OPEN ACCESS Freely available online

PLoS one

When Does Reward Maximization Lead to Matching Law?

Yutaka Sakai[1], **Tomoki Fukai**[2]*

[1] Brain Science Institute, Tamagawa University, Machida, Tokyo, Japan, [2] Laboratory for Neural Circuit Theory, Brain Science Institute, RIKEN, Wako, Saitama, Japan

Abstract

What kind of strategies subjects follow in various behavioral circumstances has been a central issue in decision making. In particular, which behavioral strategy, maximizing or matching, is more fundamental to animal's decision behavior has been a matter of debate. Here, we prove that any algorithm to achieve the stationary condition for maximizing the average reward should lead to matching when it ignores the dependence of the expected outcome on subject's past choices. We may term this strategy of partial reward maximization "matching strategy". Then, this strategy is applied to the case where the subject's decision system updates the information for making a decision. Such information includes subject's past actions or sensory stimuli, and the internal storage of this information is often called "state variables". We demonstrate that the matching strategy provides an easy way to maximize reward when combined with the exploration of the state variables that correctly represent the crucial information for reward maximization. Our results reveal for the first time how a strategy to achieve matching behavior is beneficial to reward maximization, achieving a novel insight into the relationship between maximizing and matching.

Citation: Sakai Y, Fukai T (2008) When Does Reward Maximization Lead to Matching Law? PLoS ONE 3(11): e3795. doi:10.1371/journal.pone.0003795

Editor: Tim Bussey, University of Cambridge, United Kingdom

Received April 4, 2008; **Accepted** November 4, 2008; **Published** November 24, 2008

Funding: This work was partially supported by the Grant in Aid for Priority Researches, no. 17022036 and no. 17021038. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: tfukai@brain.riken.jp

Introduction

How do animals, including humans, determine appropriate behavioral responses when their behavioral outcomes are uncertain? Decision-making is a fundamental process of the brain for organizing behaviors, and depends crucially on how subjects have been rewarded in their past behavioral responses. Mechanism of reward-driven learning has extensively been studied theoretically and experimentally. A well-known example includes the reinforcement learning theory based on the temporal difference (TD) error algorithm[1], which is powerful enough to solve difficult problems in machine control and accounts for the basal-ganglia activity representing reward expectancy in monkeys and humans[2–4]. It is generally considered that subjects attempt to choose a behavioral policy that will maximize the amount of reward under a given environmental condition [5]. In addition, many algorithms in machine learning and other brain-style computations aim at reward maximization or, somewhat more generally, optimization of a given cost function.

Nevertheless, animals often exhibit matching behavior in a variety of decision-making tasks[6–9], even if such behavior does not necessarily maximize reward. The matching law states that the frequency of choosing an option is proportional to the amount of past reward obtained from that option[6]: $N_a/(N_1+N_2+\ldots+N_n) = I_a/(I_1+N_2+\ldots+N_n)$, where N_a ($a = 1,\ldots,n$) represents the times option a has been chosen and I_a the total amount of income obtained at the option. A typical example showing this law is the alternative choice task, in which subjects have to choose one from the two options that may be rewarded at different average rates. Matching and maximizing are mathematically equivalent in simple tasks[10,11], but not in arbitrary tasks[12–15].

Decision-making models to reproduce the matching behavior have been proposed[9,16,17], and recent computational studies pointed out possible origins of matching behavior in biological neural systems[18,19]. For instance, a recent model proposed that the matching law results from the covariance learning rule in synaptic plasticity[19]. In addition, we previously demonstrated that the matching law emerges in a class of the reinforcement learning systems including the actor-critic[20,21], which has widely been used in engineering applications. However, whether matching and maximizing share a common computational principle and whether matching behavior is beneficial to decision making remain unclear. In this study, we propose a view that unifies matching behavior into the general computational framework of reward maximization.

Results

We first prove that partial maximization of reward leads to matching behavior irrespective of the mathematical algorithm used for this computation. A crucial step is to define "the matching strategy" that plays a central role in the present study. We then demonstrate how the matching strategy substitutes for the maximizing strategy in a decision-making task that is difficult to solve, when matching is combined with an appropriate utilization of available information sources.

Matching as a Sub-optimal Maximizing Strategy in Independent Choice Behaviors

The analysis is easier if we express the matching law as follows[8]:

$$N_a = 0, \text{ or } \langle r|a\rangle = \frac{I_1+I_2+\ldots+I_n}{N_1+N_2+\ldots+N_n} = \langle r\rangle \text{ for } N_a \neq 0, \quad a = 1,\cdots,n \qquad (1)$$

where $\langle r \rangle$ is the average reward per choice from all options and $\langle r|a \rangle$ the average reward conditioned on choice of option a. We can derive the above expression from the relationship $I_a \cong \langle r|a \rangle N_a$. Thus, the matching law equalizes the expected returns on all the options that are chosen sufficiently many times. Note that the matching law should not be confused with "probability matching"[22], which states that the frequency of choosing option a is proportional to $\langle r|a \rangle$ rather than I_a. Probability matching is typically observed in a task in which each expected return $\langle r|a \rangle$ is fixed and independent of subject's behavior (i.e., concurrent variable-ratio schedules). In such a simple task, the maximizing behavior satisfies the matching law, but not the probability matching. Hereafter, we focus on the matching law. Moreover, we consider the case where subjects make choices at fixed intervals. We can employ the discrete time steps without much loss of generality, since the framework describes a free-response task on continuous time if the interval is sufficiently short and choosing nothing is an available option.

We analyze the outcome of the decision process without specifying the detail of neural decision system. To this end, we assume a set of 'synapses' $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)$ that determines the behavioral policy to make decision. These variables are often called "policy parameters" in mathematical models of decision making. Then, the probability of choosing option a is given as a function $p_a(\boldsymbol{w})$ of the synaptic weights. To ensure a smooth search for an optimal set of choice probabilities, we require that arbitrary infinitesimal changes of $\{p_a(\boldsymbol{w})\}$ allowed in the space of choice probabilities can be caused by some set of infinitesimal changes $\{dw_j\}$.

