The Actor-Critic Learning Is Behind the Matching Law: Matching Versus Optimal Behaviors

Yutaka Sakai
sakai@eng.tamagawa.ac.jp
Department of Intelligent Information Systems, Tamagawa University, Machida, Tokyo 194-8610, Japan

Tomoki Fukai
tfukai@brain.riken.jp
Laboratory for Neural Circuit Theory, Brain Science Institute, RIKEN, Wako, Saitama 351-0198, Japan

The ability to make a correct choice of behavior from various options is crucial for animals' survival. The neural basis for the choice of behavior has been attracting growing attention in research on biological and artificial neural systems. Alternative choice tasks with variable ratio (VR) and variable interval (VI) schedules of reinforcement have often been employed in studying decision making by animals and humans. In the VR schedule task, alternative choices are reinforced with different probabilities, and subjects learn to select the behavioral response rewarded more frequently. In the VI schedule task, alternative choices are reinforced at different average intervals independent of the choice frequencies, and the choice behavior follows the so-called matching law. The two policies appear robustly in subjects' choice of behavior, but the underlying neural mechanisms remain unknown. Here, we show that these seemingly different policies can appear from a common computational algorithm known as actor-critic learning. We present experimentally testable variations of the VI schedule in which the matching behavior gives only a suboptimal solution to decision making and show that the actor-critic system exhibits the matching behavior in the steady state of the learning even when the matching behavior is suboptimal. However, it is found that the matching behavior can earn approximately the same reward as the optimal one in many practical situations.

1 Introduction

How animals or humans organize their behavior depends crucially on the expected return that may result from their actions. It is widely considered that they attempt to maximize the obtainable reward, and a similar concept underlies several efficient algorithms in machine learning (Sutton

& Barto, 1998). Extensive studies have been conducted to clarify whether animals' or humans' behavior obeys this conceptual rule (Rachlin, Green, Kagel, & Battalio, 1976; Sakagami, Hursh, Christensen, & Silberberg, 1989; Silberberg, Thomas, & Brendzen, 1991; Herrnstein, 1997; Daw & Touretzky, 2002; Mazur, 2005). This issue, however, still remains to be clarified.

Various types of behavioral experiments have been conducted to clarify how subjects make a particular decision according to the output of their actions. Typical examples of such decision making tasks are those with variable ratio (VR) and variable interval (VI) schedules of reinforcement (Mazur, 2005). In these tasks, subjects are typically required to choose one of several alternative behavioral responses to get a reward. In the VR schedule task, alternative choices are reinforced with different ratios to the choice frequencies. A reward is given with a probability for the response. In this case, subjects learn to select the response rewarded more frequently in a sufficiently long training time.[1] It is obvious that subjects are able to maximize the reward following this policy of decision making. This observation led us to the basic concept of reinforcement learning, in which a behavior choice must be reinforced only when it results in a reward delivery. In the VI schedule task, alternative choices are reinforced at different average intervals independent of the choice frequencies. A reward is assigned to each option at a rate independent of the current choice and the assignment to other options; once the reward is assigned, it remains available until the subject takes it. Because of this persistence of rewards, subjects must scatter their choices over the alternatives to increase the reward they will obtain. Simply choosing one of the alternatives that is rewarded more frequently does not ensure a maximal reward. In fact, in the VI schedule task, the subject's choice behavior is known to obey matching law, which says that the frequency of choosing each alternative is proportional to the size of the past reward obtained by the choice (Herrnstein, 1997). Matching behavior is widely seen in many species, including humans (Davison & McCarthy, 1987), and has been shown to approximate the best probabilistic behavior (Heyman, 1979; Baum, 1981) if the amount or the strength of the reward obtainable from each alternative is equivalent, as was the case in many previous studies of animal and human behavior. The exclusive choice behavior seen in the VR task is even consistent with matching law in a trivial sense that all the options but one are never chosen and hence produce no rewards. Therefore, we can say that the matching behavior

[1] While an obvious optimal choice behavior in the VR schedule task is to keep choosing an option that is rewarded most probably, in a realistic situation, subjects exhibit probability matching, in which the choice probability of an option is proportional to the conditional probability that the option may be rewarded when chosen (Morris, Arkadir, Nevet, Vaadia, & Bergman, 2004). The probability matching might represent an exploring behavior, since choice behavior often becomes more deterministic in extensive training. What factors, however, can make such a behavioral change remain elusive.

is observed commonly in the VR and VI tasks. Learning strategies to attain matching behavior have been proposed (Herrnstein & Vaughan, 1980; Sugrue, Corrado, & Newsome, 2004). However, the neural computations underlying the matching behavior remain unknown. In addition, the behavioral implications of the matching behavior are not necessarily obvious in a generic sense (Staddon & Hinson, 1983; Mazur, 2005).

In this study, we demonstrate that the matching law can emerge from a computational algorithm known as actor-critic learning in theories of reinforcement learning (Sutton & Barto, 1998). Reinforcement learning provides a computational framework to account for a subject's choice behavior in Markov decision processes, such as the VR task. Since the VI task is not Markovian, actor-critic learning does not necessarily ensure an optimal solution to the VI task. We propose in this study that the different behavioral policies provided by the actor-critic learning system in the VR and VI tasks may represent different outcomes of the matching law. Then we investigate how the probability of choosing each alternative is changed in actor-critic learning when the design of the VI schedule task is slightly more complicated. For instance, we may assign different amounts of reward to different alternatives at different average frequencies. Alternatively, we may withdraw a once assigned reward at some frequencies. We can prove that actor-critic learning with a sufficiently small learning rate always exhibits matching behavior in the steady state of reinforcement learning. Interestingly, however, in many cases, the matching behavior is no longer optimal. We argue possible implications of these results for interpreting the observed animal's behavior.

2 Reinforcement Schedules and the Matching Law

For simplicity in task design, we consider an alternative choice task consisting of a trial sequence with discrete time steps; free-response tasks on continuous time have been used in many previous studies concerning the matching law. At each time step, a subject is required to choose one of n available options for which rewards are set independently of the subject's choice behavior. In a variable interval (VI) schedule, a reward is assigned to option a at a rate of λ_a ($a = 1, 2, \dots, n$) independent of the assignment to other options. Once the reward is assigned, it remains available, and no additional reward is assigned until it is taken by a subject. In a variable ratio (VR) schedule, every choice of option a may result in a reward with a conditional probability of λ_a. As in many previous behavioral studies, here the amount of reward obtainable by a single choice behavior is identical for all alternatives. The VR schedule may resemble the situations that many carnivorous animals encounter during hunting. For instance, a cheetah may decide which prey, a zebra or a gazelle, she should chase according to the success of her past hunting. The VI schedule may imitate the situations that herbivorous animals meet in their foraging behavior: once they visit this grass, it will become available again only after a certain period.

