Sequential Decision Making in Biological Systems: The Role of Nonlinear Dynamical Phenomena in Working Memory and Reinforcement Learning in Long-Term Memory

（生物システムの逐次的意思決定：作業記憶の非線形
ダイナミクスの役割と長期記憶での強化学習の役割）

1997

University of Tokyo

Hiroyuki Nakahara

①

博 士 論 文

Sequential Decision Making in Biological Systems: The Role of Nonlinear Dynamical Phenomena in Working Memory and Reinforcement Learning in Long-Term Memory

（生物システムの逐次的意思決定：作業記憶の非線形ダイナミクスの役割と長期記憶での強化学習の役割）

1997年3月

東京大学 総合文化研究科
広域科学専攻 広域システム科学系

中原 裕之

Abstract

Sequential decision making is a fundamental task for any biological systems. A decision made at a time in sequential decision making has an immediate as well as long-term consequences. Both consequences, hence, should be considered to optimize sequential decisions. This thesis investigates the role of two different kinds of memory systems, working memory and long-term memory, in sequential decision making. The important difference between these systems is that working memory stores information as neural activities, whereas long-term memory stores it as synaptic strength.

First, to investigate a role of working memory in sequential decision making, the dynamics of neural activities is examined. The long-term maintenance and quick transition of neural activities is proposed as crucial in sequential decision making. Such property can be found in near saddle-node bifurcation dynamics. In simulations of foraging tasks, the proposed dynamics emerges in recurrent networks that control the movement of a creature as a functional necessity for survival in non-stationary environments.

Secondly, the functions of the loops of basal ganglia, a subsystem for one of the long-term memory systems, are investigated in sequential decision making in relation to reinforcement learning. The hypothesis is given for the functions of the loops and the performance of the model based on the hypothesis is examined in comparison with the experiment by Hikosaka and his colleagues. The model replicates their data in several aspects and predicts the monkeys' behavior in condition that is not experimentally tested yet.

Finally, the findings of above studies are summarized, followed by the discussion of their limitations and future works.

Acknowledgment

First of all, I would like to acknowledge Professor Saburo Nagano for supervising my dissertation work as my advisor. Professor Shinsuke Shimojo has kindly been my important mentor for a whole period of my Ph.D student life and has led me to develop an interest of psychophysical and computational works. Dr. Kenji Doya has patiently helped me step into computational research that turned out to be my dissertation work in close collaboration. Professor Okihide Hikosaka kindly let me in their lab activities. Discussion with him and his students, particularly S. Miyachi, K. Miyashita, K. Sakai, and X. Lu helped me get focus in the second study of the thesis. I thank to my thesis committee to refine my thesis. I thank to Dr. Kazuhiro Ueda, who has been a good "senpai", and others in Nagano lab, particularly K. Yonezawa, O. Watanabe, T. Kojima, and M. Ichinose. I appreciate the technical contribution of Reiko Nebashi as well as her encouragement, which boosted writing the thesis. I also appreciate friendship with Adrian Robert as well as his proof-reading and valuable comments on my thesis. I believe that my experience of research activities in the labs of Professor Yutaka Sayeki and Professor Kunihiko Kaneko contributed to shape the issues investigated in the thesis, though it may not be apparently obvious.

Out of University of Tokyo, I spent a fruitful time at Department of Cognitive Science, University of California, San Diego. After two years, I feel as if it is my second-home department and second-home university. Professor Edwin Hutchins kindly hosted me during the period and I appreciate his generous support for my research very much. The department provided me with an excellent research environment, through which I gained a better understanding of the fundamentals in this field. I appreciate the opportunity to interact with the many fine researchers at UCSD. As a result of this interaction and the knowledge I gained, my dissertation is richer than it would have otherwise been.

From my stay, there are too many people to whom I am grateful to name. However I would especially like to thank David Rogers, Adrian Robert, Paul Rodriguez, Michael Gray, Adam Long, Kevin Haynes and Clarence Wong who made my stay pleasant both academically and socially.

I acknowledge financial support from University of Tokyo, Rotary One Foundation, and Japan Society for the Promotion of Science.

As far as I experienced, it is hard to work on a dissertation without a cheerful circle of friends. In this respect, there are too many friends I would like to thank by name, so I won't. I wish, however, my greatest gratitude to all of them.

Finally but most importantly, I thank my parents and my brother. Without them, this thesis would not exist.

Contents

Chapter 1

INTRODUCTION

Sequential decision making is a fundamental task that any biological systems confront in interaction with the environment. It is important to note that a decision made at a time in sequential decision making has immediate as well as long-term consequences. Therefore, both consequences should be taken into account to optimize sequential decisions. This point of considering both consequences is a fundamental problem that is found in any sequential decision making tasks. This problem is known as *temporal credit assignment problem* [23, 72]. Adaptability of biological systems, including the case of sequential decision making, is central to their intelligent behaviors, and occurs at multiple time scales from subseconds to life-long time and the resulting adaptation is also preserved over different time periods. This capability is supported by multiple memory systems.

1.1 Working memory and the long-term memory for sequential decision making

Basic functional components of any memory systems are loading, storing (or maintaining) and retrieving. In theory, it is possible to make any memory system adaptive by changing any of these components if such changing fits well with the aim of the adaptation. The difficulty of changing each of these components, however, varies, depending upon type of a memory system, or the way of setting these components in a memory system.

One of the oldest and most widely accepted distinction of memory systems is the distinction between the short-term memory (STM), or working memory, and the long-term memory (LTM) [68, 77]. This distinction corresponds with two different ways of maintaining information in memory: one active and the other latent [18, 66, 82]. An active memory storage, or the STM, maintains information as neural activities, or firing in neurons, so that the information can be stored only for relatively a short time. In contrast, a latent memory storage, or the LTM, embeds information in physiological parameters such as synaptic strength in which the information can be preserved for a long time [64, 66, 82]. This difference imposes different constraints on the adaptability between the STM and the LTM in sequential decision making.

Because neural activity by itself stands for stored information at a time in the STM, active inputs are easily confused with stored information. In order to resolve temporal credit assignment problem in the STM, therefore, it is very important to investigate how such neural activities that stand for the stored information and the active inputs are maintained. The first half of this thesis focuses on this issue. It is investigated how active neural firings, activated by different input sources, should be maintained and stopped from a viewpoint of dynamical systems. By jumping up to the proposal made in Chapter 2, it is discussed that the long-term maintenance and quick transition of neural activities are important in the STM for sequential decision making. Since it is somewhat confusing to have a term such as "long-term maintenance in short term memory (STM)", working memory will, by a slight abuse of terms, substitute for the STM throughout this thesis in the following chapters. The relationship between the STM and working memory is briefly discussed in Chapter 2, and the definition of working memory in this thesis is also stated in Chapter 2.

The rest of the thesis is devoted to investigating the functions of the basal ganglia and related cortical areas in sequential decision making in relation to reinforcement learning. Learning by reinforcement signals is very common and fundamental for sequential decision making in biological systems because the information of a correct output given an input is not always available. There is a framework of reinforcement learning in machine learning to learn optimal sequential decisions that maximize rewards from the environment. The framework of reinforcement learning fits well with the characteristic of the LTM in that it provides a scheme to change synaptic weights that are the carrier of stored information in the LTM. The basal ganglia with related cortical areas has been long known to be involved in sequential motor control and has recently been recognized as being involved in sequential decision making. This study is also a natural extension of the recent hypothesis of Houk et al. [31] that reinforcement learning is a major function of the basal ganglia. In this study, an emphasis is made on the relationships of several functional loops of the basal ganglia for sequential decision making. A model with a concrete algorithm is provided, based on the scheme. The performance of the model is closely compared with recent experimental findings of Hikosaka and his colleagues [25,27,29,44,45].

1.2 Methodology

The methodology used throughout this thesis relies heavily on computer simulation. Artificial neural network techniques are employed for the investigation of the issues raised in this thesis. Specifically recurrent networks have been employed in Chapter 2 and reinforcement learning has been employed in Chapter 7. For the general review of artificial neural network techniques, see [23, 24].

1.3 Overview by Chapters

This thesis is composed as follows.

In Chapter 2, the role of working memory for sequential decision making, or goal-directed behaviors, is investigated. A specific requirement of working memory is proposed for sequential decision making from a view of dynamical systems. To embody the requirement, a mathematical analysis of a sigmoidal unit in a recurrent network is provided. It is shown through simulations of foraging tasks by use of evolutionary programming that the proposed requirement can emerge in recurrent networks that control the movement of a creature as a functional necessity for survival in non-stationary environments.

Chapter 3, 4, 5, 6 and 7 are devoted to investigate the functions of the loops of the basal ganglia in sequential decision making. In Chapter 3, we briefly summarize the inputs, outputs, and internuclear structure of the basal ganglia and several basal ganglia-thalamocortical loops. In Chapter 4, the framework of reinforcement learning is reviewed, followed by the summary of computational models of the basal ganglia by other researchers based on this framework. In Chapter 5, first, the experimental paradigm developed by Hikosaka and his colleagues, and their behavioral findings [25,27,29,44,45] are summarized. Second, neurophysiological findings on the striatum, the presupplementary motor area, and the supplementary motor area are discussed, including neurophysiological findings of Hikosaka laboratory. These behavioral and neurophysiological findings form the basis for the hypothesis in the next chapter. In Chapter 6, computational elements of the functions of the basal ganglia loops in sequential decision makings are, first, hypothesized. Based on these computational elements, a general framework of the acquisition and retrieval processes in execution is, then, proposed. Third, a model implementing an algorithm based on the general framework is explained. In Chapter 7, the performance of the model is closely examined with the experimental data of Hikosaka and his colleagues. It is shown that the performance of simulation of the model captures well the characteristics of their experimental data in several aspects and predicts the behavior of the monkeys in conditions that have not been experimentally tested yet.

Finally, the findings of above studies are summarized, and their limitations and future work are discussed in Chapter 8.

Chapter 2

NONLINEAR DYNAMICAL PHENOMENA IN WORKING MEMORY

2.1 Introduction

In this chapter, the dynamical characteristics of working memory for sequential decision making are discussed.

Before going into the details, it may be worth mentioning the general definition of working memory in relation to the short-term memory (STM) along with the definition of working memory in this study. In the psychological literature, the STM is considered as registering and retaining incoming information in a highly accessible form for a short period of time after the input [77]. The STM is traditionally distinguished from the sensory information storage in that while the the sensory information store is considered as maintaining a rather accurate and complete picture of sensory inputs for a very short time period, which is shorter than that of the STM, the STM retains the immediate interpretation of events notified by those sensory inputs [36]. Working memory is proposed as the extended concept of the STM and Baddeley [5] defines working memory as "a brain system that provides temporary storage and manipulation of information necessary for such complex cognitive tasks as language comprehension, learning, and reasoning." An important feature of working memory, or the STM, is that it actively stores information. In this study, the term, working memory, refers to a mechanism that stores information by neural activities, and that selects which information should be stored based on the currently stored information and the current sensory inputs.

As stated in Chapter 1, neural activity is required for loading, storing (maintaining) and retrieving in working memory. There are two broad classes of models proposed for working memory from a viewpoint of (artificial) neural network community [82]. In one, rapid temporally-coordinated change of synaptic strength maintains neural activities in working memory [57]. In the other, recirculating neural activities by recurrent connections serve for the maintenance of neural activities in working memory [81, 82]. In the latter scheme, which is investigated in this study, it is discussed that fixed point attractors play a role in short-term memory from a view of dynamical system [81, 82]. This view roots in that when there are fixed point attractors, neural activity pattern that is close to one of the fixed point attractors at an initial time converges to and stays with the pattern defined by the fixed point attractor in time. By this phenomena, neural activities can be temporally sustained nicely as seen in experimental results of working memory [18] as well as modeling works [82]. In addition, fixed attractors can provide the robustness of neural activities in working memory against noise [41,82]. The characteristic of the robustness is important for working memory. Both of inputs and stored entity in working memory are neural activities so that stored information would be easily interfered by inputs without such robustness.

When we consider the role of working memory in sequential decision making, however, it is not enough only to require such characteristics. In sequential decision making, a sensory information such as signaling the existence of a prey at a time evokes a goal. Presumably, the information will be temporarily stored in working memory and guide the animals to achieve the goal. However, it is not always true that the stored information leads the animals to the goal, particularly in a dynamically changing world, because the environment may change while the animals pursue the goal. In such a case, the information, stored in the working memory to set a goal, should be discarded after a while. In other cases, another information, which signals a possibly better goal, may come to the animals while the pre-set sequence is performed to achieve the pre-set goal. When, then, should the pre-set goal be still pursued or discarded to set another goal, given such another information? While it is important to have the robustness of stored neural activities in working memory, it is also important for another neural activities to be easily loaded without being compounded with the pre-stored information, if the former is more beneficial to be stored than the latter. In other words, neural activities in working memory should have the dynamics of *long-term maintenance and quick transition*. It is essential to consider these requirements for the dynamics of working memory in sequential decision making.

In the following sections, first, the dynamics of a network of sigmoidal units with self-connections will be analyzed. It is shown that both long-term maintenance and quick transition can be achieved when the system parameters are near a "saddle-node bifurcation" point. Then, it will be tested if such a dynamical mechanism can actually be helpful for a goal-seeking behavior of an autonomous agent in simulations of a foraging task similar to the one used in Nolfi et al. [50] . After optimizing neural networks that control the movement of the agents by evolutionary programming, near saddle-node bifurcation behavior is robustly found under conditions that demand efficient use of working memory. The result indicates that near saddle-node bifurcation behavior can emerge in the course of evolution as a necessity for survival in non-stationary environments. Preliminary results for this study is reported in Nakahara and Doya [49].

2.2 Near Saddle-Node Bifurcation Behavior

When a pulse-like input is given to a linear dynamical system, the rising and falling phases of the response have the same time constant. This means that long-term maintenance and quick transition cannot be simultaneously achieved by linear dynamics. Therefore, it is essential to consider a nonlinear dynamical mechanism to meet these two demands.

2.2.1 Dynamics of a self-recurrent unit

First, we consider the dynamics of a single sigmoidal unit with the self-connection weight a and the bias b in a recurrent network as in the discrete state transition framework.

$$\begin{aligned} y(t+1) &= F(ay(t)+b), && (2.1) \\ F(x) &= \frac{1}{1+\exp(-x)}, && (2.2) \end{aligned}$$

where t denotes the time and y denotes the output of a sigmoidal unit, which is considered as neural activity.

The parameters (a, b) determine the qualitative behavior of the system such as the number of fixed points and their stabilities. As we change the parameters, the qualitative behavior of the system may suddenly change. This is referred to as "bifurcation" [20].

One typical example is a "saddle-node bifurcation" in which a pair of fixed points, one stable and one unstable, emerges.

For example, as the bias b is increased in Equation (2.1), the number of fixed points changes from one (Figure 2.1A), two (B), three (C), two (D), and then back to one (not shown). A saddle node bifurcation occurs when the state transition curve $y(t+1) = F(ay(t)+b)$ is tangent to $y(t+1) = y(t)$, as in the case of Figure 2.1 B and D.

Let y^* be this point of tangency. We have the following condition for saddle-node bifurcation.

$$\begin{aligned} F(ay^* + b) &= y^* && (2.3) \\ \left.\frac{dF(ay+b)}{dy}\right|_{y=y^*} &= 1 && (2.4) \end{aligned}$$

These equations can be solved, by noting $F'(x) = F(x)(1 - F(x))$, as

$$\begin{aligned} a &= \frac{1}{y^*(1-y^*)} && (2.5) \\ b &= F^{-1}(y^*) - ay^* = F^{-1}(y^*) - \frac{1}{1-y^*} && (2.6) \end{aligned}$$

By changing the fixed point value y^* between 0 and 1, we can plot a curve in the parameter space (a, b) on which saddle-node bifurcation occurs, as shown in Figure 2.2. The system has only one stable fixed point when the parameters are outside the cusp (A) and three fixed points inside the cusp (C). A pair of stable and unstable fixed points emerges or disappears when the parameters pass across the cusp-like curve (B and D).

