. *Brain research*, 1299, 95–117.

[Singh & Sutton, 1996] Singh, S. P. & Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. *Recent Advances in Reinforcement Learning*, 22(1-3), 123–158.

[Skinner, 1938] Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. Oxford, England: Appleton-Century.

[Soltani & Wang, 2006] Soltani, A. & Wang, X. J. (2006). A biophysically based neural model of matching law behavior: melioration by stochastic synapses. *Journal of Neuroscience*, 26(14), 3731–3744.

[Stein et al., 2005] Stein, R. B., Gossen, E. R., & Jones, K. E. (2005). Neuronal variability: noise or part of the signal? *Nature Reviews Neuroscience*, 6, 389–397.

[Sugrue et al., 2004] Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science*, 304(5678), 1782–1787.

[Sugrue et al., 2005] Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: neural currencies for valuation and decision making. *Nature Reviews Neuroscience*, 6(5), 363–375.

[Sutton, 1984] Sutton, R. S. (1984). *Temporal credit assignment in reinforcement learning*. PhD thesis, University of Massachusetts, Amherst, MA.

[Sutton, 1988] Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, 3(1), 9–44.

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

[Tesauro, 1995] Tesauro, G. (1995). Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3), 58–68.

[Troyer & Doupe, 2000] Troyer, T. W. & Doupe, A. J. (2000). An associational model of birdsong sensorimotor learning i. efference copy and the learning of song syllables. *Journal of Neurophysiology*, 84(3), 1204.

[Tsodyks et al., 1998] Tsodyks, M., Pawelzik, K., & Markram, H. (1998). Neural networks with dynamic synapses. *Neural computation*, 10(4), 821–835.

[Vaughan Jr, 1981] Vaughan Jr, W. (1981). Melioration, matching, and maximization. *Journal of the Experimental Analysis of Behavior*, 36(2), 141–149.

[Wacongne et al., 2012] Wacongne, C., Changeux, J.-P., & Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. *The Journal of Neuroscience*, 32(11), 3665–3678.

[Wang, 2002] Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36(5), 955–968.

[Watkin et al., 1993] Watkin, T. L. H., Rau, A., & Biehl, M. (1993). The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2), 499–556.

[Werfel et al., 2005] Werfel, J., Xie, X., & Seung, H. S. (2005). Learning curves for stochastic gradient descent in linear feedforward networks. *Neural Computation*, 17(12), 2699–2718.

[Widrow & Lehr, 1990] Widrow, B. & Lehr, M. A. (1990). Thirty years of adaptive neural networks: Perceptron, madaline, and backpropagation. In *Proceedings of the IEEE*, volume 78 (pp. 1415–1441). New York: Wiley Online Library.

[Xie & Seung, 2004] Xie, X. & Seung, H. S. (2004). Learning in neural networks by reinforcement of irregular spiking. *Physical Review E*, 69(4), 041909.

[Yu & Cohen, 2009] Yu, A. J. & Cohen, J. D. (2009). Sequential effects: Superstition or rational behavior. In *Advances in Neural Information Processing Systems*, volume 21 (pp. 1873–1880).

Appendix A

Derivation of Ensemble Averages

In this section, we derive the ensemble averages for the single-perceptron model: $\langle \tilde{d}\boldsymbol{e} \rangle$, $\langle \tilde{d}\boldsymbol{J} \cdot \boldsymbol{e} \rangle$ and $\langle \tilde{d}^2 \|\boldsymbol{e}\|^2 \rangle$. Those for multi-perceptrons model can be derived in an analogous way.

Our model is difficult to analyze because the eligibility trace includes the past inputs and perturbations, the correlations among the past changes of $\boldsymbol{J}(m)$, the eligibility trace $\boldsymbol{e}(m)$, and the delayed instruction signal $\tilde{d}(m)$ should be considered. However, it can still be solvable. One reason is that the eligibility trace can be separated to the sum of current and past information since its kernel is exponential:

$$\boldsymbol{e}(m) \;=\; \xi(m)\boldsymbol{x}(m) + \varepsilon(1)\boldsymbol{e}(m-1). \tag{A.1}$$

This property makes the analysis relatively easy. We use following expansion to calculate the ensemble averages.