The likelihood of being rewarded by choosing an option remains constant in a VR schedule, while in a VI schedule, it increases with the time that has passed from the last choice of that option. The probability that a reward is assigned to option a by T steps after its previous choice is given as $\Pr(\text{assigned to } a|T) = 1 - (1 - \lambda_a)^T$. If a subject makes a random choice of alternatives with choice probabilities P_a, the average income from choosing option a (i.e., action a) can be derived as

$$R_a = \rho \sum_{T=1}^{\infty} \Pr(\text{assigned to } a|T)(1 - P_a)^{T-1} P_a^2 = \frac{\rho\lambda_a P_a}{1 - (1 - \lambda_a)(1 - P_a)},$$

where ρ is the amount of reward obtainable in a single choice. Note that R_a represents the average taken over the all trials rather than the trials in which option a is actually chosen by a subject. Hence, the summation over all possible choices $\sum_a R_a$ represents the average reward per a trial. The best probabilistic behavior is defined by the set of choice probabilities $\{P_a^*\}$ that maximize $\sum_a R_a$ under the constraint $\sum_a P_a = 1$. This maximization can be done using the Lagrange multiplier method; the best choice probabilities are derived as

$$P_a^* = \frac{\lambda_a/(1 - \lambda_a)}{\sum_{a'=1}^{n} \lambda_{a'}/(1 - \lambda_{a'})}, \tag{2.1}$$

which obey the matching law

$$P_a^* = R_a \bigg/ \sum_{a'=1}^{n} R_{a'} \; . \tag{2.2}$$

Thus, we find that the best choice probabilities exactly obey the matching law in the discrete time VI task with equal amounts of reward set for single choices of any option. This is consistent with the results in the continuous-time VI task (Heyman, 1979; Baum, 1981). Note that the probabilistic choice behavior, equation 2.1, is best in a limited situation that allows a subject to make a random choice on each trial with a set of constant choice probabilities. It has been known that truly optimal behavior in a VI schedule task is in general given by a perfectly periodic choice (Houston & McNamara, 1981). However, random or probabilistic choice behaviors have been observed in a wide range of VI schedule tasks (Herrnstein, 1997; Mazur, 2005; Sugrue et al., 2004), and we focus on probabilistic choice behaviors in this letter. We use the term *best* instead of *optimal* for a behavior that is optimal among probabilistic choice behaviors, reserving the term *optimal* for truly optimal behavior.

In a VR task, the optimal behavior is obviously to keep choosing the option that is rewarded at the highest rate. This behavior looks quite different from the best probabilistic choice in a VI task. Nevertheless, the complete bias toward a single option in the VR task is consistent with the matching law given by equation 2.2, since the options other than the optimal one are never chosen by the subject and hence produce no reward. It can therefore be said that the best choice probabilities satisfy the matching law in both VI and VR tasks.

3 Actor-Critic Method Without State Variables

Animals are known to exhibit the best probabilistic behavior in both VR and VI schedule tasks (Davison & McCarthy, 1987; Herrnstein, 1997; Mazur, 2005). How is the brain capable of developing the best choice probabilities in the seemingly different tasks? Is there a common computational algorithm to achieve the best choice probabilities in both cases? Here, we demonstrate that the actor-critic learning, a well-known algorithm of reinforcement learning in engineering and robotics, can account for the best choice probabilities in both VR and VI tasks.

In actor-critic learning, the "critic" predicts the rewards obtainable in the future, and the "actor" changes the system's internal states and selects an action that presumably leads to an optimized future reward according to the prediction (Sutton & Barto, 1998). For the time being, we consider actor-critic learning without state variables, since no explicitly varying state seems to exist in the alternative choice tasks. More general situations in which the state is a dynamical variable observable by a subject are discussed in appendix B.

The actor chooses an action from n alternatives according to the current choice probabilities $\{p_a\}$, where every choice is made independent of others. Note that the previous $\{P_a\}$ represents the long-term average of $\{p_a\}$. The choice probabilities are determined by the policy parameters $\{q_a\}$ as

$$p_a = f(q_a) \Bigg/ \sum_{a'=1}^{n} f(q_{a'}) ,$$

where $f(\cdot)$ is a positive, monotonically increasing function (e.g., an exponential function). A reward r_t may be given as a result of the action a_t chosen at time step t. In the VI and VR schedules introduced previously, r_t can be either ρ or zero. The critic updates the estimation of the average reward obtained by all the options, V, which decays with a time constant of $1/\alpha$ if no reward is earned by the current choice:

$$V + \Delta V \to V, \quad \Delta V = \alpha(r_t - V). \tag{3.1}$$

There exists no state variable that may change explicitly with time and may influence the estimation of V. Therefore, V always estimates the reward expected at the current time step in the environment that does not change its probabilistic structure with time.

The current choice a_t made by the actor is evaluated by the critic in terms of the error in the reward estimation, $r_t - V$. If the error is positive, that is, the obtained reward is greater than what was expected, the actor increases the probability of choosing the current action by updating the corresponding policy parameter as

$$q_{a_t} + \Delta q_{a_t} \to q_{a_t}, \quad \Delta q_{a_t} = \alpha(r_t - V). \tag{3.2}$$

Note that actor-critic learning updates only the policy parameter corresponding to the action selected in the current trial. The dopamine neurons in the basal ganglia were shown to provide the error signal, possibly serving as the critic during cognitive motor learning. The motor-related frontal cortices and the striatum, an input nucleus of the basal ganglia, are considered to be essential for motor selection, thus operating as the actor (Houk, Davis, & Beiser, 1994; Doya, 2000; Dayan & Balleine, 2002; Tanaka et al., 2004; Schultz, 2004; Haruno et al., 2004). These cortical and subcortical neural systems are the candidate loci of decision making in animals and humans.

Figure 1 demonstrates how an actor-critic system chooses alternatives ($n = 2$) in simulations of the VR and VI tasks. The rates of the reward assignment were set at different values for the two actions. The results proved that the fraction of the choice probabilities, that is, p_1 (solid curves), gradually approaches the best values (dashed lines), which are consistent with the matching law, in both tasks.

We show below why the matching behavior can be obtained by actor-critic learning. To include all the policy parameters explicitly, we rewrite the updating rule 3.2 as

$$\Delta q_a = \alpha(r_t - V)\delta_{aa_t}, \quad (a = 1, \dots, n),$$

where δ_{ij} is 1 if $i = j$, or 0 otherwise. The long-term averages of these policy parameters are given as

$$\langle \Delta q_a \rangle = \alpha \langle (r_t - V)\delta_{aa_t} \rangle = \alpha(\langle r_t \delta_{aa_t} \rangle - \langle V \delta_{aa_t} \rangle),$$

where the bracket $\langle \cdot \rangle$ means a long-term average over τ trials. The first term, $\langle r_t \delta_{aa_t} \rangle$, represents the average income obtained by choosing action a, and hence $\langle r_t \delta_{aa_t} \rangle = R_a$. In the second term, $\langle V \delta_{aa_t} \rangle$, V generally depends on the choice frequencies of option a and others. However, if the learning rate α in equation 3.1 is sufficiently small compared with the reciprocal of the averaging span τ, that is, $\alpha \ll 1/\tau$, then the reward estimation V can be regarded as constant during the averaging span τ (see appendix B). In this

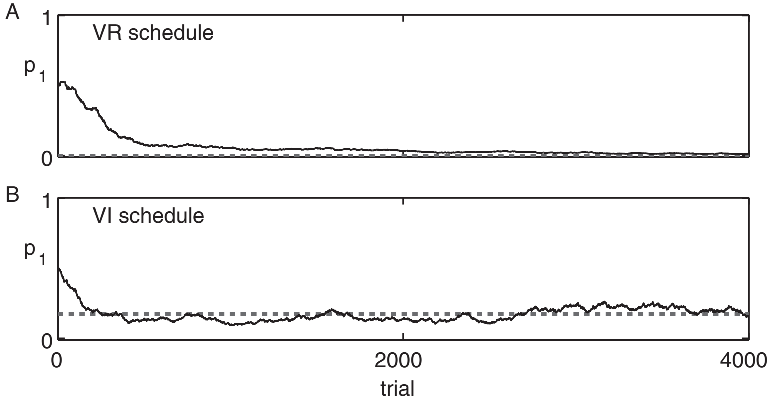

Figure 1: The choice behavior of an actor-critic system without state variables in VR and VI schedules ($n = 2$). The system achieves the best probabilistic behavior in either VR (A) or VI (B) schedule task, in which different reward rates were assigned to the alternatives: $(\lambda_1, \lambda_2) = (0.05, 0.2)$. See appendix C for other conditions of numerical simulations. Solid curves stand for the time courses of the fraction of the current choice probabilities, that is, p_1, determined by the policy parameters q_1 and q_2. Dashed lines stand for the fraction of the best choice probabilities, that is, P_1^*.