An interesting behavior can be found when the parameters are just outside the cusp, as shown in Figure 2.3 (center). The system has only one fixed point near $y = 0$, but once the unit is activated ($y \simeq 1$), because the trajectory "bounces" in the narrow channel between $y(t+1) = y(t)$ and the sigmoid activation curve, the unit stays "on" for many time steps and then goes back to the fixed point quickly. Such a mechanism

Figure 2.1

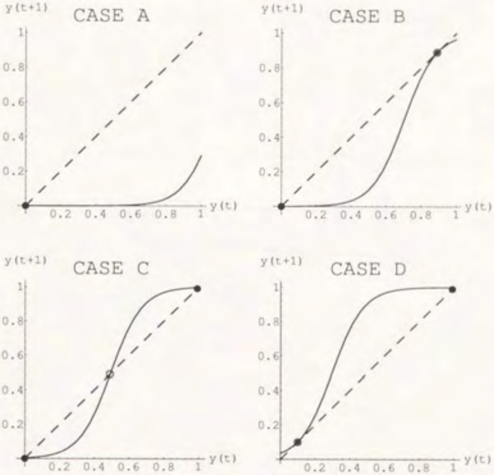


Figure 2.1: State transition diagrams of the self-recurrent unit for four different cases. A: one fixed point near $y = 0$. B: a saddle-node bifurcation at $y = 0.9$. C: three fixed points. D: another saddle-node bifurcation at $y = 0.1$. In each graph, a solid circle stands for a stable fixed point, an empty circle for an unstable fixed point, and an empty circle with a solid one inside for a saddle fixed point.

Figure 2.2

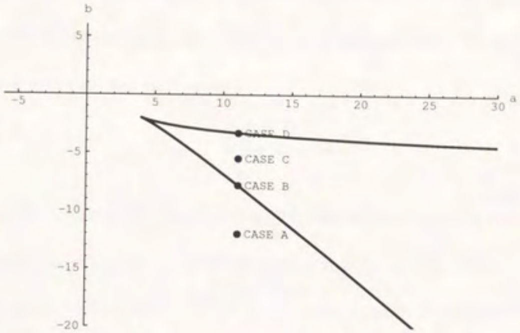


Figure 2.2: The bifurcation set in the parameter space of the self-recurrent unit. Saddle-node bifurcation is seen on the cusp-shaped curve. There are three fixed points inside and one fixed point outside the cusp.

may be useful in satisfying the requirements of the dynamics in working memory for sequential decision making: long-term maintenance and quick transition.

2.2.2 Network of self-recurrent units

Next, the dynamics of a network of the above self-recurrent units is examined.

$$y_i(t+1) = F[ay_i(t) + b + \sum_{j,j\neq i} c_{ij}y_j(t) + d_i x_i(t)], \tag{2.7}$$

where a is the self connection weight, b is the bias, c_{ij} is the lateral connection weight, d_i is the input connection weight, and $x_i(t)$ is the external input. The effect of the sum of the lateral and external inputs

$$u_i = \sum_{j,j\neq i} c_{ij}y_j + d_i x_j$$

is equivalent to the change in the bias, which slides the sigmoid curve in the state transition diagram horizontally without changing the slope. Therefore, we can analyze the behaviors of a multiple of units based on the single-unit behavior discussed above.

For example, let us consider a case in which a saddle-node bifurcation occurs at $y_1 = 0.9$. From equation (2.6), the parameters for this case is $a = 11.11$ and $b = b_1 \simeq -7.80$. As we increase b while keeping a constant, the system first has three fixed points as in Figure 2.1 C and then the lower two fixed points merge together at $y = 1 - y_1 = 0.1$ with the bias $b_2 \simeq -3.31$, which forms another saddle-node bifurcation seen as in Figure 2.1 D.

Let the bias $b_0 = -7.90$ so that the unit is near saddle-node bifurcation when there is no lateral or external inputs. If the input sum exceeds the threshold(θ), i.e. $u_i > \theta = b_2 - b_0 \simeq 4.59$, the lower fixed point at $y = 0.1$ disappears and the state jumps up to the upper fixed point near $y = 1$, quickly turning the unit "on" (Figure 2.3 left). As we

Figure 2.3

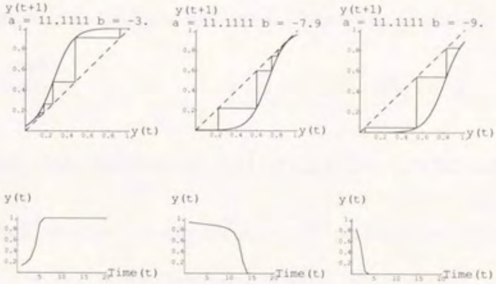


Figure 2.3: Temporal responses of self-recurrent units. Center: near saddle-node bifurcation with $a = 11.1111, b = -7.9$. Left: increased bias $b = -3.0$. Right: decreased bias $b = -9.0$.

saw above, when the input is removed, the state stays near $y = 0.9$ for many time steps (Figure 2.3 center).

If there are inhibitory lateral connections, the activation of the unit i raises the threshold for other units $k \neq i$ as $\theta' = \theta - c_{ki}y_i$, making it more difficult for other units to turn "on". On the other hand, the time course of the activated unit i is affected very little with the sub-threshold input to other units k because their activity is kept low ($y_k < 0.1$). When there is a strong input to unit k that exceeds the threshold θ', however, the unit is turned "on" and sends an inhibitory input to unit i, which is equivalent to a decrease in the bias. As a result, the activation of the unit i quickly goes down (Figure 2.3 right).

2.3 Evolution to near bifurcation dynamics

In the above section, we have theoretically shown the potential usefulness of near saddle-node bifurcation behavior for satisfying the demands of the dynamics in working memory for sequential decision making. We further hypothesize that such behavior is indeed useful in animal behaviors and can be found in the course of learning and evolution of the neural system.

To test our hypothesis, we simulated a foraging task in which a creature seeks food in a grid-like world (Figure 2.4), similar to Nolfi et al.'s [50]. Our purpose in this simulation is to see whether near bifurcation dynamics discussed in the previous section can actually improve creatures' performance in a non-stationary environment where selection and memory of sensory input is necessary. Evolutionary programming [16] was used to optimize the recurrent network that controls the movement of the creature.

Figure 2.4 shows an example of the grid-like world. There were a certain number of

Figure 2.4

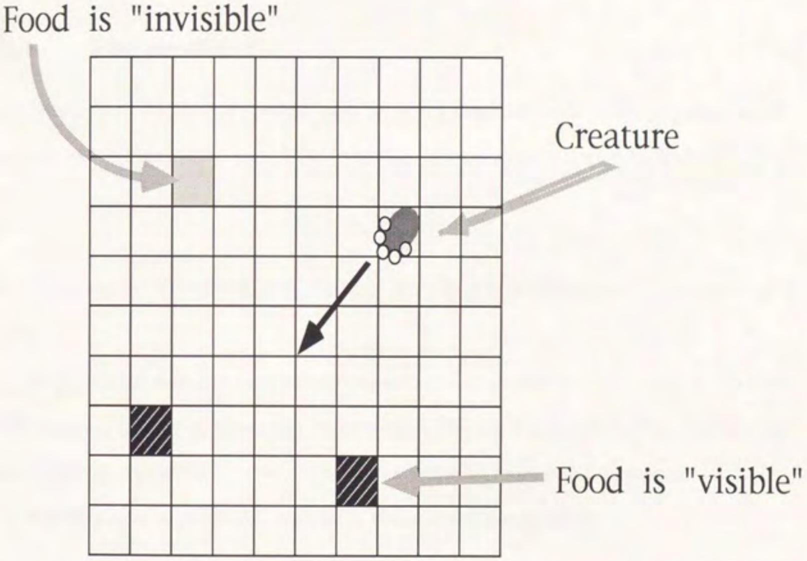


Figure 2.4: The foraging task in a grid-like world. Note that the shown example is not a 20 × 20 but a 10 × 10 grid-world.

food items in fixed positions which turned visible or invisible in a stochastic fashion, as determined by a two-state Markov system. A creature got the food when it reached the food, whether it was visible or invisible. When a food item was eaten, it was removed from that position and a new food item was placed randomly. The size of the world was 20x20 and both ends were connected as a torus. The amount of food a creature found in a certain time period was the measure of its performance.

2.3.1 The creature

A creature had five visual sensors, each of which detected food within a particular 45-degree sector(Figure 2.5, top). The activation of each sensory unit was given by

$$x_i = \sum_j \frac{1}{r_j}$$

where r_j was the distance to the j-th food item that was visible within the sector at a time.

At each time step, the creature executed one of three motor commands: L: turn left (45 degrees), C: step forward, and R: turn right (Figure 2.5, middle). The action of the creature was controlled by a two layer neural network (Figure 2.5, bottom).

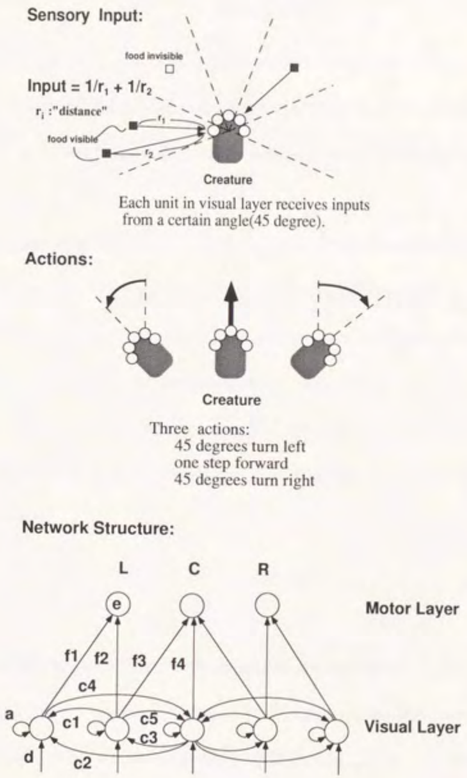
The dynamics of each of five units in visual layer was given by

$$y_i(t+1) = F(ay_i(t) + b + \sum_{j,j \neq i} c_{ij} y_j(t) + dx_i(t)) \tag{2.8}$$

where $y_i(t)$ was the output of the visual unit at time t, a was the self connection weight, b was the bias, c_{ij} was the cross connection weight, d was the input connection weight, and $x_i(t)$ was the external sensory input. Note that the self-connection a, the bias b and the input weight d were the same for all units.

Each of three units in motor layer coded the probability of taking one of the three

Figure 2.5



Each unit in visual layer receives inputs from a certain angle(45 degree).

Three actions:
45 degrees turn left
one step forward
45 degrees turn right

Figure 2.5: A creature's sensory input(Top) motor system(Middle) and network architecture(Bottom).

motor commands (L, C, R). Their output z_k was given by

$$v_k(t) = e_k + \sum_i f_{ki} y_i(t), \tag{2.9}$$

$$z_k(t) = \frac{\exp(v_k(t))}{\sum_l \exp(v_l(t))}, \tag{2.10}$$

where e_k was the bias and f_{ki} was the feedforward connection weight. Note that adding a uniform bias e_k to all the units did not affect the output because of the normalization in equation (2.10). In order to avoid the redundancy, we fixed the bias of the center unit as $e_2 = 0$.

In general, the time step of the internal operation of the network can be different from that of the external world. We chose two steps of internal time t, which corresponded to one step of external time T, i.e. $T = 2t$. This allowed an indirect effect of sensory input through the lateral connection to be utilized in taking the next action. In addition, the activation pattern in visual layer was shifted when the creature made a turn, which should give the proper mapping between the working memory and sensory input at the next external time step.

2.3.2 The world

The characteristics of the world were determined by two sets of parameters: the food density and the parameters of the Markov transition matrix. We fixed the food density 0.03, i.e. there were 12 food items randomly distributed in the 20x20 grid world.

At each world time step, each food item took one of two states "on" and "off" (visible and invisible) as given by a Markov system

$$\begin{pmatrix} P_{\text{off}}(T+1) \\ P_{\text{on}}(T+1) \end{pmatrix} = \begin{pmatrix} p_0 & (1-p_1) \\ (1-p_0) & p_1 \end{pmatrix} \begin{pmatrix} P_{\text{off}}(T) \\ P_{\text{on}}(T) \end{pmatrix} \tag{2.11}$$

where $P_{\text{on}}(T)$ and $P_{\text{off}}(T)$ were the probabilities that the food item was on and off at time T, respectively. Note that the stationary distribution $(\bar{P}_{\text{off}}, \bar{P}_{\text{on}})$ is given by

$$\bar{P}_{\text{off}} = \frac{1-p_1}{(1-p_0)+(1-p_1)} \qquad \text{and} \qquad \bar{P}_{\text{on}} = \frac{1-p_0}{(1-p_0)+(1-p_1)}.$$

Refer to Figure 2.6 to see points in the parameter space (p_0, p_1) of the Markov transition matrix that were used in the simulation.

2.3.3 Evolutionary programming

For the sake of simplicity, we put symmetric constraints on the connection weights as follows.

$$\{c_{ij}\} = \begin{pmatrix} 0 & c_1 & c_2 & 0 & 0 \\ 0 & 0 & c_3 & 0 & 0 \\ c_4 & c_5 & 0 & c_5 & c_4 \\ 0 & 0 & c_3 & 0 & 0 \\ 0 & 0 & c_2 & c_1 & 0 \end{pmatrix} \qquad \{f_{ij}\} = \begin{pmatrix} f_1 & f_2 & 0 & 0 & 0 \\ 0 & f_3 & f_4 & f_3 & 0 \\ 0 & 0 & 0 & f_2 & f_1 \end{pmatrix}$$

The bias for the motor units was also symmetric $(e_L, e_C, e_R) = (e, 0, e)$. Therefore, each creature's network was characterized by the thirteen parameters $(a, b, c_1, ..., c_5, d, e, f_1, ..., f_4)$.

A population of 60 creatures was tested on each generation. The performance was measured by the number of pieces of the food a creature obtained in $T = 400$ time steps. Each of the top twenty scoring creatures produced three offsprings; one identical copy of the parameters of the parent's and two copies of these parameters with a Gaussian fluctuation $\sim N(0, 1.5^2)$. These three offsprings of each of the top twenty scoring creatures$(3 \times 20 = 60)$, thus, become the next generation, their performance was then measured, and it continued. In preliminary experiments, we generated the initial population with random parameters whose range was $[-10.0, 10.0]$ except for the input connection weight d, whose range was $[1.0, 6.0]$. Under most conditions, the population converged to a limited range of the parameter space after evolution. Therefore, in the simulations below, a smaller range of initial parameters was used in order to speed up convergence(See Appendix A). In this paper, the result after 100 generations is reported.

2.4 Results

2.4.1 Creatures' performance

The performance of the creature after evolution under different environmental parameters (p_0, p_1) is shown in Figure 2.6. The radius of the outer circle represents the number of food items taken by the creature. Generally speaking, the performance was lower as p_0 (probability of food staying off) was increased and as p_1 (probability of food staying on) was decreased. Note that the performance was different even among the sets of the environmental parameters (p_0, p_1), which belong to the same stationary distribution.