With the above definitions, we can describe the average reward per choice as $\langle r \rangle = \sum_{a=1}^{n} \langle r|a \rangle p_a(\boldsymbol{w})$. Many decision-making algorithms attempt to maximize $\langle r \rangle$ by modifying behavioral outputs. Whatever algorithm is used, the synaptic weights to maximize $\langle r \rangle$ should satisfy the stationary condition $\partial \langle r \rangle / \partial w_j = 0$ for arbitrary j, i.e.,

$$\sum_{a=1}^{n} \langle r|a \rangle \frac{\partial p_a(\boldsymbol{w})}{\partial w_j} + \sum_{a=1}^{n} p_a(\boldsymbol{w}) \frac{\partial \langle r|a \rangle}{\partial w_j} = 0, \text{ for } \forall j. \qquad (2)$$

The first term contains the explicit dependence of the choice probability on w_j, whereas the second term the possible change in $\langle r|a \rangle$ generated implicitly by the change in subject's behavioral policy. The conditional expectation value $\langle r|a \rangle$ is obtained by taking an average over all possible patterns of past choices in which the newest choice is option a. In general, the reward probability depends not only on the current choice, but also on the history of the past choices[6,12–15]. In such a case, $\langle r|a \rangle$ depends on the choice probabilities that produced the past choices, and hence depends on w_j.

In order to maximize reward, the brain has to explore the correct dependence of the reward probability on the past choices. It seems, however, difficult to infer this dependency correctly with little knowledge on an accurate model of the environment. In such a difficult situation, the brain may simply omit the second term in Eq. 2 in its practical attempt to maximize reward,

$$\sum_{a=1}^{n} \langle r|a \rangle \frac{\partial p_a(\boldsymbol{w})}{\partial w_j} = 0, \text{ for } \forall j. \qquad (3)$$

Multiplying Eq. 3 by arbitrary variations $\{dw_j\}$ and taking a summation over j gives $\sum_{a=1}^{n} \langle r|a \rangle dp_a(\boldsymbol{w}) = \boldsymbol{R} \cdot d\boldsymbol{p}(\boldsymbol{w}) = 0$, where $dp_a(\boldsymbol{w}) \equiv \Sigma_j (\partial p_a / \partial w_j) dw_j$ represents the infinitesimal change caused by $\{dw_j\}$, and $\boldsymbol{R} \equiv (\langle r|1 \rangle, \langle r|2 \rangle, \ldots, \langle r|n \rangle)$ and $d\boldsymbol{p}(\boldsymbol{w}) \equiv (dp_1(\boldsymbol{w}), dp_2(\boldsymbol{w}), \ldots, dp_n(\boldsymbol{w}))$ are vectors in the space of multiple options. If all options have non-vanishing stationary choice probabilities, the probability changes $d\boldsymbol{p}(\boldsymbol{w})$ may occur in an arbitrary direction that satisfies the probability conservation $1 \cdot d\boldsymbol{p}(\boldsymbol{w}) = d\left(\sum_{a=1}^{n} p_a(\boldsymbol{w})\right) = 0$, where $\mathbf{1} \equiv (1, 1, \ldots, 1)$ is an n-dimensional identity vector. Therefore, the conditions $\boldsymbol{R} \cdot d\boldsymbol{p}(\boldsymbol{w}) = 0$ and $\mathbf{1} \cdot d\boldsymbol{p}(\boldsymbol{w}) = 0$ can simultaneously be satisfied only by such $\boldsymbol{R}$ that is parallel to $\mathbf{1}$. If the stationary choice probability vanishes for some option, $p_a = 0$, we can forbid the changes in this direction ($dp_a = 0$), and $\boldsymbol{R}$ should have identical components for all the options exhibiting non-zero choice probabilities. These results and Eq. 1 imply that the truncated stationary condition given by Eq. 3 is equivalent to the matching law.

Thus, the steady choice behavior exhibits matching when the decision system ignores the influence of subject's past choices on the expected outcome in aiming for the stationary condition of reward maximization. Hereafter, we call this suboptimal maximization strategy to achieve Eq. 3 "matching strategy". By contrast, we call the strategy to directly solve Eq. 2 "the maximizing strategy".

To demonstrate the above relationship between the matching and maximizing strategies, we study an alternative choice task ($n = 2$), in which the expectation value of return on each choice pattern is specified completely by the subject's current (a_t) and most recent choices (a_{t-1}) as $g_{a_t a_{t-1}} \equiv \langle r_t | a_t, a_{t-1} \rangle$ (see Methods). We consider the case where subject's current choice is independent of its past choices. Hereafter, such decision behavior is called "independent choice behavior". Since $p_2(\boldsymbol{w}) = 1 - p_1(\boldsymbol{w})$, the subject's decision system controls only the choice probability $p_1(\boldsymbol{w})$ through $\boldsymbol{w}$, and makes every choice with probability $p_1(\boldsymbol{w})$. Then the average return on the current choice $\langle r_t | a_t \rangle$ is obtained by averaging $g_{a_t a_{t-1}}$ over the possible patterns $a_{t-1} = 1,2$ as $\langle r_t | a_t \rangle = g_{a_t 1} p_1(\boldsymbol{w}) + g_{a_t 2}(1 - p_1(\boldsymbol{w}))$, and hence depends on $\boldsymbol{w}$ through the choice probability $p_1(\boldsymbol{w})$. Since $\partial \langle r_t | a_t \rangle / \partial w_j \neq 0$, the matching strategy does not maximize reward in this task. Actually, it gives $\langle r \rangle = 0.25$ whereas the maximizing strategy yields $\langle r \rangle = 0.45$ (Figure 1).

The matching strategy enables us to derive a variety of learning rules that lead to matching behavior (Supporting Text S1). For instance, such a category of learning rules includes the well-known actor-critic in the reinforcement learning theory [1,20,21], direct actor[23], melioration[16] and local matching[9]. In particular, the actor-critic and direct actor also belong to the covariance rule[19]. We numerically solved the decision task analyzed in Figure 1 to show that all these learning algorithms generate matching behavior (Figure 2A). By contrast, indirect actor [23] does not exhibit matching in the steady behavior (Figure 2B). The indirect actor belongs to Q-learning without state variables[1] (see below for the state variables). Since Q-learning determines the choice probabilities by estimating "action values", i.e., the expected returns on individual options, it does not show matching.