case, $\langle V\delta_{aa_t}\rangle$ can be factorized into $\langle V\rangle\langle\delta_{aa_t}\rangle$, and the average $\langle V\rangle$ coincides with the actual reward obtained in a steady state since the estimation error should vanish in that state. Therefore,

$$\langle V\rangle = \langle r_t\rangle = \sum_{a=1}^{n} R_a.$$

The average $\langle\delta_{aa_t}\rangle$ represents the choice frequency of individual actions, that is, $\langle\delta_{aa_t}\rangle = P_a$, which coincides with the long-term average of the instantaneous choice probability, $P_a = \langle p_a\rangle$. Thus, we obtain the slow dynamics of the policy parameters averaged over τ trials,

$$\langle\Delta q_a\rangle \simeq \alpha\left(R_a - P_a\sum_{a'=1}^{n} R_{a'}\right),$$

and $\langle\Delta q_a\rangle \simeq 0$ implies the matching law in the steady state:

$$P_a = R_a \Bigg/ \sum_{a'=1}^{n} R_{a'}\ .$$

The result shows that actor-critic learning always exhibits matching behavior in the steady state, as far as the learning rate α can be regarded as sufficiently small. We note that the above relationship could be derived regardless of the reinforcement schedule. The result does not show that the steady state defined by $\langle \Delta q_a \rangle = 0$ is attained in an arbitrary choice task. However, as far as the policy parameters stay in a finite range, the steady state $\langle \Delta q_a \rangle \simeq 0$ should be attained if the averaging span τ is sufficiently large. Divergence of a policy parameter usually leads to an exclusive choice behavior, which also satisfies the matching law. Thus, it is suggested that the actor-critic learning with a sufficiently small learning rate leads to the matching behavior in most practical choice tasks, regardless of whether the behavior is the best for a specific reinforcement schedule employed in the task.

4 VI Schedule Task with Different Reward Magnitudes

Unlike the previous cases, we can find a variety of reinforcement schedules in which the matching behavior is not longer the best. The simplest examples include an extended VI schedule task that assigns different amounts of the same reward to different options: $\rho_a \neq \rho_b$ for $a \neq b$, where ρ_a denotes the amount of reward for action a. In this case, the best choice probabilities and the average fractional incomes satisfy (see appendix A)

$$P_a^* = \frac{R_a^*/\sqrt{\rho_a}}{\sum_{a'} R_{a'}^*/\sqrt{\rho_{a'}}}.$$

This relationship deviates from the matching law by a scaling factor of $\sqrt{\rho_a}$, meaning that the matching behavior does not maximize the average reward, when the amounts of reward per choice differ for different options.

Whether actor-critic learning leads to the matching behavior was numerically tested in Figure 2E for successive four blocks of the VI schedule task with different combinations of the rates and the amounts of the reward assigned to alternatives ($n = 2$). Depending on the values of these parameters, there is a unique solution (see equation A.4) representing the matching law besides two trivial solutions representing exclusive choices: $(P_1, P_2) = (1, 0)$ and $(0, 1)$. The solid and dashed lines in each block indicate the best and the nontrivial matching choice probabilities, respectively. We find that the actor quickly learns the choice probabilities representing the matching behavior (the black curve) in all the trial blocks. Results of the simulations proved that the actor-critic learning produces the matching, but not the best, behavior in the VI task with unequal rewards. Furthermore, the learning produced no fifty-fifty random choice behavior, although it ensures the average return that is almost equivalent to the best one (the first and second blocks). We note that the fraction of the current choice probabilities of the actor follows

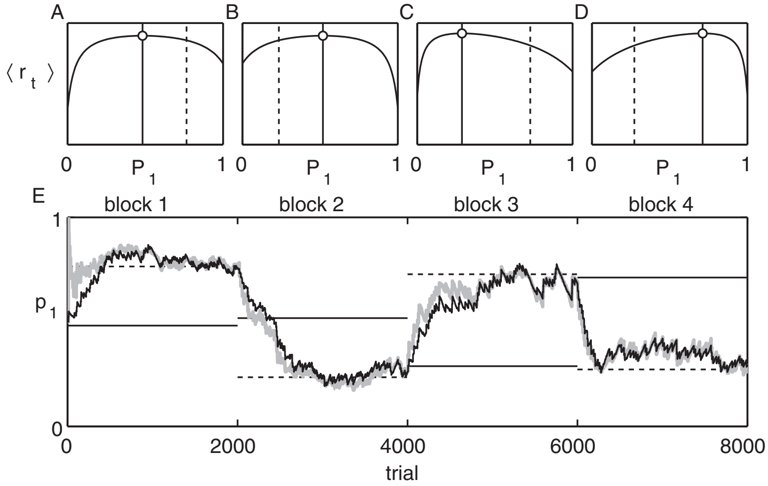

Figure 2: The matching behavior of the actor-critic system without state variables in the VI schedule task with different reward magnitudes ($n = 2$). Different reward rates and reward amounts were assigned to the alternatives in successive four blocks: $(\lambda_1, \lambda_2, \rho_1, \rho_2) = (0.05, 0.15, 2, 0.3)$, $(0.15, 0.05, 0.3, 2)$, $(0.03, 0.3, 0.3, 2)$, $(0.3, 0.03, 2, 0.3)$. Other conditions are summarized in appendix C. (A–D) The average return $\langle r_t \rangle$ is shown as a function of the choice probability P_1 for each trial block. The solid and dashed vertical lines indicate the fraction of the best and matching choice probabilities calculated analytically in equations A.3 and A.4, respectively. (E) The time courses of the fraction of the current choice probabilities of the actor; p_1 (black curve) and the fraction of the locally averaged incomes, $\hat{R}_1/(\hat{R}_1 + \hat{R}_2)$ are shown (gray curve) together with the fractions of the best (solid line) and the matching (dashed line) choice probabilities. The system always displays the matching, rather than the best, behavior in all four blocks.

the fraction of the local incomes averaged over relatively short intervals (the gray curve).

The magnitude of reward may in general differ for different options in many feasible situations in nature. It is, however, unclear how animals scale the subjective value of a reward based on its physical strength. Many factors seem to affect the evaluation of the subjective value. To avoid these complications, we give another extension of the VI schedule in which quantitative comparisons between experimental and theoretical results will be easier.

