Behavioral/Cognitive

Covariance-Based Synaptic Plasticity in an Attractor Network Model Accounts for Fast Adaptation in Free Operant Learning

Tal Neiman[1] **and Yonatan Loewenstein**[1,2]

[1]Department of Neurobiology, Alexander Silberman Institute of Life Sciences, Interdisciplinary Center for Neural Computation, Edmond and Lily Safra Center for Brain Sciences, and [2]Center for the Study of Rationality, Hebrew University, Jerusalem 91904, Israel

In free operant experiments, subjects alternate at will between targets that yield rewards stochastically. Behavior in these experiments is typically characterized by (1) an exponential distribution of stay durations, (2) matching of the relative time spent at a target to its relative share of the total number of rewards, and (3) adaptation after a change in the reward rates that can be very fast. The neural mechanism underlying these regularities is largely unknown. Moreover, current decision-making neural network models typically aim at explaining behavior in discrete-time experiments in which a single decision is made once in every trial, making these models hard to extend to the more natural case of free operant decisions. Here we show that a model based on attractor dynamics, in which transitions are induced by noise and preference is formed via covariance-based synaptic plasticity, can account for the characteristics of behavior in free operant experiments. We compare a specific instance of such a model, in which two recurrently excited populations of neurons compete for higher activity, to the behavior of rats responding on two levers for rewarding brain stimulation on a concurrent variable interval reward schedule (Gallistel et al., 2001). We show that the model is consistent with the rats' behavior, and in particular, with the observed fast adaptation to matching behavior. Further, we show that the neural model can be reduced to a behavioral model, and we use this model to deduce a novel "conservation law," which is consistent with the behavior of the rats.

Introduction

When searching for food in their natural environments, animals typically alternate between foraging locations. The decision when to leave a foraging location in favor of a different one is challenging, as natural environments are often stochastic and nonstationary. To understand the neural basis and computational principles underlying foraging in natural environments, psychologists and neuroscientists study foraging-like behavior in free operant tasks. In these experiments, an animal moves freely back and forth between different targets corresponding to different ecological patches, harvesting "rewards" that are delivered according to a predefined stochastic reward schedule (Mark and Gallistel, 1994; Gallistel et al., 2001, 2007).

In many of these experiments, whether a response on a target elicits reward depends on the animal's previous actions, such that the probability of reward increases with the time since the last response on that target, motivating subjects to switch between targets. The ensuing behavior of subjects is characterized by three regularities: (1) the distribution of dwell times in each of the targets is approximately exponential (Heyman, 1982; Gibbon, 1995); (2) the returns of the two targets, where return is defined as the number of rewards from that target per time invested in that target, are often equal; stated differently, the fraction of the total time subjects spend in a target matches the fraction of rewards harvested from that target, a behavior known as "the matching law" (Herrnstein, 1961; Davison and McCarthy, 1988; Herrnstein et al., 2000); and (3) adaptation to matching behavior can be fast. It has been estimated that the adaptation rate of rats reaches the limit set by an ideal Bayesian detector. This fast adaptation has cast doubt on the applicability of incremental processes (i.e., processes in which small changes in behavior are accumulated over time) to explain adaptation in these experiments (Gallistel et al., 2001, 2007).

Our goal here is to put forward a mechanistic explanation of these regularities of behavior observed in free operant experiments. We show that a model that is based on attractor dynamics in which transitions are induced by noise and preference is formed via covariance-based synaptic plasticity can account for the experimentally observed behavior. We demonstrate this by

Received April 30, 2012; revised Nov. 22, 2012; accepted Nov. 26, 2012.

Author contributions: T.N. and Y.L. designed research; T.N. and Y.L. performed research; T.N. and Y.L. analyzed data; T.N. and Y.L. wrote the paper.

This work was supported by the Israel Science Foundation Grant 868/08, the Ministry of Science and Technology, Israel, the Ministry of Foreign and European Affairs, the Ministry of Higher Education and Research, France, and the Gatsby Charitable Foundation. We thank C.R. Gallistel for kindly sharing the data with us and D. Hansel for his helpful comments on the manuscript.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. Yonatan Loewenstein, Department of Neurobiology, The Hebrew University, The Edmond J. Safra Campus at Givat Ram, Jerusalem 91904, Israel. E-mail: yonatan@huji.ac.il.

DOI:10.1523/JNEUROSCI.2068-12.2013

 0270-6474/13/331521-14$15.00/0

constructing a biologically plausible model for decision making in free operant conditions, in which populations of neurons compete for higher activity. We compare this model to the behavior of rats responding on two levers for rewarding brain stimulation and show that the model quantitatively accounts for the basic characteristics of behavior in free operant experiments. Moreover, for the case of two populations of neurons, we analytically derive a dynamical behavioral model from the neuronal network model. This model is used to deduce a novel behavioral "conservation law" that is shown to be consistent with the behavior of the rats.

Materials and Methods

The experimental paradigm

The data analyzed in this paper were provided by Dr. Randy Gallistel from Rutgers University. The full details of the experimental procedures appear in Gallistel et al. (2001). In short, in each experimental session, a white male Sprague Dawley rat was placed in a chamber containing two physically separated levers. In the portion of the experiment that we analyzed, each subject ($n = 6$) underwent 20 sessions, each lasting 2 h. Pressing a lever yielded a reward in the form of rewarding brain stimulation according to a concurrent variable-interval (VI) reward schedule. In this schedule, a target could be either baited or empty. Once baited, a target remained baited until it was chosen. When a subject chose a baited target, it was rewarded immediately and the target became empty. An empty target was rebaited probabilistically such that the time to rebait was drawn from a geometric distribution where the time steps were seconds. The reward schedule was characterized by the two means of the two geometric distributions, one for each of the two levers. Five different pairs of means were used in the experiment: (7.1 and 62.5 s), (8.55 and 25.64 s), (12.82 and 12.82 s), (25.64 and 8.55 s), and (62.5 s and 7.1 s). These five pairs corresponded to baiting rate ratios of 9:1, 3:1, 1:1, 1:3, and 1:9, respectively. In each session, two pairs of baiting probabilities were used and an unsignaled change in the baiting rates took place at a random point selected uniformly at random from the middle 80 min of the session. Although no explicit changeover delay was introduced, the animals took time to switch between the targets such that the minimal switching time for the fastest rat was ~1.5 s. Despite this effective changeover delay, the animals could increase the overall number of accumulated rewards by occasionally switching between targets, compared with staying in one of the targets exclusively.

In our analysis and modeling (see below) when the animal was at a target, we assume that it continuously pressed the lever corresponding to the target.

Of a total of 120 sessions, we analyzed 116 sessions: the raw data of one session were missing, and 3 more sessions were discarded because of inconsistencies in the data.

Two-population model

This section is organized in the following way. In the first subsection, we present the network model equations (Eqs. 1 and 2) and the synaptic plasticity equations (Eqs. 3 and 4). In the following three subsections, we use the model to derive a behavioral learning rule (Eq. 24) that predicts behavior based on the history of actions and rewards. In the following subsection (titled Covariance-based synaptic plasticity and the matching law), we provide a formal explanation as to why the behavior of the model obeys the matching law. For convenience, a full list of variables and functions used in the derivation of the behavioral learning rule appears in Table 1.

The network model equations and the plasticity rule. We model the behavior of the animals as resulting from competition for higher activity (e.g., a higher mean firing rate between two populations of neurons), where each population corresponds to one target in the behavioral task. The activity of each population follows a standard rate equation such that

$$\tau \dot{r}_i(t) = -r_i(t) + F(I_i(t)) + n_i(t), \quad (1)$$

Table 1. List of variables and functions used

Symbol	Description
$r_i(t)$	Activity of population i
$I_i(t)$	Input to target i
$F(x)$	The network activation function, $\tanh(\beta x)$
$n_i(t)$	Noise term
$g_i(t)$	The external input to target i
$\Delta g_i(t)$	The change to $g_i(t)$ at time t
$R(t)$	1 at time of reward delivery and 0 otherwise
$\bar{r}_i(t)$	Temporal average of the activity of population i
δr	$\frac{r_2 - r_1}{2}$
δg	$\frac{g_2 - g_1}{2}$
δn	$\frac{n_2 - n_1}{2}$
r	$\frac{r_2 + r_1}{2}$
n	$\frac{n_2 + n_1}{2}$
x	$(\omega_E + \omega_I)\delta r + \delta g$
y	$(\omega_E - \omega_I)r$
$G(x,y)$	$\frac{1 - F^2(y)}{1 - F^2(x) \cdot F^2(y)}$
$A(t)$	$\frac{(\omega_E - \omega_I)}{\tau} \int^t dt'\, F'(x(t'))$
$E(\delta r)$	$\frac{1}{2}\delta r^2 - \frac{1}{\beta(\omega_E + \omega_I)} \log(\cosh(\beta(\omega_E + \omega_I)\delta r + \beta \delta g))$
T_i	Average stay duration at target i
λ_i	Transition rate from target i
δr_{ext}	Extremum point of the energy function
δr^0_{ext}	Value of an extremum point for $\delta g = 0$
E^0	Unperturbed energy function
T^0	Escape time for the unperturbed energy function, E^0
$\Delta \delta g(t)$	The change of δg at time t
$a_i(t)$	1 if the network at time t is in state i, and 0 otherwise
$\Delta \lambda_i(t)$	The change of λ_i at time t
VC	$1/\lambda_1 + 1/\lambda_2$
$\tilde{\lambda}$	$\sqrt{\lambda_1 \cdot \lambda_2}$
f_i	The fractional investment at target i

where τ is the time constant of the dynamics, $r_i(t)$, $i \in \{1, 2\}$, denotes the activity of the neural population corresponding to target i at time t, $F(x) = \tanh(\beta x)$ is the network activation function, β is a parameter, I_i is the total synaptic input to population i and $n_i(t)$ is white noise such that $\langle n_i(t) \rangle = 0$ and $\langle n_i(t) n_j(t') \rangle = 4\sigma^2 \tau \delta_{ij} \delta(t - t')$. This noise represents stochasticity in the activity of neurons within the population of neurons or external noise.

The synaptic input to population i, I_i is given by the following:

$$\begin{aligned} I_1(t) &= \omega_E r_1(t) - \omega_I r_2(t) + g_1(t) \\ I_2(t) &= \omega_E r_2(t) - \omega_I r_1(t) + g_2(t) \end{aligned}, \quad (2)$$

where ω_E and ω_I correspond to efficacies of the self-excitation and lateral inhibition connections, respectively, and g_i is the external input to population i (Fig. 1*A*). The difference between the external inputs determines the target preference of the model as will be explained below.