$$\boldsymbol{J}(m) \;=\; \eta \sum_{p=1}^{\infty} \tilde{d}(m-p)\boldsymbol{e}(m-p). \tag{A.2}$$

In terms of minimization of an objective function, the gradient information is most important. In our model, the information is thought to be included in $\tilde{d}\boldsymbol{e}$. Its average is,

$$\begin{aligned}
&\langle \tilde{d}(m)\boldsymbol{e}(m) \rangle \\
&= -\frac{1}{2}\langle \xi^2(m-m_d)\boldsymbol{e}(m) \rangle - \langle \xi(m-m_d)(y(m-m_d) - z(m-m_d))\boldsymbol{e}(m) \rangle \\
&= \frac{1}{N}\sigma^2 \varepsilon(m_d)(\boldsymbol{B} - \boldsymbol{J}(m-m_d)).
\end{aligned} \tag{A.3}$$

Thus, $\tilde{d}\boldsymbol{e}$ contains the difference between teacher and student weights weighted by the credit and the variance of noise. Then,

$$\begin{aligned}
&\langle \tilde{d}(m)\|\boldsymbol{e}(m)\|^2\rangle \\
&= -\frac{1}{2}\langle \xi^2(m-m_d)\|\boldsymbol{e}(m)\|^2\rangle - \langle \xi(m-m_d)(y(m-m_d)-z(m-m_d))\|\boldsymbol{e}(m)\|^2\rangle \\
&= -\frac{1}{2}\left\{ \begin{array}{l} \varepsilon^2(m_d)\left[\langle \xi^4(m-m_d)\rangle - \langle \xi^2(m-m_d)\rangle^2\right] \\ +\langle \xi^2(m-m_d)\rangle\langle \|\boldsymbol{e}(m)\|^2\rangle \end{array} \right\} + \mathcal{O}\left(\frac{1}{N}\right) \\
&= -\frac{1}{2}\sigma^4 D_1 + \mathcal{O}\left(\frac{1}{N}\right), \qquad \text{(A.4)}
\end{aligned}$$

where

$$D_k \equiv 2\varepsilon^2(m_d) + kI, \quad I \equiv \sum_{p=0}^{\infty}\varepsilon^2(p).$$

From Eqs. (A.1), (A.2) and (A.4), the inner product of the student weights and the eligibility trace is

$$\begin{aligned}
\langle \boldsymbol{J}(m)\cdot\boldsymbol{e}(m)\rangle &= \eta\sum_{p=1}^{\infty}\varepsilon(p)\langle \tilde{d}(m-p)\|\boldsymbol{e}(m-p)\|^2\rangle. \\
&= -\frac{1}{2}\eta\sigma^4 D_1 S + \mathcal{O}\left(\frac{1}{N}\right), \qquad \text{(A.5)}
\end{aligned}$$

where

$$S \equiv \sum_{p=1}^{\infty}\varepsilon(p).$$

From, Eq. (A.5),

$$\begin{aligned}
&\langle \xi^2(m-m_d)\boldsymbol{J}(m)\cdot\boldsymbol{e}(m)\rangle \\
&= \varepsilon(m_d)\langle \xi^2(m-m_d)\rangle\langle \boldsymbol{J}(m-m_d)\cdot\boldsymbol{e}(m-m_d)\rangle \\
&\quad +\eta\sum_{p=1}^{m_d}\varepsilon(p)\langle \xi^2(m-m_d)\tilde{d}(m-p)\|\boldsymbol{e}(m-p)\|^2\rangle \\
&= -\frac{1}{2}\eta\sigma^6\left[D_1 S + 2F\right]. \qquad \text{(A.6)}
\end{aligned}$$

where

$$F \equiv \varepsilon(m_d)\sum_{p=0}^{m_d-1}\varepsilon(p).$$

From Eq. (A.6), we obtain the ensemble average of inner product of the student weights $\boldsymbol{J}$ and the gradient vector $\tilde{d}\boldsymbol{e}$:

$$\begin{aligned}
&\langle \tilde{d}(m)\boldsymbol{J}(m)\cdot\boldsymbol{e}(m)\rangle \\
&= -\frac{1}{2}\langle \xi^2(m-m_d)\boldsymbol{J}(m)\cdot\boldsymbol{e}(m)\rangle \\
&\quad -\langle \xi(m-m_d)\left[y(m-m_d)-z(m-m_d)\right]\boldsymbol{J}(m)\cdot\boldsymbol{e}(m)\rangle. \\
&= \frac{1}{4}\eta\sigma^6\left[D_1 S+2F\right]-\sigma^2\varepsilon(m_d)\boldsymbol{J}(m)(\boldsymbol{J}(m-m_d)-\boldsymbol{B})+\mathcal{O}\left(\frac{1}{N}\right).
\end{aligned} \tag{A.7}$$

Our next interest is $\langle \tilde{d}^2\|\boldsymbol{e}\|^2\rangle$ which is a trace of variance-covariance matrix of the vector $\tilde{d}\boldsymbol{e}$. This average involves infinite number of higher-order correlations which is of $\mathcal{O}(1)$. However, we can drop the higher-order terms by the assumption of sufficiently small η. At first, two quantities are derived:

$$\begin{aligned}
&\langle \xi^4(m-m_d)\|\boldsymbol{e}(m)\|^2\rangle \\
&= \langle \xi^4(m-m_d)\rangle\left[\langle\|\boldsymbol{e}(m)\|^2\rangle-\varepsilon^2(m_d)\langle\xi^2(m-m_d)\rangle\right] \\
&\quad +\varepsilon^2(m_d)\langle\xi^4(m-m_d)\xi^2(m-m_d)\rangle \\
&= 3\sigma^6 E_1,
\end{aligned} \tag{A.8}$$

$$\begin{aligned}
&\langle \xi^3(m-m_d)[y(m-m_d)-z(m-m_d)]\|\boldsymbol{e}(m)\|^2\rangle \\
&= \frac{1}{N}6\sigma^4\varepsilon^2(m_d)\langle[\boldsymbol{J}(m-m_d)-\boldsymbol{B}]\cdot\boldsymbol{e}(m-m_d)\rangle \;\sim\; \mathcal{O}\left(\frac{1}{N}\right),
\end{aligned} \tag{A.9}$$

where

$$E_k \;\equiv\; 4\varepsilon^2(m_d)+kI.$$

Thus, by Eqs. (A.8) and (A.9),

$$\begin{aligned}
\langle \tilde{d}^2(m)\|\boldsymbol{e}(m)\|^2\rangle \;&=\; \frac{3}{4}\sigma^6 E_1+\mathcal{O}\left(\frac{1}{N}\right) \\
&\quad +\left\langle\left\{\xi(m-m_d)[y(m-m_d)-z(m-m_d)]\right\}^2\|\boldsymbol{e}(m)\|^2\right\rangle.
\end{aligned} \tag{A.10}$$

The last term of the above equation is

$$\begin{aligned}
&\left\langle \{\xi(m-m_d)[y(m-m_d)-z(m-m_d)]\}^2 \|\boldsymbol{e}(m)\|^2 \right\rangle \\
&= \sigma^2 \left\langle \{\xi(m-m_d)[y(m-m_d)-z(m-m_d)]\}^2 \right\rangle \sum_{p=0}^{m_d-1} \varepsilon^2(p) \\
&\quad +\varepsilon^2(m_d) \left\langle \{\xi(m-m_d)[y(m-m_d)-z(m-m_d)]\}^2 \|\boldsymbol{e}(m-m_d)\|^2 \right\rangle \\
&= \sigma^4 \left[l^2(m-m_d) - 2r(m-m_d) + 1 \right] \sum_{p=0}^{m_d-1} \varepsilon^2(p) \\
&\quad +\varepsilon^2(m_d) \left[\begin{array}{l} \langle \xi^2(m-m_d) y^2(m-m_d) \|\boldsymbol{e}(m-m_d)\|^2 \rangle \\ +\langle \xi^2(m-m_d) z^2(m-m_d) \|\boldsymbol{e}(m-m_d)\|^2 \rangle \\ -2\langle \xi^2(m-m_d) y(m-m_d) z(m-m_d) \|\boldsymbol{e}(m-m_d)\|^2 \rangle \end{array} \right].
\end{aligned} \tag{A.11}$$