To examine if the recurrent connections in the creatures really contributed to the improvement in the performance, we tested the performance of feedforward network organisms, which was given by omitting the recurrent connections of the creatures, in other words, by keeping $a = c_1, ..., c_5 = 0$. The performance of this given feedforward networks were also optimized by evolutionary programming. Note that we did not try to obtain the optimized feedforward network among any possible feedforward networks, for example, feedforward network with hidden layers, but rather tested the performance of the given feedforward networks in order to see if recurrent connections had a role in improving the creature's performance or not. The radius of the gray disc in Figure 2.6 represents the food taken with the feedforward network after evolution. Performance in the feedforward case was always lower than that in the recurrent case for each set

Figure 2.6

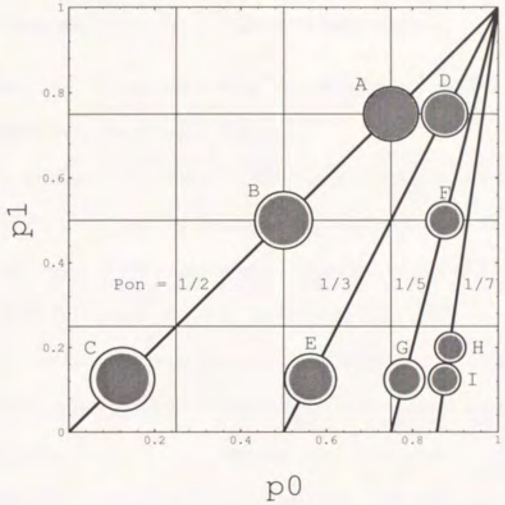


Figure 2.6: The performance of the creatures after 100 generations. Points in the parameter space (p_0, p_1), which are used for simulations, are $A = (0.75, 0.75)$, $B = (0.5, 0.5)$, $C = (0.125, 0.125)$, $D = (0.875, 0.75)$, $E = (0.562, 0.125)$, $F = (0.875, 0.5)$, $G = (0.781, 0.125)$, $H = (0.888, 0.2)$, and $I = (0.875, 0.125)$. At each point, the outer circle represents the performance in the recurrent case and the inner disc represents in the feedforward case. Oblique lines represent the parameters for the same stationary distributions: $\bar{P}_{\text{on}}(t) = 1/2$, $\bar{P}_{\text{on}}(t) = 1/3$, $\bar{P}_{\text{on}}(t) = 1/5$, and $\bar{P}_{\text{on}}(t) = 1/8$.

of environmental parameters. The difference was more marked as p_0 (probability of food staying invisible) was increased and as p_1 (probability of food staying visible) was decreased.

2.4.2 Convergence to near-bifurcation region

The self-connection and bias parameters of top ten scoring creatures under different environmental parameters are shown in Figure 2.7.

When either p_0 was small or p_1 was large, the value of the self-connection a was almost zero, as shown in Figure 2.7 for the Markov parameter setting of $D = (0.875, 0.75)$. Convergence to the region of the parameters (a, b) similar to that of D was seen also in the case of $A = (0.75, 0.75)$, $B = (0.5, 0.5)$, and $C = (0.125, 0.125)$.

As p_0 was increased and p_1 was decreased, , in other words, as the environment became more severe, meaning that the probability of food items keeping visible gets small and the probability of food items keeping invisible gets large, the convergence to a region in the vicinity of the saddle-node bifurcation boundary, which is called near saddle-node bifurcation region in this study, became more prominent. Examples of such network parameters (a, b) are shown in Figure 2.7 F, H, and I. Note that p_0 were almost constant in Figure 2.7 D, F, H, and I. It is clearly seen that the parameters found after evolution lie just underneath the saddle-node bifurcation curve. This was most prominent in the case of I, where p_1 was the smallest.

2.4.3 Dynamics of activation

We analyzed the dynamics of the network that converged to the near saddle-node bifurcation region in case of I in Figure 2.7. Figure 2.8 shows the network dynamics of the top-scoring creature. It is clearly seen that the unit in visual layer, especially the

Figure 2.7

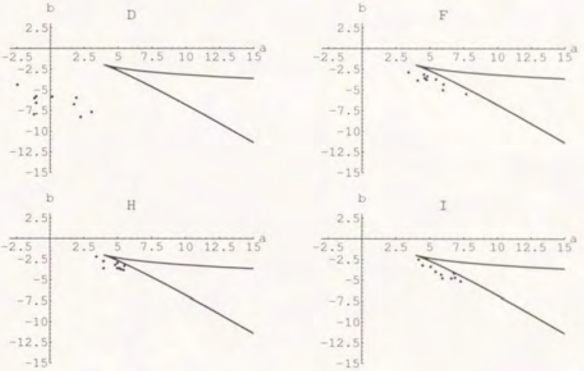


Figure 2.7: The convergence of the network parameter (a, b) with different environmental parameters (p_0, p_1) of the Markov transition matrix. Top ten scoring creature's network parameters are plotted in the bifurcation diagram. (Left:top) $D = (0.875, 0.75)$. (Right:top) $F = (0.875, 0.5)$. (Left:bottom) $H = (0.888, 0.2)$. (Right:bottom) $I = (0.875, 0.125)$.

Figure 2.8

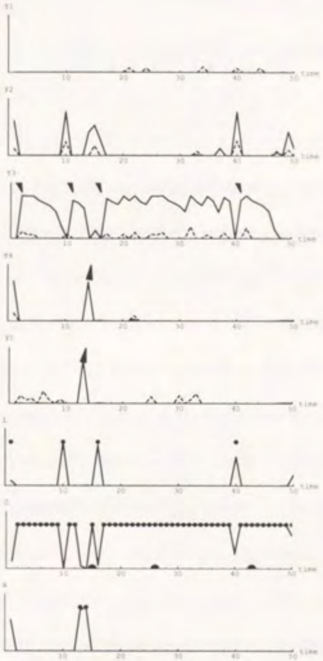


Figure 2.8: Examples of the creature's activation dynamics in the simulated environment, $I = (0.875, 0.125)$. y_1 through y_5 indicate outputs in visual layer. L, C, and R indicate outputs in motor layer, corresponding to each of motor commands, "turn left", "step forward" and "turn right" respectively. Dotted lines of units in visual layer represent the external input. Arrowheads at the top of each visual unit indicate that the activation of units in visual layer is shifted according to the creature's turn. Dots at the top of each motor unit show that the corresponding motor command is chosen at that time. Large dots on the horizontal axis of C show that the food is obtained at that time.

"center" unit $y3$, functioned as a source of working memory. When enough input came into a non-center unit, e.g. $y2$ at $T = 40$, it quickly turned on and $y3$ was immediately suppressed. The activation of $y2$ in visual layer was propagated into units in motor layer so that the creature made a left turn. According to our assumption, then, the activation of $y2$ was shifted to $y3$ and it remained active for several time steps. It should be emphasized that the near bifurcation behavior of the unit $y3$ realized the long-term maintenance and quick transition with the help of interaction with other units.

It was sometimes observed that the dynamics of the visual layer units did not realize near saddle-node bifurcation behavior. This is partly because the function of working memory could be realized not only in visual layer but also in motor layer. With a large negative bias ϵ in motor layer, the choice of motor command could be strongly biased as "center", that is, "step forward". In this case, once a creature would detect the food far away and make a turn towards the food, a creature would not have to remember the direction of food. In other words, without the long-term maintenance of memory in visual layer, a creature could use its body direction as the working memory using the fact that the default choice of motor command was to go straight ahead.

2.5 Discussion

Fixed point attractors may play a role in working memory and can give the robustness to working memory against noise. It is not enough, however, only to require working memory to maintain information, or to have such robustness, when working memory is considered in relation to sequential decision making, or goal-directed behaviors. It is important to consider the dynamics of neural activities in working memory for sequential decision making and, in this study, the characteristics of long-term maintenance and quick transition is proposed as crucial. Such dynamics can allow biological systems to stay focused on a goal, even if it isn't currently visible, and also to be able to switch quickly to another goal if the new one is likely to be more valuable or if the pre-set goal turns out to be not a good one after a while. It is very difficult to obtain such dynamics in linear dynamics in general.

We mathematically analyzed the dynamical characteristics of a self-recurrent sigmoidal unit with a bias. It was shown that both long-term maintenance and quick transition can be realized near a saddle-node bifurcation. The behavior of a network of such recurrent units can be analyzed by considering inputs as the change in the bias. Near saddle-node bifurcation behavior can be considered as a candidate of the dynamics in working memory for sequential decision making.

By the simulation of food-foraging tasks, we tested our hypothesis, which posits that the dynamics of long-term maintenance and quick transition in working memory is important and valuable in sequential decision making. Neural networks that controlled the sequential decisions of creatures in the tasks were optimized through evolutionary programming. It was shown that the performance of the creatures was lower as the probability (p_0) of food staying invisible was increased and as the probability (p_1) of food staying visible was decreased, regardless of whether the environmental parameters (p_0, p_1) belonged to the same stationary distribution. In comparison of the performance of the feedforward networks that was given by omitting recurrent connections, the performance of the recurrent network was always better than the feedforward ones. Therefore, it can be concluded that the recurrent connections contributed to the improvement in the performance of creatures. In other words, a memory provided by recurrent connections played a role in the creature's performance, although it is yet unclear whether the recurrent nature of connections or just additional number of connections have critically contributed to the performance.

The convergence of the self-connection and bias of neural network organisms to the near saddle-node bifurcation region did not occur under all conditions of tested environmental parameters. However, the convergence became more prominent as the environment became more severe, that is, as the probability (p_0) of the food staying invisible was increased and as the probability (p_1) of the food staying visible was decreased. It should be noted that the convergence of neural network organism to the near saddle-node bifurcation region was not directly designed in the simulation, but emerged indirectly in the optimization process of evolutionary programming in interaction of the given non-stationary environment. The difference of the performance between the recurrent networks and the aforementioned feedforward networks became more prominent as the probability of the food staying invisible was increased and as the probability of the food staying visible was decreased. By examining the actual dynamics of neural network in the task, whose parameters of the self-connection and bias converged to the near saddle-node bifurcation region, it was shown that there truly existed the long-term maintenance and quick transition in the network dynamics. These results may imply that near saddle-node bifurcation behavior helped a creature's survival particularly as the environment became severe, and further that this dynamics in working memory may be an emergent functional property in evolving neural systems that enables them to deal with a dynamically varying world such as the non-stationary environment as given in the simulation of this study.

In this study, the long-term maintenance and quick transition is regarded as crucial requirement of the dynamics in working memory for sequential decision making, or goal-directed behaviors. This requirement is introduced by considering two opposing demands such that information should be stored against noise as well as the stored information should be quickly discarded in some cases so that other better information can be loaded quickly without being confused with the pre-stored information. In this view, working memory is concerned not only with maintaining information but also with selecting which information should be loaded. If there are other ways to deal with these two opposing demands, the long-term maintenance and quick transition may not be necessarily required in working memory. The case stated in the section 2.4.3 is such an example, in which the dynamics of the visual layer units did not realize near saddle-node bifurcation behavior. In this case, the motor layer with the strong bias of "step forward" let a creature use its body direction as maintaining information, that is, the direction to a food item.

Limitations and future works of this study will be discussed in Chapter 8.

Chapter 3

BASAL GANGLIA IN RELATION TO SEQUENTIAL DECISION MAKING

The role of working memory in sequential decision making is investigated in the previous chapter, specifically focusing on the dynamics of neural activity in working memory. In the long-term memory (LTM), information is stored by physiological parameters such as synaptic strength. The maintenance of the information in the LTM, therefore, is more stable than in working memory and information in the LTM plays an important role in sequential decision making. In the rest of the thesis, the functions of the basal ganglia and related cortical areas for sequential decision making are investigated. The aim of this chapter is to provide accounts for this study as well as the brief review of the basal ganglia and its loops with the cerebral cortex. The overview of this chapter and following chapters is given in the end of Section 3.1.

There are a variety of conceived classifications of the LTM systems. One classification includes semantic memory, declarative memory, and *procedural memory* as part of the

LTM memory from a psychological point of view [77]. Among these, for example, the procedural memory system is referred to an *action* system: "its operations are expressed in behavior, independently of any cognition. Skillful performance of perceptual-motor tasks and conditioning of simple stimulus-response connections are examples of tasks that depend heavily on the procedural memory" by Tulving (1991,p12.) [77]. Very similarly but not identically, *skills and habits* in one classification are termed in Squire et al. (1993, p471.) [68] as:

> Skills are procedures (motor, perceptual, and cognitive) for operating in the world; habits are dispositions and tendencies that are specific to a set of stimuli and that guide behavior. Under some circumstances, skills and habits can be acquired in the absence of awareness of what has been learned and independently of long-term declarative memory for the specific episodes in which learning occurred. However, many skill-like tasks are also amenable to declarative learning strategies.

Thus, while the classification of long-term memory systems provides us with the ground for further investigation, the distinction of classifiers, as well as the correspondence of behaviors with these submemory systems, are still under the investigation and dispute. In addition, little is yet known about the correspondence of these submemory systems with underlying neural mechanisms in particular from a computational viewpoint. Furthermore, the term of sequential decision making in general is a broad term and includes a variety of our daily activities such as playing chess, playing the piano, and writing a Ph.D. thesis.

Therefore, the strategy taken in this thesis is to investigate the functions of a specific subsystem of underlying neural mechanisms for one of the LTM systems in a specific type of sequential decision making. In this study, the functions of the basal ganglia and related cortical areas are examined for skills, one of sequential decision makings, in relation to reinforcement learning. Roughly speaking, skills are composed of sequential movements to solve frequently experienced tasks and are acquired in a long run. Note that, in the learning of such skills, learning by reinforcement signal is very common because information for a desired output given an input is not always available. The basal ganglia can be considered as involved in reinforcement learning of such skills [31].

3.1 Introduction

This section aims to state the purpose of this study as well as the ground for issues investigated in terms of the LTM in the following chapters. For this purpose, the behavioral, neuroscientific, and computational accounts are provided below in their order.

Behavioral Accounts

Skills are composed of learned sequential movements. For example, to swing forehand in playing tennis, a player should move his body to an appropriate place in help of his perceptual system that follows a trajectory of a coming ball. While moving his body to a place, he should prepare to hit the ball and start to swing a racket at an appropriate timing, in help of his perception of the ball. To play well, each of joint movements should be coordinated well with each other and with a trajectory of a coming ball. If he were a novice, the coordination of his perceptual, cognitive, motor system may not work well to play yet. After hitting the ball, the player can see how well or bad his current play was. Using this information, the player tries next play. By repeating these processes many times in a day and/or across days, the player develops his skills and becomes an expert player.

Elements of this kind of processes can be schematically counted as follows:

1. *Multiple time scales in the development of skills*: Skills are eventually developed by its occasional but frequent experiences in a long run. Yet, its improvement occurs at a multiple time scales: relatively short (in a day in the above example) and relatively long (across days).

2. *Hierarchical nature and context-dependency*: Some parts of sequential movements can be transferred to another sequential movement. For example, parts of the skill in throwing a baseball can be of help for learning the serve at tennis. At the same time, there could be interference between such parts as well. Thus, there is a hierarchical nature in acquisition and execution of skills. This also means that the appropriateness of the retrieval of a part from the long-term memory depends on the context, or what a skill is being performed and what a goal is being aimed by this skill.

3. *Continuity between acquisition and retrieval in execution*: Skills are improved over a long time. The acquisition process gradually occurs in the execution process, for example, during playing tennis. At the same time, in the execution process, the learned information in the LTM should be retrieved to use. In addition, what aspects of skills to be focused should be different, depending upon what has been already acquired, that is, the information in LTM. Thus, the acquisition and retrieval cannot be isolated in the execution process.