Matching vs. Maximizing over All Possible Choice Behaviors

The quantitative analysis conducted in Figure 1 was restricted to the case where the subject generates independent choice behaviors. It was shown that the maximizing strategy earns better than the matching strategy. However, the average reward $\langle r \rangle = 0.45$ achieved by the maximizing strategy in Figure 1 is not the global maximum, but is only the best one among independent choice behaviors. For instance, an alternate choice pattern of 1212…, where the current choice depends on the most recent choice, can earn better ($\langle r \rangle = (g_{12} + g_{21})/2 = 0.6$) than the best

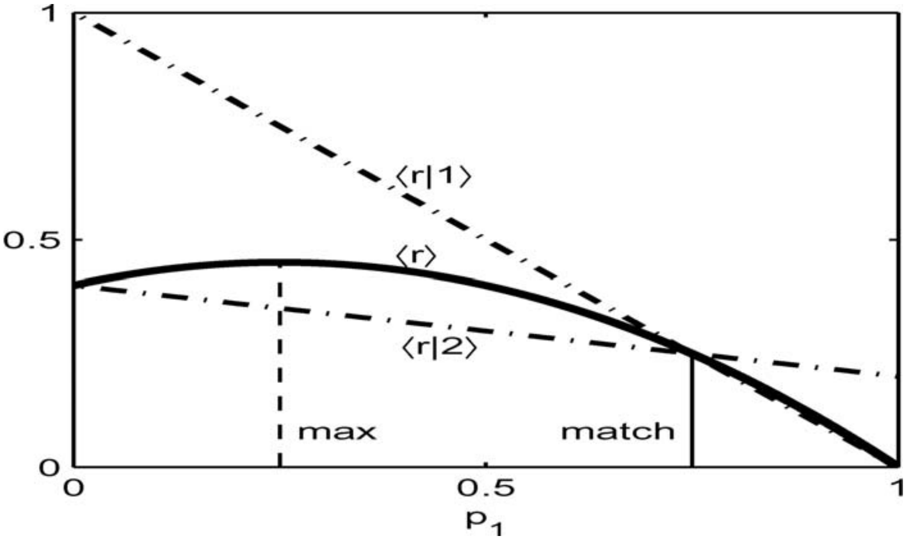

Figure 1. Dependences of the expectation values $\langle r|1\rangle$, $\langle r|2\rangle$ (dot-dashed lines) and $\langle r\rangle$ (solid curve) on p_1 in a decision task with two options (Methods). The reward probability is given as a function of the current and most recent choices, but the subject makes each choice independently of the past choices. The task parameters are set as $g_{11}=0$, $g_{21}=0.2$, $g_{12}=1$ and $g_{22}=0.4$. The expectation values are given as $\langle r|a\rangle = g_{a1}p_1 + g_{a2}(1-p_1)$ and $\langle r\rangle = \langle r|1\rangle p_1 + \langle r|2\rangle(1-p_1)$. The matching (vertical solid line) and maximizing (vertical dashed line) choice probabilities are obtained as solutions of equations $\langle r|1\rangle = \langle r|2\rangle$ and $d\langle r\rangle/dp_1=0$ respectively. The matching strategy ($\langle r\rangle=0.25$) earns less than the maximizing strategy ($\langle r\rangle=0.45$) in this task.
doi:10.1371/journal.pone.0003795.g001

independent choice behavior in that task. Thus, to produce a better outcome in some situation, the subject is required to make each choice depending on the past choices or other available information. Below, we investigate the relationship between the matching and maximizing strategies, taking all possible choice behaviors into account.

To make the argument as general as possible, we include the case where the subject may receive sensory signals $\boldsymbol{\sigma}_t$ before making a choice a_t at time t. Then, in a given task, the external and internal information available for the subject at time t consists of the histories of sensory signals, subject's past choices and the past returns: $\boldsymbol{H}_t = (\boldsymbol{\sigma}_t,\ r_{t-1},\ a_{t-1},\ \boldsymbol{\sigma}_{t-1},\ r_{t-2},\ a_{t-2},\ \boldsymbol{\sigma}_{t-2},\ldots)$. A decision-making task specifies the conditional probability distribution $P(\boldsymbol{\sigma}_{t+1},\ r_t|\,a_t,\ \boldsymbol{H}_t)$. In contrast, the general rule to determine subject's choice behavior is described by the conditional probability distribution $P(a_t|\boldsymbol{H}_t)$. The problem is how to explore an optimal behavioral policy $\hat{P}(a_t|\boldsymbol{H}_t)$ to maximize the average reward $\langle r\rangle$ in a given task.

In practice, however, it is difficult to optimize the dependence of $P(a_t|\boldsymbol{H}_t)$ on the whole history $\boldsymbol{H}_t$. Hence, subject's decision system