If β and ω_I are sufficiently large, the dynamics of the deterministic limit ($\sigma \to 0$) of Equations 1 and 2 are endowed with two attractors, one in which $r_1 > r_2$ and one in which $r_1 < r_2$. In the presence of weak noise, the activities of the two populations are characterized by two time scales. On a short time scale τ, the activities fluctuate near an attractor of the deterministic dynamics. On a longer time scale, the noise induces transitions between the attractors (Fig. 1*B*). The two attractors of the dynamics in our model correspond to the two targets and a transition from one

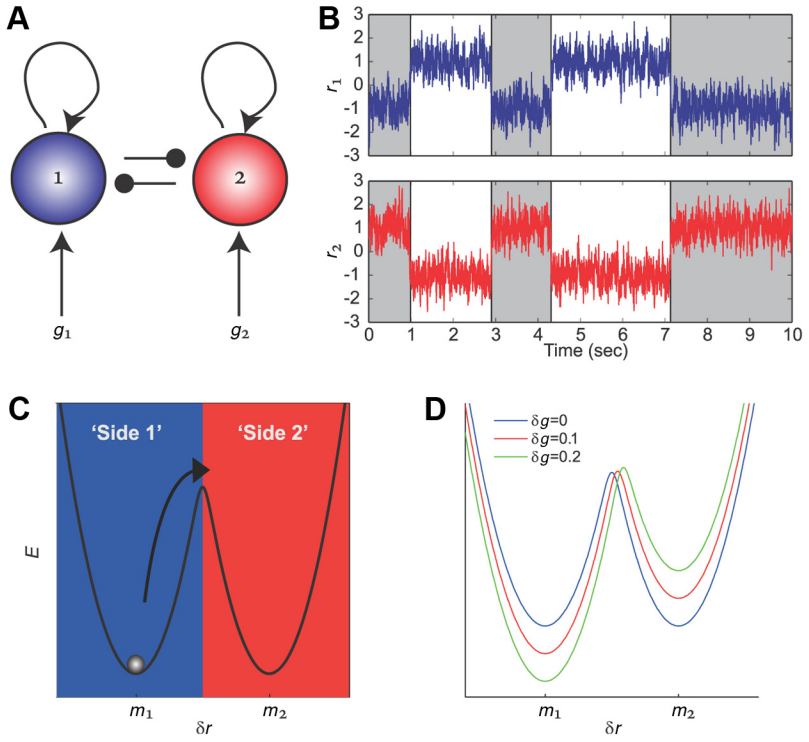

Figure 1. The two-population network model. ***A***, A schematic description of the network architecture. Curved arrows indicate excitatory connections within the populations; circled-headed lines, inhibitory connections between the populations; and vertical arrows, external inputs. ***B***, The activity of the two neuronal populations during a 10 s simulation of the model, using $g_1 = g_2 = 0$ and $\sigma = 3.25$. ***C***, The dynamics of the difference in the activity of the two populations follows the dynamics of a particle in a double-well potential, subject to white noise. ***D***, The shape of the double-well potential depends on the value of δg, depicted here for three different values of the external input.

attractor to the other corresponds to the switching of a target in the behavioral task.

The two external inputs in Equation 2, g_1 and g_2, model the animal's "preference" for the two targets. These change in our model according to a reward-dependent synaptic plasticity rule:

$$\Delta g_i(t) = \phi \cdot R(t) \cdot (r_i(t) - \bar{r}_i(t)), \quad (3)$$

where $\phi > 0$ is a parameter that denotes the magnitude of synaptic changes. $R(t) = 1$ at times of reward delivery and $R(t) = 0$ otherwise; $\bar{r}_i(t)$ is an exponentially decaying temporal average of the neural activity of population i,

$$\tau_m \frac{d\bar{r}_i(t)}{dt} = r_i(t) - \bar{r}_i(t), \quad (4)$$

where τ_m is a constant.

One-dimensional energy model for transitions between the attractors. In this subsection, we show that the network dynamics can be approximated by the dynamics of a particle in a one-dimensional double-well potential with noise.

We consider the dynamics of the difference of the activities of the two populations, $\delta r \equiv \frac{r_2 - r_1}{2}$. Substituting Equation 2 in Equation 1 and subtracting the dynamic equations of the two populations yields:

$$\tau \dot{\delta r}(t) = -\delta r(t) + \frac{F(x(t) + y(t)) + F(x(t) - y(t))}{2} + \delta n(t). \quad (5)$$

where $x \equiv (\omega_E + \omega_I)\delta r + \delta g$, $\delta g \equiv \frac{g_2 - g_1}{2}$, $y \equiv (\omega_E - \omega_I)r$, $r \equiv \frac{r_2 + r_1}{2}$, and $\delta n \equiv \frac{n_2 - n_1}{2}$.

Using the identity $\tanh(a + b) = \frac{\tanh(a) + \tanh(b)}{1 + \tanh(a) \cdot \tanh(b)}$, and the fact that $\tanh(x)$ is an antisymmetric function, Equation 5 becomes:

$$\tau \dot{\delta r}(t) = -\delta r(t) + G(x(t), y(t)) \cdot F(x(t)) + \delta n(t), \quad (6)$$

where

$$G(x, y) = \frac{1 - F^2(y)}{1 - F^2(x) \cdot F^2(y)}. \quad (7)$$

To estimate the value of $G(x, y)$, we first estimate the value of y. Substituting Equation 2 in Equation 1 and summing the dynamic equations of the two populations yields

$$\tau \dot{r} = -r + \frac{1}{2}[F(x + y) + F(y - x)] + n, \quad (8)$$

where $n = \frac{n_1 + n_2}{2}$.

Expanding F around $y = 0$ results in:

$$\tau \dot{r} \approx -r + F'(x)y + n, \quad (9)$$

where we neglected higher-order terms in y (see also below). Assuming equality in Equation 9 yields a nonhomogeneous linear differential equation, for which the solution is given by:

$$r(t) = \frac{1}{\tau} e^{\frac{-t}{\tau} + A(t)} \int^t dt' e^{\frac{t'}{\tau} - A(t')} n(t'), \quad (10)$$

where $A(t) = \frac{(\omega_E - \omega_I)}{\tau} \int^t dt' F'(x(t'))$. $r(t)$ is a stochastic variable whose value varies over time. Note that $\langle r(t) \rangle = 0$ because $\langle n(t) \rangle = 0$. To estimate the magnitude of $r(t)$, we consider its variance:

$$\langle r^2(t) \rangle = \frac{1}{\tau^2} e^{\frac{-2t}{\tau} + 2A(t)} \int^t dt' \int^t dt'' e^{\frac{t' + t''}{\tau} - A(t') - A(t'')} \langle n(t') n(t'') \rangle$$

$$= \frac{2\sigma^2}{\tau} e^{\frac{-2t}{\tau}} \int^t dt' e^{\frac{2t'}{\tau} + 2(A(t) - A(t'))}. \quad (11)$$

Note that $A(t) - A(t') = \frac{(\omega_E - \omega_I)}{\tau} \int_{t'}^t dt'' F'(x(t''))$. In our simulations, we chose ω_E and ω_I such that $\omega_E - \omega_I < 0$. Because $F'(x) \geq 0$, $t' \leq t$ implies that $A(t) - A(t') < 0$ and hence $e^{A(t) - A(t')} \leq 1$. Thus, we can use Equation 11 to set an upper limit to the variance of $r(t)$:

$$\langle r^2(t) \rangle \leq \frac{2\sigma^2}{\tau} e^{\frac{-2t}{\tau}} \int^t dt' e^{\frac{2t'}{\tau}} = \sigma^2. \quad (12)$$

Hence, $\langle y^2(t) \rangle \leq \sigma^2(\omega_E - \omega_I)^2$. In our simulations, based on the parameter fit to the behavioral data, the value of σ differed for each subject and each session (see details below) but in all cases, $\langle y^2(t) \rangle < \sigma^2(\omega_E - \omega_I)^2 < 3 \cdot 10^{-4}$. Thus,

the residual that we neglected in the Taylor expansion of F around $y = 0$ in Equation 9 is negligible. Moreover, because $\langle y^2(t)\rangle \ll 1$, we can estimate the value of $G(x,y)$ Taylor expanding Equation 7 around $y = 0$, yielding:

$$G(x, y) = 1 + (F^2(x) - 1)\beta^2 y^2 + O(y^4), \tag{13}$$

where the O notation denotes the error term in the approximation, expressing the fact that the error is smaller in absolute value than some constant times y^4 when y is close enough to 0.

Note that $|F^2(x) - 1| < 1$, in addition, in our model, $\beta = 10$. Therefore, $G(x,y) \approx 1$, where the fluctuations from 1 are expected to be smaller than 3% and are neglected in what follows. Therefore, Equation 6 is approximately one-dimensional. Rewriting Equation 6 using an energy function yields:

$$\tau\dot{\delta r} = -\frac{\partial E}{\partial \delta r} + \delta n, \tag{14}$$

where

$$E(\delta r) = \frac{1}{2}\delta r^2 - \frac{1}{\beta(\omega_E + \omega_I)} \cdot \log\left(\cosh\left(\beta(\omega_E + \omega_I)\delta r + \beta\delta g\right)\right). \tag{15}$$

The resultant energy function (Eq. 15), is characterized by two minima (Fig. 1*C*, m_1 and m_2), reflecting the two attractors of the deterministic dynamics. Thus, the dynamics of the difference in the population activity approximately follows the dynamics of a particle in a double-well potential, subject to white noise.

The transition times. The energy function E defined in Equation 15 depends on the external inputs via the value of δg. This is illustrated in Figure 1*D*, where we plot E as a function of δr for three values of δg: $\delta g = 0$ corresponds to the case in which the external inputs to the two populations are equal, $g_1 = g_2$. In this case, the energy potential is symmetric around $\delta r = 0$ (Fig. 1*C*,*D*, blue line). However, if the two populations receive different external inputs (e.g., if $g_2 < g_1$ and thus, $\delta g < 0$, the well that corresponds to target 1 is deeper than the well corresponding to target 2; Fig. 1*D*, red line). The more negative the value of δg, the deeper the well corresponding to target 1 and the shallower the well corresponding to target 2 (Fig. 1*D*, green line).

In the limit of weak noise, the particle spends most of the time near an attractor. However, occasionally, a sufficiently large fluctuation of the noise term induces a transition to the other attractor state. The deeper the well, the larger the fluctuation in the noise term required for a transition to the other attractor and, therefore, the longer, on average, the transition time. The mean transition time from the well corresponding to target 1 (Fig. 1*C*, blue) to the well corresponding to target 2 (Fig. 1*C*, red) T_1, also known as the escape time, is given by (van Kampen, 2007):

$$T_1 = \frac{\tau}{\sigma^2}\int_{m_1}^{m_2} dx e^{\frac{E(x)}{\sigma^2}} \int_{-\infty}^{x} dy e^{\frac{-E(y)}{\sigma^2}}, \tag{16}$$

where m_1 and m_2 are the two minima of the energy function E. This expression can be further simplified by using a parabolic approximation (van Kampen, 2007), resulting in:

$$T_1 \propto e^{\frac{E(\text{barrier}) - E(m_1)}{\sigma^2}}. \tag{17}$$

This parabolic approximation becomes more accurate as $\sigma^2 \to 0$. Moreover, it is well known that, in this limit of $\sigma^2 \to 0$, the transition times between the wells follow a Poisson process, with a transition rate $\lambda_i = 1/T_i$ (van Kampen, 2007). We use this result in the next section to construct a behavioral model based on transition rates.

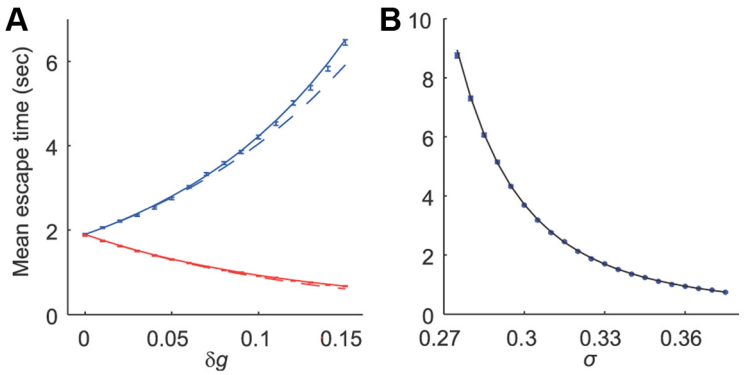

Figure 2. The analytical approximations. ***A***, Escape time as a function of δg. Blue and red dots represent the mean ± SEM escape time from targets 1 and 2, respectively, generated by simulations of Equations 1 and 2. The solid lines indicate the predicted mean escape time based on the double-well potential approximation (Eq. 16); and the hyphenated lines, the predicted mean escape times based on the parabolic approximation (Eq. 20). ***B***, Escape time as a function of the magnitude of the noise for $g_1 = g_2 = 0$. Blue dots are mean ± SEM stay duration generated by simulations of Equations 1 and 2. The black line indicates the predicted mean escape time based on the double-well potential approximation (Eq. 16). ***A***, ***B***, Each dot is based on 10^4 stays.