5 Competitive Foraging Task

Here, we introduce a task in which the best choice probabilities deviate from those given by the matching law, even if an identical reward is assigned to

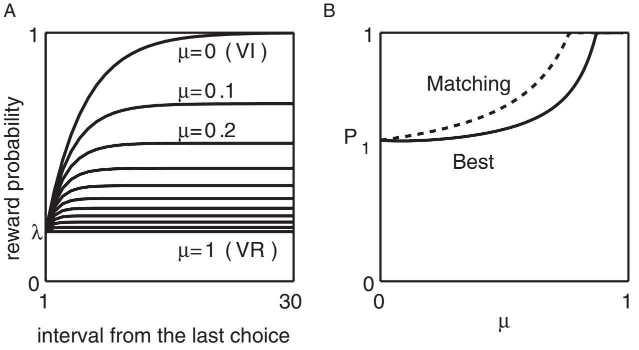

Figure 3: The competitive foraging task, in which extended VI schedules with stochastic withdrawal at a constant rate are assigned to the alternatives. (A) The probability that a reward has been assigned to an option depends on how long the subject may wait for the next choice of that option. The withdrawal rate changed by a 0.1 step interpolates choice tasks between the VI schedule ($\mu = 0$) and the VR schedule ($\mu = 1$). (B) The fractions of the best choice probabilities (solid curve) and the choice probabilities of the matching behavior (dashed curve) are shown as functions of the withdrawal rate for the schedule specified by $(\lambda_1, \lambda_2) = (0.2, 0.16)$.

every reinforcer. As in the VI schedule, identical rewards are stochastically and independently set for each option at a constant rate (λ_a for option a). However, the reward set for option a may be withdrawn at a constant rate of μ_a, if the reward has not been taken out. Different values of the withdrawal rate μ generate various mixtures of the VI and VR schedules between pure VI ($\mu = 0$) and pure VR ($\mu = 1$). The combination of the withdrawal rates, $\mu_1 = 0$ and $\mu_2 = 1$, corresponds to the concurrent VI and VR schedules, in which an option is reinforced in a VI schedule and the other in a VR schedule (Herrnstein & Heyman, 1979; Herrnstein, 1997). These schedules may better imitate the natural environment for foraging by herbivorous animals, because food may sometimes be intercepted by their competitors. Hence, we call this extended schedule task the competitive foraging task. Note that this task contains no dynamic interactions with competitors. It is said that this is a foraging task with stationarily stochastic competitors that do not change their behavior according to that of their competitors.

As in the VI schedule without the withdrawal, the likelihood of reward assignment to an option increases monotonically with the time passage from the last choice of that option (see Figure 3A). The likelihood, however, saturates to an asymptotic value that decreases with the increases in μ. The profile of the likelihood changes smoothly from the VR to VI schedule. Figure 3B displays how the best choice probabilities (solid line) and the choice probabilities satisfying the matching law (dashed line) depend on μ

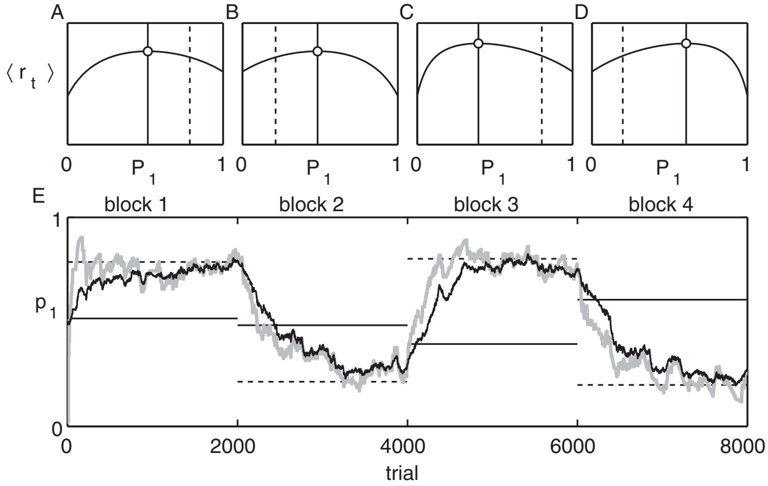

Figure 4: The matching behavior of the actor-critic system without state variables in the competitive foraging task ($n = 2$). In the four blocks, different reward rates and withdrawal rates were assigned to the alternatives: $(\lambda_1, \lambda_2, \mu_1, \mu_2) = (0.09, 0.06, 0.15, 0.4)$, $(0.06, 0.09, 0.4, 0.15)$, $(0.09, 0.06, 0.01, 0.4)$, $(0.06, 0.09, 0.4, 0.01)$. Other conditions are summarized in appendix C. (A–D) The average return $\langle r_t \rangle$ is shown as a function of choice probability P_1 for each trial block. Solid and dashed vertical lines indicate the fraction of the best and matching choice probabilities, calculated analytically in equations A.3 and A.4, respectively. (E) The time courses of the fractions of the current choice probabilities (black curve) and the locally averaged incomes (gray curve) are shown together with the fractions of the best (solid line) and the matching (dashed line) choice probabilities.

for $n = 2$. We can see that the choice probabilities differ from one another in the intermediate range $0 < \mu < 1$ but coincide at both ends representing the VI ($\mu = 0$) and VR ($\mu = 1$) schedules. In fact, the relationship between the best choice probabilities and the average incomes is explicitly given as (see appendix A)

$$P_a^* \propto \sqrt{\frac{\lambda_a + \mu_a - \mu_a \lambda_a}{\lambda_a}}\, R_a^*,$$

which differs from what is predicted by the matching law if $\mu_a \neq 0$ for some a. Figure 4E shows how the actor-critic learning system behaves in simulations of four successive blocks of the competitive foraging task with different combinations of the assignment and withdrawal rates. At given values of these parameters, there is a unique nontrivial solution representing the matching law, equation A.4. As in the conventional VI task, the fraction of the choice probabilities in the actor-critic system (black curve) follows the fraction of the locally averaged incomes (gray curve), and the choice behavior displays the nontrivial matching law (dashed line) rather than the best one (solid line). We find that the actor-critic learning again leads to the matching behavior rather than to the best one.

6 Implications of the Matching Behavior for Optimizing Rewards

We have proved that the actor-critic learning with a sufficiently small learning rate always produces the matching behavior in the steady state, regardless of whether it is the best. In addition, we have shown that the matching behavior earns an amount of rewards that is approximately equivalent to the maximum amount. We evaluated how much reward subjects may lose if they obey behavioral policies other than an optimal one. As noted previously, the optimal behavior in a VI schedule task is a periodic choice if behaviors other than random choices are allowed (Houston & McNamara, 1981). This also holds for the competitive foraging task, because the time from the last choices of the options provides sufficient information about the reward expectation (see Figure 3A). In this case, the optimal current choice at a certain pattern of the passaged time is determined by the pattern, and hence, the choice behavior becomes periodic.

Let us represent the parameter space of the competitive foraging task by a five-dimensional unit cube spanned by $(\lambda_1, \lambda_2, \mu_1, \mu_2, \rho_1/(\rho_1 + \rho_2))$. Each parameter takes its value in the range [0, 1]. In each set of task parameters, we numerically calculated the maximum rewards $\langle r_t^* \rangle$ obtainable with the optimal periodic choice and the rewards $\langle r_t \rangle$ obtainable with various choice behaviors: the best probabilistic choice, the matching choice, the alternate choice, the 50-50 random choice, and the worst probabilistic choice behaviors. By sweeping the entire space of task parameters, we can obtain the fraction of the parameter region in which each choice behavior can earn more than a ratio x of the maximum reward: $\langle r_t \rangle / \langle r_t^* \rangle > x$ (see Figure 5). We found that the matching behavior can earn more than 90% of the maximum reward in 93% of all possible task schedules of the competitive foraging task (filled circles). In 76% of all schedules, the matching behavior can earn a reward equivalent to the maximum amount (empty circle). The alternate choice and the 50-50 random choice can earn up to 50% of the maximum reward in all possible schedules. However, if we increase the acceptable range up to 90% of the maximum reward, the percentages of the successful tasks reduce to only 29% (the square) and 15% (the cross) in the alternate and 50-50 random choices, respectively. These results have proved that the matching behavior gives suboptimal behavior, which can earn rewards remarkably close to the maximum one in most schedules of the competitive foraging task.