4. *Learning type*: Learning of skills is often a type of *learning by reinforcement signals*: a learner gets a signal, which is a scalar, to tell how good or bad s/he did but not the signal to tell what was actually supposed to do. In the framework of artificial neural network, the type of learning by reinforcement signal is called *re-*

inforcement learning. If a learner would receive the signal to tell what was actually supposed to do, which is usually a vector representation in the framework, it is called *supervised learning.* Reinforcement learning (RL) is particularly important in terms of sequential decision making, because it often happens that a learner can't have a type of signals of the supervised learning in particular in early stage of learning, and further because there are the temporal credit assignment problem so that it is often the case in which reinforcement signals are at most available at each time when a learner makes a decision. As the learner experiences the world, the learner accumulates knowledge about what is supposed to do, and then, it becomes possible in theory to have a type of supervised learning. In addition, it would be desirable to store strategies of certain sequential decision making that are often required in the LTM and then it allows the learner to challenge more complex problems.

Hikosaka and his colleagues developed a new experimental paradigm to examine sequential decision making [25]. Their experimental paradigm fits well with behavioral requirements stated above in the investigation of sequential decision making, or sequential movements. In addition, their experiments include blocking specific portions of the brain so that they present interesting clues to investigate behaviors with underlying neural mechanism.

Neuroscientific Accounts

The basal ganglia has long been known as being involved in motor control [33]. The basal ganglia is also known to have a role in cognitive functions according to studies on the brain diseases such as Parkinson's and Huntington diseases [54]. It is now widely accepted that the basal ganglia is involved in a wide variety of behavioral functions, including cognitive, motor, and motivational functions [4]. A striking characteristic of the basal ganglia is that the basal ganglia receives the projections from almost the entire cortex and then projects back to a wide range of the frontal cortex via thalamus. This convergence leads many researchers to be intrigued by its functional roles and to regard the basal ganglia as functioning action selection process as well as its acquisition. However, it has not been much known about its actual computation yet from a computational viewpoint.

It is conceived that there are at least four basal ganglia-thalamocortical circuits [3]. Among them, the motor, oculomotor, and dorsolateral prefrontal circuits are of interest in this study. Even though a multiple of circuits are conceived in relation to the basal ganglia, the functional relationship among these circuits has not been investigated much from a computational viewpoint. One of the purposes of this study is to provide a computational account on the functional relationship among these circuits in terms of sequential decision making. Because of the characteristics of connections from and to the basal ganglia, the functions of the basal ganglia should not be considered in isolation.

Among the motor, oculomotor, and dorsolateral prefrontal circuits, the motor circuit is particularly interesting to us in terms of sequential decision making, that is, sequential movements. It should be noted that both of the basal ganglia and the cerebellum are the major constituents of two important subcortical loops of the motor system (Figure 3.1). While the loop of the cerebellum is closed only with motor cortical areas, the loops of the basal ganglia are connected not only with motor cortical areas but also with frontal cortical areas, which are considered as involved in higher order, or cognitive aspects, of motor movements. Thus, the cerebellum is more directly related to motor movements per se such as transforming Cartesian representations of external objects to kinematic

Figure 3.1

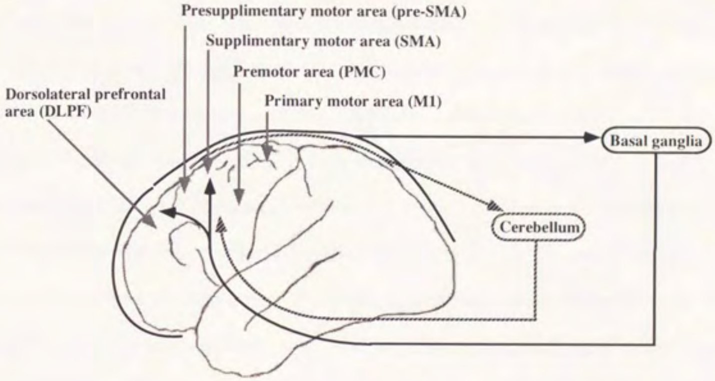

Presupplimentary motor area (pre-SMA)
Supplimentary motor area (SMA)
Premotor area (PMC)
Dorsolateral prefrontal area (DLPF)
Primary motor area (M1)
Basal ganglia
Cerebellum

Figure 3.1: Two major subcortical loops of the motor system: the loop of the basal ganglia and the loop of the cerebellum. Some cortical areas are indicated as well. Adapted from

representations, whereas the basal ganglia is more involved in cognitive aspects of motor movements [33]. In this study, the computations that are supposed to be more related to the cerebellum are omitted, which enables us to focus the functions in relation to the basal ganglia-thalamocortical loops.

Among the cortical areas included in the motor circuit, the supplementary motor area (SMA) and the presupplementary motor area (pre-SMA), which has been recently dissociated from the SMA [39,40], are of particular interest (Figure 3.1). It is because the SMA has been known to be important in programming motor sequences [33,74]. Since the dissociation of the pre-SMA from the SMA has been proposed recently, there remain much to be investigated in terms of the difference of their functions. Shima et al. [67], however, have shown very recently in their experiment that the pre-SMA is involved in error correcting activity. Hikosaka and his colleagues have also very recently shown in their experiment that the pre-SMA is involved in the acquisition process in the early stage of learning in their task (Miyashita et al. [46,47],Hikosaka et al. [28],Sakai et al [56]). Hence, from a computational viewpoint, it is very interesting and important to identify their functional roles in relation to those of the basal ganglia.

Furthermore, Schultz and his colleagues recently showed in their experiments [42, 43,58–63] that the response of dopamine neurons in the substantia nigra pars compacta in the basal ganglia shifts from primary reward to conditioned stimuli that predict reward as the conditioning establishes. This nature of responses of dopamine neurons is particularly interesting from a computational viewpoint. Because it may provide a clue to investigate the learning process of sequential movements in the basal ganglia, along with the framework of reinforcement learning in machine learning (and artificial neural networks).

Computational Accounts

As discussed in behavioral accounts, a type of learning by reinforcement signals is common and fundamental in sequential decision making of biological systems. A framework of reinforcement learning in machine learning can provide a basis for such learning [6, 71, 75].

Along with experimental findings of Schultz and his colleagues [42, 43, 58–63],Houk et al. [31] and others (e.g. Barto [7]) have recently hypothesized that dopamine neurons may encode temporal difference error in the framework of reinforcement learning. Although their hypothesis is a good starting point, the detailed correspondence of the circuitry of the basal ganglia and its computation of reinforcement learning still remains to be investigated in particular in terms of sequential decision making. These models lack the close comparison of the performance of their model with the actual behavioral data. Doya [14, 15] investigated the integration of functions between the basal ganglia and the cerebellum with an emphasis on the motor control. Berns et al. [10] proposed a competition scheme for the action selection in the basal ganglia in relation to reinforcement learning with an emphasis on the dorsolateral prefrontal circuit. Their model lacks a close comparison of the performance of their model with the behavioral data in terms of the sequential movements. In contrast with these computational models stated above, the computational model proposed later in this study rather focuses on the interaction of the basal ganglia-thalamocortical loops in terms of the behavioral aspects of the skills, discussed before, rather than on the functions of the specific loop or on the functions of motor control by itself.

The experimental paradigm of Hikosaka and his colleagues [25,27–29,37,38,44–47,55, 56] is rather complex among currently available experiments for the investigation of the acquisition, storage, and retrieval of sequential movements, along with the comparison of behavioral results with underlying neural mechanisms. Thus, the close examination of the model with their findings will provide unique opportunity in the investigation of underlying neural mechanisms in sequential movements from a computational viewpoint.

The overview of the rest of this chapter and following chapters is as follows. In the rest of this chapter, the basal ganglia, the basal ganglia-thalamocortical loops with related cortical areas, and the profile of neural activities of dopamine neurons are briefly reviewed. In the next chapter, Chapter 4, the framework of reinforcement learning is briefly summarized, followed by the discussion of computational models of the basal ganglia and related portions of the brain based on this framework. In Chapter 5, the experimental paradigm of Hikosaka and his colleagues is explained and the behavioral findings in their experiments are first examined. We, then, closely examine neurophysiological findings of the basal ganglia and related portions of the brain, including findings of Hikosaka and his colleagues. In Chapter 6, based on reviews and discussions in these above chapters, the elements of functions of the basal-ganglia thalamocortical loops for sequential decision making is first hypothesized. A general framework of the acquisition and retrieval processes in execution, then, is proposed. At last, a model implementing an algorithm based on the general framework is provided. In Chapter 7, the simulation with the model is employed. The performance of the model is investigated in comparison with experimental findings of Hikosaka and his colleagues [25,27–29,37,38,44–47,55,56]. This enables us to investigate the functions of the basal ganglia and related cortical areas in sequential decision making in close comparison with the behavioral data.

3.2 The basal ganglia consists of five nuclei

The basal ganglia consists of five large subcortical nuclei: caudate nucleus (CD), putamen (Pt), globus pallidus (GP), subthalamic nucleus (STN), and substantia nigra (SN) (See Figure 3.2). The globus pallidus is divided into two parts: the internal segment and the external segment. The substantia nigra is divided into two parts: substantia nigra pars reticulata and substantia nigra pars compacta. The internal segment of the globus pallidus (GPi) and the substantia nigra pars reticulata (SNr) can be considered as a single structure because of the striking similarities in cytology, connectivity, and function of the GPi and SNr [33]. The caudate nucleus and putamen are composed throughout of identical cell types and are fused anteriorly [33]. Together the CD and Pt are called *the striatum.*

3.3 Inputs to the basal ganglia

Almost all of the afferent connections to the basal ganglia terminate in *the striatum.* The striatum receives input from two major sources outside the basal ganglia: the cerebral cortex and the intralaminar nuclei of the thalamus [33]. The corticostriate projection consists of the most important input to the basal ganglia. It should be noted that the corticostriate projection contains fibers from the entire cerebral cortex:*motor, sensory, association, and limbic* areas. All these projections are topographically organized as well as projections from the intralaminar nuclei of the thalamus [69]. The STN also receives the projections from the prefrontal, premotor, and motor cortical areas [4].

Figure 3.2

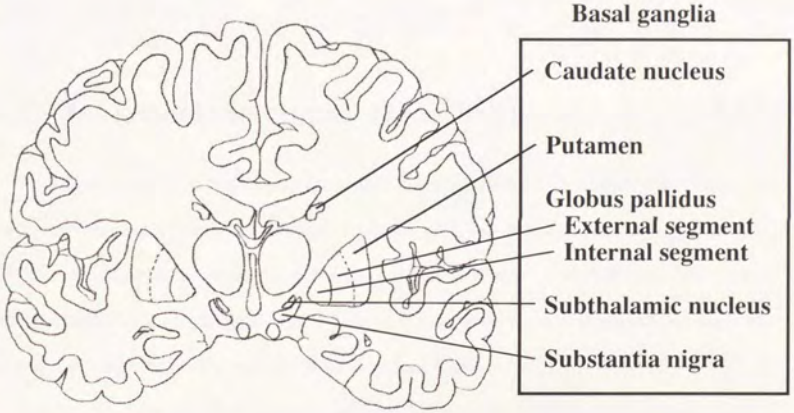


Figure 3.2: This coronal section shows the basal ganglia in relation to surrounding structures. Adapted from Kandel with modifications.

3.4 Outputs from the basal ganglia

The major outputs of the basal ganglia projects from the GPi and SNr to three nuclei in the thalamus: *the ventral lateral nuclei, the ventral anterior nuclei, and the mediodorsal nuclei* [33]. The SNr projects to the intermediate layer of the superior colliculus (SC). The GPi has an additional projection to the centromedian nucleus of the thalamus [33]. The portions of the thalamus receiving inputs from the basal ganglia project to the prefrontal (PF), premotor (PMC), supplementary (SMA), presupplementary (pre-SMA), and motor (M1) cortex [33].

3.5 Internuclear connections in the basal ganglia

The circuitry of the basal ganglia is schematically shown in Figure 3.3. There are conceived two major pathways through the basal ganglia: *direct* and *indirect* pathways [4, 33]. The *direct* pathway is the projection from the striatum to the GPi and SNr, which then projects back to the thalamus. The *indirect* pathway is the projection from the striatum through the GPe then STN to GPi and SNr.

The output from the basal ganglia, or the GPi and SNr, is mediated by the inhibitory connection to the thalamus. Inhibiting this output works to disinhibit the output of the thalamus to the cortex, which excites the cortical activities and results in behavioral movements. Thus, the direct pathway is regarded to facilitate movements by inhibitory connection from the striatum to the GPi and SNr. In contrast, the indirect pathway is considered to decrease the excitation in the cortex because the connections from the striatum to the GPe work to suppress the inhibition by the GPe on the STN, which has the excitory connection to the GPi and SNr [33].

Dopamine neurons in the SNc projects primarily to the striatum, both of the CD

Figure 3.3

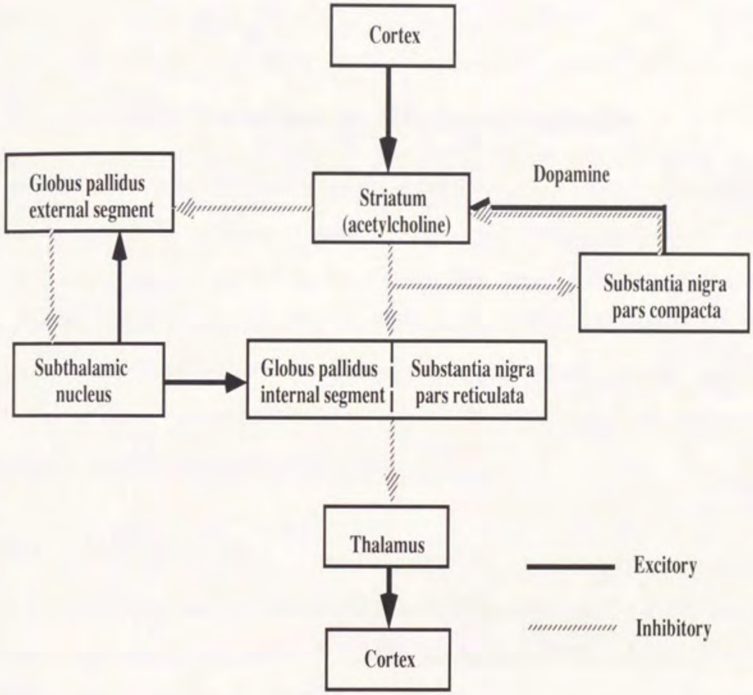

Cortex
Globus pallidus external segment
Striatum (acetylcholine)
Dopamine
Substantia nigra pars compacta
Subthalamic nucleus
Globus pallidus internal segment
Substantia nigra pars reticulata
Thalamus
Cortex
Excitory
Inhibitory

Figure 3.3: Schematic diagram of the internuclear connections in the basal ganglia, which is modified from the diagram in Kandel et al. (1991 p653) [33]. The gray arrow stands for the inhibitory connections, whereas the black arrow stands for the excitory connections.

and Pt. The loss of dopamine neurons are known to contribute to the symptoms of Parkinson's disease [33]. As examined in detail later, the dopamine neurons are regarded to contribute to reinforcement learning that occurs in the basal ganglia-thalamocortical loops.

3.6 Loops in relation to the basal ganglia

A striking characteristic of the basal ganglia is that it has the projections from almost the entire cortex. The influence of these projections goes back to most cortical areas of the frontal cortex through the thalamus [4, 33] (Seeo Figure 3.1). Though it is yet to investigate for a full understanding of functional roles of these connections, there are conceived at least four or five circuits on these "basal ganglia-thalamocortical" loops [4, 33]. Among them, we briefly review three circuits: the *motor* circuit, the *oculomotor* circuit, and the *dorsolateral prefrontal* circuit.