According to Equation 17, the mean escape time from the well is exponential with the ratio of the difference between the energies at the barrier and at the minimum, $E(\text{barrier}) - E(m_1)$, also known the energy gap, and the noise. Thus, the external inputs affect the transition times because the energies at the extrema of E depend on δg. Because of the exponential dependence of the mean escape time on the energy gaps, even a small change in the energy gap would have a large effect on the mean escape time from the well. Therefore, we focus on the effect of small changes in the value of δg. Taylor expanding the energy at the extrema around $\delta g = 0$ and denoting by δr^0_{ext} the value of an extremum point for $\delta g = 0$, it is easy to see that $\delta r_{ext} = \delta r^0_{ext} + O(\delta g)$. Expanding $E(\delta r_{ext})$ around $\delta g = 0$ yields:

$$E(\delta r_{ext}) = E^0(\delta r^0_{ext}) - \frac{\delta r^0_{ext}}{(\omega_E + \omega_I)} \cdot \delta g + O(\delta g^2), \tag{18}$$

where E^0 is the energy for $\delta g = 0$ and we used the fact that at an extremum, $\frac{\partial E^0(\delta r^0_{ext})}{\partial r} = 0$; thus,

$$\delta r^0_{ext} = \tanh(\beta(\omega_E + \omega_I)\delta r^0_{ext}). \tag{19}$$

Equation 19 has three solutions: $\delta r^0_{ext} = 0$, which corresponds to the value of δr at the barrier, and $\delta r^0_{ext} \approx \pm 1$, which corresponds to the two minima of the unperturbed energy function E^0. In our simulations, $\beta(\omega_E + \omega_I) = 12.5$ and at the minima of E^0, $|\delta r^0_{ext}| \approx 1 - 3 \cdot 10^{-11}$. Therefore, in what follows, we replace δr^0_{ext} at targets 1 and 2 with $\delta r^0_{ext} \approx -1$ and $\delta r^0_{ext} \approx 1$, respectively. Substituting Equation 18 in Equation 17 thus yields:

$$T_1 \approx T^0 \cdot e^{\frac{-\delta g}{(\omega_E + \omega_I)\sigma^2}} \tag{20a}$$

$$T_2 \approx T^0 \cdot e^{\frac{\delta g}{(\omega_E + \omega_I)\sigma^2}}, \tag{20b}$$

where T^0 is the mean escape time for the unperturbed energy function, E^0. In what follows, the value of T^0 was computed numerically using Equation 16 for $\delta g = 0$.

The derivation of Equation 20 relies on four approximations: (1) the double-well approximation (the derivation of Eqs. 14 and 15 from Eqs. 1 and 2); (2) the parabolic approximation of Equation 17; (3) the Taylor expansion of the energy function; and (4) the approximate solution to Equation 19 that replaces δr^0_{ext} with ± 1. To test these approximations, we simulated the network dynamic equations (Eqs. 1

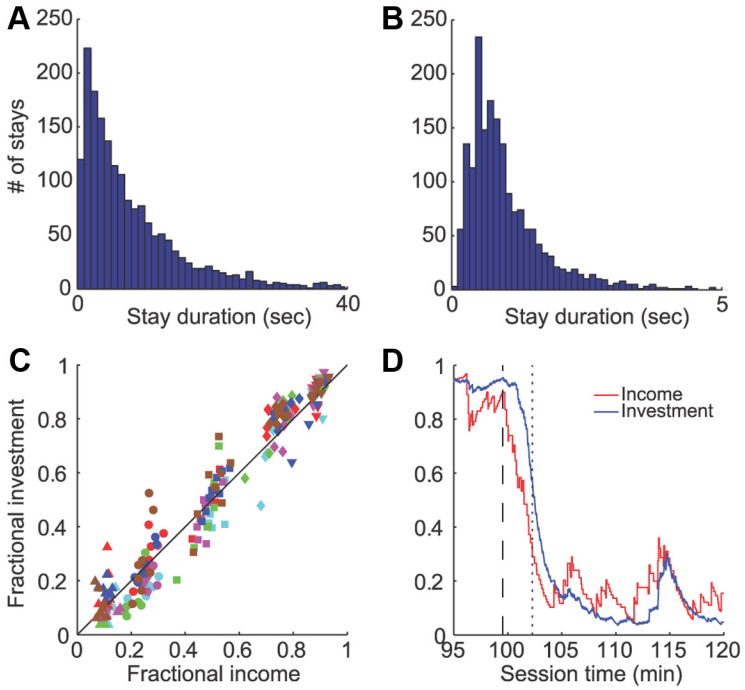

Figure 3. The behavior of the rats. ***A***, ***B***, Histogram of the stay durations of a single subject from all the stationary sections, in which the baiting rate ratio was 9:1. ***A***, Distribution of stay durations in the rich target. ***B***, Distribution of stay durations in the lean target. ***C***, Fractional investment as a function of fractional income. Each dot corresponds to one stationary section in which the baiting rates were kept constant. Colors represent the different subjects, and the different markers indicate the different baiting rate ratios (triangles, 1:9; circles, 1:3; squares, 1:1; diamonds, 3:1; inverted triangles, 9:1). The diagonal solid line indicates the behavior predicted from the matching law. ***D***, Example of instantaneous estimates of fractional income (red) and fractional investment (blue) in a single experimental session of the subject depicted with cyan. At time $t = 99.52$ min, the baiting rate ratio was changed from 9:1 to 1:9 (vertical hyphenated line). The adaptation time is defined as the time interval between the change in the baiting rates and the time at which the instantaneous fractional investment reached halfway between the fractional investments in the stationary sections before and after the change in the baiting rates (dotted vertical line, see Materials and Methods). This experimental session is the same as in Gallistel et al. (2001, their Fig. 6, top right).

and 2) and measured the average stay duration in target 1 (Fig. 2*A*, blue dots) and target 2 (Fig. 2*A*, red dots) for different values of δg. The results of these simulations are comparable to the predictions of the double-well approximation (Eq. 16, Fig. 2*A*, blue and red solid lines) and to the predictions of Equation 20 (Fig. 2*A*, blue and red hyphenated lines), supporting the analytical approximations. To further test these approximations, we compared the dependence of the mean escape time on the magnitude of the noise σ. The numerical simulations (Fig. 2*B*, dots) are comparable with the predictions of Equation 16 (Fig. 2*B*, solid lines), further supporting the analytical approximations.

Synaptic plasticity and the derivation of a behavioral model.
In the synaptic plasticity rule of Equation 3, changes in the external inputs g_1 and g_2 are driven by the product of reward and population activity, where the latter is measured relative to its exponentially weighted temporal average (Eq. 4). To derive an analytical expression in what follows, we consider an approximation of the plasticity rule, in which the running average is replaced by an ensemble average,

$$\Delta g_i(t) = \phi \cdot R(t) \cdot (r_i(t) - \langle r_i(t) \rangle), \tag{21}$$

where $\langle r_i(t) \rangle$ is the ensemble average of the neural activity of population i.

Equations 3 and 21 differ in the average term relative to which the neural activity is measured. The term $\bar{r}_i(t)$ in Equation 3 is a measure of the history of the recent activity. Approximately, it is the activity averaged over a time period of duration τ_m. By contrast, the term $\langle r_i(t) \rangle$ is a deterministic function of g_1 and g_2. It is the activity of r_i, averaged over an infinitely long period of time, in a network in which the values of the external inputs are held fixed at g_1 and g_2. Nevertheless, the temporal average is expected to converge to the ensemble average if learning is sufficiently slow (ϕ is sufficiently small), the reward schedule is stationary, and the temporal window τ_m is sufficiently long.

Subtracting the plasticity rules for g_1 and g_2 in Equation 21 yields:

$$\Delta \delta g(t) = \phi \cdot R(t) \cdot (\delta r(t) - \langle \delta r(t) \rangle). \tag{22}$$

Note that in the limit of weak noise, the value of $\delta r(t)$ fluctuates around the attractors of the deterministic dynamics, which are approximately ± 1, depending on the behavioral state of the animal. Thus, $\delta r(t) \approx a_2(t) - a_1(t)$ where $a_i(t)$ is a binary variable such that $a_i(t) = 1$ if at time t the network is in state i and $a_i(t) = 0$ otherwise. Using the same approximation, $\langle \delta r(t) \rangle \approx \Pr[a_2(t) = 1] - \Pr[a_1(t) = 1]$ and $\delta r(t) - \langle \delta r(t) \rangle \approx -2(a_1(t) - \Pr[a_1(t) = 1])$. Therefore, the plasticity rule of Equation 22 can be approximated by:

$$\Delta \delta g(t) \approx -2\phi \cdot R(t) \cdot (a_1(t) - \Pr[a_1(t) = 1]). \tag{23}$$

Substituting Equation 23 in Equation 20 and using the Poisson process approximation yields:

$$\Delta \lambda_1(t) = \lambda_1(t) \cdot \left(e^{-\eta R(t) \cdot \left(a_1(t) - \frac{\lambda_2(t)}{\lambda_1(t) + \lambda_2(t)} \right)} - 1 \right) \tag{24a}$$

$$\Delta \lambda_2(t) = \lambda_2(t) \cdot \left(e^{-\eta R(t) \cdot \left(a_2(t) - \frac{\lambda_1(t)}{\lambda_2(t) + \lambda_1(t)} \right)} - 1 \right), \tag{24b}$$

where

$$\eta = \frac{2\phi}{(\omega_E + \omega_I)\sigma^2}, \tag{25}$$

and we used the fact that in a Poisson process $\Pr[a_1(t) = 1] = \frac{\lambda_2(t)}{\lambda_1(t) + \lambda_2(t)}$.

Note that, in contrast to Equations 1–4 in which the variables denote neural activity, Equation 24 is a "behavioral model" in the sense that it relates the dynamics of the average transition rates to the history of actions and rewards with no reference to the underlying neural activity.

Covariance-based synaptic plasticity and the matching law. To see why the synaptic plasticity rule (Eqs. 3 and 4) is expected to result in matching behavior, we consider Equation 21, which is an approximation of Equations 3 and 4, in which the running average is replaced by an ensemble average. Equation 21 is an example of a "covariance" rule in which changes in the synaptic input are, on average, proportional to the covariance between the reward and the population activity. In a previous study, we proved a theorem that relates the vanishing covariance between reward and neural activity to the matching law (Loewenstein and Seung, 2006). To gain insights as to why a covariance rule leads to matching here, note that if ϕ is sufficiently small, the stochastic dynamics of Equation 21 approximately follow its average velocity approximation (Heskes and

Kappen, 1993; Kempter et al., 1999; Dayan et al., 2001), in which the right hand side of the equation is replaced by its ensemble average. To compute this average, we separately consider rewards delivered when the network is in state i and when the network is not in state i:

$$\langle \Delta g_i(t) \rangle = \phi \cdot (\langle R(t) \cdot (r_i(t) - \langle r_i(t) \rangle) \mid a_i(t) = 1 \rangle \cdot \Pr[a_i(t) = 1] + \langle R(t) \cdot (r_i(t) - \langle r_i(t) \rangle) \mid a_i(t) = 0 \rangle \cdot \Pr[a_i(t) = 0]). \quad (26)$$

In the concurrent VI reward schedule, the delivery of a reward depends on the chosen target and does not explicitly depend on the neural activity. Because the time scale of neural dynamics in our model is approximately three orders of magnitude shorter than the time scale of reward delivery, reward and neural activity are approximately independent when conditioned on the state of the network, yielding:

$$\langle \Delta g_i \rangle \approx \phi \cdot (\langle R \mid a_i = 1 \rangle \cdot (\langle r_i \mid a_i = 1 \rangle - \langle r_i \rangle) \cdot \Pr[a_i = 1] + \langle R \mid a_i = 0 \rangle \cdot (\langle r_i \mid a_i = 0 \rangle - \langle r_i \rangle) \cdot \Pr[a_i = 0]), \quad (27)$$

where for clarity we removed the dependence of the variables on t. Note that, according to Equation 27, changes in the input to population i are the sum of two products, the first corresponding to the contribution of rewards harvested when the network is in state i and the second corresponding to rewards harvested when the network is not in state i. The first term in each product is the return of the corresponding target. The second term in each product is the average population activity when in the corresponding target, measured relative to the population average activity. On average, the activity of population i is larger than its average when in state i and is lower than average otherwise, $\langle r_i | a_i = 1 \rangle > \langle r_i \rangle$ and $\langle r_i | a_i = 0 \rangle < \langle r_i \rangle$. The third term in the product corresponds to the fractional investment, the fraction of time the network spends in each of the targets.