From

$$\begin{aligned}
&\frac{1}{N}\langle \|\boldsymbol{J}(m)\|^2 \|\boldsymbol{e}(m-1)\|^2 \rangle \\
&\simeq \frac{1}{N} \left\langle \left[\|\boldsymbol{J}(m-1)\|^2 + 2\eta \tilde{d}(m-1) \boldsymbol{J}(m-1)\cdot \boldsymbol{e}(m-1) \right] \|\boldsymbol{e}(m-1)\|^2 \right\rangle \\
&= \frac{1}{N} \left\langle \|\boldsymbol{J}(m-1)\|^2 \left[\begin{array}{l} \xi^2(m-1)\|\boldsymbol{x}(m-1)\|^2 + \varepsilon^2(1)\|\boldsymbol{e}(m-2)\|^2 \\ +2\varepsilon(1)\xi(m-1)\boldsymbol{x}(m-1)\cdot \boldsymbol{e}(m-2) \end{array} \right] \right\rangle + \mathcal{O}\left(\frac{1}{N}\right) \\
&\simeq \sigma^2 l^2(m-1) + \frac{1}{N}\varepsilon^2(1)\langle \|\boldsymbol{J}(m-1)\|^2 \|\boldsymbol{e}(m-2)\|^2 \rangle \\
&\simeq \sigma^2 \sum_{p=0}^{\infty} \varepsilon^2(p) l^2(m-1-p),
\end{aligned} \tag{A.12}$$

we obtain

$$\begin{aligned}
&\langle \xi^2(m) y^2(m) \|\boldsymbol{e}(m)\|^2 \rangle \\
&= \sum_{k}^{N} \langle \xi^2(m) J_k^2(m) x_k^2(m) \|\boldsymbol{e}(m)\|^2 \rangle \\
&= \sum_{k}^{N} \left\{ \sum_{l}^{N} \langle \xi^4(m) \rangle \langle x_k^2(m) x_l^2(m) \rangle J_k^2(m) + \sigma^2 \frac{1}{N} \varepsilon^2(1) \langle J_k^2(m) \|\boldsymbol{e}(m-1)\|^2 \rangle \right\} \\
&\simeq \sigma^4 \left[3l^2(m) + \varepsilon^2(1) \sum_{p=0}^{\infty} \varepsilon^2(p) l^2(m-1-p) \right].
\end{aligned} \tag{A.13}$$

From

$$\begin{aligned}
&\frac{1}{N}\langle \boldsymbol{J}(m)\cdot\boldsymbol{B}\|\boldsymbol{e}(m-1)\|^2\rangle \\
&= \frac{1}{N}\left\langle\left[\boldsymbol{J}(m-1)+\eta\tilde{d}(m-1)\boldsymbol{e}(m-1)\right]\cdot\boldsymbol{B}\|\boldsymbol{e}(m-1)\|^2\right\rangle \\
&= \frac{1}{N}\left\langle \boldsymbol{J}(m-1)\cdot\boldsymbol{B}\left[\begin{array}{l}\xi^2(m-1)\|\boldsymbol{x}(m-1)\|^2+\varepsilon^2(1)\|\boldsymbol{e}(m-2)\|^2\\ +2\varepsilon(1)\xi(m-1)\boldsymbol{x}(m-1)\cdot\boldsymbol{e}(m-2)\end{array}\right]\right\rangle+\mathcal{O}\left(\frac{1}{N}\right) \\
&\simeq \sigma^2 r(m-1)+\varepsilon^2(1)\frac{1}{N}\langle \boldsymbol{J}(m-1)\cdot\boldsymbol{B}\|\boldsymbol{e}(m-2)\|^2\rangle \\
&\simeq \sigma^2\sum_{p=0}^{\infty}\varepsilon^2(p)r(m-1-p),
\end{aligned} \tag{A.14}$$