3.6.1 Motor circuit

Motor circuit is important in the study of this thesis because this circuit is much involved in both of acquisition and execution of sequential movements. In the *motor* circuit, most of the projections to the basal ganglia originate from the primary motor cortex (M1), the premotor area (PMC), the supplementary motor area (SMA), and the presupplementary motor area (pre-SMA). There are also the projections from the primary somatosensory cortex (S1) and from the somatosensory association cortex [4]. These projections principally terminate in the *putamen* (Pt), in particular *the bulk of the putamen*, apart from its most rostral and caudoventral extensions [4]. Then through the thalamus, these projections in the motor circuit are projected back primarily to the SMA and PMC, and to

M1. The SMA, PMC, and M1 are reciprocally interconnected with each other.

Among motor cortical areas, many cells of M1 are known to be closely related to specific muscular movements and to the degree of force exerted by the muscles [33]. Though it is still under extensive discussion, one of the prominent view on the relationship between the SMA and the PMC is the theory that the SMA takes part in self-initiated or internally-guided movement whereas the PMC is involved in movements triggered by sensory inputs [73]. Among six evidences of supporting this hypothesis in Tanji [73], it is noteworthy that (1) major subcortical inputs to the SMA via the thalamus are from the basal ganglia, whereas the PMC receives predominant inputs from the cerebellum; (2) As corticocortical inputs, the PMC receives visual information from the parietal and prefrontal cortex, whereas the SMA receives primarily somatosensory information from the parietal cortex, and limbic inputs from the cingulate cortex; (3) The patients with the lesion including the SMA had difficulties in performing sequential limb movements without external guidance by sensory information; (4) Roland et al. [52] has shown that when the subjects are asked to rehearse finger sequential movements, the blood flow increased locally in the SMA, not the PMC, whereas when the subjects performed these movements, the blood flood increased in both of the SMA and PMC. In addition to evidences listed above, Halsband et al. [21] found that the SMA is much related to the retrieval of correct movement or motor sequence from memory. Thus, it can be considered that the SMA is primarily involved in internally-generated sequential movements, whereas PMC is primarily involved in movements triggered by sensory inputs.

Note that it is recently proposed that the rostral part of the SMA is termed as the presupplementary motor area (pre-SMA), whereas the caudal part is redefined as the SMA proper, according to different physiological and anatomical characteristics of neurons between these areas [40], though some controversy still remains as to the exact boundaries of the SMA [73]. We follow this terminology through the entire thesis. For a detailed discussion, see Tanji [73], Luppino et al. [39], and Bates and Goldman-Rakic [8]. The SMA and the pre-SMA are reciprocally connected. While not the pre-SMA but the SMA is connected with M1, the pre-SMA, not the SMA, is connected with the prefrontal cortex (around the principal sulcus) [40, 73].

Matsuzaka et al. [40] found that the pre-SMA contains a significantly higher proportion of neurons with (1) cue responses, (2) preparatory activity, and (3) time-locked activity to movement trigger signal rather than the SMA proper. Halsband et al. [22] have done the experiment in which monkeys have performed the combination of three simple movement as a sequence under two experimental conditions: internally-generated (memory-guided) and externally-triggered conditions. Neural activities in the SMA, pre-SMA, PMC and M1 are analyzed in relation to time periods of one trial, which are classified as *instruction, delay, premovement, movement* and *reward* periods. Their results suggest that the pre-SMA neurons were generally more active during the delay and premovement as compared to the movement, instruction and reward periods. Thus the pre-SMA neurons are, in general, more related to the period after receiving sensory inputs and before starting movements. Neural activities in the pre-SMA were more related to the pre-movement period in the externally-triggered condition, whereas neural activities in the SMA were more related to the movement period in the internally-generated condition.

This result seems to imply that the pre-SMA was more involved in initiating movements when sequential movements were triggered by sensory inputs. In contrast, SMA neurons were more active in sessions of internally-generated movements whereas the PMC neurons were more active in externally-triggered movements. Such preferential activity was rarely found in the M1 neurons. These results coincide well with the view, discussed above, such that the SMA is important in programming sequential movements.

The question, then, arises with particular interest what different functional roles the SMA and pre-SMA have. The pre-SMA has the massive projections to the anterior striatum [51] and is interconnected with the dorsolateral prefrontal. As explained below, the pre-SMA seems much interacted with areas in the dorsolateral prefrontal circuit. It is, therefore, very interesting to address the functional role of the pre-SMA in such interaction of different circuits, which will be discussed later in detail.

3.6.2 Oculomotor circuit

In the *oculomotor* circuit, the projections from the frontal eye fields (FEF), the supplementary eye fields (SEF), the dorsolateral prefrontal cortex, and the posterior parietal cortex, all of which are interconnected, go to the body of the caudate. Each of these cortical areas projects to superior colliculus (SC) [4]. The body of the caudate nucleus (CD) projects to the SNr and GPi, which, then, projects back through the thalamus to the FEF and SEF. The SNr has at least some projections to SC via the thalamus as well [4, 33]. It may be worth mentioning that the SEF is a small area separated from either the SMA or pre-SMA. The SEF is connected to cortical and subcortical areas related to oculomotor control [73]. The SEF is known to be involved in memory-guided saccades [12, 13].

3.6.3 Dorsolateral prefrontal circuit

The dorsolateral prefrontal cortex (DLPF) locates within and around the principal sulcus and on the dorsal prefrontal convexity. This area projects to the dorsolateral head of the CD, which extends to the tail of the caudate throughout a continuous rostrocaudal expanse [4, 65] and to the rostral putamen. Posterior parietal cortex that is connected with the DLPF also projects to the dorsolateral head of the CD [4]. Many neurons in the DLPF exhibit the sustained activities in relation to the spatial sensory information. The DLPF is, thus, well known to have a short-term spatial memory characteristics [17] and has also been assumed to have a wider role in human cognition [4, 19]. With the studies of the prefrontal dysfunction and electrical recording of neuron activities, it is revealed that some neurons in the DLPF may code a preparatory set to respond with an emphasis on the temporal order of the responses [17, 19]. The fact that incremental firing precedes only correct responses is considered as supporting an idea that neurons in DLPF may be part of the internalized code to guide correct responses [19]. In addition, it is experimentally shown that some neurons may monitor the outcome of goal-directed behavior [78]. Thus, it is very likely that the dorsolateral prefrontal circuit with the DLPF is involved in getting the memory of sequences, depending upon the sensory information rather than the motor information.

In summary, it can be schematically regarded that (1) *the anterior striatum as in the prefrontal circuit* receives the massive projections from the DLPF, posterior parietal cortex, and the pre-SMA; (2) *The posterior caudate (or the body of the caudate) as in the oculomotor circuit* receives the projections from the FEF,SEF, DLPF, and the posterior parietal cortex; (3) *The posterior putamen as in the motor circuit* receives the projections from the PMC, M1, SMA and S1.

3.7 Striatum

Besides the projections from several cortical areas discussed in the previous sections, the striatum also receives limbic inputs from at least two sources: the substantia nigra pars compacta (SNc) and amygdala. These limbic inputs are known as signaling motivational states. The ventral striatum primarily receives these limbic inputs [4,59]. Even though, however, reward-related activities are more frequent in the ventral striatum, such activities are also found to surprising extents in more dorsal regions as well [59]. Apparently, both caudate and putamen appear to be involved in setting and maintaining central preparatory states related to the internal generation of individual behavioral acts on the basis of information about the environmental situation [62].

Loosely speaking, the neural activities in the striatum resemble the activity of cells in the motor, premotor and supplementary motor area at a first glance. The activities are usually related to directionally-selective, and, passive and active, movements of individual parts of the body [33]. Despite this resemblance, however, neurons in the striatum, for example in the putamen, are in general tended to be selective for the direction of limb movement than for the activation of specific muscles [33]. Though neurons in the striatum shows preparatory activities before the earliest muscle activity [53], neurons selective for movement in the striatum fire later than these in the cortical motor area in response to visually guided tracking tasks [33]. With these findings, the striatum, or the basal ganglia, is considered *not to play a significant role in the initiation of stimulus-triggered movements* and do not specify directly the muscular forces necessary for the execution of movement [33]. This view corresponds with the view of the relation between the basal ganglia and the cerebellum. The basal ganglia may be involved in the initiation of *internally generated movements* because some neurons show the time-locked activity before the monkeys start the movement voluntary and stops abruptly after starting the movements. The same neurons do not show such an activity when the movement is initiated by external cue [59]. This possibility is consistent with the striking inability to initiate movement (akinesia) exhibited by patients with Parkinson's disease [33]. Some of neurons in the striatum exhibit the expectation or preparation-related activities because the increased discharge rate continues until reward is delivered and stops abruptly thereafter [59].

Hence, the striatum is regarded to have an important role in the acquisition and execution of sequential movements, in particular, in terms of association of sensory inputs with motor outputs given the reward signal of the consequence. However, the output of the basal ganglia or the striatum is not the motor output per se but the output that requires a kind of interpreter to make itself to be motor output.

3.8 Dopamine neurons in the substantia nigra pars compacta

The profile of neural responses of dopamine (DA) neurons in the substantia nigra pars compacta (SNc) is particularly interesting for the learning process of sequential movements from a computational viewpoint. Schultz and his colleagues performed a series of experiments with monkeys, using the delayed go-nogo task, to investigate the profile of neural responses of DA neurons [42, 43, 58–63]. Their basic results are as follows (Also see Figure 3.4):

- *Optimal stimulus*: The optimal stimulus for activating DA neurons consists of *a phasically occurring unpredicted food and liquid reward* [63].
- *Transition of dopamine responses*: As DA neurons respond to a reward-predicting stimulus, they stop responding to the reward of itself [63].
- *Unpredictedness of stimuli*: The responses of dopamine neurons to rewarding or potentially rewarding liquid are due to the temporally unpredicted stimulus occur-

Figure 3.4

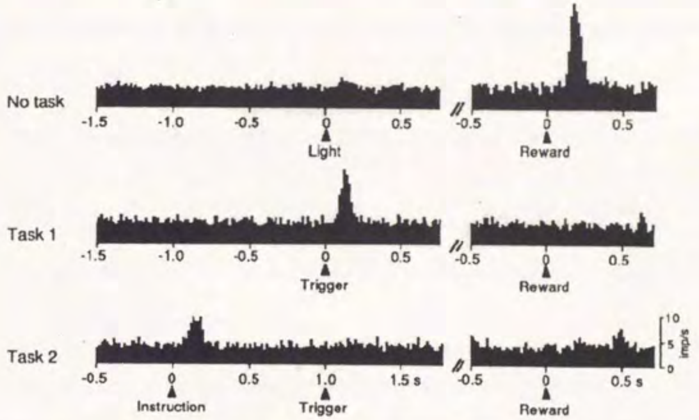


Figure 3.4: Responses of dopamine neurons to unpredicted primary reward and the transfer of this response to progressively earlier reward-predicting stimuli. Adapted from Schultz [63]. All displays show population histograms obtained by averaging the normalized perievent time histgrams of all dopamine neurons recorded in the indicated behavioral situations, independent of the presence of a response. (Top) In the absence of any behavioral task, there was no population response tested with a small light, but an average response to a drop of liquid delivered at a spout in front of the animal's mouth. (Middle) Response to a reward-predicting trigger stimulus in a spatial choice reaching task, but absence of response to reward delivered during established task performance. (Bottom) Response to an instruction cue preceding by a fixed interval of 1 second the reward-predicting trigger stimulus in an instructed spatial reaching task.

rence. A known, reward-predicting, tonic context does not prevent DA neurons from responding to the rewarding liquid. The responses during learning apparently occur because reward is not yet reliably predicted by a conditioned phasic stimulus. Only explicit phasic stimuli predicting the time of reward reduce these responses [42].

- *Population Coding*: The responses of DA neurons to these different stimuli are remarkably similar to each other. They are phasic and occur with latencies of 50 to 120 ms, last less than 200 ms. *This homogeneity of neuronal responses* suggests that dopamine neurons respond in parallel as a *population* rather than displaying differential response profiles [63].

- *Suppressed response to errored trials*: While the monkeys are learning a task and if the monkeys push a correct lever, the phasic activation of DA neurons would occur. When th monkey pushed an incorrect lever and did not receive the reward, dopamine neurons exhibit a depressed activity. [58].

- The events to which DA neurons respond belong to the most important and salient external stimuli to which a subject needs to react in order not to miss an important object [63]. Salient stimuli are *unconditioned rewards* and *aversive stimuli*, *conditioned stimuli predicting rewards or punishment*, and *high-intensity, surprising, novel stimuli* [63]. However, most DA neurons respond best to only a subset of salient stimuli, namely *primary rewards and conditioned reward-predicting stimuli* [63].

Since DA neurons discriminate between reward-predicting and nonpredictive stimuli as long as these two kinds of stimuli are sufficiently dissimilar to each other [63], and in contrast, since dopamine neurons respond to both rewarded and unrewarded stimuli when they are physically very similar [63], there are at least inputs to DA neurons from sensory information. In addition, in contrast to neurons of amygdala that respond to primary rewards in well-established tasks as an example, DA neurons rather respond to the unexpected reward. These evidences led to several researchers including Schultz et al. [63], Houk et al. [31] and others to consider DA neurons as coding the error signals, in particular, the temporal difference errors in the framework of reinforcement learning (RL) according to Houk et al. [31] and others [7, 10, 14, 15, 30, 48, 80]. From a computational viewpoint, the monkeys can learn to predict steadily the coming primary reward, and, as in the framework of RL, learn to perform sequential movements, or sequential decision making. Note, however, that the tasks employed in the experiments of Schultz and his colleagues are rather simple tasks such as the delayed go-nogo task than sequences in general. In contrast, the serial button press task developed by Hikosaka and his colleagues is more suitable to be considered as one of tasks of the sequential decision making, which will be explained in detail in Chapter 5.

The reviews of reinforcement learning and of the aforementioned computational models will be given in the next chapter.

Chapter 4

REINFORCEMENT LEARNING AND COMPUTATIONAL MODELS OF THE BASAL GANGLIA

4.1 Introduction

A brief summary of reinforcement learning (RL) is, first, provided in this chapter. Secondly, computational models of the functions of the basal ganglia based on the framework of RL are reviewed with an emphasis on the comparison to the study in this thesis.

RL has been recently paid much attention to in machine learning and neural network communities because RL has a solid mathematical ground and can be applied to various complex problems [75]. The computational framework of RL fits well with sequential decision making and provides a mathematically solid means to answer the temporal credit assignment problem.

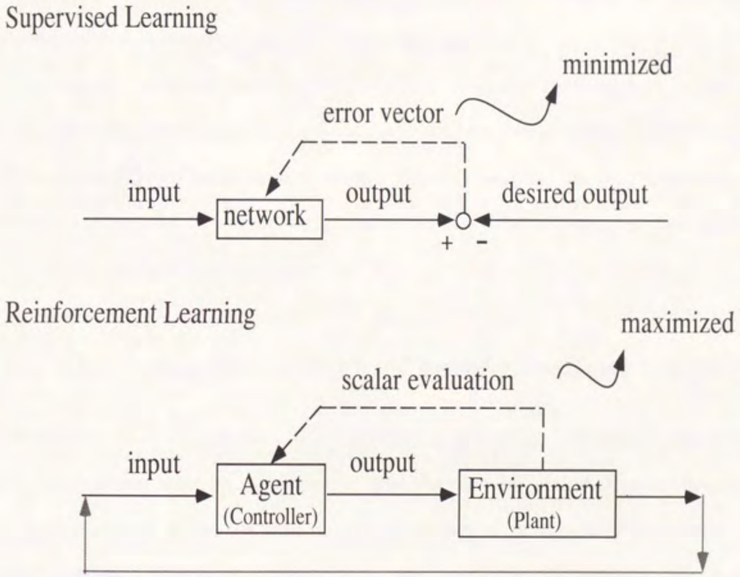
Houk et al. [31] and Barto [7] recently proposed, based on experimental findings of profiles of neural activities of dopamine neurons such as Schultz et al. [63] and others [42, 43, 58–63], that the inter-circuitry of the basal ganglia with dopamine neurons may perform a type of reinforcement learning, which is known as *Actor-Critic* scheme. The aim of this chapter is to provide a concise background of reinforcement learning to evaluate their models. Therefore, the review of RL in this chapter is far from exhaustive. For the more thorough and general review of RL, refer to Kaelbling et al. [32] and Bertsekas [11]. The functional model of the basal ganglia-thalamocortical loops that will be proposed in the following chapters is partly based on the framework of reinforcement learning.