Next, we decompose the average population activity according to the state of the network:

$$\langle r_i \rangle = \langle r_i \mid a_i = 1 \rangle \cdot \Pr[a_i = 1] + \langle r_i \mid a_i = 0 \rangle \cdot \Pr[a_i = 0]. \quad (28)$$

Equation 28 implies that the products of the second and third terms in the sum in Equation 27 are equal in their magnitude and opposite in their sign, $(\langle r_i \mid a_i = 1 \rangle - \langle r_i \rangle) \cdot \Pr[a_i = 1] = -(\langle r_i | a_i = 0 \rangle - \langle r_i \rangle) \cdot \Pr[a_i = 0]$. Thus, the relative contribution of the two products in Equation 27 depends only on the returns. Formally, substituting Equation 28 in Equation 27 yields:

$$\langle \Delta g_i \rangle = \phi \cdot \Pr[a_i = 1] \cdot \Pr[a_i = 0] \cdot (\langle r_i \mid a_i = 1 \rangle - \langle r_i \mid a_i = 0 \rangle) \cdot (\langle R \mid a_i = 1 \rangle - \langle R \mid a_i = 0 \rangle). \quad (29)$$

As long as the two targets are chosen, $\Pr[a_i=1] \cdot \Pr[a_i=0] > 0$. Moreover, $\langle r_i | a_i = 1 \rangle - \langle r_i | a_i = 0 \rangle > 0$. Thus, the synaptic plasticity rule converges, on average, only if the returns from the two targets are equal, $\langle R | a_i = 1 \rangle = \langle R | a_i = 0 \rangle$ and the network behaves according to the matching law.

Multiple-population model

To model decisions between multiple foraging locations, we consider a network composed of N populations of neurons (see Fig. 7*A* for $N = 3$).

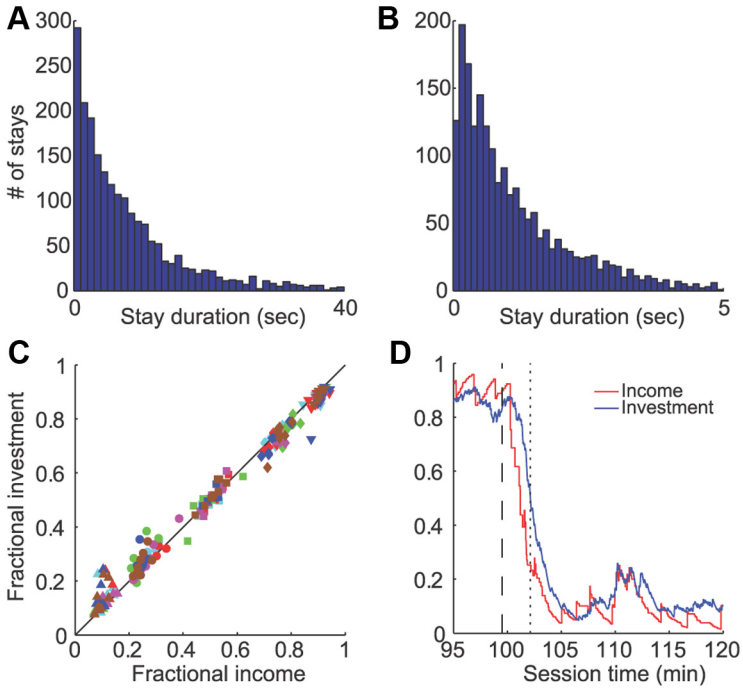

Figure 4. Simulations of the model (Eqs. 1–4). ***A***, ***B***, Histogram of the stay durations of simulations of the same sessions as in Figure 3*A*,*B*. ***C***, Fractional investment as a function of fractional income. Same as in Figure 3*C*. ***D***, Example of instantaneous estimates of fractional income (red) and fractional investment (blue) in a simulation of the experimental session depicted in Figure 3*D*.

The synaptic input to population i (I_i) is given by a generalized form of Equation 2 in which each population is self-excited and inhibited by all other populations:

$$I_i(t) = \omega_E r_i(t) - \omega_I \sum_{j=1}^{N} (1 - \delta_{ij}) r_j(t) + g_i(t). \quad (30)$$

To model decision making, the dynamics of the deterministic limit ($\sigma \to 0$) of Equations 1 and 30 should be endowed with N attractors, wherein each attractor the activity of one population is substantially larger than the activities of all other populations. It is easy to see from Equations 1 and 30 that, if $\beta \gg 1$, a sufficient condition for this multistability is that the input to every population i is in the range $\omega_E - (N-1)\omega_I < g_i(t) < -\omega_E - (N-3)\omega_I$. It should be noted that this is not a necessary condition. If we relax the requirement that all $g_i(t)$ are equal, there exists a wider range of values of $g_i(t)$ such that the dynamics is characterized by N attractors, in which one population is active and the rest are inactive.

Numerical procedures

The dynamic equations (Eqs. 1 and 2) were integrated using Euler's method with a time step of $10^{-4} \times \tau$. Recomputing with a shorter integration time step did not produce appreciable differences in the results. In the two-population model, the initial values of the inputs to the populations were $g_i^{\text{initial}} = 0, i \in \{1,2\}$. In the three-population model, the initial values of the inputs to the populations were $g_i^{\text{initial}} = -0.65, i \in \{1, 2, 3\}$. The initial values of the activities of the populations were chosen such that the activity of one population, drawn at random, was +1 and the activity/activities of the other populations were −1. To simulate saturation, we used a hard bound on the maximal synaptic efficacies such that $|g_i - g_i^{\text{initial}}| < g_{cap}$.

As described above, it took the animal time to switch between the targets. To simulate these delays, we computed the average travel time for each session and used this value as the travel time for each of the switches between the targets. A transition to target i occurred when $r_i - \sum_{j=1}^{N} (1 - \delta_{ij}) r_j(t) > N$. This definition prevented the recording of fast fluctuations during a single transition from one attractor to the other as several short visits.

Table 2. The network model parameters

Parameter	Description	Value
τ	Time constant of the dynamics	10 ms
ω_E	Strength of self-excitation	0.6
ω_I	Strength of lateral inhibition	0.65
β	Steepness of sigmoid function F	10
σ	Magnitude of the noise	Computed from behavior
ϕ	Magnitude of synaptic plasticity	Computed from behavior
g_{cap}	Maximal synaptic input	0.2
τ_m	Time constant of mean activity estimation	25 s

Data analysis

To avoid the effects of the initial conditions, the first 10 min of each session was excluded from the analysis. In analyses performed on the stationary sections, the first 10 min after the change in baiting rates was excluded from the analysis.

The fractional investment/income in Figures 3*D* and 4*D* was computed by convolving the binary vector of stays/rewards with a causal exponential filter with a decay parameter of 90 s.

In the estimation of the adaptation time, the time in which the investment fraction reached halfway was computed according to the following procedure: (1) the fraction of time the animal spent on side 1 in the 10 min before the change in the baiting rates, defined as t_{pre} was computed; (2) the fraction of time the animal spent on side 1 in the 10 min interval starting 10 min after the change in the baiting rates, defined as t_{post} was computed; and (3) the adaptation time was defined as the first crossing of the fractional investment of $(t_{\mathrm{pre}} + t_{\mathrm{post}})/2$ after the change in the baiting rates.

In the analysis of the effect of a rewarded stay on the stay durations of subsequent stays in that target (see Fig. 6*B*), the significance was determined by generating the surrogate data 10^4 times and counting the number of times the stay duration of the control exceeded that of the data.

Choice of parameters

Two populations. The behavioral model (Eq. 24) is characterized by one parameter, the learning rate η and two initial conditions, $\lambda_1(t = 0)$ and $\lambda_2(t = 0)$. Given these parameters, the time-varying transition rates $\lambda_1(t)$ are a deterministic function of the sequence of actions, $a_i(t)$ and rewards, $R(t)$. Given the transition rates, the likelihood of the sequence of stay durations can be computed trivially. For each session, we assumed that $\lambda_1(t = 0) = \lambda_2(t = 0) = \lambda_{init}$ and used the method of maximum likelihood to find the values of λ_{init} and η that best fit the behavior of the rats in that session.

In all simulations, the values of $\tau, \omega_E, \omega_I, \beta, g_{cap}$, and τ_m were held fixed (Table 2). The magnitude of the noise σ was computed from λ_{init} by numerically solving Equation 16 assuming $\delta g = 0$ (Fig. 2*B*) and finding the value of σ such that $T_1 = 1/\lambda_{init}$. The value of ϕ was extracted from Equation 25 such that $\phi = \frac{\eta \cdot (\omega_E + \omega_I)\sigma^2}{2}$.

Three populations. To keep the average stay duration in the case of three populations comparable to that of the case of two populations, we scaled all values of σ that were extracted from the data by the same factor, 0.8.

Results

Characteristics of behavior

We analyzed the switching behavior of rats in a free operant experiment in which rewards were delivered according to the concurrent VI reward schedule. To study the adaptation of animals to a change in the statistics of the reward schedule, the baiting rates changed once within each session at a random, unsignaled time in the middle part of the session. Thus, each 120 min session was composed of two sections of a stationary reward schedule, where each section lasted between 20 min and 100 min (see Materials and Methods).

The subjects constantly switched between the targets. Averaging over all subjects and sessions, animals spent on average 3.42 ± 0.01 s in a target before switching to the other side. The switching between targets was irregular, and the distribution of times that the animal dwelled in a target, which we refer to as "stay durations," rose to an early peak that increased with the corresponding target baiting rate and then tailed off in an approximately exponential manner (Gallistel et al., 2001). This is illustrated in Figure 3*A*,*B*, where we plot the histograms of the stay durations for one subject at the rich side (Fig. 3*A*) and the lean side (Fig. 3*B*), in all session parts in which the baiting rate ratio was 9:1. The approximately exponential nature of the distribution of stay durations was also manifest in the coefficient of variation, CV = 1.11 and CV = 0.97 for the stays depicted in Figure 3*A* and Figure 3*B*, respectively. Averaging over the different schedules and different animals resulted in an average CV of 0.98 ± 0.03. This CV is consistent with what is expected from an exponential distribution (for which CV = 1).