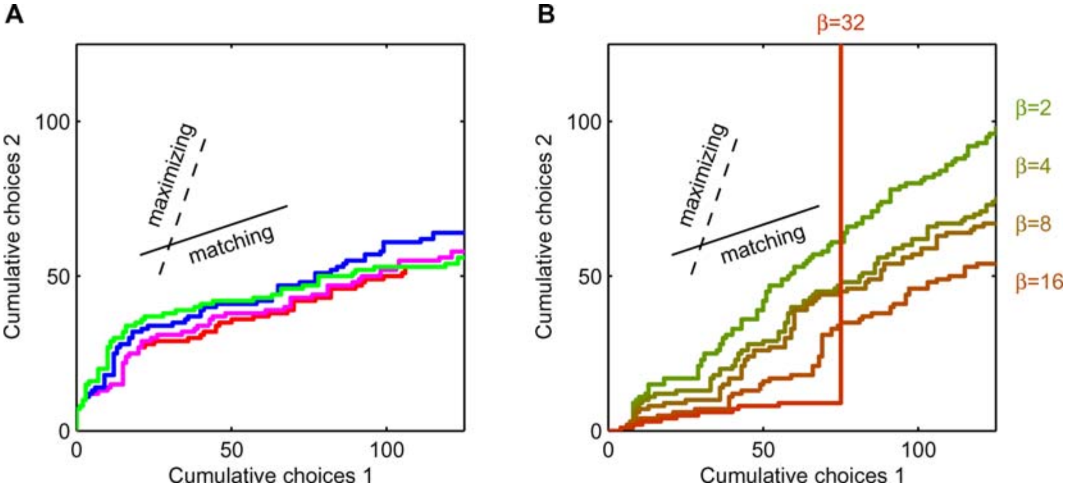

Figure 2. Decision behaviors generated by various decision-making systems. The horizontal and vertical axes indicate the cumulative numbers of choices given to option 1 and 2, respectively. Dashed and solid line segments indicate the slopes corresponding to the maximizing and matching choice probabilities, respectively. See Methods for details of the algorithms. (**A**) The actor critic (red), direct actor (magenta), local matching (blue) and melioration (green) were numerically simulated with $\beta=4$. (**B**) The Q-leaning was simulated for $\beta=2$, 4, 8, 16, and 32. At $\beta=32$, the system eventually learns to choose only option 2.
doi:10.1371/journal.pone.0003795.g002

may extract partial information $\boldsymbol{s}_t$ from $\boldsymbol{H}_t$, and restrict the behavioral policy as

$$P(a_t|\boldsymbol{H}_t)=P(a_t|\boldsymbol{s}_t). \tag{4}$$

We may call the above $\boldsymbol{s}_t$ "state variables". We assume that the decision system controls the definition of state $\boldsymbol{s}_t$, $\boldsymbol{H}_t \mapsto \boldsymbol{s}_t$, and $P(a_t|\boldsymbol{s}_t)$. In order to maximize the average reward, the decision system has to adopt an appropriate definition of state with which an optimal behavioral policy $\hat{P}(a_t|\boldsymbol{H}_t)$ satisfies Eq 4. It has been proved [24] that if a map $\boldsymbol{H}_t \mapsto \boldsymbol{s}_t$ satisfies

$$P(\boldsymbol{s}_{t+1},r_t|a_t,\boldsymbol{s}_t)=P(\boldsymbol{s}_{t+1},r_t|a_t,\boldsymbol{H}_t), \tag{5}$$

for a given task, then the maximal average reward can be obtained by a behavioral policy that satisfies Eq. 4. The average reward obtained by an arbitrary choice sequence can be expressed by $P(\boldsymbol{s}_{t+1},\ r_t|\ a_t,\ \boldsymbol{s}_t)$ that satisfies Eq. 5 and does not depend on the variables that are not reflected in $\boldsymbol{s}$. Therefore, state $\boldsymbol{s}$ that satisfies Eq. 5 represents crucial information about reward delivery in that task. The above theorem means that the optimal policy $\hat{P}(a_t|\boldsymbol{H}_t)$ depends on only the crucial information. Hereafter, we may say that a definition of state variables, $\boldsymbol{H}_t \mapsto \boldsymbol{s}_t$, is correct if and only if $\boldsymbol{s}_t$ satisfies Eq. 5. Note that the selection of the correct definition may not be unique.

Suppose that the decision system adopts a certain definition of state variables, $\boldsymbol{H}_t \mapsto \boldsymbol{s}_t$. Let $p_{a\boldsymbol{s}}=P(a_t=a\,|\,\boldsymbol{s}_t=\boldsymbol{s})$ be the choice probability with which the decision system in state $\boldsymbol{s}$ chooses option a. Each state-dependent choice probability is determined as a function of the synaptic weights $p_{a\boldsymbol{s}}(\boldsymbol{w})$. In order to explore all possible patterns of state-dependent choice probabilities smoothly, we assume that an arbitrary pattern of $\{p_{a\boldsymbol{s}}\}$ and an arbitrary direction of infinitesimal changes $\{dp_{a\boldsymbol{s}}\}$ allowed in the space of probabilities can be expressed by some pattern of $\boldsymbol{w}$ and some direction of infinitesimal changes $d\boldsymbol{w}$, respectively (see Methods).

Taking the state dependence into account, the average reward is written as $\langle r\rangle=\Sigma_{\boldsymbol{s}}\Sigma_a\langle r\,|\,a,\boldsymbol{s}\rangle\ p_{a\boldsymbol{s}}(\boldsymbol{w})P(\boldsymbol{s})$, where $\langle r\,|\,a,\ \boldsymbol{s}\rangle$ is the average reward conditioned on choice of option a in state $\boldsymbol{s}$, and $P(\boldsymbol{s})$ is the distribution of the states that the subject has visited over sufficiently many decision trials with fixed $\{p_{a\boldsymbol{s}}(\boldsymbol{w})\}$. The stationary condition for reward maximization $\partial\langle r\rangle/\partial w_j=0$ is written as

$$\sum_{\mathbf{s},a}\left(\langle r|a,\mathbf{s}\rangle P(\mathbf{s})\frac{\partial p_{a\mathbf{s}}}{\partial w_j}+\langle r|a,\mathbf{s}\rangle p_{a\mathbf{s}}\frac{\partial P(\mathbf{s})}{\partial w_j}+P(\mathbf{s})p_{a\mathbf{s}}\frac{\partial\langle r|a,\mathbf{s}\rangle}{\partial w_j}\right) \tag{6}$$
$$=0,\ \text{for}\ \forall j.$$