4.2 Reinforcement learning and supervised learning

There are basically three classes of the learning paradigm in artificial neural network: unsupervised learning, reinforcement learning, and supervised learning [23]. The problems in the class of the supervised learning are most investigated. The class of reinforcement learning (RL) is briefly discussed in contrast with that of supervised learning in this section.

See Figure 4.1. In supervised learning, the neural network is informed of the pair of the input and its corresponding output, which is called the desired output, or the target output. Hence, the neural network can know how different its current output given the input from the desired output in the supervised learning. Using the information of such difference, the learning such as the method of steepest descent can be employed in the supervised learning. Because neural networks treat the vector representation as its input

Figure 4.1



Supervised Learning

Reinforcement Learning

Figure 4.1: Schematic diagram of supervised learning (above) and reinforcement learning (below).

and output in most of cases, in other words, the neural network is fed with the desired output as the vector in the supervised learning. In contrast, in the class of reinforcement learning, when the neural network produces an output given an input, the neural network is not informed of the desired output but, if any, only of how good/bad the output is, which usually takes the form of *scalar*. The learning occurs in RL by utilizing such scalar information. It is obvious that the learning in RL has less information to improve the performance of the network than in the supervised learning.

Note that in case of sequential decision making, the neural network can't be spontaneously informed of how good/bad its output that the network produced was because of the temporal credit assignment problem. There, therefore, needs a mechanism in problems of sequential decision making to take the long-term consequence into account as well as the immediate consequence.

4.3 The basic framework of reinforcement learning

In the framework of RL, we deal with situations as following. There is an agent in a world, which is in a state, x_t, at a time, t. The agent makes an influence on the world by taking an action, a_t, at the time, t. By the action, a_t, the state of the world goes to another state, x_{t+1} at a next time, $t+1$. The agent, then, performs a next action, a_{t+1}. The world state changes again to the next state, x_{t+2}. Thus, it continues. We can represent such transition by introducing the *transition function*, denoted by $T(x_t, a_t)$ given a state, x_t, and a chosen action at the state, a_t. Then,

$$x_{t+1} = T(x_t, a_t) \tag{4.1}$$

In this transition process of the states and actions, the agent sometime, say, at the transition of a time, t, gets an *immediate*, either positive or negative, reward, r_t, such as getting food, or other times gets nothing. That is,

$$r_t = R(x_t, a_t), \tag{4.2}$$

where $R(\cdot, \cdot)$ is the function, called *reward function*, that provides an immediate reward, r_t, given a current state, x_t, and a chosen action, a_t.

Thus, the basic format of RL problems goes with the triplet data set, $(s_t, a_t, r_t) \quad (t = 0, 1, 2, ...n-1)$ and the final state, s_n with the transition function, T, and reward function, R[1].

- $s_t, s_{t+1}, ...s_{t+i}..., s_{t+n}$
- $a_t, a_{t+1}, ...a_{t+i}...a_{t+n-1}$
- $r_t, r_{t+1}, ...r_{t+i}...r_{t+n-1}$

Based on a choice of an action at a time, the transition of the world states changes, and then, rewards that the agent can obtain over the transition will be changed. The question to be asked in this framework is what action should be taken at a state. As discussed in Chapter 1, an action taken at a time has both of immediate and long-term consequences so that choosing the action at a time that gives the agent the maximum immediate reward is not necessarily the best choice. The purpose of the agent's performance is to maximize rewards that the agent can obtain over a whole transition. To increase obtained rewards, the agent must solve the temporal credit assignment problem. RL provides one way to answer this question.

In terms of the goal-directed behaviors in sequential decision making, it is often assumed that the reward is given in the transition to the final state much higher than

[1] n can be ∞ but for the sake of simplicity, we assume $n < \infty$

that to any other state. In contrast, rewards given the transition to states other than the final state are assumed to signal only *roughly* how those states are desirable to reach the final state or in some cases assumed to signal the failure of a trial by its transition. By use of such partial information, the agent must find and learn the optimal path to reach the final state.

4.4 Value function and policy

It may be obvious to solve the temporal credit assignment problem that a basic strategy is to evaluate a chosen action not only with the "immediate" reward but also with rewards over a whole transition. To do so, RL utilizes the *value function.* The value function is the function given a current state that gives the output such that *a weighted sum of all future rewards.* Let us denote the value function by $V(x_t)$ given a current state, x_t. Then the value function can be defined as:

$$V(x_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3}..... + \gamma^{n-1} r_{t+n-1} \tag{4.3}$$

or equivalently,

$$\begin{aligned} V(x_t) &= \sum_{k=0}^{n-1} \gamma^k r_{t+k} \\ &= \sum_{k=0}^{n-1} \gamma^k R(x_{t+k}, a_{t+k}), \end{aligned} \tag{4.4}$$

where γ is $0 \leq \gamma \leq 1$, called a discount factor.

When $\gamma = 1$, all future rewards are equally taken into account in the value function, whereas, if $\gamma < 1$, the later rewards are less weighted. The magnitude of γ controls how much the long-term consequences should be taken into account in the value function. In this way, the amount of the value function given a state stands the value of the state, which takes into account the long-term consequence which is weighted by the discount factor *gamma*.

Note that by the definition of the value function, the local constraint of the value function is given as follows:

$$V(x_{t+k}) = r_{t+k} + V(x_{t+k+1}) \tag{4.5}$$

The agent should perform an action at a state. It is, then, convenient to consider a function that maps a state to an action. This function is called *the policy*. Let us denote the policy by $u(x_t)$ given a state, x_t. Then,

$$a_t = u(x_t) \tag{4.6}$$

If the policy is fixed, the transition at any state is also fixed. Given the fixed transition, it is possible, theoretically at least, to obtain the value given by the value function at all states. Needless to say, the action that should be taken at any state is given by the policy.

If an action at a state, x_{t+k}, is changed, the transition can be changed at the state and, consequently, the rewards obtained over the transition can be changed at the state as well. This also means that the value function of other states may be changed. The aim of RL is to find, at each state, an action that maximizes the value function of each state. Generally speaking, it often happens that the transition function is not yet known to an agent or the reward function is not yet known to an agent either. In these cases, the agent must explore the world at first to some degree to be able to estimate the transition and reward functions and to improve its performance. Along with such estimations, the agent should also seek for the optimal policy as well. Thus, one of the most representative method in RL, called the temporal difference (TD) learning, would perform twofold of estimations; the estimation of the value function given a policy and the estimation of the optimal policy among possible ones. The TD learning is discussed in the next section.

4.5 Temporal difference learning

The method of the temporal difference (TD) learning finds the optimal value function and the optimal policy as the agent explores the world, that is, as the agent collects the information of the transition and reward functions, even if these functions are not available in advance. In this section is reviewed the simplest type of the TD learning, which is classified as TD(0) in a more generalized framework of the TD learning, called the TD(λ) learning [70].

As shown in Section 4.4, the local constraint of the value function is given as:

$$V(x_{t+k}) = r_{t+k} + V(x_{t+k+1})$$

Suppose that the estimated value function at the state x_t is denoted by $P(x_t)$, then the temporal difference (TD) error, denoted by $\hat{r}_t$, is defined as

$$\hat{r}_t = r_t + \gamma P(x_{t+1}) - P(x_t) \tag{4.7}$$

With this TD error, for example, the method of the steepest descent can be employed. Suppose that the value function is estimated by the linear function of the state, x_t, which is in a form of a vector representation, $(x_{t1}, x_{t2}, x_{t3}, ...x_{tl})$, with the weight vector,$(v_1, v_2, v_3, ...v_l)$. The value function can be given as

$$P(x_t) = \sum_{i=1}^{l} v_i(x_t)_i \tag{4.8}$$

Given the TD error $\hat{r}_t$, the estimated value function can be updated by the method of the steepest descent as follows:

$$\begin{aligned} \Delta v_i &\propto \hat{r}_t \frac{\delta P}{\delta v_i} \\ &\propto \hat{r}_t (x_t)_i, \end{aligned} \tag{4.9}$$

where Δv_i stands for change of v_i.

The TD error, $\hat{r}_t$, is also called the *effective reinforcement* , whereas, in contrast, the immediate reward, r_t, is also called the *primary reinforcement.* Note that, though it is stated in Section 4.4 that the TD learning should perform twofold of the estimations, the estimation of how the policy should be chosen is neglected in the above explanation. It is rather assumed in the above explanation that the policy is fixed. In order to obtain the optimal value function and the optimal policy, however, the estimation of the policy cannot be neglected. As a matter of fact, the actor-critic scheme provides one way to do so by use of the TD learning, which is discussed in the next section.

4.6 Actor-critic scheme

Barto et al. [6] have first shown the efficiency of the temporal difference learning in the cart-pole-balancing problem. The scheme they used in the task is called the actor-critic scheme shown in Figure 4.2.

While the critic is assumed to estimate the value function, the actor learns what action should be more preferable to be taken in each state. The procedure for learning of the critic is based on the TD learning as explained in Section 4.5. To understand how to let the actor learn the optimal policy, suppose that at a state, $x_t = (x_{t1}, x_{t2}, x_{t3}, ...x_{tl})$, there are the choice of actions, $\{a_{t1}, a_{t2}, a_{t3}, ...x_{tk}\}$, available. Then, we define the weight

Figure 4.2

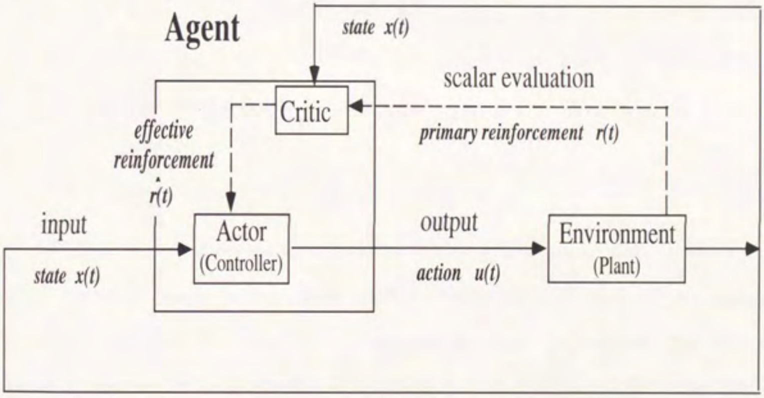


Figure 4.2: Diagram of the actor-critic scheme.

matrix, W, to provide the probability vector with respect to the choice of actions, as follows:

$$W = \{w_{ji}\} \quad (j = 1, 2, ..., k, \;\; i = 1, 2, ..., l) \tag{4.10}$$

Let us denote the probability vector with respect to the choice of actions by p_{a_t}. The probability to take an action, a_{tj}, denoted by $(p_{a_t})_j$, can be, then defined as

$$(p_{a_t})_j = \frac{\sum_{i=1}^{l} w_{ji}(x_t)_i}{\sum_{j=1}^{k} \sum_{i=1}^{l} w_{ji}(x_t)_i} \tag{4.11}$$

When an action, a_{tj}, is taken and if the TD error, $\hat{r}_t$, is obtained, the weight matrix, W, can be updated by the following rule:

$$\Delta w_{ji} \propto \hat{r}_t(x_t)_i \tag{4.12}$$

Thus, if $\hat{r}_t > 0$ by choosing an action, the action will become more preferable next time at the same state. Such preference will be iteratively and stochastically changed as the agent explores the transition and reward functions. It is proved that after an enough iterations with some conditions, the critic and actor will converge to the optimal value function and the optimal policy respectively [11]. It is worth mentioning that, from a view of competition, each of w_{ji} $(j = 1, 2, ..l)$ can be considered as competing to increase the probability of the action, $(a_t)_j$. It is, hence, often said that actors, each of which is w_{ji}, compete to each other.

In the actor-critic scheme, the critic estimates the value function through minimizing the TD error, $\hat{r}_t$, and ideally, the TD error should apprach to 0, i.e., $\hat{r}_t \rightarrow 0$. Suppose that a primary reinforcement is only given in a final state of a sequence. In this case, in estimation of value function at each state by the minimization process of the TD error by the critic, the value of the primary reinforcement at the final state is propagated to the preceding states of the sequence back in time gradually as the number of trials of the sequence increases, and, needless to say, at the same time, the value function at the final state is also gradually constructed. This nature of the TD error is pointed out as similar to the characteristics of the profile of responses of dopamine neurons, which is a basis of the computational models of the basal ganglia explained in the next section (Refer to Section 3.8 and Figure 3.4 for the experimental findings of responses of dopamine neurons.)

4.7 Review on the computational models of functions of the basal ganglia

In the preceding sections, the framework of reinforcement learning was briefly reviewed. As mentioned, based on the framework, several computational models have been proposed on various aspects of the functions of the basal ganglia with the related cortical areas in relation to dopamine (DA) neurons in the substantia nigra pars compacta (SNc). Among them, the models of Houk et al. [31], Barto [7], Berns and Sejnowski [10], Montague et al. [48] and Doya [14, 15] are of particular interest to this study in terms of sequential movements.

Barto [7] provided the comparison between the actor-critic scheme and functions of the basal ganglia. Similarly and further, Houk et al. [31] discussed that DA neurons in the SNc compute the temporal difference (TD) error with three inputs: from the *direct* pathway, the *indirect* pathway, and the primary reinforcement possibly from lateral hypothalamus (See Section 3.5). The TD error, then, is propagated into the striatum from the projection of DA neurons in SNc. Further, with the dichotomy of matrix and striosome in the striatum, they hypothesized that the striosomal module worked as the critic and the matrix module as the actor in the actor-critic scheme. Their model is important in that they pointed out the similarity between the framework of reinforcement learning and the characteristics of the responses of DA neurons. Their model, however, remains at the theoretical level and is not examined in comparison with the behavioral or neurophysiological results. Montague et al. [48] developed a theoretical framework that shows how mesencephalic DA systems, including DA neurons in the SNc, can encode the prediction errors between the expected amount of reward and the actual reward, on a basis of the TD error similar to the above models but providing a detailed comparison of behaviors between their model and physiological response of DA neurons. Their model shows that the framework of RL, or the TD learning, can capture well the characteristics of response of DA neurons.

Wise and Houk [80] and Houk and Wise [30] discussed the modular architecture linking the basal ganglia, the cerebellum, and the cerebral cortex with an emphasis on the motor control. Though their models referred to many neurophysiological evidences and the detailed characteristics of types of neurons, their models remain at the speculative level at best. Their models are not examined with the actual neural responses nor the behavioral data. The emphasis of their model is the integration of the basal ganglia and the cerebellum in motor control not at the computational level but at the abstract level.

Doya [14, 15] proposed the actor-tutor scheme as the model of the integration of functions among the basal ganglia, the cerebellum, and cerebral motor areas in terms of the motor control with an emphasis on the dynamics and kinematics. It is hypothesized that the basal ganglia learns the value function and generates the desired motor direction [14, 15], which is transformed into a motor command via the lateral part of the cerebellum. This motor command is used for control in early stage of learning. In his scheme, the direct controller can be trained by supervised learning with the use of this motor command as the teacher, instead of reinforcement learning, given some conditions as described in Doya [15]. In the simulation of cart-pole swing up task, it is shown that the learning speed has been much facilitated with this model.