According to the matching law, the fraction of total time spent in a target is equal to the fraction of rewards harvested from that target. It follows that the ratio of the means of the distribution of the stay durations ("fractional investment") should match the ratio of rewards harvested at the two targets ("fractional income"). In line with this law of behavior, when considering the subject in Figure 3, *A* and *B*, the average stay durations in the rich and lean sides were 8.7 ± 0.2 s and 1.04 ± 0.02 s, respectively, resulting in a fractional investment of 0.893 ± 0.006. The number of rewards harvested in the rich and lean sides were 2913 and 375, respectively, resulting in a fractional income of 0.886, similar to the fractional investment. The relation of the fractional income to the fractional investment in all stationary sections of the experimental sessions is shown in Figure 3*C*. Each dot in Figure 3*C* corresponds to one animal in one section of a session in which the baiting rates were kept constant. The different colors denote the different animals, and the different markers denote the different baiting rate ratios of the reward schedule used in the experiment. Note that, consistent with the matching law, all points are aligned approximately along the diagonal.

Traditionally, the matching law has been studied in stationary environments in which the parameters of the reward schedule are constant throughout the session. In contrast, as described above, each session in our dataset was composed of two stationary sections, each with a different pair of baiting rates. By analyzing the behavioral response to the unsignaled change in the reward schedule, it is possible to study how subjects adapt to matching behavior. The results of a representative experiment are shown in Figure 3*D*, where the instantaneous estimates of fractional investment (blue) and fractional income (red) are plotted as a function of time (see Materials and Methods). Initially, the baiting rate ratio was 9:1 in favor of target 1. As a result, the subject spent most of that time in that target. At time 99.52 min (hyphenated vertical line), the baiting rate ratio changed to 1:9, resulting in a decrease in the fractional income. This change in the fractional income (red) was followed by a change in the fractional investment (blue). To quantify the speed of adaptation, we computed the adaptation time, defined as the time interval between the change in the baiting rates and the time in which the investment fraction reached halfway in the adaptation process (Fig. 3*D*, dotted vertical line; see Materials and Methods). The resultant adaptation time in Figure 3*D* was 2.86 min. Averaging over all subjects and sessions, the average adaptation time was 2.77 ± 0.19 min. In a previous study, it was argued that this adaptation is as fast as the limit set by an ideal Bayesian detector (Gallistel et al., 2001), constraining the possible models for adaptation to the changing statistics of the environment.

The decision-making network model

We posit that the behavioral state of the animal is determined by the activities of two noisy populations of neurons. If population 1 is more active than population 2, the animal moves to target 1 and stays there. The opposite behavior occurs if population 2 is more active than population 1. The two populations of neurons are connected by lateral inhibition and self-excitation and receive external input that depends on the history of population activity and rewards delivered (Eqs. 1–4; Fig. 1*A*). This model is similar to previous ones proposed to explain switching between dominance periods in perceptual bistability experiments (Moreno-Bote et al., 2007; Moreno-Bote et al., 2010; Moreno-Bote et al., 2011).

Simulating the network dynamics model while assuming no external inputs (Eqs. 1 and 2), we found that the activities of the two populations (Fig. 1*B*, blue, population 1; red, population 2) alternate between two states: a state in which the activity of population 1 is high and the activity of population 2 is low (Fig. 1*B*, white background) and a state in which the activity of population 1 is low, whereas that of population 2 is high (Fig. 1*B*, gray background). This bimodal dynamics results from the lateral inhibition between the two populations of neurons: if the activity of population 1 is higher than that of population 2, the inhibition on 2 is larger than the inhibition on 1, enhancing the difference in the activities of the two populations. This difference is further enhanced by the self-excitation within each population.

The bimodal network dynamics approximately follow the dynamics of a particle in a one-dimensional double-well potential in the presence of noise (Fig. 1*C*; Materials and Methods). Most of the time, the particle fluctuates near a minimum of the potential (Fig. 1*C*, m_1 and m_2). These fluctuations correspond to the small magnitude fluctuations in the activities of the populations within a network state (Fig. 1*B*, within a single white or a gray region). However, occasionally, the noise is sufficiently strong to shift the particle beyond the energy barrier, which corresponds to a change in the state of the network. This double-well description will become useful in what follows.

The distribution of stay durations of the rat was approximately exponential (Fig. 3*A*,*B*). In comparison, a histogram of the empirical distribution of stay durations of our model was also approximately exponential, with CV = 0.99 (not shown). The exponential distribution of stay durations in our model emerges because the time scale of transitions, which is on the order of seconds, is substantially longer than the time scale of the activity of the populations, which is 10 ms. As a result, at the time of a transition between states, the network has no "memory" of the previous transition; thus, consecutive transitions are approximately independent. This intuition becomes formal when considering the double-well approximation (Materials and Methods). In the limit of weak noise, the escape time of a particle in a double-well potential is exponentially distributed (van Kampen, 2007). The rate of transitions depends on energy gap (the "depth" of the well). The larger the gap, the lower the rate of transitions.

The depths of the wells depend on the external inputs (Fig. 1*A*). If the two external inputs are equal, the corresponding double-well potential is symmetric and the transition rates between the two targets are equal (Fig. 1*C*, *D*, blue). In contrast, if the inputs to the two populations are not equal, the well associated with the larger input is deeper than the well associated with the smaller input (Fig. 1*D*, red and green). As a result, the transition rate from the target associated with the larger input is smaller than that from the target associated with the smaller input. Thus, the two external inputs determine the target preference of the model.

Synaptic plasticity

We assume that the external inputs in our model are not constant and change over time. Motivated by the finding that activity-dependent synaptic plasticity is modulated by the reward-dependent dopamine signal (Jay, 2003; Pawlak and Kerr, 2008; Wickens, 2009; Zhang et al., 2009), a number of theoretical studies have investigated the hypothesis that reward-modulated synaptic plasticity rule underlies operant learning (Seung, 2003; Xie and Seung, 2004; Loewenstein and Seung, 2006; Farries and Fairhall, 2007; Izhikevich, 2007; Law and Gold, 2009; Legenstein et al., 2010; Loewenstein, 2010). We follow a similar approach and consider a reward-modulated synaptic plasticity rule, in which changes in the inputs to the populations, Δg_i, occur only at times of reward delivery and are proportional to the difference between population instantaneous activity and its temporal average (Eqs. 3 and 4).

We simulated the network dynamic equations (Eqs. 1–4) in the same concurrent VI reward schedules as were used for the rats. To account for differences in behavior between subjects and sessions, we estimated the parameters of the network model and synaptic plasticity rule for each session separately, as described in detail below.

We found that, even when synaptic plasticity is incorporated into the model, the shape of the distribution of stay durations in the session parts in which the baiting rates were constant was approximately exponential. This is illustrated in Figure 4, *A* and *B*, where we plot the histograms of the stay durations for a simulation of the rat presented in Figure 3, *A* and *B*, at the rich side (Fig. 4*A*) and the lean side (Fig. 4*B*) in the same session parts as in Figure 3, *A* and *B*.

To quantify matching in our model, we computed the fractional investment as a function of fractional income for all stationary sections of the simulation in which the baiting rates were kept constant (Fig. 4*C*). As in Figure 3*C*, each dot in Figure 4*C* corresponds to the fractional investment as a function of fractional income of one animal model in one section. The different colors and markers denote the simulations of the different animals in the different reward schedules as in Figure 3*C*. Consistent with the matching law, all points in the simulation are aligned, approximately, along the diagonal (compare Figs. 4*C*, 3*C*), demonstrating that the plasticity rule of Equations 3 and 4, when implemented in the network model (Eqs. 1 and 2) results in matching behavior.

To illustrate the adaptation of the model to a change in the baiting rates, the dynamics of the simulation of the experimental session presented in Figure 3*D* is presented in Figure 4*D*. Similar to the behavior of the animal (Fig. 3*D*) and in line with the matching law, the instantaneous fractional investment (blue) is aligned with the instantaneous fractional income (red) in the stationary sections before and after the change in the baiting rates (at t = 99.52 min; hyphenated vertical line). Remarkably, the adaptation time of the model was as fast as that of the animal, 2.61 min (compared with 2.86 min in Fig. 3*D*). Averaging over all sessions, the adaptation time of the model was 2.50 ± 0.24 min, comparable to the adaptation time of the animals (2.77 ± 0.19 min). These simulations thus indicate that the dynamics of the network model with the synaptic plasticity rule (Eqs. 1–4) are sufficiently fast to account for the experimentally observed adaptation to matching behavior.

Convergence to matching behavior

To gain insights as to why the synaptic plasticity rule (Eqs. 3 and 4) yields matching behavior, note that changes in the inputs to a population are proportional to the difference between population instantaneous activity at the times of rewards and its average. When the network is in state 1 (Fig. 1*B*, white), the activity of population 1 is higher than its average activity. As a result, a reward delivered at that time results in a positive change in the input to population 1 and a shift of preference in favor of target 1. By contrast, a reward delivered at a time in which the network is in state 2 (Fig. 1*B*, gray), in which the activity of population 1 is lower than its average activity, results in a negative change in the input to population 1 and a shift of preference in favor of target 2. Changes in the input to population 2 are a mirror image of the changes to population 1 and equally contribute to the change in target preference.

The rates of rewards delivered when the network is in each of the states depend on the returns of the two targets. The larger the return of a target, the larger the number of rewards delivered when the network is in the corresponding state and, as a result, the larger the shift over time in preference in favor of that target. The returns, however, are not constant over time. In the concurrent VI schedule, the larger the fractional investment in a target, the lower the return associated with that target. As a result, a shift in preference in favor of a target is accompanied by a decrease in the return associated with that target and an increase in the return associated with the other target. Therefore, even if initially the return of one of the targets is larger than the return of the other target, the synaptic plasticity rule will shift the preference of the network in favor of the larger return target, thus equalizing the returns associated with the two targets. This change in preference will cease, on average, only when the returns associated with the two targets are equal. In other words, the synaptic plasticity rule will converge only when behavior follows the matching law. A more formal argumentation relating the synaptic plasticity rule with the matching law appears in Materials and Methods.

Behavioral learning rule

As discussed in the previous sections, the transition rates between the two states depend on the difference in the external inputs (Fig. 2*A*), which in turn depend on the history of rewards and actions. In the Materials and Methods, we derive an approximate analytical behavioral model that relates the transition rates to the history of actions and rewards (Eq. 24). For clarity, we rewrite Equation 24:

$$\Delta\lambda_1(t) = \lambda_1(t) \cdot \left(e^{-\eta R(t) \cdot \left(a_1(t) - \frac{\lambda_2(t)}{\lambda_1(t)+\lambda_2(t)} \right)} - 1 \right)$$

$$\Delta\lambda_2(t) = \lambda_2(t) \cdot \left(e^{-\eta R(t) \cdot \left(a_2(t) - \frac{\lambda_1(t)}{\lambda_2(t)+\lambda_1(t)} \right)} - 1 \right),$$

where λ_i denotes the transition rate from state i, $\Delta\lambda_i$ denotes the change in λ_i, $R(t)$ is a binary variable that denotes the time of reward such that $R(t) = 1$ at times of reward delivery and $R(t) = 0$ otherwise, $a_i(t)$ is a binary variable that denotes the state of the network such that $a_i(t) = 1$ at times in which the network is in state i and $a_i(t) = 0$ otherwise, and η is a parameter that depends on the parameters of the model (Eq. 25).

To gain insights into the behavior of the model, note that in the absence of reward, $R(t) = 0$, $\Delta\lambda_2(t) = \Delta\lambda_1(t) = 0$; thus, there are no changes in the transition rates. Consider a reward delivered when the animal is at target 1. In that case, $a_1(t) = 1$; thus, $a_1(t) - \frac{\lambda_2(t)}{\lambda_1(t)+\lambda_2(t)} > 0$. As a result, $\Delta\lambda_1(t) < 0$, biasing the model to spend more time at the rewarded target. Note also that at the time of reward delivery, $a_2(t) = 0$; thus, $a_2(t) - \frac{\lambda_1(t)}{\lambda_2(t)+\lambda_1(t)} < 0$, resulting in $\Delta\lambda_2(t) > 0$, biasing the model in favor of spending less time at target 2.