The maximizing strategy attempts to achieve Eq. 6 taking the whole dependence on $\boldsymbol{w}$ into account. In contrast, as in the previous case, the matching strategy ignores the dependence of the expected outcome of the current choice on $\boldsymbol{w}$ in aiming for the stationary condition. The outcome in the present case consists of the return r_t and the next state $\boldsymbol{s}_{t+1}$. Therefore, the matching strategy ignores the dependence of $P(\boldsymbol{s}_{t+1},\ r_t|\ a_t,\ \boldsymbol{s}_t)$ on $\boldsymbol{w}$, and hence ignores $\partial\langle r\,|\,a,\ \ \boldsymbol{s}\rangle/\partial w_j$ and $\partial P(\boldsymbol{s}'\,|\,a,\ \ \boldsymbol{s})/\partial w_j$, where $P(\boldsymbol{s}'\,|\,a,\ \boldsymbol{s})\equiv P(\boldsymbol{s}_{t+1}=\boldsymbol{s}'\,|\,a_t=a,\ \ \boldsymbol{s}_t=\boldsymbol{s})$. By transforming the second term repetitively with the recursive relation $P(\boldsymbol{s}')=\Sigma_{\mathbf{s},a}P(\boldsymbol{s}'\,|\,a,\boldsymbol{s})\ p_{as}(\boldsymbol{w})P(\boldsymbol{s})$ and by setting $\partial\langle r\,|\,a,\ \boldsymbol{s}\rangle/\partial w_j=\partial P(\boldsymbol{s}'\,|\,a,\ \boldsymbol{s})/\partial w_j=0$, we obtain the stationary condition of the matching strategy (Supporting Text S2):

$$\sum_{\mathbf{s},a}P(\mathbf{s})\frac{\partial p_{a\mathbf{s}}}{\partial w_j}\lim_{T\to\infty}\frac{1}{T}\sum_{\tau'=0}^{T}\sum_{\tau=0}^{\tau'}\left(\langle r_{t+\tau}\mid a_t=a,\mathbf{s}_t=\mathbf{s}\rangle-\langle r\rangle\right) \tag{7}$$
$$=0,\ \text{for}\ \forall j.$$

Note that the terms omitted in the matching strategy differ for different definitions of the state. Then, using Eq. 7 and the probability conservation, we can extend the matching law to the case of state-dependent choice behaviors (Supporting Text S2):

$$p_{a\mathbf{s}}=0$$
$$\text{or}\ \lim_{T\to\infty}\frac{1}{T}\sum_{\tau'=0}^{T}\sum_{\tau=0}^{\tau'}\left(\langle r_{t+\tau}|a_t=a,\mathbf{s}_t=\mathbf{s}\rangle-\langle r_{t+\tau}|\mathbf{s}_t=\mathbf{s}\rangle\right)=0, \tag{8}$$
$$\text{for}\quad\forall a,\boldsymbol{s}.$$

The extended matching law given as Eq. 8 depends also on the definition of the state.

We schematically illustrate the relationships between the maximizing and matching strategies with correct and incorrect definitions of the state variables (Figure 3A). The horizontal plane represents the multi-dimensional space of arbitrary choice behaviors. Defining state variables restricts the state-dependent choice behavior to a certain subspace. If state variables are correctly defined to satisfy Eq.5, the subspace (red curve) includes the optimal choice behavior (red circle). The conditional probability $P(\boldsymbol{s}_{t+1},\ r_t|\ a_t,\ \boldsymbol{s}_t)$ takes a fixed value specified by the task, which is actually independent of $\boldsymbol{w}$. Therefore, the matching strategy coincides with the maximizing strategy, which indeed earns the globally maximal average reward (red triangle) unless the choice behavior is trapped by a local stationary point. In contrast, if an incorrect definition of state variables is chosen, the set of generable choice behaviors (blue curve) does not necessarily include the optimal choice behavior. Therefore, the maximizing strategy can lead to only the best choice behavior (blue triangle) within the restricted set. The conditional probability $P(\boldsymbol{s}_{t+1},\ r_t|\ a_t,\ \boldsymbol{s}_t)$ depends on the past choices that are not reflected in state $\boldsymbol{s}_t$, and hence depends on $\boldsymbol{w}$. Therefore, the matching strategy (blue cross) in general deviates from the maximizing one (blue triangle).

To explain the above results, we conduct numerical simulations of a simple alternative task in which the reward probability is given as a function of the current and most recent two choices $(a_t,\ a_{t-1},\ a_{t-2})$ (see Methods). A correct definition of state variables for making choice a_t is $\boldsymbol{s}_t=(a_{t-1},\ a_{t-2})$. An actor-critic system (see Methods) operating on the correct state variables earns the globally maximal average reward (Figure 3B, red dashed line). In contrast, for an incorrectly defined state, such as $s_t=a_{t-1}$ or no state variable, the best average rewards (magenta and blue dashed lines, respectively) are smaller than the globally maximal one, and the average rewards earned by the actor-critic systems operating on the incorrect state variables (magenta and blue curves) are still smaller.

Thus, the matching strategy is as efficient as the maximizing one if they are combined with a mechanism to explore and select a correct definition of state variables. However, the matching strategy in general deviates from the maximizing one for the choice behaviors restricted by an incorrect definition of state variables.