we obtain

$$\begin{aligned}
\langle\xi^2(m)y(m)z(m)\|\boldsymbol{e}(m)\|^2\rangle &= \sum_k^N\langle\xi^2(m)J_k(m)B_k x_k^2(m)\|\boldsymbol{e}(m)\|^2\rangle \\
&= \sum_k^N\left[\begin{array}{l}\displaystyle\sum_l^N J_k(m)B_k\langle\xi^4(m)\rangle\langle x_k^2(m)x_l^2(m)\rangle \\ +\sigma^2\varepsilon^2(1)\dfrac{1}{N}\langle J_k(m)B_k\|\boldsymbol{e}(m-1)\|^2\rangle\end{array}\right] \\
&\simeq \sigma^4\left[3r(m)+\varepsilon^2(1)\sum_{p=0}^{\infty}\varepsilon^2(p)r(m-1-p)\right].
\end{aligned} \tag{A.15}$$

Then,

$$\begin{aligned}
&\langle\xi^2(m)z^2(m)\|\boldsymbol{e}(m)\|^2\rangle \\
&= \sum_k^N B_k^2\langle\xi^2(m)x_k^2(m)\|\boldsymbol{e}(m)\|^2\rangle \\
&= \sum_k^N B_{jk}^2\left[\sum_l^N\langle\xi^4(m)\rangle\langle x_k^2(m)x_l^2(m)\rangle+\sigma^4\varepsilon^2(1)\frac{1}{N}\sum_{p=0}^{\infty}\varepsilon^2(p)\right] \\
&= \sigma^4\left[3+\varepsilon^2(1)\sum_{p=0}^{\infty}\varepsilon^2(p)\right].
\end{aligned} \tag{A.16}$$

Following Eqs. (A.11), (A.13), (A.15) and (A.16),

$$\begin{aligned}
&\left\langle \{\xi(m-m_d)[y(m-m_d)-z(m-m_d)]\}^2 \|\boldsymbol{e}(m)\|^2 \right\rangle \\
&\simeq \sigma^4 \left[\sum_{p=0}^{m_d} \varepsilon^2(p) + 2\varepsilon^2(m_d)\right] [l^2(m-m_d) - 2r(m-m_d) + 1] \\
&\quad +\sigma^4 \sum_{p=m_d+1}^{\infty} \varepsilon^2(p) \left[l^2(m-p) - 2r(m-p) + 1\right].
\end{aligned} \tag{A.17}$$

Finally, we obtain

$$\begin{aligned}
&\langle \tilde{d}^2(m) \|\boldsymbol{e}(m)\|^2 \rangle \\
&= \frac{3}{4}\sigma^6 E_1 + \sigma^4 \left[\sum_{p=0}^{m_d} \varepsilon^2(p) + 2\varepsilon^2(m_d)\right] [l^2(m-m_d) - 2r(m-m_d) + 1] \\
&\quad +\sigma^4 \sum_{p=m_d+1}^{\infty} \varepsilon^2(p) \left[l^2(m-p) - 2r(m-p) + 1\right].
\end{aligned} \tag{A.18}$$

We have calculated the ensemble averages in discrete time domain to involve the effect of the delay. Then we transfer the obtained ensemble averages into continuous time domain. We carry out this operation according to following rules. Let X be a variable,

1. $X(m-m_d) \to X(t)$, because m_d is infinitesimally small in continuous time domain.
2. $\sum_{p=0}^{\infty} \varepsilon^2(p) X(m-p) \to IX(t)$, because X can be regarded as fixed where the kernel has sufficiently large amplitude.

Thus,

$$\boldsymbol{B} \cdot \langle \tilde{d}\boldsymbol{e} \rangle = \sigma^2 \epsilon(m_d)(1 - r(t)), \tag{A.19}$$

$$\langle \tilde{d}\boldsymbol{J} \cdot \boldsymbol{e} \rangle = \frac{1}{4}\eta\sigma^6 \left[D_1 S + 2F\right] - \sigma^2 \varepsilon(m_d)(l^2(t) - r(t)), \tag{A.20}$$

$$\langle \tilde{d}^2 \|\boldsymbol{e}\|^2 \rangle = \frac{3}{4}\sigma^6 E_1 + \sigma^4 D_1 (l^2(t) - 2r(t) + 1). \tag{A.21}$$