The behavioral model is substantially simpler than the network model. The variables of the behavioral model are directly related to the experimentally observable variables, namely, actions $a_i(t)$ and rewards $R(t)$, rather than the hidden variables of the network activity and synaptic efficacies. Moreover, the network model is characterized by 8 parameters (Table 2) and 6 initial conditions. By contrast, the behavioral model is characterized by a single parameter, the learning rate η and two initial conditions, the two initial transition rates.

Another advantage of the simpler behavioral model is that the hidden stochasticity of the network model is replaced by a deterministic model of transition rates. Given initial conditions and a learning rate η, we can compute the sequence of transition rates and use this sequence to compute the likelihood of the model. This enabled us to use the method of maximum likelihood to derive the parameters of the model that best fit the behavioral data. The parameters used in the network simulations that produced Figure 4 were extracted in this way from the data (Materials and Methods).

Model predictions

A straightforward calculation reveals that, in the behavioral model (Eq. 24), the product of the transition rates is unchanged: $(\lambda_1 + \Delta\lambda_1) \times (\lambda_2 + \Delta\lambda_2) = \lambda_1 \times \lambda_2$. Thus, we expect that, although a change in the baiting rates should affect the target preference, as manifested in the ratio of the transition rates, the product of transition rates should remain relatively unchanged. In other words, because of the identity $\log(\lambda_1 \times \lambda_2) = \log(\lambda_1) + \log(\lambda_2)$, our model predicts that the sum of the logarithm of the transition rates should remain unchanged.

To test for this "conservation law," we studied to what extent the sum of the natural logarithm of the transition rates changed after a change in the reward schedule. Assuming a stationary Poisson process, the maximum likelihood estimator of the transition rate is the number of stay durations, divided by the sum of these durations. We used this method to estimate the transition rate from each of the two targets in each of the stationary sections. The results of this analysis are shown in Figure 5*A*. Each dot in Figure 5*A* corresponds to a single session and depicts the sum of the logarithm of the two transition rates after the change in the baiting rates, as a function of this sum before the change. Different animals are depicted with different colors, as in Figure 3*C*. In line with our prediction, the dots approximately align along the diagonal (black line).

Note that caution should be exercised in this analysis. In some sessions (Fig. 5*A*, circled dots), the baiting rates in the second section of the session were a mirror image of those in the first section (ratio *x*:*y* in the first section changed to ratio *y*:*x* in the second section). In those sessions, because of the symmetry in the baiting rates, the values of λ_1 and λ_2 after the change in the baiting rates are expected to be approximately equal to λ_2 and λ_1 before that change, respectively in almost any model of learning. In those symmetrical sessions, the conservation of the product of the transition rates is a trivial outcome of the symmetry in the baiting rates. Therefore, we refined our analysis and considered only sessions in which there was no such symmetry in the ratios of

the baiting rates before and after the change (Fig. 5*A*, noncircled dots). In line with our model, the product of transition rates in the nonsymmetrical sessions remained approximately unchanged between the sections of the same session ($r = 0.81$). Importantly, the correlation between the products of transition rates was substantially larger than the correlation between the sums of transition rates, $\lambda_1 + \lambda_2$ ($r = 0.26$) or the average visit cycle (defined as the sum of the stay duration in target 1 and the following stay duration in target 2), $VC = 1/\lambda_1 + 1/\lambda_2$ ($r = 0.45$). The fact that the correlation coefficient for the product of transition rates was larger than that of the sum of transition rates or the average stay duration indicates that the conserved product of transition rates is not an artifact of heterogeneity between sessions and subjects and is not an epiphenomenon of a conservation of the sum of transition rates (Myerson and Miezin, 1980) or conservation of the average visit cycle.

Another prediction of the conservation of the product of transition rates is that the average visit cycle is expected to be a function of choice preference. When the product of transition rates is constant, a straightforward calculation reveals that the average visit cycle is given by the following:

$$VC = \frac{1}{\tilde{\lambda}} \cdot \left(\sqrt{\frac{f_1}{f_2}} + \sqrt{\frac{f_2}{f_1}} \right), \tag{31}$$

where $\tilde{\lambda}$ is the geometric mean of the transition rates ($\tilde{\lambda} = \sqrt{\lambda_1 \cdot \lambda_2}$), and f_i is the fractional investment at target i. To test this prediction, we computed the average VC and the fractional investment in each stationary section. Each dot in Figure 5*B* depicts the average VC as a function of the fractional investment in one stationary section, where the color denotes the subject tested and the symbol denotes the baiting rate ratio (same as in Fig. 3*C*). The solid lines are the prediction of Equation 31. The similarity of the line and the dots for each of the animals is consistent with the prediction of our model.

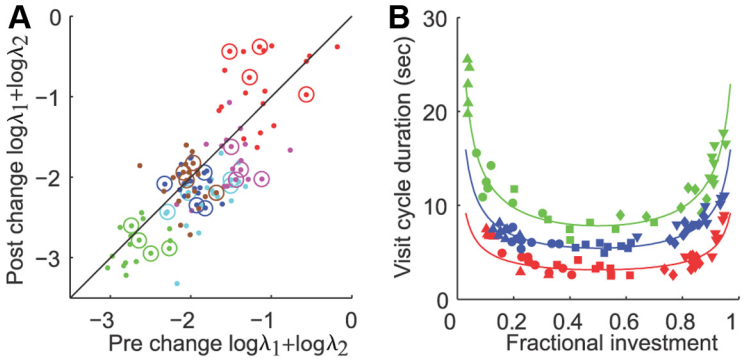

Figure 5. Predictions of the behavioral model. ***A***, The sum of the logarithm of the transition rates after the change in the baiting rates as a function of that sum before the change. Each dot corresponds to a single session. The different color codes for the different subjects. Sessions in which the baiting rates in the second section were a mirror image of those in the first section (ratio *x*:*y* in the first section changed to ratio *y*:*x* in the second section) are marked with circles. The diagonal black line indicates the prediction of the behavioral model. ***B***, Mean *VC* as a function of the fractional investment. Each dot indicates the *VC* in one stationary section as a function of the fractional investment at target 1 in that section. The 3 different colors correspond to 3 different subjects, where the subject denoted in red was the one with the shortest mean *VC*, the subject denoted in green had the longest mean *VC*, and the subject denoted in blue had an intermediate mean *VC*. The solid lines are the predictions of the behavioral model (Eq. 31). The values of $\tilde{\lambda}$ used in the prediction were the geometric means of transition rates, averaged over all sessions, $\tilde{\lambda} = 0.64\ s^{-1}$, $\tilde{\lambda} = 0.26\ s^{-1}$, and $\tilde{\lambda} = 0.37\ s^{-1}$, for the red, green, and blue subjects, respectively.

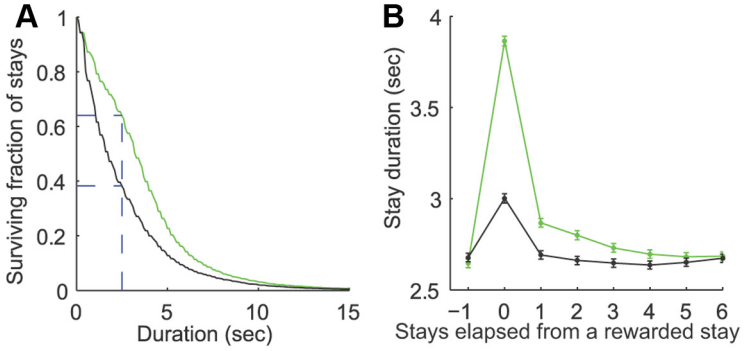

Figure 6. The effect of single rewards on stay duration. ***A***, The effect of rewards on the duration of the rewarded stay. Green represents survival plot of the distribution of rewarded stays; and black, control survival plot. The analysis was repeated for surrogate data in which the rewards were redistributed in the stays according to a concurrent VI schedule, using the same parameters of the schedule as in the experiment. The hyphenated vertical line indicates time $t = 2.5$ s; 64% of the rewarded stays of the subjects were longer than 2.5 s (upper hyphenated line), compared with only 38% of the rewarded stays in the surrogate data (lower hyphenated line). This sets a lower limit on the fraction of rewarded stays that were prolonged as a result of the reward. ***B***, The effect of rewards across stays. The mean ± SEM stay duration as a function of the number of stays elapsed from a rewarded stay in that target. Black represents the same analysis for the surrogate data. ***A***, ***B***, Stays were pooled across all subjects and taken only from the stationary sections in which the baiting rate ratio was 1:1.

The effect of single rewards on behavior

In the behavioral model (Eq. 24), every reward delivered to the animal has an immediate effect on the transition rates. This implies that the transition rates of the animals are expected to change with every reward delivered to the animal even within the stationary sections.

In contrast to this model, motivated by the fast adaptation to matching behavior, previous studies have suggested that subjects do not change their behavior in response to the sequence of rewards unless they detect a significant deviation between the predicted number of rewards they should have obtained and the actual number of rewards harvested. According to this view, the subject has an internal model of the statistics of the sequence of rewards. When the observed sequence of rewards becomes inconsistent with this internal model, the subject changes its internal model and consequently its behavior in an abrupt, stepwise manner and aligns the allocation of time at the two targets according to the newly observed statistics. In other words, subjects act as "change detectors" that change their behavior only when observing statistically significant changes in the statistics of the environment (Gallistel et al., 2001, 2007). The prediction of this model is that, during the stationary sections, single rewards will typically have no effect on behavior.

To test the different predictions of the two models, we tested whether rewards have an immediate effect on the dwell time of subjects. To do so, we considered all the stationary sections in all animals in which the two baiting rate ratio was 1:1. The green line

in Figure 6*A* depicts the fraction of stays longer than duration t as a function of t (survival plot) for all the stays in which rewards were delivered. Compared with the rewarded stays, nonrewarded stays were, on average, shorter (data not shown). However, this difference does not necessarily imply that rewards caused subjects to prolong their stays because in the concurrent VI schedule, the longer a stay, the more likely it is to be rewarded. Thus, the rewarded stays are expected to be on average longer than the nonrewarded stays even if subjects' behavior is unaffected by rewards. To account for this dependency, we created surrogate data in which the rewards were redistributed in the stays according to a concurrent VI schedule, using the same parameters of the schedule that were used in the experiment. The resultant distribution of rewarded stays (Fig. 6*A*, black line) is the expected distribution from a subject whose stays are exactly the same as observed in the experiment but whose actions are insensitive to the rewards. The difference between the green and black curves in Figure 6*A* quantifies the effect of a single reward on the immediate dwell time. Note that the green line is above the black line, indicating that, on the whole, subjects reacted to rewards by prolonging their stay duration. To be more specific as to the fraction of rewards that actually prolonged the animals' dwell times, we used Figure 6*A* to derive a lower bound on the fraction of rewarded stays that were prolonged. Consider time $t = 2.5$ in Figure 6*A* (hyphenated line). We found that 64% of the rewarded trials (green) were longer than 2.5 s, compared with only 38% as expected by chance (black). The difference ($64\% - 38\% = 26\%$) sets a lower bound such that at least 26% of the rewarded stays were prolonged by the reward. This is in contrast with the prediction of the change detection model that only a small fraction of rewards should have an effect on behavior.