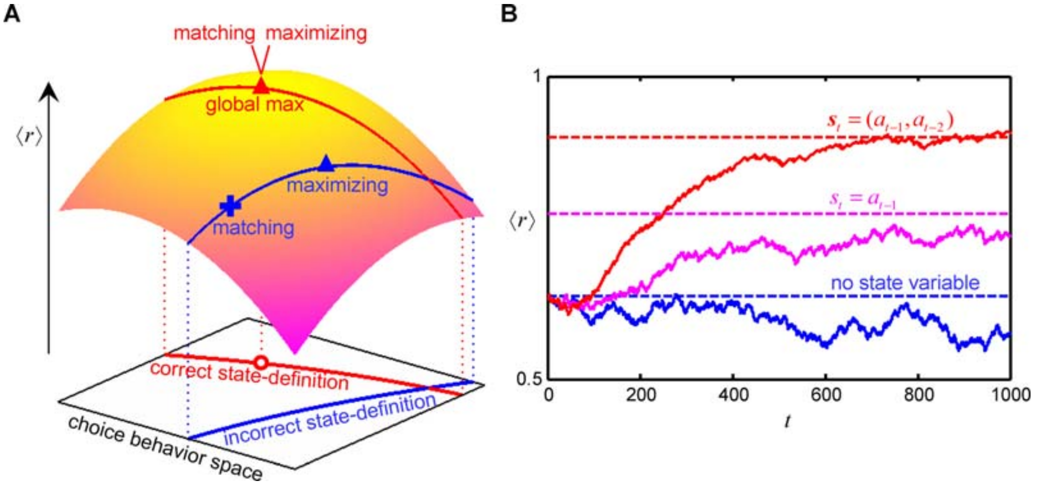

Figure 3. Relationship between the maximizing and matching strategies for state-dependent decision-making. (**A**) The performance of the matching and maximizing strategies based on correctly (red) or incorrectly (blue) defined state variables is shown schematically. (**B**) Actor-critic systems (Methods) were trained on a decision task in which the subject's current and most recent two choices, a_t, a_{t-1} and a_{t-2}, specify the reward probability according to the following task parameters: $g_{111}=0$, $g_{211}=0.6$, $g_{121}=0.9$, $g_{221}=1$, $g_{112}=1$, $g_{212}=0.6$, $g_{122}=1$, and $g_{222}=0$ (Methods). Curves and dashed lines display the local temporal averages of the rewards earned by the actor-critic systems and the best average rewards obtainable by the maximizing strategy, respectively, in three cases: no state variable (blue); an imperfect state variable $s_t=a_{t-1}$ (magenta); correct state variables $\boldsymbol{s}_t=(a_{t-1}, a_{t-2})$ (red).
doi:10.1371/journal.pone.0003795.g003

Discussion

How subjects decide behavioral responses based on their experience and reward expectancy is a current topic in neuroscience. In particular, which choice behavior, matching or maximizing, is more fundamental in decision making has long been debated. The relationship between matching and maximizing behaviors has been often discussed in the restricted case where every choice is independent of the past choices. For instance, Loewenstein and Seung [18] recently proved for independent choice behaviors that the maximizing behavior is achieved by synaptic learning rules that cancel out the infinite sum of the covariances between the current return and all of the current and past decision-related neural activities, and that the matching behavior appears when only the first term in the sum, i.e., the covariance between the current return and current decision-related neural activity, vanishes. This relationship corresponds to the relationship between Eqs. 2 and 3 when the choice probabilities are described as $p_a(\boldsymbol{w}) = e^{\beta w_a} \big/ \sum_{a'} e^{\beta w_{a'}}$ (Supporting Text S1). This study has further extended their results to derive a more general statement: any attempt to achieve the stationary condition for reward maximization results in matching behavior if it ignores the influence of the past choices on the expected outcome. This result depends on neither a specific leaning algorithm nor a specific reward schedule.

Most importantly, we have clarified the general relationship between matching and maximizing strategies among all the possible choice behaviors. We have proved that the matching strategy can lead to the optimal choice behavior when the subject's decision system correctly discovers the information sources sufficient to specify the expected outcome, and can utilize the information through state variables. Differences between the matching and maximizing strategies can arise when the decision system assigns incorrect information sources to the state variables. Our results for the first time revealed how a strategy to achieve the matching behavior is beneficial to reward maximization, and how the ignorance of the relevant information leads to the matching behavior.

The information sources relevant to the expected outcome are task-dependent. In realistic situations, the subject would have no *a priori* knowledge about the probabilistic rule of the outcomes of their behavioral responses. It seems unlikely that the brain easily identifies the relevant information sources from infinitely many combinations of the histories of past sensory inputs, returns and choices. This might explain why the matching law appears so robustly in various animal species and in various decision-making tasks as a result of ignorance of the relevant information sources. In contrast, the matching strategy with the incorrect selection of information sources may replicate various deviations from the matching behavior, such as the under/over-matching observed in various situations [25–28]. Our results provide a theoretical framework to investigate the deviations from matching on the basis of selected information sources. How the brain explores the relevant information sources remains open for further studies. Since this ability of the brain is what discriminates it from any existing artificial machine with human-like adaptive behavior, clarifying the underlying mechanism is an exciting challenge in neuroscience and its application to robotics.

Methods

Summary of assumptions

Our proof of matching law (Eq. 3) is valid for a wide class of natural learning rules, including those employing a widely-used soft-max function for choice probabilities (see below). In the following, however, we explicitly describe the assumptions necessary to make our proof mathematically rigorous. For decision-making tasks, we assumed 1) discrete time step t at which the subject is required to make decision, 2) a finite number of fixed options ($a = 1, 2, \ldots, n$) available for the subject at every time step, and 3) a scalar amount of reward given to the subject at every time step. For the decision system, we required the following assumptions: 4) the decision system can control the definition of state $\boldsymbol{s}_t$ and the state-dependent choice probabilities $\{p_{as}\}$ through a set of synapses $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)$, 5) it adopts a definition of state $\boldsymbol{s}_t$ with which the number of possible states is finite (l), and 6) on a certain definition of state, an arbitrary pattern of possible $\{p_{as}\}$ and an arbitrary direction of possible infinitesimal changes

$\{dp_{as}\}$ can be expressed by some $\boldsymbol{w}$ and $d\boldsymbol{w}$, respectively. The assumption 6 requires the following condition:

$$\begin{aligned}&\forall\left\{y_{a\mathbf{s}}\middle|y_{a\mathbf{s}}>0\ \forall a,\mathbf{s},\text{ and }\sum\nolimits_{a=1}^{n}y_{a\mathbf{s}}=1\ \forall\mathbf{s}\right\},\\ &\exists\mathbf{w}\text{ s.t. }y_{a\mathbf{s}}=p_{a\mathbf{s}}(\mathbf{w})\ \forall a,\mathbf{s},\ J(\mathbf{w})\text{ exists, and }\mathrm{rank}[J(\mathbf{w})]=l(n-1)\end{aligned}\tag{9}$$

where $\boldsymbol{q}(\boldsymbol{w})$ represents the ln-dimensional vector function consisting of the state-dependent choice probabilities $\{p_{as}(\boldsymbol{w})\}$, and $\mathcal{J}(\boldsymbol{w})$ is the Jacobian matrix of $\boldsymbol{q}(\boldsymbol{w})$: $\mathcal{J}_{ij}(\boldsymbol{w})=\partial q_i(\boldsymbol{w})/\partial w_j$. Equation 9 requires $m\geq l(n-1)$. Independent choice behaviors are generated in the case $l=1$.