Berns and Sejnowski [10] proposed the model of action selection in the basal ganglia for given cognitive, or internal, states and sensory inputs. They posit that the internal segment of the globus pallidus (GPi) works as "winner-lose-all", which is that the cells of standing the most desirable action should be turned off. This "winner-lose-all" process results in firing target cells of the thalamus to let the action of the winner be taken. To have such a mechanism in the GPi, the inputs from the *indirect* pathway, which is discussed in Section 3.5, work to gate the competition in the GPi. The subthalamic nucleus (STN) receives a prominent excitatory projection from the cortex and Berns and Sejnowski [10] proposed that the function of the indirect pathway is to inhibit actions that have recently been selected. They suggested that because the STN receives the projections from the prefrontal cortex, the selection process in the GPi can be influenced by the inputs from the indirect pathway, that is, the cognitive, or internal, states. They consider inputs from DA neurons in SNc to the striatum as the error signal of the temporal difference learning. Their model is not examined with the behavioral data.

Though the computational models discussed above of functions of the basal ganglia shed light on several important aspects of its functions, to the knowledge of the author, any of these models has not been compared closely with actual behavioral data of sequential decision making as complex as that of 2x5 task. The models of Houk et al. [31], Wise and Houk [80], and Houk and Wise [30] remain at the speculative level. Montague et al. [48] concentrated on examining the detailed correspondence between the response of DA neurons and the error signal of the model, which is hypothesized to be encoded by DA neurons. The model of Doya [14, 15] is very interesting in terms of the motor control and can be further examined at the computational level. The model of Berns and Sejnowski [10] is concerned with the action selection mechanism in relation to the dorsolateral prefrontal circuit. In contrast to these computational models, this study rather aims to investigate the functions of the basal ganglia and related cortical areas in terms of the sequential decision making without getting into details of the aspect of the motor control. The interaction among the basal ganglia-thalamocortical loops is of particular interest in relation to its functional roles in sequential decision making.

Chapter 5

REVIEW ON BEHAVIORAL AND NEUROPHYSIOLOGICAL FINDINGS IN EXPERIMENTAL WORKS

5.1 Introduction

In Chapter 3, the inputs and outputs of the basal ganglia, the internuclear structure of the basal ganglia, and some of the basal ganglia-thalamocortical loops are briefly summarized. Based on anatomical and experimental evidences, those loops of the basal ganglia are considered to be heavily involved in sequential movements, in particular in cognitive aspects in comparison to the other major subcortical loop of the cerebellum. It is, therefore, intriguing to ask how those loops of the basal ganglia work in acquisition and execution of sequential movements from a computational viewpoint. In contrast to computational researches on the functions of the cerebellum (e.g. Albus [2], Kawato [34]), the computational researches on the role of the functions of the basal ganglia in sequential movements have just started recently. In Chapter 4, the framework of the reinforcement learning and the computational models of the functions of the basal ganglia based on this framework are reviewed. Even though those computational models shed light on some aspects of the functions of the basal ganglia, there is no examination of the performance of the model in close comparison with behavioral data. Nor those models have investigated much of the functional relationship among the basal ganglia-thalamocrtical loops in sequential movements.

The aim of this chapter is twofold: to introduce in detail the experiments of the Hikosaka laboratory that have recently developed and that are very suitable to investigate the functions of the loops of the basal ganglia in sequential movements, and to discuss neurophysiological findings, including results of the Hikosaka laboratory, on the striatum, the presupplementary motor area, and the supplementary motor area, all of whose functions are very important for further investigation of the functions of the loops of the basal ganglia in sequential movements. The investigation of behavioral and neurophysiological findings in this chapter will be a basis of the hypothesis of the functions of the loops of the basal ganglia in the next chapter and the experimental task of the Hikosaka laboratory will be simulated to examine the performance of the model based on the hypothesis in Chapter 7.

5.2 The serial button press task of the Hikosaka laboratory

This section aims to introduce the experimental paradigm of serial button press task, which Hikosaka and his colleagues have developed [25, 29, 44, 45]. There is a variation in their devised serial button press task. Among them, the most representative one, called 2x5 task [25], is explained below in detail. Then, other variations are briefly explained.

A major advantage of their experiments is that their task has a hierarchical structure similar to our daily learned actions and, in addition, variations of the task can be generated practically as many as possible [25]. Thus, it can provide subjects with a situation similar to our daily life, in which we develop and acquire sequential movements, or skills, in a long run by our occasional and frequent experiences.

In the next section, their behavioral findings and implications are discussed. Their results will be compared with the model of the basal ganglia proposed later in this thesis.

5.2.1 The 2x5 task

The 2x5 task requires subjects, or monkeys, a sequential hand movement task consisting of ten button presses with many different variations [25]. The name of '2x5' originates in that there are 2 stimuli in each 'set' and 5 sets in each 'hyperset', one of which is tested in each trial. The description of their experiment in this section is quoted with a slight modification from Hikosaka et al. [25], in which a more thorough description can be found.

Behavioral paradigm

Figure 5.1 shows an example of the sequence of events in a single task trial. At the start of a trial, the home key was turned on. When the animal pressed the home key for 500 msec, two of the 16 target LEDs were turned on simultaneously, called 'set'. The animal had to press the illuminated buttons in a correct (predetermined) order which s/he had to find out by trial-and-error. If successful, another pair of LEDs, a second set, was illuminated which the monkey had to press again in a predetermined order. A total of 5 sets were presented in a fixed order for completion of a trial, called 'hyperset'. *Therefore, it should be noted that the correct order of button within a set, which displays the same two illuminating stimuli among the 16 target LEDs, may be different between different hypersets.*[1]

When the animal pressed a wrong button, all LED buttons were illuminated briefly with an unpleasant beep sound, and the trial was aborted without any reward. The animal then had to start again from the home key as a new trial. It should be emphasized that, in the following analyses, a trial was defined to be successful only when the animal completed the whole hyperset (5 sets).

After each successful set, however, the animal was given a liquid reward. The amount of the reward increased gradually from the first to the final set (from 150 to 300 ms of reward delivery duration) so that the total amount of reward was maximized by completing all sets. The duration of reward delivery was inserted between sets, during which no stimulus was presented (inter-set interval). Such a distributed delivery of reward was necessary because, when presented a new (unexperienced) hyperset, it was virtually impossible for the animal to complete the whole hyperset for the initial trials.

A major advantage of the 2x5 task is that new hypersets can be generated practically as many as possible. Since the number of possible combinations for a set is 16P2, the number of possible combinations for a hyperset is $(16P2)^5$, which amounts to about 7.96×10^{11}, an astronomical value. There have been no identical hypersets among a total of more than 1000 hypersets used for the two monkeys.

[1] This italic sentence is inserted by the author of the thesis.

Figure 5.1

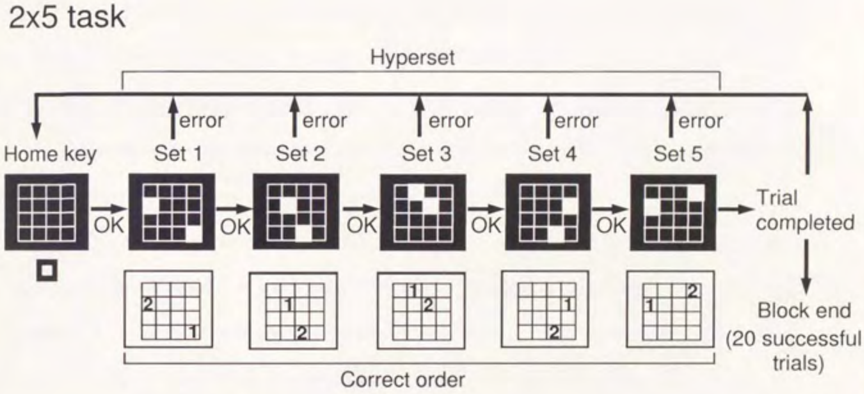


Figure 5.1: Procedure of 2x5 task with an example of a hyperset . To complete a trial, a monkey has to press 10 buttons (2 buttons x 5 sets) in a correct (predetermined) order.Taken from Miyachi et al. with a slight modification.

Experimental procedures

The same hyperset was used throughout a block of experiments so that the monkey experienced the same sequences of button press repeatedly until he completed 20 successful trials, which was defined as one block of experiment; a different hyperset was used for the next block. The monkey performed about 20 blocks for one day; a given hyperset was used usually only once in a daily experiment.

Some of the hypersets, chosen as learned hypersets, were examined daily. During the period of learning, the monkey performed the hypersets as the daily routine, one block for each hyperset. The learned hypersets were allocated to three groups in terms of the hand used: for the right hand only, for the left hand only, and for both hands. A 'both hand' hyperset was examined either once a day but the hand alternated across days or twice a day with the hand changed between the two experiments.

In addition to the learned hypersets, the monkey experienced many new hypersets, each of which was tried just once (one block). Half of the new hypersets were performed by the right hand; the other half were by the left hand. Thus, the monkey experienced roughly an equal amount of practice for each hand throughout the learning period.

For two monkeys in Hikosaka et al. [25], each of whom is termed as PI and BO, the total number of the learned hypersets was 28 for monkey PI and 14 for monkey BO. Note, however, that they were started at different stages of the monkey's experience of the 2x5 task. The total number of new hypersets was 313 for monkey PI and 92 for monkey BO. Thus, among about 20 hypersets used for a daily experiment, usually one or two were new hypersets while the others, which were among the learned hypersets, had been experienced to different degrees of learning.

In monkey PI, some of the initially learned hypersets were removed from the daily menu and, after 1 or 6 months, were re-examined to test if the procedural memory was retained after a long-term interruption. During the interrupted period, the monkey continued to learn other hypersets including new ones.

Data analysis

Two parameters were basically used to assess the performance of monkeys: the number of trials (or the number of error trials) to a criterion and the performance time. Criterion for the number of trials is set as 10 successful trials in most of cases. This value, for example, would be 25 if the animal failed in 15 trials before completing 10 trials, and then, in this example, the number of error trials to criterion becomes 15. The performance time as a measure is the time from the home key press to the second button press of the final (5th) set which was then accumulated for the initial 10 successful trials.

5.2.2 Variations of the experiments

The basic setting and procedure is the same as 2x5 task. For functional Magnetic Resonance Imaging (fMRI), the 2x10 task was employed, in which there are 2 stimuli in a set composed of a 2x2 matrix display and 10 sets in a hyperset [29].

5.3 Behavioral findings in the 2x5 task

In this section, we summarize behavioral findings of the 2x5 experiments. For more details of findings, refer to Hikosaka et al. [25], Miyashita et al. [45], Miyachi et al. [44], and Hikosaka et al. [27]. The implication of their findings contribute to the hypothesis of functions of the loops of the basal ganglia in sequential movements, proposed later in this thesis. The findings in relation to underlying neural mechanisms will be reviewed in the next section, with examinations of other experimental results.

Learning

Learning in 2x5 tasks is indicated by the decrease in the number of trials to criterion and the decrease in the performance time. There are observed three levels of learning [25]:

1. **short-term and sequence-selective**: indicated by improved performance for a particular hyperset *during a block of experiment.*
2. **long-term and sequence-selective**: indicated by improved performance for a particular hyperset *across days.*
3. **long-term and sequence-unselective**: indicated by the improvement of performance for new hypersets.

In other words, monkeys could learn, to some degree, to perform a new hyperset within a short period (≤ 5 min). In this sense, monkeys can learn each hyperset, even if it is new, *during a block of experiment.* At the same time, the number of trials as well as the performance time for a particular hyperset is decreased as monkeys experience more with the hyperset *across days.* Furthermore, by the indication of not the number of trials but the performance time, they performed gradually better with more experiences of the 2x5 tasks, regardless of any hypersets.

Maintenance of the learned skills

To examine whether the memory was retained for a long period, Hikosaka and his colleagues had the monkey learn 12 hypersets sufficiently, stopped the training, and retested them after 1 or 6 months. After the 1 month interruption, the performance was significantly better than that for new hypersets [25]. After the 6 month interruption, the performance was not different from new hypersets in terms of the number of trials, but was significantly better than new hypersets in terms of the performance time [25]. Hikosaka et al. [25] suggest that this result indicates that motor memory (measured by performance time) can be retained longer than procedural memory (measured by the number of trials).

Anticipation and the coordination between perceptual and motor system

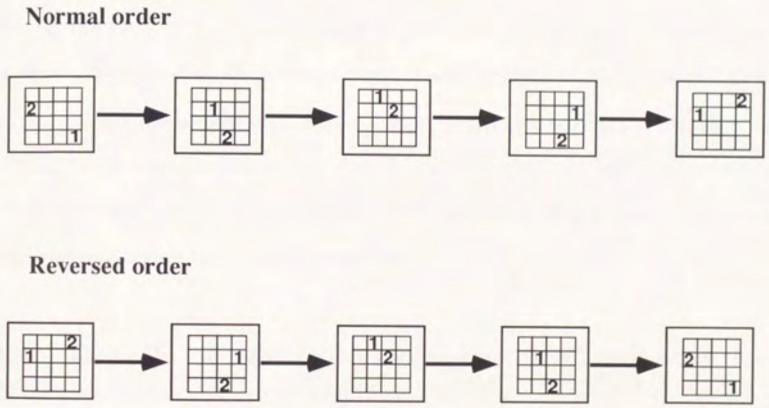
After some sufficient learning, the saccade to the first button tended to occur before the target illumination (*anticipatory saccade*). This was true only for the hyperset that monkeys extensively experienced, called learned hypersets. The likelihood of anticipatory saccade increased gradually over 20-30 days of practice of the particular hyperset. The time course was similar to the button press latency, which was the time from the set (target) onset to the time when the monkeys pressed the first button [45]. The nearly perfect performance of learned hypersets due to the extensive practice was then deteriorated by the use of the opposite hand. In addition, they found that anticipatory saccades became much less frequent when the opposite hand was used [45]. Miyashita et al. [45] suggested with these findings that the critical factor for the *extensively learned* skilled performance was the combination of the eyes and the side of the hand that was used for the practice of a given sequence [45]. This suggestion appears to indicate that the anticipatory saccade rely, at least to some degree, on the memory of motor movements as such information by the representation specific to each hand, for example, kinematic representation of joint movements of each hand. It is because the anticipatory saccade develops in the time course similar to button press latency, which is almost identical to the performance time as a measure, and because the anticipatory saccade was deteriorated by the use of the opposite hand. There is, however, another possibility such that the anticipatory saccade does not depend on the memory of motor movements but on the memory of sequential sensory inputs, and the reason that the anticipatory saccade was deteriorated by use of the opposite hand is because the temporal pattern of sequential sensory inputs are deteriorated by use of the opposite hand that is not as much familiar to a tested particular hyperset as the other hand is (Hikosaka, personal communication). The decrease of button press latency appeared to happen in most cases only when the anticipatory saccade precedes (Miyashita, personal communication). On the data in Miyashita et al. [45], the degree of the learning measured by the number of trials to criterion asymptotes faster than that measured by the ratio of anticipatory saccade. Furthermore, the degree of deterioration by use of the opposite hand is actually much more in the measure of the button press latency than that of ratio of anticipatory saccade and that of number of trials to criterion. Based on these considerations, it seems plausible to consider that the learning of anticipatory saccade happens before that of motor movements, and that the anticipatory saccade depends on the memory of the sensory inputs rather than that of the motor movements. In addition, it can be considered that a multiple of learning processes occur in order: measured by number of trials to criterion, then measured by ratio of anticipatory saccade, and then by button press latency or actually the performance time.