Our behavioral model also predicts that a reward will affect the duration of subsequent stay durations. To test this, we considered, as in Figure 6*A*, the stationary sections in which the baiting rates were equal and computed the average stay duration subsequent to rewarded stays. Figure 6*B* (green line) depicts the average stay duration in a target as a function of the number of stays (n) elapsed from a rewarded stay in that target. The black dots depict the average stay duration as a function of the number of stays elapsed from a rewarded stay in the surrogate data of Figure 6*A*. The green and black dots at $n = 0$ are the averages of the distributions of rewarded stays (the distributions described in Fig. 6*A* as a survival plots). The green dot is 29% higher than the black dot, indicating that rewards had a substantial and immediate effect on behavior. The points at $n = 1$ depict the average distribution of stays in a target, taking into account only stays in which the previous stay was rewarded. The fact that the green dot is higher than the black dot indicates that the prolonging effect of a reward persisted to the subsequent visit of the rewarded target. Further analyzing subsequent visits, we found that the effect of a rewarded stay was significant up to $n = 4$ ($p < 0.01$). Note that the black dot at $n = 0$ is substantially higher than the black dots at $n \neq 0$. This reflects the fact that, as discussed above, the longer a stay, the more likely it is that it will be rewarded. The point at $n = -1$ serves as a control; at $n = -1$, the black and green dots overlap, reflecting the fact that the effect of rewards on behavior is causal. Rewards prolong subsequent but not preceding stays.

Decision between multiple targets

Studies with choices between more than two alternatives are rare compared with the rich literature available for choices between two alternatives. Yet, most of the available data are consistent with findings from the two alternatives experiments, with matching (Miller and Loveland, 1974) or generalized matching (Hunter and Davison, 1978; Elsmore and McBride, 1994) as a phenomenological description of aggregate behavior.

The model shown in Figure 1*A* (Eqs. 1 and 2) describes the competition for higher activity between two populations of neurons. This model can be readily generalized to describe behavior in a free operant task, in which the number of targets is >2, simply by adding additional populations of neurons that correspond to the different targets (Eq. 30).

In Figure 7*A*, we consider a network model that consists of three populations of neurons. For an appropriate choice of the external inputs, the network alternates between three states, such that in each state, the activity of one population of neurons is substantially larger than that of the other two (Fig. 7*B*).

To study whether the synaptic plasticity rule (Eqs. 3 and 4) leads to matching behavior in the case of three populations, we simulated the network model in a free operant task in which rewards were delivered according to the concurrent VI schedule, with an overall reward rate identical to that of the behavioral experiment. Each simulated session lasted 120 min and was composed of two sections with fixed baiting rates chosen from ratios of 1:3:9, 1:9:9, 1:1:9, 1:3:3, 1:1:3, or 1:1:1, where the time of change in the baiting rates was as in the behavioral experiment. The parameters of the network were chosen in accordance with the parameters of a corresponding session in the behavioral experiment (see Materials and Methods). Each section of a session with fixed baiting rates is denoted in Figure 7*C* with three dots, depicting the fractional investment as a function of the fractional income for each of the three populations. The different colors denote the different "animals." Similar to the simulations with two targets, all points are aligned, approximately, along the diagonal, consistent with the matching law.

In the case of two targets, we showed analytically that in the model, the product of transitions rates is kept constant (see also Fig. 5*A*). Beyond two targets, the network model is no longer analytically tractable. Therefore, we used the numeric simulations to test for conservation of the product of transition rates in the case of three populations. Similar to Figure 5*A*, we defined the transition rate from a target to be the number of stay durations, divided by the sum of these durations. Each dot in Figure 7*D* corresponds to a single session and depicts the sum of the natural logarithm of the three transition rates after the change in the baiting rates, as a function of this sum before the change. Different animals are depicted with different colors as in Figure 5*A*. We found that the dots approximately align along the diagonal (black line), indicating that the product of transition rates is also approximately conserved, even in the case of three populations.

Discussion

In this paper, we proposed a neural network model for free operant choices that is based on competition between populations of neurons and covariance-based synaptic plasticity. We showed that the model can account for previously reported characteristics of behavior. We used the neural model to deduce a novel behavioral learning algorithm and to predict conservation of the product of transition rates, which is supported by experimental data.

The fundamentals of the model

The details of the simulations presented in this paper depend on the specific choice of parameters. However, the main results reported are invariant to a specific model implementation, provided that certain general principles are followed.

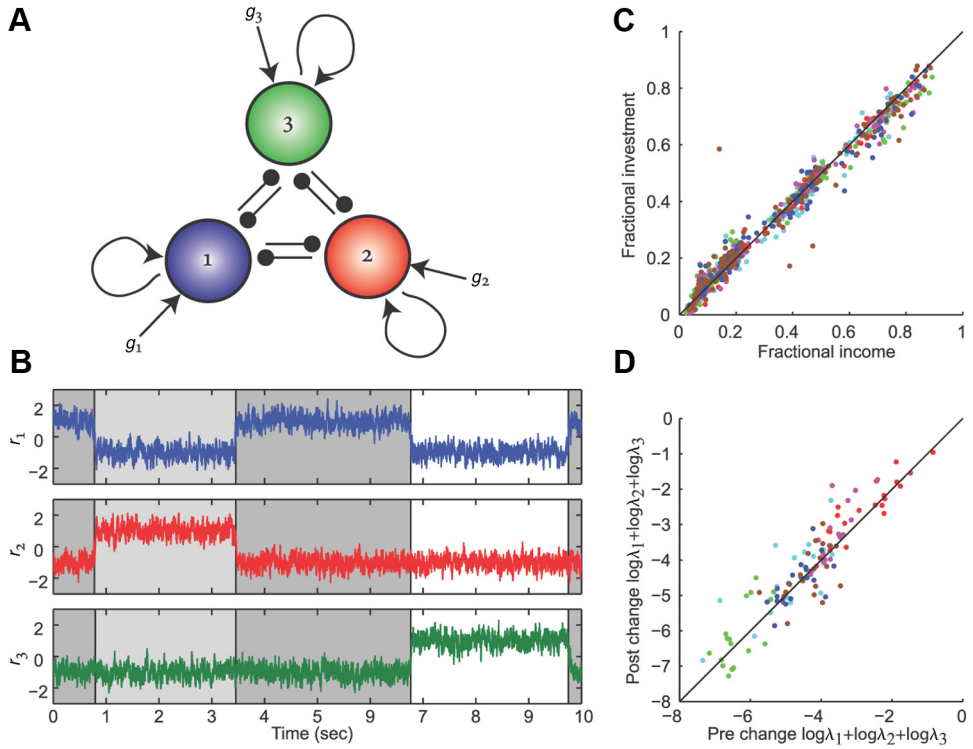

Figure 7. Generalization of the model to a network with 3 populations. ***A***, A schematic description of network architecture, same as in Figure 1*A*. ***B***, The activity of the three neuronal populations during a 10 s simulation of the model, using $g_1 = g_2 = g_3 = -0.65$ and $\sigma = 0.26$. ***C***, Fractional investment as a function of fractional income in 116 simulations. The duration of each simulation, the overall reward rate, and the time of change in the baiting rates were as in the 116 experimental sessions used in Gallistel et al. (2001). The baiting rate ratios used in the simulations were 1:3:9, 1:9:9, 1:1:9, 1:3:3, 1:1:3, and 1:1:1. Because each target could be assigned with any of the baiting rates dictated by these ratios, there were a total of 19 possible schedules. The baiting rate ratios for the first section were chosen uniformly at random from the possible 19 schedules. The baiting rate ratios for the second section were also chosen uniformly at random, with the constraint that the second schedule was not identical to, or a permutation of, the ratios in the first schedule. Each dot corresponds to one stationary section in which the baiting rates were kept constant. Colors represent the different "subjects." The diagonal solid line indicates the behavior predicted from the matching law. ***D***, The sum of the logarithm of the transition rates after the change in the baiting rates as a function of that sum before the change, for the simulations of the three populations' network. Each dot corresponds to a single session. The different color codes for the different "subjects." The diagonal black line indicates the conservation of the product of transition rates.

The approximately exponential distribution of stay duration is the outcome of noise-induced transitions in a multistable dynamic system. If the noise is sufficiently weak, the distribution of stay durations is exponential, independent of the details of the dynamics.

Matching in our model is the outcome of the plasticity rule, which approximates a covariance rule. Therefore, we expect that other approximate covariance rules will also lead to matching behavior. Deviations from a covariance rule (e.g., not subtracting the mean $\bar{r}_i$ in Eq. 3) are expected to result in deviations from matching behavior (Loewenstein, 2008).

The necessary conditions for the conservation of the product of transition rates remain unclear. Our analytical analysis reveals that, for the symmetrical two-population case, the sufficient conditions for this conservation law are that (1) the bistable network dynamics is effectively one-dimensional and (2) transitions are induced by sufficiently weak noise (such that the approximation of Eq. 20 holds). However, the simulations presented in Figure 7*D* indicate that this conservation law may emerge, even if the dynamics of the network are not one-dimensional.

Scope of the model

One aspect of behavior neglected here is the deviations of the distribution of stay duration from an exponential function. As shown in Figure 3, *A* and *B*, the mode (peak) of the distribution of stay durations is not at $t \approx 0$, as predicted by the model but shifted to the right. As noted above, this rise time depends on the baiting ratio. Incorporating this baiting-ratio-dependent rise time is a challenge to the model.

Our model also neglects the effect of motivation on behavior. In our behavioral model, the average visit cycle depends on the preference (Fig. 5*B*) and is independent of the overall rate of rewards delivered to the subject. This reflects the fact that in the experiment the sum of baiting rates was kept constant, resulting in an almost constant rate of rewards delivered to the subject (see also Sugrue et al., 2004; Corrado et al., 2005; Lau and Glimcher, 2008). Adaptation to the overall reward rate can be incorporated in our model by making the parameters that control the product of transition rates, the lateral inhibition, the self-excitation, or the variance of the noise, dynamic variables that depend on the overall rate of rewards.

In addition, our analysis reveals that the estimated learning rate η, increased substantially over the first few sessions of the experiment (data not shown; see Gallistel et al., 2001). Because the sessions we analyzed were preceded by sessions in which the bating rates were kept constant, this result indicates that subjects were influenced by the statistics of past sessions. This metaplasticity is not accounted for in our model.

Decision making in continuous time

The neural basis of decision-making and operant learning has been a subject of intense research in recent years. The experimental session is typically divided into discrete trials, where in each trial the subject chooses between two predefined actions and is

rewarded according to its choices. The discrete-trial design enables the temporal alignment of stimuli and responses to relate neural activity to decision-making variables (Schultz, 1997; Dorris and Glimcher, 2004; Morris et al., 2004; Sugrue et al., 2004; Daw and Doya, 2006; Daw et al., 2006; Padoa-Schioppa and Assad, 2006; Pessiglione et al., 2006; Kable and Glimcher, 2007; Lau and Glimcher, 2008; Niv et al., 2012).

By contrast, there is a long tradition of free operant experiments that are devoid of discrete trials. In these experiments, the subject moves freely back and forth between targets that correspond to different foraging locations. There are important conceptual differences between decisions made in discrete time and in free operant experiments. In particular, in the free operant experiments, subjects continuously choose between two asymmetric alternatives, whether to "stay" in or "leave" the target. The asymmetry in choice manifests in the fact that the probability of "leaving" at any infinitesimal interval of time is infinitesimally small. By contrast, in discrete-time experiments, subjects choose once every trial between actions that are a priori symmetrical and the probabilities of choosing all the predefined actions are typically substantial. Consequently, neural models of decision making in discrete trials (Amari and Arbib, 1977; Wang, 2002; Seung, 2003; Soltani and Wang, 2006; Fusi et al., 2007; Loewenstein, 2010) are not readily applicable to the free operant case.