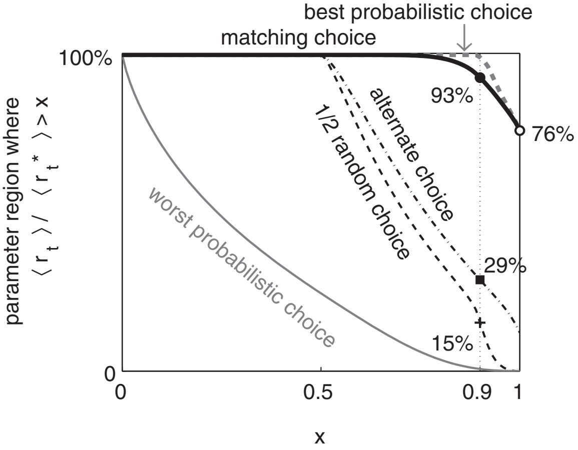

Figure 5: The comparison of various choice behaviors in the competitive foraging task. The curves represent the relative volume of the subspace of parameters in which the ratios of the average rewards to the maximum reward are greater than x, that is, $\langle r_t \rangle / \langle r_t^* \rangle > x$, for x in the best probabilistic (gray dashed), matching (black solid), regularly alternate (dot-dashed), 50-50 random (dashed), and worst probabilistic (gray solid) choice policies. Here, $\langle r_t \rangle$ is the expected reward in each behavior, and $\langle r_t^* \rangle$ is the maximum amount of reward obtainable by the optimal periodic choice. The parameter space is represented by a five-dimensional unit cube $(\lambda_1, \lambda_2, \mu_1, \mu_2, \rho_1/(\rho_1 + \rho_2))$. The marks at $x = 0.9$ of the maximum show the relative volumes occupied by the tasks in which individual choice policies can earn more than 90% of the maximum gain. The matching behavior shows a 100% gain in 76% of the competitive foraging tasks.

7 Discussion

We have proved that actor-critic learning with a sufficiently small learning rate always exhibits matching behavior in an arbitrary alternative choice task when the learning attains a steady state. We also have demonstrated that the learning develops the matching behavior with a practical value of the learning rate in several probabilistic choice tasks, including the VI schedule. It was previously shown that matching behavior approximates the best probabilistic behavior to maximize the long-term average of reward in continuous-time versions of the VI schedule task (Heyman, 1979; Baum, 1981), and here we have shown that matching behavior provides the best choice probabilities in discrete-time versions of the VI schedule task if the amount of the reward that is obtainable in single choices is identical for all alternatives. This experimental setting has been used in many behavioral studies that have demonstrated the matching law. Consequently, results of these studies were consistent with the hypothesis that the animal's behavior is reinforced to optimize the choice probabilities to maximize the reward and that the matching behavior is a consequence of this optimization process.

However, the matching behavior is not necessarily the best among the choice behaviors in some alternative choice tasks. Some previous work has shown examples in which the best choice probabilities do not satisfy the matching law. A simple example is the case where the amount of the reward obtained in a single choice is not identical for different alternatives. Consistent with the results of actor-critic learning, the matching behavior of animals was actually observed in this type of experimental setting (Baum & Rachlin, 1969; Heyman & Monaghan, 1994). However, the internal relationship between the animal's subjective value of a reward, on which the animal's decision was based, and its physical strength, on which the qualitative design of experiments was based, is unknown. Therefore, the results may not provide clear evidence that the animals exhibit matching behavior rather than the best probabilistic behavior. Another example can be found in the concurrent VI and VR schedule tasks (Herrnstein & Heyman, 1979; Herrnstein, 1997). In this case, the best choice probabilities deviate from the matching ones, even if the identical reward is assigned to different alternatives. Again, matching behavior was observed in this type of task (Herrnstein & Heyman, 1979; Vyse & Belke, 1992; Savastano & Fantino, 1994), although controversial results have been reported (Sakagami et al., 1989). The issue of matching or maximizing was examined in other types of choice tasks (Mazur, 1981; DeCarlo, 1985; Jacobs & Hackenberg, 1996), but the validity of the matching behavior remains unclear (Mazur, 2005). These experiments are certainly more complicated than those discussed in this letter due to their basic design using continuous time. In addition, these experiments introduced "changeover delays" between consecutive choices to avoid an alternate choice behavior (Stubbs, Pliskoff, & Reid, 1977; Herrnstein, 1997). These complex experimental settings make comparisons of results in the different experiments difficult.

In this study, we have extended the VI schedule by introducing the withdrawal of rewards. Such a task may imitate a general situation that herbivorous animals encounter in foraging behavior, because their food may sometimes be intercepted by their competitors. This competitive foraging task includes the concurrent VI schedules, the concurrent VR schedules, and the concurrent VI and VR schedules as extreme cases. In the discrete-time version of the competitive foraging task, the best choice probabilities deviate significantly from the matching ones in a wide range of task parameters (e.g., see Figure 4). Furthermore, we showed in Figure 5 that the alternate choice is inferior to the matching choice in many cases. Since subjects presumably avoid the alternate choice behavior in such a situation, the changeover delays may play no active role in many cases of competitive foraging. A more complicated task was examined by Montague, Dayan, and Sejnowski (1996). In their task, each choice was reinforced at a rate that was determined as an arbitrary function of the locally averaged frequency of the past choice. This task is of particular interest since an amount of obtainable reward does not depend on the local order of choosing different options. Only the total frequencies of choosing the individual options are meaningful. It can be shown that the expected reward for the optimal and matching behaviors can be controlled independently in this task. Therefore, in some cases, the matching behavior loses much reward. It is of extreme interest to test which behavior, matching or maximizing rewards, subjects may exhibit in these tasks.

Herrnstein, 1997 and his colleague studied how the matching law might be achieved in the steady state by the dynamism of behavior and proposed "melioration." In melioration, behavior should shift toward a higher local rate of reinforcement until the rates are balanced at an equilibrium point representing the matching law: $R_1/P_1 = R_2/P_2 = \cdots = R_n/P_n$ (Herrnstein, 1997; Herrnstein & Vaughan, 1980). This algorithm consists of the reward evaluation and the action choice, and hence, it is similar to actor-critic learnings (Daw & Touretzky, 2001). A crucial difference, as shown in this study, is that actor-critic learning guarantees the matching law in the steady state without evaluating the average incomes separately for the individual options. In addition, we do not explicitly require the variable V to estimate the sum of the average incomes, because we can determine the value of V from the summation of the policy parameters. Therefore, the dynamics of the actor-critic system can be described solely by the policy parameters, $\Delta q_a = \alpha(r_t - \sum_a q_a)\delta_{aa_t}$. This formulation is similar to the "direct actor," in which $\Delta q_a = \alpha(\delta_{aa_t} - p_a)r_t$ (Dayan & Abbott, 2001), The "local matching law" (Sugrue et al., 2004), by definition attains the matching behavior, because the choice probabilities are directly determined as the fractions of locally averaged incomes. The method, however, requires some mechanisms to avoid staying at singular points, at which only a single option is repeatedly chosen, as in the optimal behavior of the VR task. The different mathematical models of behavior provide different but partially similar results of decision-related computations. The different theoretical predictions can be tested by trial-by-trial fittings between theoretical and behavioral results (Sugrue et al., 2004; Samejima, Ueda, Doya, & Kimura, 2005) including transient effects (Gallistel, Mark, King, & Latham, 2001) and temporal structures (Staddon & Hinson, 1983).

If matching behavior is not necessarily optimal, is there any reason for subjects to obey such a behavior? As we have shown, the same actor-critic learning allows subjects to develop an optimal behavior in the VR task and a near-optimal behavior in the VI task and its extended versions. We conjecture that the brain may save a limited resource of neural computations by utilizing a specific algorithm that can provide at least a suboptimal solution quite efficiently for various decision tasks.