Note that the anticipatory saccade by its definition means that the monkeys could predict the next set and its corresponding correct action. Therefore, it can be said that in the case of learned hypersets, a monkey performs the actions not just by reacting to the sensory input but rather by anticipating the coming sensory input and predicting its corresponding action. The increase in the ratio of anticipatory saccades begins even in the early stage, or even first few days. This suggests that the learning of the sequence would start even in the early stage.

Context dependency of the memory retrieval for the learned hypersets

Hikosaka et al. [27] has tested the question of whether the monkeys have learned the extensively experienced hypersets (learned hypersets) as a whole of sequence or have learned only the correspondence of each set with each corresponding correct action. To differentiate these, the performance of two conditions were compared: in one condition, the monkeys tried with the learned hypersets and in the other condition, the monkeys tried with the hypersets of which all sets were the same as the learned hypersets but the sequence of the sets were reversed. See Figure 5.2. In addition, they compared the result of the reversed hyperset condition with that of the new hypersets. The results were that (1) in both criterion, the number of errors criterion and the performance time criterion , the reversed hyperset condition was significantly worse than the learned hypersets and (2) in both criterion, the performance of the reversed hypersets were not significantly different from that of the new hypersets, whereas, as discussed before, the performance of the learned hypersets is significantly better than that of both of the reversed and new hypersets. These results clearly suggest that the memory of learned hypersets are not merely the memory of the correspondence of the current sensory input with its corresponding action but the memory depending upon the information before the current sensory input (the current set), possibly upon the sets before the current set [27]. Note that if the memory of the correct action to the coming sensory input is retrieved depending upon information before the coming sensory input as discussed in case of anticipatory saccade, the result of this context dependency is understandable because the memory retrieval of the correct action to each set does not depend on the set by itself.

Figure 5.2


Normal order

Reversed order

Figure 5.2: (Above) An example of learned hypersets is given in the normal order; (below) the reversed order of the given learned hyperset above.

Summary of behavioral findings in the 2x5 task

First, all of their findings indicates the different nature of the memory between the early and late stage of learning. They showed that there were three different learning processes. The maintenance of learned skills are retained remarkably longer in the measure of the performance time. The memory of learned hypersets depends on information before the coming sensory input because the monkeys anticipate the coming set and predict its corresponding correct action for the learned hypersets. It can be, therefore, postulated that (1) the memory in the later stage of learning, that is, the memory of the learned hypersets, depends upon the memory of motor movements and (2) the memory of the correct action to the sensory input for the learned hypersets is retrieved in the way of not reacting to the sensory input but anticipating the coming sensory input and predicting its corresponding action, depending upon information such as the sensory inputs and motor outputs before the coming sensory input.

5.4 Details of neurophysiological findings with behaviors in the 2x5 task and other experiments

In this section, the correspondence of behaviors with underlying neural mechanisms in the basal ganglia-thalamocortical loops is discussed in detail with an emphasis on the motor circuit.

As discussed before, the basal ganglia has an important role in internally-generated sequential movements and the motor circuit should play a significant role in it. Among cortical areas included in the motor circuit, the relationship of functional roles of the SMA and pre-SMA is of particular interest. While Hikosaka and his colleagues provided several interesting behavioral findings as discussed in the previous section, they have also employed a series of blockade experiments by the injection of muscimol (a GABA agonist) on several interesting portions such as anterior/posterior striatum, the SMA, pre-SMA, and others. They also tested humans by functional Magnetic Resonance Imaging (fMRI) in the 2x10 task similar to the 2x5 task. These results can provide important clues to examine functional roles of the basal ganglia, the SMA, pre-SMA and other areas. There also exist several experimental findings not discussed yet, which will be further examined in connection with the functional roles of these portions of the brain. In the rest of this section, these experimental findings are examined in correspondence with the related portions of the brain.

5.4.1 Different functions in different parts of the striatum

As discussed in Section 3.7, the striatum is regarded to have an important role in the acquisition and execution of sequential movements in terms of sensory inputs and its response that will be transformed to produce motor outputs under the influence of limbic inputs, or motivational signals. Furthermore, in Section 3.6, it is pointed out that different parts of the striatum are involved in different basal ganglia-thalamocortical loops. The question, then, arises whether different parts of the striatum have different functional roles in terms of sequential movements.

Miyachi et al. [44] tested the performance of learned and new hypersets in the 2x5 task by blocking each of *anterior striatum, middle-posterior putamen*, and *middle-posterior caudate* in contrast with the control condition. Among these conditions, they compared the mean number of error trials to criterion that is a variant of one of measures they used as discussed in Section 5.2.1. First of all, as expected, the mean number of error trials is much smaller for learned hypersets than that for new hypersets. In the condition of blocking the anterior striatum, the number of error trials significantly increased for new hypersets in comparison with the control condition but not significantly increased for learned hypersets. In the condition of blocking the middle-posterior putamen, the number of error trials significantly increased for learned hypersets but not significantly increased for new hypersets. The blockade of the middle-posterior caudate produced no significant changes in terms of the number of error trials for learned hypersets and new hypersets. Based on these findings, Miyachi et al. [44] suggested that the anterior striatum contributes to the acquisition process of new hypersets but not to the retrieval of memory of learned hypersets from the long-term memory (LTM). In contrast, the middle-posterior putamen participates in the retrieval process. Apparently, the middle-posterior caudate does not specifically contribute to either acquisition or retrieval. Note that their results coincide with the correspondence of different parts of the striatum with different circuits. It is well known as discussed before that the anterior striatum has massive projections from prefrontal and the pre-SMA and is involved in the dorsolateral prefrontal circuit. The middle-posterior putamen is involved in the motor circuit and the middle-posterior caudate is involved in the oculomotor circuit. It is, however, noteworthy that the number of error trials for the learned hypersets by the blockade of the posterior putamen did not become as many as that for the new hypersets in the normal condition, even though the increase of the number of error trials by this blockade is statistically significant in comparison with the number of error trials in the control condition for the learned hypersets. It is, therefore, not clear to what extent the posterior putamen is involved in the retrieval process of information from the LTM. Because Miyachi et al. [44] only tested with the new hypersets and the learned hypersets that have already been extensively experienced, there is a possibility such that the stored information of the learned hypersets in relation to the posterior putamen may be already transferred to other portions of the brain at a lower level such as the cerebellum [37]. It should be interesting to employ their blockade experiment with 'half-learned' hypersets.

In summary, the anterior striatum is much involved in the early stage of acquisition process of sequential movements. The posterior putamen can be considered as involved in the retrieval process of information from the LTM, but the degree of the involvement is questioned.

5.4.2 The presupplementary motor area

The pre-SMA receives inputs from the prefrontal cortex(in and around the principal sulcus) and the rostral cingulate motor area [8]. In addition, the pre-SMA receives modest projections from the inferior parietal lobule [39, 73], which is connected with dorsolateral prefrontal cortex(DLPF), supplementary eye field(SEF), and frontal eye field (FEF). The pre-SMA has reciprocal connections with the SEF as well and has the projections to the anterior striatum. Accordingly, the pre-SMA is heavily interacted with the cortical areas involved in the dorsolateral prefrontal and oculomotor circuit. Furthermore, the pre-SMA is reciprocally connected with the SMA as well. Thus the pre-SMA is a pivotal cortical area to investigate the interaction of the basal ganglia-thalamocortical loops. In addition, it must be noted that the pre-SMA is the area particularly well connected with the cerebellum. The thalamic terminal fields of inputs from the cerebellar nuclei overlap considerably with the distribution of thalamic neurons projecting to the pre-SMA, but not much with that projecting to the SMA [73]. Therefore, via the thalamus, the pre-SMA has the input from the cerebellum as well. Furthermore, a wide range of the medial frontal cortex, including the pre-SMA, is known to receive the massive projections from the dopamine neurons of the mesocrtical dopaminegic system which is different from the nigrostriatal dopaminegic system discussed in Section 3.8 [33, 79]. In short, the pre-SMA has rich sensory, motor, and motivational inputs from the cortical and subcortical areas. These connectivities suggest that the pre-SMA may be involved in a kind of reinforcement learning. There are several experimental evidences to support this view. In fMRI study of 2x10 task in the 2x2 matrix display, Hikosaka et al. [29] showed that the pre-SMA is particularly active for learning of new sequences, not movements per se. In contrast, the SMA proper was active for sequential movements, not learning. In addition, Shima et al. [67] have shown that many cells of the pre-SMA showed the increase of activities when monkey had to discard motor plans, using the external cue from the environment. In the 2x5 task experiment by blocking the pre-SMA, Miyashita et al. [46, 47] has shown that the performance of the new hypersets was significantly worse in the blockade of the pre-SMA by the number of error trials to criterion than that of the new hypersets in the control condition, whereas that of the learned hypersets was not different from that of the learned hypersets in the control condition. The pre-SMA, hence, is considered as much involved in the acquisition process of sequential movements particularly in the early stage.

In the 2x10 task using 2x2 matrix display in fMRI studies [28, 56], two conditions of sequence, that is, 'color' and 'place' conditions, are tested. The experimental procedure is similar to that of the 2x5 task. The difference of 'color' condition lies in that the sensory feature of defining the sequence is not spatial information, or 'place' as in the 2x5 task, but 'color'.

It should be noted that in 'place' condition, sensory information, that is, spatial information, can be tied to the sequence of hand-movements. In other words, the memory of sequential movements by themselves can be learned in 'place' condition. In contrast, even when the subject learned the sequence in 'color' condition, the subject should convert sensory information, or 'color' information, into spatial information to move hand in each trial. In other words, it is impossible in 'color' condition to form the memory of sequential movements but possible to learn the sequence of color, that is, sensory inputs. Surprisingly, the activity transition in the pre-SMA is the same between two conditions. The transition of the activities in the pre-SMA is that, in early stage, the activities are high and are gradually decreasing till the end. This suggests that the learning in the pre-SMA depends more on sensory inputs than on motor outputs.

Then the question to be asked is whether the pre-SMA is related either to the association between a sensory input and its corresponding response, which is not transformed into an actual motor output yet, or to the learning of such an association as in sequence. Sakai and Hikosaka [55] has tested the 2x1 task with the 4x4 matrix display and the cue signal. In their experiment, the cue signal is illuminating two locations, either of which a monkey should push first and the 'go' signal is illuminating the same two locations with a different color from that of the cue signal. It was observed that the pre-SMA neural activities increased during after-error-response compared to after-correct-response, but never changed their phasic pattern response. They are remarkably different from that of rostral cingulate motor area (rCMA). Neural activities in the rCMA exhibited sustained activities between two trials, one in which the monkey made the mistake and the other in which the monkey made the correction of the action. This result suggests, as Sakai and Hikosaka [55] indicated, that not the pre-SMA but the rCMA may be responsible to keep the error signal, whereas the pre-SMA is responsible to associate sensory inputs with motor outputs and to renew the association by use of the error signal from the rCMA. Furthermore, it should be noted that the profile of neural activities in the pre-SMA in their experiment is remarkably similar to that of Shima et al. [67]. Given that this 2x1 task is not a sequential task, this result may suggest that the function of pre-SMA is related to the association aspect rather than the sequential, or temporal organization aspect in terms of sequential movements [55].

In summary, the pre-SMA seems to have a dual role: (1) associating the sensory inputs with its corresponding response that requires the transformation to produce the motor output particularly in the early stage of the acquisition process of sequential movements, with an emphasis on the association aspect rather than the aspect of the sequences, and (2) transforming the acquired result to somewhere else such as the SMA.

5.4.3 The supplementary motor area

The SMA is considered as being involved in internally-generated sequential movements, as discussed in Section 3.6. In other words, the SMA can be considered as the storage of sequential movements as one of the long-term memory (LTM) systems. The view such that the SMA has its evolutionary origin in the hippocampus and has a limbic cortical root in the anterior cingulate cortex [73] also supports the idea that the plasticity in the SMA can contribute to store the memory of sequential movements, along with experimental evidences such as Aizawa et al. [1].

The SMA and the pre-SMA has a rich reciprocal connections. As pointed out in Section 3.6, the pre-SMA not the SMA has the projection from the prefrontal cortex, whereas the SMA is richly linked with M1 and has modest inputs from the PMC [73]. Because it is difficult to consider the striatum as initiating such sequential movements as discussed in Section 3.7, the SMA may be involved in initiating these movements. The prefrontal cortex may be involved in engaging sequential movements, depending upon the sensory information, as discussed in Section 3.7. Hence, it is possible for the SMA to be influenced by the prefrontal cortex via the projection from the pre-SMA.

There, however, exist some cautions to consider the SMA by itself as the long-term memory of sequential movements. First of all, it is worth noting, according to Tanji (1994, p258.) [73], that " SMA neurons are active during the performance of the same motor task, when observed 2-6 months after the training. Surprisingly, when we kept training a monkey for 12 months in the same task, no premovement activity was found in either the left or the right SMA." Secondly, when the part of the SMA that is regarded as hand part topographically in this area, ipsilaterally or contralaterally, was blocked in the 2x5 task [46,47], the number of error trials for learned hyperset did not increase significantly than that of the learned hypersets in the control condition. These evidences are against the view such that the SMA is the storage of sequential movements. Yet, since the motor sequence task of Tanji [73] is relatively a simple task, it may be possible that the memory of their task can be transferred somewhere such as the PMC after overtraining. In addition, according to observations of Miyashita (personal communication) in her experiment [46,47], the bilateral blockade in the SMA disrupted more learned hypersets. It was also observed that the blockade of the SMA caused monkeys to mistakenly push a different position for the *second* stimulus of a set, which is very rare in the control condition and that the performance of 'half-learned' hypersets were more disrupted (Miyashita, personal communication). Therefore, it is still possible at least to some degree to consider the SMA as the storage of the sequential movements, provided that after overtraining, the stored information can be transferred to a lower level such as the premotor cortex (PMC), primary motor cortex (M1) and/or the cerebellum [37,38].

It should be mentioned, in addition, that in the blockade of the SMA in the 2x5 task, the number of error trials for the new hypersets also increased significantly than that for the new hypersets in the control condition, though the increase is less significant than in the case of the comparison between the blockade of the pre-SMA and the control condition [46,47]. This result suggests that the SMA may have a role even in the sequential movements for the new hypersets. One possibility is that the output of the pre-SMA may be through the SMA to be transformed to produce the motor output so that the blockade of the SMA interferes the performance for new hypersets.

In summary, the SMA may have a dual role: (1) storing the memory of sequential movements, provided that after overtraining, the memory is transferred to other areas, and (2) transforming the output of the pre-SMA as motor output.

5.5 Summary

The implications of behavioral findings of the 2x5 task and of neurophysiological findings on the striatum, the presupplementary motor area (pre-SMA), and the supplementary motor area (SMA), which are concluded in this chapter, are summarized in this section for readers' convenience, sacrificing the redundancy.

The behavioral findings of the 2x5 task indicated the different nature of the memory between the early and late stage of learning. The memory in the late stage of learning depends more on the memory of motor movements than that in the early stage. The memory of the correct action to the sensory input in the late stage is retrieved in the way of not reacting to the sensory input but anticipating the coming sensory input and predicting its corresponding action, depending upon information such as the sensory inputs and motor outputs before the coming sensory input.

The anterior striatum is much involved in the early stage of acquisition process of sequential movements. The posterior putamen can be considered as involved in the retrieval process of information in the late stage, but the degree of the involvement is questioned. The pre-SMA is considered as playing a dual role: (1) mapping the sensory inputs with its corresponding response in the early stage, with an emphasis on the mapping aspect rather than the aspect of the sequences, and (2) transferring this acquired result to somewhere else such as the SMA. The SMA is considered as playing a dual role: (1) storing the memory of sequential movements, provided that after overtraining, the memory is transferred to other areas, and (2) transforming the output of the pre-SMA as motor output.