Decision-making task for demonstrations

To examine the performance of the matching and maximizing strategies, we introduced a decision-making task in which reward is given ($r_t=1$) or not given ($r_t=0$) to the subject according to the probability determined by the subject's current (a_t) and most recent one or two choices (a_{t-1} and a_{t-2}). Each choice should be taken from one of two options ($a=1$, 2), although it is straightforward to extend the present results to more general tasks with more than two options. The conditional expectation value of return on each choice pattern is given as a task parameter: $g_{a_t a_{t-1}}\equiv\langle r_t|a_t,a_{t-1}\rangle$ or $g_{a_t a_{t-1} a_{t-2}}\equiv\langle r_t|a_t,a_{t-1},a_{t-2}\rangle$. The values of these parameters are given in figure legends. For given task parameters $\{g_{a_t a_{t-1} a_{t-2}}\}$, we can calculate the maximum of the average reward $\langle r\rangle=\Sigma_{a,a',a''}g_{aa'a''}p_{aa'a''}P(a',a'')$, where $p_{aa'a''}$ is the conditional choice probability $p_{aa'a''}\equiv P(a_t=a\,|\,a_{t-1}=a',\ a_{t-2}=a'')$, and $P(a',\ a'')$ is the probability distribution $P(a',\ a'')\equiv P(a_{t-1}=a',\ a_{t-2}=a'')$ obtained as a solution of equation $P(a,a')=\Sigma_{a''}p_{aa'a''}P(a',a'')$. The best average rewards obtainable by the restricted choice behaviors with state-definition $s_t\equiv a_{t-1}$ and no state variable can be calculated by restricting $p_{aa'a''}$ as $p_{aa'1}=p_{aa'2}=p_{aa'}$ and $p_{a1}=p_{a2}=p_a$, respectively.

Learning rules for independent choice behaviors

Synapse-updating rules can be described by change Δw_j in w_j at time t, $w_j(t+1)=w_j(t)+\Delta w_j(t)$. Melioration[16] proposes to increase the choice probability of the option that has the largest expectation value of return. An implementation of melioration is described as $p_1(\boldsymbol{w})=w_0$, $p_2(\boldsymbol{w})=1-w_0$, $\Delta w_0=\alpha(w_1-w_2)$ and $\Delta w_a=\alpha\delta_{aa_t}(r_t-w_a)$, where α is a positive constant, and $\delta_{aa_t}=1$ if $a_t=a$, and $\delta_{aa_t}=0$ otherwise. The average returns $\langle r|1\rangle$ and $\langle r|2\rangle$ are estimated as w_1 and w_2, and the choice probabilities are determined by w_0 updated by the estimated average returns. Local matching[9] is designed to directly achieve the matching law as $p_a(\mathbf{w})=w_a/\sum\nolimits_{a'=1}^{n}w_{a'}$ and $\Delta w_a=\alpha(\delta_{aa_t}r_t-w_a)$. For actor-critic[1], direct actor[23] and Q-learning[1], we used a soft-max function as each choice probability: $p_a(\mathbf{w})=e^{\beta w_a}/\sum\nolimits_{a'=1}^{n}e^{\beta w_{a'}}$, where β is a positive constant. Individual updating rules are described as $\mathit{\Delta}w_a=\alpha\beta_{aa_t}(r_t-u)$ and $\Delta u=\alpha(r_t-u)$ (actor-critic), $\Delta w_a=\alpha r_t(\delta_{aa_t}-p_a)$ (direct actor) and $\Delta w_a=\alpha\delta_{aa_t}(r_t-w_a)$ (Q-learning). The details of the algorithms and the relations to the matching strategy and the covariance rule[19] are discussed in Supporting Text S1.

Actor-critic model with state variables

An iterative method to achieve Eq. 7 was shown in [29,30]. Assuming a set of synapses corresponding to individual options in individual states $\{w_{as}\}$ and defining the choice probabilities in each state as $p_{as}(\boldsymbol{w})=e^{\beta w_{as}}/\sum\nolimits_{a'=1}^{n}e^{\beta w_{a's}}$, we can obtain the stochastic gradient ascent rule for Eq. 7 as $\langle\Delta w_{as}\rangle=\lambda\beta P(\boldsymbol{s})p_{as}(Q_{as}-V_s)$, where λ is a positive constant, and $Q_{as}\equiv\lim\limits_{T\to\infty}\frac{1}{T}\sum\limits_{\tau'=0}^{T}\sum\limits_{\tau=0}^{\tau'}(\langle r_{t+\tau}|a_t=a,$ $\boldsymbol{s}_t=\boldsymbol{s}.\rangle-\langle r\rangle)$ and $V_{\boldsymbol{s}}\equiv\Sigma_a Q_{a\boldsymbol{s}}p_{a\boldsymbol{s}}$ represent the relative values of choosing a in state $\boldsymbol{s}$ (relative action-value) and of state $\boldsymbol{s}$ (relative state-value), respectively. Using the relations $P(\mathbf{s})p_{as}Q_{as}=\langle\delta_{\mathbf{s}\mathbf{s}_t}\delta_{aa_t}(r_t-\langle r\rangle+V_{\mathbf{s}_{t+1}})\rangle$ and $P(\mathbf{s})p_{as}V_{\mathbf{s}}=\langle\delta_{\mathbf{s}\mathbf{s}_t}\delta_{aa_t}V_{\mathbf{s}_t}\rangle$, we can obtain the actor-critic model as an implementation of the matching strategy:

$$\begin{cases}\Delta u=\alpha(r_t-u),\\ \Delta v_{\mathbf{s}}=\alpha\delta_{\mathbf{s}\mathbf{s}_t}(r_t-u+v_{\mathbf{s}_{t+1}}-v_{\mathbf{s}}),\\ \Delta w_{a\mathbf{s}}=\lambda\beta\delta_{\mathbf{s}\mathbf{s}_t}\delta_{aa_t}(r_t-u+v_{\mathbf{s}_{t+1}}-v_{\mathbf{s}}),\end{cases}\tag{10}$$

where $\delta_{\mathbf{s}\mathbf{s}_t}=1$ if $\boldsymbol{s}_t=\boldsymbol{s}$, and $\delta_{\mathbf{s}\mathbf{s}_t}=0$ otherwise. The variable u estimates the average reward and the variable v_s represents the state-value of $\boldsymbol{s}$ estimated with the temporal difference (TD) error algorithm. While the actor-critic system is usually designed for maximizing a discounted sum of future rewards[1], the updating rule in Eq. 10 was derived to maximize the average reward[24,29,30].