The neural mechanism of decision making has been studied extensively in various behavioral tasks (Barraclough, Conroy, & Lee, 2004; Breiter, Aharon, Kahneman, Dale, & Shizgal, 2001; Knutson, Adams, Fong,

& Hommer, 2001; Montague & Berns, 2002). Accumulating evidence suggests that the errors in predicting future rewards are signaled by dopamine neurons in the ventral tegmental area and the basal ganglia (Montague et al., 1996; Schultz, Dayan, & Montague, 1997; McClure, Berns, & Montague, 2003). It is likely that these nuclei play a role of critic in the behavioral learning. By contrast, activity of prefrontal or parietal neurons has been shown to represent the probability that a sensory stimulus might appear the probability that a stimulus might predict a future reward (Platt & Glimcher, 1999). These activities might represent the value function of a sensory stimulus, and synaptic mechanisms to generate such neural activity have been suggested by modeling studies (Seung, 2003; Wang, 2002). The results of these experimental and theoretical studies must be combined into theories of decision-making behavior; our results provide a useful platform to study these theories.

Appendix A: Competitive Foraging Task

By introducing the stochastic withdrawal of assigned rewards, we get an extended VI schedule that incorporates the VR schedule. Suppose that the rewards are independently assigned to and withdrawn from each option at rates λ_a and μ_a, respectively. We introduce dynamical variables that represent the current status of the rewards set for individual options. These states are updated at time step t as follows:

1. Assignment of a reward to an option turns its state *assigned*.
2. Reward r_{a_t} can be given to the current choice a_t if the corresponding state is *assigned* and the state is changed to *unassigned*.
3. Withdrawal of a reward from an option turns its state *unassigned*.

Thus, $\mu_a = 0$ corresponds to the VI schedule and $\mu_a = 1$ to the VR schedule.

Let T and S denote the time steps from the last choice of a and the last of withdrawal from a, respectively. The conditional probability that the state of option a is *assigned* for given T and S is given as

$$\Pr(\text{'}\textit{assigned}\text{'}\ |T, S) = 1 - (1 - \lambda_a)^{\min\{S,T\}}.$$

The probability that the state is *assigned* at step T is derived by averaging the above expression over S:

$$\begin{aligned}\Pr(\textit{assigned}|T) &= \sum_{S=1}^{\infty} \Pr(\textit{assigned}|T, S) \Pr(S) \\ &= \sum_{S=1}^{\infty} \left(1 - (1 - \lambda_a)^{\min\{S,T\}}\right) (1 - \mu_a)^{S-1} \mu_a \\ &= \frac{\lambda_a \left(1 - (1 - \lambda_a)^T (1 - \mu_a)^T\right)}{1 - (1 - \lambda_a)(1 - \mu_a)} . \end{aligned} \tag{A.1}$$

If the subject's step-by-step choice is made independently with constant probability P_a, then the probability that the state is *assigned* when option a is chosen at an arbitrary time step can be derived as

$$\begin{aligned}\Pr(\textit{assigned}) &= \sum_{T=1}^{\infty} \Pr(\textit{assigned}|T)(1 - P_a)^{T-1} P_a \\ &= \frac{\lambda_a}{1 - (1 - \lambda_a)(1 - \mu_a)(1 - P_a)} .\end{aligned}$$

Therefore, the average fractional income from option a, R_a, is expressed as

$$R_a = \rho_a \Pr(\textit{'assigned'})P_a = \frac{\rho_a \lambda_a P_a}{1 - (1 - \lambda_a)(1 - \mu_a)(1 - P_a)} , \tag{A.2}$$

where ρ_a represents the amount of reward obtainable from option a in a single trial. Note that R_a is the average income defined over all trials rather than the trials restricted to the choices of option a.

The best choice probabilities $\{P_a^*\}$ can be obtained by solving the following optimization problem:

$$\max_{\{P_a\}} \sum_{a=1}^{n} R_a, \qquad \text{under a constraint } \sum_{a=1}^{n} P_a = 1, \quad P_a \geq 0.$$

Lagrange's method of multipliers gives candidates for the best choice probabilities $\{P_a^*\}$ as

$$P_a^* = \frac{C\sqrt{\rho_a \lambda_a(\lambda_a + \mu_a - \mu_a \lambda_a)} - (\lambda_a + \mu_a - \mu_a \lambda_a)}{(1 - \mu_a)(1 - \lambda_a)} \quad \text{or} \ 0, \tag{A.3}$$

where C is a constant determined by $\sum_a P_a = 1$. Due to the convexity of the total average return $\sum_a R_a$, if a solution that satisfies $P_a^* > 0$ for arbitrary a exists, it gives the global maximum. Otherwise a solution that admits $P_a^* = 0$ for some options must be selected to maximize the average return.

In the VR schedule task ($\mu_1 = \mu_2 = \cdots = \mu_n = 1$), equation A.3 gives $C\sqrt{\rho_a \lambda_a} = 1$ or $P_a^* = 0$. If $\rho_a \lambda_a$, the long-term average of the reward assigned to an option, is not identical for all options, the best choice is given by $P_a^* = 1$ for $a = \arg\max_a \rho_a \lambda_a$ and $P_a^* = 0$ otherwise. Substituting equation A.3 for the denominator of equation A.2 leads to the relationship between the best choice probabilities and the average fractional incomes:

$$R_a^* = \frac{1}{C}\sqrt{\frac{\rho_a \lambda_a}{\lambda_a + \mu_a - \mu_a \lambda_a}}\ P_a^*.$$

Thus, the best choice probabilities generally differ from the prediction of the matching behavior.

In the VI schedule task ($\mu_1 = \mu_2 = \cdots = \mu_n = 0$) with different amounts of reward, the relationship between the best choice probability and the average fractional income is scaled by $\sqrt{\rho_a}$ and hence differs from the matching law:

$$R_a^* \propto \sqrt{\rho_a}\, P_a^*.$$

For an equal amount of reward, $\rho_1 = \rho_2 = \cdots = \rho_n$, the matching law holds. In this case, a solution that satisfies $P_a^* > 0$ for arbitrary a always exists, so the best choice probabilities are determined as

$$P_a^* \propto \frac{\lambda_a}{1 - \lambda_a} \propto R_a^*.$$

As is shown above, the best choice probabilities do not generally satisfy the matching law. The choice probabilities satisfying the matching law are derived from the proportional relationship between P_a and R_a in equation A.2,

$$P_a = \frac{C\rho_a\lambda_a - \lambda_a - \mu_a + \lambda_a\mu_a}{(1-\lambda_a)(1-\mu_a)} \quad \text{or} \quad 0, \tag{A.4}$$

where C is a constant determined by $\sum_a P_a = 1$. The matching law provides multiple solutions corresponding to possible sets of options for which $P_a = 0$ if the values of $\{P_a\}$ determined by equation A.4 are all nonnegative. The maximum number of solutions is $2^n - 1$. In the case of two alternatives, $n = 2$, a nontrivial solution other than exclusive choices $(P_1, P_2) = (0, 1)$ and $(1, 0)$ can be uniquely determined if it exists.

Appendix B: The Matching Law with State Transitions

Here, we extend our analysis to actor-critic learning with state variables and show that the matching law holds if the law is applied to the choices made in the same state. State variables may represent any information available for subjects' decision making, such as sensory inputs or the history of actions and rewards. The likelihood of the future return is often employed as a state variable. In this case, the learner may use a state transition itself as a reinforcer without an actual reward, so the matching law must be extended to include such an indirect reinforcer. This extension is straightforward. Let us consider the future return by a choice made in the state $s_t = s$. The average income obtained in the term after choosing option a in state s is given as

$$R_a(s) \equiv \left\langle \left(r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots\right) \delta_{aa_t} | s_t = s \right\rangle,$$

where γ is a factor that discounts rewards in the remote future $(0 < \gamma < 1)$, and $\langle\cdot|\cdot\rangle$ represents a conditional averaging operation. The average income $R_a(s)$ represents the correlation between the choice a in state s and a weighted sum of the incomes after the choice. Then the matching law can be interpreted as

$$P_a(s) = R_a(s) \Bigg/ \sum_{a'} R_{a'}(s) \ , \tag{B.1}$$

in terms of the likelihood of the future return, where $P_a(s)$ denotes the choice probability of action a in state s. We wish to prove this relationship in the steady state of the actor-critic learning with state variables.