Switching behavior in free operant experiments bears similarities to switching of perceptual states in response to ambiguous stimuli that have two distinct interpretations. In both cases, behavior is characterized by "spontaneous" transitions that cannot be directly linked to a sensory cue. In both cases, the timings of transitions between the states are described as renewal processes, and the distributions of dominance durations (in perceptual bistability)/stay durations (in free operant experiments) are well approximated by a Γ function (Levelt, 1968; Heyman, 1982; Gibbon, 1995; Tong et al., 1998). Moreover, the time scale of the two processes is similar, with transitions between the states every several seconds. Finally, the overall rate of alternations is largest when the preference for the targets is similar (Fig. 5*B*) (see also Moreno-Bote et al., 2010), yet alternations occur even if one of the alternatives is substantially more dominant than the other.

Motivated by the similarity in behavior, our neural model for decision making resembles previous models of perceptual bistability (Riani and Simonotto, 1994; Salinas, 2003; Moreno-Bote et al., 2007) but with two main differences. First, there is no adaptation in our model, reflecting the fact that the distribution of stay durations in free operant experiments is well approximated by an exponential function. More importantly, our model incorporates reward-dependent synaptic plasticity, reflecting the observed sensitivity of animals' behavior to the delivery of rewards. Yet, in light of the similarities pointed out above, it would be intriguing to test whether the product of dominance durations is conserved in perceptual bistability experiments, similar to the conservation of the product of transition rates in the free operant experiment.

The role of noise

According to our network model, transitions between the two targets are induced by noise. In the absence of noise, the model will remain in one target, even if it is associated with the smaller input. The larger the noise, the larger is the rate of transitions (Fig. 2*B*). This sensitivity of the model to the magnitude of the noise is similar to the sensitivity to noise in models of perceptual bistability. By contrast, models of decision making in discrete trials are less sensitive to the magnitude of noise because of the lack of hysteresis in these models.

In the limit of weak noise, the ratio of transition rates, which determines the extent of compliance with the matching law, is independent of the magnitude of the noise. The magnitude of noise only determines the product of transition rates (Eq. 17) and the learning rate (Eq. 25). This is reminiscent of our previous studies in which we showed that the adaptation to matching in discrete trial experiments is independent of the magnitude of noise (Loewenstein, 2010).

The role of incremental learning in operant conditioning

Incremental learning is a family of learning algorithms, in which small changes are made gradually and iteratively. These algorithms are widely used in artificial intelligence and are relatively easy to implement in biological hardware. However, it has been argued that such learning is too slow to account for operant learning. In particular, the fast adaptation of the rats to a change in the baiting rates tends to be regarded as an indication that cognitively challenging Bayesian inference is required to account for adaptation to matching behavior in free operant experiments. By contrast, we have shown that the fast adaptation of these animals to matching behavior can be accounted for by a simple incremental mechanism in which synapses stochastically decorrelate reward from neural activity. Moreover, consistent with the idea of incremental learning, we showed that rats do make iteratively small changes in their foraging behavior triggered by the harvesting of rewards (Fig. 6).

Clearly, operant learning is likely to be mediated by multiple mechanisms, implemented by different brain modules. These mechanisms range from high-level processes in which complex cognitive reasoning determines individual actions (Baron, 2000), through simpler incremental learning of state-action values (Sutton and Barto, 1998; Doya, 2008; Sakai and Fukai, 2008) to even simpler adaptation (Seung, 2003). In this paper, we demonstrated that incremental adaptation can account for behavior that was previously believed to require more complex reasoning.

Notes

Supplemental material for this article, animations of the network dynamics, is available at http://bio.huji.ac.il/yonatanlab/movies. This material has not been peer reviewed.

References

Amari S, Arbib MA (1977) Competition and cooperation in neural nets. Systems Neurosci 2:72–120.

Baron J (2000) Thinking and deciding. Cambridge, United Kingdom: Cambridge University.

Corrado GS, Sugrue LP, Seung HS, Newsome WT (2005) Linear-nonlinear-Poisson models of primate choice dynamics. J Exp Anal Behav 84:581–617. CrossRef Medline

Davison M, McCarthy D (1988) The matching law: a research review. Hillsdale, NJ: Lawrence Erlbaum.

Daw ND, Doya K (2006) The computational neurobiology of learning and reward. Curr Opin Neurobiol 16:199–204. CrossRef Medline

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876–879. CrossRef Medline

Dayan P, Abbott LF, Abbott L (2001) Theoretical neuroscience: computational and mathematical modeling of neural systems. Cambridge, MA: MIT Press.

Dorris MC, Glimcher PW (2004) Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. Neuron 44:365–378. CrossRef Medline

Doya K (2008) Modulators of decision making. Nat Neurosci 11:410–416. CrossRef Medline

Elsmore TF, McBride SA (1994) An eight-alternative concurrent schedule: foraging in a radial maze. J Exp Anal Behav 61:331–348. CrossRef Medline

Farries MA, Fairhall AL (2007) Reinforcement learning with modulated

spike timing-dependent synaptic plasticity. J Neurophysiol 98:3648–3665. CrossRef Medline

Frémaux N, Sprekeler H, Gerstner W (2010) Functional requirements for reward-modulated spike-timing-dependent plasticity. J Neurosci 30:13326–13337. CrossRef Medline

Fusi S, Asaad WF, Miller EK, Wang XJ (2007) A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple time scales. Neuron 54:319–333. CrossRef Medline

Gallistel CR, Mark TA, King AP, Latham PE (2001) The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect. J Exp Psychol 27:354–372. CrossRef Medline

Gallistel CR, King AP, Gottlieb D, Balci F, Papachristos EB, Szalecki M, Carbone KS (2007) Is matching innate? J Exp Anal Behav 87:161–199. CrossRef Medline

Gibbon J (1995) Dynamics of time matching: arousal makes better seem worse. Psychonom Bull Rev 2:208–215. CrossRef

Herrnstein RJ (1961) Relative and absolute strength of response as a function of frequency of reinforcement. J Exp Anal Behav 4:267–272. CrossRef Medline

Herrnstein RJ, Rachlin H, Laibson DI (2000) The matching law: papers in psychology and economics. Cambridge, MA: Harvard University.

Heskes TM, Kappen B (1993) On-line learning processes in artificial neural networks: math foundations of neural networks, pp 199–233. Amsterdam: Elsevier.

Heyman GM (1982) Is time allocation unconditioned behavior. Quant Anal Behav 2:459–490.

Hunter IW, Davison MC (1978) Response rate and changeover performance on concurrent variable-interval schedules. J Exp Anal Behav 29:535–556. CrossRef Medline

Izhikevich EM (2007) Solving the distal reward problem through linkage of STDP and dopamine signaling. Cereb Cortex 17:2443–2452. CrossRef Medline

Jay TM (2003) Dopamine: a potential substrate for synaptic plasticity and memory mechanisms. Prog Neurobiol 69:375–390. CrossRef Medline

Kable JW, Glimcher PW (2007) The neural correlates of subjective value during intertemporal choice. Nat Neurosci 10:1625–1633. CrossRef Medline

Kempter R, Gerstner W, Van Hemmen JL (1999) Hebbian learning and spiking neurons. Phys Rev 59:4498.

Lau B, Glimcher PW (2008) Value representations in the primate striatum during matching behavior. Neuron 58:451–463. CrossRef Medline

Law CT, Gold JI (2009) Reinforcement learning can account for associative and perceptual learning on a visual-decision task. Nat Neurosci 12:655–663. CrossRef Medline

Legenstein R, Chase SM, Schwartz AB, Maass W (2010) A reward-modulated hebbian learning rule can explain experimentally observed network reorganization in a brain control task. J Neurosci 30:8400–8410. CrossRef Medline

Levelt WJM (1968) On binocular rivalry. Paris: Mouton.

Loewenstein Y (2008) Robustness of learning that is based on covariance-driven synaptic plasticity. PLoS Comp Biol 4:e1000007. CrossRef Medline

Loewenstein Y (2010) Synaptic theory of replicator-like melioration. Front Comput Neurosci 4.

Loewenstein Y, Seung HS (2006) Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity. Proc Natl Acad Sci U S A 103:15224–15229. CrossRef Medline

Mark TA, Gallistel CR (1994) Kinetics of matching. J Exp Psychol Anim Behav Process 20:79–95. CrossRef Medline

Miller HL, Loveland DH (1974) Matching when the number of response alternatives is large. Learn Behav 2:106–110. CrossRef

Moreno-Bote R, Rinzel J, Rubin N (2007) Noise-induced alternations in an attractor network model of perceptual bistability. J Neurophysiol 98:1125–1139. CrossRef Medline

Moreno-Bote R, Shpiro A, Rinzel J, Rubin N (2010) Alternation rate in perceptual bistability is maximal at and symmetric around equi-dominance. J Vis 10(11):1. CrossRef

Moreno-Bote R, Knill DC, Pouget A (2011) Bayesian sampling in visual perception. Proc Natl Acad Sci U S A 108:12491–12496. CrossRef Medline

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. CrossRef Medline

Myerson J, Miezin FM (1980) The kinetics of choice: an operant systems analysis. Psychol Rev 87:160. CrossRef

Niv Y, Edlund JA, Dayan P, O'Doherty JP (2012) Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. J Neurosci 32:551–562. CrossRef Medline

Padoa-Schioppa C, Assad JA (2006) Neurons in the orbitofrontal cortex encode economic value. Nature 441:223–226. CrossRef Medline

Pawlak V, Kerr JN (2008) Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. J Neurosci 28:2435–2446. CrossRef Medline

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 442:1042–1045. CrossRef Medline

Riani M, Simonotto E (1994) Stochastic resonance in the perceptual interpretation of ambiguous figures: a neural network model. Phys Rev Lett 72:3120–3123. CrossRef Medline

Sakai Y, Fukai T (2008) The actor-critic learning is behind the matching law: matching versus optimal behaviors. Neural Comput 20:227–251. CrossRef Medline

Salinas E (2003) Background synaptic activity as a switch between dynamical states in a network. Neural Comput 15:1439–1475. CrossRef Medline

Schultz W (1997) Dopamine neurons and their role in reward mechanisms. Curr Opin Neurobiol 7:191–197. CrossRef Medline

Seung HS (2003) Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. Neuron 40:1063–1073. CrossRef Medline

Soltani A, Wang XJ (2006) A biophysically based neural model of matching law behavior: melioration by stochastic synapses. J Neurosci 26:3731–3744. CrossRef Medline

Sugrue LP, Corrado GS, Newsome WT (2004) Matching behavior and the representation of value in the parietal cortex. Science 304:1782–1787. CrossRef Medline

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, United Kingdom: Cambridge University.

Tong F, Nakayama K, Vaughan JT, Kanwisher N (1998) Binocular rivalry and visual awareness in human extrastriate cortex. Neuron 21:753–759. CrossRef Medline

van Kampen NG (2007) Stochastic processes in physics and chemistry. Amsterdam: North Holland.

Wang XJ (2002) Probabilistic decision making by slow reverberation in cortical circuits. Neuron 36:955–968. CrossRef Medline

Wickens JR (2009) Synaptic plasticity in the basal ganglia. Behav Brain Res 199:119–128. CrossRef Medline

Williams RJ (1992) Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning 8:229–256. CrossRef

Xie X, Seung HS (2004) Learning in neural networks by reinforcement of irregular spiking. Phys Rev 69:041909. CrossRef Medline

Zhang JC, Lau PM, Bi GQ (2009) Gain in sensitivity and loss in temporal contrast of STDP by dopaminergic modulation at hippocampal synapses. Proc Natl Acad Sci U S A 106:13028–13033. CrossRef Medline