Numerical simulations

In the simulations shown in Figures 2 and 3B, model parameters were set as $\alpha=\lambda\beta=0.05$, and the initial values of all dynamical variables were set to 1. The value of β was set as $\beta=4$ by default, while it was varied for the Q-learning simulations (Figure 2B). To show the time evolution of reward in Figure 3B, we updated the local average y according to $\Delta y=(r_t-y)/200$ from an initial value of 0.64, which is the average reward obtained with even choice probabilities: $p_1=p_2=0.5$.

Supporting Information

Text S1 Strategies of different learning rules. Several well-known learning algorithms are categorized into the matching, maximizing and other strategies.

Found at: doi:10.1371/journal.pone.0003795.s001 (0.19 MB DOC)

Text S2 Matching strategy in state-dependent choice behaviors. The extensions of the stationary condition and the matching law are derived.

Found at: doi:10.1371/journal.pone.0003795.s002 (0.19 MB DOC)

Acknowledgments

We thank Helena Wang for careful reading of the manuscript.

Author Contributions

Conceived and designed the experiments: YS. Performed the experiments: YS. Wrote the paper: TF.

References

1. Sutton RS, Barto AG (1998) Reinforcement Learning. Cambridge, MA (USA): MIT press. 322 p.
2. Houk JC, Davis JL, Beiser DG (1994) Models of Information Processing in the Basal Ganglia (Computational Neuroscience) Bradford Books. 382 p.
3. Schultz W (1998) Predictive reward signal of dopamine neurons. J Neurophsiol 80: 1–27.
4. Tanaka SC, Doya K, Okada G, Ueda K, Okamoto Y, et al. (2004) Prediction of immediate and future rewards differentially recruits corticobasal ganglia loops. Nature Neurosci 7: 887–893.

5. Mazur JE (2005) Learning and Behavior. Prentice Hall. 464 p.
6. Herrnstein RJ (1961) Relative and absolute strength of response as a function of frequency of reinforcement. J. Exp. Anal. Behav 4: 267–272.
7. Davison M, McCarthy D (1987) The Matching Law: A Research Review Lawrence Erlbaum Assoc Inc. 296 p.
8. Herrnstein RJ (1997) The Matching Law: Papers in Psychology and Economics. Cambridge, MA (USA): Harvard Univ Press. 334 p.
9. Sugrue LP, Corrado GS, Newsome WT (2004) Matching behavior and the representation of value in the parietal cortex. Science 304: 1782–1787.
10. Heyman GM (1979) A Markov model description of changeover probabilities on concurrent variable-interval schedules. J Exp Anal Behav 31: 41–51.
11. Baum WM (1981) Optimization and the matching law as accounts of instrumental behavior. J Exp Anal Behav 36: 387–402.
12. Herrnstein RJ, Heyman GM (1979) Is matching compatible with reinforcement maximization on concurrent variable interval, variable ratio? J Exp Anal Behav 31: 209–223.
13. Mazur J (1981) Optimization theory fails to predict performance of pigeons in a two-response situation. Science 214: 823–825.
14. Vaughan WJ (1981) Melioration, matching, and maximization. J Exp Anal Behav 36: 141–149.
15. DeCarlo LT (1985) Matching and maximizing with variable-time schedules. J Exp Anal Behav 43: 75–81.
16. Herrnstein RJ, Vaughan WJ (1980) Melioration and behavioral allocation. In Staddon J, ed. Limits to action: The allocation of individual behavior. New York (USA): Academic Press.
17. Corrado GS, Sugrue LP, Newsome WT (2005) Linear-nonlinear-Poisson models of primate choice dynamics. J Exp Anal Behav 84: 581–617.
18. Soltani A, Wang X (2006) A biophysically based neural model of matching law behavior: melioration by stochastic synapses. J Neurosci 26: 3731–3744.
19. Loewenstein Y, Seung H (2006) Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity. Proc Natl Acad Sci 103: 15224–15229.
20. Sakai Y, Okamoto H, Fukai T (2006) Computational algorithms and neuronal network models underlying decision processes. Neural Netw 19: 1091–1105.
21. Sakai Y, Fukai T (2008) The actor-critic learning is behind the matching law: Matching vs. optimal behaviors. Neural Comput 20: 227–251.
22. Shanks DR, Tunney RJ, McCarthy JD (2002) A re-examination of probability matching and rational choice. J Behav Dec Making 15: 233–250.
23. Dayan P, Abbott L (2001) Theoretical Neuroscience. Cambridge, MA (USA): MIT press. 360 p.
24. Bertsekas DP, Tsitsiklis JN (1996) Neuro-Dynamic Programming. Belmont, MA (USA): Athena Scientific. 491 p.
25. Baum WM (1979) Matching, undermatching, and overmatching in studies of choice. J Exp Anal Behav 32: 269–81.
26. Davison M, Kerr A (1989) Sensitivity of time allocation to an overall reinforcer rate feedback function in concurrent interval schedules. J Exp Anal Behav 51: 215–231.
27. Alsop B, Davison M (1991) Effects of varying stimulus disparity and the reinforcer ratio in concurrent-schedule and signal-detection procedures. J Exp Anal Behav 56: 67–80.
28. Davison M, Nevin J (1999) Stimuli, reinforcers, and behavior: an integration. J Exp Anal Behav 71: 439–482.
29. Marbach P, Tsitsiklis JN (2001) Simulation-based optimization of Markov reward processes. IEEE Trans Automat Contr 46: 191–209.
30. Konda VR, Tsitsiklis JN (2003) On actor-critic algorithms. SIAM J Contr Optim 42: 1143–1166.