In this case, the critic must make reasonable predictions of the likelihood of the future reward. Several methods have been proposed for this online estimation. Here, we adopt so-called temporal difference (TD) learning:

$$\Delta V(s_t) = \alpha\big(r_t + \gamma V(s_{t+1}) - V(s_t)\big).$$

The function $V(s)$ is called as the state value function and represents how much reward (discounted by a timescale of $1/\gamma$) is expected to follow state s.

The choice probability $p_a(s)$ of the actor in state s is described by the policy parameters $q_a(s)$, which are updated in terms of the state value function as

$$\begin{aligned} p_a(s) &= f(q_a(s)) \Bigg/ \sum_{a'} f(q_{a'}(s)) \ , \\ \Delta q_a(s_t) &= \alpha(r_t + \gamma V(s_{t+1}) - V(s_t))\delta_{aa_t}. \end{aligned}$$

Since $V(s) = \sum_a q_a(s)$, if the relation is satisfied by the initial conditions, the dynamics are described only by the policy parameters,

$$\Delta q_a(s_t) = \alpha\left(r_t + \gamma \sum_{a'} q_{a'}(s_{t+1}) - \sum_{a'} q_{a'}(s_t)\right)\delta_{aa_t}.$$

Therefore, the explicit representation of the state value function is not required in the actor-critic algorithm.

In the steady state, $\langle \Delta V(s)\rangle = 0$, and we have

$$\langle V(s)\rangle = \langle r_t + \gamma V(s_{t+1}) | s_t = s\rangle = \big\langle r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots | s_t = s\big\rangle .$$

The average changes in the policy parameters in $s_t = s$ can be calculated as

$$\begin{aligned}\frac{\langle \Delta q_a(s)\rangle}{\alpha} &= \langle (r_t + \gamma V(s_{t+1}) - V(s))\delta_{aa_t}\rangle \\ &= \left\langle \left(r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots\right)\delta_{aa_t}\right\rangle - \langle V(s)\delta_{aa_t}\rangle \\ &= R_a(s) - \langle V(s)\delta_{aa_t}\rangle,\end{aligned}$$

where the condition $s_t = s$ is omitted from the form $\langle \cdot | s_t = s\rangle$ for brevity. The second term satisfies an inequality,

$$\left|\langle V(s)\delta_{aa_t}\rangle - \langle V(s)\rangle\langle\delta_{aa_t}\rangle\right| \le \sqrt{\mathrm{Var}[V(s)]\ \mathrm{Var}[\delta_{aa_t}]},$$

where Var[$\cdot$] represents the conditional variance on $s_t = s$. The value of δ_{aa_t} is 0 or 1. Therefore, the variance is not larger than $\frac{1}{4}$. The maximum deviation of $V(s)$ in the time span τ is not larger than the following upper bound,

$$\max V(s) - \min V(s) \le \alpha\tau(1 + \gamma + \gamma^2 + \cdots)\max_t r_t = \frac{\alpha\tau \max_t r_t}{1-\gamma},$$

from which we can obtain the inequality

$$\left|\langle V(s)\delta_{aa_t}\rangle - \langle V(s)\rangle\langle\delta_{aa_t}\rangle\right| \le \frac{\alpha\tau \max_t r_t}{4(1-\gamma)}\ .$$

Since the value $V(s)$ is not larger than $\max_t r_t/(1-\gamma)$, and since $\langle\delta_{aa_t}\rangle$ represents the choice frequency at $s_t = s$, $P_a(s)$, we have

$$|\varepsilon_a(s)| \le \frac{\alpha\tau}{4}, \quad \text{where } \varepsilon_a(s) \equiv P_a(s) - \frac{\langle V(s)\delta_{aa_t}\rangle}{\langle V(s)\rangle}. \tag{B.2}$$

The average changes in the policy parameters are described with this $\varepsilon_a(s)$ as

$$\frac{\langle \Delta q_a(s)\rangle}{\alpha} = R_a(s) - \langle V(s)\rangle(P_a(s) - \varepsilon_a(s)).$$

The steady state defined by the condition $\langle \Delta q_a(s)\rangle = 0$ ensures the following choice frequencies,

$$P_a(s) = R_a(s) \Bigg/ \sum_{a'} R_{a'}(s) + \varepsilon_a(s),$$

which implies that $\varepsilon_a(s)$ represents the deviation from the extended matching law, equation B.1, and is diminished in the limit of an infinitesimally small learning rate, $\alpha\tau \to 0$. The upper bound of the deviation can be derived from the worst evaluation in which rewards happen to be given on every trial in every averaging span τ. The deviation is far smaller in practical cases than the upper bound $\alpha\tau/4$. Thus, we find that the actor-critic learning exhibits the extended matching law in the steady state for a sufficiently small learning rate.

Other learning rules leading to the matching law can also be extended to include state variables. The local matching law proposed by Sugrue et al. (2004) determines the current choice probabilities directly from the fractional incomes in the immediate past. The learning rule with state variables can be achieved by online TD learning for estimating the state-dependent reward expectation and the choice probability as

$$\Delta\hat{R}_a(s_t) = \alpha\left[\left(r_t + \gamma\sum_a \hat{R}_a(s_{t+1})\right)\delta_{aa_t} - \hat{R}_a(s_t)\right],$$

$$p_a(s) = \hat{R}_a(s) \Bigg/ \sum_{a'} \hat{R}_a(s)\ .$$

"Melioration," introduced by Herrnstein (1997), proposes to update the current choice probabilities so as to increase the choice probability associated with the greatest value of $Q_a = R_a/P_a$. This learning rule gives a steady state in which all options have the same average value of these ratios, hence leading to the matching law. To incorporate state transitions, melioration can be reformulated as TD learning that estimates the value of the state, $Q_a(s)$:

$$\Delta\hat{Q}_a(s_t) = \alpha\left[\left(r_t + \gamma\sum_a \hat{Q}_a(s_{t+1})p_a(s_{t+1}) - \hat{Q}_a(s_t)\right)\delta_{aa_t}\right], \tag{B.3}$$

and the update of the choice probabilities can be described by using policy parameters as

$$\Delta q_{a^*}(s_t) = \alpha\left(\hat{Q}_{a^*}(s_t) - \frac{1}{n}\sum_{a=1}^{n}\hat{Q}_a(s_t)\right),\quad a^* \equiv \arg\max_a \hat{Q}_a(s_t),$$

$$p_a(s) = f(q_a(s)) \Bigg/ \sum_{a'} f(q_{a'}(s))\ .$$

Melioration is implemented by the two parts responsible for evaluation and choice and hence is interpreted as a class of the actor-critic learnings in a broad sense (Daw & Touretzky, 2001). If the choice is deterministically made on the basis of the maximum value of $Q_a(s_t)$, namely, $a_t = \arg\max_a \hat{Q}_a(s_t)$, then equation B.3 is described as

$$\Delta \hat{Q}_a(s_t) = \alpha \left[\left(r_t + \gamma \max_a \hat{Q}_a(s_{t+1}) - \hat{Q}_a(s_t) \right) \delta_{aa_t} \right].$$

Thus, melioration includes "greedy Q-learning"(Sutton & Barto, 1998) as an extreme case.

Reinforcement learning rules based on the Q-values, for example, Q-learning and Sarsa (Sutton & Barto, 1998), do not generally exhibit the matching law. In Q-based learning, the relationships between the choice probabilities and the average returns, is determined as

$$p_a(s) = f(Q_a(s)) \Bigg/ \sum_a f(Q_a(s)) .$$

Therefore, the matching law can be obtained only in the greedy case defined at the limit, $f(q) = \lim_{\beta \to \infty} e^{\beta q}$.

Appendix C: Details of Simulations

In Figures 1, 2, and 4, simulations were conducted in the actor-critic system defined with $\alpha = 0.005$ and $f(q) = e^{10q}$, and the initial conditions at $t = 0$ were set as $V = q_1 = q_2 = 0$. In the simulations displayed in Figures 2 and 4, the locally averaged incomes $\hat{R}_1$ and $\hat{R}_2$ were updated on each trial according to

$$\Delta \hat{R}_a = \alpha (r_t \delta_{aa_t} - \hat{R}_a) ,$$

with the initial conditions $\hat{R}_1 = \hat{R}_2 = 0$ at $t = 0$. The timescale for the averaging is set as $1/\alpha = 200$ trials. In order to see the local matchings, the fraction $\hat{R}_1/(\hat{R}_1 + \hat{R}_2)$ are shown with gray curves in Figures 2E and 4E.

Acknowledgments

We express our sincere thanks to Takayuki Sakagami of Keio University for fruitful discussions on animals' choice behavior at the early stage of this study. This work was partially supported by Grants in Aid for Priority Researches 17022036 and 17021038.

References

Barraclough, D., Conroy, M., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, *7*(4), 404–410.

Baum, W. M. (1981). Optimization and the matching law as accounts of instrumental behavior. *Journal of the Experimental Analysis of Behavior*, *36*, 387–402.

Baum, W., & Rachlin, H. (1969). Choice as time allocation. *Journal of the Experimental Analysis of Behavior*, *12*, 861–874.

Breiter, H. C., Aharon, I., Kahneman, D., Dale, A., & Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron*, *30*, 619–639.

Davison, M., & McCarthy, D. (1987). *The matching law: A research review*. Mahwah, NJ: Erlbaum.

Daw, N. D., & Touretzky, D. S. (2001). Operant behavior suggests attentional gating of dopamine system inputs. *Neurocomputing*, *38–40*, 1161–1167.

Daw, N. D., & Touretzky, D. S. (2002). Long-term reward prediction in TD models of the dopamine system. *Neural Computation*, *14*, 2567–2583.

Dayan, P., & Abbott, L. (2001). *Theoretical Neuroscience*. Cambridge, MA: MIT press.

Dayan, P., & Balleine, B. W. (2002). Reward, motivation, and reinforcement learning. *Neuron*, *36*, 285–298.

DeCarlo, L. T. (1985). Matching and maximizing with variable-time schedules. *Journal of the Experimental Analysis of Behavior*, *43*, 75–81.

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, *10*, 732–739.

Gallistel, C., Mark, T., King, A., & Latham, P. (2001). The rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect. *J. Exp. Psychol. Anim. Behav. Processes*, *27*, 354–372.

Haruno, M., Kuroda, T., Doya, K., Toyama, K., Kimura, M., Samejima, K., Imamizu, H., & Kawato, M. (2004). A neural correlate of reward-based behavioral learning in caudate nucleus: A functional magnetic resonance imaging study of a stochastic decision task. *J. Neurosci.*, *24*(7), 1660–1665.

Herrnstein, R. J. (1997). *The matching law: Papers in psychology and economics*. Cambridge, MA: Harvard University Press.

Herrnstein, R. J., & Heyman, G. M. (1979). Is matching compatible with reinforcement maximization on concurrent variable interval, variable ratio? *Journal of the Experimental Analysis of Behavior*, *31*, 209–223.

Herrnstein, R. J., & Vaughan, W. J. (1980). Melioration and behavioral allocation. In J. Staddon (Ed.), *Limits to action: The allocation of individual behavior*. New York: Academic Press.

Heyman, G. M. (1979). A Markov model description of changeover probabilities on concurrent variable-interval schedules. *Journal of the Experimental Analysis of Behavior*, *31*, 41–51.

Heyman, G., & Monaghan, M. (1994). Reinforcer magnitude (sucrose concentration) and the matching law theory of response strength. *Journal of the Experimental Analysis of Behavior*, *61*, 505–516.

Houk, J. C., Davis, J. L., & Beiser, D. G. (1994). *Models of information processing in the basal ganglia (computational neuroscience)*. Cambridge, MA: Bradford Books, MIT Press.

Houston, A. I., & McNamara, J. (1981). How to maximize reward rate in two variable-interval paradigms. *Journal of the Experimental Analysis of Behavior*, *35*, 367–396.

Jacobs, E. A., & Hackenberg, T. D. (1996). Humans' choices in situations of time-based diminishing returns: Effects of fixed-interval duration and progressive-interval step size. *Journal of the Experimental Analysis of Behavior*, *65*, 5–19.

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. J. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J. Neuroscience*, *15*, 1–5.

Mazur, J. (1981). Optimization theory fails to predict performance of pigeons in a two-response situation. *Science*, *214*(4522), 823–825.

Mazur, J. E. (2005). *Learning and behavior*.(6th ed.). Upper Saddle River, NJ: Prentice Hall.

McClure, S., Berns, G. S., & Montague, P. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339–346.

Montague, P., & Berns, G. (2002). Neural economics and the biological substrates of valuation. *Neuron*, *36*(2), 265–284.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neuroscience*, *16*, 1936–1947.

Morris, G., Arkadir, D., Nevet, A., Vaadia, E., & Bergman, H. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron*, *43*, 133–143.

Platt, M., & Glimcher, P. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233–238.

Rachlin, H., Green, L., Kagel, J., & Battalio, R. (1976). Economic demand theory and psychological studies of choice. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 10, pp. 129–154). New York: Academic Press.

Sakagami, T., Hursh, S. R., Christensen, J., & Silberberg, A. (1989). Income maximizing in concurrent interval-ratio schedules. *Journal of the Experimental Analysis of Behavior*, *52*, 41–46.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific-reward value in the striatum. *Science*, *310*, 1337–1340.

Savastano, H. I., & Fantino, E. (1994). Human choice in concurrent ratio-interval schedules of reinforcement. *Journal of the Experimental Analysis of Behavior*, *61*, 453–463.

Schultz, W. (2004). Neural coding of basic reward terms of animal learning theory, game theory, microeconomics and behavioural ecology. *Current Opinion in Neurobiology*, *14*, 139–147.

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

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

Silberberg, A., Thomas, J., & Brendzen, N. (1991). Human choice on concurrent variable-interval, variable-ratio schedules. *Journal of the Experimental Analysis of Behavior*, *56*, 575–584.

Staddon, J., & Hinson, J. (1983). Optimization: A result or a mechanism? *Science*, *221*, 976–977.

Stubbs, D. A., Pliskoff, S. S., & Reid, H. M. (1977). Concurrent schedules: A quantitative relation between changeover behavior and its consequences. *Journal of the Experimental Analysis of Behavior*, *27*, 85–96.

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

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning*. Cambridge, MA: MIT press.

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, *7*, 887–893.

Vyse, S. A., & Belke, T. W. (1992). Maximizing versus matching on concurrent variable-interval schedules. *Journal of the Experimental Analysis of Behavior*, *58*, 325–334.

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

Received June 27, 2005; accepted March 31, 2007.